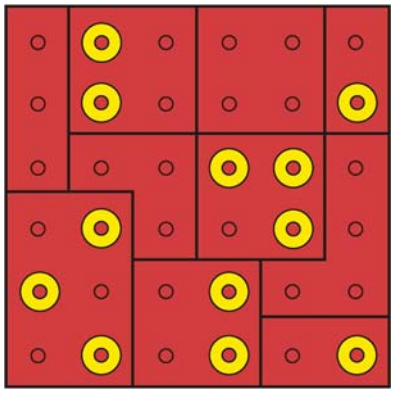
Nº 12-587-X au catalogue

Méthodes et pratiques d'enquête



Statistique Canada Statistics Canada

Canada

Comment obtenir d'autres renseignements

Pour toute demande de renseignements au sujet de ce produit ou sur l'ensemble des données et des services de Statistique Canada, visiter notre site Web à www.statcan.gc.ca. Vous pouvez également communiquer avec nous par courriel à infostats@statcan.gc.ca ou par téléphone entre 8 h 30 et 16 h 30 du lundi au vendredi aux numéros suivants :

Centre de contact national de Statistique Canada

Numéros sans frais (Canada et États-Unis) :

Service de renseignements	1-800-263-1136
Service national d'appareils de télécommunications pour les malentendants	1-800-363-7629
Télécopieur	1-877-287-4369

Appels locaux ou internationaux :

Service de renseignements	1-613-951-8116
Télécopieur	1-613-951-0581

Programme des services de dépôt

Service de renseignements	1-800-635-7943
Télécopieur	1-800-565-7757

Comment accéder à ce produit

Le produit nº 12-587-X au catalogue est disponible gratuitement sous format électronique. Pour obtenir un exemplaire, il suffit de visiter notre site Web à www.statcan.gc.ca et de parcourir par « Ressource clé » > « Publications ».

Normes de service à la clientèle

Statistique Canada s'engage à fournir à ses clients des services rapides, fiables et courtois. À cet égard, notre organisme s'est doté de normes de service à la clientèle que les employés observent. Pour obtenir une copie de ces normes de service, veuillez communiquer avec Statistique Canada au numéro sans frais 1-800-263-1136. Les normes de service sont aussi publiées sur le site www.statcan.gc.ca sous « À propos de nous » > « Notre organisme » > « Offrir des services aux Canadiens ».

Statistique Canada

Méthodes et pratiques d'enquête

Publication autorisée par le ministre responsable de Statistique Canada

Publiée pour la première fois en octobre 2003

Nº 12-587-X au catalogue

ISBN 978-1-100-95206-2

Périodicité : hors série

Ottawa

This publication is also available in English.

Note de reconnaissance

Le succès du système statistique du Canada repose sur un partenariat bien établi entre Statistique Canada et la population, les entreprises, les administrations canadiennes et les autres organismes. Sans cette collaboration et cette bonne volonté, il serait impossible de produire des statistiques précises et actuelles.

Données de catalogage avant publication de la Bibliothèque nationale du Canada

Vedette principale au titre:

Méthodes et pratiques d'enquête

Publié aussi en anglais sous le titre : Survey methods and practices
ISBN 0-660-96826-6
CS12-587-XPF

1. Enquêtes – Méthodologie. 2. Ménages (Statistique) – Enquêtes – Méthodologie.
3. Questionnaires – Design. 4. Échantillonnage (Statistique) – Méthodologie.
I. Statistique Canada. II. Statistique Canada. Division des méthodes d'enquêtes. III. Title.

HA37.C3 S8714 2003 001.4'33
C2003-988001-X

Préface

Je suis très fier de la publication des *Méthodes et Pratiques d'enquête* de Statistique Canada. Ce réel accomplissement couronne les efforts d'un grand nombre d'employés de Statistique Canada, en particulier des divisions de méthodologie d'enquête, auxquels je souhaite exprimer ma gratitude.

Cette publication a profité de cours donnés aux employés de Statistique Canada, d'ateliers offerts à nos clients, et de cours sur les recensements et sondages présentés aux statisticiens d'Afrique et d'Amérique latine. Le *Cours de base sur les enquêtes*, unique et innovateur, déjà offert à plus de 80 reprises à quelque 2000 employés de Statistique Canada et à des employés d'autres agences statistiques nationales, a été une influence notable sur cette publication. Finalement, la réalisation du *Survey Skills Development Manual* pour le compte du Bureau national de la statistique de Chine sous les auspices du Programme de coopération statistique Canada - Chine *(Canada – China Statistical Co-operation Program*) a donné une impulsion particulière à ce projet.

Cette publication servira de support au *Cours de base sur les enquêtes* et je crois qu'elle deviendra une lecture obligée et une référence pour tous les employés de Statistique Canada associés de près ou de loin à une enquête. Je souhaite qu'elle soit aussi utile aux statisticiens d'autres agences nationales et aux étudiants de cours sur la méthodologie d'enquête qui y trouveront un aperçu de la pratique.

Ottawa
Octobre 2003

Dr. Ivan P. Fellegi
Statisticien en chef du Canada

Avant-propos

Ce manuel est avant tout un guide pratique pour la planification, la conception, et la réalisation d'enquêtes. Il aborde les nombreux concepts d'enquête et de nombreuses méthodes élémentaires qui peuvent être utilisées à profit lors de la conception et la réalisation d'une enquête. Ce manuel ne remplace cependant pas le jugement éclairé et l'expertise; il vise plutôt à y contribuer en donnant un aperçu de ce qui est nécessaire à la conception d'enquêtes efficientes et de grande qualité, et de la façon d'utiliser les données d'enquête de façon efficace et pertinente pour l'analyse.

Ce manuel prend sa source dans le Programme de coopération statistique Canada – Chine, financé par l'Agence canadienne de coopération internationale. La manuel qui avait été préparé pour ce programme en vue de contribuer au programme national de formation statistique du Bureau national de la Statistique de Chine. Une *Étude de cas* accompagnait le manuel, en en illustrant les principaux points à l'aide d'une enquête fictive. Ces deux documents ont été revus et modifiés afin de mieux répondre aux besoins de Statistique Canada, particulièrement comme outil de référence pour son *Cours de base sur les enquêtes*.

Bien que ce manuel se concentre sur les aspects fondamentaux des enquêtes utiles à tous les lecteurs, certains chapitres sont plus techniques. Le généraliste pourra étudier ces chapitres en passant outre les points techniques soulignés ci-dessous.

Les cinq premiers chapitres couvrent les aspects généraux du plan d'enquête, notamment :
- une introduction aux concepts de l'enquête et à ses étapes (Chapitre 1),
- la formulation des objectifs d'une enquête (Chapitre 2),
- des considérations générales sur le plan d'enquête (Chapitre 3), par exemple,
    - le choix entre une enquête - échantillon et un recensement,
    - la méthode de définition de la population qui sera observée,
    - les divers genres de base de sondage,
    - les sources d'erreurs dans une enquête,
- les méthodes de collecte des données de l'enquête (Chapitre 4), par exemple,
    - l'autodénombrement, l'interview sur place ou l'interview téléphonique,
    - les questionnaires sur support papier ou électronique,

et
- la conception d'un questionnaire (Chapitre 5).

Les Chapitres 6, 7 et 8 couvrent les points plus techniques du plan de l'enquête - échantillon :
- comment choisir un échantillon (Chapitre 6),
- comment estimer les caractéristiques de la population (Chapitre 7),
- comment déterminer la taille de l'échantillon et répartir l'échantillon entre les strates (Chapitre 8).

Au Chapitre 7, la matière technique plus approfondie commence à la Section 7.*3 Estimation de l'erreur d'échantillonnage des estimations de l'enquête*. Au chapitre 8, la formule utilisée pour déterminer la taille de l'échantillon fait appel à une compréhension plus technique et elle commence à la Section *8.1.3 Formule de calcul de la taille de l'échantillon*.

Le Chapitre 9 couvre les principales opérations de collecte des données et précise comment organiser les opérations de collecte.

Le Chapitre 10 traite de la transformation des réponses à un questionnaire d'enquête en un fichier complet de données d'enquête. La matière technique plus approfondie commence à la Section *10.4.1 Méthodes d'imputation*.

Le Chapitre 11 porte sur l'analyse des données. La matière technique plus approfondie commence à la Section *11.4 Vérification des hypothèses au sujet d'une population : variables continues.*

Le Chapitre 12 traite la diffusion des données aux utilisateurs et le contrôle de la divulgation de données individuelles ou d'un groupe d'individus.

Le Chapitre 13 traite de questions pertinentes à la planification et à la gestion d'une enquête. Ce chapitre non technique vise les gestionnaires d'enquête éventuels ou ceux qui participent à la planification et à la gestion d'une enquête, ou qui s'intéressent à ces sujets.

Deux annexes sont ajoutées à ces 13 chapitres. L'Annexe A porte sur l'utilisation de données administratives dont la collecte a été faite par des organismes gouvernementaux, des hôpitaux, des écoles, etc., à des fins administratives plutôt que statistiques. L'Annexe B couvre le contrôle qualitatif et l'assurance de la qualité, deux méthodes qui peuvent être appliquées à diverses étapes de l'enquête pour minimiser et vérifier les erreurs.

Remerciements

Nous remercions les nombreux employés de Statistique Canada qui ont collaboré à la préparation de *Méthodes et pratiques d'enquête*, en particulier:

Éditrices : Sarah Franklin et Charlene Walker.

Réviseurs : Jean-René Boudreau, Richard Burgess, David Dolson, Jean Dumais, Allen Gower, Michel Hidiroglou, Claude Julien, Frances Laffey, Pierre Lavallée, Andrew Maw, Jean-Pierre Morin, Walter Mudryk, Christian Nadeau, Steven Rathwell, Georgia Roberts, Linda Standish, Jean-Louis Tambay.

Réviseur de la traduction française: Jean Dumais.

Nous remercions aussi tous ceux qui ont collaboré à la préparation de la version originale du *China Survey Skills Manual* (Manuel des notions élémentaires d'enquête en Chine), et en particulier :

Équipe du projet : Richard Burgess, Jean Dumais, Sarah Franklin, Hew Gough, Charlene Walker.

Comité directeur : Louise Bertrand, David Binder, Geoffrey Hole, John Kovar, Normand Laniel, Jacqueline Ouellette, Béla Prigly, Lee Reid, M.P. Singh.

Rédacteurs (membres de l'équipe du projet et) : Colin Babyak, Rita Green, Christian Houle, Paul Kelly, Frances Laffey, Frank Mayda, Dave Paton, Sander Post, Martin Renaud, Johanne Tremblay.

Réviseurs : Benoît Allard, Mike Bankier, Jean-François Beaumont, Julie Bernier, Louise Bertrand, France Bilocq, Gérard Côté, Johanne Denis, David Dolson, Jack Gambino, Allen Gower, Hank Hofmann, John Kovar, Michel Latouche, Yi Li, Harold Mantel, Mary March, Jean-Pierre Morin, Eric Rancourt, Steven Rathwell, Georgia Roberts, Alvin Satin, Wilma Shastry, Larry Swain, Jean-Louis Tambay.

Mise en page: Nick Budko et Carole Jean-Marie.

Nous remercions aussi le *Statistical Education Centre* (Centre de l'enseignement de la statistique) du *NBS* (Bureau national de la statistique) pour leurs apports et rétroaction, et nous apprécions le travail préliminaire de Jane Burgess, Owen Power, Marc Joncas et Sandrine Prasil.

Finalement, nous souhaitons souligner le travail de Hank Hofmann, Marcel Brochu, Jean Dumais et Terry Evers, l'équipe responsable du développement et du lancement du *Cours de base sur les enquêtes* à l'automne 1990 en anglais et à l'automne 1991 en français.

Des publications et des documents variés de Statistique Canada ont servi à l'élaboration de ce manuel. Voici certains documents importants :

- *L'Échantillonnage, Un guide non mathématique*, par A. Satin et W. Shastry,
- *Statistique Canada, Lignes directrices concernant la qualité,*
- Matériel de cours pour *Enquêtes : du début à la fin (416),*
- Matériel de cours pour *Introduction aux techniques d'échantillonnage (412)*,

- Matériel de cours pour *Cours de base sur les enquêtes (CBE)*.

D’autres documents de Statistique Canada sont énumérés à la fin de chaque chapitre, le cas échéant.

PUBLICATIONS ÉLECTRONIQUES
DISPONIBLES À

www.statcan.gc.ca

Table des matières

1. Introduction aux enquêtes 1
2. Formulation de l'énoncé des objectifs 11
3. Introduction au plan d'enquête 21
4. Méthodes de collecte des données 41
5. Conception du questionnaire 63
6. Plans d'échantillonnage 97
7. Estimation 133
8. Calcul de la taille de l'échantillon et répartition 165
9. Opérations de collecte des données 191
10. Traitement 217
11. Analyse des données de l'enquête 247
12. Diffusion des données 283
13. Planification et gestion de l'enquête 303

Annexe A: Données administratives 329

Annexe B: Contrôle qualitatif et assurance de la qualité 335

Étude de cas 351

Index 415

PUBLICATIONS ÉLECTRONIQUES
DISPONIBLES À

www.statcan.gc.ca

Chapitre 1 - Introduction aux enquêtes

1.0 Introduction

Qu'est-ce qu'une enquête? ***Une enquête est une activité organisée et méthodique de collecte de données sur des caractéristiques d'intérêt d'une partie ou de la totalité des unités d'une population à l'aide de concepts, de méthodes et de procédures bien définis. Elle est suivie d'un exercice de compilation permettant de présenter les données recueillies sous une forme récapitulative utile.*** Une enquête commence habituellement s'il y a un besoin d'information et s'il n'y a pas de données ou si elles sont insuffisantes. C'est parfois l'organisme statistique lui-même qui en a besoin ou un client à l'externe, peut-être un ministère, un organisme gouvernemental ou un organisme privé. L'organisme statistique ou le client veut habituellement étudier les caractéristiques d'une population, assembler une base de données à des fins analytiques ou vérifier une hypothèse.

Une enquête comprend plusieurs étapes liées entre elles, notamment, la définition des objectifs, la sélection d'une base de sondage, le choix du plan d'échantillonnage, la conception du questionnaire, la collecte et le traitement des données, l'analyse et la diffusion des données, et la documentation de l'enquête.

La durée d'une enquête peut être répartie en plusieurs phases. La première est la planification, viennent ensuite les phases de la conception et de l'élaboration puis, celle de la mise en œuvre. En bout de ligne, tout le processus de l'enquête est examiné et évalué.

L'objectif de ce chapitre est de donner un aperçu des activités comprises dans le déroulement d'une enquête statistique, et les détails seront versés aux chapitres suivants et en annexes. Afin d'aider à illustrer les points pertinents à l'enseignement dans ce manuel, le lecteur est invité à lire le manuel de l'étude de cas qui est un cheminement de la planification jusqu'à la conception et à la mise en œuvre d'une enquête statistique fictive.

1.1 Étapes d'une enquête

À première vue peut-être, le déroulement d'une enquête consiste simplement à poser des questions et à compiler les réponses pour obtenir des statistiques. Il faut cependant faire une enquête étape par étape, appliquer des procédures et des formules précises pour que les résultats donnent de l'information exacte et significative. Il faut bien connaître les tâches particulières, leurs liens et leur pertinence pour comprendre le processus complet.

Voici les étapes d'une enquête :
- formulation de l'énoncé des objectifs,
- sélection d'une base de sondage,
- choix d'un plan d'échantillonnage,
- conception du questionnaire,
- collecte des données,
- saisie et codage des données,
- vérification et imputation,
- estimation,
- analyse des données,
- diffusion des données,

- documentation.

Voici maintenant une brève description de chaque étape.

1.1.1 Formulation de l'énoncé des objectifs

La formulation de l'énoncé des objectifs est l'une des plus importantes tâches d'une enquête. Elle établit non seulement les besoins d'information de l'enquête dans l'ensemble, mais aussi les définitions opérationnelles à utiliser, les sujets à considérer en particulier et le plan d'analyse. Cette étape de l'enquête détermine ce qu'elle comprendra ou non, ce que le client a besoin de savoir plutôt que ce qui serait intéressant d'apprendre.

Le **Chapitre 2 - Formulation de l'énoncé des objectifs** explique comment formuler les objectifs et déterminer la matière de l'enquête.

1.1.2 Sélection d'une base de sondage

La base du sondage donne les moyens d'identifier les unités de la population de l'enquête et de communiquer avec elles. La base prend la forme d'une liste, par exemple,

- une liste physique, notamment, un fichier de données, un imprimé d'ordinateur ou un annuaire téléphonique,
- une liste conceptuelle, par exemple une liste de tous les véhicules qui entrent au stationnement d'un centre commercial entre 9 h et 20 h pendant une journée en particulier,
- une liste géographique dont les unités correspondent à des secteurs géographiques et dont les unités composantes sont des ménages, des fermes, des entreprises, etc.

Un organisme statistique peut habituellement utiliser, approfondir ou créer une base de sondage. La base choisie détermine la définition de la population de l'enquête et peut avoir des répercussions sur les méthodes de collecte des données, de sélection et d'estimation de l'échantillon, ainsi que sur le coût de l'enquête et la qualité des résultats. Les bases de sondage sont présentées au **Chapitre 3 - Introduction au plan d'enquête**.

1.1.3 Choix d'un plan d'échantillonnage

Il y a deux genres d'enquête : l'enquête-échantillon et le recensement. ***Au cours d'une enquête-échantillon, la collecte des données est faite pour une partie seulement (habituellement très petite) des unités de la population, mais lors d'un recensement, la collecte des données est faite pour toutes les unités de la population.*** Il y a deux types d'échantillonnage : l'échantillonnage non probabiliste et probabiliste. L'échantillonnage non probabiliste est un moyen rapide, facile et bon marché de sélectionner des unités de la population, mais la méthode de sélection est subjective. Afin de faire des déductions sur la population à partir d'un échantillon non probabiliste, l'analyste des données doit supposer que l'échantillon est représentatif de la population. Cette supposition est souvent risquée à cause de la méthode de sélection subjective. L'échantillonnage probabiliste est plus complexe, demande plus de temps et coûte habituellement plus cher que l'échantillonnage non probabiliste. Étant donné cependant que la sélection des unités de la population est aléatoire et que la probabilité de sélection de chaque unité peut être calculée, des estimations fiables sont possibles, ainsi que des estimations d'erreur d'échantillonnage et des déductions sur la population. L'échantillonnage non probabiliste est

habituellement inapproprié pour un organisme statistique et le présent manuel cible donc l'échantillonnage probabiliste.

Il y a de nombreuses méthodes de sélection d'un échantillon probabiliste. Il faut tenir compte de certains éléments pour choisir le plan d'échantillonnage, notamment, la base de sondage, la variabilité des unités de la population et les coûts de l'enquête sur la population. Le plan d'échantillonnage détermine en partie la taille de l'échantillon qui a des répercussions directes sur les coûts de l'enquête, le temps et le nombre d'intervieweurs nécessaires pour conclure l'enquête et sur d'autres considérations opérationnelles importantes. Il n'y a ni solution magique ni recette parfaite pour déterminer la taille de l'échantillon. Il s'agit plutôt d'essayer de répondre au plus grand nombre de besoins possibles dont l'un des plus importants est la qualité des estimations, tout en tenant compte des contraintes opérationnelles.

Les points forts et les points faibles relatifs de l'enquête-échantillon et du recensement sont expliqués au **Chapitre 3 - Introduction au plan d'enquête**. Les plans d'échantillonnage non probabiliste et probabiliste sont présentés au **Chapitre 6 - Plans d'échantillonnage**. Les lignes directrices formulées pour déterminer la taille nécessaire d'un échantillon sont inscrites au **Chapitre 8 - Calcul de la taille de l'échantillon et répartition.**

1.1.4 Conception du questionnaire

Un questionnaire (ou un formulaire) est un groupe ou une séquence de questions formulées pour obtenir d'un répondant de l'information sur un sujet. Les questionnaires sont au cœur du processus de collecte des données parce qu'ils ont des répercussions importantes sur la qualité des données et une incidence sur l'image de marque que projette l'organisme statistique dans le grand public. Les questionnaires sont sur support papier ou électronique.

La conception d'un questionnaire suscite des interrogations : quelles questions poser, comment les formuler au mieux et comment organiser les questions pour obtenir l'information voulue? Le but est d'obtenir de l'information et, à cette fin, les répondants doivent comprendre les questions et donner facilement les réponses exactes en un format qui convient au traitement ultérieur et à l'analyse des données. Il y a des principes bien établis de conception d'un questionnaire, mais la création d'un bon questionnaire est un art qui demande de l'ingéniosité, de l'expérience et des mises à l'essai. Si les besoins de données ne sont pas transformés correctement en un instrument de collecte des données structuré de qualité élevée, un « bon » échantillon peut donner de « mauvais » résultats.

Ce sujet est approfondi au **Chapitre 5 - Conception du questionnaire**.

1.1.5 Collecte des données

La collecte des données est le processus appliqué pour obtenir l'information nécessaire de chaque unité sélectionnée dans l'enquête. Les méthodes élémentaires de collecte des données sont l'autodénombrement, c'est-à-dire que les répondants remplissent le questionnaire sans l'aide d'un intervieweur, et l'intervention de l'intervieweur (par l'intermédiaire de l'interview téléphonique ou sur place). D'autres méthodes de collectes de données comprennent l'observation directe, la déclaration électronique des données et l'utilisation des données administratives.

La collecte des données peut être faite sur support papier ou électronique. Si une méthode de collecte sur support papier est privilégiée, les réponses sont inscrites dans des questionnaires imprimés. Si on opte plutôt pour une méthode assistée par ordinateur, le questionnaire est affiché à l'écran de l'ordinateur et les

réponses sont entrées directement au clavier. Les méthodes assistées par ordinateur ont un avantage : la ***saisie des données ou transformation des réponses en format lisible par la machine*** est faite pendant la collecte, éliminant ainsi cette activité du traitement après la collecte. Autre avantage : les données non valables ou incohérentes peuvent être identifiées plus rapidement que celles des questionnaires sur support papier.

Les méthodes de collecte des données sont considérées au **Chapitre 4 - Méthodes de collecte des données**. Le recours aux données administratives est examiné en **Annexe A - Données administratives.** Les activités de collecte des données, y compris certaines interventions de l'intervieweur, notamment l'énumération, le repérage et les méthodes d'organisation de la collecte des données, sont précisées au **Chapitre 9 - Opérations de collecte des données**.

1.1.6 Saisie et codage des données

Si les données n'ont pas été collectées au moyen d'une méthode assistée par ordinateur, elles doivent être codées et saisies . ***Le codage est le processus d'affectation d'une valeur numérique aux réponses pour faciliter la saisie et le traitement des données en général.*** Certaines questions sont parfois précodées sur le questionnaire même, mais d'autres sont codées après la collecte pendant le traitement manuel ou automatisé. La saisie et le codage des données sont des activités qui coûtent cher et qui demandent beaucoup de temps, mais elles sont essentielles à la qualité des données parce que les erreurs entrées peuvent avoir des répercussions sur les résultats finals de l'enquête. Il faut donc mettre l'accent sur la prévention des erreurs dès les premières étapes. L'assurance de la qualité et le contrôle qualitatif sont deux méthodes de surveillance et de vérification des erreurs. L'objectif de l'assurance de la qualité est de prévoir et d'empêcher les problèmes, et celui du contrôle qualitatif est de garantir que le nombre d'erreurs est restreint aux limites acceptables.

Le **Chapitre 10 - Traitement** porte sur la saisie et le codage des données. Les questions de qualité sont considérées en **Annexe B - Contrôle qualitatif et assurance de la qualité.**

1.1.7 Vérification et imputation

La vérification est l'application de mesures pour repérer les entrées manquantes, non valables ou incohérentes qui indiquent des enregistrements de données éventuellement erronées. L'objectif de la vérification est de mieux comprendre les processus et les données de l'enquête pour garantir que les données finales de l'enquête sont complètes, convergentes et valables. Les vérifications peuvent être de simples mesures de contrôle manuel qu'appliquent les intervieweurs sur place ou des vérifications complètes exécutées par un programme informatique. L'importance de la vérification faite est un compromis entre l'objectif, c'est-à-dire que tous les enregistrements sont « parfaits », et une somme raisonnable de ressources affectées (temps et argent) pour atteindre cet objectif.

Certaines lacunes de vérification sont comblées à l'aide d'un suivi auprès du répondant ou d'un examen manuel du questionnaire, mais il est à peu près impossible de corriger toutes les erreurs ainsi, et l'imputation est souvent utilisée pour régler les autres cas. ***L'imputation est un processus appliqué pour déterminer et attribuer des valeurs de remplacement, afin de résoudre les problèmes de données manquantes, non valables ou incohérentes.***

L'imputation peut améliorer la qualité des données finales, mais il faut choisir prudemment une méthodologie d'imputation appropriée. Certaines méthodes d'imputation ne protègent pas les liens entre les variables ou peuvent en fait susciter une distorsion des liens sous-jacents des données. Il faut tenir

compte du genre d'enquête, de ses objectifs et des caractéristiques de l'erreur pour choisir la méthode convenable.

Le **Chapitre 10 - Traitement** reprend en détail la vérification et l'imputation.

1.1.8 Estimation

Après la collecte, la saisie, le codage, la vérification et l'imputation des données, l'étape suivante est l'estimation. ***Il s'agit d'un moyen que l'organisme statistique applique pour obtenir des valeurs de la population d'intérêt et tirer des conclusions sur cette population à partir de l'information obtenue d'un échantillon seulement de la population.*** Une estimation peut être un total, une moyenne, un ratio, un pourcentage, etc.

Le fondement de l'estimation dans une enquête-échantillon est la pondération qui indique le nombre moyen d'unités de la population représentée par une unité de l'échantillon. Un total de la population peut être estimé, par exemple, en additionnant les valeurs pondérées des unités de l'échantillon. Le plan de sondage dicte la pondération initiale. Des modifications sont parfois apportées à cette pondération pour compenser, par exemple, pour les unités qui ne répondent pas à l'enquête (c.-à-d. non-réponses totales) ou pour tenir compte de l'information secondaire. Les modifications apportées pour les non-réponses peuvent aussi s'appliquer aux données d'un recensement.

Une enquête-échantillon peut accuser une erreur d'échantillonnage parce qu'une partie seulement de la population est dénombrée et que les unités échantillonnées n'ont pas exactement les mêmes caractéristiques que toutes les unités de la population représentée. Il faudrait toujours ajouter une estimation de l'ampleur de l'erreur d'échantillonnage pour chaque estimation, afin d'indiquer aux utilisateurs la qualité des données.

Le **Chapitre 7 - Estimation** traite de l'estimation des statistiques simples. L'estimation de l'erreur d'échantillonnage est couverte au **Chapitre 7- Estimation** et au **Chapitre 11 - Analyse des données de l'enquête**.

1.1.9 Analyse des données

L'analyse des données comprend le sommaire des données et l'interprétation de leur signification pour obtenir des réponses claires aux questions qui ont motivé l'enquête. L'analyse des données devrait nouer un lien entre les résultats de l'enquête et les questions et problèmes mentionnés dans l'énoncé des objectifs. Il s'agit de l'une des étapes les plus cruciales de l'enquête parce que la qualité de l'analyse peut avoir des répercussions substantielles sur l'utilité de l'enquête dans l'ensemble.

L'analyse des données peut être restreinte aux données de l'enquête ou établir une comparaison entre les estimations de l'enquête et les résultats d'autres enquêtes ou sources de données. Elle consiste souvent à examiner des tableaux, des graphiques et diverses mesures sommaires, par exemple, les moyennes et les répartitions des fréquences pour résumer les données. L'inférence statistique peut servir à vérifier les hypothèses ou étudier les liens entre des caractéristiques, par exemple, à l'aide de tests de régression, d'analyses de l'écart ou du chi au carré.

Le **Chapitre 11 - Analyse des données de l'enquête** reprend ce sujet en détail.

1.1.10 Diffusion des données

La diffusion des données est la distribution des données de l'enquête aux utilisateurs par l'intermédiaire de divers médias, par exemple, un communiqué, une interview radio ou télédiffusée, une réponse téléphonique ou télécopiée à une demande spéciale, la publication d'un document, une microfiche, un média électronique, y compris Internet, ou un fichier de microdonnées sur CD, etc.

La prestation et la présentation des résultats finaux sont très importantes. Les utilisateurs devraient trouver, interpréter, comprendre et utiliser correctement et facilement les résultats de l'enquête. Il faudrait résumer les résultats de l'enquête, indiquer les points forts et les points faibles des données, et mettre en évidence les détails importants dans un rapport écrit qui comprend des tableaux et des graphiques.

Avant de diffuser les données, il faudrait en évaluer la qualité pour aider à considérer et interpréter les résultats et la qualité de l'enquête, et informer les utilisateurs, afin qu'ils jugent par eux-mêmes de l'utilité des données. Cette activité peut aussi donner des renseignements précieux pour améliorer l'enquête (si elle est prévue de nouveau) ou d'autres enquêtes. Cette évaluation et le rapport subséquent devraient comprendre une description de la méthodologie de l'enquête, ainsi que les mesures et les sources d'erreur.

Au volet du processus de diffusion, la loi oblige de nombreux organismes statistiques à protéger la confidentialité de l'information des répondants. ***Le contrôle de la divulgation englobeles mesures appliquées pour protéger les données diffusées, afin d'empêcher toute infraction à la vie privée des répondants.*** Il s'agit, notamment, d'identifier et d'éliminer (ou de modifier) les cases des tableaux qui risquent de révéler de l'information sur une personne. Certaines données doivent habituellement être supprimées ou modifiées. Avant de choisir une méthode de contrôle de la divulgation, il faudrait comparer diverses méthodes, compte tenu de leurs répercussions sur les résultats de l'enquête et du risque de divulgation pour une personne.

De nombreux autres aspects de la diffusion sont couverts au **Chapitre 12 - Diffusion des données**.

1.1.11 Documentation

La documentation donne un dossier de l'enquête et devrait comprendre chaque étape et phase de l'enquête. Elle peut comprendre divers aspects de l'enquête et cibler différents groupes, notamment, la direction, le personnel technique, les concepteurs d'autres enquêtes et les utilisateurs. Un rapport sur la qualité des données, par exemple, donne aux utilisateurs un contexte pour l'utilisation informée des données. Un rapport d'enquête qui comprend, non seulement les décisions prises, mais aussi leurs justifications, donne à la direction et au personnel technique de l'information utile pour l'élaboration et l'application ultérieures d'enquêtes semblables. Au cours de la mise en œuvre, la documentation des procédures à l'intention du personnel aide à garantir un déroulement efficace.

Le **Chapitre 12 - Diffusion des données** précise comment organiser un rapport et donne des lignes directrices sur la rédaction.

1.2 Cycle de vie utile d'une enquête

Les étapes de l'enquête présentées ci-dessus ne sont pas nécessairement séquentielles : certaines se déroulent en parallèle, d'autres, par exemple la vérification, sont réitérées à divers moments pendant le

processus de l'enquête. Chaque étape doit d'abord être planifiée, conçue et élaborée, mise en œuvre ensuite et évaluée en bout de ligne. Les phases de la vie utile d'une enquête sont décrites ci-dessous.

1.2.1 Planification de l'enquête

La planification est la première phase du processus de l'enquête. Il faut cependant sélectionner et appliquer auparavant une structure de planification et de gestion. Une structure habituellement utilisée est l'approche de l'équipe de l'enquête ou du projet, c'est-à-dire qu'une équipe interdisciplinaire est chargée de la planification, de la conception, de la mise en œuvre et de l'évaluation de l'enquête et de ses aboutissants prévus. L'équipe interdisciplinaire est formée de membres qui ont des aptitudes techniques différentes, par exemple, un statisticien, un programmeur, un expert dans le domaine de l'étude, un expert de la collecte des données, etc.

La planification d'une enquête devrait se dérouler par étapes d'exactitude et de détails croissants. À l'étape préliminaire ou de proposition de l'enquête, seules les notions les plus générales des besoins de données du client peuvent être connues. Lorsque la proposition d'enquête a été formulée, il est important de déterminer si une nouvelle enquête est nécessaire, sans oublier les options, les coûts et les priorités du client et de l'organisme statistique. Il est parfois possible d'obtenir, en tout ou en partie, l'information voulue dans les dossiers administratifs d'administrations publiques, d'institutions et d'organismes. Autrement, il peut être possible d'ajouter des questions à un questionnaire d'enquête existant ou de refondre une enquête existante.

S'il est déterminé que les sources de données de rechange ne peuvent répondre aux besoins d'information, l'équipe passe à la formulation d'un énoncé des objectifs et elle approfondit sa compréhension des choix de base de sondage, de la taille générale de l'échantillon, des besoins de précision, des options de collecte des données, de l'échéancier et des coûts. La faisabilité de l'enquête est habituellement déterminée à cette étape.

Lorsque les objectifs de l'enquête sont évidents, chaque membre de l'équipe prépare les plans de la composante pertinente à sa responsabilité dans l'équipe. La planification devient plus complexe au cours de cette étape. Les avantages et inconvénients des méthodologies de rechange devraient être examinés et comparés, compte tenu des points suivants : couverture, mode de collecte des données, fréquence, détails géographiques, fardeau de la réponse, qualité, coût, ressources nécessaires et rapidité d'exécution.

Au cours des étapes ultérieures du processus de l'enquête, les plans sont élaborés, révisés et améliorés, et des aspects plus détaillés sont examinés. Chaque activité et opération exige un certain plan de conception, d'élaboration et d'application. La planification continue pendant tout le processus de l'enquête et des modifications sont apportées au besoin.

Les détails de la planification sont expliqués au **Chapitre 13 - Planification et gestion de l'enquête.**

1.2.2 Conception et élaboration

Après avoir établi un grand cadre méthodologique, il est possible d'accomplir un travail détaillé sur les diverses étapes d'une enquête à la phase intitulée conception et élaboration. L'objectif général de cette phase est de déterminer l'ensemble des méthodes et procédures qui permettront d'établir un équilibre approprié entre les objectifs de qualité et les limites des ressources.

Au cours de cette phase, les essais préliminaires ou les enquêtes pilotes nécessaires sont exécutés pour évaluer, par exemple, si le questionnaire est approprié, si la base de sondage convient, si les procédures opérationnelles sont bien choisies, etc. Tout le matériel sur place (p. ex., manuels d'instruction et de formation des intervieweurs, documents de contrôle des échantillons) est préparé pour l'étape de la collecte des données. Les programmes logiciels pour les questionnaires administrés par ordinateur sont élaborés, modifiés ou mis à l'essai. La touche finale est apportée aux procédures de sélection et d'estimation de l'échantillon pour établir des spécifications. Les spécifications sur le codage, la saisie des données, la vérification et l'imputation sont préparées pour le traitement des données.

Des procédures devraient être conçues pour contrôler et mesurer la qualité à chaque étape de l'enquête par souci d'efficacité (à l'aide de procédures de contrôle qualitatif et d'assurance de la qualité) et pour évaluer la qualité des produits statistiques en bout de ligne.

1.2.3 Mise en œuvre

Après avoir vérifié si tous les systèmes sont en place, l'enquête peut maintenant être lancée. C'est la phase de la mise en œuvre. Les manuels et les formules de contrôle de l'enquête sont imprimés, ainsi que le questionnaire (s'il s'agit d'un questionnaire sur support papier). Les intervieweurs sont formés, l'échantillon est sélectionné, la collecte de l'information est faite, et tout est réalisé comme prévu pendant la phase de l'élaboration. Le traitement des données commence après ces activités. Il comprend la saisie, le codage, la vérification et l'imputation des données. Le résultat est un ensemble de données complet bien structuré qui permet de produire les totalisations nécessaires et d'analyser les résultats de l'enquête. Ces résultats sont ensuite vérifiés aux fins de la confidentialité puis, diffusés. À chaque étape, la qualité des données devrait être mesurée et surveillée à l'aide des méthodes conçues et élaborées au cours de l'étape précédente.

1.2.4 Évaluation de l'enquête

L'évaluation est un processus continu au cours de l'enquête. Chaque étape de l'enquête devrait être évaluée pour déterminer l'efficience, l'efficacité et les coûts, en particulier dans le cas des enquêtes réitérées, afin d'apporter avec le temps des améliorations à sa conception et à la mise en œuvre. Ce processus comprend des examens des méthodes appliquées, ainsi que des évaluations de l'efficacité opérationnelle et de la rentabilité. Ces évaluations sont un test pour déterminer si les pratiques techniques sont convenables. Elles servent aussi à améliorer et orienter l'application de concepts particuliers ou de composantes de la méthodologie et des opérations au cours d'une enquête et d'une enquête à l'autre. Elles soutiennent les activités et fournissent des mesures et des examens des limites de la qualité des données du programme. Chaque étape de l'enquête est aussi évaluée pour donner un aperçu des lacunes ou des problèmes à d'autres étapes de l'enquête. La vérification et l'imputation peuvent donner, par exemple, de l'information sur les problèmes que posent les questionnaires.

Les évaluations d'enquêtes précédentes ou d'enquêtes pilotes sont importantes lors de la planification d'une nouvelle activité statistique : elles peuvent aider à formuler des objectifs d'enquête réalistes, donner une idée de la qualité des données que l'on veut obtenir et de l'information essentielle à la conception de l'enquête et au traitement des données.

1.3 Sommaire

Qu'est-ce qu'une enquête? Toute activité organisée et méthodique de collecte d'information est une enquête. Elle est habituellement motivée par le besoin d'étudier les caractéristiques d'une population, d'implanter une base de données à des fins analytiques ou de vérifier une hypothèse.

Quelles sont les étapes de l'enquête? Une enquête est une procédure beaucoup plus complexe que la simple activité de poser des questions et de compiler les réponses pour produire des statistiques. Il faut franchir de nombreuses étapes et appliquer des méthodes et procédures précises pour que les résultats donnent de l'information exacte. Ces étapes comprennent la formulation des objectifs de l'enquête, le choix de la conception de l'échantillon, la conception du questionnaire, la collecte, le traitement et la totalisation des données puis, la diffusion des résultats.

Comment les étapes sont-elles franchies? L'exécution d'une enquête peut être décrite comme un cycle de vie utile à quatre phases. La première est la planification qui permet d'établir les objectifs de l'enquête, la méthodologie, le budget et l'échéancier des activités. La deuxième est la conception et l'élaboration des étapes de l'enquête. La troisième consiste à franchir les étapes de l'enquête. La qualité est mesurée et surveillée pendant la troisième phase pour garantir que le processus fonctionne comme prévu. En dernier lieu, les étapes de l'enquête sont examinées et évaluées.

Bibliographie

Cochran, W.G. 1977. *Sampling Techniques.* John Wiley and Sons, New York.

Des Raj. 1972. *The Design of Sample Surveys.* McGraw-Hill Series in Probability and Statistics, New York.

Moser C.A. et G. Kalton. 1971. *Survey Methods in Social Investigation*. Heinemann Educational Books Limited, London.

Särndal, C.E., B. Swensson et J. Wretman. 1992. *Model Assisted Survey Sampling*. Springer-Verlag, New York.

Satin, A. et W. Shastry. 1993. *Échantillonnage statistique : un guide non mathématique – Deuxième édition.* Statistique Canada. 12-602F.

Statistique Canada. 1987. *Lignes directrices concernant la qualité.* Deuxième édition.

Statistique Canada. 1998. *Statistique Canada – Lignes directrices concernant la qualité.* Troisième édition. 12-539-X1F.

PUBLICATIONS ÉLECTRONIQUES
DISPONIBLES À

www.statcan.gc.ca

Chapitre 2 - Formulation de l'énoncé des objectifs

2.0 Introduction

La première tâche de la planification d'une enquête est de préciser les objectifs le mieux et le plus clairement possible. Un énoncé clair des objectifs oriente toutes les étapes ultérieures de l'enquête. Ces étapes devraient être planifiées de façon à garantir que les résultats en bout de ligne correspondent aux objectifs originaux.

Supposons que vous prévoyez une enquête sur la pauvreté. Il n'est pas suffisant d'indiquer que l'objectif de l'enquête est d'obtenir, par exemple, de l'information sur les « conditions de logement des pauvres ». Ce genre d'énoncé vague peut être une description globale du thème général de l'enquête, mais en bout de ligne, il faut approfondir en une formulation plus spécifique. Que signifie « conditions de logement »? S'agit-il de l'édifice, de l'âge du bâtiment, de la nécessité de rénover ou de la densité (p. ex., le nombre de personnes par mètre carré)? Que signifie précisément le terme « pauvre »? La pauvreté est-elle mesurée selon les revenus, les dépenses, les dettes, ou les trois?

L'organisme statistique, en consultation avec le client, doit d'abord définir les besoins d'information, les principaux utilisateurs et les principales utilisations des données plus complètement et précisément. En général, quels renseignements sont nécessaires sur les conditions de logement des pauvres? Qui a besoin des données et pourquoi? Supposons que le client qui demande l'enquête soit le conseil municipal. Celui-ci a l'impression que les conditions de logement des pauvres laissent à désirer et prévoit qu'il devra bâtir de nouveaux logements subventionnés. Il voudra peut-être savoir combien de nouvelles résidences seront nécessaires et combien elles coûteront. Il pourrait demander aux pauvres où ils veulent les nouveaux logements. La Ville devra peut-être modifier la subvention, compte tenu de la pauvreté de la famille, et elle aura donc besoin de données sur les divers niveaux de pauvreté.

Il faut ensuite formuler des définitions opérationnelles particulières, y compris une définition de la population cible. Ces définitions indiquent qui (ou quoi) sera observé et ce qui sera mesuré. Dans le cas des « pauvres », la définition peut comprendre toutes les familles dont le revenu brut est inférieur à un certain seuil. Il faut aussi définir les termes « famille » et « revenu ». Il faut préciser la couverture de la population : quel secteur géographique intéresse le client, quels secteurs de la ville? Quelle est la période de référence, la semaine dernière, l'année dernière?

L'organisme statistique doit aussi connaître les sujets particuliers qui seront examinés dans l'enquête. Le client veut-il de l'information sur le revenu par tranche, le genre de logement (p. ex., immeubles d'appartements, maisons individuelles, etc.), l'âge du logement, le nombre de personnes qui y habitent, etc.? À quel point chaque sujet doit-il être détaillé et quelle sera la mise en forme des résultats? Le tout fait habituellement l'objet d'une proposition de tableaux d'analyse. Dans une enquête-échantillon, le niveau de détail possible est fonction de la taille de l'erreur d'échantillonnage dans les estimations, ainsi que des contraintes opérationnelles, notamment, le temps, le budget, le personnel et le matériel disponibles. Ces cibles de qualité et contraintes opérationnelles auront des répercussions profondes sur la portée de l'enquête.

L'organisme statistique, en consultation avec le client, peut réviser plusieurs fois l'énoncé des objectifs pendant la planification, la conception et l'élaboration de l'enquête.

L'objectif de ce chapitre est d'illustrer comment formuler l'énoncé des objectifs.

2.1 Processus d'élaboration de l'énoncé des objectifs par étapes successives

L'élaboration de l'énoncé des objectifs est un processus itératif qui engage l'organisme statistique, le client et les utilisateurs (s'ils ne sont pas le client). Les étapes du processus visent à déterminer :

- les besoins d'information,
- les utilisateurs et les utilisations des données,
- les principaux concepts et les définitions opérationnelles,
- la matière de l'enquête,
- le plan d'analyse.

Considérons l'exemple suivant pour illustrer ces étapes. Le conseil municipal a demandé à la Régie des transports en commun de la région (RTCR) d'appliquer des mesures pour faciliter l'utilisation des transports en commun par les citoyens âgés (c.-à-d. les « personnes âgées »). La RTCR n'a pas d'information à jour sur les besoins ou les habitudes de déplacement des personnes âgées et elle a donc communiqué avec l'organisme statistique pour obtenir de l'aide à la collecte de nouvelles données. Le paragraphe suivant est l'énoncé initial de la RTCR sur la situation :

> *La RTCR considère modifier son service actuel pour faciliter l'utilisation des transports en commun par les personnes âgées. Les changements possibles comprennent, par exemple, l'achat d'autobus spéciaux, la modification des autobus actuels, l'ajout de nouveaux itinéraires ou peut-être des tarifs subventionnés. Avant de procéder à des achats et des modifications qui coûtent cher, la RTCR demande de l'information sur les besoins de transport des personnes âgées pour établir un budget et apporter des améliorations selon leurs besoins.*

2.1.1 Besoins d'information (énoncer le problème)

La première étape est la description en termes génériques des besoins d'information du client. L'organisme statistique devrait commencer par identifier le problème et l'énoncer en termes généraux. Pourquoi l'enquête a-t-elle été suggérée? Quelles sont les questions sous-jacentes et dans quel contexte sont-elles posées?

Dans l'exemple de la RTCR, le conseil municipal lui a demandé « d'appliquer des mesures pour faciliter l'utilisation des transports en commun par les personnes âgées ». Dans l'énoncé initial, la RTCR a interprété cette demande comme un besoin de modifier le service actuel pour « faciliter l'utilisation des transports en commun par les personnes âgées ». Quel est en fait l'objectif qu'il faut considérer directement pour aider la RTCR à atteindre cet objectif?

> La RTCR demande de l'information sur les *besoins de transport des personnes âgées*, et veut savoir si l'on *répond* actuellement à ces besoins et comment.

Les besoins d'information de l'enquête dans l'ensemble sont maintenant identifiés. Il est important de revenir à cet énoncé à chaque étape de l'enquête pour garantir que les objectifs de l'enquête sont atteints.

2.1.2 Utilisateurs et utilisations des données

Les deux questions suivantes se posent : Qui sont les principaux utilisateurs des données? À quoi servira l'information? L'organisme statistique a besoin de savoir qui sont les utilisateurs parce que leur rétroaction est très importante pendant la phase de planification de l'enquête. (Les utilisateurs des données en bout de ligne ne sont pas toujours le client, mais c'est souvent le cas.) Il faut déterminer les

utilisations des données pour préciser davantage les besoins d'information. Cette étape est franchie en consultation avec le client et les utilisateurs des données. Quel genre de questions stratégiques faut-il considérer? L'information de l'enquête servira-t-elle à décrire une situation ou à analyser des relations? Quel genre de décisions peuvent être prises à l'aide des données et quelles peuvent être les conséquences? Il faudrait aussi consulter les répondants éventuels si possible parce qu'ils pourraient mentionner des questions et des préoccupations importantes pour eux et qui pourraient avoir des répercussions sur la matière de l'enquête.

À son avis, la « RTCR demande de l'information sur les besoins de transport des personnes âgées pour établir un budget et apporter des améliorations selon leurs besoins ». L'information peut servir en particulier aux planificateurs des transports de la RTCR aux fins suivantes :
- achat d'autobus spéciaux,
- modification des autobus actuels,
- ajout de nouveaux itinéraires,
- subvention des tarifs.

Les besoins d'information de l'enquête sont maintenant identifiés, ainsi que les utilisateurs et les utilisations des données. Voilà qui est particulièrement important. Supposons, par exemple, que la RTCR prévoit qu'il faudra ajouter de nouveaux itinéraires, elle voudra peut-être demander aux personnes âgées où devraient être aménagés ces itinéraires. Si la RTCR prévoit modifier les autobus actuels, elle voudra peut-être savoir quelles modifications préfèrent les personnes âgées. Si la RTCR considère acheter des autobus spéciaux, elle voudra peut-être savoir de quel genre d'autobus ont besoin les personnes âgées. Si la RTCR compte percevoir des tarifs subventionnés, elle voudra peut-être demander aux personnes âgées quels tarifs elles considèrent raisonnables. Les résultats prévus et les conséquences de ces résultats déterminent donc la matière de l'enquête.

2.1.3 Définitions opérationnelles et des concepts

L'organisme statistique a besoin de définitions précises et claires pour déterminer les données nécessaires, afin d'atteindre les objectifs de l'enquête. Ces définitions peuvent préciser des exclusions, notamment, les personnes sans abri ou qui habitent dans des institutions, etc. Il faudrait utiliser des définitions standard reconnues dans la mesure du possible. Elles faciliteront la communication entre les utilisateurs des données et les répondants et garantiront l'uniformité entre les enquêtes. L'organisme statistique devra peut-être élaborer certaines définitions standard, par exemple, pour le logement, le ménage, la famille, etc.

Il faut poser trois questions pour déterminer les définitions opérationnelles : Quoi ou quoi? Où? et Quand? L'un des premiers concepts à définir est la population cible de l'enquête. ***La population cible est la population dont on veut obtenir de l'information.*** C'est l'ensemble des unités que le client est intéressé à étudier. Selon les caractéristiques et l'objectif de l'enquête, ces unités sont habituellement des personnes, des ménages, des écoles, des hôpitaux, des fermes, des entreprises, etc. Reprenons l'exemple de la RTCR. Il faudrait poser les questions suivantes pour définir la population cible de l'enquête:

i. À qui ou à quoi le client s'intéresse-t-il?

L'organisme statistique doit, dans ce cas, considérer le genre d'unités que comprend la population cible et les caractéristiques qui définissent les unités. Aux fins de l'enquête de la RTCR, il est établi que le client s'intéresse à l'utilisation des transports en commun par les personnes âgées et à leurs besoins. Des définitions explicites de *personnes âgées*, *transport en commun* et *utilisation* sont nécessaires. Supposons que les personnes âgées sont les 65 ans ou plus selon la définition. (Le client doit vérifier auprès de la RTCR quelle est sa définition de personnes âgées pour les transports urbains). Il peut y avoir divers

transports en commun : autobus, train, métro et véhicules pour besoins spéciaux. Supposons que le client s'intéresse seulement aux autobus. Autre question : le client s'intéresse-t-il seulement aux personnes âgées qui utilisent actuellement les autobus ou à toutes les personnes âgées? Le client peut s'intéresser à toutes les personnes âgées.

ii. Quelles sont les unités d'intérêt?

La question cible le lieu géographique des unités (c.-à-d. les personnes âgées). Le client s'intéresse peut-être seulement à l'utilisation des autobus de transport en commun qui se déplacent dans le secteur métropolitain de la ville (selon la définition d'un recensement récent, par exemple, et de nouveau, une définition claire est nécessaire) ou peut-être même au territoire de la RTCR (c.-à-d. le territoire que sert le réseau actuel des itinéraires des autobus de transport en commun). Le client doit donc décider si toutes les personnes âgées font partie de la population cible ou si celle-ci comprend seulement celles qui habitent dans une région en particulier.

iii. Quelle est la période de référence de l'enquête? (Quand?)

Sur quelle période les données portent-elles? (Quand?) La réponse semble être « maintenant » parce que l'énoncé de la RTCR cible les besoins actuels. Voilà qui pourrait signifier en pratique que des questions seront posées aux personnes âgées sur leur utilisation des autobus de transport en commun pendant une période de référence récente (semaine, mois, etc.). Faudrait-il faire enquête auprès des personnes âgées pour plus d'une période ou leur poser des questions sur plusieurs périodes de référence différentes?

Une importante considération sur la période de référence est la saisonnalité. Certaines activités seront liées à une période en particulier de la semaine, du mois ou de l'année. Les conclusions peuvent donc viser une période en particulier, mais elles ne sont pas nécessairement valables pour d'autres périodes. Si la RTCR pose des questions aux personnes âgées dans son questionnaire, par exemple, sur leur utilisation du réseau de transport en commun en semaine, les résultats de l'enquête ne seront peut-être pas valables pour les fins de semaine.

Après la population cible, de nombreux autres concepts doivent être définis. Voici les exemples de trois concepts connexes habituellement utilisés dans les enquêtes auprès des ménages à Statistique Canada :

> ***Un logement est un ensemble de pièces d'habitation structurellement distinctes qui a une entrée privée à l'extérieur de l'édifice ou à partir d'un couloir commun ou d'un escalier à l'intérieur de l'édifice.***
>
> ***Un ménage est une personne ou un groupe de personnes qui habitent un logement. Un ménage peut être une personne qui habite seule, une famille ou plus, un groupe de personnes sans lien de parenté, mais qui habitent le même logement.***
>
> ***Une famille est un groupe de deux personnes ou plus qui habitent le même logement et qui ont des liens de parenté par le sang, le mariage (y compris l'union libre) ou l'adoption. Une personne qui habite seule ou qui n'a de lien avec personne d'autre dans le logement où elle habite est classée comme personne hors famille.***

Le **Chapitre 3 - Introduction au plan d'enquête** donne davantage de détails pour définir la population cible et celle du sondage.

2.1.4 Matière du sondage

Un énoncé des objectifs évident garantit que la matière de l'enquête est appropriée et clairement définie. Après avoir déterminé les besoins d'information dans l'ensemble, les utilisateurs et les utilisations, ainsi que les définitions opérationnelles, l'organisme statistique doit ensuite considérer le genre de sujets en particulier qui seront étudiés dans l'enquête. Il s'agit souvent d'un processus itératif. Le processus de précision de la matière de l'enquête révèle souvent que les besoins d'information et les utilisations sont incomplets, ou même qu'il est impossible de répondre à certains besoins pour des raisons opérationnelles ou à cause des définitions.

Revenons à l'exemple de la RTCR. L'information nécessaire à un échelon raisonnablement général a été identifiée. L'organisme statistique doit maintenant en apprendre davantage à ce sujet.

Le client voudra peut-être aussi déterminer diverses caractéristiques des personnes âgées, notamment :
- l'âge,
- le sexe,
- les incapacités,
- le revenu du ménage,
- le lieu géographique (les personnes âgées habitent-elles surtout dans des secteurs restreints en ville, notamment un foyer de retraite, ou sont-elles réparties sur tout le territoire de la ville?),
- le genre de logement (p. ex., maisons de retraite, appartements, résidences),
- la composition du ménage (avec qui habitent-elles?).

Le client peut avoir besoin de renseignements sur les points suivants pour déterminer les besoins de transport :
- nombre de déplacements la semaine dernière,
- fréquence des déplacements (par heure de la journée, en semaine et en fin de semaine),
- modes de transport utilisés,
- problèmes d'utilisation des autobus de transport en commun,
- nombre de déplacements locaux.

Vouloir de l'information sur les caractéristiques des déplacements peut susciter des questions sur les points suivants :
- raison des déplacements,
- point de départ géographique et la destination des déplacements,
- limites au déplacement,
- aides spéciales ou l'assistance nécessaire,
- nombre de déplacements annulés à cause du manque de transport.

Le client devra peut-être comprendre certains points, pour déterminer si les besoins sont satisfaits ou non, notamment :
- l'accès (combien de personnes âgées ont une automobile, une bicyclette, etc.?),
- l'utilisation des autobus de transport en commun,
- la somme dépensée pour les autobus de transport en commun,
- les moyens d'améliorer le service,
- les moyens d'inciter les personnes âgées à utiliser (ou utiliser plus souvent) les autobus de transport en commun.

Tus les concepts qui ne sont pas déjà définis devront l'être. Que signifie, par exemple, une incapacité? Qu'est-ce qu'un déplacement?

Les sujets à couvrir en particulier déterminent les variables à obtenir, la conception du questionnaire et même le plan d'échantillonnage. Ces points ont aussi des répercussions sur le choix de la méthode de collecte des données, par exemple, faudrait-il retenir les services d'intervieweurs ou non, et quels seront donc les coûts de l'enquête?

L'organisme statistique doit couvrir tous les aspects des besoins d'information, mais si elle veut éviter des frais superflus ou un fardeau de réponse excessif pour la population de l'enquête, il devrait éliminer tous les articles qui ne sont pas directement liés aux objectifs de l'enquête.

Au cours d'une étape ultérieure, cette description de la matière de l'enquête doit être formulée en questions et mise en forme dans un questionnaire. Ce sujet est couvert en détail au **Chapitre 5 - Conception du questionnaire**.

2.1.5 Plan d'analyse (totalisations proposées)

Lorsque tous les articles à mesurer sont identifiés, la tâche suivante consiste à déterminer combien de détails seront nécessaires pour chaque article et la mise en forme des résultats. Quelles mesures, calculs, indices, etc., sont nécessaires? Faut-il obtenir des estimations pour les sous-populations? Le plan détaillé de la méthode d'analyse et la présentation des données est le plan d'analyse, et aux analyses prévues s'ajoute la création nécessaire de totalisations proposées. Un plan d'analyse facilite énormément la conception du questionnaire.

Dans le cas des détails des résultats finaux, par exemple, est-il nécessaire de faire une distinction entre les divers groupes d'âge des personnes âgées? Le client doit-il faire la différence entre les hommes et les femmes, ou entre divers types de transport (autobus, automobile, bicyclette, etc.)? Faut-il utiliser des données nominales ou en continu? Le client a-t-il besoin de savoir, par exemple, le revenu exact d'une personne âgée ou le revenu par tranche est-il suffisant? (Si le client est intéressé à calculer les moyennes, le revenu exact est plus approprié.)

Remarquez que le plan d'analyse peut comprendre le retour et des retouches aux définitions opérationnelles et à la matière de l'enquête. Dans l'exemple de la RTCR, voici certaines possibilités pour le genre de détails des résultats, par ordre croissant de détail :

Revenu du ménage :
- tranches de revenu du ménage (p. ex., moins de 15 000 $, de 15 000 $ à 29 999 $, de 30 000 $ à 49 999 $, etc.),
- revenu total exact du ménage,
- revenu exact de chaque source (traitement ou rémunération, régime de retraite, investissements).

Incapacités :
- une seule question pour déterminer si la personne âgée a une condition physique qui limite sa capacité de déplacement local,
- une seule question sur plusieurs incapacités distinctes,
- une série de questions à poser pour déterminer la présence, les caractéristiques et la gravité de chaque incapacité.

Composition du ménage :
- personnes âgées qui vivent seules – qui ne vivent pas seules,
- nombre de personnes dans les ménages,

- catégories de ménage (personne seule, couple, deux adultes ayant des liens autres que ceux d'un couple, trois adultes ou plus ayant des liens, etc.),
- âge de chaque adulte et sa relation avec la personne de référence pour déterminer la composition exacte du ménage.

Nombre de déplacements la semaine dernière :
- tranches (p. ex., de 0 à 3, de 4 à 6, etc.),
- nombre exact,
- nombre exact par jour et heure du jour.

Fréquence des déplacements :
- pourcentage de déplacements en semaine ou en fin de semaine,
- nombre exact de déplacements chaque jour de la semaine.

Modes de transport utilisé :
- mode de transport utilisé le plus souvent pendant la période de référence (p. ex., la semaine dernière),
- tous les modes de transport utilisés (transport en commun et véhicule privé),
- nombre de déplacements en autobus de transport en commun seulement,
- mode de transport utilisé pour chaque déplacement.

Problèmes d'utilisation des autobus de transport en commun :
- élément qui cause la plus importante difficulté,
- tous les éléments qui causent une difficulté,
- énumération des éléments par ordre de difficulté causée,
- cote de la difficulté que pose chaque élément.

Dans les cas présentés ci-dessus, la première répartition la moins détaillée peut être suffisante, ou elle ne contient pas suffisamment de détails pour répondre aux besoins d'information du client. La dernière répartition la plus détaillée peut donner exactement le bon niveau de détails, ou elle peut être trop détaillée et, en fait, trop difficile à répondre. L'information détaillée donne une plus grande souplesse pour l'analyse et permet la comparaison avec d'autres sources d'information, mais l'organisme statistique devrait toujours essayer de demander l'information suffisamment détaillée pour répondre aux besoins de l'analyse, et sans plus, afin d'éviter un fardeau excessif aux répondants.

Il est bon de préparer un ensemble préliminaire de totalisations proposées et d'autres principaux résultats voulus. Déterminer comment les résultats seront présentés aide à définir non seulement le niveau de détail, mais aussi la portée complète de l'enquête. Sans un plan d'analyse clair, il peut être possible à la fin de l'enquête de produire des centaines de tableaux d'analyse, mais seulement quelques-uns pourraient être directement liés aux objectifs de l'enquête.

Les totalisations proposées devraient préciser chaque variable qui sera présentée dans un tableau et ses catégories. L'objectif de cette étape est de créer et de retenir des spécimens de ces tableaux qui formeront l'analyse. La spécification à ce niveau permet à l'organisme statistique de commencer à formuler la version préliminaire des questions du questionnaire de l'enquête.

Aux fins de l'enquête de la RTCR, par exemple, la population devrait être répartie en deux groupes ou plus (p. ex., pour comparer les personnes âgées ayant une incapacité à celles qui n'en n'ont pas).

Des sommaires d'articles distincts (répartitions des fréquences, moyennes, médianes, etc.) peuvent être produits, notamment,
- le pourcentage de déplacements chaque jour de la semaine (Tableau 1),

- le nombre moyen de déplacements en transport en commun,
- la somme moyenne dépensée pour les transports la semaine dernière,
- le pourcentage de personnes âgées par raison de déplacement la plus fréquente.

Tableau 1 : Déplacements par jour de la semaine

Jour de la semaine	Nombre de déplacements	% du total des déplacements
Dimanche		
Lundi		
Mardi		
Mercredi		
Jeudi		
Vendredi		
Samedi		
Total		

Les totalisations croisées d'intérêt éventuel peuvent comprendre :
- le nombre de déplacements par mode de transport (Tableau 2),
- le nombre d'autobus utilisés par points de départ et d'arrivée,
- la répartition des raisons pour ne pas utiliser le transport en commun par caractéristique de personne (p. ex., personne ayant une incapacité, etc.).

D'autres liens peuvent faire l'objet d'une enquête, notamment :
- la somme moyenne dépensée pour les transports par tranche de revenu,
- le revenu médian des personnes âgées confinées à la maison.

Tableau 2 : Nombre de déplacements par mode de transport

Mode de transport	Nombre de déplacements	% du total des déplacements
Transport en commun		
Autobus		
Métro		
Autre		
Transport privé		
Automobile – camion		
Bicyclette		
Marche		
Autre		
Total		

2.2 Contraintes ayant des répercussions sur l'énoncé des objectifs

De nombreuses exigences et contraintes peuvent avoir des répercussions sur l'énoncé des objectifs de l'enquête. L'une est liée à la qualité des estimations. À quel point les résultats de l'enquête devraient-ils être précis? La question fait référence à l'ampleur de l'erreur d'échantillonnage acceptable pour les variables les plus importantes. Les résultats détaillés et précis exigent souvent de très larges échantillons qui sont parfois au-delà des moyens du client. Celui-ci peut donc décider d'exiger moins de précision ou d'obtenir des données plus agrégées, moins détaillées.

Les éléments qui ont des répercussions sur la précision et donc, sur la taille de l'échantillon comprennent ceux-ci :
- la variabilité de la caractéristique d'intérêt de la population,
- la taille de la population,
- le plan d'échantillonnage et la méthode d'estimation,
- le taux de réponse.

Les contraintes opérationnelles ont aussi des répercussions sur la précision. Ces éléments sont parfois les plus influents :
- Quelle taille d'échantillon le client a-t-il les moyens d'utiliser?
- Combien de temps peut être réservé au travail d'élaboration?
- Combien de temps peut être réservé au déroulement de l'enquête au complet?
- Les résultats sont-ils rapidement nécessaires après la collecte?
- Combien d'intervieweurs sont nécessaires? Combien sont disponibles?
- Combien d'ordinateurs sont disponibles? Combien de membres du personnel de soutien informatique sont disponibles?

La précision est élaborée davantage au **Chapitre 3 - Introduction au plan d'enquête,** au **Chapitre 6 - Plans d'échantillonnage,** au **Chapitre 7 - Estimation** et au **Chapitre 8 - Calcul de la taille de l'échantillon et répartition**.

Voici d'autres éléments qui ont des répercussions sur l'énoncé des objectifs :
- Les variables nécessaires peuvent-elles être mesurées à l'aide des techniques disponibles?
- Faudra-t-il imposer aux répondants un fardeau trop lourd pour obtenir les résultats voulus?
- La vie privée du répondant sera-t-elle compromise à cause du niveau de détail des résultats diffusés?
- L'enquête aura-t-elle des répercussions négatives sur la réputation de l'organisme d'enquête?

Toutes ces considérations sont des points de la planification d'une enquête. Les différents aspects de la gestion d'une enquête sont couverts au **Chapitre 13 - Planification et gestion de l'enquête**.

2.3 Sommaire

S'il n'a pas une idée claire des besoins d'information, l'organisme statistique risque de cibler un problème différent, d'obtenir des résultats incomplets ou hors de propos, et de perdre du temps et des ressources. Les activités de l'enquête pourraient simplement ennuyer ou perturber de nombreux répondants sans donner de renseignements utiles. Les objectifs de l'enquête doivent donc être clairement définis pendant la phase de planification.

Voici un résumé des questions les plus importantes et des points à considérer lors de l'élaboration des besoins d'information et des objectifs de l'enquête :
- Quels sont les besoins d'information de l'enquête dans l'ensemble?
- Qui utilisera les données et comment?
- Quelles définitions serviront à l'enquête?
- Quel genre de sujets en particuliers seront considérés pendant l'enquête?
- Un plan d'analyse a-t-il été préparé avec totalisations proposées?
- À quel point les estimations doivent-elles être précises?
- Quelles sont les contraintes opérationnelles?

La formulation des objectifs de l'enquête peut être peaufinée davantage pendant la conception et l'élaboration du questionnaire en particulier (voir le **Chapitre 5 - Conception du questionnaire)**.

Bibliographie

Brackstone, G.J. 1991. Shaping Statistical Services to Satisfy User Needs. *Statistical Journal of the United Nations*. ECE 8: 243-257.

Brackstone, G.J. 1993. Data Relevance: Keeping Pace with User Needs. *Journal of Official Statistics*. 9: 49-56.

Fink, A. 1995. *The Survey Kit*. Sage Publications, California.

Fowler, F.J. 1984. *Survey Research Methods*. 1. Sage Publications, California.

Kish, L. 1965. *Survey Sampling*. John Wiley and Sons, New York.

Levy, P. et S. Lemeshow. 1991. *Sampling of Populations*. John Wiley and Sons, New York.

Moser C.A. et G. Kalton. 1971. *Survey Methods in Social Investigation*. Heinemann Educational Books Limited, London.

Satin, A. et W. Shastry. 1993. *Échantillonnage statistique : un guide non mathématique – Deuxième édition.* Statistique Canada. 12-602F.

Statistique Canada. 1998. Politique sur les Normes. *Manuel des politiques.* 2.10.

Chapitre 3 - Introduction au plan d'enquête

3.0 Introduction

Lorsque les objectifs de l'enquête sont clairement définis, il faut considérer le plan d'enquête. Voici les questions importantes : faut-il faire une enquête-échantillon ou un recensement? La population qui intéresse le client peut-elle faire l'objet d'une enquête? Quelles peuvent être les principales sources d'erreur dans l'enquête et leurs répercussions sur les résultats?

De nombreux éléments aident à déterminer s'il faut faire une enquête-échantillon ou un recensement, notamment, le budget et les ressources disponibles, la taille de la population et des sous-populations d'intérêt, et l'échéancier des résultats de l'enquête.

La base de sondage définit en bout de ligne la population observée qui peut être différente de celle que cible le client. Avant de choisir une base de sondage en particulier, il faut évaluer la qualité de diverses bases éventuelles pour déterminer en particulier laquelle couvre le mieux à la population cible.

Une enquête peut présenter deux genres d'erreur : l'erreur d'échantillonnage et l'erreur non due à l'échantillonnage. L'erreur d'échantillonnage est possible seulement dans l'enquête-échantillon. L'erreur non due à l'échantillonnage est possible dans l'enquête-échantillon et le recensement, et un certain nombre de raisons peuvent l'expliquer : la base de sondage est incomplète, certains répondants n'ont pas déclaré correctement les données, des données de certains répondants peuvent manquer, etc.

L'objectif de ce chapitre est de présenter ces considérations importantes pour le plan d'enquête. Davantage d'information à propos de la planification d'une enquête-échantillon est donnée au **Chapitre 6 - Plans d'échantillonnage**.

3.1 Recensement et enquête-échantillon

Il y a deux genres d'enquête, l'enquête-échantillon et le recensement. La différence est que le ***recensement cible la collecte de renseignements pour toutes les unités de la population***, mais ***l'enquête-échantillon retient à cette fin une partie seulement (habituellement très petite) des unités de la population***. Dans les deux cas, l'information sert à établir des statistiques pour la population dans l'ensemble et, habituellement, pour des sous-groupes de la population.

La principale raison de préférer l'enquête-échantillon au recensement est que l'enquête - échantillon est souvent un moyen plus économique et rapide d'obtenir de l'information de qualité suffisante pour les besoins du client. Étant donné qu'une enquête-échantillon est une opération à plus petite échelle qu'un recensement, elle est aussi plus facile à contrôler et à surveiller. Dans certains cas cependant, un recensement peut être préférable ou nécessaire. (Pour une définition formelle de la qualité, voir l'**Annexe B - Contrôle qualitatif et assurance de la qualité**).

La liste suivante englobe les éléments les plus importants à considérer avant de choisir un recensement ou une enquête par échantillonnage :

i. Erreurs d'enquête

Il y a deux genres d'erreurs d'enquête, l'erreur d'échantillonnage et l'erreur non due à l'échantillonnage.

L'erreur d'échantillonnage est propre à toute enquête-échantillon. ***Il y a erreur d'échantillonnage lorsqu'on estime une caractéristique en mesurant seulement une partie de la population au lieu de la population au complet.***

L'erreur d'échantillonnage est habituellement mesurée en déterminant dans quelle mesure les estimations de l'échantillon sont différentes l'une de l'autre, compte tenu de tous les échantillons possibles de la même taille et en appliquant la même méthode d'échantillonnage (plan d'échantillonnage). L'ampleur de l'erreur d'échantillonnage peut être limitée par la taille de l'échantillon (elle diminue dans la mesure où augmente la taille de l'échantillon), le plan d'échantillonnage et la méthode d'estimation.

Il n'y a pas d'erreur d'échantillonnage dans un recensement parce que tous les membres de la population sont dénombrés. Les résultats du recensement devraient donc être plus précis, semble-t-il, que ceux de l'enquête-échantillon. Toute enquête peut cependant comporter ***des erreurs non dues à l'échantillonnage, c.-à-d. toutes les erreurs qui ne sont pas liées à l'échantillonnage***, et le recensement, encore plus que l'enquête-échantillon, parce qu'il est possible d'affecter davantage de ressources à l'enquête-échantillon pour réduire les erreurs non dues à l'échantillonnage. Ces erreurs peuvent donner des résultats d'enquête biaisés. Les erreurs de mesure et de traitement sont des exemples d'erreurs non dues à l'échantillonnage.

La Section 3.4 donne des détails sur les sources d'erreur d'enquête, alors que le **Chapitre 7 - Estimation** et le **Chapitre 11 - Analyse des données de l'enquête** abordent la méthode de calcul de l'erreur d'échantillonnage.

ii. Coût

Étant donné que tous les membres de la population font l'objet de l'enquête, le recensement coûte plus cher que l'enquête-échantillon (la collecte des données est l'activité la plus chère de l'enquête). Dans le cas d'une grande population, il est habituellement possible d'obtenir des résultats précis à partir d'échantillons relativement modestes. L'Enquête sur la population active canadienne, par exemple, est faite chaque mois auprès de 130 000 résidents environ. La population canadienne compte approximativement 30 millions de citoyens et la taille de l'échantillon est donc de moins de 0,5 % de la population. Un recensement coûterait considérablement plus cher.

iii. Rapidité d'exécution

Il faut souvent obtenir et traiter les données, puis diffuser les résultats, au cours d'une période relativement brève. Étant donné que le recensement saisit des données pour toute la population, la collecte et le traitement des données d'un recensement demandent considérablement plus de temps que pour une enquête-échantillon.

iv. Taille de la population

Le recensement peut être préférable pour une petite population. En effet, pour faire des estimations ayant une petite erreur d'échantillonnage, il peut être nécessaire de tirer un large échantillon de la population. Dans ce cas et pour des frais supplémentaires minimes, les données peuvent être disponibles pour toute la population, au lieu d'une fraction seulement. Le recensement d'une grande population d'autre part coûte très cher et l'enquête-échantillon est donc habituellement préférable.

Les éléments qui ont des répercussions sur la taille de l'échantillon sont repris au **Chapitre 8 - Calcul de la taille de l'échantillon et répartition.**

v. Estimation pour un petit domaine

Compte tenu du point précédent, le recensement peut être préférable lorsque des estimations d'enquête sont nécessaires pour des secteurs géographiques restreints ou des secteurs ayant une petite population. Une enquête nationale peut être nécessaire, par exemple, pour obtenir des statistiques sur chaque ville au pays. L'enquête-échantillon peut donner des statistiques nationales dont l'erreur d'échantillonnage est minime, mais, compte tenu de la taille de l'échantillon, il peut y avoir trop peu de répondants pour donner des estimations dont l'erreur d'échantillonnage est minime pour toutes les villes. Étant donné que le recensement cible chacun et qu'il n'y a pas d'erreur d'échantillonnage, il peut donner des estimations pour tous les sous-groupes possibles de la population.

Il n'est pas toujours nécessaire de faire le recensement ***ou*** l'enquête-échantillon. Il est parfois possible de combiner les deux. Si vous voulez des estimations sur de petits domaines, par exemple, l'enquête-échantillon peut se dérouler dans les plus grandes villes et le recensement, dans les plus petites.

vi. Prédominance des attributs

Si l'objectif de l'enquête est d'estimer la proportion de la population ayant une certaine caractéristique, et si la caractéristique est commune, une enquête-échantillon devrait être suffisante. Si la caractéristique est rare cependant, le recensement peut être nécessaire. La taille de la sous-population ayant la caractéristique détermine le choix.

Supposons, par exemple, que le client veut déterminer le pourcentage de personnes âgées dans la population et que ce pourcentage, à son avis, est d'environ 15 %. L'enquête-échantillon devrait permettre d'estimer ce pourcentage avec une petite erreur d'échantillonnage. Si les attributs sont plus rares cependant, et s'ils touchent moins de 1 % de la population, le recensement peut être plus approprié. (L'hypothèse est que la base du sondage n'a pu identifier ces personnes auparavant.)

Il est bien entendu possible qu'avant de procéder à l'enquête, absolument personne n'ait de donnée sur la prédominance de l'attribut en question. Il est conseillé dans ce cas de procéder à une étude préliminaire, c.-à-d. une étude de faisabilité ou une enquête pilote.

vii. Besoins spécialisés

Il arrive que l'information voulue par enquête ne peut être demandée directement au répondant ou elle peut être un fardeau pour lui. Une enquête sur la santé, par exemple, peut demander des données sur la tension artérielle, le groupe sanguin et la condition physique des répondants, données qui peuvent être déterminées avec précision par un professionnel de la santé seulement. Si le genre de données visées demande du personnel chevronné, du matériel de mesure qui coûte cher, ou s'il faut imposer un fardeau relativement lourd aux répondants, il peut être impossible de faire un recensement. Dans certains domaines en particulier (contrôle qualitatif d'un processus de fabrication par exemple), le caractère destructif de certains tests peut indiquer que l'enquête-échantillon est la seule option logique.

viii. Autres éléments

Il y a d'autres raisons de faire le recensement. La création d'une base de sondage en est une. De nombreux pays, par exemple, font le recensement quinquennal ou décennal de la population. Les données tirées de ce genre de recensement peuvent servir de base de sondage à une enquête-échantillon ultérieure qui cible la même population.

Obtenir de l'information comparative est une autre raison de faire le recensement. L'information comparative peut être le dénombrement connu de la population, par exemple, le nombre d'hommes et de femmes. L'information peut servir à améliorer les estimations de l'enquête-échantillon (voir le **Chapitre 7 - Estimation)**.

3.2 Population cible et population d'enquête

Au **Chapitre 2 - Formulation de l'énoncé des objectifs,** nous avons expliqué comment formuler les définitions opérationnelles et des concepts. L'un des premiers concepts à définir, y est-il mentionné, est la ***population cible, c.-à-d. la population dont on veut obtenir de l'information***.

Les éléments suivants sont essentiels à la définition de la population cible et aux définitions opérationnelles en général :
- genre d'unités que comprend la population et caractéristiques particulières de ces unités (qui ou quoi?),
- localisation des unités (où?),
- période de référence considérée (quand?).

L'organisme statistique commence avec une population conceptuelle, pour laquelle il n'y a peut-être aucune liste concrète, afin de définir la population cible. La population conceptuelle peut être, par exemple, l'ensemble des agriculteurs. Il faut définir le terme « agriculteur » pour cerner la population cible. Celui qui a un petit jardin dans la cour arrière est-il un agriculteur? Quelle est la distinction entre un agriculteur et un jardinier occasionnel? Qu'en est-il si un exploitant agricole n'a vendu aucun de ses produits? La définition de la population cible peut englober, en bout de ligne, tous les agriculteurs au Canada dont les revenus sont supérieurs à un certain seuil au cours d'une année de référence en particulier.

La population d'enquête est en fait la population que couvre l'enquête. Elle peut être différente de la population cible, mais idéalement, les deux devraient être très semblables. Il est important de souligner que les conclusions tirées des résultats de l'enquête s'appliquent seulement à la population de l'enquête. Voilà pourquoi la population d'enquête devrait être clairement définie dans la documentation de l'enquête.

Diverses raisons peuvent expliquer les différences entre les deux populations. La difficulté et le coût élevé de la collecte des données dans les régions isolées, par exemple, peut motiver la décision d'exclure ces unités de la population d'enquête. De même, les membres de la population cible qui vivent à l'étranger ou qui sont dans des institutions peuvent être exclus de la population d'enquête s'il est trop difficile ou coûteux de les intégrer.

Les exemples suivants illustrent les différences possibles entre la population cible et la population d'enquête.

Exemple 3.1 :

Enquête sur les revenus et les dépenses des ménages

Population cible :	Toute la population résidant au Canada le 30 avril 1997.
Population d'enquête :	La population du Canada au 30 avril 1997, à l'exception de ceux qui habitent dans des institutions ou qui n'ont aucune adresse permanente.

Aux fins de cette enquête, il a été décidé qu'il serait trop difficile de faire enquête auprès des gens sans adresse permanente (les expériences précédentes ont eu peu de succès). De plus, ceux qui habitent en institution peuvent être mentalement ou physiquement incapables de répondre aux questions. Nombre de ces gens peuvent être indisposés à répondre, et même s'ils l'étaient, souvent, les questions posées ne s'appliquent pas à leur situation, et il faudrait donc élaborer des instruments d'enquête modifiés. Il faudrait aussi prévoir des dispositions particulières pour avoir accès à certaines institutions en particulier.

3.3 Base de sondage

Lorsque la définition de la population cible satisfait le client et l'organisme statistique, certains moyens d'accès aux unités de la population sont nécessaires. ***La base de sondage donne les moyens d'identifier les unités de la population d'enquête et de communiquer avec elles***. Cette base de sondage définit en bout de ligne la population d'enquête : si la base de sondage ne comprend pas les numéros de téléphone non publiés, par exemple, ils sont aussi exclus de la population d'enquête.

Exemple 3.2 :

Recensement du secteur de la fabrication

Population cible :	Tous les établissements de fabrication en exploitation au Canada en avril 2002.
Population d'enquête :	Tous les établissements de fabrication où des employés travaillaient au Canada en avril 2002.

Le propriétaire peut exploiter un établissement de fabrication, avec employés ou non. Dans cet exemple, la seule base de sondage disponible s'applique aux établissements qui ont des employés et ceux qui n'en n'ont pas sont donc exclus de la population d'enquête.

(La population cible est souvent redéfinie pour correspondre à la population qui peut en pratique faire l'objet d'une enquête. Voilà l'approche dorénavant appliquée dans ce manuel : la population cible fait référence à la population que l'enquête prévoit couvrir, compte tenu des contraintes opérationnelles et pratiques et de la base de sondage utilisée.)

Une base de sondage est nécessaire, non seulement comme véhicule d'accès aux unités de la population d'enquête, mais aussi parce que dans certaines enquêtes, l'organisme statistique doit être en mesure de calculer la ***probabilité d'inclusion*** que présente une unité de la population dans l'échantillon. Si on a recours à l'échantillonnage probabiliste, ces probabilités permettent de tirer des conclusions sur la population observée, et c'est l'objectif de l'enquête. (Consulter le **Chapitre 6 - Plans d'échantillonnage** pour obtenir une définition de l'échantillonnage probabiliste.)

On a déjà fait référence aux unités de l'enquête dont on peut distinguer trois types :
- l'*unité d'échantillonnage* (l'unité qui fait l'objet de l'échantillonnage),
- l'*unité de référence* (l'unité sur laquelle l'information est fournie),
- l'*unité déclarante* (l'unité qui donne l'information).

Dans certaines enquêtes, ces unités sont toutes les mêmes, mais il en est souvent autrement. Dans le cas d'une enquête auprès des enfants, par exemple, il n'est peut-être pas pratique que l'unité de référence, un enfant, soit l'unité déclarante. Un plan d'échantillonnage commun pour les enquêtes auprès des ménages est le recours à une base de sondage qui énumère les ménages dans la population de l'enquête (une telle

base peut donner la meilleure couverture de tous les enfants de la population cible). Dans une enquête qui applique ce genre de base de sondage, on procéderait à l'échantillonnage des ménages et demanderait à un parent de répondre au nom de l'unité de l'analyse, c'est-à-dire l'enfant.

La base de sondage devrait comprendre les renseignements suivants, en tout ou en partie :

i. Données d'identification

Des données d'identification sont les renseignements de la base de sondage qui identifient sans ambiguïté chaque unité de l'échantillon, par exemple, le nom, l'adresse exacte et un numéro d'identification unique.

ii. Données de communication

Les données de communication sont les renseignements nécessaires pour situer les unités de l'échantillon pendant la collecte, par exemple, l'adresse postale ou le numéro de téléphone.

iii. Données de classification

Les données de classification servent à la sélection de l'échantillon et, éventuellement, à l'estimation. Si les gens qui habitent dans des appartements, par exemple, font l'objet d'une enquête différente de ceux qui habitent dans des résidences, la base de sondage doit donc classer différents types de logement (c.-à-d. appartements, maisons individuelles, etc.). Les données de classification peuvent aussi comprendre une mesure de la taille à utiliser pour l'échantillonnage, par exemple, le nombre d'employés qui travaillent dans une entreprise ou le nombre d'acres d'une ferme. Voici d'autres exemples de données de classification : classification géographique (p. ex., province, division ou subdivision du recensement), classification type des professions (CTP) ou classification type des industries (p. ex., CTI ou Système de classification des industries de l'Amérique du Nord, SCIAN).

iv. Données de mise à jour

Les données de mise à jour sont nécessaires si l'enquête doit être réitérée, par exemple, dates des ajouts ou des modifications apportées aux données de la base de sondage.

v. Données de couplage

Les données de couplage sont utilisées pour lier les unités de la base de sondage à une source de données plus à jour, par exemple, pour mettre à jour la base de sondage.

La base de sondage est en résumé un ensemble de renseignements qui donnent le moyen d'avoir accès aux unités sélectionnées de la population de l'enquête. Les données d'identification et de communication sont le minimum nécessaire pour faire l'enquête. Les données de classification, de mise à jour et de couplage sont cependant aussi souhaitables. Les données de la base de sondage sont un outil d'échantillonnage, mais nous constaterons aussi dans les chapitres ultérieurs qu'elles peuvent servir à vérifier et imputer des données manquantes ou incohérentes, et à améliorer l'échantillonnage et l'estimation.

Les différents aspects des plans d'échantillonnage sont repris au **Chapitre 6 - Plans d'échantillonnage** et au **Chapitre 7 - Estimation.** Le **Chapitre 10 - Traitement** porte sur la vérification et l'imputation.

3.3.1 Types de base de sondage

Il y a deux principales catégories de base de sondage : les listes et les bases aréolaires. Si aucune base de sondage n'est appropriée, des bases multiples peuvent être utilisées.

3.3.1.1 Liste

Une liste peut être définie comme une liste conceptuelle ou physique de toutes les unités de la population de l'enquête. Une liste conceptuelle est souvent utilisée pour une population qui existe seulement au cours de l'enquête. Un exemple serait la liste de tous les véhicules qui entrent dans le stationnement d'un centre commercial entre 9 h et 20 h pendant une journée en particulier.

Il est possible d'obtenir des listes physiques, ou listes réelles des unités de la population, de différentes sources. Divers organismes et paliers de l'administration publique maintiennent des listes à des fins administratives. Ces données administratives sont souvent les sources les plus efficientes de données de mise à jour de la base de sondage. Voici des exemples de liste :

- registre des statistiques de l'état civil (p. ex., une liste de toutes les naissances ou de tous les décès dans la population, ou les deux),
- registre des entreprises (p. ex., une liste de toutes les entreprises en exploitation),
- registre des adresses (p. ex., une liste des ménages et des adresses municipales),
- annuaire téléphonique (c.-à-d. une liste de tous les ménages dont le numéro de téléphone est publié),
- listes de clients (c.-à-d. une liste de tous les clients d'une entreprise),
- listes de membres (c.-à-d. une liste de tous les membres d'un organisme).

Il faut tenir compte des éléments suivants lorsqu'on utilise des données administratives pour établir une liste :

i. Coût

Les sources administratives offrent souvent un point de départ bon marché pour établir la base de sondage. Elles sont aussi une source d'information pour la mise à jour de cette base.

ii. Couverture

La source administrative devrait couvrir correctement la population cible.

iii. Mise à jour

Il est important de déterminer à quel point une information administrative est à jour. Il faudrait considérer le temps nécessaire pour traiter les mises à jour et le délai de communication des données à l'organisme statistique parce qu'ils peuvent être des critères décisifs pour déterminer s'il faut utiliser ou non une source administrative en particulier.

iv. Définitions

Les définitions qu'utilise la source administrative devraient correspondre le plus possible aux concepts de l'enquête. La définition d'un logement ou d'une entreprise, par exemple, peut être différente de celle de l'enquête.

v. Qualité

La qualité des données que fournit la source administrative devrait correspondre à l'ensemble des normes de qualité de l'enquête. (Si les données administratives ont un taux de rejet élevé à la vérification, par exemple, l'organisme statistique peut décider que les données sont de qualité insuffisante. La vérification est couverte en détail au **Chapitre 10 - Traitement.**)

vi. Stabilité de l'information de la source

Lorsque les sources administratives sont utilisées pour établir une base de sondage, l'ensemble des variables que fournit la source devrait être aussi stable que possible dans le temps. Les modifications des concepts, des classifications ou de la matière à la source peuvent causer des problèmes graves de mise à jour de la base de sondage.

vii. Relations officielles et juridiques

Il devrait idéalement y avoir une relation (par exemple, un contrat signé) entre l'organisme statistique et la source de l'information administrative. Voilà qui peut être important pour garantir la confidentialité des données. Il est aussi important d'avoir un dialogue ouvert et de favoriser la collaboration entre les deux partenaires.

viii. Documentation

Les dossiers de données devraient être documentés du point de vue des variables qu'ils contiennent et de leur configuration. C'est particulièrement important si les dossiers sont tenus dans différents secteurs de compétence.

ix. Accessibilité – facilité d'utilisation

L'information est-elle disponible sur support électronique? Comment l'information est-elle organisée? Faut-il combiner différentes listes avant de pouvoir les utiliser?

Pour en savoir davantage à propos de l'utilisation des données administratives, on peut consulter l'**Annexe A - Données administratives**.

3.3.1.2 Base aréolaire

Une base aréolaire est une liste spéciale dont les unités sont des secteurs géographiques. La population observée est située dans ces secteurs géographiques. Les bases aréolaires peuvent servir lorsque l'enquête a un caractère géographique (mesurer les peuplements de la faune, par exemple, en comptant le nombre d'animaux par kilomètre carré) ou lorsqu'une liste appropriée n'est pas disponible, auquel cas la base aréolaire peut être un moyen de créer une liste. Une liste inappropriée est souvent un problème. C'est parce que les populations peuvent changer avec le temps, des unités naissent, meurent, déménagent ou changent de nom, de composition ou de caractère, et n'importe quelle liste peut devenir désuète. Les limites géographiques sont cependant plus stables et il est souvent plus facile de maintenir une base aréolaire.

Les bases aréolaires sont habituellement composées d'une hiérarchie d'unités géographiques. Des unités de base de sondage à un niveau peuvent être subdivisées pour former des unités au niveau suivant. Les grandes régions géographiques comme les provinces peuvent être composées de districts ou de

municipalités qui peuvent aussi être divisés en plus petits secteurs, par exemples, les îlots d'une ville. Dans les plus petits secteurs géographiques échantillonnés, la population peut être listée pour échantillonner les unités de ce secteur.

L'échantillonnage à partir d'une base aréolaire est souvent effectué en plusieurs étapes. Supposons, par exemple, qu'il faut tirer un échantillon des logements d'une ville en particulier pour l'enquête, mais qu'il n'y a pas de liste à jour. Une base aréolaire peut servir à créer une liste à jour des logements, comme suit : à la première étape de l'échantillonnage, des secteurs géographiques sont échantillonnés, par exemple, les îlots d'une ville. Ensuite, pour chaque îlot sélectionné, une liste est établie en énumérant tous les logements des îlots échantillonnés en ville. À la deuxième étape de l'échantillonnage, un échantillon de logements est ensuite sélectionné. Ce genre d'approche a un avantage : elle maintient les coûts de création d'une base de sondage dans des limites raisonnables et elle restreint l'échantillon à un nombre limité de secteurs géographiques, moyen rentable de faire des enquêtes par interview sur place.

Il est important que les unités géographiques à échantillonner dans une base aréolaire soient identifiables uniquement sur une carte et que les intervieweurs puissent repérer facilement les limites. Voilà pourquoi les îlots des villes, les routes principales et les rivières sont souvent utilisés pour délimiter les unités géographiques d'une base aréolaire.

L'examen de l'échantillonnage à partir des bases aréolaires est plus approfondi au **Chapitre 6 - Plans d'échantillonnage**. L'établissement d'une liste pour une base aréolaire est expliqué au **Chapitre 9 - Opérations de collecte des données**.

3.3.1.3 Base de sondage multiple

Une base de sondage multiple est une combinaison de deux bases ou plus (des listes et des bases aréolaires ou deux listes ou plus).

Les bases de sondage multiples sont habituellement utilisées lorsqu'aucune base unique ne peut fournir la couverture nécessaire de la population cible. Pendant l'Enquête sur la santé dans les collectivités canadiennes (ESCC), on utilise la base aréolaire de l'Enquête sur la population active (EPA) et une base de composition aléatoire (CA).

Le principal avantage d'une base multiple est que la couverture de la population cible peut être meilleure. L'un des principaux inconvénients cependant est que la même unité d'échantillonnage peut paraître plusieurs fois dans la base de sondage. Idéalement, une unité devrait paraître une fois seulement dans les bases utilisées pour établir la base de sondage multiple. En pratique toutefois, une unité est souvent entrée dans plus d'une de ces bases. Il y a plusieurs moyens de traiter le chevauchement entre les bases de composantes :

- éliminer le chevauchement pendant la création de la base de sondage,
- résoudre le problème pendant la sélection de l'échantillon (ou sur place),
- corriger le problème à l'étape de l'estimation.

Bankier (1986) approfondit ce sujet. La composition aléatoire est étudiée au **Chapitre 4 - Méthodes de collecte des données**.

3.3.2 Défauts de la base de sondage

Plusieurs défauts de base éventuels sont décrits ci-dessous :

i. Sous-dénombrement

Le sous-dénombrement est le résultat de l'exclusion de la base de sondage de certaines unités qui font partie de la population cible. C'est souvent dû au laps de temps entre la collecte et le traitement des données utilisées pour implanter la base de sondage. Entre le moment où la base est achevée et celui où se déroule l'enquête, certaines unités sont « nées » dans la population. Toute unité qui arrive dans la population cible après l'achèvement de la base de sondage n'a aucune chance d'être sélectionnée pour l'enquête. Il en résulte une sous-estimation de la taille de la population cible et les estimations peuvent être biaisées. Des procédures sont nécessaires pour mesurer l'ampleur du sous-dénombrement et corriger au besoin.

ii. Surdénombrement

Le surdénombrement est le résultat de l'ajout à la base de sondage de certaines unités qui ne font pas partie de la population cible. C'est souvent dû à un laps de temps lors du traitement des données de la base de sondage. Entre le moment où la base est achevée et celui où se déroule l'enquête, certaines unités de la population « meurent » (une unité est morte si elle ne fait plus partie de la population cible). Toute unité qui est dans la base de sondage, y compris ces unités mortes hors du champ de l'enquête, peuvent être sélectionnées pour l'enquête. Si ces unités ne sont pas correctement classées hors du champ de l'enquête dans la base de sondage, la stratégie d'échantillonnage peut être moins efficiente du point de vue statistique et les résultats peuvent être biaisés.

iii. Répétition

Il y a répétition lorsque la même unité paraît plus d'une fois dans la base de sondage. Dans une base d'entreprise, par exemple, la même entreprise peut être énumérée une fois sous sa raison sociale et une fois sous son nom commercial. Voilà un problème fréquent des bases de sondage multiples. La répétition a tendance à donner une surestimation de la taille de la population cible et les estimations peuvent être biaisées. Souvent, les unités en double sont repérées seulement à l'étape de la collecte des données de l'enquête.

iv. Classification erronée

Les erreurs de classification sont des valeurs inexactes attribuées à des variables de la base de sondage. Un homme est inscrit par erreur à la catégorie femme, par exemple, ou une entreprise de détail est classée grossiste. Le résultat peut être un échantillonnage inefficient, ou se traduire par le sous-dénombrement (ou le surdénombrement) parce que si l'échantillon comprend seulement des détaillants, par exemple, ceux qui auront été classés grossistes par erreur seront oubliés. Les erreurs de données d'identification ou de communication peuvent susciter des difficultés de repérage du répondant pendant la collecte.

Le **Chapitre 6 - Plans d'échantillonnage** donne davantage d'information sur l'efficience statistique et les plans d'échantillonnage.

3.3.3 Qualités d'une bonne base de sondage

Quatre critères déterminent la qualité d'une base de sondage :

i. Pertinence

La pertinence devrait être mesurée en déterminant à quel point la base de sondage correspond et permet l'accès à la population cible. Plus elle est différente de la population cible, plus l'écart s'élargit entre la population d'enquête et la population cible. Il faudrait aussi évaluer à quel point elle permet la comparaison des résultats des données entre divers programmes d'enquête. L'utilité de la base de sondage pour d'autres enquêtes qui couvrent la même population cible est aussi une mesure essentielle de sa pertinence.

ii. Précision

Il faudrait évaluer la précision en tenant compte de différentes caractéristiques. Il faudrait d'abord évaluer les erreurs de dénombrement (sous-dénombrement, surdénombrement et répétition). Quelle est l'importance des unités manquantes, hors du champ de l'enquête ou en double dans la base de sondage? Il faudrait ensuite vérifier les erreurs de classification. Les unités sont-elles toutes classées? Si oui, le sont-elles correctement? Il faudrait être très attentif aux données de communication. Sont-elles complètes? Si oui, sont-elles exactes et précises? Les répercussions de la précision des données se manifesteront pendant les étapes de la collecte et du traitement dans l'enquête. La précision des données de la base de sondage a des répercussions profondes sur la qualité des résultats de l'enquête.

iii. Actualité/ Fraîcheur

Il faudrait mesurer l'actualité / la fraîcheur des renseignements en vérifiant à quel point la base est à jour, compte tenu de la période de référence de l'enquête. Si l'information de la base est loin d'être à jour (à cause de la source des données utilisée pour implanter la base de sondage ou de la période nécessaire pour établir la base), il faut alors appliquer certaines mesures pour améliorer la rapidité d'exécution.

iv. Coût

Les coûts peuvent être calculés de différentes façons. Il faudrait d'abord déterminer le total des frais engagés pour obtenir et implanter la base de sondage. Il faudrait ensuite comparer le coût de la base de sondage et le coût total de l'enquête. Il faudrait enfin comparer les frais de mise à jour de la base de sondage au total du budget du programme d'enquête. Les bases de sondage servent souvent à plusieurs enquêtes pour accentuer la rentabilité.

Les caractéristiques souhaitables suivantes s'ajoutent à ces importants critères :

a. Procédures et concepts normalisés

Il faudrait appliquer à l'information entrée dans la base de sondage des définitions, procédures, classifications et concepts normalisés que comprennent le client et l'utilisateur des données. Voilà qui est particulièrement important si ces définitions, procédures, classifications et concepts servent à d'autres enquêtes. La base de sondage devrait aussi permettre une stratification efficiente (du point de vue statistique et des frais de collecte).

b. La base de sondage devrait être facile à mettre à jour à l'aide des sources administratives et de l'enquête.

C'est un moyen de garantir qu'elle est tenue à jour et que la couverture est complète.

c. La base de sondage devrait être facile à utiliser

Les bases de sondage qui répondent à toutes les exigences ci-dessus sont peu nombreuses. Le but est de choisir la base qui répond le mieux à ces critères. Il est important de savoir que la base de sondage a des répercussions directes sur de nombreuses étapes de l'enquête. Elle a, notamment, des répercussions sur la méthode de collecte des données. Si la base de sondage ne donne pas les numéros de téléphone, il ne peut y avoir d'interviews téléphoniques. Elle a aussi des répercussions sur la méthode d'échantillonnage. La qualité de la base de sondage a donc, bien entendu, des répercussions sur les résultats finals de l'enquête.

3.3.4 Conseils et lignes directrices

Voici des conseils et lignes directrices utiles pour choisir et utiliser au mieux la base de sondage :

i. Lorsqu'il faut choisir une base de sondage (si plusieurs sont disponibles), évaluer différentes bases possibles à l'étape de la planification de l'enquête pour déterminer leur pertinence et leur qualité.

ii. Éviter les bases de sondage multiples si possible. Lorsqu'aucune base unique n'est appropriée, cependant, considérer une base multiple.

iii. Utiliser la même base de sondage pour les enquêtes qui ont la même population ou le même sous-ensemble de la population cible. Voilà qui évitera les résultats non convergents entre les enquêtes et qui diminuera les coûts liés à la mise à jour et à l'évaluation de la base de sondage.

iv. Intégrer des procédures pour éliminer les répétitions, mettre à jour les naissances, les décès et les unités hors du champ de l'enquête, ainsi que les modifications apportées à tout autre renseignement de la base de sondage pour améliorer ou maintenir la qualité de la base de sondage.

v. Intégrer les mises à jour de la base de sondage le plus rapidement possible.

vi. Insister sur l'importance de la couverture et appliquez des procédures d'assurance de la qualité efficaces aux activités liées à la base de sondage. Voilà qui aidera à minimiser les erreurs dans cette base.

vii. Surveiller périodiquement la qualité de la couverture de la base de sondage en nouant des liens avec d'autres sources ou en vérifiant l'information pendant la collecte des données.

viii. Déterminer et surveiller la couverture des sources administratives par l'intermédiaire de la communication avec le gestionnaire de la source, en particulier lorsque ces sources sont hors du contrôle de l'enquête.

ix. Ajouter des descriptions de la population cible et de celle de l'enquête, de la base de sondage et de la couverture dans la documentation de l'enquête.

x. Procéder à des vérifications cartographiques pour les bases aréolaires à l'aide de vérifications sur place ou d'autres sources cartographiques pour obtenir une délimitation claire et sans chevauchement des secteurs géographiques utilisés dans le plan d'échantillonnage.

3.4 Erreurs d'enquête

Dans un monde parfait, il serait possible de sélectionner un échantillon parfait, de concevoir un questionnaire parfait, d'avoir des intervieweurs parfaits qui obtiendraient de l'information parfaite de répondants parfaits. Il n'y aurait donc pas d'erreurs de saisie de l'information ou de conversion en une mise en forme traitable par ordinateur.

Évidemment, le monde n'est pas parfait et même l'enquête la plus simple pose des problèmes. S'ils ne sont pas prévus et contrôlés, ces problèmes peuvent intégrer de telles erreurs, que les résultats de l'enquête seront inutiles. Il faut donc faire tous les efforts possibles au cours des phases de planification, de conception et d'élaboration de l'enquête pour prévoir les erreurs d'enquête et appliquer les mesures nécessaires pour les éviter. Au cours de la phase de mise en œuvre, il faudrait utiliser des techniques de contrôle qualitatif pour cerner et minimiser les répercussions des erreurs d'enquête. On peut consulter à cette fin l'**Annexe B - Contrôle qualitatif et assurance de la qualité.**

Diverses sources expliquent les erreurs d'enquête. Elles peuvent être classées en deux principales catégories : *erreur d'échantillonnage* et *erreur non due à l'échantillonnage.*

3.4.1 Erreur d'échantillonnage

L'erreur d'échantillonnage a déjà été définie. Elle est le résultat de l'estimation d'une caractéristique de la population en mesurant une partie au lieu de toute la population. Étant donné que toute enquête-échantillon peut comprendre une erreur d'échantillonnage, l'organisme statistique doit donner une certaine indication de la portée de l'erreur aux utilisateurs éventuels des données de l'enquête. Des méthodes de calcul de l'erreur d'échantillonnage s'appliquent à l'enquête-échantillon probabiliste. Ces méthodes découlent directement du plan d'échantillonnage et de la méthode d'estimation de l'enquête.

La mesure appliquée le plus souvent pour quantifier l'erreur d'échantillonnage est la variance d'échantillonnage. ***La variance d'échantillonnage détermine à quel point l'estimation d'une caractéristique de divers échantillons possibles de même taille et de même conception est différente l'une de l'autre.*** Dans le cas des plans d'échantillonnage qui utilisent l'échantillonnage probabiliste, l'ampleur de la variance d'échantillonnage d'une estimation peut être déterminée en tenant compte des différences de la caractéristique observées entre les unités de l'échantillon (c.-à-d. compte tenu des différences observées dans l'échantillon obtenu). La variance d'échantillonnage estimée est donc fonction de l'échantillon sélectionné et varie d'un échantillon à l'autre. Le point principal est l'ampleur de la variance d'échantillonnage estimée d'une estimation relativement à la taille de l'estimation de l'enquête : si la variance est relativement grande, la précision de l'estimation est donc médiocre et n'est pas fiable.

Les éléments qui ont des répercussions sur l'ampleur de la variance d'échantillonnage comprennent :

i. La variabilité de la caractéristique d'intérêt dans la population

Plus la caractéristique dans la population est variable, plus la variance d'échantillonnage est grande.

ii. La taille de la population

En général, la taille de la population a des répercussions sur la variance d'échantillonnage seulement pour les populations de petite taille ou de taille moyenne.

iii. Le plan d'échantillonnage et les méthodes d'estimation

Certains plans d'échantillonnage sont plus efficients que d'autres parce que, pour la même taille d'échantillon et la même méthode d'estimation, un plan peut donner une variance d'échantillonnage moindre que l'autre.

iv. Le taux de réponse

La variance d'échantillonnage augmente dans la mesure où la taille de l'échantillon diminue. Étant donné que les non-répondants diminuent en fait la taille de l'échantillon, les non-réponses augmentent la variance d'échantillonnage. Les non-réponses peuvent aussi biaiser les résultats (voir 3.4.2.3).

Les détails sur les plans d'échantillonnage et l'échantillonnage probabiliste sont couverts au **Chapitre 6 - Plans d'échantillonnage**. La méthode d'estimation de la variance d'échantillonnage, le biais et l'erreur quadratique moyenne sont étudiés au **Chapitre 7 - Estimation,** au **Chapitre 8 - Calcul de la taille de l'échantillon et répartition** et au **Chapitre 11 - Analyse des données de l'enquête.**

3.4.2 Erreurs non dues à l'échantillonnage

Outre l'erreur d'échantillonnage, un large éventail d'erreurs qui ne sont pas liées au processus d'échantillonnage peuvent être repérées dans une enquête. Ces erreurs sont habituellement intitulées erreurs non dues à l'échantillonnage. ***Les erreurs non dues à l'échantillonnage peuvent être définies comme des erreurs possibles pendant à peu près toutes les activités d'enquête, mis à part l'échantillonnage***. Ces erreurs se retrouvent dans l'enquête-échantillon et le recensement (contrairement à l'erreur d'échantillonnage qui est présente seulement dans l'enquête-échantillon). Les erreurs non dues à l'échantillonnage peuvent être réparties en deux groupes :

i. Erreurs aléatoires

Les erreurs aléatoires ont des répercussions qui s'éliminent approximativement si l'échantillon est suffisamment grand, le résultat étant une variabilité accrue.

ii. Erreurs systématiques

Les erreurs systématiques ont tendance à avoir la même orientation, elles s'accumulent donc dans tout l'échantillon et les résultats finaux sont biaisés. Contrairement à la variance d'échantillonnage et aux erreurs aléatoires, ce biais ne diminue pas malgré l'augmentation de la taille de l'échantillon. Les erreurs systématiques sont la principale cause de préoccupation au chapitre de la qualité des données de l'enquête. Malheureusement, les erreurs non dues à l'échantillonnage sont souvent très difficiles et parfois même impossibles à mesurer.

Voici les principales sources d'erreurs non dues à l'échantillonnage :
- couverture,
- mesure,
- non-réponse,

- traitement.

3.4.2.1 Erreur de couverture

Les erreurs de couverture sont des omissions, des ajouts erronés, des répétitions et des erreurs de classification d'unités dans la base de sondage. Elles ont des répercussions sur chaque estimation de l'enquête et sont donc l'un des plus importants types d'erreur. Elles peuvent même être la principale source d'erreurs du recensement. Les erreurs de couverture peuvent susciter des estimations biaisées et les répercussions peuvent varier pour différents sous-groupes de la population. Ces erreurs ont tendance à être systématiques et sont habituellement dues au sous-dénombrement. Voilà pourquoi les organismes statistiques essaient d'en diminuer l'incidence le plus possible.

3.4.2.2 Erreur de mesure

L'erreur de mesure est la différence entre la réponse inscrite à une question et la « vraie » valeur. Le répondant, l'intervieweur, le questionnaire, la méthode de collecte des données et l'outil de mesure peuvent susciter ce genre d'erreur.

L'une des principales causes de l'erreur de mesure est l'incompréhension du répondant ou de l'intervieweur. Voici des sources possibles d'incompréhension :
- recours au jargon technique,
- manque de clarté des concepts (c.-à-d. utilisation de concepts non standard),
- formulation médiocre des questions,
- formation inappropriée de l'intervieweur,
- communication de renseignements erronés (c.-à-d. erreur de mémoire ou manque de sources d'information disponibles),
- problème de langue,
- traduction médiocre (si l'enquête est multilingue).

La méthode de collecte des données peut aussi avoir des répercussions sur l'erreur de mesure. Les méthodes assistées par intervieweurs (recours à des intervieweurs bien formés), par exemple, peuvent donner des erreurs de mesure plus petites que les méthodes d'enquête par autodénombrement qui ne donnent pas d'aide aux répondants pour remplir le questionnaire.

Dans les enquêtes avec mesure directe, les intervieweurs font la collecte des données par observation ou en prenant les mesures (p. ex., enquêtes sur les prix). L'erreur de mesure peut être due à l'intervieweur ou à l'outil de mesure. Lors d'une enquête sur le poids des gens, par exemple, si la balance n'est pas bien calibrée, les poids ne seront pas correctement déterminés.

Les erreurs de mesure éparpillées aléatoirement autour de la vraie valeur auront des répercussions sur la précision des estimations de l'enquête : la précision diminue dans la mesure où augmente la variabilité. Si les erreurs de mesure reflètent systématiquement certaines valeurs ou catégories , un biais se glissera et les estimations de l'enquête seront trompeuses. Il y a erreur systématique, par exemple, si l'intervieweur doit mesurer la taille des enfants à l'école et si les enfants portent des souliers pendant la mesure, auquel cas, toutes les tailles sont systématiquement surestimées.

Les expressions « erreur de mesure » et « erreur de réponse » sont souvent utilisées sans distinction. Les méthodes de collecte des données sont considérées au **Chapitre 4 - Méthodes de collecte des données**. L'erreur de réponse fait l'objet d'un examen détaillé au **Chapitre 5 - Conception du questionnaire**.

3.4.2.3 Erreur due à la non-réponse

Il y a deux genres de non-réponse : la non-réponse partielle (à une ou à quelques questions) et la non-réponse totale. ***Il y a non-réponse partielle lorsque l'information est disponible pour certaines questions seulement***, notamment, parce que le répondant répond à une partie seulement du questionnaire. ***Il y a non-réponse totale en l'absence de toutes les données ou presque d'une unité d'échantillonnage***.

La non-réponse peut causer plusieurs problèmes dans une enquête. Le principal problème est que les non-répondants ont souvent des caractéristiques différentes de celles des répondants, et les estimations de l'enquête seront biaisées si les non-réponses ne sont pas corrigées. Lors d'une enquête sur l'alphabétisation, par exemple, les résultats de l'enquête peuvent être biaisés si la majorité des non-répondants sont analphabètes. Si le taux de non-réponse est élevé, le biais peut être suffisamment marqué pour que les résultats de l'enquête soient inutiles. La non-réponse totale pose un deuxième problème : elle diminue la taille réelle de l'échantillon parce qu'il était prévu que davantage d'unités répondraient à l'enquête. La variance d'échantillonnage augmente donc au détriment de la précision des estimations. S'il est possible de prévoir le taux de réponse, la taille initiale de l'échantillon devrait augmenter pour en tenir compte. L'augmentation de la taille de l'échantillon diminue la variance de l'échantillonnage et permet donc d'apporter une correction pour les non-réponses qui sont réparties au hasard, mais elle ne diminue pas le biais de la non-réponse systématique.

Certaines raisons peuvent expliquer la non-réponse totale : il n'y avait personne à la maison, la personne sélectionnée a refusé ou était incapable de participer à l'enquête. Une explication médiocre de l'objectif de l'enquête ou de son utilisation prévue peut aussi susciter une non-réponse. Des données de base de sondage médiocres ou périmées sont un autre élément : les données d'identification de l'unité de l'enquête peuvent être inappropriées et ne permettent pas de la situer. De plus, une unité en particulier est parfois sélectionnée pour de nombreuses enquêtes différentes ou pour une enquête réitérée et, à la longue, l'unité en vient à refuser de répondre aux enquêtes à cause du fardeau de réponse. Enfin, si les données du répondant sont considérées inutilisables, elles peuvent être traitées comme une non-réponse.

Il peut y avoir non-réponse à une question si le répondant ne connaît pas la réponse, refuse de répondre, oublie de répondre ou adopte un cheminement erroné pendant le questionnaire. Parfois, le répondant ne peut répondre parce qu'il est malade ou parce qu'il éprouve des difficultés à communiquer dans la langue de l'enquête. La conception médiocre du questionnaire peut aussi favoriser la non-réponse à certaines questions. Les concepts présentés au répondant dans le questionnaire ou pendant l'interview peuvent être difficiles à comprendre ou mal définis. L'interview peut se prolonger inutilement ou le débit des questions peut être illogique. Les répondants peuvent donc se décourager et cesser de répondre avant la fin de l'interview ou ils peuvent simplement suivre un cheminement erroné dans le questionnaire.

Les intervieweurs peuvent aussi avoir une incidence sur la non-réponse totale ou partielle. Des techniques d'interview médiocres empêchent certains intervieweurs d'établir une bonne relation avec le répondant qui peut donc refuser de participer ou, s'il le fait, perdre rapidement tout intérêt pour l'enquête. Certains intervieweurs indiquent des erreurs dues à la non-réponse à une question parce qu'ils ne suivent pas les instructions ou ne lisent pas les questions telles qu'elles sont formulées.

Enfin, les méthodes de collecte des données peuvent être une source de non-réponse. Les intervieweurs font souvent le suivi d'une non-réponse pour obtenir certaines réponses (p. ex., renverser un refus). Le suivi inapproprié des non-répondants ou le suivi au mauvais moment peut empêcher de corriger la non-réponse. La perte des données d'un fichier ou d'un questionnaire peut aussi donner des erreurs « dues à la non-réponse ». (Les données perdues, même si le nombre réel de cas est mince, sont une importante source de préoccupations à cause de l'infraction éventuelle à la confidentialité des données du répondant.)

Le **Chapitre 5 - Conception du questionnaire** révélera les détails de la conception du questionnaire. Le traitement de la non-réponse totale est couvert au **Chapitre 7 - Estimation** alors que la non-réponse partielle est traitée au **Chapitre 10 - Traitement**. Les procédures sur le terrain sont précisées au **Chapitre 9 - Opérations de collecte des données.**

3.4.2.4 Erreur de traitement

Le traitement transforme les réponses de l'enquête obtenues pendant la collecte en une mise en forme qui convient à la totalisation et à l'analyse des données. Il comprend toutes les activités de manutention des données après la collecte et avant l'estimation. Il s'agit d'un ensemble d'activités manuelles et automatisées qui demandent beaucoup de temps et de ressources, et ce volet est donc une source éventuelle d'erreurs. Des erreurs de traitement peuvent se produire, par exemple, pendant le codage ou la saisie des données, la vérification ou l'imputation. Elles peuvent être aléatoires comme toutes les autres erreurs et accroître ainsi la variance des estimations de l'enquête, ou elles peuvent être systématiques et ajouter un biais.

Le codage est le processus d'affectation d'une valeur numérique aux réponses pour faciliter la saisie des données et le traitement en général. Le codage comprend l'attribution d'un code (p. ex., le code de profession) à une réponse donnée ou la comparaison de la réponse avec un ensemble de codes et la sélection de celui qui décrit le mieux la réponse.

Dans le cas des questions fermées (questions ayant des catégories de réponses prédéterminées), les codes sont souvent attribués avant l'interview. Quant aux questions ouvertes (le répondant répond dans ses propres mots), le codage peut être manuel ou automatisé. L'intégralité et la qualité de la réponse à une question ouverte, ainsi que la méthode de codage de la réponse, déterminent la qualité du codage. Le codage manuel des questions ouvertes demande l'interprétation et du jugement, et l'erreur est donc possible. Deux codeurs différents peuvent coder la même réponse différemment. Les codeurs de peu d'expérience et de formation médiocre sont particulièrement exposés aux erreurs de codage. Au cours d'une opération de codage automatisé, un problème de programmation peut susciter des erreurs ou il est possible que le programme ne tienne pas compte correctement de toute l'information disponible. Si le codage est programmé et exécuté automatiquement, un problème de programmation sera systématiquement répété et introduira un biais (p. ex., erreur de classification de la profession).

La saisie des données est la mise en forme des réponses lisible à la machine. Il y a erreur de saisie des données si elles ne sont pas entrées à l'ordinateur exactement comme elles paraissent dans le questionnaire. La complexité des données alphanumériques et le manque de clarté des réponses fournies peuvent expliquer ce problème. La présentation physique du questionnaire ou les documents de codage peuvent susciter des erreurs de saisie des données. La méthode de saisie des données peut aussi occasionner des erreurs (la saisie des données peut être une activité manuelle ou automatisée, par exemple, à l'aide d'un lecteur optique de caractères).

La vérification consiste à inscrire des coches pour identifier des entrées manquantes, erronées ou incohérentes qui révèlent l'enregistrement de données éventuellement erronées. L'imputation est un processus qui détermine et attribue des valeurs de remplacement, afin de résoudre les problèmes de données manquantes, erronées ou incohérentes. Les erreurs de vérification et d'imputation sont souvent simultanées parce que les deux processus sont très étroitement liés.

La structure complexe ou la qualité médiocre des données originales peut expliquer les erreurs de vérification et d'imputation. Lorsque les processus de vérification et d'imputation sont automatisés, les défaillances des programmes insuffisamment mis à l'essai peuvent aussi expliquer les erreurs. Le choix

d'une méthode d'imputation inappropriée peut susciter des biais. La modification inexacte des données considérées erronées ou la modification erronée de données exactes peuvent aussi expliquer les erreurs.

Les activités de traitement sont expliquées en détail au **Chapitre 10 - Traitement**. Les erreurs de traitement sont souvent surveillées et contrôlées à l'aide de techniques de contrôle qualitatif. L'**Annexe B - Contrôle qualitatif et assurance de la qualité** donne davantage de détails.

3.5 Sommaire

Ce chapitre a présenté certains points importants à considérer lors de la planification de l'enquête. La distinction entre une enquête et un recensement a été expliquée, ainsi que les avantages et les inconvénients de chacun. La différence entre la population cible et la population de l'enquête a ensuite été précisée. Une section sur les bases de sondage a exposé les divers types de bases qui peuvent être utilisées dans une enquête, les qualités d'une bonne base, ainsi que les défauts souvent manifestes et des moyens de les éliminer. Enfin, les différents types et les sources diverses d'erreurs dans une enquête ont été considérés. L'erreur d'échantillonnage a été brièvement définie (la question sera approfondie dans les chapitres ultérieurs) et l'accent a été mis sur les erreurs non dues à l'échantillonnage : erreurs de couverture, de mesure ou de traitement et erreur due à la non-réponse.

Comment planifier et gérer une enquête en général, quelles sont les étapes de la planification de l'enquête? C'est le sujet du **Chapitre 13 - Planification et gestion de l'enquête**. Quelle est la conception qui permettra de déterminer comment sélectionner l'échantillon de la population ciblée? Le **Chapitre 6 - Plans d'échantillonnage** répond à cette question**.**

Bibliographie

Bankier, M. 1986. Estimators Based on Several Stratified Samples with Applications to Multiple Frame Surveys. *Journal of the American Statistical Association.* 81-396.

Biemer, P.P., Groves, R.M., Lyberg, L.E., Mathiowetz, N.A. et S. Sudman, Éds. 1991. *Measurement Errors in Surveys*. John Wiley and Sons, New York.

Cialdini, R., M. Couper et R.M. Groves. 1992. Understanding the Decision to Participate in a Survey. *Public Opinion Quarterly*. 56: 475-495.

Cox, B.G., D.A. Binder, B.N. Chinnappa, A. Christianson, M.J. Colledge, P.S. Kott, Éds. 1995. *Business Survey Methods.* John Wiley and Sons, New York.

Food and Agriculture Organization of the United Nations (FAO). 1996. *Multiple Frame Agriculture Surveys. Volume 1: Current Surveys Based on Area and List Sampling Methods*. FAO, Rome.

Fuller, W. 1987. *Measurement Error Models*. John Wiley and Sons, New York.

Gosselin, J.-F., B.N. Chinnappa, P.D. Ghangurde et J. Tourigny. 1978. Coverage. *A Compendium of Methods of Error Evaluation in Censuses and Surveys*. Statistics Canada. 13-546E: 7-9.

Groves, R.M. 1989. *Survey Errors and Survey Costs*. John Wiley and Sons, New York.

Hartley, H.O. 1962. Multiple Frame Surveys. *Proceedings of the Social Statistics Section.* American Statistical Association. 203-206.

Laniel, N. et H. Finlay. 1991. Data Quality Concerns with Sub-Annual Business Survey Frames. *Proceedings of the Section on Survey Research Methods.* American Statistical Association. 202-207.

Lessler, J.T. et W.D. Kalsbeek. 1992. *Nonsampling Errors in Surveys*. John Wiley and Sons, New York.

Linacre, S.J. et D.J. Trewin. 1989. Evaluation of Errors and Appropriate Resource Allocation in Economic Collections. *Proceedings of the Annual Research Conference*. U.S. Bureau of the Census. 197-209.

Lyberg, L., P. Biemer, M. Collins, E. de Leeuw, C. Dippo, N. Schwarz et D. Trewin, Éds. 1997. *Survey Measurement and Process Quality*. John Wiley and Sons, New York.

Statistique Canada. 1998. *Statistique Canada - Lignes directrices concernant la qualité*. Troisième édition. 12-539-XIF.

Swain, L., J.D. Drew, B. Lafrance et K. Lance. 1992. La Création d'un registre des adresses résidentielles pour améliorer la couverture du recensement du Canada de 1991. *Techniques d'enquête*. 18(1): 139-156.

Swain, L. et D. Dolson. 1997. Current Issues in Household Survey Nonresponse at Statistics Canada. *Statistics in Transition*. 3: 439-468.

PUBLICATIONS ÉLECTRONIQUES
DISPONIBLES À

www.statcan.gc.ca

Chapitre 4 - Méthodes de collecte des données

4.0 Introduction

La collecte des données est le processus qui permet d'obtenir l'information nécessaire pour chaque unité sélectionnée de l'enquête. Pendant la collecte des données, les intervenants de l'enquête déterminent où sont les membres de la population, c'est-à-dire des particuliers ou des organismes, ils communiquent avec eux et leur demandent de participer à l'enquête. Un questionnaire est ensuite administré et les réponses sont enregistrées. Ce processus coûte cher, demande beaucoup de temps et énormément de ressources, et il a des répercussions directes sur la qualité des données. Étant le principal moyen de communication du grand public avec l'organisme statistique, il contribue à l'image de marque de l'organisme et a de grandes répercussions sur la pertinence de l'organisme et la qualité de ses données. Au cours de la phase de planification de l'enquête, il faut prendre de nombreuses décisions sur la méthode de collecte des données. Un intervieweur devrait-il administrer le questionnaire? Si oui, faut-il faire une interview téléphonique ou sur place? Faut-il appliquer une combinaison de méthodes, les répondants devraient-ils remplir le questionnaire eux-mêmes et faut-il faire le suivi auprès des non-répondants au cours d'une interview téléphonique? Le questionnaire devrait-il être sur support papier ou électronique? Faut-il utiliser des données administratives pour obtenir certaines données de l'enquête? La collecte des données pour plusieurs enquêtes devrait-elle être combinée?

La méthode de collecte des données choisie devrait donner un taux de participation élevé et les données obtenues devraient être les plus complètes et précises possibles, mais la méthode devrait aussi minimiser le fardeau pour les répondants et tenir compte du budget et des limites opérationnelles du client.

L'objectif de ce chapitre est de présenter les diverses méthodes de collecte des données, y compris l'enquête par autodénombrement, les méthodes assistées par intervieweur ou ordinateur et le recours aux données administratives, ainsi que les critères qui aident à déterminer quelle méthode est la plus appropriée. Les opérations de collecte des données en général (l'accent est mis sur les méthodes assistées par intervieweur), par exemple, comment repérer les unités de l'échantillonnage, susciter la collaboration et saisir les réponses sont exposées au **Chapitre 9 - Opérations de collecte des données.**

4.1 Méthodes élémentaires de collecte des données

Voici les méthodes élémentaires de collecte des données :

i. Autodénombrement

Le répondant remplit le questionnaire d'enquête par autodénombrement sans l'aide d'un intervieweur. Divers moyens peuvent servir à envoyer le questionnaire au répondant et à le retourner à l'expéditeur : le service postal, le télécopieur, un moyen électronique (y compris Internet) ou un enquêteur. (Si le questionnaire est retourné par télécopieur ou sur support électronique, une ligne sécuritaire ou le chiffrage est alors nécessaire pour garantir la confidentialité des données du répondant). La méthode sur support papier est intitulée interview papier et crayon (IPC) et la méthode sur support électronique est intitulée auto-interview assistée par ordinateur (AIAO).

ii. Questionnaire assisté par intervieweur (interviews téléphoniques ou sur place)

a. Interviews sur place

Un intervieweur aide le répondant à remplir le questionnaire. L'interview se déroule sur place, habituellement à la résidence du répondant ou en milieu de travail, même si elle peut avoir lieu dans un endroit public (p. ex., aéroport, centre commercial). La méthode sur support papier est intitulée interview papier et crayon (IPC) et la méthode assistée par ordinateur est intitulée interview sur place assistée par ordinateur (IPAO).

b. Interviews téléphoniques

Un intervieweur aide le répondant à remplir le questionnaire au téléphone. La méthode sur support papier est intitulée interview papier et crayon (IPC) et la méthode assistée par ordinateur est intitulée interview téléphonique assistée par ordinateur (ITAO).

Ce chapitre commence par un exposé des méthodes élémentaires de collecte des données, et surtout de la collecte sur support papier (même si de nombreux commentaires ciblent aussi les méthodes assistées par ordinateur). Les avantages et les inconvénients de la collecte assistée par ordinateur sont expliqués à la Section 4.2. D'autres méthodes de collecte des données, notamment l'utilisation de données administratives, sont définies à la Section 4.3. Le tableau à la fin du chapitre présente une comparaison des méthodes de collecte des données.

4.1.1 Autodénombrement

Les méthodes d'enquête par autodénombrement exigent un questionnaire très bien structuré, facile à suivre et donnant des instructions claires au répondant. Il peut y avoir un numéro de téléphone pour obtenir de l'aide, afin de remplir le questionnaire. Celui-ci a habituellement une présentation visuelle plus élaborée qu'un questionnaire assisté par intervieweur et ce, pour susciter la participation du répondant. (Le **Chapitre 5 - Conception du questionnaire** donne davantage de détails sur la méthode de conception du questionnaire.)

Comparativement à la gestion des interviews, l'administration de l'enquête par autodénombrement est relativement facile. Elle coûte aussi habituellement moins cher que les méthodes assistées par intervieweur et des échantillons de plus grande taille peuvent être sélectionnés. Cette méthode est utile pour les enquêtes qui exigent de l'information détaillée parce que le répondant peut consulter des dossiers personnels. Voilà qui peut diminuer les erreurs de réponse parce que le répondant n'a pas à faire appel uniquement à la mémoire. L'une des applications de l'autodénombrement comprend le journal ou le carnet de notes. Au cours d'une enquête par journal, le répondant prend des notes pendant la période de référence de l'enquête, par exemple, un journal d'auditeur pour les enquêtes sur la radio et la télévision pendant une semaine en particulier, ou un carnet de notes sur les achats d'essence pour une enquête sur la consommation d'essence des véhicules. L'autodénombrement est aussi utile pour les questions à caractère délicat parce que le questionnaire peut être rempli en privé, sans intervieweur.

L'un des inconvénients de l'enquête par autodénombrement est que les répondants doivent avoir des connaissances ou une bonne scolarité, ou le sujet d'enquête doit être très simple. Autre inconvénient : les taux de réponse sont habituellement plus faibles que ceux des méthodes assistées par intervieweur parce qu'il n'y a pas de pression exercée pour que le répondant réponde entièrement au questionnaire. Le taux de réponse aux enquêtes par autodénombrement à Statistique Canada est habituellement inférieur à 70 %. (Le Recensement de la population est une exception, le taux de retour par la poste étant de 85 %, à cause

d'une vaste campagne de publicité et du caractère obligatoire de l'enquête.) On a souvent recours à de nombreux suivis, pour demander aux répondants de remplir entièrement le questionnaire, ou à des intervieweurs au téléphone pour obtenir un bon taux de réponse (voir la Section 4.3.4). De plus, même si le questionnaire peut contenir beaucoup de matériel de référence sur les concepts de l'enquête et des guides pour aider à remplir le questionnaire, le résultat n'est habituellement pas aussi bon qu'en présence d'un intervieweur parce que de nombreux répondants ne lisent pas les instructions. La qualité peut donc être médiocre, comparativement aux méthodes assistées par intervieweur, parce que le répondant peut manquer les instructions « passez à », mal interpréter l'information, etc. Voilà pourquoi l'enquête par autodénombrement exige le suivi après la collecte pour corriger les erreurs.

Il y a plusieurs moyens de livrer et de reprendre les questionnaires d'enquête par autodénombrement. Il faut examiner de près le choix du moyen de livraison et de ramassage des questionnaires, considérer attentivement la qualité des données, les coûts de la collecte, la durée de la période de collecte, les taux de réponse, etc., avant de choisir la combinaison qui convient le mieux. La base de sondage utilisée et l'information de la collecte disponible dans la base auront des répercussions sur ce choix : pour envoyer les questionnaires par la poste, il faut avoir le nom et l'adresse postale, et le système postal doit être fiable. La base de sondage doit contenir les numéros de télécopieur des répondants pour utiliser ce moyen. Voici les moyens les plus habituels de livraison et de retour des questionnaires :

- livraison – ramassage du questionnaire sur support papier en personne,
- envoi par la poste – ramassage du questionnaire sur support papier en personne,
- livraison en personne – retour du questionnaire sur support papier par la poste,
- envoi – retour du questionnaire sur support papier par la poste.

Lorsqu'un questionnaire sur support papier est livré et retourné par la poste, l'enquête par autodénombrement demande une longue période de collecte parce que c'est la méthode la plus lente de collecte des données. Au Canada, c'est aussi la méthode la moins onéreuse.

4.1.2 Méthodes assistées par intervieweur

Le principal avantage des méthodes assistées par intervieweur est que l'interview est personnalisée, les questions et les concepts de l'enquête peuvent être interprétés, et l'intervieweur peut augmenter le taux de réponse et la qualité des données dans l'ensemble. Les méthodes assistées par intervieweur sont particulièrement utiles pour les populations d'enquête dont les taux d'alphabétisation sont modestes, lorsque le questionnaire ou les concepts sont complexes, ou chaque fois que l'enquête par autodénombrement serait difficile.

L'intervieweur peut augmenter le taux de réponse en suscitant l'intérêt pour l'enquête et il peut répondre aux questions du répondant sur la confidentialité des données, l'objectif de l'enquête, ce qui lui est demandé pendant l'interview, la longueur de l'interview, l'utilisation des résultats de l'enquête, etc. Certains répondants peuvent avoir l'impression, par exemple, que l'information qu'ils donnent pourra être utilisée à leur détriment, ils peuvent soutenir que la matière du sujet a un caractère délicat ou ils peuvent craindre de ne pas avoir la « bonne » réponse. L'intervieweur peut garantir au répondant que les données seront en sécurité et que l'organisme statistique respectera toutes les procédures de sécurité pour maintenir la confidentialité.

La qualité des données dans l'ensemble peut être améliorée parce que l'intervieweur peut obtenir une formation approfondie sur les concepts et les définitions de l'enquête et aider le répondant en cas de problème d'interprétation du questionnaire. L'enquêteur peut empêcher les erreurs de réponse et la non-réponse partielle en repérant immédiatement les erreurs et en les corrigeant en présence du répondant. Cette intervention diminue aussi le nombre de suivis qui peut demander beaucoup de temps à l'organisme

qui fait enquête et représenter un fardeau pour le répondant. Enfin, l'intervieweur peut améliorer la qualité des données en vérifiant s'il y a eu communication avec l'unité d'échantillonnage choisie et si c'est bien elle qui est interviewée.

Autre avantage de l'interview : elle permet des périodes de collecte plus souples. Si la collecte des données est trop lente et s'il faut accélérer le processus, d'autres intervieweurs peuvent être engagés. Cette solution ne s'applique pas aux méthodes d'enquête par autodénombrement parce qu'il est pratiquement impossible de déterminer quand le répondant remplit et retourne le questionnaire.

Voici les deux principaux inconvénients des méthodes assistées par intervieweur : elles peuvent coûter cher et la gestion peut être difficile. Certaines dépenses comprennent la rémunération de l'intervieweur, sa formation, les frais de transport et d'hébergement (pour les interviews sur place) ou la superficie de bureau et les téléphones pour les interviews téléphoniques centralisées. Autres inconvénients des méthodes assistées par intervieweur : la formation médiocre de l'intervieweur peut occasionner des erreurs de réponse et, dans le cas des sujets à caractère délicat, le répondant peut hésiter à répondre aux questions (même si l'interview téléphonique permet un certain anonymat). Si un intervieweur bien formé n'est pas disponible et qu'un biais lié à un intervieweur devient un problème sérieux, l'enquête par autodénombrement peut être préférable.

Les erreurs de réponse ont été étudiées au **Chapitre 3 - Introduction au plan d'enquête**; le sujet est repris au **Chapitre 5 - Conception du questionnaire.** Les sections suivantes exposent les avantages et les inconvénients des interviews téléphoniques et sur place.

4.1.2.1 Interviews sur place

L'interview sur place se déroule en présence du répondant. Celle-ci est habituellement faite à la résidence de la personne ou en milieu de travail. C'est la seule méthode réaliste de collecte des données pour certaines populations cibles, par exemple, lorsque l'interview téléphonique est impossible ou que l'enquête exige une visite pour échantillonner ou repérer des membres de la population (p. ex., pour achever le listage d'une base aréolaire).

Les interviews sur place donnent souvent les taux de réponse les plus élevés (habituellement, de 80 % à 95 % pour Statistique Canada), mais c'est la méthode de collecte la plus onéreuse d'habitude, compte tenu des frais de transport et d'hébergement des intervieweurs. Cette méthode de collecte peut donc inciter à sélectionner des échantillons de plus petite taille que celle des interviews téléphoniques ou d'enquête par autodénombrement. L'interview sur place pose un autre problème : il peut être difficile de rencontrer la personne à la maison ou au travail et l'intervieweur devra peut-être visiter la résidence ou le lieu de travail plusieurs fois avant de réussir à communiquer avec le répondant. Celui-ci est parfois présent, mais l'heure ne convient pas, et l'intervieweur doit convenir d'une nouvelle rencontre pour l'interview.

Voici d'autres avantages de l'interview sur place :
- l'intervieweur peut faire des observations directes (qui sont impossibles pendant l'interview téléphonique),
- l'intervieweur réussit généralement mieux à convaincre une personne qui refuse de répondre,
- l'intervieweur peut inspirer confiance en montrant au répondant ses pièces d'identité officielles.

Voici d'autres inconvénients de l'interview sur place :
- il est parfois difficile de retenir les services d'un intervieweur raisonnablement qualifié dans tous les domaines enquêtés,

- il est difficile de confier des charges de travail à des intervieweurs moins débordés,
- il est difficile d'appliquer un programme de contrôle qualitatif au processus de l'interview.

4.1.2.2 Interviews téléphoniques

L'interview téléphonique offre un taux de réponse raisonnable à un coût raisonnable. Ce genre d'interviews donne des taux de réponse de moyens à élevés au Canada, inférieurs à ceux de l'interview sur place, mais supérieurs à ceux de l'enquête par autodénombrement (le taux de réponse habituel est de 70 % à 85 % à Statistique Canada). Certaines enquêtes par autodénombrement comprennent l'interview téléphonique de suivi pour obtenir un meilleur taux de réponse. L'interview téléphonique coûte habituellement moins cher que l'interview sur place parce qu'il n'y a pas de frais de déplacement de l'intervieweur et la collecte est habituellement plus rapide que celle de l'interview sur place ou de l'enquête par autodénombrement. L'interview téléphonique permet aussi de poser des questions à caractère délicat, mais cette méthode n'est pas aussi anonyme que celle de l'enquête par autodénombrement. Cette méthode de collecte est plus sécuritaire que l'interview sur place parce que l'intervieweur n'a pas besoin d'aller dans des endroits dangereux ou isolés. Si le répondant est absent ou s'il veut reporter l'interview, la communication avec celui-ci demande aussi moins de temps que dans le cas de l'interview sur place. Enfin, le contrôle qualitatif du processus de l'interview peut être appliqué facilement parce que la surveillance de l'interview téléphonique est moins difficile.

L'enquête téléphonique a un inconvénient : la longueur de l'interview et la complexité du questionnaire sont limitées parce que le répondant a moins de patience pendant une longue interview complexe au téléphone. Il est mieux disposé pendant une interview sur place. Comme un intervieweur administre le questionnaire, celui-ci peut donc être plus complexe que celui de l'enquête par autodénombrement. Les observations directes sont impossibles au téléphone, et c'est un autre inconvénient.

Voici d'autres inconvénients de l'interview téléphonique :
- il peut être difficile d'établir une base de sondage avec une bonne couverture des numéros de téléphone,
- l'échantillonnage des numéros de téléphone est souvent inefficiente (c.-à-d. qu'il est possible de téléphoner à de nombreuses unités hors du champ de l'enquête),
- la confidentialité peut être un problème si une autre personne peut entendre les réponses du répondant (p. ex., lignes téléphoniques partagées),
- l'interview téléphonique est moins personnelle que l'interview sur place et il peut être plus difficile de convaincre les gens de l'importance de l'enquête,
- l'interview téléphonique peut coûter cher en interurbains.

L'échantillonnage des numéros de téléphone, y compris la composition aléatoire, est approfondi au **Chapitre 6 - Plans d'échantillonnage.**

4.1.2.2.1 Échantillonnage des interviewés par téléphone

Le processus de l'interview téléphonique comprend souvent le processus de sélection de l'échantillon : l'intervieweur sélectionne souvent l'échantillon à la première étape de l'interview. Nous décrirons maintenant des questions pertinentes à l'échantillonnage des interviewés au téléphone qui peuvent déterminer si l'interview téléphonique est la méthode de collecte des données appropriée pour une enquête en particulier.

L'échantillonnage des interviewés par téléphone et l'interview téléphonique sont souvent choisis pour les enquêtes auprès des ménages comme compromis pratique entre les échantillons de base aréolaire avec interviews sur place plus onéreuses, mais de qualité supérieure, et les enquêtes par questionnaire d'autodénombrement envoi-retour par la poste moins onéreuses, mais de qualité inférieure. Le recours à l'échantillonnage des interviewés par téléphone et à l'interview téléphonique permet d'éviter les coûts élevés de mise à jour des bases aréolaires et du temps de déplacement pour les interviews sur place, tout en obtenant des taux de réponse raisonnablement élevés. Il est important que la population dont on tire l'échantillon soit représentative de la population cible, au moins du point de vue des caractéristiques d'intérêt pour l'enquête, comme c'est toujours le cas pour toute méthode d'échantillonnage. L'enquête téléphonique peut poser un problème en ce sens si un pourcentage important de la population cible n'a pas le téléphone ou s'il y a des écarts importants entre les sous-populations. (Certains biais éventuels à cause des écarts des taux de service téléphonique peuvent être diminués de la même façon que les taux différentiels de non-réponse peuvent être pris en compte.)

Un exemple de biais que peut susciter l'utilisation d'une base de sondage non représentative est le cas maintenant classique de l'enquête d'opinion du Literary Digest effectuée pendant les élections présidentielles de 1936 aux É.-U. Le sondage soutenait que le candidat républicain Alf Landon l'emporterait sur le président Roosevelt :

Landon	55 %,
Roosevelt	41 %.

Les résultats ont cependant été très différents :

Landon	37 %,
Roosevelt	61 %.

La base utilisée pour cette enquête-échantillon par envoi et retour par la poste était axée surtout sur les adresses trouvées dans les répertoires téléphoniques et les listes d'enregistrement des automobiles. Les Américains propriétaires d'automobiles et de téléphones en 1936 étaient généralement bien nantis et votaient pour le Parti républicain. Un pourcentage important de l'électorat n'avait cependant ni téléphone ni automobile et ces citoyens avaient tendance à voter pour le Parti démocratique. (Au Canada, aux États-Unis et dans de nombreux pays d'Europe de nos jours, à peu près tous les ménages ont le service téléphonique, et la possibilité d'obtenir des résultats biaisés pour de nombreuses estimations des enquêtes auprès des interviewés par téléphone est beaucoup moindre que ce n'est le cas dans cet exemple de 1936.)

Le choix de la base de sondage est une importante question lors de la sélection d'un échantillon des interviewés par téléphone. La base de sondage de l'enquête par téléphone serait complète si elle comprenait tous les numéros de téléphone utilisés et, pour qu'elle soit efficiente, elle devrait contenir le moins possible de numéros non utilisés. L'amélioration de l'exhaustivité d'une base de sondage d'enquête par téléphone en réduit généralement l'efficience. Il est important d'essayer d'obtenir autant l'efficience que l'exhaustivité. Le recours aux annuaires téléphoniques (en direct ou sur support papier) comme base de sélection d'échantillons d'interviewés par téléphone peut être très efficient, mais le manque d'exhaustivité augmente le risque de résultats biaisés. Les numéros de téléphone non publiés ne sont pas dans ces annuaires et c'est évident, les annuaires sont toujours périmés depuis plusieurs mois, ou même plusieurs années, et les gens qui ont de nouveaux numéros n'y sont pas inscrits non plus. Des techniques de composition aléatoire (CA) sont habituellement appliquées pour améliorer l'exhaustivité de la couverture d'un échantillon d'interviewés par téléphone. Voici un exemple de ce qui peut être fait au Canada pour obtenir l'efficience et l'exhaustivité.

La composition des numéros de téléphone varie d'un pays à l'autre, mais au Canada, le modèle nord-américain est utilisé, c.-à-d. le numéro de téléphone à dix chiffres : un indicatif régional à trois chiffres, suivi d'un préfixe à trois chiffres auxquels s'ajoutent quatre chiffres supplémentaires. Il y a actuellement 21 indicatifs régionaux utilisés au Canada, et une seule base de sondage consisterait en l'annexion de 10 millions de numéros à sept chiffres possibles pour chacun des 21 indicatifs régionaux, le résultat donnant 210 millions de numéros de téléphone dans la base de sondage. Celle-ci serait absolument complète (jusqu'à ce qu'un nouvel indicatif régional soit ajouté), mais, au Canada actuellement, environ 13 millions de numéros de téléphone seulement sont attribués aux ménages, c'est-à-dire que 94 % des numéros sélectionnés au hasard dans cette base de sondage ne permettraient pas de communiquer avec les ménages. Il serait donc peu efficient d'utiliser la base. Il est cependant possible de l'améliorer. Il est possible d'acheter une liste de toutes les combinaisons d'indicatifs régionaux et de préfixes utilisés en Amérique du Nord. Il y en a actuellement 8 600 au Canada, à partir desquelles une base peut être établie en annexant à chacun les 10 000 numéros à quatre chiffres possibles pour obtenir une base de 86 millions de numéros dont 85 % seulement ne permettraient pas de communiquer avec les ménages. La méthode Mitofsky-Waksberg peut être appliquée pour améliorer davantage l'efficience opérationnelle de cette base de sondage au risque de compliquer les procédures de terrain et de diminuer éventuellement l'efficience statistique en ayant recours aux grappes. Statistique Canada améliore davantage la base d'échantillonnage des interviewés par téléphone à l'aide des données administratives.

Une banque de centaine comprend les 100 numéros de téléphone dont l'indicatif régional, le préfixe et les deux numéros suivants sont identiques. À l'aide des listes administratives de numéros de téléphone publiés, il est possible d'identifier toutes les banques de centaine qui contiennent au moins un numéro de téléphone de ménage publié. Il y a environ 260 000 de ces ***banques actives*** qui donnent une base de sondage contenant 26 millions de numéros de téléphone dont environ la moitié seulement ne permettent pas de communiquer avec les ménages. Les deux premières bases de sondage sont complètes à un moment donné (et elles le sont jusqu'à ce que de nouveaux indicatifs régionaux ou de nouvelles combinaisons d'indicatifs régionaux et de préfixes soient activés), mais la base de sondage fondée sur les banques de centaine actives peut être incomplète. Si une banque de centaine ne contient pas de numéros de ménage publiés, mais si elle contient certains numéros de ménage non publiés, cette caractéristique ne paraîtra pas dans la base comme elle le devrait. Les sociétés de téléphone utilisent aussi de nouvelles banques de centaine beaucoup plus souvent que des indicateurs régionaux et des préfixes nouveaux, et les nouveaux indicatifs régionaux et préfixes sont publiés avant d'être activés. Cette dernière méthode de CA est intitulée troncation des banques sans numéros listés ou élimination des banques inutiles.

Même si les banques inutiles sont éliminées pour améliorer l'efficience, environ la moitié des numéros de téléphone d'un échantillon obtenu par CA pour une enquête auprès des ménages au Canada seront des numéros hors du champ de l'enquête. Avant de faire une interview au téléphone, l'intervieweur doit donc confirmer que le numéro est dans le champ de l'enquête. Dans le cas des enquêtes par CA au Canada, l'intervieweur passe de quatre à six minutes de son temps sur des numéros hors du champ de l'enquête pour chaque interview achevée. Ces minutes peuvent représenter un pourcentage important du temps total que l'intervieweur passe à chaque interview si le questionnaire est bref. C'est néanmoins relativement peu, comparativement au temps de déplacement nécessaire pour les interviews sur place.

Le plan d'échantillonnage par composition aléatoire est moins souple que les plans de base aréolaire du point de vue de la stratification. (La stratification répartit la population de l'enquête en sous-populations, par exemple, en provinces. Trois principales raisons justifient la stratification : faire en sorte que la stratégie d'échantillonnage soit efficiente, garantir des tailles d'échantillon appropriées pour les sous-populations particulières qui font l'objet de l'analyse et éviter de tirer un échantillon « erroné ».) Les bases aréolaires donnent une souplesse pour le choix de la strate géographique, mais pour les enquêtes par CA, la géographie de la stratification doit être axée sur l'indicatif régional et le préfixe (ou des concepts semblables pour les réseaux téléphoniques hors de l'Amérique du Nord). Des secteurs géographiques

correspondent généralement à ces indicatifs et préfixes, mais ils ne correspondent peut-être pas aux limites municipales ou à d'autres limites d'importance pour l'enquête. L'échantillonnage par CA permet la sélection d'échantillons de ménages non constitués en grappe pour compenser et donne des échantillons qui ont tendance à être plus efficients du point de vue statistique (effets de plan moindres) que les échantillons des bases aréolaires.

Dillman (1978), Groves et coll. (1979), Groves et coll. (1988) et Lavrakis (1987) donnent davantage d'information sur l'échantillonnage des ménages par téléphone. Les problèmes de couverture de la base de sondage ont été vus au **Chapitre 3 - Introduction au plan d'enquête.** Le **Chapitre 6 - Plans d'échantillonnage** expose des considérations détaillées sur les plans d'échantillonnage. Le **Chapitre 7 - Estimation** donne de l'information sur les corrections à apporter pour les non-réponses.

4.1.3 Choix d'une méthode d'enquête par autodénombrement ou assistée par intervieweur

Il faut considérer diverses questions pour sélectionner une méthode de collecte des données :
- l'information pour la collecte disponible dans la base de sondage,
- les caractéristiques de la population cible,
- le genre de questions posées,
- les ressources disponibles (p. ex., les intervieweurs),
- la facilité à remplir le questionnaire,
- les considérations sur la vie privée,
- les exigences de qualité des données.

L'information pour la collecte disponible dans la base de sondage est un élément important pour déterminer la méthode de collecte des données la plus appropriée. Si la base ne comprend pas les adresses postales, les questionnaires d'enquête par autodénombrement ne peuvent être envoyés aux répondants par la poste. Si les numéros de téléphone à jour ne sont pas disponibles et si la composition aléatoire est considérée inappropriée, les interviews ne peuvent donc être faites par téléphone.

Les caractéristiques de la population cible ont des répercussions sur la méthode de collecte des données. Si le taux d'alphabétisation de la population est faible ou si les difficultés de communication sont un problème (p. ex., les immigrants), les méthodes assistées par intervieweur peuvent être la seule option. La répartition géographique de la population et de l'échantillon sont aussi importantes. Si la population et l'échantillon sont largement dispersés au pays, les interviews sur place pourraient coûter trop cher et être trop difficiles à accomplir. (La répartition de la population et le coût de la collecte des données sont des éléments qui aident à déterminer la méthode d'échantillonnage la plus appropriée comme on l'explique au **Chapitre 6 - Plans d'échantillonnage**.)

Le genre de questions de l'enquête a des répercussions sur la collecte des données. Dans le cas de la matière à caractère délicat, une méthode de collecte axée sur l'anonymat, notamment les interviews téléphoniques et d'enquête par autodénombrement, peut être la plus appropriée. Si des questions complexes sont posées, un intervieweur peut être nécessaire pour expliquer les questions et les concepts. Si l'intervieweur doit faire des observations ou prendre des mesures (p. ex., administration d'un examen d'alphabétisation aux enfants) ou présenter le matériel aux répondants (p. ex., graphiques ou diagrammes), l'interview sur place peut alors être nécessaire.

Les ressources disponibles ont des répercussions profondes sur le choix de la méthode de collecte des données, Ces ressources comprennent le budget, le personnel, le matériel et le temps disponibles. L'application d'une méthode assistée par intervieweur exige un budget suffisant pour l'embauche, la formation et les déplacements des intervieweurs. L'organisme statistique doit aussi être en mesure

d'obtenir le nombre d'intervieweurs nécessaires. Si une méthode assistée par ordinateur est sélectionnée, des programmeurs chevronnés seront nécessaires, ainsi que le matériel informatique approprié.

Certaines méthodes de collecte des données sont plus faciles à administrer que d'autres. Les interviews téléphoniques centralisées (c.-à-d. que tous les intervieweurs sont installés au même endroit pour téléphoner), par exemple, sont plus faciles à organiser que les interviews sur place et géographiquement dispersées. La période limite de collecte des données est aussi importante : les méthodes par autodénombrement sont habituellement plus lentes que les méthodes assistées par intervieweurs et les méthodes manuelles sur support papier sont normalement plus lentes que celles assistées par ordinateur.

Il faudrait en bout de ligne considérer les exigences de qualité des données lors de la sélection d'une méthode de collecte des données. Les intervieweurs bien formés aux concepts utilisés dans l'enquête peuvent réduire les erreurs de réponse et les non-réponses. Il faudrait considérer les exigences de précision : les échantillons plus nombreux donnent généralement des estimations plus précises (c.-à-d. des estimations comprenant une erreur d'échantillonnage de moindre importance), mais plus la méthode de collecte des données coûte cher, plus l'échantillon à la portée des moyens du client est réduit. Les interviews sur place sont souvent la méthode la plus chère et les enquêtes par autodénombrement, la moins chère. La capacité de mesurer la qualité et d'appliquer les procédures de contrôle qualitatif peut aussi être importante. Il est plus facile de surveiller la qualité des interviews téléphoniques, par exemple, que celle des interviews sur place.

Le tableau suivant affiche une comparaison entre les méthodes de collecte des données d'enquête par autodénombrement, par interview sur place et par interview téléphonique, compte tenu du temps nécessaire pour achever la collecte des données et déterminer les taux de réponse.

Tableau 1 : Méthodes de collecte des données d'enquête par autodénombrement et assistée par intervieweur

	Autodénombrement	Intervieweur	
		Sur place	**Téléphonique**
Coût	Faible	Élevé	Raisonnable
Temps	Plus long	Moyen	Moins long
Taux de réponse	Faible	Élevé	Moyen - élevé

Les détails pour déterminer comment sélectionner un plan d'échantillonnage sont couverts au **Chapitre 6 - Plans d'échantillonnage**. Les éléments qui déterminent la taille de l'échantillon sont étudiés au **Chapitre 8 - Calcul de la taille de l'échantillon et répartition.**

4.2 Collecte des données assistée par ordinateur

Un élément important du plan d'enquête est de déterminer si la collecte comprend des méthodes assistées par ordinateur ou une approche traditionnelle sur support papier, auquel cas les réponses sont inscrites dans un questionnaire sur support papier. La méthode intitulée interview papier et crayon (PAPI pour *paper and pencil interview*[1]) est encore appliquée parfois, mais la collecte des données assistée par ordinateur devient prédominante.

[1] Dans ce manuel, on utilisera l'acronyme anglais PAPI plutôt que le français IPC pour éviter toute confusion avec l'Indice des prix à la consommation.

Si l'enquête doit se dérouler une seule fois, les méthodes sur support papier coûtent souvent moins cher et demandent moins de temps d'élaboration que les méthodes assistées par ordinateur. La saisie des données, c.-à-d. le transfert des réponses en une mise en forme interprétable par la machine, devient cependant une opération distincte après la collecte. La saisie des données est une étape de l'enquête nécessaire parce que toutes les données d'enquête doivent en bout de ligne être entrées et sauvegardées dans un ordinateur.

Voici d'autres inconvénients de la collecte sur support papier :
- la collecte manuelle des données demande beaucoup de temps et la lecture optique du questionnaire n'est peut-être pas une option,
- le questionnaire ne devrait pas comprendre des instructions « passez à » ou des vérifications compliquées,
- l'impression des questionnaires peut coûter cher,
- l'envoi des questionnaires par la poste peut coûter cher,
- les questionnaires remplis doivent être rangés et protégés en toute sécurité pour garantir la confidentialité des données des répondants.

L'avantage principal des méthodes assistées par ordinateur est la simultanéité de la collecte et de la saisie des données, le résultat étant un processus de saisie et de collecte intégré, plus rapide et plus efficient. N'importe quelle méthode de collecte des données peut servir au déroulement de l'interview assistée par ordinateur (IAO) :
- autodénombrement (auto-interview assistée par ordinateur, AIAO),
- téléphone (interview téléphonique assistée par ordinateur, ITAO),
- sur place (interview sur place assistée par ordinateur, IPAO).

L'AIAO est une technologie relativement récente et peu utilisée. Le questionnaire, ainsi qu'un programme de vérification pour repérer les entrées manquantes, erronées ou incohérentes, et des caractéristiques d'aide, sont envoyés au répondant en mise en forme électronique. Le répondant remplit le questionnaire à l'aide de son ordinateur. Cette méthode permet au répondant qui a le matériel informatique et le logiciel nécessaires de saisir et de vérifier directement les données à l'aide de son ordinateur pendant que le système l'incite à passer d'une question de l'enquête à l'autre. Le questionnaire sur disquette peut être envoyé par la poste ou par messager, ou le fichier électronique peut être acheminé par modem à l'aide d'une ligne protégée.

Statistique Canada procède actuellement à des projets pilotes de collecte des données en mise en forme structurée standard en toute sécurité sur Internet. Le recours à Internet a des avantages : les coûts de collecte et de saisie des données diminuent et la rapidité d'exécution est à la hausse. Internet à un inconvénient : le questionnaire doit être compatible avec les différentes versions logicielles de la toile (p. ex., Explorer, Netscape, Windows, UNIX, etc.). Autre inconvénient : le nombre d'utilisateurs de l'Internet à haute vitesse est relativement faible (même si ce nombre pourrait augmenter rapidement au cours des prochaines années) et cette option est donc peu probable pour les enquêtes auprès des ménages, mais elle est plus réaliste pour les enquêtes auprès des entreprises.

L'AIAO a un avantage en général : elle est souple et pratique pour le répondant qui a le matériel informatique et le logiciel, mais tous les répondants n'ont pas le matériel informatique nécessaire, et c'est le principal inconvénient.

Pour l'ITAO et l'IPAO, chaque intervieweur dispose d'un ordinateur. L'intervieweur lit un scénario affiché à l'écran et entre les réponses directement dans l'ordinateur. L'ITAO et l'IPAO deviennent prédominantes au Canada à mesure que la technologie informatique évolue.

L'IAO a généralement de nombreux avantages comparativement au PAPI, surtout à cause de la collecte et de la saisie simultanées des données. Il est plus facile de faire la surveillance et le contrôle qualitatif des données parce que la collecte, la vérification automatisée et la saisie des données sont plus uniformes et contrôlées que dans le cas des méthodes sur support papier. La vérification automatisée signifie aussi que les rejets à la vérification peuvent être résolus immédiatement, ce qui diminue le fardeau de réponse et la nécessité de suivi. Il est plus facile d'appliquer le contrôle qualitatif du processus d'interview et de produire des rapports de gestion sur le statut des interviews (p. ex., taux de réponse, nombre d'interviews achevées, nombre d'interviews en instance, durée de chaque interview, etc.). Le questionnaire peut être plus complexe du point de vue des instructions « passez à » et des vérifications. Les résultats sont souvent plus rapides que dans le cas des enquêtes sur support papier (en particulier les questionnaires envoyés et retournés par la poste).

La collecte assistée par ordinateur a un inconvénient : la personne qui remplit le questionnaire, le répondant ou l'intervieweur, doit être formée et bien connaître l'application logicielle. (S'il connaît bien l'application, l'intervieweur peut cependant réserver plus de temps aux aptitudes interpersonnelles.) Les enquêtes assistées par ordinateur exigent aussi un travail de développement approfondi et coûteux de la part des experts en programmation informatique (problème qui peut être relativement amenuisé s'ils peuvent adapter à l'enquête l'application informatique d'une autre enquête). Le questionnaire doit être programmé pour que chaque question soit affichée à l'écran de l'ordinateur selon la séquence appropriée. L'application doit être soigneusement mise à l'essai pour garantir que les écrans sont affichés dans l'ordre approprié et qu'ils orientent l'intervieweur ou le répondant sur la voie prédéterminée. Les vérifications, l'aide en direct et les fonctions supplémentaires qui aident les intervieweurs ou les répondants doivent aussi être programmées et mises à l'essai. Le coût du matériel est aussi un autre inconvénient des enquêtes assistées par intervieweur parce que chacun d'eux a besoin d'un ordinateur (même si ce coût peut être réparti entre les enquêtes). Le rangement en toute sécurité des questionnaires sur support papier pour protéger la confidentialité des données des répondants n'est pas nécessaire, un avantage dans ce cas, mais il faut protéger les ordinateurs contre le vol.

L'élaboration et la mise à l'essai de méthodes assistées par ordinateur, en particulier pour une nouvelle enquête, peut être un long processus qui coûte cher. S'il s'agit d'enquêtes comprenant un échantillon de grande taille cependant, la collecte par ordinateur peut réduire énormément les coûts de saisie et de vérification des données. S'il s'agit d'enquêtes réitérées, la collecte assistée par ordinateur peut coûter moins cher à long terme que la collecte sur support papier, compte tenu des économies d'impression, et parce que le coût d'élaboration peut être réparti sur plusieurs cycles de collecte.

Voici quelques autres avantages de la collecte assistée par ordinateur :
- elle est écologiquement conviviale (les questionnaires ne sont pas imprimés),
- l'interview connexe peut se dérouler facilement pour les enquêtes réitérées, afin de réduire les erreurs de réponse (c.-à-d. les renseignements d'un répondant fournis au cours d'un cycle précédent de l'enquête peuvent être utilisés au cours de cycles ultérieurs)…

et quelques autres inconvénients de la collecte assistée par ordinateur :
- le transfert entre ordinateurs (p. ex., de l'ordinateur de l'intervieweur à celui du bureau central) doit être fait à l'aide d'une ligne protégée pour garantir la confidentialité des données des répondants,
- elle est vulnérable aux difficultés techniques (vie utile de la pile, problèmes de transfert des fichiers, etc.), il faut beaucoup de temps pour les régler, et il est possible de perdre ou d'endommager des données,
- des experts informatiques sont nécessaires pour élaborer le logiciel et régler les problèmes techniques.

La saisie des données est considérée plus en détail au **Chapitre 9 - Opérations de collecte des données**. Le **Chapitre 10 - Traitement** cerne le traitement des données, y compris la lecture optique pour la saisie des données et la vérification.

4.3 Autres méthodes de collecte

Outre les méthodes assistées par intervieweur et d'enquête par autodénombrement, d'autres méthodes de collecte des données comprennent l'observation directe, la déclaration électronique des données, les données administratives, les méthodes combinées et les enquêtes omnibus ou supplémentaires.

4.3.1 Observation directe

Cette méthode consiste à observer ou mesurer directement les caractéristiques d'intérêt sur place ou en laboratoire. Elle peut être la seule possibilité pour certains concepts (p. ex., des données médicales) et elle est souvent appliquée aux enquêtes sur les prix. La télédétection est une forme d'observation directe qui interprète les images satellites. Elle est utilisée dans certaines enquêtes sur les exploitations agricoles pour estimer les types et les secteurs de culture. Ce genre de collecte ne peut être appliquée à la majorité des données parce qu'elles ne peuvent être observées ou mesurées directement.

La mesure directe est habituellement précise et, lorsque seules les observations sont faites, il n'y a pas de fardeau de réponse. Dans le cas de la mesure des gens, cependant, les sujets à l'étude peuvent considérer que ces mesures sont un tracas et un fardeau, par exemple, au cours d'une étude médicale lorsqu'il faut prélever des échantillons de sang des patients. Les taux de participation peuvent être faibles.

La mesure directe pose une difficulté, c.-à-d. qu'elle peut coûter cher parce qu'il faut former tous les intervieweurs à l'observation et à la mesure des données, et il pourrait être nécessaire d'embaucher des spécialistes (p. ex., des infirmières pour mesurer la tension artérielle). Si des spécialistes sont nécessaires et si seulement quelques-uns peuvent être engagés, le plan d'échantillonnage et la taille de l'échantillon peuvent être énormément restreints.

4.3.2 Déclaration électronique des données (DED)

Certaines enquêtes permettent aux répondants de fournir des données électroniques (p. ex., sur disquette, bande d'ordinateur ou cartouche), selon leur propre mise en forme. La DED est une forme d'autodénombrement qui peut être très pratique pour le répondant, mais elle est habituellement offerte seulement s'il n'y a aucun autre moyen d'obtenir les données. Certaines entreprises, par exemple, peuvent fournir leurs données seulement de cette façon.

Lorsque les données sont transférées de l'ordinateur du répondant à celui de l'organisme statistique à l'aide d'une ligne réservée (de modem à modem), il y a transfert de données d'ordinateur à ordinateur. On évite souvent la DED si une mise en forme standard ne peut être convenue avec le répondant parce qu'il faut énormément de travail pour vérifier et traiter les données, afin qu'elles conviennent à la mise en forme utilisée par l'organisme statistique.

4.3.3 Données administratives

Il est possible d'obtenir l'information nécessaire à certaines enquêtes à partir des données administratives. ***Les données administratives sont celles qui ont été obtenues à des fins administratives (p. ex., pour administrer, réglementer ou imposer des activités d'entreprises ou de particuliers), et non à des fins statistiques (pour étudier des groupes de particuliers, d'entreprises, d'exploitations agricoles, etc.).***
Les dossiers administratifs ont un énorme avantage parce qu'ils permettent d'éviter la majeure partie des coûts de collecte des données et du fardeau des répondants. Des résultats d'enquête rapides sont aussi possibles parce que les données existent déjà. L'objectif du programme administratif peut cependant être très différent de celui de l'enquête, et il faut donc évaluer prudemment les définitions et les concepts (p. ex., la population cible et la couverture de cette population). Il y a aussi un manque de contrôle qualitatif des données (déterminé par l'administrateur et non l'organisme statistique). Le suivi des rejets à la vérification est habituellement impossible. Il y a aussi un travail de traitement habituellement considérable à faire pour garantir la mise en forme des données administratives selon les exigences de l'organisme statistique. Enfin, la confidentialité peut susciter des préoccupations quant à l'utilisation de données administratives à des fins statistiques.

L'**Annexe A - Données administratives** offre davantage de détails à ce sujet.

4.3.4 Méthodes combinées

L'une des stratégies de collecte les plus satisfaisantes est d'offrir aux répondants un choix de méthode de collecte des données. Les avantages des méthodes combinées comprennent des taux de réponse améliorés, un nombre moins élevé d'erreurs de réponse et une collecte plus rapide. Les méthodes combinées ont un inconvénient, c.-à-d. que la collecte peut être plus complexe et coûter plus cher. Autre inconvénient : elles produisent des données hétérogènes qui peuvent compliquer le traitement et l'analyse. Si une enquête par questionnaire postal d'autodénombrement se déroule, par exemple, et s'il y a suivi téléphonique auprès des non-répondants à l'aide d'un questionnaire plus bref, les deux versions du questionnaire doivent être rapprochées pendant le traitement. Les résultats peuvent être biaisés si les données des interviews téléphoniques sont de meilleure qualité que celle des questionnaires d'enquête par autodénombrement, ce qui compliquera l'analyse des données.

L'enquête mensuelle est un autre exemple de méthodes combinées : il peut être rentable de communiquer avec le répondant en personne pour la première interview et au téléphone pour les interviews ultérieures. C'est le cas de l'Enquête canadienne sur la population active. Lors du recensement de la population canadienne, la collecte des données est faite avant tout à l'aide d'un questionnaire envoyé et retourné par la poste. Le suivi est fait au cours d'interviews téléphoniques et sur place. Le taux de réponse des enquêtes par questionnaire postal d'autodénombrement est amélioré à l'aide du suivi auprès des non-répondants en deux étapes, d'abord au téléphone, puis sur place. Une autre solution de plus en plus populaire : les données obtenues par questionnaire sont combinées aux données administratives, afin de réduire l'erreur de mémoire, le fardeau des répondants et les coûts de l'enquête.

4.3.5 Enquêtes omnibus et supplémentaires

Il est parfois possible d'appliquer un moyen de collecte à une autre enquête, à l'aide d'une enquête omnibus ou supplémentaire. Lors d'une enquête supplémentaire, le nouveau questionnaire est ajouté en supplément au questionnaire de l'autre enquête. Celle-ci est habituellement une enquête à grande échelle. Les enquêtes supplémentaires sont communes dans les organismes gouvernementaux. Un exemple à Statistique Canada est l'Enquête sur les voyages des Canadiens qui collecte de l'information sur les

déplacements et les caractéristiques des Canadiens qui voyagent, et elle se déroule en supplément à l'Enquête sur la population active.

Les questions de plusieurs enquêtes différentes sont combinées en un seul questionnaire pour une enquête omnibus. Le questionnaire de l'enquête omnibus est ainsi composé de plusieurs sections, chacune traitant d'un sujet différent pour un client différent. Les clients partagent les coûts de l'enquête proportionnellement à l'effort de collecte et de traitement à faire pour les différentes sections de l'enquête. Ce moyen peut donc être efficace pour réduire les coûts d'élaboration et l'exécution de l'enquête. Des organismes statistiques et des entreprises de recherche du secteur privé procèdent régulièrement à ce genre d'enquête qui engage plusieurs partenaires ayant divers besoins de recherche. Le principal avantage de cette approche est la diminution des coûts, souvent importante, comparativement au déroulement d'une enquête distincte pour chaque sujet.

Cette approche a un inconvénient, c.-à-d. que le répondant est aux prises avec un imposant questionnaire composé d'une variété de sujets qui peuvent être décousus et avoir parfois un caractère délicat. Le questionnaire peut donc être un fardeau qui n'incitera pas le répondant à y répondre. Le manque de contrôle sur la disposition des questions du questionnaire peut aussi avoir des répercussions sur les réponses.

4.4 Sommaire

Il y a trois principales méthodes de collecte des données : les interviews d'enquête par autodénombrement, les interviews sur place et les interviews téléphoniques. L'enquête par questionnaire postal d'autodénombrement est habituellement la méthode de collecte de données la moins chère. Malheureusement, l'enquête par autodénombrement donne souvent le taux de réponse le plus bas et peut demander le plus de temps, surtout si le questionnaire est envoyé et retourné par la poste.

L'interview sur place donne habituellement le taux de réponse le plus élevé, mais elle peut aussi être la plus chère. Elle est souvent appliquée aux enquêtes qui ont des questions complexes ou élaborées, lorsque l'échantillon demande une visite sur place pour situer et sélectionner les répondants, et en cas de couverture médiocre de la population cible à l'aide du téléphone, ou lorsque le taux d'alphabétisation est faible dans la population.

Les interviews téléphoniques donnent habituellement des taux de réponse moyens, elles coûtent moins cher que les interviews sur place et c'est la méthode de collecte la plus rapide. Elles peuvent être particulièrement avantageuses si la population et l'échantillon sont géographiquement éparpillés, si les interviews sur place coûtent très cher et il serait difficile de les réaliser. Le principal problème des interviews téléphoniques est l'échantillonnage des interviewés par téléphone : les listes de numéros de téléphone sont rapidement périmées (ce qui donne un sous-dénombrement dans la base) et la composition aléatoire est inefficiente.

Toutes ces méthodes de collecte des données peuvent être appliquées sur support papier ou électronique. Le principal avantage des méthodes assistées par ordinateur est que la collecte et la saisie des données sont combinées. Le principal inconvénient des méthodes assistées par ordinateur est l'application informatique qui demande du temps et des sommes considérables.

Voici d'autres méthodes de collecte des données : l'observation directe, la déclaration électronique des données, les données administratives, les méthodes combinées et les enquêtes supplémentaires ou omnibus. L'observation directe peut être précise, mais elle ne peut être appliquée à toutes les données et elle exige souvent le recours à des spécialistes. La déclaration électronique des données est pratique pour

les répondants qui peuvent déclarer leurs données sur support électronique, mais elle demande un travail considérable pour convertir les données du répondant en une mise en forme voulue. Les données administratives peuvent servir comme méthode de collecte des données indirecte pour certaines enquêtes. Cette méthode peut éliminer le fardeau de réponse, réduire énormément les coûts de l'enquête et accélérer la rapidité d'exécution, mais l'organisme statistique doit examiner attentivement les concepts utilisés par les sources administratives et la qualité des données. Une combinaison des méthodes est souvent un bon moyen de diminuer les coûts, d'améliorer les taux de réponse et d'accélérer la rapidité d'exécution. La collaboration à d'autres enquêtes à l'aide d'une enquête supplémentaire ou omnibus est un autre moyen de diminuer les coûts.

Bibliographie

Cox, B.G., D.A. Binder, B.N. Chinnappa, A. Christianson, M.J. Colledge et P.S. Kott, Éds. 1995. *Business Survey Methods.* John Wiley and Sons, New York.

Couper, M.P., R.P. Baker, J. Bethlehem, C.Z.F. Clark, J. Martin, W.L. Nicholls II et J.M. O'Reilly, Éds. 1998. *Computer Assisted Survey Information Collection*. John Wiley and Sons, New York.

Dielman, L. et M.P. Couper. 1995. Data Quality in a CAPI Survey: Keying Errors. *Journal of Official Statistics,* 11: 141-146.

Dillman, D.A. 1978. *Mail and Telephone Surveys: The Total Design Method.* John Wiley and Sons, New York.

Dillman, D.A. 2000. *Mail and Internet Surveys: The Tailored Design Method*. John Wiley and Sons, New York.

Dufour, J., R. Kaushal, C. Clark et J. Bench, eds. 1995. *Converting the Labour Force Survey to Computer-Assisted Interviewing*. Statistics Canada. HSMD-95-009E.

Groves, R.M. 1989. *Survey Errors and Survey Costs*. John Wiley and Sons, New York.

Groves, R.M., P.P. Biemer, L.E. Lyberg, J.T. Massey, W.L. Nicholls et J. Waksberg, Éds. 1988. *Telephone Survey Methodology*. John Wiley and Sons, New York.

Groves, R.M. et R.L. Kahn. 1979. *Surveys by Telephone: A National Comparison with Personal Interviews*. Academic Press, New York.

Kasprzyk, D., G.J. Duncan, G. Kalton et M.P. Singh, Éds. 1989. *Panel Surveys*. John Wiley and Sons, New York.

Lavrakis, P. J. 1987. *Telephone Survey Methods: Sampling, Selection and Supervision*. Applied Social Research Methods Series. 7. Sage Publications, California.

Tableau 2 : Comparaison des méthodes de collecte des données

Méthode	Avantages	Inconvénients
A. Support papier	- elle peut contenir du matériel de référence imprimé pour réduire les erreurs de réponse - un ordinateur n'est pas nécessaire pour la collecte des données - elle peut demander moins de temps pour élaborer les procédures de collecte que les méthodes assistées par ordinateur - pour les enquêtes uniques ou les petites enquêtes, la collecte peut coûter moins cher que les méthodes assistées par ordinateur	- la saisie des données est distincte de la collecte - l'impression des questionnaires peut coûter cher - les questionnaires ne peuvent avoir un cheminement des questions – des instructions « passez à » complexes - la collecte des données demande beaucoup de travail manuel - seulement quelques vérifications manuelles simples du questionnaire sont possibles - l'interview connexe est difficile pour les enquêtes réitérées
A.1 Autodénom-brement	- elle est facile à administrer - elle peut réduire le taux d'erreur parce que le répondant peut consulter des dossiers personnels - méthode habituellement la moins chère, le client a donc les moyens d'avoir un échantillon plus important et d'obtenir une plus grande précision - le questionnaire peut être rempli sans la présence d'un intervieweur, une caractéristique positive pour les questions à caractère délicat - il n'est pas nécessaire d'avoir un grand nombre d'intervieweurs formés	- les erreurs de réponse peuvent augmenter parce que le répondant ne lira probablement pas le matériel de référence - les données peuvent être de moins bonne qualité que dans le cas des méthodes assistées par intervieweur - le questionnaire doit être bien conçu et convivial pour le répondant, et donner des instructions claires pour susciter la participation et diminuer les erreurs de réponse - le questionnaire ne peut être trop long ou complexe - la méthode devrait seulement être appliquée pour des sujets simples et directs, ou à des populations ayant une bonne scolarité - les taux de réponse sont inférieurs à ceux des méthodes assistées par intervieweur (de nombreux suivis peuvent être nécessaires pour améliorer le taux de réponse) - il faut faire le suivi des rejets à la vérification - si les questionnaires sont envoyés par la poste, il faut régler l'affranchissement - les questionnaires envoyés et retournés par la poste peuvent être la méthode de collecte des données la plus lente

Méthode	Avantages	Inconvénients
A2. Assistée par intervieweur	- les taux de réponse sont meilleurs que ceux des enquêtes par autodénombrement parce que l'intervieweur peut susciter l'intérêt du répondant et être sensibilisé à ses préoccupations - l'intervieweur peut améliorer la qualité des données en expliquant les concepts et en aidant à régler les problèmes : il peut diminuer les erreurs de réponses et le nombre de questionnaires répondus en partie seulement - l'intervieweur peut garantir que l'unité exacte fait l'objet de l'enquête - il peut diminuer le fardeau de réponse parce que l'intervieweur peut vérifier immédiatement et faire le suivi des rejets à la vérification - l'intervieweur peut expliquer les méthodes utilisées pour garantir la sécurité et la confidentialité des données - la méthode permet une période de collecte plus souple que celle de l'autodénombrement (davantage d'intervieweurs peuvent être engagés pour accélérer la collecte) - la méthode peut être reliée à des populations ayant des capacités de lecture et d'écriture très limitées	- les interviews peuvent coûter cher : coût de la formation des intervieweurs, rémunération des intervieweurs, logement et transport des intervieweurs sur place – espaces de bureau pour les interviews téléphoniques - il faut avoir du temps pour former les intervieweurs - de bonnes aptitudes à la gestion sont nécessaires pour coordonner toutes les interviews - les erreurs de réponse peuvent augmenter : les intervieweurs de formation médiocre peuvent augmenter les erreurs de réponse, les répondants peuvent hésiter à répondre à des questions à caractère délicat (en particulier si l'intervieweur est engagé à l'échelon local) ou donner simplement des réponses socialement convenables - il peut être difficile d'obtenir une base de sondage et une bonne couverture de tous les numéros de téléphone - l'échantillonnage des interviewés au téléphone est inefficient
A2.1. Interviews sur place	- la méthode offre une interview très personnalisée (plus que celle des interviews téléphoniques) qui permet habituellement d'obtenir des taux de réponse plus élevés que ceux des interviews téléphoniques - l'intervieweur peut inspirer la confiance au répondant en lui montrant ses pièces d'identité officielles - l'intervieweur peut faire des observations directes - il est possible d'administrer un questionnaire plus complexe que celui des interviews téléphoniques et de l'enquête par autodénombrement	- c'est habituellement la méthode de collecte des données la plus chère et les tailles d'échantillon sont donc plus petites - il est difficile d'exercer le contrôle qualitatif des interviews, comparativement aux interviews téléphoniques qui sont plus faciles à surveiller - réussir à communiquer avec les gens à la maison ou au travail peut être difficile et demander beaucoup de temps - il peut être difficile de découvrir et de maintenir en poste les intervieweurs convenablement qualifiés dans tous les domaines de l'enquête - il est difficile de confier la charge de travail des intervieweurs à des collègues moins débordés

Méthode	Avantages	Inconvénients
A2.2. Interviews téléphoniques	- les taux de réponse sont habituellement plus élevés que ceux des enquêtes par autodénombrement - le contrôle qualitatif de l'interview est facile - elles coûtent moins cher que les interviews sur place (il n'est pas nécessaire de payer les déplacements et le logement) - on obtient les réponses plus rapidement (comparativement aux interviews sur place ou d'enquête par autodénombrement) - il est possible de procéder aux interviews dans des régions difficiles ou inaccessibles - elles sont plus anonymes que les interviews sur place : il est possible de poser des questions à caractère délicat - elles coûtent moins cher que les interviews sur place et permettent d'utiliser des échantillons plus importants, si nécessaire	- les taux de réponse sont moins élevés que ceux des interviews sur place - il faut payer l'espace de bureau pour les intervieweurs - il faudra peut-être régler les interurbains - les questionnaires ne peuvent être trop longs ou complexes - des observations directes sont impossibles - des problèmes de confidentialité sont possibles si les lignes téléphoniques sont partagées - l'échantillonnage des interviewés par téléphone est inefficient : le résultat de la composition d'un numéro de téléphone peut ne donner aucune communication, produire une non-réponse ou la communication avec une unité hors de la portée de l'enquête - le résultat peut être biaisé à cause du sous-dénombrement lorsque l'échantillonnage est fait à partir d'une liste administrative de numéros de téléphone - l'échantillonnage par CA a des variables de stratification géographique limitée

Méthode	Avantages	Inconvénients
B. Assistée par ordinateur	- la vérification pendant la collecte peut être automatisée, il est donc possible de régler immédiatement les rejets à la vérification et de diminuer le fardeau de réponse à cause du suivi - la collecte, la vérification et la saisie des données sont intégrées, c.-à-d. plus rapides (temps de réponse plus rapide), efficientes et faciles à surveiller que celles des méthodes sur support papier - les données peuvent être de meilleure qualité - il est possible d'administrer des questionnaires à enchaînement complexe (instructions « passez à ») - il est facile de produire des rapports de gestion (p. ex., sur les taux de réponse) - la collecte peut être moins chère que la collecte sur support papier pour les grandes enquêtes ou les enquêtes réitérées - il est possible de réduire les coûts d'élaboration en adaptant un logiciel élaboré pour une enquête semblable - la protection de la confidentialité des questionnaires remplis est meilleure - ces méthodes sont écologiquement conviviales (moins de papier gaspillé) - les interviews connexes sont possibles pour les enquêtes réitérées	- l'élaboration de l'application informatique peut demander beaucoup de temps et coûter cher - il faut faire l'essai approfondi de l'application informatique - des experts en informatique seront nécessaires pour élaborer – modifier l'application informatique - cette méthode est à la merci des difficultés techniques - les intervieweurs et les répondants doivent savoir comment utiliser l'application informatique - la méthode a des exigences d'infrastructure (p. ex., il faut remettre aux intervieweurs des ordinateurs portatifs) - il faut être en mesure de transmettre les données en toute sécurité d'un ordinateur à l'autre (p. ex., acheminement des données acquises sur place ou au bureau central)
B1. Auto-interview assistée par ordinateur (AIAO)	- méthode souple et pratique pour les répondants qui ont un ordinateur	- les répondants doivent utiliser facilement les ordinateurs et l'application, avoir les logiciels et le matériel informatique nécessaires

Méthode	Avantages	Inconvénients
B2. Interview assistée par ordinateur (IAO)	- il est possible d'automatiser la gestion des interviews (p. ex., automatisation de l'ordonnancement des interviews) - il est plus facile de gérer les interviews que dans le cas de la collecte sur support papier - la méthode peut accentuer la qualité des données : les intervieweurs qui utilisent facilement l'application informatique peuvent réserver plus de temps aux aptitudes interpersonnelles	- il faut déterminer les coûts du matériel informatique pour les intervieweurs - les intervieweurs peuvent avoir besoin d'information supplémentaire sur l'application informatique - il faut considérer les questions de sécurité (p. ex., l'ordinateur peut être volé)
C. Autres méthodes de collecte des données		
C1. Observation directe	- lorsque toutes les données sont observées, il n'y a pas de fardeau de réponse - les données obtenues sont habituellement plus précises que celles des méthodes d'enquête par interview et par autodénombrement	- la méthode peut coûter très cher si des spécialistes sont nécessaires pour prendre des mesures, et les échantillons pourraient donc être assez restreints - la méthode ne peut être appliquée à la plupart des enquêtes - si des mesures sont prises, les participants peuvent les considérer comme un tracas et le taux de participation sera faible
C2. Déclaration électronique des données (DED)	- la méthode est pratique pour les répondants qui ont un ordinateur	- la mise en forme des données des répondants peut varier et le traitement peut donc coûter cher et demander beaucoup de temps
C2.1. Internet	- la collecte et la saisie des données coûtent moins cher - la rapidité d'exécution est à la hausse	- le nombre d'internautes est encore faible

Méthode	Avantages	Inconvénients
C3. Données administratives	- les données sont rapidement disponibles et il est souvent possible d'obtenir des résultats d'enquête rapides et à prix raisonnable - le fardeau de réponse est faible	- l'objectif du programme administratif peut être différent de celui de l'enquête : il faut évaluer l'utilité de la source administrative du point de vue des concepts et des définitions de l'enquête (p. ex., problèmes de population cible et de couverture, périodes de référence, etc.) - l'organisme statistique a peu de contrôle sur la qualité des données - il est difficile ou impossible de faire le suivi des rejets à la vérification - le traitement des données administratives peut demander beaucoup de temps et coûter cher (p. ex., il faudra peut-être modifier le format de l'enregistrement) - l'utilisation des données administratives peut susciter des préoccupations de confidentialité
C4. Méthodes combinées	- taux de réponse amélioré - diminution des erreurs de réponse - collecte des données plus rapide	- la collecte peut être plus complexe et coûter plus cher - la méthode produit des données hétérogènes qui peuvent compliquer le traitement
C5. Enquêtes supplémentaires et omnibus	- les coûts sont partagés entre plusieurs enquêtes	- il faut déterminer les enquêtes appropriées avec lesquelles travailler – les auteurs de l'enquête ne voudront peut-être pas être liés à certaines enquêtes - cette méthode peut être un fardeau pour le répondant et se traduire par des taux de réponse moindres - le manque de contrôle de l'ordre des sections du questionnaire peut avoir des répercussions sur les réponses

PUBLICATIONS ÉLECTRONIQUES
DISPONIBLES À

www.statcan.gc.ca

Chapitre 5 - Conception du questionnaire

5.0 Introduction

Un questionnaire (ou formule) est un groupe ou une séquence de questions conçues pour obtenir d'un répondant de l'information sur un sujet. Les questionnaires comprennent les formules utilisées pour les enquêtes-échantillons et les recensements, ainsi que les formules administratives. Les questionnaires sont au coeur du processus de collecte des données. Ils ont des répercussions importantes sur la qualité des données parce qu'ils constituent le moyen de collecte des données. Ils ont aussi des répercussions sur l'image de marque que l'organisme statistique projette dans le public.

Les questions posées doivent être conformes à l'énoncé des objectifs de l'enquête et permettre la collecte d'information utile pour l'analyse des données. Elles doivent répondre à tous les besoins d'information, mais chaque question devrait avoir une justification explicite pour être inscrite dans le questionnaire. Il faut savoir pourquoi chaque question est posée et à quoi servira l'information. La formulation de la question doit être claire. Les questions doivent être réparties en séquences logiques pour le répondant. Les questions doivent être formulées pour être faciles à comprendre et permettre au répondant d'y répondre précisément. Enfin, le questionnaire devrait être mis à l'essai avant son application, à l'aide d'un essai cognitif, de groupes de discussion, d'un prétest et d'autres méthodes décrites dans ce chapitre.

Un questionnaire bien conçu devrait :
- permettre la collecte des données avec efficience et le résultat devrait comprendre un nombre minimal d'erreurs et de données incohérentes,
- être convivial pour l'intervieweur et le répondant (s'il s'agit d'une enquête assistée par intervieweur),
- diminuer dans l'ensemble le coût et le temps de la collecte des données.

L'objectif de ce chapitre est de donner un aperçu des étapes de la conception d'un questionnaire, y compris l'élaboration et la mise à l'essai des questionnaires. Les principaux types de questions, ouvertes et fermées, sont décrites, ainsi que leurs avantages et inconvénients. On donnera aussi quelques lignes directrices pour l'élaboration des questions. Enfin, les problèmes d'erreur de réponse et de traitement pertinents à la conception d'un questionnaire font l'objet d'un examen.

5.1 Processus de conception du questionnaire

Le processus de conception du questionnaire commence par la formulation des objectifs de l'enquête et des besoins d'information (**Chapitre 2 - Formulation de l'énoncé des objectifs**) et continue avec les étapes suivantes :
- consultation avec les utilisateurs des données et les répondants,
- examen des questionnaires précédents,
- version provisoire du questionnaire,
- examen et révision du questionnaire,
- mise à l'essai et révision du questionnaire,
- touche finale apportée au questionnaire.

5.1.1 Consultation avec les utilisateurs des données et les répondants

Le processus de consultation avec les utilisateurs des données commence lors de la formulation des objectifs de l'enquête au cours de la phase de planification et continue pendant la conception et

l'élaboration du questionnaire. Cette consultation approfondie est particulièrement importante pour les grandes enquêtes, sinon toutes, d'un organisme statistique. Une compréhension approfondie de l'utilisation des données devrait permettre à l'organisme statistique d'élaborer un questionnaire bien conçu qui répond aux besoins des utilisateurs.

Il faudrait consulter non seulement les utilisateurs des données, mais aussi les répondants, les experts de la matière de l'étude et ceux qui ont procédé à des enquêtes semblables auparavant, avant de formuler la version provisoire du questionnaire. Ils devraient pouvoir donner une rétroaction sur le genre d'information que les répondants peuvent fournir et aider à préciser les concepts à étudier. Rencontrer les répondants peut aider à identifier les questions et les préoccupations importantes pour eux et à obtenir des répercussions sur les décisions pertinentes à la matière du questionnaire. Cette intervention peut aussi aider à identifier les expressions et le langage qu'utilisent les répondants pour décrire les concepts de l'enquête, et donner une bonne idée de la façon dont les catégories de questions et réponses devraient être formulées. Ces discussions peuvent se dérouler pendant des consultations approfondies ou en *groupe de discussion* (voir la Section 5.1.5.3).

5.1.2 Examen des questionnaires précédents

D'autres enquêtes sont une bonne source d'information pour l'élaboration d'une enquête. L'examen des questions posées dans d'autres enquêtes sur le même sujet ou un sujet semblable peut être un bon point de départ lorsqu'il faut formuler une question (c.-à-d. rédiger une question). Lorsque l'on souhaite comparer les résultats de différentes enquêtes, il est préférable d'utiliser les mêmes questions. Il faudrait aussi examiner la documentation sur la qualité des données de ces enquêtes pour évaluer l'efficacité du questionnaire (p. ex., les problèmes de rédaction des questions, le fardeau de réponse, les taux de refus, etc.).

5.1.3 Formulation du questionnaire

La prochaine étape est l'élaboration d'une version préliminaire du questionnaire au complet. Étant donné que la conception globale et les objectifs de l'enquête ont des répercussions sur le questionnaire, il faut considérer les volets suivants :

i. Méthode de collecte des données

La méthodologie de collecte des données a une incidence sur la longueur du questionnaire et la formulation des questions. Les questionnaires d'enquête par autodénombrement devraient être moins complexes et plus brefs que ceux des méthodes assistées par intervieweur et ils devraient de préférence être autonomes, c.-à-d. que toute l'information pertinente (p. ex., instructions, information sur les personnes-ressources, exemples) est comprise dans le questionnaire. Dans le cas des méthodes assistées par intervieweur, la formulation d'une question est souvent différente de celle des questionnaires d'enquête par autodénombrement. La question posée de vive voix devrait sembler neutre. Les interviews sur place et les enquêtes par autodénombrement permettent davantage de catégories de réponses que les interviews téléphoniques qui devraient être brèves.

La présentation, l'organisation et la structure d'enregistrement des données seront aussi très différentes d'un questionnaire à l'autre, par exemple, un questionnaire d'enquête par autodénombrement, un questionnaire d'interview téléphonique ou sur place, ou encore un questionnaire papier et crayon ou assisté par ordinateur.

ii. Caractéristiques des répondants

Les caractéristiques des répondants influencent la formulation des questions. Elles peuvent avoir des répercussions sur la terminologie ou la complexité du langage utilisé pour poser les questions. Les questions destinées au grand public devraient être faciles à comprendre pour tous les répondants, mais il est possible, dans une enquête qui cible des professionnels, d'utiliser un langage technique ou professionnel pertinent au travail des répondants.

iii. Fardeau de réponse

Le fardeau de réponse du questionnaire, le temps et l'effort nécessaires pour répondre aux questions et la possibilité que le répondant consulte des dossiers ou d'autres personnes doivent être pris en considération. Il faudrait minimiser le nombre de questions, et chaque question inscrite au questionnaire devrait être justifiable. (Le but de certaines questions peut être de faciliter la compréhension d'une question ultérieure ou elle peut servir à l'évaluation.)

iv. Complexité des données qui font l'objet de la collecte

Une formulation attentive des questions est nécessaire lors de la collecte de données complexes. Des instructions devraient être intégrées aux questions qui couvrent des sujets complexes. Voilà qui aidera l'intervieweur à expliquer les questions, et le répondant, à y répondre précisément.

v. Confidentialité et caractère délicat de l'information

Ces deux points peuvent avoir des répercussions directes sur la formulation des questions. Le questionnaire devrait comprendre des énoncés d'introduction qui précisent comment la confidentialité des données du répondant sera protégée. Il faudrait aussi expliquer à quoi serviront les données, qui y aura accès, la durée de vie utile des données, etc. Si des questions à caractère délicat sont posées (questions qui peuvent mettre certains répondants mal à l'aise), il peut être nécessaire d'appliquer des techniques qui amenuiseront les répercussions de ces questions. Cette mesure accentue la possibilité d'une réponse (voir la section 5.3.8 pour en apprendre davantage).

vi. Traduction

Le questionnaire devrait être traduit dans toutes les langues couramment parlées dans la population cible. Il faut être attentif lors de la traduction de questions formulées dans une autre langue pour tenir compte, non seulement de la langue, mais aussi des différentes coutumes et cultures. Une « rétro-traduction » (la traduction du texte traduit dans la langue d'origine) peut souvent aider à identifier des erreurs.

vii. Comparabilité des résultats avec ceux d'autres enquêtes

Si les résultats de l'enquête sont comparés avec ceux d'autres enquêtes, les questions doivent être rédigées de la même façon. Chaque version de la question doit cerner le point de la même façon et avoir la même signification dans le contexte de la question. Afin de garantir la comparabilité des résultats avec ceux d'autres enquêtes, il faudrait utiliser la même formulation de la question après avoir confirmé la qualité des résultats précédents. Certaines questions peuvent aussi être étroitement liées à celles qui les précèdent immédiatement.

viii. Cohérence

La formulation de la question doit avoir la même signification pour tous les répondants, soit celle que cible l'organisme statistique. Si le questionnaire est traduit dans différentes langues, il est particulièrement important de mettre à l'essai chaque version dans toutes les langues.

ix. Autres éléments

Voici d'autres éléments à considérer lors de la formulation des questions :
- la disponibilité des données voulues,
- la disposition du répondant à répondre,
- la possibilité d'une non-réponse,
- les exigences administratives,
- le genre de questions,
- la formulation de chaque question,
- la présentation du questionnaire,
- les sources de mesure ou d'erreur de réponse,
- le traitement du questionnaire.

Les exigences administratives de l'organisation de l'enquête comprennent les ententes d'échange des données, un énoncé informant les répondants de la confidentialité de leurs réponses, des versions bilingues du questionnaire, etc.

Les questions peuvent être ouvertes ou fermées (les questions fermées donnent des catégories de réponse). Les divers genres de questions sont examinés en détail à la section 5.2. Les lignes directrices appliquées à la formulation des questions de l'enquête sont considérées à la section 5.3. Les sources d'erreur de réponse sont mentionnées à la section 5.4. Les considérations sur la présentation et le traitement du questionnaire sont précisées aux sections 5.5 et 5.6.

5.1.4 Examen et révision du questionnaire

Il est essentiel que le questionnaire soit examiné à l'interne avant la mise à l'essai. Cet examen devrait identifier tous les problèmes évidents du questionnaire, par exemple, les erreurs d'orthographe ou de grammaire, ou la rédaction maladroite. Il est aussi utile à cette étape de demander à des intervenants qui ne sont pas directement engagés dans le projet d'examiner le questionnaire. Ceux-ci peuvent comprendre des experts du domaine à l'étude, des gens qui ont l'expérience de la conception des questionnaires, des intervieweurs ou des membres de la population à l'étude. Ils peuvent souvent faire des commentaires et des suggestions utiles qui susciteront la révision des questions et des catégories de réponse.

5.1.5 Mise à l'essai et révision du questionnaire

Il est important de procéder à la mise à l'essai de toutes les versions (c.-à-d. les versions dans toutes les langues) du questionnaire auprès de répondants « représentatifs » bien avant le début de la collecte des données (c.-à-d. représentatifs de la population cible, peut-être des répondants d'un certain âge, d'un sexe ou l'autre, ou ayant une scolarité en particulier). Il peut être aussi important de faire l'essai du questionnaire auprès de sous-populations en particulier qui peuvent avoir des problèmes avec certaines questions.

Répondre à une question est un processus complexe. Les répondants doivent d'abord comprendre la question. Ils doivent ensuite faire un effort de mémoire ou fouiller des dossiers pour extraire l'information

demandée. Ils doivent aussi réfléchir à la réponse exacte à la question et déterminer s'ils sont disposés à révéler l'information, en tout ou en partie. Ils répondent alors à la question. Chacun de ces processus peut être une source d'erreur. (Tourangeau *et coll.*, 2000)

Les méthodes de mise à l'essai des questions visent à identifier les difficultés et les erreurs possibles. La mise à l'essai permet aussi de déterminer si les instructions sont claires ou si l'ordre des questions a des répercussions sur l'interprétation de ces questions et d'obtenir les impressions des répondants sur la présentation du questionnaire. L'un des avantages de la mise à l'essai du questionnaire est la production d'un questionnaire convivial pour le répondant et l'intervieweur qui facilite la collecte de données précises en une mise en forme propice à la saisie et au codage des données. Enfin, la mise à l'essai aide aussi à minimiser les erreurs et à diminuer le coût et le temps de la collecte, de la saisie et du traitement des données.

Les méthodes appliquées aux mises à l'essai des questionnaires (matière, présentation, etc.) sont habituellement axées sur de petits échantillons subjectifs non probabilistes de répondants tirés de la population cible. Voici les méthodes décrites dans les sections suivantes :
- prétest,
- méthodes cognitives,
- groupes de discussion,
- compte rendu des intervieweurs,
- codage comportemental des interactions entre l'intervieweur et le répondant,
- essai d'échantillons fractionnés,
- essai pilote.

5.1.5.1 Prétest (essai préliminaire)

Le prétest (parfois intitulé essai préliminaire) est facile, le coût est raisonnable, et c'est une étape fondamentale de l'élaboration d'un questionnaire. S'il n'y a pas d'autres mises à l'essai du questionnaire, il faudrait au moins faire un prétest. La taille de l'échantillon du prétest peut varier de 20 à 100 répondants ou plus. Si le principal objectif est de repérer des problèmes de rédaction ou de séquence, très peu d'interviews sont nécessaires. Il faut en faire davantage (de 50 à 100) pour déterminer les catégories de réponse aux questions fermées, à partir des réponses aux questions ouvertes du prétest. Le questionnaire devrait être administré de la même façon que prévu pour la principale enquête (p. ex., assistée par intervieweur ou ordinateur, sur place, au téléphone ou sur support papier). Il faudrait cependant avoir recours à un intervieweur pour la mise à l'essai des questionnaires d'enquête par autodénombrement.

Lors du prétest, le répondant n'est pas informé, il remplit simplement le questionnaire ou répond à l'interview pour refléter la situation lors de la collecte réelle des données. Le prétest indique seulement là où il y a un problème. Sans aller plus loin, il ne détermine pas pourquoi il y a un problème ou comment le corriger. La mise à l'essai non officiel n'identifiera peut-être pas non plus tous les problèmes du questionnaire.

Voici à quoi sert le prétest d'un questionnaire :
- découvrir l'ordre ou la rédaction médiocres des questions,
- repérer les erreurs de présentation ou d'instructions du questionnaire,
- identifier les problèmes d'application logicielle d'un questionnaire assisté par ordinateur,
- déterminer les problèmes éventuels si le répondant ne peut ou ne veut répondre aux questions,
- suggérer des catégories de réponse supplémentaires qui peuvent être codées d'avance dans le questionnaire,

- donner une indication préliminaire de la longueur de l'interview et du taux de réponse (y compris la non-réponse partielle).

5.1.5.2 Méthodes cognitives

Les méthodes cognitives sont particulièrement utiles pour l'essai des questionnaires parce qu'elles sont conçues pour faire enquête sur les étapes du processus de réponse. Les méthodes cognitives donnent les moyens d'examiner les processus de réflexion du répondant lorsqu'il répond aux questions de l'enquête. Les méthodes cognitives aident donc à évaluer la validité des questions et à identifier les sources éventuelles d'erreur de réponse et de non-réponse.

Les interviews cognitives donnent l'occasion d'évaluer le questionnaire du point de vue du répondant. Elles ciblent des points comme la compréhension et les réactions à la formule. Cette mesure permet d'intégrer la perspective du répondant directement dans le processus de conception du questionnaire et d'en arriver à la conception d'un questionnaire convivial pour le répondant parce qu'il est facile à comprendre et à remplir avec précision.

Les interviews cognitives se déroulent souvent en « laboratoire » ou dans une salle munie d'un miroir d'observation. La taille de l'échantillon est relativement minime. De 12 à 15 interviews cognitives seulement peuvent se dérouler, mais parfois jusqu'à 100 et plus, pour mettre à l'essai la version préliminaire d'un questionnaire. Étant donné la taille de l'échantillon relativement minime, une approche itérative est parfois appliquée et des modifications sont apportées au questionnaire après quelques interviews cognitives avant de donner suite à la mise à l'essai.

Voici certaines méthodes cognitives de mise à l'essai :

i. L'observation des répondants

Le répondant est observé pendant qu'il répond au questionnaire. L'observation donne des renseignements sur le comportement du répondant, notamment :
- les sections du questionnaire qu'il lit,
- la séquence de réponse aux questions,
- le répondant se reporte aux instructions ou non,
- le genre de dossiers qu'il examine,
- le répondant consulte quelqu'un ou non,
- le temps qu'il prend à répondre à chaque section,
- les corrections ou modifications qu'il apporte aux réponses.

ii. Les interviews « penser tout haut »

Le répondant est invité à « penser tout haut » lorsqu'il répond aux questions, à faire des commentaires sur chaque question et à expliquer comment il a choisi la réponse en bout de ligne. Ce genre d'interview penser tout haut est intitulé interview simultanée « penser tout haut ». Si le répondant explique son processus de réflexion après coup, pendant une discussion de suivi, l'interview est alors intitulée interview rétrospective « penser tout haut ». Ces deux méthodes sont très utiles pour la mise à l'essai des questionnaires et l'identification des sources éventuelles d'erreur et des améliorations qui peuvent être apportées.

Des techniques particulières, notamment les *questions d'approfondissement, la reformulation* et *la notation de la confiance* sont appliquées pendant les interviews cognitives.

a. Questions d'approfondissement

Les questions d'approfondissement servent à cibler des aspects en particulier du processus de réponse (c.-à-d. compréhension, extraction, réflexion ou réponse). L'intervieweur peut demander, par exemple, comment et pourquoi un répondant a choisi une réponse ou comment il a interprété les concepts, les mots.

b. Reformulation

Le répondant est invité à répéter les instructions ou la question dans ses propres mots, ou à expliquer la signification des termes et des concepts. La reformulation permet de déterminer si un répondant a lu et bien compris les instructions et les questions.

c. Notation de la confiance

Le répondant cote le degré de confiance en la précision de ses réponses. Cette technique révèle à quel point le répondant a eu de la difficulté à formuler une réponse à une question ou s'il a essayé de deviner.

5.1.5.3 Groupes de discussion

Un groupe de discussion considère un sujet sélectionné par les participants choisis dans la population d'intérêt. Le groupe de discussion donne l'occasion de consulter les membres de la population cible, les utilisateurs des données ou les intervieweurs pour intégrer leur point de vue dans le processus de conception du questionnaire. Au cours des premières étapes de l'élaboration du questionnaire, les groupes de discussion peuvent aider à préciser les objectifs de l'enquête et les besoins de données, et identifier les concepts, définitions et questions saillantes de la recherche. Les groupes de discussion servent aussi à la mise à l'essai des questionnaires. On fait appel à eux pour évaluer la compréhension du langage et de la rédaction des questions et des instructions de la part du répondant, ainsi que d'autres formulations et mises en forme des questions.

Un animateur qui connaît bien les techniques d'interview des groupes et l'objectif de la discussion oriente le groupe de discussion. Chaque groupe comprend habituellement de six à douze personnes et la taille optimale est de sept à neuf personnes. Une séance en groupe de discussion demande habituellement deux heures environ. Le groupe de discussion est enregistré sur bande sonore (et parfois sur bande vidéo) que les observateurs peuvent entendre dans une salle contiguë derrière un miroir d'observation. Il est recommandé que ceux qui élaborent le questionnaire observent le groupe de discussion. Les observateurs n'interviennent pas dans la discussion du groupe, mais leurs observations peuvent servir à l'animateur à la fin de la séance du groupe de discussion.

Si le questionnaire de l'enquête par autodénombrement est mis à l'essai, il peut être achevé immédiatement avant la discussion du groupe (si le temps le permet) ou le répondant peut le remplir d'avance et l'apporter à la séance du groupe de discussion. S'il s'agit d'un questionnaire assisté par intervieweur, ce dernier peut l'administrer quelques jours avant la réunion du groupe de discussion.

Lancer la discussion du groupe en demandant aux participants d'exprimer leur réaction au questionnaire dans l'ensemble est une technique utile. Le groupe discute ensuite des questions et problèmes particuliers que suscite le questionnaire. L'animateur du groupe de discussion examine le questionnaire au complet, question par question, ou cible des questions d'intérêt en particulier. L'animateur devrait avoir des aptitudes à approfondir la matière parce que certains participants du groupe de discussion peuvent hésiter à faire des commentaires négatifs, même s'ils sont pertinents. L'animateur devrait aussi donner à chaque membre l'occasion de s'exprimer pour éviter qu'une personne ou deux domine(nt) la discussion. La séance du groupe

de discussion peut être conclue en demandant aux participants de recommander l'amélioration la plus importante, à leur avis, qu'il faudrait apporter au questionnaire.

5.1.5.4 Compte rendu des intervieweurs

Le compte rendu de l'intervieweur se déroule souvent après la séance en groupe de discussion ou pendant les prétests. L'intervieweur discute de l'expérience acquise pendant l'interview des répondants et approfondit ainsi la compréhension des résultats du questionnaire. Sa perspective peut aider à déterminer les améliorations à apporter au questionnaire. L'intervieweur fait habituellement son compte rendu dans un groupe très semblable au groupe de discussion. Autrement, des formules de notation ou des questionnaires de compte rendu peuvent servir à obtenir de l'information sur les problèmes que posent le questionnaire et des suggestions d'amélioration.

5.1.5.5 Codage comportemental des interactions entre l'intervieweur et le répondant

Le codage comportemental peut être fait pendant que l'intervieweur administre le questionnaire. Ce genre de mise à l'essai comprend le codage systématique par un tiers de l'interaction entre l'intervieweur et le répondant. Le tiers cible comment l'intervieweur pose les questions et la réaction du répondant. L'interview de mise à l'essai est souvent enregistrée sur bande sonore et la relation entre l'intervieweur et le répondant est ensuite analysée. Le codage comportemental aide à identifier certains problèmes, par exemple, l'intervieweur n'a pas lu les questions telles qu'elles sont formulées ou le répondant a demandé des précisions. Si le codage comportemental révèle qu'une question pose des difficultés, une mesure corrective peut être justifiée. En général, le codage comportemental ne donne cependant pas d'information sur les raisons du problème ou la solution possible. Un large échantillon est souvent nécessaire pour analyser les résultats du codage comportemental, surtout si le questionnaire comprend de nombreuses instructions « passez à » qui orientent le répondant dans un questionnaire à cheminements variés.

5.1.5.6 Essai d'échantillons fractionnés

Les mises à l'essai d'échantillons fractionnés servent à déterminer les deux « meilleures » versions ou plus d'un questionnaire ou d'une question. La mise à l'essai d'un échantillon fractionné est parfois intitulée expérience du « questionnaire à deux formes » ou du « panel fractionné ». Elle comprend un plan d'échantillonnage expérimental intégré au processus de collecte des données. S'il s'agit d'un seul plan d'échantillonnage fractionné, la moitié de l'échantillon reçoit une version du questionnaire et l'autre moitié, l'autre version.

La mise à l'essai de l'échantillon fractionné permet non seulement de comparer les variations des questions, mais aussi les différentes méthodes de collecte des données pour déterminer la meilleure méthode. Un plan d'échantillonnage probabiliste et des tailles d'échantillons appropriées sont nécessaires pour analyser les différences entre les échantillons.

5.1.5.7 Enquête pilote

Une enquête pilote se déroule pour observer toutes les étapes du processus de l'enquête, y compris l'administration du questionnaire. Une enquête pilote est une « simulation » qui applique la version finale du plan d'enquête à petite échelle du début à la fin, y compris le traitement et l'analyse des données. Elle permet à l'organisme statistique de considérer les résultats du questionnaire pendant toutes les étapes de

l'enquête (collecte, vérification, imputation, traitement, analyse des données, etc.). Le questionnaire est habituellement soumis à des essais approfondis à l'aide des méthodes susmentionnées avant l'enquête pilote.

5.1.6 Touche finale apportée au questionnaire

La conception du questionnaire est un processus itératif : des modifications sont continuellement apportées pendant l'élaboration et la mise à l'essai du questionnaire. Les objectifs et les besoins d'information sont formulés et réévalués, les répondants et les utilisateurs des données sont consultés, la version préliminaire des questions proposées est formulée et mise à l'essai, les questions sont examinées et révisées jusqu'à la formulation de la version finale du questionnaire. Lorsqu'il est décidé qu'il n'y aura pas d'autres modifications apportées au questionnaire, l'étape finale du processus est franchie. La touche finale est alors apportée au questionnaire et il est imprimé ou programmé, selon la méthode de saisie des données appliquée.

5.2 Genres de questions : ouvertes et fermées

Il y a deux genres de questions : ouvertes et fermées. ***Les questions ouvertes ne donnent pas les catégories de réponse au répondant.*** Le répondant donne un chiffre exact ou une réponse à la question dans ses propres mots par écrit dans le cas d'un questionnaire d'enquête par autodénombrement ou l'intervieweur enregistre la réponse intégralement. Une question ouverte devrait comprendre un espace suffisant pour inscrire la réponse.

Voici un exemple de question ouverte :

Quel est le plus important problème au Canada?

_____

_____

_____

Les questions fermées comprennent des catégories de réponse inscrites sous la question. On répond à la question fermée en cochant une case ou en encerclant la réponse exacte dans l'énumération. Les réponses possibles énumérées pour une question sont intitulées catégories de réponse.

Voici un exemple de questions fermées :

Quel est le plus important problème au Canada? (Cochez une réponse seulement)

- ❍ *Chômage*
- ❍ *Économie – récession*
- ❍ *Déficit fédéral*
- ❍ *Impôts*
- ❍ *Unité nationale*
- ❍ *Crime – violence*
- ❍ *Environnement*
- ❍ *Autre*

Une question ouverte permet au répondant d'exprimer une réponse sans l'influence des catégories de réponse inscrites sous une question fermée. Ce choix permet cependant d'interpréter la question de différentes façons. Une question ouverte élargit donc la portée de la question en général et la version fermée donne au répondant des indices sur la manière d'interpréter la question. Une question fermée ramène aussi le répondant à un ensemble de réponses déterminées.

Les questions ouvertes ont plusieurs applications. L'un des avantages est qu'elles donnent au répondant l'occasion de s'exprimer ou d'élaborer. Elles sont importantes lorsqu'il faut examiner une question mal comprise ou très large. Les questions ouvertes sont donc souvent utilisées pendant l'élaboration et la mise à l'essai du questionnaire. Elles sont posées à des groupes de discussion, par exemple, pour obtenir des commentaires et des opinions sur la question posée et pour susciter la discussion. Une question ouverte permet aussi à l'organisme statistique d'obtenir la formulation « naturelle » personnelle du répondant. Elle est importante lors de l'examen de la rédaction d'une question et des catégories de réponse à une question fermée.

Un autre avantage des questions ouvertes est qu'elles peuvent servir à obtenir des données numériques exactes, par exemple, l'âge précis du répondant. Les intervenants des enquêtes auprès des entreprises demandent souvent les sommes exactes des revenus et dépenses déclarés. Les données numériques exactes sont nécessaires pour certaines analyses des données (p. ex., calculer une moyenne ou une médiane).

Les questions ouvertes ont une autre utilité, elles permettent de faire le suivi des questions fermées. Une question fermée peut demander, par exemple :

Avez-vous des suggestions pour améliorer notre service à la clientèle?
❍ *Non*
❍ *Oui*

Si le répondant coche « Oui », une question ouverte de suivi pourrait être la suivante :

Si oui, quelles sont vos suggestions?

_____

_____

_____

Les questions ouvertes comme celle-ci : « Avez-vous des commentaires supplémentaires? » sont souvent posées à la fin des sections de questions ou à la fin du questionnaire. Le répondant a donc l'occasion d'ajouter tout ce qui est pertinent, à son avis, aux questions considérées dans le questionnaire. Certains répondants pourraient vouloir ajouter de l'information supplémentaire pour préciser une réponse. Il est important de prévoir ce genre d'information dans le questionnaire.

Les questions ouvertes dont les réponses sont rédigées (au lieu d'être numériques) ont des inconvénients. Elles sont un fardeau parce que le répondant doit déterminer l'intention de la question et formuler une réponse sans l'aide des catégories de réponse. Dans une enquête par autodénombrement, l'inscription de la réponse demande du temps. Lors d'une enquête assistée par intervieweur, la collecte, la saisie et le codage des données sont un fardeau. Il est souvent difficile pour l'intervieweur de saisir intégralement la réponse du répondant et, après la collecte, toutes les réponses différentes sont habituellement réparties en catégories et un code numérique leur est attribué pour faciliter le traitement et l'analyse des données. Les

questions ouvertes se traduisent donc habituellement par un traitement plus cher, plus exposé aux erreurs et qui demandent plus de temps que les questions fermées.

S'il s'agit de données numériques, Il est plus difficile de répondre à des questions ouvertes qu'à des questions fermées, et la saisie des données est plus difficile et exposée aux erreurs.

Exemples de questions ouvertes qui demandent des réponses rédigées :

Quels produits ou services offre votre entreprise?

Que pensez-vous faire dans cinq ans?

Veuillez faire davantage de commentaires sur les questions ou problèmes considérés dans ce questionnaire.

Exemples de questions ouvertes qui demandent des réponses numériques :

Quelle est votre meilleure estimation du revenu total avant impôts et déductions de tous les membres du ménage et de toutes sources depuis 12 mois?

Depuis combien d'années le propriétaire actuel exploite-t-il l'établissement?

Au cours d'un mois normal, combien de fois les membres de votre ménage utilisent-ils Internet à la maison?

Les questions fermées ont de nombreux avantages dont le plus important est qu'elles sont moins un fardeau pour les répondants, et la collecte et la saisie des données sont plus faciles et moins chères. Le répondant réagit plus rapidement et facilement parce qu'il choisit simplement la catégorie de réponse appropriée au lieu de formuler une réponse et de l'inscrire dans ses propres mots. Il répondra correctement sans doute plus souvent parce que les catégories de réponse indiquent la cible des questions. Il est plus facile d'analyser les données obtenues à l'aide de questions fermées parce que les réponses sont plus cohérentes et déjà regroupées. Si une question est posée dans plusieurs enquêtes, l'utilisation des mêmes catégories de réponse facilite la comparaison entre les enquêtes.

Les questions fermées ont plusieurs inconvénients. Pendant la formulation de la version préliminaire des questions, il faut souvent faire des efforts pour élaborer des catégories de réponse (c.-à-d. que le codage est fait avant la collecte, mais l'activité peut quand même être difficile). Les catégories de réponse doivent être mutuellement exclusives et exhaustives comme dans tout codage. Si les catégories de réponse ne sont pas clairement formulées, le répondant pourrait avoir davantage de problèmes que si la question posée était ouverte. Les questions fermées suscitent une autre préoccupation, à savoir que les catégories de réponse étant énumérées, le répondant peut se sentir obligé de choisir une catégorie de réponse, peu importe s'il ou si elle a formulé une réponse ou a même les connaissances nécessaires pour répondre à la question. Si la question demande une opinion, le répondant peut être obligé de choisir une catégorie qui ne correspond pas à son opinion, ou d'exprimer une opinion lorsqu'en fait, il n'en n'a pas. (Une catégorie « Ne sais pas » ou « Sans objet » est parfois ajoutée pour éviter la situation. Dans le cas d'un questionnaire assisté par intervieweur, il est pratique commune d'ajouter une catégorie de réponse « Refus ».) Autre problème éventuel : les catégories de réponse peuvent simplifier exagérément un point en confinant le répondant à une réponse possible.

Il y a plusieurs genres de questions fermées : les plus fréquemment utilisées sont les questions dichotomiques, à choix multiples, avec réponses à cocher, avec classement et avec échelle d'évaluation. Elles sont décrites ci-dessous.

Le **Chapitre 10 - Traitement** donne davantage d'information sur le codage des questions ouvertes et fermées.

5.2.1 Questions dichotomiques

La question dichotomique est la version la plus simple d'une question fermée. Il s'agit souvent d'une question oui – non et elle sert à répartir les répondants en deux groupes distincts. La question dichotomique permet aussi la sélection pour éviter de demander aux répondants une série de questions qui ne s'appliquent pas à eux. La directive « Passez à la question X » est ensuite inscrite immédiatement après l'une des catégories de réponse et les répondants passent outre à certaines questions. Cette instruction est parfois appelée « aiguillage ».

Par exemple :

Avez-vous fumé des cigarettes hier?
- ❍ *Oui*
- ❍ *Non ----------Passez à la question 14*

5.2.2 Questions à choix multiples et avec réponses à cocher

La *question à choix multiples* demande au répondant de sélectionner *une réponse* dans une liste de choix et la *question avec réponses à cocher* demande au répondant de choisir *au moins une réponse* dans la liste. Déterminer s'il s'agit d'une question à choix multiples ou avec réponses à cocher n'est peut-être pas évident pour le répondant. Il faut donc ajouter des instructions. Remarquez qu'une catégorie « Autre (précisez) » est habituellement ajoutée pour garantir l'exhaustivité de la liste.

Exemple de questions à choix multiples :

De quel genre de logement s'agit-il? (Cochez une réponse seulement.)
- ❍ *Maison individuelle*
- ❍ *Maison jumelée (en parallèle)*
- ❍ *Maison sur jardin, en bande ou en rangée*
- ❍ *Duplex (superposé)*
- ❍ *Immeuble à hauteur restreinte (moins de cinq étages)*
- ❍ *Crime – violence*
- ❍ *Tour d'habitation (cinq étages ou plus)*
- ❍ *Autre (veuillez préciser)*_____

Exemple de question avec réponses à cocher :

Quel genre d'hébergement avez-vous choisi pendant vos vacances? (Cochez toutes les réponses appropriées.)
- ❍ *Hôtel (y compris maison de chambres pour touristes)*
- ❍ *Motel*
- ❍ *Camping ou parc de roulottes*
- ❍ *Résidence d'amis ou de parents*
- ❍ *Cabine ou chalet à louer*
- ❍ *Autre (centre d'hébergement, université, etc.)*

Les catégories de réponse des questions à choix multiples et avec réponses à cocher demandent une formulation attentive. La liste des catégories de réponse devrait être mutuellement exclusive et exhaustive. Les catégories de l'exemple qui suit ne sont pas mutuellement exclusives, elles se chevauchent :

Quel âge avez-vous?
- ❍ *de 20 à 30 ans*
- ❍ *de 30 à 40 ans*
- ❍ *de 40 à 50 ans*
- ❍ *50 ans ou plus*

Un répondant qui a 30, 40 ou 50 ans peut choisir deux catégories de réponse. L'analyse des données serait difficile parce qu'il est impossible de savoir quelle catégorie choisira ce répondant. La liste des catégories de réponse n'est pas exhaustive et c'est un autre problème. Si les moins de 20 ans font partie de la population cible, comment répondront-ils à cette question? Voici un meilleur choix de catégories de réponse :

Quel âge avez-vous?
- ❍ *moins de 20 ans*
- ❍ *de 20 à 29 ans*
- ❍ *de 30 à 39 ans*
- ❍ *de 40 à 49 ans*
- ❍ *50 ans ou plus*

5.2.3 Questions avec classement

La question avec classement est un autre genre de question fermée et elle demande au répondant d'établir l'ordre des catégories de réponse, par exemple :

Voici une liste de certains moyens dont les gens se servent pour trouver un emploi. Veuillez les classer par ordre d'efficacité en inscrivant « 1 » à la méthode qui serait la plus utile, selon vous, « 2 » à la méthode qui serait la plus utile en second lieu, et ainsi de suite.

____ *Envoi de curriculum vitae par la poste*
____ *Annonces dans les journaux ou les revues*
____ *Centres d'emploi du gouvernement*
____ *Vérification auprès d'amis*
____ *Service de placement privé*
____ *Communication directe avec des employeurs*
____ *Autre (veuillez préciser)* _____

Les répondants considèrent souvent que le classement des catégories est un fardeau, surtout si les points à classer sont très différents l'un de l'autre ou si l'interview se déroule au téléphone. Les questions avec classement posent un autre problème : les écarts d'importance des réponses classées sont inconnus et ne sont probablement pas équivalents, c'est-à-dire que l'écart entre 1 et 2 ne peut être considéré comme équivalent à l'écart entre 2 et 3. Voilà qui complique l'analyse des données. Si trois réponses sont inscrites pour être classées, par exemple, le répondant les classera 1, 2 et 3, mais il peut considérer que les deux premières sont très proches et que la troisième est loin derrière. Il est impossible d'obtenir ce genre d'information simplement à partir du classement. Autre inconvénient : le répondant peut attribuer la même cote à deux réponses ou plus. Dans l'exemple ci-dessus, le répondant peut attribuer la cote 1 à la

réponse « Annonces dans les journaux ou les revues » et à « Centres d'emploi du gouvernement ». Les questions avec classement posent une autre difficulté parce que les répondants ne pourront peut-être pas classer tous les choix de la liste. Il peut être raisonnable de prévoir qu'ils en classeront seulement quelques-uns (p. ex., cinq ou moins).

Voici un exemple de question avec classement qui demande au répondant de sélectionner les plus importants éléments et de classer seulement ceux qui sont importants.

> *Veuillez classer les cinq éléments les plus importants qui influencent votre entreprise lorsqu'elle choisit un transporteur. L'information nous aidera à cibler notre attention et nos ressources sur les secteurs qui sont essentiels pour répondre à vos besoins de service.*
>
> *Veuillez classer leur importance en inscrivant le chiffre « 1 » à l'élément le plus important, selon vous, « 2 » à l'élément le plus important en second lieu, et ainsi de suite.*
>
> ___ *Transport sans dommage*
> ___ *Prix*
> ___ *Marketing et représentants des ventes*
> ___ *Représentants du service à la clientèle*
> ___ *Solution rapide des problèmes de service*
> ___ *Traitement des réclamations de marchandise*
> ___ *Uniformité du service*
> ___ *Fréquence du service*
> ___ *Période en transit*
> ___ *Communication rapide des avis de retard de service*
> ___ *Précision de la facturation*
> ___ *Autre (veuillez préciser)*

5.2.4 Questions avec échelle d'évaluation

Les questions avec l'échelle d'évaluation demandent au répondant d'évaluer leur réponse, par exemple :

> *Êtes-vous satisfait de notre service à la clientèle?*
> ❍ *Très satisfait*
> ❍ *Satisfait*
> ❍ *Insatisfait*
> ❍ *Très insatisfait*

La formulation d'une question avec échelle d'évaluation demande plusieurs considérations. Premièrement, combien de catégories devrait avoir l'échelle d'évaluation? Elle pourrait en avoir seulement deux – d'accord, pas d'accord – ou jusqu'à 10, à partir de 1 (sans importance) jusqu'à 10 (extrêmement important).

Deuxièmement, une question se pose, à savoir si l'échelle d'évaluation devrait avoir ou non un choix neutre, par exemple, ni satisfait ni insatisfait. En l'absence d'une possibilité neutre, le répondant doit faire un choix. D'autre part, les répondants ont tendance à choisir la réponse neutre si elle est ajoutée. Il est possible d'ajouter le choix neutre dans un questionnaire assisté par intervieweur, mais sans l'offrir au répondant. Il est alors sélectionné seulement si le répondant l'exprime spontanément.

L'exemple ci-dessus n'offre pas de choix neutre comme celui ci-dessous.

Êtes-vous satisfait de notre service à la clientèle?
- ❍ *Très satisfait*
- ❍ *Satisfait*
- ❍ *Ni satisfait ni insatisfait*
- ❍ *Insatisfait*
- ❍ *Très insatisfait*

Troisièmement, lors de la formulation d'une question avec échelle d'évaluation, il faut considérer l'ajout de la catégorie « Ne sais pas – pas d'opinion » ou « Sans objet », compte tenu de la question posée. Lorsque vous posez une question sur un service en particulier que le répondant n'a peut-être jamais utilisé, par exemple, il faut ajouter la catégorie « Sans objet ».

Dans chacun de ces cas (nombre de catégories de l'échelle d'évaluation, ajout d'un choix neutre, recours à la catégorie « Sans objet »), la solution sera déterminée en tenant compte des objectifs de l'enquête, du point à coter, de la méthode de collecte des données et des préférences de l'organisme statistique.

La question avec classement présentée à la section 5.2.3 sur le choix d'un transporteur peut être reformulée en question avec échelle d'évaluation, comme suit :

Voici un certain nombre d'éléments qui influencent une entreprise lorsqu'elle choisit un transporteur. Certains peuvent être plus importants que d'autres pour votre entreprise. Compte tenu des priorités de votre entreprise, veuillez coter l'importance de chaque élément de 1 à 10, 1 équivalant à la cote « Sans importance » et 10 équivalant à « Extrêmement important ».

___ *Transport sans dommage*
___ *Prix*
___ *Marketing et représentants des ventes*
___ *Représentants du service à la clientèle*
___ *Solution rapide des problèmes de service*
___ *Traitement des réclamations de marchandise*
___ *Uniformité du service*
___ *Fréquence du service*
___ *Période en transit*
___ *Communications rapides des avis de retard de service*
___ *Précision de la facturation*

La version de la question avec échelle d'évaluation demandera plus de temps en interview, mais il est plus facile pour le répondant de coter chaque catégorie de réponse au lieu de les classer. C'est particulièrement vrai pour les interviews téléphoniques.

5.3 Lignes directrices sur la rédaction des questions de l'enquête

La formulation des questions devrait être claire et significative pour les répondants. Les données de l'enquête seront de qualité supérieure si les répondants peuvent facilement comprendre la signification des mots. Ils seront aussi davantage disposés à donner de l'information, et en mesure de le faire, s'ils comprennent clairement la question posée. Il est aussi essentiel que la compréhension des questions de l'enquête de la part des répondants corresponde à l'intention du concepteur du questionnaire.

La formulation d'une question peut donner des résultats faussés et des données d'enquête inexactes si les répondants :
- ne comprennent pas la signification des mots dans une question,
- n'interprètent pas les mots selon l'intention du concepteur,
- ne connaissent pas les concepts véhiculés dans la formulation d'une question.

Les sections suivantes décrivent certaines lignes directrices générales à considérer pour éviter ces problèmes.

5.3.1 La simplicité est de rigueur

Le meilleur moyen de communiquer clairement avec les répondants est d'utiliser des mots simples, quotidiens, et de vérifier si tous les termes sont appropriés pour la population qui fait l'objet de l'enquête. Le langage de la question suivante n'est ni simple ni quotidien.

> *Êtes-vous conscient de la fusion imminente des circonscriptions à proximité de la nouvelle région métropolitaine?*

De nombreux répondants de l'enquête pourraient ne pas connaître ou comprendre la signification des termes et des expressions *fusion imminente, circonscriptions* ou *nouvelle région métropolitaine*.

Il faut toujours considérer les aptitudes linguistiques des répondants lors de l'élaboration des questions. IL est préférable de choisir des mots faciles à comprendre pour tout le monde. Si l'enquête cible une population ayant une scolarité plus poussée, par exemple des avocats, des enseignants ou d'autres professionnels, il est possible d'avoir recours à un langage plus complexe. En bout de ligne, le langage utilisé devrait correspondre à la compréhension moyenne de la population cible.

Les termes techniques ou le jargon spécialisé que les répondants ne connaissent pas sont à éviter. Si ces termes sont nécessaires, cependant, il faudrait ajouter des précisions ou des définitions à l'intention des répondants. Il faut définir les concepts nouveaux ou complexes pour que tous les répondants aient la même compréhension de la question. Les définitions peuvent être ajoutées à la question, aux instructions à l'intention des répondants imprimées ailleurs dans le questionnaire ou à un cahier d'instructions distinct (un cahier distinct a cependant moins de chance d'être lu).

Les termes de la question suivante ne sont ni simples ni communs.

> *Le vaccin antipneumococcique vous a-t-il été administré?*

La majorité des citoyens dans la population en général ne connaissent probablement pas le terme médical *antipneumococcique* et il sera donc difficile de répondre à la question. De nombreuses personnes ne pourront probablement pas donner une réponse précise. Voici une meilleure formulation :

> *Avez-vous été vacciné contre la grippe?*

5.3.2 Définition des acronymes et des abréviations

Les textes techniques et scientifiques sont souvent truffés d'acronymes et d'abréviations, ainsi que d'expressions juridiques et d'entreprise. Il vaut mieux les utiliser dans les contextes où les lecteurs connaissent bien la matière. Lors des enquêtes auprès du grand public, il faudrait les éviter, sauf s'ils sont bien connus. Il sera probablement plus facile de comprendre clairement les questions si elles comprennent

la rédaction complète du mot, du terme ou de l'organisme ciblé, au lieu d'une abréviation. Il faut toujours définir d'abord les abréviations et les acronymes utilisés.

La question suivante comprend un acronyme qui peut semer la confusion chez les répondants.

Savez-vous où est situé le bureau de l'ARAP le plus près?

De nombreux répondants ne sauront pas que l'ARAP est l'acronyme de l'Administration du rétablissement agricole des Prairies.

5.3.3 Vérification de la pertinence des questions

Il est important de faire un effort pour minimiser le fardeau des répondants. Un important moyen à cette fin est de faire en sorte que seules les questions pertinentes soient posées aux répondants. Cette mesure diminue la longueur des interviews, le temps de participation des répondants et les coûts de l'enquête.

La question suivante, par exemple, ne s'applique pas à tous les répondants, seulement à ceux qui ont un emploi.

Quelle est votre occupation actuelle dans la population active?

Cette question devrait suivre une question de sélection conçue pour déterminer si un répondant a un emploi et elle devrait être posée seulement à ceux qui ont indiqué qu'ils en ont un. Même si la question semble anodine, elle pourrait irriter ceux qui n'ont pas d'emploi.

Les concepteurs de questionnaire devraient aussi déterminer si les répondants ont suffisamment de connaissances pour répondre à la question posée. Sinon, ils peuvent choisir de ne pas répondre ou donner une réponse erronée. Peu de citoyens dans le grand public ont des connaissances suffisamment spécialisées, par exemple, pour donner une réponse informée à la question suivante.

L'incinération à 1 600 ^{o}C pendant 30 minutes est-elle suffisante, à votre avis, pour éliminer les biphényles polychlorés?

5.3.4 La précision est de rigueur

La rédaction des questions de l'enquête doit être aussi précise que possible pour garantir que les répondants comprennent exactement ce qu'on attend d'eux. Un processus semblable à celui qui est appliqué pour définir les concepts, présenté au **Chapitre 2 - Formulation de l'énoncé des objectifs**, est appliqué ici à cette fin. Le concepteur du questionnaire doit demander : Qui? Quoi? Où? et Quand? Il faut préciser clairement pour chaque question :

- À qui s'applique-t-elle?
- Quelle information faut-il ajouter à la réponse ou y retrancher?
- Quelles unités doit donner la réponse (p. ex., kg ou lb)?
- La question vise quelle période (Quand?)?

La question suivante peut sembler simple et directe à première vue.

Quel est votre revenu?

À la réflexion cependant, il n'est pas si facile d'y répondre. Premièrement, *votre* désigne qui? Ce n'est pas évident. Il faudrait préciser s'il s'agit-il du revenu personnel du répondant, de celui de la famille ou du ménage. Deuxièmement, pour quelle *période de référence* le répondant devrait-il donner l'information sur le revenu? La semaine dernière, le mois dernier, l'an dernier? Enfin, qu'est-ce que le répondant devrait considérer comme revenu? Le salaire et les traitements seulement? Le salaire et les traitements, y compris les gratifications? Le salaire, les traitements et les revenus d'autres sources? Autre chose?

Voici deux exemples de formulation améliorée de la question (si le terme « ménage » a été défini pour le répondant).

> *Quel a été le revenu total de toute source de votre ménage avant impôt et déductions l'an dernier?*
>
> *Quel a été le revenu total de votre ménage avant déductions l'an dernier? Ajoutez les revenus tirés des traitements, des salaires et de toute autre source.*

La question suivante illustre le problème possible lorsque la formulation d'une question n'est pas suffisamment précise. On a présenté au répondant une bouteille de boisson à l'orange avant qu'il réponde à cette question d'une enquête (*Poursuite au civil 47LL* (1945), U.S. D.C. N.J., U.S. c. 88 cas – boisson à l'orange Bireley).

> *Combien de jus d'orange contient cette boisson à votre avis?*
> Voici des exemples de nombreuses réponses différentes possibles :
>
> - ❍ *une orange, un peu d'eau et de sucre*
> - ❍ *25 % de jus d'orange et 75 % d'eau gazéifiée*
> - ❍ *jus d'une demi-douzaine d'oranges*
> - ❍ *trois onces de jus d'orange*
> - ❍ *concentration intégrale*
> - ❍ *un quart de tasse de jus d'orange*
> - ❍ *aucun*
> - ❍ *très peu*
> - ❍ *un quart de jus d'orange*
> - ❍ *très peu de jus d'orange, sinon aucun*
> - ❍ *ne sais pas*
> - ❍ *pas beaucoup*
> - ❍ *de trois à quatre onces de jus d'orange*
> - ❍ *une chopine*
> - ❍ *en majeure partie*
> - ❍ *environ un verre et demi*

Voici des formulations plus précises de la question sur le jus d'orange :

> *Cette bouteille contient 300 ml d'une boisson. Combien de mililitres de jus d'orange contient-elle à votre avis? __ ml*
>
> *Cette boisson contient quel pourcentage de jus d'orange à votre avis? ___ %*
>
> *Quelle proportion de cette boisson – un quart, une demie, trois quarts, ou laquelle – est du jus d'orange, à votre avis? ___*

Chacune de ces questions demande une réponse en unités particulières : mililitres, pourcentage, fraction. L'organisme statistique qui pose des questions ainsi formulées obtiendra davantage de réponses en unités mentionnées dans la question.

5.3.5 Les questions à deux volets

Une question à deux volets est en fait une question qui en pose deux. Elle couvre plus d'un concept en général, par exemple :

> *Prévoyez-vous laisser votre automobile à la maison et emprunter l'autobus pour aller au travail l'année prochaine?*

Certaines personnes auront de la difficulté à répondre à cette question parce que leur situation personnelle ne correspond peut-être pas simplement à une réponse par oui ou non. Un répondant peut prévoir, notamment,

- d'utiliser parfois l'automobile et d'emprunter l'autobus à d'autres occasions,
- de toujours laisser l'automobile à la maison et d'aller au travail à bicyclette,
- d'aller au travail en automobile, mais parfois à bicyclette,
- de toujours laisser l'automobile à la maison et de se rendre au travail par d'autres moyens,
- d'aller au travail en automobile parfois et d'emprunter autrement divers moyens,
- de choisir une autre combinaison.

La question est réellement double : *Prévoyez-vous laisser l'automobile à la maison l'année prochaine?* et *Prévoyez-vous emprunter l'autobus pour aller au travail l'année prochaine?* La meilleure solution peut être de formuler deux questions.

Les concepteurs de questionnaire devraient examiner toutes les questions qui contiennent les mots *et* et *ou* pour vérifier si elles pourraient semer la confusion chez les répondants. Il serait bon d'examiner l'objectif de ces questions pour déterminer si une question unique est appropriée ou s'il vaudrait mieux :

- formuler au moins deux questions :
- mettre en évidence les principaux mots dans la question,
- ajouter des instructions pour préciser,
- donner des exemples,
- poser seulement les questions pertinentes aux objectifs de l'enquête.

Ceci dit, il est important de savoir que les questions qui contiennent les mots *et* et *ou* ne sont pas nécessairement toutes des questions à deux volets, par exemple :

> *Quelle est la première langue que vous avez apprise et que vous comprenez toujours?*

L'objectif de cette question est de déterminer, parmi les langues que comprend le répondant, celle qu'il a apprise en premier. La réponse pertinente est la langue qui répond aux deux conditions de la question. Voilà qui peut sembler évident pour le concepteur du questionnaire, mais certains répondants pourraient hésiter à répondre. Il serait bon de donner des instructions avec des exemples pour aider le répondant à comprendre ce qu'on lui demande, et d'insister sur le mot *et* dans la question, par exemple :

> *Quelle est la langue que vous avez apprise en premier **et** que vous comprenez toujours?*
>
> *(Instructions au répondant : Cette question est posée pour déterminer la langue qui répond aux deux conditions, la langue **que vous avez apprise en premier** et **que vous comprenez toujours**. Une personne peut avoir appris le chinois d'abord, mais ne plus le comprendre parce qu'elle a immigré très jeune au Canada. Le chinois serait donc une réponse inexacte parce qu'elle ne répond pas aux deux conditions de la question. La deuxième langue apprise était l'anglais et la personne le comprend toujours. Dans ce cas, la réponse exacte à la question est l'anglais, langue que le répondant a appris en premier lieu et qu'il comprend toujours.)*

5.3.6 Les questions suggestives

Une question suggestive ou insidieuse suggère une certaine réponse ou incite le répondant à en choisir une en particulier. Autrement dit, la formulation de la question a des répercussions sur les réponses. Les questions suggestives peuvent fausser les réponses et avoir des répercussions sur les résultats de l'enquête.

Question suggestive :

> *Veuillez préciser si vous êtes d'accord avec l'énoncé suivant, si vous n'êtes pas d'accord ou si vous n'avez aucune opinion : « Le tourisme est avantageux pour le comté de Northumberland et il faudrait donc en faire la promotion ».*

Question neutre :

> *Veuillez préciser si vous êtes d'accord avec l'énoncé suivant, si vous n'êtes pas d'accord ou si vous n'avez aucune opinion : « Il faudrait faire la promotion du tourisme pour le comté de Northumberland ».*

Les questions d'enquête devraient être formulées pour que toutes les possibilités soient évidentes pour le répondant. Autrement, la question pourrait être suggestive et avoir des répercussions négatives sur les résultats de l'enquête. Il y a une seule réponse possible à la question suivante (Payne, 1951).

> *Pensez-vous que la majorité des entreprises de fabrication qui mettent à pied des travailleurs pendant les périodes creuses devraient prendre des dispositions pour éviter les mises à pied et donner du travail régulier pendant toute l'année?*
> ❍ *Oui*
> ❍ *Non*
> ❍ *Aucune opinion*

Résultats

63 % *Oui, les entreprises peuvent éviter les mises à pied*
22 % *Non, les entreprises ne peuvent éviter les mises à pied*
15 % *Aucune opinion*

La seule possibilité offerte aux répondants dans cette question est de préciser, à leur avis, si les entreprises *peuvent prendre des dispositions pour éviter les mises à pied*. Lorsqu'il y a une seule possibilité, les répondants ont souvent tendance à en convenir. Dans cet exemple, 63 % des répondants sont d'avis que les *entreprises peuvent éviter les mises à pied*, et c'est la seule option présentée dans la question. Voici une autre formulation possible de la même question.

> *Pensez-vous que la majorité des entreprises de fabrication qui mettent à pied des travailleurs pendant les périodes creuses pourraient prendre des dispositions pour éviter les mises à pied et donner aux employés du travail régulier pendant toute l'année, ou pensez-vous que les mises à pied sont inévitables?*
> ❍ *Oui, les entreprises peuvent éviter les mises à pied*
> ❍ *Non, les mises à pied son inévitables*
> ❍ *Aucune opinion*

Résultats

35 % *Oui, les entreprises peuvent éviter les mises à pied*

41 % Non, les mises à pied sont inévitables
24 % Aucune opinion

La question comprend deux possibilités évidentes : *les entreprises peuvent éviter les mises à pied* et *les mises à pied sont inévitables*. Les résultats de cette question sont mieux répartis que ceux de la question précédente entre *oui, non* et *aucune opinion.*

La présentation d'autres réponses possibles à la question incite davantage les gens, en théorie, à réfléchir à la réponse avant de répondre et la réponse est donc plus fiable.

5.3.7 Les négations doubles

Il faudrait éviter les structures de phrase qui contiennent des négations doubles parce que le répondant ne saura pas s'il est d'accord ou pas. Voici un exemple :

Seriez-vous pour ou contre l'interdiction de la vente d'alcool dans les dépanneurs?

Le répondant devra déterminer, pour répondre à la question, que s'il est *pour* l'interdiction de la vente d'alcool dans les dépanneurs, il est *contre* l'autorisation. De même, s'il est *contre* l'interdiction de la vente, il est donc *pour* l'autorisation.

La question est difficile parce qu'elle comprend une négation double : *contre* et *interdiction* sont deux négations. Les questions formulées à l'aide d'une négation double sèment souvent la confusion chez les répondants qui, à leur insu, peuvent donner une réponse qui contredit leurs convictions. Il vaut mieux reformuler la question qui devrait contenir une seule négation. Voici une version plus claire de la question :

Seriez-vous pour ou contre l'autorisation de la vente d'alcool dans les dépanneurs?

5.3.8 Les répercussions des questions à caractère délicat

Les questions personnelles, menaçantes ou à caractère délicat, de l'avis du répondant, peuvent donner une *réponse biaisée socialement convenable*. Les répondants ont tendance à choisir la réponse la plus favorable pour l'estime de soi, ou qui convient aux normes sociales, au lieu d'exprimer une conviction ou de révéler la vérité. Le résultat possible est une sous-déclaration des caractéristiques ou comportements mesurés.

Les questions suivantes, par exemple, peuvent donner des réponses biaisées socialement convenables :

Y a-t-il eu une période où vous n'avez pas été en mesure de garantir la subsistance de votre famille?

Avez-vous déjà conduit un véhicule automobile sous l'influence de l'alcool?

Quel est votre revenu?

Combien pesez-vous?

Combien de fois avez-vous participé à des groupes de discussion sur Internet le mois dernier?

Avez-vous déjà considéré le suicide?

Il est mentionné au **Chapitre 4 - Méthodes de collecte des données** que certaines méthodes (c'est-à-dire les questionnaires d'enquête par autodénombrement et les enquêtes téléphoniques) sont plus anonymes que d'autres et les questions à caractère délicat sont donc moins menaçantes pour les répondants. Si un intervieweur administre le questionnaire, les questions à caractère délicat, en particulier, ne devraient pas être posées à un répondant en présence d'autres personnes.

La formulation prudente peut aussi aider à diminuer les répercussions de questions à caractère délicat sur les réponses de l'enquête. Il y a plusieurs techniques à appliquer pour poser une question à caractère délicat de façon moins menaçante. Une approche à appliquer avant de poser la question est de suggérer que le comportement à caractère délicat n'est pas inhabituel. Certaines expressions, notamment *de nombreuses personnes* ou *la majorité des gens*, peuvent aider à poser la question. Si cette technique est appliquée, il faut éviter les biais (c.-à-d. que la question ne devrait pas inciter le répondant à déclarer un comportement qu'il n'a jamais eu). Poser des questions préliminaires est une autre technique qui permet d'en arriver à poser la question à caractère délicat après un certain nombre de questions pertinentes à caractère moins délicat. Une troisième technique est le recours à une question fermée ayant un éventail de catégories de réponses. Dans le cas des renseignements personnels, c'est-à-dire l'âge, le revenu ou la fréquence du comportement indésirable notamment, le répondant peut être mieux disposé à répondre à la question si un éventail de réponses est ajouté. Voici un exemple :

Quel a été votre revenu total avant déductions l'an dernier? (Ajoutez les revenus tirés des traitements, des salaires et de toute autre source.)

- *moins de 20 000 $*
- *de 20 000 $ à 39 999 $*
- *de 40 000 $ à 59 999 $*
- *de 60 000 $ à 79 999 $*
- *de 80 000 $ à 99 999 $*
- *100 000 $ ou plus*

5.3.9 La lisibilité des questions

Les questions de l'enquête devraient être aussi concises que possible et en langage quotidien pour que la population cible n'ait pas de problèmes de compréhension. Le questionnaire devrait être rédigé à la deuxième personne (*vous*) pour que les répondants le considèrent moins froid, plus personnel, et il faudrait respecter les règles de grammaire.

Le plus important test est de vérifier la réaction lorsque les questions sont lues à haute voix. Elles devraient sembler naturelles, avoir un ton de dialogue et être faciles à suivre pour celui qui écoute. La question suivante ne respecte pas cette ligne directrice.

Quelle cote attribueriez-vous à l'utilité de la prestation de l'information sur les caractéristiques psychologiques et sociologiques de la transition, notamment, l'accès au programme informatisé d'orientation professionnelle interactive qu'offre le bureau régional du ministère aux employés qui prennent leur retraite, lorsqu'il est disponible et conformément à la décision de l'agent du personnel régional?

Cette question est trop longue, le langage est complexe, la construction est compliquée, elle semble rigide et bureaucratique, et il est donc difficile de la comprendre et d'y répondre à cause de ces caractéristiques.

5.4 Erreur de réponse

Au **Chapitre 3 - Introduction au plan d'enquête**, l'une des sources d'erreur non due à l'échantillonnage qui a été considérée était l'erreur de mesure qui est la ***différence entre la réponse enregistrée à une question et la « vraie » valeur***. Dans la documentation sur la conception du questionnaire, cette erreur est plus souvent intitulée erreur de réponse. Le questionnaire étant un moyen de collecte des données, il est donc une source importante d'erreurs de réponse. Il est donc essentiel de concevoir le questionnaire et de le mettre à l'essai pour minimiser ces erreurs.

5.4.1 Sources d'erreur de réponse

Les erreurs de réponse sont possibles n'importe où dans le processus d'enregistrement des questions et réponses. Les erreurs peuvent être attribuées au questionnaire, au répondant, à l'intervieweur, à la méthode de collecte des données ou à l'outil de mesure (dans le cas d'une enquête avec mesure directe).

Les sources d'erreur de réponse due au questionnaire ont déjà été mentionnées aux sections précédentes. Les questions fermées, par exemple, peuvent inciter le répondant à choisir une réponse, peu importe s'il a une opinion ou non, ou s'il a même les connaissances suffisantes pour répondre à la question, et les réponses biaisées socialement convenables peuvent être un problème dans le cas des questions à caractère délicat. Toute question mal formulée peut être mal interprétée. Voici en général les explications des erreurs de réponse occasionnées par le questionnaire :

- le genre de question (ouverte ou fermée),
- la formulation de la question,
- la longueur du questionnaire (peut fatiguer le répondant),
- la présentation du questionnaire (p. ex., les instructions « Passez à » compliquées peuvent occasionner des erreurs, en particulier dans les questionnaires sur support papier) (voir la Section 5.5),
- le traitement du questionnaire (voirla Section 5.6).

Le répondant peut aussi avoir de la difficulté à se remémorer des comportements ou des événements antérieurs. Cette source d'erreur de réponse est intitulée *erreur de mémorisation*. L'une des erreurs de mémorisation est *l'erreur de mémoire*, c'est-à-dire que le répondant ne se souvient pas de tous les événements qui se sont déroulés au cours de la période de référence. Le résultat est une sous-déclaration des comportements ou des événements. La situation inverse est aussi possible. Le répondant peut déclarer des activités qui se sont déroulées hors de la période de référence pensant qu'elles en faisaient partie. Cette source d'erreur est intitulée *erreur de télescopage* et le résultat est habituellement une surdéclaration des comportements manifestés ou des événements. La situation s'explique ainsi : le répondant à tendance à déclarer que des comportements se sont manifestés ou des événements ont eu lieu plus récemment que ce n'est le cas en réalité. Il s'agit de *télescopage en aval*. Le répondant peut déclarer à l'occasion que des comportements se sont manifestés ou des événements ont eu lieu plus longtemps auparavant que ce n'est le cas en réalité. Cette erreur est intitulée *télescopage en amont*. En général, plus la période de référence est longue, plus grande est la perte de mémoire (et ainsi, la possibilité d'erreurs de mémoire). Les périodes de référence plus brèves ont cependant tendance à augmenter les erreurs de télescopage.

Les enquêtes réitérées peuvent poser ce qu'on appelle un *problème de concordance* lorsqu'un nombre particulièrement important de changements sont déclarés à la lisière de deux périodes de référence comparativement au nombre de changements pendant la période de référence. La situation peut être corrigée à l'aide de l'interview connexe.

Voici des exemples de questions qui exigent que le répondant se souvienne d'événements ou de comportements antérieurs :

Combien de fois avez-vous visité le médecin depuis 12 mois?

Quelles revues avez-vous lues le mois dernier?

Quelles émissions de télévision avez-vous écoutées la semaine dernière?

Les intervieweurs peuvent aussi être une source d'erreur de réponse. Chaque intervieweur doit poser la question de la même manière à chaque interview. S'il y a plusieurs interviews et si un intervieweur modifie la formulation d'une question, la signification de la question peut alors changer. Les intervieweurs peuvent aussi faire erreur lorsqu'ils enregistrent la réponse, par négligence ou délibérément (convaincus que le répondant aurait dû répondre différemment), ou en interprétant mal la réponse. Dans les enquêtes avec mesure directe, l'intervieweur peut mesurer la caractéristique (p. ex., tension artérielle) et faire erreur. L'intervieweur, compte tenu de sa réaction aux réponses, peut aussi influencer le comportement du répondant. Si l'intervieweur exprime son étonnement, par exemple, lorsque le répondant précise combien il dépense en vêtements, celui-ci peut déclarer des montants moindres pour les autres questions sur les dépenses.

5.4.2 Techniques de réduction des erreurs de réponse

Il est possible d'identifier les sources d'erreur de réponse et d'appliquer des techniques pour réduire les répercussions de ce genre d'erreurs.

La longueur des questions peut avoir des répercussions sur les erreurs de réponse. Les questionnaires couvrent souvent divers sujets. Si l'intervieweur administre le questionnaire, il est difficile pour le répondant de prévoir la question suivante. Le recours à des questions plus longues, mais quand même précises, simples et claires, est une technique qui aide le répondant à cibler un nouveau sujet. Une version plus longue d'une question donne davantage de temps au répondant pour formuler une réponse. La recherche suggère qu'une question plus longue peut inciter le répondant à s'exprimer davantage, ce qui peut raviver des souvenirs. Le répondant peut aussi avoir davantage de temps pour réfléchir et donner une réponse plus complète.

Question brève :

Quels problèmes de santé avez-vous eus l'an dernier?

Longue question :

La question suivante porte sur les problèmes de santé l'an dernier. Nous posons la question à chacun dans l'enquête. Quels problèmes de santé avez-vous eus l'an dernier?

Afin de réduire les erreurs de réponse des intervieweurs, ils devraient être bien formés et des procédures de contrôle qualitatif, notamment des techniques de réinterview, devraient être appliquées, pour identifier les problèmes et donner une nouvelle formation aux intervieweurs, au besoin.

Il faut faire tous les efforts possibles pour produire un questionnaire bien conçu, selon la description dans ce chapitre, afin de réduire les erreurs de réponse que peut susciter le questionnaire.

Si des problèmes de mémoire sont repérés dans un questionnaire, les techniques suivantes peuvent être appliquées, en tout ou en partie :

i. La période de référence peut être abrégée s'il est déterminé que le répondant a de la difficulté à se remémorer tous les événements qui se sont déroulés pendant cette période.

Ce problème est possible quand les occurrences sont fréquentes. Si la question demande le nombre de visites du répondant chez le médecin l'an dernier, par exemple, il peut être difficile de se souvenir de chaque occurrence s'il a visité souvent le médecin. Si la période de référence est plus courte, les réponses peuvent être plus précises. Il faut cependant éviter une période de référence trop brève parce que le nombre d'événements déclarés serait insuffisant. La longueur optimale de la période de référence peut être déterminée pendant l'évaluation du questionnaire.

ii. Un calendrier ou des points de repère comme les congés fériés peuvent aussi aider à minimiser les erreurs de mémoire.

iii. Le rappel borné est une technique de diminution des erreurs de télescopage.

Les répondants sont interviewés au début et à la fin de la période de référence. Les événements identifiés à la première interview peuvent être retranchés s'ils sont déclarés de nouveau pendant la deuxième interview.

iv. L'interview connexe est aussi un moyen de diminuer les erreurs de réponse dans les enquêtes réitérées.

Au cours de l'interview connexe, l'information que le répondant a donnée pendant un cycle précédent de l'enquête est disponible pour les cycles ultérieurs. Cette mesure peut aider le répondant à situer les événements dans la période de référence voulue et l'empêcher de déclarer des événements mentionnés auparavant.

v. Si le répondant a de la difficulté à déclarer un événement avec précision, il peut être possible de consulter des dossiers.

Si la question demande au répondant de déclarer son revenu l'an dernier, par exemple, il pourrait confirmer sa réponse en consultant sa déclaration de revenus. Le répondant peut aussi tenir des dossiers dans d'autres situations. Certaines personnes ont des dossiers des dépenses du ménage, notamment, les factures mensuelles de téléphone ou les reçus d'achat d'essence. La consultation des dossiers pour diminuer les erreurs de mémoire est probablement davantage appliquée au questionnaire de l'enquête par autodénombrement.

vi. Un autre moyen utile pour les questionnaires de l'enquête par autodénombrement est le journal.

Lorsqu'il est important d'obtenir de l'information détaillée sur une période prolongée, le répondant peut utiliser un journal pour entrer les événements à mesure. Le journal a tendance à servir aux enquêtes sur les dépenses des ménages, la consommation des aliments, l'emploi du temps, l'écoute de la télévision et de la radio.

Dans le cas des questionnaires assistés par intervieweur, d'autres techniques peuvent aider le participant à répondre précisément aux questions. Si vous demandez au répondant de déclarer les aliments consommés sur une période de 24 heures, il peut être difficile d'indiquer les portions. L'intervieweur peut avoir

recours à des moyens visuels qui indiqueraient la taille des diverses portions et le répondant pourrait sélectionner celle qui convient.

D'autres points du plan d'enquête, notamment le délai d'exécution de la collecte des données, peuvent aussi améliorer la mémoire, par exemple, l'ordonnancement d'une enquête sur les revenus en avril.

5.5 Présentation du questionnaire

Au genre de questions et à leur formulation s'ajoutent l'ordre des questions, les énoncés de transition, les instructions et la mise en forme du questionnaire qui sont aussi des éléments importants pour créer un questionnaire de qualité.

5.5.1 Ordonnancement des questions

L'ordre des questions devrait être conçu pour maintenir l'intérêt du répondant et l'inciter à remplir le questionnaire ou à répondre à l'interview. La séquence des questions devrait être logique pour le répondant et faciliter le rappel à la mémoire. Les questions devraient couler doucement de l'une à l'autre. Il faudrait regrouper les questions sur un même sujet.

i. Introduction

L'introduction à l'enquête, que lit le répondant ou qui lui est lue, est très importante parce qu'elle donne le ton à tout le questionnaire. L'introduction du questionnaire devrait :

- donner le titre ou le sujet de l'enquête,
- identifier le commanditaire de l'enquête,
- exprimer l'objectif de l'enquête,
- demander la collaboration du répondant,
- expliquer pourquoi il est important de remplir le questionnaire,
- garantir que le répondant comprend clairement la valeur de ses renseignements,
- souligner comment seront utilisées les données de l'enquête,
- préciser comment le répondant peut avoir accès aux résultats de l'enquête,
- indiquer que les réponses seront confidentielles et ajouter toute entente d'échange de données avec d'autres organismes statistiques, ministères, clients, etc.,
- donner l'adresse et la date de retour pour le questionnaire d'enquête envoyé par la poste.

ii. Questions d'entrée en matière

Les questions d'entrée en matière sont importantes pour inciter le répondant à participer à l'enquête. La première question devrait porter directement sur l'objectif de l'enquête et cibler tous les répondants, autrement, le répondant remettra en question la pertinence de l'enquête. Les premières questions devraient aussi être faciles à répondre. Commencer par une question ouverte qui demande une réponse détaillée peut donner une non-réponse si le questionnaire est considéré comme un fardeau trop lourd pour y donner suite.

iii. Répartition des questions à caractère délicat

Il faudrait considérer attentivement où intégrer les questions à caractère délicat. Si elles sont posées trop tôt, le répondant peut hésiter à y répondre, mais si elles sont posées à la fin d'un long questionnaire, la fatigue du répondant peut avoir des répercussions sur la qualité des réponses. Il faudrait donc poser des

questions à caractère délicat au moment où le répondant est probablement le plus à l'aise pour y répondre et lorsqu'elles sont les plus significatives dans le contexte des autres questions. Les questions à caractère délicat sur la santé, par exemple, devraient être posées à la section où sont posées les autres questions pertinentes à la santé.

iv. Répartition des questions démographiques et de classification

Ces renseignements sont souvent utilisés à des fins de regroupement pour analyser les données et faire des comparaisons entre des enquêtes. Il faudrait expliquer pourquoi ce genre de questions est posée, par exemple, « les quelques questions suivantes aideront à comparer l'information sur votre santé à celle d'autres personnes ayant des antécédents semblables. » Dans le cas des enquêtes sur les ménages et d'autres enquêtes sociales, l'information démographique est reportée à l'occasion à la fin du questionnaire.

5.5.2 Énoncés de transition

Les énoncés de transition des questionnaires servent à présenter des sections de questions connexes. Ils sont importants dans les questionnaires assistés par intervieweur parce qu'ils indiquent au répondant qu'un nouveau sujet sera considéré, par exemple :

> *Partie A - Nous voulons d'abord obtenir des renseignements généraux sur votre exploitation agricole.*
>
> *Partie B - Nous voulons maintenant obtenir de l'information sur votre superficie en culture l'an dernier.*
>
> *Partie C - Les questions suivantes portent sur les déclarations de revenus de votre exploitation agricole l'an dernier.*

5.5.3 Instructions

Le questionnaire assisté par intervieweur ou celui de l'enquête par autodénombrement devrait comprendre des instructions claires, brèves et faciles à trouver. Ces instructions peuvent être inscrites directement au-dessus des questions ciblées, au début du questionnaire, dans un guide distinct qui accompagne les questions, dans un encart, etc.

Les instructions de l'exemple suivant sont ajoutées en caractères gras à la deuxième question.

> *Vous avez travaillé pour qui?*
>
> *De quel genre d'entreprise, d'industrie ou de service s'agissait-il? Donnez une description complète.* ***Fabrication de boîtes en carton, par exemple, voirie, vente de chaussures au détail, etc.***

Si les instructions sont entrées ailleurs dans le questionnaire, le répondant ou l'intervieweur doit savoir où les trouver. La question pourrait, par exemple, préciser au répondant que les instructions sont dans un guide de référence. Les instructions sont parfois ajoutées au début du questionnaire ou au début d'une section du questionnaire, par exemple :

Nota : Les questions suivantes ciblent votre travail ou votre entreprise la semaine dernière. Si vous n'aviez pas de travail ou d'entreprise la semaine dernière, répondez en tenant compte de l'emploi qui a duré le plus longtemps depuis le 1er janvier. Si vous avez eu plus d'un emploi la semaine dernière, répondez selon l'emploi où vous avez travaillé pendant le plus grand nombre d'heures.

Dans l'exemple précédent, les directives sont inscrites avant les questions posées et elles peuvent être présentées dans un style de caractères différent de celui des questions.

Les définitions devraient être inscrites au début du questionnaire si elles sont pertinentes à l'ensemble des questions, autrement, elles peuvent être ajoutées à certaines questions en particulier, au besoin. L'utilisation des caractères gras met l'accent sur les points importants, par exemple les périodes de référence ou de déclaration, et le répondant réfléchira probablement alors davantage en tenant compte de la période de référence de la question. S'il est nécessaire de préciser des points en particulier à inclure ou à exclure, il vaut mieux ajouter ces remarques aux questions directement, et non dans les instructions distinctes, par exemple :

Combien de pièces y a-t-il dans ce logement?
- Comptez la cuisine, les chambres à coucher, les pièces habitables au grenier ou au sous-sol, etc.

L'an dernier, un membre actif de cette exploitation agricole a-t-il été atteint d'une lésion liée aux activités agricoles qui a demandé l'attention médicale d'un professionnel de la santé (médecin, infirmière, etc.) ou qui a occasionné une perte de temps de travail?
- Comptez seulement les lésions des membres actifs de cette exploitation agricole.
- N'inscrivez pas les problèmes de santé chroniques.

Quelle est la superficie totale des grandes cultures ciblées pour la récolte cette année, même si elle a été cultivée ou ensemencée au cours d'une année précédente?
- Comptez toutes les grandes cultures, peu importe si la superficie vous appartient, si elle est louée ou si vous l'avez louée à bail.
- Comptez toutes les terres qui seront ensemencées, même si ce n'est pas déjà fait.
- Déclarez les secteurs seulement une fois, même si plus d'une culture sera récoltée cette année.

Combien de semaines par année travaillez-vous habituellement à ce poste? Veuillez compter les congés annuels et autres congés payés.

Les instructions « Passez à » devraient être clairement indiquées dans les questionnaires sur support papier. Des flèches en gras bien situées devraient orienter le répondant ou l'intervieweur vers la question appropriée suivante. Les instructions « Passez à » devraient être clairement liées à la case de réponse pertinente (p. ex., à l'aide de lignes tracées directement vers la case ou le cercle de réponse). Enfin, il faudrait minimiser les instructions « Passez à » des questionnaires d'enquête par autodénombrement.

5.5.4 Considérations sur la mise en forme

Il y a de nombreuses considérations à ne pas oublier lors de l'organisation des mots imprimés sur support papier ou affichés à l'écran. Il faudrait maintenir l'uniformité du style et de la police de caractères des questions, instructions, entêtes et énoncés de transition. Le recours à des polices et styles de caractères différents pour les questions et les instructions permet au répondant ou à l'intervieweur d'identifier facilement les questions. Les titres et entêtes de section ont habituellement une police de caractères plus

larges que celle des questions et des catégories de réponse. Il faudrait énumérer consécutivement les questions d'un bout à l'autre du questionnaire. Des nombres, titres ou lettres peuvent indiquer les sections. Les codes d'entrée des données imprimés dans le questionnaire ou affichés à l'écran devraient être clairement distincts des questions ou de la numérotation des questions.

Il serait bon d'inscrire un titre ou une entête à chaque section du questionnaire, par exemple :

INFORMATION AUX RÉPONDANTS

SECTION 1 : Information générale
SECTION 2 : Déclaration des revenus
SECTION 3 : Dépenses d'immobilisations
SECTION 4 : Population active
SECTION 5 : Commentaires

Il faut considérer toutes les caractéristiques du questionnaire pertinentes à sa présentation. La couverture avant d'un questionnaire d'enquête par autodénombrement est extrêmement importante parce qu'elle doit attirer l'attention du répondant. Il faut prendre des décisions sur le genre de papier et la taille du papier utilisé pour le questionnaire.

La couleur du questionnaire peut avoir plusieurs utilités. Différentes versions du questionnaire (p. ex., selon la langue) peuvent être imprimées sur du papier de couleurs variées. Si le questionnaire est imprimé sur papier couleur, les cases de réponse sont blanches ou d'un ton plus pâle de la même couleur. Voilà qui aide le répondant ou l'intervieweur à déterminer correctement où répondre à chaque question.

Le recours à des cases de réponse aux questions ouvertes et à des cercles de réponse pour les questions fermées est une convention qui aide aussi le répondant ou l'intervieweur. Il est plus facile de déterminer où entrer la réponse à l'aide de cette convention. Le cercle des catégories de réponse aux questions fermées devrait être disposé uniformément avant ou après la réponse. Des graphiques peuvent servir à améliorer le questionnaire. Les graphiques, s'ils sont appropriés, peuvent aider à indiquer les sujets de la section, mais ils ne devraient pas empêcher de remplir le questionnaire.

Dans le cas des applications d'interviews assistées par ordinateur, Statistique Canada a élaboré des normes pour présenter une interface commune à tous les intervieweurs et réduire les coûts de développement, de mise à l'essai et de formation. Ces normes s'appliquent à certains points, notamment,

- l'utilisation de couleurs ou du noir et blanc,
- les polices de caractères,
- les clés de fonction,
- les clés de navigation,
- les écrans de question standard,
- l'interface Windows de Microsoft.

Ces normes sont données en détail dans *Screen Display and Functionality Standards for Social Survey Full BLAISE Applications* (2001) – Normes relatives aux affichages et aux fonctions complètes des applications BLAISE pour les enquêtes sociales.

5.6 Considérations sur le traitement lors de la conception du questionnaire

Le traitement est la mise en forme convenable des réponses de l'enquête obtenues pendant la collecte des données aux fins de la totalisation et de l'analyse des données. Il comprend toutes les activités de traitement des données après la collecte et avant l'estimation. Certaines activités, c'est-à-dire la saisie, la vérification et le codage des données, peuvent être faites pendant la collecte des données à l'aide d'une application assistée par ordinateur pour rationaliser le traitement.

Il faudrait considérer les tâches de traitement pendant la conception et l'élaboration du questionnaire. Le programme de codage devrait être élaboré en même temps que la formulation des questions. Il faudrait imprimer les codes des questions fermées sur le questionnaire sur support papier. Il faudrait aussi considérer la saisie des données lorsque les décisions sont prises sur la présentation du questionnaire sur support papier.

La présentation du questionnaire a des répercussions sur la facilité de la saisie des données des questionnaires sur support papier. L'inscription uniforme de codes numériques après des catégories de réponse et l'alignement des questions en colonnes facilitent la saisie des données. Toutes les étapes de traitement pertinentes au questionnaire (saisie des données, codage, etc.) devraient être mises à l'essai pour garantir l'efficacité du questionnaire aux fins de ces opérations.

Il faudrait considérer les répercussions de la formulation des questions sur la saisie des données. Chaque questionnaire devrait comprendre un numéro d'identification unique pour faciliter la vérification de la saisie des données. Il est parfois nécessaire de revenir au questionnaire original pour déterminer si l'information a été saisie correctement. Il faudrait saisir les données des questionnaires sur support papier le plus rapidement possible après les avoir reçus. Cette mesure permet la mise en œuvre de systèmes utilisés pour vérifier si l'information entrée au fichier correspond à celle du questionnaire.

Le **Chapitre 10 - Traitement** donne davantage de détails à ce sujet.

5.7 Sommaire

La conception et l'élaboration d'un questionnaire ont été considérées dans ce chapitre. La première étape est la formulation des objectifs de l'enquête. Les répondants et les utilisateurs des données sont ensuite consultés et les questionnaires d'enquêtes semblables font l'objet d'un examen. Vient ensuite la formulation de la version préliminaire du questionnaire qui doit être mise à l'essai et révisée soigneusement avant d'y apporter la touche finale. La mise à l'essai peut comprendre le prétest, la mise à l'essai cognitif, les groupes de discussion, les comptes rendus des intervieweurs, le codage comportemental, les mises à l'essai d'échantillons fractionnés et un essai pilote.

Il y a deux genres de questions : fermées ou ouvertes. Les questions fermées peuvent être des questions dichotomiques, à choix multiples, avec classement ou avec échelle d'évaluation. Les questions ouvertes permettent l'expression personnelle, mais elles peuvent être un fardeau, demander du temps et être difficiles à analyser. Les questions fermées sont habituellement un fardeau moindre pour le répondant, et la collecte et la saisie des données coûtent moins cher et sont plus faciles. Un choix médiocre de catégories de réponse peut cependant occasionner l'erreur de réponse.

Il faudrait respecter les lignes directrices suivantes lors de la formulation d'un questionnaire d'enquête :
- être simple (la simplicité est de rigueur),
- définir les acronymes et les abréviations,

- vérifier si les questions sont pertinentes,
- être précis (la précision est de rigueur),
- éviter les questions à deux volets,
- éviter les questions suggestives,
- éviter les négations doubles,
- amenuiser les répercussions des questions à caractère délicat,
- vérifier s'il est facile de lire les questions.

Le questionnaire devrait être conçu pour minimiser les erreurs de réponse possibles. La présentation du questionnaire est aussi importante. L'introduction et la répartition séquentielle des questions peuvent susciter ou réprimer la participation des répondants. Il faudrait utiliser des énoncés de transition présentant les nouveaux sujets, et les instructions au répondant ou à l'intervieweur devraient être claires, brèves et faciles à trouver. Il faudrait évaluer la mise en forme générale du questionnaire pour en déterminer les répercussions sur le répondant et l'intervieweur : police de caractères, entête de section, couleur du questionnaire, mise en forme des catégories de réponse, etc. Enfin, il faudrait considérer le traitement du questionnaire : il devrait être conçu pour faciliter la collecte et la saisie des données.

Bibliographie

Advertising Research Foundation. 1985. *Focus Groups: Issues and Approaches.* Advertising Research Foundation, Inc., New York, New York. 10022.

American Statistical Association. 1993. How to Conduct Pretesting. *The Section on Survey Research Methods.* American Statistical Association.

Babyak, C., A. Gower, L. Gendron, J. Mulvihill et R.A. Zaroski. 2000. Testing of Questionnaires for Statistics Canada's Unified Enterprise Survey. *Proceedings of the International Conference on Establishment Surveys II*. American Statistical Association.

Biemer, P.P., R.M. Groves, L.E. Lyberg, N.A. Mathiowetz et S. Sudman, Éds. 1991. *Measurement Errors in Surveys.* John Wiley and Sons, New York.

Bishop, G.F. 1987. Experiments with the Middle Response Alternative in Survey Questions. *Public Opinion Quarterly,* 51: 220-232.

Bureau of the Census. *Pretesting Policy and Options: Demographic Surveys at the Census Bureau*. U.S. Department of Commerce, Washington, D.C.

Carlson, L.T., J.L. Preston et D.K. French. 1993. Using Focus Groups to Identify User Needs and Data Availability. *Proceedings of the International Conference on Establishment Surveys.* American Statistical Association. 300-308.

Converse, J.M. et S. Presser. 1986. *Survey Questions: Handcrafting the Standardized Questionnaire.* Sage University Paper Series on Quantitative Applications in the Social Sciences. 07-063. Sage Publications, Thousand Oaks, California.

Couper, M.P. 2001. Web Surveys. *Public Opinion Quarterly,* 64(4): 464-494.

Desvousges, W.H. et J.H. Frey. 1989. Integrating Focus Groups and Surveys: Examples from Environmental Risk Studies. *Journal of Official Statistics,* 5(4): 349-363.

Dillman, D.A. 1978. *Mail and Telephone Surveys: The Total Design Method.* John Wiley and Sons, New York.

Dillman, D.A., M.D. Sinclair et J.R. Clark. 1993. Effects of Questionnaire Length, Respondent-friendly Design, and a Difficult Question on Response Rates for Occupant-addressed Census Mail Surveys. *Public Opinion Quarterly,* 57(3): 289-304.

Esposito, J.L., P.C. Campanelli, J.M. Rothgeb et A.E. Polivka. 1991. Determining Which Questions are Best: Methodologies for Evaluating Survey Questions. *Proceedings of the Section on Survey Research Methods*, American Statistical Association. 46-57.

Fowler, F.J., Jr. 1995. *Improving Survey Questions: Design and Evaluation.* Applied Social Research Methods Series. 38. Sage Publications, Thousand Oaks, California.

Fowler, F.J., Jr. et T.W. Mangione. 1990. *Standardized Survey Interviewing.* Applied Social Research Methods Series. 18, Sage Publications, Thousand Oaks, California.

Gower, A.R. 1994. Conception des questionnaires d'enquêtes-entreprises. *Téchniques d'enquête,* 20(2): 129-142.

Gower, A.R. 1997. Présentation des questions sous forme séquentielle, matricielle, de feuillet unique et de livret pour le questionnaire du recensement au Canada. *Comptes-rendus du Symposium 97 de Statistique Canada: nouvelles directions pour les enquêtes et les recensements,* Statistique Canada. 251-256.

Gower, A.R. et G. Haarsma. 1997. A Comparison of Two Methods in a Test of the Canadian Census Questionnaire: Think-aloud Interviews vs. Focus Groups. *Proceedings of the Minimum Standards in Questionnaire Testing Workshop.* Statistics Sweden. 35-37.

Gower, A.R., B. Bélanger et M.-J. Williams. 1998. Using Focus Groups with Respondents and Interviewers to Evaluate the Questionnaire and Interviewing Procedures after the Survey Has Taken Place. *Proceedings of the 1998 Joint Statistical Meetings, Section on Survey Research Methods.* American Statistical Association. 404-409.

Gower, A.R., K. McClure, A. Paletta et M.-J. Williams. 1999. When to Use Focus Groups versus Cognitive Interviews in the Development and Testing of Questionnaires: The Statistics Canada Experience. *Proceedings: Quality Issues in Question Testing (QUEST 99).* Office for National Statistics, England. 51-66.

Jabine, T., E. Loftus, M. Straf, J. Tanur, et R. Tourangeau, Éds. *Cognitive Aspects of Survey Methodology: Building a Bridge Between Disciplines.* National Academy of Science, Washington, DC.

Kalton, G. and H. Schuman. 1982. The Effect of the Question on Survey Responses: A Review. *Journal of the Royal Statistical Society,* 145(1): 42-73.

Krueger, R.A. 1997. *Analyzing and Reporting Focus Group Results.* Focus Group Kit. 6. Sage Publications, Thousand Oaks, California.

Krueger, R.A. 1997. *Developing Questions for Focus Groups.* Focus Group Kit. 3. Sage Publications, Thousand Oaks, California.

Morgan, D.L. 1997. *Planning Focus Groups.* Focus Group Kit. 2. Sage Publications, Thousand Oaks, California.

Morgan, D.L. 1997. *The Focus Group Guidebook.* Focus Group Kit. 1. Sage Publications, Thousand Oaks, California.

Oppenheim, A.N. 1992. *Questionnaire Design, Interviewing and Attitude Measurement.* Pinter Publishers, London.

Payne, S.L. 1951. *The Art of Asking Questions*, Princeton University Press, Princeton, New Jersey

Platek, R., F.K. Pierre-Pierre et P. Stevens. 1985. *Élaboration et conception des questionnaires d'enquête.* Statistique Canada. 12-519F.

Presser, S. et J. Blair. 1994. Survey Pretesting: Do Different Methods Produce Different Results? *Sociological Methodology,* 24: 73-104.

Statistique Canada. 1994. Politique concernant l'examen et la mise à l'essai des questionnaires. *Manuel des politiques*. 2.8.

Statistique Canada. 1996a. Politique d'information des répondants aux enquêtes, *Manuel des politiques*. 1.1.

Statistics Canada. 2001. Screen Display and Functionality Standards for Social Survey Full BLAISE Applications.

Statistics Canada. 2001. Standard Question Blocks for Social Survey Full BLAISE Applications.

Tourangeau, R., L.J. Rips et K. Rasinski, 2000, *The Psychology of Survey Response*, Cambridge University Press, Cambridge, U.K.

PUBLICATIONS ÉLECTRONIQUES
DISPONIBLES À

www.statcan.gc.ca

Chapitre 6 - Plans d'échantillonnage

6.0 Introduction

Le **Chapitre 3 - Introduction au plan d'enquête** précise qu'au cours de la phase de planification, l'organisme statistique doit déterminer s'il fait un recensement ou une enquête-échantillon. Si la décision est une enquête-échantillon, l'organisme doit donc prévoir comment sélectionner l'échantillon. ***L'échantillonnage est un moyen de sélectionner un sous-ensemble d'unités dans une population aux fins de la collecte de l'information sur ces unités pour formuler des inférences sur l'ensemble de la population.***

Il a deux genres d'échantillonnage : l'échantillonnage probabiliste et non probabiliste. Il faut savoir si des inférences fiables seront faites au sujet de la population pour choisir l'un ou l'autre. Dans l'échantillonnage non probabiliste considéré à la Section 6.1, une méthode subjective de sélection des unités est appliquée à une population. C'est un moyen rapide, facile et bon marché de sélectionner un échantillon. Cependant, s'il veut formuler des inférences au sujet de la population à partir de l'échantillon, l'analyste des données doit supposer que l'échantillon est représentatif de la population. Cette supposition est souvent risquée si l'échantillon est non probabiliste.

L'échantillonnage probabiliste, considéré à la Section 6.2, comprend la sélection d'unités dans une population selon le principe du choix aléatoire ou au hasard. L'échantillonnage probabiliste est plus complexe, demande davantage de temps et coûte habituellement plus cher que l'échantillonnage non probabiliste. Étant donné que les unités de la population sont sélectionnées au hasard, et que la probabilité d'inclusion de chaque unité peut être calculée, il est cependant possible de faire des estimations fiables, ainsi que des estimations de l'erreur d'échantillonnage, et de formuler des inférences au sujet de la population.

Un échantillon probabiliste peut être sélectionné de plusieurs façons différentes. Il faut considérer un certain nombre de facteurs pour choisir le plan, notamment, la base de sondage disponible, les caractéristiques des différences entre les unités de la population (c.-à-d. leur variabilité) et les frais qu'il faudrait engager pour faire enquête sur les membres de la population. Il est possible d'établir un équilibre entre l'erreur d'échantillonnage, les coûts et la rapidité d'exécution en choisissant le plan et la taille de l'échantillon pour une population donnée.

L'objectif de ce chapitre est de présenter différents facteurs à considérer pour déterminer quel plan d'échantillonnage probabiliste est approprié à une enquête en particulier. Le **Chapitre 8 - Calcul de la taille de l'échantillon et répartition** donne des détails sur les facteurs qui ont des répercussions sur la taille de l'échantillon.

6.1 Échantillonnage non probabiliste

L'échantillonnage non probabiliste est un moyen de sélectionner des unités d'une population à l'aide d'une méthode subjective (c.-à-d. non aléatoire). Il n'est pas nécessaire d'avoir une base de sondage complète pour l'échantillonnage non probabiliste qui est donc un moyen rapide, facile et bon marché d'obtenir des données. L'échantillonnage non probabiliste pose un problème : il n'est pas évident qu'il est possible de généraliser et d'appliquer les résultats de l'échantillon à toute la population. La raison de cette constatation est que la sélection d'unités dans une population pour un échantillon non probabiliste peut donner des biais d'importance.

Par exemple, il est courant que l'intervieweur décide subjectivement qui doit être échantillonné. Étant donné que l'intervieweur sélectionnera probablement les membres de la population les plus amicaux ou faciles d'accès, une partie importante de la population n'aura aucune chance d'être sélectionnée et celle-ci sera peut-être systématiquement différente des membres sélectionnés. Non seulement la situation peut biaiser les résultats de l'enquête, mais elle peut aussi diminuer erronément la variabilité apparente de la population à cause d'une tendance à sélectionner des unités « typiques » et d'éliminer les valeurs extrêmes. L'échantillonnage probabiliste évite justement ce genre de biais à cause de la sélection aléatoire des unités (voir la Section 6.2).

Étant donné le biais de sélection et (habituellement) l'absence de base de sondage, la probabilité d'inclusion d'une personne ne peut être calculée pour les échantillons non probabilistes et il est donc impossible de faire des estimations fiables ou des estimations de leur erreur d'échantillonnage. Il faut supposer que l'échantillon est représentatif de la population pour faire des inférences sur celle-ci. Il faut habituellement supposer que les caractéristiques de la population correspondent à un certain modèle ou qu'elles sont également ou aléatoirement réparties dans la population. C'est souvent dangereux à cause de la difficulté d'évaluer si oui ou non ces suppositions sont fondées.

Les études de marché utilisent souvent l'échantillonnage non probabiliste comme mesure de rechange rapide à prix raisonnable, comparativement à l'échantillonnage probabiliste, mais ce n'est pas un substitut valable de l'échantillonnage probabiliste pour les raisons mentionnées ci-dessus. Dans ce cas, pourquoi choisir un échantillonnage non probabiliste? Celui-ci peut être appliqué à des études qui servent :

- d'outil pour donner des idées,
- d'étape préliminaire à l'élaboration d'une enquête par échantillonnage probabiliste,
- d'étape de suivi pour aider à comprendre les résultats d'une enquête par échantillonnage probabiliste.

L'échantillonnage non probabiliste peut donner, par exemple, de l'information importante au cours des premières étapes d'une enquête. Il peut servir à des études diagnostiques ou de recherche pour acquérir un aperçu des attitudes, certitudes, motivations et comportements des gens. L'échantillonnage non probabiliste est parfois la seule option viable; par exemple, l'échantillonnage des bénévoles peut être le seul moyen d'obtenir des données pour des expériences médicales.

L'échantillonnage non probabiliste est souvent utilisé pour sélectionner des personnes pour des groupes de discussion ou des interviews approfondies. Statistique Canada utilise l'échantillonnage non probabiliste, par exemple, pour faire l'essai des questions du Recensement de la population, afin de vérifier si les questions posées et les concepts utilisés sont clairs pour les répondants. Si la matière d'une question est considérée controversée, des sous-populations peuvent aussi être sélectionnées et mises à l'essai. Si ces questions peuvent être formulées de sorte qu'elles soient acceptables pour ces personnes, par l'intermédiaire de groupes de discussion, elles peuvent être acceptables pour tous les membres de la population. (Les groupes de discussion sont étudiés au **Chapitre 5 - Conception du questionnaire**.)

Les études préliminaires sont un autre exemple qui motive l'utilisation de l'échantillonnage non probabiliste. Si une nouvelle enquête est planifiée pour couvrir un domaine très peu connu, des plans d'échantillonnage non probabilistes sont souvent utilisés dans les enquêtes pilotes. Considérons, par exemple, l'industrie relativement nouvelle de la conception des pages Web. Supposons qu'il n'y a pas de renseignements sur le nombre de personnes qui travaillent dans l'industrie, leurs revenus ou d'autres détails de la profession. Une enquête pilote serait planifiée et des questionnaires seraient envoyés à quelques personnes qui conçoivent des pages Web. L'examen des questionnaires retournés peut donner une idée sur leurs revenus et révéler que de nombreux concepteurs de pages Web travaillent à domicile, qu'ils ont uniquement un numéro de téléphone personnel et qu'ils annoncent exclusivement sur Internet.

Voici les **avantages** de l'échantillonnage non probabiliste :

i. Il est rapide et pratique.

Règle générale, les échantillons non probabilistes sont obtenus en peu de temps et l'enquête est rapide : il est très facile de simplement sortir et poser des questions à la première centaine de personnes rencontrées dans la rue.

ii. Il est relativement bon marché.

Il faut habituellement quelques heures seulement du temps d'un intervieweur pour faire ce genre d'enquête. De plus, les échantillons non probabilistes ne sont généralement pas dispersés géographiquement et les frais de déplacement des intervieweurs sont donc minimes.

iii. Une base de sondage n'est pas nécessaire.

iv. Il peut être utile pour les études de recherche et d'élaboration d'enquête.

Voici les **inconvénients** de l'échantillonnage non probabiliste :

i. Il faut avoir des hypothèses solides sur la représentativité de l'échantillon pour formuler des inférences sur la population. Étant donné que tous les échantillons non probabilistes comportent un biais de sélection, il est souvent dangereux de formuler ces hypothèses. Il vaudrait mieux procéder à un échantillonnage probabiliste si des inférences sont nécessaires.

ii. Il est impossible de déterminer la probabilité qu'une unité de la population soit sélectionnée pour l'échantillon, et des estimations fiables et des estimations de l'erreur d'échantillonnage ne peuvent donc être faites.

Les sections suivantes décrivent cinq différents types de méthodes d'échantillonnage non probabilistes : l'échantillonnage à l'aveuglette, l'échantillonnage à participation volontaire, l'échantillonnage au jugé, l'échantillonnage par quotas et l'échantillonnage probabiliste modifié. L'échantillonnage de réseaux ou boule de neige moins souvent utilisé est présenté à la Section 6.3.

6.1.1 Échantillonnage à l'aveuglette

Les unités sont sélectionnées de façon arbitraire, sans idée préconçue, et la planification est minime, sinon nulle. Celui qui fait l'échantillonnage à l'aveuglette présume que la population est homogène : si les unités de la population sont toutes semblables, n'importe quelle unité peut être choisie pour l'échantillon. L'interview de « l'homme de la rue » est un exemple d'échantillonnage à l'aveuglette parce que l'intervieweur choisit n'importe quel passant. Sauf si la population est vraiment homogène, les biais de l'intervieweur et du passant au moment de l'échantillonnage peuvent malheureusement avoir des répercussions sur la sélection.

6.1.2 Échantillonnage à participation volontaire

Cette méthode fait appel à des répondants volontaires. Les volontaires doivent généralement faire l'objet d'un examen pour obtenir un ensemble de caractéristiques qui convient aux objectifs de l'enquête (p. ex.,

les personnes atteintes d'une maladie en particulier). Cette méthode peut être marquée d'un important biais de sélection, mais elle est parfois nécessaire. Pour des raisons de déontologie, on peut faire appel, par exemple, à des volontaires ayant des conditions médicales particulières pour procéder à certaines expériences médicales.

Voici un autre exemple d'échantillonnage à participation volontaire : au cours d'une émission radio ou télédiffusée, une question fait l'objet d'une discussion et les citoyens à l'écoute sont invités à téléphoner pour exprimer leurs opinions. Seuls ceux que le sujet intéresse vraiment d'une façon ou d'une autre ont tendance à répondre. La majorité silencieuse ne répond habituellement pas et nous avons donc un biais de sélection marqué. L'échantillonnage à participation volontaire sert souvent à sélectionner des particuliers pour des groupes de discussion ou des interviews approfondies (c.-à-d. une mise à l'essai qualitative qui exclut la généralisation appliquée à la population complète).

6.1.3 Échantillonnage au jugé

À l'aide de cette méthode, l'échantillonnage est fait en tenant compte des idées préalables sur la composition et le comportement de la population. Un expert qui connaît la population décide quelles unités devraient être choisies. Autrement dit, l'expert sélectionne à dessein ce qui est considéré comme un échantillon représentatif. Les biais du chercheur peuvent marquer l'échantillonnage au jugé qui peut être encore plus biaisé qu'un échantillonnage à l'aveuglette. Étant donné que les idées préconçues du chercheur sont reflétées dans l'échantillon, des biais importants peuvent être intégrés si ces idées préconçues sont inexactes. Il peut cependant être utile aux études de recherche, par exemple, lors de la sélection de personnes pour des groupes de discussion ou des interviews approfondies, afin de vérifier des aspects particuliers d'un questionnaire.

6.1.4 Échantillonnage par quotas

Voilà l'un des échantillonnages non probabilistes les plus communs. L'échantillonnage est fait jusqu'à ce qu'un nombre déterminé d'unités (quotas) soient sélectionnées dans diverses sous-populations. L'échantillonnage par quotas est un moyen d'atteindre les objectifs de taille d'échantillon pour les sous-populations.

Les quotas peuvent être établis selon des proportions de population. S'il y a 100 hommes et 100 femmes dans la population, par exemple, et s'il faut tirer un échantillon de 20 personnes, 10 hommes et 10 femmes peuvent être interviewés. L'échantillonnage par quotas peut être considéré préférable à d'autres formes d'échantillonnage non probabiliste (p. ex., échantillonnage au jugé) parce qu'il faut inclure des membres de sous-populations différentes.

L'échantillonnage par quotas ressemble à l'échantillonnage stratifié parce que des unités semblables sont regroupées (des détails sur l'échantillonnage stratifié sont donnés à la Section 6.2.6). La méthode de sélection des unités est cependant différente. Les unités sont sélectionnées aléatoirement dans l'échantillonnage probabiliste, mais dans l'échantillonnage par quotas, une méthode non aléatoire est appliquée, c'est-à-dire que l'intervieweur décide habituellement qui est ajouté à l'échantillon. Les unités sollicitées qui ne sont pas disposées à participer sont simplement remplacées par d'autres qui le sont, et l'on ignore en fait le biais de non-réponse.

Les études de marché utilisent souvent l'échantillonnage par quotas (en particulier pour les enquêtes au téléphone) au lieu de l'échantillonnage stratifié pour faire enquête auprès de citoyens ayant des profils

socioéconomiques particuliers parce qu'il est relativement meilleur marché que l'échantillonnage stratifié, il est facile à administrer et il a la caractéristique souhaitable de respecter les proportions de la population. Il masque cependant un biais de sélection éventuellement important.

Dans ce cas comme dans tous les autres plans d'échantillonnage non probabiliste il faut présumer que les personnes sélectionnées sont semblables aux autres pour formuler des inférences sur la population. Ces fortes présomptions sont rarement valables.

6.1.5 Échantillonnage probabiliste modifié

L'échantillonnage probabiliste modifié est une combinaison d'échantillonnage probabiliste et non probabiliste. Les premières étapes sont habituellement axées sur l'échantillonnage probabiliste (voir la section suivante). La dernière étape est un échantillon non probabiliste, habituellement un échantillon par quotas. Des secteurs géographiques peuvent être sélectionnés, par exemple, à l'aide d'un plan d'échantillonnage probabiliste et ensuite, dans chaque région, un échantillon de personnes peut être choisi par quotas.

6.2 Échantillonnage probabiliste

L'échantillonnage probabiliste est une méthode qui permet de formuler des inférences sur la population, compte tenu des observations tirées de l'échantillon. Celui-ci devrait être libre de tout biais de sélection pour formuler les inférences. L'échantillonnage probabiliste évite ce biais par la sélection aléatoire d'unités de la population (à l'aide d'un ordinateur ou d'un tableau de nombres aléatoires). Il ne faut pas oublier que le terme aléatoire ne signifie pas arbitraire. En particulier, les intervieweurs ne choisissent pas arbitrairement les répondants parce que leur biais personnel aurait des répercussions sur l'échantillonnage. Le terme aléatoire signifie que la sélection n'est pas biaisée, c'est un tirage au sort. L'échantillonnage probabiliste ne permet pas à l'intervieweur de décider subjectivement qui doit être choisi.

Voici les deux principaux critères de l'échantillonnage probabiliste : la sélection des unités est aléatoire, toutes les unités de la population de l'enquête ont une probabilité d'inclusion différente de zéro dans l'échantillon et il est possible de calculer ces probabilités. Il n'est pas nécessaire que toutes les unités aient la même probabilité d'inclusion et, en fait, dans les enquêtes les plus complexes, la probabilité d'inclusion varie d'une unité à l'autre.

Il y a de nombreux types différents de plans d'échantillonnage probabiliste. Le plus élémentaire est l'échantillonnage aléatoire simple et la complexité des plans s'accentue ensuite pour englober l'échantillonnage systématique, l'échantillonnage avec probabilité proportionnelle à la taille, l'échantillonnage par grappes, l'échantillonnage stratifié, l'échantillonnage à plusieurs degrés, l'échantillonnage à plusieurs phases et l'échantillonnage par répliques. Chacune de ces techniques d'échantillonnage est utile dans différentes situations. Si l'objectif de l'enquête est simplement d'obtenir des estimations de la population en général, et si la stratification serait inappropriée ou impossible, l'échantillonnage aléatoire simple pourrait alors être le meilleur choix. Si le coût de la collecte des données de l'enquête est élevé et si les ressources sont disponibles, l'échantillonnage par grappes est souvent le choix. Si des estimations de sous-populations sont aussi demandées (p. ex., des estimations par province, groupe d'âge ou taille d'entreprise), l'échantillonnage stratifié est habituellement appliqué.

La majorité des plans plus complexes ont recours à l'information auxiliaire de la base de sondage pour améliorer l'échantillonnage. Si la base a été créée à partir d'un recensement précédent ou de données administratives, il peut y avoir une mine de renseignements supplémentaires qui peuvent servir à l'échantillonnage. Dans le cas d'une enquête sur les exploitations agricoles (fermes), par exemple, l'organisme statistique peut avoir la taille de chaque exploitation en hectares tirée du recensement agricole le plus récent. S'il s'agit d'une enquête sur les citoyens, l'information (p. ex., âge, sexe, origine ethnique, etc.) peut être disponible pour chacun dans le plus récent recensement de la population. Lors d'une enquête sur les entreprises, l'organisme statistique peut avoir de l'information administrative, notamment, sur le genre d'industrie (p. ex., détaillant, grossiste, fabricant), le genre d'entreprise (p. ex., magasin d'aliments), le nombre d'employés, etc. L'information auxiliaire améliore l'échantillonnage s'il y a corrélation entre les données auxiliaires et les variables de l'enquête.

Voici le principal **avantage** de l'échantillonnage probabiliste : la sélection de chaque unité est aléatoire, la probabilité d'inclusion de chaque unité peut être calculée, il est possible de faire des estimations fiables et d'estimer l'erreur d'échantillonnage de chaque estimation. On peut donc formuler des inférences sur la population. Un plan d'échantillonnage probabiliste permet en fait souvent d'utiliser un échantillon relativement petit pour formuler des inférences sur une grande population.

Voici les principaux **inconvénients** de l'échantillonnage probabiliste : il est plus difficile, il demande plus de temps et il coûte habituellement plus cher que l'échantillonnage non probabiliste. Les frais de création et d'entretien d'une base de sondage de bonne qualité sont substantiels en général. Étant donné que les échantillons probabilistes ont tendance à être géographiquement répartis plus largement dans la population que les échantillons non probabilistes, les tailles d'échantillon sont habituellement plus grandes, la collecte des données coûte souvent plus chère et sa gestion est plus difficile. Pour un organisme statistique, la capacité de formuler des inférences à partir d'un échantillon probabiliste surpasse habituellement ses inconvénients.

On a vu au **Chapitre 3 - Introduction au plan d'enquête** les qualités d'un bon plan. L'utilisation des données administratives est couverte à l'**Annexe A - Données administratives.**

6.2.1 Efficience statistique

L'échantillonnage aléatoire simple (EAS) est une référence pour l'évaluation de l'efficience d'autres stratégies d'échantillonnage. Voici certaines définitions pour comprendre le concept de l'échantillonnage efficient.

Un paramètre est une caractéristique de la population que le client ou l'utilisateur des données est intéressé à estimer, par exemple, la moyenne, la proportion ou le total de la population. ***Un estimateur est une formule de calcul d'une estimation du paramètre dans l'échantillon et une estimation est la valeur de l'estimateur calculé à l'aide des données de l'échantillon obtenu. La stratégie d'échantillonnage est la combinaison du plan d'échantillonnage et de l'estimateur utilisé.***

Le paramètre d'intérêt peut être, par exemple, la moyenne de la population, $\overline{Y}$, calculée comme suit :

$$\overline{Y} = \sum_{i \in U} \frac{y_i}{N}$$

où y_i est la valeur de la variable y de la i^e unité, U est l'ensemble des unités de la population et il y a N unités dans la population.

Dans le cas d'un EAS dont le taux de réponse est de 100 %, l'estimateur habituel, mais il n'est pas le seul, pour la moyenne de la population est le suivant :

$$\hat{\bar{Y}} = \sum_{i \in S_r} \frac{y_i}{n}$$

où S_r est l'ensemble des répondants de l'échantillon qui comprend n unités. La valeur que prend $\sum_{i \in S_r} \frac{y_i}{n}$ pour un échantillon en particulier est une estimation.

Les estimations calculées à partir d'échantillons différents sont différentes l'une de l'autre. La ***distribution d'échantillonnage d'un estimateur est la répartition de toutes les valeurs différentes que l'estimateur peut avoir pour tous les échantillons possibles du même plan d'échantillonnage de la population***. La stratégie d'échantillonnage détermine donc cette répartition.

Les estimateurs ont certaines caractéristiques souhaitables. L'estimateur devrait, par exemple, être non biaisé ou approximativement non biaisé. ***Un estimateur n'est pas biaisé si l'estimation moyenne, compte tenu de tous les échantillons possibles, est équivalente à la valeur réelle du paramètre.*** La répartition de l'échantillonnage le plus près possible de la moyenne (c.-à-d. que l'erreur d'échantillonnage est minime) est une autre caractéristique souhaitable d'un estimateur. L'erreur d'échantillonnage d'un estimateur est mesurée par sa variance d'échantillonnage déterminée comme fluctuation de sa moyenne calculée en tenant compte de tous les échantillons possibles tirés du plan d'échantillonnage. Un estimateur ayant une variance d'échantillonnage minime est considéré *précis.* La précision augmente quand la variance d'échantillonnage diminue. Il faut noter qu'un estimateur peut être précis et biaisé. L'*exactitude* tient compte à la fois de la variance et du biais; un estimateur exact jouit d'une bonne précision et est peu entaché de biais.

Une stratégie d'échantillonnage est plus efficiente qu'une autre si la variance d'échantillonnage de l'estimateur est plus petite que celle d'une autre stratégie d'échantillonnage. Afin de ne pas semer la confusion au sujet de ce genre d'efficience avec d'autres, par exemple le coût unitaire, cette notion sera donc intitulée efficience statistique. L'efficience statistique est une considération importante si vous comparez divers plans d'échantillonnage possibles parce que les économies peuvent être considérables si un plan peut donner une précision équivalente ou meilleure et si la taille de l'échantillon est plus petite. Les plans d'échantillonnage suivants donnent une comparaison de leur efficience comparativement à l'EAS. Celle-ci est formellement mesurée en calculant l'effet de plan es répercussions du plan dont les détails sont expliqués à la section 7.3.3 du **Chapitre 7 - Estimation**.

On trouvera au **Chapitre 7 - Estimation** davantage de détails sur l'estimation, les facteurs qui ont des répercussions sur la précision et l'estimation de la précision.

6.2.2 Échantillonnage aléatoire simple (EAS)

L'échantillonnage aléatoire simple (EAS) est le point de départ de tout plan d'échantillonnage probabiliste. L'EAS est une méthode de sélection en une étape qui garantit que chaque échantillon possible de taille n a une chance égale d'être sélectionné. Chaque unité de l'échantillon a donc la même probabilité d'inclusion. Cette probabilité, π, est égale à n/N, où N est le nombre d'unités dans la population.

L'échantillonnage peut être fait avec ou sans remise. L'échantillonnage avec remise permet à une unité d'être sélectionnée plus d'une fois. L'échantillonnage sans remise signifie que lorsqu'une unité a été

sélectionnée, elle ne peut l'être de nouveau. L'échantillonnage aléatoire simple avec remise (EASAR) et l'échantillonnage aléatoire simple sans remise (EASSR) sont pratiquement identiques si la taille de l'échantillon est une très petite fraction de la taille de la population parce que la possibilité que la même unité apparaisse plus d'une fois dans l'échantillon est minime. L'échantillonnage sans remise donne généralement des résultats plus précis et est plus pratique du point de vue opérationnel. Aux fins de ce chapitre, l'échantillonnage est supposé être sans remise, sauf avis contraire.

Considérons une population de cinq personnes et supposons qu'un échantillon de trois est sélectionné (EASSR). Étiquetons les personnes de la population 1, 2, 3, 4 et 5 et précisons que la population est la série {1, 2, 3, 4, 5}. Il y a dix échantillons possibles de trois personnes : {1, 2, 3}, {1, 2, 4}, {1, 2, 5}, {1, 3, 4}, {1, 3, 5}, {1, 4, 5}, {2, 3, 4}, {2, 3, 5}, {2, 4, 5} et {3, 4, 5}. Chacun de ces échantillons a une chance égale d'être sélectionné et chaque personne est sélectionnée dans six des dix échantillons possibles, chaque personne a donc une probabilité d'inclusion de $\pi = 6/10 = n/N = 3/5$.

L'organisme statistique qui veut sélectionner un échantillon aléatoire simple a habituellement établi une base de sondage complète (une liste ou une base aréolaire) avant l'échantillonnage. Dans une liste, les unités sont généralement numérotées de 1 à *N*, mais la méthode d'affectation d'un chiffre unique à chaque unité n'est pas importante. Ensuite, *n* unités de la liste sont choisies au hasard à l'aide d'un tableau de nombres aléatoires ou de nombres aléatoires produits par ordinateur et les unités correspondantes forment l'échantillon.

Considérons une enquête auprès des élèves d'une école pour illustrer la technique de l'EASSR. Supposons qu'une liste convenable d'élèves est disponible ou peut être dressée à partir de sources existantes. Cette liste sert de base d'échantillonnage ou de sondage. Supposons maintenant que la liste de la population contient N=1530 élèves dont un échantillon de la taille n=90 est nécessaire. La prochaine étape est de décider comment sélectionner 90 élèves.

La sélection de l'échantillon peut être faite à l'aide d'un tableau de nombres aléatoires (voir le tableau 1). La première étape comprend la sélection d'un nombre à quatre chiffres (parce que c'est le nombre de chiffres de 1530). Commençons l'échantillonnage en sélectionnant un nombre n'importe où dans le tableau et en procédant dans n'importe quelle direction. Les premiers 90 nombres à quatre chiffres qui ne sont pas supérieurs à 1530 sont sélectionnés.

Supposons que la ligne 01 et la colonne 85 - 89 sont sélectionnées au départ. En procédant vers le bas de cette colonne, les nombres aléatoires sélectionnés sont 189, 256, 984, 744, 1441, 617, etc. La sélection continue jusqu'à ce qu'on obtienne 90 nombres différents. Le résultat est un échantillon d'élèves et de nombres correspondants dans la liste de la population. (Étant donné que la méthode considérée est l'EASSR, les nombres qui apparaissent plus d'une fois ne sont pas retenus). Un tableau de nombres aléatoires a été utilisé ci-dessus pour illustrer la sélection manuelle d'un échantillon aléatoire simple, mais en pratique, un programme informatique sélectionnerait les unités au hasard.

Tableau 1 : Extrait d'un tableau de nombres aléatoires

	50-54	**55-59**	**60-64**	**65-69**	**70-74**	**75-79**	**80-84**	**85-89**	**90-94**	**95-99**
00	59311	58030	52098	87024	14194	82848	04190	96574	90464	29065
01	98567	76364	77204	27062	53402	96621	43918	01896	83991	51141
02	10363	97518	51400	98342	24830	61891	27101	37855	06235	33516
03	86852	19558	64432	99612	53537	59798	32803	67708	15297	28612
04	11258	24591	36863	31721	81305	94335	34936	02566	80972	08188
05	95068	84628	35911	33020	70659	80428	39936	31855	34334	64865
06	54463	47437	73804	36239	18739	72824	83671	39892	60518	37092
07	16874	62677	57412	31389	56869	62233	80827	73917	82402	84420
08	92484	63157	76593	03205	84869	72389	96363	52887	01087	66591
09	15669	56689	35682	53256	62300	81872	35213	09840	34471	74441
10	99116	75486	84989	23476	52967	67104	39495	39100	17217	74073
11	15696	10703	65178	90637	63110	17622	53988	71087	84148	11670
12	97720	15369	51269	69620	03388	13699	33423	67453	43269	56720
13	11666	13841	71681	98000	35979	39719	81899	07449	47985	46967
14	71628	73130	78783	75691	41632	09847	61547	18707	85489	69944
15	40501	51089	99943	91843	41995	88931	73631	69361	05375	15417
16	22518	55576	98215	82068	10798	82611	36584	67466	69377	40054
17	75112	30485	62173	02132	14878	92879	22281	16783	86352	00077
18	08327	02671	98191	84342	90813	49268	95441	15496	20168	09271
19	60251	45548	02146	05597	48228	81366	34598	72856	66762	17002
20	57430	82270	10421	00540	43648	75888	66049	21511	47676	33444
21	73528	39559	34434	88596	54086	71693	43132	14414	79949	85193
22	25991	65959	70769	64721	86413	33475	42740	06175	82758	66248
23	78388	16638	09134	59980	63806	48472	39318	35434	24057	74739
24	12477	09965	96657	57994	59439	76330	24596	77515	09577	91871
…	…	…	…	…	…	…	…	…	…	…
45	12900	71775	29845	60774	94924	21810	38636	33717	67598	82521
46	75086	23537	49639	33595	31484	97588	28617	17979	78749	35234
47	99445	51434	29181	09993	38190	42553	68922	52125	91077	40197
48	26075	31671	45386	36583	93459	48599	52022	41330	60650	91321
49	13636	93596	23377	51133	95126	61496	42474	45141	46660	42338

L'EAS a un certain nombre d'**avantages** comparativement à d'autres techniques d'échantillonnage probabiliste, notamment :

i. C'est la technique d'échantillonnage la plus simple.

ii. Il n'est pas nécessaire d'avoir de l'information supplémentaire (auxiliaire) dans la base de sondage pour tirer l'échantillon.

Les seuls renseignements nécessaires sont une liste complète de la population de l'enquête et de l'information permettant d'entrer en communication avec les personnes choisies.

iii. L'élaboration technique n'est pas nécessaire.

La théorie sous-jacente à l'EAS est bien établie et il y a des formules standard pour déterminer la taille de l'échantillon, les estimations de la population et de la variance, et ces formules sont faciles à appliquer.

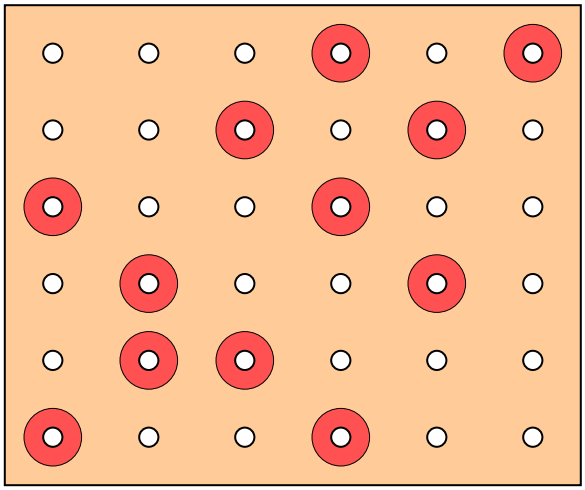
Voici les **inconvénients** de l'EAS :

i. L'information auxiliaire n'est pas utilisée même si cette information existe dans la base de sondage. Les résultats peuvent donc donner des estimations statistiquement moins efficientes que celles d'un autre plan d'échantillonnage.

ii. Il peut coûter cher s'il y a des interviews sur place parce que l'échantillon peut être largement étalé géographiquement.

iii. L'échantillon de l'EAS peut être « mauvais ». Tous les échantillons de taille n ont une chance égale d'être ajoutés à l'échantillon et il est donc possible d'obtenir un échantillon qui n'est pas bien réparti et qui représente peu la population.

Échantillon aléatoire simple (illustré, *n*=12)



6.2.3 Échantillonnage systématique (SYS)

Les unités d'un échantillonnage systématique (SYS) sont sélectionnées à intervalles réguliers dans la population. L'échantillonnage systématique sert parfois si l'organisme statistique veut utiliser un EAS, mais s'il n'y a pas de liste disponible, ou si l'ordre de la liste est approximativement aléatoire, auquel cas, le SYS est encore plus simple à faire que l'EAS. Un intervalle d'échantillonnage et une origine choisie au hasard sont nécessaires. Si une liste est utilisée et si la taille de la population, N, est un multiple de la taille de l'échantillon, n, chaque k^e unité est sélectionnée lorsque l'intervalle k est égal à N/n. Un seul nombre, l'origine r, est choisi au hasard entre 1 et k inclusivement. Les unités sélectionnées sont donc : $r, r+k, r+2k, ..., r+(n-1)k$. Chaque unité, comme dans l'EAS, a une probabilité d'inclusion, π, égale à n/N, mais, contrairement à l'EAS, chaque combinaison de n unités n'a pas une chance égale d'être sélectionnée : dans un SYS, nous pouvons uniquement sélectionner les échantillons dont les unités sont séparées par k. Seulement k échantillons possibles peuvent donc être tirés de la population à l'aide de cette méthode.

Supposons, pour illustrer le SYS, qu'une population contienne N=54 unités et qu'un échantillon de taille n=9 unités soit sectionné. L'intervalle d'échantillonnage serait $k = N/n = 54/9 = 6$. Un nombre aléatoire entre 1 et $k = 6$, disons 2, est ensuite choisi. Les unités de la population sélectionnées pour l'échantillon sont ensuite numérotées : 2, 8, 14, 20, 26, 32, 38, 44 et 50. En présence d'un intervalle d'échantillonnage de 6 et d'une population dont la taille est de 54 unités, il y a seulement six échantillons SYS possibles, mais il y a plus de 25 millions d'échantillons aléatoires simple de taille 6 possibles.

Un avantage de l'échantillonnage systématique est qu'il peut être utilisé lorsqu'il n'y a pas de liste disponible des unités de la population. Une base de sondage peut être établie dans ce cas en choisissant chaque k^e personne jusqu'à la fin de la population.

Le SYS pose un problème : la taille de l'échantillon, n, est connue seulement après la sélection de l'échantillon. Il peut y avoir un autre problème si l'intervalle d'échantillonnage, k, correspond à une certaine périodicité dans la population. Supposons, par exemple, qu'une enquête sur la circulation est faite dans un secteur et qu'une journée seulement de la semaine peut être échantillonné, autrement dit, k est chaque 7^e jour. Les débits de la circulation dans l'enquête seront extrêmement différents si les jours échantillons sont toujours le dimanche au lieu d'être toujours le mardi. Bien entendu, si la période d'échantillonnage est le 5^e jour, chaque jour de la semaine peut alors être visé par l'enquête. Malheureusement, dans la plupart des cas, la périodicité n'est pas connue d'avance.

Si N ne peut être également divisée par n, l'intervalle de l'échantillonnage SYS n'est pas un nombre entier. Dans cette occurrence, k peut être considéré égal au nombre entier le plus près, mais la taille de l'échantillon variera d'un échantillon à l'autre. Supposons, par exemple, que N=55 et n=9, alors k=55/9=6,1. Supposons que k est 6 et r=2, l'échantillon contient donc les unités numérotées : 2, 8, 14, 20, 26, 32, 38, 44 et 50. Si l'origine choisie au hasard est $r=1$ et si chaque sixième unité est sélectionnée, l'échantillon comprend donc les unités : 1, 7, 13, 19, 25, 31, 37, 43, 49 et 55. Dans ce cas, l'échantillon est de taille 10, et non 9. Une autre approche est d'arrondir chaque valeur $r, r+k, r+2k, ..., r+(n-1)k$ au nombre entier le plus près. Dans cette approche, la taille de l'échantillon obtenu est fixe. Supposons de nouveau, par exemple, que N=55 et n=9, c'est-à-dire que k=55/9=6,1. Si r=1, l'échantillon comprend les unités 1, 7, 13, 19, 25, 31, 38, 44 et 50.

D'autre part, si N ne peut être divisé également par n, on pourra alors faire un *échantillonnage systématique circulaire* pour éviter une taille de l'échantillon variable. Dans cette méthode, il est considéré que les unités de la population existent sur un cercle et on y compte « modulo N ». La valeur attribuée à k est égale au nombre entier le plus près de N/n, mais l'origine choisie au hasard, r, peut être entre 1 et N, au lieu de 1 et k (c.-à-d. que la première unité peut être n'importe où dans la liste). Les unités

sélectionnées, comme auparavant, sont : r, $r+k$, $r+2k$, ..., $r+(n-1)k$. Si la j^e unité est telle que $r+(j-1)k > N$, l'unité choisie est donc $r+(j-1)k - N$. Cela veut dire qu'à la fin de la liste, l'échantillonnage continue au début de la liste. L'avantage de la méthode circulaire est que chaque unité a une chance égale d'être dans l'échantillon. À l'aide de l'exemple suivant, supposons, par exemple, que N=55, n=9 et k=6. Une origine choisie au hasard, r, est sélectionnée entre 1 et 55, disons r=42. Les unités de la population sélectionnées sont donc : 42, 48, 54, 5, 11, 17, 23, 29 et 35.

L'échantillonnage SYS a un certain nombre d'**avantages**, selon les circonstances et l'objectif de l'enquête :

i. C'est un substitut de l'EAS lorsqu'il n'y a pas de base de sondage.

ii. Contrairement à l'EAS, l'information auxiliaire de la base de sondage n'est pas nécessaire.

iii. Il peut donner un échantillon mieux réparti que celui de l'EAS (compte tenu de l'intervalle d'échantillonnage et de la méthode de tri de la liste).

iv. C'est une théorie aussi bien établie que celle de l'EAS et les estimations sont faciles à calculer.

v. Il est plus simple que l'EAS parce qu'un seul nombre aléatoire est nécessaire.

Voici les **inconvénients** du SYS :

i. Il peut donner un « mauvais » échantillon si l'intervalle d'échantillonnage correspond à une certaine périodicité dans la population.

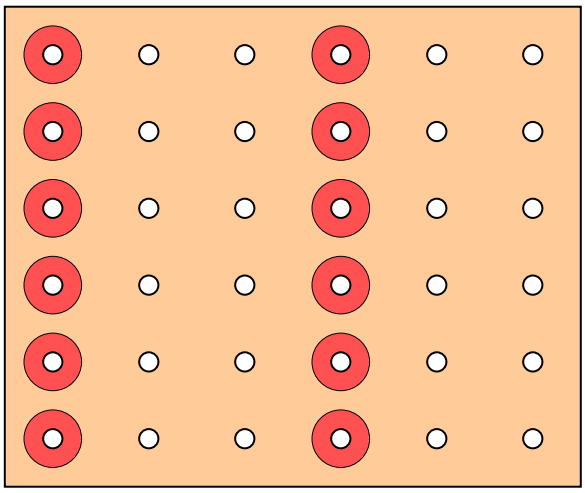
ii. L'information auxiliaire qui peut être disponible dans la base de sondage n'est pas utilisée, comme dans le cas de l'EAS, et le résultat peut être une stratégie d'échantillonnage inefficiente.

iii. La taille de l'échantillon final n'est pas connue d'avance lorsqu'une base de sondage conceptuelle est utilisée.

iv. Il n'a pas d'estimateur non biaisé de la variance d'échantillonnage. L'échantillon systématique est souvent traité comme un échantillon aléatoire simple pour faire l'estimation de variance. C'est approprié seulement lorsque la liste est triée au hasard. (Pour davantage d'information sur l'estimation de la variance pour un échantillon systématique, on consultera Cochran (1977) ou Lohr (1999).)

v. Il peut donner une taille d'échantillon variable si la taille de la population, N, ne peut être divisée également par la taille de l'échantillon voulue, n (mais il est possible d'éviter cela en utilisant le SYS circulaire).

Échantillon systématique (illustré, *n*=12, *N*=36, *k*=3)



L'EAS et le SYS circulaire sont deux plans d'échantillonnage probabiliste à probabilité égale parce que chaque échantillon possible a exactement la même chance d'être sélectionné. Les techniques d'échantillonnage ne donnent pas toutes des probabilités égales. Les plans d'échantillonnage décrits dans les sections suivantes peuvent donner des probabilités inégales. On se rappellera que dans un échantillonnage probabiliste, le critère n'est pas que toutes les unités aient la même probabilité d'inclusion, mais plutôt qu'elles aient une probabilité d'inclusion connue différente de zéro. L'échantillonnage avec probabilités inégales peut souvent améliorer l'efficience statistique de la stratégie d'échantillonnage.

6.2.4 Échantillonnage avec probabilité proportionnelle à la taille (PPT)

L'échantillonnage avec probabilité proportionnelle à la taille (PPT) est une technique qui utilise des données auxiliaires et donne des probabilités d'inclusion inégales. Si les tailles des unités de la population varient et si ces tailles sont connues, l'information peut servir pendant l'échantillonnage pour accentuer l'efficience statistique. L'échantillonnage PPT peut augmenter énormément la précision si les mesures des tailles sont précises et si les variables d'intérêt sont corrélées avec la taille de l'unité. Quand on dispose de mesures de tailles moins précises, il vaut mieux créer des groupements de tailles et procéder à l'échantillonnage stratifié (Section 6.2.6).

Un bon exemple d'une variable de taille de l'échantillonnage PPT est la superficie. L'échantillonnage PPT est souvent utilisé dans les enquêtes sur les exploitations agricoles et la mesure de la taille est la taille de l'exploitation agricole (ferme) en hectares. La taille d'une exploitation agricole peut, bien entendu, augmenter (ou diminuer) si l'exploitant achète ou vend une terre, mais dans la majorité des cas, la taille de l'exploitation agricole est constante d'année en année. De plus, des questions typiques aux enquêtes sur les exploitations agricoles, notamment les revenus, les récoltes, le bétail et les dépenses, sont souvent corrélées avec la propriété foncière. D'autres mesures de taille pour les enquêtes sur les entreprises comprennent le nombre d'employés, les ventes annuelles et le nombre d'emplacements, mais ces variables risquent davantage de changer d'année en année.

Dans un échantillonnage PPT, la taille de l'unité détermine la probabilité d'inclusion. Dans le cas d'une exploitation agricole ayant une superficie de 200 hectares, par exemple, la probabilité d'être sélectionnée est donc deux fois celle d'une exploitation de 100 hectares.

Aux fins de l'illustration, supposons une population de six exploitations agricoles (fermes) et le client est intéressé à estimer les dépenses totales de cette population à l'aide d'un échantillon d'une exploitation. (Un échantillon de taille 1 est utilisé pour illustration, mais en pratique, un organisme statistique sélectionne rarement une seule unité.) Supposons qu'il y a une mesure de taille stable pour chaque exploitation agricole (la taille de l'exploitation en hectares) et, pour illustrer l'efficience accrue comparativement à l'EAS, supposons aussi que les dépenses de chaque exploitation agricole sont connues. (Bien entendu, en réalité, si les dépenses étaient connues, il ne serait pas nécessaire de procéder à l'enquête.)

Considérons la liste d'exploitations agricoles suivante :

Tableau 2 : Valeurs de la population

Unité d'échantillonnage : Ferme	Information auxiliaire de la base : Taille de la ferme en hectares	Variable d'intérêt de l'enquête : Dépenses ($)
1	50	26 000
2	1 000	470 000
3	125	63 800
4	300	145 000
5	500	230 000
6	25	12 500
Total	**2 000**	**947 300**

Le total réel des dépenses est 947 300 $ pour cette population de six fermes. Un échantillon aléatoire simple peut être sélectionné, chaque échantillon contenant une unité et chaque unité ayant une probabilité d'inclusion de 1/6. Six échantillons d'EAS différents de taille n=1 sont possibles. Considérons les résultats obtenus de l'EAS. Il faut invoquer à cette fin certains concepts d'estimation (expliqués en détail au **Chapitre 7 - Estimation**). Dans le cas d'un échantillon de taille un, le total des dépenses pour la population est estimé en multipliant les dépenses de l'unité échantillonnée par le poids de l'unité. Ce poids est le nombre moyen d'unités de la population de l'enquête que l'unité échantillonnée représente et est l'inverse de la probabilité d'inclusion.

Tableau 3 : Échantillons possibles de taille n=1 de l'EAS

Échantillon (Ferme sélectionnée)	Probabilité d'inclusion (π)	Poids ($1/\pi$)	Dépenses ($)	Estimation du total des dépenses de la population ($)
Échantillon 1 (Ferme 1)	1/6	6	26 000	156 000
Échantillon 2 (Ferme 2)	1/6	6	470 000	2 820 000
Échantillon 3 (Ferme 3)	1/6	6	63 800	382 800
Échantillon 4 (Ferme 4)	1/6	6	145 000	870 000
Échantillon 5 (Ferme 5)	1/6	6	230 000	1 380 000
Échantillon 6 (Ferme 6)	1/6	6	12 500	75 000
		Estimation moyenne de l'échantillon		**947 300**

On remarquera la grande variabilité d'échantillonnage dans les estimations de l'EAS qui passe de 75 000 $ à 2,8 millions de dollars. L'échantillonnage PPT peut donner des estimations avec variabilité d'échantillonnage beaucoup plus petite.

Tableau 4 : Échantillons PPT possibles de taille *n*=1

Échantillon (Ferme sélectionnée)	Taille de la ferme	Probabilité d'inclusion (π)	Poids ($1/\pi$)	Dépenses ($)	Estimation du total des dépenses de la population ($)
Échantillon 1 (Ferme 1)	50	50/2 000	2 000/50	26 000	1 040 000
Échantillon 2 (Ferme 2)	1 000	1 000/2 000	2 000/1000	470 000	940 000
Échantillon 3 (Ferme 3)	125	125/2 000	2 000/125	63 800	1 020 800
Échantillon 4 (Ferme 4)	300	300/2 000	2 000/300	145 000	966 667
Échantillon 5 (Ferme 5)	500	500/2 000	2 000/500	230 000	920 000
Échantillon 6 (Ferme 6)	25	25/2 000	2 000/25	12 500	1 000 000
			Estimation moyenne de l'échantillon		**947 300**

La variabilité d'échantillonnage est beaucoup plus faible pour un échantillon PPT. Les estimations tirées des six échantillons possibles passent maintenant d'un seuil de 920 000 $ à un plafond de 1,4 million de dollars seulement, un résultat meilleur que celui de l'EAS. (La probabilité d'inclusion de l'échantillonnage PPT est calculée ainsi : taille de l'exploitation agricole divisée par la taille totale de toutes les exploitations).

Il est supposé y avoir un lien, dans cet exemple, entre les dépenses et la taille de l'exploitation agricole, une supposition valable de toute évidence dans ce cas ou l'échantillonnage PPT n'aurait pas eu autant de succès. En fait, si les variables d'intérêt et la variable de la taille n'avaient pas été corrélées, l'échantillonnage PPT n'aurait peut-être pas été meilleur que l'EAS et pourrait même avoir été pire.

Le principal **avantage** de l'échantillonnage PPT est qu'il peut améliorer l'efficience statistique de la stratégie d'échantillonnage à l'aide de l'information auxiliaire. Le résultat peut être une diminution importante de la variance de l'échantillonnage comparativement à l'EAS ou même à l'échantillonnage stratifié (Section 6.2.6).

Voici les **inconvénients** de l'échantillonnage PPT :

i. Il faut avoir une base de sondage qui contient de l'information auxiliaire à jour de bonne qualité pour toutes les unités de la base qui peuvent servir de mesures de la taille.

ii. Il est inapproprié si les mesures de la taille ne sont pas précises ou stables. Dans ces circonstances, il vaut mieux créer des groupements de tailles et faire un échantillonnage stratifié.

iii. Il n'est pas toujours applicable parce que chaque population n'a pas nécessairement une mesure de la taille stable mise en corrélation avec les principales variables de l'enquête.

iv. Le résultat peut être une stratégie d'échantillonnage statistiquement moins efficiente que celle de l'EAS pour les variables de l'enquête qui ne sont pas corrélées avec les variables de la taille.

v. L'estimation de la variance d'échantillonnage d'une estimation est plus complexe.

vi. La création d'une base de sondage coûte plus cher et est plus complexe que celle de l'EAS ou du SYS parce que la taille de chaque unité dans la population doit être mesurée et sauvegardée.

6.2.4.1 Méthodes d'échantillonnage PPT

Comment obtient-on un échantillon PPT? Il y a de nombreuses méthodes d'échantillonnage PPT, mais trois techniques sont habituellement utilisées sont la méthode aléatoire, la méthode systématique et la méthode systématique aléatoire. (Il est supposé dans ce qui suit que les mesures de la taille sont des valeurs entières.)

i. Méthode aléatoire d'échantillonnage PPT :

- pour chaque unité de la population, faire le calcul cumulatif des mesures de la taille des unités jusqu'à l'unité elle-même comprise,

- déterminer l'étendue correspondant à chaque unité dans la population, c'est-à-dire à partir de la somme cumulative de l'unité précédente (mais sans l'inclure) jusqu'à la somme cumulative de l'unité courante,

- sélectionner un nombre aléatoire entre 0 (si les mesures de taille ne sont pas des nombres entiers) ou 1 (si les mesures de taille sont des nombres entiers) et la taille cumulative totale, et sélectionner l'unité dont l'étendue comprend le nombre aléatoire,

- répéter l'étape précédente jusqu'à ce que *n* unités soient sélectionnées.

Illustrons en utilisant en exemple des exploitations agricoles :

Tableau 5 : Échantillonnage PPT à l'aide de la méthode aléatoire

Ferme	Taille	Taille cumulative	Étendue
1	50	50	1-50
2	1000	1050	51-1050
3	125	1175	1051-1175
4	300	1475	1176-1475
5	500	1975	1476-1975
6	25	2000	1976-2000

Trois nombres aléatoires entre 1 et 2000 sont sélectionnés pour obtenir un échantillon de trois unités. Supposons que ces nombres sont : 1697, 624 et 1109. Les exploitations agricoles (fermes) sélectionnées sont donc : les fermes 5, 2 et 3.

Dans le cas de la méthode aléatoire d'échantillonnage PPT sans remise, si plus d'une unité est sélectionnée, essayer de maintenir les probabilités directement proportionnelles à la taille et estimer les variances d'échantillonnage des estimations de l'enquête peuvent susciter des complications. La situation devient encore plus compliquée si plus de deux ou trois unités sont sélectionnées avec PPT sans remise et, en fait, fait l'objet d'un nombre considérable de travaux de recherche. La majeure partie de cette recherche est contenue dans les ouvrages de Horvitz et Thompson (1952), Yates et Grundy (1953), Rao, Hartley et Cochran (1962), Fellegi (1963), Brewer et Hanif (1983).

ii. Méthode systématique :

- pour chaque unité de la population, faire le calcul cumulatif des mesures de taille des unités jusqu'à l'unité elle-même comprise,

- déterminer l'étendue correspondant à chaque unité dans la population, c'est-à-dire à partir de la somme cumulative de l'unité précédente (mais sans l'inclure) jusqu'à la somme cumulative de l'unité courante,

- déterminer l'intervalle d'échantillonnage, k = (taille cumulative totale)/n,

- déterminer une origine choisie au hasard, r, entre 0 (si les mesures de taille ne sont pas des nombres entiers) ou 1 (si les mesures de taille sont des nombres entiers) et k,

- sélectionner les unités dont l'étendue contient les nombres aléatoires $r, r+k, r+2k, \ldots r+(n-1)k$.

iii. Méthode systématique aléatoire :

La liste est établie au hasard dans cette méthode avant l'application de l'échantillonnage systématique. Si la liste est utilisée dans l'ordre original, comme dans le cas de l'échantillonnage systématique, certains échantillons possibles peuvent être éliminés. Lorsque la liste est établie au hasard, le nombre d'échantillons éventuels qui peuvent être tirés est à la hausse.

On se souviendra des problèmes que posent ces méthodes. Dans le cas des méthodes systématiques aléatoires et systématiques, par exemple, si la taille d'une unité est plus grande que l'intervalle, elle peut être sélectionnée plus d'une fois. Ce problème peut être résolu uniquement en répartissant ces grandes unités en strates distinctes et en en faisant l'échantillonnage à part (Section 6.2.6). La difficulté d'estimation des variances d'échantillonnage est un autre problème.

6.2.5 Échantillonnage par grappes

L'échantillonnage par grappes est le processus de sélection aléatoire de groupes complets (grappes) d'unités de la population dans la base de sondage. C'est habituellement une stratégie d'échantillonnage statistiquement moins efficiente que l'EAS et elle est appliquée pour plusieurs raisons. Premièrement, l'échantillonnage par grappes peut réduire énormément le coût de la collecte, surtout si la population est largement dispersée et si on a recours à des interviews sur place. Deuxièmement, il n'est pas toujours pratique d'échantillonner des unités distinctes de la population. Il est parfois plus facile de faire l'échantillonnage de groupes d'unités de la population (p. ex., ménages complets). Troisièmement, elle permet de faire des estimations pour les grappes elles-mêmes (p. ex., revenu moyen par ménage).

L'échantillonnage par grappes est un processus en deux étapes. Premièrement, la population est regroupée en grappes (il peut s'agir de grappes naturelles, p. ex., ménages, écoles). La deuxième étape est la sélection d'un échantillon de grappes et l'interview de toutes les unités des grappes sélectionnées.

La base de sondage peut déterminer la méthode d'échantillonnage. Jusqu'à maintenant, la cible a été l'échantillonnage d'unités individuelles de la population à partir d'une liste. Si les unités de la population sont naturellement regroupées, il est souvent plus facile d'établir une base de sondage pour ces groupes et d'en faire l'échantillonnage, plutôt que d'essayer d'établir une liste de toutes les unités individuelles de la

population. Le client peut être intéressé, par exemple, à échantillonner les enseignants, mais avoir seulement une liste des écoles. Dans le cas des enquêtes sur les ménages ou les exploitations agricoles, de nombreux pays n'ont pas de listes complètes et à jour des gens, des ménages ou des exploitations agricoles dans aucune grande région géographique, mais ils ont des cartes des régions. Il est alors possible d'établir une base aréolaire et de répartir les secteurs géographiques en régions (grappes), de faire l'échantillonnage des régions et d'interviewer chacun dans la région. Divers plans d'échantillonnage peuvent servir pour sélectionner les grappes, notamment, l'EAS, le SYS ou le PPT. Un plan commun utilise le PPT dont l'échantillonnage est proportionnel à la taille de la grappe.

Il ne faut pas oublier un certain nombre de considérations pour l'échantillonnage par grappes. Les estimations seront statistiquement efficientes si les unités d'une grappe sont aussi différentes que possible. Autrement, si les unités d'une grappe sont semblables, elles donnent toutes de l'information semblable et il suffirait d'interviewer une unité.

Les unités d'une grappe ont souvent des caractéristiques malheureusement semblables et elles sont donc plus homogènes que les unités sélectionnées au hasard dans la population en général. Le résultat est une procédure d'échantillonnage moins efficiente que celle de l'EAS. Supposons, par exemple, que deux échantillons sont tirés d'une ville de 100 000 personnes. L'échantillonnage par grappes est utilisé pour le premier échantillon et un îlot de la ville englobant 400 résidents est sélectionné au hasard. L'EAS est appliqué au deuxième échantillon pour sélectionner 400 personnes dans une liste de 100 000 résidents. L'échantillon de 400 résidents de l'EAS sera probablement beaucoup plus diversifié aux volets revenus, âge, occupation et scolarité (pour nommer seulement quelques variables) que l'échantillon par grappes de 400 personnes qui habitent toutes le même îlot en ville.

La qualité de l'homogénéité des unités des grappes, le nombre d'unités de la population dans chaque grappe et le nombre de grappes de l'échantillon déterminent l'efficience statistique de l'échantillonnage par grappes. Si les unités voisines sont semblables, il est statistiquement plus efficient de sélectionner de nombreuses petites grappes plutôt que quelques-unes plus larges. Lors des interviews sur place cependant, plus l'échantillon est dispersé, plus l'enquête coûte cher. L'organisme statistique doit établir un équilibre entre le nombre optimal et la taille des grappes et le coût.

L'échantillonnage par grappes peut poser des difficultés logistiques. Si la base de sondage est une base aréolaire tirée d'une carte et si l'unité d'échantillonnage est une grappe de logements, il peut être difficile de déterminer si un logement est dans une grappe ou une autre. Il faudrait établir certaines règles élémentaires pour déterminer quelles unités font partie d'une grappe. Si la règle suivante est établie, par exemple, à savoir que *les logements font partie de la grappe où se trouve leur entrée principale (porte à l'avant)*, la majorité des problèmes seraient éliminés (habituellement, le logement complet est à l'intérieur ou à l'extérieur des limites d'une grappe). Si un logement semble également réparti entre plus d'une grappe, tirez au sort pour éviter un biais. Dans l'Enquête canadienne sur la population active (EPA), les grappes sont déterminées en tirant une ligne au milieu de la rue. Il est donc facile de déterminer si un logement est dans l'échantillon ou non. (Le lecteur trouvera davantage d'information sur ces considérations pratiques au **Chapitre 9 - Opérations de collecte des données).**

Voici les **avantages** de l'échantillonnage par grappes :

i. Il peut réduire énormément le coût de la collecte parce que l'échantillon est moins dispersé que celui de l'EAS. C'est particulièrement important si la population est largement répartie et si l'enquête comprend des interviews sur place parce qu'il est possible d'économiser en diminuant le temps de déplacement des intervieweurs, en particulier pour les populations en milieu rural.

ii. Il est plus facile à appliquer que l'EAS ou le SYS aux populations regroupées naturellement par grappes (p. ex., ménages, écoles) et à certaines populations conceptuelles, par exemple, les personnes qui traversent une frontière pendant une période déterminée. Il peut être difficile, coûteux ou impossible d'établir une liste de toutes les unités individuelles de ce genre de population comme l'exige l'EAS.

iii. Il permet de faire des estimations pour les grappes elles-mêmes. Les estimations du nombre moyen d'enseignants par école sont un exemple (lorsque les écoles sont en grappes).

iv. Il peut être statistiquement plus efficient qu'un EAS si les unités des grappes sont hétérogènes (différentes) du point de vue des variables de l'étude et si les grappes sont homogènes (semblables). Ce n'est cependant pas le cas en pratique, habituellement.

Voici les **inconvénients** de l'échantillonnage par grappes :

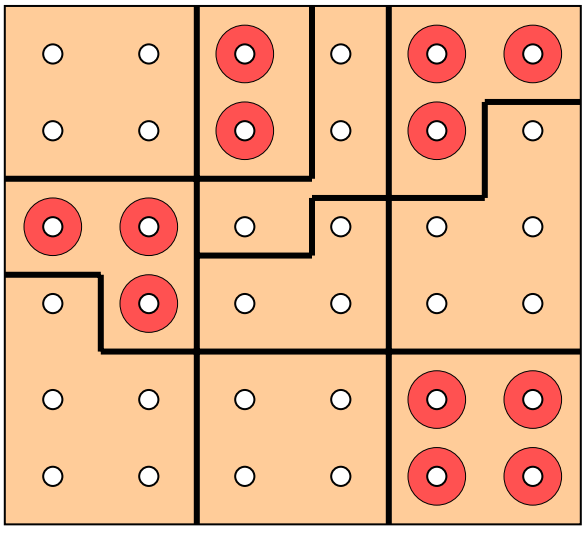
i. Il peut être statistiquement moins efficient que l'EAS si les unités des grappes sont homogènes du point de vue des variables de l'étude. C'est souvent le cas parce que les unités d'une grappe ont tendance à avoir des caractéristiques semblables. Le nombre de grappes sélectionnées peut cependant être augmenté pour éliminer cette perte d'efficience statistique.

ii. La taille finale de l'échantillon n'est pas connue d'avance parce que le nombre d'unités d'une grappe est déterminé seulement à la conclusion de l'enquête.

iii. L'organisation de l'enquête peut être plus complexe que dans le cas d'autres méthodes.

iv. L'estimation de la variance peut être plus complexe que celle de l'EAS si les grappes sont échantillonnées sans remise.

Échantillon par grappes (illustré, quatre grappes sont échantillonnées)



6.2.6 Échantillonnage stratifié (STR)

Au cours de l'échantillonnage stratifié, la population est répartie en groupes homogènes mutuellement exclusifs intitulés strates et des échantillons indépendants sont ensuite sélectionnés dans chaque strate. N'importe quel plan d'échantillonnage mentionné dans ce chapitre peut servir à l'échantillonnage d'une strate, à partir de méthodes plus simples comme l'EAS ou le SYS, jusqu'aux méthodes plus complexes comme l'échantillonnage PPT, par grappes, à plusieurs degrés ou à plusieurs phases (considérés plus loin dans ce chapitre). Dans l'échantillonnage par grappes, par exemple, il est très commun de stratifier d'abord et de tirer ensuite l'échantillon par grappes. Cette méthode est intitulée échantillonnage par grappes stratifié.

Une population peut être stratifiée par n'importe quelle variable disponible pour toutes les unités de la base de sondage avant de procéder à l'enquête. Cette information, par exemple, peut être simplement l'adresse de l'unité qui permettra la stratification par province, ou les données sur les revenus entrées dans la base de sondage qui permettront la stratification par groupe de revenu, les variables de stratification souvent utilisées comprennent : l'âge, le sexe, la géographie (p. ex., province), le revenu, les revenus de toute source, la taille du ménage, la taille de l'entreprise, le genre d'entreprise, le nombre d'employés, etc.).

Trois principales raisons justifient la stratification. Premièrement, elle permet d'obtenir une stratégie d'échantillonnage plus efficiente que celle de l'EAS ou du SYS. Deuxièmement, elle donne des tailles d'échantillon suffisantes pour des domaines d'intérêt en particulier qui motivent l'analyse à effectuer. Troisièmement, elle aide à éviter de tirer un « mauvais » échantillon.

D'une part, pour une taille d'échantillon et un estimateur donnés, la stratification peut diminuer l'erreur d'échantillonnage ou, d'autre part, pour une erreur d'échantillonnage donnée, la taille de l'échantillon peut être plus petite. Bien que les grappes et les strates soient toutes deux des regroupements d'unités de la population, un échantillon est tiré de chaque strate mais les grappes sont enquêtées intégralement. La stratification est en général plus précise que l'EAS, mais l'échantillonnage par grappes l'est généralement moins (parce que les unités voisines sont habituellement semblables).

Il faut observer une forte homogénéité dans une strate (c.-à-d. que les unités d'une strate devraient être semblables quant à la variable d'intérêt) pour améliorer l'efficience statistique d'une stratégie d'échantillonnage de l'EAS et les strates elles-mêmes doivent être différentes le plus possible (quant à la même variable d'intérêt). On peut généralement obtenir ce résultat si les variables de la stratification sont corrélées avec la variable d'intérêt de l'enquête. Cochran (1977) explique pourquoi la stratification peut augmenter la précision des estimations par rapport à l'EAS :

> Si chaque strate est homogène, c'est-à-dire si les mesures varient peu d'une unité à l'autre, il est possible d'obtenir une estimation précise de n'importe quelle moyenne de strate à l'aide d'un petit échantillon de cette strate. Ces estimations peuvent être intégrées en une estimation précise de la population dans l'ensemble.

La stratification est particulièrement importante si les *populations sont asymétriques* (c.-à-d. lorsque la répartition des valeurs d'une variable n'est pas symétrique et qu'elle affiche une tendance vers la droite ou la gauche). Les enquêtes auprès des entreprises et des exploitations agricoles, par exemple, ont souvent des populations fortement asymétriques : quelques grandes entreprises et exploitations agricoles peu nombreuses ont souvent de grandes valeurs pour les variables d'intérêt (p. ex., revenus, dépenses, nombre d'employés). Quelques unités de la population peuvent alors avoir d'importantes répercussions sur les estimations, si elles sont sélectionnées dans l'échantillon, elles peuvent augmenter énormément

l'estimation et, si elles ne sont pas sélectionnées, l'estimation peut être beaucoup plus faible. Autrement dit, ces unités peuvent augmenter la variabilité d'échantillonnage de l'estimation. Ces unités devraient donc former une strate distincte pour garantir qu'elles ne représentent pas d'autres unités éventuellement plus petites de la population.

Une variable de la taille, dérivée du nombre d'employés, par exemple, est souvent utilisée pour la stratification des entreprises. Si la variable de la taille a trois valeurs, petite, moyenne et grande, l'efficience statistique est améliorée si les grandes entreprises ont des ventes semblables, les moyennes entreprises ont des ventes semblables et les petites entreprises ont des ventes semblables, et si les moyennes et grandes entreprises, et les moyennes et petites entreprises, ont des ventes très différentes. De même, dans un plan d'échantillonnage qui utilise des bases aréolaires, la représentation appropriée des grandes villes peut être garantie en les intégrant dans une strate distincte et en faisant l'échantillonnage de chaque strate séparément.

Dans l'exemple précédent, il était raisonnable de stratifier par nombre d'employés parce que c'est une mesure de la taille de l'entreprise et elle est probablement étroitement liée aux ventes. D'autre part, si une enquête cible l'âge de ces employés, il est insensé de stratifier par nombre d'employés parce qu'il n'y a pas de corrélation. De plus, la stratification statistiquement efficiente pour une variable de l'enquête peut fonctionner moins bien pour d'autres. Les variables de la stratification sont habituellement choisies selon leur corrélation avec les plus importantes variables de l'enquête. Dans le cas des variables moins importantes de l'enquête qui n'ont pas de corrélation avec les variables de la stratification, cela signifie que les estimations pour un échantillon stratifié peuvent être moins efficientes que celles de l'EAS.

La deuxième raison de la stratification est de garantir des tailles d'échantillon appropriées pour les ***domaines d'intérêt*** connus. Au cours de la conception d'une enquête, l'objectif général est souvent d'estimer un total. Combien de personnes n'avaient pas d'emploi le mois dernier? Quel était le total des ventes au détail le mois dernier? Souvent, le client veut, non seulement les totaux dans l'ensemble, mais aussi des estimations pour les sous-groupes de la population intitulés domaines.

Le client veut, par exemple, savoir combien d'hommes étaient sans emploi et comparer ce résultat au nombre de femmes sans emploi. De même, le client veut peut-être avoir les résultats des ventes le mois dernier pour les magasins de vêtements ou pour tous les magasins de détail dans une province en particulier. Établir des estimations pour les sous-groupes est intitulé estimation du domaine. Si des estimations de domaines sont nécessaires, la capacité de les calculer à l'aide d'un échantillon suffisamment large dans chaque domaine devrait être intégrée au plan d'échantillonnage. Si l'information est disponible dans la base de sondage, le moyen le plus facile d'y arriver est de garantir que les strates correspondent exactement aux domaines d'intérêt.

La troisième raison de la stratification est l'application d'une mesure de protection contre le tirage d'un « mauvais » échantillon. Dans le cas de l'EAS, la sélection de l'échantillon est laissée entièrement à la chance. L'échantillonnage stratifié tente de restreindre les échantillons possibles aux moins extrêmes en garantissant qu'au moins certaines parties de la population seront représentées dans l'échantillon. La base de sondage devrait être stratifiée par sexe (en supposant que cette variable auxiliaire est disponible dans la base), par exemple, pour garantir que les hommes et les femmes sont inclus dans l'échantillon.

Ajoutons à ces raisons que la stratification est souvent utilisée parce qu'elle est pratique du point de vue opérationnel ou administratif. Elle peut permettre à l'organisme statistique de contrôler la répartition du travail sur le terrain entre ses bureaux régionaux. Si la collecte des données est faite par province, par exemple, la stratification par province est appropriée et le bureau régional provincial peut obtenir sa part de l'échantillon.

Lorsque la population a été répartie en strates, l'organisme statistique doit déterminer combien d'unités il faut échantillonner dans chaque strate. Cette étape est intitulée répartition de l'échantillon et elle est considérée au **Chapitre 8 - Calcul de la taille de l'échantillon et répartition.**

Les probabilités d'inclusion varient habituellement d'une strate à l'autre, compte tenu de la répartition de l'échantillon entre les strates. Il faut considérer la taille de l'échantillon et la taille de la population dans chaque strate pour calculer les probabilités d'inclusion de la majorité des plans d'échantillonnage. Considérons une population de N=1000 unités stratifiées en deux groupes pour illustrer : une strate a N_1=250 unités et l'autre, N_2=750 unités. Supposons que l'EAS est utilisé pour sélectionner n_1=50 unités à la première strate et n_2=50 unités à la deuxième strate. La probabilité, π_1, qu'une unité de la première strate soit sélectionnée est donc $\pi_1 = 50/250 = 1/5$ et la probabilité, π_2, qu'une unité de la deuxième strate soit sélectionnée est $\pi_2 = 50/750 = 1/15$. Les unités ont donc différentes probabilités d'inclusion, c'est-à-dire qu'une unité de la première strate a plus de chance d'être sélectionnée que celle de la deuxième.

Voici les **avantages** de l'échantillonnage stratifié :

i. Il peut accroître la précision des estimations de la population dans l'ensemble et la stratégie d'échantillonnage est donc plus efficiente. Un échantillon plus petit peut éviter une dépense considérable pendant l'enquête, en particulier lors de la collecte des données.

ii. Il aide à garantir que les sous-groupes importants, répartis en strates, sont bien représentés dans l'échantillon et les estimateurs de domaines sont alors statistiquement efficients.

iii. Il peut être pratique du point de vue opérationnel et administratif.
iv. Il aide à éviter la sélection d'un « mauvais » échantillon.

v. Il permet d'appliquer différents plans d'échantillonnage et diverses procédures à des strates différentes (p. ex., EAS pour une strate, PPT pour une autre).

Voici les **inconvénients** de l'échantillonnage stratifié :

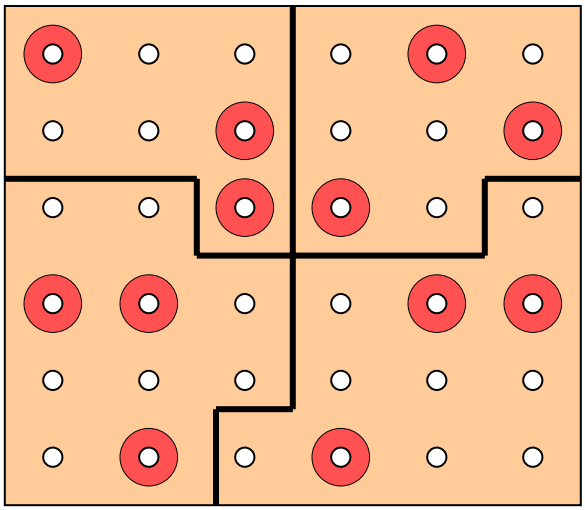
i. Le plan d'échantillonnage doit contenir de l'information auxiliaire de haute qualité pour toutes les unités du plan, et non pas seulement celles de l'échantillon, qui peuvent être utilisées pour la stratification.

ii. L'établissement de la base de sondage coûte plus cher et est plus complexe que dans le cas de l'EAS ou du SYS parce que la base exige une bonne information auxiliaire.

iii. Il peut donner une stratégie d'échantillonnage stratégiquement moins efficiente que celle de l'EAS pour les variables de l'enquête qui ne sont pas corrélées avec les variables de la stratification.

iv. L'estimation est légèrement plus complexe que celle de l'EAS ou du SYS.

Échantillon stratifié (illustré, quatre strates, trois unités sélectionnées par strate)



6.2.7 Échantillonnage à plusieurs degrés

Nos considérations ont été axées jusqu'à maintenant sur les plans d'échantillonnage à un degré. L'échantillonnage à plusieurs degrés est le processus de sélection d'un échantillon à deux degrés successifs ou plus. Les unités sélectionnées au premier degré sont intitulées unités primaires d'échantillonnage (UPÉ), les unités sélectionnées au deuxième degré sont intitulées unités secondaires d'échantillonnage (USÉ), etc. Les unités à chaque degré ont une structure différente et sont hiérarchiques (p. ex., les personnes qui habitent dans un logement, les logements qui forment un îlot en ville, les îlots qui forment une ville, etc.). Les USÉ sont souvent les unités individuelles de la population dans un échantillonnage à deux degrés.
Un plan d'échantillonnage commun à plusieurs degrés comprend l'échantillonnage par grappes à deux degrés à l'aide d'une base aréolaire au premier degré pour sélectionner des régions (l'UPÉ) et d'un échantillon systématique de logements (l'USÉ) dans une région, au deuxième degré. Compte tenu de l'échantillonnage par grappes à un degré présenté auparavant, chaque unité d'une grappe échantillonnée est comprise dans l'échantillon. Dans l'échantillonnage à deux degrés, seulement certaines unités de chaque UPÉ sélectionnée sont sous-échantillonnées.

L'échantillonnage à plusieurs degrés est habituellement utilisé dans des bases aréolaires pour pallier les inefficiences de l'échantillonnage par grappes à un degré qui est en fait rarement utilisé. Si les unités voisines dans une grappe sont semblables, il est statistiquement plus efficient d'échantillonner quelques USÉ de nombreuses UPÉ que d'échantillonner de nombreuses USÉ de moins d'UPÉ.

Les échantillons à plusieurs degrés peuvent avoir n'importe quel nombre de degrés, mais, étant donné que la complexité du plan (et de l'estimation) augmente avec le nombre de degrés, les plans d'échantillonnage sont souvent restreints à deux ou trois degrés. Il faut souligner que la base de sondage pour le premier degré est généralement très stable. Une base aréolaire qui couvre de grands secteurs géographiques, par exemple, ne change pas rapidement avec le temps. Les bases du deuxième degré (et des degrés suivants) nécessaires pour échantillonner des unités à des degrés ultérieurs sont habituellement moins stables. Ces bases sont souvent des listes établies sur place pendant la collecte des données. Dans le cas des secteurs géographiques échantillonnés au premier degré, par exemple, une liste de tous les logements des secteurs échantillonnés peut être établie. Moins d'efforts sont nécessaires pour lister seulement les secteurs

échantillonnés plutôt que toute la population. (Le **Chapitre 9 - Opérations de collecte des données** couvre en détails le listage.)

Chaque degré d'un échantillon à plusieurs degrés peut être accompli à l'aide de n'importe quelle technique d'échantillonnage. La souplesse est donc l'un des principaux avantages de l'échantillonnage à plusieurs degrés. Un échantillon aléatoire simple peut être tiré, par exemple, d'une UPÉ sélectionnée au premier degré. Il peut y avoir, pour une autre UPÉ, une mesure de la taille corrélée avec les principales variables de l'enquête et l'échantillonnage PPT peut être utilisé pour cette UPÉ.

L'échantillon de l'Enquête canadienne sur la population active (EPA) est un exemple d'échantillon stratifié à plusieurs degrés. Le pays est réparti en plus de 1 100 strates. Chaque strate comprend un groupe de secteurs de dénombrement (SD). Les SD sont des secteurs géographiques définis dans le Recensement de la population et la région couverte peut être dénombrée par un *recenseur* (ils sont délimités en tenant compte de la taille du territoire et de la densité de la population). Le premier degré de l'échantillonnage est un échantillon stratifié de grappes (SD ou groupes de SD) tiré de ces strates. Au deuxième degré, les grappes sont cartographiées, tous les logements de ces grappes sont listés et le recenseur sélectionne un échantillon systématique de logements dans chaque liste. Toutes les personnes d'un logement sélectionné sont ensuite interviewées pour l'enquête.

N'oubliez pas que les exemples présentés jusqu'à maintenant appliquent une base aréolaire au premier degré, mais ce n'est pas une exigence de l'échantillonnage à plusieurs degrés. Un exemple d'échantillon à plusieurs degrés qui appliquerait un genre différent de base est une enquête sur les voyages dans un aéroport. L'unité d'échantillonnage primaire pourrait être le temps, les jours dans un mois, et l'unité au deuxième degré pourrait être les voyageurs eux-mêmes. Dans le cas d'une enquête plus complexe sur les voyages, l'unité du deuxième degré pourrait être les avions de passagers à l'arrivée et l'unité au troisième degré pourrait être les sièges occupés dans l'avion.

Voici les **avantages** de l'échantillonnage à plusieurs degrés :

i. Il peut donner une stratégie d'échantillonnage statistiquement plus efficiente que celle du plan d'échantillonnage par grappes à un degré lorsque les grappes sont homogènes quant aux variables d'intérêt (c.-à-d. une réduction de la taille de l'échantillon).

ii. Il peut réduire énormément le temps de déplacement et les coûts des interviews sur place parce que l'échantillon est moins dispersé que celui d'autres formes d'échantillonnage, notamment l'EAS.

iii. Il n'est pas nécessaire d'avoir une liste de toute la population. Il faut simplement avoir une bonne base à chaque degré de sélection de l'échantillon.

Voici les **inconvénients** de l'échantillonnage à plusieurs degrés :

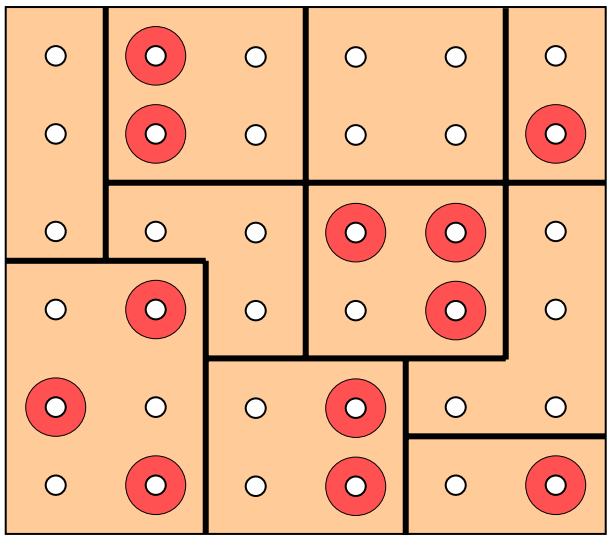
i. L'efficience statistique est habituellement moindre que celle de l'EAS (même s'il peut être plus efficient qu'une stratégie par grappes à un degré).

ii. La taille finale de l'échantillon n'est pas toujours connue d'avance parce que le nombre d'unités d'une grappe est habituellement connu seulement à la conclusion de l'enquête. (La taille de l'échantillon peut être contrôlée, cependant, si un nombre déterminé d'unités est sélectionné dans chaque grappe.)

iii. L'organisation de l'enquête est plus complexe que celle d'un échantillonnage par grappes à un degré.

iv. Ses formules de calcul des estimations et de la variance d'échantillonnage peuvent être complexes.

Échantillon à plusieurs degrés (illustré, plan d'échantillonnage par grappes à deux degrés, six UPÉ sélectionnées et jusqu'à trois USÉ sélectionnées dans chaque UPÉ)



6.2.8 Échantillonnage à plusieurs phases

Les appellations se ressemblent, mais l'échantillonnage à plusieurs phases est très différent de l'échantillonnage à plusieurs degrés. L'échantillonnage à plusieurs phases comprend aussi la sélection de deux échantillons ou plus, mais les échantillons sont tirés de la même base et les unités ont la même structure à chaque phase. La collecte pour l'échantillon à plusieurs phases cible surtout l'information d'un large échantillon d'unités et ensuite, l'information plus détaillée pour un sous-échantillon de ces unités. L'échantillonnage à plusieurs phases le plus commun est l'échantillonnage à deux phases (ou échantillonnage double), mais trois phases ou plus sont aussi possibles. Plus il y a de phases, cependant, plus les estimations et le plan d'échantillonnage sont complexes, tout comme dans l'échantillonnage à plusieurs degrés.

L'échantillonnage à plusieurs phases est utile lorsque la base de sondage manque d'information auxiliaire qui pourrait servir à stratifier la population ou à en retrancher une partie. Supposons, par exemple, que l'on ait besoin d'information sur les éleveurs de bovins, mais la base de sondage comprend seulement une liste d'exploitations agricoles, sans information auxiliaire. On pourrait procéder à une enquête simple en posant seulement une question : « Votre exploitation agricole est-elle axée, en tout ou en partie, sur l'élevage de bovins? » Cette enquête à une seule question devrait coûter très peu par interview (surtout si elle est faite au téléphone) et l'organisme devrait donc pouvoir obtenir un important échantillon. Lorsque le premier échantillon est tiré, un deuxième échantillon plus petit peut être sélectionné dans la population des éleveurs de bovins et vous pouvez leur poser des questions plus détaillées. L'organisme statistique qui applique cette méthode évite les frais de sondage des unités hors du champ de l'enquête (c.-à-d. ceux qui ne sont pas éleveurs de bovins).

L'échantillonnage à plusieurs phases peut aussi servir à la collecte de l'information plus détaillée à partir d'un sous-échantillon lorsque le budget n'est pas suffisant pour obtenir de l'information de tout l'échantillon ou lorsque le fardeau de réponse serait excessif. L'Enquête trimestrielle sur les marchandises vendues au détail (ETMVD) est un exemple. La première phase de l'enquête est l'Enquête mensuelle sur le commerce de gros et de détail (EMCGD). Les enquêteurs de l'EMCGD demandent chaque mois deux variables aux grossistes et aux détaillants : les ventes et les stocks mensuels. Les enquêteurs de l'ETMVD sous-échantillonnent les détaillants et leur demandent de faire rapport sur leurs ventes par produits de détail, par exemple, les vêtements, les articles électroniques, les denrées alimentaires, etc.

L'échantillonnage à plusieurs phases peut aussi servir lorsque les frais de collecte des données sont très différents pour diverses questions d'une enquête. Considérons une enquête sur la santé qui pose des questions élémentaires sur le régime alimentaire, le tabagisme, l'exercice et la consommation d'alcool. Supposons de plus que les enquêteurs demandent aux répondants de se prêter à certaines mesures directes, notamment, marcher sur un tapis roulant, faire prendre une mesure de leur tension artérielle et de leur taux de cholestérol. Poser quelques questions coûte relativement peu, mais les examens médicaux demandent le temps d'un praticien formé en soins de santé et l'utilisation d'un laboratoire équipé qui coûtent relativement cher. L'enquête peut être faite à l'aide d'un échantillon à deux phases, les questions élémentaires sont posées à la première phase et les mesures directes sont prises seulement auprès de l'échantillon plus petit de la deuxième phase.

Les données obtenues à la première phase peuvent servir à la stratification ou à l'information de sélection, mais aussi pour améliorer l'efficience de l'estimation (p. ex., pour l'estimation par régression). Ces notions seront reprises au **Chapitre 7 - Estimation**.

Voici les **avantages** de l'échantillonnage à plusieurs phases :

i. Il peut augmenter énormément la précision des estimations (comparativement à l'EAS).

ii. Il peut servir à obtenir de l'information auxiliaire qui n'est pas dans le plan d'échantillonnage (en particulier de l'information sur la stratification pour l'échantillonnage à la deuxième phase).

iii. Il peut être utile si les frais de collecte pour certaines variables de l'enquête sont excessifs ou si le fardeau des répondants est trop lourd.

Voici les **inconvénients** de l'échantillonnage à plusieurs phases :

i. Il faut plus de temps pour obtenir des résultats que le temps nécessaire pour une enquête à une phase si les résultats de la première phase sont nécessaires pour procéder à la deuxième phase.

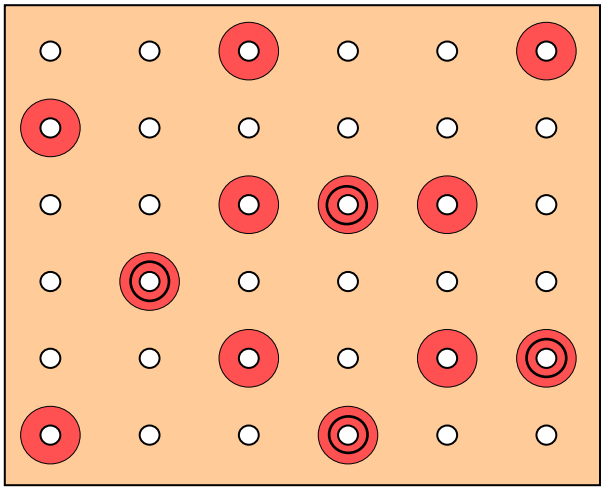
ii. Il peut coûter plus cher qu'une enquête à une phase parce qu'il faut interviewer une unité échantillonnée plus d'une fois.

iii. Si la population est mobile ou si les caractéristiques d'intérêt changent souvent, la période écoulée entre les phases peut poser des problèmes.

iv. L'organisation de l'enquête peut être complexe.

v. Ses formules de calcul des estimations et de la variance de l'échantillonnage peuvent être très complexes.

Échantillon à plusieurs phases (illustré, 12 unités sélectionnées à la première phase, quatre à la deuxième)



6.2.9 Échantillonnage par répliques

L'échantillonnage par répliques comprend la sélection d'un nombre d'échantillons indépendants dans une population et non dans un seul échantillon. Au lieu d'un échantillon global, un certain nombre d'échantillons plus petits, de taille à peu près égale, intitulés répliques, sont sélectionnés indépendamment, chacun à partir du même plan d'échantillonnage. L'échantillonnage par répliques peut servir lorsque les résultats préliminaires sont demandés rapidement. Ces résultats préliminaires peuvent être tirés du traitement et de l'analyse d'une seule réplique.

La principale raison d'un échantillonnage par répliques est de faciliter le calcul de la variance d'échantillonnage des estimations d'une enquête (la variance d'échantillonnage est une mesure de l'erreur d'échantillonnage). Il est généralement possible de calculer la variance d'échantillonnage à l'aide d'échantillons probabilistes, mais ces calculs peuvent être extrêmement difficiles selon la complexité du plan d'échantillonnage. Certaines expressions mathématiques pour la variance de l'échantillonnage sont difficiles à déterminer, fastidieuses à programmer, coûtent cher, et c'est un problème. Dans le cas de l'échantillonnage systématique en particulier, les estimations de la variance ne peuvent être calculées directement, sauf si des hypothèses sont formulées sur la disposition des unités dans la liste.

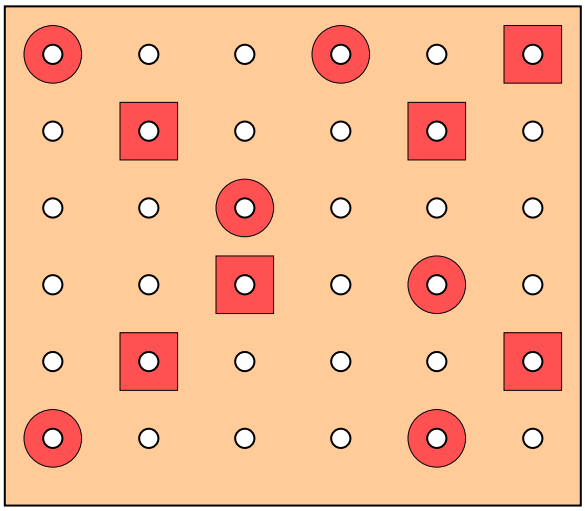
On obtient des mesures de l'erreur d'échantillonnage en examinant à quel point les estimations varient d'un échantillon à l'autre, compte tenu de tous les échantillons possibles de la même taille et du même plan d'échantillonnage. L'échantillonnage par répliques simule ce concept. Au lieu de tirer tous les échantillons possibles (ce qui n'est pas pratique), un nombre raisonnable d'échantillons plus petits est sélectionné à l'aide de méthodes identiques. Au lieu de sélectionner un échantillon de taille 10 000, par exemple, dix échantillons indépendants de taille 1 000 peuvent être sélectionnés. En comparant les estimations de chacun de ces dix échantillons, on peut obtenir des estimations de la variance d'échantillonnage. La fiabilité des estimations de la variance d'échantillonnage augmente avec le nombre de répétitions sélectionnées. (Un exemple d'échantillonnage par répliques pour estimation de la variance est donné à la Section 7.3.4 du **Chapitre 7 - Estimation**.)

Un certain nombre d'autres procédures appliquent le ré-échantillonnage pour estimer la variance d'échantillonnage lorsque les plans d'échantillonnage sont complexes. Ils comprennent les répliques équilibrées (méthode BRR), la méthode du Jackknife et la méthode d'auto-amorçage (*Bootstrap*). Ces

techniques sont toutes des ramifications de l'idée élémentaire de l'échantillonnage par répliques, mais elles sont différentes l'une de l'autre quant à la précision de la mesure de la variance d'échantillonnage de divers genres d'estimations d'enquête, de leur complexité opérationnelle et des situations auxquelles elles s'appliquent le mieux.

Cette approche a des inconvénients, par exemple, les estimations de la variance d'échantillonnage ont tendance à être moins précises en général que si elles étaient directement basées sur des expressions statistiques qui intègrent des caractéristiques de plan d'échantillonnage, notamment, l'échantillonnage à plusieurs degrés, la stratification, etc.

Échantillonnage par répliques (illustré, deux échantillons sélectionnés de taille 6)



6.3 Sujets spéciaux en échantillonnage

Les plans d'échantillonnage sont parfois modifiés pour répondre aux besoins spéciaux d'une enquête en particulier. Cette mesure peut être nécessaire si la population cible est particulièrement difficile à situer, si la caractéristique d'intérêt est très rare dans la population, ou à cause des besoins analytiques de l'enquête ou de la méthode de collecte des données. Le **Chapitre 4 - Méthodes de collecte des données** exposait les plans d'échantillonnage des interviewés au téléphone, y compris la composition aléatoire (CA). Les sections suivantes décrivent d'autres applications particulières des plans d'échantillonnage pour répondre à des besoins d'enquête spéciaux.

6.3.1 Enquêtes répétées

Les enquêtes uniques ont de nombreuses différences comparativement aux enquêtes répétées. Le but d'une enquête répétée est souvent d'étudier les tendances ou les modifications des caractéristiques d'intérêt au fil du temps.

Avant de prendre des décisions sur le plan d'échantillonnage d'enquêtes répétées, il faudrait tenir compte de la possibilité de détérioration de l'efficience statistique de la stratégie d'échantillonnage au cours du temps. Un organisme statistique peut choisir, par exemple, d'utiliser des variables de stratification plus

stables et d'éviter celles qui peuvent être statistiquement plus efficientes à court terme, mais qui peuvent changer rapidement avec le temps.

Beaucoup de renseignements sont en général disponibles et utiles aux fins des plans ultérieurs, et c'est une autre caractéristique de l'enquête répétée. Il est possible d'examiner au cours du temps si les diverses caractéristiques du plan d'échantillonnage sont suffisantes, par exemple, la pertinence des limites et des variables de stratification, la méthode de répartition de l'échantillon et la taille des unités à diverses étapes du plan d'échantillonnage à plusieurs degrés, afin d'accentuer l'efficience statistique. L'information nécessaire pour établir un plan d'enquête unique efficient est souvent très limitée.

Quand on élabore un plan d'enquête répétée, il faut prévoir des dispositions pour tenir compte de certains événements, par exemple, les naissances, les décès et les modifications de la mesure de la taille. Les méthodes d'estimation et d'échantillonnage appliquées aux enquêtes répétées devraient intégrer ces modifications de façon statistiquement efficiente pour interrompre le moins possible les opérations d'enquête en cours.

Un type particulier d'enquête répétée est l'échantillon constant (panel) ou *enquête longitudinale*, c'est-à-dire que les données sont obtenues des mêmes unités de l'échantillon à plusieurs occasions. Ces enquêtes permettent habituellement de mesurer les modifications des caractéristiques d'une population donnée et d'obtenir une précision plus grande que celle d'une série d'échantillons indépendants de taille comparable. Si une enquête est répétée, le recours à un échantillon longitudinal a des **avantages**, comparativement à une série d'échantillons indépendants spéciaux. Voici certains avantages :

i. Il diminue la variance d'échantillonnage pour les estimations du changement (c.-à-d. $\hat{Y}_2 - \hat{Y}_1$, où $\hat{Y}_1$ est une estimation du total à l'occasion 1 et $\hat{Y}_2$ est une estimation du total à l'occasion 2). Vous pouvez obtenir, par exemple, une mesure du changement du nombre de personnes sans emploi d'un mois à l'autre.

ii. Il peut servir à obtenir de l'information sur le comportement des répondants avec le temps.

iii. Il peut diminuer les erreurs de réponse (parce que les répondants approfondissent leur compréhension du questionnaire).

iv. Les coûts peuvent diminuer avec le temps (l'élaboration de l'enquête, la programmation des systèmes informatiques, la formation du personnel, etc., sont faites au cours d'une longue période).

Voici certains **inconvénients** de l'utilisation de l'échantillon longitudinal au lieu de plusieurs échantillons indépendants :

i. Les estimations, le traitement des non-réponses, etc., sont plus complexes.

ii. Il faut que le budget de l'enquête soit garanti pendant toute la vie utile de l'échantillon constant. Un engagement financier pour couvrir les coûts est donc nécessaire pendant une longue période.

iii. Il est plus difficile de maintenir la représentativité au cours de périodes prolongées à cause des changements qui se produisent dans la population avec le temps, notamment, l'ajout de nouvelles unités et le retrait d'autres.

iv. Le nombre d'erreurs de réponse peut augmenter (p. ex., la connaissance du questionnaire peut inciter certains répondants à répondre incorrectement aux questions pour accélérer l'interview).

v. Le nombre de non-réponses peut augmenter avec le temps (à cause de la fatigue des répondants, la même personne faisant l'objet d'une enquête réitérée dans le temps, le repérage est difficile, etc.).

vi. Son organisation est plus complexe que celle d'une enquête unique.

vii. Il peut susciter un comportement motivé par l'enquête. Les questions réitérées sur les visites au médecin, par exemple, peuvent inciter un répondant à visiter un médecin à la suite de l'enquête.

viii. Il peut être difficile de définir certains concepts (p. ex., la composition du ménage peut changer avec le temps et alors, comment définir un ménage longitudinal?).

ix. Si l'échantillon sélectionné au départ est un « mauvais » échantillon, l'organisme statistique peut continuer de l'utiliser.

Le plan d'échantillonnage intermédiaire entre les échantillons indépendants utilisés à des occasions successives et l'échantillon longitudinal est intitulé plan d'échantillonnage avec renouvellement, c'est-à-dire qu'une partie de l'échantillon est remplacée chaque fois que le sondage est fait.

L'Enquête sur la population active (EPA) applique, par exemple, un plan d'échantillonnage avec renouvellement. Des ménages forment l'échantillon pendant six mois consécutifs et, chaque mois, un sixième de l'échantillon est remplacé par un nouveau groupe de ménages. L'échantillon de l'EPA est réparti en six panels (ou groupes de rotation). Chaque panel fait l'objet de l'enquête une fois par mois pendant six mois. À la fin du sixième mois, un groupe de rotation est supprimé de l'enquête (renouvelé) et un nouveau est ajouté. Le fardeau du répondant est ainsi limité (l'interview moyenne de l'EPA demande moins de dix minutes) et on conserve un bon chevauchement de l'échantillon chaque mois. Le renouvellement mensuel de l'échantillon est un avantage supplémentaire. Si l'échantillon n'est jamais mis à jour, les membres de l'échantillon vieillissent et les familles des nouveaux logements n'ont jamais la chance d'être ajoutés à l'échantillon. Dans ce cas, l'échantillon ne reflète plus la population actuelle et devient biaisé avec le temps.

Ce plan d'échantillonnage a l'avantage qu'il permet de mesurer les changements chaque mois avec une plus grande précision, il coûte moins cher et il y a moins d'interruptions des opérations sur place, comparativement aux échantillons indépendants. Il amenuise aussi le problème du fardeau des répondants lié aux études avec échantillons constants. (Afin de refléter les changements de la taille et de la structure de la population, ainsi que les besoins de données, néanmoins, des modifications périodiques sont apportées au plan d'échantillonnage de l'EPA, habituellement à l'occasion du recensement décennal.)

Ces plans d'échantillonnage servent non seulement à l'EPA, mais aussi aux enquêtes auprès des entreprises. Il ne faut pas oublier que les plans d'échantillonnage avec renouvellement exigent un plan d'échantillonnage élémentaire, mais aussi une méthodologie de renouvellement de l'échantillon. Il s'agit de notions complexes hors de la portée de ce manuel. On trouvera dans Kalton *et coll.* (1992) et dans Kasprzyk (1989) une présentation détaillée des plans à rotation partielle et des enquêtes longitudinales.

6.3.2 Enquêtes entrée-sortie

Les enquêtes entrée-sortie s'appliquent aux populations qui traversent une frontière, par exemple, les gens qui entrent dans un pays (ou qui le quittent) ou les utilisateurs d'une route à péage. Établir une liste à jour de ces populations avec information sur les personnes-ressources pour interviewer les unités ou leur envoyer un questionnaire pose un problème. Supposons, par exemple, que le client veuille interviewer les étrangers en visite au Canada et qu'il soit possible d'obtenir des douanes une liste de tous les visiteurs arrivés au pays à une date en particulier. Comment trouver ces gens pour les interviewer? Voilà un problème. Dès que la base de sondage est créée, les voyageurs sont probablement déjà retournés chez eux et l'entrevue n'est pas pratique. S'ils sont toujours au Canada, il est peu probable qu'une adresse pour communiquer avec eux soit disponible.

Voilà pourquoi une base de sondage conceptuelle et l'échantillonnage systématique, ou l'échantillonnage par grappes à deux degrés avec échantillonnage systématique dans les grappes échantillonnées, est(sont) souvent utilisé(s) pour faire enquête sur ces populations. La base de sondage conceptuelle peut être une liste des unités de la population énumérées au cours d'une certaine période à certains endroits en particulier. La base de sondage aura une couverture complète si ces endroits sont les secteurs où la population cible est concentrée. Des points d'entrée et de sortie sont souvent utilisés. Les points de sortie sont plus populaires parce que la plupart des enquêtes ciblent les activités de l'unité avant qu'elle quitte le secteur.

Il est important de tenir compte dans le plan d'échantillonnage, comme dans tout plan d'échantillonnage, des procédures sur place. Le défi à relever à l'échelon opérationnel et du plan d'échantillonnage est le recours optimal aux travailleurs sur place, tout en maintenant un échantillon probabiliste. L'achalandage inégal des visiteurs donne une charge de travail extrêmement variable et la répartition efficiente du personnel est donc difficile. L'utilisation la plus efficace du temps d'un intervieweur est l'interview du k^e visiteur après avoir *achevé* l'interview en cours, mais le plan d'échantillonnage serait alors non probabiliste. Il est préférable d'appliquer l'échantillonnage systématique, c'est-à-dire qu'une personne compte les gens et une petite équipe d'intervieweurs remet des questionnaires ou procède à des interviews. La densité de l'achalandage et de la longueur de l'interview, s'il y a interview, déterminent la taille de l'équipe.

La collecte des données peut se faire par autodénombrement, interviews ou observation directe lorsque c'est approprié. Dans le cas d'un questionnaire par autodénombrement, le taux de réponse est meilleur si le répondant remplit le questionnaire sur place au lieu de le retourner à l'organisme statistique par la poste. Les interviews demandent évidemment davantage de personnel sur place, mais elles donnent des taux de réponse plus élevés. L'observation directe est très précise et souhaitable, mais elle n'est pas toujours applicable.

Le principal **avantage** de l'enquête entrée-sortie est que la base de sondage pour l'étape finale peut être créée pendant que l'enquêteur est sur place.

Voici les **inconvénients** de l'enquête entrée-sortie :

i. Il peut être difficile de nouer un lien entre la population de l'enquête et une population habituellement comprise. Les enquêtes entrée-sortie mesurent des visiteurs, et non des personnes, voilà pourquoi. Si une enquête est faite à un magasin, par exemple, celui qui visite le magasin plus d'une fois au cours de la période sera compté plus d'une fois.

ii. Il peut être difficile de gérer les opérations sur place à cause des débits variables de la population. Voilà pourquoi de brèves interviews sont recommandées.

iii. Les taux de réponse sont typiquement faibles.

6.3.3 Échantillonnage boule de neige

Supposons que le client veut trouver des particuliers rares dans une population, qu'il en connaît déjà certains et qu'il peut communiquer avec eux. Une approche possible est de communiquer avec ceux-là et demander simplement s'ils connaissent quelqu'un comme eux, puis de communiquer avec ces personnes, etc. L'échantillon prend de l'ampleur comme une boule de neige qui descend une colline pour englober éventuellement à peu près tous ceux qui ont cette caractéristique. L'échantillonnage boule de neige est utile pour des populations petites ou spécialisées, notamment, les aveugles, les sourds, d'autres personnes qui ne font peut-être pas partie d'un groupe organisé ou, par exemple, des musiciens, peintres ou poètes qui ne sont pas déjà identifiés dans une liste de sondage. L'échantillonnage boule de neige est cependant une méthode d'échantillonnage non probabiliste : certains particuliers ou sous-groupes pourraient n'avoir aucune chance d'être échantillonnés. Il faut faire de solides hypothèses de modélisation (qui ne se concrétisent habituellement pas) pour formuler des inférences.

L'échantillonnage de réseaux et l'échantillonnage adaptatif par grappes sont des plans d'échantillonnage semblables utilisés pour cibler des populations rares ou spécialisées.

6.4 Sommaire

Ce chapitre a ciblé les notions élémentaires de l'échantillonnage. Les deux principaux types d'échantillonnage sont l'échantillonnage probabiliste et non probabiliste. L'utilité de l'échantillonnage non probabiliste est limitée pour les enquêtes des organismes statistiques parce que la sélection biaisée des unités ne permet pas de formuler immédiatement des inférences sur la population de l'enquête. Il est cependant facile et rapide et il peut être utile pour les études de recherche ou pendant la phase d'élaboration d'une enquête (p. ex., pour faire l'essai du questionnaire).

L'échantillonnage probabiliste devrait être utilisé lorsqu'il faut formuler des inférences sur la population, compte tenu des résultats de l'enquête. Dans un échantillon probabiliste, chaque unité de la base de sondage a une probabilité différente de zéro d'être sélectionnée et la sélection des unités est aléatoire. La sélection n'est donc pas biaisée et il est possible de calculer les probabilités d'inclusion et la variance d'échantillonnage des estimations, puis de formuler des inférences sur la population. Voici les principaux inconvénients de l'échantillonnage probabiliste : il demande plus de temps et coûte plus cher que l'échantillonnage non probabiliste, et la base d'échantillonnage doit être de qualité élevée.

Les plans d'échantillonnage probabiliste les plus simples sont l'échantillonnage aléatoire simple et l'échantillonnage systématique qui donnent des probabilités d'inclusion égales. Des plans d'échantillonnage plus complexes peuvent donner des probabilités d'inclusion inégales et la majorité d'entre eux exigent de l'information auxiliaire, y compris les échantillonnages avec probabilité proportionnelle à la taille, stratifiés, par grappes, à plusieurs degrés et à plusieurs phases. Les plans d'échantillonnage probabiliste inégaux sont typiquement utilisés pour améliorer l'efficience statistique de la stratégie d'échantillonnage ou pour diminuer les coûts de l'échantillonnage. La base d'échantillonnage justifie parfois leur utilisation.

Lorsque l'on choisit entre divers plans d'échantillonnage possibles, il faut d'abord déterminer quels plans d'échantillonnage sont réalistes, compte tenu de la base de sondage, des unités de la base de sondage, des domaines d'intérêt, du fardeau de la réponse, de la méthode de collecte des données, du budget, etc.

Voici certains points à considérer :

- Y a-t-il des données auxiliaires dans la base de sondage qui pourraient servir à améliorer l'efficience de l'échantillonnage (avec stratification ou PPT)?
- La base de sondage manque-t-elle d'information auxiliaire qui pourrait servir à la stratification ou à éliminer certaines unités? La collecte des données coûte-t-elle cher ou est-elle un fardeau (considérez deux phases)?
- La population est-elle naturellement répartie par grappes ou les unités de la base de sondage sont-elles des grappes? La population est-elle répartie géographiquement et y aura-t-il des interviews sur place (échantillonnage à un degré ou par grappes à plusieurs degrés)?

En bout de ligne, plusieurs applications spéciales de plans d'échantillonnage sont possibles, selon les besoins particuliers de l'enquête.

Pour apprendre comment déterminer la taille de l'échantillon nécessaire pour obtenir un degré de précision donné et comment comparer l'efficience de différents plans d'échantillonnage en comparant les effets de plan, le lecteur consultera le **Chapitre 8 - Calcul de la taille de l'échantillon et répartition.**

Bibliographie

Bebbington, A.C. 1975. A Simple Method of Drawing a Sample without Replacement. *Applied Statistics,* 24(1).

Binder, D.A. 1998. Les enquêtes longitudinales : Pourquoi ces enquêtes sont-elles différentes de toutes les autres ? *Techniques d'enquête,* 24(2): 107-115.

Brewer K.R.W et M. Hanif. 1983. Sampling with Unequal Probabilities. Springer-Verlag, New York.

Cochran, W.G. 1977. *Sampling Techniques*. John Wiley and Sons, New York.

Conner, W.S. 1966. An Exact Formula for the Probability that Two Specified Sample Units Will Occur in a Sample Drawn with Unequal Probabilities and Without Replacement. *Journal of the American Statistical Association,* 61: 385-390.

Cox, B.G., D.A. Binder, B.N. Chinnappa, A. Christianson, M.J. Colledge et P.S. Kott, Éds. 1995. *Business Survey Methods.* John Wiley and Sons, New York.

Droesbeke, J.-J., B. Fichet et P. Tassi, (1987). *Les Sondages*. Economica, Paris.

Fellegi, I.P. (1963). Sampling with Varying Probabilities Without Replacement Rotating and Non-Rotating Samples. *Journal of the American Statistical Association,* 58: 183-201.

Fink, A. (1995). *The Survey Kit*. Sage Publications, California.

Fowler, F.J. 1984. *Survey Research Methods*. 1. Sage Publications, California.

Gambino, J.G., M.P. Singh, J. Dufour, B. Kennedy et J. Lindeyer. 1998. *Méthodologie de l'enquête sur la population active du Canada.* Statistique Canada. 71-526.

Gray, G.B. 1971. Joint Probabilities of Selection of Units in Systematic Samples. *Proceedings for the American Statistical Association*. 271-276.

Hidiroglou, M.A. 1994. Sampling and Estimation for Establishment Surveys: Stumbling Blocks and Progress. *Proceedings of the Section on Survey Research Methods*. American Statistical Association. 153-162.

Hidiroglou, M.A. et G.B. Gray. 1980. Construction of Joint Probabilities of Selection for Systematic P.P.S. Sampling. *Applied Statistics,* 29(1): 663-685.

Hidiroglou, M.A. et K.P. Srinath. 1993. Problems Associated with Designing Sub-Annual Business Surveys. *Journal of Economic Statistics,* 11: 397-405.

Horvitz, D.G. et D.J. Thompson. 1952. A Generalization of Sampling Without Replacement from a Finite Universe. *Journal of the American Statistical Association*. 47: 663-685.

Kalton, G., J. Kordos et R. Platek, Éds. 1992. *Small Area Statistics and Survey Designs*. Central Statistical Office, Warsaw. 31-75.

Kasprzyk, D., G.J. Duncan, G. Kalton et M.P. Singh, Éds. 1989. *Panel Surveys*. John Wiley and Sons, New York.

Kish, L. 1965. *Survey Sampling*. John Wiley and Sons, New York.

Lavallée, P. 1998. *Théorie et Applications des enquêtes longitudinales*, Notes de cours 411F, Statistique Canada.

Levy, P. et S. Lemeshow. 1991. *Sampling of Populations*. John Wiley and Sons, New York.

Lohr, S. 1999. *Sampling: Design and Analysis*. Duxbury Press, U.S.A.

McLeod, A.I. et D.R. Bellhouse. 1983. A Convenient Algorithm for Drawing a SRS. *Applied Statistics,* 32(2).

Moser C.A. et G. Kalton. 1971. *Survey Methods in Social Investigation*. Heinemann Educational Books Limited, London.

Rao, J.N.K, H.O. Hartley et W.G. Cochran. 1962. On a Simple Procedure of Unequal Probability Sampling Without Replacement. *Journal of the Royal Statistical Society,* B, 27: 482-490.

Särndal, C.E., B. Swensson et J. Wretman. 1992. *Model Assisted Survey Sampling*. Springer-Verlag, New York.

Satin, A. et W. Shastry. 1993. *L'échantillonnage : un guide non mathématique – Deuxième édition.* Statistique Canada. 12-602F.

Stuart, A. 1968. *Basic Ideas of Scientific Sampling*. Charles Griffin and Company Limited, London.

Thompson, M. 1997. *Theory of Sample Surveys*. Chapman and Hill, United Kingdom.

Thompson, S.K. 1992. *Sampling*. John Wiley and Sons, New York.

Yates, F. et P.M. Grundy. 1953. Selection Without Replacement from Within Strata with Probability-proportional-to-size. *Journal of the Royal Statistical Society*. B, 15: 235-261.

PUBLICATIONS ÉLECTRONIQUES
DISPONIBLES À

www.statcan.gc.ca

Chapitre 7 - Estimation

7.0 Introduction

Il est indiqué au **Chapitre 1 - Introduction à l'enquête** que l'étude des caractéristiques d'une population est habituellement la motivation du client. ***L'estimation est un moyen que l'organisme statistique utilise pour obtenir des valeurs de la population d'intérêt et tirer des conclusions sur cette population à partir de l'information obtenue d'un échantillon .***

Le principe sous-jacent à l'estimation dans une enquête probabiliste est que chaque unité de l'échantillon représente non seulement elle-même, mais aussi plusieurs unités de la population de l'enquête. Le nombre moyen d'unités de la population que représente une unité de l'échantillon est souvent intitulé *poids de base* ou *pondération d'après le plan* pour cette unité. Déterminer la pondération est un important volet du processus d'estimation. Les poids de base peuvent servir à l'estimation, mais la majorité des enquêtes produisent une série de poids d'estimation en ajustant les poids de base. Tenir compte des non-réponses et utiliser les données auxiliaires sont les deux justifications habituelles des ajustements.

Lorsque les d'estimation ont été calculés, ils sont appliqués aux données de l'échantillon pour déterminer les *estimations*. Des mesures sommaires de la population, par exemple les *totaux, moyennes* et *proportions,* sont habituellement estimées pour un large éventail de caractéristiques obtenues des unités de l'échantillon. Ces caractéristiques, souvent intitulées variables en théorie statistique, peuvent être qualitatives, par exemple le sexe ou l'état civil, ou quantitatives, notamment l'âge ou le revenu. Différentes formules sont appropriées pour l'estimation des mesures sommaires, selon le genre de données.

Déterminer l'importance de l'erreur d'échantillonnage dans l'estimation est un volet important de l'estimation. Elle donne une mesure de la qualité des estimations de l'enquête pour un plan d'échantillonnage en particulier. L'erreur d'échantillonnage peut être estimée seulement si l'échantillonnage est probabiliste.

L'objectif de ce chapitre est d'illustrer comment calculer les poids, établir des estimations des mesures sommaires et des estimations de leur erreur d'échantillonnage pour les enquêtes avec échantillonnage probabiliste.

7.1 Pondération

La première étape de l'estimation est l'attribution d'un poids à chaque unité échantillonnée ou à chaque unité échantillonnée répondante. La *poids de base* peut être considéré comme ***le nombre moyen d'unités dans la population de l'enquête que chaque unité échantillonnée représente*** et elle est déterminée par le plan d'échantillonnage. La pondération du plan, w_d (où *d* représente le *plan, design* en anglais), pour une unité de l'échantillon est l'inverse de sa probabilité d'inclusion, π. On se souviendra que la probabilité de sélection d'une unité, pour un plan d'échantillonnage à plusieurs degrés ou à plusieurs phases est le produit des probabilités de sélection à chaque degré ou phase. Dans un échantillon à deux phases où la probabilité de sélection d'une unité est π_1 à la première phase et π_2 à la deuxième phase, la pondération du plan pour une unité de l'échantillon est donc :

$$w_d = \frac{1}{\pi_1} \times \frac{1}{\pi_2}.$$

Les données de l'échantillon d'une enquête sont généralement entrées dans un fichier comprenant un enregistrement pour chaque unité échantillonnée. Nous savons que chaque unité de l'échantillonnage probabiliste a une probabilité connue, π, d'être échantillonnée. Si cette probabilité d'inclusion est, par exemple, une sur 50, chaque unité sélectionnée représente donc une moyenne de 50 unités de la population de l'enquête et le poids de base est $w_d = 50$. Si le poids est un nombre entier, un moyen de produire des estimations pour la population serait de recopier chaque enregistrement pour qu'il y ait 50 copies de chacun et de calculer ensuite les valeurs d'intérêt (par exemple, les moyennes, les totaux, les ratios, etc.) de ce fichier. La répétition devient plus difficile si le poids n'est pas une valeur entière. (Si deux unités sur cinq sont sélectionnées à l'aide de l'échantillonnage aléatoire simple, par exemple, le poids de base est donc $w_d = 2{,}5$). Il est en général plus facile d'ajouter une variable de pondération à l'enregistrement de chaque unité de l'échantillon.

L'étude de la pondération commencera par les plans d'échantillonnage avec probabilité égale qui sont le cas de pondération le plus simple.

7.1.1 Pondération pour plans d'échantillonnage avec probabilité égale

Les plans d'échantillonnage sont considérés autopondérés lorsque les poids de base sont les mêmes pour toutes les unités de l'échantillon. C'est le cas lorsque chaque unité a la même probabilité d'inclusion. Dans un plan d'échantillonnage autopondéré, si aucun ajustement ultérieur n'est apporté aux poids de base (p. ex., pour les non-réponses ou les données auxiliaires), les poids peuvent être ignorés pour produire *certaines* statistiques comme les proportions et les moyennes. Le calcul des totaux exige simplement que le total de l'échantillon soit multiplié par le poids de base.

Quels plans d'échantillonnage à un degré sont autopondérés? Les échantillons aléatoires simples (EAS) et les échantillons systématiques sont autopondérés parce que chaque unité a une chance égale d'être incluse dans l'échantillon. Dans un plan stratifié, un plan autopondéré est obtenu, par exemple, si un EAS est sélectionné dans chaque strate et si la taille de l'échantillon de chaque strate est proportionnelle à la taille de la population de la strate. La fraction d'échantillonnage est donc la même dans chaque strate et toutes les unités de la population ont la même probabilité d'inclusion. (Cette répartition de l'échantillon entre les strates est intitulée répartition proportionnelle à N et fait l'objet d'une étude plus détaillée au **Chapitre 8 - Calcul de la taille de l'échantillon et répartition**.)

Exemple 7.1 : EAS stratifié avec répartition proportionnelle à N

Supposons qu'une population de $N = 1\,000$ personnes est répartie en deux strates dans la base de sondage. La première strate est composée de $N_1 = 400$ hommes et la deuxième, de $N_2 = 600$ femmes. Un échantillon total de $n = 250$ est tiré des deux strates et l'échantillon est réparti proportionnellement à la taille de chaque strate. La fraction d'échantillonnage de chaque strate est donc équivalente à $n/N = 250/1\,000 = 1/4$.

Tableau 1 : EAS stratifié avec répartition proportionnelle à N

Strate	Taille de la population	Taille de l'échantillon
Homme	$N_1 = 400$	$n_1 = 100$
Femme	$N_2 = 600$	$n_2 = 150$
Total	$N = 1\,000$	$n = 250$

Voici les probabilités d'inclusion dans chaque strate :

Strate 1, Hommes :
$$\pi_1 = \frac{n_1}{N_1} = \frac{100}{400} = \frac{1}{4}$$

Strate 2, Femmes :
$$\pi_2 = \frac{n_2}{N_2} = \frac{150}{600} = \frac{1}{4}$$

La probabilité d'être sélectionné est donc la même pour chacun, ainsi que le poids de base, $w_d = 1/\pi = 4$.

Dans un plan d'échantillonnage à plusieurs degrés, l'autopondération est obtenue en sélectionnant des grappes avec probabilité proportionnelle à la taille (PPT) à tous les degrés, à l'exception du dernier. Un nombre déterminé d'unités dans une grappe sont sélectionnées au dernier degré (p. ex., choisir toujours n=5 au dernier degré). L'échantillonnage PPT est souvent utilisé dans les plans à plusieurs degrés parce qu'il peut donner un échantillon autopondéré et permettre de contrôler la taille de l'échantillon.

Un exemple de plan d'échantillonnage autopondéré à deux phases serait un EAS, un échantillon systématique ou un échantillon stratifié avec répartition proportionnelle à N obtenu à chaque phase.

Les plans d'échantillonnage ont été étudiés au **Chapitre 6 - Plans d'échantillonnage**.

7.1.2 Pondération pour plans d'échantillonnage avec probabilités inégales

La simplicité des plans autopondérés est intéressante, mais il n'est pas toujours possible ou souhaitable de sélectionner un échantillon autopondéré. Dans un plan stratifié aux fins d'une enquête nationale, par exemple, pour des régions plus petites, la répartition proportionnelle à N peut donner des tailles d'échantillon insuffisantes et, pour les régions plus grandes, elle peut donner des échantillons trop gros.

L'exemple suivant illustre les poids de base pour un plan d'échantillonnage stratifié dont la taille de l'échantillon de chaque strate n'est pas proportionnelle à la taille de la population de la strate. (Le **Chapitre 8 - Calcul de la taille de l'échantillon et répartition** donne davantage de détails sur la répartition non proportionnelle.)

Exemple 7.2 : EAS stratifié avec répartition non proportionnelle

Aux fins d'une enquête sur les transports en commun, la population de N=1 100 personnes est répartie en deux strates géographiques. Étant donné que les personnes qui habitent en milieu rural et urbain peuvent être très différentes quant à l'information obtenue pour l'enquête, un plan d'échantillonnage stratifié est utilisé. La strate urbaine est de la taille N_1=1 000 et la strate rurale, N_2=100. Un échantillon de n=250 personnes est sélectionné : n_1=200 dans la strate urbaine et n_2=50 dans la strate rurale. Quelles sont les poids de base pour les personnes échantillonnées?

Tableau 2 : EAS stratifié avec répartition non proportionnelle

Strate	Taille de la population	Taille de l'échantillon
Urbain	N_1 = 1 000	n_1 = 200
Rural	N_2 = 100	n_2 = 50
Total	N = 1 100	n = 250

Les probabilités d'inclusion de chaque strate sont calculées comme suit:

Strate 1, Urbain :
$$\pi_1 = \frac{n_1}{N_1} = \frac{200}{1\,000} = \frac{1}{5}$$

Strate 2, Rural :
$$\pi_2 = \frac{n_2}{N_2} = \frac{50}{100} = \frac{1}{2}$$

Dans le fichier de l'échantillon, chaque répondant de la strate du milieu urbain a un poids de base de $w_{d,1} = 5$ et chaque répondant de la strate du milieu rural a un poids de base de $w_{d,2} = 2$.

Dans un échantillonnage à plusieurs degrés ou phases, la pondération du plan dans l'ensemble est calculée en multipliant la probabilité de sélection à chaque degré ou phase et en appliquant ensuite l'inverse. Dans un échantillon par grappes à deux degrés, par exemple, supposons que vous sélectionnez un EAS de n_1=10 dans un ensemble de N_1=100 grappes au premier degré et un EAS de n_2=30 unités dans chaque grappe au deuxième degré, le nombre d'unités dans chaque grappe étant N_2=60.

La probabilité de sélection au premier degré est donnée par:

$$\pi_1 = \frac{n_1}{N_1} = \frac{10}{100} = \frac{1}{10},$$

et la probabilité au deuxième degré par:

$$\pi_2 = \frac{n_2}{N_2} = \frac{30}{60} = \frac{1}{2}.$$

Le poids de base est donc :

$$w_d = \frac{1}{\pi_1} \times \frac{1}{\pi_2} = 10 \times 2 = 20.$$

7.1.3 Ajustement de la pondération pour les non-réponses

Les non-réponses sont un problème dans toutes les enquêtes et elles se produisent lorsque, pour certaines raisons, l'information demandée aux unités échantillonnées n'est pas disponible, en tout ou en partie. Il est mentionné au **Chapitre 3 - Introduction au plan d'enquête** qu'il y a deux principaux types de non-réponse, la non-réponse partielle et la non-réponse totale. Il y a ***non-réponse partielle lorsque l'information est disponible pour certaines questions seulement***, par exemple, lorsque la personne répond à une partie seulement du questionnaire. L'imputation des valeurs manquantes est l'approche la plus commune dans ce cas. (Diverses approches d'imputation pour les non-réponses à une question ou partielles sont considérées au **Chapitre 10 - Traitement**.)

Cette section traite de la ***non-réponse totale, c'est-à-dire lorsque toutes les données ou presque d'une unité échantillonnée sont manquantes.*** Il s'agit de cas où l'unité de l'échantillon refuse de participer, où il est impossible d'établir un contact, où l'unité ne peut être repérée ou encore si l'information obtenue est inutile. La façon la plus facile de traiter ces non-réponses est de les ignorer. Dans certaines circonstances exceptionnelles, des proportions ou des moyennes estimées sans ajustement pour les non-réponses totales sont les mêmes que celles produites en appliquant un ajustement pour les non-réponses. Si l'on vous néglige de compenser pour les unités non répondantes, les totaux sont généralement sous-estimés (p. ex., la taille d'une population, le total des revenus ou le total d'acres récoltés).

La façon la plus commune de traiter la non-réponse totale est d'ajuster les poids de base en supposant que les unités répondantes représentent les unités répondantes et non répondantes. Cette mesure est raisonnable si l'on considère que les non-répondants sont équivalents aux répondants pour les caractéristiques mesurées dans l'enquête. Les poids de base pour les non-répondants sont ensuite redistribuées entre les répondants. Cette mesure est souvent appliquée à l'aide d'un facteur d'ajustement pour les non-réponses qui est multiplié par la poids de base, afin d'obtenir une pondération ajustée pour les non-réponses, ceci étant illustré dans l'exemple 7.3 ci-dessous.

On remarquera que les données de recensement peuvent aussi avoir un ajustement de pondération pour les non-réponses et les poids de base seraient alors équivalents à un, $w_d = 1$. Le biais de non-réponse a été étudié au **Chapitre 3 - Introduction au plan d'enquête** et au **Chapitre 5 - Conception du questionnaire**.

7.1.3.1 Facteurs d'ajustement de la pondération pour les non-réponses

Le facteur d'ajustement pour les non-réponses est habituellement défini comme le rapport entre la somme des poids dans l'échantillon original et la somme des poids des unités répondantes. Dans un plan d'échantillonnage autopondéré, il est équivalent au rapport entre le nombre d'unités de l'échantillon original et le nombre d'unités répondantes, et il est illustré ci-dessous.

Exemple 7.3 : Facteur d'ajustement pour les non-réponses d'un EAS, un groupe de non-réponses

Un EAS de n=25 personnes est sélectionné dans une population de N=100 personnes. Disons que le nombre d'unités répondantes est n_r. Seulement n_r=20 personnes ont donné toute l'information demandée. Quelles sont les poids ajustés pour les non-réponses des unités de l'échantillon?

1. La première étape est le calcul des probabilités d'inclusion pour un EAS :

$$\pi = \frac{n}{N} = \frac{25}{100} = \frac{1}{4}.$$

Le poids de base pour chaque unité échantillonnée est donc w_d=4.

2. La deuxième étape est le calcul du facteur d'ajustement pour les non-réponses.

Seulement n_r = 20 personnes sur n = 25 personnes sélectionnées ont donné l'information demandée et la taille de l'échantillon final est donc de 20 unités. Si les unités répondantes représentent les unités répondantes et non répondantes, le facteur d'ajustement pour les non-réponses est donné par:

$$\frac{n}{n_r} = \frac{25}{20} = 1{,}25.$$

3. La dernière étape est le calcul des poids ajustés pour les non-réponses.

Les poids ajustés pour les non-réponses, w_{nr}, sont le produit des poids de base et du facteur d'ajustement pour les non-réponses :

$$w_{nr} = w_d \frac{n}{n_r} = 4 \times 1{,}25 = 5.$$

Chaque répondant représente donc cinq personnes dans la population de l'enquête. *Un poids final de 5* est attribué à chaque unité dans le fichier des données.

Si l'hypothèse selon laquelle tous les non-répondants sont équivalents aux répondants pour les caractéristiques mesurées dans l'enquête est appropriée, le même facteur d'ajustement pour les non-réponses peut être utilisé pour toutes les unités répondantes (comme ci-dessus). Il y a habituellement des sous-groupes, cependant, qui ont tendance à répondre différemment et qui ont différentes caractéristiques, et un ajustement identique pour tous les répondants peut donc biaiser les résultats. Les ménages unifamiliaux, par exemple, ont souvent des taux de réponse inférieurs à ceux des ménages multiples et ils ont des caractéristiques différentes : il faudrait donc procéder à des ajustements distincts pour les non-réponses.

Un facteur d'ajustement différent pour les non-réponses est appliqué dans l'exemple ci-dessous à chacune des deux strates : urbaine et rurale. Des caractéristiques d'intérêt différentes entre les strates justifient cette mesure.

Exemple 7.2 (suite) : Facteur d'ajustement pour les non-réponses de l'EAS stratifié (où le groupe des non-réponses correspond à la strate)

Seulement $n_{r,1}$=150 personnes dans la strate du milieu urbain et $n_{r,2}$=40 personnes dans la strate du milieu rural ont donné l'information demandée pendant la collecte des données. Quelles sont les poids ajustés pour les non-réponses de ces répondants?

Tableau 3 : EAS stratifié avec les non-réponses

Strate	Taille de la population	Taille de l'échantillon	Nombre de répondants
Urbain	$N_1 = 1\,000$	$n_1 = 200$	$n_{r,1} = 150$
Rural	$N_2 = 100$	$n_2 = 50$	$n_{r,2} = 40$

1. La pondération du plan d'échantillonnage dans chaque strate est $w_{d,1}$=5 pour la strate du milieu urbain et $w_{d,2}$=2 pour la strate du milieu rural.

2. Un facteur d'ajustement pour les non-réponses est calculé à chaque strate, comme suit :

Strate 1, Urbain :
$$\frac{n_1}{n_{r,1}} = \frac{200}{150} = 1{,}33$$

Strate 2, Rural :
$$\frac{n_2}{n_{r,2}} = \frac{50}{40} = 1{,}25$$

3. La pondération ajustée pour les non-réponses dans chaque strate est le produit de la pondération du plan d'échantillonnage et du facteur d'ajustement pour les non-réponses.

Strate 1, Urbain :
$$w_{nr,1} = w_{d,1}\frac{n_1}{n_{r,1}} = 5 \times 1{,}33 = 6{,}67$$

Strate 2, Rural :
$$w_{nr,2} = w_{d,2}\frac{n_2}{n_{r,2}} = 2 \times 1{,}25 = 2{,}5$$

Dans le fichier de l'échantillon, on attribue à chaque répondant de la strate du milieu urbain un poids final de 6,67 et à chaque répondant de la strate du milieu rural, un poids final de 2,5.

Lors du calcul du facteur d'ajustement pour les non-réponses, il est important de tenir compte du fait que certaines unités échantillonnées peuvent se révéler hors du champ de l'enquête (c.-à-d. qu'elles ne font pas partie de la population cible). Dans une enquête sur les entreprises de détail, par exemple, certains renseignements dans la base de sondage peuvent être inexacts et une unité échantillonnée peut en fait être un grossiste. L'ajustement pour les non-réponses devrait être calculé seulement pour les unités admissibles parce que les unités hors du champ de l'enquête représentent habituellement d'autres unités hors du champ de l'enquête dans la base de sondage. Il n'est pas prévu que les unités hors du champ de l'enquête répondront au sondage et il faut donc présumer que leur taux de non-réponse sera 100 %. Dans l'exemple ci-dessus, il est supposé que tous les non-répondants sont admissibles, mais les facteurs d'ajustement pertinents pourraient être très différents selon le nombre de non-répondants admissibles considéré exact.

Il vaut mieux regrouper, pour toutes ces méthodes, les enregistrements semblables comme c'est le cas pour les ajustements de pondération pour les non-réponses (voir le **Chapitre 7 - Estimation**). Ces regroupements sont intitulés classes d'imputation.

L'ajustement pour les non-réponses devrait être fait distinctement pour des groupes de répondants semblables lorsque chaque groupe de répondants peut représenter les non-répondants de ce groupe. Ces groupements peuvent être par strate ou par strate a posteriori (voir la section suivante), ou une analyse peut être faite pour déterminer les groupements (p. ex., test du chi au carré ou régression logistique). Groves et Couper (1998) couvrent en détail la non-réponse dans les enquêtes auprès des ménages. .

Le test du khi carré et la régression logistique sont étudiés au **Chapitre 11 - Analyse des données de l'enquête.**

7.1.4 Recours à l'information auxiliaire pour ajuster les pondérations

Les poids de base multipliés par le facteur d'ajustement pour les non-réponses peuvent servir à déterminer les pondérations finales et les estimations des caractéristiques voulues de l'enquête. L'information sur la population de l'enquête peut cependant être disponible à d'autres sources, par exemple, à la suite d'un recensement précédent. Cette information peut aussi être intégrée au processus de pondération.

Il y a deux principales raisons pour utiliser les données auxiliaires lors de l'estimation. Premièrement, il est souvent important que les estimations de l'enquête correspondent aux totaux connus de la population ou aux estimations d'une autre enquête plus fiable. De nombreuses enquêtes sociales, par exemple, ajustent leurs estimations de l'enquête pour qu'elles soient conformes aux estimations (âge, répartition des sexes, etc.) du recensement de la population le plus récent. L'information auxiliaire peut aussi être obtenue à partir de données administratives ou d'une autre enquête considérée plus fiable parce que sa taille d'échantillon est plus large ou parce qu'il faut respecter ses estimations publiées.

Deuxièmement, les données auxiliaires sont utilisées pour améliorer la précision des estimations. En effet, un estimateur ayant une petite variance d'échantillonnage, une mesure de l'erreur d'échantillonnage, est considéré précis. Au **Chapitre 6 - Plans d'échantillonnage**, nous avons considéré l'importance de l'utilisation des données auxiliaires lors de la conception de l'échantillon, afin d'améliorer l'efficience statistique de la stratégie d'échantillonnage. À l'étape du plan d'échantillonnage cependant, l'information auxiliaire doit être disponible pour toutes les unités de la base de sondage. À l'étape de l'estimation, les données auxiliaires peuvent servir à améliorer la précision des estimations si les variables auxiliaires ont été mesurées pour les unités de l'enquête et si les totaux ou les estimations de la population pour ces variables auxiliaires sont disponibles à une autre source fiable.

Si l'âge d'une personne n'est pas disponible dans la base de sondage, par exemple, il ne peut servir à stratifier la population. Si cette information est obtenue lors de l'enquête cependant, les estimations de l'enquête peuvent être ajustées pour correspondre à la répartition des âges dans le recensement. Si l'âge est corrélé avec d'autres variables obtenues pendant l'enquête (p. ex., les variables sur la santé), son utilisation comme données auxiliaires peut améliorer la précision des estimations. Cette notion est illustrée aux sections suivantes.

L'information auxiliaire peut aussi servir pour apporter d'autres corrections lorsqu'il y a des taux de non-réponses différents dans divers sous-groupes de la population. Elle peut aussi aider à ajuster s'il y a des défauts de couverture parce que la population du sondage est différente de la population cible.

Il y a trois exigences élémentaires pour utiliser avec succès des données auxiliaires à l'étape de l'estimation:

- les données auxiliaires doivent être bien corrélées avec les variables de l'enquête,

- les sources externes d'information sur la population doivent être exactes,

- il faut faire la collecte de l'information auxiliaire pour toutes les unités répondantes de l'échantillon quand on ne connaît que les totaux de la population.

Plus les variables de l'enquête sont étroitement corrélées avec les données auxiliaires disponibles, plus les estimations qui utilisent les données auxiliaires seront efficientes. La précision de l'information auxiliaire est importante. Non seulement les données doivent être fiables, mais il est aussi important que la source de données externe couvre la même population cible et qu'elle soit établie selon des concepts, définitions et périodes de référence comparables, etc., à ceux de l'enquête.

Les sections suivantes décrivent comment les données auxiliaires peuvent servir aux estimations. L'**Annexe A - Données administratives** explore plus avant l'utilisation des données administratives.

7.1.4.1 Stratification *a posteriori*

La *stratification a posteriori* est utilisée pour ajuster les poids de l'enquête à l'aide de variables qui conviennent à la stratification, mais qui ne pouvaient être utilisées à l'étape de la conception du plan parce que les données n'étaient pas disponibles, ou parce que de l'information plus fiable et à jour est devenue disponible après la sélection de l'échantillon. La stratification *a posteriori* est utilisée lorsque les données auxiliaires sont numériques, par exemple, le nombre d'hommes et de femmes dans la population. Elle est plus efficace pour diminuer la variance d'échantillonnage lorsque les moyennes des variables d'intérêt de la population sont aussi différentes que possible entre les strates *a posteriori*. Il ne faut pas oublier qu'il est préférable, si possible, de stratifier à l'étape de la conception du plan au lieu d'appliquer la stratification *a posteriori*.

L'exemple suivant révèle comment utiliser la stratification *a posteriori* pour améliorer l'estimation du nombre de fumeurs dans une entreprise.

Exemple 7.4 : Ajustement de la pondération pour stratification *a posteriori* de l'EAS

Supposons qu'une enquête est faite pour obtenir de l'information sur le tabagisme des employés dans une petite entreprise. Un EAS de n=25 personnes est sélectionné dans une liste de N=78 employés. Il n'y a pas d'information auxiliaire disponible qui peut servir à la stratification à l'étape de la conception du plan.

La collecte de l'information est faite sur le tabagisme, et l'âge et le sexe de chaque répondant sont aussi disponibles. Sur n=25 personnes à l'origine, n_r=15 répondent et la répartition suivante est faite :

Tableau 4 : EAS, non-réponse et stratification *a posteriori*

	Strate *a posteriori* 1, hommes	**Strate *a posteriori* 2, femmes**	**Nombre de répondants**
Tous les employés	3	12	15
Employés fumeurs	1	7	8

1. La probabilité d'inclusion de chaque unité échantillonné est donnée par:

$$\pi = \frac{n}{N} = \frac{25}{78} = 0{,}32$$

Le poids de base est donc w_d=1/π=3,12.

2. Le facteur d'ajustement pour les non-réponses, en supposant que chacun dans l'enquête a la même probabilité de répondre au sondage (c.-à-d., un groupe de non-réponses) est donné par:

$$\frac{n}{n_r} = \frac{25}{15} = 1{,}67$$

3. On obtient les poids ajustés pour les non-réponses en faisant:

$$w_{nr} = w_d \frac{n}{n_r} = 3{,}12 \times 1{,}67 = 5{,}2$$

Tous les répondants ont donc la même pondération ajustée pour les non-réponses, w_r=5,2. Les estimations suivantes de l'enquête sont faites à l'aide de ces poids (consultez la section 7.2 pour obtenir des détails et apprendre comment faire des estimations d'enquête) :

Tableau 5 : Estimation de l'enquête et ajustement pour les non-réponses

	Hommes	**Femmes**	**Total**
Nombre d'employés	15,6	62,4	78,0
Nombre de fumeurs	5,2	36,4	41,6
Proportion de fumeurs	0,33	0,59	0,53

Les pondérations ajustées pour les non-réponses donnent une estimation d'environ 16 hommes et 62 femmes qui travaillent dans l'entreprise, ainsi qu'une estimation de 33 % de fumeurs et de 59 % de fumeuses dans l'entreprise. Supposons que l'information auxiliaire suivante devienne disponible après

l'enquête : 42 hommes et 36 femmes travaillent dans l'entreprise. Les estimations de l'enquête sont très différentes de ces valeurs réelles.

L'organisme statistique veut que les estimations de l'enquête soient conformes au nombre connu d'hommes et de femmes. L'organisme soupçonne aussi que le tabagisme est lié au sexe d'une personne et la stratification *a posteriori* pourrait améliorer la précision des estimations sur le tabagisme. Si cette information avait en fait été disponible au moment de la conception du plan, l'organisme statistique aurait stratifié par sexe. Que faire?

Il est possible de stratifier l'échantillon après le fait et de déterminer des pondérations stratifiées *a posteriori* à utiliser au moment de l'estimation. La pondération stratifiée *a posteriori*, w_{pst}, est le produit de la pondération ajustée pour les non-réponses, w_{nr}, et du facteur d'ajustement stratifié *a posteriori*.

Le facteur d'ajustement stratifié *a posteriori* est déterminé pour chaque strate *a posteriori*. Ce facteur correspond au rapport entre le nombre d'unités de la population dans la strate *a posteriori*, N, et le nombre estimé d'unités de la population dans la strate *a posteriori*, $\hat{N}$, qui est estimé à l'aide des pondérations du plan ajusté pour les non-réponses.

4. Le facteur d'ajustement pour stratification *a posteriori* se calcule comme suit :

Strate *a posteriori* 1, Hommes :

$$\frac{N_{hommes}}{\hat{N}_{hommes}} = \frac{42}{15{,}6} = 2{,}69$$

Strate *a posteriori* 2, Femmes :

$$\frac{N_{femmes}}{\hat{N}_{femmes}} = \frac{36}{62{,}4} = 0{,}58$$

(Remarque : Cet exemple vaut pour l'EAS, mais la même formule, $N / \hat{N}$, est utilisée pour des pondérations de plan d'échantillonnage plus complexes.)

Lorsqu'il est appliqué à la pondération ajustée pour les non-réponses, il donne les pondérations stratifiées *a posteriori* finales :

Strate *a posteriori* 1, Hommes :

$$w_{pst,hommes} = w_{nr} \times \frac{N_{hommes}}{\hat{N}_{hommes}} = 5{,}2 \times 2{,}69 = 14$$

Strate *a posteriori* 2, Femmes :

$$w_{pst,femmes} = w_{nr} \times \frac{N_{femmes}}{\hat{N}_{femmes}} = 5{,}2 \times 0{,}58 = 3$$

Voici maintenant les estimations de l'enquête à l'aide des pondérations stratifiées *a posteriori* :

Tableau 6 : Estimations de l'enquête avec ajustement pour les non-réponses et la stratification *a posteriori*

	Hommes	Femmes	Total
Nombre de personnes	42	36	78
Nombre de fumeurs	14	21	35
Proportion de fumeurs	0,33	0,59	0,45

Les estimations du nombre d'hommes et de femmes sont maintenant conformes aux totaux connus d'hommes et de femmes dans l'entreprise et, dans la mesure où le sexe est lié au nombre et à la proportion de fumeurs, il est possible d'améliorer énormément la précision. Remarquez que la proportion de fumeurs dans chaque strate *a posteriori* n'a pas changé, mais que la proportion de fumeurs dans la population totale qui comprend plus d'une strate *a posteriori* a changé.

7.1.4.2 Estimation par quotient

Une méthode souvent appliquée dans les enquêtes pour intégrer de l'information auxiliaire, afin d'améliorer les estimations de l'enquête, est *l'estimation par quotient*. Lorsque les données sont des nombres, l'estimation par quotient correspond à la stratification *a posteriori*. Dans le cas d'une estimation par quotient en général, les poids des enregistrements dans un groupe de classification sont ajustées par un facteur multiplicatif. Ce facteur est le rapport entre l'estimation tirée des données auxiliaires et l'estimation de l'enquête pour la même variable, pour le groupe de classification.

Si l'objectif d'une enquête est d'estimer le nombre d'acres de blé dans des régions en particulier, par exemple, le nombre total d'acres dans chaque région pourrait être une variable auxiliaire utile. Si le nombre d'acres de blé est fortement corrélé avec le total des terres dans la région, cette information auxiliaire pourrait améliorer les estimations du nombre d'acres de blé. À l'aide de l'estimation par quotient, le facteur d'ajustement pour chaque région serait la superficie totale des terres de la région divisée par l'estimation dans l'échantillon du total de la superficie des terres de la région (c.-à-d. que le facteur d'ajustement serait $Y / \hat{Y}$) . Dans l'exemple 7.4, les quotients sont le nombre d'hommes divisé par le nombre estimé d'hommes et le même quotient pour les femmes et, ceux-ci étant des nombres, ils correspondent à la stratification *a posteriori*. Tout comme les méthodes précédentes, l'estimation par quotient peut être appliquée distinctement par strate si l'information auxiliaire est disponible à ce niveau et si la taille de l'échantillon de chaque strate est suffisante.

7.1.4.3 Ajustements de la pondération plus complexes : calibration et régression généralisée

L'estimation par quotient est fondée sur l'hypothèse selon laquelle il y a une simple relation multiplicative entre les caractéristiques de l'enquête et une variable auxiliaire (c.-à-d. que 2,7 fois plus d'hommes signifie 2,7 fois plus de fumeurs). Cette hypothèse peut cependant ne pas être vraie dans toutes les situations. Premièrement, le recours à une seule variable auxiliaire peut être insuffisant pour donner de bonnes estimations. Deuxièmement, la relation entre la variable estimée et la variable auxiliaire peut être plus complexe qu'une simple relation multiplicative. Dans ce cas, l'*estimation par régression* peut être utilisée. L'estimation par régression est une approche plus complexe qui permet à l'analyste de considérer des modèles plus perfectionnés, y compris des modèles ayant plus d'une variable auxiliaire.

L'estimation par quotient pose un autre problème : il peut être nécessaire de garantir que les totaux de l'échantillon pondéré correspondent aux totaux connus de la population pour *plus d'une* caractéristique. Si les totaux de l'échantillon pondéré doivent correspondre aux totaux de référence pour plus d'une caractéristique, il faut appliquer une méthode intitulée *calibration*. La situation se produit, par exemple, lorsque des strates *a posteriori* sont formées à l'aide de plus d'une variable et que seuls les totaux marginaux de la population pour chaque variable sont connus. Si les données sur la population étaient disponibles en nombre par groupe d'âge et par sexe, par exemple, mais si le nombre d'hommes et de femmes dans chaque groupe d'âge est inconnu, la méthode de stratification *a posteriori* décrite auparavant ne pourrait être appliquée en utilisant les deux caractéristiques. (L'estimation par quotient et la stratification *a posteriori* sont une calibration qui utilise une seule variable.)

Un prolongement de la méthode par quotient intitulée *méthode itérative du quotient* permet à l'organisme statistique d'établir les pondérations ajustées pour que les estimations soient très près des nombres de référence marginaux pour chaque caractéristique. Vous obtiendrez une description de cette méthode en consultant Deming et Stephan (1940), Arora et Brackstone (1977), Bankier (1978), Brackstone et Rao (1978), Binder (1988).

Des procédures d'estimation généralisées ont aussi été élaborées pour obtenir des estimations qui conviennent aux totaux de référence pour de nombreuses caractéristiques simultanément. Lorsque ces techniques générales sont appliquées, le processus qui garantit que les nombres correspondent aux totaux de référence est intitulé *calibration* et les ajustements de pondération obtenus sont intitulés *facteurs de calibration*.

Ces techniques, comme l'estimation par quotient et la stratification *a posteriori*, donnent des ajustements des poids de base. Les poids finaux utilisés pour calculer les estimations deviennent le produit des poids ajustés pour les non-réponses et des facteurs de calibration. Les procédures d'estimation généralisées sont hors de la portée de ce document. Le lecteur découvrira la théorie de l'estimation par régression généralisée dans Särndal, Swensson et Wretman (1992), Deville et Särndal (1992) et Hidiroglou et Särndal (1998). Le **Chapitre 11 - Analyse des données de l'enquête** présente une introduction à la régression linéaire.

7.2 Production d'estimations simples (totaux, moyennes et proportions)

Les exemples ont illustré jusqu'à maintenant comment calculer les poids de base et les ajuster pour les non-réponses et pour les données auxiliaires. Cette section explique comment obtenir des estimations à l'aide des poids finaux (poids d'estimation).

On a noté au **Chapitre 6 - Plans d'échantillonnage** qu'avec l'échantillonnage probabiliste, il est possible de déterminer la distribution d'échantillonnage de l'estimateur. Dans l'échantillonnage non probabiliste, étant donné que les probabilités de tirage des divers échantillons sont inconnues, la distribution d'échantillonnage ne peut être calculée. Pouvoir mesurer l'erreur d'échantillonnage est un volet important de l'estimation de l'enquête et l'une des principales raisons de procéder à un échantillonnage probabiliste.

7.2.1 Estimateurs pour divers genres de données

Des statistiques descriptives simples comme les totaux, moyennes et proportions, sont produites pour à peu près toutes les enquêtes. Des statistiques et des analyses plus complexes sont aussi habituellement nécessaires. Dans la majorité des enquêtes, des données sont obtenues pour un large éventail de variables qui peuvent être qualitatives (aussi intitulées nominales) ou quantitatives.

Quelques catégories seulement sont possibles pour certaines variables qualitatives, par exemple, le sexe ou l'état civil. Quant aux questions d'opinion, les réponses des participants sont souvent obtenues à l'aide d'une échelle d'agrément, par exemple, *vraiment d'accord, d'accord, ni pour ni contre, pas d'accord, vraiment pas d'accord*. Remarquez qu'avec les données nominales, chaque unité correspond à une seule catégorie.

Si l'unité de mesure indique des quantités comme des mètres ou des années, les données sont *quantitatives*. Les données quantitatives sont habituellement des réponses aux questions du genre *quelle quantité?* ou *quel nombre?*, c.-à-d. combien? Certains exemples sont l'âge, le nombre d'enfants, le nombre d'heures travaillées, les dépenses et les revenus, la tension artérielle.

Différents types d'estimateurs sont appropriés pour ces divers genres de variables. On produit habituellement des proportions et des comptes totaux pour des variables qualitatives, tandis que les moyennes et les totaux sont estimés pour des variables quantitatives. Dans cette section, les procédures

appliquées pour obtenir des estimations seront présentées distinctement pour les données qualitatives et quantitatives.

Outre le genre de données, une autre considération pendant l'estimation est la caractéristique déterminante de la population que ciblent les estimations. Des estimations peuvent être établies pour toute la population de l'enquête ou pour des sous-groupes ou *domaines* de la population en particulier (p. ex., les provinces). Si la classification originale des unités de l'échantillonnage a changé pendant la période écoulée entre l'échantillonnage et l'estimation, la nouvelle classification devrait être utilisée pour l'estimation des domaines.

Les réponses aux questions suivantes devraient aider à déterminer comment les estimations de l'enquête sont calculées :
- Quel genre de statistiques sont demandées? Un total, une moyenne, une proportion?
- Quel genre de données sont utilisées? Qualitatives ou quantitatives?
- Quelles sont les poids finaux ?
- Quels sont les domaines d'intérêt?

Les procédures d'estimation des totaux, moyennes et proportions, pour toute la population d'enquête et pour des domaines, sont décrites ci-dessous pour les données qualitatives et quantitatives.

Les estimateurs suivants peuvent être appliqués à tout plan d'échantillonnage probabiliste simple (p. ex., EAS, SYS) ou plus complexe. Il est important surtout que la pondération finale de chaque unité corresponde au correctement le plan d'échantillonnage.

i. Estimation d'un total de la population

L'estimation du *nombre total d'unités dans la population d'enquête* est calculée, pour les données qualitatives et quantitatives, en additionnant les poids finaux (ajustés) des unités répondantes :

$$\hat{N} = \sum_{i \in S_r} w_i$$

où i est la i^e unité répondante de l'échantillon, w_i, son poids final et S_r, l'ensemble des unités répondantes.

L'estimation d'une *valeur totale* pour les données quantitatives (p. ex., les dépenses totales) est le produit du poids final , w_i, et de la valeur, y_i, pour chaque unité répondante dont on fait la somme pour toutes les unités répondantes :

$$\hat{Y} = \sum_{i \in S_r} w_i y_i$$

ii. Estimation d'une moyenne de la population

L'estimation d'une valeur moyenne dans la population pour les données quantitatives est obtenue en additionnant le produit de la valeur observée et du poids final pour chaque unité répondante, et en divisant cette somme par celle des poids. Autrement dit, l'estimation de la moyenne dans la population est l'estimation de la valeur totale des données quantitatives divisée par l'estimation du nombre total d'unités dans la population.

$$\hat{\bar{Y}} = \frac{\sum_{i \in S_r} w_i y_i}{\sum_{i \in S_r} w_i} = \frac{\hat{Y}}{\hat{N}}$$

Remarque : Pour l'EAS ou le SYS ayant un taux de réponse de 100 % sans ajustement pour la pondération, l'estimateur se simplifie ainsi :

$$\hat{\bar{Y}} = \frac{\sum_{i \in S_r} y_i}{n}$$

iii. Estimation d'une proportion de la population

L'estimation de la proportion des unités dans la population de l'enquête ayant une caractéristique donnée, pour les données qualitatives, est obtenue en additionnant les poids des unités ayant cette caractéristique, et en divisant ce total par la somme des poids pour tous les répondants. Autrement dit, l'estimation de la proportion dans la population est l'estimation du nombre total des unités qui ont la caractéristique donnée divisée par l'estimation du nombre total d'unités dans la population.

$$\hat{P} = \frac{\sum_{i \in S_r \cap C} w_i}{\sum_{i \in S_r} w_i} = \frac{\hat{N}_C}{\hat{N}}$$

où C est l'ensemble des unités ayant la caractéristique donnée.

iv. Estimation pour les domaines de la population

Des estimations peuvent être demandées pour certains domaines, notamment le groupe d'âge, le type de logement, la taille du ménage ou la tranche de revenu.

- L'estimation de la taille de la population pour un domaine d'intérêt, tant pour les données qualitatives que quantitatives se calcule ainsi :

$$\hat{N}_{domaine} = \sum_{i \in S_r \cap domaine} w_i$$

- L'estimation d'un total de domaines pour les données quantitatives est donnée par :

$$\hat{Y}_{domaine} = \sum_{i \in S_r \cap domaine} w_i y_i$$

- On en déduit l'estimation d'une moyenne de domaines pour les données quantitatives :

$$\hat{\bar{Y}}_{domaine} = \frac{\sum_{i \in S_r \cap domaine} w_i y_i}{\sum_{i \in S_r \cap domaine} w_i} = \frac{\hat{Y}_{domaine}}{\hat{N}_{domaine}}$$

- De façon équivalente, l'estimation d'une proportion de domaines pour les données qualitatives ou quantitatives est donnée par :

$$\hat{P}_{domaine} = \frac{\sum_{i \in S_r \cap domaine \cap C} w_i}{\sum_{i \in S_r \cap domaine} w_i} = \frac{\hat{N}_{domaine \cap C}}{\hat{N}_{domaine}}$$

Ces procédures d'estimation sont illustrées dans les sections suivantes.

7.2.2 Estimations des totaux, moyennes et proportions

L'utilisation correcte des poids d'estimation est au cœur du processus d'estimation.

Exemple 7.6 : Estimation lorsque les poids finaux sont inégaux, EAS

Une enquête est menée pour obtenir de l'information sur une population d'exploitations agricoles (fermes). Un échantillon de n=10 exploitations est sélectionné à l'aide d'un plan d'échantillonnage stratifié. Les 10 exploitations agricoles répondent et il n'y a pas d'ajustement aux poids, le poids final étant donc égal au poids de base. Il faut obtenir des estimations à partir du fichier de données suivant :

Strate	Pondération finale	Genre de ferme	Revenu ($)
1	5,67	1	75 000
1	5,67	2	15 000
1	5,67	1	125 000
1	5,67	1	67 000
1	5,67	2	80 000
1	5,67	1	40 000
2	16,5	1	30 000
2	16,5	1	14 000
2	16,5	2	48 000
2	16,5	1	22 000

Genre de ferme
1=culture ($N_1 = 34$, $n_1 = 6$)
2=élevage ($N_2 = 66$, $n_2 = 4$)

- Le nombre d'exploitations agricoles dans la population est estimé à :

$$\begin{aligned}\hat{N} &= \sum_{i \in S_r} w_i \\ &= 5,67 + 5,67 + 5,67 + 5,67 + 5,67 + 5,67 + 16,5 + 16,5 + 16,5 + 16,5 \\ &= 100\end{aligned}$$

Remarque : Si les pondérations avaient été omises, le calcul erroné de l'estimation aurait donné 10.

- Le nombre estimé d'exploitations agricoles d'élevage (eae) est donnée par :

$$\hat{N}_{eae} = \sum_{i \in S_r \cap eae} w_i = 5,67 + 5,67 + 16,5 = 28$$

- On en déduit la proportion estimée d'exploitations agricoles d'élevage :

$$\hat{P} = \frac{\sum_{i \in S_r \cap eae} w_i}{\sum_{i \in S_r} w_i} = \frac{28}{100} = 0{,}28$$

- Le revenu total de la population entière d'exploitations agricoles est estimé à:

$$\begin{aligned}\hat{Y} &= \sum_{i \in S_r} w_i y_i \\ &= 5{,}67 \times 75\,000 + 5{,}67 \times 15\,000 + \ldots. + 16{,}5 \times 22\,000 \\ &= 4\,160\,340\end{aligned}$$

- On estime le revenu moyen pour toute la population d'exploitations agricoles en faisant :

$$\hat{\bar{Y}} = \frac{\sum_{i \in S_r} w_i y_i}{\sum_{i \in S_r} w_i} = \frac{4\,160\,340}{100{,}02} = 41\,595$$

- L'estimation du revenu total des exploitations agricoles d'élevage est donnée par :

$$\begin{aligned}\hat{Y}_{eae} &= \sum_{i \in S_r \cap eae} w_i y_i \\ &= 5{,}67 \times 84\,000 + 5{,}67 \times 48\,000 + 16{,}5 \times 23\,000 \\ &= 1\,330\,650\end{aligned}$$

- L'estimation du revenu moyen des exploitations agricoles d'élevage est :

$$\begin{aligned}\hat{\bar{Y}}_{eae} &= \frac{\sum_{i \in S_r \cap eae} w_i y_i}{\sum_{i \in S_r \cap eae} w_i} \\ &= \frac{5{,}67 \times 84\,000 + 5{,}67 \times 48\,000 + 16{,}5 \times 23\,000}{5{,}67 + 5{,}67 + 16{,}5} \\ &= \frac{1\,330\,650}{27{,}8} = 47\,796\end{aligned}$$

Remarque : Si les pondérations de l'échantillonnage étaient ignorées, les estimations seraient inexactes. Le tableau ci-dessous montre la comparaison :

Tableau 7 : Comparaison des estimations calculées avec et sans pondération

Paramètre estimé	Estimation exacte avec pondération	Estimation inexacte sans pondération
N	100	10
N_{eae}	28	3
P	0,28	0,30
Y	4 160 340 $	516 000 $
$\overline{Y}$	41 595 $	51 600 $
$\hat{Y}_{eae}$	1 330 650 $	155 000 $
$\overline{Y}_{eae}$	47 796 $	51 667 $

Il est possible d'établir des estimations pour les données qualitatives à l'aide de techniques habituellement réservées aux variables quantitatives. Une *variable indicatrice* peut être définie pour chaque catégorie de la variable qualitative qui prend la valeur 1 si l'unité appartient à la catégorie, et 0 autrement. L'estimation du nombre total d'unités ayant la caractéristique est obtenue en calculant le produit de la valeur de la variable indicatrice (1 ou 0) et du poids pour chaque unité répondante, et ensuite, en faisant la somme pour toutes les unités répondantes. Compte tenu de cette approche, les procédures d'estimation des données qualitatives et quantitatives sont les mêmes.

7.2.3 Questions d'estimation

7.2.3.1 Estimation pour les petits domaines

Le plan d'échantillonnage devrait tenir compte des domaines d'intérêt par l'intermédiaire de la stratification lorsque c'est possible. Cette mesure garantit une précision et une taille de l'échantillon appropriées. Des restrictions appliquées à la taille de l'échantillon et à d'autres critères de plan d'échantillonnage (notamment l'information de la base de sondage) peuvent cependant signifier que seul un nombre minime de strates peuvent être formées et, pour certains domaines, en particulier les petits, la taille de l'échantillon peut donc être insuffisante.

Une taille d'échantillon insuffisante dans un domaine peut poser un problème au moment de l'estimation. Diverses techniques sont disponibles pour obtenir des estimations dans ces cas. Elles comprennent l'estimation synthétique, l'estimation composite et d'autres encore. Ces méthodes exigent habituellement de l'information corrélée d'une autre source ou le recours à de bons modèles. L'application de ces techniques peut devenir complexe et elle est hors de la portée de ce manuel. Le lecteur intéressé obtiendra davantage d'information sur ce sujet avancé en consultant Särndal, Swensson et Wretman (1992), Ghosh et Rao (1994), Singh, Gambino et Mantel (1994).

7.2.3.2 Valeurs aberrantes

Selon la définition de Barnett et Lewis (1995), une valeur aberrante est ***une observation ou un sous-ensemble d'observations qui semble(nt) être incohérente(s), compte tenu des autres séries de données.*** Il y a diverses méthodes disponibles pour diminuer les répercussions des valeurs aberrantes sur les estimations de l'enquête. Les ignorer simplement peut diminuer la précision, leur donner une pondération de un ou de zéro peut biaiser les résultats. D'autre part, l'information auxiliaire et la stratification *a posteriori* peuvent être utilisées pour garantir que les valeurs aberrantes n'ont pas de répercussions

excessives sur les estimations. Les valeurs aberrantes sont abordées au **Chapitre 10 - Traitement**. Ce sujet avancé est aussi étudié dans Kish (1965), et Hidiroglou et Srinath (1981).

7.3 Estimation des erreurs d'échantillonnage des estimations de l'enquête

Des erreurs peuvent se glisser dans les estimations d'une enquête. Au **Chapitre 3 - Introduction au plan d'enquête**, nous mentionnons deux types élémentaires d'erreurs, l'erreur d'échantillonnage et les erreurs non dues à l'échantillonnage. Les erreurs non dues à l'échantillonnage se traduisent souvent par un biais et sont difficiles à mesurer. L'erreur d'échantillonnage donne la variabilité, elle mesure à quel point une estimation de différents échantillons possibles de la même taille et du même plan d'échantillonnage, à l'aide du même estimateur, donne des résultats différents l'un de l'autre.

L'importance d'une estimation de la variance d'échantillonnage à l'étape de la conception du plan, afin de comparer l'efficience statistique de différents plans d'échantillonnage, est expliquée au **Chapitre 6 - Plans d'échantillonnage**. Le **Chapitre 8 - Calcul de la taille de l'échantillon et répartition** révèle comment une estimation de la variance d'échantillonnage est utilisée, afin de déterminer la taille de l'échantillon nécessaire pour obtenir un niveau de précision donné.

L'objectif de cette section est d'illustrer comment la variance d'échantillonnage est mesurée et l'importance de la prise en compte du plan d'échantillonnage. Cette section présente seulement les estimateurs de la variance pour une moyenne ou un total estimé pour un EAS ou un EAS stratifié en supposant qu'il n'y a pas d'ajustement des poids de base. L'estimation de la variance pour une proportion estimée d'un EAS et des plans d'échantillonnage plus complexes (à l'aide d'un effet de plan) sont expliqués au **Chapitre 8 - Calcul de la taille de l'échantillon et répartition.**

Chaque stratégie d'échantillonnage a sa formule particulière d'estimation de la variance d'échantillonnage et elle peut être compliquée. Il faudrait consulter un statisticien d'enquête qui connaît bien ce genre de problème pour estimer correctement la variance d'échantillonnage pour des données complexes (c.-à-d. pour les plans d'échantillonnage plus complexes et lorsqu'il y a ajustements de la pondération).

Les formules présentées dans ce chapitre se retrouvent dans tout ouvrage d'introduction à la théorie des sondages, par exemple, Cochran (1977) ou Lohr (1999).

7.3.1 Variance d'échantillonnage

Du point de vue mathématique, la variance d'échantillonnage d'une estimation est la déviation quadratique moyenne par rapport à la valeur moyenne de l'estimateur pour tous les échantillons possibles. Une liste de facteurs qui ont des répercussions sur l'importance de la variance d'échantillonnage a été donnée au **Chapitre 3 - Introduction au plan d'enquête** :

- la variabilité de la caractéristique d'intérêt dans la population,
- la taille de la population,
- le taux de réponse,
- le plan d'échantillonnage et la méthode d'estimation.

Les répercussions de ces facteurs sur la variance d'échantillonnage sont illustrées ci-dessous à l'aide de l'estimateur pour la variance d'échantillonnage d'une moyenne de la population estimée à partir de

l'EASSR avec un taux de réponse de 100 %. (Voir aussi le **Chapitre 8 - Calcul de la taille de l'échantillon et répartition.)**

La valeur de chaque variable, y_i, varie habituellement d'une unité à l'autre dans la population. La variance de la population, σ^2, de toutes les unités, y_i, dans la population est définie comme suit :

$$\sigma^2 = \frac{(N-1)}{N} S^2$$

où

$$S^2 = \sum_{i \in U} \frac{(y_i - \bar{Y})^2}{N-1}$$

U est l'ensemble des unités de la population et il y a N unités dans la population.

Un estimateur non biaisé de la moyenne de la population pour un EASSR de taille n avec un taux de réponse de 100 % est donné par :

$$\hat{\bar{Y}} = \sum_{i \in S_r} \frac{y_i}{n}$$

où S_r est l'ensemble des répondants de l'échantillon et il y a n unités dans l'échantillon.

L'estimation, $\hat{\bar{Y}}$, varie d'un échantillon à l'autre. La variance d'échantillonnage de $\hat{\bar{Y}}$ pour un EASSR de taille n peut être exprimée ainsi :

$$Var(\hat{\bar{Y}}) = (1 - \frac{n}{N}) \frac{S^2}{n}$$

Il est évident, compte tenu de l'équation ci-dessus, qu'une estimation pour une caractéristique ayant de grandes différences d'une unité à l'autre, c.-à-d. une variabilité élevée dans la population, a une variance d'échantillonnage plus grande que celle d'une estimation pour une caractéristique ayant une faible variabilité dans la population. Habituellement, S^2 est inconnue et doit être estimée (voir la section 7.3.2.3.).

Il est aussi évident que la taille de la population a des répercussions sur la variance d'échantillonnage : l'équation $f = n/N$ est appelée la fraction de sondage et l'équation $(1-f) = 1-n/N$ est le facteur de correction de la population finie (cpf, aussi parfois appelé *facteur d'exhaustivité*). La variance d'échantillonnage diminue dans la mesure où la taille de l'échantillon, n, augmente et, lors d'un recensement (où $n=N$), il n'y a pas de variance d'échantillonnage. Lorsque la fraction de sondage est petite (c.-à-d. que la taille de l'échantillon est petite comparativement à la population), on peut ignorer la cpf. (Selon Cochran (1977), ce facteur peut être ignoré s'il n'est pas supérieur à 5 % et, dans de nombreux cas , même s'il est aussi élevé que 10 %.) Toutefois, les non-réponses augmentent la variance d'échantillonnage en diminuant en fait la taille de l'échantillon.

Il est expliqué au **Chapitre 6 - Plans d'échantillonnage** que certaines stratégies d'échantillonnage sont plus efficientes que d'autres. La stratification, par exemple, et l'utilisation d'estimateurs par le ratio peuvent améliorer la précision des estimations.

Dans l'étude de la variance d'échantillonnage qui suit, il est supposé que l'estimateur n'est pas biaisé. Dans certains cas cependant, il vaut mieux avoir un estimateur biaisé (p. ex., lorsque sa précision est

meilleure que tout autre estimateur non biaisé). S'il y a un biais, peu importe la raison, à cause d'un estimateur biaisé ou d'une erreur non due à l'échantillonnage, les formules de variance de l'échantillonnage présentées dans les sections suivantes permettent de calculer l'erreur quadratique moyenne (EQM) qui est une mesure de la variance d'échantillonnage et du biais. Le résultat peut susciter des problèmes d'intervalles de confiance et ceci sera repris au **Chapitre 11 - Analyse des données de l'enquête.**

7.3.1.1 Calcul de la variance d'échantillonnage réelle

L'exemple suivant illustre comment calculer la variance d'échantillonnage réelle des dépenses moyennes estimées pour les articles vidéo dans un EASSR de taille n=2.

Exemple 7.7 :

Les dépenses pour les articles vidéo d'une population de quatre ménages sont inscrites ci-dessous. Dans un EASSR de taille n=2, quelle est la variance d'échantillonnage réelle des dépenses moyennes estimées?

Tableau 8 : Dépenses pour articles vidéo par ménage

Ménage	Dépenses pour articles vidéo ($)
1	10
2	20
3	30
4	40

Remarquez d'abord que la valeur du paramètre des *dépenses moyennes de la population pour les articles vidéo* est la suivante :

$$\overline{Y} = \sum_{i \in U} \frac{y_i}{N}$$
$$= \frac{10 + 20 + 30 + 40}{4} = 25$$

Voici l'estimateur habituel pour la *moyenne estimée* dans un EAS :

$$\hat{\overline{Y}} = \sum_{i \in S_r} \frac{y_i}{n} = \sum_{i \in S_r} \frac{y_i}{2}$$

Nous pouvons calculer la variance d'échantillonnage réelle de la moyenne estimée, $Var(\hat{\overline{Y}})$, pour un EASSR de taille n=2 en considérant les résultats de tous les échantillons possibles de taille 2 de l'EASSR. Ils sont affichés au tableau ci-dessous :

Tableau 9 : Calcul de la variance d'échantillonnage réelle de $\hat{\bar{Y}}$

Échantillon	Unités de l'échantillon	Estimation de l'échantillon ($) $\hat{\bar{Y}}$	$(\hat{\bar{Y}} - \bar{Y})$	$(\hat{\bar{Y}} - \bar{Y})^2$
1	(1,2)	15	-10	100
2	(1,3)	20	-5	25
3	(2,3)	25	0	0
4	(1,4)	25	0	0
5	(2,4)	30	5	25
6	(3,4)	35	10	100
	Moyenne	**25**	**0**	**41.7**

1. D'abord, calculer la moyenne de toutes les moyennes possibles de l'échantillon :

$$\overline{\hat{\bar{Y}}} = \frac{\hat{\bar{Y}}_{(1)} + \hat{\bar{Y}}_{(2)} + \hat{\bar{Y}}_{(3)} + \hat{\bar{Y}}_{(4)} + \hat{\bar{Y}}_{(5)} + \hat{\bar{Y}}_{(6)}}{6}$$
$$= \frac{15 + 20 + 25 + 25 + 30 + 35}{6} = 25 = \bar{Y}$$

On remarque que la valeur moyenne de l'estimation pour tous les échantillons possibles est égale à la moyenne de la population, $\bar{Y}$. Voilà qui est prévisible parce que l'estimateur pour $\hat{\bar{Y}}$ n'est pas biaisé.

2. Ensuite, calculer la différence entre chaque estimation de l'échantillon et l'estimation moyenne de tous les échantillons (c.-à-d. $\hat{\bar{Y}}_j - \bar{Y}$ pour le $j^{ème}$ échantillon) inscrite dans la quatrième colonne du tableau ci-dessus.

3. Calculer le carré de ces différences (c.-à-d. $(\hat{\bar{Y}}_j - \bar{Y})^2$) inscrit dans la cinquième colonne du tableau.

4. Pour l'ensemble, *J*, de tous les échantillons de la population, calculer la moyenne des différences au carré :

$$Var(\hat{\bar{Y}}) = \sum_{j \in J} \frac{(\hat{\bar{Y}}_j - \bar{Y})^2}{6}$$
$$= \frac{100 + 25 + 0 + 0 + 25 + 100}{6} = 41.7$$

La variance d'échantillonnage réelle des dépenses moyennes estimées pour les articles vidéo d'un EASSR de taille *n*=2 pour cette population est donc $Var(\hat{\bar{Y}}) = 41{,}7$.

Le problème de l'approche ci-dessus est qu'il n'est pas pratique de sélectionner tous les échantillons possibles de la population. Une solution de rechange est de sélectionner de nombreux échantillons par répliques, comme il est mentionné à la Section 6.3.9 du **Chapitre 6 - Plans d'échantillonnage** et à la section 7.3.4. D'autre part, l'équation présentée plus tôt pourrait être utilisée directement :

$$Var(\hat{\bar{Y}}) = (1 - \frac{n}{N})\frac{S^2}{n}$$

où :

$$\begin{aligned} S^2 &= \sum_{i \in U} \frac{(y_i - \bar{Y})^2}{N-1} \\ &= \frac{(10-25)^2 + (20-25)^2 + (30-25)^2 + (40-25)^2}{(4-1)} \\ &= 166{,}7 \end{aligned}$$

donc :

$$Var(\hat{\bar{Y}}) = \left(1 - \frac{n}{N}\right)\frac{S^2}{n} = \left(1 - \frac{2}{4}\right)\frac{166{,}7}{2} = 41{,}7$$

L'équation ci-dessus pose un problème : sauf s'il y a eu recensement auparavant, la variabilité de la population, S^2, est inconnue et doit être estimée à partir d'un seul échantillon. Si l'échantillonnage probabiliste est utilisé, la distribution d'échantillonnage de l'estimateur peut être calculée et la variance de la population peut être estimée à partir d'un seul échantillon.

Les formules pour $\hat{S}^2$ se trouvent dans n'importe quel ouvrage théorique sur l'échantillonnage pour les plans d'échantillonnage standard (EAS, échantillonnage stratifié, etc.). Lorsque le plan d'échantillonnage ou la procédure d'estimation est complexe, d'autres méthodes peuvent servir, notamment celles qui sont décrites à la Section 7.3.4.

Un estimateur sans biais de la variance d'échantillonnage de la moyenne estimée, $\hat{\bar{Y}}$, pour un EASSR, est donné par :

$$\hat{Var}(\hat{\bar{Y}}) = \left(1 - \frac{n}{N}\right)\frac{\hat{S}^2}{n}$$

où :

$$\hat{S}^2 = \sum_{i \in S} \frac{(y_i - \bar{y})^2}{n-1}$$

et où :

$$\bar{y} = \sum_{i \in S} \frac{y_i}{n}$$

Cette formule sera illustrée à la Section 7.3.2.3.

Un estimateur sans biais de la variance d'échantillonnage du total estimé, $\hat{Y}$, pour un EASSR, est donné par :

$$\hat{Var}(\hat{Y}) = \hat{Var}(N \times \hat{\bar{Y}}) = N^2 \hat{Var}(\hat{\bar{Y}}) = N^2\left(1 - \frac{n}{N}\right)\frac{\hat{S}^2}{n}.$$

On verra au **Chapitre 8 - Calcul de la taille de l'échantillon et répartition** l'estimation de la variance d'échantillonnage d'une proportion estimée, $\hat{P}$.

7.3.2 Autres mesures de l'erreur d'échantillonnage

Avant d'illustrer comment *estimer* la variance d'échantillonnage, d'autres mesures communes de l'erreur d'échantillonnage seront présentées, notamment :
- l'erreur-type,
- le coefficient de variation,
- la marge d'erreur,
- l'intervalle de confiance.

Ce sont des expressions connexes et il est possible de passer de l'une à l'autre en appliquant des opérations mathématiques simples.

7.3.2.1 Erreur-type et coefficient de variation

L'erreur-type d'un estimateur est la racine carrée de sa variance d'échantillonnage. Cette mesure est plus facile à interpréter parce qu'elle donne une indication de l'erreur d'échantillonnage à l'aide de la même échelle que l'estimation, tandis que la variance est basée sur les différences quadratiques.

Même l'erreur-type peut cependant être difficile à interpréter lorsqu'on pose la question « Quelle ampleur d'erreur-type est acceptable? » C'est l'importance de l'estimation qui détermine la largeur. Une erreur-type de 100, par exemple, serait considérée grande pour mesurer la moyenne du poids des gens, mais pas pour estimer le revenu annuel moyen.

Il est plus utile dans de nombreuses situations d'évaluer la taille de l'erreur-type par rapport à l'estimation de la caractéristique mesurée. Le ***coefficient de variation*** donne ce genre de mesure. C'est le ***rapport entre l'erreur-type de l'estimation de l'enquête et la valeur moyenne de l'estimation elle-même, pour tous les échantillons possibles***. Le coefficient de variation est habituellement calculé comme l'estimation de l'erreur-type de l'estimation de l'enquête par rapport à l'estimation elle-même. Cette mesure relative de l'erreur d'échantillonnage est habituellement exprimée en pourcentage (10 % au lieu de 0,1). Elle est très utile pour comparer la précision des estimations de l'échantillon lorsque leurs tailles ou échelles sont différentes l'une de l'autre. Elle est cependant moins utile pour les estimateurs des caractéristiques dont la valeur réelle peut être zéro ou négative, y compris les estimations de changement (p. ex., le changement du revenu moyen depuis l'an dernier).

7.3.2.2 Marge d'erreur et intervalle de confiance

Il n'est pas rare de lire les résultats d'une enquête publiés dans un journal comme suit :

> Selon une enquête récente, 15 % des résidents d'Ottawa assistent à des services religieux chaque semaine. Les résultats, tirés d'un échantillon de 1 345 résidents, sont considérés précis à plus ou moins 3 %, 19 fois sur 20.

Que signifie cet énoncé? Il révèle que la valeur réelle, le pourcentage réel des gens qui assistent à des services religieux chaque semaine, se situe probablement à trois points de l'estimation (15 %). Dans l'exemple ci-dessus, la marge d'erreur est de plus ou moins trois points, ou simplement 3 %, et l'intervalle de confiance correspond à la plage de 12 % à 18 %. Les marges d'erreur comprennent toujours un énoncé sur la confiance, c'est-à-dire le degré de confiance que suscite l'intervalle. Dans cet exemple, l'énoncé sur la confiance est *19 fois sur 20*. Si l'enquête était répétée de nombreuses fois, cela

signifie que 19 fois sur 20 (ou 95 % des fois), l'intervalle de confiance couvrirait la valeur réelle de la population.

La théorie sous-jacente à l'établissement des intervalles de confiance peut être décrite comme suit. Supposons une estimation de la moyenne de la population, $\hat{\bar{Y}}$, pour un échantillon de grande taille, et une estimation de l'erreur-type, $S\hat{E}(\hat{\bar{Y}})$. En vertu du théorème central limite et de la distribution normale, les chances sont donc :

- de 0,10 que l'erreur absolue $|\hat{\bar{Y}} - \bar{Y}|$ soit supérieure à $1{,}65 \times S\hat{E}(\hat{\bar{Y}})$ (ce qui correspond à un intervalle de confiance de 90 %),
- de 0,05 que l'erreur absolue $|\hat{\bar{Y}} - \bar{Y}|$ soit supérieure à $1{,}96 \times S\hat{E}(\hat{\bar{Y}})$ (ce qui correspond à un intervalle de confiance de 95 %),
- de 0,01 que l'erreur absolue $|\hat{\bar{Y}} - \bar{Y}|$ soit supérieure à $2{,}58 \times S\hat{E}(\hat{\bar{Y}})$ (ce qui correspond à un intervalle de confiance de 99 %).

Ces formules s'appliquent à tous les estimateurs normalement distribués. Il ne faut pas oublier que les erreurs-types sont utiles, non seulement pour le calcul des intervalles de confiance, mais aussi pour l'analyse inférentielle des données, par exemple, les tests d' hypothèse (voir le **Chapitre 11 - Analyse des données de l'enquête**).

Le lecteur intéressé trouvera voulez davantage de détails sur la théorie sous-jacente aux intervalles de confiance dans les ouvrages sur la théorie de l'échantillonnage (p. ex., Cochran (1977), Lohr (1999), Särndal, Swensson et Wretman (1992), Stuart (1968)). point superflu en anglais

7.3.2.3 Estimation de la variance d'échantillonnage et autres mesures de l'erreur d'échantillonnage de l'EASSR

L'exemple suivant illustre comment estimer les mesures de l'erreur d'échantillonnage à partir d'un seul échantillon réalisé à l'aide de l'EASSR, (en supposant un taux de réponse de 100 % et aucun ajustement pour les données auxiliaires.)

Exemple 7.8 : Estimation de la variance d'échantillonnage, de l'erreur-type, du coefficient de variation, de la marge d'erreur et de l'intervalle de confiance pour $\hat{\bar{Y}}$, EASSR

Un EASSR de n=10 personnes (taux de réponse de 100 %) est sélectionné dans une population de N=500 personnes. L'âge de chaque unité échantillonnée est inscrit dans le tableau ci-dessous (trié par âge). Quelle est la variance d'échantillonnage estimée de l'âge moyen estimé? Quels sont l'erreur-type et le coefficient de variation estimés? Quels sont la marge d'erreur et l'intervalle de confiance pour un niveau de confiance de 95 %?

Tableau 10 : Calcul de la variance d'échantillonnage estimée de $\hat{\bar{Y}}$

Personne	Âge de l'unité de l'échantillon, y_i	$(y_i - \bar{y})$	$(y_i - \bar{y})^2$
1	21	-13,4	179,56
2	26	-8,4	70,56
3	27	-7,4	54,76
4	32	-2,4	5,76
5	34	-0,4	0,16
6	37	2,6	6,76
7	38	3,6	12,96
8	40	5,6	31,36
9	42	7,6	57,76
10	47	12,6	158,76

1. Estimation de l'âge moyen de la population :

$$\hat{\bar{Y}} = \sum_{i \in S_r} \frac{y_i}{n}$$
$$= \frac{21 + 26 + 27 + 32 + 34 + 37 + 38 + 40 + 42 + 47}{10} = 34,4$$

L'âge moyen estimé est donc de 34,4 ans. On notera que la moyenne estimée de la population est la moyenne de l'échantillon simple pour un EAS (sans facteur d'ajustement pour les non-réponses ou les données auxiliaires).

2. Estimation de la variance d'échantillonnage de $\hat{\bar{Y}}$ pour un EASSR :

$$\hat{Var}(\hat{\bar{Y}}) = \left(1 - \frac{n}{N}\right)\frac{\hat{S}^2}{n}$$

où $\hat{S}^2$ est :

$$\hat{S}^2 = \sum_{i \in S_r} \frac{(y_i - \bar{y})^2}{n-1}$$
$$= \frac{179,56 + 70,56 + 54,76 + ... + 31,36 + 57,76 + 158,76}{10 - 1}$$
$$= 64,3$$

donc :

$$\hat{Var}(\hat{\bar{Y}}) = \left(1 - \frac{n}{N}\right)\frac{\hat{S}^2}{n} = \left(1 - \frac{10}{500}\right)\frac{64.3}{10} = 6,3$$

La variance d'échantillonnage estimée est donc 6,3.

3. Estimation de l'erreur-type, $\hat{SE}(\hat{\bar{Y}})$, et du coefficient de variation, $\hat{CV}(\hat{\bar{Y}})$:

$$\hat{SE}(\hat{\bar{Y}}) = \sqrt{\hat{Var}(\hat{\bar{Y}})} = \sqrt{6,3} = 2,5$$

$$\hat{CV}(\hat{\bar{Y}}) = \frac{\hat{SE}(\hat{\bar{Y}})}{\bar{y}} = \frac{2,5}{34,4} = 0,073 = 7,3\ \%$$

4. Calcul de la marge d'erreur et de l'intervalle de confiance avec niveau de confiance de 95 % :

Marge d'erreur $= 1,96 \times \hat{SE}(\hat{\bar{Y}})$
$= 1,96 \text{ x } 2,5$
$= 4,9$

Intervalle de confiance $= \hat{\bar{Y}} \pm 1,96 \times \hat{SE}(\hat{\bar{Y}})$
$= 34,4 \pm 4,9$
$= (29,5,\ 39,3)$

On peut donc affirmer avec un taux de confiance de 95 % que l'âge moyen réel de la population se situe entre 29,5 et 39,3 ans. (À proprement parler, l'interprétation exacte est que l'intervalle de confiance dans un échantillonnage répété comprendrait la valeur réelle de la population en moyenne 95 % des fois.)

7.3.2.4 Estimation de la variance d'échantillonnage de l'EASSR stratifié

Illustrons maintenant les répercussions si l'on néglige de prendre en compte le plan d'échantillonnage réel et que la variance d'échantillonnage est simplement calculée à l'aide de l'équation pour un échantillon aléatoire simple. Le marché des logiciels offre un large éventail de logiciels de statistique et de traitement des données sur ordinateur personnel, mais très peu tiennent compte correctement du plan d'enquête, pas même ceux dont la publicité soutient qu'ils sont spécialisés en traitement des enquêtes. Un certain nombre d'examens des logiciels statistiques a été fait depuis dix ans, et il serait prudent et judicieux d'en lire quelques-uns; un répertoire est tenu à jour au http://www.fas.harvard.edu/~stats/survey-soft/survey-soft.html.

Exemple 7.6 (suite) : Estimation de la variance d'échantillonnage

Si l'échantillon est le résultat d'un échantillon aléatoire simple de taille n=10 (taux de réponse de 100 %) tiré d'une population de taille N=100, il est alors possible d'obtenir les estimations suivantes du revenu moyen et de la variance d'échantillonnage du revenu moyen estimé.

$$\hat{\bar{Y}} = \sum_{i \in S_r} \frac{y_i}{n} = 51\,600$$

(comparativement à 41 595 le plan d'échantillonnage est pris en compte).

La variance d'échantillonnage estimée (en milliers) :

$$\hat{Var}_{EAS}(\hat{\bar{Y}}) = \left(1 - \frac{n}{N}\right)\frac{\hat{S}^2}{n} = \left(1 - \frac{10}{100}\right)\frac{1\,247}{10} = 112,2$$

et l'erreur-type est (en milliers) $\sqrt{\hat{Var}_{EAS}(\hat{\bar{Y}})} = \hat{SE}_{EAS}(\hat{\bar{Y}})$=10,6.

Afin d'estimer correctement la variance d'échantillonnage de la moyenne à partir d'un échantillon stratifié, il faut déterminer la variance d'échantillonnage estimée de chaque strate *h* et faire la somme des résultats de chaque strate pour obtenir une estimation complète (en milliers de $) :

$$\hat{V}_{STR}(\hat{\bar{Y}}) = \frac{1}{N^2}\sum_h N_h^2\left(1-\frac{n_h}{N_h}\right)\frac{\hat{S}_h^2}{n_h}$$
$$= \frac{1}{100^2}\left\{34^2\left(1-\frac{6}{34}\right)\frac{1406}{6} + 66^2\left(1-\frac{4}{66}\right)\frac{211{,}6}{4}\right\} = 44$$

et l'erreur-type est (en milliers) $\sqrt{\hat{Var}_{STR}(\hat{\bar{Y}})} = \hat{SE}_{STR}(\hat{\bar{Y}})$ = 6,6.

Si variance d'échantillonnage avait été estimée sans tenir compte du plan d'échantillonnage, et que l'estimateur pour un EAS avait été utilisé, la variance d'échantillonnage aurait été surestimée. En revanche, si le plan d'échantillonnage avait été un plan par grappes et la formule de l'EAS avait été utilisée, la variance d'échantillonnage réelle aurait probablement été sous-estimée.

7.3.3 Effet de plan

L'***effet de plan*** compare la variance des estimateurs entre un plan d'échantillonnage et un EAS. Il s'agit du ***rapport entre la variance d'échantillonnage d'un estimateur, selon un plan d'échantillonnage donné, et la variance d'échantillonnage de l'estimateur d'un EAS de même taille***.

Il est mentionné au **Chapitre 6 - Plans d'échantillonnage** que cette mesure est souvent appliquée pour comparer l'efficience des estimateurs de divers plans d'échantillonnage. Si le ratio est inférieur à un, le résultat indique que le plan d'échantillonnage est plus efficient que l'EAS, s'il est supérieur à un, le plan d'échantillonnage est moins efficient que l'EAS.

Dans le cas de l'exemple présenté plus tôt,

$$deff = \frac{\hat{Var}_{STR}(\hat{\bar{Y}})}{\hat{Var}_{EAS}(\hat{\bar{Y}})} = \frac{44\ 000}{112\ 200} = 0{,}39$$

c'est-à-dire que la stratification améliore énormément la précision de la moyenne estimée de l'enquête.

Les effets du plan d'échantillonnage aident aussi à obtenir des estimations approximatives de la variance pour des plans d'échantillonnage complexes. Si une estimation de l'effet du plan d'échantillonnage est disponible dans une enquête précédente qui a utilisé le même plan d'échantillonnage, elle peut servir à déterminer la taille de l'échantillon nécessaire de l'enquête. (Ce point sera considéré au **Chapitre 8 - Calcul de la taille de l'échantillon et répartition**).

On consultera Kish (1965) pour davantage d'information sur les effets de plan.

7.3.4 Estimation de la variance d'échantillonnage à l'aide de l'échantillonnage par répliques

Les variances des statistiques simples, notamment les moyennes et les totaux, peuvent être estimées à l'aide de procédures mentionnées auparavant pour les plans d'échantillonnage simples. Si les plans d'échantillonnage ou les estimateurs sont plus complexes cependant (p. ex., des fonctions des totaux), il peut être difficile de déterminer la formule de la variance d'échantillonnage.

Des méthodes de rechange (autres que l'application d'un effet de plan) permettent d'estimer la variance d'échantillonnage pour une série sélectionnée de procédures d'estimation et de plans d'échantillonnage. L'*échantillonnage par répliques* en est une. Dans un échantillonnage par répliques, au lieu de sélectionner un échantillon de taille *n, k* échantillons indépendants de taille *n / k* sont sélectionnés. Une estimation de la caractéristique d'intérêt est faite pour chacun de ces échantillons *k*. La variabilité des estimations des échantillons *k* sert ensuite à estimer la variance d'échantillonnage. L'estimation, *t*, de la caractéristique d'intérêt est obtenue à l'aide de la moyenne des estimations faites pour chaque échantillon :

$$t = \sum_{j \in K} \frac{t_j}{k}$$

où *K* est l'ensemble des échantillons, *k* est le nombre d'échantillons et t_j est l'estimation du j^e échantillon.

La variance d'échantillonnage estimée de *t*, $\hat{Var}(t)$, est le résultat de l'équation :

$$\hat{Var}(t) = \sum_{j \in K} \frac{(t_j - t)^2}{k(k-1)}$$

Exemple 7.9 : Estimation de la variance d'échantillonnage de $\hat{\bar{Y}}$ à l'aide de l'échantillonnage par répliques, EAS

Dans l'exemple 7.8, au lieu de sélectionner un échantillon de taille *n*=10 et d'appliquer l'équation précédente pour estimer $Var(t) = Var(\hat{\bar{Y}})$, deux échantillons de taille *n*=5 sont sélectionnés. Les résultats sont inscrits au tableau suivant.

Tableau 11 : Calcul de la variance d'échantillonnage estimée de $\hat{\bar{Y}}$ à l'aide de l'échantillonnage par répliques

Échantillon 1	Âge des unités de l'échantillon	Échantillon 2	Âge des unités de l'échantillon
1	21	1	26
2	27	2	32
3	34	3	37
4	38	4	40
5	42	5	47
Âge moyen	32,4		36,4

1. L'âge moyen de la population est estimé à :

$$\hat{\bar{Y}} = \sum_{j \in K} \frac{\hat{\bar{Y}}_j}{k}$$
$$= \frac{32{,}4 + 36{,}4}{2} = 34{,}4$$

2. Voici la variance d'échantillonnage estimée de l'âge moyen à l'aide de la méthode d'échantillonnage réitéré :

$$\hat{Var}(\hat{\bar{Y}}) = \sum_{j \in K} \frac{(\hat{\bar{Y}}_j - \hat{\bar{Y}})^2}{k(k-1)}$$
$$= \frac{(32{,}4 - 34{,}4)^2 + (36{,}4 - 34{,}4)^2}{2} = 4$$

L'erreur-type estimée, $\hat{SE}(\hat{\bar{Y}})$, est 2.

D'autres méthodes de ré-échantillonnage, notamment la méthode du *Jackknife* et celle du *Bootstrap* (auto-amorçage), sont aussi souvent utilisées dans les enquêtes ayant des plans complexes. Rust et Rao (1996), Wolter (1985) ou Efron (1981) donnent une description de ces méthodes. Gambino *et coll*. (1998) donnent un exemple de la méthode du Jackknife appliquée dans une enquête sur les ménages comprenant un estimateur et un plan d'échantillonnage complexes. D'autres techniques qui ne sont pas basées sur le ré-échantillonnage, notamment l'approximation par séries de Taylor, peuvent aussi servir lorsque le plan d'échantillonnage est complexe. Hidiroglou et Paton (1987), Binder (1996), Särndal, Swensson et Wretman (1992) et Wolter (1985) constituent d'excellentes sources.

7.4 Sommaire

La liste suivante donne un aperçu des points importants à considérer pour estimer les données d'une enquête :

1. L'estimation doit tenir compte du plan d'échantillonnage. Il faudrait intégrer à cette fin les poids de base au processus d'estimation.

2. Les poids de base devraient être ajustés pour les non-réponses.

3. Il faudrait utiliser, si possible, l'information auxiliaire, si elle est de qualité appropriée et corrélée avec les principales variables de l'enquête, pour améliorer l'uniformité et la précision des estimations.

4. Il faudrait utiliser le plan d'échantillonnage et la répartition des échantillons pour répondre aux exigences des domaines d'intérêt. Si ce n'est pas possible à l'étape de la conception du plan d'échantillonnage, il faudrait considérer des méthodes d'estimation spéciales à l'étape de l'estimation.

5. Les valeurs aberrantes peuvent donner une grande variabilité d'échantillonnage dans les estimations. Il faudrait considérer le repérage et le traitement des valeurs aberrantes à l'étape de l'estimation.

6. Les estimations de l'enquête devraient comprendre une estimation de leur erreur d'échantillonnage, sous forme de variance d'échantillonnage, d'erreur-type, de coefficient de variation, de marge d'erreur ou d'intervalle de confiance.

On propose au **Chapitre 11 - Analyse des données de l'enquête** des utilisation de données pour fins d'analyse qui vont au-delà des simples statistiques descriptives..

Bibliographie

Arora, H.R. et G.J. Brackstone. 1977. An Investigation of the Properties of Raking Ratio Estimators: I, With Simple Random Sampling. *Survey Methodology,* 3(1): 62-83.

Bankier, M.D. 1978. An Estimate of the Efficiency of Raking Ratio Estimators under Simple Random Sampling. *Survey Methodology*. 4(1): 115-124.

Barnett, V. et T. Lewis. 1995, *Outliers in Statistical Data*. John Wiley and Sons, Chichester.

Binder, D.A. 1983. On the Variance of Asymptotically Normal Estimators from Complex Surveys. *International Statistical Review,* 51: 279-292.

Binder, D.A. 1996. Méthodes de linéarisation pour les échantillons à une et deux phases: une approche de type "recette". *Techniques d'enquête*. 22(1): 17-22.

Binder, D.A. 1998. Estimating the Variance of Raking Ratio Estimators. *Canadian Journal of Statistics,* 16: 47-55.

Brackstone, G. et J.N.K. Rao. 1979. An Investigation of Raking Ratio Estimators. *Sankhyà*, Series C, 42: 97-114.

Chambers, R.L. 1986. Outlier Robust Finite Population Estimation. *Journal of the American Statistical Association,* 81: 1063-1069.

Cochran, W.G. 1977. *Sampling Techniques*. John Wiley and Sons, New York.

Cox, B.G., D.A. Binder, B.N. Chinnappa, A. Christianson, M.J. Colledge et P.S. Kott, Éds. 1995. *Business Survey Methods.* John Wiley and Sons, New York.

Deming, W.E. et F.F. Stephan. 1940. On the least squares adjustment of a sampled frequency table when the expected marginal totals are known. *Annals of Mathematical Statistics,* 11: 427-444.

Deville, J.C. et C.E. Särndal. 1992. Calibration estimators in survey sampling. *Journal of the American Statistical Association,* 87: 376-382.

Efron, B. 1981. *The Jackknife, the Bootstrap and Other Resampling Plans*. SIAM. 38. Philadelphia.

Eltinge, J.L. et I.S. Yansaneh. 1997. Méthodes diagnostiques pour la construction de cellules de correction pour la non-réponse, avec application à la non-réponse aux questions sur le revenu dans la "U.S. Consumer Expenditure Survey". *Techniques d'enquête*, 23(1): 37-45.

Estevao, V., M.A. Hidiroglou, and C.E. Särndal. 1995. Methodological Principles for a Generalized Estimation System at Statistics Canada. *Journal of Official Statistics,* 11: 181-204.

Fink, A. 1995. *The Survey Kit*. Sage Publications, California.

Fowler, F.J. 1984. *Survey Research Methods*. 1. Sage Publications, California.

Gambino, J.G., M.P. Singh, J. Dufour, B. Kennedy et J. Lindeyer. 1998. *Méthodologie de l'enquête sur la population active du Canada*. Statistique Canada. 71-526.

Ghosh, M. et J.N.K. Rao. 1994. Small Area Estimation: An Appraisal. *Statistical Science,* 9: 55-93.

Groves, R. et M.P. Couper. 1998. *Nonresponse in Household Interview Surveys*. John Wiley and Sons, New York.

Hidiroglou, M.A. et D.G. Paton. 1987. Some Experiences in Computing Estimates and their Variances Using Data from Complex Survey Designs. Dans *Applied Probability, Stochastic Processes, and Sampling Theory*. I.B. MacNeill et G.J. Umphrey, Éds. D. Riedel Publishing.

Hidiroglou, M.A. et C.-E. Särndal. 1998. Emploi de données auxiliaires dans l'échantillonnage à deux phases. *Techniques d'enquête,* 24(1): 11-20.

Hidiroglou, M.A. et K.P. Srinath. 1981. Some Estimators of Population Total Containing Large Units. *Journal of the American Statistical Association,* 47: 663-685.

Holt, D. et T.M.F. Smith. 1979. Post-Stratification. *Journal of the Royal Statistical Society,* A, 142: 33-46.

Kalton, G. et D. Kasprzyk. 1986. Le traitement des données d'enquête manquantes. *Techniques d'enquête,* 12(1): 1-17.

Kish, L. 1965. *Survey Sampling*. John Wiley and Sons, New York.

Kovar, J.G., J.N.K. Rao et C.F.J. Wu. 1988. Bootstrap and Other Methods to Measure Error in Survey Estimates. *Canadian Journal of Statistics,* 16, Supplement: 25-45.

Lehtonen, R. et E.J. Pahkinen. 1995. *Practical Methods for Design and Analysis of Complex Surveys*. John Wiley and Sons, New York.

Levy, P. et S. Lemeshow. 1999. *Sampling of Populations*, John Wiley and Sons, New York.

Lohr, S. 1999. *Sampling: Design and Analysis*. Duxbury Press, U.S.A.

Madow, W.G., H. Nisselson, I. Olkin et D.B. Rubin, Éds. 1983. *Incomplete Data in Sample Surveys, Volume 1*. Academic Press, New York.

Madow, W.G., I. Olkin et D.B. Rubin, Éds. 1983. *Incomplete Data in Sample Surveys, Volume 2*. Academic Press, New York.

Madow, W.G. et I. Olkin, Éds. 1983. *Incomplete Data in Sample Surveys, Volume 3*. Academic Press, New York.

Moser, C.A. et G. Kalton. 1971. *Survey Methods in Social Investigation*. Heinemann Educational Books Limited, London.

Platek, R., J.N.K. Rao, C.E. Särndal et M.P. Singh, Éds. 1987. *Small Area Statistics*. John Wiley and Sons, New York.

Pollock, K.H., S.C. Turner et C.A. Brown. 1994. Techniques de saisie - resaisie pour l'estimation de la taille de la population et de totaux de population lorsqu'on ne dispose pas d'une base de sondage complète. *Techniques d'enquête,* 20(2): 121-128.

Rancourt, E., H. Lee et C.E. Särndal. 1993. Variance Estimation Under More than One Imputation Method. *Proceedings of the International Conference on Establishment Surveys*. American Statistical Association. 374-379.

Rao, J.N.K. et C.F.J. Wu. 1988. Resampling Inference with Complex Survey Data. *Journal of the American Staistical Association,* 83: 231-241.

Rao, J.N.K. 1996. On the Estimation with Imputed Survey Data. *Journal of the American Statistical Association,* 91: 499-506.

Rao, J.N.K, C.F.J. Wu et K. Yue. 1992. Quelques travaux récents sur les méthodes de rééchantillonnage applicables aux enquêtes complexes. *Techniques d'enquête,* 18(2): 209-217.

Rust, K.F. et J.N.K. Rao. 1996. Variance Estimation for Complex Surveys using Replication Techniques. *Statistical Methods in Medical Research,* 5: 283-310.

Särndal, C.E., B. Swensson et J. Wretman. 1992. *Model Assisted Survey Sampling*. Springer-Verlag, New York.

Satin, A. et W. Shastry. 1993. *L'échantillonnage : un guide non mathématique – Deuxième édition.* Statistique Canada. 12-602F.

Schnell, D., W.J. Kennedy, G. Sullivan, H.J. Park et W.A. Fuller. 1988. Logiciel d'ordinateur personnel pour l'estimation de variance dans les enquêtes complexes. *Techniques d'enquête,* 14(1): 63-73.

Singh, A.C. 1996. Combining Information in Survey Sampling by Modified Regression. Proceedings of the Section on Survey Research Methods. *American Statistical Association*. 120-129.

Singh, M.P., J. Gambino et H.J. Mantel. 1994. Les petites régions: problèmes et solutions. *Techniques d'enquête,* 20(1): 3-23.

Skinner, C.K., D. Holt et T.M.F. Smith. 1989. *Analysis of Complex Surveys*. John Wiley and Sons, Chichester.

Stuart, A. 1968. *Basic Ideas of Scientific Sampling*. Charles Griffin and Company Limited, London.

Thompson, M. 1997. *Theory of Sample Surveys*. Chapman and Hill, United Kingdom.

Thompson, S.K. 1992. *Sampling*. John Wiley and Sons, New York

Wolter, K.M. 1985. *Introduction to Variance Estimation*. Springer-Verlag, New York.

Yung, W. et J.N.K. Rao. 1996. Linéarisation des estimateurs de variance Jackknife dans un échantillonnage stratifié à degrés multiples. *Techniques d'enquête.* 22(1): 23-31.

Chapitre 8 - Calcul de la taille de l'échantillon et répartition

8.0 Introduction

Voici l'une des questions les plus souvent posées à un statisticien : Quelle taille devrait avoir l'échantillon? Les gestionnaires sont anxieux d'obtenir une réponse à cette question fondamentale pendant la phase de la planification de l'enquête parce qu'elle a des répercussions directes sur les considérations opérationnelles, notamment, le nombre d'intervieweurs nécessaires.

Il n'y a pas de solution magique ou de recette parfaite pour déterminer la taille de l'échantillon. Il s'agit plutôt d'un processus de compromis au cours duquel les besoins de précision des estimations sont pondérés en tenant compte de diverses contraintes opérationnelles, par exemple, le budget, le temps et les ressources disponibles.

Il faut se rappeler que les facteurs qui ont des répercussions sur la précision (présentés au **Chapitre 7 - Estimation**) comprennent la variabilité et la taille de la population, le plan d'échantillonnage, l'estimateur et le taux de réponse. Il faut tenir compte de ces facteurs pour établir la formule de calcul de la taille de l'échantillon nécessaire pour obtenir un degré de précision en particulier.

Des contraintes opérationnelles s'ajoutent à ces facteurs et ont des répercussions sur la taille de l'échantillon. Ces facteurs ont parfois davantage d'influence. Quelle taille d'échantillon le client peut-il s'offrir? Combien de temps est-il prévu pour procéder à l'enquête au complet? Combien d'intervieweurs sont disponibles? Ces contraintes sont souvent exclues de la formule de calcul de la taille de l'échantillon, mais il faut en tenir compte.

Si un plan d'échantillonnage stratifié est utilisé, l'organisme statistique devra savoir, non seulement quelle taille doit avoir l'échantillon, mais aussi comment il devrait être réparti entre les strates. Ce point est intitulé répartition de l'échantillon. Deux stratégies sont possibles. La première est de déterminer la taille totale de l'échantillon et de la répartir ensuite entre les strates. La deuxième est de déterminer la précision voulue et ensuite, la taille de l'échantillon nécessaire dans chaque strate.

De nombreuses méthodes de répartition différentes sont disponibles. La répartition proportionnelle à N donne la même proportion d'unités de la population échantillonnée dans chaque strate. Dans la répartition non proportionnelle, les strates sont échantillonnées à différents taux. Les méthodes de répartition non proportionnelles comprennent la répartition proportionnelle à Y, la répartition proportionnelle à $\sqrt{N}$, la répartition proportionnelle à $\sqrt{Y}$, la répartition optimale, la répartition de Neyman et la répartition optimale lorsque les variances sont égales.

L'objectif de ce chapitre est d'illustrer comment calculer la taille de l'échantillon, compte tenu d'un degré cible de précision, comment répartir un échantillon stratifié, et de donner des conseils pour déterminer quelle méthode de répartition convient le mieux.

8.1 Choix de la taille de l'échantillon

Nous avons expliqué au **Chapitre 7 - Estimation** que la précision des estimations de l'enquête et la taille de l'échantillon sont liées. Étant donné que la variance d'échantillonnage diminue lorsque la taille de

l'échantillon augmente, plus les estimations doivent être précises, plus la taille d'échantillon nécessaire doit donc être grande. La précision ciblée des estimations de l'enquête détermine donc la taille appropriée de l'échantillon. Cette précision d'une estimation, t, peut être intitulée erreur-type admissible, *SE(t)*, marge d'erreur, $z \times SE(t)$, ou coefficient de variation $SE(t)/t$. Le choix de la taille de l'échantillon pour une enquête comprend souvent la spécification de la précision voulue à l'aide de l'une ou l'autre de ces mesures.

Le choix de la taille de l'échantillon vise à limiter les erreurs d'échantillonnage et les non-réponses aléatoires. Il ne vise pas à limiter d'autres erreurs non dues à l'échantillonnage. Pour obtenir des résultats d'enquête exacts, il faudrait minimiser le biais dû aux erreurs non dues à l'échantillonnage lorsque c'est possible (voir le **Chapitre 5 - Conception du questionnaire** et le **Chapitre 3 - Introduction au plan d'enquête** pour obtenir davantage de détails sur les erreurs non dues à l'échantillonnage).

Avant de présenter les formules de calcul de la taille de l'échantillon pour un degré donné de précision, nous considérerons dans ce chapitre comment déterminer le degré de précision approprié et les facteurs ayant des répercussions sur la précision.

8.1.1 Considérations sur le besoin de précision

L'organisme statistique devrait considérer plusieurs questions pertinentes avant de déterminer le degré approprié de précision pour les estimations de l'enquête d'un client. L'organisme et le client devraient examiner ce qui est demandé des estimations de l'enquête du point de vue des totalisations des données, des analyses et des décisions qui seront prises, compte tenu des estimations de l'enquête.

i. À quoi serviront les estimations de l'enquête? Quelle variance d'échantillonnage est acceptable dans les estimations de l'enquête?

Quel degré d'incertitude le client peut-il tolérer dans les estimations de l'enquête? Une marge d'erreur de ±6 % et un niveau de confiance de 95 % sont-ils convenables pour les objectifs du client, ou des estimations plus (ou moins) précises sont-elles nécessaires pour atteindre les objectifs de l'enquête?

Si les résultats de l'enquête servent à prendre des décisions importantes qui auront de grandes répercussions ou qui comprennent un risque marqué, le client peut exiger des estimations d'enquête plus précises que s'il veut simplement obtenir une estimation approximative d'une caractéristique d'intérêt.

ii. Des estimations sont-elles nécessaires pour des sous-groupes (domaines) de la population de l'enquête?

Les résultats de l'enquête peuvent comprendre des estimations pour de nombreux sous-groupes ou domaines. Après avoir obtenu des estimations d'enquête à l'échelon national, par exemple, des estimations provinciales ou sous-provinciales peuvent être nécessaires, ou le client peut avoir besoin d'estimations pour d'autres sous-groupes importants dans la population de l'enquête, selon le sexe, l'âge, la scolarité, etc. Il faudrait déterminer le degré de précision approprié pour répondre à ces besoins de données. Un degré différent de précision peut être déterminé pour divers domaines. Dans une enquête nationale par exemple, le commanditaire de l'enquête peut demander une marge d'erreur de ±3 % pour toutes les estimations nationales, mais une marge d'erreur de ±5 % pour les estimations provinciales peut le satisfaire, ainsi qu'une marge d'erreur de ±10 % pour les estimations sous-provinciales. Des strates sont habituellement formées pour chaque domaine d'intérêt dans ce cas.

iii. Quelle est l'ampleur de la variance d'échantillonnage comparativement à l'estimation de l'enquête?

Il faudrait déterminer les besoins de précision après avoir considéré la taille de l'estimation. Disons par exemple qu'à la suite d'une nouvelle politique, les bureaux du gouvernement fédéral doivent offrir un service dans la langue officielle d'une minorité si au moins P=0,05 (ou 5 %) des demandes sont présentées dans cette langue. Supposons que divers bureaux du gouvernement décident de procéder à une enquête auprès de leur clientèle pour estimer la demande de services dans la langue officielle d'une minorité. À première vue, une marge d'erreur de ±0,05 semble élevée si une estimation de l'enquête doit se situer à 0,05 environ. Il faudrait déterminer dans ce cas une plus petite marge d'erreur, peut-être ±0,01 ou ±0,02 au plus (c.-à-d. que l'intervalle de confiance devrait être de 0,05 ± 0,01 ou 0,05 ± 0,02).

Le client devrait en fait considérer la taille de la plus petite estimation nécessaire pour déterminer les besoins de précision. Supposons que l'enquête sert à estimer des proportions. Certaines de ces proportions pourraient être P = 0,50 ou plus, mais d'autres pourraient être aussi minimes que P = 0,50 ou P = 0,10. Si la plus petite proportion à estimer doit être P = 0,05 et si cette proportion est importante pour les objectifs de l'enquête, l'organisme statistique (et le client) voudront obtenir une marge d'erreur de moins de 0,05.

iv. Quelles sont les répercussions pratiques du besoin de précision? Quel degré de précision est obtenu si l'on augmente la taille de l'échantillon?

Plus la taille de l'échantillon augmente, plus le degré de précision est élevé. Le gain en précision n'est cependant pas directement proportionnel à l'augmentation de la taille de l'échantillon. Considérons une variable qualitative qui a deux modalités possibles, A et B (p. ex., *hommes* et *femmes*) et le client est intéressé à estimer la proportion de la population dans la catégorie A.

Le tableau 1 illustre la marge d'erreur obtenue dans la proportion estimée d'un échantillon aléatoire simple (EAS) pour diverses tailles d'échantillon et un taux de confiance de 95 %. La proportion réelle de la population de la catégorie A est P=0,5 (ou 50 %) et la taille de la population est N=100 000. (Consultez la Section 8.1.3 pour obtenir la formule de calcul de la variance d'échantillonnage d'une proportion estimée.)

Tableau 1 : Taille d'échantillonnage et marge d'erreur d'une estimation de P, à l'aide d'un EAS, lorsque P=0,5

Taille de l'échantillon	Marge d'erreur
50	±0,139
100	±0,098
500	±0,044
1 000	±0,031

Le tableau ci-dessus montre que la taille de l'échantillon double pour passer de 50 à 100 et la marge d'erreur de l'estimation de la proportion s'améliore pour passer de ±0,14 à ±0,10. La marge d'erreur n'a cependant pas diminué de moitié pour s'établir à ±0,07, comme on pourrait s'y attendre. Doubler la taille de l'échantillon pour qu'elle passe de 500 à 1 000 ne diminue pas non plus la marge d'erreur de moitié. Malgré l'impression de la plupart des gens, il n'y a pas de lien linéaire entre la taille de l'échantillon et la marge d'erreur.

Cet exemple fait valoir que l'organisme statistique et le client doivent décider s'il vaut la peine de faire les efforts et d'investir les ressources nécessaires pour interviewer 1 000 personnes au lieu de 500, afin d'améliorer la précision d'une marge d'erreur de ±0,045 à ±0,032.

La meilleure solution n'est peut-être pas toujours de choisir la plus grande taille d'échantillon possible donnant la plus petite marge d'erreur. Il est parfois possible d'obtenir des résultats suffisamment précis en acceptant une marge d'erreur plus large et en utilisant des ressources avec plus d'efficience. Choisir un échantillon de plus petite taille pour réserver de l'argent à d'autres facteurs qui ont des répercussions sur l'exactitude des résultats de l'enquête, par exemple, pour réduire l'erreur non due à l'échantillonnage, peut être plus efficace (p. ex., faire le suivi auprès des non-répondants, faire l'essai du questionnaire, former les intervieweurs, etc.).

8.1.2 Facteurs ayant des effets sur la précision

Nous avons présenté au **Chapitre 3 - Introduction au plan d'enquête** et au **Chapitre 7 - Estimation** les divers facteurs ayant des effets sur la précision. Cette section illustre les répercussions de ces facteurs et présente des considérations lorsqu'il faut déterminer la taille de l'échantillon pour un degré de précision en particulier.

8.1.2.1 Variabilité de la population

La caractéristique, ou variable d'intérêt, est typiquement différente d'une personne, d'un ménage, d'une entreprise, d'une exploitation agricole, etc., à l'autre dans la population de l'enquête. Cette variabilité ne peut être contrôlée, mais son ampleur a des répercussions sur la taille de l'échantillon nécessaire pour obtenir un degré de précision en particulier pour une caractéristique d'intérêt.

Considérez le Tableau 2 ci-dessous. Supposons qu'une nouvelle enquête vise à estimer la proportion de clients satisfaits des services d'une certaine entreprise et qu'il y a seulement deux valeurs possibles pour la variable *satisfaction de la clientèle : satisfait* ou *insatisfait*. Certaines valeurs possibles servant à déterminer la proportion réelle de clients satisfaits et insatisfaits sont énumérées ci-dessous :

Tableau 2 : Répartition possible de la satisfaction de la clientèle pour la population réelle

1.	100 % Satisfaits	0 % Insatisfait
2.	90 % Satisfaits	10 % Insatisfaits
3.	80 % Satisfaits	20 % Insatisfaits
4.	70 % Satisfaits	30 % Insatisfaits
5.	60 % Satisfaits	40 % Insatisfaits
6.	**50 % Satisfaits**	**50 % Insatisfaits**
7.	40 % Satisfaits	60 % Insatisfaits
8.	30 % Satisfaits	70 % Insatisfaits
9.	20 % Satisfaits	80 % Insatisfaits
10.	10 % Satisfaits	90 % Insatisfaits
11.	0 % Satisfait	100 % Insatisfaits

Du point de vue de la variabilité de la satisfaction de la clientèle dans la population, les nombres 1 et 11 dans la liste de possibilités ci-dessus sont les mêmes, c'est-à-dire qu'il n'y a pas de variabilité, tous les clients ont la même opinion. Les nombres 2 et 10 de la liste reflètent une très petite variabilité, 90 % des clients ont la même opinion et seulement 10 % ont une opinion contraire. Chaque série de nombres

suivants, 3 et 9, 4 et 8, 5 et 7, a la même variabilité. À partir des nombres 1 à 6 ou, de même, des nombres 11 à 6, la variabilité de la caractéristique *satisfaction de la clientèle* augmente. Dans la situation que représente le nombre 6, c'est-à-dire une *répartition moitié-moitié*, où 50 % des clients sont satisfaits et 50 % des clients sont insatisfaits, nous avons ici le point de *variabilité maximale* dans la population quant à la satisfaction de la clientèle. Si tous les clients étaient satisfaits des services obtenus, il n'y aurait donc pas de variabilité de la satisfaction de la clientèle et un échantillon *d'un seul* client donnerait une estimation fiable de la satisfaction de la clientèle. Dans la mesure où la variabilité réelle d'une caractéristique d'intérêt augmente dans la population de l'enquête, cependant, la taille de l'échantillon doit aussi augmenter pour donner une estimation de cette caractéristique avec une bonne précision.

Il est difficile de mesurer précisément les caractéristiques qui ont des taux élevés de variabilité. Il faut des tailles d'échantillon de plus en plus larges pour obtenir des estimations précises de ces variables. Si vous considérez la précision des estimations, la taille de l'échantillon nécessaire est la plus large lorsque la variabilité de la caractéristique d'intérêt est à son point maximal. Si la caractéristique a deux valeurs seulement, la situation se produit lorsqu'il y a une répartition moitié-moitié dans la population. Si vous voulez déterminer la taille de l'échantillon pour une enquête, il faut donc obtenir auparavant une estimation de la variabilité d'une caractéristique dans la population de l'enquête parce que la variabilité réelle n'est généralement pas connue d'avance. Vous pouvez l'obtenir à l'aide d'une étude précédente sur le même sujet ou d'une enquête pilote.

Après l'enquête, si l'organisme statistique réalise que la caractéristique d'intérêt varie plus que prévu au moment de déterminer la taille de l'échantillon, les estimations de l'enquête seront moins précises que prévu. D'autre part, si la variabilité de la caractéristique d'intérêt est moins marquée que la variabilité prévue, la taille de l'échantillon nécessaire sera surestimée et les estimations de l'enquête seront plus précises que celles demandées. Pour obtenir la précision demandée pour une enquête, il est habituellement recommandé de faire une estimation raisonnable de la variabilité de la caractéristique de la population lors du calcul de la taille de l'échantillon demandé. Autrement dit, en pratique, si la variabilité de la caractéristique à mesurer dans l'enquête n'est pas connue d'avance, supposer la plus grande variabilité est souvent une bonne idée. Il faudrait donc supposer une répartition moitié-moitié de la population lorsqu'une variable a seulement deux modalités possibles.

Les enquêtes par échantillon mesurent habituellement plus d'une caractéristique, chacune ayant une variabilité différente. Un échantillon suffisamment large pour une caractéristique peut être trop restreint pour une autre qui a une plus grande variabilité. Pour obtenir une taille d'échantillon suffisamment grande pour les principales caractéristiques, la taille de l'échantillon devrait être déterminée selon la caractéristique ayant la plus grande variabilité à votre avis, ou celle jugée la plus importante.

8.1.2.2 Taille de la population

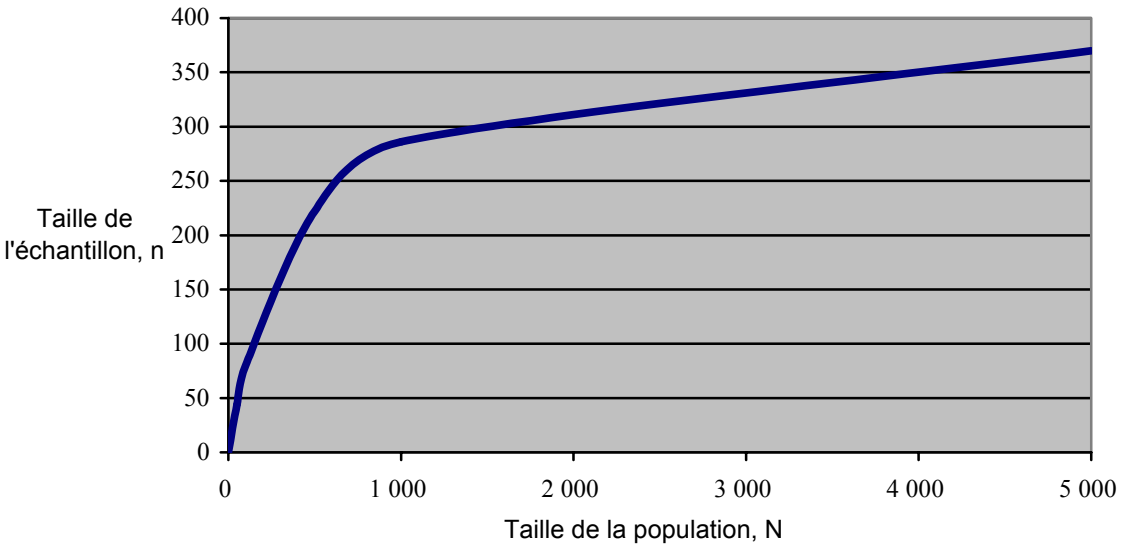
L'importance de la taille de la population sur la taille de l'échantillon varie selon la taille de la population.. Elle est très importante pour une petite population, moyennement importante pour une population de taille moyenne et peu importante pour une grande population.

Revenons, par exemple, à l'enquête sur la satisfaction de la clientèle et disons que la proportion réelle de clients satisfaits est P=0,5 (50 %). Supposons que l'organisme statistique veut tirer un échantillon de la population à l'aide d'un EAS et qu'il veut, pour l'estimation de P, une marge d'erreur de ±0,05 et un taux de confiance de 95 % (c.-à-d., un intervalle de confiance de 0,50±0,05). Le tableau et le graphique suivants illustrent la taille de l'échantillon nécessaire pour différentes tailles de population.

Tableau 3 : Taille de l'échantillon nécessaire pour estimer *P* avec une marge d'erreur de 0,05 et un taux de confiance de 95 %, à l'aide d'un EAS, lorsque *P*=0,5

Taille de la population	Taille de l'échantillon nécessaire
50	44
100	80
500	222
1 000	286
5 000	370
10 000	385
100 000	398
1 000 000	400
10 000 000	400

Graphique 1 : Taille de l'échantillon nécessaire pour estimer *P* avec une marge d'erreur de 0,05 et un niveau de confiance de 95 %, à l'aide d'un EAS, lorsque *P*=0,5



On constate, pour obtenir le degré de précision demandé, que la taille de l'échantillon augmente à un taux qui diminue à mesure qu'augmente la taille de la population. L'organisme statistique a besoin d'une taille d'échantillon de 44 questionnaires remplis pour une population de 50, mais il n'a pas besoin de doubler la taille de l'échantillon à 88 si la population de l'enquête double. La taille de l'échantillon nécessaire approche rapidement n=400 pour des populations d'enquête de N=5 000 et plus. Pour un EAS, 400 questionnaires remplis seraient donc suffisants pour répondre aux besoins d'une précision donnée pour des populations de plus de 5 000 lorsque la proportion de la population réelle est P=0,5.

Une proportion substantielle de la population doit souvent faire l'objet d'une enquête pour obtenir la précision voulue si la population est très petite. Voilà pourquoi, en pratique, on fait souvent le recensement des petites populations.

8.1.2.3 Plan d'échantillonnage et estimateur

La stratégie d'échantillonnage, c'est-à-dire le plan d'échantillonnage et l'estimateur utilisé, ont des répercussions sur la précision. Les techniques de calcul de la taille de l'échantillon pour un degré donné de précision appliquent souvent la formule de la variance d'échantillonnage pour un EAS. Nous avons expliqué au **Chapitre 6 - Plans d'échantillonnage** et au **Chapitre 7 - Estimation** que des plans d'échantillonnage plus complexes utilisant le même estimateur et une taille d'échantillon équivalente peuvent donner des estimations plus ou moins précises. Si la formule de calcul de la taille de l'échantillon suppose l'EAS, un ajustement est donc nécessaire pour tenir compte du plan d'échantillonnage.

En général, si la formule de calcul de la taille de l'échantillon suppose un EAS, mais si un plan d'échantillonnage plus complexe est utilisé, la taille de l'échantillon nécessaire pour obtenir un degré donné de précision doit être multipliée par un facteur intitulé *effet de plan (deff)*. Mentionnons un point tiré du **Chapitre 7 - Estimation :** ***l'effet de plan est le rapport entre la variance d'échantillonnage d'un estimateur, selon un plan d'échantillonnage donné, et la variance d'échantillonnage de l'estimateur d'un EAS ayant la même taille***. Dans un plan d'échantillonnage aléatoire simple, *deff* = 1, et habituellement, *deff* ≤ 1 pour un plan d'échantillonnage stratifié et *deff* ≥ 1 pour un plan d'échantillonnage par grappes.

Il est habituellement possible d'obtenir une estimation des répercussions du plan d'échantillonnage pour les principales variables de l'enquête à partir d'une enquête précédente comprenant le même plan d'échantillonnage, ou un très semblable, et le même genre de matière à l'étude. Obtenir l'effet de plan d'une enquête pilote est une autre option. Si l'organisme statistique prévoit utiliser un plan d'échantillonnage stratifié et s'il n'y a pas d'estimation convenable de l'effet de plan disponible et tirée d'une enquête précédente, *deff* = 1 peut servir à calculer la taille de l'échantillon (c.-à-d. que nous supposons un EAS). La précision des estimations de l'enquête devrait être de qualité comparable à celle obtenue avec un échantillon aléatoire simple et, si la stratification est efficace, la précision sera meilleure. Il est beaucoup plus difficile de décider quel devrait être l'effet du plan d'échantillonnage si un plan d'échantillonnage par grappes est prévu et s'il n'y a pas de connaissances préalables des répercussions des grappes sur la variance d'échantillonnage. Un effet de plan d'au moins deux pourrait être appliqué dans ce cas, mais l'effet de plan peut atteindre jusqu'à six ou sept quand les grappes sont très homogènes.

8.1.2.4. Taux de réponse à l'enquête

S'il veut atteindre la précision voulue pour les estimations de l'enquête, l'organisme statistique doit ajuster la taille de l'échantillon pour le taux de réponse prévu. Il sélectionne à cette fin un large échantillon, compte tenu d'un taux de réponse prévu et estimé à partir d'enquêtes semblables ou d'une enquête pilote dans la même population.

Si la taille de l'échantillon initial calculée est de 400, par exemple, et si un taux de réponse de 75 % est prévu, l'organisme statistique devrait alors sélectionner l'échantillon suivant :

$$n = \frac{400}{0{,}75} = 533.$$

Lorsque l'organisme statistique et le client ont choisi un certain taux de réponse voulu, l'organisme doit faire tous les efforts possibles pour obtenir au moins ce taux de réponse. S'il n'obtient pas le taux de réponse prévu, il y aura des répercussions sur la précision des résultats de l'enquête. Un taux de réponse

inférieur donnera une taille d'échantillon plus petite que celle qui est nécessaire pour atteindre la précision voulue et, d'autre part, un taux de réponse supérieur aura l'effet contraire.

Cet ajustement est appliqué en supposant que les unités manquantes sont aléatoires, c.-à-d. que les non-répondants ont des caractéristiques semblables à celles des répondants. Simplement augmenter la taille de l'échantillon est insuffisant pour réagir correctement à une non-réponse totale. Un biais éventuel est toujours possible si les non-répondants sont différents des répondants du point de vue des caractéristiques d'intérêt de l'enquête. (Voir le **Chapitre 7 - Estimation** et le **Chapitre 10 - Traitement** pour déterminer comment traiter le biais dû à la non-réponse.)

8.1.3 Formules de calcul de la taille de l'échantillon

Les formules suivantes peuvent servir à calculer la taille de l'échantillon nécessaire pour obtenir un degré donné de précision pour une moyenne ou proportion estimée.

i. Précision d'une moyenne estimée, $\hat{\bar{Y}}$, pour un échantillon aléatoire simple (taux de réponse de 100 %)

La marge d'erreur et la formule appliquée à l'erreur-type d'une estimation pour un EAS servent souvent à déterminer la taille de l'échantillon. Voici l'équation pour une erreur-type estimée d'une moyenne estimée, $\hat{\bar{Y}}$, d'un EAS sans remise :

$$\hat{SE}(\hat{\bar{Y}}) = \sqrt{\left(1 - \frac{n}{N}\right)} \frac{\hat{S}}{\sqrt{n}} \quad (1)$$

où $\hat{S}$ est la racine carrée de l'estimation de la variance de la population de y_i (voir aussi le **Chapitre 7 - Estimation**).

Notons e la marge d'erreur nécessaire :

$$e = z\sqrt{\left(1 - \frac{n}{N}\right)} \frac{\hat{S}}{\sqrt{n}} \qquad (2)$$

où z est déterminé selon le niveau de confiance. La solution pour n donne :

$$n = \frac{z^2 \hat{S}^2}{e^2 + \frac{z^2 \hat{S}^2}{N}} \qquad (3)$$

Les étapes suivantes sont donc nécessaires pour déterminer n :
- une marge d'erreur voulue, e,
- une valeur correspondante à un niveau de confiance voulu, z,
- la taille de la population, N,
- une estimation de la variabilité de la population, $\hat{S}^2$.

Ce dernier point est plus difficile à obtenir et une approximation est souvent faite à l'aide d'études précédentes d'une population semblable. (Il est aussi possible de calculer la taille de l'échantillon

nécessaire à l'aide d'un coefficient donné de variation. Ce point est considéré à la Section 8.2.1.2 pour un total estimé d'un EAS stratifié.)

ii. Précision d'une proportion estimée, $\hat{P}$, pour un échantillon aléatoire simple (taux de réponse de 100 %)

La précision nécessaire sera déterminée dans ce cas selon la marge d'erreur et la caractéristique d'intérêt sera la proportion de la population, *P*, qui fait partie de l'une des deux catégories. Nous savons que la proportion estimée, $\hat{P}$, pour les grandes populations, est approximativement distribuée normalement et la variabilité de la caractéristique binaire, y_i, de la population peut être estimée comme suit :

$$\hat{S}^2 = \hat{P}(1-\hat{P})$$

L'équation (3) devient donc :

$$n = \frac{z^2 \hat{P}(1-\hat{P})}{e^2 + \frac{z^2 \hat{P}(1-\hat{P})}{N}}$$

Si une bonne estimation de la proportion, $\hat{P}$, est disponible avant l'enquête, il faudrait l'utiliser dans l'équation ci-dessus. Autrement, s'il n'y a pas de données sur la population, $\hat{P} = 0{,}5$ peut-être utilisée, le résultat étant la taille d'échantillon maximale, étant donné les autres suppositions.

Noues expliquerons dans la section suivante qu'il faut faire une estimation de l'effet de plan si le plan n'est pas un EAS et une estimation du taux de réponse, *r*, est nécessaire si le taux de réponse à l'enquête est inférieur à 100 %.

iii. Approche étape par étape pour déterminer la taille de l'échantillon, compte tenu de la précision d'une proportion estimée, $\hat{P}$, pour tout plan d'échantillonnage (lorsque le taux de réponse est <100 %)

Une approche étape par étape est appliquée dans les exemples suivants pour calculer la taille de l'échantillon. Une taille d'échantillon initiale est d'abord calculée et elle est ensuite ajustée, compte tenu de la taille de la population, de l'effet du plan d'échantillonnage et du taux de réponse.

1. Taille de l'échantillon initial

Remarquez l'utilisation dans l'équation (1) du facteur de correction d'échantillonnage pour population finie $(1 - n / N)$, afin d'apporter une correction, compte tenu de la taille de la population. Si ce facteur est omis, une estimation préliminaire de la taille de l'échantillon, n_1, peut être obtenue simplement comme suit :

$$n_1 = \frac{z^2 \hat{P}(1-\hat{P})}{e^2}$$

On remarquera la formule pour n_1 est aussi valable si *e* et $\hat{P}$ sont exprimées en pourcentage, et non en proportions.

2. Ajustement pour la taille de la population à l'aide de l'équation suivante (le résultat aura des répercussions seulement pour les populations de petite taille ou de taille moyenne) :

$$n_2 = n_1 \frac{N}{N + n_1}$$

3. Si le plan d'échantillonnage n'est pas un échantillon aléatoire simple, la formule suivante peut servir à ajuster la taille de l'échantillon pour l'effet du plan d'échantillonnage :

$$n_3 = Deff \times n_2$$

où *deff* est l'effet du plan d'échantillonnage et, habituellement :
deff = 1 pour les plans d'échantillonnage aléatoires simples,
deff < 1 pour les plans d'échantillonnage stratifiés,
deff > 1 pour les plans d'échantillonnage par grappes ou à plusieurs degrés.

4. En bout de ligne, ajustement pour le taux de réponse, afin de déterminer la taille finale de l'échantillon, *n* :

$$n = \frac{n_3}{r}$$

où *r* est le taux de réponse prévu.

8.1.3.1 Exemples de choix de taille de l'échantillon

Les exemples suivants illustrent l'approche étape par étape du calcul de la taille de l'échantillon.

Exemple 8.1 : EAS

L'éditeur d'une revue veut obtenir une estimation de la satisfaction des lecteurs en général. Il serait possible de communiquer avec les 2 500 abonnés à l'aide d'un questionnaire envoyé par la poste, mais l'éditeur a décidé d'interviewer un échantillon aléatoire simple par téléphone à cause des contraintes de temps. Combien de lecteurs faudrait-il interviewer?

Voici certaines hypothèses:
- l'éditeur sera satisfait si la proportion de la population réelle est à ±0,10 de la proportion de la population estimée, compte tenu des résultats de l'échantillon, c.-à-d. que la marge d'erreur nécessaire, $e = 0{,}10$;
- l'éditeur veut obtenir un niveau de confiance de 95 % dans les estimations de l'enquête (c.-à-d. qu'il y aurait seulement une chance sur 20 d'obtenir un échantillon qui donne une estimation hors de l'étendue $\hat{P} \pm 0{,}10$, donc $z = 1{,}96$);
- un EAS sera utilisé;
- un taux de réponse de 65 % environ est prévu, c.-à-d. que r=0,65;
- étant donné qu'il n'y a pas d'estimation de $\hat{P}$ disponible, le degré de satisfaction de la clientèle est donc supposé être $\hat{P} = 0{,}5$.

Voici le calcul de la taille de l'échantillon nécessaire :

1. Calcul de la taille de l'échantillon initial, n_1 :

$$n_1 = \frac{z^2 \hat{P}(1-\hat{P})}{e^2}$$
$$= \frac{(1{,}96)^2 (0{,}50)(0{,}50)}{(0{,}10)^2} = 96$$

2. Ajustement de la taille de l'échantillon pour tenir compte de la taille de la population :

$$n_2 = n_1 \frac{N}{N + n_1}$$
$$= 96 \frac{2500}{(2500 + 96)} = 92$$

3. Ajustement de la taille de l'échantillon, compte tenu de l'effet de plan :

$$n_3 = Deff \times n_2$$
$$= n_2 = 92$$

Dans cet exemple, *deff* = 1 parce qu'on suppose qu'un EAS sera utilisé.

4. Ajustement pour le taux de réponse, afin de déterminer la taille de l'échantillon final, *n* :

$$n = \frac{n_3}{r}$$
$$= \frac{92}{0{,}65} = 142$$

Remarque : Si un taux de réponse d'au moins 65 % n'est pas réalisé pendant l'enquête, la taille de l'échantillon final sera plus petite que prévu et les estimations de l'enquête pourraient donc être moins précises que l'exige la planification. Si un taux de réponse plus élevé est obtenu, l'échantillon sera plus large que prévu et les estimations de l'enquête pourraient être plus précises.

Après ces étapes, l'éditeur devrait tirer un EAS de 142 des 2 500 abonnés pour estimer le niveau de satisfaction des lecteurs de la revue avec une marge d'erreur de 0,10 et un niveau de confiance de 95 %, compte tenu d'un taux de réponse prévu de 65 %.

Exemple 8.2 : EAS stratifié

Une enquête d'opinion publique est prévue pour déterminer la proportion de la population en faveur de l'aménagement d'un nouveau parc provincial. La population comprend tous les adultes dans deux villes et en milieu rural. Un échantillon aléatoire simple des adultes dans chaque ville et un autre pour le milieu rural seront sélectionnés. Il faut déterminer la taille de l'échantillon nécessaire dans chaque strate.

La taille de la population est de 657 500 et la répartition est la suivante :

Tableau 4 : Population des trois strates

H	Strate	Population (N_h)
1	Ville 1	400 000
2	Ville 2	250 000
3	Milieu rural	7 500
Total		**657 500**

Les besoins de données particuliers de l'enquête déterminent la taille de l'échantillon nécessaire. Les deux options suivantes peuvent être considérées.

Option 1 : Marge d'erreur pour les estimations de la population dans l'ensemble

Supposons que des estimations précises pour chaque strate *ne* sont *pas* nécessaires. Une estimation avec marge d'erreur de ± 0,05 et un niveau de confiance de 95 % pour le *secteur dans l'ensemble* sont suffisants. Une estimation préliminaire de la proportion n'est pas disponible et nous supposons que $\hat{P} = 0{,}5$. Un taux de réponse de 50 % est prévu.

1. Calcul de la taille de l'échantillon initial, n_1 :

$$n_1 = \frac{z^2 \hat{P}(1-\hat{P})}{e^2}$$
$$= \frac{(1{,}96)^2 (0{,}50)(0{,}50)}{(0{,}05)^2} = 384$$

2. Calcul de la taille de l'échantillon modifiée, n_2 :

$$n_2 = n_1 \frac{N}{N + n_1}$$
$$= 384 \frac{657\,500}{657\,500 + 384} = 384$$

(Remarque : Si la valeur n_1/N est négligeable, on peut supposer que $n_2 = n_1$)

3. Ajustement pour l'effet de plan :

$$n_3 = Deff \times n_2$$
$$= n_2 = 384$$

Habituellement, *deff* < 1 pour un échantillonnage aléatoire stratifié. Dans le présent exemple, il n'y a pas d'estimation disponible de *deff* et, si on pose que *deff* = 1, le résultat devrait vous donner une estimation plus raisonnable de la taille de l'échantillon (c.-à-d. plus large).

4. Ajustement pour le taux de réponse, afin de déterminer la taille de l'échantillon final, *n* :

$$n = \frac{n_3}{r}$$
$$= \frac{384}{0{,}50} = 768$$

La taille de l'échantillon nécessaire est 768. On verra à la Section 8.2 comment répartir ces 768 unités échantillonnées sur trois strates.

Option 2 : Marge d'erreur pour chaque estimation de strate

Supposons que le client demande des résultats ayant une marge d'erreur de "0,05 et un taux de confiance de 95 % pour *chaque* strate. Il faut maintenant calculer la taille de l'échantillon individuel pour chaque strate (c.-à-d. que chaque strate est traitée comme une population en soi).

Remarquez que les Villes 1 et 2 ont de larges populations et que la taille de leur population ne devrait pas avoir de répercussion sur la taille de l'échantillon. Compte tenu des hypothèses ci-dessus, la taille de l'échantillon de chacune de ces deux strates est donc 768. La population plus petite du milieu rural devrait cependant avoir des répercussions sur la taille de l'échantillon.

Milieu rural :

$$n_1 = \frac{z^2 \hat{P}(1-\hat{P})}{e^2} = \frac{(1{,}96)^2 (0{,}50)(0{,}50)}{(0{,}05)^2} = 384$$

$$n_2 = n_1 \frac{N}{N + n_1} = 384\left(\frac{7\,500}{7\,500 + 384}\right) = 366$$

$$n_3 = 366$$

$$n = \frac{n_3}{r} = \frac{366}{0{,}50} = 732$$

La taille *totale* de l'échantillon est donc 768 (Ville 1) + 768 (Ville 2) + 732 (milieu rural) = 2 268.

En comparant les options 1 et 2, la taille de l'échantillon total de 2 268 pour l'option 2 est près de trois fois plus grande que la taille de l'échantillon de 768 pour l'option 1. Autrement dit, si une seule estimation pour la population totale des trois strates est nécessaire, la taille de l'échantillon nécessaire est inférieure à celle qu'il faudra déterminer si des estimations précises par strate sont demandées parce qu'il faudrait alors établir des tailles d'échantillon suffisantes dans chaque strate.

Cet exemple illustre clairement l'importance de l'examen des besoins de précision pour chaque domaine distinct. Si de nombreux domaines sont nécessaires, les répercussions sur la taille de l'échantillon total peuvent être importantes et donner éventuellement une taille d'échantillonnage supérieure au budget et aux ressources opérationnelles du client. En général, plus on demande d'estimations de domaines, plus la taille de l'échantillon doit être grande. Il faut donc éventuellement en venir à des compromis pour obtenir des niveaux d'erreur acceptables. On peut choisir d'augmenter les niveaux tolérables d'erreur dans chaque strate, ou combiner deux domaines ou plus. Ceci sera repris à la Section 8.2.2.2.

8.1.4 Contraintes opérationnelles, de coûts et de temps

Nous avons considéré un seul aspect de la taille de l'échantillon jusqu'à maintenant, c'est-à-dire la taille de l'échantillon nécessaire, afin d'obtenir un degré de précision en particulier pour les estimations de l'enquête les plus importantes. En pratique, le temps, les coûts et d'autres restrictions opérationnelles sont aussi au premier plan.

Dans de nombreux sondages, les fonds sont attribués et les délais sont déterminés avant même que les décisions soient prises sur les particularités de l'enquête. La taille de l'échantillon nécessaire pour procéder au sondage peut se révéler plus grande que l'échantillon qu'il est possible d'obtenir, compte tenu des fonds disponibles. S'il est impossible d'obtenir d'autres fonds, il faudra peut-être réduire la taille de l'échantillon et diminuer ainsi la précision des estimations. On pourrait aussi renoncer aux estimations de certains domaines. La question se pose aussi pour les considérations de temps. Si le temps attribué est insuffisant, il faudra peut-être limiter la taille et l'envergure de l'enquête pour respecter les délais.

Les contraintes opérationnelles qu'impose la méthode de collecte des données choisie, la disponibilité du personnel sur place, la disponibilité du personnel de codage et de vérification et les installations de traitement ont aussi des répercussions sur la taille de l'échantillon. Il peut s'agir en fait des points les plus importants pour déterminer la taille de l'échantillon. Nous avons considéré au **Chapitre 4 - Méthodes de collecte des données**, par exemple, que les interviews sur place permettent d'obtenir de l'information plus complexe et des taux de réponse plus élevés, mais elles coûtent cher. Il n'est donc pas toujours pratique de les appliquer à de gros échantillons.

8.2 Répartition de l'échantillon pour des plans d'EAS stratifié

Pour déterminer l'efficience de l'échantillonnage stratifié, il est important de considérer comment la taille totale de l'échantillon, n, est répartie dans chaque strate. Nous avons expliqué au **Chapitre 6 - Plans d'échantillonnage** que, dans un plan d'échantillonnage stratifié, le nombre total d'unités de la population, c.-à-d. N, est divisé en L strates sans chevauchement de taille $N_1, N_2, \ldots, N_L$, respectivement. La taille de la population est donc égale à la somme, pour toutes les strates, du nombre d'unités dans la population : $N = N_1 + N_2 + \ldots + N_L$. Un échantillon est tiré indépendamment de chaque strate. La taille de l'échantillon dans chaque strate est n_h ($h = 1, 2, \ldots, L$), où $n = n_1 + n_2 + \ldots + n_L$.

La répartition de l'échantillon, n, en L strates est possible en appliquant l'un ou l'autre des critères suivants. La taille totale de l'échantillon peut être déterminée à l'aide des méthodes décrites auparavant dans ce chapitre et répartie ensuite entre les strates (ou taille de l'échantillon fixe). On peut aussi déterminer la taille de l'échantillon nécessaire dans chaque strate pour obtenir la précision voulue et faire la somme, afin d'obtenir la taille de l'échantillon total (ou coefficient de variation fixe, si la précision voulue est exprimée en coefficient de variation).

8.2.1 Critères de répartition

Cette section décrit en détail la différence entre les répartitions selon une taille d'échantillon fixe et un coefficient de variation fixe.

8.2.1.1 Taille d'échantillon fixe

Une taille d'échantillon fixe n est attribuée aux strates d'une façon particulière dans ce cas. La proportion de l'échantillon attribuée à la h^e strate est $a_h = n_h / n$, où chaque a_h se situe entre 0 and et 1 inclusivement (c.-à-d. $0 \le a_h \le 1$) et la somme des a_h est égale à 1 (c.-à-d. $\sum_{h=1}^{L} a_h = 1$).

Dans chaque strate h, la taille de l'échantillon n_h est donc égale au résultat de la taille de l'échantillon total n et de la proportion a_h de l'échantillon tiré de cette strate en particulier :

$$n_h = n \times a_h \qquad (4)$$

Si la strate a une proportion a_h = ½, par exemple, la moitié de l'échantillon complet est donc attribué à cette strate.

Compte tenu de ce critère de répartition, la taille de l'échantillon n dans l'ensemble étant connue, la taille de l'échantillon n_h pour chaque strate peut être calculée dès que la valeur a_h est déterminée pour chaque strate. Il y a de nombreuses façons de déterminer a_h : l'une d'elle consiste à déterminer les valeurs de a_h qui minimisent la variance d'échantillonnage des caractéristiques d'intérêt. La Section 8.2.2. explique comment déterminer la valeur de a_h.

8.2.1.2 Coefficient de variation fixe

La solution de rechange à l'établissement de la taille de l'échantillon, n, est le calcul de la taille de l'échantillon nécessaire dans chaque strate, n_h, compte tenu d'un certain degré de précision pour les estimations *dans l'ensemble*. Il faut alors trouver la taille de l'échantillon n_h (h = 1, 2, … , L) pour chaque strate, afin que le coefficient de variation des estimations dans l'ensemble ne soit pas supérieur à la valeur voulue CV.

Considérons, par exemple, l'estimation d'un total, $\hat{Y}$, à partir d'un échantillon aléatoire simple stratifié. L'équation permettant d'obtenir le coefficient de variation d'un total estimé à partir d'un échantillon stratifié peut être exprimé de la façon suivante pour la taille de l'échantillon total, n[1] :

$$n = \frac{\sum_{h=1}^{L} N_h^2 S_h^2 / a_h}{CV^2 Y^2 + \sum_{h=1}^{L} N_h S_h^2}$$

où :

N_h est la taille de la strate,

S_h^2 est la variabilité des unités, y_i, de la strate h de la population,

a_h est la proportion de l'échantillon attribuée à la strate,

[1] Consulter la Section 7.3.2.4 pour obtenir de l'information sur la variance d'échantillonnage d'un total estimé pour un échantillon stratifié. L'équation ci-dessus peut être obtenue en établissant que $CV(Y) = \sqrt{Var(\hat{Y})} / Y$ où $Var(Y) = N^2 Var(\hat{\bar{Y}})$ et $N = \sum_h N_h$.

CV est le coefficient de variation exigé pour *Y*,
Y est le total.

Remarque : Dans la formule ci-dessus, nous supposons que $n_h = n \times a_h < N_h$, c.-à-d. que la taille de l'échantillon attribué par strate est inférieure à la taille de la population par strate. Consultez à cette fin la Section 8.2.3. La variance de la population, S_h^2, peut être estimée à l'aide de $\hat{S}_h^{\,2}$, comme suit :

$$\hat{S}_h^2 = \frac{\sum_{i=1}^{n_h} (y_{hi} - \bar{y}_h)^2}{n_h - 1}$$

où $\bar{y}_h$, la moyenne de la strate de l'échantillon, est :

$$\bar{y}_h = \frac{\sum_{i=1}^{n_h} y_{hi}}{n_h}$$

(Remarque : Si y_{hi} est une variable binaire, la moyenne de la strate est une proportion, c.-à-d. que $\bar{y}_h = \hat{P}_h$, et $\hat{S}_h^{\,2} = \hat{P}_h(1 - \hat{P}_h)$).

Substituant $n_h = n \times a_h$, $\hat{S}_h^2$ et $\hat{Y}$ dans l'équation précédente pour *n*, on obtient le résultat suivant pour n_h :

$$n_h = a_h \frac{\sum_{h=1}^{L} N_h^2 \hat{S}_h^2 / a_h}{CV^2 \hat{Y}^2 + \sum_{h=1}^{L} N_h \hat{S}_h^2} \qquad (5)$$

Après avoir déterminé la valeur de a_h pour chaque strate, on peut calculer chaque taille d'échantillon n_h. N'oubliez pas : nous avons expliqué auparavant dans ce chapitre comment déterminer la taille de l'échantillon et, pour trouver n_h, il faut établir la précision nécessaire (sous forme de coefficient de variation dans ce cas), la variabilité estimée de la population, $\hat{S}_h^2$, et la taille de la population, N_h. Il faudrait aussi apporter un ajustement pour les non-réponses à la taille d'échantillon n_h finale.

L'approche du coefficient de variation fixe pour répartir l'échantillon est plus compliquée que l'approche de la taille de l'échantillon fixe et seulement cette dernière sera utilisée pour illustrer la répartition de l'échantillon.

8.2.2 Méthodes de répartition de l'échantillon

Les équations (4) et (5) sont les outils élémentaires de répartition de l'échantillon stratifié. Chaque équation peut être appliquée dès que les valeurs ont été déterminées pour chaque a_h. Le choix d'une a_h pour chaque strate peut être classé en deux genres de méthodes : répartition proportionnelle ou non proportionnelle. Ces méthodes dépendent de certaines quantités : la taille de la population de la strate, une autre mesure de la taille de la strate, la variabilité de la population de la strate ou le coût de l'enquête dans la strate.

8.2.2.1 Répartition proportionnelle

Dans la répartition proportionnelle, ou répartition proportionnelle à N, la taille de l'échantillon, n_h, de chaque strate est proportionnelle à la taille de la population, N_h, de la strate. Une part plus importante de l'échantillon est donc attribuée à une strate plus grande qu'à une strate plus petite. On obtient ainsi un taux de sondage, $f_h = n_h / N_h$, semblable dans chaque strate et égal au taux de sondage dans l'ensemble, $f = n / N$. On obtient donc l'équation suivante :

$$n_h = \frac{N_h}{N} n$$

Le résultat de la répartition proportionnelle à N est donc $a_h = n_h / n = N_h / N$. Autrement dit, le facteur de répartition a_h pour chaque strate est égal au ratio de la taille de la population de la strate à la taille de la population entière. Ce genre de répartition est illustré au **Chapitre 7 - Estimation**.

La répartition proportionnelle à N est souvent utilisée lorsque l'information sur les variances de strate de la population ne sont pas disponibles. Elle n'est donc pas utilisée pour calculer les tailles d'échantillon pour une variance de coefficient fixe parce que l'application de cette approche demande des connaissances sur la variance de chaque strate. La répartition proportionnelle à N sert aussi à l'autopondération du plan d'échantillonnage (c.-à-d. que toutes les unités ont la même probabilité d'inclusion, π, et la même pondération du plan d'échantillonnage, $1 / \pi$, évidemment).

La répartition proportionnelle à N appliquée à l'échantillonnage stratifié est considérablement plus efficiente que l'échantillonnage aléatoire simple de la population complète si les moyennes de strate, $\overline{Y}_h$, sont considérablement différentes l'une de l'autre. Si les strates sont cependant formées de sorte que leurs moyennes, $\overline{Y}_h$, soient à peu près les mêmes, la stratification avec répartition proportionnelle à N donne seulement une légère diminution de la variance d'échantillonnage. La répartition proportionnelle à N n'est jamais pire que l'échantillonnage aléatoire simple et n'a donc jamais d'effet du plan d'échantillonnage, *deff*, plus grand que 1.

L'exemple suivant illustre une répartition proportionnelle à N à l'aide d'une taille d'échantillon fixe, n.

Exemple 8.2 (suite) :

Dans l'option 1 de l'exemple 8.2, le calcul de la taille de l'échantillon n donne 768 personnes. La répartition proportionnelle à N pour une taille d'échantillon fixe est utilisée, afin de déterminer comment répartir 768 personnes en trois strates.

1. Calcul de la valeur du facteur de répartition a_h pour chaque strate à l'aide de la répartition proportionnelle à N.

Ville 1 :	Ville 2 :	Milieu rural :
$a_1 = \frac{N_1}{N}$	$a_2 = \frac{N_2}{N}$	$a_3 = \frac{N_3}{N}$
$= \frac{400\,000}{657\,500}$	$= \frac{250\,000}{657\,500}$	$= \frac{7\,500}{657\,500}$
$= 0{,}6084$	$= 0{,}3802$	$= 0{,}0114$

2. Calcul de la taille de l'échantillon n_h pour chaque strate.

Ville 1 :	Ville 2 :	Milieu rural :
$n_1 = na_1$	$n_2 = na_2$	$n_3 = na_3$
$= 768 \times 0{,}6084$	$= 768 \times 0{,}3802$	$= 768 \times 0{,}0114$
$= 467$	$= 292$	$= 9$

On constate que la majorité de l'échantillon est réparti entre les strates plus larges, Ville 1 et Ville 2 où 467 et 292 personnes sont échantillonnées respectivement. La plus petite strate, le milieu rural, obtient une plus petite portion de l'échantillon complet, soit un échantillon de neuf personnes seulement. Les résultats sont résumés au tableau suivant.

Tableau 5 : Répartition proportionnelle à N

H	**Strate**	**Population (N_h)**	a_h	n_h	$f_h = n_h/N_h$
1	Ville 1	400 000	0,6084	467	0,0012
2	Ville 2	250 000	0,3802	292	0,0012
3	Milieu rural	7 500	0,0114	9	0,0012
Total		**657 500**	**1**	**768**	**0,0012**

La répartition proportionnelle à N du tableau ci-dessus donne un plan d'échantillonnage autopondéré parce que le taux de sondage, f_h, est égal à 0,0012 dans les trois strates.

La différence entre la répartition ci-dessus et la taille de l'échantillon déterminée à l'option 2 de l'exemple 8.2 est remarquable : la répartition ci-dessus répond à un besoin de précision pour une estimation de la population *dans l'ensemble* et l'option 2 de l'exemple 8.2 répond à un besoin de précision pour *chaque* strate.

8.2.2.2 Répartition non proportionnelle

Les taux de sondage de la répartition non proportionnelle sont différents d'une strate à l'autre. Les méthodes de répartition non proportionnelles suivantes seront présentées et expliquées : répartition proportionnelle à Y, répartition proportionnelle à la $\sqrt{N}$, répartition proportionnelle à la $\sqrt{Y}$, répartition optimale, répartition de Neyman et répartition optimale lorsque les variances sont égales. La terminologie peut semer la confusion parce que certaines méthodes de répartition non proportionnelles sont intitulées méthodes de répartition proportionnelle (p. ex., répartition proportionnelle à Y). Il ne rappeler que la méthode de répartition est considérée non proportionnelle dès que le taux de sondage est différent entre au moins deux strates.

8.2.2.2.1 Répartition proportionnelle à Y

Étant donné une variable d'enquête, y_{hi}, vue comme une mesure de la taille pour la i^e unité de la h^e strate, les tailles de l'échantillon, n_h, peuvent être calculées comme proportions de Y_h, une mesure agrégée de la taille de la strate h. Ce genre de répartition est intitulée répartition proportionnelle à Y. Dans ce cas, $a_h = Y_h / Y$. Cela signifie que le facteur de répartition a_h pour chaque strate est équivalent au ratio de la mesure de la taille de la strate à la mesure de la taille de la population entière.

La répartition proportionnelle à Y est une méthode très populaire pour les enquêtes sur les entreprises où l'on trouve souvent que la distribution des y_{hi} est asymétrique (c.-à-d. qu'elle a des valeurs extrêmes à une

queue de la distribution). Des exemples typiques sont *l'emploi* dans les industries de fabrication et les *ventes* dans les industries de détail. Dans chaque cas, un petit nombre d'entreprises peuvent représenter un pourcentage élevé du *total de l'emploi* ou du *total des ventes*. D'autre part, les autres entreprises en plus grand nombre peuvent représenter seulement une petite fraction de *l'emploi total* ou du *total des ventes*.

Dans les enquêtes sur les entreprises, les strates sont habituellement établies selon la mesure de la taille disponible (p. ex., le nombre d'employés, le revenu brut de l'entreprise, les ventes nettes). La mesure de la taille peut servir, notamment, à créer trois strates pour les petites, moyennes et grandes entreprises. La strate qui comprend le plus grand nombre d'unités est souvent plus variable que d'autres. Dans un cas extrême, la répartition proportionnelle à Y se traduit par l'échantillonnage avec certitude des plus importantes unités d'une population asymétrique.

La répartition proportionnelle à Y donne une meilleure précision que la répartition proportionnelle à N pour les estimations d'enquête qui sont plus fortement corrélées avec Y_h qu'avec la taille de la strate, N_h.

8.2.2.2.2 Répartition proportionnelle à $\sqrt{N}$

Toutes les méthodes de répartition présentées jusqu'à maintenant ciblent uniquement la précision de l'estimation globale $\hat{Y}$. Le client peut cependant être intéressé à obtenir aussi une bonne précision pour les estimations de la strate, $\hat{Y}_h$. Si les strates sont des provinces, par exemple, les estimations provinciales sont probablement aussi importantes que les estimations nationales. La répartition par strate à l'aide de la répartition proportionnelle à la $\sqrt{N}$ peut améliorer la précision des estimations de la strate. Le paramètre de répartition a_h est alors calculé ainsi :

$$a_h = \frac{\sqrt{N_h}}{\sum_{h=1}^{L} \sqrt{N_h}}$$

Autrement dit, le paramètre de répartition a_h est égal au ratio de la racine carrée de la taille de la population de la strate à la somme de la racine carrée de la taille de la population de toutes les strates.

La répartition proportionnelle à $\sqrt{N}$ n'est pas aussi efficace que d'autres méthodes de répartition quant à la précision maximale dans l'ensemble. Elle peut cependant donner de meilleures estimations au niveau de la strate. Elle est souvent utilisée comme compromis entre la répartition optimale (voir 8.2.2.2.4) et la répartition pour répondre à toutes les contraintes des domaines (où les domaines sont définis comme des strates). La répartition optimale pour les estimations nationales, par exemple, peut donner de grandes variances d'échantillonnage pour des domaines d'intérêt plus petits (p. ex., provinces) et la répartition de l'échantillon total pour répondre aux contraintes des domaines (comme dans l'option 2 de l'exemple 8.2) peut donner une répartition inefficiente de l'échantillon total. La répartition proportionnelle à la $\sqrt{N}$ est un compromis entre la répartition dans l'ensemble et au niveau des domaines.

L'exemple suivant illustre l'application de la répartition proportionnelle à la $\sqrt{N}$ pour une taille d'échantillon fixe, n.

Exemple 8.2 (suite) :

Dans l'exemple précédent, un échantillon fixe de 768 personnes a été réparti en trois strates à l'aide de la répartition proportionnelle à N. La répartition par strate ci-dessous est faite à l'aide de la méthode de la répartition proportionnelle à la $\sqrt{N}$.

1. Calcul de la valeur du facteur de répartition a_h pour chaque strate à l'aide de la répartition proportionnelle à la $\sqrt{N}$.

Ville 1 :

$$a_1 = \frac{\sqrt{N_1}}{\sum_{h=1}^{3}\sqrt{N_h}} = \frac{632{,}46}{1\,219{,}06} = 0{,}5188$$

Ville 2 :

$$a_2 = \frac{\sqrt{N_2}}{\sum_{h=1}^{3}\sqrt{N_h}} = \frac{500}{1\,219{,}06} = 0{,}4102$$

Milieu rural :

$$a_3 = \frac{\sqrt{N_3}}{\sum_{h=1}^{3}\sqrt{N_h}} = \frac{86{,}60}{1\,219{,}06} = 0{,}0710$$

2. Calcul de la taille de l'échantillon n_h pour chaque strate.

Ville 1 :

$$n_1 = na_1 = 768 \times 0{,}5188 = 398$$

Ville 2 :

$$n_2 = na_2 = 768 \times 0{,}4102 = 315$$

Milieu rural :

$$n_3 = na_3 = 768 \times 0{,}0710 = 55$$

Le tableau suivant résume les résultats et compare la répartition proportionnelle à N et la répartition proportionnelle à la $\sqrt{N}$.

Tableau 6 : Comparaison de la répartition proportionnelle à N et de la répartition proportionnelle à $\sqrt{N}$

		Répartition proportionnelle à N				Répartition proportionnelle à la $\sqrt{N}$			
h	**Strate**	**Population (N_h)**	A_h	n_h	f_h	$\sqrt{N_h}$	a_h	n_h	f_h
1	Ville 1	400 000	0,6084	467	0,0012	632,46	0,5188	398	0,0010
2	Ville 2	250 000	0,3802	292	0,0012	500	0,4102	315	0,0013
3	Milieu rural	7 500	0,0114	9	0,0012	86,60	0,0710	55	0,0073
8.2	**Total**	**657 500**	**1**	**768**	**0,0012**	**1 219,06**	**1**	**768**	**0,0012**

La répartition proportionnelle à la $\sqrt{N}$ donne une taille d'échantillon plus petite pour la Ville 1 que la répartition proportionnelle à N. D'autre part, elle donne un échantillon plus grand pour la Ville 2 et le Milieu rural. La précision de l'estimation pour la Ville 2 et le Milieu rural est donc meilleure avec la répartition proportionnelle à la $\sqrt{N}$ qu'avec la répartition proportionnelle à N parce que la taille de l'échantillon est plus grand. (Il serait difficile d'obtenir une bonne estimation du milieu rural à partir de neuf unités seulement.) La diminution de la taille de l'échantillon de la Ville 1 aura de légères répercussions sur la précision de l'estimation. L'augmentation de la taille de l'échantillon de la Ville 2

aura simplement une répercussion légèrement positive sur la précision des résultats. L'augmentation de la taille d'échantillon du Milieu rural améliore cependant beaucoup la précision des estimations. La précision à la hausse en Milieu rural surpasse la perte de précision dans la Ville 1.

8.2.2.2.3 Répartition proportionnelle à $\sqrt{Y}$

Un autre moyen de garantir que l'estimation dans l'ensemble et les estimations de la strate sont raisonnablement fiables est le recours à la répartition proportionnelle à la $\sqrt{Y}$, où y_{hi} est une mesure de la taille. Il s'agit d'une autre mesure plus précise que la répartition proportionnelle à la $\sqrt{N}$ pour les estimations de l'enquête corrélées davantage avec la variable de la taille, Y_h, qu'avec la taille de la strate, N_h. Voici le paramètre de répartition a_h :

$$a_h = \frac{\sqrt{Y_h}}{\sum_{h=1}^{L} \sqrt{Y_h}}$$

Cela signifie que le paramètre de répartition a_h est égal au rapport entre la racine carrée de la mesure de la taille de la strate et la somme de la racine carrée de la mesure de la taille de toutes les strates.
Tout comme dans le cas de la répartition proportionnelle à la $\sqrt{N}$, le recours à la répartition proportionnelle à la $\sqrt{Y}$ pour calculer les valeurs de a_h (et ultérieurement les valeurs de n_h) n'est pas aussi efficient que l'application d'autres méthodes de répartition quant à la précision dans l'ensemble. Cette répartition donne cependant des estimations plus précises à l'échelon de la strate.

Les répartitions proportionnelles à la $\sqrt{N}$ et à la $\sqrt{Y}$ sont parfois intitulées *répartitions par puissance* où l'attribution d'une puissance à Y, par exemple, est définie plus généralement comme suit :

$$a_h = \frac{Y_h{}^p}{\sum_{h=1}^{L} Y_h{}^p}$$

où p est habituellement une fraction (p. ex., ½). On trouvera dans Bankier (1988) davantage de détails sur les répartitions par puissance.

8.2.2.2.4 Répartition optimale

Lorsque le coût de l'interview par unité est différent d'une strate à l'autre et que les variances de la population, S_h^2, varient énormément, une méthode de répartition non proportionnée intitulée répartition optimale peut être considérée. C'est la seule méthode de répartition présentée ici qui tient compte des coûts.

Afin d'utiliser la répartition optimale, l'organisme statistique a besoin d'une fonction pour modéliser le coût. La plus simple fonction du coût total est exprimée comme suit :

$$Coût = C = c_o + \sum_{h=1}^{L} c_h n_h$$

où c_h est le coût par unité de sondage dans la strate h (h = 1, 2, … , L) et c_0 est un coût général fixe. Cette fonction coût est meilleure lorsque le principal article du coût est celui de l'interview ou de la mesure de chaque unité.

Le paramètre de répartition a_h utilisé pour la répartition optimale est calculé comme suit :

$$a_h = \frac{N_h S_h \Big/ \sqrt{c_h}}{\sum_{h=1}^{L} N_h S_h \Big/ \sqrt{c_h}}$$

La répartition optimale minimise la variance de l'estimation pour un coût donné et, de même, elle minimise le coût de l'échantillon total pour une variance globale en particulier. Afin d'atteindre ce but, l'échantillonnage est augmenté dans les strates qui ont de grandes variances ou tailles de population et il est diminué dans les strates dont les interviews coûtent cher. Règle générale pour la répartition optimale, un grand échantillon est sélectionné dans une strate donnée si :

- la strate est plus nombreuse,
- la strate témoigne d'une plus grande variabilité interne,
- le déroulement de l'enquête coûte moins cher dans la strate.

Il faut obtenir de l'information précise sur les variances par strate et les coûts à l'unité pour appliquer la répartition optimale. En pratique, les variances et les coûts peuvent être inconnus. Un moyen de surpasser cette limite est d'estimer les variances et les coûts d'un échantillon préliminaire ou d'une enquête précédente. Une difficulté de la répartition optimale est que les variances et les coûts estimés de la strate peuvent être imprécis. En pratique donc, le plan d'échantillonnage n'est peut-être pas optimal.

Lorsque les variances et les coûts sont égaux pour toutes les strates, la répartition optimale se réduit à la répartition proportionnelle à N. La variance de l'estimation est minimisée pour cette répartition. Si seulement les coûts sont équivalents pour toutes les strates, la répartition optimale est ramenée à ce qui est généralement intitulé *répartition de Neyman* expliquée ci-dessous.

8.2.2.2.5 Répartition de Neyman

Cette répartition optimale particulière intitulée *répartition de Neyman* est appliquée lorsque le coût d'une interview est identique à chaque strate. C'est une répartition de la taille de l'échantillon total en strates qui minimise la variance de l'estimation dans l'ensemble. La répartition de Neyman attribue davantage d'unités de l'échantillon aux strates plus larges, aux strates qui affichent les variances les plus élevées, ou aux deux. De nouveau, comme dans le cas de la répartition optimale, les variances peuvent être inconnues et des estimations sont habituellement utilisées.

Voici l'expression du paramètre de répartition a_h :

$$a_h = \frac{N_h S_h}{\sum_{h=1}^{L} N_h S_h}.$$

C'est-à-dire que le paramètre de répartition a_h est égal au ratio du résultat de la taille de la strate et de la racine carrée de la variance au résultat de la taille de la strate et de la racine carrée de la variance de toutes les strates.

S'il est impossible d'obtenir une valeur précise pour les variances, le ratio de la variance de la strate à la moyenne de la strate ($S_h / \overline{Y}_h$) peut être considéré constant entre les strates. Cette supposition ramène la répartition de Neyman à une répartition proportionnelle à Y. La répartition de Neyman pose une difficulté, comme la répartition optimale, c'est-à-dire que l'estimation des variances de la strate n'est peut-être pas précise, et ainsi, le plan d'échantillonnage n'est peut-être pas optimal.

8.2.2.2.6 Répartition optimale lorsque les variances sont égales

La répartition optimale, occurrence particulière, est faite si les variances sont égales dans toutes les strates, ce qui est inhabituel, et cette répartition est donc rarement appliquée. Elle l'est cependant s'il n'y a pas d'information sur les variances de la population ou lorsqu'on peut supposer que ces variances sont approximativement égales et que le facteur de répartition prédominant est le coût, auquel cas, ce genre de répartition attribue davantage d'unités de l'échantillon aux strates plus larges, à celles qui coûtent moins cher, ou les deux. Le paramètre de répartition a_h est défini comme suit :

$$a_h = \frac{N_h / \sqrt{c_h}}{\sum_{h=1}^{L} N_h / \sqrt{c_h}}.$$

8.2.3 Considérations particulières pendant la répartition

Il ne faut pas oublier les préoccupations suivantes pendant la répartition :

i. Utilisation des données auxiliaires pour la répartition proportionnelle à la $\sqrt{Y}$ et à Y

Lors de la mise en œuvre d'un plan d'échantillonnage stratifié et de la répartition proportionnelle à Y ou à la racine carrée de Y, en pratique, la valeur de Y est inconnue pour toutes les unités de la population et, si elle était connue, il ne serait pas nécessaire de procéder à un sondage pour cette variable. Lorsque ces méthodes de répartition sont appliquées, des données auxiliaires fortement corrélées avec Y sont donc utilisées et elles sont habituellement tirées d'enquêtes précédentes ou de données administratives. Il n'est pas évident que le coefficient de variation précisé pour la variable de l'enquête sera obtenu parce que l'organisme statistique applique une mesure auxiliaire de la taille. La puissance de la corrélation entre la variable de l'enquête et la variable auxiliaire utilisée déterminent donc l'efficience taille-stratification et la précision de la répartition.

ii. Répartition excessive

Dans un plan d'échantillonnage stratifié qui applique la répartition optimale, la répartition de Neyman, la répartition proportionnelle à Y ou la répartition proportionnelle à la $\sqrt{Y}$, il est possible que la valeur n_h attribuée dépasse la taille de la population N_h. Il s'agit d'une répartition excessive. Il faudrait alors procéder à un recensement des strates qui demandent des échantillons excessifs. La taille de l'échantillon globale obtenue à la suite de ce genre de répartition excessive sera ensuite plus petite que la taille de l'échantillon original et il serait possible de ne pas obtenir la précision demandée dans l'ensemble. La solution est d'augmenter l'échantillon dans les autres strates où n_h est plus petit que N_h à l'aide du surplus dans les tailles d'échantillon tiré des strates recensées.

iii. Taille minimale de l'échantillon de la strate

Il est habituellement recommandé d'attribuer au moins deux unités à chaque strate. Ces deux unités sont en fait le nombre minimal possible pour obtenir une estimation non biaisée de la variance des estimations. Remarquez que la taille minimale de l'échantillon de la strate devrait être supérieure à deux, compte tenu de la non-réponse totale.

Malheureusement, toutes les méthodes de répartition de l'échantillon examinées à la Section 8.2.2 peuvent donner des tailles d'échantillon inférieures à deux, ou même à un. La solution la plus habituelle dans ce cas est d'augmenter la taille de l'échantillon à deux dans les strates qui posent ce problème. Cette mesure augmentera la taille totale de l'échantillon. Une autre solution serait de répartir un échantillon de taille deux à toutes les strates et d'attribuer ensuite la taille de l'échantillon qui reste à toutes les strates à l'aide de l'une des méthodes de répartition présentées plus tôt. Cette solution a l'avantage de ne pas augmenter la taille totale de l'échantillon.

iv. Répartition selon plusieurs variables

La répartition qui convient à une variable pourrait ne pas convenir à une autre variable de l'enquête. Afin de répartir l'échantillon selon plus d'une variable, il faut appliquer une répartition *intermédiaire*. Des méthodes de répartition multidimensionnelle (certaines d'entre elles appliquent la programmation linéaire) ont été élaborées pour résoudre ce genre de problème (Bethel (1989)).

Il ne faut surtout pas oublier que l'organisme statistique veut répartir l'échantillon afin de répondre aux besoins de précision pour les principales variables d'intérêt de l'enquête. Cela signifie habituellement que les estimations pour les variables moins importantes de l'enquête ne seront pas aussi précises que celles des principales variables.

8.3 Sommaire

Déterminer la taille de l'échantillon est un processus de compromis et de choix pratiques entre des besoins de précision souvent concurrents et des contraintes opérationnelles, par exemple le budget dans l'ensemble, le coût de l'enquête pour chaque strate, le temps disponible et le nombre d'intervieweurs nécessaires et disponibles. Les décisions à prendre sur la taille de l'échantillon peuvent demander un nouvel examen et une modification éventuelle des objectifs, des besoins de données, des degrés de précision, des éléments du plan d'enquête, des activités sur place, etc., déterminés au point de départ. L'organisme statistique et le client ciblent souvent la rentabilité pour que le client puisse obtenir la taille de l'échantillon nécessaire. Ils prévoient, notamment, des interviews plus brèves, appliquent une autre méthode de collecte des données, oublient certains domaines d'intérêt ou considèrent un autre plan d'échantillonnage.

Si l'échantillonnage stratifié est utilisé, l'échantillon doit être réparti entre les strates. Il y a deux façons d'y arriver : déterminer la taille de l'échantillon total et la répartir entre les strates pour minimiser la variabilité ou, compte tenu d'une précision demandée, déterminer la taille de l'échantillon nécessaire dans chaque strate. Il faut une formule de répartition, a_h, dans chaque strate pour ces deux méthodes. Il y a diverses méthodes différentes de répartition. La répartition proportionnelle à N est la méthode de répartition proportionnelle qui donne des fractions d'échantillonnage égales dans chaque strate. Les méthodes de répartition non proportionnelles distribuent l'échantillon entre les strates, compte tenu de la taille de la population dans la strate ou d'une autre mesure de la taille de la strate, de la variabilité de la population de la strate ou du coût de l'enquête dans la strate.

Bibliographie

Bankier, M. 1988. Power Allocations: Determining Sample Sizes for Subnational Areas. *The American Statistician,* 42: 174-177.

Bethel, J. 1989. Répartition de l'échantillon dans les enquêtes à plusieurs variables. *Techniques d'enquête,* 15(1):49-60.

Cochran, W.G. 1977. *Sampling Techniques*. John Wiley and Sons, New York.

Fink, A. 1995. *The Survey Kit*. Sage Publications, California.
Fowler, F.J. 1984. *Survey Research Methods*. 1. Sage Publications, California.

Hidiroglou, M. 1986. The Construction of a Self-Representing Stratum of Large Units in Survey Design. *The American Statistician,* 40: 27-31.

Hidiroglou, M. 1993. *Quelques méthodes pour calculer les tailles d'échantillon ainsi que leur allocation pour les enquêtes-entreprises*. Statistique Canada.

Glasser, G.J. 1962. On the Complete Coverage of Large Units in a Statistical Study. *Review of the International Statistical Institute,* 30: 28-32.

Gower, A. et K. Kelly. 1993. *How Big Should the Sample Be?* Statistics Canada.

Kish, L. 1965. *Survey Sampling*. John Wiley and Sons, New York.

Latouche, M. 1988. *Détermination, allocation et sélection de l'échantillon.* Statistique Canada. 88-021F.

Lavallée, P. et M.A. Hidiroglou. 1988. Sur la stratification des populations asymétriques. *Techniques d'enquête,* 14(1): 35-45.

Lehtonen, R. et E.J. Pahkinen. 1995. *Practical Methods for the Design and Analysis of Complex Surveys, Statistics in Practice*. John Wiley and Sons, New York.

Levy, P. et S. Lemeshow. 1999. *Sampling of Populations*. John Wiley and Sons, New York.

Lohr, S. 1999. *Sampling: Design and Analysis*. Duxbury Press, U.S.A.

Moser C.A. et G. Kalton. 1971. *Survey Methods in Social Investigation*. Heinemann Educational Books Limited, London.

Särndal, C.E., B. Swensson et J. Wretman. 1992. *Model Assisted Survey Sampling*. Springer-Verlag, New York.

Satin, A. et W. Shastry. 1993. *L'échantillonnage : un guide non mathématique – Deuxième édition.* Statistique Canada. 12-602F.

Sethi, Y.K. 1963. A Note on Optimum Stratification of Populations for Estimating the Population Means. *Australian Journal of Statistics,* 5: 20-33.

Thompson, M. 1997. *Theory of Sample Surveys*. Chapman and Hill, United Kingdom.

Thompson, S.K. 1992. *Sampling*. John Wiley and Sons, New York.

PUBLICATIONS ÉLECTRONIQUES
DISPONIBLES À

www.statcan.gc.ca

Chapitre 9 - Opérations de collecte des données

9.0 Introduction

La collecte des données est habituellement le volet d'une enquête qui coûte le plus cher. C'est pour cette raison, et parce qu'il coûte très cher de résoudre les problèmes qui surviennent durant la collecte – et qui peuvent faire échouer tout le projet – qu'il faut bien réfléchir à cette étape de l'enquête et la planifier attentivement. Les diverses méthodes de collecte des données sont considérées au **Chapitre 4 - Méthodes de collecte des données**. Ce chapitre expose les diverses activités qui se déroulent pendant la collecte des données et explique comment elles devraient être organisées et accomplies. Les enquêtes assistées par intervieweur sont ciblées parce qu'elles ont les exigences opérationnelles les plus complètes.

La collecte des données devrait être organisée le plus efficacement possible, tout en maintenant des pratiques d'interview uniformes pour tous les intervieweurs. Une méthode d'organisation, appliquée à Statistique Canada et présentée dans ce chapitre, fait appel aux bureaux régionaux qui font rapport au Bureau central.

La collaboration des répondants s'obtient souvent au prix d'importantes relations publiques. L'organisme statistique doit notamment maintenir une bonne réputation. Il faut aussi organiser des campagnes publicitaires et utiliser divers outils, par exemple, des lettres de présentation, des brochures sur l'enquête et du matériel d'enquête préparé pour radiotélédiffusion et publication dans les journaux, afin de susciter un intérêt pour l'enquête et d'encourager la participation des répondants.

Retenir les services de bons intervieweurs et d'autres membres du personnel est essentiel au succès de la collecte des données. La persévérance et la qualité de ces travailleurs déterminent la qualité de la collecte des données et des résultats de l'enquête. La formation et les manuels sont donc aussi importants.

On procède parfois au listage et au dépistage avant de faire les interviews ou de distribuer les questionnaires par autodénombrement. Le listage est nécessaire quand on a recours à des bases aréolaires. L'objectif du listage est d'établir une liste d'unités à échantillonner (p. ex., logements ou entreprises) dans un secteur géographique en particulier. Le dépistage est fait si une unité de l'échantillon ne peut être repérée à l'aide de l'information de la base de sondage. Les numéros de téléphone ne sont peut-être plus à jour, par exemple, dans la base de sondage.

L'interview ou l'autodénombrement peut commencer lorsque le répondant a été repéré et qu'on a pu établir le contact avec lui. Il ne s'agit pas simplement de poser des questions, il faut aussi établir le calendrier des interviews, obtenir la collaboration des répondants, minimiser les erreurs de réponse, faire le suivi des rejets à la vérification, coder les réponses, contrôler les documents et surveiller la qualité de la collecte des données. Celle-ci est considérée complète seulement à la conclusion de ces activités.

9.1 Organisation de la collecte des données

Il y a de nombreuses méthodes d'organisation des activités de collecte des données, mais l'une des plus habituelles est la répartition du pays en régions, chacune ayant un bureau régional qui fait rapport au Bureau central. Les bureaux régionaux peuvent tirer pleinement avantage des connaissances locales à l'aide de cette structure pour améliorer les relations avec les répondants et augmenter l'efficacité en diminuant le coût du suivi des questionnaires incomplets, des refus, des non-réponses, etc. Les bureaux régionaux sont chargés de la gestion des activités de collecte et de saisie des données dans leurs secteurs (la saisie des données consiste à transformer les réponses pour les rendre lisibles à la machine; à ce

propos, on peut consulter le **Chapitre 4 - Méthodes de collecte des données** et le **Chapitre 10 - Traitement)**. Le Bureau central veille à ce que des procédures et concepts normalisés et uniformes soient appliqués dans toutes les régions. Il est aussi chargé de la gestion de l'enquête dans l'ensemble et de la conception des procédures de collecte des données.

Il faut considérer un certain nombre de points pour structurer les bureaux régionaux. S'il s'agit d'une grande enquête ou d'un recensement, il peut être nécessaire d'ouvrir des bureaux de district qui feront rapport à un bureau régional. Les points suivants influenceront le nombre de régions ou de bureaux :

- la taille de l'enquête,
- la taille de l'échantillon et le lieu où sont situées les unités de l'échantillon,
- l'éloignement de l'endroit,
- la difficulté de communiquer avec les répondants,
- la collaboration des répondants qui peut être difficile à obtenir,
- la langue des répondants,
- la structure des administrations locales, régionales ou provinciales,
- la population active (p. ex., disponibilité du personnel, scolarité, profils linguistiques),
- les moyens de transport (p. ex., autoroutes, ports, centres d'expédition).

9.1.1 Bureau central

Le Bureau central est généralement chargé de la conception et de la planification des activités et outils suivants :

i. Procédures de collecte des données

Le Bureau central conçoit et élabore habituellement des manuels pour les intervieweurs, les surveillants et les autres membres du personnel de la collecte des données. Ces procédures comprennent les interviews, le listage des unités échantillonnées et la mise à jour de la base de sondage, le dépistage des répondants, le suivi auprès des non-répondants, ainsi que la vérification et le codage sur place. Certaines de ces tâches sont détaillées au **Chapitre 10 - Traitement**.

ii. Traitement des données

Le traitement transforme les réponses du sondage obtenues pendant la collecte pour qu'elles conviennent à la totalisation et à l'analyse des données. Les activités de traitement comprennent le codage et la saisie des données, la vérification et l'imputation. Le Bureau central est chargé de l'élaboration de procédures et programmes de traitement, ainsi que des procédures de contrôle qualitatif et d'assurance de la qualité qui seront appliquées dans les bureaux régionaux. Les activités de traitement sont considérées au **Chapitre 10 - Traitement**. Le contrôle qualitatif et l'assurance de la qualité sont approfondis à l'**Annexe B - Contrôle qualitatif et assurance de la qualité**.

iii. Procédures de formation

Elles comprennent l'élaboration d'exercices, de scénarios d'interview simulée, de matériel audiovisuel et de guides de formation.

iv. Échéancier de la collecte des données

Un calendrier de collecte des données est établi, les étapes de l'enquête et les périodes de rapport sont précisées, afin d'atteindre la date visée. L'échéancier comprend les taux ciblés de cas résolus pour chaque période de rapport, ainsi que les taux de réponse voulus (voir les détails à la Section 9.5.3).

v. Systèmes de contrôle et de rapport

Des formules de contrôle sont élaborées pour l'échantillon au complet et pour chaque intervieweur (la formule est intitulée tâche de l'intervieweur), et des procédures de rapport régulier sont appliquées pour mettre à jour l'état de chaque unité échantillonnée, afin de garantir que toutes les activités de collecte des données se déroulent comme prévu. Cet outil est habituellement intitulé « Système d'information de gestion » (SIG). Le repérage de l'état d'une unité et la préparation de rapports de surveillance appropriés sont programmés dans un logiciel pour l'interview assistée par ordinateur.

Le SIG devrait avoir la capacité de suivre les mesures de la qualité, les dépenses et d'autres mesures du rendement pendant la collecte des données. Il faudrait suivre tous les coûts de la collecte des données, par exemple, l'affranchissement postal, les appels téléphoniques, les déplacements, l'informatique et la consommation par personne par jour. Il faudrait évaluer et surveiller d'importantes mesures de la qualité pendant le processus de la collecte, y compris les taux de réponse, les taux de suivi et le calcul des non-réponses totales pour chaque raison. Des mesures de la qualité et de la productivité peuvent servir simultanément à déterminer les pointes logiques de la collecte des données (p. ex., lorsque le taux de réponse a atteint une cible prévue ou lorsque l'amélioration du taux de réponse pour obtenir un taux supérieur déterminé coûterait trop cher) et les modifications à apporter s'il est impossible de respecter la date limite de la collecte. Ces mesures de la qualité servent aussi à évaluer les méthodes qui pourraient être appliquées à d'autres enquêtes et à obtenir de l'information pour l'évaluation de la qualité des données.

vi. Campagne de relations publiques

LA campagne de relations publiques comprend l'élaboration et la production de brochures, d'affiches, etc. Certains éléments de la campagne de relations publiques sont pris en charge directement au Bureau central pour les grandes enquêtes. Étant donné cependant que l'impression du public peut varier considérablement d'une région à l'autre dans un grand pays, les bureaux régionaux sont chargés de la majeure partie de ce travail.

9.1.2 Bureau régional

La collecte et la saisie des données, ainsi que les relations publiques sont les principales responsabilités des bureaux régionaux. Ceux-ci font souvent la saisie des données parce qu'il est plus facile de contrôler, gérer et télécharger au Bureau central des fichiers électroniques que d'envoyer des questionnaires sur support papier. Il est aussi plus facile pour un bureau régional de faire la saisie de ses lots restreints de questionnaires, comparativement au Bureau central qui devrait faire la saisie des données de toutes les régions.

Le recours à l'infrastructure des bureaux régionaux permet de faire la gestion quotidienne des activités de collecte des données le plus près possible de la scène des opérations, et il est plus facile d'identifier et de résoudre les problèmes au fur et à mesure.

La hiérarchie suivante des employés de la collecte des données est établie pour la plupart des enquêtes qui se déroulent à partir des bureaux régionaux :

i. Chef de projet régional

Le chef de projet régional veille dans l'ensemble à ce que la collecte des données soit achevée à temps et qu'elle réponde aux normes de qualité déterminées. Il est aussi chargé du budget régional. Il adopte habituellement l'échéancier de l'enquête globale et y ajoute les étapes et les points de repère régionaux détaillés qui sont nécessaires pour maintenir l'enquête dans la bonne voie. Le chef de projet régional est la personne-ressource du Bureau central et l'intervenant chargé de toutes les expéditions du Bureau central et vers celui-ci. La distribution du matériel et la prestation de l'information aux surveillants lui sont aussi confiées. S'il s'agit de très grandes enquêtes, par exemple le Recensement de la population canadienne, des chefs régionaux adjoints et des chefs de district sont ajoutés à l'équipe.

ii. Surveillant

S'il y a plus d'un surveillant à cause de la taille de l'enquête, chacun supervise une équipe d'intervieweurs. Le ratio de surveillants à intervieweurs varie selon les points suivants :

- la complexité de l'enquête,
- l'expérience des surveillants et du personnel chargé des interviews,
- l'endroit où est situé l'échantillon,
- les difficultés de déplacement,
- la facilité d'utilisation du Système d'information de gestion.

Le surveillant retient les services des intervieweurs, les forme, obtient et distribue le matériel et les articles, en collaboration avec le chef de projet régional. Le surveillant est chargé de la gestion quotidienne de la collecte des données, y compris la supervision des intervieweurs et la surveillance étroite de l'état d'avancement et de la qualité. Le suivi des refus (voir la Section 9.4.6) est une autre tâche importante du surveillant.

iii. Intervieweur

L'intervieweur procède à la collecte des données en soi et fait rapport régulièrement au surveillant sur les problèmes et l'état d'avancement. Le rôle de l'intervieweur est étudié à la Section 9.4.

9.2 Relations publiques

L'objectif de la campagne de relations publiques est de sensibiliser la population à l'enquête, afin d'éliminer la méfiance des gens envers les étrangers qui frappent à leurs portes dans le quartier, de susciter leur intérêt, d'accroître ainsi le taux de réponse et de rehausser la précision des réponses.

La meilleure stratégie de relations publiques dans un organisme statistique est l'acquisition et le maintien d'une réputation professionnelle indéniable. Il faut garantir à cette fin que les données obtenues sont fiables, tout à fait disponibles, utilisées et appréciées, et que le respect de la confidentialité des réponses des répondants est incontestable.

L'organisme doit, non seulement avoir bonne réputation si elle veut obtenir un bon taux de réponse, mais il doit aussi donner aux intervieweurs les outils nécessaires pour répondre aux questions et aux plaintes. D'autres outils sont aussi utiles, notamment, une lettre de présentation, une brochure de l'enquête et une

brochure sur l'organisme en général. Des campagnes de publicité sont aussi fréquentes pour les grandes enquêtes.

Cette section sur les relations publiques couvre seulement la communication avec le grand public. Il faudrait aussi prévoir d'autres communications avec les gens à l'extérieur de l'équipe pendant le processus de l'enquête. Au début de la phase de la planification, par exemple, les planificateurs de l'enquête devraient déterminer des questions que l'enquête proposée pourrait éventuellement susciter, prévoir comment y répondre et décider de procéder ou non à l'enquête. Un sujet proposé de l'enquête peut, par exemple, avoir un caractère trop délicat pour une partie des répondants. Il est aussi important, pendant le processus de planification, de consulter des intervenants, des répondants, des représentants d'administrations locales et d'autres intéressés pour garantir l'élaboration d'une matière appropriée dans le questionnaire. Il faudrait consulter ces intéressés ou les informer des résultats à la fin du processus de l'enquête.

La confidentialité et la planification d'une enquête sont détaillées au **Chapitre 12 - Diffusion des données** et au **Chapitre 13 - Planification et gestion de l'enquête**, respectivement.

9.2.1 Campagnes publicitaires

Les campagnes publicitaires peuvent comprendre les points suivants, en tout ou en partie, selon le sujet de l'enquête, le budget et la population cible :

- des relations actives avec les médias (messages d'intérêt public, faits et porte-parole pour les journaux et les stations de radio et de télévision) et des relations avec les médias pendant la collecte des données pour régler les problèmes qui se posent pendant le processus,

- la communication avec des groupes qui feront la promotion de l'enquête pendant la collecte des données et avec ceux qui se prononceront contre,

- des lettres aux importants représentants du public (ou aux représentants d'associations ou d'établissements qui ont un lien avec la population cible de l'enquête) pour demander leur soutien et leur fournir des encarts pour un discours, un bulletin, etc. (ces représentants prendront souvent la parole pour soutenir activement l'enquête),

- des lettres aux représentants des forces de l'ordre pour les informer de l'enquête au cas où des répondants communiqueraient avec eux s'ils se méfient du personnel de l'enquête,

- des affiches dans les endroits publics, notamment les bureaux de poste et les bibliothèques, ou dans des lieux où la population cible les remarquera probablement.

9.2.2 Relations avec les répondants

Les campagnes de relations publiques ciblent habituellement les répondants pour les sensibiliser davantage et obtenir leur collaboration. Voici les articles utilisés le plus souvent :

i. Lettre de présentation

Cette lettre précise l'objectif, les dates et la méthode de collecte, et explique l'importance de l'enquête. Le premier représentant régional supérieur (directeur régional) signe généralement les lettres de présentation.

Celles-ci (et les brochures si elles sont appropriées) sont envoyées avec les questionnaires dans les enquêtes par la poste. On envoie habituellement les lettres (et les brochures le cas échéant) une semaine à l'avance si des intervieweurs font l'enquête. Les répondants ne seront donc pas surpris et collaboreront probablement davantage lorsque l'intervieweur se présentera à leur domicile.

ii. Brochure de l'enquête

Il s'agit simplement de renseignements qui décrivent l'importance de l'enquête pour le bien public. Elle devrait comprendre des exemples d'utilisation des données et une source précisant où obtenir les données. Les brochures peuvent être envoyées à tous les répondants ou utilisées seulement si les répondants sont réticents.

iii. Brochure générale de l'organisme

Document général à distribuer qui illustre la variété des données que l'organisme obtient et diffuse, la brochure accentue la crédibilité de l'organisme et sert souvent aussi aux relations avec les répondants.

iv. Soutien pendant la collecte pour régler les problèmes imprévus avec des répondants

v. Manuel de l'intervieweur

Un manuel de l'intervieweur bien élaboré devrait donner les réponses aux questions et objections prévues.

vi. Spécialistes formés pour convaincre ceux qui refusent de répondre

Les intervieweurs confient habituellement les refus à leurs surveillants qui sont formés pour appliquer des méthodes, afin de convertir les refus en réponses.

Voici un exemple de lettre de présentation que Statistique Canada a utilisé pour le British Columbia Farm Resource Management Survey (Enquête sur la gestion des ressources agricoles en Colombie-Britannique) en 1998. La lettre précise en vertu de quelle loi ou quelle autorité l'enquête se déroule, donne une garantie de confidentialité et insiste sur l'importance de la participation du répondant. Le nom et le numéro de téléphone d'une personne-ressource sont ajoutés à la fin de la lettre au cas où le répondant aurait des questions, et elle porte la signature appropriée, dans ce cas, celle du directeur régional.

Monsieur, Madame,

Votre exploitation agricole a été sélectionnée au hasard pour participer à l'Enquête sur la gestion des ressources agricoles en Colombie-Britannique, une importante étude sur les pratiques de gestion agricole visant trois principales ressources : le sol, l'eau et le fumier – l'engrais. Cette enquête à participation volontaire cible en particulier l'élevage du bétail, et c'est la première d'une série d'enquêtes détaillées qui produiront en définitive des données uniformes pour tous les groupes de produits dans la province. Cette étude se déroule en collaboration avec le ministère de l'Agriculture et de l'Alimentation de la Colombie-Britannique pour veiller à ce que les programmes agricoles reflètent les méthodes changeantes de la gestion des ressources à la ferme aujourd'hui.

Entre le 5 et le 24 octobre, un intervieweur de Statistique Canada vous téléphonera pour procéder à une interview de cinq à dix minutes au téléphone. Aucune question financière détaillée ne sera posée et vous n'aurez pas besoin de consulter vos dossiers. Nous demanderons cependant le nombre de têtes et le genre de bétail dans votre exploitation pour obtenir une perspective sur les pratiques de gestion appliquées.

Toute information obtenue à Statistique Canada est strictement confidentielle et protégée par la loi. Elle sera utilisée uniquement pour dresser des tableaux statistiques qui ne permettent pas d'identifier un répondant en particulier ou ses renseignements.

Statistique Canada reconnaît l'effort énorme que font les répondants du secteur agricole pour répondre aux questionnaires des enquêtes. Cette collaboration signifie que des données pertinentes et à jour sont disponibles sur ce secteur en changement rapide. J'apprécie sincèrement votre collaboration aux enquêtes précédentes et je vous remercie d'avance de votre participation à cette importante étude.

Si vous voulez davantage d'information sur cette enquête, veuillez téléphoner à M^me^ Unetelle, gestionnaire des enquêtes sur l'agriculture, Bureau de la région du Pacifique (Vancouver), en composant le numéro sans frais 1 800 555-5555.

Le directeur,
Région du Pacifique

Jean Ixe

9.3 Préparation des procédures de collecte des données

Il y a de nombreuses tâches à accomplir avant la collecte des données, par exemple :
- rédiger des manuels,
- embaucher et former du personnel,
- concevoir des procédures de listage,
- concevoir des procédures de dépistage.

Ces tâches sont considérées dans cette section.

9.3.1 Manuels

Des intervieweurs et d'autres membres du personnel de qualité sont la clé du succès de la collecte des données. De bons intervieweurs et membres du personnel de l'enquête doivent avoir les capacités et les qualités personnelles nécessaires pour être efficaces. L'uniformité et la qualité de leur travail déterminent la qualité des résultats de l'enquête. Des manuels décrivent les procédures normalisées et donnent des instructions pour régler des problèmes imprévus. L'équipe de l'enquête prépare généralement un manuel de l'intervieweur, un manuel du surveillant et, si nécessaire, des instructions de listage pour les bases aréolaires.

9.3.1.1 Manuel des intervieweurs

Le manuel des intervieweurs est la principale et parfois la seule source d'information que l'intervieweur peut consulter pour obtenir des renseignements sur son travail. Il est réparti en sections ou chapitres sur les sujets suivants :

i. Information générale

Cette section énonce l'objectif et l'importance de l'enquête, les utilisations prévues des données et les règles de collecte des données de l'organisme (confidentialité, langue de l'intervieweur, mandat de l'organisme, etc.). Une copie de la lettre de présentation envoyée aux répondants y est habituellement ajoutée, ainsi que de l'information élémentaire sur la méthode de sélection de l'échantillon.

ii. Présentation

Cette section explique comment établir la première communication avec un répondant, vérifier s'il s'agit du répondant voulu, examiner ou corriger l'information de la base de sondage (numéro de téléphone, etc.) et les lignes directrices d'interview des substituts (Section 9.4.7).

iii. Matière du questionnaire

Cette section comprend une copie du ou des questionnaire(s), la définition des concepts de l'enquête et la terminologie. Il est important que l'intervieweur comprenne la signification et l'objectif de chaque question. Cette section porte aussi sur les questions des répondants, les problèmes éventuels et les interventions appropriées.

iv. Vérification sur place – prétraitement des questionnaires

Les vérifications sont des règles appliquées pour identifier les entrées manquantes, invalides ou incohérentes qui indiquent des données éventuellement erronées. Les intervieweurs doivent faire des vérifications sur place (c.-à-d. vérifications faites pendant l'interview ou peu après). Les règles de vérification doivent être clairement décrites et préciser comment les appliquera l'intervieweur.

v. Gestion des unités d'échantillonnage

Cette section porte sur le rejet à la vérification, le suivi des non-réponses et le nombre de tentatives que doit faire l'intervieweur pour essayer d'obtenir une réponse. Elle précise aussi comment attribuer un code d'état définitif à chaque questionnaire (p. ex., questionnaire rempli, refus, etc.). Vous obtiendrez davantage de détails à la Section 9.5.2.

vi. Gestion des tâches

Cette section couvre certains détails administratifs, par exemple, comment les intervieweurs font rapport sur l'état d'avancement de leurs questionnaires, comment ils retournent les questionnaires au bureau régional, comment ils présentent les documents des dépenses sur place (p. ex., dépenses de déplacement, d'hébergement, etc.), comment ils sont rémunérés et comment le matériel et les articles sont distribués et retournés.

vii. Sûreté et sécurité sur place

Cette section porte sur la santé et la sécurité au travail, ainsi que sur les systèmes de contrôle efficaces pour garantir la sécurité des questionnaires et du transfert des données des bureaux régionaux au Bureau central.

viii. Questions et réponses

Cette dernière section comprend une liste des questions que posent habituellement les répondants (par exemple : Comment ai-je été choisi pour l'enquête?) et les réponses appropriées.

Les aptitudes à l'interview et les techniques d'interview en général peuvent aussi être intégrées au manuel des intervieweurs avec exemples pertinents à l'enquête en particulier.

9.3.1.2 Manuel des surveillants

Les surveillants doivent très bien connaître la matière du manuel des intervieweurs. Un manuel spécial des surveillants est aussi prévu pour donner des instructions sur la gestion de l'enquête.

Les sujets suivants sont habituellement ajoutés au manuel des surveillants :
- embauche et formation des intervieweurs,
- conception des tâches des intervieweurs,
- santé et sécurité au travail,
- contrôle de la qualité et du rendement (c.-à-d. observation des interviews, surveillance de l'état d'avancement de l'enquête comparativement à des mesures déterminées de la qualité, des dépenses et des délais d'exécution),
- logistique (p. ex., distribution et retour des articles, rémunération des intervieweurs, retour et présentation des questionnaires pour la saisie des données, etc.),
- sécurité et protection des renseignements personnels,
- autres méthodes de collecte des données pour tenir compte des personnes ayant une incapacité, des problèmes de langue, des cas dont l'inclusion à la population cible est ambiguë (p. ex., étrangers, visiteurs), etc.,
- intervention pour convaincre ceux qui refusent de répondre au questionnaire.

9.3.2 Embauche et formation des intervieweurs

Les intervieweurs sont essentiels au succès d'une enquête assistée par intervieweur. Il est important de vérifier si ceux qui sont engagés ont les qualités personnelles et les capacités nécessaires, et s'ils ont la formation et les outils appropriés.

L'organisme statistique devrait tenir à jour une liste d'intervieweurs d'expérience qui servira au moment de l'embauche. Si les besoins de l'enquête sont nombreux ou très importants, il peut être nécessaire d'obtenir du personnel supplémentaire. Des avis peuvent être affichés ou des annonces peuvent être diffusées dans les journaux locaux ou à la radio pour inviter les candidats éventuels, ou le personnel approprié peut être recruté (par exemple, le personnel de la livraison du courrier).

Il faut préciser les qualifications nécessaires pour faire l'enquête et établir les critères d'embauche. La scolarité, les aptitudes interpersonnelles, la capacité de s'exprimer dans les langues locales, les aptitudes à l'organisation et l'intégrité sont des éléments importants à considérer lors de l'embauche des intervieweurs (il y a habituellement une vérification de sécurité). S'il s'agit d'interviews sur place, l'endroit et la connaissance du secteur peuvent aussi être importants. Une équipe, comprenant habituellement le surveillant et le chef régional principal, interviewe les candidats éventuels.

La formation des intervieweurs doit être soigneusement planifiée pour qu'ils aient tous un rendement uniforme et la même compréhension des concepts de l'enquête. Les surveillants sont habituellement

formés en premier. Ceux-ci forment ensuite les intervieweurs. Des représentants du Bureau central observent souvent la formation et donnent des conseils. Plusieurs jours de formation intensive sont généralement offerts à l'aide des techniques énumérées ci-dessous :

i. Études à domicile

Les intervieweurs examinent attentivement les manuels et (éventuellement) font les exercices écrits.

ii. Formation en classe

Les surveillants et les intervieweurs étudient en classe ce qu'ils peuvent faire pour établir de bonnes relations avec les répondants et obtenir ainsi des réponses. De bonnes techniques et pratiques d'interview, ainsi que des aptitudes à l'interview sont présentées. Les surveillants examinent et corrigent les erreurs dans les exercices faits à domicile. Les intervieweurs examinent ensuite la matière complète du questionnaire pour bien comprendre les concepts et les questions (écran par écran pour l'interview assistée par ordinateur). Les cas spéciaux et à problèmes sont revus en classe afin de laisser suffisamment de temps pour les questions et les précisions.

iii. Interviews simulées

Les interviews simulées donnent l'occasion aux intervieweurs de mettre en pratique leurs techniques avant d'intervenir sur place. Elles donnent aussi aux intervieweurs l'occasion d'observer les aptitudes et les techniques appliquées par leurs pairs et de faire des commentaires. Le surveillant ou un autre intervieweur intervient à cette étape à titre de répondant. Divers scénarios sont mis à l'essai, y compris les cas typiques et à problèmes.

iv. Interviews concrètes

Quand cela est possible, on fait aussi des interviews avec des répondants réels avant de procéder à l'enquête sur le terrain. Les répondants sont parfois des membres du personnel de l'organisme qui ne sont pas informés de l'enquête, ou autrement, ce sont des répondants échantillonnés dans la population cible (mais qui ne font pas partie de l'échantillon qui servira à l'enquête réelle). Les interviews concrètes devraient aussi être un volet d'un essai pilote (voir le **Chapitre 5 - Conception du questionnaire**).

v. Examen des premières interviews

Le surveillant aura avantage à rencontrer chaque intervieweur pour examiner les premières interviews achevées. Si l'intervieweur a des problèmes, ils peuvent être identifiés et corrigés rapidement.

Les procédures administratives (p. ex., rapports hebdomadaires, formules de contrôle, etc.) pour la gestion des tâches sont habituellement le dernier sujet couvert pendant la formation. À la conclusion de la formation, chaque intervieweur se voit confier sa tâche.

9.3.3 Listage

Le listage est nécessaire lorsqu'une base aréolaire sert à l'échantillonnage. On a vu au **Chapitre 6 - Plans d'échantillonnage** qu'un plan d'échantillonnage habituel pour une base aréolaire est un plan d'échantillonnage par grappes à deux degrés, les secteurs géographiques étant échantillonnés au premier degré dans une base aréolaire (ce sont les unités primaires d'échantillonnage ou UPÉ). On peut ensuite tirer de ces UPÉ un échantillon systématique de logements (unités secondaires d'échantillonnage ou

USÉ). Afin d'échantillonner les logements, il faut d'abord établir une liste de tous les logements dans le champ de l'enquête de l'UPÉ (c.-à-d. que les logements admissibles à l'échantillonnage doivent être listés, l'admissibilité étant définie selon la population cible de l'enquête).

Il est avantageux de bien connaître le secteur géographique (UPÉ) pour faire les interviews et le même groupe d'intervieweurs est donc souvent chargé du listage et des interviews dans l'UPÉ. Le listage est d'autant plus exact qu'il est fait peu de temps avant les interviews.

Avant le listage, chaque intervieweur (ou celui qui fait le listage) devrait obtenir les articles suivants et la formation nécessaire pour les utiliser :

i. Une carte de l'UPÉ aux limites clairement définies

Les limites de chaque UPÉ doivent être clairement définies pour éviter le chevauchement des UPÉ ou les segments manquants. La carte devrait être la plus détaillée et à jour qui soit disponible et comprendre des points de référence bien inscrits (voies ferrées, ponts, cours d'eau, noms de rue, etc.). Ces données viennent parfois de sources municipales ou d'arpentage.

ii. Instructions sur le listage

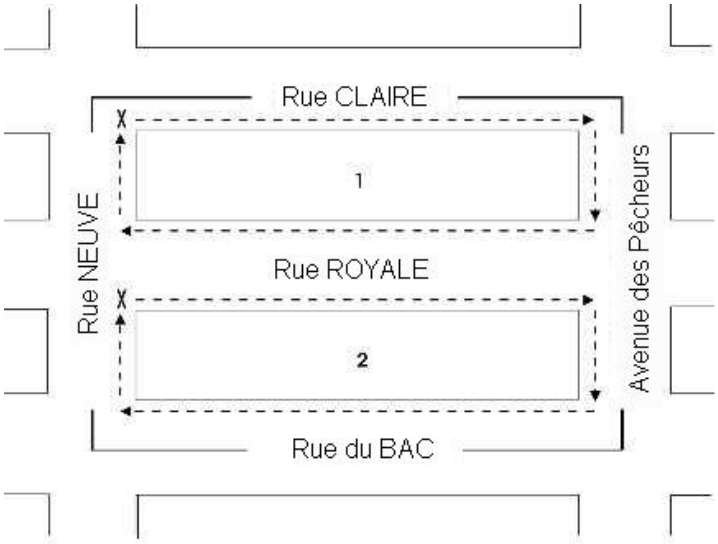
Elles comprennent des instructions sur la méthode à appliquer pour tracer l'itinéraire d'une UPÉ, afin de couvrir le secteur complet sans rebrousser chemin (pour éviter les risques de dédoublement) ou oublier des secteurs. Il y a aussi des instructions sur la méthode à appliquer pour identifier une unité d'échantillonnage dans le champ de l'enquête. Si l'unité de l'échantillon est un logement, par exemple, il devrait y avoir des instructions précisant comment trouver et identifier des logements confinés ou dans un immeuble à unités multiples, ainsi que la définition de logement inoccupé, etc.

iii. Une formule de listage et l'information à obtenir pour chaque unité de l'échantillon dans le champ de l'enquête

Le genre d'enquête détermine la définition d'une unité dans le champ de l'enquête et le nombre de renseignements à obtenir. Les données obtenues pendant le listage sont habituellement limitées à l'information nécessaire pour situer l'unité (adresse ou endroit sur la rue, nom, numéro de téléphone, etc.) et toute donnée nécessaire pour passer à l'étape suivante de l'échantillonnage.

Lorsque le listage est achevé, les données des formules de listage sont généralement saisies électroniquement et envoyées au Bureau central pour passer à l'étape suivante de l'échantillonnage.

Pour illustrer le listage, on trouvera ci-dessous une carte de grappe de l'Enquête sur la population active de Statistique Canada; on y trouve les limites de l'UPÉ tracées en ligne continue, un « X » inscrit au point de départ du listage et une ligne de tirets (---) trace l'itinéraire. On demande à l'intervieweur de commencer à l'intersection nord-ouest, de procéder dans le sens des aiguilles d'une montre autour de chaque îlot et de lister tous les logements habitables à sa droite. La même méthode générale de listage (à partir d'un point de départ déterminé en continuant dans le sens des aiguilles d'une montre pour lister les unités à droite) est appliquée en milieu rural. Le listage est plus compliqué si le logement est une exploitation agricole parce qu'elle peut empiéter sur les limites de l'UPÉ, et la solution est alors de lister l'exploitation agricole dans l'UPÉ qui englobe la voie ou l'entrée principale.



Si la méthode d'échantillonnage est très simple, l'intervieweur peut parfois faire le listage, l'échantillonnage et les interviews simultanément. Lors du Recensement de la population canadienne, par exemple, quatre ménages sur cinq dans un secteur de dénombrement (SD) reçoivent un bref questionnaire et le cinquième ménage obtient un questionnaire plus long et plus détaillé. À l'aide de l'échantillonnage systématique avec origine choisie au hasard dans chaque SD, l'enquêteur liste les ménages et remet le long questionnaire au cinquième ménage de chaque tranche de cinq ménages.

On a vu au **Chapitre 3 - Introduction au plan d'enquête** et au **Chapitre 6 - Plans d'échantillonnage** les détails sur les bases aréolaires.

9.3.4 Dépistage

Le dépistage est nécessaire quand l'information disponible dans la base de sondage est insuffisante pour situer le répondant. Dans les enquêtes téléphoniques, par exemple, certains numéros de téléphone dans la base de sondage ne sont peut-être plus à jour. Dans les sondages par la poste, le bureau de poste peut retourner certains questionnaires parce que l'adresse est incomplète ou inexacte, l'établissement n'existe plus ou le répondant a déménagé. Le dépistage peut être fait avant de procéder à l'enquête si l'on soupçonne que l'information dans la base de sondage n'est plus à jour.

Voici donc l'objectif du dépistage :
- situer l'unité échantillonnée,
- mettre à jour l'information d'identification élémentaire dans la base de sondage (p. ex., nom, adresse, numéro de téléphone, personne-ressource ou, dans une base aréolaire, indiquer le lieu géographique d'une exploitation agricole sur une carte, etc.),
- déterminer si l'unité est toujours dans le champ de l'enquête (p. ex., L'établissement a-t-il fermé ses portes? Le répondant a-t-il déménagé hors du champ géographique de l'enquête?).

Les outils de dépistage les plus souvent utilisés comprennent ceux-ci :
- annuaires téléphoniques à jour, répertoires d'entreprises, assistance-annuaire,
- information tirée d'autres bases de sondage plus à jour,
- dossiers d'autres organismes gouvernementaux (p. ex., listes de permis d'entreprise dans une municipalité en particulier, listes de permis de conduire dans une province, etc.),

- pour les enquêtes réitérées, repérage de l'information que l'unité échantillonnée a donnée à une occasion précédente (p. ex., adresse ou numéro de téléphone de parents qui peuvent aider à repérer le répondant).

Les intervieweurs peuvent faire le dépistage jusqu'à un certain point, mais il est souvent plus efficace pour le bureau régional d'avoir une équipe d'employés affectés au dépistage et qui ont accès à tous les répertoires et dossiers nécessaires. Après le dépistage, l'information de communication dans la base de sondage peut être mise à jour ou, si le dépistage est fait après le lancement de la collecte, les unités dépistées peuvent être retournées aux intervieweurs originaux, ou elles peuvent être confiées à un intervieweur « de rappel » en particulier.

La qualité de l'information auxiliaire dans la base de sondage, ainsi que le talent et l'esprit d'initiative du dépisteur, déterminent le succès du dépistage. Statistique Canada procède, par exemple, à une enquête sur les diplômés d'université deux ans après la collation des grades. La base de sondage comprend l'adresse et le numéro de téléphone les plus récents de chaque étudiant selon les dossiers des universités. Étant donné que les diplômés récents sont extrêmement mobiles, nombre d'entre eux ont déménagé depuis. Si les données auxiliaires comprennent aussi le nom et l'adresse des parents, l'intervieweur peut dépister l'étudiant en communiquant avec eux. Il est très important de veiller à ce que les intervieweurs ne donnent pas d'information confidentielle pendant le dépistage.

9.4 Déroulement des interviews

Après avoir planifié toutes les activités de collecte des données, préparé les manuels, embauché et formé le personnel, fait le listage et le dépistage préliminaire, les interviews peuvent commencer. Cette étape ne se limite pas à communiquer simplement avec les répondants et à poser des questions. L'intervieweur est chargé des activités suivantes :

- préparer les interviews et en établir le calendrier,
- veiller à ce que l'information soit obtenue de l'unité d'échantillonnage choisie,
- susciter la collaboration des répondants pour minimiser les non-réponses,
- poser les questions et inscrire les réponses précisément pour éviter les erreurs,
- vérifier les réponses,
- appliquer toutes les procédures de sécurité pour garantir la confidentialité des données.

L'intervieweur et d'autres membres du personnel de la collecte des données sont aussi chargés des tâches suivantes :

- faire le suivi des rejets à la vérification et des non-réponses,
- coder les données (si les questionnaires ne sont pas entièrement codés d'avance),
- exercer un contrôle sur les documents (formules de repérage pour le cheminement pendant le processus, par exemple, le nombre de questionnaires envoyés par la poste, retournés, en instance, etc.),
- surveiller la qualité de la collecte des données.

Les Sections 9.4.1 à 9.4.8 suivantes ciblent la préparation des interviews et l'établissement du calendrier, ainsi que les techniques d'interview à appliquer, y compris les techniques de présentation, d'utilisation du questionnaire, d'approfondissement pour obtenir des réponses, de conclusion de l'interview, de réaction aux refus ou à d'autres situations à caractère délicat et d'interview d'un substitut. La surveillance de la qualité de la collecte des données est étudiée à la Section 9.5.

La vérification et le codage des données sont approfondis au **Chapitre 10 - Traitement**.

9.4.1 Préparation des interviews et établissement de l'horaire

L'intervieweur est plus efficace s'il a planifié la journée de travail, s'il a établi l'horaire des appels ou des visites et s'il a une connaissance approfondie du questionnaire, des formules de contrôle et du matériel de l'enquête. Voici certaines lignes directrices utiles pour l'organisation de la tâche de l'intervieweur :

i. Organisation du temps

L'intervieweur qui organise le travail chaque jour sait exactement combien d'appels il prévoit faire. Il doit réserver suffisamment de temps entre les appels pour compléter les notes nécessaires prises pendant l'interview et ajouter les corrections au besoin pour les questionnaires sur support papier.

ii. Entrée des commentaires

L'intervieweur devrait entrer tous les commentaires à l'écran approprié de l'ordinateur ou les inscrire dans l'espace prévu au questionnaire. Il devrait ajouter certains renseignements, notamment, le meilleur moment pour téléphoner au répondant, le genre de suivi convenu, quand il sera achevé et le nom de la personne la mieux ou la plus informée avec qui il pourra communiquer.

iii. Rendez-vous à prévoir

L'intervieweur devrait toujours avoir à la main un calendrier ou un journal pour inscrire l'heure des interviews et il devrait entrer le rendez-vous à l'écran approprié de l'ordinateur ou l'inscrire au questionnaire. L'intervieweur ne devrait pas inscrire l'heure et la date des rendez-vous prévus sur des morceaux de papier qu'il perdrait probablement.

Lorsqu'il établit le calendrier des interviews, l'intervieweur ne devrait pas oublier les points suivants :

a. Évitez les appels très tôt ou très tard quand vous communiquez avec un ménage. Nombre de personnes n'aiment pas recevoir des appels téléphoniques ou répondre à la porte tard en soirée (certains répondants seront effrayés si vous frappez à la porte en soirée). Pour les interviews sur place, l'intervieweur peut déposer une lettre de présentation dans la boîte aux lettres et une note personnelle précisant qu'il prévoit revenir, ou il peut ajouter un numéro de téléphone que le répondant peut composer pour confirmer l'heure de l'interview ou en prévoir une autre. Si l'intervieweur dérange quelqu'un pendant l'heure du repas, présenter des excuses est un bon moyen de susciter une réaction positive de la part du répondant.

b. Les répondants dans les entreprises sont souvent occupés lorsque l'intervieweur téléphone la première fois et il peut être nécessaire de prévoir une heure qui convient ou de laisser un numéro de téléphone que le répondant pourra composer pour déterminer une heure propice avec l'intervieweur. Le répondant ne devrait pas avoir de difficulté à communiquer avec l'intervieweur qui devrait donner un numéro de téléphone où il est toujours possible de le rejoindre ou préciser les heures pendant lesquelles il n'est pas disponible.

c. Si le répondant n'est pas disponible lorsque l'intervieweur téléphone la première fois, ce dernier devrait établir une relation amicale avec la personne qui répond à l'appel et déterminer le moment propice pour communiquer avec le répondant.

d. S'il ne peut communiquer avec un répondant, l'intervieweur devrait téléphoner un autre jour et à une heure différente. Si l'intervieweur a téléphoné deux fois pour les interviews sur place et s'il ne peut repérer le répondant, il ou elle peut demander à un voisin quel est le moment propice pour

communiquer avec quelqu'un au logement sélectionné. Il faut essayer de communiquer au moins trois fois pour les enquêtes par interviews sur place à Statistique Canada et jusqu'à dix fois pour les enquêtes téléphoniques.

9.4.2 Techniques de présentation

Il est important que l'intervieweur établisse une bonne relation avec le répondant au début de l'interview. La première impression que donne l'intervieweur influence énormément le résultat de l'interview. Il est essentiel d'avoir une attitude professionnelle, mais amicale, pour donner la meilleure impression. Cette disposition aidera à nouer un lien qui incitera le répondant à donner des réponses complètes et précises.

La présentation est la pierre angulaire d'une bonne relation entre l'intervieweur et le répondant. La présentation devrait être brève (en particulier au téléphone) et sincère. Elle devrait comprendre ce qui suit :

- le nom de l'intervieweur et de l'organisme,
- le titre et l'objectif de l'enquête,
- l'utilisation des données (pour établir l'importance de l'enquête),
- la loi en vertu de laquelle les données sont demandées,
- une garantie de confidentialité.

S'il s'agit d'interviews sur place, une lettre de présentation, une brochure sur l'enquête, ou les deux, livrée(s) avant l'interview sont un bon moyen de présenter l'enquête et d'en établir la légitimité. Au moment de l'interview, les répondants se souviennent alors avoir reçu quelque chose au sujet de l'enquête. L'intervieweur devrait porter et présenter sa carte d'identité d'intervieweur lorsqu'il visite un répondant.

Il est essentiel d'établir une bonne relation au départ pour le succès de l'interview. L'intervieweur doit écouter le répondant et être prêt à répondre à ses questions (le manuel de l'intervieweur devrait comprendre les réponses aux questions habituelles). Si le répondant hésite à aller de l'avant, l'intervieweur devrait essayer de déterminer les principales préoccupations du répondant et y répondre. Les préoccupations exprimées peuvent être l'une des suivantes :

i. Pourquoi m'avez-vous choisi?

Donner une explication simple de la sélection aléatoire convaincra le répondant qu'il a été choisi au hasard et que ses réponses sont importantes parce qu'il représente en fait d'autres personnes dans la population.

ii. Qui consultera mes données? Comment utiliserez-vous mes réponses?

La principale préoccupation est maintenant la confidentialité de l'information que peut donner le répondant. L'intervieweur devrait informer le répondant que ses réponses et celles d'autres répondants seront agrégées et utilisées uniquement pour produire des tableaux statistiques ou des résultats agrégés (sommaires). Les tableaux statistiques peuvent aider les auteurs de politiques et les décideurs à déterminer si la situation considérée dans l'enquête est satisfaisante ou si une certaine intervention est nécessaire.

iii. Je n'ai pas le temps maintenant.

L'intervieweur doit préciser honnêtement la durée de l'interview. Il devrait être disposé à faire l'interview immédiatement. Il ne devrait jamais supposer que le répondant n'a pas le temps. Si le répondant ne peut

répondre immédiatement à l'interview, l'intervieweur devrait suggérer une autre heure et prendre des dispositions fermes pour déterminer le moment de l'interview. Certains refuseront de participer s'ils ne comprennent pas clairement l'importance de l'enquête et l'utilisation des données. L'intervieweur devrait être certain que ces points sont clairement expliqués pendant la présentation.

9.4.3 Utilisation du questionnaire

La collecte des données doit être uniforme pour toutes les interviews, c'est-à-dire qu'il faut poser les mêmes questions de la même façon à tous les répondants. Les lignes directrices suivantes expliquent comment utiliser le questionnaire pour faire la collecte uniforme des données :

i. Il faut respecter la formulation lorsque vous posez les questions.

La recherche révèle que la modification, même très légère ou par mégarde, de la formulation peut changer la réponse obtenue.

ii. Il faut poser les questions dans l'ordre.

La séquence des questions est planifiée aux fins de la continuité. La séquence est aussi disposée de façon à ce que les premières questions n'aient pas de répercussions négatives sur les réponses du répondant aux questions ultérieures.

iii. Il faut poser chaque question pertinente.

Lorsque le répondant répond à une question, il répond aussi parfois à une autre question ultérieure dans l'interview. Il est quand même important que l'intervieweur pose la question ultérieure au moment opportun. Le répondant peut affirmer : *Vous m'avez déjà dit quelque chose à ce sujet, mais…* la situation indique que l'intervieweur est conscient de la réponse précédente et qu'il demande la collaboration du répondant pour répondre de nouveau à la question.

iv. Il faut poser les questions positivement.

Un intervieweur peut être mal à l'aise lorsqu'il pose certains questions et sembler s'excuser, par exemple : *Vous refuserez peut-être de répondre à cette question, mais . . .*ou *Cette question vous semblera probablement insensée . . .* Ces affirmations ont des répercussions négatives sur le débit de l'interview et elles ont tendance à modifier les réponses du répondant. Si l'intervieweur pose la question sur un ton positif ou neutre, le répondant comprend qu'il s'agit simplement d'une autre question et qu'il peut y répondre sans crainte d'être jugé.

v. Il faut expliquer les délais entre les questions, en particulier pendant les interviews téléphoniques.

Le temps d'entrée est plus long pour certaines réponses. L'intervieweur peut expliquer au répondant en ajoutant : *Veuillez excuser le délai, j'inscris – j'entre votre réponse.*

vi. Il faut poser de nouveau les questions mal comprises ou interprétées.

Les questions devraient être formulées de façon à ce que chacun les comprenne et la majorité des répondants les comprendront (si le questionnaire est bien conçu). À l'occasion cependant, un répondant peut mal comprendre ou interpréter une question. L'intervieweur doit alors répéter la question en

respectant la formulation. Si la réponse est toujours inappropriée, l'intervieweur devra peut-être approfondir (voir la Section 9.4.4).

vii. Il faut être attentif en particulier aux instructions « passez à ».

Une question filtre ou « passez à » détermine si les questions ultérieures s'appliquent et détermine le cheminement de l'interview. L'intervieweur doit être particulièrement attentif aux questions filtres et remarquer les « instructions à l'intervieweur » dans le questionnaire. L'instruction « passez à » est programmée pour l'interview assistée par ordinateur (IAO), mais l'intervieweur doit quand même bien connaître les caractéristiques du cheminement.

viii. Il faut avoir une attitude neutre, peu importe l'information obtenue.

Le répondant peut donner des réponses socialement acceptables, à son avis, s'il a l'impression que l'intervieweur porte un jugement. Le répondant ne doit pas avoir l'impression que certaines réponses sont plus acceptables que d'autres. Rien dans l'attitude de l'intervieweur ou dans son ton ne devrait laisser soupçonner la critique, la surprise, l'approbation ou la désapprobation, l'accord ou le désaccord lorsque la personne répond aux questions. L'intervieweur accepte la réponse du répondant si elle correspond à l'éventail des réponses acceptables.

La conception du questionnaire et les erreurs de réponse ont été étudiées au **Chapitre 5 - Conception du questionnaire**.

9.4.4 Approfondissement

L'approfondissement est une technique utilisée lorsque l'intervieweur remarque que la réponse n'atteint pas l'objectif de la question. Le répondant ne sait peut-être pas la réponse ou peut mal interpréter ou comprendre la question et sa réponse est donc incomplète, obscure ou incohérente, compte tenu d'autres renseignements. L'intervieweur doit donc approfondir sur un ton neutre pour obtenir l'information nécessaire.

Avant d'approfondir cependant, il faut poser de nouveau la question en respectant sa formulation au cas où le répondant n'aurait simplement pas entendu la question. L'intervieweur devrait utiliser les définitions s'il doit préciser la question. S'il n'obtient toujours pas une réponse satisfaisante, il peut utiliser un énoncé neutre pour demander davantage d'information, notamment :

Je ne suis pas certain de ce que vous voulez dire…

ou

Pouvez-vous m'en dire un peu plus?

ou

Autre chose?

L'intervieweur peut aussi aider le répondant en ciblant la catégorie de réponse exacte :

Quel nombre est le plus près, selon vous?

ou

Est-il plus grand, ou moins grand que…? (pour les réponses numériques)

ou

Était-ce le printemps, l'été, l'automne ou l'hiver?

Il est possible de reformuler la question, mais il faut être très prudent. Les questions ne devraient pas être reformulées de façon à suggérer une réponse. Si la question est *Combien de semaines avez-vous travaillé l'an dernier?,* par exemple, il ne faudrait pas la reformuler ainsi : *Avez-vous travaillé toute l'année?*, mais plutôt comme suit : *Avez-vous travaillé l'an dernier?* et si oui, *Pendant combien de semaines?*

9.4.5 Conclusion de l'interview

La dernière étape du processus de l'interview est de vérifier si vous avez obtenu toute l'information nécessaire et si elle est écrite lisiblement. À la fin de chaque interview, l'intervieweur examine le questionnaire attentivement et apporte les vérifications nécessaires. Il ou elle n'aura peut-être pas suffisamment de temps pour le faire en présence du répondant. Il est donc important de remercier poliment le répondant pour toute l'information donnée, mais d'ajouter qu'un suivi téléphonique est possible si une précision est nécessaire. L'intervieweur devrait offrir de répondre aux questions du répondant sur l'enquête, s'il en a d'autres. Il est important que le répondant ait l'impression d'avoir bien rempli son temps et que sa participation à l'enquête est importante et valable.

9.4.6 Refus et autres situations délicates

Un manque d'information sur l'enquête ou l'organisme statistique, ou un moment inopportun, expliquent habituellement le refus de participer d'un répondant. Les lignes directrices suivantes peuvent aider l'intervieweur à intervenir en cas de refus ou dans une situation délicate :

i. Dans le cas d'enquêtes auprès d'entreprises ou d'institutions, le chargé d'enquête devrait s'assurer que l'intervieweur communique avec la personne appropriée dans l'organisme au moment opportun et que l'information est facilement disponible. Dans la mesure du possible, on peut offrir à ces répondants de fournir les données selon une méthode et une présentation qui leur convient.

ii. Si l'heure de l'interview ne convient pas, l'intervieweur devrait présenter des excuses (au lieu de risquer un refus) et suggérer une heure pour téléphoner de nouveau.

iii. Il est peut-être possible de négocier avec un répondant réticent. L'intervieweur peut suggérer que le répondant réponde à quelques questions et, lorsque l'interview est lancée, le répondant peut décider de continuer. L'intervieweur peut informer le répondant, par souci de courtoisie, qu'il ou qu'elle peut refuser de répondre à des questions en particulier s'il considère qu'il essuierait autrement un refus total.

iv. Si l'intervieweur obtient un refus catégorique de vive voix, il devrait se retirer poliment et déclarer l'incident au surveillant pour suivi. Insister sur l'interview peut remettre en question le succès du surveillant qui tentera de convertir un refus en réponse.

v. Ne demandez pas au répondant de répondre devant d'autres personnes. L'intervieweur devrait prendre des dispositions pour téléphoner de nouveau au moment opportun si le répondant le préfère ou réserver un moment en privé pour l'interview.

vi. Si le répondant a une difficulté linguistique et accepte le recours à un interprète, un membre de la famille peut parfois interpréter chaque question et réponse.

vii. Si le répondant éprouve un problème personnel, par exemple une maladie grave ou un deuil, l'intervieweur doit évaluer la situation et déterminer s'il continue l'interview, prend des dispositions pour téléphoner de nouveau à un moment opportun ou met fin à l'interview s'il était déplacé de continuer ou s'il n'a aucune chance de succès.

viii. Communiquer avec un répondant dans une tour d'habitation pose parfois un problème parce que l'interphone n'est pas un bon moyen d'obtenir une interview. L'intervieweur peut essayer d'établir la communication avec le surintendant, le bailleur ou le propriétaire de l'immeuble pour expliquer le but de la visite et demander la permission d'entrer dans l'immeuble pour pouvoir faire une présentation sur place.

ix. L'intervieweur ne doit jamais argumenter avec un répondant ou le menacer, directement ou implicitement. L'information complète et fiable exige la collaboration de plein gré. L'intervieweur ne doit jamais se lancer dans des sujets de conversation controversés, par exemple, la politique.

x. L'intervieweur ne doit jamais avoir recours à des pratiques qui contreviennent à l'éthique pour procéder à une interview. Si le répondant n'est pas à domicile, l'intervieweur (après s'être identifié) peut demander à un voisin quel moment serait opportun pour téléphoner. L'intervieweur devrait cependant être prudent, éviter de susciter la méfiance et limiter ses questions lorsqu'il demande quand communiquer avec le répondant.

xi. En bout de ligne, et c'est aussi important, l'intervieweur ne devrait pas oublier ses droits. S'il est menacé de mauvais traitements, victime de menaces de vive voix, de harcèlement physique ou de violence, l'intervieweur devrait quitter immédiatement et déclarer l'incident au surveillant.

9.4.7 Interview d'un substitut (par procuration)

L'intervieweur peut obtenir l'information pour un répondant absent, dans certaines enquêtes, en interviewant une autre personne informée, et cette mesure est intitulée réponse d'un substitut ou réponse par procuration. La réponse par procuration convient aux enquêtes qui collectent des données généralement connues d'autres personnes que le répondant ciblé, et elle est donc habituellement inappropriée pour les questions personnelles, d'opinion personnelle ou à caractère délicat.

Il faudrait informer l'intervieweur pendant la formation si l'interview de substituts est permise et, si oui, il faudrait préciser qui sont les substituts acceptables. L'intervieweur devrait supposer en général qu'un substitut ne convient pas à l'interview, sauf avis contraire. Si un substitut ne convient pas et si que le répondant éprouve des difficultés à communiquer dans l'une ou l'autre des langues officielles, d'autres membres de la famille peuvent intervenir à titre de traducteurs avec la permission du répondant.

L'interview sans substitut exige généralement un effort plus grand que celui de l'interview avec substitut et le taux de réponse est moins élevé. L'intervieweur ne doit pas oublier que le nombre de rappels et de rendez-vous nécessaires pour procéder à des interviews sans substitut devrait être soigneusement déterminé pour éviter le fardeau de réponse et limiter les coûts de l'enquête.

9.4.8 Principaux points de l'interview efficace

Voici les principaux points de l'interview efficace :

i. Confiance

L'intervieweur doit avoir confiance en ses capacités. Il peut y arriver seulement s'il comprend bien l'enquête et le rôle de l'intervieweur.

ii. Aptitudes à écouter

L'intervieweur devrait attendre que le répondant ait fini de parler avant de cesser de l'écouter. L'intervieweur peut indiquer qu'il écoute en ajoutant à l'occasion *Oui, je vois*. L'intervieweur ne devrait cependant pas supposer qu'il sait ce que dira le répondant et finir la phrase à sa place. Il vaut mieux poser des questions s'il a l'impression que le répondant ou lui-même est passé à côté de la question.

iii. Compassion

L'intervieweur devrait être sensibilisé à la situation du répondant au moment de la visite ou de l'appel téléphonique. Si le répondant décrit un incident personnel, l'intervieweur devrait faire preuve d'intérêt (sans juger) et essayer ensuite d'orienter de nouveau le répondant vers l'interview.

iv. Élocution

L'expression de vive voix est importante, en particulier pour l'interview téléphonique. L'intervieweur devrait s'exprimer très clairement, à un rythme modéré. Si l'intervieweur s'exprime trop rapidement, le répondant peut manquer une partie de la question. S'il s'exprime trop lentement, le répondant peut commencer à répondre avant qu'il ait fini de poser la question. Baissez la tête et le ton de la voix baisse. Un ton de voix plus bas est plus clair et s'entend mieux, en particulier au téléphone. Il faudrait donner des exemples du rythme et du ton appropriés pendant la formation.

v. Connaissance du questionnaire

L'intervieweur doit connaître le questionnaire, les concepts et la terminologie utilisés dans l'enquête. Il n'aura pas le temps pendant l'interview de consulter les définitions ou les réponses aux questions dans le manuel. Rien ne peut rompre la communication plus rapidement que de longues pauses, en particulier pendant les interviews téléphoniques.

9.5 Surveillance de la qualité et du rendement

Dans les enquêtes avec interview assistée par ordinateur (IAO), la gestion des tâches de l'intervieweur, ainsi que la surveillance de l'état d'avancement dans l'ensemble, et de nombreux indicateurs de la qualité et du rendement sont automatisés à l'aide d'un logiciel. Les mesures de la qualité et du rendement sont intégrées à la programmation de l'IAO, mais les principes sont les mêmes que ceux des enquêtes sur support papier. Nous utilisons dans cette section les exigences des enquêtes sur support papier pour illustrer les contrôles nécessaires à appliquer dans toute enquête. Le lecteur trouvera davantage de détails au **Chapitre 10 - Traitement** et à l'**Annexe B - Contrôle qualitatif et assurance de la qualité**.

9.5.1 Surveillance de la qualité de la collecte des données

Le surveillant devrait surveiller la qualité de la collecte des données comme suit :

i. Surveillance étroite des intervieweurs

Le surveillant écoute les interviews concrètes, en particulier celles des nouveaux intervieweurs et des premières étapes de la collecte, pour vérifier si le questionnaire est utilisé correctement et si les techniques d'interview sont efficaces et uniformes d'une interview à l'autre. Étant donné que les intervieweurs peuvent donner une rétroaction valable sur les procédures de collecte des données et la conception du questionnaire, il faudrait les inviter à suggérer des améliorations à apporter à l'enquête.

ii. Vérification (ou vérification au hasard) des questionnaires achevés

Cette mesure garantit que l'intervieweur applique correctement les vérifications sur place et que l'information manquante peut être obtenue des répondants, en temps opportun, pendant que l'enquête se déroule toujours sur le terrain. Si le surveillant révise les vérifications des données, l'équipe chargée de l'enquête peut obtenir d'avance des renseignements sur le genre de rejets à la vérification possibles pendant l'étape de la vérification informatique après la collecte.

iii. Surveillance des mesures de la qualité et du rendement

Ce point est considéré en détail en 9.5.3. Ces mesures donnent une idée de la qualité pendant la collecte des données. Si le surveillant repère et règle les problèmes le plus tôt possible, il peut gérer la collecte des données de façon à atteindre, ou mieux, dépasser les taux de réponse et les autres indicateurs de qualité cibles.

iv. Contrôle strict des documents

Il faut exercer un contrôle sur chaque questionnaire à chaque étape de la collecte des données à l'aide de certaines entrées, par exemple, « reçu de l'intervieweur le (date) », « envoyé au dépistage le (date) », « envoyé à la saisie le (date) », etc. La formule de contrôle de l'échantillon du surveillant (et son lien avec les identificateurs de chaque intervieweur) est essentielle au contrôle efficace des documents et de l'échantillon. Les intervieweurs peuvent inscrire l'état d'avancement de leurs tâches sur papier, mais l'automatisation du contrôle de l'échantillon dans l'ensemble est recommandée pour simplifier le travail du surveillant. Un code d'état définitif (p. ex., achevé, refus, etc.) doit être appliqué à chaque questionnaire à la fin de la collecte.

v. Séance d'information des intervieweurs

Une séance d'information du personnel de l'enquête à la fin de la collecte des données peut aider à découvrir les problèmes du processus de collecte des données. Ces problèmes peuvent être des renseignements importants pour le traitement après la collecte (c.-à-d. identifier les vérifications qui sont nécessaires après la collecte). Des améliorations peuvent aussi être apportées au Système d'information de gestion, aux campagnes de relations publiques, etc., dans le cas des enquêtes réitérées.

vi. Repérage des modifications apportées aux données

Le chargé d'enquête voudra peut-être repérer les modifications apportées aux données pendant les processus d'enquête ultérieurs. La fréquence des rejets à la vérification après la collecte, ainsi que le nombre et le genre de corrections apportées aux données, peuvent se traduire par des renseignements utiles sur la qualité et servir d'indications précisant que les outils ou les procédures de collecte devraient être modifiés au cours des cycles ultérieurs de l'enquête.

9.5.2 Gestion des tâches de l'intervieweur

Périodiquement pendant la collecte (habituellement une fois par semaine), l'intervieweur doit faire rapport sur l'état d'avancement dans l'ensemble sur une feuille de contrôle des tâches. Le code d'état « en instance » (réparti ensuite en deux catégories : « tentative à faire » et « tentative faite ») est attribué aux unités de l'échantillon toujours en cours ou avec lesquelles l'intervieweur n'a pas encore communiqué. Lorsque l'intervieweur a traité une unité de l'échantillon au mieux de sa capacité, un code d'état « résolu » lui est attribué comme suit :

ii. Achevé : L'intervieweur a entièrement achevé l'interview.

iii. Achevé en partie : Le répondant n'a pas répondu à tout le questionnaire, mais il a répondu aux principales questions. Avant d'envoyer les intervieweurs sur le terrain, on identifie l'ensemble minimal de questions auxquelles les réponses constituent un questionnaire utilisable.

iv. Incomplet – non-réponse totale pour l'une des raisons suivantes :
- refus,
- absent pendant toute la période de l'enquête,
- dépistage impossible,
- hors du champ de l'enquête (p. ex., décédé, n'est plus en affaires, logement démoli),
- temporairement hors du champ de l'enquête (p. ex., logement inoccupé),
- pas de communication (p. ex., personne à domicile).

Les cas de refus et de « dépistage impossible » peuvent être référés pour suivi. Les rapports sommaires des tâches permettent de surveiller le nombre total d'unités de l'échantillon dans chaque catégorie. L'intervieweur remet des rapports sommaires hebdomadaires et envoie aussi au bureau régional tous les questionnaires résolus chaque semaine pour permettre le suivi des refus et la saisie des données en temps opportun.

9.5.3 Surveillance des surveillants

Les cibles de qualité et de rendement sont établies au début de la collecte des données. Les surveillants devraient se charger de la surveillance et de la gestion de leurs opérations, autant du point de vue des tâches de l'intervieweur que dans l'ensemble, pour garantir que les cibles sont atteintes.

La cible de rendement est déterminée selon la proportion d'enregistrements résolus :

$$\text{taux d'unités résolues} = \frac{\text{nombre d'unités résolues}}{\text{échantillon total (c.-à-d. résolues + en instance)}}$$

Ce taux donne une indication de la somme de travail prévu qu'a achevé l'intervieweur. Le taux d'unités résolues chaque semaine est comparé au taux cible pour vérifier si l'enquête sera conclue à temps. Les dépenses sont habituellement intégrées dans cette évaluation pour déterminer si l'enquête est toujours dans les limites du budget. Deux mesures habituelles du rendement par rapport au coût sont le *coût par unité résolue* et le *solde du budget par unité non résolue*.

Le taux de réponse est un autre indicateur de rendement. Les répercussions des non-réponses sont considérées au **Chapitre 3 - Introduction au plan d'enquête** et au **Chapitre 7 - Estimation**. Dans la plupart des enquêtes, la non-réponse est un élément important de l'erreur non due à l'échantillonnage (sous forme de biais) et de l'erreur d'échantillonnage (qui se traduit par une perte de précision des

estimations). La qualité ciblée est axée surtout sur le taux de réponse qui peut être déterminé comme suit pendant la collecte :

$$\text{taux de réponse} = \frac{\text{nombre d'unités répondantes (c.-à-d. complètes + partielles)}}{\text{unités résolues admissibles + unités non résolues}} \quad [1]$$

Supposons, par exemple, qu'un échantillon de 1 000 unités a été sélectionné dont 800 sont résolues (complètes, partielles, refus, hors du champ de l'enquête, etc.) après une semaine de collecte des données. Du nombre d'unités résolues, 700 sont dans le champ de l'enquête. Du nombre d'unités dans le champ de l'enquête, 550 répondent au questionnaire (réponse complète ou partielle). Le taux de réponse après la première semaine de l'enquête est donc 550/(700+200)= 61,1 %.

Un facteur d'ajustement est parfois appliqué aux unités non résolues, étant donné que certaines pourraient être hors du champ de l'enquête. Cela signifie dans l'exemple ci-dessus que, des 200 unités non résolues, environ 175 seraient probablement dans le champ de l'enquête (si l'on retient la même proportion que celle des unités résolues). Le taux de réponse ajusté serait donc 550/(700+175) = 62,9 %.

Outre les taux de réponse et la proportion d'enregistrements résolus, les surveillants devraient aussi surveiller d'autres indicateurs qui peuvent révéler d'éventuels problèmes de qualité. Des taux élevés de refus ou de non-communication dans une tâche peuvent indiquer que l'intervieweur a des problèmes. Si certains codes (en particulier « inoccupé ») sont plus fréquents chez certains intervieweurs, il peut y avoir un problème (p. ex., le logement était-il vraiment inoccupé ou les résidents étaient-ils temporairement absents? L'intervieweur ne fait peut-être pas la différence entre les deux). Des taux élevés de refus dans un échantillon complet révèlent la résistance du public et il peut être difficile d'obtenir la qualité de données voulue compte tenu de la période de l'enquête et de l'enveloppe budgétaire. Un nombre plus élevé que prévu d'unités « hors du champ de l'enquête » peut révéler des problèmes de base de sondage. Dans le cas des enquêtes-entreprises, les grandes entreprises peuvent être surveillées distinctement parce qu'elles peuvent avoir des répercussions sur les estimations définitives de l'enquête.

Pour plus d'information sur les normes et lignes directrices de mesure des non-réponses de Statistique Canada, consulter les Normes et lignes directrices de déclaration des taux de non-réponse (2001).

9.5.4 Techniques perfectionnées de mesure de la qualité

Les taux de réponse et certaines des autres mesures considérées ci-dessus sont les seuls indicateurs de la qualité de la collecte des données dans de nombreuses enquêtes. Dans les très grandes enquêtes réitérées, il est possible de concevoir des expériences pour essayer de mesurer l'ampleur du biais que suscitent les activités de collecte des données. Voici les études les plus habituelles :

Vérifications du listage : Les unités admissibles (dans le champ de l'enquête) des UPÉ ont-elles été toutes listées? Tous les membres admissibles d'un ménage ont-ils été listés? Y a-t-il des répétitions? Quelles sont les caractéristiques des unités manquantes ou réitérées? Voilà une tentative de mesure des erreurs de couverture.

[1] Ce résultat est équivalent au nombre total d'unités dans l'échantillon moins le nombre d'unités hors du champ de l'enquête.

i. Vérifications des logements inoccupés

Les unités listées inoccupées (ou hors du champ de l'enquête) étaient-elles réellement inoccupées ou y a-t-il eu erreur de classification? Quelles sont les caractéristiques des unités classées par erreur? On essaie ainsi de mesurer le sous-dénombrement dans la base de sondage.

ii. Nouvel interview pour mesurer les erreurs de réponse

Un intervieweur différent (parfois un surveillant) interviewe de nouveau un sous-échantillon de répondants pour déterminer si les réponses originales aux principales questions sont exactes. Les réponses à la nouvelle interview sont comparées aux réponses originales. Si les réponses sont différentes, certains cas, sinon tous, sont rapprochés pour déterminer lesquels sont corrects.

iii. Suivi des non-réponses

Une étude spéciale des non-réponses peut être faite pour évaluer le biais qui découle de la non-réponse totale à l'aide d'interviews spéciales de suivi avec des non-répondants de l'enquête (c.-à-d. essayer d'obtenir des réponses d'un sous-ensemble de non-répondants).

Étant donné que leur objectif est de mesurer le biais que suscite l'erreur non due à l'échantillonnage, ces études sont des enquêtes complexes en soi, elles peuvent coûter cher, et il faut les concevoir selon les principes considérés aux chapitres précédents.

Les erreurs non dues à l'échantillonnage ont été étudiées au **Chapitre 3 - Introduction au plan d'enquête**.

9.6 Sommaire

Ce chapitre explique comment les opérations de collecte des données peuvent être organisées, ainsi que les divers échelons de responsabilité, d'organisation et de contrôle. Une attention spéciale a été apportée aux interventions respectives du surveillant et de l'intervieweur. Les questions de relations publiques ont été considérées, y compris les campagnes publicitaires de l'enquête et les relations avec les répondants et le grand public. La préparation de la collecte des données, notamment les manuels pertinents, l'embauche et la formation des intervieweurs, le listage, le dépistage et les mises à jour de la base de sondage, ainsi que les techniques d'interview, ont fait l'objet d'un examen. Les méthodes de surveillance de la qualité et du rendement ont été considérées en définitive.

Il faudrait appliquer les lignes directrices suivantes pour garantir que les données obtenues pendant l'enquête sont complètes et précises, le plus possible :

i. Les intervieweurs sont essentiels au succès des enquêtes assistées par intervieweur. Ceux qui sont embauchés devraient avoir les capacités et les qualités personnelles nécessaires, ainsi que la formation et les outils appropriés.

ii. Les procédures de collecte des données devraient être appliquées uniformément à toutes les unités échantillonnées et les erreurs devraient être extraites le plus possible de ces procédures : tous les intervieweurs devraient recevoir la même formation et les mêmes manuels, tous les codeurs devraient recevoir les mêmes instructions, etc.

iii. Il faudrait appliquer les procédures de contrôle de l'échantillon appropriées à toutes les opérations de collecte des données. Ces procédures permettent de repérer l'état d'avancement des questionnaires, à partir du début jusqu'à la conclusion de la collecte et de l'entrée des données.

iv. Afin d'optimiser les taux de réponse et la qualité de l'information obtenue des entreprises et des établissements, le chargé d'enquête devrait veiller à ce qu'un intervenant communique avec la personne appropriée dans l'organisme, au moment opportun, pour que l'information soit facilement disponible. Il faudrait permettre à ces répondants de communiquer les données selon une méthode et une présentation qui leur conviennent, lorsque c'est possible.

v. Il faudrait établir des systèmes de contrôle efficaces pour garantir la sécurité des questionnaires et de la communication des données des bureaux régionaux au Bureau central.

vi. Il faudrait implanter un Système d'information de gestion pour repérer les mesures de la qualité, les dépenses et d'autres mesures du rendement pendant la collecte des données.

vii. Le chargé d'enquête voudra peut-être repérer les modifications apportées aux données pendant les processus ultérieurs de l'enquête. La fréquence des rejets à la vérification après la collecte, ainsi que le nombre et le genre de corrections apportées aux données, peuvent donner de l'information utile sur la qualité et servir d'indication révélant que les outils et procédures de collecte devraient être modifiés dans les cycles ultérieurs de l'enquête.

Bibliographie

Cialdini, R., M. Couper et R.M. Groves. 1992. Understanding the Decision to Participate in a Survey. *Public Opinion Quarterly,* 56: 475-495.

Couper, M.P. et R.M. Groves. 1992. Le rôle de l'intervieweur dans la participation aux enquêtes, *Techniques d'enquête*, 18(2): 279-294.

Statistique Canada. 2001. *Normes et lignes directrices pour la déclaration des taux de non-réponse*.

Statistique Canada. 1995. *Les techniques d'interview d'une enquête : un guide pour mener des interviews efficaces.* 12F0047XPF.

Statistique Canada.1998. Politique d'information des répondants aux enquêtes. *Manuel des politiques*. 1.1.

Statistique Canada. 1998. Lignes directrices concernant la qualité. 12-539-XIF.

PUBLICATIONS ÉLECTRONIQUES
DISPONIBLES À

www.statcan.gc.ca

Chapitre 10 - Traitement

10.0 Introduction

Le traitement transforme les réponses du questionnaire obtenues pendant la collecte pour qu'elles conviennent à la totalisation et à l'analyse des données. Il comprend toutes les activités de traitement des données, automatisé et manuel, après la collecte et avant l'estimation. Le traitement demande beaucoup de temps et de ressources, et il a des répercussions sur la qualité et le coût des données définitives. Il est donc important de bien le planifier, de surveiller la qualité de sa mise en œuvre et d'apporter des mesures correctives au besoin.

Le genre de données à obtenir, la méthode de collecte, le budget et les objectifs de l'enquête du point de vue de la qualité des données, notamment, déterminent la portée et l'ordre des activités de traitement. Le codage, par exemple, peut être fait avant ou après la saisie des données, mais la vérification est habituellement faite tout au long de l'enquête. Voici un exemple des activités de traitement d'un questionnaire sur support papier :

i. Vérification des données du questionnaire après la collecte. Cette étape garantit que toute l'information nécessaire a été obtenue et enregistrée lisiblement, que les notes de l'intervieweur ont été examinées et que certaines vérifications préliminaires ont été faites pour déterminer s'il y a des incohérences et des erreurs grossières.

ii. Codage de toutes les données du questionnaire qui doivent être codées (p. ex., réponses aux questions ouvertes rédigées à la main).

iii. Saisie des données. D'autres activités de codage peuvent suivre la saisie des données.

iv. Vérification détaillée et ensuite, imputation. Les questionnaires rejetés après une vérification ou plus sont retirés du lot pour examen ultérieur, suivi auprès des répondants ou imputation.

v. Détection des valeurs aberrantes pour identifier les valeurs extrêmes ou suspectes.

vi. Sauvegarde dans une base de données pour faciliter l'utilisation des données pendant les activités après le traitement.

Plusieurs activités ci-dessus, notamment la saisie, la vérification et le codage, peuvent être intégrées par automatisation à l'aide de méthodes de collecte assistée par ordinateur pour rationaliser le traitement.

Étant donné que des erreurs sont probables à chaque étape du traitement, en particulier pour les activités répétitives et manuelles, par exemple le codage, la saisie et la vérification, il faudrait surveiller le traitement et apporter des mesures correctives au besoin pour maintenir ou améliorer la qualité. Cette intervention est possible en appliquant les procédures de contrôle qualitatif et d'assurance de la qualité.

L'objectif de ce chapitre est de couvrir les principales activités de traitement : codage, saisie des données, vérification, imputation, détection des valeurs aberrantes, traitement et implantation d'une base de données. Le lecteur obtiendra des détails sur le contrôle qualitatif et l'assurance de la qualité en consultant l'**Annexe B - Contrôle qualitatif et assurance de la qualité**.

10.1 Codage

Le codage est le processus d'attribution d'une valeur numérique aux réponses pour faciliter la saisie des données et le traitement en général. Il est mentionné au **Chapitre 3 - Introduction au plan d'enquête** que le codage comprend l'attribution d'un code à une réponse donnée ou la comparaison de la réponse à un ensemble de codes et la sélection de celui qui décrit le mieux la réponse.

Déterminer un ensemble de catégories de réponse à une question pose des difficultés qui ont été considérées au **Chapitre 5 - Conception du questionnaire**. Les catégories de réponse sont déterminées avant la collecte pour les questions fermées et le code numérique est habituellement affiché sur le questionnaire à côté de chaque catégorie de réponse. Le codage des réponses aux questions ouvertes est fait après la collecte, et il peut être manuel ou automatisé. Le codeur doit interpréter et faire preuve de jugement pour le codage manuel et les résultats peuvent varier d'un codeur à l'autre.

Lorsque vous choisissez la méthode de codage, l'objectif devrait être de classer les réponses en un ensemble significatif de catégories exhaustives et mutuellement exclusives qui font ressortir les caractéristiques essentielles des réponses. Le codage de certaines questions peut être direct (p. ex., état matrimonial). Autrement, un autre système de codage standard peut exister, par exemple pour la géographie, la branche d'activité et la profession. Il n'y a cependant pas de système de codage standard appliqué à de nombreuses autres questions et le choix d'une bonne méthode de codage n'est pas une tâche triviale. La méthode de codage devrait être uniforme et logique. Il faut déterminer à quel point les codes doivent être détaillés, compte tenu de l'objectif de l'enquête, des totalisations et des analyses de données à faire. Il vaut mieux commencer avec une liste assez large parce qu'un nombre insuffisant de catégories peut être trompeur et une grande catégorie *autre* peut être démunie d'information. Les catégories peuvent toujours être regroupées, mais il est difficile de les ventiler après coup.

Nous avons précisé au **Chapitre 5 - Conception du questionnaire** qu'il faudrait considérer le genre d'interview pour déterminer les catégories de réponse à une question fermée. Remettre une liste de 50 catégories sur support papier aux répondants est faisable (mais pas idéal) pour les enquêtes par autodénombrement, mais ce n'est pas pratique d'énumérer les 50 catégories de réponse pendant une interview téléphonique. Idéalement, toutes les questions d'un questionnaire seraient fermées et il y aurait une brève liste de catégories de réponse pour simplifier le codage. Ce n'est pas toujours possible en pratique et les questions ouvertes sont parfois nécessaires.

10.1.1 Codage préalable des questions fermées

Les catégories de réponse aux questions fermées peuvent être codées d'avance dans le questionnaire. Des cases pour les codes peuvent être disposées à côté de la réponse à coder ou dans la marge pour les questionnaires sur support papier. Voilà qui améliore énormément l'efficacité de la saisie des données après la collecte : au lieu de dactylographier la catégorie de réponse sélectionnée, un code numérique est entré (il est aussi plus facile d'analyser des codes numériques qu'une suite de mots). Les codes des méthodes de collecte assistée par ordinateur sont automatiquement saisis lorsque l'intervieweur ou le répondant choisit une réponse.

Le système de codage suivant, par exemple, a été utilisé dans le Sondage auprès des fonctionnaires fédéraux en 2002 :

Combien de promotions avez-vous eues depuis trois ans?
- ❍ *aucune*
- ❍ *une*
- ❍ *plus d'une*

Les avantages des questions fermées ont été considérés au **Chapitre 5 - Conception du questionnaire** : elles sont un fardeau moindre pour les répondants, et la collecte, la saisie et l'analyse des données coûtent moins cher, elles sont plus rapides et faciles que les questions ouvertes. La formulation naturelle du répondant est cependant inconnue, un inconvénient des questions fermées. Il peut donc être difficile de vérifier la qualité du codage. Si une question ouverte est posée pour déterminer la profession d'une personne, par exemple, la description du travail du répondant peut donner un code de profession différent de celui que le répondant ou l'intervieweur aurait sélectionné dans une énumération de codes de profession ajoutée à une question fermée.

10.1.2 Codage manuel des questions ouvertes

Lors du codage manuel des questions ouvertes, le codeur (habituellement après la collecte) doit lire, interpréter et convertir à la main une réponse par écrit à une question ouverte en un code numérique. Ce code numérique est ensuite inscrit dans le questionnaire ou entré à l'ordinateur. Le codeur devra peut-être simplement remarquer si la réponse contient un mot clé ou une référence à un élément en particulier pour attribuer un code. Parfois le codage est déterminé à partir de la réponse à une question seulement, parfois à partir des réponses à plusieurs questions connexes. La clarté et l'exhaustivité de la réponse écrite, la qualité de la vérification initiale, la logique de la méthode de codage et l'aptitude du codeur influencent énormément la qualité du codage dans ce cas.

Les codeurs doivent être bien formés parce qu'il faut tenir compte des points suivants pour appliquer la méthode de codage :
- le nombre de réponses possibles,
- la complexité (jugement),
- l'ambiguïté possible de la réponse (c.-à-d. la qualité de la réponse).

La variabilité entre les codeurs est inévitable. Une vérification détaillée du premier lot de questionnaires d'un codeur est nécessaire pour repérer les erreurs et déterminer si une formation supplémentaire est nécessaire. On peut ensuite faire des vérifications périodiques de la qualité du codage et apporter des mesures correctives au besoin. Cette mesure est souvent appliquée à l'aide des méthodes de contrôle qualitatif (voir l'**Annexe B - Contrôle qualitatif et assurance de la qualité**).

10.1.3 Codage automatisé des questions ouvertes

Le codage des questions ouvertes est habituellement une opération manuelle. Compte tenu de l'évolution technologique, des ressources restreintes et, plus encore, des exigences de rapidité et de qualité, le processus de codage est cependant de plus en plus automatisé.

Deux fichiers sont habituellement entrés dans un système de codage automatisé. Un fichier contient les réponses de l'enquête qu'il faut coder, intitulé fichier de réponse en lettres. Le deuxième fichier est intitulé fichier de référence et contient les réponses (ou phrases) écrites typiques et leurs codes numériques correspondants.

Le parsage est le plus souvent la première étape du codage automatisé. Le parsage est le processus de normalisation d'une phrase pour permettre à l'ordinateur de repérer les phrases équivalentes. Le parsage comprend habituellement la suppression des caractères superflus, par exemple la ponctuation, les mots en double, les mots sans importance, certains suffixes et préfixes, etc. Le parsage est appliqué aux fichiers de référence et de réponses en lettres avant d'aller de l'avant.

L'étape suivante comprend la recherche d'une entrée dans le fichier de référence qui correspond exactement à une réponse en lettres dans le fichier de l'enquête. S'il y en a une, le code du fichier de référence est copié dans le fichier de l'enquête et l'enregistrement[1] est considéré codé. S'il n'y a pas de correspondance exacte cependant, on essaie de trouver les enregistrements du fichier de référence qui correspondent le plus possible. Une cote est attribuée à chaque enregistrement du fichier de référence pour indiquer à quel point la phrase du fichier de référence est semblable à la réponse du questionnaire. Les cotes sont évaluées selon des paramètres déterminés (lesquels sont précisés pour réduire le risque d'erreur) et si une cote est suffisamment élevée, le code est transféré à la réponse du questionnaire et l'enregistrement est considéré codé.

Plusieurs enregistrements du fichier de référence ayant des cotes semblables sont parfois repérés, mais parfois aussi, le fichier de référence ne contient aucun enregistrement qui correspond suffisamment à la réponse du questionnaire. Dans ces situations, les enregistrements sont généralement envoyés à une petite équipe de codage manuel dotée de codeurs experts chargés de coder les enregistrements non codés à la fin de l'étape automatisée et de vérifier la qualité du produit du système automatisé (voir l'**Annexe B - Contrôle qualitatif et assurance de la qualité**).

10.2 Saisie des données

La saisie des données consiste à transformer les réponses pour les rendre lisibles à la machine. La saisie est faite après la collecte (habituellement après le prétraitement et certaines vérifications préliminaires du questionnaire) pour les méthodes de collecte sur support papier. Dans ce cas par exemple, un commis (un opérateur de saisie des données) entre au clavier de l'ordinateur les valeurs déclarées dans le questionnaire. La saisie se fait au moment de la collecte pour les méthodes de collecte par ordinateur.

Il y a plusieurs moyens d'améliorer l'efficacité de la saisie des données. Les méthodes de collecte des données assistées par ordinateur sont un moyen. La collecte et la saisie étant simultanées, le processus de collecte et de saisie des données est donc plus rapide et efficace que celui des méthodes sur support papier. Cependant, les programmes des logiciels demandent beaucoup de développement et de mises à l'essai. (Les avantages et les inconvénients de la collecte des données assistée par ordinateur sont étudiées au **Chapitre 4 - Méthodes de collecte des données**.)

Le codage préalable des questions fermées peut améliorer énormément l'efficacité de la saisie des données manuelle pour les méthodes de collecte sur support papier. La lecture optique des questionnaires remplis est une autre option. La lecture optique fonctionne au mieux pour les questions fermées et elle est moins fiable pour la saisie des réponses aux questions ouvertes. La lecture optique peut réduire les erreurs de saisie des données comparativement à la saisie manuelle, mais les erreurs de lecture optique sont possibles et doivent être évaluées et minimisées. La logistique de la lecture optique demande plus de travail pour les longs questionnaires parce qu'il faut enlever les agrafes, ajouter des identificateurs de questionnaire à chaque page, réinitialiser les lecteurs pour lire les différentes pages, etc. Coder toutes les

[1] Dans ce chapitre, le *questionnaire* est généralement le document sur support papier et l'*enregistrement* est la version électronique du questionnaire rempli.

réponses sur une seule feuille de papier est une autre option. La lecture optique est simplifiée, mais l'intervieweur devra faire davantage d'efforts pour lire une question sur une feuille et inscrire la réponse sur une autre. Cette méthode est aussi restreinte aux questions fermées et, si l'intervieweur a en main une grande feuille remplie de cases de réponse, il est plus facile de coder la mauvaise réponse ou de coder la réponse dans la mauvaise case. L'intervieweur aura aussi de la difficulté à consulter une réponse d'un répondant parce que les questions et réponses sont inscrites sur des feuilles distinctes.

Il est particulièrement important d'appliquer les procédures de contrôle qualitatif et d'assurance de la qualité aux méthodes de collecte sur support papier pour minimiser et corriger les erreurs pendant la saisie des données (voir l'**Annexe B - Contrôle qualitatif et assurance de la qualité**).

10.3 Vérification

Dans un monde idéal, chaque questionnaire serait rempli sans erreur. Les réponses à certaines questions peuvent malheureusement être absentes, incomplètes ou inexactes. ***La vérification est l'examen des réponses pour identifier les entrées manquantes, non valables ou incohérentes qui indiquent des enregistrements de données éventuellement erronées***. La vérification permet habituellement d'identifier les erreurs non dues à l'échantillonnage que suscitent les erreurs de mesure (réponses), les non-réponses ou le traitement. La vérification vise à :

- mieux comprendre les processus et les données de l'enquête,
- repérer les données erronées ou manquantes,
- faire le suivi auprès du répondant,
- acheminer un enregistrement pour imputation,
- supprimer un enregistrement.

Des règles de vérification sont appliquées pour identifier les enregistrements erronés. Voici des exemples de règles de vérification :

- chaque question doit avoir une réponse et seulement une,
- les réponses valides à la question X sont *1* ou *2*,
- la somme des parties pour la question X ne peut être moindre que la réponse à la question Y.

Des vérifications peuvent être faites à plusieurs étapes pendant le processus de l'enquête et elles passent des simples vérifications préliminaires des intervieweurs sur place aux vérifications automatisées plus complexes exécutées par un programme informatique après la saisie des données. Les règles de la vérification sont généralement formulées selon ce qui peut être logique ou valide, compte tenu :

- des connaissances de l'expert en la matière,
- d'autres enquêtes ou données connexes,
- de la structure du questionnaire et de ses questions,
- d'une théorie statistique.

Les experts en la matière devraient savoir comment les variables sont liées l'une à l'autre et quelles réponses sont raisonnables. Leur intervention est importante pour préciser le genre de règles appropriées. Ces analystes ont habituellement l'expérience du genre de données vérifiées. Un analyste des transports, par exemple, peut être conscient de l'étendue des valeurs acceptables pour les taux de consommation d'essence des divers modèles et marques de véhicule. L'analyse d'autres enquêtes ou ensembles de données pertinents aux mêmes genres de variables que celles qui sont vérifiées peut être utile pour établir certaines règles de vérification.

Point tout aussi important, la configuration et la structure du questionnaire ont des répercussions sur les règles de la vérification. Les vérifications devraient déterminer si les réponses correspondent au

cheminement logique des questions. Il est souvent révélé à l'aide des instructions sur *l'enchaînement des questions* ou « *passez à* » qui sous-entendent que certaines questions du questionnaire ne s'appliquent pas à certaines catégories de répondants et le répondant doit alors passer à une autre question.

Il y a trois principales catégories de vérification : les vérifications de *validité*, de *cohérence* et de *distribution*. Les vérifications de validité et de cohérence sont appliquées à un questionnaire à la fois. Les vérifications de validité ciblent la syntaxe des réponses et comprennent la vérification des caractères non numériques entrés dans les champs numériques et le repérage des valeurs manquantes. Les deux premiers exemples de règles de vérification ci-dessus correspondent à des vérifications de validité. Celles-ci peuvent aussi déterminer si les données codées s'inscrivent dans l'étendue permise des valeurs. Une vérification de l'étendue peut être faite, par exemple, pour l'âge déclaré d'un répondant, afin de vérifier s'il se situe entre 0 et 125 ans.

Les *vérifications de cohérence* déterminent si les liens entre les questions sont respectés. Le troisième exemple de règle de vérification ci-dessus est la vérification de cohérence. Les vérifications de cohérence peuvent utilisent des liens logiques, juridiques, comptables ou structurels entre les questions ou entre les volets d'une question. Le lien entre la date de naissance et l'état matrimonial est un exemple auquel la vérification de cohérence peut être appliquée : « l'état matrimonial d'une personne de moins de 15 ans peut seulement être *jamais marié* ». Les vérifications de cohérence peuvent aussi porter sur le cheminement logique des questions, par exemple, « si le répondant inscrit *non* à la question X, il ne peut répondre à la question Y ». Les vérifications de cohérence peuvent aussi comprendre le recours aux données chronologiques (p. ex., les ratios d'une année à l'autre). Dans le cas des enquêtes-ménages, les vérifications peuvent être faites entre les membres du ménage.

Les *vérifications de distribution* sont faites en observant les données entre les questionnaires. Elles tentent de déterminer les enregistrements qui sont des valeurs aberrantes du point de vue de la distribution des données. Les vérifications de distribution sont parfois considérées comme des vérifications statistiques (Hidiroglou et Berthelot, 1986) ou la détection de valeurs aberrantes (voir la Section 10.5). Les erreurs non dues à l'échantillonnage sont considérées au **Chapitre 3 - Introduction au plan d'enquête**.

10.3.1 Vérifications pendant la collecte des données

Les vérifications pendant la collecte des données sont souvent intitulées vérifications sur le terrain (sur place), ce sont en général des vérifications de validité et, parfois, de simples vérifications de cohérence. Voici pourquoi la vérification est faite pendant la collecte des données :

- déterminer s'il faut améliorer la méthode de collecte des données,
- décider s'il faut davantage de formation,
- détecter les erreurs évidentes et faire le suivi immédiat auprès du répondant,
- épurer les entrées.

Les intervenants suivants peuvent faire la vérification pendant la collecte des données :

- le répondant (enquête par autodénombrement),
- l'intervieweur pendant l'interview,
- l'intervieweur immédiatement après l'interview,
- le surveillant de l'intervieweur,
- le personnel de bureau.

Les vérifications sur place sont faites pour déterminer les problèmes que posent les procédures de collecte des données et la conception du questionnaire, ainsi que le besoin d'approfondir la formation de

l'intervieweur. Elles servent aussi à détecter les erreurs que l'intervieweur ou le répondant ont commises pendant l'interview, ainsi que l'information manquante pendant la collecte, afin d'amenuiser le besoin de suivi ultérieur. La vérification pendant la collecte est beaucoup plus facile à faire si elle est incorporée à une méthode de collecte assistée par ordinateur.

Les répondants peuvent vérifier leurs propres réponses à un questionnaire d'autodénombrement. Presque toutes les enquêtes assistées par intervieweur comprennent une certaine vérification pendant l'interview, les intervieweurs ont des instructions et sont formés pour examiner les réponses qu'ils inscrivent dans un questionnaire immédiatement à la fin de l'interview, après avoir quitté le logement ou raccroché le combiné du téléphone. Ils ont ainsi l'occasion de détecter et de traiter les enregistrements rejetés après l'application des règles de la vérification, soit parce qu'ils ont toujours l'information exacte à la mémoire, soit parce qu'ils peuvent facilement faire le suivi à peu de frais auprès du répondant pour déterminer les valeurs exactes. Les rejets à la vérification toujours non résolus sont habituellement réglés plus tard par imputation.

Les vérifications sur place servent aussi à épurer des réponses. L'intervieweur inscrit souvent de brèves notes en marge du questionnaire pendant l'interview ou dans la section des notes de l'application de l'ITAO. L'intervieweur prend des notes parce qu'il ne connaît pas le programme de codage des questions ouvertes ou il veut consulter le manuel de l'intervieweur pour interpréter une réponse. L'intervieweur vérifie alors ces questionnaires après l'interview pour épurer ces notes.

L'une des tâches confiées aux surveillants est la vérification du travail de ces intervieweurs pour détecter les erreurs et les en informer. Les genres de rejets détectés sont habituellement semblables à ceux que pourrait repérer l'intervieweur immédiatement après l'interview et l'intervieweur a habituellement l'occasion de faire le suivi auprès du répondant pour déterminer les valeurs exactes. Les surveillants devraient aussi chercher les caractéristiques des erreurs commises. Il faudrait communiquer à toute l'équipe les leçons apprises d'un intervieweur.

Dans de nombreuses enquêtes, le répondant ou l'intervieweur envoie les questionnaires remplis au bureau régional pour téléchargement et *prétraitement* par les préposés aux activités de bureau. Ce prétraitement comprend souvent les mêmes vérifications des intervieweurs ou des surveillants, ou des vérifications supplémentaires. Le prétraitement comprend le déchiffrage des réponses inscrites à la main, l'interprétation des remarques de l'intervieweur, la normalisation des échelles de mesure (p. ex., calculer en mètres une valeur inscrite en pieds), etc. Il permet aussi de vérifier si l'intervieweur a rempli tous les champs administratifs du questionnaire, notamment, les codes d'état des réponses (p. ex., qui indiquent si le questionnaire est rempli en tout ou en partie). Ce processus donne une vérification ou un examen autonome systématique des données du questionnaire avant de les envoyer à la saisie des données. La vérification des codes d'identification du questionnaire peut aussi être un élément important de cet exercice parce que les questionnaires ne peuvent être entrés ou les données ne peuvent être saisies sans identification complète. L'ampleur des vérifications dépend du budget disponible et jusqu'à quel point le personnel de bureau affecté à la vérification peut repérer et résoudre les problèmes. Ce genre de vérification est intégré, si possible, au codage, au pointage ou à la répartition en lots des questions du questionnaire qui peuvent être nécessaires avant de lancer la saisie des données. Le personnel du bureau régional peut faire le suivi auprès du répondant, dans certains cas, pour résoudre d'importants rejets à la vérification.

10.3.2 Vérifications après la collecte des données

Les vérifications les plus détaillées et compliquées sont faites au cours d'une étape distincte de vérification et d'imputation après la collecte des données. Les opérateurs de la saisie des données peuvent

faire des vérifications pendant la saisie, ou des programmes informatiques s'en chargent automatiquement, ou c'est l'application informatique qui les fait dans le cas des méthodes de collecte assistées par ordinateur. S'il s'agit de la saisie manuelle des données des questionnaires sur support papier, il est économique de profiter de l'occasion pour appliquer les règles et épurer les données suffisamment, afin que les étapes de traitement ultérieures soient plus efficientes. La vérification pendant la saisie des données est généralement minimisée parce que l'intervention après un rejet à la vérification ralentit la saisie des données. À cette étape du traitement, ce sont surtout des vérifications de validité et de simples vérifications de cohérence.

Les règles de vérification plus complexes sont généralement réservées à l'étape distincte de vérification après la saisie des données, ainsi que des vérifications de validité et des vérifications plus complexes de la cohérence souvent faites en même temps que la vérification sélective et la détection des valeurs aberrantes (voir la Section 10.5).

Au volet des rejets à la vérification après la collecte des données, la procédure habituelle est d'indiquer le champ rejeté à la vérification et de l'imputer, ou d'extraire l'enregistrement du traitement ultérieur.

La majorité des rejets à la vérification à cette étape sont marqués pour imputation. Il est utile d'entrer un code spécial pour les valeurs rejetées à la vérification, afin d'indiquer qu'une valeur inacceptable ou un blanc invalide a été repéré. Ces indications sont particulièrement utiles pour évaluer la qualité des données de l'enquête. Dans certains cas, l'enregistrement ou le questionnaire peut être rejeté après avoir appliqué tellement de règles de vérification (ou un petit nombre de vérifications critiques) qu'il devient inutile au traitement ultérieur. L'enregistrement est alors habituellement considéré comme celui d'un non-répondant, il est retiré du circuit du traitement et il y a ajustement de la pondération pour non-réponse (consulter le **Chapitre 7 - Estimation** pour obtenir des détails sur les ajustements de la pondération).

10.3.3 Vérification sélective

En vérification, il faut faire un compromis entre la perfection voulue pour chaque enregistrement et l'affectation de ressources raisonnables (c.-à-d. temps et argent) pour bien épurer les données. Beaucoup de temps et d'efforts ont été investis auparavant pour identifier toutes les erreurs d'enquête. La survérification des données est non seulement une utilisation médiocre des ressources, mais elle peut aussi donner des résultats biaisés. Les données doivent habituellement correspondre à un modèle défini d'avance ou sinon, elles sont rejetées à la vérification. Si les données sont modifiées chaque fois qu'elles sont rejetées à la vérification, elles peuvent devenir énormément biaisées comparativement au modèle et ne plus refléter la situation réelle. La survérification et les suivis réitérés auprès des répondants peuvent aussi accroître le fardeau de réponse et miner la collaboration des répondants à l'avenir.

Des pratiques de vérification sélective sont recommandées, en particulier pour les enquêtes-entreprises (c.-à-d. si la population est asymétrique et si quelques entreprises dominent les estimations), afin d'éviter de réserver trop de temps et d'épuiser des ressources pour vérifier des données qui ont peu de répercussions sur les estimations définitives. L'approche de la vérification sélective repose sur l'idée selon laquelle seuls les rejets critiques à la vérification doivent être traités. La vérification sélective s'applique en général aux données quantitatives. Une procédure qui modifie les enregistrements individuels selon leurs répercussions éventuelles sur les estimations de l'enquête, ou par l'intermédiaire de l'analyse des données agrégées, est un exemple de l'application de la vérification sélective. Les résultats éventuels de la vérification sélective des rejets sont le suivi auprès du répondant, le retrait de l'enregistrement du traitement ultérieur ou une indication des enregistrements ciblés pour imputation.

La vérification sélective permet une :
- diminution des coûts,
- amélioration de la qualité des données si les ressources sont réacheminées vers les enregistrements ayant des répercussions importantes ou vers d'autres activités,
- amélioration de la rapidité d'exécution lorsque diminue le temps de traitement,
- diminution du fardeau de réponse lorsque diminue le nombre de suivis.

Cependant, avec la vérification sélective :
- la qualité des données obtient moins d'attention au niveau de l'unité individuelle,
- il peut rester des données incohérentes et les utilisateurs pourraient avoir l'impression que la qualité des données est médiocre,
- l'erreur non due à l'échantillonnage pour les petits domaines peut être plus grande si tous les questionnaires ne sont pas vérifiés individuellement,
- les préposés au traitement des données, les experts en la matière, la direction ou les utilisateurs des données peuvent être réticents et faire moins confiance aux données.

Voici certaines approches de la vérification sélective :

i. Approche descendante

Si cette méthode est appliquée, les valeurs des données pondérées les plus influentes sont listées de haut en bas pour un domaine d'estimation donné et elles sont examinées une par une. La vérification et l'examen des données prennent fin lorsque la valeur suivante de la donnée la plus influente n'a pas de répercussions importantes sur l'estimation du domaine. Considérons, par exemple, un échantillon de cinq entreprises tirées d'une population de 100 si on veut estimer dans l'enquête le nombre total d'employés dans la population. L'estimation du nombre total d'employés dans l'enquête est 737. L'analyste a l'impression que cette estimation est trop élevée (parce qu'il prévoit que le nombre moyen d'employés par entreprise est de trois). L'analyste examine la contribution relative de chaque enregistrement à l'estimation totale. On peut constater au tableau 1 que le premier enregistrement atteint 81,4 % de l'estimation du total. Compte tenu de son influence sur l'estimation, cet enregistrement est examiné de plus près. Il devient vite évident que le nombre d'employés déclarés dans cette entreprise est plus élevé que prévu et la pondération est plus élevée que celle des autres enregistrements (peut-être à cause d'un ajustement pour les non-réponses). Cet enregistrement est donc traité comme une observation influante (voir la Section 10.5). Étant donné que les autres valeurs pondérées représentent seulement une petite proportion du total dans l'ensemble, elles ne sont pas examinées de plus près.

Tableau 1 : Exemple de vérification descendante

Enregistrement	Nombre d'employés	Pondération	Proportion du total
1	12	50	81,4 %
2	7	8	7,6 %
3	3	12	4,9 %
4	2	15	3,3 %
5	1	15	2,0 %

ii. Méthode agrégée

La méthode agrégée permet d'identifier les *estimations pour un domaine* qui paraissent suspectes. Les données pondérées de *tous* les enregistrements du domaine sont ensuite examinées. Dans une enquête estimant la taille moyenne des ménages, par exemple, si la taille moyenne dans un village en particulier

est de 23, tous les enregistrements individuels pondérés de ce village seraient examinés pour déterminer si certaines valeurs semblent être substantiellement plus élevées que les autres.

iii. Méthode graphique

Les données sont disposées en graphique pour identifier les valeurs suspectes. La distribution des données peut être présentée en graphique, par exemple, pour identifier les queues improbables de la distribution.

iv. Cote du questionnaire

Berthelot et Latouche (1992) proposent l'utilisation d'une fonction de cotation, c'est-à-dire qu'une cote est attribuée à chaque répondant selon une certaine mesure de la taille, le nombre d'éléments de données suspects dans le questionnaire et l'importance relative des variables. Seuls les enregistrements ayant une cote élevée sont examinés.

10.3.4 Vérifications manuelle et automatisée

La vérification peut être automatisée au moyen d'un programme informatique. L'ampleur de la vérification à faire (c'est-à-dire le nombre d'éléments de données ou de questionnaires), les caractéristiques et la complexité des règles de vérification appliquées, les répercussions de l'unité, l'importance des variables et l'étape du traitement du questionnaire à laquelle s'appliquent les règles de la vérification déterminent si le traitement manuel ou automatisé est approprié. Plus les règles de la vérification sont complexes, plus le traitement manuel est difficile et exposé aux erreurs. Dans certaines enquêtes d'autre part (p. ex., sur support papier), il est difficile, sinon impossible, d'intégrer les vérifications automatisées pendant la collecte des données. D'autres éléments qui ont des répercussions sur le choix de la vérification manuelle ou automatisée comprennent la nécessité de surveiller les interviews et de laisser une piste de vérification. La vérification après la saisie des données est cependant automatisée d'habitude. Selon un principe généralement accepté pour cette étape de vérification, et l'étape d'imputation connexe, il ne devrait pas être nécessaire de revenir au questionnaire individuel sur support papier pour référence, sauf si cette intervention est absolument nécessaire. Autrement dit, les enregistrements électroniques obtenus après la saisie des données devraient contenir tous les renseignements nécessaires pour faire la vérification et l'imputation ultérieures.

10.3.5 Contraintes de la vérification

La vérification des données est assujettie :
- aux ressources disponibles (temps, budget et personnes),
- au logiciel disponible,
- au fardeau du répondant,
- à l'utilisation prévue des données,
- à la coordination avec l'imputation.

i. Ressources (temps, budget et personnes)

Avec une approche de vérification manuelle, le processus de vérification peut coûter cher en main-d'œuvre. Il faut :
- élaborer et documenter les règles de vérification à appliquer et les interventions nécessaires en présence d'un rejet à la vérification,
- former les vérificateurs,

- établir un mécanisme de surveillance et d'examen du travail des vérificateurs (c.-à-d. appliquer les procédures de contrôle qualitatif et d'assurance de la qualité),
- appliquer une méthode d'évaluation des répercussions de la vérification sur les données originales.

En milieu informatique, les répercussions aux volets temps, coûts et ressources pour l'élaboration au premier plan peuvent être énormes. Les tâches comprennent celles-ci :
- élaboration et documentation des règles de la vérification,
- rédaction d'un programme informatique ou adaptation d'un logiciel pour identifier les rejets à la vérification,
- mise à l'essai du programme informatique,
- vérification des données de l'enquête en exécutant le programme.

Il est important de déterminer dans les deux cas si l'investissement en vérification vaut la peine. Des ressources sont gaspillées si on applique une stratégie de vérification qui coûte cher et demande beaucoup de temps pour repérer quelques enregistrements dont les répercussions sur les résultats de l'enquête sont négligeables. Il est risqué d'autre part d'appliquer uniquement une stratégie rudimentaire de vérification pour découvrir en bout de ligne des erreurs et des incohérences majeures dans les réponses au questionnaire. Combien d'enregistrements seront probablement rejetés après l'application des règles de vérification? Quelles seront les répercussions de ces rejets sur la qualité des données obtenues? Les enregistrements ont-ils tous la même valeur? Des questions du genre sont importantes, mais il n'est pas toujours facile d'y répondre. La qualité de la conception du questionnaire, ainsi que la compréhension approfondie ou non de l'enquête chez les répondants et la qualité de la formation des intervieweurs, notamment, déterminent les réponses à ces questions.

Il est souvent préférable d'analyser les données brutes (c.-à-d. avant la vérification), surtout si l'enquête est réitérée, avant d'appliquer une stratégie de vérification. L'organisme statistique peut ainsi déterminer d'avance le nombre probable de rejets à la vérification et le genre de liens entre les questions. Il faudrait en fait considérer que la vérification est un processus continu qui n'a pas nécessairement un point de départ et d'arrivée. C'est un processus d'apprentissage qui cible l'amélioration constante de tout le déroulement de l'enquête à la longue.

ii. Logiciel

Certaines progiciels spécialisés servent à la vérification et à l'imputation des données d'un questionnaire (p. ex., le Système généralisé de vérification et d'imputation de Statistique Canada, SGVI, ou le Système canadien de contrôle et d'imputation du recensement, SCANCIR). Ces trousses peuvent permettre l'application de règles de vérification approfondies en contrepartie d'un investissement préalable assez raisonnable en conception de systèmes. D'autre part, les organismes statistiques peuvent programmer leur propre stratégie de vérification.

iii. Fardeau du répondant

L'une des conséquences de la vérification des questionnaires est la possibilité de suivi auprès des répondants pour traiter des données manquantes ou erronées. Dans la plupart des situations, le répondant est la source la plus précise d'information pour les questions du questionnaire. Le suivi est cependant un fardeau pour le répondant et il coûte cher à l'organisme statistique. Une période relativement longue peut aussi s'écouler entre l'interview et le suivi, et le répondant peut avoir oublié la réponse exacte. Ces considérations signifient que le suivi (pour traiter des rejets à la vérification) est généralement limité aux rejets à la vérification identifiés pendant la collecte ou repérés après la vérification sélective. Étant donné que le suivi après la collecte n'est en général ni pratique ni souhaitable, l'imputation est nécessaire.

iv. Utilisation prévue des données

L'utilisation des données obtenues devrait déterminer, dans une large mesure, l'ampleur de la vérification. Il n'est peut-être pas nécessaire de vérifier rigoureusement les ensembles ou éléments de données qui serviront d'abord aux examens qualitatifs, au cours desquels les décisions ne seront pas prises selon des mesures précises. Peut-être vaudrait-il mieux vérifier de plus près les ensembles ou éléments de données qui auront une importance stratégique dans la prise de décisions. De plus, dans un ensemble de données en particulier, certains éléments peuvent être beaucoup plus importants que d'autres, et il peut donc être préférable de réserver davantage de temps et de ressources pour en faire l'épuration.

D'autre part, certains enregistrements d'un ensemble de données peuvent avoir plus d'importance que d'autres et contribuer énormément aux estimations de l'enquête. C'est particulièrement le cas dans les enquêtes-entreprises où 5 % des entreprises peuvent afficher 95 % du total des gains dans une branche d'activité en particulier. Cibler les enregistrements ou les champs les plus influents est l'une des raisons d'être de la vérification sélective (Section 10.3.3) et de la détection des valeurs aberrantes (Section 10.5).

v. Coordination avec l'imputation

La vérification en soi a une valeur minimale sans une certaine intervention pour traiter des éléments rejetés après l'application des règles de vérification. S'il n'y a pas de suivi auprès du répondant, cette mesure corrective est généralement intitulée imputation. Les interventions simultanées de vérification et d'imputation sont étroitement liées. Il est donc important de considérer comment l'imputation sera faite pendant l'élaboration des spécifications de la vérification. Dans de nombreux cas, l'imputation est faite lorsque le rejet à la vérification est détecté (avant de passer à l'examen des règles suivantes). Il est préférable d'appliquer cette approche lorsque l'intervention nécessaire devient évidente, étant donné le genre de questions ou de réponses à des questions connexes. L'imputation est souvent faite cependant au cours d'une étape distincte lorsque toutes les données ont été traitées après application de toutes les règles de vérification.

10.3.6 Lignes directrices à propos de la vérification

Voici certaines lignes directrices à propos de la vérification :

i. Le personnel qui a l'expertise de la matière, de la conception des questionnaires, de l'analyse des données et d'autres enquêtes semblables devrait élaborer les vérifications.

ii. La vérification devrait être faite à plusieurs étapes de l'enquête.

iii. La vérification appliquée à chaque étape ne devrait pas contredire la vérification à une autre étape (les vérifications faites pendant la collecte et le traitement devraient être uniformes).

iv. La vérification devrait être appliquée pour obtenir de l'information sur le processus de l'enquête, soit sous forme de mesures de la qualité de l'enquête en cours ou pour suggérer des améliorations aux enquêtes ultérieures.

v. Certaines hypothèses sont formulées sur les données au début d'une enquête. Il est possible de mettre à l'épreuve la validité de ces hypothèses pendant la vérification. Il peut devenir évident, par exemple, que certaines vérifications d'étendue étaient trop strictes ou que certaines vérifications séquentielles ont donné trop souvent un rejet, et les règles de vérification se révèlent

donc inappropriées (ou le questionnaire pose certains problèmes). Cette information devrait servir à ajuster les vérifications à l'avenir (ou à améliorer la maquette du questionnaire).

vi. Il faudrait communiquer aux utilisateurs l'information sur le genre de vérifications faites et leurs répercussions sur les données de l'enquête.

vii. Il faudrait appliquer les procédures de contrôle qualitatif et d'assurance de la qualité pour minimiser et corriger les erreurs ajoutées pendant la vérification (voir l'**Annexe B - Contrôle qualitatif et assurance de la qualité**).

10.4 Imputation

L'imputation est un processus utilisé pour déterminer et attribuer des valeurs de remplacement, afin de résoudre les problèmes que suscitent les données manquantes, invalides ou incohérentes. Il faut à cette fin changer certaines des réponses et toutes les valeurs manquantes de l'enregistrement vérifié pour créer un enregistrement plausible et cohérent en soi. Certains problèmes sont corrigés auparavant lorsqu'on communique avec le répondant ou qu'on étudie le questionnaire à la main, mais, nous l'avons mentionné auparavant, il est habituellement impossible de résoudre tous les problèmes de cette façon et l'imputation est appliquée pour régler les autres rejets à la vérification.

Laisser l'utilisateur traiter les données manquantes, invalides ou incohérentes est une solution de rechange à l'imputation. Cette approche n'est pas recommandée. Si l'utilisateur décide d'ignorer ou de supprimer tous les enregistrements qui affichent des rejets à la vérification, un bon nombre de données peuvent être perdues si de nombreux enregistrements sont touchés. Si l'utilisateur essaie de remplacer les données manquantes, le résultat peut déboucher sur des estimations incohérentes de différents utilisateurs et entacher la réputation de l'organisme statistique chargé de l'enquête. L'utilisateur a accès à moins de variables que l'organisme statistique pour l'imputation et il est donc probable que l'utilisateur ne puisse traiter aussi bien les rejets à la vérification.

L'ajustement de la pondération pour les non-réponses est une approche souvent utilisée dans le cas d'une non-réponse totale ou lorsque la collecte a produit peu de données, sinon aucune, (voir le **Chapitre 7 - Estimation)**.

10.4.1 Méthodes d'imputation

Les méthodes d'imputation peuvent être réparties en deux catégories, stochastique ou déterministe. L'*imputation déterministe* signifie qu'il y a seulement une valeur imputée possible, compte tenu des données du répondant. L'*imputation stochastique* a un caractère aléatoire : si l'imputation était réitérée pour le même ensemble de données, les méthodes déterministes imputeraient la même valeur chaque fois, mais les méthodes stochastiques pourraient imputer une valeur différente chaque fois.

Les méthodes d'imputation déterministe comprennent l'imputation :
- déductive,
- de la valeur moyenne,
- par ratio-régression,
- séquentielle par donneur de l'enquête (hot-deck),
- séquentielle par donneur d'une autre source (cold-deck),
- selon le voisin le plus proche.

Chaque méthode déterministe a une contrepartie stochastique, à l'exception de l'imputation déductive. Pendant l'imputation des données quantitatives, on peut obtenir un résultat en ajoutant à la valeur imputée un résidu aléatoire tiré d'une distribution ou d'un modèle approprié. La contrepartie stochastique de l'imputation séquentielle hot-deck est l'imputation aléatoire hot-deck. L'imputation stochastique protège peut-être mieux la structure de la fréquence de l'ensemble des données et peut rétablir une variabilité plus réaliste dans les valeurs imputées que les méthodes déterministes.

À l'exception des méthodes d'imputation par donneur où un donneur peut servir à imputer toutes les données manquantes ou incohérentes pour un enregistrement destinataire, les méthodes suivantes considèrent l'imputation d'un élément à la fois.

10.4.1.1 Imputation déductive

L'application de la méthode *d'imputation déductive* permet de déduire avec certitude une valeur manquante ou incohérente. La déduction est souvent basée sur les caractéristiques des réponses données à d'autres questions du questionnaire. L'imputation déductive est habituellement faite avant d'appliquer toute autre méthode. Dans une somme de quatre articles, par exemple, si le total déclaré est 100, si deux articles valent 60 et 40 et si les deux autres sont laissées en blanc, on peut déduire que les deux valeurs manquantes sont zéro.

L'imputation doit plus souvent remplacer une valeur qui n'est pas considérée vraie en toute certitude. La matière ci-dessous donne une brève description de certaines méthodes habituelles d'imputation. Il vaut mieux regrouper des enregistrements semblables pour toutes ces méthodes, comme c'est le cas dans les ajustements de la pondération pour les non-réponses (voir le **Chapitre 7 - Estimation**). Ces regroupements sont intitulés classes d'imputation.

10.4.1.2 Imputation de la valeur moyenne

Lorsque la méthode d'imputation de la valeur moyenne est utilisée, la valeur manquante ou incohérente est remplacée par la valeur moyenne pour la classe d'imputation. Supposons, par exemple, qu'un questionnaire d'enquête sur le logement n'a pas la valeur du loyer mensuel d'un appartement. La valeur manquante peut être imputée en déterminant le loyer mensuel moyen des répondants qui ont déclaré correctement leur loyer mensuel (la classe d'imputation pourrait comprendre les répondants du même secteur géographique que celui du questionnaire qu'il faut imputer).

L'imputation de la valeur moyenne pour la donnée manquante est équivalente à l'application du même facteur d'ajustement pour la non-réponse à tous les répondants d'une même classe d'imputation. On considère que la non-réponse est uniforme et que les non-répondants ont des caractéristiques semblables à celles des répondants.

L'imputation de la valeur moyenne peut donner des estimations ponctuelles raisonnables (c.-à-d. les estimations des totaux, des moyennes, etc.), mais elle détruit les distributions et les liens multidimensionnels en créant une pointe artificielle à la moyenne de la classe. Le résultat diminue artificiellement la variance d'échantillonnage estimée des estimations définitives si des formules conventionnelles de calcul de variance sont utilisées.

L'imputation de la valeur moyenne est souvent utilisée en dernier recours pour éviter de perturber la distribution des données s'il n'y a pas d'information auxiliaire disponible ou si l'imputation cible très peu d'enregistrements.

10.4.1.3 Imputation par ratio-régression

L'information auxiliaire ou les réponses valides d'autres enregistrements sont utilisées dans l'imputation par ratio-régression pour concevoir un modèle de ratio ou de régression qui utilise les liens entre deux variables ou plus. Le modèle suivant est souvent utilisé pour l'imputation par ratio :

$$y_i = Rx_i + \varepsilon_i$$

où

y_i est la valeur de la variable y pour la i^e unité,
x_i est la valeur d'une variable x auxiliaire pour la i^e unité,
R est la pente de la droite (c.-à-d. le changement en y_i lorsque x_i augmente d'une unité),
ε_i est considérée être une variable de l'erreur aléatoire de moyenne 0 et de variance σ^2.

Autrement dit, on suppose pour ce modèle que y_i est approximativement linéaire par rapport à x_i et que les valeurs observées de y_i s'écartent de part et d'autre de cette ligne d'une grandeur aléatoire ε_I.

Les valeurs de y_i peuvent ensuite être imputées, comme suit :

$$\tilde{y}_i = \frac{\bar{y}}{\bar{x}} x_i$$

où

$\tilde{y}_i$ est la valeur imputée pour la variable y de l'enregistrement i,
$\bar{x}$ est la valeur x moyenne déclarée pour la classe d'imputation,
$\bar{y}$ est la valeur y moyenne déclarée pour la classe d'imputation.

Supposons, par exemple, qu'un questionnaire sur l'emploi, la masse salariale et les heures de travail contienne une entrée invalide pour la masse salariale, y_i, sur une période de deux semaines, mais que le nombre d'employés rémunérés, x_i, soit entré correctement et que nous sachions dans quelle branche d'activité l'entreprise est exploitée. À l'aide d'autres questionnaires de la même enquête et de la même branche d'activité (c.-à-d. la classe d'imputation) où les données sur la masse salariale et le nombre d'employés rémunérés sont déclarées correctement, il est possible de déterminer le ratio entre la masse salariale et le nombre d'employés. Ce ratio (de la masse salariale au nombre d'employés) peut ensuite être appliqué au nombre d'employés du questionnaire qu'il faut imputer, afin de déterminer une valeur pour la masse salariale.

L'hypothèse dans ce cas est que le modèle de régression ou de ratio ajusté aux questionnaires ayant des données valides (c.-à-d. qui ont passé toutes les vérifications) dans la classe d'imputation s'applique aussi bien aux questionnaires de la classe d'imputation qui ont été rejetés aux vérifications. Si cette hypothèse est fausse, il peut y avoir un biais marqué.

La présence de variables étroitement liées à la variable imputée, le degré de complexité des calculs mathématiques et le calcul restreint ou non à une classe d'imputation, ou appliqué ou non à tout l'ensemble des données, déterminent largement la précision des valeurs imputées. Cette méthode a un avantage, c'est-à-dire qu'elle peut protéger les liens entre les variables. Les estimateurs du ratio et de la régression donneront probablement aussi des valeurs imputées plus stables que de simples moyennes. Cette méthode d'imputation peut cependant ajouter artificiellement des liens à l'étape de l'analyse des données. Tout comme la plupart des autres méthodes d'imputation (à l'exception de l'imputation déductive), elle diminue la variance d'échantillonnage estimée des estimations définitives si des formules conventionnelles de calcul de la variance sont appliquées.

L'imputation de la valeur précédente, aussi intitulée imputation par report ou par report en aval, est un cas particulier d'imputation par ratio-régression, c'est-à-dire que la valeur de l'occurrence présente est imputée en ajustant la valeur de l'occurrence précédente aux fins de la croissance. Elle est souvent utilisée pour les variables quantitatives dans les applications des enquêtes-entreprises.

L'estimation par ratio et régression est expliquée plus en détail au **Chapitre 11 - Analyse des données de l'enquête**.

10.4.1.4 Imputation par donneur de l'enquête (hot-deck)

L'*imputation hot-deck* utilise l'information de l'enregistrement d'un donneur qui a habituellement passé toutes les vérifications pour remplacer des valeurs manquantes ou incohérentes d'un enregistrement destinataire. Afin de trouver un enregistrement donneur semblable à l'enregistrement destinataire, des variables liées à celles qui ont besoin d'imputation sont identifiées pour établir des classes d'imputation. L'ensemble des enregistrements dans la classe d'imputation qui ont passé toutes les vérifications est le groupe donneur pour les enregistrements de la classe d'imputation qui ont besoin d'imputation. L'imputation hot-deck peut servir à l'imputation de données qualitatives ou quantitatives, mais elle utilise généralement des variables qualitatives pour établir les classes d'imputation. Les deux principaux types d'imputation hot-deck sont l'imputation hot-deck *séquentielle* et *aléatoire*.

Dans le cas de l'imputation hot-deck séquentielle, les données font l'objet du traitement séquentiel dans la classe d'imputation, un enregistrement à la fois (c.-à-d. trié dans un certain ordre). L'imputation est faite en remplaçant l'article manquant d'un questionnaire par la valeur épurée du donneur précédent dans la classe d'imputation. L'imputation hot-deck séquentielle est une méthode d'imputation déterministe si la même méthode de tri est appliquée chaque fois. Lors de l'imputation hot-deck aléatoire, des donneurs sont sélectionnés au hasard dans la classe d'imputation. L'imputation hot-deck aléatoire est une méthode d'imputation stochastique.

Considérons l'exemple de l'imputation du statut de fumeur d'un répondant pour illustrer l'imputation hot-deck. Supposons qu'il y a deux réponses possibles : fumeur et non-fumeur. Des classes d'imputation sont établies selon le groupe d'âge et le sexe pour trouver un enregistrement donneur parce que ces variables sont liées au statut de fumeur d'une personne. Supposons que l'enregistrement ayant besoin d'imputation est celui d'une femme de la catégorie des 15 à 24 ans. L'ensemble des donneurs comprend toutes les répondantes âgées de 15 à 24 ans qui ont déclaré leur statut de fumeur. La sélection d'un donneur peut être aléatoire (c.-à-d. hot-deck aléatoire) ou séquentielle si l'on dresse la liste des donneurs et que l'on en sélectionne un (c.-à-d. hot-deck séquentielle).

Les méthodes de l'imputation par donneur ont un avantage (imputation hot-deck et par le plus proche voisin, voir la Section 10.4.1.6), c'est-à-dire que les donneurs semblables (entreprises, ménages, etc.) devraient avoir des caractéristiques semblables et la valeur imputée devrait donc être assez près de la valeur réelle. Dans l'imputation par donneur, de plus, il est habituellement possible de maintenir la distribution multidimensionnelle des données.

Il y a cependant certains inconvénients. En voici un : l'utilisation multiple du même donneur est fréquente dans l'imputation hot-deck séquentielle. L'utilisation réitérée d'un donneur peut susciter une distorsion de la distribution des données et diminuer artificiellement la variance d'échantillonnage estimée. Autre inconvénient : une bonne information auxiliaire et au moins une réponse partielle (p. ex., revenu du ménage, âge, sexe, etc.) sont nécessaires pour établir les classes d'imputation et ces données ne sont pas toujours disponibles pour les enregistrements qui ont besoin d'imputation. Il faut aussi être prudent si la classe d'imputation est petite ou si le taux de non-réponse dans la classe d'imputation est élevé parce qu'il

pourrait n'y avoir aucun donneur. (Cette constatation est vraie pour toutes les méthodes qui utilisent des classes d'imputation.)

L'imputation hot-deck *hiérarchique* peut être utilisée pour qu'il soit toujours possible de trouver un enregistrement donneur. L'imputation hiérarchique utilise plus d'un niveau de classe d'imputation. S'il n'y a pas de donneur dans la première classe d'imputation la plus détaillée, les classes sont regroupées en une structure hiérarchique suffisante pour obtenir un donneur.

L'imputation par donneur est étudiée à la Section 10.4.3.

10.4.1.5 Imputation par donneur d'une autre source (cold-deck)

L'imputation cold-deck est semblable à l'imputation hot-deck, mais cette dernière utilise les donneurs de l'enquête courante et la première utilise les donneurs d'une autre source. L'imputation cold-deck utilise souvent les données chronologiques de la même enquête réalisée auparavant ou d'un recensement. Si la sélection des donneurs est aléatoire, l'imputation est stochastique, autrement, elle est déterministe.

10.4.1.6 Imputation par voisin le plus proche

Dans les enquêtes dont les données sont largement quantitatives (p. ex., enquêtes-entreprises comprenant la déclaration des ventes et de l'inventaire), il peut être nécessaire ou préférable de trouver un enregistrement donneur par appariement avec des données quantitatives. L'imputation par le plus proche voisin est la sélection d'un enregistrement donneur selon des variables d'appariement. Lorsque cette méthode d'imputation est utilisée, le but n'est pas nécessairement de trouver un enregistrement donneur qui corresponde exactement au destinataire pour chacune des variables d'appariement. Le but est plutôt de trouver le donneur le plus près du destinataire du point de vue des variables d'appariement dans la classe d'imputation, c.-à-d. de trouver le « voisin » le plus près. Cette proximité est définie par une mesure de l'écart entre deux observations calculé à l'aide des variables d'appariement (p. ex., pour imputer un inventaire manquant, trouver le plus proche voisin du point de vue des ventes déclarées dans la classe d'imputation).

L'application des méthodes d'imputation par le plus proche voisin exige de la prudence si l'échelle des variables d'appariement est très différente (p. ex., monnaie et territoire). Il faut transformer les variables d'une certaine façon dans la plupart des cas pour normaliser l'échelle.

10.4.1.7 Imputation déterministe avec résidus aléatoires

Les méthodes déterministes appliquées aux données quantitatives peuvent devenir stochastiques en ajoutant des résidus aléatoires, par exemple, en imputant la valeur moyenne et en ajoutant un résidu aléatoire :

$$\tilde{y}_i = \bar{y} + e_i^*$$

où

$\tilde{y}_i$ est la valeur imputée pour la variable y de l'enregistrement i,

$\bar{y}$ est la moyenne pour la classe d'imputation,

e_i* est un résidu modèle aléatoire sélectionné parmi les répondants ou tiré d'une distribution.

Pour choisir un résidu e_i*, il suffit de calculer les résidus comme suit pour l'ensemble des répondants dans une classe d'imputation :

$$e_{i(r)} = y_{i(r)} - \bar{y}_r$$

où

$y_{i(r)}$ est la valeur y déclarée pour le i^e répondant,
$\bar{y}_r$ est la valeur y moyenne déclarée pour la classe d'imputation.

On peut ensuite déterminer e_i* en sélectionnant au hasard parmi toutes les valeurs de $e_{i(r)}$ dans la classe d'imputation.

Voir Kalton et Kasprzyk (1986) pour en apprendre davantage sur les approches de l'imputation stochastique.

10.4.2 Choix des valeurs à imputer

Après application d'une règle de vérification, les champs rejetés à cause de non-réponses ou de données invalides qui ne sont pas résolus par l'intermédiaire d'un suivi auprès du répondant devraient faire l'objet d'une imputation. L'imputation n'est pas recommandée pour tous les autres rejets à la vérification parce qu'il est préférable de conserver le plus possible les données du répondant. Il vaut mieux imputer un ensemble minimal de champs pour un enregistrement. La structure Fellegi-Holt (Fellegi et Holt (1976)) est l'une de ces méthodes d'identification des champs qui ont besoin d'imputation. Trois critères sont appliqués pour déterminer quels champs ont besoin d'imputation :

- il faut changer le moins d'éléments possibles des données (champs) dans chaque enregistrement pour que chacun passe toutes les vérifications,
- il faut maintenir le plus possible la structure de la fréquence du fichier des données,
- les règles d'imputation devraient découler des règles de vérification correspondantes sans spécification explicite.

Une caractéristique importante de l'approche de la vérification de Fellegi-Holt est que les règles de vérification ne sont pas spécifiques à une méthode d'imputation en particulier. Il y a d'abord, pour chaque enregistrement rejeté à la vérification, une étape de localisation d'erreurs qui permet de déterminer l'ensemble minimal de variables (champs) à imputer, ainsi que l'étendue acceptable (ou les étendues) des valeurs à imputer. Dans la majorité des applications de cette approche, un seul donneur est sélectionné dans les enregistrements qui ont passé la vérification, à l'aide de l'appariement, compte tenu d'autres variables comprises dans les vérifications, mais qui n'exigent pas d'imputation. La méthode comprend la recherche d'un seul appariement exact et elle peut être élargie pour tenir compte d'autres variables qui ne font pas explicitement partie des vérifications. Parfois, il peut n'y avoir aucun donneur convenable et il faut donc prévoir une méthode d'imputation par défaut.

Considérons deux règles de vérification d'une enquête quelconque, par exemple, une vérification « état matrimonial – âge » pour identifier ceux qui sont mariés et qui ont moins de 16 ans, et une vérification « degré de scolarité – âge » pour identifier ceux qui ont une scolarité universitaire et qui ont moins de 18 ans. Supposons qu'un enregistrement des données de l'enquête est rejeté à ces deux vérifications : une femme de dix ans est mariée et a une scolarité universitaire. L'état matrimonial et le degré de scolarité de cette personne pourraient être changés, ou simplement l'âge, pour passer les deux vérifications. La structure Fellegi-Holt recommande de changer l'âge.

10.4.3 Questions d'imputation par donneur

Il faut considérer les points suivants pour élaborer un système d'imputation par donneur (c.-à-d. imputation hot-deck, cold-deck ou par le plus proche voisin) :

i. Comment trouver un enregistrement donneur pour un destinataire?

Le but est de trouver un enregistrement donneur semblable au destinataire pour chaque destinataire. LA création des classes d'imputation mérite une étude sérieuse : il est important que les variables qui ont besoin d'imputation et celles utilisées pour sélectionner les donneurs soient étroitement liées. Il est important, pour les méthodes qui exigent l'établissement de classes d'imputation, que celles-ci soient assez larges pour que des donneurs éventuels soient disponibles en nombre suffisant, mais sans être trop larges parce que les enregistrements d'un groupe de donneurs pourraient être différents.

ii. Tous les champs d'un enregistrement destinataire devraient-ils être imputés à partir d'un seul donneur?

Il est préférable de le faire et d'utiliser tous les champs d'un enregistrement pour maintenir les distributions conjointes entre les variables. Dans une enquête sur la population active, par exemple, si la profession et le revenu personnel sont marqués pour imputation, il est évidemment avantageux d'imputer ces deux variables à l'aide du même enregistrement donneur pour maintenir le lien entre le revenu et la profession. L'imputation à l'aide d'un seul donneur a un autre avantage : étant donné que le donneur doit avoir passé avec succès toutes les vérifications, il peut servir à imputer toutes les valeurs manquantes (c.-à-d. que l'imputation est plus facile).

L'imputation par donneur pose cependant un problème : si les variables d'appariement utilisées sont trop nombreuses (p. ex., les variables utilisées pour établir des classes d'imputation dans le cas des imputations hot-deck et cold-deck), il est possible de ne trouver aucun donneur convenable. Autre problème : les variables d'appariement utilisées pour imputer un champ ne conviennent pas nécessairement à un autre, en particulier si les variables qui ont besoin d'imputation ne sont pas liées. Considérons une enquête à objectifs multiples sur la santé et supposons que la taille des personnes et le nombre de cigarettes fumées chaque jour sont marqués pour imputation. Dans ce cas, un ensemble différent de variables d'appariement pourrait être approprié pour chaque champ qui a besoin d'imputation.

Lorsque des procédures d'imputation par donneur sont appliquées, l'imputation est souvent répartie en plusieurs étapes et certains ensembles de champs sont imputés à chaque étape. Plusieurs donneurs peuvent donc être engagés pour compléter un seul enregistrement déficient. Si cette situation pose un problème, certains des principaux champs imputés peuvent servir à établir des classes d'imputation aux étapes ultérieures pour maintenir l'intégrité interne.

iii. Un enregistrement donneur peut-il servir à imputer plus d'un destinataire?

Si l'imputation de plusieurs enregistrements destinataires est faite avec le même donneur, les répercussions sur les estimations définitives de l'enquête peuvent être importantes. Limiter l'utilisation répétée d'un enregistrement donneur à une fin permet d'en élargir l'utilisation ailleurs et d'éviter la surutilisation. Si le taux de réponse dans une classe d'imputation en particulier est très faible, limiter le recours à un donneur peut déboucher sur certains appariements médiocres (c.-à-d. que l'enregistrement donneur peut être très peu semblable à celui du destinataire) et il pourrait n'y avoir aucun donneur pour certains destinataires. D'autre part, la surutilisation d'un donneur (surtout si le donneur a des caractéristiques uniques et s'il est donc très différent des autres dans la population) peut avoir des

répercussions substantielles sur les estimations de l'enquête. Si l'utilisation d'un enregistrement donneur n'est pas limitée, il devrait y avoir une méthode d'identification des enregistrements donneurs souvent utilisés. Si certains de ces enregistrements ont des champs suspects ou aberrants, il peut être nécessaire d'examiner les processus de traitement pour déterminer si les résultats définitifs de l'enquête affichent une distorsion due au processus d'imputation.

iv. Y a-t-il une intervention ultérieure si aucun donneur convenable n'est repéré pour certains destinataires?

Un enregistrement donneur peut ne pas être trouvé pour certains destinataires. Une procédure de rechange est habituellement appliquée pour ces destinataires (p. ex., imputation hot-deck ou cold-deck hiérarchique ou imputation de la valeur moyenne).

v. Les données considérées dans l'enquête sont-elles qualitatives ou quantitatives?

Certaines méthodes d'imputation sont plus appropriées pour les variables qualitatives et d'autres conviennent mieux aux variables quantitatives. Les méthodes hot-deck ont été élaborées pour traiter les données qualitatives et l'imputation selon le plus proche voisin est davantage approprié pour les données quantitatives. Les deux méthodes sont maintenant utiles dans chaque situation, y compris pour les problèmes mixtes.

10.4.4 Estimation de la variance pour les données imputées

Toutes les méthodes d'imputation présentées donnent une seule valeur imputée pour chaque valeur manquante ou incohérente. Elles altèrent, jusqu'à un certain point, la distribution originale des valeurs pour une variable et peuvent donner des estimations de la variance inappropriées lorsque des estimateurs standard de variance sont utilisés. Le résultat peut donner des intervalles de confiance trop étroits et des rejets d'hypothèse nulle erronés. La portée de la distorsion varie considérablement selon l'ampleur de l'imputation faite et la méthode appliquée.

Lorsque l'imputation est faite, s'il n'y a pas d'autres erreurs non dues à l'échantillonnage, la variance d'une estimation a deux composantes : l'une est due à l'échantillonnage (la variance d'échantillonnage) et l'autre, à l'imputation (la variance due à l'imputation). La composante variance d'échantillonnage est habituellement sous-estimée en présence de données imputées parce que les formules traditionnelles sont basées sur un taux de réponse de 100 %. Les méthodes d'imputation stochastiques ajoutent une certaine perturbation à l'ensemble des données achevées, et c'est un avantage. Si l'imputation stochastique est utilisée, la variance d'échantillonnage d'une estimation peut donc être correctement estimée la plupart du temps à l'aide des méthodes traditionnelles. La variance d'imputation doit cependant être estimée quand même pour déterminer la variance totale de l'estimation.

Il est important d'estimer les composantes échantillonnage et imputation de la variance totale, non seulement pour formuler des inférences exactes, mais aussi pour déterminer l'importance relative de la variance d'échantillonnage et de la variance d'imputation. Cette mesure peut aider à informer les utilisateurs sur la qualité des données et aider à attribuer des ressources d'enquête entre la taille de l'échantillon et les processus de vérification – d'imputation.

Proposition de Rubin (1987), l'imputation multiple est une méthode qui permet de considérer ce problème si l'on impute « correctement » plusieurs, disons m, fois chaque valeur ayant besoin d'imputation (voir Rubin (1987) ou Binder et Weimin (1996) pour obtenir une définition de l'imputation « correcte »). Il est possible d'obtenir m estimations pour l'article à partir de l'ensemble des données achevées. Une seule

estimation combinée en est tirée, ainsi qu'une estimation de la variance qui exprime l'incertitude au sujet de la valeur à imputer. L'imputation multiple exige cependant davantage de travail pour le traitement des données, la sauvegarde et le calcul des estimations.

Les méthodes importantes d'estimation de la variance ont été élargies pour englober des fichiers contenant les données imputées dans le cas de l'imputation simple. Les approches sont décrites dans Särndal (1992), Rao et Shao (1992), Rao et Sitter (1995) et Gagnon et coll. (1996). Une comparaison des méthodes est présentée dans Lee, Rancourt et Särndal (1994, 2001).

10.4.5 Lignes directrices à propos de l'imputation

L'imputation peut améliorer la qualité des données définitives, mais il faut choisir prudemment une méthode d'imputation appropriée. L'imputation est un risque parce qu'elle peut détruire les données déclarées pour créer des enregistrements qui correspondent à des modèles préconçus qui peuvent se révéler ultérieurement inexacts. L'enquête, ses objectifs, l'information auxiliaire disponible et le genre d'erreur déterminent l'à-propos de la méthode d'imputation.

Voici certaines lignes directrices pour l'imputation :

i. Les enregistrements imputés devraient ressembler de près à l'enregistrement rejeté à la vérification. Cet objectif est habituellement atteint en imputant le nombre minimal de variables pour sauvegarder le plus de données possible du répondant. L'hypothèse sous-jacente (qui n'est pas toujours vraie en pratique) est qu'un répondant fera probablement une erreur ou deux au lieu d'en faire plusieurs.

ii. Une bonne imputation comprend une piste de vérification aux fins de l'évaluation. Les valeurs imputées devraient être indiquées et les méthodes et les sources d'imputation, clairement identifiées. Les valeurs imputées et non imputées des champs de l'enregistrement devraient être retenues, afin d'évaluer l'ampleur et les répercussions de l'imputation.

iii. Les enregistrements imputés devraient passer toutes les vérifications.

iv. Les méthodes d'imputation doivent être choisies avec soin, compte tenu du genre de données à imputer.

v. La méthode d'imputation devrait favoriser le plus possible la diminution du biais de non-réponse et le maintien des liens entre les éléments (c.-à-d. qu'il faut évaluer si le modèle sous-jacent à l'imputation est adéquat).

vi. Le système d'imputation devrait être conçu, spécifié, programmé et mis à l'essai d'avance.

vii. Le processus devrait être automatisé, objectif, reproductible et efficace.

viii. Le système d'imputation devrait être en mesure de traiter toute caractéristique des champs manquants ou incohérents.

ix. Si une méthode d'imputation par donneur est utilisée, l'enregistrement imputé devrait ressembler de près aux donneurs sélectionnés. La combinaison des réponses imputées et non imputées pour l'enregistrement imputé aura ainsi plus de chance de passer les vérifications et d'être plausible.

10.4.6 Évaluation des procédures d'imputation

La taille de l'enquête et le budget déterminent la somme de travail à accomplir pour mesurer les répercussions de l'imputation. Les utilisateurs des données de l'enquête devraient cependant toujours avoir certains renseignements élémentaires sur l'ampleur de la modélisation ou de l'estimation par imputation des données de l'enquête. Lors de l'évaluation de la procédure d'imputation, les préoccupations les plus pertinentes sont le biais et la variance d'imputation des estimations de l'enquête.

Si le budget de l'enquête est suffisamment élevé, l'une des options est de faire une étude complète des répercussions de l'imputation et d'examiner les estimations de l'enquête avec et sans imputation. Les écarts importants devraient être examinés et essayer de découvrir un biais éventuel dû à l'imputation.

Si cette mesure est impossible, il faudrait au moins surveiller l'imputation pour informer les utilisateurs de l'importance de l'imputation et préciser où elle a été faite. Il peut être utile, à la fin de l'imputation, de produire les résultats suivants (certains sont spécifiques à une méthode en particulier) :

- le nombre d'enregistrements imputés (c.-à-d. le nombre d'enregistrements destinataires),
- le nombre d'imputations dans chaque champ et la méthode utilisée,
- le nombre d'enregistrements qui peuvent servir de donneurs,
- le nombre d'enregistrements utilisés en fait comme donneurs et le nombre de destinataires ainsi imputés,
- une liste (ou un fichier) énumérant les donneurs utilisés pour chaque bénéficiaire (pour repérer les sources des enregistrements imputés inhabituels),
- une liste de tous les enregistrements rejetés à l'imputation (p. ex., parce qu'aucun donneur n'a été découvert).

Remarquez que l'information ci-dessus est utile pour la refonte d'une enquête ou la conduite d'une enquête semblable. Ces renseignements peuvent aider à améliorer le système de vérification et d'imputation, le questionnaire de l'enquête et les procédures de collecte. Si la réponse à une question a un taux d'imputation élevé, par exemple, la question peut être mal formulée (et la qualité des données peut être médiocre).

10.5 Identification et traitement des valeurs aberrantes

L'identification des valeurs aberrantes peut être considérée comme un genre de vérification parce que les enregistrements suspects sont identifiés. Au **Chapitre 7 - Estimation**, ***on a défini une valeur aberrante comme une observation ou un sous-ensemble d'observations qui semble(nt) incohérente(s) par rapport aux autres données de l'ensemble***. Il faudrait aussi faire la distinction entre les observations extrêmes et influentes. Une observation est influente si la combinaison de la valeur déclarée et la pondération définitive de l'enquête ont une grande influence sur l'estimation. Une valeur extrême n'est cependant pas nécessairement influente, et vice versa.

Il est possible de faire la distinction entre des valeurs aberrantes unidimensionnelles (à une variable) et multidimensionnelles (à plusieurs variables). Une observation est une valeur aberrante unidimensionnelle si elle est aberrante par rapport à une seule variable. Une observation est une valeur aberrante multidimensionnelle si elle est aberrante par rapport à deux variables ou plus. Il est peut-être facile, par exemple, de trouver une personne mesurant deux mètres *ou* une personne pesant 45 kg, mais quelqu'un qui mesure deux mètres *et* pèse seulement 45 kg est un exemple de valeur aberrante multidimensionnelle.

Chaque enquête comprend des valeurs aberrantes pour à peu près chaque variable d'intérêt. De nombreuses raisons expliquent les valeurs aberrantes :

i. Il y a des erreurs dans les données (p. ex., erreurs de saisie des données).

ii. On peut considérer que les valeurs aberrantes sont tirées d'un autre modèle ou d'une autre distribution. Vous pouvez penser, par exemple, que la majorité des données sont tirées d'une distribution normale, mais que les valeurs aberrantes peuvent provenir d'une distribution exponentielle.

iii. La valeur aberrante peut être due à la variabilité inhérente des données. Une valeur peut sembler suspecte, mais découler simplement de la variabilité inhérente de l'ensemble des données, autrement dit, il peut s'agir d'une observation extrême, mais légitime, de la distribution. La situation est possible si la population est asymétrique et c'est souvent le cas dans les enquêtes-entreprises. La répartition des ventes selon la taille de l'entreprise, par exemple, est typiquement asymétrique, c.-à-d. que quelques très grandes entreprises affichent souvent la majeure partie des ventes dans l'ensemble.

10.5.1 Identification des valeurs aberrantes

Les méthodes de détection des valeurs aberrantes les plus populaires sont les méthodes à une variable parce qu'elles sont plus simples que les méthodes à plusieurs variables. Les valeurs aberrantes sont habituellement détectées en mesurant leur distance relative par rapport au centre des données. Si $y_1, y_2, \ldots, y_n$ sont les données de l'échantillon observé, par exemple, et m et s sont des mesures de la tendance centrale et de l'étendue des données, respectivement, la distance relative, d_i, de y_i par rapport au centre des données peut être déterminé ainsi :

$$d_i = \frac{|y_i - m|}{s}$$

Si d_i surpasse une valeur limite déterminée, l'observation est alors considérée comme une valeur aberrante.

D'autre part, un intervalle de tolérance peut être attribué comme suit :

$$(m - c_L s, m + c_U s)$$

où c_L et c_U sont les valeurs limites inférieure et supérieure prédéterminées. Si la population est asymétrique, des valeurs inégales de c_L et de c_U sont utilisées. Les observations hors de cette intervalle sont déclarées valeurs aberrantes.

La moyenne et la variance de l'échantillon sont les statistiques les plus souvent utilisées pour estimer le centre et l'étalement des données. Étant donné qu'elles sont sensibles aux valeurs aberrantes cependant, elles sont un choix médiocre pour la détection de ces valeurs. La moyenne de l'échantillon se déplace vers les valeurs aberrantes, par exemple, si elles sont en grappes d'un côté et les valeurs aberrantes augmentent énormément la variance de l'échantillon. Les valeurs relatives de la distance de certaines valeurs aberrantes peuvent donc sembler négligeables et la procédure de détection peut échouer. Ce problème est intitulé *effet de dissimulation*.

Voilà pourquoi l'une des méthodes les plus populaires de détection des valeurs aberrantes est la *méthode par quartile* qui utilise la médiane pour estimer le centre et les étendues quartiles pour estimer l'étalement des données pondérées parce que ces statistiques résistent mieux (c.-à-d. qu'elles ne réagissent pas) aux valeurs aberrantes. Les quartiles répartissent les données en quatre parties : 25 % des données simples

sont inférieures au premier quartile, $q_{0,25}$, 50 % des données simples sont inférieures au deuxième quartile (ou la médiane), $q_{0,5}$, et 75 % des données simples sont inférieures au troisième quartile, $q_{0,75}$. (La médiane et les étendues des quartiles sont considérées davantage au **Chapitre 11 - Analyse des données de l'enquête**).

Les étendues des quartiles inférieur et supérieur, h_L et h_U, sont définies comme suit :

$$h_L = q_{0,5} - q_{0,25}$$
$$h_U = q_{0,75} - q_{0,5}$$

L'intervalle de tolérance devient donc :

$$(q_{0,5} - c_L h_L, q_{0,5} + c_U h_U)$$

et certaines valeurs déterminées sont attribuées à c_L et c_U en examinant les données précédentes ou selon l'expérience acquise. Toute observation hors de cet intervalle est considérée comme une valeur aberrante.

Voir Barnett et Lewis (1995) pour obtenir davantage d'information sur les méthodes de détection des valeurs aberrantes.

10.5.2 Traitement des valeurs aberrantes

Les valeurs aberrantes détectées à l'étape de la vérification dans le processus de l'enquête peuvent être traitées de différentes façons. Dans le contexte d'un système de vérification manuel, les valeurs aberrantes éventuelles sont examinées, les répondants relancés, et les données aberrantes sont modifiées si elles se révèlent en erreur. Dans un contexte automatisé, les valeurs aberrantes sont souvent imputées. Dans les cas où les données aberrantes n'ont pas d'influence sur les résultats finaux, il n'y a pas nécessité de traitement particulier.

Les valeurs aberrantes non traitées à la vérification peuvent être considérées à l'estimation. Ignorer simplement les valeurs aberrantes non traitées peut donner des estimations médiocres et accroître la variance d'échantillonnage des estimations. Attribuer une pondération de un à une valeur aberrante (pour diminuer ses répercussions sur les estimations) peut biaiser les résultats. Le but du traitement des valeurs aberrantes est d'en diminuer les répercussions sur la variance d'échantillonnage de l'estimation sans trop biaiser les résultats.

Les approches suivantes peuvent être appliquées pour traiter les valeurs aberrantes pendant l'estimation :
- changer la valeur,
- changer la pondération,
- utiliser une estimation robuste.

i. Changement de valeur

La winsorisation est un exemple de traitement d'une valeur extrême. La winsorisation est le recodage des k valeurs les plus grandes.

Le lecteur se rappellera que, dans un échantillonnage aléatoire simple (si le taux de réponse est de 100 %), l'estimateur habituel non biaisé du total de la population Y est obtenu ainsi :

$$\hat{Y} = \frac{N}{n} \sum_{i=1}^{n} y_i$$

où i est la i^e unité d'un échantillon de taille n.

Supposons que y_i , $i=1,2, \ldots, n$ sont les valeurs ordonnées de y_i dans un échantillon de taille n d'une population de taille N et que les k valeurs les plus grandes sont considérées être des valeurs aberrantes, l'estimateur unilatéral windsorisé d'ordre k est défini en remplaçant ces valeurs aberrantes par la n-k^e plus grande valeur, y_{n-k}, c.-à-d. :

$$\hat{Y}_W = \frac{N}{n}\left(\sum_{i=1}^{n-k} y_i + k y_{n-k}\right)$$

On remarque que la winsorisation est plutôt appliquée dans les situations à une variable et rarement donc dans les enquêtes-échantillons à plusieurs variables.

ii. Changement de pondération

La diminution des pondérations des valeurs aberrantes pour en amenuiser les répercussions est une autre option. Attribuer une valeur de zéro ou de un à la pondération d'une valeur aberrante est un exemple. Cette mesure est rarement appliquée à cause de ses répercussions marquées sur les estimations, en particulier pour les populations asymétriques. Elle peut donner un biais grave, habituellement une sous-estimation. Si deux grandes entreprises affichent la majorité des ventes au détail, par exemple, et si l'une des deux est identifiée comme une valeur aberrante, le retrait de cette entreprise des estimations donnera une sous-estimation importante du total des ventes au détail. Plusieurs estimateurs ayant des pondérations réduites pour les données aberrantes ont été proposés, voir Rao (1970), Hidiroglou et Srinath (1981).

iii. Estimateurs robustes

En théorie classique de l'estimation, l'estimateur d'un paramètre de population est axé sur la supposition d'une certaine distribution. On suppose habituellement que la distribution d'échantillonnage est normale pour l'estimateur (voir le **Chapitre 7 - Estimation** pour la définition de distribution d'échantillonnage). Les estimateurs habituels de la moyenne et de la variance de l'échantillon sont optimaux en situation normale. Ces estimateurs sont cependant extrêmement sensibles aux valeurs aberrantes. Des estimateurs robustes sont moins sensibles aux hypothèses sur la distribution. La médiane est plus robuste que la moyenne, par exemple, les étendues interquartiles sont plus robustes que l'estimateur habituel de la variance. De nombreux estimateurs robustes complexes ont été proposés au cours des années, y compris les estimateurs M de Huber, Huber (1964).

Voir Kish (1965), Barnett et Lewis (1995), Rousseeuw et Leroy (1987), Lee et coll. (1992) ou Lee (1995) pour obtenir davantage d'information sur les estimateurs robustes et la détection des valeurs aberrantes en général. Voir le **Chapitre 11 - Analyse des données de l'enquête** pour obtenir davantage d'information sur la moyenne comparativement à la médiane.

10.6 Production des résultats – création d'une base de données

Après le codage, la saisie des données, la vérification, l'imputation et la détection des valeurs aberrantes, les données sont relativement prêtes pour l'estimation, l'analyse et la diffusion. Avant de procéder cependant, il faut déterminer la mise en forme pour la sauvegarde des données. Une base de données et un fichier non hiérarchique sont les deux principaux choix. La disposition bidimensionnelle informatisée des enregistrements et de leurs valeurs correspondantes donne un fichier non hiérarchique. Il est facile de le télécharger d'une plateforme à l'autre et il peut être consulté à l'aide d'un logiciel tableur ou statistique.

La majorité des logiciels statistiques doivent avoir des données sauvegardées en mise en forme spéciale pour faciliter le traitement rapide, et c'est le principal inconvénient d'un fichier non hiérarchique. Avec un tel fichier, cette mise en forme particulière est continuellement modifiée, une inefficacité inutile. Si les données sont sauvegardées sous forme de base de données, il est possible d'utiliser certains logiciels statistiques et de base de données sans nécessairement créer de nouveau le fichier. Les demandes peuvent être entrées directement dans la base de données. Le choix de format de base de données peut cependant restreindre le choix du logiciel statistique et d'exploitation de base de données qui peuvent servir à la totalisation et à l'analyse. Il vaut peut-être mieux créer un fichier non hiérarchique et plusieurs fichiers différents avec les résultats de l'enquête dans la base de données.

Lorsque le format de sauvegarde des données a été sélectionné, les poids finaux (pour l'estimation) sont calculés et les totalisations prévues sont faites (voir le **Chapitre 7 - Estimation** pour obtenir une description de la méthode de calcul des poids finaux). Les programmes informatiques sont habituellement rédigés pour calculer les pondérations et produire les totalisations. Vous pouvez aussi faire une analyse des données plus perfectionnée. Il faut examiner les données avant de les diffuser pour vérifier si elles respectent les critères de confidentialité des répondants. Ce processus intitulé contrôle de la divulgation peut déboucher sur la suppression de certaines données de l'enquête. Voir le **Chapitre 12 - Diffusion des données** pour obtenir davantage d'information sur l'analyse des données et le contrôle de la divulgation.

10.7 Traitement manuel ou automatisé

Le traitement de tous les volets, ou presque, d'une enquête était auparavant manuel. Les ordinateurs permettent maintenant le traitement automatisé des données.

Les avantages de l'automatisation du codage et de la saisie des données, de la lecture optique des caractères, des méthodes de collecte des données assistées par ordinateur et du codage préalable du questionnaire ont déjà été considérés. Les arguments en faveur de l'utilisation des ordinateurs pour la collecte des données s'appliquent aussi au traitement. L'expérience révèle qu'en général les ordinateurs sont bien meilleurs que les personnes pour traiter un nombre élevé de renseignements. L'automatisation peut améliorer la qualité des données à tous les points de vue, et en particulier la rapidité : elle donne des résultats plus rapidement et exige moins de ressources. Elle garantit aussi que les procédures appliquées (p. ex., la vérification et l'imputation) sont uniformes et elle diminue les erreurs non dues à l'échantillonnage. Elle permet aussi d'appliquer des méthodes plus complexes (p. ex., pour la vérification, l'imputation, le codage, le contrôle qualitatif, etc.), de suivre le traitement et de faire rapport sur chacune de ses étapes (p. ex., le nombre de vérifications et d'imputations faites). L'automatisation facilite aussi la surveillance et le contrôle qualitatif du traitement.

L'automatisation a cependant certains inconvénients, par exemple :

- la formulation de spécifications pour chaque système qui sera automatisé et l'élaboration d'un programme informatique pour chaque procédure (p. ex., l'imputation) sont nécessaires et peuvent demander beaucoup de temps,
- la formation des opérateurs qui utiliseront le logiciel est obligatoire,
- le codage, la vérification et l'imputation automatisés n'englobent pas les renseignements supplémentaires de l'opérateur.

Il est sage d'automatiser les procédures le plus possible, malgré ces inconvénients. L'investissement supplémentaire en temps au départ devient très avantageux plus tard pendant le processus de l'enquête (surtout si l'enquête est réitérée). Il faudra au moins toujours saisir les données, et en faire la pondération et l'estimation, à l'ordinateur. L'automatisation permet l'uniformité qui est importante pour obtenir des

résultats précis et mesurables. Tirer avantage des systèmes et processus existant, des systèmes automatisés de codage, etc., est aussi une bonne décision.

10.8 Sommaire

Le traitement est une importante activité de l'enquête qui convertit les réponses des questionnaires en une mise en forme qui convient à l'analyse des données et à la totalisation. Le traitement coûte cher, demande beaucoup de temps et de ressources, et a des répercussions sur la qualité définitive des données. L'automatisation peut en augmenter l'efficacité et améliorer la qualité définitive des données.

Le traitement commence normalement par une épuration préliminaire du questionnaire, suivie du codage et de la saisie des données. L'étape suivante est habituellement une vérification plus détaillée pour identifier les données manquantes ou incohérentes, et ensuite, l'imputation est faite pour intégrer des substituts plausibles à ces valeurs. La détection des valeurs aberrantes est aussi utile pour identifier les valeurs suspectes. Lorsque les données sont complètes, convergentes et valides, elles sont habituellement sauvegardées dans une base de données.

Bibliographie

Bankier, M., M. Lachance et P. Poirier. 1999. A Generic Implementation of the Nearest neighbour imputation method. *Proceedings of the Survey Research Methods Section*. American Statistical Association. 548-553.

Barnett, V. et T. Lewis. 1995. *Outliers in Statistical Data*. John Wiley and Sons, Chichester.

Binder, D. et S. Weimin. 1996. Frequency Valid Multiple Imputation for Surveys with a Complex Design. *Proceedings for the Section on Survey Research Methods of the American Statistical Association,* 1: 281-286.

Boucher, L, J.-P. S. Simard et J.-F. Gosselin. 1993. Macro-Editing, a Case Study: Selective Editing for the Annual Survey of Manufacturers Conducted by Statistics Canada, *Proceedings of the International Conference on Establishment Surveys.* American Statistical Association. Virginia.

Brick, J.M. et G. Kalton. 1996. Handling Missing Data in Survey Research. *Statistical Mathematics in Medical Research,* 5: 215-238.

Chambers, R.L. 1986. Outlier Robust Finite Population Estimation. *Journal of the American Statistical Association,* 81: 1063-1069.

Cox, B.G., D. A. Binder, B.N. Chinnappa, A. Christianson, M.J. Colledge et P.S. Kott, Éds. 1995. *Business Survey Methods.* John Wiley and Sons, New York.

Dielman, L. et M.P. Couper. 1995. Data Quality in a CAPI Survey: Keying Errors. *Journal of Official Statistics,* 11(2): 141-146.

Dolson, D. 1999. *Imputation Methods*. Statistics Canada.

Fay, R.E. 1996. Alternative Paradigms for the Analysis of Imputed Survey Data. *Journal of the American Statistical Association,* 91: 490-498.

Fellegi, I.P. et D. Holt. 1976. A Systematic Approach to Automatic Edit and Imputation. *Journal of the American Statistical Association,* 71: 17-35.

Gagnon, F., H. Lee, E. Rancourt and C.E. Särndal. 1996. Estimating the Variance of the Generalized Regression Estimation in the Presence of Imputation for the Generalized Estimation System. *Proceedings of the Survey Methods Section*. Statistical Society of Canada. 151-156.

Granquist, L. 1984. On the Role of Editing. *Statistisk tidskrift,* 2: 105-118.

Granquist, L. et J. Kovar. 1997. Editing of Survey Data: How Much is Enough? In Lyberg, L., et al., eds. 1997. *Survey Measurement and Process Quality*. John Wiley and Sons, New York. 415-436.

Hidiroglou, M.A. 1999. Notes de cours *Methods for Designing Business Survey*.

Hidiroglou, M.A. 1999. Notes de cours *Methods for Designing Business Survey*. Commandité par l'AISE, 52è session de l'IIS, Université de Jyvaskylä, Finlande.

Hidiroglou, M.A. et J.-M. Berthelot. 1986. Contrôle statistique et imputation dans les enquêtes-entreprises périodiques, *Techniques d'enquête*, 12(1): 79-89.

Hidiroglou, M.A. et K.P. Srinath. 1981. Some Estimators of a Population Total Containing Large Units. *Journal of the American Statistical Association,* 78: 690-695.

Huber, P.J. 1964. Robust Estimation of a Location Parameter. *Annals of Mathematical Statistics,* 35: 73-101.

Kalton, G. et D. Kasprzyk. 1982. Imputation for Missing Survey Responses. *Proceedings of the Section on Survey Research Methods*. American Statistical Association. 23-31.

Kalton, G. et D. Kasprzyk, D. 1986. Le traitement des données d'enquête manquantes. *Techniques d'enquête*. 12(1): 1-18.

Kish, L. 1965. *Survey Sampling*. John Wiley and Sons, New York.

Kovar, J.G., J. MacMillan et P. Whitridge. 1988. *Système généralisé de vérification et d'imputation – Aperçu et stratégie (Mis à jour en février 1991).* Statistique Canada. BSMD-88-007 E/F.

Latouche, M. et J.-M. Berthelot. 1992. Use of a Score Function to Prioritize and Limit Recontacts in Editing Business Surveys. *Journal of Official Statistics*, 8(3): 389-400.

Lee, H., E. Rancourt et C.E. Särndal. 1994. Experiments with Variance Estimation from Survey Data with Imputed Values. *Journal of Official Statistics,* 10(3): 231-243.

Lee, H., E. Rancourt et C.E. Särndal. 2001. Variance Estimation from Survey Data under Single Value Imputation. *Survey Nonresponse*. John Wiley and Sons, New York.

Lee, H. 1995. Outliers in Business Surveys. Dans *Business Survey Methods.* Cox, B.G., D. A. Binder, B.N. Chinnappa, A. Christianson, M.J. Colledge et P.S. Kott, Éds. John Wiley and Sons. New York. 503-526.

Lyberg, L. et P. Dean. 1992 Automated Coding of Survey Responses: An International Review. Presented at the Conference of European Statisticians. Washington, D.C.

Moser, C.A. et G. Kalton. 1971. *Survey Methods in Social Investigation*. Heinemann Educational Books Limited, London.

Raj, D. 1972. *The Design of Sample Surveys*. McGraw-Hill Series in Probability and Statistics, New York.

Rancourt, E., H. Lee et C.E. Särndal 1993. *Variance Estimation Under More than One Imputation Method*. Proceedings of the International Conference on Establishment Surveys, American Statistical Association, 374-379.

Rao, C.R. 1970. Estimation of Heteroscedastic Variances in Linear Models. *Journal of the American Statistical Association,* 65: 161-172.

Rao, J.N.K. et J. Shao. 1992. Jackknife Variance Estimation with Survey Data under Hot-deck Imputation. *Biometrika,* 79: 811-822.

Rao, J.N.K. et R.R. Sitter. 1995. Variance Estimation under Two-Phase Sampling with Application to Imputation for Missing Data. *Biometrika,* 82: 453-460.

Rao, J.N.K. 1996. On Variance Estimation with Imputed Survey Data. *Journal of the American Statistical Association,* 91: 499-506.

Rousseeuw, P.J. et A.M. Leroy. 1987. *Robust Regression and Outlier Detection*. John Wiley and Sons, New York.

Rubin, D.B. 1987. *Multiple Imputation for Nonresponse in Surveys*. John Wiley and Sons, New York.

Rubin, D.B. 1996. Multiple Imputation after 18+ Years. *Journal of the American Statistical Association,* 91: 473-489.

Sande, I.G. 1979. A Personal View of Hot-deck Imputation Procedures. *Survey Methodology,* 5(2): 238-258.

Sande, I.G. 1982. Imputation in Surveys: Coping with Reality. *The American Statistician,* 36(3). Part 1: 145-152.

Särndal, C.E. 1992. Méthodes pour estimer la précision des estimations d'une enquête ayant fait l'objet d'une imputation. *Techniques d'enquête*, 18(2): 257-268.

Särndal, C.E., B. Swensson et J. Wretman. 1992. *Model Assisted Survey Sampling*. Springer Verlag, New York.

Shao, J. et R.R. Sitter. 1996. Bootstrap for Imputed Survey Data. *Journal of the American Statistical Association,* 94: 254-265.

Statistique Canada. 1990. Notes de cours, *Cours de base sur les enquêtes*.

Statistique Canada. 1998. *Lignes directrices concernant la qualité*. 12-539-XIF.

Statistique Canada. 1998. Notes de cours STC416 "*Les Enquêtes de A à Z*".

Wenzowski, M.J. 1988. Advances in Automated Coding and Computer-Assisted Coding Software at Statistics Canada. Proceedings of the 1996 Annual Research of the U.S. Census Bureau.

Yung, W. et J.N.K. Rao. 2000. Jackknife Variance Estimation under Imputation for Estimators using Poststratification Information. *Journal of the American Statistical Association,* 95: 903-915.

Chapitre 11 - Analyse des données de l'enquête

11.0 Introduction

L'analyse des données comprend le résumé des données et l'interprétation de leur signification pour donner des réponses claires aux questions qui ont motivé l'enquête. Il faut souvent interpréter des tableaux et diverses mesures de récapitulation, par exemple, des distributions de fréquences, des moyennes et des étendues de valeurs, ou des analyses plus approfondies peuvent être faites. L'analyste voudra peut-être décrire seulement les unités observées et, dans ce cas, tous les outils de la statistique élémentaire et intermédiaire sont disponibles (tableaux, diagrammes et graphiques, mesures élémentaires de la position et de dispersion, modélisation de base, modèles de classification, etc.). L'analyste voudra plus souvent décrire la population et vérifier les hypothèses formulées à ce sujet, et il faudra correctement tenir compte du plan d'échantillonnage pendant l'analyse.

L'objectif de ce chapitre est de considérer l'analyse des statistiques sommaires (distributions des fréquences, mesures de la tendance centrale et mesures de l'étalement), de présenter des méthodes plus analytiques qui comprennent l'analyse par inférence pour les échantillons probabilistes et de déterminer comment ces mesures s'appliquent à des plans d'échantillonnage simples ou complexes.

Le chapitre commence avec l'analyse de données d'enquête simples, sans stratification, grappes, ajustements aux poids, etc. L'analyse des données d'enquête plus complexes devient rapidement compliquée et il vaut mieux consulter un spécialiste. La matière plus approfondie dans ce chapitre exige des cours de premier cycle en statistique pour bien comprendre et elle commence à la Section 11.3.2.2.1**.**

11.1 Paramètres, estimations, erreur d'échantillonnage

Tout d'abord, rappelons certaines définitions présentées au **Chapitre 6 - Plans d'échantillonnage.** ***Un paramètre est une caractéristique de la population que le client ou l'utilisateur des données est intéressé à estimer***, par exemple, la *moyenne* de la population, $\overline{Y}$. ***Un estimateur est une formule de calcul de l'estimation du paramètre et l'estimation est la valeur de l'estimateur déterminée à l'aide des données de l'échantillon réalisé.*** Les estimations calculées à partir d'échantillons différents sont différentes l'une de l'autre. ***La distribution d'échantillonnage d'un estimateur est la distribution de toutes les valeurs différentes que l'estimateur peut avoir pour tous les échantillons possibles du même plan d'échantillonnage.*** L'estimateur et le plan d'échantillonnage déterminent cette distribution. Un estimateur non biaisé ou approximativement non biaisé et la distribution de l'échantillonnage concentrée le plus près possible de la moyenne (c.-à-d. que l'erreur d'échantillonnage est petit) sont deux caractéristiques souhaitables. Dans le cas des échantillons probabilistes, cette erreur peut être mesurée, habituellement en estimant la variance d'échantillonnage, l'erreur-type, le coefficient de variation ou la marge d'erreur.

11.2 Genres de données

Une enquête permet la collecte d'un éventail de caractéristiques ou variables. Nous avons mentionné au **Chapitre 7 - Estimation** qu'une enquête unique peut comprendre des variables qualitatives et quantitatives. Les variables qualitatives sont codées (nominales) et les variables quantitatives indiquent un nombre. D'autres décompositions sont possibles : les variables qualitatives peuvent être nominales ou ordinales et les variables quantitatives peuvent être discrètes ou continues.

i. Variables nominales

Une variable nominale est une série de catégories qui sont simplement des étiquettes ou des noms sans lien mathématique entre eux. On ne peut affirmer qu'une catégorie en particulier est plus grande qu'une autre, égale ou inférieure à une autre, par exemple, si le *genre de sport* est la variable nominale, *cricket* < *soccer* n'a aucun sens.

ii. Variables ordinales

Une variable ordinale est une série de catégories ordonnées ou classées selon une échelle ou un continuum déterminé, et une catégorie en particulier peut précéder ou suivre une autre. Les différences entre les catégories ne sont pas nécessairement équivalentes. Des nombres peuvent être attribués aux variables ordinales, mais uniquement pour ordonner les matières, et les additionner ou faire d'autres opérations arithmétiques est inapproprié. Voici un exemple de données ordinales : *vraiment d'accord, d'accord, ni pour ni contre, pas d'accord, vraiment pas d'accord*. D'autres exemples : la collecte de l'âge à l'aide d'une question fermée, par exemple, *moins de 15 ans, de 15 ans à 34 ans, …, 75 ans ou plus*, ou la tension artérielle qualifiée de *basse*, *normale*, *élevée*.

iii. Variables discrètes

Une variable discrète est une variable quantitative ayant des valeurs dénombrables. Voici un exemple de ce genre de variables : une variable dont les valeurs possibles sont entières et il ne peut y avoir de valeur intermédiaire entre deux valeurs entières. La taille d'un ménage peut être, par exemple, 1, 2, 3, et des valeurs comme 1,5 ou 4,75 sont impossibles. Il n'est cependant pas nécessaire que les variables discrètes soient entières : un autre exemple de variable discrète est la taille des souliers qui peut être 6, $6^1/_2$, 7, $7^1/_2$, etc., mais $6^3/_4$ est impossible.

iv. Variables continues

Une variable continue est une variable quantitative dont toute valeur dans une certaine étendue est possible (contrairement à une variable discrète dont certaines valeurs en particulier seulement sont possibles). La taille et le poids sont donc des variables continues, mais le nombre de buts d'une équipe de hockey est une variable discrète. Il est possible pour une personne d'avoir n'importe quelle taille, jusqu'à un certain point, par exemple, 1,68 mètre, mais une équipe de hockey ne peut compter 2,3 buts parce que le nombre de but est discret et entier. Remarquez que les variables continues peuvent être transformées en variables nominales, par exemple, les mesures de la tension artérielle peuvent être qualifiées de basse, normale ou élevée.

Le type de données détermine le genre de procédures analytiques qui peuvent être appliquées et la question est expliquée aux sections suivantes.

11.3 Mesures de récapitulation

Dans *Analysis of Complex Surveys* (Analyse d'enquêtes complexes) (1989), Skinner, Holt et Smith affirment que les données d'une enquête-échantillon peuvent servir à des fins descriptives ou analytiques. Les utilisations descriptives ciblent l'estimation des mesures récapitulatives de la population, par exemple les moyennes et les fréquences, mais les utilisations analytiques surpassent les mesures récapitulatives et donnent une explication des processus sous-jacents aux mesures descriptives.

Cette section présente les mesures récapitulatives suivantes :
- distributions de fréquences (en tableau ou graphique),
- mesures de tendance centrale (c.-à-d. moyenne, médiane ou mode),
- mesures de l'étalement de la distribution (p. ex., variance, étendue).

Il est important de bien tenir compte du plan d'échantillonnage pendant l'analyse de la population. Ce chapitre présente d'abord le cas de données simples à la Section 11.3.1 et les lignes directrices pour afficher les résultats de l'enquête en graphique. Les mesures de position et d'étalement pour les données plus complexes sont ensuite exposées à la Section 11.3.2.

11.3.1 Données d'enquête simple

Les statistiques sommaires pour les données d'enquête simple (p. ex., le recensement, l'échantillon aléatoire simple (EAS), ou l'échantillon systématique (SYS), sans ajustements de pondération) sont considérées dans les cours de premier cycle en statistique et présentées brièvement dans ce chapitre. Le lecteur intéressé peut consulter de nombreux ouvrages pour obtenir davantage d'information (p. ex., Lohr (1999), Cochran (1977)).

11.3.1.1 Estimation et présentation des distributions de fréquences

La *distribution de fréquences* est la représentation le plus simple d'une variable. Les distributions de fréquences d'une variable qualitative donnent la fréquence de chaque catégorie, le nombre d'observations dans chaque catégorie, et les résultats peuvent être présentés en tableau ou en graphique (p. ex., un graphique à barres). Les distributions de fréquences pour les variables quantitatives sont habituellement présentées en graphiques parce qu'un tableau des fréquences de chaque valeur de la variable pourrait être peu pratique.

Nous avons présenté au **Chapitre 7 - Estimation** les estimateurs utilisant les pondérations d'échantillonnage qui s'appliquent aux données simples et complexes. Des estimateurs de domaines sont utilisés pour estimer les fréquences et le domaine est une catégorie (pour une donnée qualitative) ou une valeur (pour une donnée quantitative).

Par exemple, l'estimateur habituel pour la taille de la population dans un domaine d'intérêt pour les données qualitatives s'écrit :

$$\hat{N}_{domaine} = \sum_{i \in S_r \cap domaine} w_i$$

où w_i est le poids final ajusté du i^e répondant et S_r est l'ensemble des répondants. L'estimateur habituel du total d'un domaine pour les données quantitatives s'écrit :

$$\hat{Y}_{domaine} = \sum_{i \in S_r \cap domaine} w_i y_i \ .$$

Illustrons l'estimation des distributions de fréquences pour une enquête-échantillon : considérez un EAS de 100 employés sélectionnés dans une population de 1 000 hommes. L'une des variables de l'enquête est la variable nominale *genre de travail* qui comprend deux catégories : *travailleurs manuels* et *travailleurs de bureau*. Une autre variable de l'enquête est la variable continue *tension artérielle systolique* qu'une infirmière diplômée a mesurée directement et inscrite en millimètres de mercure (mm Hg). Après

l'enquête, les lectures de tension artérielle sont aussi catégorisées en trois groupes : *basse, moyenne* ou *élevée*.

Voici la distribution des fréquences pondérées de la variable qualitative *genre de travailleurs* en tableau :

Tableau 1 : Nombre d'hommes estimé par genre de travailleurs

Genre de travailleurs	Nombre d'hommes estimés $\hat{N}$
Manuel	550
Bureau	450
Total	1 000
Source : Enquête fictive auprès des travailleurs, Canada, 2002.	

(Remarque : chaque estimation dans ces tableaux devrait comprendre une estimation de l'erreur d'échantillonnage.)

La distribution des fréquences pondérées de la variable qualitative *tension artérielle* est présentée dans le tableau suivant:

Tableau 2 : Nombre d'hommes estimé par tension artérielle

Tension artérielle	Nombre d'hommes estimés $\hat{N}$
Basse	320
Normale	630
Élevée	50
Total	1 000
Source : Enquête fictive auprès des travailleurs, Canada, 2002.	

Les distributions conjointes sont utiles pour identifier les combinaisons inhabituelles. La *distribution conjointe* pondérée du *genre de travailleurs* et de la *tension artérielle* ci-dessous n'indique aucune incidence de tension artérielle élevée chez les travailleurs manuels (pour la population visée par l'enquête) :

Tableau 3 : Nombre d'hommes estimé par genre de travailleurs et tension artérielle

Genre de travailleurs	Tension artérielle			Nombre d'hommes estimé $\hat{N}$
	Basse	Normale	Élevée	
Manuel	240	310	0	550
Bureau	80	320	50	450
Total	320	630	50	1000
Source : Enquête fictive auprès des travailleurs, Canada, 2002.				

(Les tableaux des distributions conjointes sont souvent analysés avant la diffusion des données dans le public pour se garantir de la divulgation des données confidentielles, c.-à-d. que les tableaux qui révèlent des particuliers sont supprimés. Le **Chapitre 12 - Diffusion des données** donne davantage d'information sur les méthodes de contrôle de la divulgation.)

Les distributions de fréquences peuvent aussi être représentées à l'aide de graphiques ou de diagrammes. L'analyse des données devrait en fait commencer par une analyse visuelle des données. L'affichage graphique est important pour de nombreuses raisons, notamment :

- les représentations graphiques des données sont supérieures aux représentations simplement numériques pour découvrir la structure caractéristique de la distribution,
- la forme de la distribution est au moins aussi importante que l'étalement et le centre de la distribution,
- la forme de la distribution devrait déterminer le choix du paramètre (p. ex., moyenne, médiane ou mode) pour décrire les données d'une seule variable.

Les graphiques et diagrammes suivants pourraient être ajoutés à un rapport sur les résultats de l'enquête :

- diagramme à secteurs,
- diagramme à colonnes,
- graphique à barres,
- graphique linéaire,
- diagramme à boîte et moustaches.

Ceux-ci sont examinés à la section suivante.

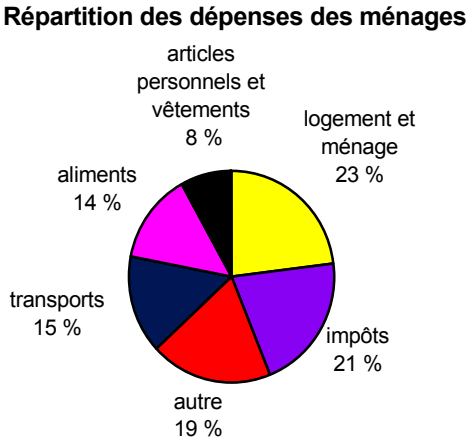
11.3.1.1.1 Diagrammes et schémas

Le genre de diagramme à utiliser est déterminé par les données qu'il faut représenter et par le message qu'on veut souligner : ordre de grandeur, taille ou tendance.

i. Diagrammes à secteurs

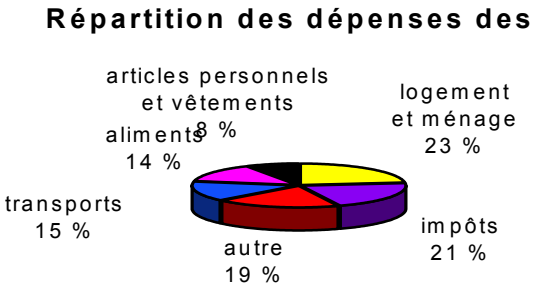
Un diagramme à secteurs est un cercle divisé en pointes comme une tarte pour afficher le pourcentage de la population dans différentes catégories d'une variable qualitative. Un diagramme est utile si la population doit être répartie en groupes distincts (p. ex., la langue maternelle est le français ou l'anglais) et, de préférence, seules quelques unités sont entrées dans la catégorie *autre* ou *sans objet*. Les diagrammes à secteurs sont utilisés pour répondre à des questions sur les proportions relatives de composantes mutuellement exclusives.

Lorsque l'on trace un diagramme à secteurs, il faudrait répartir les secteurs (pointes de tarte) selon la taille, la pointe la plus large à 12 h, et ainsi de suite dans le sens des aiguilles d'une montre, les pointes diminuant graduellement. Le nombre de secteurs devrait être limité à cinq ou six en général. S'il y a de nombreux petits secteurs, il vaudrait peut-être mieux les regrouper. Les étiquettes devraient être à l'extérieur des pointes et il faut éviter les flèches et les légendes. Un bon exemple de diagramme à secteurs affichant les dépenses des ménages est illustré ci-dessous.



Source : Enquête fictive sur les revenus et dépenses des ménages, Canada, 2002

Le tracé tridimensionnel des diagrammes à secteurs (voir ci-dessous) peut semer la confusion parce qu'il déforme les secteurs et il vaut mieux l'éviter.

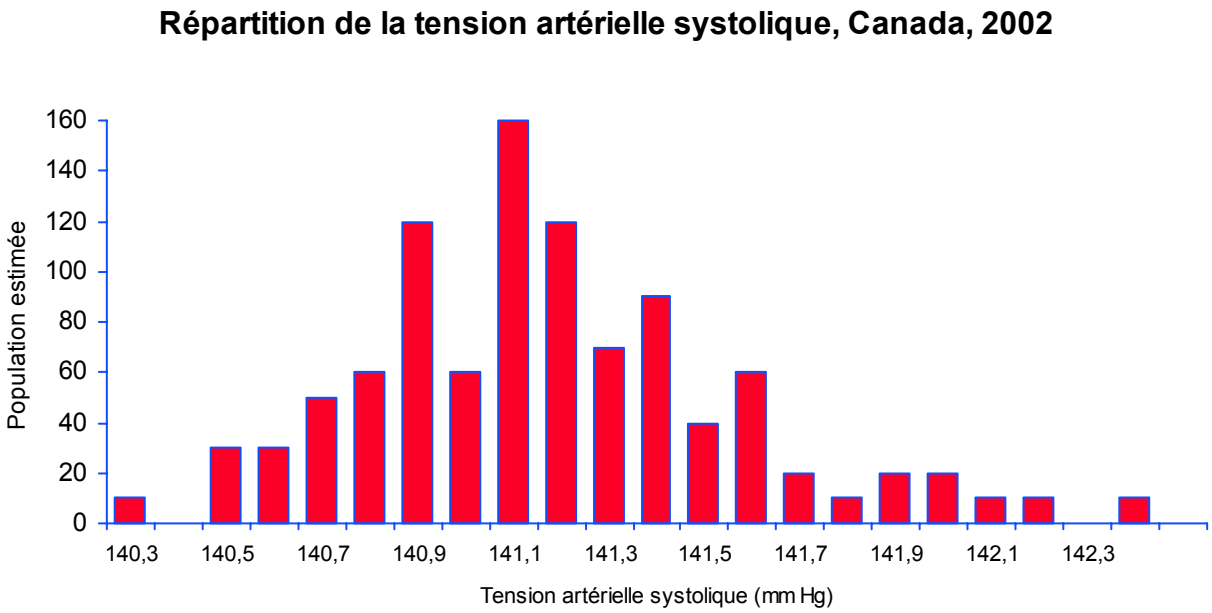


Source : Enquête fictive sur les revenus et dépenses des ménages, Canada,

ii. Diagrammes à colonnes

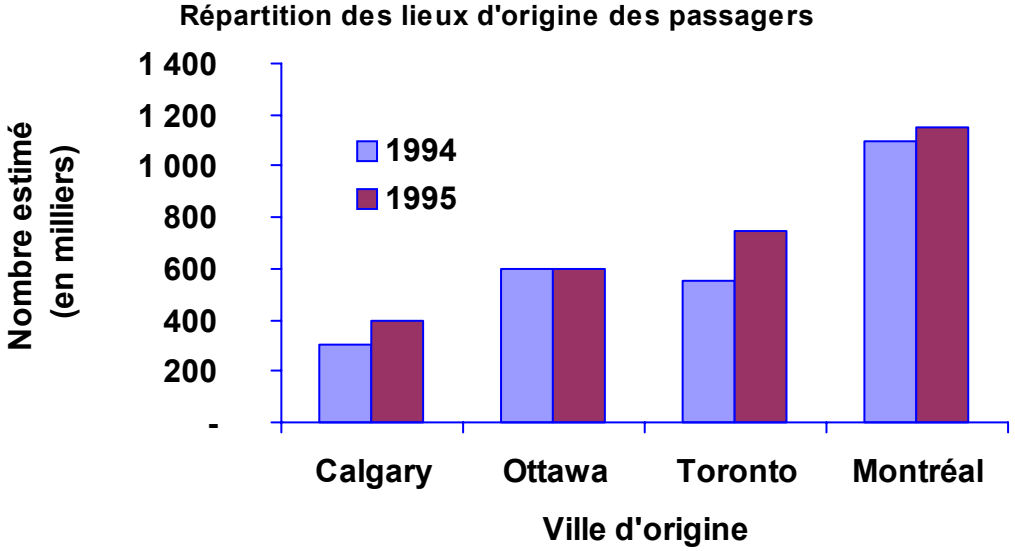
Un diagramme à colonnes comprend une série de colonnes dont les hauteurs représentent les ordres de grandeur (p. ex., totaux, moyennes ou proportions). Le diagramme à colonnes peut être utilisé pour les variables qualitatives ou quantitatives. Le diagramme à colonnes devrait être utilisé pour quelques points seulement et les colonnes devraient avoir la même largeur.

La distribution des tensions artérielles systoliques (une variable quantitative), par exemple, pourrait être estimée dans la population à l'aide des données d'enquête d'un échantillon pondéré et elle est présentée dans le diagramme à colonnes suivant :



Source : Enquête fictive auprès des travailleurs (hommes), Canada, 2002.

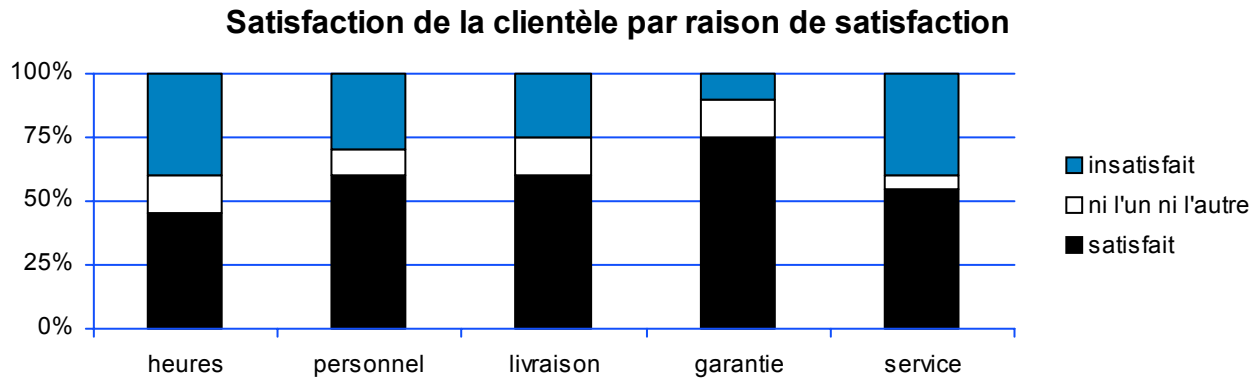
Le diagramme à colonnes comprend un certain nombre de variations. Un *diagramme à colonnes regroupées* a plusieurs variables regroupées en barres côte à côte. Il ne devrait pas y avoir plus de trois barres dans un groupe. L'analyste voudra peut-être comparer, par exemple, le revenu total, les ventes totales et le revenu net au cours d'une certaine période. Voici un exemple d'un diagramme à colonnes regroupées dont les colonnes côte à côte représentent les années consécutives et chaque groupe de colonnes, le nombre de passagers qui visitent une région donnée par ville d'origine des passagers.



Source : Enquête fictive sur le tourisme du régime intérieur, Canada, 1995,1996.

Un *diagramme à colonnes proportionnelles (ou à colonnes empilées)* donne la proportion de la population dans chaque catégorie d'une variable qualitative et chaque colonne représente un domaine différent. Les colonnes ont toutes la même hauteur et la proportion ayant le plus d'intérêt devrait être la

plus proche de la ligne de base pour faciliter la comparaison. La variable comparée entre différents domaines ne devrait pas avoir plus de trois catégories parce que le diagramme à colonnes empilées sera presque illisible s'il y en a trop. Voilà pourquoi, dans l'exemple suivant, les cinq catégories (très satisfait, satisfait, ni l'un ni l'autre, insatisfait et très insatisfait) de la variable *satisfaction* ont été ramenées à trois (satisfait, ni l'un ni l'autre et insatisfait) et comparées pour cinq domaines d'intérêt (heures, personnel, livraison, garantie, service) :



Source : Sondage fictif sur la satisfaction de la clientèle, endroit, année.

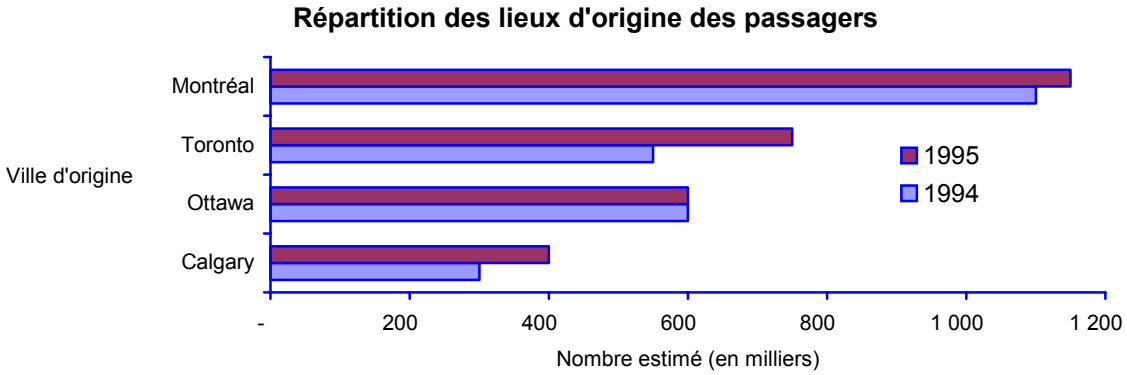
Les diagrammes à colonnes sont généralement utilisés pour des valeurs positives seulement (p. ex., dénombrement de la population, proportions, etc.). Un *graphique à tuyaux d'orgue plus-moins* affiche cependant des valeurs positives et négatives au cours d'une certaine période. Une valeur négative pointe simplement vers le bas sous la ligne de base au lieu de pointer vers le haut.

iii. Graphiques à barres

Un graphique à barres est un graphique à « colonnes » horizontales. Lorsque l'on trace un graphique à barres, les barres devraient être disposées par ordre de longueur (de la plus longue à la plus courte, ou vice versa).

Si des valeurs exactes sont inscrites, le graphique devrait être annoté (c.-à-d. que la valeur exacte devrait être inscrite à la fin de chaque barre). Si ces vedettes de la colonne de titres sont longues, un graphique à barres peut être plus facile à lire et paraître moins encombré qu'un diagramme à colonnes. Il y a de nombreuses variations sur le graphique à barres élémentaire qui correspondent à différents types de diagrammes à colonnes (p. ex., à colonnes empilées, groupées, etc.).

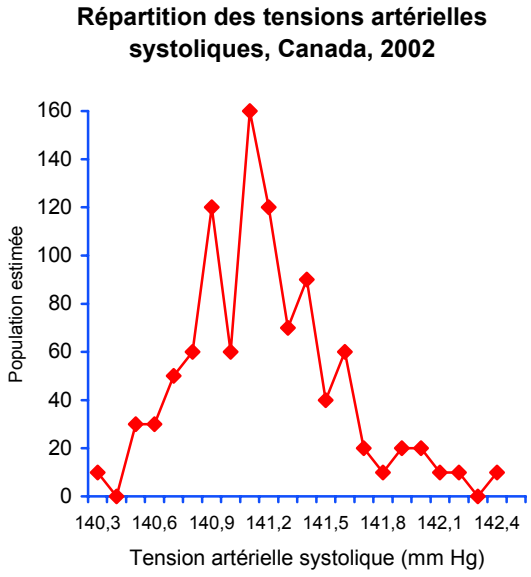
Voici un exemple de graphique à barres groupées :



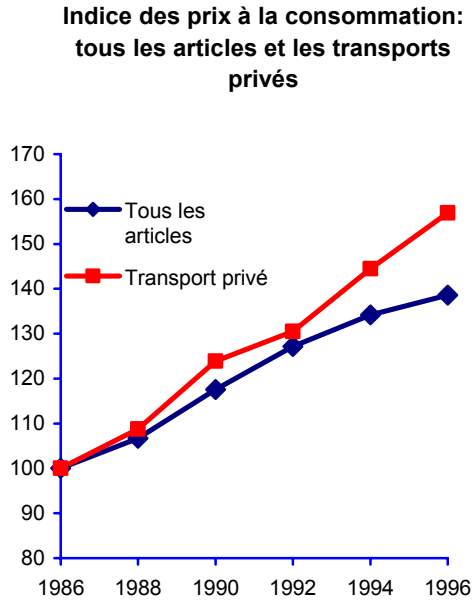
Source : Enquête fictive sur le tourisme du marché intérieur, Canada 1995, 1996.

iv. Graphiques linéaires

Un graphique linéaire affiche une variation dans l'ordre de grandeur d'une variable au cours d'une certaine période (p. ex., totaux, moyennes ou proportions dans le temps). Le temps (la variable explicative) est placé sur l'axe horizontal. L'étendue des valeurs de la variable d'intérêt est placée sur l'axe vertical. Un point (c.-à-d. une mesure de l'ordre de grandeur) est tracé pour cette variable pour chaque unité de temps et les points sont liés en séquence. Les lignes sont droites d'un point à l'autre ou elles peuvent être des courbes peu prononcées. Voici des exemples de graphiques linéaires :



Source : Enquête fictive auprès des travailleurs (hommes), Canada, 2002.

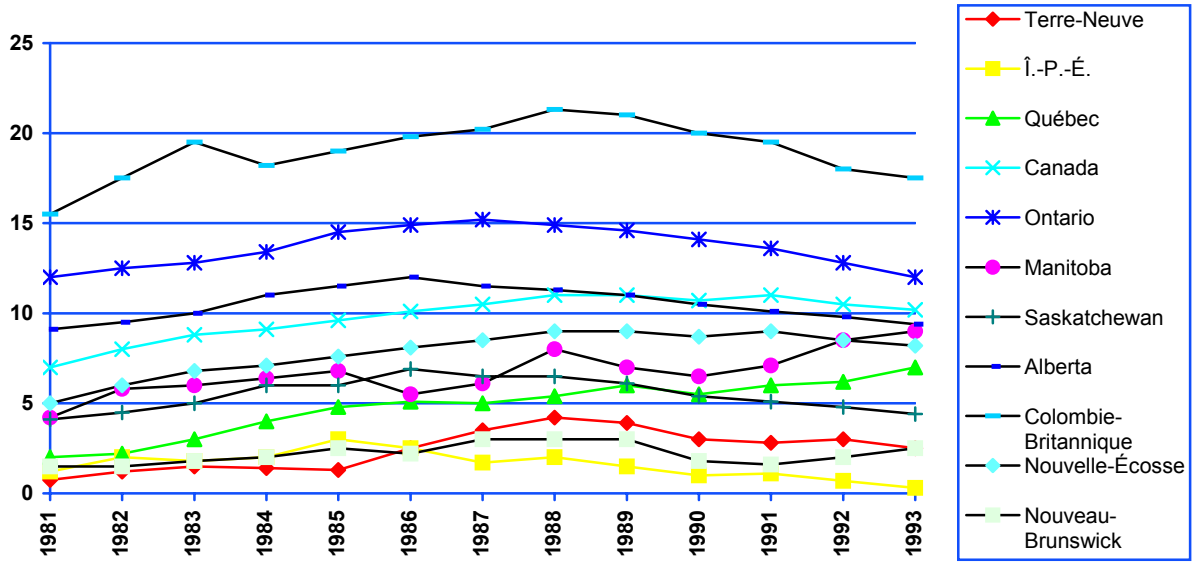


Les graphiques linéaires devraient servir à démontrer les tendances ou le mouvement. Le graphique linéaire est préférable au diagramme à colonnes pour les séries de temps ayant un grand nombre de points. Le graphique linéaire est le meilleur moyen de mettre en évidence les différences ou les ressemblances entre des groupes pour comparer plusieurs séries de données. Si les données révèlent des tendances évidentes, le graphique linéaire donne à l'utilisateur une certaine capacité prédictive. Les tendances

constantes à la hausse ou à la baisse, ou la périodicité évidente, permettent à l'observateur d'interpoler ou d'extrapoler des données.

Il vaut mieux ne pas comparer de trop nombreuses séries simultanément pour éviter la confusion. Voici un exemple de graphique linéaire médiocre :

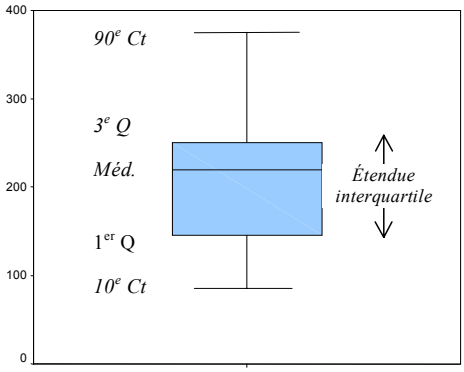
Avortements thérapeutiques par tranche de 10 000 femmes en âge de procréer, Canada et provinces, 1981, 1993



v. Diagrammes à boîte et moustaches

Les statistiques sommaires peuvent aussi être présentées en un seul graphique récapitulatif : le diagramme à boîte et moustaches. Celui-ci est utilisé pour étudier la distribution et l'étalement des données. La boîte elle-même se prolonge à partir du premier quartile (c.-à-d. le 25^{e} centile) jusqu'au troisième quartile (c.-à-d. le 75^{e} centile) et une ligne est tracée à la médiane (c.-à-d. le 50^{e} centile). Les « extrémités » ou pointes des lignes liées à la boîte représentent les valeurs minimales et maximales. Certaines trousses statistiques affichent aussi la moyenne et l'erreur-type de la moyenne (s'il s'agit d'un EAS) dans le tracé en boîte, mais ni l'une ni l'autre n'est affichée ici.

Distribution des prix de vente des maisons, juin 2002
(Prix en milliers de dollars)



Source : Enquête fictive sur les maisons vendues en juin 2002.

Les définitions de la médiane, du premier quartile, du troisième quartile, du 10e et du 90e centile sont données à l'exemple 11.1 dans la section suivante pour les données simples et aux Sections 11.3.2.1 et 11.3.2.2 pour les données complexes. On trouvera davantage d'information sur les diagrammes à boîte et moustaches dans Tukey (1977).

11.3.1.2 Position : moyenne, médiane et mode

Il y a trois mesures communes de la position : la moyenne, la médiane et le mode. Dans les analyses statistiques, la *moyenne* est de loin la plus souvent utilisée pour les données quantitatives. ***La moyenne de la population pour un recensement est simplement la moyenne arithmétique pour les données quantitatives : la somme de toutes les valeurs d'une variable divisée par le nombre de valeurs.*** Voici l'estimateur habituel pour estimer la moyenne de la population à l'aide d'un échantillon aléatoire simple dont le taux de réponse atteint 100 % :

$$\hat{\bar{Y}} = \frac{\sum_{i \in S_r} y_i}{n_r}$$

où y_i est la valeur déclarée pour la i^e unité répondante et n est la taille de l'échantillon.

La moyenne a plusieurs avantages comparativement aux autres mesures de la position. Premièrement, elle est facile à calculer et à comprendre. Elle a la caractéristique souhaitable d'être un estimateur non biaisé de la moyenne de la population pour de nombreux plans d'échantillonnage probabiliste et de grands échantillons en général. La moyenne a cependant plusieurs inconvénients. Lorsque vous considérez des valeurs entières, notamment le nombre d'enfants par ménage, la moyenne peut être une faction. Le nombre moyen d'enfants par ménage, par exemple, peut être 1,8. La moyenne ne peut servir de mesure de position pour les variables qualitatives. De plus, les valeurs extrêmes peuvent avoir une grande influence sur la moyenne (elle se déplace vers les valeurs extrêmes). Dans une enquête sur les revenus par exemple, si quelques membres de la population ont des revenus extrêmement élevés, ceux-ci gonfleront la moyenne de la population. Si l'utilisateur veut une estimation de la valeur centrale, il préférera peut-être une mesure de position moins sensible aux distributions asymétriques ou aux valeurs extrêmes.

La *médiane* est une autre mesure de la position. ***La médiane est la valeur du milieu d'une série de données disposées en ordre numérique (à partir de la plus petite jusqu'à la plus grande ou de la plus grande jusqu'à la plus petite).*** Si les données ont un nombre pair de points, la médiane est la moyenne des deux valeurs du milieu. La médiane peut servir pour les données quantitatives et numériques ordinales, et elle est la meilleure mesure de la tendance centrale d'une variable ordinale.

Les valeurs extrêmes ont moins de répercussions sur la médiane que sur la moyenne et c'est son principal avantage. Dans une enquête sur les revenus, par exemple, les revenus très élevés ont moins d'incidence sur la médiane. Celle-ci serait en fait inchangée même si le revenu le plus élevé était en millions ou en milliards. Dans le cas des données d'enquête d'un échantillon, le principal inconvénient de la médiane est qu'il est habituellement plus difficile d'en calculer la variance d'échantillonnage et, évidemment, de l'utiliser pour l'analyse par inférence.

La troisième mesure de la position est le *mode*. ***Le mode est la valeur des données la plus fréquente***. C'est la plus générale des trois mesures de la tendance centrale. Il peut être appliqué à tous les genres de données, mais il est le plus approprié pour les données qualitatives et c'est la seule mesure sensée de la tendance centrale pour les données nominales. Au cours d'un recensement agricole par exemple, si vous demandez aux agriculteurs d'inscrire la culture qui couvre la majeure partie de leur terre, et si 38 des 50 agriculteurs de la population inscrivent que cette récolte est le blé, le blé est donc le mode.

Le principal avantage du mode est sa simplicité parce qu'il peut être déterminé à partir d'un tableau ou d'un graphique de la distribution des fréquences des données. Le mode a cependant un certain nombre d'inconvénients. Premièrement, il ne décrit peut-être pas suffisamment les données parce que la catégorie la plus commune peut être peu fréquente. Ce problème se pose habituellement lorsqu'il y a de nombreuses valeurs de données possibles. Dans un recensement sur la migration interurbaine par exemple, vous pouvez faire la collecte de données nominales en demandant à 2 000 personnes quelle ville elles ont quitté et vous pouvez obtenir 1 999 réponses différentes, deux personnes seulement ayant le même point d'origine. Cette ville en commun serait le mode, mais il ne serait pas très significatif. Voilà pourquoi le mode est rarement appliqué aux données quantitatives qui ont habituellement de nombreuses valeurs possibles. Contrairement à la médiane et à la moyenne, le mode n'est pas nécessairement unique non plus. Plusieurs catégories peuvent être égales lorsque vous déterminez le rang le plus commun.

Une question se pose évidemment : « Quelle mesure devrait-on utiliser? » Il est important que la mesure soit significative, appropriée, et qu'elle réponde aux besoins de l'utilisateur. Le mode devrait en général être utilisé pour les données nominales, la médiane, pour les données numériques ordinales et quantitatives asymétriques (c.-à-d. qui ne sont pas symétriques par rapport à la moyenne), et la moyenne, pour les données quantitatives réparties symétriquement. Si on considère des données quantitatives, la distribution des valeurs de la variable devrait déterminer le choix. Si la distribution est symétrique et s'il y a seulement un sommet (p. ex., distribution normale) la moyenne, la médiane et le mode sont identiques. Le choix n'a pas d'importance dans ce cas, mais si l'analyste sait qu'ils sont identiques, les données sont donc symétriques. Si la distribution est asymétrique, une estimation des trois mesures donne un indicateur de l'ampleur de l'asymétrie.

D'autres mesures de la position sont parfois utilisées dans les analyses statistiques descriptives. ***Les quartiles sont des mesures de la position et, comme dans le cas de la médiane, il faut d'abord inscrire les valeurs des données en ordre, mais au lieu de séparer la distribution en deux parties (comme dans le cas de la médiane), les quartiles ont quatre parties, chacune contenant 25 % de la distribution en ordre. Les centiles établissent aussi les valeurs des données en ordre, mais ils divisent la distribution en 100 entrées égales.*** Le 10^{e}, le 50^{e} (la médiane) et le 90^{e} centiles sont des statistiques souvent utilisées.

Exemple 11.1 : Mesures de la position pour un recensement des ventes de maisons

Supposons qu'une enquête est faite aux fins du recensement de toutes les maisons vendues au mois de juin dans une ville en particulier et que l'on obtienne les chiffres de vente suivants (en milliers de dollars) : 85, 235, 146, 295, 96, 250, 235, 205, 195 et 375. Triés en ordre : 85, 96, 146, 195, 205, 235, 235, 250, 295 et 375.

Tableau 4 : Mesures de position pour un recensement des ventes de maisons

Mesure de la position	Valeur
Moyenne	211 700 $
Médiane	220 000 $ (moyenne de 205 000 $ et 235 000 $)
Mode	235 000 $
1^{er} quartile (ou 25^{e} centile)	146 000 $ (plus petite valeur plus grande que la première tranche de 25 % des valeurs)
3^{ed} quartile (ou 75^{e} centile)	250 000 $ (plus petite valeur plus grande que la première tranche de 75 % des valeurs).
90^{e} centile	375 000 $ (plus petite valeur plus grande que la première tranche de 90 % des valeurs)

11.3.1.3 Étalement

L'étalement est la variabilité ou la dispersion des données. Une mesure de l'étalement est présentée au **Chapitre 7 - Estimation,** c'est-à-dire la *variance* qui est calculée comme le carré des différences par rapport à la valeur de la moyenne. La variance de deux distributions différentes a été considérée : celle de la population et celle de l'estimateur. La variance de la population mesure l'étalement de la distribution de toutes les données y_i de la population (où y est une variable d'intérêt et y_i est la valeur de la i^e unité). La variance d'échantillonnage mesure l'étalement de la distribution des estimations de différents échantillons à l'aide du même estimateur et du même plan d'échantillonnage. Afin de donner aux utilisateurs de l'information sur la qualité de l'enquête-échantillon, toutes les estimations de l'échantillon devraient comprendre une certaine mesure de l'erreur d'échantillonnage (variance d'échantillonnage, erreur-type, coefficient de variation ou marge d'erreur).

Outre la variance de la population, d'autres mesures de l'étalement de la population comprennent *l'étendue* et *l'étendue interquartile*. L'étendue est l'écart entre la plus grande et la plus petite valeur. Étant donné que cette mesure utilise seulement deux valeurs de la distribution, elle donne seulement une idée générale de l'étalement et les valeurs extrêmes ont d'énormes répercussions sur elle.

L'*étendue interquartile* donne l'étendue de la tranche de 50 % au milieu des données. C'est l'écart entre le troisième et le premier quartile (ou le 75^e et le 25^e centile). Cette mesure est moins fragile aux valeurs extrêmes et elle est donc plus utile que la simple étendue pour mesurer l'étalement. L'étendue interquartile peut servir à toutes les données quantitatives.

Exemple 11.1 (suite) : Étendue et étendue interquartile pour un recensement des ventes de maisons

Pour le recensement des ventes de maisons, l'étendue vaut 290 000 $ (c.-à-d. 375 000 $ - 85 000 $) et l'étendue interquartile vaut 104 000 $ (c.-à-d. 250 000 $ - 146 000 $).

Considérons le cas hypothétique suivant pour comprendre l'importance de la combinaison de l'information sur l'étalement et de l'information sur la position. Un employé dans une banque vend des produits financiers pour la retraite et essaie de déterminer le meilleur endroit pour ouvrir un nouveau bureau. Les travailleurs de 45 ans environ sont le marché ciblé parce qu'ils ne sont pas trop loin de la retraite, mais ils n'ont probablement pas commencé à planifier et ils ont de l'argent disponible. Le bureau pourrait être ouvert dans deux villes éventuellement. Un rapport statistique sur un recensement des villes révèle que l'âge moyen des travailleurs est 45 ans dans les deux. Sans autre information, l'employé voudra peut-être ouvrir un bureau dans chaque ville. En considérant l'étalement des données cependant, il constate que les travailleurs de la ville A ont tous entre 40 et 50 ans, et ceux de la ville B ont de 15 à 65 ans, les deux valeurs modales étant de 20 et 60 ans. Le tableau est maintenant très différent et il peut être plus avantageux pour la banque d'ouvrir un bureau dans la ville A (il faudra quand même obtenir davantage d'information, par exemple, combien de résidents de 40 à 50 ans habitent dans chaque ville).

11.3.2 Données d'enquête complexe

Les mêmes estimateurs de domaines présentés au **Chapitre 7 - Estimation** et à la Section 11.3.1 ci-dessus peuvent être utilisés pour estimer les distributions de fréquences, les moyennes, les totaux et les proportions des sondages ayant des données complexes. Les estimations pour les statistiques d'ordre comme la médiane et l'étendue interquartile sont plus compliquées.

11.3.2.1 Médiane

La médiane de la population est la valeur de la variable d'intérêt sous laquelle se trouve la moitié de la distribution de la population. L'estimation de la médiane de la population est plus compliquée pour les plans d'enquête complexes que pour les plans d'enquête simples. Des tailles de classes estimées sont nécessaires pour les variables ordinales. Les valeurs des données sont triées par ordre croissant pour les variables continues et les poids d'estimations sont additionnés jusqu'à ce que la somme atteigne 50 % de la taille estimée de la population. L'exemple suivant illustre comment estimer la médiane de la population en estimant la distribution de la population à l'aide de calculs pondérés.

Exemple 11.2 : Médiane estimée pour un EAS stratifié (voir l'exemple 7.6 du **Chapitre 7 - Estimation**)

Rappelons pas que l'échantillon de cet exemple a dix observations et deux strates. Voici les points de l'échantillon (trié) :

Tableau 5 : Distribution estimée des revenus de la population d'un EAS stratifié

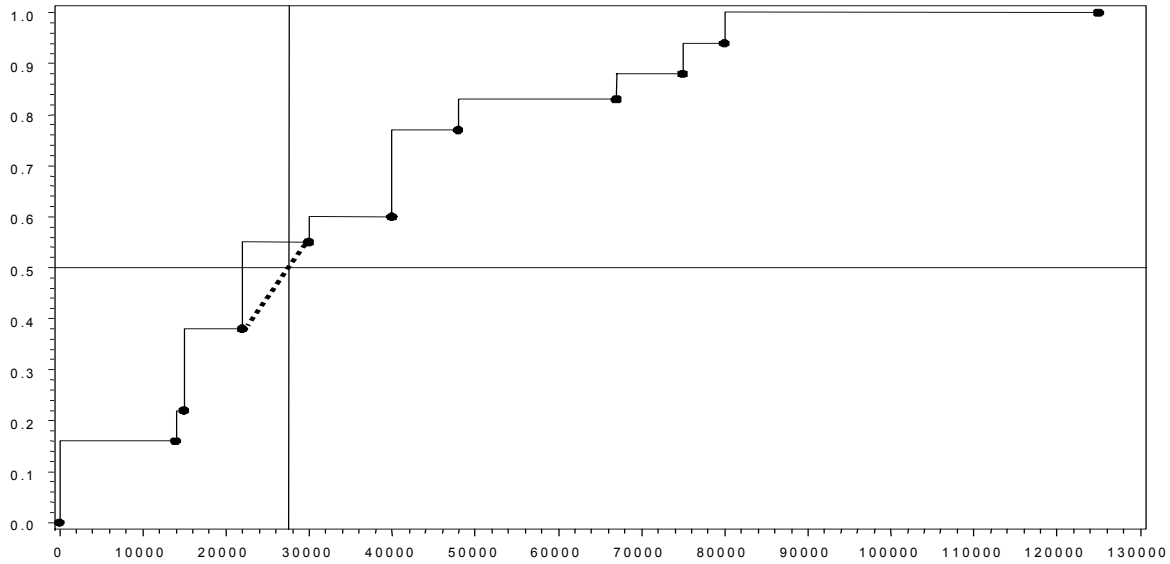
Strate	N° d'id. de l'unité	Poids final	Poids cumulés	Poids cumulés relatifs	Genre de ferme	Revenu déclaré ($)
2	8	16,5	16,5	0,16	1	14 000
1	2	5,67	22,17	0,22	2	15 000
2	10	16,5	38,67	0,38	1	22 000
2	7	16,5	55,17	0,55	1	30 000
1	6	5,67	60,84	0,60	1	40 000
2	9	16,5	77,34	0,77	2	48 000
1	4	5,67	83,01	0,83	1	67 000
1	1	5,67	88,68	0,88	1	75 000
1	5	5,67	94,35	0,94	2	80 000
1	3	5,67	100,02	1,0	1	125 000

Genre d'exploitation agricole 1= culture (N_1=34, n_1=6)
2= élevage (N_2=66, n_2=4)

Le revenu moyen de la population entière d'exploitations agricoles est estimé à :

$$\hat{\bar{Y}} = \frac{\sum_{i \in S_r} w_i y_i}{\sum_{i \in S_r} w_i} = \frac{4{,}160{,}340}{100.02} = 41{,}595.1 .$$

Estimation de la médiane à partir de la distribution estimée



La médiane estimée se situe entre 22 000 $ et 30 000 $ parce que les pondérations cumulées sont de 0,38 à 0,55 pour ces deux chiffres. Il est pratique commune, pour obtenir une seule valeur, de faire une interpolation linéaire entre deux points (22 000 $, 0,38) et (30 000 $, 0,55) pour obtenir les coordonnées du point médian (*Méd.,* 0,50), cette explication étant illustrée ci-dessus (l'estimation non pondérée de la médiane est 44 000).

$$Méd. = 22\,000 + \frac{30\,000 - 22\,000}{0{,}55 - 0{,}38}(0{,}50 - 0{,}38) = 27\,647\,.$$

11.3.2.2 Étalement

Il est plus facile de présenter le cas des données d'enquête d'un EAS ou d'un EAS stratifié sans ajustement de pondération comme celui du **Chapitre 7 - Estimation** pour illustrer le concept de la variance d'échantillonnage. En pratique cependant, à peu près toutes les enquêtes ont des données plus complexes, même si le plan d'échantillonnage est un EAS ou un plan systématique (SYS), un ajustement de pondération pour les non-réponses est habituellement appliqué, et les formules de l'EAS ou de l'EAS stratifié ne s'appliquent donc pas.

Le plan d'échantillonnage et l'estimateur ponctuel déterminent la formule de la variance exacte (c.-à-d. que l'estimateur de la moyenne détermine l'estimateur pour la variance d'échantillonnage d'une moyenne). L'estimation de la variance pour des données complexes devient rapidement compliquée. Afin d'estimer correctement l'erreur d'échantillonnage pour un sondage ayant des données complexes, il est préférable de consulter un statisticien d'enquête qui connaît bien ce genre de problème. Il n'est pas recommandé d'utiliser simplement un logiciel, même un logiciel statistique, parce qu'un EAS implicite sans ajustement de pondération y est souvent intégré.

Pour estimer les étendues interquartiles pour des données complexes, on peut appliquer l'approche expliquée ci-dessus pour la médiane, afin d'estimer le 25^{e} et le 75^{e} centile.

11.3.2.2.1 Intervalles de confiance en présence de biais

L'étude de l'estimation et de l'analyse des données d'enquête a supposé jusqu'à maintenant qu'il n'y avait pas de biais. Au **Chapitre 3 - Introduction au plan d'enquête**, nous avons énuméré quatre sources d'erreurs non dues à l'échantillonnage qui peuvent causer un biais : la couverture, la mesure, la non-réponse et les erreurs de traitement. L'estimateur peut aussi causer un biais : l'analyste peut préférer utiliser un estimateur ayant un petit biais, mais une bonne précision, au lieu d'un estimateur non biaisé ayant une précision médiocre.

La variation totale par rapport à la valeur réelle d'un paramètre, θ, est intitulée *erreur quadratique moyenne :*

$$\begin{aligned} MSE(t) &= E(t-\theta)^2 \\ &= E(t-E(t))^2 + (E(t)-\theta)^2 \\ &= Var(t) + \big(Biais(t)\big)^2 \end{aligned}$$

où t est l'estimation de θ pour un échantillon réalisé, $E(t)$ est la valeur prévue, ou l'estimation moyenne de tous les échantillons possibles et $Var(t)$ est la variance d'échantillonnage de t.

En présence d'un biais, $E(t) = \theta + B$. S'il n'y a pas de biais, $E(t) = \theta$, et la variation totale par rapport à la valeur réelle, θ, est simplement la variance d'échantillonnage :

$$\begin{aligned} MSE(t) &= E(t-\theta)^2 \\ &= E(t-E(t))^2 + (E(t)-\theta)^2 \\ &= Var(t). \end{aligned}$$

Les intervalles de confiance (considérées auparavant au **Chapitre 7 - Estimation** et au **Chapitre 8 - Calcul de la taille de l'échantillon et répartition**) sont souvent utilisés pour présenter les résultats d'enquêtes probabilistes. Étant donné une estimation t et son erreur-type, $S\hat{E}(t) = \sqrt{V\hat{a}r(t)}$, un intervalle de confiance peut être établi comme suit :

$$(t - z \times S\hat{E}(t), \quad t + z \times S\hat{E}(t))$$

où z est la valeur correspondant au niveau de confiance (p. ex., z=1,96 pour un intervalle de confiance de 95 %) dans un tableau type de distribution normale. On reconnaît la théorie standard enseignée dans les cours de statistique de premier cycle. Elle s'applique aux moyennes, aux proportions, aux paramètres de régression et à de nombreuses autres statistiques. Son assise théorique est le théorème central limite dans les populations infinies. Il faut cependant un échantillon suffisamment large pour que la théorie asymptotique s'applique et c'est sa limite pratique.

Un intervalle de confiance de 95 % est parfois décrit ainsi :

> Selon une enquête récente, 15 % des résidents d'Ottawa assistent à des services religieux chaque semaine. Les résultats, tirés d'un échantillon de 1 345 résidents, sont considérés précis à plus ou moins 3 %, 19 fois sur 20.

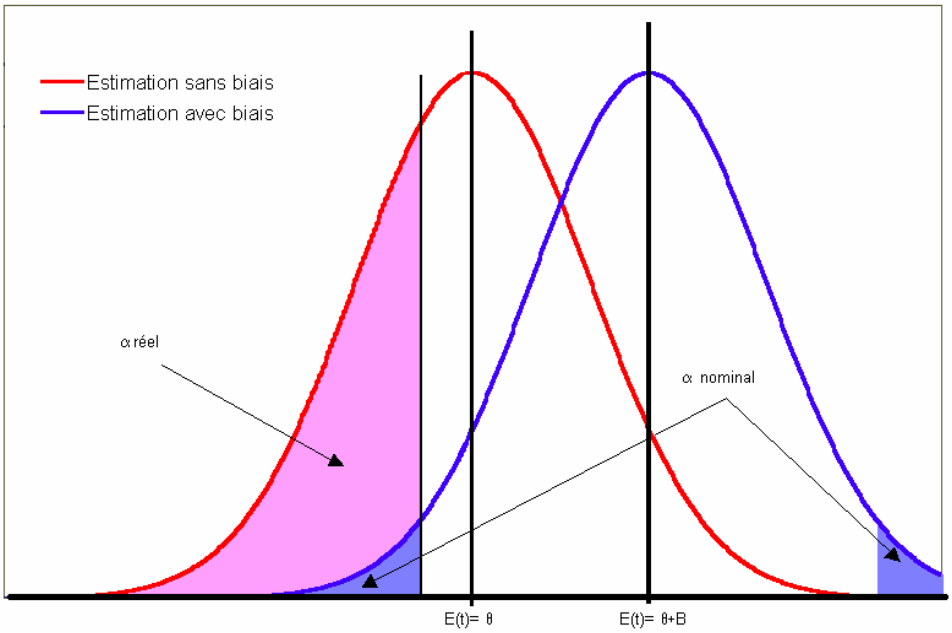
Un intervalle de confiance de 95 %, pour des estimateurs non biaisés qui ont des distributions d'échantillonnage normales ou approximativement normales, signifie que si l'enquête est répétée de

nombreuses fois, environ 19 fois sur 20 (ou 95 % des occasions), l'intervalle de confiance couvrirait la valeur de la population réelle.

En présence d'un biais, il n'y a habituellement pas de mesure du *Biais(t),* et s'il y en avait une, $t + B\hat{i}ais(t)$ serait une estimation non biaisée de θ et un intervalle de confiance serait établi par rapport à cette valeur, mais un intervalle de confiance est plutôt établi par rapport à t à l'aide de $S\hat{E}(t) = \sqrt{V\hat{a}r(t)}$ au lieu de $\sqrt{M\hat{S}E(t)}$.

Les répercussions du biais sur l'intervalle de confiance sont remarquées surtout dans la probabilité de couverture (« est-ce vraiment 95 %? »). L'intervalle de confiance (le secteur entre les zones ombrées de la courbe de droite) est décalé du point de vue de la valeur réelle. La probabilité de couverture pour un intervalle de confiance autour de θ est la zone ombrée sous la courbe de gauche.

Distribution d'un estimateur *t* avec et sans biais
B / SE(*t*) =1



Särndal *et coll.* (1992) donnent le tableau suivant de la probabilité de couverture réelle comme une fonction du biais relatif, c.-à-d. le ratio $B / \sqrt{Var(t)}$.

Tableau 6 : Probabilité de couverture, compte tenu de B/√V

Biais relatif	Probabilité de couverture
0,00	0,95
0,05	0,9497
0,10	0,9489
0,30	0,9396
0,50	0,9210
1,00	0,8300

Nous l'avons déjà mentionné, *B* est en pratique inconnu, mais le concepteur et l'utilisateur de l'enquête doivent être conscients de son existence et de ses répercussions préjudiciables.

Le rapport d'enquête ou le rapport d'analyse des données comprend souvent des tableaux des erreurs-types ou des coefficients de variation estimés, c.-à-d. le ratio de l'erreur d'échantillonnage à l'estimation (en pour cent), et les utilisateurs peuvent donc calculer leurs propres intervalles de confiance et procéder à leurs vérifications d'hypothèses. Ces tableaux devraient comprendre une explication de la méthode à appliquer pour faire des tests d'hypothèses, ainsi que de l'information sur le biais et ses répercussions.

11.4 Test d'hypothèses sur une population : variables continues

Rappelons que les enquêtes-échantillons sont habituellement faites pour étudier les caractéristiques d'une population, établir une base de données à des fins analytiques ou vérifier une hypothèse. La théorie et les méthodes considérées jusqu'à maintenant dans ce manuel ciblent surtout la description de la population et de ses caractéristiques : Il y a combien d'hommes et de femmes dans la population? Combien sont fumeurs? Quelle proportion de la population les familles à faible revenu forment-elles? Quel est le revenu médian des ménages ?

Cette section cible les tests d'hypothèses au sujet de la population : la proportion des fumeurs est-elle différente de celle des fumeuses? La proportion des familles à faible revenu est-elle la même dans toutes les provinces? L'espérance de vie varie-t-elle d'une province à l'autre?

11.4.1 Introduction : les éléments d'un test

Un test d'hypothèse est une procédure appliquée pour déterminer si les données de l'échantillon soutiennent les énoncés formulés au sujet de la population. Une hypothèse est un énoncé, ou une théorie, sur la valeur réelle de la population d'une caractéristique. Un test d'hypothèse comprend la vérification d'une *hypothèse nulle,* H_o, compte tenu d'une *hypothèse alternative,* H_1. Si vous tirez à pile ou face de nombreuses fois, par exemple, l'hypothèse nulle peut être H_o : *la pièce n'est pas biaisée* et l'hypothèse alternative est H_1 : *la pièce est biaisée*.

La probabilité que les valeurs observées soient le résultat fortuit de l'échantillonnage, en supposant que l'hypothèse nulle est vraie, est calculée à l'aide des données d'un échantillon. Si cette probabilité se révèle être plus petite que le *niveau de signification* du test, l'hypothèse nulle est rejetée.

Un test d'hypothèse a quatre composantes : les hypothèses *nulle* et *alternative*, la *statistique du test* et le *niveau de signification*. On devrait ajouter un cinquième élément : une *conclusion*.

i. Hypothèse nulle

L'hypothèse nulle est un énoncé au sujet d'un paramètre de la population que l'analyste veut vérifier et son symbole est H_o. Voici des exemples éventuels d'hypothèse nulle :

- les revenus moyens de deux provinces sont semblables, $H_O : \overline{Y}_1 = \overline{Y}_2$,

- la proportion de fumeurs de la population est de 40 %, $H_O : P = 0{,}4$,

- l'âge moyen de la population est de 38 ans, $H_O : \overline{Y} = 38$.

ii. Hypothèse alternative

L'hypothèse nulle est testée par rapport à l'hypothèse alternative dont le symbole est souvent H_1 ou H_A. L'hypothèse alternative est souvent un énoncé sur la population qui devrait être vrai. L'hypothèse alternative peut être acceptée seulement si les données d'un niveau de signification en particulier ne peuvent soutenir l'hypothèse nulle. Les hypothèses alternatives aux hypothèses nulles ci-dessus pourraient être, par exemple,

- les revenus moyens de deux provinces sont différents, $H_1 : \overline{Y}_1 \neq \overline{Y}_2$,

- la proportion de fumeurs dans la population est supérieure à 40 %, $H_1 : P > 0{,}4$,

- l'âge moyen de la population est de moins de 38 ans. $H_1 : \overline{Y} < 38$.

iii. Statistique du test

La statistique du test est une valeur calculée à partir d'un échantillon (ou de plusieurs échantillons) pour tester une hypothèse sur la population d'où l'échantillon est tiré. Les données, l'hypothèse vérifiée, le niveau de signification et l'estimateur utilisés pour estimer le paramètre déterminent la valeur de la statistique. Celle-ci exige habituellement que l'estimateur ne soit pas biaisé (ou qu'il soit approximativement sans biais) et que la distribution de l'échantillonnage de l'estimateur soit connue. Une statistique « *z* » est distribuée normalement, par exemple, une statistique « khi carré » a une distribution khi carré et une statistique « *F* » a une distribution *F* de Fisher-Snedecor.

iv. Niveau de signification

Les seuls résultats possibles d'un test d'hypothèse sont *rejeter l'hypothèse nulle* ou *ne pas rejeter l'hypothèse nulle*. Rejeter l'hypothèse nulle ne signifie pas toujours qu'elle est fausse et ne pas la rejeter ne signifie pas qu'elle est nécessairement vraie. Il y a en fait deux genres de conclusions erronées : conclure que l'hypothèse nulle est fausse lorsqu'elle est vraie et conclure qu'elle est vraie lorsqu'elle est fausse.

Tableau 7 : Types d'erreur

		L'hypothèse nulle est en fait :	
		VRAIE	FAUSSE
Il est conclu après vérification que l'hypothèse nulle est :	VRAIE		II
	FAUSSE	I	

Ces deux genres de conclusion erronée sont intitulées erreur de type I et erreur de type II respectivement. Le niveau de signification d'un test, soit α, est le risque accepté de commettre une erreur de type I, autrement dit, de rejeter une hypothèse nulle vraie. La valeur, $\alpha = 0{,}05$, par exemple, est souvent utilisée. Si un risque moindre est exigé, on peut attribuer une valeur inférieure à α, disons $\alpha = 0{,}01$. Si un risque plus grand est acceptable, on peut utiliser $\alpha = 0{,}10$.

Le dictionnaire de la statistique de Cambridge (Everitt, 1998) illustre les niveaux de signification comme suit : on tire à pile ou face 100 fois et on obtient face à chaque fois. On peut soupçonner avec raison que

la pièce est biaisée, mais il y a une mince possibilité qu'elle ne soit pas biaisée et qu'elle tombe simplement de cette façon. Nous savons cependant que la probabilité qu'une bonne pièce tombe de la même façon 100 fois sur 100 est très mince : $2 \times (½)^{100}$, ou $1{,}6 \times 10^{-30}$ (c'est la valeur de la statistique du test). Compte tenu de ces points, l'analyste peut rejeter en toute confiance l'hypothèse nulle, *H_o : la pièce n'est pas biaisée* pour adopter l'hypothèse alternative, *H_1 : la pièce est biaisée*, sachant qu'il y a seulement une mince possibilité que sa conclusion soit inexacte. Supposons cependant que la pièce est tirée six fois seulement et qu'elle donne face chaque fois. La probabilité qu'une pièce équilibrée tombe de cette façon est : $2 \times (½)^6$, c.-à-d. 0,031. C'est peu probable, mais pas impossible. Si le niveau de signification est $\alpha = 0{,}05$, l'analyste rejetterait l'hypothèse nulle, mais avec un niveau de signification plus strict de $\alpha = 0{,}01$, l'analyste ne pourrait pas rejeter l'hypothèse nulle.

Il y a deux genres de tests d'hypothèse : les tests unilatéral et bilatéral. Un test est unilatéral lorsque la région de rejet pour l'hypothèse nulle, exprimée graphiquement, consiste en une queue de distribution de l'échantillonnage de l'estimateur. (La région de rejet est l'ensemble des valeurs de la statistique du test qui inciteraient à rejeter l'hypothèse nulle.) Dans un test bilatéral, la région de rejet comprend les deux queues de distribution. Les tests bilatéraux sont habituellement utilisés avec des estimateurs normalement distribués. L'hypothèse alternative ci-dessus, par exemple, selon laquelle les revenus moyens des deux provinces sont différents (p. ex., $H_1 : \overline{Y}_1 \neq \overline{Y}_2$), utiliserait un test bilatéral, alors que les deux autres hypothèses alternatives appliqueraient des tests unilatéraux.

11.4.2 Données d'enquête simples

La matière considérée dans cette section est habituellement le sujet des cours de statistique de premier cycle et nous éviterons intentionnellement les détails et les complications. Le lecteur intéressé peut consulter des ouvrages élémentaires (p. ex., Snedecor et Cochran (1989), Wonnacott et Wonnacott (1977)).

11.4.2.1 Essai pour une moyenne unique

Compte tenu d'une série de données obtenues à l'aide d'un plan d'échantillonnage aléatoire simple d'une population, la moyenne de la population estimée, $\hat{\overline{Y}}$, n'est pas biaisée et (si l'échantillon est suffisamment grand) elle est distribuée presque normalement avec une moyenne, $\overline{Y}$, et une erreur-type estimée, $\hat{SE}(\hat{\overline{Y}})$. Si l'analyste veut tester l'hypothèse selon laquelle la valeur de $\overline{Y}$ est k (c.-à-d. que $H_O : \overline{Y} = k$), la statistique du test suivante peut être utilisée :

$$z = \frac{\hat{\overline{Y}} - k}{\hat{SE}(\hat{\overline{Y}})}.$$

Cette statistique du test est intitulée statistique z parce que, si H_o est vraie, z a donc une distribution type approximativement normale, une moyenne égale à 0 et une erreur-type égale à 1. C'est la même statistique z que celle utilisée pour établir les intervalles de confiance pour la moyenne (voir Section 7.3.2.2).

Parce qu'il connaît la distribution de z, l'analyste connaît la probabilité que z s'écarte de sa moyenne d'un certain nombre d'erreurs-types; il détermine ainsi le niveau de signification pour un test. Il est connu, par exemple, que 5 % des valeurs (absolues) de z sont supérieures à 1,96. Afin de faire un test bilatéral (p. ex., $H_1 : \overline{Y} \neq k$) à l'aide d'une statistique z et de $\alpha = 0{,}05$, la région de rejet serait donc les valeurs de z

inférieures à –1,96 ou supérieures à 1,96. Dans le cas d'un test unilatéral (p. ex., $H_1 : \overline{Y} > k$), pour établir un test dont $\alpha = 0{,}05$, la région de rejet serait les valeurs de *z* supérieures à 1,65.

Il est souvent raisonnable de supposer pour les grands échantillons que $\hat{\overline{Y}}$ suit une distribution normale. C'est parce que dans certaines conditions, selon le théorème central limite, la distribution de la moyenne de l'échantillon approche la distribution normale quand augmente la taille de l'échantillon.

Exemple 11.3 : Test sur une moyenne d'un EAS

Supposons qu'un organisme statistique procède à une enquête sur la santé et fait la collecte des données à l'aide d'un échantillon probabiliste. L'organisme veut vérifier l'hypothèse selon laquelle il y a un problème d'embonpoint dans la population, lequel est défini comme le poids moyen de la population étant supérieur à 100 kg. L'organisme sait que l'estimateur habituel pour la moyenne de la population n'est pas biaisé et est normalement distribué. Une statistique *z* est donc utilisée, et le niveau de signification est de $\alpha = 5\%$. Étant donné qu'un test unilatéral est approprié et, afin d'obtenir un taux de certitude de 95 % pour rejeter l'hypothèse nulle, la région de rejet comprend toutes les valeurs *z* supérieures à 1,65.

Voici la vérification de l'hypothèse :

$$\begin{cases} H_0 : \overline{Y} \le 100kg \\ H_1 : \overline{Y} > 100kg \end{cases}$$

La statistique du test :

$$z = \frac{\hat{\overline{Y}} - 100}{S\hat{E}(\hat{\overline{Y}})}$$

Si les estimations de l'enquête sont $\hat{\overline{Y}} = 102{,}1$ et $S\hat{E}(\hat{\overline{Y}}) = 1{,}5$, alors :

$$z = \frac{102{,}1 - 100}{1{,}5} = \frac{2{,}1}{1{,}5} = 1{,}4 .$$

Étant donné que 1,4 est inférieur à 1,645, la donnée n'est pas dans la région de rejet. L'évidence n'est donc pas suffisante pour rejeter l'hypothèse nulle.

11.4.2.2 Comparaison entre deux moyennes de (sous-)populations

L'approche appliquée au test d'une moyenne peut facilement être appliquée à deux moyennes : soit la différence entre deux groupes d'intérêt, soit le même groupe mesuré à deux points dans le temps. Supposons maintenant qu'un échantillon est tiré de chaque groupe, que les échantillons sont indépendants et que chaque échantillon est suffisamment large pour justifier l'application du théorème central limite.

Le premier groupe a une moyenne inconnue, $\overline{Y}_1$, le deuxième groupe a une moyenne inconnue, $\overline{Y}_2$, et si l'hypothèse nulle est vraie, ces moyennes inconnues sont égales. Leur différence est donc zéro et leurs estimations devraient être très près l'une de l'autre. Toute grande différence observée entre les estimations sont dues à des échantillons malheureusement mauvais (mais il n'y a pas de bonnes raisons pour cela) ou bien, H_0 est faux. Compte tenu de cette explication, le test peut être fait comme suit :

$$\begin{cases} H_0 : \bar{Y}_1 = \bar{Y}_2 \\ H_1 : \bar{Y}_1 \neq \bar{Y}_2 \end{cases}$$

et la statistique du test asymptotiquement normale est :

$$z = \frac{\hat{\bar{Y}}_1 - \hat{\bar{Y}}_2}{\sqrt{\hat{Var}(\hat{\bar{Y}}_1) + \hat{Var}(\hat{\bar{Y}}_2)}}.$$

Le niveau de test décidé d'avance est habituellement 5 %. Étant donné qu'il s'agit d'un test bilatéral, cela correspond à rejeter l'hypothèse nulle si la statistique du test est à l'extérieur de la fourchette (-1,96, 1,96). (Remarquez que cette statistique du test est correcte seulement si les deux échantillons sont indépendants.)

Exemple 11.4 : Test de deux moyennes d'un EAS

Supposons qu'un analyste est intéressé à déterminer si les gens dans une province ont un poids plus élevé, en moyenne, que ceux d'une autre province. Un échantillon aléatoire simple est tiré dans chaque province et les résultats sont $\hat{\bar{Y}}_1 = 95$, $\hat{\bar{Y}}_2 = 105$, $\hat{SE}(\hat{\bar{Y}}_1) = 1{,}4$, $\hat{SE}(\hat{\bar{Y}}_2) = 2{,}2$. Voici donc la statistique du test :

$$z = \frac{95 - 105}{\sqrt{1{,}4^2 + 2{,}2^2}} = \frac{-10}{\sqrt{6{,}80}} = -3{,}83$$

et les deux groupes sont jugés significativement différents.

11.4.2.3 Comparaison entre de nombreuses moyennes de (sous)-populations : modèles d'analyse de la variance (ANOVA) à une dimension et de régression linéaire

Le prolongement naturel de la théorie ci-dessus est l'élaboration d'un essai pour comparer les moyennes de nombreux groupes. Dans le cas de l'ANOVA (*an*alyse (*o*f) de la *va*riance), aucune supposition n'est faite sur le lien éventuel entre les moyennes et, pour les *modèles linéaires*, une hypothèse est formulée sur les liens linéaires entre les moyennes. Les *modèles linéaires* sont dans une catégorie de techniques statistiques utilisées pour déterminer si une variable de réponse a des liens linéaires avec une ou plusieurs variables explicatives. Les effets des diverses variables explicatives sont additifs, une importante caractéristique des modèles linéaires.

11.4.2.3.1. Analyse de la variance (ANOVA)

L'*ANOVA* sert à évaluer l'effet d'une ou de plusieurs variables qualitatives (intitulées *facteurs*) sur une variable de réponse continue. Les différences entre les moyennes sont vérifiées en étudiant la variabilité d'un ensemble d'observations pour déterminer si la variabilité est aléatoire ou si elle peut être attribuée à un ou plusieurs facteurs.

L'ANOVA la plus simple est un plan à un facteur pour lequel un échantillon est tiré de chacun des k différents groupes d'un seul facteur (c.-à-d. que k moyennes différentes sont comparées et, selon l'hypothèse nulle, elles sont toutes égales). L'analyste voudra peut-être, par exemple, vérifier l'hypothèse nulle selon laquelle il n'y a pas de différence entre les revenus moyens des dix provinces :

$$H_O : \overline{Y}_1 = \overline{Y}_2 = \overline{Y}_3 = \overline{Y}_4 = \overline{Y}_5 = \overline{Y}_6 = \overline{Y}_7 = \overline{Y}_8 = \overline{Y}_9 = \overline{Y}_{10}\,.$$

Cette hypothèse s'écrit comme un modèle d'ANOVA :

$$y_{gi} = \gamma_0 + \gamma_g + \varepsilon_i$$

où y_{gi} est la valeur de la variable de réponse, le revenu, pour la i^e unité de la g^e province, γ_0 est le revenu moyen de toutes les provinces, γ_g est la différence entre le revenu moyen de la province g et le revenu moyen national; si toutes les moyennes sont égales, $\Sigma\gamma_g = 0$; finalement, ε_i est une variable d'erreur aléatoire, de moyenne nulle et de variance σ^2.

La variation totale dans la population est répartie en variation due aux différences entre les k groupes et la variation due aux différences entre les sujets dans un même groupe. Cette décomposition peut s'écrire :

$$\sum_g \sum_i (y_{gi} - \overline{\overline{Y}})^2 = \sum_g N_g (\overline{Y}_g - \overline{\overline{Y}})^2 + \sum_g \sum_i \left(y_{gi} - \overline{Y}_g\right)^2$$

où N_g est le nombre d'unités du groupe g, $\overline{Y}_g$ est la moyenne du groupe g et $\overline{\overline{Y}}$ est la moyenne générale.

Si les échantillons aléatoires indépendants ont été tirés de g populations distribuées normalement, cette variation peut être estimée comme suit :

$$\sum_g \sum_{i \in S} (y_{gi} - \overline{\overline{y}})^2 = \sum_g n_g (\overline{y}_g - \overline{\overline{y}})^2 + \sum_g \sum_{i \in} \left(y_{gi} - \overline{y}_g\right)^2$$

$$SS(total) = SS(Modèle) + SS(Résiduel)$$

où n_g est le nombre d'unités échantillonnées du groupe g, $\overline{y}_g$ est la moyenne de l'échantillon du groupe g, $\overline{\overline{y}}$ est la moyenne générale de l'échantillon et *SS* est la « somme des carrés ».

Si les moyennes de l'échantillon k sont toutes les mêmes, elles sont aussi égales à la moyenne générale $\overline{\overline{y}}$. Dans les limites de la variation aléatoire, la variance entre les groupes, c.-à-d. :

$$MS(Modèle) = \frac{SS(Modèle)}{g-1}$$

devrait donc être près de zéro.

Il est possible de tester cette hypothèse à l'aide du test F établi comme suit :

$$\begin{cases} H_0 : \gamma_g = 0, \forall g \\ H_1 : \gamma_j \neq 0, \text{pour certaines } j \end{cases} \equiv \begin{cases} H_0 : \overline{Y}_1 = \cdots = \overline{Y}_{10} \\ H_1 : \overline{Y}_j \neq \overline{Y}_k, \text{pour certaines } j, k \end{cases}$$

et la statistique du test est

$$F = \frac{MS(Modèle)}{MS(Résiduel)} = \frac{SS(Modèle) \big/ (g-1)}{SS(Résiduel) \big/ g(n_g - 1)} \sim F_{g-1;g(n_g-1)}.$$

Cette statistique a une distribution F de Fisher-Snedecor à $(g-1)$ et $g(n_g-1)$ degrés de liberté. Les valeurs critiques sont lues à partir de « tableaux F », avec les degrés de liberté et α de niveau approprié. On considère qu'il existe une différence importante entre les moyennes quand la statistique F calculée est suffisamment grande, c.-à-d. plus grande que la valeur critique donnée par la table F.

Nous décrivons ici un cas approprié au plan d'échantillonnage le plus simple, c.-à-d. que nous supposons des échantillons de taille égale et un échantillonnage aléatoire simple dans chaque groupe. Ce n'est pas une situation typique des grandes enquêtes et cette stratégie n'est pas efficace dans les applications pratiques des plans expérimentaux.

Le lecteur intéressé par l'ANOVA peut consultez des ouvrages d'introduction à la statistique (p. ex., Lohr (1999), Wonnacott et Wonnacott (1977)), ou des ouvrages sur les plans expérimentaux (Box, Hunter, Hunter (1978)).

11.4.2.3.2. Régression linéaire

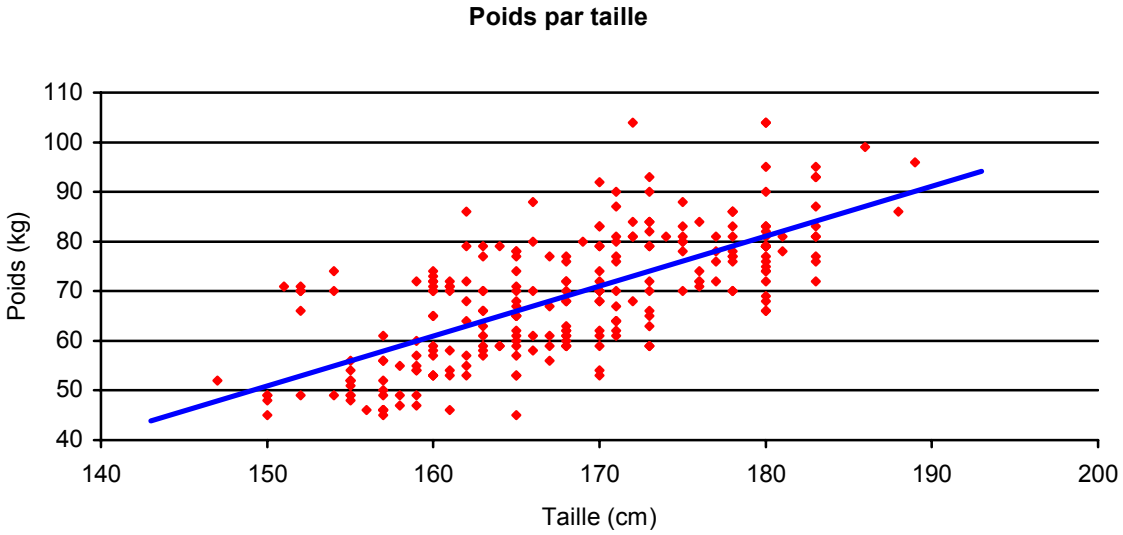
La *régression linéaire* est probablement le modèle linéaire le mieux connu. L'ANOVA aide à déterminer si la moyenne d'un groupe est très différente des autres et la régression sert à identifier ou modéliser les liens entre les différentes moyennes de groupe. Faire des prédictions ou des prévisions de la variable de réponse pour les valeurs des variables explicatives connexes connues est une autre application de la régression linéaire. La variable de réponse est habituellement une variable continue (p. ex., âge, poids, taille) en régression linéaire et les variables explicatives peuvent être qualitatives ou quantitatives. Si une seule variable explicative est utilisée, la régression est *simple* et si plusieurs sont utilisées, elle est *multiple*.

Supposons par exemple qu'une enquête a été faite pour obtenir des données sur la taille et le poids, et l'analyste est intéressé à déterminer comment ces variables sont liées. Compte tenu du graphique de données suivant, il semble y avoir un lien linéaire entre les deux variables.

Le modèle mathématique de ce lien est exprimé ainsi :

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

où y_i est la valeur de la variable de réponse continue, le poids, pour la i^e unité, x_i est la valeur de la variable explicative, la taille, pour la i^e unité, β_0 est l'ordonnée à l'origine (valeur de y lorsque x_i=0), β_1 est la pente de la ligne (le changement en y_i pour un changement d'une unité en x_i), ε_i est une variable d'erreur aléatoire, de moyenne nulle et de variance σ^2. Autrement dit, on suppose que y_i est approximativement lié linéairement à x_i et que les valeurs observées de y_i dévient d'un nombre aléatoire, ε_i, au-dessus et au-dessous de cette ligne. β_0 et β_1 sont les paramètres inconnus estimés à l'aide des données de l'échantillon. Afin de déterminer si les deux variables sont liées linéairement ou non, les intervalles de confiance peuvent être établis pour β_1 et les tests d'hypothèses peuvent être faits au sujet de sa vraie valeur.



Les estimations de paramètres peuvent être déterminées à l'aide des données observées (en supposant ici un échantillonnage aléatoire simple), comme suit :

$$\hat{\beta}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}, \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}.$$

Bien entendu, On peut aussi obtenir les erreurs d'échantillonnage de ces estimations. Le test est appliqué à β_1 pour déterminer si le lien est significatif, c'est-à-dire si la ligne n'est pas horizontale, ou $\beta_1 \neq 0$. Voici les hypothèses nulle et alternative :

$$\begin{cases} H_0 : \beta_1 = 0 \\ H_1 : \beta_1 \neq 0 \end{cases},$$

la statistique du test est la z bien connue :

$$z = \frac{\hat{\beta}_1}{S\hat{E}(\hat{\beta}_1)}$$

qui a une distribution type normale, compte tenu des habituelles hypothèses d'asymptoticité. Le critère de décision est identique à celui observé auparavant, c.-à-d. que l'on rejette H_0 si les valeurs de z sont à l'intérieur de la région de rejet pour un niveau α choisi.

Exemple 11.5 : Régression linéaire pour les données sur la taille et le poids, cas d'un EAS

Supposons que les données sur la taille et le poids ci-dessus ont été obtenues à l'aide d'un EAS et que les estimations suivantes ont été calculées :

Tableau 8 : Valeurs estimées pour β_o et β_1

	Estimation	SÊ(estimation)	z
$\hat{\beta}_0$	-90,88	7,66	
$\hat{\beta}_1$	0,95	0,04	21,09

L'analyste conclurait qu'il y a un lien important entre le poids et la taille des gens dans la population.

La régression est aussi utilisée dans les enquêtes pendant l'estimation ou l'imputation pour améliorer la qualité des estimations (consultez le **Chapitre 7 - Estimation** et le **Chapitre 10 - Traitement**). Draper et Smith (1981) expliquent en détail la théorie et les applications des modèles de régression linéaire.

11.4.3 Données d'enquête complexe

11.4.3.1 Test pour une seule moyenne

Le test pour une moyenne peut être facilement étendu à des données d'enquête complexe. Les exigences asymptotiques pour le test sont en effet couvertes par la version pour population finie du théorème central limite. L'estimation exacte de l'erreur d'échantillonnage de l'estimateur de la moyenne (c.-à-d. tenant compte de la stratification des données et des effets de grappe) remplace les exigences traditionnelles, à savoir que les données doivent être indépendantes et identiquement distribuées.

11.4.3.2 Comparaison entre de nombreuses moyennes de sous-populations : adaptation de l'ANOVA et de la régression

Des modèles d'ANOVA et de régression peuvent être appliqués aux populations et des analyses par inférence peuvent être faites lorsque des échantillons de plans d'enquête complexes sont tirés de ces populations. Peu d'ouvrages traitent de l'estimation et du test des paramètres des modèles d'ANOVA et de régression linéaire avec plans complexes. Ceux qui le font les regroupent en modèles linéaires et les traitent simultanément.

L'intérêt théorique et les difficultés dépassent la portée de ce manuel. Le lecteur devrait maintenant très bien savoir que les trousses statistiques ordinaires ne tiendront pas compte correctement des complexités éventuelles du plan d'intérêt et donneront probablement des résultats trompeurs. Le lecteur intéressé devrait consulter le Chapitre 8 de Thompson (1992), le Chapitre 7 de Särndal *et coll.* (1992), et le Chapitre 8 de Lehtonen et Pahkinen (1995).

11.5 Tests d'hypothèses sur une population : variables discrètes

Nous avons étudié à la Section 11.4 les test d'hypothèses à propos de variables continues. L'analyse des données nominales est fréquente (p. ex., analyse des dénombrements pour différentes catégories). Dans la distribution conjointe au Tableau 3, par exemple, l'analyste voudra peut-être vérifier si la proportion de la population de travailleurs de bureau qui a une tension artérielle basse est différente de la proportion de travailleurs manuels qui a une tension artérielle basse.

11.5.1 Tests d'indépendance et d'homogénéité avec données d'enquête simple

Les liens entre les variables discrètes d'une population, en particulier les variables discrètes ayant un petit nombre de valeurs distinctes, sont souvent examinés et mis à l'essai à l'aide de tableaux de contingence d'effectifs et de proportions.

11.5.1.1 Tests d'indépendance

Dans un *tableau de contingence* à deux entrées, il est souvent intéressant de déterminer si deux caractéristiques qui définissent les lignes et les colonnes du tableau sont indépendantes. Disons que la variable *A,* ayant *r* valeurs différentes est la caractéristique définissant les lignes du tableau et la variable *B* ayant *c* valeurs différentes est la caractéristique définissant les colonnes. Les proportions observées (ou effectifs) dans l'échantillon sont affichées dans un tableau $r \times c$, la valeur dans la i^e ligne et la j^e colonne étant la proportion (ou nombre) de particuliers qui ont simultanément la valeur *i* pour la variable *A* et la valeur *j* pour la variable *B*.

Tableau 9 : Effectifs observés dans un tableau de contingence à deux entrées ayant *r* lignes et *c* colonnes

Variable A	Variable B				Tailles d'échantillon
	1	2	…	*c*	
1	n_{11}	n_{12}		n_{1c}	n_{1+}
2	n_{21}	n_{22}		n_{2c}	n_{2+}
3					
…					
r	n_{r1}	n_{r2}		n_{rc}	n_{r+}
	n_{+1}	n_{+2}		n_{+c}	n_{++}

Disons que p_{ij} représente la proportion de la population dans la case (i, j) et p_{i+} et p_{+j} représentent les proportions de la i^e ligne de la j^e colonne respectivement. Leur estimateur est $\hat{p}_{ij} = \frac{n_{ij}}{n_{++}}$, $\hat{p}_{i+} = \frac{n_{i+}}{n_{++}}$ et $\hat{p}_{+j} = \frac{n_{+j}}{n_{++}}$, respectivement. Les hypothèses d'indépendance à vérifier peuvent ensuite être formulées comme suit :

$$\begin{cases} H_0 : p_{ij} = p_{i+} p_{+j}, \quad i = 1...r; j = 1...c \\ H_1 : p_{ij} \neq p_{i+} p_{+j}, \quad \text{pour certaines } i \text{ et } j \end{cases}.$$

Remarquons que $\sum_i \sum_j p_{ij} = 1$ parce que chaque individu de la population à l'étude fait partie d'une case seulement. Pour un ensemble de données obtenues en appliquant à la population un plan d'échantillonnage aléatoire simple, les tests d'indépendance reposent ou bien sur la statistique X^2 de Pearson :

$$X^2 = n \sum_{i,j} \frac{\left(\hat{p}_{ij} - \hat{p}_{i+} \hat{p}_{+j}\right)^2}{\hat{p}_{i+} \hat{p}_{+j}},$$

ou bien sur le rapport de vraisemblance G^2 :

$$G^2 = 2n\sum_{i,j} \hat{p}_{ij} \ln\left(\frac{\hat{p}_{ij}}{\hat{p}_{i+}\hat{p}_{+j}}\right), \text{et } \hat{p}_{ij} = \frac{n_{ij}}{n_{++}},$$

où n_{ij} est le dénombrement de l'échantillon dans la case (i, j) et $\hat{p}_{ij}$ est la proportion de l'échantillon correspondante.

Compte tenu de H_o et des autres hypothèses sur l'échantillonnage, les deux statistiques ont une distribution de khi carré asymptotique à $(r\text{-}1)(c\text{-}1)$ degrés de liberté. L'hypothèse nulle est rejetée pour un niveau de signification α donné si X^2 (ou G^2) est plus grand que la valeur critique totalisée $\chi^2_{(1-\alpha);(r-1)(c-1)}$.

Exemple 11.6 : Test d'indépendance pour les données d'un tableau de contingence tirées d'un EAS

On veut vérifier si la fréquence de lecture du journal quotidien (caractéristique *A*, i=*1* pour *chaque jour*, *i*=2 pour *parfois,* i=*3* pour *jamais*) est indépendante des opinions politiques du lecteur (caractéristique *B*, *j*=1 pour *extrême droite*, *j*=2 pour *droite modérée*, *j*=3 pour *gauche modérée*, *j*=4 pour *extrême gauche*). Supposons qu'un échantillon aléatoire simple de n=500 est sélectionné et que les deux caractéristiques sont mesurées pour tous les individus. Les résultats de l'enquête sont :

Tableau 10 : Estimations de l'enquête

Lecture du journal…		**Opinion politique**				
		Gauche		Droite		
		Extrême	Modérée	Modérée	Extrême	**Total**
Chaque jour	Effectif	n_{11} = 35	50	36	6	127
	Proportion(%)	$\hat{p}_{11}$ = 7,0	10,0	7,2	1,2	$\hat{p}_{1+}$ = 25,4
Parfois	Effectif	46	124	72	16	258
	Proportion(%)	9,2	24,8	14,40	3,2	51,6
Jamais	Effectif	28	50	33	4	115
	Proportion(%)	5,6	10,0	6,6	0,8	23,0
Total	Effectif	109	224	141	26	500
	Proportion(%)	$\hat{p}_{+1}$ = 21,8	44,8	28,2	5,2	100,0

Les résultats des tests :

Tableau 11 : Statistiques du test

Variable	**Df**	**Valeur**	**valeur p**
X^2 de Pearson	6	6,86	0,334
Rapport des vraisemblances G^2	6	6,90	0,329

Étant donné que les valeurs des tests sont bien inférieures à la valeur critique pour α=0,05, $\chi^2_{0,95;6}$= 12,59, l'évidence statistique n'est pas suffisante pour confirmer que les opinions politiques et la fréquence de lecture d'un journal sont liées. D'autre part, on peut comparer la probabilité d'obtenir un résultat au moins aussi extrême que celui obtenu (p. ex., $\Pr(X^2 \geq 6{,}86) = 0{,}334$) avec α, le niveau du test (ici, $\alpha = 0{,}05$). Cette probabilité est intitulée *valeur p*. Si la *valeur p* est plus grande que α, en supposant toujours que H_0 est vraie, on devrait affirmer que ce qui a été observé n'était pas suffisamment extrême pour rejeter l'hypothèse nulle.

11.5.1.2 Tests d'homogénéité

Un autre cas simple avec variables discrètes est le test d'homogénéité des proportions entre les populations lorsqu'un échantillon indépendant a été sélectionné dans chaque population. La comparaison entre les provinces de la proportion de personnes qui ne parlent aucune langue officielle, qui en parlent une ou les deux, par exemple, à l'aide d'une enquête nationale et d'échantillons indépendants dans chaque province, serait un test d'homogénéité.

Supposons dans cette situation que r populations sont comparées. Disons que $p_{j|i}$ est une proportion d'individus dans la i^e population ayant la j^e valeur d'une variable discrète de c catégories. Étant donné que chaque particulier de la i^e population doit être dans l'une des c catégories, $\sum_j p_{j|i} = 1$. Voici l'hypothèse à vérifier :

$$\begin{cases} H_0 : p_{j|1} = p_{j|2} = \ldots = p_{j|r}, \quad j = 1 \ldots c \\ H_1 : p_{j|i} \neq p_{j|k}, \quad \text{pour certaines } i \text{ et } k \text{ et pour au moins une } j \end{cases}.$$

Supposons que des échantillons aléatoires simples indépendants de taille n_{i+}, i=1, …, r, sont choisis dans chaque population. Disons que n_{ij} est le nombre de particuliers dans la i^e population ayant la j^e valeur de la variable discrète. Évidemment, $\sum_j n_{ij} = n_{i+}$, la taille de l'échantillon. Les dénombrements peuvent être entrés dans un tableau $r \times c$ et la proportion $p_{j|i}$ peut être estimée par $\hat{p}_{j|i} = \dfrac{n_{ij}}{n_{i+}}$.

La variable de Pearson pour un test d'homogénéité s'écrit :

$$X^2_{(H)} = n \sum_i^r \frac{n_{i+}}{n_{++}} \sum_j^c \frac{\left(\hat{p}_{j|i} - \hat{p}_{+j}\right)^2}{\hat{p}_{+j}}, \quad \text{où} \quad \hat{p}_{+j} = \frac{n_{+j}}{n_{++}}.$$

Il y a aussi une variable correspondante du rapport de vraisemblance.

Des calculs directs révéleront que $X^2_{(H)}$ peut aussi être formulée comme suit :

$$X^2_{(H)} = n \sum_i^r \sum_j^c \frac{\left(\hat{p}_{ij} - \hat{p}_{i+}\hat{p}_{+j}\right)^2}{\hat{p}_{i+}\hat{p}_{+j}},$$

la formulation étant semblable à la variable X^2 de Pearson pour le test d'indépendance. Selon l'hypothèse nulle de l'homogénéité, $X^2_{(H)}$ a aussi une distribution de khi carré asymptotique à (r-1)(c-1) degrés de liberté.

Exemple 11.6 (suite) : Test d'homogénéité

Dans l'enquête sur les opinions politiques examinées ci-dessus, au lieu d'un échantillon aléatoire simple, supposons que nous avons quatre échantillons aléatoires simples indépendants, chacun pour un groupe d'opinion politique différent. Le test d'homogénéité consisterait à vérifier si la fréquence de lecture du journal est la même pour chaque groupe politique.

11.5.1.3 Application de modèles log-linéaires lors de tests d'hypothèses

Les liens entre les proportions dans les cases d'un tableau de contingence peuvent souvent être exprimés sous forme d'un modèle linéaire logarithmique. Dans un tableau à deux entrées, par exemple, un modèle linéaire logarithmique saturé prend la forme d'un modèle ANOVA à deux facteurs avec interaction :

$$\ln(p_{ij}) = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}$$

$$\text{et} \quad \sum \alpha_i = \sum \beta_j = 0$$

$$\sum_i (\alpha\beta)_{ij} = \sum_j (\alpha\beta)_{ij} = 0$$

L'hypothèse d'indépendance précédente est équivalente au test de l'absence d'interaction et peut être reformulée ainsi :

$$H_0 : (\alpha\beta)_{ij} = 0, \forall i, j.$$

De nombreuses hypothèses différentes au sujet des paramètres des modèles log-linéaires peuvent être formulées et à chacune correspond une statistique donnée. On les retrouvera, notamment, dans Agresti (1996). La statistique du test pour vérifier l'indépendance (qui n'est pas précisée ici) a une distribution de khi carré asymptotique à $(r\text{-}1)(c\text{-}1)$ degrés de liberté lorsque l'hypothèse est vraie.

11.5.2 Tests d'indépendance et d'homogénéité avec données d'enquête complexe

Des tests de propriété, comme l'indépendance ou l'homogénéité de variables discrètes dans la population, peuvent aussi être effectués à l'aide des données obtenues d'une enquête ayant un plan complexe. Les tests faits dans une enquête simple ne sont cependant pas applicables au plan complexe sans modification. Tout d'abord, un tableau de contingence de dénombrements d'échantillons ou de proportions simples ignorant les probabilités de sélection ne donnera pas un aperçu précis du lien entre les variables discrètes qui déterminent les cases du tableau. De même, l'utilisation sans modification des variables à tester pour l'indépendance et l'homogénéité développées en 11.5.1.1 pourraient donner des conclusions inexactes (parce que ces variables à tester ne suivent plus une distribution de khi carré centrale lorsque l'hypothèse est vraie). Dans la matière qui suit, seule le test d'indépendance sera considéré, mais des approches semblables sont disponibles pour le test d'homogénéité.

De nombreuses approches différentes ont été proposées pour tenir compte d'un plan d'enquête complexe dans un test d'indépendance. Thomas *et coll.* (1996) décrivent plus de 25 méthodes et donnent une bibliographie approfondie, ils comparent aussi les résultats de ces méthodes à l'aide d'une étude de simulation. Lohr (1999) donne un compte rendu clair des principales méthodes intégrées à des logiciels pour analyse des données d'enquêtes complexes.

Une catégorie d'approches consiste à apporter des ajustements aux statistiques semblables à celles de Pearson et du rapport de vraisemblance décrites ci-dessus pour les données de plans d'enquête simples. La première étape de ces approches consiste à modifier les statistiques X^2 et G^2 définies en 11.5.1.1 en remplaçant $\hat{p}_{ij} = \dfrac{n_{ij}}{n_{++}}$ par sa version pondérée $\hat{p}_{ij} = \dfrac{\sum_{k \in S} w_k y_{kij}}{\sum_{k \in S} w_k}$

où $y_{kij} = \begin{cases} 1, & \text{si } y_k \text{ est dans la case } (i,j) \\ 0, & \text{autrement} \end{cases}$

et w_k est le poids du k^e individu échantillonné.

Cette modification seule n'est pas suffisante, compte tenu de H_0, parce que ces variables modifiées qui seront X^2_m et G^2_m ne suivent pas la distribution $\chi^2_{(r\text{-}1)(c\text{-}1)}$. D'autres ajustements sont nécessaires, par exemple, une multiplication des variables X^2_m et G^2_m par une constante pour obtenir des variables qui peuvent suivre approximativement une distribution de khi carré. Deux ajustements bien connus qui ont été intégrés à certains progiciels sont décrits ci-dessous.

Les corrections de premier ordre apportées à X^2_m et G^2_m (Rao et Scott (1981) (1984)), souvent intitulées « *corrections de premier ordre de Rao-Scott* », consistent à faire correspondre la moyenne asymptotique des statistiques de test à la moyenne d'une distribution $\chi^2_{(r\text{-}1)(c\text{-}1)}$. Les statistiques corrigées sont exprimées ainsi : $X^2_{RS} = \frac{X^2_m}{\hat{\delta}}$ et $G^2_{RS} = \frac{G^2_m}{\hat{\delta}}$, où $\hat{\delta}$ est une fonction des effets de plan pour estimer les proportions conjointes p_{ij} et les proportions marginales p_{i+} et p_{+j}. La correction exige donc la capacité de faire une estimation de la variance pour les proportions estimées comprises dans les formules qui s'appliquent à X^2_m et G^2_m. X^2_{RS} et G^2_{RS} peuvent ensuite être comparées à une distribution $\chi^2_{(r\text{-}1)(c\text{-}1)}$.

Les corrections de premier ordre ajustent seulement X^2_m et G^2_m, de sorte que leurs moyennes sont les mêmes que celle d'une variable aléatoire avec distribution $\chi^2_{(r\text{-}1)(c\text{-}1)}$. Rao et Scott (1981) et (1984) ont aussi proposé une correction de deuxième ordre, souvent intitulée « *correction de Satterthwaite* », qui fait correspondre les moyennes et la variance de la statistique du test à la moyenne et à la variance d'une distribution χ^2. Cette correction de deuxième ordre est peut-être plus difficile à calculer que la correction de premier ordre, mais le résultat peut être meilleur si les effets du plan d'échantillonnage varient énormément d'une case du tableau à l'autre.

Exemple 11.7 : Test d'indépendance pour les données d'un échantillon stratifié par grappes (Lohr, 1999, p. 332-334)

Au cours d'une enquête sur les jeunes et la criminalité (*Survey of Youth in Custody - Enquête sur le placement sous garde des jeunes*) du Département de la justice des É.-U. en 1987, on a sélectionné un échantillon de 2 621 adolescents et jeunes adultes résidant dans des établissements de longue durée pour les jeunes sous la gouverne de l'État. Il s'agit d'un échantillon stratifié par grappes avec probabilités inégales sélectionné dans 52 établissements. Les interviews ont permis d'obtenir de l'information sur le contexte familial, les antécédents criminels et la consommation de drogue et d'alcool. À l'aide des données de l'enquête, le tableau suivant a été dressé pour établir un lien possible entre l'âge et le caractère violent ou non de l'infraction criminelle. Voici les proportions pondérées :

Tableau 12 : Proportions estimées (à l'aide des pondérations de l'enquête)

		Groupe d'âge			
		≤ 15	16 ou 17	≥ 18	
Infraction avec violence?	Non	0,1698	0,2616	0,1275	0,5589
	Oui	0,1107	0,1851	0,1453	0,4411
		0,2805	0,4467	0,2728	1,0000

Si le plan d'échantillonnage par grappes et les probabilités de sélection inégales avaient été omis, les proportions estimées auraient été les suivantes :

Tableau 13 : Proportions non pondérées

		Groupe d'âge			
		≤ 15	16 ou 17	≥ 18	
Infraction avec violence?	Non	0,1389	0,2823	0,1328	0,5540
	Oui	0,0908	0,1969	0,1583	0,4460
		0,2297	0,4792	0,2911	1,0000

Il est évident que le plan d'échantillonnage a un effet sur les estimations et qu'il ne peut être négligé ou rejeté.

De même, la simple statistique du test de Pearson pour l'indépendance définie en 11.5.1.1 aurait une valeur de 34. Étant donné que la valeur critique de $\chi^2_{(2-1)(3-1)}$ est 5,99 au niveau de 5 %, l'hypothèse de l'indépendance serait rejetée. Il est maintenant évident que les contrevenants ne sont pas distribués au hasard entre les établissements correctionnels. En particulier, tous les résidents de l'établissement n^o 31 sont des délinquants violents. Les effets de grappes peuvent être constatés après avoir déterminé les effets du plan d'enquête pour le tableau précédent :

Tableau 14 : Effets du plan d'échantillonnage

		Groupe d'âge			
		≤ 15	16 ou 17	≥ 18	
Infraction avec violence?	Non	20,2	1,9	2,8	5,7
	Oui	5,3	8,4	2,4	5,7
		22,0	9,7	4,3	

La statistique du test ajustée du premier ordre a une valeur de X^2_{RS} = 16,2. Les effets de plan sont remarquables, même si la conclusion est la même.

11.6 Sommaire

L'analyse des données est l'une des étapes les plus délicates d'une enquête parce que la qualité de l'analyse et la méthode de communication efficace peuvent avoir des répercussions substantielles sur l'utilité de l'enquête dans l'ensemble. L'analyse des données devrait établir un lien entre les résultats de l'enquête et les questions et préoccupations identifiées au cours de la première étape de l'enquête.

L'analyse des données peut être restreinte aux données de l'enquête seulement ou elle peut comprendre une comparaison entre les résultats de l'enquête et les résultats tirés d'autres enquêtes ou sources de données. L'analyse consiste souvent à examiner des tableaux et graphiques de diverses mesures de récapitulation, notamment, les distributions de fréquences, les moyennes et les étendues. D'autres genres d'analyses de données plus perfectionnées sont aussi possibles, et l'inférence statistique peut être appliquée pour vérifier des hypothèses ou étudier des liens entre des caractéristiques.

Il faut correctement tenir compte du plan d'échantillonnage pour formuler des inférences au sujet de la population. Bien qu'on puisse obtenir des formules normalisées dans des ouvrages statistiques pour les données d'enquête simple, il est préférable de consulter un spécialiste si les données sont plus complexes.

Bibliographie

Agresti, A. 1996. *An Introduction to Categorical Data Analysis*, John Wiley and Sons. New York.

Aldrich, J.H. et F.D. Nelson. 1984. *Linear probability, Logit and Probit Models*, Quantitative Applications in the Social Sciences Series. 07-045. Sage Publications, California.

Ardilly, P. 1994. *Les Techniques de sondage*. Editions Technip, Paris.

Bausch, T. et U. Bankhofer. 1992. Statistical Software Packages for PCs - A Market Survey. *Statistical Papers [anciennement: Statistischen Hefte],* 33: 283-306.

Binder, D.A. 1984. Analyse de données qualitatives d'enquêtes complexes: quelques expériences canadiennes. *Techniques d'enquête,* 10(2): 155-170.

Box, G.E.P., Hunter, W.G. et J.S. Hunter. 1978. *Statistics for Experimenters.* John Wiley and Sons, New York.

Bouroche, J.-M. et G. Saporta. 1980. *L'Analyse des données*. Collection Que sais-je? 1854, Presses Universitaires de France, Paris.

Brogan, D.J. 1998. Pitfalls of Using Standard Statistical Software Packages for Sample Survey Data. *Encyclopedia of Biostatistics*. John Wiley and Sons, New York.

Brackstone, G. 1999. La gestion de la qualité des données dans un bureau de statistique. *Techniques d'enquête,* 25(2):159-171.

Carlson, B.L. 1998. Software for Statistical Analysis of Sample Survey Data. *Encyclopedia of Biostatistics.* John Wiley and Sons, New York.

Chambers, R.L. and C.J. Skinner. 2003. *Analysis of Survey Data.* John Wiley and Sons.

Cohen, S. B. 1997. An Evaluation of Alternative PC-Based Packages for the Analysis of Complex Survey Data. *The American Statistician,* 51: 285-292.

Draper, N.R. et H. Smith. 1981. *Applied Regression Analysis*. Second Edition. John Wiley and Sons, New York.

Dubois, J.-L. et D. Blaizeau. 1989. *Connaître les conditions de vie des ménagesdans les pays en voie de développement : Analyser les résultats*. Collection Méthodologies. Ministère de la coopération et du développement, Paris.

Dufour, J. 1996. *Qualité des données à l'enquête sur la population active*. Statistique Canada. HSMD-96-002E/F.

Ehrenberg, A.S.C. 1982, *A Primer in Data Reduction – An Introductory Statistics Textbook*. John Wiley and Sons, Great Britain.

Everitt, B.S. 1998. *The Cambridge Dictionary of Statistics*. Cambridge University Press, United Kingdom.

Fellegi, I.P. 1980. Approximate Tests of Independence and Goodness of Fit Based on Stratified Multistage Samples. *Journal of the American Statistical Association,* 75: 261-268.

Fink, A. et J. Kosecoff. 1998. *How to Conduct Surveys: a Step-by-Step Guide*. Sage Publications, California.

Freund, J.E. et R.E. Walpole. 1987. *Mathematical Statistics*. Fourth edition. Prentice Hall, New Jersey.

Friendly, M. 1995. *Categorical Data Analysis with Graphics*. Statistical Consulting Service Short Course, York University, Toronto.

Hidiroglou, M.A. et J.N.K. Rao. 1987. Chi-squared Tests with Categorical Data from Complex Surveys, I and II. *Journal of Official Statistics,* 3: 117-140.

Holt, D., T.M.F. Smith et P.D. Winter. 1980. Regression Analysis of Data from Complex Surveys. *Journal of the Royal Statistical Society, Series A (General),* 143(4): 474-487.

Johnson, S., N.L. Kotz et C.B. Read. 1982. *Encyclopedia of Statistical Sciences.* John Wiley and Sons, Toronto.

Lee, E.S., R.N. Forthofer et R.J. Lorimor. 1989. *Analyzing Complex Survey Data*, Quantitative Applications in the Social Sciences Series. 07-071. Sage Publications, California.

Lehtonen, R. et E.J. Pahkinen. 1995. *Practical Methods for the Design and Analysis of Complex Surveys, Statistics in Practice*. John Wiley and Sons, New York.

Lepkowski, J. et J. Bowles. 1996. Logiciels pour ordinateurs personnels pour l'estimation des erreurs d'échantillonnage. *Statisticien d'enquêtes*, 35:12-20.

Levy, P. S. et S. Lemeshow. 1999. *Sampling of Population: Methods and Applications*. Third edition. John Wiley and Sons, New York.

Lohr, S. 1999. *Sampling: Design and Analysis*. Duxbury Press, U.S.A.

Mendenhall, W. 1991. *Introduction to Probability and Statistics*. Eighth edition. PWS-Kent Press, Boston.

Nathan, G. et D. Holt. 1980. The Effect of Survey Design on Regression Analysis. *Journal of the Royal Statistical Society, Series B (Methodological),* 42(3): 377-386.

Porkess, R. 1991. *The Harper Collins Dictionary of Statistics*. Harper Collins, New York.

Rao, J.N.K. et A.J. Scott. 1981.The Analysis of Categorical Data from Complex Sample Surveys: Chi-Squared Tests for Goodness of Fit and Independence in Two-Way Tables (in Applications). *Journal of the American Statistical Association,* 76(374): 221-230.

Rao, J.N.K. et A.J. Scott. 1987. On Simple Adjustments to Chi-square Tests with Sample Survey Data. *Annals of Statistics,* 15: 385-397.

Rao, J.N.K., S. Kumar et G. Roberts. 1989. Analyse de données d'enquête avec variables de réponse qualitatives: méthodes et logiciels. *Techniques d'enquête,* 15(1): 169-196.

Särndal, C.E., B. Swensson et J. Wretman. 1992. *Model Assisted Survey Sampling*. Springer-Verlag, New York.

Skinner, C.K., D. Holt et T.M.F. Smith. 1989. *Analysis of Complex Surveys*. John Wiley and Sons, Chichester.

Snedecor, G. et Cochran, W.G., 1989, *Statistical Methods*, Eighth edition, Iowa State University Press, Ames Iowa.

Steel, R.G.D. et J.H. Torrie. 1980. *Principles and Procedures of Statistics – A Biometrical Approach*, Second edition. McGraw-Hill, U.S.A.

Thompson, S. K., 1992, *Sampling*, John Wiley and Sons, New York.

Tillé, Y. 2001. *Théorie des sondages : Échantillonnage et estimation en populations finies*. Dunod, Paris.

Tukey, J.W. 1977. *Exploratory Data Analysis*. Addison-Wesley Publishing Company, Reading, MA.

Wonnacott, T.H. et R.J. Wonnacott. 1977. *Introductory Statistics*. John Wiley and Sons, New York.

Wonnacott, T.H. et R.J. Wonnacott. 1991. *Statistique: Économie - gestion - sciences – médecine.* Economica, Paris.

PUBLICATIONS ÉLECTRONIQUES
DISPONIBLES À

www.statcan.gc.ca

Chapitre 12 - Diffusion des données

12.0 Introduction

La majorité des gens jugent l'enquête complète en définitive selon les données ou les rapports diffusés. Nous considérons dans ce chapitre des moyens de diffuser les résultats de l'enquête aux utilisateurs et nous mettons un accent particulier sur l'une des principales méthodes de diffusion : un rapport d'enquête avec tableaux et graphiques.

La qualité des données doit faire l'objet d'une évaluation pour en informer les utilisateurs, afin qu'ils déterminent par eux-mêmes l'utilité des données. Cette mesure peut aussi donner des renseignements utiles pour améliorer l'enquête (si elle est répétée) ou d'autres enquêtes. Cette évaluation et le rapport connexe devraient comprendre une description des techniques d'enquête, ainsi que les mesures et les sources d'erreurs d'échantillonnage et non dues à l'échantillonnage.

Avant la diffusion des données, on s'assurera qu'elles ne violent pas l'anonymat des répondants. Ce processus, intitulé contrôle de la divulgation, peut se traduire par la suppression ou la modification de certaines données.

L'objectif de ce chapitre est de présenter différentes méthodes de diffusion, de donner des conseils sur l'organisation d'un rapport sur support papier et d'expliquer des méthodes de contrôle de la divulgation des données en tableau et des fichiers de microdonnées à grande diffusion.

12.1 Diffusion des données

La diffusion des données est la communication des données de l'enquête aux utilisateurs à l'aide de divers médias. La communication des résultats de l'enquête aux utilisateurs comprend la réduction d'un grand ensemble de renseignements en détails concis et importants, tout en indiquant les points forts et les points faibles des données. Lors de la prestation des résultats aux utilisateurs, il est important de vérifier si l'information est précise, complète, accessible, compréhensible, utilisable, actuelle, conforme aux exigences de la confidentialité et facturée correctement. Les responsables de la diffusion devraient exploiter les progrès technologiques pour permettre aux utilisateurs de traiter l'information statistique au moindre coût et avec efficience dans leur propre milieu de travail.

Annoncer d'avance les dates de diffusion des résultats de l'enquête permet de susciter l'intérêt, de rester neutre, et d'être perçu comme tel. Cependant, il faut prévoir une explication du retard en cas de circonstances imprévues.

Les données peuvent être diffusées à l'aide de divers médias : communiqué, interview à la télévision ou à la radio, réponse à une demande spéciale par télécopieur ou au téléphone, publication d'un document, microfiches, média électronique, y compris Internet, ou fichier de microdonnées à grande diffusion sur disque compact. (Un fichier de microdonnées à grande diffusion est un fichier anonyme qui contient les enregistrements individuels des réponses de chaque répondant au questionnaire.)

Plusieurs genres de rapports peuvent être publiés, notamment :
- un rapport principal de l'enquête qui comprend les méthodes, ainsi que les principales totalisations et constatations,
- un rapport d'analyse des données,

- un rapport d'évaluation de la qualité des données,
- un rapport sur les techniques d'enquête,
- des rapports spécialisés sur les procédures de traitement et de collecte des données, des études méthodologiques, etc.

La section suivante cible la méthode d'organisation d'un rapport d'enquête.

12.2 Principal rapport de l'enquête

Le rapport principal est l'un des produits les plus importants de l'enquête. C'est habituellement le premier rapport préparé et diffusé aux utilisateurs, et il contient donc de l'information sur les buts et les techniques de l'enquête, la documentation des concepts et définitions, ainsi que les principales totalisations et constatations. L'uniformité de la structure du rapport de l'enquête aide les utilisateurs à chercher et à trouver des renseignements particuliers sur l'enquête.

L'organisme statistique a probablement des politiques, normes et lignes directrices particulières sur la matière, l'organisation et la présentation de ces rapports. Compte tenu de ces points, voici une structure suggérée du rapport de l'enquête qui s'applique dans la plupart des situations.

i. Page titre

Cet élément est nécessaire. Les organismes statistiques élaborent habituellement une norme de mise en page qui comprend les logos et l'identification numérique pour les contrôles d'impression.

ii. Table des matières

Cet élément est nécessaire. Il aide les utilisateurs à trouver l'information voulue.

iii. Liste des tableaux et graphiques

Cet élément est nécessaire. De nombreux utilisateurs veulent consulter une représentation visuelle des résultats pour leur propre présentation ou pour comparer avec d'autres sources.

iv. Faits saillants ou sommaire

Cet élément est optionnel, mais fortement recommandé. Un sommaire de deux ou trois pages au plus révèle les constatations ou faits saillants les plus importants de l'enquête. Il s'agit d'une référence rapide pour ceux qui n'ont peut-être pas le temps d'étudier tous les détails du rapport principal, mais qui doivent connaître les points essentiels des constatations de l'enquête. Il donne parfois une brève description des objectifs de l'enquête, précise quand et où elle a eu lieu, et ajoute les principaux sujets couverts. Il devrait ensuite décrire, chacune dans un bref paragraphe, les constatations les plus intéressantes, en commençant par les résultats plus généraux pour mettre ensuite en évidence certaines constatations plus particulières ou imprévues. Le sommaire est parfois simplement une énumération en points des plus importantes constatations. Les faits saillants devraient être explicites. Voici les exemples : *La superficie totale des grandes cultures a diminué de 3 % depuis 1986, les interviews téléphoniques préoccupent davantage les répondants, la majorité ont affirmé qu'ils refuseraient de donner de l'information financière au téléphone.*

v. Introduction

L'introduction devrait donner de l'information contextuelle pertinente à l'élaboration de l'enquête, par exemple, les antécédents du projet, les commanditaires, les objectifs de l'étude, un aperçu de la méthodologie et la raison d'être du rapport. Elle peut aussi donner un synopsis des activités accomplies et des leçons apprises en termes généraux seulement parce que les résultats détaillés seront formulés dans les sections ultérieures. Elle donne un aperçu des sections à venir et des liens entre elles.

vi. Objectifs

Cette section est optionnelle. Si l'enquête comprend un grand nombre de clients ou d'utilisateurs et si elle couvre un large éventail de sujets, il serait bon d'avoir une section entièrement réservée à l'explication des objectifs de l'enquête. Ces détails sont cependant couverts dans l'introduction d'habitude.

vii. Corps du texte

Le corps du rapport est réparti en plusieurs sections. L'une des premières sections devrait donner la définition des concepts et des principales variables (davantage de détails peuvent être insérés en annexe) et expliquer les techniques d'enquête, les procédures de collecte, le traitement, etc. Les résultats de l'enquête et les totalisations suivent habituellement (y compris les mesures de la qualité dont la précision).

Tous les détails des principales conclusions se trouvent dans le corps du rapport. Chaque section qui présente les résultats devrait commencer par les constatations et résultats les plus importants suivis de renseignements plus détaillés. Les idées devraient être disposées logiquement par ordre d'importance. Les tableaux, les graphiques, ainsi que l'explication des résultats et de leur importance, se trouvent dans le corps du rapport. L'un des plus grands défis de la rédaction d'un rapport est de soutenir l'attention du lecteur. Les messages principaux devraient être disposés par ordre d'importance, aux fins de la lecture et de la compréhension. L'information devrait aussi être présentée en langage le plus simple possible pour les lecteurs ciblés.

viii. Conclusions

Cette section présente un synopsis des répercussions et des constatations. Toutes les conclusions ou les recommandations pertinentes sur l'intervention nécessaire devraient être entrées ici. Une analyse appropriée et la considération des répercussions éventuelles des erreurs d'échantillonnage et non dues à l'échantillonnage devraient soutenir les conclusions. Les organismes donateurs, les ministères qui financent l'enquête et les analystes stratégiques prendront sans doute les grandes décisions, et cette section offre une occasion unique à ceux qui sont le plus proches de l'enquête d'exposer leur compréhension des répercussions des leçons apprises.

ix. Recommandations

Cette section est optionnelle. Dans les rapports techniques, des recommandations peuvent être formulées pour résoudre des problèmes constatés pendant toute activité de l'enquête. Elles sont faites à l'avantage des intervenants d'autres enquêtes qui peuvent se trouver dans la même situation.

x. Bibliographie – liste des références

Toutes les références devraient être clairement identifiées.

xi. Personnes-ressources

Cette section est obligatoire. Il faut donner aux utilisateurs, dans toute enquête, un moyen de communication clair et direct avec une personne informée du projet. Il faut donner un numéro de téléphone, un numéro de télécopieur, une adresse de courrier électronique et une adresse postale. Il est de plus en plus fréquent de renvoyer à un site Web où l'information est téléchargée.

xii. Annexes

L'objectif des annexes est de donner une idée brève, mais précise, des sujets qui ne sont pas décrits dans le corps du rapport ou d'ajouter des détails essentiels qui alourdiraient trop le texte. Les annexes peuvent contenir des détails sur les objectifs de l'enquête, la population à l'étude et le questionnaire, d'autres détails sur les techniques d'enquête, des mesures supplémentaires de la qualité des données, y compris la formule appliquée pour estimer la variance d'échantillonnage, une description des essais statistiques, etc. Les procédures de collecte de données appliquées sur place sont parfois ajoutées (p. ex., la méthode de collecte des données, la formation et la supervision du personnel sur place). Le matériel ou le logiciel utilisé peut être mentionné, ainsi que de l'information sur la mise au point des systèmes informatiques.

12.2.1 Lignes directrices sur la rédaction

Le choix d'une présentation et d'un style appropriés pour le rapport dépend énormément de la clientèle cible et de l'objectif du rapport. Les rapports peuvent être rédigés pour le personnel de l'enquête, d'autres réalisateurs d'enquête, des analystes stratégiques et des spécialistes du sujet, des leaders politiques ou le grand public.

Poser une question intéressante, organiser logiquement les faits pour répondre à des questions et ajouter la réponse est une structure efficace souvent utilisée pour chaque section du corps du rapport.

Voici certaines lignes directrices sur la rédaction :

i. Expression claire et concise

Un bon rapport communique un certain nombre de messages particuliers, sans être encombré de détails inutiles. Un langage concis est souhaitable, mais il n'est pas toujours préférable d'être plus bref et la clarté devrait toujours avoir la préséance. Voici un exemple : l'expression *les fonds distincts de placement à long terme axés sur la retraite et l'actif des compagnies d'assurance-vie* n'est pas mauvaise, mais l'expression *l'actif des compagnies d'assurance-vie et leurs fonds distincts (placements à long terme axés sur la retraite)* est définitivement plus claire.

ii. Formulation active ou passive

Le sujet responsable de l'action qu'exprime le verbe est habituellement (et de préférence) mentionné en premier dans la phrase, par exemple, *Le Manitoba et l'Alberta ont enregistré les plus importantes augmentations provinciales des recettes monétaires pendant la période.* Voilà une formulation active qui donne à la phrase un caractère actif et convaincant, ainsi qu'une impression de confiance, que n'expriment pas les verbes à la forme passive. Comparons avec la même phrase rédigée à la forme passive : *Les augmentations les plus prononcées des recettes monétaires pour la période ont été enregistrées à l'échelon provincial au Manitoba et en Alberta.*

iii. Abréviations et acronymes

On utilisera les abréviations et les acronymes avec parcimonie et, dans le doute, on donnera l'expression au complet. Le lecteur n'en connaît peut-être pas la signification. À la première apparition, on donnera l'expression au complet et on ajoutera l'abréviation entre parenthèses, par exemple : *l'Indice des prix à la consommation (IPC)*. Dans le reste du texte, on pourra ensuite utiliser l'abréviation ou l'acronyme seulement.

iv. Terminologie conforme

La terminologie utilisée dans les divers éléments du rapport doit être uniforme. Si le titre et le texte font référence à *l'actif et au passif*, par exemple, le tableau ou le graphique ne devrait pas être intitulé *bilan*. Nous avons aussi expliqué au **Chapitre 2 - Formulation de l'énoncé des objectifs** que la terminologie est plus utile si elle est conforme à celle d'autres enquêtes.

v. Catégories résiduelles

Les catégories résiduelles sont souvent larges et dignes de mention, mais le terme *autre* est vague et ambigu. Définissez la catégorie ou identifiez ses composantes, si possible, au lieu de les intituler simplement *autre(s)*.

vi. Dates

Il faudrait éviter les références ambiguës aux dates, par exemple, l'an dernier ou le mois dernier. Il faudrait inscrire à la première mention le mois et l'année en particulier, par exemple, *la chute de près de 26 % des expéditions intérieures en juillet 1993 semble impressionnante, mais elle est comparable à la diminution de 23 % enregistrée en juillet 1992.*

vii. Période de référence

Il faudrait inscrire la période de référence immédiatement sous le principal titre descriptif de chaque diffusion et annonce de données, par exemple, *Enquête mensuelle sur les industries manufacturières, décembre 2002*. Si les données sont provisoires, il faudrait l'indiquer à la ligne de la période de référence pour éviter de répéter « provisoire » partout dans le texte.

viii. Ratios simples

Il est acceptable d'utiliser une demie, un tiers, un quart ou un cinquième pour exprimer les résultats. Les ratios suivants deviennent difficiles à comprendre. On s'efforcera de rester uniforme, en évitant de mélanger les ratios dans la même phrase. On utilisera des nombres entiers si possible, en décrivant, par exemple, *deux femmes sur trois, comparativement à un homme sur trois…* On utilisera des ratios simples, p. ex., *il y a deux fois plus de chances que les femmes …*

ix. Pourcentages

Les pourcentages sont donnés entre parenthèses, p. ex., *environ deux tiers (66 %) des avocats et un tiers (32 %) des dentistes…* et sont donnés en entiers, *p. ex., 45 % au lieu de 45,3 %*, sauf si une donnée plus détaillée est justifiée et précise. Il est préférable d'éviter de surcharger la phrase de pourcentages ou de catégories. Au lieu d'inscrire les résultats comme suit, par exemple, *…définitivement oui (17 %), probablement oui (25 %), probablement non (27 %) et définitivement non (14 %)…* il peut être plus clair

et plus simple de déclarer que les *répondants étaient répartis à peu près également entre oui (42 %) et non (41 %)*, si les détails sont inscrits dans un tableau.

x. Chiffres trop nombreux dans le texte : lecteur confus et message obscur

Voici un exemple de texte qui sème la confusion : *Le taux national d'infractions totales par tranche de 100 000 personnes a augmenté régulièrement de 1979 à 1981 pour afficher ensuite des diminutions annuelles consécutives entre 1982 et 1985. Une augmentation de 4,1 % a cependant été remarquée en 1986 comparativement à 1985, suivie d'une augmentation de 2,5 % entre 1986 et 1987. Comparativement à 1987, une diminution de 1,2 % a été enregistrée en 1988. Ce taux a augmenté de 9,1 % au cours de la période de 1979 à 1988.*

xi. Titres, rubriques et sous-titres

Le principal titre peut comprendre des références aux dates et aux années de base, par exemple, *Produit intérieur brut réel au coût des facteurs par branche d'activité, juillet 1993 (données provisoires)*. Il faudrait utiliser des sous-titres par la suite comme guide ou indication pour orienter le lecteur dans le texte, par exemple, *industries manufacturières*. Les sous-titres trop nombreux, trop longs et truffés de jargon perdent leur influence. Ils devraient donner des renseignements sur ce qui suit, et non pas être simplement des étiquettes dénuées d'information. Il faut veiller à ce que chaque mot et chaque sous-titre ait son importance.

xii. Services de communication

De nombreux organismes peuvent utiliser les services de professionnels des communications qui relèvent les difficultés pour les lecteurs ciblés et font des suggestions pour améliorer le texte. Ces intervenants peuvent examiner le rapport ou même aider à le préparer avant la rédaction. Les corrections et la révision de la traduction font partie des services.

12.2.2 Tableaux

Nous avons expliqué au **Chapitre 2 - Formulation de l'énoncé des objectifs** que les tableaux sont l'un des principaux résultats d'une enquête. Il faudrait considérer directement dans les tableaux l'objectif énoncé et les exigences particulières des produits de l'enquête. Les tableaux servent à illustrer ou examiner les caractéristiques quantitatives des données. Ils peuvent rapidement révéler les liens entre plusieurs variables et permettre la comparaison directe des sommes.

Pendant l'analyse et avant la documentation des résultats, l'analyste devrait vérifier les estimations et les tableaux produits. Les résultats sont-ils uniformes à l'interne? Cela signifie que les totaux marginaux au sujet des mêmes variables devraient être les mêmes dans différents tableaux. Les calculs des sous-populations devraient être équivalents au total de la population, etc. Les totaux correspondent-ils à ceux d'autres sources?

Les lignes directrices suivantes devraient être appliquées lors de la préparation des tableaux :

i. Les tableaux devraient être simples et afficher seulement les principaux renseignements pour justifier un point à la fois en général. Il vaut mieux avoir deux tableaux simples qu'un tableau trop compliqué.

ii. La mise en forme, les espaces et la formulation dans l'ensemble, la disposition et l'apparence des titres, les vedettes des lignes et des colonnes, ainsi que d'autres mesures d'étiquetage, devraient aider à mettre en évidence les données des tableaux et à empêcher les erreurs d'interprétation.

iii. Les tableaux devraient être clairs, logiques et uniformes.

iv. Les titres devraient être clairs et succincts et il faudrait éviter les abréviations.

v. La présentation des rubriques devrait soutenir le message de l'analyse dans un ordre rationnel et il faudrait énoncer clairement toutes les unités de mesure.

vi. La conception des tableaux devrait permettre, le plus possible, la lecture des photocopies. Il devrait aussi y avoir suffisamment d'information dans le tableau (titre, notes en bas de page, etc.) pour ne pas perdre sa valeur à la photocopie.

vii. Arrondir les données aidera le lecteur à comprendre la précision des estimations.

viii. Si les données sont tirées d'une enquête-échantillon, les estimations et les mesures de la précision devraient être faites à l'aide des poids finaux (c.-à-d. les pondérations du plan qui peuvent être ajustées pour les non-réponses et les données auxiliaires comme on l'a vu au **Chapitre 7 - Estimation**).

12.2.3 Graphiques

Les graphiques et diagrammes servent à la présentation visuelle des données. Ils ciblent les caractéristiques, formes ou distributions relatives, et les ordres de grandeur. De bons graphiques devraient ajouter au texte et aux tableaux, et non simplement répéter l'information. Ils devraient servir à expliquer ou soutenir les principaux points dans le texte. Celui-ci devrait reporter aux graphiques qui devraient être disposés après la référence.

La présentation des graphiques et diagrammes est aussi considérée au **Chapitre 11 - Analyse des données de l'enquête**. L'ensemble des règles élémentaires s'appliquent en majorité autant aux graphiques et diagrammes qu'aux tableaux. Ils devraient être simples et afficher seulement les principaux renseignements pour justifier un point à la fois en général. Les explications détaillées devraient être superflues. Il faudrait utiliser les graphiques pour diffuser une interprétation visuelle et intuitive des faits saillants ou des tendances, et ils devraient donc être épurés et ordonnés. Toute tentative de communiquer trop d'information dans un seul graphique ou tableau peut simplement semer la confusion chez les lecteurs. Si les graphiques sont utilisés pour illustrer des points sur la population, il faut avoir recours aux pondérations définitives de l'échantillon pour les données d'une enquête-échantillon.

Il faudrait prendre garde de ne pas induire le lecteur en erreur. Les graphiques sont très efficaces pour communiquer l'information, mais il est facile de les utiliser erronément. Les titres, légendes et axes étiquetés négligemment, le recours inapproprié aux espaces en gris, les échelles faussées ou non uniformes, etc., sont des erreurs fréquentes. Il faudrait en général éviter les présentations tridimensionnelles, mais elles peuvent être appropriées dans certaines circonstances (p. ex., le tracé d'une surface).

On devrait utiliser des intervalles uniformes pour les graphiques linéaires. Il vaut mieux éviter, par exemple, ce genre de progression : 1, 2, 5, 8, même s'il n'y a pas de données simples pour les autres

valeurs; il est préférable d'inscrire plutôt 1, 2, 3, 4, 5, 6, 7, 8. Si la série commence par un nombre élevé, ou s'il y a un écart dans les valeurs de l'un des axes, on ajoutera un symbole pour indiquer l'écart.

12.3 Rapport d'analyse des données

Le principal rapport de l'enquête comprend certains résultats et constatations élémentaires, mais il faudra faire un rapport d'analyse des données ou d'autres rapports si une analyse plus approfondie est prévue. Les activités analytiques devraient déboucher en bout de ligne sur un rapport sur support papier qui réponde clairement aux questions qui ont suscité l'enquête. Le rapport d'analyse des données peut être structuré en général comme le principal rapport de l'enquête présenté à la Section 12.2. Il devrait y avoir un sommaire des méthodes analytiques dans le rapport d'analyse des données, ainsi qu'une description et une considération des répercussions éventuelles des erreurs d'échantillonnage et non dues à l'échantillonnage, des mises en garde et des hypothèses sur les résultats et leur signification statistique.

Voir le **Chapitre 11 - Analyse des données de l'enquête** pour obtenir des détails sur les méthodes d'analyse des données.

12.4 Rapport sur la qualité des données

L'évaluation de la qualité des données est une étape importante de toute enquête et il faudrait la documenter dans le principal rapport de l'enquête ou dans un rapport distinct sur la qualité des données. ***L'évaluation de la qualité des données est le processus d'évaluation du produit définitif, compte tenu des objectifs originaux de l'activité statistique du point de vue de la précision ou de la fiabilité des données***. Ce genre d'information permet aux utilisateurs de procéder à une interprétation et à une utilisation mieux informées des résultats de l'enquête. Les utilisateurs doivent être en mesure d'évaluer à quel point les erreurs dans les données en restreignent l'utilisation, mais peu d'utilisateurs peuvent évaluer eux-mêmes la précision des données tirées d'une enquête. L'organisme statistique est donc chargé des évaluations de la qualité des données nécessaires et de la diffusion des résultats de ces évaluations aux utilisateurs au moment opportun et en présentation facile à utiliser. Les évaluations de la qualité des données sont aussi à l'avantage de l'organisme statistique. Dans la mesure où les erreurs peuvent être détectées à des étapes particulières du processus de l'enquête, ces évaluations peuvent servir à améliorer la qualité des occurrences ultérieures si l'enquête est réitérée ou s'il y a d'autres enquêtes semblables.

L'analyste devrait aussi considérer l'application de *méthodes d'attestation*. L'*attestation* comprend un examen approfondi des indicateurs de qualité des données, diverses analyses à une variable et à plusieurs variables et de nombreuses études comparatives, afin de comparer les résultats de l'enquête à d'autres sources et à des séries chronologiques. Toutes les données du recensement canadien passent, par exemple, par une évaluation et un examen critique rigoureux de la qualité pour en déterminer la pertinence et la fiabilité aux fins de la diffusion publique à des échelons particuliers du territoire de déclaration. La pertinence signifie que les données doivent répondre à des normes particulières de qualité et de confidentialité.

La documentation sur la qualité des données devrait comprendre l'information sur les techniques d'enquête et les indicateurs de qualité des données. Les éléments suivants déterminent la précision de la documentation sur la qualité des données nécessaire :

- le genre d'enquête (recensement, échantillon, données administratives, etc.) et la fréquence (unique ou réitérée),
- le genre de données tirées de la collecte,

- le genre d'analyse faite,
- les utilisations prévues des données (c.-à-d. répercussions sur les politiques, planification économique et sociale, etc.),
- l'éventualité d'erreurs et les répercussions sur l'utilisation des données,
- l'utilisation prévue du rapport sur la qualité (p. ex., information générale, améliorations ultérieures, etc.),
- l'auditoire ciblé dans le rapport sur la qualité,
- le moyen de diffusion (c.-à-d. publication, fichier de microdonnées à grande diffusion, etc.),
- le budget total du programme et le coût de l'évaluation de la qualité par rapport au coût dans l'ensemble.

Les éléments suivants donnent une liste partielle des articles qui peuvent être ajoutés à un document sur la qualité des données. Cette information devrait normalement être jointe au produit statistique. S'il est impossible d'ajouter cette documentation pour une raison ou pour une autre, il faudrait expliquer en référence comment trouver cette documentation sur la qualité des données.

i. Nota aux utilisateurs (le cas échéant)

Il faudrait ajouter cet élément le cas échéant. Il peut s'agir de faits saillants de l'information formulée dans une section suivante ou plus, d'explications particulières ou de mises en garde dont il faut informer les utilisateurs.

ii. Sources de données et méthodologie

Cette sous-section devrait couvrir les points suivants :
- la méthodologie générale (différences entre population cible et population observée, sources de données, méthodes de collecte, traitement, estimation et validation),
- la période de référence,
- les révisions, le cas échéant,
- les ajustements, le cas échéant.

iii. Concepts et variables mesurées

Cette sous-*section devrait couvrir les variables, classifications et concepts utilisés les plus importants.*

iv. Précision des données

Il devrait y avoir un énoncé sur les questions de précision, y compris la taille de l'échantillon, les genres d'erreurs non dues à l'échantillonnage et leurs sources (p. ex., taux de non-réponse, taux d'imputation, etc.) par région géographique et caractéristique.

v. Comparabilité des données et sources connexes

Cette sous-section indique, le cas échéant, si les données sont comparables ou non dans le temps et elle précise pourquoi (p. ex., en cas de modification de la formulation d'une question).

vi. Autres évaluations et indicateurs de qualité (le cas échéant)

Cette sous-section offre un sommaire des approches ou méthodes analytiques, pour tous les résultats analytiques, ainsi qu'une brève description et une considération des répercussions éventuelles des questions de précision, des hypothèses et des mises en garde sur les résultats et leur signification

statistique. Elle comprend aussi une description d'autres sources d'erreurs éventuelles importantes ou de tout autre événement (p. ex., une grève) qui peut éventuellement avoir des répercussions sur l'exactitude, l'actualité et l'interprétation ou l'utilisation des données.

vii. Annexes (si nécessaires)

12.5 Contrôle de la confidentialité et de la divulgation

La majorité des organismes statistiques doivent protéger la confidentialité de l'information du répondant en vertu de la loi. Voici certaines mesures qui garantissent la confidentialité :
- protéger les questionnaires pendant la collecte des données et l'acheminement,
- exiger que tous les employés prêtent serment de non-divulgation de l'information confidentielle,
- restreindre l'accès aux bâtiments et aux serveurs où sont sauvegardés les renseignements confidentiels,
- appliquer des méthodes de contrôle de la divulgation.

La protection de la confidentialité de l'information des répondants est essentielle à un organisme statistique pour maintenir la confiance du public et obtenir ainsi des taux de réponse élevés et des données de bonne qualité.

Le contrôle de la divulgation englobe l'ensemble des mesures prises pour protéger les données diffusées, afin d'empêcher les infractions contre l'anonymat des répondants. Il comprend, par exemple, la suppression de cases dans un tableau qui pourraient révéler de l'information confidentielle sur un répondant en particulier. L'application du contrôle de la divulgation a parfois des répercussions négatives sur la qualité des données parce qu'il faut supprimer ou modifier certaines données pendant le processus. Le but du contrôle de la divulgation est de garantir le respect de la confidentialité de l'information donnée par un répondant, tout en maintenant l'utilité des données dans la mesure du possible.

12.5.1 Divulgation

Deux principaux types de divulgation suscitent des préoccupations : la divulgation de l'identité et celle des attributs. ***Il y a divulgation d'identité lorsqu'un répondant en particulier (personne, ménage, entreprise, etc.) peut être identifié à partir des données diffusées. Une information confidentielle est ainsi révélée.*** Le problème se pose surtout dans le cas des microdonnées parce que l'identification de l'enregistrement d'un répondant débouche presque certainement sur la révélation des caractéristiques de ce répondant. ***Il y a divulgation d'attribut lorsqu'il est possible, à partir des données diffusées, de révéler l'information confidentielle au sujet d'un répondant.*** Le problème se pose surtout dans le cas du recensement ou des données administratives parce que l'erreur d'échantillonnage aide à protéger les résultats contre la divulgation. Un recensement peut, par exemple, donner une étendue étroite (précise) du revenu des médecins dans un certain secteur.

Il y a divers genres de divulgation qui comportent différents degrés de gravité. Les définitions suivantes ne sont pas mutuellement exclusives :

i. Données sur soi-même déduites par soi-même

Si un répondant peut déduire l'information qu'il a donnée, mais s'il est seul à pouvoir le faire, il n'y a donc pas de problème de divulgation. Il y a cependant perception d'un problème si le répondant a l'impression que d'autres peuvent aussi déduire l'information, même si ce n'est pas le cas. Voilà pourquoi

les organismes statistiques devraient essayer d'éviter de divulguer des résultats en une présentation qui permettrait cette occurrence.

ii. Données d'un répondant déduites par un tiers

Quelqu'un peut, dans ce cas, déduire l'information d'un répondant. Ce problème le plus grave est l'objet du contrôle de la divulgation.

iii. Données d'un tiers sur soi déduites par soi-même

Ce problème est particulier aux enquêtes à répondants multiples liés entre eux. Une enquête sur les enfants, par exemple, peut comprendre une section pour les parents, une pour les enseignants et une pour les enfants. Si un père peut s'identifier à l'aide d'un fichier de microdonnées, il peut déterminer quelles réponses ont donné ses enfants ou les enseignants.

iv. Constatation directe

La situation est possible si l'information confidentielle peut être déduite en observant simplement une case. Voici un exemple très simple : un tableau affiche les revenus moyens par profession pour un secteur donné. Si une profession comprend seulement une personne dans le secteur, son salaire est la moyenne. S'il y en a deux, chaque personne peut utiliser la moyenne pour déduire l'information sur l'autre (un cas de divulgation résiduelle).

v. Divulgation résiduelle

Il y a divulgation résiduelle si d'autres renseignements diffusés ou autrement disponibles permettent une estimation précise de l'information supprimée. Si une composante d'un total est supprimée, par exemple, il y a divulgation résiduelle parce que la composante manquante peut être estimée en soustrayant les autres composantes du total.

vi. Divulgation exacte

Une personne peut, dans ce cas, déduire la valeur exacte de l'information confidentielle. L'exemple donné au paragraphe de la constatation directe ci-dessus est un exemple de divulgation exacte.

vii. Divulgation approximative

Dans ce cas, une personne ne peut déduire la valeur exacte de l'information confidentielle, mais elle peut déterminer un intervalle qui pourrait fort probablement contenir la valeur confidentielle. Il y a divulgation si l'intervalle est suffisamment restreint pour causer éventuellement des actions préjudiciables au répondant. Si un fabricant domine, par exemple, la production totale d'un certain produit (disons plus de 95 %), la diffusion d'une estimation de la production totale de ce produit donne une estimation approximative de la production de ce fabricant.

12.5.2 Techniques de non-divulgation

Les méthodes de restriction de l'accès et de restriction des données sont deux approches de protection de la confidentialité des données. Les méthodes de restriction de l'accès empêchent ou restreignent l'accès aux données, notamment comme suit : l'accès à un emplacement ou à un serveur est limité au personnel autorisé, les fichiers sont protégés à l'aide d'un mot de passe ou du chiffrage, les données sont échangées

conformément aux modalités d'un contrat de licence, etc. Ces méthodes ne sont pas considérées dans ce texte. Les méthodes de restriction des données protègent les données elles-mêmes. Ces méthodes sont classées en méthodes de réduction des données (c.-à-d. que l'information diffusée est réduite) et en méthodes de perturbation des données (c.-à-d. que les données sont modifiées).

Les techniques élaborées pour éviter la divulgation varient selon le produit statistique, les trois produits les plus habituels étant les tableaux de fréquences (calculs ou calculs pondérés), les tableaux de données quantitatives (pour les données quantitatives) et les fichiers de microdonnées à grande diffusion. Les techniques sont décrites ci-dessous.

12.5.2.1 Protection des données des tableaux

Les cases d'un tableau de données quantitatives donnent des valeurs numériques (habituellement non négatives), par exemple, les moyennes, les totaux des valeurs en dollars ou le nombre d'employés. Ces tableaux posent un risque de divulgation, en particulier lorsque les données sont tirées d'une enquête-entreprise. Étant donné le caractère asymétrique des données des entreprises, les données d'une grande entreprise ou deux peuvent dominer certaines cases et la divulgation des valeurs de ces cases peut donner une estimation raisonnable des valeurs de ces importants répondants. La première étape est donc l'identification des cases à caractère délicat (c.-à-d. les cases à risque de divulgation). De nombreuses règles ont été élaborées pour identifier les cases à caractère délicat. Voici quelques exemples de règles qui définissent une case à caractère délicat dans un tableau de données quantitatives :

i. Règles de la limite

Il doit y avoir un nombre minimal de répondants dans une case (p. ex., au moins trois répondants). Sous ce seuil, elle devient une case à caractère délicat.

ii. La règle (n,k)

Une case est considérée à caractère délicat si trop peu de répondants englobent une trop grande partie du total de la case, c.-à-d. si les n plus importants répondants représentent au moins k % de la valeur totale de la case. Selon la règle (2,90), si les deux plus importants répondants représentent plus de 90 % du total de la case, il s'agit d'une case à caractère délicat.

iii. La règle p-pour cent

Nous avons une case à caractère délicat si sa diffusion permetait à quelqu'un d'estimer la contribution d'un répondant à moins de p-pour cent de sa valeur. Le risque devient maximal si le deuxième plus important répondant de la case essaie d'estimer la contribution du plus important répondant en soustrayant sa propre valeur du total diffusé. Supposons que la case comprend m répondants, la divulgation est équivalente si : $x_3 + ... + x_m < (p/100)x_1$, où x_1 est la valeur du plus important répondant, x_3 est la valeur du troisième répondant par ordre d'importance, etc.

Les règles (n,k) et p-pour cent sont deux exemples de règles appliquées pour identifier les cases où dominent une ou deux grandes unités. Ce ne sont pas les seuls choix de règles disponibles, il y en a d'autres. Les règles de Duffett sont parfois appliquées à Statistique Canada. Les règles de Duffett sont des ensembles de règles (n,k) et le nombre d'unités dans la case détermine les paramètres. Peu importe la ou les règle(s) appliquée(s), il vaut généralement mieux garder confidentielles les valeurs de leurs paramètres.

Les cases d'un tableau de fréquences donnent le nombre réel ou estimé d'unités ayant les caractéristiques de la case. Les tableaux de fréquences peuvent poser un risque de divulgation lorsqu'ils révèlent les caractéristiques d'un répondant. Nous avons mentionné auparavant que le risque de divulgation d'attributs est le plus grand dans le cas d'un recensement ou des données administratives, même si les données d'un échantillon peuvent poser des risques de divulgation lorsque les unités de l'échantillon des participants à l'enquête sont connues (p. ex., les autres membres du ménage). Trois problèmes éventuels sont les cases pleines, les cases dont le total est zéro et les cases de faibles fréquences.

Une case est pleine lorsqu'une seule catégorie de réponse englobe tous les répondants, par exemple, lorsqu'une seule case d'une ligne ou d'une colonne a une valeur différente de zéro. Les cases pleines posent un risque de divulgation si elles permettent d'obtenir de l'information confidentielle sur une sous-population. Un tableau peut révéler, par exemple, que tout le personnel de soutien d'une institution a un « diplôme d'études secondaires » à la case scolarité. Si un tableau de répartition des revenus les englobe sans exception dans la tranche « 20 000 $ à 29 999 $ », l'information sur la rémunération est alors divulguée.

Les cases de valeur zéro, c.-à-d. les cases sans unité ou dont le total est zéro, peuvent aussi poser un risque de divulgation pour des raisons semblables à celles des cases pleines (ces dernières sont le résultat des cases de valeur zéro). Le tableau ci-dessus peut révéler, par exemple, que le personnel de soutien est réparti en deux catégories de scolarité seulement : « études inachevées » et « diplôme d'études secondaires ». Un tableau de répartition des revenus dans un établissement dont les employés sont divisés en trois catégories de revenu : « de 20 000 $ à 29 999 $ », « de 30 000 $ à 39 999 » et « de 90 000 $ à 99 999 $ » peut donner une bonne estimation de la rémunération des ingénieurs de l'établissement (la dernière tranche).

Les cases de faibles fréquences comptent peu de répondants, p. ex., moins de trois ou cinq. Les cases de faibles fréquences peuvent poser un risque de divulgation si elles permettent l'identification de leurs répondants et révèlent certaines de leurs caractéristiques. Un tableau de fréquences d'un recensement dans un secteur restreint qui affiche, par exemple, deux ménages monoparentaux dont le chef est un divorcé peut identifier ces ménages. Les caractéristiques supplémentaires révélées sur les membres de la case peuvent constituer une divulgation d'attributs. Les cases de faible valeur posent un autre problème : elles peuvent donner l'impression qu'il y a eu divulgation, même si ce n'est pas le cas. S'il est révélé, par exemple, qu'il y a eu seulement un nouveau cas de cancer du côlon détecté à l'Île-du-Prince-Édouard en 2001, la situation ne révèle quand même rien sur l'intéressé.

Il y a plusieurs moyens de traiter les cases à caractère délicat.

1. Les méthodes de réduction des données comprennent les suivantes :

i. Regroupement de cases

Il s'agit de regrouper les catégories pour augmenter le nombre d'entrées par case (p. ex., réduire le niveau de détails fournis dans la branche d'activité). Cette méthode simple peut réduire de beaucoup l'information en supprimant les détails des données.

ii. Suppression de cases

Lorsque les cases à caractère délicat sont supprimées, il faut habituellement supprimer des cases sans caractère délicat pour éviter que les valeurs des cases à caractère délicat soient déduites du total marginal. Ces autres cases sont intitulées *cases de suppression complémentaire* et il y a de nombreuses règles pour choisir lesquelles supprimer. Le genre de variable et le degré de protection voulu déterminent le choix des

règles à appliquer à une case en particulier. La suppression complémentaire de cases peut être optimisée en minimisant le nombre de cases supprimées, la somme des valeurs des cases supprimées et le nombre de répondants supprimés, ou en appliquant une méthode qui se traduit par un compromis entre ces besoins. D'autres règles peuvent être appliquées, par exemple, la préférence peut être accordée à l'identification de suppressions complémentaires dans le même regroupement de branches d'activité.

2. Voici certaines méthodes de perturbation des données :

i. Arrondissement déterministe

Les données d'une case sont arrondies selon une règle déterministe (p. ex., arrondies à la baisse au multiple de 10 précédent si le dernier chiffre de l'unité est inférieur à cinq et arrondi à la hausse autrement). Cette mesure peut cependant donner un biais et l'équivalence entre les valeurs arrondies et les totaux marginaux arrondis peut être rompue.

ii. Arrondissement aléatoire

L'orientation de l'arrondissement est déterminée au hasard. Cette méthode offre une meilleure protection que l'arrondissement déterministe, la même base d'arrondissement étant utilisée, parce qu'il est plus difficile d'estimer la valeur originale. De plus, il n'y pas de biais, mais le maintien des totaux marginaux peut aussi être rompu.

iii. Arrondissement aléatoire contrôlé

L'arrondissement aléatoire contrôlé permet de conserver les marges agrégées définies d'avance. L'application de cette méthode aux tableaux multidimensionnels n'est pas une mince affaire. Il est possible de trouver des solutions pour les tableaux à trois dimensions au plus, mais il n'y en n'a pas pour les tableaux ayant davantage de dimensions.

iv. Ajout d'interférences

On peut ajouter des interférences aléatoires aux résultats des tableaux pour susciter davantage d'incertitude et diminuer le risque de divulgation.

v. Méthodes de contrôle de la divulgation des microdonnées (voir la section suivante).

Après avoir appliquer des méthodes de contrôle de la divulgation à un fichier de microdonnées, on peut ensuite procéder en toute sécurité à toutes les totalisations à partir de ce fichier.

Lorsque de multiples tableaux sont produits à partir de la même enquête, la protection de la confidentialité ne peut se faire indépendamment pour chaque tableau parce que la combinaison de l'information de différents tableaux peut déboucher sur la divulgation. Idéalement, il faut donc appliquer les techniques décrites ci-dessus en considérant les tableaux déjà publiés et ceux qui ne le sont pas encore. Les totalisations devraient être définies d'avance le plus tôt possible pour les tableaux de données quantitatives parce que la protection des tableaux spéciaux est particulièrement difficile. La combinaison de centaines de tableaux peut facilement donner des milliers ou des dizaines de milliers de cases et l'automatisation est donc nécessaire.

Si l'enquête est réitérée régulièrement, il faut élaborer avec une attention toute particulière une caractéristique de suppression à appliquer à la série complète. Cependant, après un certain nombre de

répétitions, la suppression d'une case ou d'un enregistrement peut devenir facultative (p. ex., la taille d'une case peut grossir jusqu'à ce qu'elle perde son caractère délicat). Le contraire est aussi possible.

12.5.2.2 Protection des fichiers de microdonnées à grande diffusion

Les fichiers de microdonnées à grande diffusion, qui contiennent des enregistrements individuels, sont diffusés parce qu'ils permettent aux utilisateurs de procéder à des analyses des données de l'enquête qui sont difficiles à faire à partir des résultats des tableaux. Les fichiers de microdonnées à grande diffusion ont un caractère très délicat et il faut en considérer la confidentialité de près avant la diffusion pour éviter de révéler l'identité des répondants. La divulgation des fichiers à grande diffusion cible habituellement les données tirées de l'échantillon seulement parce que le risque d'identification des répondants augmente avec le taux d'échantillonnage (dans le cas d'un recensement, un fichier à grande diffusion peut être divulgué pour un échantillon de répondants). Il y a habituellement des identificateurs directs ou personnels au fichier principal de l'enquête qui peuvent seuls identifier un particulier (p. ex., nom, adresse, numéro d'identification). Il faut les éliminer. Il y a aussi des identificateurs indirects, des variables qui peuvent servir à identifier les répondants (p. ex., secteur géographique, âge, profession, race, ou même revenu dans certains cas).

Les identificateurs indirects d'un fichier à grande diffusion sont examinés pour déterminer s'ils peuvent servir à identifier les répondants. Voici certaines vérifications :

i. Vérification des identificateurs indirects pour les enregistrements uniques.

On peut chercher, par exemple, des particuliers ayant des revenus très élevés ou des tailles de ménage exceptionnelles.

ii. Analyse des tableaux d'identificateurs indirects à deux et trois dimensions (p. ex., âge, sexe, scolarité, etc.).

On peut étudier, par exemple, les tableaux d'âge par sexe par degré de scolarité et y chercher les combinaisons uniques (p. ex., une personne très âgée qui a toujours un emploi).

D'autres vérifications peuvent être faites selon le genre d'enquête (structure de la population, plan d'échantillonnage, collecte ou utilisation des données, …). L'information sur le plan d'échantillonnage et les pondérations de l'enquête sont examinées, par exemple, pour vérifier si elles révèlent des renseignements à caractère délicat sur le secteur géographique des unités de l'échantillon. Si les données de l'enquête sont hiérarchiques (p. ex., ménage-personne), les liens entre les unités sont alors examinés (p. ex., recherche de combinaisons rares d'âges des conjoints). Si les données sont tirées d'une source administrative, la probabilité de nouer avec succès des liens entre les enregistrements des fichiers à grande diffusion et la base de données administratives est examinée, etc.

Plusieurs méthodes sont disponibles pour réduire les risques de divulgation. Les méthodes ont toutes un coût du point de vue de l'utilité analytique des données obtenues. Il faut appliquer les méthodes avec prudence pour maintenir le plus possible la valeur analytique des données (p. ex., corrélations et moyennes des variables).

1. Les méthodes de réduction des données comprennent les suivantes :

i. Suppression des identificateurs directs (il faudrait toujours le faire).

ii. Suppression des variables des identificateurs indirects qui accroissent le risque de divulgation (p. ex., pays d'origine, questions sur les troubles de la vue).

iii. Suppression d'enregistrements individuels (p. ex., pour une personnalité bien connue).

iv. Suppression de données individuelles d'un enregistrement en particulier (p. ex., une appartenance ethnique très rare dans une région en particulier).

v. Nouveau codage des données :

- réduction des détails géographiques, l'information géographique peut augmenter énormément le risque de divulgation et elle devrait être ajoutée seulement à des niveaux très agrégés,

- données tronquées par le haut et par le bas (p. ex., les revenus supérieurs à 100 000 $ pourraient être tronqués par le haut à 100 000 $),

- variables quantitatives réparties en catégorie, par exemple, l'âge ou le revenu,

- variables catégoriques agrégées davantage.

vi. Échantillonnage du fichier de microdonnées

On construit un sous-échantillon des données de l'enquête (et on ajuste conformément les pondérations d'échantillonnage). Cette mesure sert à susciter l'incertitude pour les unités de l'échantillon qui ont des caractéristiques uniques. Il s'agit d'une précaution nécessaire si le fichier original est un recensement.

vii. Enregistrements microagrégés

La microagrégation est le regroupement d'enregistrements, par exemple trois à la fois, et le remplacement des valeurs des variables quantitatives par les valeurs moyennes des groupes. Les variables catégoriques peuvent servir à définir les groupes d'unités semblables.

2. Les méthodes de perturbation des données comprennent les suivantes :

i. Arrondissement des microdonnées et ajouts d'interférences

On peut faire l'arrondissement déterministe ou aléatoire des données, par exemple, ou ajoute aux valeurs des données des interférences aléatoires normalement distribuées.

ii. Échange de données

Des enregistrements correspondants à un échantillon d'enregistrements de microdonnées sont identifiés selon un ensemble déterminé de variables et les valeurs d'autres variables sont échangées entre les enregistrements correspondants. L'échange de données peut servir à échanger des variables d'identificateurs indirects ou des variables à caractère éventuellement délicat. Il peut être possible de sélectionner un petit échantillon d'enregistrements de microdonnées, par exemple, et les valeurs de leurs revenus pourraient être échangées avec des enregistrements qui ont des valeurs similaires pour la géographie, l'âge et le sexe. Cette technique peut cependant avoir des répercussions sur l'analyse de la corrélation.

iii. Suppression de l'information et remplacement par des données imputées

On peut remplacer les valeurs déclarées par des valeurs moyennes, par exemple, pour des petites populations.

Si des bases de données externes, par exemple des fichiers de données administratives, et l'enquête ont des variables communes, les utilisateurs peuvent essayer de nouer des liens entre leurs données et le fichier de données à grande diffusion. Il faut accroître dans ces cas la portée de la perturbation.

12.5.3 Autres considérations sur la confidentialité

L'équilibre entre le besoin d'information pour utilisation publique et la nécessité de protéger les renseignements confidentiels des unités de l'échantillon est l'un des défis que doit relever un organisme statistique. Celui-ci fait appel à la bonne volonté des répondants, à leur générosité et à l'utilisation de leur temps non rémunéré, et un contrat implicite est donc convenu avec les répondants. L'organisme statistique doit considérer quatre éléments pendant la planification d'une enquête pour respecter ce contrat implicite :

i. L'intrusion dans la vie privée devrait être évitée. La société a-t-elle vraiment besoin de l'information? Est-il possible de trouver l'information sans la demander aux particuliers? Une taille d'échantillon inférieure convient-elle?

ii. Il ne doit y avoir aucun risque de préjudice indirect pour le répondant. Les particuliers qui répondent aux questions peuvent-ils être éventuellement en danger? Au cours d'une enquête sur la violence dans les ménages, par exemple, le persécuteur peut entendre le répondant pendant l'interview et le blesser après le départ de l'intervieweur.

iii. Il faudrait garantir la confidentialité aux répondants. Ils devraient être informés que toute l'information sera diffusée dans le grand public en une mise en forme qui empêchera la divulgation de l'information personnelle à leur sujet.

iv. Les répondants peuvent donc donner leur consentement informé (c.-à-d. qu'ils comprennent à quoi serviront les données et qu'ils sont d'accord). Il faut donc les informer des points suivants :
   - l'objectif de l'enquête (y compris les utilisations et les utilisateurs prévus des données de l'enquête),
   - le pouvoir (loi) qui autorise la collecte des données,
   - les détails sur l'enregistrement de la collecte (pour suivi),
   - le caractère obligatoire ou volontaire de l'enquête,
   - la protection de la confidentialité,
   - les plans de liaison des données avec d'autres fichiers,
   - l'identité des parties à toute entente d'échange de l'information.

Ces conditions sont essentielles pour obtenir de l'information fiable des répondants. Afin de garantir que les répondants donnent leur consentement informé, l'organisme statistique devrait appliquer une politique uniforme pour renseigner les répondants sur la nécessité de l'enquête et sur leurs droits et responsabilités. L'information demandée en vertu de cette politique doit être rédigée sur support papier pour toutes les enquêtes, et communiquée aux répondants au moment de la collecte ou avant. S'il s'agit d'une enquête téléphonique sans matériel de présentation, l'information doit être communiquée de vive voix et envoyée par écrit sur demande. (Voir le **Chapitre 5 - Conception du questionnaire** qui donne l'information à ajouter au questionnaire.)

Voici des considérations supplémentaires sur la confidentialité :

a. Il faut protéger les questionnaires pendant la collecte, leur acheminement, la sauvegarde et l'extraction des données.

b. Les intervieweurs ne devraient pas interviewer en public (parcs, restaurants, etc.) parce que d'autres pourraient entendre les réponses. Il faut éviter d'interviewer sur des sujets à caractère délicat lorsque d'autres peuvent entendre.

c. Des particuliers sont sélectionnés à partir d'une liste pour certaines enquêtes et une procédure de repérage est appliquée si la personne a déménagé. Les intervieweurs doivent être conscients du risque de divulgation durant le repérage (au cours d'une enquête de suivi qui cible les répondants atteints d'asthme, par exemple, seul le répondant doit être informé de la raison du suivi).

d. Les moyens de communication électronique, par exemple les téléphones sans fil, les téléphones cellulaires et l'Internet, utilisent une longueur d'onde publique et toute l'information communiquée par ces moyens est accessible à tous ceux qui s'en donnent la peine. Il faudrait donc éviter de transmettre l'information confidentielle par ces moyens, sauf après chiffrage sécuritaire. Il faudrait informer les répondants qui utilisent ces téléphones sans fil ou cellulaires pendant une interview que ce moyen pose un risque et leur demander d'utiliser un téléphone filaire si possible. Les réponses aux enquêtes faites sur Internet devraient être chiffrées. La majorité des programmes de courrier électronique et des navigateurs de la toile (Web) ont une capacité de chiffrage.

e. Après la compilation des données en mise en forme lisible à la machine et lorsque le traitement est achevé, le questionnaire devrait être détruit (déchiqueté, brûlé, etc.), si cette mesure est conforme à la politique de l'organisme statistique sur la confidentialité.

12.6 Sommaire

L'évaluation et la diffusion des données sont des étapes très importantes d'une enquête. L'objectif est de communiquer l'information aux utilisateurs pour qu'ils soient en mesure de comprendre les résultats de l'enquête et de prendre des décisions. Pendant l'évaluation des données, il faudrait évaluer les résultats définitifs, compte tenu des objectifs originaux de l'enquête. Ils devraient indiquer les points forts et les points faibles de l'enquête pour que les utilisateurs déterminent à quel point les erreurs dans les données en restreignent l'utilisation.

Les méthodes de diffusion comprennent les rapports sur support papier avec tableaux et graphiques, un fichier de microdonnées à grande diffusion, ou les deux. Avant la diffusion des résultats (ou données) de l'enquête cependant, il faut en faire une mise à l'essai approfondie pour vérifier le respect de la confidentialité des répondants.

Bibliographie

Ardilly, P. 1994. *Les Techniques de sondage*. Editions Technip, Paris.

Boudreau, J.R. 1996. Évaluation et réduction du risque de divulgation dans les fichiers de microdonnées à variables discrètes. *Symposium 95: Des données à l'information : méthodes et systèmes : recueil.*. Statistique Canada. 155-168.

Brogan, D.J. 1998. Pitfalls of Using Standard Statistical Software Packages for Sample Survey Data. *Encyclopedia of Biostatistics*. John Wiley and Sons, New York.

Brackstone, G. 1999. La gestion de la qualité des données dans un bureau de Statistique. *Techniques d'enquête,* 25(2): 159-172.

Carlson, B.L., A.E. Johnson, and S.B. Cohen. 1993. An Evaluation of the Use of Personal Computers for Variance Estimation with Complex Survey Data. *Journal of Official Statistics,* 9(4): 795-814.

Cox, B.G., D.A. Binder, B.N. Chinnappa, A. Christianson, M.J. Colledge et P.S. Kott, Éds. 1995. *Business Survey Methods.* John Wiley and Sons, New York.

Doyle, P., Lane, J.I., Theeuwes, J.M. et L.V. Zayatz, Éds. 2001. *Confidentiality, Disclosure, and Data Access – Theory and Practical Applications for Statistical Agencies.* North-Holland.

Dufour, J. 1996. *Labour Force Survey Data Quality*. Statistics Canada. HSMD-96-002E/F.

Ehrenberg, A.S.C. 1982. *A Primer in Data Reduction – An Introductory Statistics Textbook*. John Wiley and Sons, Great Britain.

Everitt, B.S. 1998. *The Cambridge Dictionary of Statistics*. Cambridge University Press. United Kingdom.

Fink, A. et J. Kosecoff. 1998. *How to Conduct Surveys: a Step-by-Step Guide*. Sage Publications, California.

Freund, J.E. et R.E. Walpole. 1987. *Mathematical Statistics*. Fourth edition. Prentice Hall, New Jersey.

Groves, R.M. 1989. *Survey Errors and Survey Costs*. John Wiley and Sons, New York.

Johnson, S., N.L. Kotz et C.B. Read. 1982. *Encyclopedia of Statistical Sciences.* John Wiley and Sons, Toronto.

Levy, P.S. et S. Lemeshow. 1999. *Sampling of Population: Methods and Applications*. Third edition. John Wiley and Sons, New York.

Lohr, S. 1999. *Sampling: Design and Analysis*. Duxbury Press, U.S.A.

MacNeill, I.B. et G.J. Humphrey, Éds. 1987. *Applied Probability, Statistics and Sampling Theory*. Reidel, Boston.

Mendenhall, W. 1991. *Introduction to Probability and Statistics*. Eighth edition. PWS-Kent Press, Boston.

Mood, A.M., F.A. Graybill et D.C. Boes. 1974. *Introduction to the Theory of Statistics*. Third edition, McGraw-Hill Series in Probability and Statistics, McGraw-Hill, U.S.A.

Travaux publics et services gouvernementaux Canada, Bureau de la traduction, 1996. *Le Guide du rédacteur*. Ottawa.

Schackis, D. 1993. *Manual for Disclosure Control*. Eurostat, Luxembourg.

Steel, R.G.D. et J.H. Torrie. 1980. *Principles and Procedures of Statistics – A Biometrical Approach*, Second edition. *McGraw-Hill, U.S.A.*

Statistique Canada. 1993. *Normes et lignes directrices pour la déclaration des taux de réponse.*

Statistique Canada. 2000. Politique visant à informer les utilisateurs sur la qualité et la méthodologie. *Manuel des politiques,* Politique 2.3

Willenborg, L. et T. de Wall. 1996. *Statistical Disclosure Control in Practice*. Lecture Notes in Statistics 111. Springer-Verlag, New York.

Willenborg, L. et T. de Wall. 2001. *Elements of Statistical Disclosure Control*. Lecture Notes in Statistics 155. Springer-Verlag, New York.

Wilson, J.R. et M. Reiser. 1993. Transforming Hypotheses for Test of Homogeneity in Survey Data. *Journal of Official Statistics,* 9(4): 815-824.

Chapitre 13 - Planification et gestion de l'enquête

13.0 Introduction

La planification et la gestion d'une enquête sont essentielles à son succès. Il est impossible de bien comprendre le but à atteindre et les moyens d'y parvenir sans structure de gestion claire et efficace. De nombreuses structures peuvent être appliquées à la gestion d'une enquête. Une structure souvent utilisée est l'approche par équipe de projet ou d'enquête. La planification, la conception, la mise en œuvre et l'évaluation d'une enquête et de ses résultats prévus sont confiés à une équipe interdisciplinaire. Celle-ci est formée de personnes exerçant les divers métiers nécessaires à la planification et à la mise en œuvre de l'enquête. Tous les membres de l'équipe de l'enquête se chargent de la planification, de la gestion et de la coordination des activités dans leur domaine d'expertise et de responsabilité, et ils coordonnent ces activités avec toutes les autres activités de l'enquête.

La planification et la gestion sont les activités clés qui permettent d'atteindre les objectifs de l'enquête. La planification détermine la stratégie que l'organisme statistique applique pour atteindre les objectifs de l'enquête. À l'étape de la planification d'une activité statistique éventuelle, les exigences du client, les moyens à consacrer pour répondre à ces exigences et la satisfaction recherchée font l'objet d'un examen (voir le **Chapitre 2 - Formulation de l'énoncé des objectifs**). Les besoins de financement et de ressources, ainsi que l'échéancier des activités, sont aussi déterminés à l'étape de la planification. Une étape de planification de la qualité est essentielle si l'on veut déterminer la qualité et le coût du projet dans l'ensemble. Une certaine planification continue pendant presque tout le cycle de l'enquête et elle prend fin seulement lorsque les données définitives demandées ont été livrées à la satisfaction de l'utilisateur.

Une bonne planification exige une bonne gestion et des intervenants informés et chevronnés. Peu importe la structure de gestion utilisée, il faudrait nommer un chargé d'enquête et lui confier le bon fonctionnement de tous les aspects de l'activité d'enquête. Ce chargé d'enquête devrait rendre compte à la direction, habituellement un Comité directeur qui donne orientation et conseils. Les principales fonctions de gestion comprennent l'organisation, l'orientation, la surveillance et le contrôle de l'enquête.

L'objectif de ce chapitre est de considérer comment planifier une enquête, l'accent étant mis sur l'approche de l'équipe de l'enquête. Une liste de vérification des activités ou méthodes qui devraient être considérées pendant la planification (voir la Liste de vérification de la planification) est ajoutée à la fin du chapitre.

13.1 Approches de la planification et de la gestion de l'enquête

Les questions élémentaires à considérer lors de la planification et de la gestion d'une enquête dans un organisme statistique ressemblent beaucoup aux questions qui se posent pour tout projet dans tout organisme, c.-à-d. comment identifier clairement les besoins, les communiquer efficacement et y répondre le plus rapidement possible, dans les limites du budget et en obtenant la meilleure qualité possible conformément aux besoins et à prix raisonnable? Les particularités sont très différentes, mais les éléments communs essentiels qui permettent d'atteindre les objectifs de tout projet sont les mêmes : communication, connaissances, aptitudes, engagement, efficience et efficacité.

Les ressources disponibles dans l'organisme statistique, son organisation fonctionnelle, la répartition des responsabilités dans l'organisme et dans le système de la statistique nationale déterminent en partie le style et le genre de planification et de gestion d'une enquête. Il faut prévoir l'organisation en tenant compte des responsabilités, de la culture et des pratiques opérationnelles de l'organisme statistique, des bureaux de la

statistique provinciale, de l'organisation statistique dans l'ensemble au pays à tous les échelons et des groupes professionnels pertinents.

Une équipe de projet est une structure efficace habituellement utilisée pour la planification et la gestion d'une enquête. L'engagement de la direction et des intervenants appropriés permet à l'équipe de l'enquête de mettre en commun des connaissances et des aptitudes, d'inciter à l'engagement, de simplifier et d'améliorer la communication, et de donner l'occasion d'accentuer l'efficience et l'efficacité. D'autres structures ou approches sont possibles et peuvent être nécessaires dans certaines situations. Si la structure de l'équipe du projet n'est pas explicitement appliquée cependant, il faut insister et compter davantage sur des spécifications précisément rédigées comme moyen de communication, mais elles laissent peu de place à la modification. Il faut aussi mettre davantage l'accent sur les aptitudes particulières des participants, mais il y a moins d'occasions de surveillance et de contrôle efficaces.

13.1.1 L'équipe de l'enquête

Une équipe d'enquête interdisciplinaire est souvent chargée de la planification, de la conception, de la mise en œuvre et de l'évaluation de l'enquête et de ses produits prévus. Elle est composée de membres ayant différentes aptitudes techniques nécessaires pour planifier l'enquête et la mettre en œuvre. Tous les membres de l'équipe de l'enquête se chargent de la planification, de la gestion et de la coordination d'activités dans leur domaine d'expertise et de responsabilité. Ils coordonnent aussi ces activités avec toutes les autres activités de l'enquête. Chaque membre de l'équipe a ses responsabilités particulières, mais tous sont chargés d'atteindre les objectifs de l'enquête. Chacun examine les propositions, plans, budgets, options, stratégies et principaux instruments ou spécifications qui font ensuite l'objet d'une discussion en équipe pour déterminer le meilleur moyen de procéder. Chaque membre d'équipe de la plupart des enquêtes obtient le soutien d'autres membres du personnel à qui sont confiées des activités à accomplir au nom de l'équipe. D'autre part, chaque unité organisationnelle engagée dans une enquête n'a pas besoin d'être directement représentée à l'équipe de l'enquête. Il n'est pas nécessaire de représenter à l'équipe du projet, par exemple, les services de logistique, d'imprimerie, de saisie des données ou d'administration.

L'équipe de l'enquête n'est pas un comité ou un ensemble de particuliers ayant chacun un objectif et un programme distinct. C'est un groupe de personnes qui travaillent ensemble et qui collaborent à un seul projet. Exception faite du chargé de projet, il ne devrait pas y avoir de hiérarchie dans l'équipe, seulement des interventions. La communication, la solution de problèmes et la réponse à des questions avec efficacité, ainsi que le soutien de l'innovation, de l'efficience, de l'engagement et de la qualité, sont les caractéristiques ciblées de l'approche en équipe. La discussion ouverte et constante, des commentaires constructifs, une rétroaction positive, la souplesse et une disposition à considérer toutes les exigences et les questions, et tous les points de vue, sont des éléments essentiels. L'équipe doit aussi être minutieuse, réfléchie, autonome et déterminée. Les questions, problèmes et différends doivent être résolus correctement, sans équivoque, au moment opportun, les propositions doivent être considérées de la même manière, et il faut bien comprendre les répercussions des décisions.

Les équipes d'enquête sont habituellement composées d'un chargé d'enquête, d'un expert de la matière à l'étude, d'un statisticien d'enquête, d'un analyste des systèmes informatiques et d'un spécialiste des opérations et de la collecte des données.

i. Chargé d'enquête

La gestion de l'enquête est confiée au chargé d'enquête. Il ou elle veille à ce que chacun se conforme aux objectifs, au budget et à l'échéancier. Le chargé d'enquête doit habituellement déterminer les ressources nécessaires à l'enquête, tracer un plan préliminaire, coordonner la préparation et la mise à jour du plan, préparer le budget et surveiller l'utilisation des ressources et l'état d'avancement. Il établit aussi la liaison avec la direction et le client, et leur fait rapport sur l'état d'avancement. Il veille au respect des engagements envers les dispositions réglementaires, législatives et juridiques du Bureau, et à l'application de ses politiques, normes, lignes directrices et règlements. Le chargé d'enquête préside les réunions de l'équipe et y représente le client, peu importe ses relations fonctionnelles ou organisationnelles avec ce client.

ii. Coordonnateur de la matière

Le coordonnateur de la matière est chargé du contenu de l'enquête. S'il y a plus d'un domaine à l'étude (p. ex., une enquête visant à étudier les données sur la démographie, la scolarité, la population active et le revenu), le coordonnateur se charge des activités et des entrées de ceux qui participent à la matière, mais qui ne sont pas membres de l'équipe de l'enquête. Il ou elle veille à ce que la matière soit clairement et correctement représentée, à ce que les utilisations des données prévues soient évidentes dans l'énoncé des objectifs, ainsi qu'à la conception et à la mise en œuvre, par l'intermédiaire de discussions avec le client et l'équipe. Le coordonnateur de la matière se charge en particulier de la préparation des définitions et des concepts, de la collecte et de l'analyse des données chronologiques pertinentes (pour la planification et l'élaboration), de l'élaboration de la matière et de la mise à l'essai du questionnaire, de la préparation de toute matière qui exige la collecte des données et des spécifications de traitement, de la conception des sorties statistiques, de l'élaboration et de l'application de l'analyse des données, et de la préparation du texte analytique, ou il coordonne ces activités. Il coordonne aussi la validation ou l'attestation des résultats de l'enquête et donne son expertise en la matière pour l'évaluation de la qualité des données et la préparation de la documentation pertinente à la matière.

iii. Coordonnateur de la méthodologie statistique

Un statisticien d'enquête chevronné (ou un spécialiste de la méthodologie) est chargé d'orienter et de coordonner la conception et l'élaboration de la méthodologie statistique qui sera appliquée à l'enquête. Il ou elle est chargé(e) en particulier du plan d'échantillonnage, de la pondération et de l'estimation, de la conception de méthodes de contrôle qualitatif, de la conception et de mesures de l'évaluation de la qualité des données, de la conception de mécanismes ou de stratégies d'imputation et de vérification, et des aspects statistiques de la diffusion et de l'analyse des données. Le spécialiste des techniques d'enquête intervient aussi à titre de consultant et de conseiller auprès de tous les autres membres de l'équipe sur les questions de méthodologie statistique et garantit l'application constante de méthodes statistiques efficientes et logiques.

iv. Coordonnateur des systèmes informatiques

L'analyste des systèmes est chargé de la conception et de la mise au point de programmes et de systèmes informatiques, ainsi que de l'identification, l'intégration et la modification des logiciels commerciaux ou de ceux déjà sur place qui sont nécessaires pour procéder à l'enquête. Il ou elle veille à ce que ces systèmes fonctionnent selon les spécifications d'autres participants et membres de l'équipe. L'analyste des systèmes intervient aussi à titre de consultant ou de conseiller auprès de tous les autres membres de l'équipe de l'enquête sur des questions d'informatique et veille à l'application constante de méthodes efficientes, vérifiables, compatibles et logiques pendant tout l'exercice. Il coordonne aussi certains aspects du traitement statistique (p. ex., vérification et imputation, pondération et estimation, tabulation), ou en garantit la mise en œuvre efficace, compte tenu de l'intervention du chargé des opérations.

v. Chargé des opérations et de la collecte des données

Le chargé des opérations et de la collecte des données veille à l'élaboration de la collecte des données, à la saisie et au codage des spécifications et procédures. Il est aussi chargé de la planification et de la coordination du recrutement, de la formation, de la surveillance et du contrôle du personnel de la collecte des données, notamment les intervieweurs et les surveillants, ainsi que du personnel de codage et de saisie des données, le cas échéant. Ses responsabilités comprennent l'élaboration, la mise en œuvre et la gestion des opérations de collecte et des programmes de relations publiques, la préparation du matériel (p. ex., les manuels et les cartes) et les besoins de soutien logistique. Il ou elle intervient à titre de consultant et de conseiller auprès de tous les autres membres de l'équipe de l'enquête sur des questions opérationnelles pertinentes à son expertise et il veille à ce que les spécifications et exigences que d'autres membres de l'équipe ont élaborées, p. ex., les spécifications sur la vérification et le suivi de l'intervieweur, soient correctement intégrées aux procédures. L'intervention du chargé des opérations peut englober la collecte sur place par l'intermédiaire des bureaux régionaux, des opérations distinctes de saisie et de codage des données, ainsi que le déroulement d'activités opérationnelles manuelles ou automatisées accomplies au Bureau central. Ces interventions peuvent être confiées à deux personnes pour les plus grands projets, un chargé de la collecte des données et un chargé du traitement.

vi. Autres membres

Un bon nombre de chargés de tâches ou de coordonnateurs peuvent être nécessaires pour les grands projets, par exemple, un recensement de la population. Dans ce genre de projets, des chargés de tâches distincts peuvent être affectés aux communications ou à la publicité, aux données d'entrée et de sortie géographiques (liées à la base de sondage, aux produits et variables géographiques) et aux résultats de l'enquête. Les grandes enquêtes ou les recensements peuvent aussi être subdivisés en tâches (p. ex., élaboration et conception du questionnaire, collecte des données, vérification et imputation) et en sous-tâches particulières (p. ex., contrôle qualitatif de la collecte, codage, etc.). Une équipe est affectée à chaque tâche et sous-tâche. Les caractéristiques des membres des équipes peuvent être différentes, mais chaque équipe a un chef d'équipe qui fait rapport à un chef à l'échelon supérieur de la structure de l'équipe de l'enquête.

La gestion a deux dimensions. L'équipe de l'enquête donne une dimension, la gestion des ressources. Elle est habituellement intitulée *gestion matricielle*. L'organisation fonctionnelle donne l'autre, à l'aide de la prestation de services à contrat ou de produits intermédiaires, de l'affectation de membres de l'équipe, de l'examen technique et de la supervision de leur contribution à l'enquête.

13.1.2 Comités supplémentaires de planification et de gestion de l'enquête

La structure de la gestion et de la planification exige plus qu'une équipe de projet ou d'enquête pour être efficace. Dans les grandes enquêtes ou les groupes d'enquêtes qui forment un programme statistique cohérent, un certain nombre de comités peuvent servir à orienter et à conseiller.

i. Un comité de gestion de l'organisme statistique (il peut avoir divers noms, p. ex., Comité stratégique) veille à ce que l'enquête soit pertinente dans le contexte global du programme statistique de l'organisme et à ce que la gestion en soit efficace. Ce comité (ou un chef fonctionnel d'un échelon équivalent) approuve en définitive le lancement de l'enquête, ainsi que la répartition du budget et des ressources de l'enquête.

ii. Un Comité directeur, composé d'intervenants chargés d'un groupe d'enquêtes semblables, affecte aux enquêtes des services ou des ressources techniques ou spécialisés. Le Comité directeur approuve l'énoncé des objectifs, l'échéancier, la répartition des ressources dans les limites du budget approuvé,

la matière détaillée, les sorties et la méthodologie. Il donne aussi l'orientation générale et des conseils à l'équipe de l'enquête, surveille l'état d'avancement et règle les problèmes hors de la capacité de solution de l'équipe. Peu importe le commanditaire ultime de l'enquête, le Comité directeur est souvent considéré comme le client pour l'équipe du projet. Le chargé d'enquête devrait être membre de ce comité ou lui faire rapport directement et assister aux réunions. Si un ministère ou un organisme externe finance l'enquête, il peut être utile qu'un représentant de cet organisme siège au comité.

iii. Un comité spécialisé donne des conseils et l'orientation sur la matière, les concepts, les définitions et les classifications. Ce comité coordonne les fonctions entre les programmes statistiques pour promouvoir l'uniformité et les normes de la matière. En l'absence d'un tel comité, le coordonnateur de la matière devrait nouer des liens avec ses collègues techniciens pour obtenir une orientation et un contexte.

iv. Dans certains cas, un comité consultatif ayant des membres régionaux, provinciaux ou de divers ministères peut donner des conseils sur des questions générales, priorités, matières, exigences des utilisateurs et sorties particulières. Ce comité reflète l'intérêt dévolu des membres envers l'enquête ou ses résultats. En l'absence d'un tel comité, le chargé de projet et le coordonnateur de la matière doivent garantir la liaison efficace avec les utilisateurs et les groupes qui peuvent avoir un intérêt dévolu envers l'enquête.

v. Un comité consultatif professionnel formé d'experts autonomes (à l'externe) peut finalement donner des conseils à caractère technique et faire un examen constant des méthodes statistiques et des sorties. En l'absence d'un tel comité, le spécialiste de la méthodologie d'enquête devrait demander des conseils, et vérifier s'il procède à une recherche et à des essais des méthodes appropriés.

Ces comités peuvent servir à l'étape du processus de planification et de gestion. Leur engagement reflète et concrétise le soutien du programme et détermine la priorité de l'enquête dans le programme statistique de l'organisme dans l'ensemble. Voilà qui garantit l'engagement à tous les paliers de la direction de l'enquête et l'accès aux ressources spécialisées nécessaires. L'importance de l'enquête, la structure du système statistique national et les pratiques de gestion de l'organisme statistique déterminent le genre de structure des comités.

L'organisation fonctionnelle et de l'infrastructure de l'organisme statistique donne à l'équipe un soutien très varié. L'organisation fonctionnelle devrait être chargée de la surveillance technique, de la conception et de l'élaboration, c.-à-d. que les surveillants des secteurs fonctionnels examinent les contributions des subalternes qui travaillent en équipe. Les membres de l'équipe, pour leur part, devraient demander une orientation ou des conseils techniques à leur surveillant fonctionnel et aux collègues dans leur infrastructure. Le chef de l'équipe devrait être en mesure de supposer que les entrées obtenues pour son projet ont l'approbation et le soutien du surveillant des membres (en supposant que l'organisation fonctionnelle prévoit la surveillance selon une expertise technique ou professionnelle) ou reflètent une certaine forme d'évaluation et de soutien de la part des pairs.

L'équipe devrait aussi demander de l'aide ou des services qui peuvent être disponibles dans l'organisme, p. ex., services et installations de collecte et de saisie des données, services informatiques, d'imprimerie, de communication publique ou avec les médias, services et installations de formation, des services de liaison interministérielle ou fédérale-provinciale, etc. Il serait inhabituel que l'équipe se charge de tous ces aspects d'une enquête. Le surveillant ou chef fonctionnel noue des liens avec le membre approprié de l'équipe de l'enquête et lui fait rapport sur des questions liées à l'enquête, mais la gestion de ces opérations se fait habituellement dans l'unité ou l'organisme fonctionnel.

13.1.3 Autres approches de la gestion

L'approche de l'équipe de l'enquête est extrêmement efficace pour l'élaboration de nouvelles enquêtes et les grandes modifications des plans d'enquête. Tous les organismes ne sont cependant pas en mesure de fonctionner ainsi, en particulier pour les très grands projets statistiques. Plusieurs variantes sont décrites ci-dessous :

i. Grande enquête ou recensement

Si le projet est un recensement de la population ou si l'organisme statistique a peu de sources centralisées ou de centres d'expertise technique, la structure de l'équipe de l'enquête peut permettre seulement la gestion, et non la gestion, la conception et la mise en œuvre. Dans ces situations, l'équipe de l'enquête doit compter sur plusieurs équipes de projets de composantes importantes, des équipes chargées de tâches et des équipes chargées de sous-tâches pour procéder à la planification, à la conception et à la mise en œuvre détaillées. Une conséquence probable de cette structure est que l'équipe serait composée de chargés de projets de composantes qui veilleraient en majeure partie à la gestion d'une série d'opérations ou de modules de l'enquête globale (p. ex., la collecte des données). Les interventions des coordonnateurs des systèmes informatiques et spécialisés, et de la méthodologie statistique, pourraient être accomplies seulement à l'échelon du projet de composante ou de l'équipe chargée d'une tâche. Une perte de communication, d'efficience et de qualité en serait le résultat, mais cette perte doit être évaluée du point de vue des exigences de contrôle et de gestion efficaces. La perte éventuelle est beaucoup plus importante dans le cas d'une nouvelle enquête. Cette perte devrait être moindre pour une enquête en cours, par exemple un recensement de la population, s'il y a une évaluation suffisante et une longue période d'élaboration intercensitaire.

ii. Fournisseur de compétences et de services à l'externe

Si l'organisme n'a pas l'infrastructure ou les installations techniques nécessaires pour tous les aspects de l'enquête, il devra obtenir ces compétences de sources externes, par exemple, à l'aide d'un contrat à court terme ou en impartissant les fonctions à un organisme à l'externe. Si l'organisme n'a pas le personnel ou les installations appropriées pour l'impression des questionnaires et des manuels de l'intervieweur, ou pour la saisie des données, par exemple, il devra chercher des services à l'externe pour accomplir ces activités. Les intervenants du projet devront préparer les spécifications des fonctions ou services à obtenir, et préciser les conditions et attentes, et l'entrepreneur devra les accepter. L'organisme externe est ensuite chargé de l'application de ces spécifications conformément aux modalités du contrat. Dans la majorité des enquêtes, certains volets sont impartis à l'interne (hors de l'équipe du projet) ou à des fournisseurs de services à l'externe.

Certaines enquêtes nouvelles sont suffisamment simples du point de vue opérationnel pour que ses intervenants appliquent des méthodes habituelles ou des établissements commerciaux peuvent faire le travail sans avoir vraiment besoin d'un engagement direct avec l'équipe de l'enquête (par exemple, pour l'impression). S'il s'agit de fonctions complexes, d'enquêtes plus vastes et de recensements, l'impartition ajoute un risque et exige des contrôles particuliers. Dans le cas d'un organisme statistique national, l'entrepreneur à l'externe peut être un ministère, une institution, un organisme du secteur privé ou un particulier. Il faut appliquer les règles et règlements de l'organisme statistique, ainsi que ceux du gouvernement, y compris les règles et règlements des marchés publics et de l'impartition. L'équipe du projet doit vérifier attentivement si elle a accès à la gestion interne, à l'expérience et aux connaissances administratives et juridiques qui lui seront nécessaires. Il faut faire preuve d'un empressement proportionnellement approprié à l'importance et au risque lors de la sélection de l'entrepreneur, de la préparation et de l'approbation des spécifications et exigences, de la surveillance et de la gestion de la mise en œuvre et de l'accomplissement de ces activités.

iii. Prestation de services particuliers au comité du projet

La planification et la gestion de certaines enquêtes pourraient être faites par l'intermédiaire d'un comité directeur ou de projet et les participants à l'élaboration, au plan d'enquête et à sa mise en œuvre pourraient travailler distinctement par l'intermédiaire des membres désignés du comité qui n'ont pas d'intervention en équipe. L'enquête est habituellement un volet seulement du mandat du comité et ne fait pas directement partie de son objectif. Les enquêtes faites par un ministère qui n'est pas un organisme statistique (mais l'organisme peut apporter un certain soutien au plan d'enquête) sont souvent organisées ainsi. Dans ces cas, le comité est à l'intérieur du ministère d'accueil. Les enquêtes qui exigent des aptitudes spécialisées à la mesure et à l'observation directe (p. ex., les analyses du sang ou d'autres mesures médicales ou dentaires) peuvent aussi fonctionner de cette façon.

iv. Modification limitée du plan d'enquête

Dans de nombreuses situations, le travail de conception englobe seulement une composante d'une enquête en cours, par exemple, la modification de la conception du questionnaire ou du plan d'échantillonnage. Ces cas sont souvent réglés sans avoir recours à une équipe d'enquête, même si de nombreux aspects de l'enquête peuvent exiger une mise à jour ou une amélioration. Le temps et le coût expliquent habituellement pourquoi l'activité est accomplie sans équipe d'enquête.

Dans tous ces cas, les diverses aptitudes et connaissances nécessaires pour planifier et faire l'enquête ou planifier et appliquer une composante en particulier sont obtenues à contrat. La majorité des organismes statistiques fonctionnent ainsi pour divers projets, par exemple, pour donner des services consultatifs à des ministères, afin qu'ils procèdent à leurs propres enquêtes uniques (par exemple, sur la satisfaction des clients – utilisateurs – employés ou sur l'opinion publique) ou à des projets de modification partielle de la conception ou du plan d'enquête.

L'impartition peut être l'approche la plus rentable pour les composantes qui peuvent être précisées exactement si une unité organisationnelle ou un entrepreneur à l'externe a les connaissances et les ressources techniques nécessaires pour concevoir, élaborer ou produire à temps la composante demandée. La communication doit cependant être efficace, et la capacité, les aptitudes, l'état d'avancement et le respect des spécifications devront être évidents.

v. Le chef fonctionnel est le chargé d'enquête

Dans certains cas, la majorité des compétences nécessaires sont disponibles dans l'unité fonctionnelle qui a lancé l'enquête. Même s'il y a une distinction significative, le chargé d'enquête est aussi le chef fonctionnel de la majorité des participants, sinon tous. L'équipe des participants devrait néanmoins fonctionner comme une équipe d'enquête décrite ci-dessus. Nombre des attributs positifs de l'équipe d'enquête, par exemple la volonté de collaboration, l'ouverture d'esprit, la responsabilité partagée, l'autonomie et l'objectivité, pourraient cependant se révéler peu réalistes. Les divers genres d'expériences et de connaissances peuvent être différents et les différences aux niveaux fonctionnels peuvent se traduire plus souvent en conflits. Les participants hors du secteur fonctionnel peuvent avoir moins d'influence, une cible de responsabilité plus étroite peut leur être confiée et ils peuvent avoir moins d'interventions en équipe.

13.2 Planification de l'enquête

Il est évident, espérons-le, que diverses décisions doivent être prises pendant la préparation et la mise en œuvre du plan d'enquête pour garantir qu'elle atteint ses objectifs. L'enquête fait aussi partie d'un système statistique plus large. Elle doit donc atteindre aussi des objectifs plus larges et respecter les exigences plus

grandes de l'organisme statistique, compte tenu notamment des mérites des objectifs de l'enquête, du programme et du mandat de l'organisme dans l'ensemble et du coût de l'option de la production des données ou des renseignements. Ces objectifs et exigences forment cinq ensembles élémentaires de critères qu'il faut respecter lors de la planification, de la conception et de la mise en œuvre de toute enquête ou projet statistique.

i. Les données de l'enquête doivent être « aptes à l'utilisation ».

Il n'y a pas de définition communément acceptée entre les organismes statistiques de ce qui constitue « l'aptitude à l'utilisation ». On peut cependant évaluer si les données de l'enquête et l'information statistique sont aptes à l'utilisation, selon les six caractéristiques suivantes : pertinence, exactitude, actualité, accessibilité, intelligibilité et cohérence (pour une définition de ces termes, voir l'**Annexe B - Contrôle qualitatif et assurance de la qualité**).

ii. Il faut justifier le fardeau de réponse imposé ou la réaction probable du public.

Combien de temps faut-il pour remplir le questionnaire? Combien de temps faut-il au répondant pour vérifier ses dossiers et obtenir l'information d'autres membres du ménage ou de l'entreprise? À quel point les questions sont-elles indiscrètes? La période de collecte de l'enquête empiétera-t-elle sur le travail du répondant (p. ex., faire une enquête en milieu rural pendant les semailles ou la récolte)? Les définitions élémentaires sont-elles différentes de celles d'autres enquêtes qui appliquent des concepts semblables? La population a-t-elle trop souvent fait l'objet d'une enquête auparavant? L'enquête nuira-t-elle à la réputation de l'organisme ou aura-t-elle des répercussions négatives sur d'autres enquêtes (p. ex., taux de réponse réduit à cause de la controverse ou parce que l'enquête se déroule simultanément à une autre)? La valeur sociale inhérente des données qui seront tirées de l'enquête justifiera-t-elle le fardeau de réponse et le coût de l'enquête, et sera-t-il possible de le démontrer aux répondants et au public?

iii. Les résultats diffusés de l'enquête devraient refléter tous les résultats valides.

L'analyse de l'organisme statistique et la diffusion ne doivent pas être limitées au point de refléter, supposer ou soutenir indûment en fait une perspective, une intention, une conclusion ou un point de vue en particulier. (Voilà des répercussions de la diffusion incomplète ou des résultats analytiques limités qui ne sont pas inhabituels et involontaires.) Les résultats de l'enquête ne devraient pas servir à soutenir une perspective, un point de vue ou une conclusion en particulier, sauf si d'autres résultats plausibles ou contraires, ou si d'autres conclusions ont été mis à l'essai ou rejetés à l'aide de vérifications et de preuves statistiques évidentes. La vraisemblance ou la fiabilité statistique de ces essais, ainsi que les résultats ou les conclusions, doivent aussi être présentés clairement avec les résultats.

iv. Il faut respecter les exigences des politiques, règlements, procédures administratives, normes et lignes directrices de l'organisme et du gouvernement, et appliquer des méthodes et pratiques logiques.

Diverses « règles » s'appliquent au déroulement d'une enquête, à partir de l'autorisation officielle de la collecte des données en particulier, jusqu'à l'application de méthodes valides et efficientes du point de vue statistique. Des méthodes et pratiques logiques sont nécessaires pour éviter de gaspiller les ressources, pour répondre plus efficacement et de toute évidence aux critères d'« aptitude à l'utilisation » et pour soutenir en fait la réputation professionnelle de l'organisme.

v. Une enquête qui répond à tous ces critères doit être réalisable dans les limites du budget convenu et approuvé, à l'aide des moyens et ressources disponibles.

13.2.1 Étapes du plan d'enquête

La planification de l'enquête devrait se dérouler par phases d'exactitude et de détails croissants. À l'étape préliminaire, ou étape de proposition de l'enquête, seule la notion de faisabilité et les besoins de données de l'enquête les plus généraux peuvent être connus. En consultation avec les utilisateurs et le client, l'équipe précise davantage les concepts analytiques et les besoins de données, et elle commence à considérer le choix de la base de sondage, la taille générale de l'échantillon et la précision nécessaires, les options de collecte des données, l'échéancier et le coût. Elle se fait simultanément une idée des limites que le client imposera au coût et à l'échéancier, et elle en apprend davantage sur les ressources qui peuvent être disponibles pour l'enquête. Les plans sont révisés, élaborés et peaufinés, et des aspects plus détaillés sont examinés pendant les étapes ultérieures. Un certain genre de plan pour la conception, l'élaboration et la mise en œuvre est nécessaire pour chaque activité et opération. La planification continue quand même. Il faut faire des ajustements et apporter des modifications, et il peut être nécessaire d'établir des mesures correctives et des plans d'urgence.

Une enquête ou un projet statistique passe par les cinq étapes de planification suivantes :

i. formulation de la proposition d'enquête,

ii. établissement du plan d'enquête et détermination de la faisabilité,

iii. préparation de plans pour les composantes de l'enquête,

iv. touche finale apportée aux plans pendant la conception et l'élaboration,

v. ajustement des plans et ajout pendant la mise en œuvre et l'évaluation.

13.2.1.1 Première étape : proposition d'enquête

La première étape de la planification d'une enquête est le repérage d'une lacune d'information et la préparation d'une proposition d'enquête. Le passage d'un besoin identifié à une enquête exige d'abord l'approbation ou l'accord pour procéder à la planification préliminaire. Les membres du Comité de direction de l'organisme ou certains membres du Comité directeur affecteront probablement un chargé d'enquête par intérim qui donnera l'information initiale sur les solutions de rechange à l'enquête (p. ex., des données d'une autre enquête ou d'une source administrative sont disponibles), le coût et la faisabilité de l'enquête. L'étape préliminaire de la modification d'un plan d'enquête ou d'une demande d'enquête d'un organisme à l'externe serait semblable.

Il faudra peut-être faire rapidement ces études initiales et le résultat pourrait être un peu superficiel. L'examen préliminaire devrait néanmoins être axé sur la consultation avec des experts de diverses disciplines qui participeraient probablement à l'équipe de l'enquête éventuelle (même si ces intervenants pourraient, en bout de ligne, ne pas être membres de l'équipe réelle de l'enquête). Il faudrait entreprendre un processus d'entente et de soutien de ces experts avant le processus de prise de décisions, mais qui en ferait partie, pour déterminer s'il faut procéder à une élaboration plus substantielle aux fins de l'énoncé des objectifs et, si oui, préciser comment. Il faut aussi faire une étude complète de faisabilité et un plan d'enquête.

Lorsqu'une proposition d'enquête a été préparée, examinée et fait l'objet d'une discussion, la direction est en position de décider si elle procède à la planification et à l'élaboration ultérieures. Si oui, l'équipe de l'enquête doit élaborer l'énoncé des objectifs et confirmer rapidement la faisabilité, ainsi que certaines grandes options ou solutions de rechange.

13.2.1.2 Deuxième étape : détermination de la faisabilité et établissement du plan d'enquête

Voilà une étape critique de la planification parce qu'il faut déterminer le coût de l'enquête (de très près). C'est particulièrement important si le coût estimé approche ou dépasse le coût maximal que l'organisme de financement a établi. Voici les principaux buts de cette étape de la planification :

i. formuler (la version préliminaire de) l'énoncé des objectifs, déterminer les plafonds des coûts et les cibles de qualité, et donner un aperçu de l'échéancier,

ii. déterminer et évaluer la pertinence et l'accessibilité des sources concrètes de données et repérer les lacunes d'information (données administratives et enquêtes déjà réalisées),

iii. choisir la base de sondage, les unités statistiques éventuelles pour l'échantillonnage (le cas échéant) et la méthode de collecte des données,

iv. préciser les approches méthodologiques appliquées à d'autres enquêtes sur la même population cible et aux enquêtes d'autres organismes statistiques sur le même sujet général,

v. donner une évaluation préliminaire du coût, de l'échéancier, de la faisabilité et du fardeau de réponse, ainsi qu'une évaluation de la pertinence et des répercussions de l'échantillonnage du point de vue du coût et des exigences de qualité,

vi. rédiger un rapport de faisabilité et de planification, y compris les options, ainsi que les questions, besoins et limites en particulier (p. ex., du point de vue des politiques et règlements, ainsi que des pratiques, limites et exigences juridiques), qui couvrira chaque étape du déroulement de l'enquête, y compris l'énoncé des objectifs, ainsi que les options pour la base de sondage, le plan d'échantillonnage, la collecte des données, le traitement, le contrôle de la divulgation, les mises à l'essai, la diffusion, le budget, etc.

Il est possible, à chaque volet de cette étape, de faire rapport au Comité directeur et de mettre fin au processus si l'équipe détermine, selon une indication suffisante, qu'une enquête ne serait pas réaliste, compte tenu des modalités de la version préliminaire de l'énoncé des objectifs. Si la planification continue jusqu'à la formulation d'un rapport de faisabilité et de planification, le Comité directeur devrait examiner et évaluer la proposition d'enquête. Une décision sur le déroulement de l'enquête ou non, ainsi que sur ses paramètres et le plan général, est prise en tenant compte de la proposition d'enquête. Toute décision prise pour entreprendre d'autres activités de planification ou de conception aboutit habituellement à la troisième étape.

13.2.1.3 Troisième étape : préparation des plans des composantes

Chaque membre de l'équipe prépare les composantes du plan lié à sa responsabilité dans l'équipe. Chacun donne aussi une rétroaction sur les plans des autres et y exerce son expertise. Les équipes de tâche et de sous-tâche préparent aussi des plans et les coordonnent avec le membre responsable de l'équipe de l'enquête. Celle-ci examine et approfondit tous les plans. Voici les étapes de la préparation de ces plans :

i. voir à l'élaboration, la conception, la mise en œuvre et l'évaluation des plans d'activités, des échéanciers, des estimations des besoins de ressources et des estimations détaillées des coûts aux fins de la mise en œuvre pour chaque composante et étape de l'enquête ou du projet statistique,

ii. examiner tous les plans des composantes, identifier les entrées et les sorties pour chaque composante et les dépendances,

iii. procéder à l'élaboration nécessaire comme principale entrée aux plans des autres composantes,

iv. nouer les liens et établir l'uniformité à l'intérieur des composantes et entre celles-ci,

v. modifier les échéanciers au besoin,

vi. préparer les principales étapes et l'échéancier général,

vii. formuler la version définitive de la proposition et du plan pour la mise à l'essai,

viii. réviser les budgets et ajuster les plans au besoin.

La planification devient plus complexe au cours de cette étape. Afin de planifier les composantes opérationnelles (collecte, saisie et traitement des données), il faut accomplir un travail significatif intégré à la planification pour le plan d'échantillonnage (il faudrait déterminer la taille et la répartition), la matière du questionnaire, la méthodologie détaillée de la collecte et les exigences de vérification et d'assurance de la qualité. Les plans de collecte n'ont aucun sens, par exemple, sans une estimation précise de la longueur de l'interview.

À la fin de cette étape, et en supposant que le Comité directeur ait donné son approbation, l'équipe prend des dispositions, ou apporte la touche finale aux dispositions prises, pour obtenir les ressources nécessaires. Il faudrait maintenant déterminer la date de référence, la date de collecte des données, le budget et les besoins de ressources.

13.2.1.4 Quatrième étape : achèvement des plans pour la conception, l'élaboration et la mise en œuvre

À cette étape, il ne s'agit plus de décider que faire, mais plutôt de passer à l'action. Les questions de planification en instance devraient donc être de menus détails seulement et bien se situer dans les limites des plans concrets (pour le coût, le temps et les ressources). La touche finale peut être apportée aux plans de mise à l'essai et de mise en œuvre à cette étape seulement. Nous avons mentionné auparavant que divers aspects de la conception et de l'élaboration commencent à des moments différents, et les méthodes, procédures et systèmes qui seront utilisés sont déterminés à des degrés distincts de certitude. Dans certains cas, ceux qui obtiennent les spécifications et qui doivent les appliquer peuvent avoir une compréhension générale seulement de ce qu'ils doivent transformer en spécifications, procédures ou systèmes informatiques plus détaillés. Quelques modifications de dernière minute apportées aux spécifications pendant l'élaboration ou à la suite de la mise à l'essai sont toujours possibles. Il faut faire des compromis pour s'en tenir au coût convenu, compte tenu des contraintes de temps et de ressources.

13.2.1.5 Cinquième étape : ajustements et plans supplémentaires

Au cours de la conception, de la mise en œuvre et de l'évaluation de la qualité, il est possible de découvrir que tous les aspects de l'enquête ne se déroulent pas comme prévu. Les taux de réponse peuvent être supérieurs ou

inférieurs. Le pistage peut coûter plus cher. Une proportion plus élevée du travail des intervieweurs peut être rejetée pendant le contrôle qualitatif, ce qui cause des retards. Le taux de rejet à la vérification d'une variable en particulier peut être excessivement élevé. À l'étape de l'attestation de la qualité des données, il est possible de découvrir que de nombreux répondants ont mal interprété une question, etc. L'équipe de l'enquête devrait examiner ces situations et préparer rapidement des plans.

Si le coût augmente, si des ressources supplémentaires sont nécessaires, si un retard ou des répercussions sur les objectifs de l'enquête ou les exigences de qualité sont prévus, le plan supplémentaire devrait comprendre des options et des conséquences. Il faut aussi obtenir l'approbation du Comité directeur.

Même sans ces problèmes graves, il peut être nécessaire d'apporter des ajustements quotidiens aux plans. À mesure que l'écart se referme entre la date de référence et la date d'achèvement de l'enquête, les petits problèmes deviennent rapidement énormes.

13.2.2 Estimation du temps, des coûts et des ressources nécessaires

Pendant la planification de l'enquête, l'estimation des coûts (budget) et des besoins de ressources et de temps (échéancier) est faite par étapes de plus en plus détaillées et précises. Les estimations sont faites au départ selon des hypothèses générales sur la méthodologie qui sera appliquée, le nombre et le genre de membres du personnel et d'autres ressources nécessaires pour planifier, concevoir, mettre en œuvre et évaluer l'enquête, ainsi que les besoins logistiques, de matériel, d'articles, de transport, etc. Ces estimations doivent être plus exactes et détaillées à chaque étape de la planification.

Les experts de secteurs fonctionnels particuliers qui fournissent des ressources ou des services doivent préparer ou examiner et soutenir les estimations. Celles-ci devraient cibler l'information chronologique sur l'utilisation des ressources, la durée, le coût (d'enquêtes précédentes ou courantes, l'utilisation jusqu'à maintenant dans l'enquête en élaboration) et l'information administrative sur les coûts actuels à l'unité. Cette information doit ensuite servir à l'application particulière de l'enquête en élaboration.

Les activités ou méthodologies pertinentes parmi celles énumérées dans la Liste de vérification de la planification (voir à la fin de ce chapitre) représentent au moins une liste partielle des entrées qui sont relatives aux coûts et qui ont besoin de ressources et de temps pour préparation, achèvement ou prestation. Aux volets ressources, temps et estimation des coûts, cependant, il faudrait considérer les points suivants :

- les principales utilisations des données et les exigences sur la qualité,
- les caractéristiques de la population cible et la matière de l'enquête,
- la longueur et la complexité du questionnaire et de l'entrevue (le cas échéant),
- la complexité du plan d'échantillonnage et le genre de base de sondage (p. ex., base aréolaire, liste, composition aléatoire ou listes téléphoniques),
- la taille et la répartition de l'échantillon,
- la méthode de collecte des données (interview sur place, interview téléphonique, questionnaire envoyé par la poste, etc.),
- les procédures sur place (interview avec – sans substitut, stratégie et exigences de pistage et de suivi),
- le nombre et la complexité des vérifications intégrées,
- le taux de réponse prévu,
- le personnel avec – sans expérience, les besoins de recrutement et de formation,
- le matériel informatique et le logiciel, ainsi que les frais d'informatique,
- les spécifications, procédures et systèmes qu'il faut concevoir, élaborer et mettre à l'essai, lesquels peuvent être réutilisés ou modifiés,
- les besoins de ressources (matériel et personnel, coût par type et niveau de personnel),
- les exigences de rapport administratif et de gestion,

- la fréquence et la durée des réunions (équipe et Comité directeur, etc.).

Les besoins aux volets coûts, échéanciers et ressources doivent intégrer les activités du chargé d'enquête, des membres de l'équipe et de tous les autres participants. Les estimations devraient comprendre toutes les activités, depuis le début jusqu'à la prestation du dernier produit et rapport.

13.3 Gestion de l'enquête

Peu importe la structure de gestion appliquée, la gestion d'une enquête, après la planification, comprend l'organisation, l'orientation, la surveillance et le contrôle de l'enquête.

i. Organisation

L'organisation est la fonction de gestion qui permet de réunir les intervenants, les fonctions et les éléments physiques pour atteindre les objectifs de l'organisme. Le chargé d'enquête est responsable de la gestion de l'enquête et il a l'obligation de rendre compte, mais il doit aussi faire appel à l'équipe de l'enquête (et aux chefs des groupes fonctionnels de prestation des services) pour partager cette responsabilité. Les membres de l'équipe participent à cette fin à l'affectation des responsabilités et ils en conviennent. Les responsabilités devraient être affectées selon l'expertise, l'expérience et les ensembles particuliers d'activités ou de composantes compatibles de l'enquête. Il faut couvrir toutes les activités en collaboration avec les intervenants respectifs chargés des entrées et des sorties de chaque activité de l'enquête. L'une des interventions du chargé d'enquête est de veiller à ce qu'il n'y ait ni lacunes ni conflits.

ii. Orientation

L'orientation de l'enquête, ou plus particulièrement du travail des participants, comprend la prise de décisions, la prestation de conseils et l'acquisition ou la prestation d'aide au besoin. Les chefs doivent faire preuve de leadership, offrir des occasions de formation et de perfectionnement, susciter et maintenir de bonnes communications. Il faut résoudre les conflits clairement et rapidement. Si le chef commence avec un bon plan, la confiance, une bonne compréhension évidente des objectifs, le personnel motivé et bien formé garantira qu'il atteindra les objectifs de l'enquête.

iii. Surveillance et contrôle

La surveillance et le contrôle sont une fonction de la gestion qui demande d'être constamment bien informé et de réagir à tous les problèmes pour maintenir l'état d'avancement de l'enquête selon le plan. L'équipe de l'enquête doit vérifier si les ressources affectées à l'enquête sont disponibles et si elles sont utilisées avec efficience et efficacité. Elle doit vérifier si les plans d'enquête sont appliqués correctement et apporter les corrections et les ajustements nécessaires. Le chargé d'enquête doit vérifier si les plans, politiques et procédures sont appliqués à la lettre, et si les participants ciblent toujours les objectifs. Il formule et communique les instructions et en vérifient l'application, détermine les normes de rendement et le suivi pour accomplir les tâches, et vérifie le respect des échéanciers.

La surveillance est faite par l'intermédiaire de réunions régulières de l'équipe, de discussions, de communications quotidiennes avec les participants, et à l'aide de plans et de divers rapports d'information de gestion. Le rapport de planification, le budget et l'échéancier sont les principales références. Il faut repérer l'utilisation des ressources, les dépenses et l'état d'avancement, et faire rapport. Chaque membre de l'équipe devrait régulièrement présenter un rapport (de vive voix ou par écrit, selon la situation) sur l'état d'avancement, l'utilisation et les dépenses. Il faudrait faire rapport sur les données opérationnelles, par exemple les taux de réponse, les taux d'achèvement de l'intervieweur, les taux de suivi, l'information des

rapports de production et les opérations de contrôle qualitatif et d'assurance de la qualité, et examiner toutes ces données. La fréquence des réunions et des rapports devrait être déterminée selon l'urgence éventuelle de l'intervention en cas de problème.

Les prévisions aux volets de l'échéancier et des coûts pour les activités critiques d'un échéancier strict, par exemples les interviews, devraient être réparties jusqu'au niveau le plus bas des étapes de la composante. Ces étapes devraient être inscrites à un calendrier quotidien et surveillées si possible et si cette mesure est logique. Il serait autrement difficile de déterminer combien de temps il faudra pour réaliser les activités, si l'état d'avancement correspond à l'échéancier et si des mesures correctives, ajustements ou modifications des plans sont nécessaires.

iv. Communication, coordination et examen

L'équipe de l'enquête, et en particulier le chargé d'enquête, est responsable de la coordination et de la communication. L'équipe de l'enquête et les chefs fonctionnels sont chargés des examens et des communications dans leur secteur d'activité et de la prestation d'une rétroaction à l'équipe et au chargé d'enquête. Les plans, budgets, échéanciers, énoncés de responsabilité et mandats sont les principaux outils de communication et de coordination. Une trousse complète devrait être à la disposition de tous les participants. Ceux-ci doivent être informés de leurs interventions et des objectifs de celles-ci, et ils doivent connaître leur position exacte dans l'enquête en général. L'équipe de l'enquête doit aussi intervenir pour recevoir les communications appropriées aux fins de la surveillance et de l'obligation de rendre compte.

L'horaire de l'équipe devrait comprendre les principales étapes de l'enquête pour faciliter la communication et la coordination. Il devrait y avoir un dossier des décisions (une liste des décisions pertinentes prises en équipe). Il faudrait prévoir une procédure d'avis de problème pour les opérations et systèmes les plus importants (un bref rapport sur les erreurs, les incohérences et les solutions qui exigent des modifications à apporter aux procédures, opérations ou systèmes, ou d'autres changements qui ne sont pas reflétés dans les plans ou spécifications approuvés). Il faut distribuer largement le dossier des décisions et les avis de problème. Il faudrait préparer et remettre aux chefs et aux membres d'équipe des graphiques de cheminement ou des articles semblables qui affichent l'échéancier et les liens entre les activités.

Nous avons mentionné auparavant que le surveillant ou le chef fonctionnel devrait examiner les entrées de son personnel dans l'enquête. L'équipe de l'enquête a aussi sa responsabilité. Elle doit vérifier si les entrées (les spécifications, procédures et manuels, le questionnaire, etc.) correspondent aux besoins de l'enquête, afin de garantir que toutes les composantes atteignent leurs objectifs particuliers, sont harmonieuses et conformes aux objectifs de l'enquête et aux plans.

13.3.1 Fonctions du chargé d'enquête

Le chargé d'enquête a des responsabilités hiérarchiques et de coordination qu'il ne partage pas avec l'équipe de l'enquête dans l'ensemble. Outre les tâches énumérées à la section précédente, quelqu'un doit être en charge et il faut lui confier l'obligation de rendre compte et la responsabilité générale pour qu'il prenne au moment opportun des décisions conformes au mandat déterminé au Comité directeur. Quelqu'un doit avoir une interaction directe avec le Comité directeur. Il doit représenter l'équipe de l'enquête auprès des utilisateurs en général et des utilisateurs qui versent les fonds en particulier. Ces fonctions sont des volets du rôle de gestion du chargé d'enquête.

Le chargé d'enquête et d'autres membres de l'équipe collaborent avec le client et les utilisateurs pour déterminer leurs besoins analytiques et de données. Le chargé d'enquête doit cependant veiller à ce que les décisions prises pendant le déroulement de l'enquête ne compromettent pas l'aptitude à l'utilisation

fondamentale des données définitives. Il répond aux questions du client sur l'état d'avancement, justifie les décisions, communique les préférences du client à l'équipe de l'enquête et vérifie si l'argent du client est réparti correctement et dans les limites du budget. Il est aussi un intermédiaire entre le client et l'équipe de l'enquête. L'équipe du projet peut donc faire son travail sans interruption ou interférence. Le chargé d'enquête a la même intervention auprès du Comité directeur. Si le client n'est pas membre du Comité directeur, le chargé d'enquête doit garantir qu'il y a communication trilatérale efficace des exigences, décisions et résultats.

Il est essentiel que le chargé d'enquête soit informé personnellement et directement en tout temps que le client et les principaux utilisateurs savent ce qu'ils veulent, comprennent ce qu'ils obtiennent, connaissent les limites et déterminent comment les données répondront ou non à leurs besoins. Les coûts, conséquences et solutions de rechange appropriées doivent aussi être évidents. Une condition semblable s'applique au Comité directeur.

Le chargé d'enquête coordonne les activités de l'équipe de l'enquête et vérifie si les plans, spécifications, décisions, etc., sont correctement communiqués aux membres de l'équipe. Il réagit à tout problème imprévu et veille à ce que les intervenants appropriés soient informés pour prendre les mesures nécessaires. Le chargé d'enquête doit garantir qu'il est possible de surveiller l'état d'avancement et la qualité, et de repérer les nouveaux problèmes. Il ou elle doit avoir suffisamment d'information en tout temps pour pouvoir soutenir personnellement la crédibilité de l'enquête et de ses résultats, et en comprendre les limites.

Le chargé d'enquête doit veiller à ce que les activités de l'équipe de l'enquête soient correctement coordonnées avec les groupes fonctionnels et de l'infrastructure ou les fournisseurs de services de l'organisme statistique ou à l'externe. Il doit aussi être informé des enquêtes semblables et des nouvelles techniques et méthodes. Il doit veiller à ce que les participants soient conscients de la portée et de l'à-propos de leur engagement dans l'enquête et à ce qu'ils soient rapidement informés de toute modification apportée au plan. Il doit surveiller leur engagement et obtenir des preuves évidentes du rendement demandé. Une bonne partie de ce genre d'activités peut être déléguée aux membres de l'équipe de l'enquête, mais le chargé d'enquête devrait être en mesure de procéder à une vérification autonome de l'état d'avancement, habituellement par l'intermédiaire de communications mensuelles avec les chefs correspondants.

Le chargé d'enquête peut représenter l'organisme auprès du public et prendre la parole sur l'enquête. Il est la personne-ressource définitive pour les répondants de l'enquête, ceux qui veulent obtenir davantage d'information sur l'enquête et ceux qui portent plainte ou qui ont des questions. Lorsque les données de l'enquête sont diffusées, le chargé d'enquête est une personne-ressource (ainsi que le coordonnateur de la matière) qui répond aux questions des médias, des analystes des données et des chercheurs.

Le chargé d'enquête a surtout la responsabilité non exclusive de prévoir, d'empêcher et de résoudre les problèmes. Il doit avoir judicieusement recours à son expérience et à ses connaissances. Il ne doit jamais oublier d'être sceptique, réaliste et favorable simultanément. Il doit être pragmatique pour prendre des décisions ou adopter des positions qui ne sont pas toujours idéales du point de vue des relations avec le personnel et des demandes des utilisateurs. Il doit éviter d'ajuster ou de modifier ce qui ne tourne pas rond, car il ne ferait que perpétuer, voiler ou aggraver les problèmes. Il doit éviter le rafistolage, la perturbation et l'interférence.

Le chargé d'enquête doit surveiller la participation des membres de l'équipe. Afin d'éviter les problèmes, il doit essayer de maintenir la cohésion dans l'équipe pendant la démarche aussi longtemps que cette collaboration aide à obtenir le succès de l'enquête. Il ne doit cependant pas supposer que tous les participants resteront en poste jusqu'à la fin de l'enquête. Le chargé d'enquête devrait considérer une stratégie de relève non officielle et réfléchir à certaines options ou mesures de rechange. Il faut être disposé à prendre des mesures en cas d'absence d'un participant ou d'un membre de l'équipe à cause d'une maladie prolongée,

d'une promotion, d'une nouvelle affectation ou d'un départ de l'organisme. Il faut aussi prendre des dispositions pour remplacer le membre ou le participant qui perturbe l'enquête ou dont la contribution est inappropriée.

13.3.2 Compréhension des complications

Le chargé d'enquête devrait en définitive avoir l'expérience et les connaissances suffisantes pour comprendre les complications. En voici quelques-unes qu'il ne faudrait pas oublier (sans ordre particulier) :

- les membres de l'équipe n'ont pas les mêmes niveaux relatifs d'aptitudes ou la même expérience et certains n'ont pas les aptitudes appropriées,

- la répartition des responsabilités dans l'équipe peut être inappropriée ou disproportionnée,

- des communications médiocres ou des lacunes marquées dans les communications sont possibles à l'occasion,

- le président du Comité directeur et le chargé d'enquête mènent l'enquête (l'équipe n'est pas clairement informée sur l'orientation, ils ne consultent pas certains membres de l'équipe avant de prendre des décisions, ils réagissent aux problèmes ou aux questions à résoudre sans obtenir d'information contextuelle de l'équipe de l'enquête),

- les objectifs changent ou sont vagues (le client ou l'utilisateur ne sait pas ce qu'il veut ou ne comprend pas les questions, il ajoute des exigences par la suite, il essaie d'en faire trop dans une seule enquête),

- les méthodes, concepts ou questions sont excessivement complexes,

- les mises à l'essai sont inappropriées,

- il y a des erreurs de planification (imposer ou accepter un plan rigide, avoir un échéancier irréaliste ou affecter des ressources inappropriées, omettre les examens suffisamment détaillés des plans et des spécifications (examen seulement si quelque chose ne tourne pas rond), être trop optimiste pour déterminer combien de temps prendront les activités ou quelles sont les complications possibles, ou n'avoir aucune idée sur la question, constater que les ressources disponibles sont moindres que celles prévues),

- il y a interférence de l'externe,

- la compréhension des causes et effets manque lors de l'évaluation des problèmes et de la conception de solutions,

- l'accent est mis sur la méthodologie (comme une fin en soi) et non sur les objectifs,

- l'engagement ou la participation manque lorsque le Comité directeur ou la direction en a besoin,

- il n'y a pas d'engagement des membres de l'équipe (les affectations – engagements des participants sont trop nombreux ou ils sont distraits par d'autres activités hors de l'enquête ou des activités supplémentaires imposées dans l'enquête, par exemple, dépannage ou réponse aux demandes d'information hors du champ prévu de l'enquête, des membres de l'équipe considèrent des affectations

ultérieures ou acceptent des affectations avant l'achèvement de l'enquête en cours, il est impossible d'obtenir des participants pour faire ce qu'ils ne veulent pas faire ou ce dont ils doutent entièrement).

13.4 Sommaire

Ce chapitre couvre les principaux sujets de la planification et de la gestion d'une enquête, notamment :

i. Les méthodes d'organisation de la planification et de la gestion de l'enquête

Nous avons été particulièrement attentifs à l'approche de l'équipe de l'enquête à cause de sa capacité de mettre en commun des connaissances et aptitudes, de susciter l'engagement, de simplifier et d'améliorer les communications, et de donner ainsi une occasion d'obtenir une meilleure efficience et efficacité.

ii. Les étapes et le processus de la planification

La planification de l'enquête doit être faite par phases de plus en plus détaillées et précises, à partir de la formulation de la proposition de l'enquête pour en déterminer la faisabilité et établir le plan de l'enquête, en passant par la préparation de plans de composantes de l'enquête et la touche finale apportée aux plans pendant la conception et l'élaboration, jusqu'à l'ajustement et aux plans complémentaires pendant la mise en œuvre et l'évaluation.

iii. La méthode de gestion de l'enquête pour atteindre ses objectifs

Une bonne planification exige une bonne gestion, ainsi que des intervenants chevronnés et bien informés. Il faudrait nommer un chargé d'enquête responsable du fonctionnement approprié de tous les aspects de l'activité de l'enquête. Les principales fonctions de gestion comprennent l'organisation, l'orientation, la surveillance et le contrôle de l'enquête.

Bibliographie

Amabile, T.M. 1998. How to Kill Creativity. *Harvard Business Review*. September-October 1998: 65-74.

Biemer, P.P., R.M. Groves, L.E. Lyberg, N.A. Mathiowetz et S. Sudman, Éds. 1991. *Measurement Errors in Surveys*. John Wiley and Sons, New York.

Brackstone, G.J. 1993. Data Relevance: Keeping Pace with User Needs. *Journal of Official Statistics,* 9: 49-56.

Brackstone, G. 1999. La gestion de la qualité des données dans un bureau statistique. *Techniques d'enquête,* 25(2): 159-172.

Cialdini, R., M. Couper et R.M. Groves. 1992. Understanding the Decision to Participate in a Survey. *Public Opinion Quarterly,* 56: 475-495.

Collins, J. 1999. Turning Goals into Results: The Power of Catalytic Mechanisms. *Harvard Business Review*. July-August 1999: 71-82.

Cox, B.G., D.A. Binder, B.N. Chinnappa, A. Christianson, M.J. Colledge et P.S. Kott, Éds. 1995. *Business Survey Methods*. John Wiley and Sons, New York.

Dinsmore, P.C., Éd. 1993. The American Management Association Handbook of Project Management. AMACON, American Management Association, New York.

Drucker, P.F. 1999. Managing Oneself. *Harvard Business Review*. March-April 1999: 65-74.

Early, J.F. 1990. La gestion de la qualité dans les programmes statistiques nationaux. *Symposium 1990: Mesure et amélioration de la qualité des donnée,* Ottawa.

Eisenhardt, K.M., J.L. Kahwajy et L.J. Bourgeois III. 1997. How Management Teams Can Have a Good Fight. *Harvard Business Review*. July-August 1997: 77-85.

Fellegi, I.P. 1992. Planning and Priority Setting – the Canadian Experience. *Statistics in the Democratic Process at the End of the 20th Century*; *Anniversary publication for the 40th Plenary Session of the Conference of European Statisticians.* Federal Statistical Office, Federal Republic of Germany, Wiesbaden.

Fellegi, I.P. 1996. Characteristics of an Effective Statistical System. *International Statistical Review,* 64(2).

Freedman, D.H. 1992. Is Management Still a Science? *Harvard Business Review*. November-October 1992: 26-38.

Goleman, D. 1998. What Makes a Leader? *Harvard Business Review*. November – December 1998: 93-102.

Groves, R.M. 1989. *Survey Errors and Survey Costs*. John Wiley and Sons, New York.

Kish, L. 1965. *Survey Sampling*. John Wiley and Sons, New York.

Linacre, S.J. et D.J. Trewin. 1989. Evaluation of Errors and Appropriate Resource Allocation in Economic Collections. *Proceedings of the Annual Research Conference*. U.S. Bureau of the Census. 197-209.

Lyberg, L., P. Biemer, M. Collins, E. de Leeuw, C. Dippo, N. Schwarz and D. Trewin, Éds. 1997. *Survey Measurement and Process Quality*. John Wiley and Sons, New York.

Pinto, J.K., Éd. 1998. *The Project Management Institute Project Management Handbook*. Jossey-Bass Inc, San Francisco.

Project Management Institute. 2000. *A Guide to the Project Management Body of Knowledge*. 2000 Edition. Project Management Institute, Newton Square, PA.

Smith, T.M.F. 1995. Problématique de l'affectation des ressources.*Symposium 95, Des données à l'information: méthodes et systèmes: recueil.* 115-122.

Statistique Canada. 2000. Politique visant à informer les utilisateurs de la qualité des données et de la méthodologie. *Manuel des poliqitues*. Politique 2.3.

Statistics Canada. 1987. *Quality Guidelines*. Deuxième édition.

Statistique Canada. 1998. *Lignes directrices concernant la qualité*. Troisième édition. 12-539-XIF.

Statistique Canada. 2002. *Le Cadre d'assurance de la qualité*.

Sull, D.N. 1999. Why Good Companies Go Bad? *Harvard Business Review*. July-August 1999: 42-52.

Wang, R.Y. et D.M. Strong. 1996. Beyond Accuracy: What Data Quality Means to Data Consumers. *Journal of Management Information Systems,* 12(4): 5-34.

Liste de vérification de la planification

Voici certaines considérations à ne pas oublier lors de la planification des étapes de l'enquête :

1. Formulation de l'énoncé des objectifs

- besoins d'information de l'enquête,
- principales utilisations et principaux utilisateurs des données,
- définitions et concepts opérationnels,
- matière de l'enquête,
- plan d'analyse (c.-à-d. la structure et le niveau de détail des résultats de l'enquête).

Voir le **Chapitre 2 - Formulation de l'énoncé des objectifs**.

2. Sélection d'une base de sondage

- définition de la population cible,
- définition de la population observée selon les bases de sondage disponibles,
- base de sondage aréolaire, liste ou base de sondage multiple,
- utilisation des bases de sondage déjà créées,
- coût d'établissement d'une nouvelle base de sondage,
- données disponibles pour établir une nouvelle base de sondage,
- unités de la base de sondage,
- qualité de la base de sondage,
  - sous-dénombrement,
  - surdénombrement,
  - doubles,
- base de sondage à jour?

Voir le **Chapitre 3 - Introduction au plan d'enquête.**

3. Choix du plan d'échantillonnage

- recensement ou échantillon,
- échantillonnage probabiliste ou non probabiliste pour l'enquête-échantillon,
- si l'échantillonnage probabiliste est choisi :
  - échantillonnage aléatoire simple,
  - échantillonnage aléatoire simple stratifié,
  - échantillonnage par grappes,
  - échantillonnage à plusieurs degrés,
  - échantillonnage à plusieurs phases,
- si l'échantillonnage stratifié est choisi :
  - variables de stratification,
  - méthode de répartition de l'échantillon en strates,
- taille de l'échantillon,
- méthode d'estimation,
- degré de précision voulu (variance) des estimations,
- enquête unique ou réitérée?

Voir le **Chapitre 6 – Plans d'échantillonnage,** le **Chapitre 7 - Estimation** et le **Chapitre 8 – Calcul de la taille de l'échantillon et répartition**.

4. Conception du questionnaire

- méthodes de collecte :
  - assistée par intervieweur, autodénombrement ou observation directe,
- si la méthode assistée par intervieweur est appliquée :
  - interview sur place ou téléphonique,
- si la méthode par autodénombrement est appliquée :
  - méthode de distribution et de collecte des questionnaires,
- utilisation de certaines données administratives pour une partie de la collecte des données?
- recours à des substituts à titre de répondants?
- matière du questionnaire,
- formulation des questions,
- genres de questions :
  - réponses ouvertes ou fermées,
- ordre des questions.

Voir le **Chapitre 4 – Méthodes de collecte des données** et le **Chapitre 5 – Conception du questionnaire**.

5. Collecte des données

- comment situer les unités sélectionnées et communiquer avec elles,
- sélection des intervieweurs,
- formation des intervieweurs,
- supervision des intervieweurs,
- contrôle des documents (numéro de repérage des questionnaires complétés, etc.),
- procédures de travail sur place,
- contrôle de la qualité du travail sur place :
  - observation des intervieweurs,
  - nouvelles interviews,
- vérifications sur place,
- suivi des non-réponses.

Voir le **Chapitre 9 – Opérations de collecte des données.**

6. Saisie et codage des données

- saisie des données,
- codage des données :
  - méthode de classification à appliquer,
- méthodes de mesure, de contrôle et de correction des erreurs :
  - assurance de la qualité,
  - contrôle qualitatif.

Voir le **Chapitre 10 – Traitement** et l'**Annexe B – Contrôle qualitatif et assurance de la qualité.**

7. Vérification et imputation

- vérifications à faire sur place,
- vérifications à faire après la collecte sur place (vérifications manuelles et automatisées),
- genre de vérifications à faire :
  - vérifications de la validité,

  - vérifications de l'uniformité,
- uniformité des vérifications,
- méthodes d'imputation à appliquer,
- uniformité de l'imputation,
- préparation et mise à l'essai des systèmes d'imputation et de vérification.

Voir le **Chapitre 10 – Traitement.**

8. Estimation

- calcul des pondérations du plan d'échantillonnage,
- ajustements possibles pour :
  - le total des non-réponses,
  - les données auxiliaires,
- paramètres à estimer (estimations ponctuelles) :
  - totaux,
  - ratios,
  - proportions, etc.,
- estimateurs pour les estimations ponctuelles,
- estimateurs pour la variance d'échantillonnage des estimations ponctuelles.

Voir le **Chapitre 7 – Estimation**.

9. Analyse des données et présentation des résultats de l'enquête

- mesures de l'erreur d'échantillonnage,
- mesures de l'erreur non due à l'échantillonnage :
  - erreur de couverture,
  - non-réponses (p. ex., taux de non-réponses),
  - erreur de mesure,
  - erreur de traitement (p. ex., taux de rejets à la vérification),
- méthodes de mesure, de contrôle et de correction des erreurs :
  - assurance de la qualité,
  - contrôle qualitatif,
- évaluation de toutes les opérations de l'enquête,
- type d'analyses à faire,
- totalisation des données,
- rapports à produire,
- méthodes de contrôle de la divulgation,
- suppression ou modification de données.

Voir le **Chapitre 11 – Analyse des données de l'enquête,** le **Chapitre 12 – Diffusion des données** et l'**Annexe B – Contrôle qualitatif et assurance de la qualité.**

10. Diffusion des données

- utilisateurs et utilisations,
- moyens de diffusion :
  - publication sur support papier,
  - discours ou présentation en public,
  - interview à la radio ou à la télévision,

  - microfiches,
  - médias électroniques :
    - internet;
    - fichier de microdonnées,
- méthodes de contrôle de la diffusion.

Voir le **Chapitre 12 – Diffusion des données**.

11. Documentation

- auditoire cible :
  - direction,
  - personnel technique,
  - planificateurs d'autres enquêtes,
  - etc.,
- rapport d'enquête,
- rapports sur la méthodologie,
- rapports d'évaluation de la qualité des données,
- manuels de formation (p. ex., pour les interviews),
- rapports de rendement des intervieweurs,
- manuels d'instruction (p. ex., pour les répondants),
- échéancier des activités,
- spécifications pour les programmes des systèmes,
- rapport de faisabilité,
- rapports d'état d'avancement,
- rapport d'enquête (qui documente l'application de toutes les étapes de l'enquête),
- rapport d'analyse des données,
- rapport général ou rapports techniques.

Voir le **Chapitre 9 – Opérations de collecte des données** pour la documentation des opérations sur place et le **Chapitre 12 – Diffusion des données** pour la documentation en général.

Liste de vérification des coûts

Il faut tenir compte de certains éléments pour évaluer les coûts de l'enquête, notamment :

- la planification,
- la conception et l'élaboration :
  - le plan d'enquête,
  - les procédures de l'enquête (p. ex., la collecte des données),
  - le traitement après l'enquête,
- l'évaluation de l'enquête,
- la documentation,
- la formation du personnel.

Les coûts du **traitement de l'enquête** comprendraient le temps du personnel, l'achat ou la location du matériel et des logiciels, et d'autres services, p. ex., les bureaux, les meubles, les articles.

Les critères suivants déterminent la conception et l'élaboration d'une **application de l'interview assistée par ordinateur (IAO)** :

- la longueur et la complexité des questionnaires,
- le nombre et la complexité des vérifications intégrées,
- le progiciel utilisé,
- les exigences d'entrée préalable de l'information tirée d'un cycle précédent,
- les fonctionnalités nécessaires, par exemple les rapports de gestion, de pistage, etc.,
- le système d'échantillonnage, c.-à-d. les listes téléphoniques ou la composition aléatoire, etc.,
- la trousse de formation à intégrer à l'application,
- le nombre de révisions à apporter aux spécifications,
- la période de référence pour l'élaboration,
- les exigences de la mise à l'essai,
- la taille de l'échantillon (c.-à-d. si des mesures spéciales sont nécessaires à cause du nombre important de données).

Les éléments suivants déterminent le coût de la **formation du personnel** :

- le genre de formation (en classe, études à domicile, etc.),
- les degrés de formation (qui forme qui),
- la durée (heures, jours),
- l'endroit,
- le nombre d'intervieweurs formés,
- la location de matériel (p. ex., téléphone, ordinateurs, visualiseur d'OP).

Les éléments suivants déterminent le coût de la **collecte des données** :

- la taille de l'échantillon,
- la répartition de l'échantillon,
- la base d'échantillonnage,
- la durée de l'interview,
- la méthode de collecte des données (sur place, au téléphone, par la poste, etc.) :
  - p. ex., le nombre d'intervieweurs s'il s'agit d'une enquête assistée par intervieweur,
- le taux de réponses prévu,
- la stratégie de suivi,

- la population cible (interview de substituts ou non),
- exigences de pistage,
- échéancier de la collecte des données,
- vérifications manuelles et codage,
- exigences de la saisie des données et de la vérification,
- location de matériel (téléphone, ordinateurs),
- location de superficies,
- coût des déplacements des intervieweurs.

PUBLICATIONS ÉLECTRONIQUES
DISPONIBLES À

www.statcan.gc.ca

Annexe A - Données administratives

1.0 Introduction

Les demandes de statistiques sur de nombreux aspects de la société se maintiennent à la hausse. Une méthode à appliquer pour obtenir des données statistiques est, bien entendu, l'enquête. Les contraintes budgétaires et les préoccupations que suscite le fardeau des répondants ont cependant incité les organismes statistiques à examiner des méthodes de rechange pour obtenir des données statistiques. L'utilisation des données administratives est une méthode de rechange. ***Les données administratives sont celles qui ont été obtenues à des fins administratives (p. ex., pour administrer, réglementer ou percevoir des impôts auprès des entreprises ou des particuliers) et non à des fins statistiques (pour étudier des groupes de particuliers, d'entreprises, d'exploitations agricoles ou fermes, etc.).***

2.0 Utilisation des données administratives

Voici les principales utilisations statistiques des données administratives :

i. Totalisation directe ou analyse

Les données administratives sont, dans ce cas, la principale source de données pour les unités de l'échantillon, en tout ou en partie. Les données sont agrégées, analysées et diffusées de la même façon que les données d'enquête. Si des données administratives sont utilisées au lieu d'une enquête-échantillon, l'organisme statistique doit franchir certaines étapes de l'enquête étudiées dans ce manuel, mais pas toujours. L'organisme administratif ferait normalement, par exemple, la collecte, la saisie et le codage des données, mais l'organisme statistique devrait quand même procéder à la vérification, à l'imputation et à l'analyse des données. Dans certains cas, les données administratives peuvent être la seule source pratique (p. ex., information détaillée sur les frais des soins de santé).

ii. Estimation indirecte

L'estimation indirecte comprend l'utilisation de données administratives comme entrées dans le système d'estimation par l'intermédiaire de la régression, de l'estimation, du calibrage, etc., par exemple, l'utilisation de données administratives comme variables auxiliaires d'un modèle, comme on l'a vu au **Chapitre 7 - Estimation**. Elle comprend aussi la combinaison de données de plusieurs sources administratives pour produire des estimations.

iii. Bases de sondage

Les données administratives servent souvent à créer, compléter ou mettre à jour des bases de sondage (voir le **Chapitre 3 - Introduction au plan d'enquête**).

iv. Évaluation de l'enquête

Les données administratives peuvent servir à évaluer les données de l'enquête à l'échelon des microdonnées ou des données agrégées (consulter le **Chapitre 12 - Diffusion des données**).

Il y a six principales sources de données administratives :

i. Les enregistrements maintenus pour réglementer le cheminement des biens et des particuliers qui franchissent les frontières, y compris les dossiers d'importation, d'exportation, d'immigration et d'émigration.

ii. Les dossiers exigés par la loi pour enregistrer des événements, notamment les naissances, décès, mariages, divorces, constitutions en personne morale, octrois de permis, etc.

iii. Les dossiers nécessaires pour administrer les avantages sociaux ou les obligations, notamment les impôts, l'assurance-emploi, les régimes de retraite, l'assurance-santé, les prestations familiales, les listes électorales, etc.

iv. Les dossiers nécessaires pour administrer les établissements publics, par exemple les écoles, universités, établissements de santé, tribunaux, prisons, etc.

v. Les dossiers ouverts à la suite de la réglementation d'une branche d'activité par le gouvernement, notamment les transports, les activités bancaires, la radiotélédiffusion, les télécommunications, etc.

vi. Les dossiers ouverts pour la prestation de services publics, par exemple l'électricité, le téléphone, l'eau, etc.

Les concepts, les définitions, la couverture (et la mesure dans laquelle ces éléments restent constants), la qualité de la déclaration et du traitement des données, ainsi que la rapidité de leur disponibilité déterminent l'utilité des données administratives. Ces éléments peuvent varier énormément selon la source administrative et le genre d'information. Avant de décider d'utiliser les données administratives, il est nécessaire de les évaluer minutieusement, en prenant en compte les considérations suivantes :

i. Rapidité

Compte tenu de la source d'information, les intervenants d'une enquête qui utilisent seulement des données administratives peuvent être en mesure de produire des résultats plus rapidement que s'ils avaient recours à une enquête-échantillon. D'autre part, le programme administratif peut produire les données plus lentement qu'une enquête-échantillon (surtout si les données administratives constituent un recensement ou si elles sont tirées de plusieurs secteurs de compétence gouvernementale). Le traitement des données administratives après réception peut être particulièrement lent s'il faut combiner de nombreux fichiers.

ii. Coût

De nombreuses étapes de l'enquête peuvent être éliminées (en particulier la collecte des données) et les coûts diminuent donc.

iii. Fardeau de réponse

Il n'y a pas de fardeau de réponse si on utilise des données administratives au lieu d'administrer un questionnaire.

iv. Couverture

Les exigences administratives, qui peuvent être différentes des exigences statistiques, définissent la population cible.

v. Matière

Étant donné que les exigences administratives définissent la matière, les données administratives ne couvrent peut-être pas tous les sujets d'intérêt.

vi. Concepts et définitions

Le programme administratif, conçu aux fins d'autres objectifs, peut utiliser des définitions et concepts différents de ceux que le réalisateur de l'enquête aurait choisis. Les concepts de la source administrative pourraient en fait ne pas convenir au problème de la recherche.

vii. Erreur d'échantillonnage

Si les données administratives couvrent la population cible au complet (c.-à-d. qu'elles constituent un recensement), il n'y a donc pas d'erreur d'échantillonnage. Si les données administratives remplacent certaines données d'un échantillon de la population, l'erreur d'échantillonnage est toujours possible.

viii. Erreurs non dues à l'échantillonnage

Il est souvent plus difficile de contrôler les erreurs non dues à l'échantillonnage que dans le cas d'une enquête-échantillon. Il peut y avoir davantage d'erreurs ou d'omissions dans les données administratives que dans les données d'enquête (la vérification et l'imputation sont donc essentielles). Lorsque des particuliers ou des entreprises sont avantagés ou désavantagés, selon l'information fournie à la source administrative, l'information peut aussi être biaisée. Dans certains cas, les données administratives peuvent contenir moins d'erreurs que les données d'enquête, par exemple, lorsque l'erreur de mémoire peut amenuiser la capacité du répondant de répondre précisément aux questions ou lorsqu'il pourrait arrondir sa réponse à une question d'enquête (revenu).

ix. Contrôle qualitatif

Le contrôle exercé sur le programme administratif détermine la qualité de la collecte, de la saisie et du codage des données, et il peut être moins strict que celui d'un organisme statistique. Il faut donc procéder à des évaluations continues ou périodiques de la qualité des données reçues.

x. Fiabilité de la source administrative

La source administrative n'est peut-être pas fiable du point de vue de la prestation uniforme des données lorsqu'on en a besoin. La couverture, la matière et les concepts peuvent aussi changer avec le temps. Il faudrait donc collaborer avec les concepteurs du système administratif et maintenir la communication pour se tenir à jour sur les modifications proposées des concepts, des définitions, de la couverture, de la fréquence et de l'actualité qui peuvent avoir des répercussions sur leur utilisation statistique, et il faudrait intervenir en faveur de modifications à apporter qui amélioreront au lieu d'amenuiser leur utilisation statistique.

xi. Mise en forme des données

La mise en forme des données n'est peut-être pas pratique. Les données pourraient être agrégées seulement, par exemple, et l'organisme statistique préférerait des enregistrements individuels pour chaque unité. Les données peuvent provenir de plus d'une source, un problème éventuel de correspondance et d'uniformisation des données entre différentes mises en forme. Les fichiers ne sont peut-être pas bien documentés non plus.

xii. Questions de renseignements personnels

L'utilisation des données administratives peut susciter des préoccupations au sujet de la protection des renseignements personnels dans le grand public, surtout si les dossiers administratifs sont liés à d'autres sources de données. Il faudrait donc considérer les répercussions de la protection des renseignements personnels et les problèmes de contrôle de la divulgation, surtout lorsque les données sont liées à d'autres fichiers.

Bibliographie

Brackstone, G.J. 1987. Utilisation des dossiers administratifs à des fins statistiques. *Techniques d'enquête,* 13(1): 35-51.

Brackstone, G.J. 1988. Utilisations statistiques des données administratives: questions et défis. *Symposium 87: Les utilisations statistiques des données administratives: recueil*. 5-18. Ottawa

Cox, L.H. et R.F. Boruch. 1988. Record Linkage, Privacy and Statistical Policy. *Journal of Official Statistics,* 4: 3-16.

Hidiroglou, M.A., M. Latouche, B. Armstrong et M. Gossen. 1995. Improving Survey Information Using Administrative Records: The Case of the Canadian Employment Surveys. *Proceedings of the Annual Research Conference*. U.S. Bureau of the Census. 171-197.

Internal Revenue Service. 1999. *Statistics of Income: Turning Administrative Systems into Information Systems*. Washington, D.C.

Internal Revenue Service. 2000. *Statistics of Income Bulletin,* 19(4). Washington, D.C.

Kilss, B. et W. Alvey, Éds. 1984. *Statistical Uses of Administrative Records: Recent Research and Present Prospects*. 1. Department of the Treasury. Internal Revenue Service. Statistics of Income Division.

Kilss, B. and W. Alvey, Éds. 1984. *Statistical Uses of Administrative Records: Recent Research and Present Prospects*. 2. Department of the Treasury. Internal Revenue Service. Statistics of Income Division.

Konschnik, C.A., J.S. Johnson et J.N. Burton. 1998. The Use of Administrative Records in Current Business Surveys and Censuses. *Proceedings of the Section on Survey Research Methods*. American Statistical Association. 202-207.

Michaud, S., D. Dolson, D. Adams et M. Renaud. 1995. Combining Administrative and Survey Data to Reduce Respondent Burden in Longitudinal Surveys. *Proceedings of the Section on Survey Research Methods*. American Statistical Association. 11-20.

Monty, A. et H. Finlay. 1994. Strengths and Weaknesses of Administrative Data Sources: Experiences of the Canadian Business Register. *Statistical Journal of the United Nations,* ECE 11: 205-210.

Singh, M.P., J. Gambino et H.J. Mantel. 1994. Les petites régions: problèmes et solutions. *Techniques d'enquête,* 20(1): 3-15.

Statistique Canada 1996. Politique d'information des répondants aux enquêtes. *Manuel des politiques*. Politique 1.1

Statistique Canada 1996. Politique relative au couplage d'enregistrements. *Manuel des politiques*. Politique 4.1

Statistique Canada. 1998. *Lignes directrices concernant la qualité*. Troisième édition. 12-539-XIF.

Sweet, E.M. 1997. Using Administrative Record Persons in the 1996 Community Census. *Proceedings of the Section on Survey Research Methods*. American Statistical Association. 416-421.

Wolfson, M., S. Gribble, M. Bordt, B. Murphy et G. Rowe. 1987. La base de données de simulaiton de politique sociale: un exemple d'intégration de données d'enquête et de données administratives. *Symposium 87: Les utilisations statistiques des données administratives: recueil*. 233-268.

PUBLICATIONS ÉLECTRONIQUES
DISPONIBLES À

www.statcan.gc.ca

Annexe B - Contrôle qualitatif et assurance de la qualité

1.0 Introduction

S'ils ne sont pas prévus et maîtrisés pendant le déroulement de l'enquête, un bon nombre de problèmes peuvent ajouter des erreurs non dues à l'échantillonnage au point où les résultats de l'enquête seront inutiles. Réserver une partie du budget global de l'enquête aux programmes de contrôle qualitatif et d'assurance de la qualité est une bonne pratique pour minimiser et contrôler les erreurs qui peuvent être ajoutées à diverses étapes de l'enquête.

Le contrôle qualitatif est une technique employée pour garantir que la qualité est supérieure à un seuil déterminé en mesurant les caractéristiques d'intérêt de la qualité, en les comparant à une norme et en appliquant des mesures correctives si la norme n'est pas atteinte.

L'assurance de la qualité comprend toutes les activités qui visent à obtenir la qualité. Le but de l'assurance de la qualité est d'empêcher, de réduire ou de limiter l'occurrence d'erreurs dans une enquête pour obtenir les résultats voulus la première fois. L'assurance de la qualité applique une approche holistique pour obtenir la qualité nécessaire à la planification, à la mise à l'essai et à la surveillance appropriées de tous les processus et systèmes, afin qu'ils fonctionnent comme prévu. L'assurance de la qualité aide donc à prévoir les problèmes et le contrôle qualitatif, à répondre aux problèmes observés.

Nous devons définir la qualité avant d'expliquer plus en détail les procédures de contrôle qualitatif et d'assurance de la qualité. Statistique Canada considère que « l'aptitude à l'usage » des données définit la qualité. « L'aptitude à l'usage » comprend non seulement les concepts statistiques de variance et de biais, mais aussi d'autres aspects comme la pertinence et l'actualité qui déterminent dans quelle mesure les informations statistiques peuvent jouer pleinement leur rôle.

Statistique Canada définit la qualité, ou l'aptitude à l'usage » de l'information statistique en fonction de six paramètres : la pertinence, l'exactitude, l'actualité, l'accessibilité, l'intelligibilité et la cohérence (Statistique Canada, 2002).

Par ***pertinence*** des données statistiques, on entend la mesure dans laquelle les besoins réels des clients sont satisfaits. Pour être qualifiées de pertinentes, les données doivent éclairer les utilisateurs sur les questions les plus importantes à leurs yeux. L'évaluation de la pertinence est subjective, car celle-ci dépend des divers besoins des utilisateurs. Le défi du Bureau est de jauger les besoins conflictuels des utilisateurs actuels et potentiels afin de concevoir un programme qui réponde le mieux aux principaux besoins compte tenu des contraintes en matière de ressources.

Par ***exactitude*** des données statistiques, on entend la mesure dans laquelle l'information décrit bien le phénomène qu'elle doit mesurer. Elle est habituellement exprimée en terme d'erreur dans les estimations statistiques et est traditionnellement décomposée en biais (erreur systématique) et variance (erreur aléatoire). On peut aussi la définir par rapport aux sources principales d'erreurs susceptibles de mener à des données imprécises (par exemple, couverture, échantillonnage, non-réponse, réponse).

L'***actualité*** des données statistiques correspond au délai entre le point de référence (ou la fin de la période de référence) auquel se rapporte l'information et la date à laquelle les données sont disponibles. Il doit souvent y avoir compromis entre l'*actualité* et l'*exactitude* de l'information. L'*actualité* influera sur la *pertinence*.

Par ***accessibilité*** des données statistiques, on entend la facilité avec laquelle on peut se les procurer du Bureau. Il s'agit notamment de la facilité avec laquelle on peut constater que l'information existe de même que le caractère approprié de la présentation de l'information ou du média au moyen duquel on peut avoir accès aux données. Pour certains utilisateurs, le coût peut également être un aspect de l'*accessibilité*.

Par ***intelligibilité*** des données statistiques, on entend la disponibilité de renseignements supplémentaires et de métadonnées nécessaires à l'interprétation et à l'utilisation appropriée de ces données. Il s'agit en général de renseignements sur les variables, les classifications et les concepts sous-jacents utilisés, sur les méthodes de collecte et de traitement des données et sur les indicateurs de l'exactitude des données statistiques.

Par ***cohérence*** des données statistiques, on entend la mesure dans laquelle celles-ci peuvent être jumelées à d'autres renseignements statistiques dans un vaste cadre analytique au fil du temps. L'utilisation de concepts, de classifications et de populations cibles types favorise la cohérence, tout comme l'utilisation de méthodes d'enquêtes communes. Par ***cohérence***, on n'entend pas nécessairement la concordance numérique parfaite.

Ces éléments de la qualité se chevauchent et sont interreliés. Il est très difficile de construire un modèle statistique efficace qui permettrait d'intégrer toutes les caractéristiques de la qualité en un seul indicateur. Il n'y a eu que quelques essais (par exemple, Linacre et Trewin, 1993), sauf dans les cas simples ou unidimensionnels, de développement de modèles statistiques pour déterminer si un ensemble de caractéristiques de la qualité obtenues en particulier donnerait en général une qualité supérieure à un autre ensemble.

Considérer, gérer et équilibrer dans le temps les divers facteurs ou éléments qui constituent la qualité permettent d'obtenir un degré acceptable de qualité, mais il faut être attentif aux objectifs du programme, aux principales utilisations des données, aux coûts, ainsi qu'aux conditions et circonstances, qui ont des répercussions sur la qualité et les attentes des utilisateurs. Les éléments de la qualité nouent des liens complexes et toute intervention visant à considérer ou modifier un aspect de la qualité aura donc tendance à avoir des répercussions sur les autres éléments de la qualité. L'équilibre de ces facteurs peut donc changer de façons qui ne peuvent être facilement modélisées ou quantifiées correctement d'avance. Les décisions et les interventions qui permettent d'obtenir cet équilibre sont basées sur les connaissances, l'expérience, les examens, la rétroaction, la consultation et, inévitablement, le jugement.

2.0 Contrôle qualitatif

On a recours au contrôle qualitatif pour mesurer le rendement réel, le comparer aux normes et réagir à l'écart. Ainsi, le contrôle qualitatif se concentre sur un aspect de la qualité : l'exactitude. Le contrôle qualitatif est généralement appliqué, à l'étape du traitement de l'enquête, au travail habituellement accompli par des personnes ayant divers niveaux de formation et de capacité, et lorsque la tâche est répétitive et manuelle. Il s'applique donc à certaines activités, notamment, le codage, la saisie des données, les corrections manuelles (pendant ou après la collecte) et la vérification.

Le contrôle qualitatif ne se préoccupe de ce qui peut être mesuré et jugé acceptable ou non; si on ne peut pas mesurer, on ne peut pas faire de contrôle qualitatif. Parfois, bien que la mesure soit possible, il peut être trop onéreux, en temps et en argent, de faire un contrôle qualitatif (p. e. déterminer si la réponse à une question ouverte a été codée correctement).

Le nombre et l'importance des erreurs varient habituellement entre les activités et les particuliers qui travaillent à la même activité. Le contrôle qualitatif peut servir à identifier les éléments importants qui contribuent à l'erreur et garantir des degrés de qualité acceptables à la sortie.

Le contrôle qualitatif statistique est l'application des techniques statistiques aux fins de la comparaison avec des normes et pour obtenir un degré donné de qualité. Les programmes de contrôle qualitatif statistique permettent de limiter aux taux précisés les erreurs ajoutées à la suite d'une opération d'enquête, sous inspection minimale.

Les extrants d'une activité de travail peuvent être considérés selon deux perspectives différentes de la qualité. D'une part, les extrants représentent les produits individuels (ou services) conformes aux normes ou non. D'autre part, le travail peut être considéré comme une séquence de tâches accomplies dans des conditions relativement stables pour produire les extrants voulus (c.-à-d. une perspective du processus). Les deux points de vue sont valables et nécessaires selon les hypothèses formulées sur le processus et l'objectif de la procédure du contrôle qualitatif. Ces deux points de vue donnent lieu à deux méthodes principales de contrôle qualitatif : le contrôle du produit statistique et le contrôle du processus statistique.

2.1 Contrôle statistique du produit

Le contrôle statistique du produit utilise l'échantillonnage et des règles de prise de décisions pour déterminer les lots de travail acceptables et ceux qui ne le sont pas. L'objet du contrôle du produit est le lot individuel et non le processus. L'objectif du contrôle qualitatif est de déterminer quelles unités individuelles ou lots d'unités sont conformes aux exigences de qualité établies. Le contrôle du produit est une mesure corrective parce que les lots étiquetés « médiocres » sont améliorés (retravaillés). De plus, les erreurs trouvées dans le lot dont on a mesuré la qualité sont corrigées. (Les erreurs dans les lots « acceptés » ne sont pas toujours corrigées, par exemple les erreurs de collecte qui demanderaient une relance auprès du répondant.) Bien que ce soit une bonne pratique, il n'est pas toujours nécessaire d'identifier et de corriger les causes de la qualité médiocre. *L'échantillonnage d'acceptation* est le principal outil du contrôle qualitatif.

2.1.1 Échantillonnage d'acceptation

L'échantillonnage d'acceptation est une technique de contrôle qualitatif qui établit le plan d'échantillonnage et les règles de décisions pour déterminer quels lots sont acceptables ou non. Dans sa forme la moins compliquée, l'échantillonnage d'acceptation comprend la répartition du travail en lots, la sélection et la vérification d'un échantillon probabiliste dans chaque lot, et l'acceptation ou le rejet du lot, selon l'ampleur des erreurs relevées dans l'échantillon. Les autres lots rejetés font habituellement l'objet d'une inspection complète et ils sont rectifiés au besoin.

En particulier :
- la production est répartie en lots d'unités de taille N,
- un échantillon de taille n est sélectionné dans chaque lot,
- les unités de travail de l'échantillon font l'objet d'une inspection,
- le nombre total d'erreurs, d, de l'échantillon est comparé à une limite déterminée et le nombre acceptable est c,
- si $d > c$, le lot est rejeté et il fait l'objet d'une inspection complète, si $d <= c$, le lot est accepté sans autre inspection.

En créant les lots, on tente généralement de faire des lots de qualité homogène. Un lot contient habituellement le travail d'une seule personne sur une courte période de temps. Si cette personne travaille sur plusieurs objets simultanément (en codant deux variables différentes pour chaque questionnaire, par exemple), les lots ne devraient contenir qu'un seul objet. Cependant, plus les lots sont volumineux, moins on les inspecte, et on doit donc arriver à un compromis.

Le plan d'échantillonnage est précisé par les deux nombres n et c qui peuvent être calculés de diverses façons, selon le but que vise le contrôle qualitatif. Il y a plusieurs versions de l'échantillonnage d'acceptation. Dans le contexte du traitement des données d'enquête, les valeurs de n et de c sont fixées de sorte que le taux d'erreur à la sortie soit inférieur à une certaine borne appelée *qualité moyenne en sortie* (*average outgoing quality limit* ou *AOQL*), tout en minimisant le nombre d'inspections requises. Cette méthode assure que le niveau de qualité globale sur l'ensemble des lots dépasse un seuil minimal. C'est une assurance qu'à la fin du contrôle qualitatif, le nombre d'unités en erreur soit inférieur à *AOQL*.

Les valeurs de n et de c dépendent de :
- la qualité prévue des intrants (avant l'inspection),
- la qualité voulue des extrants,
- la taille du lot, N,
- le risque (probabilité) de prise de décisions erronées,
    - la probabilité de rejet d'un bon lot (erreur du type I),
    - la probabilité d'acceptation d'un mauvais lot (erreur du type II).

Voici d'autres considérations qui ont des répercussions sur la méthode de contrôle qualitatif :
- la définition de l'unité d'échantillonnage (p. ex., une partie du questionnaire, tout le questionnaire),
- la formation des lots (p. ex., le travail d'une journée),
- la mesure de la qualité (p.ex. le taux d'erreur, ou le nombre d'unités défectueuses par centaine produite),
- la définition et la classification des erreurs,
- la méthode de sélection de l'échantillon (p. ex., échantillon aléatoire simple (EAS), échantillonnage systématique ou par grappes),
- les procédures de rétroaction.

Le lecteur consultera avec profit Duncan (1986), Dodge et Romig (1959), Hald (1981) ou Smith et Mudryk (1989) pour en savoir davantage à propos de l'échantillonnage d'acceptation et sur l'établissement d'un plan d'échantillonnage.

La rétroaction en est habituellement une partie intégrale de tout programme de contrôle qualitatif officiel. La rétroaction, de vive voix ou par écrit, est habituellement faite à l'aide de rapports, de tableaux ou de graphiques sur les évaluations et les résultats de la qualité compilés pendant le processus d'inspection. Ces résultats sont ensuite réacheminés régulièrement à divers échelons du personnel affecté à l'opération de l'enquête. La rétroaction peut participer à l'amélioration de la qualité, ce qui se traduit par une diminution des taux d'échantillonnage (réduction de n) et des coûts.

Voici des exemples de rétroaction :

i. Donner aux opérateurs (p. e. commis au traitement) de l'information sur le rendement du groupe et leur rendement personnel (actuel et antérieur) et sur les causes les plus fréquentes de leurs erreurs. Les opérateurs peuvent ensuite suivre leur propre progrès, comparer leur rendement à celui de leurs pairs et déterminer explicitement où ils font des erreurs. Ce genre de rétroaction améliore la capacité de l'opérateur, le moral et la productivité.

ii. Donner aux surveillants une rétroaction sur le rendement des opérateurs. L'information comprend les taux d'erreur, les taux d'inspection et de rejet, ainsi que les estimations de la qualité des données de sortie. Cette information aide les surveillants à gérer efficacement les opérateurs, attribuer les ressources et répartir le travail, identifier les opérateurs et les secteurs à problème, et déterminer les besoins de formation.

iii. Remettre à la direction des sommaires des principaux indicateurs de qualité. Cette mesure aide la direction à repérer le progrès de l'application du point de la qualité et des coûts, à recommander des modifications à apporter aux objectifs opérationnels et à obtenir une assurance de la qualité pour le processus de l'enquête. Au cours d'une période soutenue, cette mesure peut inciter à modifier la méthodologie, les procédures ou les plans d'échantillonnage pour diminuer ensuite le nombre d'inspections.

2.2 Contrôle statistique du processus (CSP)

Un processus est une séquence d'activités planifiée orientée vers un résultat ou un but voulu, par exemple, la fabrication d'une pièce d'automobile. Chaque étape du déroulement d'une enquête peut être considérée comme un processus, par exemple, la sélection d'une base d'échantillonnage, la sélection de l'échantillon, la collecte des données, le traitement des données, etc. Tout processus comprend des intrants et des extrants. Les intrants peuvent comprendre des gens, du matériel, des méthodes, de l'équipement, un milieu, la direction. Les extrants du processus sont le produit ou le service.

Lors du contrôle statistique du processus, on suppose que les extrants sont les résultats d'un processus uniforme, bien défini, raisonnablement prévisible du point de vue de ces extrants, et qui produit des biens qui atteignent ou dépassent le niveau de qualité visé. Un tel processus est dit « sous contrôle ». Selon cette approche, l'objectif du contrôle qualitatif est d'échantillonner occasionnellement le processus qui fonctionne bien (c.-à-d. à des intervalles déterminés) pour vérifier si quelque chose a changé dans le processus (c.-à-d. s'il s'est détérioré).

Le contrôle statistique du processus est l'application de techniques statistiques pour mesurer et analyser la variation dans les processus. Il y a toujours une variation parce que les extrants que produit le même processus varient d'une certaine façon. Le plan d'échantillonnage (hasard simple, stratifié, en grappes, etc.) et les règles de décision servent à surveiller la qualité du processus et à lancer une intervention lorsqu'il est évident que le processus est hors contrôle. Les fluctuations mineures dans les mesures qui peuvent être dues à la variabilité de l'échantillonnage n'ont pas de répercussions sur cette procédure. Toutefois, lorsque les mesures dévient suffisamment, le processus est interrompu, les causes de la déviation sont déterminées et le processus est ajusté.

Le contrôle du processus est une mesure préventive parce que le processus est interrompu lorsqu'il devient hors contrôle, ce qui évite de produire des nombres importants d'extrants défectueux. On ne fait aucun effort visant améliorer directement la qualité en corrigeant des erreurs. Il s'agit d'identifier et de tarir les sources d'erreurs. Si possible, étant donné la chaîne d'opérations, le processus devrait être interrompu jusqu'à ce qu'on ait remédié à l'augmentation des défauts.

Il est habituellement possible d'identifier la cause profonde de la plupart des problèmes, mais il peut être difficile de le faire dans certains cas. Plusieurs outils disponibles aident à y arriver, y compris l'analyse Pareto, les graphiques de contrôle, les diagrammes cause-effet, les séances de remue-méninges, etc. Juran et Godfrey (1998) discutent des analyses de Pareto et des diagrammes cause-effet.

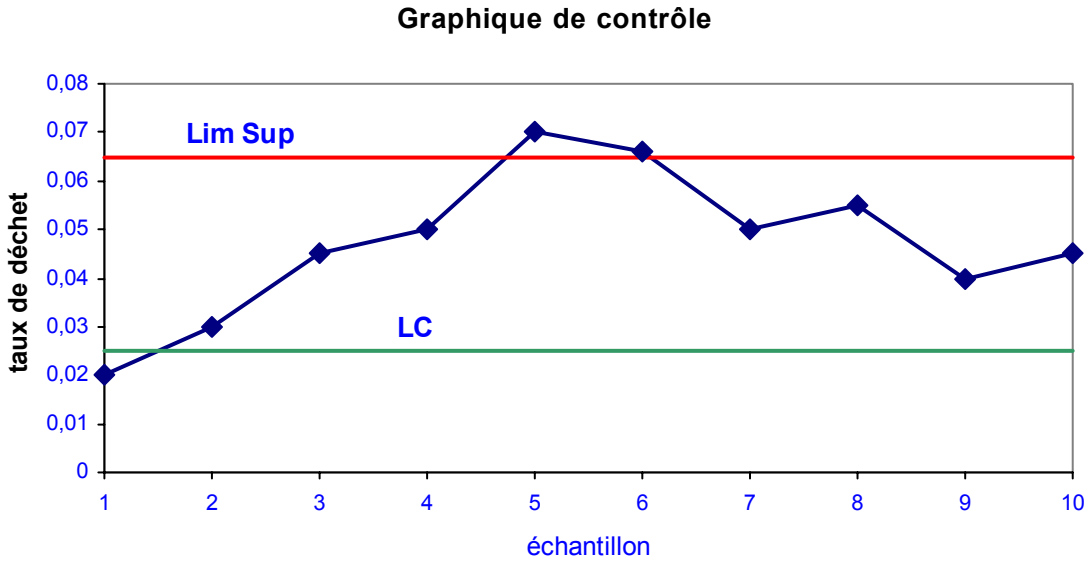
Tout comme le contrôle du produit, la rétroaction sur le contrôle du processus devrait être communiquée aux opérateurs, aux surveillants et à la direction.

2.2.1 Graphiques de contrôle

Le graphique de contrôle est le principal outil du contrôle statistique du processus. Un graphique de contrôle trace les mesures consécutives de l'échantillon tirées périodiquement d'un processus pour examiner si elles sont dans les limites établies par la variation du processus. L'ensemble des limites est intitulé limites de contrôle. Les limites supérieure et inférieure de contrôle peuvent être établies en fonction du jugement d'en expert, mais elles sont habituellement établies à trois écarts-types de la ligne du centre qui représente la valeur moyenne ou moyenne du processus. C'est l'équivalent de calculer les bornes d'un intervalle de confiance (voir **Chapitre 8 - Calcul de la taille de l'échantillon et répartition**). Les lignes de contrôle sont les valeurs à l'extérieur desquelles le processus est considéré hors contrôle. L'axe vertical représente la mesure de la qualité (p. ex., pour cent des données défectueuses) et l'axe horizontal affiche la valeur de chaque observation de l'échantillon en séquence chronologique.

Si toutes les observations de l'échantillon sont dans les limites de contrôle, le processus est considéré sous contrôle. Si une observation ou plus est(sont) hors des limites de contrôle, il faudrait interrompre le processus et faire enquête pour déterminer les causes de la perte de contrôle. Même quand aucune observation ne se pose hors des limites, mais qu'on observe une tendance, par exemple plusieurs lots à la suite se positionnent au-dessus de la ligne centrale, ou on observe une augmentation constante, il y aurait lieu d'inspecter le processus.

Divers graphiques de contrôle sont disponibles selon la mesure de la qualité appliquée et la taille de l'échantillon fixe ou variable. Le graphique de contrôle des attributs le plus commun est le « graphique de contrôle » illustré ci-dessous.



On remarquera que le graphique ne porte pas de limite inférieure; dans le traitement des données d'enquête, la mesure d'intérêt est le taux d'erreur. Le fait que le taux d'erreur baisse ne constitue pas une source d'inquiétude.

Pour des détails sur d'autres graphiques de contrôle et déterminer comment en calculer la ligne du centre et les limites de contrôle, consulter Duncan (1986), Schilling (1982) ou Wheeler (1986).

2.3 Contrôle statistique du *produit* et contrôle statistique du *processus*

Le contrôle statistique du *produit* (habituellement par échantillonnage d'acceptation) se préoccupe des extrants : le produit. Dans le contexte du traitement d'une enquête, le but est de détecter et corriger assez d'erreurs pour que le produit satisfasse aux exigences minimales de la qualité. Des plans de contrôle qualitatif peuvent aider à minimiser la probabilité qu'un lot de qualité médiocre soit néanmoins accepté (probabilité d'une erreur de type II), ou que la qualité globale de l'ensemble des lots soit acceptable. Le contrôle statistique du *processus* ne cherche pas à détecter les erreurs; on part de l'hypothèse que la qualité est déjà acceptable – on n'a pas à corriger d'erreurs – et on vérifie la qualité de certains lots pour s'assurer que celle-ci est toujours acceptable. Si les données contredisent l'hypothèse, il ne s'agit pas de corriger les erreurs, mais de corriger le processus.

On a rarement recours au contrôle statistique du processus dans les traitements d'enquête parce que cela suppose un processus qui a fonctionné avec consistance et fiabilité au niveau de qualité espéré. Ce genre de processus est rare dans les enquêtes dont les opérations sont en partie manuelles. Avec l'expérience, les opérateurs deviennent davantage « fiables » et obtiennent des promotions à des postes d'encadrement, ou passent à un autre projet. De plus, quand la qualité du travail d'un opérateur commence à décliner, il est difficile de relever rapidement la qualité; la cause de la baisse de qualité est souvent liée à l'ennui, la fatigue, etc.

Cependant, ne serait-ce que parce qu'il est moins onéreux, le contrôle statistique du processus est préférable à l'échantillonnage d'acceptation; les échantillons de contrôle statistique du processus sont plus petits que ceux exigés par l'échantillonnage d'acceptation.

Le contrôle statistique du processus s'est montré particulièrement bien adapté aux opérations de saisie automatisées (*ICR = Intelligent Character Recognition* ou reconnaissance intelligente de caractères). On a soumis la saisie du Recensement de l'agriculture de 2001 au contrôle statistique du processus. Comme le travail est exécuté par une machine, une fois le niveau de qualité requis atteint, on pouvait s'attendre à ce qu'il soit maintenu.. Si les taux d'erreurs augmentaient subitement, l'ajustement ou le remplacement d'une pièce suffisait à corriger la situation. La surveillance des interviews par ITAO se prête aussi au contrôle statistique du processus. Dans ce genre d'application, on échantillonne des appels et les erreurs (question mal posée, impolitesse, etc.) totalisées. Comme il est impossible de corriger de telles erreurs, c'est-à-dire faire du contrôle qualitatif, on doit plutôt surveiller le processus, c'est-à-dire faire du contrôle statistique de processus.

2.4 Contrôle d'acceptation

Voici une importante question à considérer dans les opérations d'enquête : quand doit-on appliquer quelles méthodes de contrôle statistique de la qualité, en particulier l'*échantillonnage d'acceptation* ou le *contrôle statistique du processus* (CSP)? Comme on l'a expliqué plus haut, le début de plusieurs opérations d'enquête commence de façon assez imprévisible parce qu'elles mettent en jeu beaucoup de personnel avec un taux élevé de roulement. Cependant, la formation, l'expérience et la rétroaction

permettent de stabiliser ces processus. Tirer avantage de cette stabilisation et modifier les procédures d'inspection est donc souvent une bonne pratique, afin de réduire éventuellement le nombre des inspections et les coûts connexes.

Diverses stratégies d'inspection sont disponibles à cette fin, y compris : l'inspection réduite (c.-à-d. prendre de plus petits échantillons et augmenter le risque d'accepter un lot de moindre qualité), l'inspection plus serrée (c.-à-d. prendre des échantillons plus grands et réduire le risque d'accepter des lots médiocres), l'inspection normale, l'inspection à 100 % et les vérifications au hasard. Il peut aussi s'agir d'abaisser le niveau de qualité visé si on doit lui consacrer beaucoup de temps et de ressources.

L'ampleur de la stabilité du processus qui est évidente détermine la méthode qu'il faudra appliquer. En termes généraux, plus un processus est stable et prévisible, moins l'inspection est nécessaire (c.-à-d. qu'un risque plus grand à l'échantillonnage peut être justifié).

Schilling (1982) a formulé le postulat de l'*approche du contrôle d'acceptation qui comprend une stratégie continue de sélection, d'application et de modification des procédures d'échantillonnage d'acceptation en milieu d'inspection changeant.* Les procédures d'inspection modifiées périodiquement sont une fonction du *degré de qualité* atteint et des *antécédents de la qualité* disponibles. Le principe prédominant du contrôle d'acceptation est d'adapter continuellement les procédures d'acceptation aux conditions présentes (qui changent généralement avec le temps). La structure qui sert à déterminer quand changer de procédures d'inspection est affichée dans le tableau suivant. Ce tableau est considéré plus en détail au chapitre 19 de l'ouvrage de Schilling (1982).

Tableau 1 : Contrôle d'acceptation – Procédure d'inspection à l'aide des antécédents de la qualité et des résultats précédents

	Antécédents de la qualité relative		
Résultats précédents	**Minimes**	**Moyens**	**Approfondis**
	< 10 lots	de 10 à 50 lots	> 50 lots
Excellents	Plan normal	Vérifications réduites – de lots non successifs	CSP – vérifications au hasard
Moyens	Plan normal	Plan normal	Vérifications réduites – de lots non successifs
Médiocres	100 %	100 % – vérifications plus étroites	100 % – vérifications plus étroites

On peut remarquer dans le tableau ci-dessus que le processus d'inspection du contrôle d'acceptation devient dynamique et change à mesure que le processus s'améliore ou se détériore. En général, lorsque la qualité s'améliore et que les antécédents de la qualité qui soutiennent cette constatation s'approfondissent, les plans d'échantillonnage sont modifiés pour passer des plans normaux à des plans avec inspections réduites et ensuite, à des plans avec inspections par sauts (*skip-lot sampling* échantillonnage d'acceptation où l'on laisse passer des lots sans les inspecter si la qualité. des lots précédents est élevée), à des plans avec CSP, puis à des vérifications périodiques au hasard. L'objectif ultime de la stratégie du contrôle d'acceptation est de réduire continuellement les inspections et les coûts connexes, tout en maintenant les degrés de qualité déterminés.

3.0 Assurance de la qualité

Les erreurs peuvent coûter chères et être difficiles à corriger, et il faudrait donc insister sur la prévention des erreurs aux premières étapes de l'enquête. L'assurance de la qualité couvre tous les aspects de la qualité; son but est d'empêcher les erreurs de se produire en premier lieu.

Par exemple, une stratégie générale d'assurance de la qualité aux fins du contrôle des erreurs non dues à l'échantillonnage est de prévoir les problèmes avant qu'ils ne se posent, et prendre les mesures pour les empêcher ou les minimiser, idéalement aux étapes de la planification et de la conception de l'enquête.

Voici des exemples de l'assurance de la qualité :
- élaborer une planification intensive,
- procéder à une étude de faisabilité,
- faire une enquête pilote (c.-à-d. mise à l'essai du système d'enquête complet, du début à la fin, à petite échelle),
- former les intervieweurs, les surveillants, les opérateurs de la saisie des données, les codeurs, etc.,
- organiser des séances d'information,
- améliorer la base d'échantillonnage,
- améliorer le plan d'échantillonnage,
- améliorer la conception du questionnaire,
- modifier la méthode de la collecte des données (p. ex., passer de la collecte sur support papier à la collecte assistée par ordinateur),
- prévoir de meilleurs suivis de routine,
- formuler des procédures de traitement plus claires,
- faire des essais approfondis de tous les systèmes de traitement avant de les utiliser,
- vérifier au hasard la collecte des données et les résultats des activités d'autres grandes enquêtes.

Les lignes directrices concernant la qualité (1998) de Statistique Canada recommandent les activités d'assurance de la qualité suivantes pendant l'étape de la conception et de la mise en œuvre d'une enquête :

i. L'implantation d'un régime comprenant un comité directeur et une direction du projet pour garantir que les programmes statistiques se déroulent selon leur mandat. Cette mesure donne un mécanisme d'examen, de surveillance et de rapport sur l'état d'avancement, les problèmes et les questions, elle garantit l'interprétation appropriée du mandat et de l'objectif, ainsi que l'expression de jugements appropriés.

ii. L'application d'une approche par direction du projet-équipe de projet interdisciplinaire pour la conception et la mise en œuvre, afin de garantir que les considérations sur la qualité obtiennent l'attention appropriée.

iii. Lorsque des méthodes particulières sont appliquées, elles devraient correspondre à l'ensemble des pratiques statistiques acceptées et justifiables, compte tenu des circonstances. Il faudrait favoriser le recours à de nouvelles technologies et aux innovations pour améliorer la qualité et l'efficience après les avoir mis à l'essai pour minimiser le risque. Il faudrait mettre les questionnaires à l'essai pour vérifier si les répondants comprennent les questions et peuvent donner les réponses voulues, selon un degré de qualité acceptable. Il est important de surveiller la qualité, d'intervenir efficacement en cas de problèmes imprévus, de vérifier ou de soutenir la crédibilité des résultats et d'en comprendre les limites.

iv. À l'étape de la conception ou de la nouvelle conception et dans le cadre des examens en cours, il devrait y avoir des évaluations techniques des méthodes proposées ou appliquées, ainsi que de l'efficacité opérationnelle et des coûts par rapport au rendement. Cette mesure permettra de vérifier si les pratiques ou les propositions techniques sont convenables. Elle aidera aussi à améliorer et à orienter la mise en œuvre de composantes particulières de la méthodologie et des opérations dans les programmes et entre eux.

v. L'analyse des données sert à décrire les phénomènes statistiques, à informer en ce sens, et à découvrir les lacunes des données, mais elle devrait aussi être un moyen d'évaluer ou de mesurer l'exactitude et la convergence des données. Dans ce contexte, les résultats de l'analyse peuvent déboucher, par exemple, sur des procédures supplémentaires ou modifiées de vérification, des changements apportés à la conception du questionnaire, des procédures de collecte de données supplémentaires, d'autres séances de formation du personnel, l'application de nouvelles méthodes, procédures ou systèmes, ou une nouvelle conception.

Du point de vue du travail, il est important de favoriser un milieu qui suscite l'intérêt pour la qualité et l'atteinte de la meilleure qualité possible dans les limites opérationnelles et budgétaires. Ce volet comprend :

- le recrutement de personnes talentueuses et leur perfectionnement pour qu'elles apprécient les questions de qualité,
- un réseau de communication interne ouvert et efficace,
- des mesures explicites pour élaborer des partenariats et approfondir la compréhension des fournisseurs de l'organisme (en particulier les répondants),
- l'élaboration et le maintien de définitions, classifications, structures et outils méthodologiques standard pour soutenir l'intelligibilité et la cohérence.

Il faudrait enfin documenter toutes les procédures de contrôle qualitatif et d'assurance de la qualité. Cette documentation devrait comprendre :

i. Les options, le choix éventuel et la justification : Le choix des procédures de contrôle qualitatif et de l'assurance de la qualité en particulier n'est pas évident pour toute opération et les éléments pris en considération devraient faire l'objet d'une discussion.

ii. Les procédures : Il faudrait prévoir des instructions ou un manuel à l'intention des surveillants et des vérificateurs.

iii. Les rapports : Il faudrait produire des rapports périodiques sur les résultats des procédures de contrôle qualitatif et sur le rendement de chaque opérateur, afin de faire rapport sur la qualité ou d'identifier les opérateurs qui ont besoin davantage de formation.

Bibliographie

Brackstone, G. 1999. La gestion de la qualité des données dans un bureau statistique. *Techniques d'enquête,* 25(2): 159-172.

Dodge, H.F. et H.G. Romig. 1959. *Sampling Inspection Tables: Single and Double Sampling*. Second edition. John Wiley and Sons, New York.

Dufour, J. 1996. *Labour Force Survey Data Quality*. Statistics Canada. HSMD-96-002E/F.

Duncan, A.J. 1986. *Quality Control and Industrial Statistics*. Fifth edition. R.D. Irwin Inc., Illinois

Fellegi, I.P. 1996. Characteristics of an Effective Statistical System. *International Statistical Review,* 64(2).

Groves, R.M. 1989. *Survey Errors and Survey Costs*. John Wiley and Sons, New York.

Hald, A. 1981. *Statistical Theory of Sampling Inspection by Attributes.* Academic Press, New York.

Juran, J.M. et A.B. Godfrey. 1998. *Juran's Quality Handbook*. Fifth Edition. McGraw-Hill, New York.

Linacre, S.J. et D.J. Trewin. 1989. Evaluation of Errors and Appropriate Resource Allocation in Economic Collections. *Proceedings of the Annual Research Conference*. U.S. Bureau of the Census. 197-209.

Linacre, S.J. et D.J. Trewin. 1993. Total Survey Design – An Application to a Collection of the Construction Industry, *Journal of Official Statistics,* 9(3): 611-621.

Lyberg, L. 1997. *Survey Measurement and Process Quality*. John Wiley and Sons, New York.

Mudryk, W. 2000. Note de cours STC446 *Méthodes statistiques pour le contrôle de la qualité*. Ottawa.

Mudryk, W., M.J. Burgess et P. Xiao. 1996. Quality Control of CATI Operations in Statistics Canada. *Proceedings of the Section on Survey Research Methods*. American Statistical Association. 150-159.

Schilling, E.G., 1982. *Acceptance Sampling in Quality Control*. Marcel Dekker, New York.

Smith, J., W. Mudryk, et R. Stankewich. 1989. *Standardization of QC Sampling Plans for Survey Operations, Part 1: Guidelines and Rationale*, Quality Control Section, Business Survey Methods Division, Statistics Canada.

Statistique Canada. 1998. *Lignes directrices concernant la qualtié*. Troisième édition. 12-539-XIF.

Wheeler, D.J. et D.S. Chambers. 1986. *Understanding Statistical Process Control.* SPC Press, Knoxville, TN.

Williams, K,C. Denyes, M. March et W. Mudryk. 1996. Mesure de la qualité durant le traitement des données d'enquête. *Symposium 96: Erreurs non dues à l'échantillonnage : recueil.* Statistique Canada. 131-142.

PUBLICATIONS ÉLECTRONIQUES
DISPONIBLES À

www.statcan.gc.ca

Méthodes et Pratiques d’enquête - Étude de cas

PUBLICATIONS ÉLECTRONIQUES
DISPONIBLES À

www.statcan.gc.ca

Préface

Cette Étude de cas propose une enquête fictive conçue pour suivre pas à pas le développement d'une enquête générale auprès de ménages. On y retrouvera les méthodes et principes énoncés dans les chapitres correspondants des *Méthodes et pratiques d'enquête*. De cette façon, un seul exemple suffit à illustrer toute la matière du livre. Le processus d'élaboration de l'enquête de l'Étude de cas est décrit du point de vue de l'équipe de projet à qui l'on en aurait confié la responsabilité.

Table des matières

1. Introduction aux enquêtes 351

2. Formulation de l'énoncé des objectifs 354

3. Introduction au plan d'enquête 364

4. Méthodes de collecte de données 367

5. Conception du questionnaire 369

6. Plans d'échantillonnage. 377

7. Estimation 380

8. Calcul de la taille d'échantillon et répartition 385

9. Opérations de collecte de données 392

10. Traitement 397

11. Analyse des données 401

12. Diffusion des données 405

13. Planification et gestion de l'enquête 408

Chapitre 1 - Introduction aux enquêtes

1.0 Introduction

Un changement économique rapide est remarqué depuis quelques années au pays de Belleterre qui passe par un processus de réforme économique. Le Bureau de statistique de Belleterre (BSB) est bien conscient qu'il est de plus en plus urgent d'obtenir de l'information à jour sur l'état de l'économie et la situation socioéconomique de la population. Il a entrepris des efforts pour moderniser plusieurs aspects de son programme statistique.

Le BSB convient en particulier qu'il faut obtenir de l'information pertinente, objective et précise sur la situation des ménages en milieu rural et urbain. Il veut à cette fin obtenir des ressources pour lancer une enquête permanente sur les ménages qui pourrait être intitulée Enquête générale sur les ménages (EGM). Le gouvernement national affirme soutenir l'enquête et détermine actuellement les ressources qu'il faudrait réserver à cet effort.

Les études préliminaires des demandes d'information de divers ministères utilisateurs indiquent que l'EGM devrait avoir lieu une fois par année et être représentative de la population à l'échelon du pays, des importants centres urbains et des régions économiques infranationales. Au nombre des grands sujets ciblés par l'enquête, on compte :

- les caractéristiques sociodémographiques,
- l'activité du marché du travail,
- les caractéristiques des revenus et dépenses,
- les indicateurs des conditions de vie.

Un certain nombre d'importants ministères nationaux veulent aussi obtenir de l'information de l'EGM. Le ministère de la Santé apprécierait de l'information détaillée sur la santé de la population et le ministère de l'Agriculture a demandé des données sur les activités agricoles des ménages en milieu rural et urbain. Plusieurs ministères économiques voudraient de l'information sur les petites entreprises. Toutes ces activités supplémentaires sont considérées, mais aucune décision définitive n'a encore été prise sur les points, s'il en est, qui seront ajoutés à l'EGM.

Dans le contexte de ces grands objectifs, une équipe de projet est formée pour concevoir et mettre en œuvre la nouvelle enquête. Vous avez été choisi membre de l'équipe du projet et commencerez bientôt à participer à l'élaboration de l'enquête. La première réunion de l'équipe aura lieu sous peu et vous réservez du temps pour examiner l'information disponible sur Belleterre à partir du dernier recensement et d'autres enquêtes sur les ménages. Vous réalisez rapidement que la majeure partie de l'information, non seulement n'est plus à jour, mais qu'elle est aussi insuffisante, parce que les données disponibles ne reflètent pas les nouvelles réalités économiques. Voir l'Annexe 1.1 pour un aperçu de Belleterre.

<u>Questions de récapitulation</u> :

Pourquoi une enquête a-t-elle été proposée?

Quels sont les points élémentaires à considérer dans l'enquête?

Annexe 1.1 : Profil statistique de Belleterre

Voici un sommaire de l'information disponible sur la situation économique et démographique de Belleterre. Les données sont tirées des recensements de 1970 et de 1994. Les notes comprennent certaines projections et estimations préliminaires formulées à l'aide du recensement de 1994.

La population de Belleterre s'élève actuellement à environ 44 millions de personnes, comparativement à 30 millions environ en 1970 et à un peu moins de 41 millions lors du recensement de 1994. Le taux de croissance modéré de la population se maintient et elle devrait atteindre 55 millions de personnes d'ici 20 ans. Cette croissance de la population, ainsi que le taux élevé de la migration du milieu rural en milieu urbain, ont accéléré les récents changements dans la structure économique du pays.

L'urbanisation est à la hausse depuis deux décennies à Belleterre où la migration de la campagne vers les centres urbains est importante, en particulier dans les trois plus grandes villes. Environ 40 % seulement de la population habitent maintenant en milieu urbain.

La capitale, Ville A, est établie dans la région de la vallée centrale. Sa population en croissance rapide devrait atteindre près de quatre millions de citoyens au milieu 2005. Dans les deux autres principaux centres urbains, Ville B au sud du pays et Ville C au nord, le rythme de croissance de chacune est semblable et inférieur à celui de Ville A. Ces trois villes qui englobaient un peu moins du douzième de la population de Belleterre en 1970 comprennent maintenant près du cinquième d'un total plus élevé et elles continuent d'attirer un débit constant de migrants de la campagne à proximité.

L'économie du pays est toujours largement axée sur les ressources naturelles et agricoles, les principaux produits agricoles étant le riz et le café. L'exploitation minière, en particulier le cuivre et le charbon, favorisent les échanges avec l'étranger et contribuent ainsi à l'économie. L'infrastructure de la fabrication est en expansion rapide, en particulier pour les textiles et les composantes électroniques, et les produits sont exportés à l'étranger en majeure partie. Le revenu par personne dans l'ensemble, selon les estimations, a augmenté de 6,4 % par année en moyenne depuis 1990.

Les tableaux suivants affichent un sommaire des indicateurs économiques et démographiques importants.

Tableau 1.1 : Population de Belleterre

Année	Population (au milieu de l'année)	Source
1970	30 110 000	Recensement de 1970
1994	40 850 000	Recensement de 1994
2000	43 840 000	Estimation du BSB
2005	47 200 000	Projection démographique du BSB
2010	50 000 000	Projection démographique du BSB
2015	52 800 000	Projection démographique du BSB
2020	55 300 000	Projection démographique du BSB

Tableau 1.2 : Belleterre : Population (en milliers de personnes), par région

Région	Source		
	Recensement de 1970	**Recensement de 1994**	**Estimation 2000 du BSB**
Ville A	1 760	3 250	4 080
Ville B	925	1 675	2 060
Ville C	2 145	3 189	3 625
District D	1 885	2 467	2 600
District E	3 400	4 450	4 690
District F	3 670	4 800	5 045
District G	3 085	3 975	4 160
District H	2 300	2 965	3 080
District I	3 200	4 120	4 320
District J	4 260	5 480	5 640
District K	3 480	4 470	4 540
Total	**30 110**	**40 850**	**43 840**

Tableau 1.3 : Belleterre : Produit intérieur brut par personne

Année	PIB par personne (prix de 1990)	Source
1990	5150	Estimation intercensitaire (BSB)
1994	6175	Données corrigées du recensement
2000	9600	Estimation préliminaire

Chapitre 2 - Formulation de l'énoncé des objectifs

2.0 Formulation de l'énoncé des objectifs

L'équipe de projet chargée de l'élaboration de la nouvelle Enquête générale sur les ménages (EGM) est officiellement nommée et se réunit pour la première fois pour commencer son travail. Elle fera rapport au Comité directeur formé de représentants de la direction et de professionnels chevronnés, et elle doit préparer rapidement un plan de travail détaillé pour présentation au Comité directeur.

L'équipe sait que l'EGM doit couvrir un large éventail de sujets et donner de l'information au moment opportun plusieurs fois par année. Elle convient aussi qu'il y a des millions de personnes au pays et constate rapidement qu'il ne serait pas réaliste ou nécessaire de faire un recensement. Une enquête-échantillon sera suffisante et probablement préférable à un recensement (voir le Chapitre 3 de l'étude de cas), même si le genre et la taille de l'échantillon ne sont pas encore précis.

L'équipe décide qu'il faut réserver les premières réunions à l'élaboration d'un énoncé d'objectifs pour approbation au Comité directeur avant d'entreprendre l'élaboration détaillée de l'enquête.

2.1 Détermination des besoins d'information

L'équipe de projet commence à travailler à l'énoncé des objectifs pendant la deuxième réunion. Compte tenu de la longue liste de demandes d'information dont l'équipe est consciente, l'objectif général de l'EGM pourrait être énoncé librement comme suit : étudier les conditions économiques et sociales de la population. Cet énoncé est beaucoup trop vague pour l'appliquer directement en pratique et vous ne pouvez certainement pas espérer obtenir de l'information utile en posant simplement des questions aux gens sur leurs « conditions économiques et sociales ».

L'équipe doit donc relever deux défis. Le premier est de convertir l'énoncé général vague des besoins en sujets particuliers pour l'étude et le deuxième est de déterminer lesquels de ces sujets peuvent et devraient être couverts dans l'enquête.

Les quatre grands thèmes de l'information sociodémographique, de l'activité du marché du travail, des caractéristiques des revenus et dépenses, et des conditions de vie sont déjà considérés comme des priorités élevées. Divers ministères ont aussi demandé des données sur la santé, la production agricole et l'activité des petites entreprises.

L'équipe doit vérifier s'il est possible ou nécessaire d'intégrer une telle variété de sujets dans une seule enquête, et elle doit faire des recommandations au Comité directeur sur la faisabilité, les avantages et les risques de l'intégration de chacun des principaux domaines.

L'équipe doit essayer d'identifier et de consulter les principaux utilisateurs des données et de s'informer sur les définitions et les concepts pertinents de la matière pour déterminer les besoins particuliers d'information.

2.2 Utilisateurs et utilisations des données

Une liste des principaux utilisateurs des données est dressée à partir de conseils des membres du Comité directeur. Certains d'entre eux sont au BSB et travaillent dans des domaines spécialisés, notamment les divisions de la statistique du secteur de la fabrication, de l'analyse démographique et des comptes nationaux. D'autres sont des analystes de politiques des ministères des programmes centraux qui s'intéressent à certains domaines, notamment, la politique industrielle et de l'emploi, la politique de la construction des logements et résidences, le secteur de l'enseignement et l'expansion des transports.

Un membre de l'équipe est affecté à chaque thème proposé de l'enquête, afin de consulter le plus grand nombre possible des principaux utilisateurs des données dans son domaine et de préparer un sommaire de ses discussions pour la prochaine réunion. Après les discussions préliminaires avec les utilisateurs des données, l'équipe du projet se réunit pour étudier les rapports sur chaque consultation. Les membres en ont tellement appris en fait qu'il faudra prévoir trois autres réunions.

Le membre de l'équipe affecté aux caractéristiques sociodémographiques fait rapport en premier et affirme que les principaux utilisateurs sont les experts intéressés avant tout aux événements démographiques élémentaires de la famille, notamment, les naissances, décès et mariages, la composition de la famille et la migration (p. ex., l'immigration, l'émigration et la migration entre régions et milieux urbain et rural). Il est convenu après discussion que le membre de l'équipe essaiera d'obtenir de l'information plus détaillé sur chacun de ces sujets, en particulier sur la disponibilité de renseignements pertinents dans les sources actuelles, par exemple, les registres de l'état civil.

Le deuxième membre informe l'équipe que les principaux sujets pertinents à l'activité du marché du travail comprennent le statut de la population active (employé, sans emploi, hors de la population active), le travail salarié ou le travail autonome, le travail à plein temps ou à temps partiel, la branche d'activité, la profession, le nombre d'heures travaillées, etc. Plusieurs utilisateurs, en particulier les planificateurs des ministères de l'Emploi, de l'Éducation et de la Construction, soutiennent fermement qu'aucune des enquêtes actuelles ne répond suffisamment aux besoins d'information sur le marché du travail dans le contexte des changements rapides actuels dans la structure économique du pays. Ils ont donc besoin, par exemple, d'estimations précises et objectives du nombre de personnes qui travaillent dans des activités en particulier, notamment la construction de bâtiments, dans chaque région ou ville du pays. Les analystes veulent aussi déterminer le nombre de travailleurs qui ont plus d'un emploi, leurs heures réelles de travail et leurs gains. Le membre de l'équipe est chargé de franchir l'étape suivante et d'obtenir de l'information plus détaillée sur les besoins de données et d'essayer d'obtenir les données éventuellement disponibles, par exemple, celles des enquêtes précédentes.

Votre rapport porte sur les revenus et dépenses des ménages. Vous avez appris que la définition de revenu des ménages peut comprendre tous les revenus (bruts ou nets) en espèces ou en nature de tous les membres du ménage au cours d'une période de référence, par exemple, le mois précédent ou l'année dernière. Les dépenses peuvent comprendre les sommes versées pour les aliments, les vêtements, le logement, les transports, la scolarité, les soins de santé, etc., ainsi que les biens ou services échangés directement (troc) pour d'autres biens et services reçus. L'actif et le passif du ménage peuvent aussi être considérés pertinents aux fins analytiques de l'enquête.

Vous avez aussi constaté qu'il y a de nombreux utilisateurs éventuels des données sur ces sujets. Le BSB voudrait obtenir des données plus détaillées sur les revenus et dépenses pour renforcer certaines composantes des comptes nationaux. Les comptes actuels sont moins complets au chapitre des revenus du secteur privé, en particulier pour la main-d'œuvre et les entreprises. Ceux qui s'intéressent à la mesure du commerce de détail, du pouvoir d'achat et de la construction résidentielle prévue veulent en apprendre davantage sur l'évolution des revenus de la famille et les changements des caractéristiques des dépenses

pour renforcer les rapports qu'ils présentent aux décideurs de l'économie. On a aussi constaté que les estimations du revenu disponible intéressent beaucoup le secteur de la fabrication et les administrations du tourisme. Il reste beaucoup de travail à faire pour en arriver au niveau de détails approprié et vous convenez d'entreprendre le deuxième volet de l'enquête dans le domaine. Vous examinerez donc de plus près la disponibilité des données pertinentes actuelles.

Le quatrième membre de l'équipe fait rapport et, selon lui, les plus importants indicateurs des conditions de vie semblent faire référence aux conditions du logement, à l'accès et au recours aux transports, à l'accès aux services d'enseignement, ces volets étant tous très importants pour les planificateurs de l'infrastructure municipale et nationale. Une nouvelle phase de recherche est lancée sur ce sujet.

Le chargé de projet a déjà eu certaines discussions préliminaires sur les thèmes supplémentaires éventuels, et il fait aussi brièvement rapport :

i. Le recours aux services de soins de santé des ménages l'an dernier pourrait comprendre l'achat de produits pharmaceutiques, les visites aux cliniques ou aux médecins locaux, les soins en milieu hospitalier, par exemple, la chirurgie dans les hôpitaux. Ces sujets intéressent particulièrement le ministère de la Santé qui veut mesurer le taux de changement de la demande pour les différents types de services de soins de santé.

ii. L'intérêt pour l'activité agricole du ménage peut comprendre les cultures ou l'élevage du bétail sur une terre louée ou qui appartient au ménage, ou sur une terre communautaire, que la production soit pour la vente au marché ou la consommation personnelle. Les analystes du ministère de l'Agriculture veulent de l'information à jour sur les genres et les quantités de produits agricoles envoyés aux marchés urbains, afin de faciliter la planification et la formulation de politiques.

iii. L'activité des entreprises à domicile (autres que la vente de produits agricoles) peut comprendre la fabrication à petite échelle, les restaurants non officiels, la coiffure et un grand nombre d'autres activités, par exemple, la boulangerie et la fabrication de chaussures. Les analystes de la planification économique sont intéressés à obtenir cette information pour comprendre la structure de l'économie et repérer le taux de mouvement vers les petites entreprises du secteur privé.

Les ministères qui s'intéressent à des thèmes supplémentaires ont obtenu une estimation générale du coût de l'intégration de leurs sujets dans une grande enquête polyvalente et, aux dernières nouvelles, ils semblent hésiter à libérer les ressources nécessaires, au moins pour cette année. Intégrer tant de sujets dans la première édition de l'EGM inquiète aussi l'équipe à cause de l'énorme fardeau de réponse et des répercussions négatives éventuelles sur la qualité des données. L'équipe informe le Comité directeur qui lui conseille de cibler la définition de l'énoncé des objectifs selon les quatre principaux thèmes et de réserver les ajouts éventuels aux occasions d'enquête ultérieures.

L'équipe continuera donc d'approfondir l'énoncé des objectifs pour chacun des quatre principaux sujets (caractéristiques sociodémographiques, activités du marché du travail, revenus et dépenses, conditions de vie), tout en considérant l'utilité des sources de données actuelles.

Compte tenu des discussions prolongées au cours des récentes réunions de l'équipe, vous continuez votre recherche dans la matière qui vous est confiée, c.-à-d. les revenus et dépenses des ménages. Plusieurs utilisateurs et diverses utilisations des données sont déjà identifiés. Les ministères centraux ont aussi besoin davantage d'information à jour sur les revenus des familles, et en particulier sur les dépenses pour les aliments, afin de déterminer si des subventions de l'État sont nécessaires pour protéger les familles à faible revenu. D'autres utilisateurs voudraient des données pour établir des modèles économétriques, afin d'estimer les hausses de demandes de biens de consommation, de denrées alimentaires superflues et de

logements améliorés. Ces estimations donneraient aux secteurs de la fabrication et de la construction de l'information qui les aiderait à planifier de nouveaux projets et à prendre des décisions sur l'embauche de travailleurs supplémentaires.

Vous n'avez cependant pas eu le temps d'aller de l'avant à cette deuxième étape de la définition et déjà, l'équipe est convoquée à une réunion imprévue avec le Comité directeur. L'équipe est informée que le budget prévu pour la première édition (première année) de l'EGM est réduit de beaucoup (plus de 50 %). L'équipe devra cibler les thèmes essentiels les plus importants, tout en continuant de préparer une infrastructure d'enquête qui pourrait immédiatement prendre de l'expansion pendant la deuxième année si les ressources nécessaires sont disponibles.

Le Comité directeur et l'équipe du projet considèrent l'information obtenue à ce jour et tirent les conclusions suivantes :

i. Presque toutes les mesures démographiques demandées, même imparfaites, peuvent être produites avec satisfaction à partir des données actuelles, notamment les statistiques de l'état civil et les registres du logement, directement ou autrement, pour au moins une autre année.

ii. Plusieurs ministères affirment qu'aucune des sources actuelles ne donne de l'information appropriée sur le marché du travail dans le contexte des circonstances économiques changeantes.

iii. Plusieurs membres du Comité directeur soutiennent que l'information sur les revenus et dépenses tirée des enquêtes actuelles peut servir à moyen terme, surtout si elle est renforcée à l'aide de l'EGM par de meilleurs renseignements sur l'évolution du marché du travail.

iv. Au volet de l'information demandée sur les conditions de vie, il faut obtenir de meilleures données sur le marché du travail pour améliorer les projections sur les besoins de logements, de transports et de services d'enseignement, afin de répondre aux principaux besoins immédiats.

Il devient évident que la conception initiale de l'EGM devra être axée sur la composante marché du travail et qu'il faudra reporter à plus tard les autres sujets de la liste initiale.

2.3 Concepts et définitions opérationnelles

La portée de l'EGM est maintenant définie de plus près et la prochaine tâche de l'équipe est de préciser les concepts et les définitions opérationnelles pour les sujets de l'enquête, afin de détailler la matière de l'enquête. La charge de travail est de nouveau répartie entre les membres de l'équipe.

L'équipe commence à définir certains des concepts essentiels à la description de l'activité du marché du travail de Belleterre : population active, employé, sans emploi. Les membres de l'équipe réfléchissent aux concepts, mais formulent davantage de questions que de réponses, par exemple :

i. Population active

Qui doit-elle englober? Le concept de la population active s'applique-t-il également en milieu urbain et rural? Que faire avec ceux qui travaillent ou résident dans une région, mais dont la résidence permanente est ailleurs? À des fins pratiques, faudra-t-il considérer que cette personne fait partie de la population active de son lieu de résidence permanente ou de la région où elle travaille actuellement? Comment classer les personnes qui peuvent travailler, qui ne sont pas encore à la retraite, mais qui ne travaillent pas ou ne cherchent pas de travail?

ii. Employé

La définition de l'emploi comprend-elle seulement le travail rémunéré en argent ou faudrait-il ajouter le travail en échange de biens et services? Comment définir le travail à plein temps et à temps partiel? Une personne peut-elle avoir un emploi à plein temps et un autre à temps partiel, si oui, est-ce le nombre d'heures travaillées ou un autre critère qui détermine la définition? Y a-t-il des définitions convenables d'employeur, d'employé et de travailleur autonome? De nouveau, les définitions de l'emploi devraient-elles être différentes si la personne habite en milieu rural ou urbain? Si une personne travaille ailleurs que dans le secteur de sa résidence permanente, faudrait-il considérer qu'elle y est employée, et faudrait-il établir qu'elle est employée, sans emploi, ou simplement absente de son lieu de résidence permanente?

iii. Sans emploi

Une personne est-elle considérée sans emploi simplement parce qu'elle ne travaille pas? Qu'en est-il si elle ne veut pas travailler? Ou si elle a cherché du travail tellement longtemps qu'elle a abandonné, découragée de ne jamais trouver un emploi?

L'équipe a beaucoup de difficultés à répondre à toutes ces nouvelles questions. Certains membres cherchent des définitions utilisées dans d'autres pays et découvrent les définitions standard suivantes de l'Organisation internationale du travail (OIT) :

i. Population active *:* Une personne est considérée membre de la population active si elle est employée ou sans emploi (voir ci-dessous).

ii. Employé : La personne employée est celle qui, au cours de la période de référence :

a. accomplit n'importe quelle tâche à un poste ou dans une entreprise, c'est-à-dire un travail rémunéré dans le contexte d'une relation employeur-employé, ou qui est travailleur autonome. Cette catégorie comprend aussi le travail non rémunéré dans la famille, c'est-à-dire selon la définition, un travail non rémunéré qui contribue directement au fonctionnement d'une exploitation agricole, d'une entreprise ou d'une pratique professionnelle exploitée par un parent du même ménage et dont il est propriétaire,

ou

b. a un emploi, mais n'est pas au travail pour certaines raisons, notamment une maladie ou une incapacité de sa personne, des responsabilités personnelles ou familiales, les congés annuels, un différend employeur-employés ou pour d'autres raisons (à l'exception de la personne mise à pied, entre deux emplois occasionnels ou qui a un emploi commençant à une date ultérieure).

iii. Sans emploi : La personne sans emploi est celle qui, au cours de la période de référence :

a. est mise à pied temporairement, sauf si elle sera rappelée au travail et si elle est disponible pour travailler,

ou

b. est sans travail, a activement cherché du travail depuis quatre semaines et est disponible pour travailler,

ou

c. a un nouvel emploi qui commencera dans les quatre semaines suivant la période de référence et est disponible pour travailler.

iv. Hors de la population active : La personne hors de la population active est celle qui, au cours de la période de référence, n'est pas disposée à offrir ou fournir des services de main-d'œuvre, ou est incapable de le faire, compte tenu des conditions de son marché du travail, c'est-à-dire qu'elle n'est ni employée ni sans emploi.

L'équipe décide de concevoir le questionnaire de sorte que les données obtenues serviront à estimer les mesures selon les concepts de l'OIT. Il sera donc possible de comparer avec les mesures internationales appliquées à l'activité du marché du travail. L'équipe remarque que les définitions de l'OIT restreignent l'analyse aux personnes âgées de 15 ans et plus. Elle décide d'appliquer cette norme en général.

L'équipe remarque aussi qu'il est difficile de définir certains concepts essentiels, par exemple, le logement, le ménage et la famille. Après une certaine recherche, l'équipe décide d'adopter les définitions standard suivantes :

v. Logement : tout ensemble de pièces d'habitation de structure distincte comprenant une entrée privée à l'extérieur de l'édifice ou qui donne sur un vestibule ou un escalier commun dans l'édifice.

vi. Ménage : toute personne ou groupe de personnes qui habitent dans un logement. Un ménage peut comprendre tout ensemble des éléments suivants : une personne qui vit seule, une famille ou plus, un groupe de personnes sans lien de parenté, mais qui partagent le même logement.

vii. Famille : un groupe de deux personnes ou plus qui habitent dans le même logement et qui sont liées par le sang, le mariage (union libre comprise) ou l'adoption. Une personne qui habite seule ou qui n'a de lien avec aucune autre personne dans le logement où elle habite est classée « hors famille ».

L'équipe constate qu'une enquête ciblant strictement les activités de la population active ne permettra pas aux analystes de tracer un profil très détaillé de la population active à Belleterre. Elle réalise qu'il faudra faire la collecte de données, non seulement sur l'activité, mais aussi sur la démographie, la scolarité, les revenus, etc., aux fins de la classification.

2.4 Matière de l'enquête et plan d'analyse

L'équipe commence à tracer certains tableaux préliminaires de données de sortie pour aider à préciser la matière de l'enquête demandant à chaque étape quelles questions analytiques elle peut aider à répondre. Il semble évident, par exemple, que l'EGM devra permettre de produire certains tableaux, par exemple, «La population active selon le degré de scolarité » et l'« Emploi selon l'âge et le sexe ». Il faut donc faire la collecte de renseignements sur l'emploi et les caractéristiques démographiques du ménage.

Vous tracez un tableau fictif de la première rubrique :

Tableau 2.1 : Situation vis-à-vis de l'activité selon le degré de scolarité

Degré de scolarité	Situation vis-à-vis de l'activité			
	Employé	**Sans emploi**	**Population active**	**Hors de la population active**
Sous la moyenne				
Au-dessus de la moyenne				
Total				

Résultat tout à fait insuffisant. Dans le cas des étiquettes « sous la moyenne » et « au-dessus de la moyenne » de la colonne Degré de scolarité, s'agit-il du degré de scolarité moyen qui sera déterminé à partir des données de l'enquête ou d'un certain concept exogène de la « moyenne du degré de scolarité ». Quelles valeurs seront inscrites dans les cases du tableau? S'agira-t-il d'un calcul (nombre estimé de personnes), de proportions, de pourcentages?

Vous n'avez pas oublié que des questions plus détaillées (dans le fichier des données évidemment) peuvent toujours être regroupées pour totalisation et vous considérez l'autre extrême :

Tableau 2.2 : Situation vis-à-vis de l'activité selon le nombre d'années de scolarité (nombre de personnes)

Années de scolarité	Situation vis-à-vis de l'activité				
	Employé	**Sans emploi**	**Population active**	**Hors de la population active**	**Total**
1					
2					
3					
…					
99					
Total					

La collecte d'autant de détails et l'interprétation sensée seront difficiles. Vous décidez que les degrés de scolarité devraient être regroupés de façon significative pour la totalisation, même si le nombre exact d'années de scolarité est demandé pour permettre l'analyse détaillée de la variable de la scolarité. Dans le tableau ci-dessus, les degrés de scolarité pourraient être regroupés ainsi : études primaires, études secondaires de premier cycle, études secondaires de deuxième cycle, études collégiales techniques et études universitaires (la question pourrait cibler le nombre de plus élevé d'années d'études suivies ou achevées).

Il a été suggéré que des estimations fiables sont nécessaires pour chacune des 11 régions de Belleterre (trois villes et huit districts). Vous oubliez pour l'instant le besoin éventuel de détails plus approfondis et vous considérez des totalisations qui permettront d'afficher distinctement les 11 régions précisées.

Tableau 2.3 : Situation vis-à-vis de l'activité selon la région, pour la population adulte de Belleterre, (données pondérées)

Région	Situation vis-à-vis de l'activité				
	Employé	Sans emploi	Population active	Hors de la population active	Total
Ville A					
Ville B					
Ville C					
District D					
District E					
District F					
District G					
District H					
District I					
District J					
District K					
Total					

Vous ne savez pas vraiment comment présenter ce tableau le plus efficacement, mais cette décision peut être reportée parce qu'elle n'aura pas de répercussions sur les questions à poser.

Vous tracez plusieurs tableaux semblables et essayez dans chaque cas de déterminer les détails appropriés. Vous êtes enfin prêt pour la prochaine réunion de l'équipe où les suggestions de chaque membre feront l'objet d'une discussion et seront comparées. L'équipe a préparé près de 20 tableaux, par exemple :

- situation vis-à-vis de l'activité selon l'âge et le sexe,
- situation vis-à-vis de l'activité selon le degré de scolarité,
- emploi par branche d'activité,
- situation vis-à-vis de l'activité par région,
- nombre moyen d'heures habituelles de travail des employés selon quelques caractéristiques choisies,
- traitements moyens des employés selon quelques caractéristiques choisies.

La liste détaillée des sujets prend maintenant forme et l'équipe commence à rédiger l'énoncé des objectifs qui comprend son interprétation des besoins d'information pour l'enquête, l'identification des principaux utilisateurs connus, les définitions provisoires des principaux concepts et la proposition d'un certain nombre de tableaux pour l'analyse descriptive préliminaire. L'équipe n'a pas l'impression d'être actuellement en position de faire des commentaires sur la précision parce que ni le plan d'échantillonnage ni la fréquence de l'enquête n'ont été déterminés.

L'équipe envoie l'Annexe 2.1 aux membres du Comité directeur pour discussion à la prochaine réunion. (Il est convenu de la présenter avec mention qu'il s'agit là d'une version préliminaire.) Voilà qui donnera du temps aux membres du Comité directeur pour préparer des commentaires détaillés et, simultanément, l'équipe du projet continuera de travailler aux détails de la matière de l'enquête proposée.

Questions de récapitulation :

Donnez des définitions de la population cible et de la population observée.

Expliquez dans vos propres mots pourquoi l'EGM couvrira une matière beaucoup plus restreinte que celle considérée au départ.

Essayez de formuler un énoncé plus explicite des besoins de données et de leurs utilisations pour la Division de la démographie et le ministère de la Planification économique.

Proposez cinq tableaux à y ajouter.

Quelles définitions de population active, employé et sans emploi proposeriez-vous au Comité directeur? Auraient-elles des répercussions sur les sujets proposés?

Annexe 2.1 : Ébauche de l'énoncé des objectifs de l'Enquête générale sur les ménages de Belleterre

Introduction

À la demande et sous la direction du Comité directeur de l'Enquête générale sur les ménages (EGM), l'équipe de projet a préparé l'ébauche suivante de l'énoncé des objectifs pour la première édition de l'EGM qui se déroulera l'an prochain.

L'enquête portera sur les activités du marché du travail (emploi et autres activités génératrices de revenu, recherche d'emploi, heures de travail, traitements, etc.), ainsi que sur certaines caractéristiques sociodémographiques. Il faudra mentionner en contexte que le but original était d'élaborer une enquête ayant une couverture thématique plus large, y compris de nombreux indicateurs liés à la démographie, aux revenus et dépenses des ménages, aux activités des entreprises, et qui comprendrait éventuellement des données supplémentaires sur la santé et l'agriculture. Ce genre d'enquête est toujours l'objectif à moyen terme du BSB. Le projet initial sera cependant moins ambitieux et plus étroitement ciblé.

La recherche préliminaire a révélé que les domaines spécialisés de la démographie et des revenus et dépenses peuvent être couverts correctement à l'aide des sources actuelles (registres de l'état civil et enquêtes sur les ménages en milieu urbain et rural, respectivement) pour au moins une autre année. L'information sur les domaines à priorité élevée des autres sujets d'importance peut au moins être renforcée à l'aide des données obtenues avec cette version de l'EGM. Voilà pourquoi les objectifs énoncés visent surtout les activités du marché du travail.

Principaux utilisateurs des données

Les principaux utilisateurs des données de l'EGM sont la Division de l'analyse de la population active du BSB et les planificateurs économiques des ministères de l'Emploi, de l'Éducation et de la Construction du secteur des ménages. D'autres utilisateurs comprendront la Division de la démographie du BSB, le ministère de la Planification économique et la Commission nationale du travail. Des exemples de communication avec ces utilisateurs sur leurs besoins de données sont joints en Annexe A *(non insérée)*.

Principaux concepts et définitions

Nous aurons recours à des définitions normalisées de certains concepts, notamment le logement, le ménage et la famille, afin de maintenir l'uniformité avec d'autres produits statistiques du BSB.

L'équipe du projet recommande d'adopter les définitions largement utilisées de population active, employé, sans emploi et hors de la population active de l'Organisation internationale du travail (OIT) pour faciliter la comparaison internationale.

Proposition de contenu

Les sujets suivants seront ajoutés à l'enquête :

Logement – ménage
Mode d'occupation (propriété ou location)
Composition du ménage
- Âge
- Sexe
- Scolarité achevée (degré et nombre d'années d'études)

Activité et Population active
Situation vis-à-vis de l'activité (employé, sans emploi, hors de la population active),
Branche d'activité (secteur primaire, fabrication, ventes, services, etc.),
Profession (directeur, superviseur, professionnel, manœuvre, etc.),
Nombre d'heures travaillées,
Revenu d'emploi,
Autre activité économique,
- Emploi autonome,
  - Secteur officiel,
  - Secteur non officiel (c.-à-d. « économie clandestine »).

Plan d'analyse préliminaire

La première analyse comprendra les tableaux des nombres et des pourcentages estimés pour chacun des principaux articles énumérés ci-dessus, ainsi qu'un certain nombre de totalisations croisées. Environ 20 totalisations principales sont proposées en Annexe B *(non insérée)*.

Il faut encore apporter d'autres détails de l'analyse, mais ils comprendront probablement la production de tableaux de répartitions régionales et de branches d'activité plus détaillées.

Chapitre 3 - Introduction au plan d'enquête

3.0 Introduction

Le Comité directeur a approuvé l'ébauche de l'énoncé des objectifs et l'équipe du projet se réunit maintenant pour entreprendre le travail substantiel de conception de l'EGM.

La première question, à savoir s'il faut faire un recensement ou une enquête-échantillon, est déjà résolue. Un recensement serait hors de prix, même un seul par année. De plus, même si l'argent était disponible, l'opération serait si énorme et compliquée, et il y aurait tant de problèmes logistiques et de gestion que les erreurs non dues à l'échantillonnage (comme les erreurs de traitement) surpasseraient les avantages tirés de la prévention de l'erreur d'échantillonnage (que donnerait un plan d'enquête-échantillon approprié).

Seule une enquête-échantillon donnera de l'information annuelle ou infra-annuelle sur une population de près de 44 millions de personnes. Vous avez en fait été informé que le plan d'enquête devrait être conçu pour application quatre fois par année, afin d'obtenir des estimations utiles à chaque trimestre de l'année civile. L'équipe préparera donc un plan d'échantillonnage pratique, compte tenu de ces échéanciers, à l'aide d'un échantillon suffisamment large pour obtenir des résultats fiables à chaque trimestre.

3.1 Populations cible et observée

À première vue, la population cible semble facilement définie et l'équipe considère qu'il s'agit provisoirement de la population adulte de Belleterre. L'équipe du projet identifie cependant plusieurs problèmes :

i. Résidents temporaires

Faut-il inclure dans la population active ceux qui ont emménagé temporairement au pays? Ils ne font pas officiellement partie de l'économie, mais ils ont un emploi qui se traduit par des produits. D'autre part, ceux qui ont déménagé dans un autre pays ont un emploi qui se traduit par des produits dans ce pays. Il peut être impossible de communiquer avec eux et il n'est pas évident qu'ils devraient faire partie de l'activité économique de Belleterre, même si nombre d'entre eux peuvent envoyer une partie de leur revenu au pays.

Où faudrait-il dénombrer ceux qui sont passés d'un milieu rural à un milieu urbain au pays? (Ils habitent habituellement dans des logements temporaires.) Ils ont des répercussions importantes et croissantes sur la dynamique sociale et économique des secteurs urbains au pays au chapitre des augmentations ou diminutions imprévues de la main-d'œuvre disponible dans un secteur urbain comparativement à un autre (et, pourrait-on ajouter, des répercussions sur les secteurs ruraux quant à la diminution de la population active agricole disponible).

L'équipe décide de considérer membres du ménage ceux qui habitent habituellement dans le logement, autrement dit, chacun qui considère que le logement est son lieu habituel de résidence. L'équipe constate qu'il faudra définir beaucoup plus clairement ce concept pour l'appliquer, mais c'est au moins un point de départ.

ii. Logements collectifs

L'enquête devrait-elle couvrir seulement les résidents des logements individuels ou ceux des logements collectifs aussi (c'est-à-dire les logements où habitent plus d'un ménage)? Les logements collectifs peuvent tout englober, à partir des petites maisons de chambres ou des pensions, jusqu'aux hôpitaux et aux prisons. Voilà des situations évidemment très différentes. Dans les grands instituts à logements collectifs, notamment les hôpitaux, les prisons, les bases militaires, etc., même si l'équipe décidait que l'enquête devrait comprendre ces résidents, serait-il réaliste de faire des interviews avec eux? D'autre part, les petits logements collectifs ressemblent de près aux logements privés et l'équipe se demande s'il ne faudrait pas les traiter comme des logements privés aux fins de la collecte des données.

iii. Régions éloignées

L'équipe considère maintenant ceux qui habitent dans certains régions éloignées et inaccessibles, car il coûterait beaucoup trop cher de les intégrer à l'enquête. (Ce groupe est cependant minime à Belleterre, c'est-à-dire moins de 1 % de la population.)

L'équipe établira provisoirement que la population cible est la population adulte dont le lieu habituel de résidence est à Belleterre. Ceux qui habitent dans des régions éloignées seront exclus, ainsi que les résidents des instituts à logements collectifs, notamment les hôpitaux, les prisons, les bases militaires, etc. L'équipe réalise que la base d'échantillonnage utilisée influera sur la période de référence, aidera à déterminer les créneaux de la population qu'il faudrait exclure pour des raisons pratiques (p. ex., région trop éloignée, coût trop élevé, enquête trop difficile à faire), etc., et déterminera la définition de la population cible.

3.2 Base de sondage

Les enquêtes sur les ménages à Belleterre ont habituellement été faites à l'aide des listes des registres des ménages et de la population qui ont servi de base d'échantillonnage. Étant donné l'ampleur de la documentation administrative sur les mouvements de la population, ces listes ont généralement été considérées très complètes et à jour. Les augmentations récentes des taux d'émigration et de migration dans les régions et entre elles signifient cependant que cette considération n'est peut-être plus exacte.

Étant donné que les listes des registres disponibles sont largement maintenues à l'échelon local et dans des bureaux publics de la ville, il pourrait y avoir chevauchement jusqu'à un certain point et il n'est pas évident que les listes sont mises à jour aussi souvent ou précisément partout. Il faudrait intégrer toutes ces listes en une seule grande base de sondage pour éliminer le chevauchement entre les bases. Bien entendu, il faudrait ensuite tenir cette base à jour et, à cette fin, obtenir continuellement de l'information de mise à jour de plusieurs de centaines de bureaux différents, à chaque trimestre au moins. D'autre part, l'équipe remarque que l'établissement de nouvelles listes pour tout le pays coûterait très cher.

Compte tenu de ces éléments, l'équipe convient qu'elle doit étudier d'autres sources possibles d'information pour établir la base de l'EGM.

Certains pays qui ont de bons registres de la population utilisent des bases d'échantillonnage aréolaire pour leurs enquêtes sur les ménages. L'équipe fait des recherches sur la documentation disponible dans d'autres pays pour déterminer si cette approche lui permettrait éventuellement de couvrir une population plus entièrement représentative.

Les membres de l'équipe savent qu'une base aréolaire peut offrir, en théorie, une couverture presque complète, mais la préparer peut aussi coûter cher. Étant donné qu'ils devraient commencer au point de départ, ils ne sont pas certains de pouvoir élaborer à temps une bonne base aréolaire et un plan d'échantillonnage connexe pour la première édition de l'enquête. Au cours de leurs discussions avec des représentants de plusieurs autres équipes d'enquête du BSB, l'équipe de l'EGM a cependant appris qu'ils ont récemment décidé de faire l'expérience des bases aréolaires. Ils ont déjà fait une recherche substantielle sur la question. Étant donné que les résultats de cette recherche préliminaire semblent très prometteurs, l'équipe de l'EGM propose aux autres équipes d'enquête et au Comité directeur de combiner leurs efforts pour produire et appliquer une base aréolaire. Le Comité directeur approuve l'idée parce que les coûts de l'implantation et de la mise à jour de la base seraient partagés avec d'autres enquêtes.

Il est décidé d'appliquer cette approche et l'équipe entreprend donc l'évaluation de l'état actuel des cartes nécessaires et d'autres renseignements cartographiques pour commencer à préparer la base aréolaire et obtenir une estimation raisonnable des ressources nécessaires pour achever cette tâche à temps pour l'enquête.

Les cartes topographiques à grande échelle et à jour généralement en très bonne condition détermineront les niveaux plus élevés de la base aréolaire (c.-à-d. qui définissent les unités d'échantillonnage primaires). Les niveaux inférieurs (c.-à-d. qui définissent les unités d'échantillonnage au deuxième et troisième degré) devront être déterminés à partir de l'identification des rues et, éventuellement, des logements en milieu urbain, et des logements et villages en milieu rural. L'équipe convient que la maintenance de la base exigera un effort permanent, éventuellement par roulement, et il faudra être particulièrement attentif aux secteurs à croissance élevée.

L'équipe doit donc commencer à élaborer une suite d'étapes pour délimiter les unités d'échantillonnage à divers échelons de la hiérarchie géographique, à partir de l'identification des limites naturelles à grande échelle comme les principales rivières, jusqu'aux plans à formuler pour établir la liste inévitable des logements dans les secteurs sélectionnés. L'équipe n'est évidemment pas encore en mesure de passer aux détails de ces étapes parce qu'il faut auparavant prendre des décisions sur le plan d'échantillonnage concret.

Questions de récapitulation :

Pourquoi une enquête-échantillon est-elle la seule solution pratique pour répondre aux besoins de données?

Expliquez pourquoi ceux qui habitent dans les logements collectifs devraient ou non être intégrés à la population cible. Ajoutez des considérations sur l'accès à ces personnes, ainsi que sur la pertinence et la qualité éventuelle de leurs réponses.

Expliquez les autres différences possibles entre la population cible et la population définitive observée.

Si l'approche de la liste était maintenue, quelles auraient été les sources et les répercussions éventuelles de l'erreur non due à l'échantillonnage?

Quelles sont, à votre avis, les trois plus importantes sources probables d'erreurs non dues à l'échantillonnage liées à l'utilisation d'une base aréolaire? Suggérez un moyen ou deux de réduire ou de contrôler chacune d'elle.

Combien de cartes ou de tracés distincts à l'échelle faudrait-il, à votre avis, pour réaliser le plan d'échantillonnage de la base aréolaire?

Chapitre 4 - Méthodes de collecte des données

4.0 Méthode de collecte des données

L'équipe du projet se réunit pour choisir une méthode de collecte des données. Les membres de l'équipe discutent des avantages et inconvénients des trois méthodes élémentaires : l'interview sur place, l'interview téléphonique et l'autodénombrement (p. ex., envoi et retour du questionnaire par la poste).

L'interview téléphonique ne semble pas très pratique parce que le pourcentage de ménages sans téléphone est très élevé, surtout en milieu rural. La population n'est pas habituée à traiter avec les autorités gouvernementales au téléphone et de nombreuses personnes hésiteraient beaucoup à répondre à une enquête du genre. De plus, certains concepts à considérer dans l'enquête sont complexes et les gens pourraient avoir de la difficulté à comprendre les explications au téléphone. Il n'y a pas non plus d'annuaire téléphonique complet et à jour couvrant les ménages abonnés parce que la couverture téléphonique augmente rapidement dans certaines régions. Voilà pourquoi l'équipe décide de ne pas donner suite à l'option téléphonique.

Étant donné que l'approche par base aréolaire dépendra nécessairement de l'information sur l'adresse des logements, l'équipe considère brièvement la possibilité de faire une enquête par la poste. L'équipe a cependant appris que les quelques études de marché faites par la poste ont donné des taux de réponse très faibles et les questionnaires retournés comprenaient de nombreuses réponses incomplètes ou incohérentes. Ces problèmes s'aggravent si les formules d'enquête sont simplement envoyées au « chef du ménage » et non à une personne en particulier. Étant donné que l'EGM comprendra beaucoup plus de questions et que certains sujets sont très complexes, l'envoi par la poste ne semble pas en mesure de fournir des données de qualité convenable.

L'interview sur place pourrait être la seule possibilité, malgré le coût élevé de l'affectation d'un grand nombre d'intervieweurs à l'enquête. L'équipe commence à discuter de la logistique de l'embauche et de la formation d'un nombre suffisant d'intervieweurs, à planifier la préparation des manuels de soutien nécessaires et à étudier les besoins de véhicules ou autre matériel.

Même si le plan d'échantillonnage n'est pas encore très avancé, le méthodologiste d'enquête recommande de répartir les interviews des ménages sélectionnés sur plusieurs trimestres pour obtenir de bonnes estimations du changement de la situation vis-à-vis de l'activité. L'équipe considère la possibilité de faire une première interview sur place et d'utiliser ensuite le téléphone ou l'envoi par la poste pour les autres interviews. Le problème du sous-dénombrement persiste cependant pour l'interview téléphonique et il est à craindre que les taux de réponse des retours par la poste seront plutôt faibles, même après la communication personnelle de la première interview. L'équipe continue néanmoins de considérer l'option de l'envoi et retour par la poste comme stratégie possible d'interviews ultérieures.

Un autre élément incite l'équipe à reconsidérer l'option de l'autodénombrement. Il est à craindre que les répondants jugent certaines questions très personnelles ou à caractère délicat et ils peuvent donc hésiter à répondre à l'intervieweur. Un questionnaire par la poste pourrait obtenir des réponses à ces questions. D'autre part, les répondants devraient mieux répondre à la majorité du questionnaire pendant une interview sur place. Un membre de l'équipe suggère une approche en mode mixte : faire une interview sur place pour la majeure partie du questionnaire, mais demander au répondant d'inscrire dans une page distincte les réponses aux questions à caractère délicat, en privé, et de la déposer dans une boîte scellée à dont disposera l'intervieweur ou de la retourner par la poste. L'équipe décide que la mise à l'essai du questionnaire et de la méthode de collecte des données sera nécessaire pour résoudre la question.

Il reste encore à prendre une importante décision, à savoir s'il faut utiliser un questionnaire sur support-papier (interview papier et crayon ou *PAPI*) ou appliquer une approche informatique (c.-à-d. interview sur place assistée par ordinateur, IPAO). On connaît les principaux avantages de la méthode *PAPI*: il n'est pas nécessaire d'acheter du matériel très cher et cette approche est bien maîtrisée parce que cette méthode a été appliquée à toutes les enquêtes du BSB jusqu'à maintenant. D'autre part, les membres de l'équipe de l'EGM réalisent qu'ils pourraient fournir à faible coût un ordinateur de poche aux intervieweurs et, avec un minimum de programmation, qu'ils pourraient éliminer une étape distincte de saisie des données et garantir qu'une grande partie de la vérification des données est faite au moment de l'interview (lorsque les corrections sont le plus facile à faire).

Les membres de l'équipe décident de calculer le coût éventuel de chacune de ces approches, ils essaient simultanément d'évaluer les améliorations possibles de la qualité des données s'ils choisissent l'IPAO et les résultats sont pondérés par rapport à la complexité supplémentaire de l'élaboration d'une application informatique de l'IPAO. Des renseignements concrets sur la qualité des données seront bien entendus disponibles seulement après avoir fait l'enquête. Certaines approximations sont quand même possibles maintenant pour aider à prendre la décision appropriée.

Après examen de l'information disponible (le coût et la période de préparation, la possibilité de changer le questionnaire après le premier cycle), l'équipe décide de procéder à la première édition de l'EGM à l'aide de *PAPI* et de réexaminer la question ultérieurement.

Questions de récapitulation :

L'équipe aurait-elle dû considérer la possibilité d'appliquer une approche en mode mixte, y compris l'interview téléphonique dans les régions des principales villes où les abonnés au téléphone sont très nombreux? Discutez des avantages et des inconvénients de ce genre d'approche.

Discutez des avantages et des inconvénients des différentes possibilités suggérées pour régler le problème des questions à caractère délicat.

Considérez les mesures de sécurité qui seront nécessaires pour protéger le matériel si l'IPAO est appliquée.

Considérez les mesures nécessaires dans chaque cas pour protéger la sécurité et la confidentialité des données (questionnaires ou fichiers électroniques).

Chapitre 5 - Conception du questionnaire

5.0 Conception du questionnaire

L'équipe du projet de l'EGM est consciente que la conception du questionnaire demandera beaucoup de travail et elle se réunit pour répartir les responsabilités de la rédaction de la première version des diverses sections.

Elle examine d'abord l'énoncé des objectifs parce qu'il contient déjà le noyau de chaque question ou groupe de questions. Il faut maintenant formuler chaque point en une question claire qui aura la meilleure possibilité d'inciter les répondants à répondre correctement.

Le premier groupe de questions portera sur les caractéristiques du ménage et l'information démographique voulue. Un membre de l'équipe (A) est affecté à la préparation de la première version de cette section et il comptera énormément sur le genre de questions posées dans les enquêtes déjà réalisées sur les ménages.

La principale section du questionnaire portera sur les activités dans la population active des membres admissibles du ménage, c'est-à-dire toutes les personnes âgées de 15 ans et plus. La matière de cette partie est étendue et la responsabilité est donc répartie entre plusieurs membres de l'équipe :

i. Un membre (B) se chargera des questions visant à déterminer le situation vis-à-vis de l'activité (employé, sans emploi ou hors de la population active). Il faudra bien entendu poser plus d'une question.

ii. Un autre membre (C) formulera les versions préliminaires d'une série de questions sur le genre d'activités de la personne employée.

iii. Le membre D préparera des questions sur d'autres activités économiques, par exemple, le travail autonome (y compris l'agriculture) et le travail dans le secteur non officiel de l'économie.

iv. Le membre E s'intéressera aux questions secondaires de l'activité dans la population active qui visent les heures travaillées et les gains.

5.1 Structure du questionnaire

Avant que le travail sur le questionnaire ne soit trop avancé, le membre A demande une réunion de l'équipe du projet pour discuter de certaines questions relevées pendant l'élaboration des questions démographiques. Il souligne que l'équipe doit déterminer comment structurer le questionnaire avant de formuler la version préliminaire des questions définitives. Devrait-il y avoir un questionnaire pour chacun dans le ménage? Qui devrait répondre aux questions? Devrait-il y avoir un questionnaire différent pour les logements collectifs?

Un membre de l'équipe suggère qu'il devrait y avoir deux questionnaires différents, un pour le logement et un pour les répondants du logement. Un autre affirme qu'il devrait y en avoir trois : un pour le logement, un autre pour chaque famille et un troisième pour chaque personne.

Ces approches suscitent un certain débat. D'une part, il est jugé important de comprendre la structure familiale dans le logement, cette information étant pertinente pour déterminer combien de personnes

comptent sur le revenu de chaque travailleur. D'autre part, il est vital d'identifier tous ceux qui habitent dans le logement et certains membres de l'équipe craignent que l'identification des familles débouche sur l'omission de personnes qui forment le ménage, mais qui ne sont pas membres d'une famille.

Il est éventuellement décidé d'essayer d'utiliser deux formules : une pour le logement (intitulée Formule F1) et une pour chaque personne âgée de 15 ans et plus (F2).

La Formule F1 permettra d'obtenir de l'information sur le logement, de dresser une liste complète de tous ceux qui habitent dans le logement, de déterminer qu'elle est l'unité familiale de chacun et d'obtenir l'information démographique sur ces personnes. Ces renseignements serviront à décider qui devrait recevoir le questionnaire destiné à la personne. La Formule F2 servira à la collecte de l'information sur l'activité de la personne dans la population active que l'équipe veut obtenir pour l'enquête. Celle-ci décide que toutes les questions de la Formule F1 seront posées à une personne bien informée sur le ménage et celles de la Formule F2 seront posées à la personne ciblée. L'équipe remarque qu'il faut appliquer un identificateur de logement à chaque formule, afin de lier l'information sur le logement à l'information sur la personne après la collecte.

Un membre de l'équipe souligne que, dans un grand ménage (ou dans un logement collectif) la personne qui répond à la Formule F1 pourrait ne pas connaître les renseignements démographiques de chacun dans le ménage. L'équipe discute de cette question et décide qu'une personne de chaque unité familiale sera consultée pour compléter l'information si celui qui répond à la formule n'a pas les renseignements démographiques pour tout le ménage.

Les membres de l'équipe discutent pour déterminer si les réponses par procuration devraient être permises dans le questionnaire destiné à la personne et ils décident que les intervieweurs devraient d'abord tenter de communiquer avec chacun qui doit remplir une Formule F2, mais s'ils n'y arrivent pas, ils peuvent procéder à l'interview d'un substitut qui répondra au nom de la personne choisie.

L'équipe considère ensuite la création de questionnaires distincts pour les logements collectifs et les logements privés. Il est déjà décidé que l'enquête ne couvrira pas les grands logements collectifs (établissements militaires, hôpitaux, prisons, etc.). L'équipe a donc l'impression que le questionnaire pour les logements privés englobera correctement les plus petits logements collectifs de l'itinéraire des intervieweurs. De plus, si cette information est nécessaire, une variable peut être tirée des logements privés-collectifs après la collecte, compte tenu du nombre de familles ou de personnes sans liens de parenté qui habitent dans le logement.

5.2 Ébauche du questionnaire

Chaque membre de l'équipe continue la consultation et la recherche entreprises pendant la préparation de l'énoncé des objectifs, communique avec les principaux utilisateurs des données et consulte des questionnaires existant. L'équipe prévoit préparer une ébauche raisonnablement complète de tout le questionnaire avant d'entreprendre une mise à l'essai approfondie ou une recherche cognitive sur n'importe quelle section. Cette mesure est nécessaire parce que l'ordre des questions ne sera pas évident avant que la majeure partie des questions soient inscrites au moins sous forme préliminaire. Les membres ont néanmoins l'impression qu'une mise à l'essai de certaines questions pourrait être nécessaire plus tôt s'il devient difficile de décider comment les formuler.

L'équipe se réunit bientôt pour commencer l'examen de l'ébauche des questions de la Formule F2.

Les membres B et D de l'équipe, confiants que leurs questions seraient très faciles, se sont portés volontaires pour entreprendre plusieurs tâches liées à la base et au plan d'échantillonnage qui demandent beaucoup de temps. Leurs ébauches sont donc très incomplètes. La discussion sur leurs sections est reportée jusqu'à la prochaine réunion.

L'équipe considère les questions de C sur le genre d'activités des gens dans la population active :

Si vous êtes identifié employé

C1 Votre employeur est-il (veuillez cocher une seule option)?

- ❍ *une société d'État*
- ❍ *un établissement public (p. ex., hôpital, école, etc.)*
- ❍ *une entreprise privée non familiale*
- ❍ *un membre de la famille*
- ❍ *autre (veuillez préciser)*_____

C2 Quel est le titre de votre poste (p. ex., balayeur, ingénieur, gérant des ventes, conducteur)?

_____

C3 Dans quel secteur se déroule la principale activité économique de votre employeur?

- ❍ *Agriculture, pêches et foresterie*
- ❍ *Industries de l'extraction*
- ❍ *Fabrication*
- ❍ *Transports*
- ❍ *Construction*
- ❍ *Vente au détail*
- ❍ *Vente en gros*
- ❍ *Tourisme*
- ❍ *Secteur des arts et de la culture*
- ❍ *Autres services*
- ❍ *Administration publique (y compris la sécurité)*

C est évidemment un expert de la classification des professions et des branches d'activité. Malheureusement pour l'ébauche du questionnaire, les autres membres de l'équipe ne le sont pas et ils posent rapidement des questions. Presque tous s'interrogent sur la question C3 ou s'y opposent, ils affirment que les répondants ne comprendront pas le terme « secteur » ou l'expression « principale activité économique », mais ils conviennent que la question « Que fait votre employeur? » est trop vague. L'équipe ajoute que les activités énumérées sont la pierre angulaire d'un système de classification que seuls quelques spécialistes connaissent bien après des années de travail. Qu'est-ce que l'industrie de l'extraction? L'expression peut évoquer un dentiste pour certains, alors qu'il s'agit en fait de l'exploration et de l'exploitation minières et pétrolières. De même, un journaliste peut déclarer que son employeur fait partie de la branche des arts et de la culture, mais l'opérateur de presses à imprimer qui travaille pour la même entreprise répondra probablement « fabrication ». Plusieurs membres de l'équipe soutiennent qu'il faudrait poser une question ouverte et la coder au bureau. C n'est pas contre l'idée, mais il ne sait comment formuler la question pour donner suffisamment de détails aux codeurs pour qu'ils fassent leur travail précisément. Le débat sur la formulation de la question ou des questions continue pendant des heures.

L'ébauche des questions du membre E de l'équipe sur les heures de travail et les traitements fait ensuite l'objet d'un examen.

E1 Avez-vous un travail

- ❍ *à plein temps (35 heures par semaine ou plus)*
- ❍ *à temps partiel (moins de 35 heures par semaine)*
- ❍ *saisonnier*

E2 Combien êtes-vous rémunéré?

_____

E3 L'employeur offre-t-il les avantages suivants?

a. Assurance-santé ou soins de santé	❍ *Oui*	❍ *Non*
b. Subvention du loyer	❍ *Oui*	❍ *Non*
c. Régime de retraite	❍ *Oui*	❍ *Non*

La question E2 soulève plusieurs objections. La question « Combien êtes-vous rémunéré? » est non seulement vague, mais elle semble aussi indiscrète, et de nombreuses personnes hésiteront à y répondre, même si elle est bien formulée. Les membres de l'équipe mentionnent des exemples constatés auparavant et soutiennent qu'il faut préciser davantage. La question devrait porter sur le revenu total de l'emploi d'une personne, en argent, avant retenues à la source pour participation à des programmes à frais partagés (p. ex., régimes de retraite à cotisation partagée). Comment englober le tout en une seule question cependant? Il est évident qu'une mise à l'essai est nécessaire pour cette question.

Un membre de l'équipe a l'impression que le revenu est une question à caractère trop délicat et qu'il faut ajouter une incitation, par exemple, « Nous posons la question suivante à tous les répondants de cette enquête pour mieux comprendre la situation de l'emploi au pays.» Il suggère aussi de déplacer la question à la fin du questionnaire.

Un autre membre est d'avis que la formulation de la question E1 est trop vague, mais il ne peut suggérer d'amélioration. Un autre encore affirme que la catégorie « travail saisonnier » ne fait pas partie de la distinction entre l'emploi à plein temps et à temps partiel, et qu'il faudrait poser une question distincte.

5.3 Examen à l'interne

La prochaine réunion avec le Comité directeur est maintenant annoncée et les discussions tenues jusqu'à maintenant y sont présentées. Le Comité directeur convient de l'approche générale visant à utiliser les Formules F1 et F2. Il fait aussi des commentaires détaillés sur les questions qui sont conformes à la rétroaction précédente de l'équipe du projet.

Compte tenu des commentaires de l'équipe du projet et du Comité directeur, l'équipe révise le questionnaire comme suit. (Nota : seules les sections des membres C et E feront l'objet d'un suivi par l'intermédiaire d'un processus d'examen.)

Si vous êtes identifié employé

C1 Quel est le titre de votre poste (p. ex., balayeur, ingénieur, gérant des ventes, conducteur)?

_____

C2 Quel est le genre d'entreprise, de branche d'activité ou de service?

_____

E1 Pendant combien d'heures par semaine travaillez-vous habituellement?

_____

Je vais maintenant poser quelques brèves questions sur vos revenus.

E2 Quel est votre taux horaire de rémunération (avant impôts et autres retenues à la source)?

_____

E3 L'employeur offre-t-il les avantages suivants?

a. Assurance-santé ou soins de santé	❍ *Oui*	❍ *Non*
b. Subvention du loyer	❍ *Oui*	❍ *Non*
c. Régime de retraite	❍ *Oui*	❍ *Non*

5.4 Mise à l'essai du questionnaire

L'équipe décide d'animer un groupe de discussion pour la mise à l'essai du questionnaire actuel, afin de déterminer si les répondants le comprennent facilement et s'ils peuvent donner les réponses exactes.

Le BSB n'engage pas d'animateurs qualifiés de groupes de discussion et l'équipe du projet décide de retenir les services d'un expert de la mise à l'essai en groupe de discussion, M. F. Il a une formation spécialisée en animation de groupes de discussion et il devrait pouvoir intégrer la mise à l'essai à son horaire.

Plusieurs intervieweurs chevronnés sont détachés des bureaux régionaux pour administrer le questionnaire aux répondants du groupe de discussion. L'équipe du projet sélectionne des dizaines de répondants « typiques » et il y aura plusieurs séances en groupe de discussion pendant une semaine. L'équipe sélectionne ceux qui représenteront l'éventail complet des répondants, certains en milieu urbain, d'autres en milieu rural, ainsi que des travailleurs, étudiants et retraités.

Les intervieweurs administrent le questionnaire aux répondants avant le début des discussions en groupe. M. F oriente ensuite des discussions distinctes avec les intervieweurs et les répondants pour leur poser des questions sur le questionnaire et les problèmes qu'ils ont rencontrés.

Compte tenu des discussions en groupe, l'équipe découvre que les questions sur les heures de travail posent les problème suivants :

i. Plusieurs répondants déclarent qu'ils ont de la difficulté à répondre à la question sur le nombre d'heures de travail hebdomadaires habituelles parce que les heures varient d'une semaine à l'autre.

ii. De nombreux répondants demandent s'il faut ajouter le temps supplémentaire aux heures habituelles.

iii. Les questions sèment la confusion chez les répondants qui ont plus d'un emploi.

Les questions sur les revenus posent les problèmes suivants :

i. De nombreux répondants ne sont pas rémunérés à taux horaire, ils affirment donc souvent ne pas connaître leur taux horaire et ils donnent plutôt leur traitement pour une période de référence différente. Les intervieweurs inscrivent habituellement une note en ce sens en marge du questionnaire.

ii. De nombreux répondants ne savent pas si l'employeur offre des avantages sociaux. Cette question n'est pas très sensée non plus pour les personnes qui ont un travail autonome.

L'équipe est cependant très heureuse de constater que les questions sur la profession ou la branche d'activité ne semble pas poser de difficulté aux répondants ou aux intervieweurs. C décide cependant, par mesure de précaution, d'envoyer les réponses aux codeurs formés au codage des professions et des branches d'activité. Il constate après plusieurs heures que les réponses sont trop générales, qu'elles ne donnent pas suffisamment d'information et que les codeurs ne peuvent les coder en détail. Il faudra ajouter plusieurs questions sur la profession et la branche d'activité pour donner suffisamment d'information aux codeurs, afin qu'ils prennent une décision sur le code exact à attribuer.

L'équipe a, de toute évidence, encore beaucoup de travail à faire sur le questionnaire.

5.5 Ébauche définitive

Après intégration au questionnaire des commentaires découlant de la mise à l'essai en groupe de discussion, voici l'ébauche définitive des sections de C et de E (nota : ces questions seront posées seulement aux répondants employés) :

DESCRIPTION DE FONCTIONS

Les questions suivantes concernent votre emploi principal (c'est-à-dire que vous accomplissez la majeure partie de vos heures de travail à ce poste).

JD1 *Êtes-vous employé ou avez-vous un travail autonome?*
- ❍ *Employé (passez à JD5)*
- ❍ *Travail autonome*

JD2 *Avez-vous une entreprise constituée en personne morale?*
- ❍ *Oui*
- ❍ *Non*

JD3 *Avez-vous des employés?*
- ❍ *Oui*
- ❍ *Non*

JD4 *Quelle est la raison sociale de votre entreprise? (Passez à JD6)*

_____

JD5 *Pour qui travaillez-vous?*

_____

JD6 *De quel genre d'entreprise, branche d'activité ou service s'agit-il? (p. ex.., voirie, école primaire, riziculture, magasin de chaussures, garage)*

_____

JD7 *Quel est votre travail ou profession? (p. ex., secrétaire juridique, plombier, guide de pêche, enseignant)*

_____

JD8 *Quelles sont vos principales activités ou tâches à ce poste? (p. ex., préparation de documents juridiques, installation de plomberie résidentielle, orientation de groupes de pêche, enseignement des mathématiques)*

_____

HEURES DE TRAVAIL (EMPLOI PRINCIPAL)

Les questions suivantes portent sur les heures de travail à votre emploi principal (le poste où vous travaillez le plus grand nombre d'heures).

WH1 *Le nombre d'heures de travail varie-t-il d'une semaine à l'autre?à*
- ❍ *Oui*
- ❍ *Non (passer à WH3)*

WH2 *Combien d'heures par semaine travaillez-vous habituellement en moyenne? (Passez à la section suivante)*

_____

WH3 *Pendant combien d'heures avez-vous travaillé la semaine dernière?*

_____

WH4 *Pendant combien d'heures rémunérées en temps supplémentaire avez-vous travaillé à ce poste la semaine dernière?*

_____

WH5 *Pendant combien d'heures supplémentaires non rémunérés avez-vous travaillé à ce poste la semaine dernière?*

_____

TRAITEMENT

Je vais maintenant poser quelques brèves questions sur votre traitement.

E1 *Êtes-vous rémunéré à taux horaire?*
- ❍ *Oui (passez à E2)*
- ❍ *Non (passez à E3)*

E2 *Quel est votre taux horaire? (Passez à la section suivante)*

_____

E3 *Comment pouvez-vous le plus facilement exprimer votre traitement ou rémunération, avant impôt et autres retenues à la source? Est-ce par année, par mois, par semaine ou autrement? Quel est votre traitement ou rémunération?*

_____*/par année*
OU _____*/par mois*
OU _____*/par semaine*
OU _____*/autre (précisez la période de référence _____)*

Questions de récapitulation :

Étant donné qu'il s'agit de la première version du questionnaire de l'EGM destiné à une enquête complète, quel genre de mises à l'essai proposeriez-vous?

Proposez d'autres versions des questions présentées à la Section 5.5.

Chapitre 6 - Plans d'échantillonnage

Nota au lecteur : Les chapitres 6, 7 et 8 couvrent respectivement le choix du plan d'échantillonnage, de la méthode d'estimation et le calcul de la taille et la répartition de l'échantillon. Ces composantes d'un plan d'enquête sont étroitement liées entre elles et leur élaboration est en fait une seule opération très complexe. Les trois sujets sont considérés distinctement dans ce document pour respecter l'ordre de présentation de la matière dans les chapitres correspondants.

6.0 Plan d'échantillonnage

Il n'y a pas de liste à jour de la population de Belleterre, il serait exorbitant de dresser cette liste et une base aréolaire est donc choisie. Avec une base aréolaire, l'échantillonnage des logements est un préalable à la sélection des membres des ménages (c.-à-d. que l'unité d'échantillonnage ultime est le ménage et que l'interview se déroulera avec un membre du ménage qui fera rapport pour chacun dans le ménage).

La formule exacte de sélection des logements n'est pas immédiatement évidente, mais afin d'éviter le listage de tous les logements, il semble logique d'utiliser un plan d'échantillonnage par grappes à deux ou trois degrés (et d'énumérer seulement les logements dans les secteurs échantillonnés au dernier degré). L'équipe réalise que l'échantillonnage à trois degrés, bien qu'économique, peut être très compliqué en pratique, et elle fait donc tous les efforts pour identifier les unités convenables qui permettront un plan à deux degrés.

L'équipe convient d'appliquer la stratification géographique à plusieurs niveaux de détail dans le plan d'échantillonnage. Belleterre est naturellement réparti en 11 régions, c'est-à-dire trois principales villes et huit districts supplémentaires. Les districts sont répartis en plusieurs autres villes et grandes municipalités considérées milieux urbains et en un bon nombre de villages et régions périphériques qualifiés de milieu rural (voir la description au Chapitre 1 de l'étude de cas).

Étant donné les changements récents dans la dynamique de la population active et de la population du pays, et les différences entre les principales villes et le reste du pays, le Comité directeur a fait savoir qu'il veut obtenir le même degré de précision dans l'EGM (variance d'échantillonnage) pour chaque grande ville et chaque district. Compte tenu de cette exigence, il plus efficient de traiter les trois villes et les huit districts comme domaines planifiés, c'est-à-dire de faire un plan d'échantillonnage par strates, et de prévoir un échantillon suffisant dans chacun d'eux. Le résultat est en fait 11 strates de premier niveau et pour lesquelles il faut obtenir la même précision.

L'équipe veut aussi avoir un échantillon le plus représentatif possible pour les secteurs raisonnablement larges dans chaque ville et district, afin de stratifier davantage. Ces strates secondaires auront des populations de taille à peu près semblables et elles seront établies selon les limites de la municipalité ou du comté.

Des raisons administratives pratiques motivent aussi cette stratification géographique parce que les strates proposées correspondent en majeure partie aux différents échelons des unités administratives, notamment, les villes, districts et comtés.

Il y a d'autres variables de stratification souhaitables, tant démographiques qu'économiques, mais il ne semble pas réaliste de les utiliser dans ce plan d'échantillonnage parce qu'elles ne seront pas disponibles dans la base aréolaire avant la sélection de l'échantillon. La possibilité de la stratification *a posteriori* à l'étape de l'estimation sera cependant considérée plus tard et les intervenants examineront la situation de près pour garantir la collecte des variables voulues dans le questionnaire de l'enquête.

Un plan d'échantillonnage par grappes stratifié à deux degrés est donc proposé, les trois principales villes et les huit districts sont chacun une strate et des sous-strates seront créées dans chacune. L'identification de sous-strates relativement petites facilitera la conformité au plan d'échantillonnage à deux degrés.

Afin d'éviter la confusion entre les différents niveaux, il est convenu d'intituler les principales strates « région » (Villes A, B, C et les huit districts) et de réserver le terme « strate » aux sous-strates inférieures. Les 11 régions sont identifiées comme suit pour faciliter la référence :

Tableau 6.1 : Strates régionales

Région	
1	Ville A
2	Ville B
3	Ville C
4	District D
5	District E
6	District F
7	District G
8	District H
9	District J
10	District K
11	District L

Un nombre d'unités primaires d'échantillonnage (UPÉ), ou grappes, sera défini dans chaque strate et, au premier degré de l'échantillonnage, une UPÉ ou plus sera(ont) sélectionnée(s) dans la strate. Les UPÉ ne devraient pas être trop grandes par souci d'efficience. Il serait en fait souhaitable qu'elles soient d'une taille convenable pour que l'équipe des intervieweurs les couvre efficacement, compte tenu du temps prévu pour la collecte des données. Les UPÉ devraient donc être de tailles à peu près égales et contenir plusieurs centaines de logements.

Au deuxième degré de l'échantillonnage, tous les logements de l'UPÉ seront listés et échantillonnés (un membre du ménage sera interviewé). Le nombre de logements échantillonnés par UPÉ devrait être raisonnable pour une équipe d'interview. À la suite de discussions avec les bureaux régionaux, la taille a été déterminée à 40 logements.

L'équipe sait qu'il est souhaitable de sélectionner au moins deux grappes par strate pour permettre l'estimation exacte de la variance d'échantillonnage et que, selon les ouvrages classiques, cette estimation devient plus complexe si l'on sélectionne plus de deux grappes à l'aide de l'échantillonnage avec probabilité proportionnelle à la taille (PPT), ce qu'elle considère nécessaire. Elle prend donc des dispositions pour sélectionner deux grappes dans chaque strate.

Il faut ensuite lister les logements dans les grappes sélectionnées et tirer un échantillon de logements au deuxième degré. Lorsque les listes sont compilées, les logements peuvent être sélectionnés à l'aide de l'échantillonnage aléatoire simple (EAS) ou de l'échantillonnage aléatoire systématique (SYS). L'équipe décide de recommander le SYS pour des raisons pratiques, par exemple, essayer de répartir l'échantillon le mieux possible entre toutes les grappes sélectionnées.

Les détails des taux de sondage ne sont pas encore déterminés, mais il semble déjà évident qu'ils seront raisonnablement faibles dans les grappes sélectionnées (peut-être 40 sur 400 logements, ou un sur dix) et le SYS est donc très pratique. Simultanément, même si cette mesure répartit l'échantillon dans un certain secteur, le territoire couvert par une seule grappe ne devrait pas être très vaste en général et les coûts de

déplacement à l'intérieur d'une grappe sélectionnée ne seront donc pas très importants comparativement au coût qu'il faudrait engager pour y arriver en premier lieu.

Après un certain travail préliminaire, il devient évident que la taille des grappes variera probablement beaucoup s'il faut respecter les limites naturelles. Ces limites naturelles sont cependant très importantes pour le contrôle efficace des opérations sur place et la sélection des grappes et des logements appropriés selon les cartes et les listes.

Il est éventuellement décidé de sélectionner les grappes à l'aide de l'échantillonnage avec probabilité proportionnelle à la taille (PPT) en utilisant les estimations de la population les plus récentes pour chaque grappe comme mesure de la taille. Dans chaque grappe sélectionnée, un nombre déterminé de logements sera choisi à l'aide du SYS. (On verra au Chapitre 7 que cette mesure signifie que toutes les unités de la même strate ont la même probabilité de sélection.)

Ayant établi la structure générale du plan d'échantillonnage, l'équipe commence à tracer les détails du plan pour la région 1 (Ville A) et la région 4 (district D), prévoyant appliquer la même approche aux autres régions.

Questions de récapitulation

Y a-t-il une solution de rechange réaliste à la stratification à deux niveaux (région et strate) qu'a élaborée l'équipe de l'EGM?

Un plan d'échantillonnage à trois degrés, y compris des UPÉ plus larges et un degré intermédiaire d'unités secondaires d'échantillonnage (USÉ) comme grappes, aurait-il été plus efficient? Aurait-il été réaliste en pratique?

Le plan d'échantillonnage appliquera la méthode d'échantillonnage avec PPT pour la sélection des grappes. Lequel serait le plus approprié : l'échantillonnage aléatoire ou systématique avec PPT? Si vous choisissez l'échantillonnage systématique avec PPT, comment suggérez-vous de trier les listes des grappes?

Étant donné qu'il est considéré d'utiliser le SYS pour la sélection des logements dans les grappes sélectionnées, faites des commentaires sur les inconvénients éventuels de cette approche en milieu vraiment rural. Suggérez une stratégie de rechange dans ces secteurs.

Le plan d'échantillonnage est autopondéré dans chaque strate. Est-il raisonnablement possible que l'échantillon soit autopondéré à l'échelon régional, c.-à-d. de garantir que tous les logements sélectionnés dans une région complète aient la même pondération du plan d'échantillonnage? Est-ce un objectif souhaitable?

Chapitre 7 – Estimation

7.0 Estimation

L'équipe de l'EGM cible maintenant les procédures nécessaires pour obtenir des estimations d'enquête sur les caractéristiques obtenues à l'aide du questionnaire.

L'échantillon comprendra de plusieurs milliers de ménages (voir le Chapitre 8 de l'étude de cas pour les détails), mais l'information tirée des interviews de ces ménages devraient représenter la population entière de près de 11 millions de ménages au pays.

Compte tenu du plan d'échantillonnage probabiliste élaboré au Chapitre 6 de l'étude de cas, de la taille de l'échantillon et de la répartition décrite au Chapitre 8, les résultats de l'échantillon donneront en fait des estimations représentatives de la population du pays, et ils auront la variance d'échantillonnage minime voulue, dans la mesure où les non-réponses ne posent pas de problèmes graves.

7.1 Pondération

L'EGM doit donner des estimations pour un grand nombre de variables, mais toutes ces estimations seront basées sur un ensemble de pondérations liées à chaque enregistrement de données des ménages qui répondent à l'enquête, et elles seront déterminées selon le résultat de trois calculs assez simples :

- la pondération du plan d'échantillonnage déterminée selon la probabilité de sélection des ménages,
- un élément d'ajustement pour les non-réponses,
- un ajustement de la stratification *a posteriori* selon les données démographiques comparatives.

7.1.1 Pondération du plan d'échantillonnage

La première étape est de calculer la pondération du plan d'échantillonnage obtenue directement, comme l'expression le suggère, du plan d'échantillonnage, c'est-à-dire que la pondération du plan d'échantillonnage est l'inverse de la probabilité de sélection. Le plan d'échantillonnage dans ce cas est un plan stratifié à deux degrés dont les UPÉ sont sélectionnées à l'aide de la PPT au premier degré et les USÉ (logements) sont sélectionnés à l'aide de l'échantillonnage systématique au deuxième degré.

La pondération du plan d'échantillonnage est calculée pour le logement sélectionné. La même pondération du plan d'échantillonnage sera attribuée au ménage qui habite dans ce logement et à chaque personne du ménage. Dans un plan d'échantillonnage probabiliste à plusieurs degrés, la probabilité de sélection d'une unité au dernier degré est simplement le résultat des probabilités de sélection à chaque degré. De même, la pondération du plan d'échantillonnage peut-être considérée comme le résultat des pondérations à chaque degré parce que chacune d'elle est l'inverse de la probabilité correspondante.

La pondération complète du plan de l'EGM est donc simplement le résultat des pondérations au premier et au deuxième degré. Chaque logement i sélectionné dans la grappe j de la strate h a donc une pondération du plan d'échantillonnage équivalente à :

$$w_{d,hji} = \frac{1}{\pi_{1,hj}} \times \frac{1}{\pi_{2,hji}}$$

où $\pi_{1,hj}$ est la probabilité de sélection de la grappe *j* dans la strate *h* au premier degré et $\pi_{2,hji}$ est la probabilité de sélection du logement *i* dans la grappe *j* de la strate *h* au deuxième degré (si la grappe *j* est sélectionné au premier degré).

N'oubliez pas que le numérotage de la strate a deux niveaux. Le premier numéro est l'identification de la ville ou du district (strate du premier niveau) et le deuxième est le numéro d'identification de la strate dans cette ville ou ce district. Cette particularité est indiquée pour des raisons pratiques par un seul indice *h*.

Étant donné que les grappes sont sélectionnées avec une probabilité proportionnelle à la taille (PPT), ces probabilités doivent être calculées en mesures de la taille utilisées à la conception du plan d'enquête. La mesure de la taille utilisée est le nombre de logements dans la grappe, cette mesure sera représentée par *z* et la grappe *j* de la strate *h* a une taille z_{hj}. Voici la mesure de la taille totale de toute strate *h* :

$$z_h = \sum_j^{m_h} z_{hj}$$

où m_h est le nombre de grappes dans la strate *h*.

Si k_h grappes sont sélectionnées dans la strate *h*, la probabilité de sélection de la grappe *j* est $k_h z_{hj}/z_h$, et la pondération du plan d'échantillonnage au premier degré pour cette grappe est donc :

$$\begin{aligned} w_{d1,hj} &= \frac{1}{\pi_{1,hj}} \\ &= \frac{z_h}{k_h \times z_{hj}} \end{aligned}$$

Au deuxième degré, 40 logements sont sélectionnés, c'est-à-dire que tous les logements de la grappe *j* ont une probabilité de sélection au deuxième degré de $40/z_{hj}$. La pondération du plan au deuxième degré pour les logements sélectionnés est donc :

$$\begin{aligned} w_{d2,hji} &= \frac{1}{\pi_{2,hji}} \\ &= \frac{z_{hj}}{40} \end{aligned}$$

La mesure de taille z_{hj} se rapporte à la taille de la grappe, le nombre de logements, au moment de la création de la base. La taille observée au listage peut être différente. La quantité $z_{hj}/40$ est le pas de sondage utilisé pour le tirage systématique au dernier degré du plan. L'utilisation de ce pas fixe à la taille réelle de la grappe donnera un échantillon de taille aléatoire qu'on espère proche de 40.

L'équipe de l'EGM convient que les grappes devront être sélectionnées à l'aide des tailles estimées parce qu'il est impossible de dénombrer tout le pays. Les membres de l'équipe devront ensuite garder ces mesures de la taille pour le calcul de la pondération au deuxième degré. Cela signifie en fait que les estimations postcensitaires de la population par strate servent d'ancrage, ce qui est logique si les mêmes chiffres sont utilisés pour la stratification à posteriori explicite (voir ci-dessous).

Nous obtenons donc :

$$w_{d,hji} = \frac{1}{\pi_{1,hj}} \times \frac{1}{\pi_{2,hji}}$$
$$= \frac{z_h}{k_h \times z_{hj}} \times \frac{z_{hj}}{40}$$
$$= \frac{z_h}{k_h \times 40}$$

et l'expression peut être représentée simplement par w_{dh}.

N'oubliez pas que tous les logements sélectionnés dans les grappes sélectionnées de la strate *h* ont la même pondération du plan d'échantillonnage et que le plan est autopondéré à l'échelon de la strate.

7.1.2 Pondération ajustée pour les non-réponses

La composante suivante de la pondération est l'ajustement pour le total des non-réponses. Même si un taux de réponse élevé est prévu, il ne sera certainement pas 100 % et l'omission d'un ajustement pour le nombre réel de non-réponses donnerait des sous-estimations des totaux.

Un groupe de non-répondants pourrait aussi, bien entendu, être différent des répondants du point de vue de certaines des variables importantes de l'enquête et, dans ce cas, ajuster les pondérations des répondants pour représenter les non-répondants pourrait donner un biais.

L'équipe de l'EGM étudie ce risque attentivement. Le méthodologiste de l'équipe fait remarquer que le biais de non-réponse dans toute estimation est essentiellement un résultat de deux éléments : le nombre de non-réponses et l'ampleur de la différence entre les répondants et les non-répondants.

L'équipe prévoit un taux de non-réponse raisonnablement faible et elle espère qu'une composante de ce produit sera suffisamment petite. Les membres de l'équipe se demandent si les non-répondants pourraient vraiment être très différents des répondants en ce qui a trait à la majorité des variables importantes de l'enquête. Il semble très probable que les non-répondants pourraient en fait avoir des caractéristiques très différentes de celles des répondants. L'équipe craint, par exemple, que les personnes mises à pied, les migrants récents et ceux qui sont très mobiles pourraient en fait être beaucoup moins nombreux à répondre que les personnes qui ont un emploi régulier et stable dans leur secteur d'enregistrement permanent. Voilà qui se traduirait par une contribution marquée au biais de non-réponse.

L'équipe n'a cependant pas de preuve tangible et décide qu'elle doit pour l'instant prévoir un ajustement des pondérations comme si les non-réponses étaient aléatoires. Les membres conviennent simultanément de recommander des études de suivi d'un sous-échantillon de non-répondants pour essayer d'obtenir une mesure de la taille du biais qu'il ne faudrait pas ignorer simplement, à leur avis. Ils considèrent aussi qu'une autre étude devrait être prévue pour examiner les non-réponses partielles (lorsque certaines questions seulement obtiennent une réponse), afin de déterminer s'il est possible de tracer ainsi un profil des répondants réticents.

Il y a plusieurs méthodes possibles d'ajustement de la pondération pour les non-réponses, selon le plan d'échantillonnage utilisé. Il peut être fait à l'échelon de la grappe, de la strate ou de la ville – du district. Il semble peu raisonnable d'apporter un seul ajustement pour tout le pays parce qu'on sait que la population n'est pas homogène et, en fait, même l'ajustement à l'échelon de chaque district ou principale ville

semble peu conseillé pour la même raison. Le choix peut être important parce que les strates sont habituellement des comtés et les grappes sont très petites. Les grappes peuvent être plus homogènes que la strate complète, mais elles sont si petites que les ajustements pourraient être peu stables à cause des tailles d'échantillon minimes et des taux de réponse qui peuvent être très variables à cet échelon.

L'échantillon de la strate englobe plusieurs fois celui de chaque grappe, la pondération du plan d'échantillonnage est la même pour les deux grappes dans chaque cas et l'équipe est donc d'avis qu'il est plus logique d'appliquer l'ajustement à l'échelon de la strate.

Si le nombre de ménages répondants dans la strate est n_{rh}, la pondération ajustée pour les non-réponses sera équivalente à :

$$\begin{aligned} w_{nr,h} &= w_{d,h}\frac{n_h}{n_{rh}} \\ &= \frac{z_h}{k_h \times 40} \times \frac{k_h \times 40}{n_{rh}} \\ &= \frac{z_h}{n_{rh}} \end{aligned}$$

et toutes les personnes et tous les ménages répondants dans la strate ont de nouveau la même pondération.

Soulignons que cette affirmation sera strictement vraie seulement si toutes les personnes admissibles dans un ménage répondant sont également des répondants. L'équipe de l'EGM considère que l'hypothèse sera vraie aux fins de la planification, mais elle est disposée à proposer un autre ajustement des pondérations des personnes si elles ne répondent pas toutes.

L'équipe remarque que les valeurs numériques des pondérations du plan varieront considérablement d'une strate à l'autre parce que les tailles des strates varient et k ne sera peut-être pas semblable dans toutes les strates. L'étendue des valeurs n'est peut-être pas très large, mais elle l'est suffisamment pour qu'il soit impossible de considérer que le plan est autopondéré à un échelon supérieur à celui de la strate.

7.1.3 Pondération définitive stratifiée a posteriori

L'équipe propose d'ajouter un ajustement aux pondérations pour garantir que les estimations reproduisent les totaux importants connus et améliorer la précision des estimations par stratification *a posteriori* selon des groupements homogènes. Les plus importants de ceux-ci du point de vue de la participation à la population active sont la taille de la population pour chaque sexe et pour les principaux groupes d'âge pertinents (les 15 à 24 ans, les 25 à 49 ans, les 50 à 64 ans et les 65 ans et plus) parce que ces groupes ont des profils très différents dans la population active.

Les estimations postcensitaires de la population par âge et par sexe sont ancrées sur les données tirées du recensement le plus récent, mises à jour à l'aide des registres des logements et des statistiques de l'état civil sur les naissances et les décès. Ces données sont considérées très précises pour chaque grande ville ou district, mais elles tiennent peu compte de la migration interne, elles ont été projetées pendant près de 10 ans et l'équipe de l'EGM n'est pas convaincue qu'elles sont bonnes à l'échelon de la strate.

Les membres de l'équipe consultent les représentants de la Division des études démographiques du BSB et concluent qu'ils devraient recommander l'ajustement à l'échelon du district ou de la ville seulement,

mais que l'ajustement devrait être fait à l'échelon de la strate lorsque les résultats du prochain recensement seront disponibles.

La pondération définitive pour chaque personne qui répond à l'enquête sera donc :

$$w_{f,hza} = w_{nr,h} \times \frac{N_{za}}{\hat{N}_{za}}$$

où N_{za} est l'estimation postcensitaire de la population pour le groupe d'âge-sexe a de la ville ou du district z, $\hat{N}_{za}$ est la somme des pondérations (ajustées pour les non-réponses) pour tous les répondants du secteur z qui appartiennent au groupe d'âge-sexe a, et elle correspond à l'estimation directe de la population de ce groupe d'âge-sexe.

Remarquez que vous obtenez maintenant différentes pondérations pour les personnes d'un même ménage. L'échantillon est cependant autopondéré dans le groupe d'âge-sexe et la strate.

7.2 Estimation de la variance d'échantillonnage

Certains ouvrages standard contiennent les formules d'estimation de la variance pour les plans d'échantillonnage stratifiés à deux degrés qui sont autopondérés à l'échelon de la strate, même si elles sont relativement complexes. Étant donné les ajustements proposés pour les totaux de non-réponse et par groupe d'âge-sexe cependant, les expressions algébriques deviendront complexes et l'équipe convient de faire une recherche sur l'application d'une méthode par rééchantillonnage, par exemple, la méthode du jacknife ou du « bootstrap ». Elle n'a pas encore fait beaucoup de travail sur cet aspect jusqu'à maintenant.

Questions de récapitulation :

L'équipe de l'EGM est-elle justifiée de supposer qu'un biais de non-réponse est peu potable? Considérez les taux de réponse probables pour les ménages privés comparativement à ceux des ménages collectifs, en particulier ceux qui comptent de nombreux travailleurs de passage ou migrants récents. Serait-il possible de stratifier par type de ménages (logements) avant d'apporter l'ajustement pour les non-réponses dans ces strates?

Faites un commentaire sur le choix de l'échelon où est apporté l'ajustement pour les non-réponses. L'équipe a-t-elle pris la bonne décision, à votre avis, sur l'ajustement de la pondération pour les non-réponses? Vaudrait-t-il mieux ajuster les pondérations à l'échelon de la grappe, de la strate ou de la ville – du district? Expliquez.

Faites la même chose pour l'ajustement stratifié a posteriori pour l'âge et le sexe.

Chapitre 8 - Calcul de la taille de l'échantillon et répartition

8.0 Calcul de la taille de l'échantillon et répartition

L'équipe a établi la structure générale du plan d'échantillonnage et commence à examiner les détails du plan, c'est-à-dire la taille de l'échantillon, la répartition de l'échantillon entre les strates et le nombre de grappes par strate.

L'équipe considère les plus importantes variables à mesurer : les pourcentages ou proportions, par exemple, la proportion d'adultes dans la population active, la proportion de ceux qui ont un emploi, la proportion des personnes âgées de 55 ans et plus qui sont toujours économiquement actives.

En consultation avec le Comité directeur, l'équipe considère les facteurs dont il faudrait tenir compte pour déterminer la taille de l'échantillon et sa répartition entre les régions. Le Comité directeur précise qu'un degré élevé de précision est nécessaire à l'échelon national, mais qu'il est très important de contrôler la précision pour chaque région parce que les décisions sur la planification seront d'abord prises à l'échelon régional.

L'équipe discute du recours à une fonction de répartition optimale pour tenir compte de l'écart des coûts ou des variances entre les régions. Il n'y a cependant pas de bonnes indications des différences relatives dans les variances et l'équipe a l'impression que les coûts de la collecte des données, composante la plus importante du coût total de l'enquête, ne varieront pas énormément d'une région à l'autre. De plus, la répartition optimale ne garantit pas nécessairement la précision suffisante des estimations régionales. Il est donc décidé de cibler la même marge d'erreur pour toutes les régions.

Le représentant de la méthodologie au Comité directeur souligne que si nous obtenons une bonne précision pour chacune des 11 régions, c'est-à-dire une petite marge d'erreur pour les estimations les plus importantes, la marge d'erreur des estimations nationales ne sera donc pas supérieure à un tiers des marges d'erreur régionales, c'est-à-dire que le degré de précision devrait être très bon.

Les besoins de précision énoncés indiquent que la marge d'erreur (à un niveau de confiance de 95 %) devrait être de 2,5 % au plus pour les principales variables dans chacune des 11 régions. Étant donné que chaque région a une grande population, cela signifie que la taille de l'échantillon devrait être la même dans chaque région.

8.1 Calcul de la taille de l'échantillon par région

1. Taille initiale de l'échantillon dans chaque région

Voici une estimation préliminaire de la taille de l'échantillon, n_1, :

$$n_1 = \frac{z^2 \hat{P}(1-\hat{P})}{e^2}$$

où e est la marge d'erreur (0,025) et $\hat{P}$ est la proportion ciblée (supposons que P=0,5).

$$n_1 = \frac{(1.96)^2[.5(1-.5)]}{(0.25)^2}$$
$$= 1\,537$$

2. Le méthodologiste souligne qu'il n'est probablement pas nécessaire d'ajuster la taille de l'échantillon selon la taille de la population parce que la plus petite région – la Ville B – est très large (plus de deux millions de personnes).

3. La meilleure estimation de l'effet du plan d'échantillonnage dans l'ensemble est *deff=2* (selon des enquêtes semblables faites dans d'autres pays) :

$$n_3 = deff \times n_2$$
$$= 2 \times 1{,}537$$
$$= 3{,}074$$

4. La taille de l'échantillon est ajustée pour le taux de non-réponse qui ne devrait pas être supérieur à 20 % (de nouveau, compte tenu d'enquêtes semblables qui se sont déroulées ailleurs). Ce taux de non-réponse représente un modeste pourcentage de logements libres ou inoccupés et un nombre raisonnable de non-réponses à cause des familles temporairement absentes ou simplement difficiles à repérer. Voici l'estimation de la taille définitive de l'échantillon par région :

$$n = \frac{n_3}{r}$$
$$= \frac{3{,}074}{.8}$$
$$= 3{,}843$$

Sur l'ensemble des régions, la taille de l'échantillon total est donc 11 x 3 843 = 42 473 personnes ciblées dans l'enquête. L'équipe suppose qu'il y a environ deux adultes par ménage, c'est-à-dire 21 237 ménages à Belleterre, soit 1 922 par région.

Plusieurs membres de l'équipe ne comprennent pas pourquoi il est important de supposer qu'il y a deux adultes par ménage. Le méthodologiste explique que l'unité d'échantillonnage pour l'enquête est le logement, mais que toutes les estimations seront faites pour la population adulte au pays parce que ce sont les adultes, et non les logements, qui forment la population active. Étant donné que les besoins de précision sont liés aux adultes, ces calculs doivent donc d'abord être faits selon le nombre d'adultes nécessaires traduits ensuite en nombre de logements nécessaires dans l'échantillon.

Le méthodologiste confirme qu'il n'est pas nécessaire de considérer la correction d'échantillonnage pour population finie parce qu'elle est de 3 843 divisé par deux millions, soit 0,0019 pour la Ville B (la plus petite région).

L'équipe prévoit surveiller les taux de réponse pour chaque grappe sélectionnée, inscrire le genre de ménage et tout autre renseignement pertinent, afin de repérer les différences dans les taux de réponse par genre de ménage ou par région, ou d'autres variables dont on pourrait tenir compte pour modifier la stratégie d'ajustement de la pondération pour les non-réponses au cours des années ultérieures.

8.2 Définition de la stratification au deuxième degré et des grappes

On estime que la population du pays atteindra environ 44 millions de personnes vers 2005 et l'équipe est informée qu'elle devrait supposer une moyenne de quatre personnes par logement (compte tenu des enfants, des parents âgés et un petit pourcentage de logements ayant plus d'un ménage sans être des logements collectifs). Le résultat est un total d'environ 11 millions de logements considérés, aux fins de la planification, comme correspondant à 11 millions de ménages.

Compte tenu de la taille de l'échantillon de 21 237 ménages, cela signifierait un taux d'échantillonnage moyen de 1 sur 518. Cela peut sembler très faible, mais la taille de l'échantillon est très grande. Les populations varient énormément à l'échelon régional et les taux d'échantillonnage exigés par la taille de l'échantillon régional ciblé de 1 922 logements varient donc aussi beaucoup. Les taux d'échantillonnage entre les régions varieront de 1 sur 268 environ dans la Ville B à 1 sur 754 dans la région la plus populeuse, le District J.

Le plan d'échantillonnage a été établi pour permettre la définition d'un grand nombre de strates dont chacune aura un nombre substantiel d'unités primaires d'échantillonnage (UPÉ) intitulées grappes. Deux (à l'occasion trois) grappes seront sélectionnées par strate pour faciliter l'application de techniques d'estimation de la variance simple.

Les tailles des grappes et des strates afficheront bien entendu des différences dans chaque région et les taux d'échantillonnage dans chaque strate d'une région varieront donc aussi relativement. L'équipe n'oublie pas non plus qu'il n'est pas souhaitable que les taux d'échantillonnage dans les grappes deviennent trop petits parce qu'ils susciteraient une dispersion géographique indue, au moins en milieu rural.

Nous l'avons mentionné auparavant, les strates dans chaque région ont été définies selon les limites administratives ou municipales.

Tableau 8.1 : Stratification

Région	Comtés	Municipalités	Strates
Ville A	3	21	21
Ville B	3	17	17
Ville C	3	16	19
District D	11	23	23
District E	11	26	26
District F	13	29	28
District G	10	26	26
District H	7	21	23
District I	11	26	26
District J	13	34	34
District K	11	28	30
Total	**96**	**267**	**273**

Répartition proportionnelle à N entre les strates dans chaque région

L'échantillon régional sera fractionné entre les strates à l'aide de la répartition proportionnelle à N. N'oubliez pas que, selon la répartition proportionnelle à N,

$$n_h = n \times \frac{N_h}{N}$$

où n=1 922.

Des grappes sont créées dans chaque strate. Les considérations liées à l'efficience des opérations sur place limitent les étendues des options lorsqu'on détermine les tailles des grappes. Une équipe de trois intervieweurs et d'un surveillant devrait pouvoir faire entre 75 et 100 interviews par semaine. Compte tenu des circonstances imprévues, il est décidé que la taille de l'échantillon dans chaque strate sera de 40, pour que l'équipe puisse couvrir les deux grappes prévues dans une strate en une semaine de travail, pour un total de 80 interviews si le taux de réponse atteint 100 %. Si trois grappes sont sélectionnées dans une strate, il est prévu d'affecter quatre intervieweurs à l'équipe ou de prendre une journée ou deux de plus pour accomplir la tâche.

La création des grappes et la répartition de l'échantillon pour la Ville A donne ceci :

Tableau 8.2 : Répartition de l'échantillon entre les strates de la Ville A

Strate	Comté	Population de logements (N_h)	Nombre de grappes dans la strate	Taille moyenne de la grappe (logements)	Taille de l'échantillon de logements (n_h)
1	1	39 836	52	766	75
2	1	42 481	50	849,6	80
3	2	58 411	43	1 358,4	110
4	2	52 039	45	1 156,4	98
5	3	55 800	48	1 162,5	105
…	…	…	…	…	…
21	9	50 900	52	978,8	96
Total		1 020 600	1 092	934,6	1 922

Remarquez que les taux d'échantillonnage dans chaque strate sont à peu près égaux (parce que les strates sont de tailles à peu près égales). Les taux d'échantillonnage varient de 1 sur 529 à 1 sur 535 dans la Ville A.

Étant donné que l'échantillon prévu dans chaque grappe est de 40 logements et que nous pouvons sélectionner seulement un nombre entier de grappes, la taille de l'échantillon prévue dans la Ville A est de 2 000 (c.-à-d. 50 grappes de 40 ménages par grappe), selon le tableau ci-dessous.

Tableau 8.3 : Nombre de grappes à échantillonner selon les caractéristiques ci-dessus

Strate	Population de logements (N_h)	Nombre de grappes par strate	Nombre de grappes échantillonnées	Taille de l'échantillon de logements (n_h)
1	39 836	52	2	80
2	42 481	50	2	80
3	58 411	43	3	120
4	52 039	45	3	120
5	55 800	48	3	120
…	…	…	…	
21	50 900	52	3	120
Total	1 020 600	1 092	50	2 000

Le District D a 11 comtés et 23 strates ont été définies au total. La structure est relativement plus compliquée que celle de la Ville A parce que les populations des comtés varient largement et il est souhaitable que les limites des strates n'empiètent pas sur les limites du comté. Deux petits comtés ont été fusionnés dans un cas pour faire une seule strate et, dans un autre, le comté lui-même est une strate. Les comtés sont devenus deux ou trois strates dans la plupart des cas (réparties selon les limites municipales).

Un économiste membre de l'équipe a indiqué que dans un comté, les caractéristiques économiques et de la population active en milieu urbain et rural seront probablement très différentes et que l'échantillonnage par grappes peut être très peu efficient. Après étude, le méthodologiste suggère de former deux ou trois strates dans un comté, afin que l'une contienne les secteurs les plus urbains et l'autre (ou les deux autres), les secteurs les plus ruraux. Deux ou trois grappes seront ensuite sélectionnées dans chaque strate et les autres membres de l'équipe conviennent que l'échantillon devrait être raisonnablement bien équilibré.

Les tailles des strates sont maintenant de 19 400 à 37 600 logements, la moyenne étant de 28 300 logements par strate. Les grappes sont en moyenne relativement plus petites que celles de la Ville A, le nombre de grappes par strate est de 24 à 44 et, de nouveau, deux ou trois grappes seront sélectionnées par strate pour donner 80 ou 120 logements par strate.

Le tableau suivant affiche l'échantillon du District D réparti entre ces 23 strates à l'aide de la répartition proportionnelle à N.

Tableau 8.4 : Répartition de l'échantillon entre les strates dans le District D

Strate	Comté	Population de logements (N_h)	Nombre de grappes dans la strate	Taille moyenne de la grappe (logements)	Taille de l'échantillon de logements (n_h)
1 urbain	1	22 400	25	896	66
2 rural	1	26 200	32	818,8	77
3 urbain	2	30 200	25	1 208	89
4 rural	2	24 400	28	871,4	72
5 rural	2	30 600	38	805,3	90
…	…				…
9 urbain	4	21 800	26	838,5	64
10 rural	4	28 900	32	903,1	85
11 rural	4	32 200	36	894,4	95
12 rural	5	19 400	24	808,3	57
13 rural	6+7	24 200	26	930,8	72
14 urbain	8	28 900	34	850	85
15 rural	8	29 400	41	717,1	87
…,					
22 urbain	11	30 800	44	700	91
23 rural	11	22 900	29	789,7	68
Total		650 100	856	759,4	1 922

Tableau 8.5 : Répartition de l'échantillon entre les strates du District D

Strate	Population de logements (N_h)	Nombre de grappes dans la strate	Nombre de grappes échantillonnées	Taille de l'échantillon de logements (n_h)
1	22 400	25	2	80
2	26 200	32	2	80
3	30 200	25	3	120
4	24 400	28	2	80
5	30 600	38	3	120
…				…
9	21 800	26	2	80
10	28 900	32		
11	32 200	36	3	120
12	19 400	24	2	80
13	24 200	26	2	80
14	28 900	34	3	120
15	29 400	41	3	120
….				
22	30 800	44	3	120
23	22 900	29	2	80
Total	650 100	856	50	2 000

La taille de l'échantillon des adultes prévue dans l'ensemble est donc de 4 000 pour la Ville A et le District D. Le méthodologiste procède à la répartition pour toutes les villes et districts, et constate que la taille prévue de l'échantillon dans l'ensemble est de 22 000.

<u>Questions de récapitulation</u> :

Discutez de l'hypothèse de l'équipe de l'EGM, c'est-à-dire « un ménage, deux adultes ». L'équipe peut-elle faire mieux avant d'obtenir des données de la première édition de l'EGM?

Quelles seront les répercussions de l'ajout des logements collectifs sur l'hypothèse formulée au point précédent et quelle est votre réaction?

Pouvez-vous donner certaines raisons pratiques expliquant pourquoi la taille de l'échantillon dans la grappe devrait être plus large, ou plus petite, que celle suggérée par l'équipe de l'EGM? (Considérez les conditions sur place dans lesquelles travailleront les équipes d'intervieweurs). Ne vaudrait-il pas mieux former de plus petites grappes et ensuite, de plus petits échantillons de logements dans les grappes en combinant plusieurs grappes sélectionnées, afin de composer une tâche pour l'équipe d'intervieweurs? Faites des commentaires sur les avantages et les inconvénients de ce genre de modification.

Étant donné que le choix de la taille de l'échantillon cible une marge d'erreur de 2,5 % pour chacune des 11 régions dont les populations varient beaucoup, quelle marge d'erreur approximative obtiendrez-vous à l'échelon national, à votre avis?

Exprimez-vous sur l'hypothèse établissant en moyenne quatre personnes par ménage. Comment cette hypothèse variera-t-elle en milieu rural et en milieu urbain?

Chapitre 9 - Opérations de collecte des données

9.0 Opérations de collecte des données

L'équipe du projet a commencé à préparer les opérations de collecte des données immédiatement après avoir décidé d'appliquer l'interview sur place pour la collecte des données de l'EGM.

Le recours à un plan d'échantillonnage à deux degrés exige que la définition des grappes et la répartition de la taille de l'échantillon entre les strates et les grappes soit statistiquement efficientes, mais aussi rentables et réalistes d'un point de vue opérationnel. Étant donné ces considérations, la majeure partie du travail préparatoire aux opérations de collecte des données a été fait parallèlement à l'élaboration du plan d'échantillonnage.

9.1 Organisation régionale du projet de l'EGM

Le BSB est chargé de la conception et de la gestion de l'EGM dans l'ensemble. Il a cependant besoin de la collaboration active des organismes statistiques à l'échelon des districts et des sous-districts pour faire la collecte et le traitement des données.

Il y a un bureau de la statistique du district dans les trois principales villes. Ces bureaux seront intensivement engagés dans la collecte des données et les premières étapes du traitement des données, et ils travailleront sous l'orientation générale du BSB. Chacune des huit autres régions (Districts D à K) a aussi au moins un bureau de district et, dans certains cas, de sous-district. Chaque district regroupe de sept à treize comtés et ces comtés correspondent en majeure partie à deux ou trois strates. Il est donc toujours pratique d'organiser les équipes de collecte des données à l'échelon du comté, sous la coordination du bureau du district et, bien entendu du BSB. Les coûts de la collecte des données seront réduits parce que les membres des équipes d'intervieweurs seront probablement résidents dans les secteurs où ils seront affectés, ou ils habiteront à proximité.

Le BSB, qui sera le Bureau central de l'enquête, est chargé avant tout de l'établissement de l'échéancier de la collecte des données et des taux de réponse cibles. Le BSB, en collaboration avec les bureaux de la statistique du district, élabore aussi les systèmes de rapport et les formules de contrôle nécessaires pour garantir le listage et la sélection exacts des unités d'échantillonnage, ainsi que l'acheminement au moment opportun des questionnaires remplis aux bureaux de la statistique.

L'équipe du BSB prépare, de la même façon, les manuels des opérations et de formation des intervieweurs et des surveillants, et elle remet les ébauches de ces documents aux bureaux de district pour commentaires et révisions. Le BSB prépare également des exercices de formation et d'autre matériel, même si la majorité des séances de formation se dérouleront dans les bureaux de district ou de sous-district. Après avoir apporté la touche finale aux manuels et autres formules, le BSB les imprimera et les distribuera par l'intermédiaire des bureaux de district et de sous-district.

Étant donné l'ampleur des activités à accomplir et à coordonner, chaque bureau de district a nommé un chargé de projet régional de l'EGM. Celui-ci est chargé de la majeure partie de l'organisation du travail du bureau pertinent à l'EGM et il est la principale personne-ressource de l'équipe chargée de l'EGM au BSB. Il a été décidé que la saisie des données se déroulera aux bureaux de district et le chargé de projet régional est aussi responsable de la coordination de ces activités (voir le Chapitre 10 de l'étude de cas). Les chargés de projet régionaux seront responsables de la circulation du matériel entre le BSB et les bureaux locaux, ainsi que du retour des questionnaires remplis à leur propre bureau pour la saisie des

données et l'envoi des fichiers électroniques au BSB pour traitement final (vérification et imputation, repérage des valeurs aberrantes).

À l'aide des lignes directrices élaborées en collaboration avec l'équipe du BSB, les chargés de projet régionaux travailleront avec leurs homologues locaux pour identifier des candidats convenables qu'ils embaucheront et formeront à titre d'intervieweurs et de surveillants. L'équipe du BSB surveillera et observera ces étapes pour garantir l'uniformité et recevra aussi des rapports d'état d'avancement réguliers des chargés de projet régionaux.

Dans la plupart des cas, une équipe d'un surveillant et de trois ou quatre intervieweurs embauchés à l'échelon du comté (ou du bureau du sous-district) sera chargée de la collecte des données dans les grappes de sa strate de l'EGM. Dans certains cas, une équipe d'intervieweurs couvrira plus d'une strate, mais cette situation devrait être rare étant donné qu'il est prévu de faire toutes les interviews en un peu plus d'une semaine. L'équipe travaillera dans une grappe jusqu'à ce qu'elle soit achevée et passera à la suivante.

L'équipe du projet de l'EGM remarque qu'il faudra donc environ 275 surveillants au total (un pour chacune des 25 strates dans chacune des 11 régions) et près de 900 intervieweurs.

9.2 Relations publiques

L'EGM est une nouvelle enquête importante et le BSB a l'intention d'obtenir les données de la meilleure qualité possible, y compris les données des groupes de la population qui peuvent avoir des raisons d'être réticents à répondre aux questions, et l'équipe décide donc, avec l'approbation du Comité directeur de l'EGM, de lancer une vaste campagne de relations publiques au pays.

Un membre du personnel de la Division des communications du BSB est affecté à cette fin à l'équipe du projet de l'EGM pour préparer du matériel d'information convenable. Il comprendra une lettre de présentation expliquant les objectifs et l'importance de l'EGM. Le directeur général du BSB signera la lettre, ainsi qu'une personnalité appropriée dans la collectivité, soit le chef du bureau de la statistique du district ou du sous-district, ou encore un représentant bien connu de l'administration municipale.

Voici la première ébauche de la lettre :

Monsieur, Madame,

Votre ménage a été sélectionné pour participer à l'Enquête générale sur les ménages de Belleterre. Cette nouvelle enquête publique importante donnera de l'information essentielle sur les activités de la population de notre pays et sur la situation familiale en période actuelle de changement économique rapide.

Le Bureau de la statistique de Belleterre se charge de l'enquête. La confidentialité de l'information obtenue dans l'enquête est protégée en vertu de la loi. Vos réponses aux questions de l'enquête seront strictement confidentielles. Toutes les données obtenues dans cette enquête serviront à des fins statistiques seulement. Vos réponses seront combinées à celles de nombreux autres citoyens pour tracer un profil statistique fiable des conditions au pays.

Il faudra environ 20 minutes pour achever l'interview. S'il n'est pas pratique de réserver ce temps lorsque l'intervieweur communiquera avec vous la première fois, il(elle) prendra volontiers un rendez-vous pour procéder à l'interview au moment qui vous convient le mieux.

Le Bureau de la statistique de Belleterre reconnaît votre importante contribution et celle d'autres citoyens qui prennent le temps de nous faire part de ces renseignements essentiels. Nous vous remercions d'avance de votre collaboration précieuse.

Si vous avez des questions auxquelles l'intervieweur ne peut répondre, n'hésitez pas à communiquer avec le représentant local de l'EGM au Bureau de la statistique de la direction – du comté situé à ----------, ou en composant le numéro 1-23-456-7899.

M. Untel	*C. Lacase*
Directeur, BSB	*Agent responsable*
	(Nom du bureau local)

Un communiqué soigneusement formulé est aussi préparé et sera envoyé aux journaux locaux, stations de radio, bureaux de police et bureaux de l'administration municipale les informant des activités prochaines de l'EGM et demandant leur collaboration pour informer le grand public. Plusieurs des principaux paragraphes sont rédigés pour faciliter cette tâche et ils seront lus en ondes ou imprimés intégralement dans les journaux. Le même message sera ainsi diffusé partout au pays.

9.3 Préparation pour la collecte

L'équipe de l'EGM doit élaborer trois manuels : un pour le listage des logements, un pour l'interview et un pour les surveillants qui superviseront les deux opérations.

L'équipe de l'EGM consulte des manuels utilisés dans des enquêtes précédentes sur les ménages pour rédiger les manuels de l'intervieweur et du surveillant, et elle en tire des sections complètes à intégrer aux manuels de l'EGM. La majeure partie du manuel de listage doit cependant être élaborée au complet parce que le BSB n'a pas fait de listage auparavant. L'équipe peut consulter des manuels d'autres pays qui utilisent souvent des bases aréolaires pour élaborer le manuel.

Les surveillants seront embauchés et formés en premier lieu, comme nous l'avons mentionné ci-dessus. Ils devraient de préférence avoir une expérience préalable de l'enquête. Ils doivent aussi avoir les qualités et les aptitudes personnelles nécessaires pour orienter efficacement une équipe de plusieurs intervieweurs pendant les activités de listage et d'interview et pour intervenir en présence de répondants difficiles, tout en garantissant la communication fiable et à temps des données et d'autres renseignements au bureau de la statistique responsable. Tout le personnel embauché doit être sérieux et digne de confiance pour que le travail soit accompli selon les directives et pour protéger la confidentialité des données obtenues.

Les chargés de projet régionaux formeront les 275 surveillants environ dont ils auront besoin. L'équipe de l'EGM prévoit faire un effort énorme pour former suffisamment les chargés de projets régionaux aux volets des procédures de collecte, des concepts, des objectifs de l'enquête, etc., pour qu'ils soient en mesure de travailler avec des petits groupes de représentants de district ou de sous-district, afin d'interviewer les candidats aux postes de surveillant et de former ensuite ceux qui sont embauchés. Ceux-ci collaboreront avec les représentants des bureaux de district et de sous-district sous l'orientation de l'équipe de l'EGM et des chargés de projets régionaux pour interviewer et embaucher un nombre suffisant d'intervieweurs qualifiés.

L'équipe de l'EGM préférerait affecter au projet des membres du personnel actuel de la surveillance et de l'interview si possible. Cependant, si certains postes ne peuvent être dotés à l'interne en collaboration avec les divers bureaux régionaux, des annonces seront diffusées à la radio, dans les journaux locaux et les salles communautaires pour énumérer les qualifications essentielles au travail et obtenir des candidatures.

Les intervieweurs auront une formation approfondie pour les préparer à leurs tâches. Chacun d'eux doit d'abord étudier les manuels à domicile et rédiger plusieurs exercices. Ils auront ensuite plusieurs jours de formation en classe, y étudieront les techniques d'interview et les aptitudes à cette fin, et ils seront très bien informés du questionnaire. Les surveillants discuteront des exercices des intervieweurs rédigés à domicile et y apporteront des corrections pendant la formation, et il y aura des interviews fictives aux fins de la pratique avec d'autres intervieweurs et des répondants qui ne feront pas partie de l'échantillon de l'EGM.

9.4 Listage

Étant donné que le plan d'échantillonnage utilise des grappes délimitées sur les cartes, la première tâche des équipes d'intervieweurs est d'identifier les grappes sélectionnées et d'aller sur place pour lister les logements dans chaque grappe en ajoutant une description suffisamment détaillée pour que chacun soit identifié uniquement et facilement repéré s'il est sélectionné pour l'interview. Les équipes se déplacent à pied, à bicyclette ou en automobile, selon la taille géographique de la grappe.

Lorsque les listes sont complètes, le surveillant et le représentant du bureau du district ou du sous-district (selon les instructions obtenues de l'équipe de l'EGM par l'intermédiaire du chargé de projet régional) sélectionnent le nombre approprié de logements à l'aide de l'échantillonnage systématique des listes lorsque le chargé de projet régional leur a communiqué les origines choisies au hasard et les intervalles d'échantillonnage.

9.5 Collecte des données

La collecte des données commence après l'identification des logements sélectionnés. Les intervieweurs communiquent avec les ménages ou les citoyens qui habitent ces logements et, si possible, interviewent les membres admissibles du ménage. Si aucun membre admissible à l'interview n'est présent à domicile, l'intervieweur demande quand il(elle) peut communiquer de nouveau et, si personne n'est présent, il laisse une note précisant qu'il communiquera de nouveau plus tard. Si quelqu'un est présent, mais s'il ne peut passer l'interview, l'intervieweur essaie de prendre des dispositions pour procéder à l'interview dans les jours suivants et insiste de nouveau sur l'importance de la contribution de cette personne à l'enquête. Si le chef du ménage refuse carrément l'interview, l'intervieweur essaie de le persuader une fois de plus, mais se retire ensuite et mentionne le cas au surveillant chargé du suivi pour convertir un refus en réponse complète.

L'intervieweur vérifie si tous les documents d'interview sont complets et si le statut de chaque interview est correctement entré dans les formules de contrôle à la fin de chaque journée. Certaines vérifications sur place sont faites et si des erreurs sont détectées, il y a communication avec le ménage le jour suivant pour régler les incohérences ou les omissions. Les formules sont retournées au bureau lorsqu'elles sont complètes.

Le surveillant observe une certaine partie des interviews de chaque intervieweur, en particulier au début de la période de collecte des données, pour obtenir des données de très bonne qualité. Les membres du

personnel du bureau municipal et de l'équipe de l'EGM du BSB observent aussi certaines interviews dans divers secteurs du district. Les surveillants donnent une rétroaction aux intervieweurs pour corriger les erreurs et améliorer leurs aptitudes à l'interview.

L'équipe de l'EGM applique aussi un programme de ré-interviews, communique de nouveau avec un sous-échantillon des unités échantillonnées une semaine après la collecte des données pour vérifier des renseignements critiques du questionnaire. Des précisions sont apportées en collaboration avec le répondant s'il y a des différences par rapport à l'information originale. Les données tirées de la ré-interview serviront à estimer l'ampleur des divers types d'erreurs non dues à l'échantillonnage (par exemple, l'erreur de couverture, de mesure, de non-réponse ou de traitement) dans les données.

Les surveillants vérifient également si les questionnaires ont été remplis correctement et si les codes de statut attribués sont exacts. Ils doivent aussi garantir que tous les questionnaires et autres formules sont repérés et en ordre avant de les envoyer au bureau pour traitement.

Le surveillant et le personnel du bureau municipal organisent des séances d'information à l'intention des intervieweurs lorsque les interviews sont achevées, afin d'obtenir une rétroaction précieuse pour le traitement des données et la révision du questionnaire et des manuels en vue de la prochaine édition de l'EGM.

Questions de récapitulation :

Est-il pratique d'embaucher un aussi grand nombre d'intervieweurs et de surveillants qui travailleront pendant une période aussi brève à chaque trimestre? Quelles seraient les répercussions si l'on faisait appel à moins d'équipes et si l'on répartissait les interviews sur plusieurs semaines? (Considérez divers aspects, par exemple, les coûts, la qualité des données et les autres opérations liées à la collecte des données, notamment le listage et le dénombrement.)

Pouvez-vous suggérer des améliorations à apporter à l'ébauche de la lettre aux répondants pour les aider à comprendre à quel point l'enquête est importante et pour les convaincre que leurs réponses seront confidentielles?

Il a été mentionné que la saisie des données serait faite aux bureaux de district. Est-ce la disposition la plus efficace ou serait-il préférable de faire la saisie des données au Bureau central?

Chapitre 10 - Traitement

10.0 Traitement

Dès leur retour aux bureaux de la direction des districts ou des sous-districts, les questionnaires sont traités pour obtenir un fichier de données d'où seront dressés des tableaux et les résultats de l'enquête feront l'objet d'une analyse. Les étapes à franchir à ce volet de l'enquête comprennent le codage, la saisie des données, la vérification et l'imputation, la détection et le traitement des valeurs aberrantes, ainsi que la création d'une base de données. Diverses procédures de contrôle qualitatif et d'assurance de la qualité sont aussi élaborées et appliquées.

10.1 Saisie des données et codage

Les questions fermées sont codées d'avance sur le questionnaire et les opérateurs de la saisie des données entrent simplement les numéros de code inscrits à côté des cercles ou des cases qui correspondent à la réponse. Les réponses aux questions ouvertes qui portent sur les nombres (âge du répondant en années, heures travaillées, traitement, etc.) peuvent aussi être saisies directement à partir du questionnaire.

D'autre part, le questionnaire contient plusieurs questions, par exemple le genre de travail accompli, qui ont tellement de réponses possibles que les catégories et les codes ne peuvent être ajoutés au questionnaire. Dans ces cas, une liste de code a été préparée et remise au personnel du bureau qui inscrit le code approprié sur le questionnaire avant la saisie des données. La question JD6 demande, par exemple, au répondant dans quelle branche se déroule la principale activité économique de son employeur. Une liste de codes de secteurs économiques a donc été remise au personnel du bureau qui attribuera la valeur la plus appropriée à la réponse avant la saisie des données.

Il y a aussi quelques questions avec case « autre, veuillez préciser » qui ne peuvent être codées d'avance. L'équipe de l'EGM a élaboré pour ses cases une liste préliminaire des sujets les plus probables qui sont inscrits en réponse et y ont attribué des codes. Un membre de l'équipe étudie un certain nombre de questionnaires (peut-être 100 environ) pour déterminer si d'autres sujets deviennent assez fréquents pour mériter un code. Toutes les réponses qui peuvent être codées à l'aide de la liste révisée le sont. (On a demandé aux opérateurs de la saisie des données d'entrer la réponse donnée dans une zone de texte pour toutes les réponses qui ne peuvent être codées avant la saisie des données.) Ces cases sont étudiées après la saisie des données et le personnel essaie de nouveau d'élaborer une liste complète de codes. Toutes les cases qui ne peuvent toujours pas être codées reçoivent le code implicite pour « autre ».

À la conclusion du codage, les lots de questionnaires sont envoyés aux opérateurs de la saisie des données qui entrent les réponses codées dans l'ordinateur pour créer un fichier de données préliminaires. L'équipe de soutien informatique de l'équipe de l'EGM a préparé des écrans de saisie des données à cette fin.

Au cours de l'enquête, à partir de la collecte des données jusqu'à la préparation des tableaux, l'équipe applique un certain nombre de procédures de contrôle qualitatif et d'assurance de la qualité. Le programme de contrôle qualitatif sert à vérifier un pourcentage du travail de chaque opérateur de l'entrée des données et à examiner de nouveau son travail si le nombre d'erreurs est supérieur à une limite déterminée. Une procédure semblable est appliquée pour vérifier le codage. L'échantillonnage d'acceptation est fait dans les deux cas et la production quotidienne de chaque opérateur est traitée comme un lot.

Une rétroaction tirée des résultats de l'échantillonnage d'acceptation est communiquée à chaque opérateur et, si nécessaire, une formation supplémentaire est offerte (ou, si les cibles ne sont pas atteintes, l'opérateur est libéré de sa tâche). Les surveillants et les chargés d'enquête sont aussi informés des indicateurs pertinents.

10.2 Vérification

La vérification commence lorsque les intervieweurs, et ensuite les surveillants, ont prétraité les questionnaires pour vérifier s'ils sont correctement remplis et si le suivi auprès du répondant est nécessaire. Un autre prétraitement est fait au bureau avant la saisie des données et le codage.

D'autres vérifications sont faites pendant la saisie des données parce que les écrans sont programmés pour détecter certaines erreurs à l'entrée au clavier. Elles comprennent les vérifications de validité pour les codes inadmissibles. Si l'opérateur entre une valeur de trois à huit pour le sexe du répondant, par exemple, l'ordinateur émet un signal sonore et attend une correction parce que cette variable devrait être codée 1 pour homme, 2 pour femme ou 9 pour une non-réponse.

Les programmes de saisie des données comprennent aussi des vérifications de convergence pour les caractéristiques erronées de l'instruction « passez à ». À la section de la description de fonctions, par exemple, seuls les travailleurs autonomes sont sensés répondre aux questions JD2 à JD4. Si le répondant n'est pas un travailleur autonome, il devrait passer outre les questions JD2 à JD4 et il faudrait entrer un code (p. ex., « 8 ») indiquant que les questions ne s'appliquent pas. Si l'opérateur de la saisie des données essaie d'entrer une réponse à l'une de ces questions pour une personne qui n'est pas travailleur autonome, l'écran de saisie des données émet un signal sonore pour que l'opérateur vérifie la donnée.

Une vérification plus complète est faite lorsque les données sont sur support électronique. Dès que la touche finale est apportée au questionnaire, l'équipe commence à formuler des règles de vérification des combinaisons inacceptables de codes pour les groupes de questions connexes (deux ou trois questions ou plus). Un enregistrement serait rejeté à l'application d'une règle de vérification, par exemple, si le traitement semble trop élevé ou trop faible étant donné le nombre d'heures travaillées. Ces règles de vérification sont appliquées automatiquement au fichier de données. L'équipe a cependant été informée des risques de la survérification et de la nécessité d'apporter seulement les modifications minimales nécessaires aux données du répondant, et les rejets à la vérification seront imputés seulement si l'enregistrement a des répercussions importantes sur l'estimation. Des zones de « signalisation » particulières sont donc prévues pour chaque question, afin d'indiquer si la zone a été rejetée à la vérification et si elle devrait être entrée dans le système d'imputation.

Étant donné les étapes de prétraitement et de vérification manuelle, très peu d'enregistrements ont encore tellement de rejets à la vérification à cette étape qu'ils doivent être déclarés inutilisables. Ces cas sont traités comme une non-réponse totale. Ils sont traités comme les refus et d'autres occurrences de non-réponse totale à l'étape de la collecte des données, et un ajustement est apporté aux pondérations de l'enquête pour redressement.

10.3 Imputation

Les zones rejetées à l'application d'une règle de vérification et les non-réponses à une question des enregistrements utilisables du questionnaire sont analysées pour imputation éventuelle. La méthode d'imputation varie selon le genre de question.

La redondance dans les questions de quelques zones permet l'imputation déterministe. La section sur la composition des ménages comprend, par exemple, une zone pour la taille totale du ménage, ainsi que le nom, l'âge, etc. de tous les membres du ménage. Si ces zones sont toujours incohérentes malgré la vérification précédente, la taille totale du ménage est signalée inexacte et l'imputation déterministe est appliquée à la valeur qui devient le nombre de personnes inscrites dans les autres zones.

D'autre part, si un répondant déclare un revenu de la vente de produits agricoles, mais n'inscrit pas une somme, la valeur est imputée à l'aide de la méthode hot-deck (donneur de l'enquête) aléatoire en utilisant comme donneurs éventuels tous les autres répondants dans la même strate qui tirent un revenu de cette activité. S'il n'y a pas de donneur convenable, le groupe de donneurs éventuels est élargi pour comprendre tous les répondants de la même région.

Lorsque l'imputation est achevée, les programmes de vérification peuvent de nouveau être appliqués aux fichiers de données pour garantir que les données sont entièrement cohérentes. Il faut appliquer de nouveau l'imputation au fichier pour éliminer les quelques enregistrements toujours rejetés après l'application de règles de vérification et le fichier est ensuite vérifié de nouveau pour en déterminer la cohérence.

Les indicateurs de diagnostic, par exemple, le nombre de cases imputées dans chaque zone, le genre d'imputation appliquée, le nombre de donneurs admissibles, la fréquence de leur utilisation et d'autres mesures, sont simultanément enregistrés comme entrées au processus d'évaluation de l'enquête. Ces indicateurs serviront à l'étape de l'évaluation pour calculer les taux de rejet à la vérification et les taux d'imputation pour les principales variables de l'enquête et les variables qui affichent le plus grand nombre de problèmes. Les signalisations d'imputation et de vérification sont aussi maintenues au fichier pour aider à déterminer la qualité de la base de données définitive dans l'ensemble.

10.4 Détection et traitement des valeurs aberrantes

Étant donné le genre de questions, relativement peu de zones du fichier de données de l'EGM sont des valeurs aberrantes, c'est-à-dire des observations extrêmes ou influentes. Il y a cependant plusieurs variables, par exemple les traitements et même les heures travaillées, auxquelles peuvent être attribuées des valeurs extrêmes, mais possibles. Une personne déclare, par exemple, avoir travaillé 96 heures la semaine dernière, information qui peut en fait être vraie. Un autre répondant déclare un traitement de 75 000 $ la semaine dernière, ce qui peut aussi être vrai, mais le cas serait très rare, surtout si cette personne inscrit à l'entrée profession qu'elle est enseignante à l'école élémentaire. Son revenu réel serait probablement plus près de 750 $ dans ce cas. Même si la réponse était vraie, elle pourrait avoir une influence indue sur les estimations de l'enquête si, pour une certaine raison, son ménage avait une pondération de l'échantillonnage inhabituellement élevée (par exemple, à cause d'ajustements de la pondération pour les non-réponses).

Afin de détecter et de régler les cas du genre, l'équipe de l'EGM a élaboré des approches systématiques de la détection et du traitement des valeurs aberrantes axées sur les quantiles de l'échantillon (quartiles et médiane). Les membres de l'équipe n'étant pas certains du choix des valeurs critiques, ils ont préféré des valeurs relativement faibles (c'est-à-dire qui permettront probablement de détecter les « valeurs aberrantes trop nombreuses ». Les analystes examinent ensuite toutes les valeurs signalées pour mieux comprendre les répartitions des données avant de prendre des décisions définitives, à savoir quels cas devraient être définis comme valeurs aberrantes et comment il faudrait les traiter. Compte tenu de cette expérience, l'équipe élaborera ensuite une approche plus systématique de la détection et du traitement des valeurs aberrantes pour la deuxième édition de l'EGM.

10.5 Création de la base de données

Les dernières étapes du traitement sont la création d'un fichier non hiérarchique qui servira de fichier élémentaire de données de l'enquête, le calcul des pondérations et leur ajout au fichier. L'équipe de l'EGM entre les résultats au fichier et le sauvegarde ensuite en une présentation qui convient au principal logiciel de traitement statistique du BSB (afin d'accélérer la préparation des tableaux et d'autres sorties de l'enquête).

Questions de récapitulation :

Il est mentionné ci-dessus que les questions qui ont de nombreuses catégories de réponses possibles doivent être codées à la main avant la saisie des données. Pouvez-vous suggérer un autre moyen de traiter ces cas? Serait-il éventuellement plus ou moins efficient que le codage manuel? Expliquez.

Est-il réaliste d'appliquer l'échantillonnage d'acceptation suggéré pour le contrôle qualitatif du codage et de l'entrée des données décrit ci-dessus, étant donné que 300 codeurs et opérateurs de l'entrée des données environ sont répartis entre les divers bureaux de district? Pouvez-vous suggérer une autre approche du traitement des données et du contrôle qualitatif qui serait plus efficace? Quels en seraient les inconvénients?

Suggérez d'autres approches de l'imputation qui pourraient servir à l'EGM.

Quelles variables auxiliaires faudrait-il considérer lors de la création des classes d'imputation?

Chapitre 11 - Analyse des données

11.0 Analyse des données

L'équipe de l'EGM a maintenant une base de données complète et épurée qu'elle doit analyser pour communiquer de l'information utile aux utilisateurs, afin de les aider à répondre aux questions qui ont motivé l'enquête.

11.1 Mesures sommaires

L'analyse préliminaire est surtout descriptive et comprend les distributions de fréquence à une variable (tris à plat), ainsi que les moyennes, proportions et totaux estimés, pour toutes les variables produites directement à partir du questionnaire ou qui en sont tirées pendant le traitement. Les estimations sont accompagnées de leurs erreurs-types estimées. Il y a aussi un nombre substantiel de totalisations croisées (tris croisés), nombre d'entre elles sont planifiées, mais d'autres sont élaborées à la suite de l'analyse préliminaire des données. Après l'analyse préliminaire, par exemple, un analyste décide d'étudier le genre d'emplois des hommes et des femmes. Voici un extrait de son analyse :

Quels genres d'emplois occupent les hommes et les femmes?

Il y a une grande différence de répartitions des professions entre les hommes et les femmes. Malgré un déplacement apparent des femmes vers les professions de gestion et des hommes vers les secteurs des ventes et services, les structures professionnelles traditionnelles des hommes et des femmes se maintiennent :

i. *Il y a plus de femmes que d'hommes qui travaillent dans les secteurs suivants : ventes et services, affaires, finances et administration, sciences sociales, enseignement, service gouvernemental et religion, arts, culture, loisirs et sports, santé.*

ii. *D'autre part, il y a plus d'hommes que de femmes affectés à des postes de col bleu, de gestion, de sciences naturelles et appliquées.*

Tableau 11.1a : Emploi par profession et sexe

Profession	Répartition entre les professions (%)	
	Hommes	Femmes
Gestion	11,9	8,6
Affaires, finances et administration	9,5	26,9
Sciences naturelles et appliquées	8,6	2,4
Santé	2,0	9,1
Sciences sociales, enseignement, service gouvernemental et religion	4,8	8,9
Arts, culture, loisirs et sports	2,4	3,2
Ventes et services	19,6	31,5
Commerce, transport et fonctionnement du matériel	24,5	2,2
Industrie primaire	6,0	2,0
Traitement, fabrication et services publics	10,5	5,2
Total	100,0	100,0

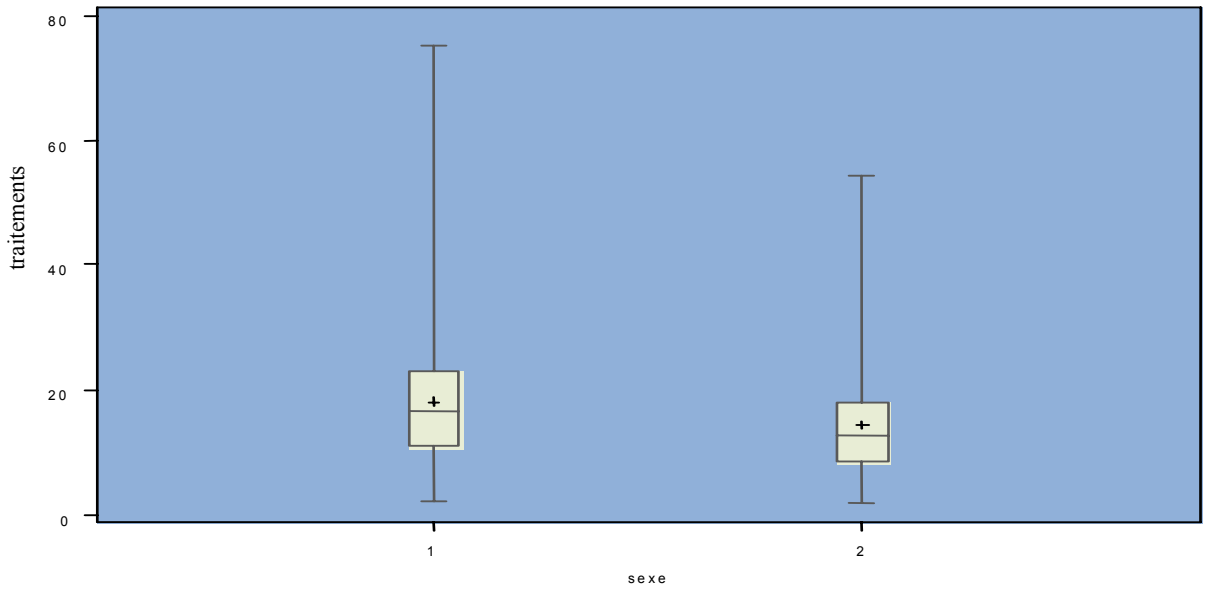
Tableau 11.1b : Emploi par profession et par sexe

Profession	Répartition entre les hommes et les femmes (%)	
	Hommes	Femmes
Gestion	62,6	37,4
Affaires, finances et administration	29,8	70,2
Sciences naturelles et appliquées	81,4	18,6
Santé	20,7	79,3
Sciences sociales, enseignement, service gouvernemental et religion	39,5	60,5
Arts, culture, loisirs et sports	47,4	52,6
Ventes et services	42,6	57,4
Commerce, transport et fonctionnement du matériel	93,0	7,0
Industrie primaire	78,1	21,9
Traitement, fabrication et services publics	70,7	29,3
Total	54,5	45,5

11.2 Tests d'hypothèses sur la population

Même si des mesures sommaires sont nécessaires comme point de départ de l'analyse, la majorité des utilisateurs et des analystes veulent tester certaines hypothèses sur la population. Une analyste veut, par exemple, étudier la différence de traitement des employés de 15 à 65 ans selon le sexe. Les hommes sont mieux rémunérés que les femmes, selon son hypothèse. Elle examine d'abord l'estimation des taux horaires moyens des deux et remarquent une différence substantielle : 19 $ pour les hommes et 15 $ pour les femmes. Elle trace un graphique des données et constate aussi que les répartitions affichent une asymétrie marquée, c'est-à-dire que la médiane pour les hommes est de 17 $ et celle des femmes est de 13 $. Ceci est représenté à la Figure 11.1.

Figure 11.1 Taux horaires des hommes (1) et des femmes (2)



L'analyste teste l'hypothèse nulle selon laquelle les traitements moyens des hommes et des femmes sont les mêmes, comparativement à l'hypothèse alternative, selon laquelle ils sont différents. Elle constate qu'à un niveau de confiance de 95 %, ils sont différents.

Après avoir réfléchi au lien entre les traitements et le sexe cependant, l'analyste se demande si d'autres variables pourraient aussi avoir des répercussions. Elle soupçonne que l'âge, la scolarité, la branche d'activité et la profession peuvent aussi être liés aux gains. Elle décide de faire une analyse de variance de la variable « traitements » posée comme variable dépendante, mais elle doit d'abord décider comment traiter certaines variables indépendantes.

Premièrement, l'âge n'est pas le bon genre de données à utiliser dans une analyse de la variance et, avant de faire cette analyse, l'analyste doit d'abord grouper la variable d'âge. Elle décide de la grouper en tranches de 10 ans et d'intégrer une variable calculée à la base de données intitulée « groupe d'âge » qui comprendra la valeur 1 pour une personne âgée de moins de 25 ans, 2 pour une personne âgée de 25 à 34 ans, etc.

Elles doit ensuite déterminer que faire des variables profession et branche d'activité. Celles-ci sont établies à l'aide des systèmes de classification type qui comprennent un code à quatre chiffres. Le premier chiffre indique la branche d'activité générale ou le groupe de professions et chaque chiffre suivant précise davantage. Il y a des milliers de groupes au total pour chaque système de codage et ils sont trop nombreux pour les utiliser dans une analyse de la variance. Il y a trop peu de répondants dans certains groupes pour donner des résultats significatifs. Les systèmes de codage sont aussi très subjectifs et l'on a impression que les taux d'erreur sont très élevés au niveau de quatre chiffres. Compte tenu de tous ces points, l'analyste décide de tronquer les codes au seul premier chiffre pour l'analyse.

Après avoir calculé les variables nécessaires, l'analyste fait l'analyse de la variance à l'aide des traitements comme variable dépendante et du groupe d'âge, du sexe, de la scolarité, de la branche d'activité et de la profession comme variables indépendantes. Le modèle se révèle significatif, c'est-à-dire qu'il explique une variation marquée dans les traitements et chacune des variables indépendantes devient aussi un élément important de la variation des traitements.

L'analyste soupçonne cependant qu'il peut y avoir une certaine interaction entre les variables indépendantes. Qu'en est-il, par exemple, si l'âge de la personne détermine les répercussions du sexe sur les traitements? Afin d'étudier ces questions, l'analyste décide d'ajouter toutes les répercussions de l'interaction au modèle pour examiner leur influence sur les traitements. Elle constate que toutes les répercussions de l'interaction sont importantes à l'exception du groupe d'âge par branche d'activité. Cela signifie que, non seulement le groupe d'âge, le sexe, la scolarité, la branche d'activité et la profession déterminent les traitements, mais que la plupart des combinaisons de ces variables ont aussi des répercussions sur les traitements.

L'analyste réalise que la question est très compliquée et qu'il faut faire davantage de recherche.

11.3 Autre analyse

Outre l'analyse déjà faite, l'analyste de l'EGM veut aussi analyser certaines caractéristiques dans le temps et apporter éventuellement des corrections en fonction des variations saisonnières lorsque suffisamment de données seront disponibles.

Questions de récapitulation :

Quel genre de données sont les variables âge et groupe d'âge?

Quelles autres variables peuvent servir à l'analyse de la variance?

Quels autres genres d'analyses suggéreriez-vous?

Chapitre 12 - Diffusion des données

12.0 Diffusion des données

L'équipe de l'EGM a analysé les données et elle doit maintenant faire rapport sur les résultats aux utilisateurs pour les aider à répondre aux questions qui ont motivé l'enquête. L'équipe évaluera simultanément les données selon les objectifs de l'enquête.

12.1 Principal rapport de l'enquête

Le corps du rapport commence par une introduction donnant une brève description des objectifs de l'EGM, les divers organismes engagés dans le plan d'enquête et l'exécution, les principaux utilisateurs prévus et un aperçu de la méthodologie utilisée.

La majeure partie du principal rapport de l'enquête comprend des tableaux prévus tirés directement de la base de données des réponses au questionnaire. Cependant, afin que ces tableaux soient le plus utile possible pour les utilisateurs, l'équipe prépare un rapport textuel résumant les plus importants résultats, des considérations sur leur signification selon les objectifs originaux de l'enquête et un commentaire sur les points forts et les points faibles des données. Le texte du rapport comprend un certain nombre de graphiques et de tableaux pour illustrer les principaux points. Le rapport comprend aussi des commentaires sur des points autres que les données de la principale base de données, par exemple, il mentionne les problèmes particuliers relevés pendant le travail de terrain ou les questions dont les réponses ont exigé des taux particulièrement élevés d'imputation. D'autres rapports, qui seront diffusés plus tard, sont aussi mentionnés (Section 12.3).

La dernière section du rapport principal tire des conclusions sur la situation du marché du travail dans l'ensemble à Belleterre et comprend des recommandations pour étudier davantage plusieurs points importants qui ne sont pas approfondis dans cette première édition de l'EGM.

Les tableaux des annexes affichent les distributions de fréquence à une variable (tris à plat), ainsi que les moyennes, proportions et totaux estimés, pour toutes les variables produites directement à partir du questionnaire ou qui en sont tirées pendant le traitement. Les estimations sont accompagnées de leurs erreurs-types estimées.

12.2 Rapport sur la qualité des données

L'équipe du projet de l'EGM prépare aussi un rapport complet sur la qualité des données de la première édition de l'enquête, afin de faciliter l'élaboration de l'enquête pour les années ultérieures. Le rapport comprendra des mesures de la variabilité de l'échantillonnage, notamment, les coefficients de variation ou les effets du plan d'échantillonnage. Il comprendra aussi les taux d'inoccupation, les taux de réponse (totale et partielle) et un bon nombre de mesures et d'indices obtenus à chaque étape du contrôle qualitatif appliqué au codage, à la saisie des données et à d'autres phases de l'enquête. Il décrira le processus de vérification et d'imputation, et considérera les taux d'imputation et tout autre problème découvert dans les données pendant ce processus. Il y aura de plus un commentaire sur l'à-propos de la comparaison des résultats de l'EGM avec d'autres sources de données disponibles, l'accent étant mis sur la comparabilité ou les différences des concepts et des outils de mesure utilisés et on précisera si l'étalonnage, selon une source plus fiable, a été utilisé pour certaines des données.

Voici une liste de certaines des valeurs examinées dans le rapport :
- coefficients de variation par région,
- effets du plan d'échantillonnage par région,
- taux d'inoccupation par milieu urbain – rural,
- taux de non-réponse par région et milieu urbain – rural dans la région,
- taux de non-réponse par genre de non-réponse,
- données utilisées pour la stratification *a posteriori*,
- taux de rejet à la vérification par question,
- taux d'imputation par question,
- taux d'erreur de codage,
- taux d'erreur de saisie des données,
- nombre moyen de communications pour conclure un cas de réponse,
- nombre moyen de communications pour conclure un cas de non-réponse,
- durée moyenne de l'interview des cas de réponse.

12.3 Autres rapports

L'équipe de l'EGM produira probablement plusieurs rapports supplémentaires, y compris :

i. Rapport d'évaluation de l'enquête. Il comprendra les recommandations de l'équipe sur les améliorations à apporter au processus de l'enquête pour que l'édition suivante de l'EGM fonctionne encore mieux et avec davantage d'efficience.

ii. Rapports d'analyse détaillée. L'équipe de l'EGM prévoit produire une série de rapports analytiques à rédiger en collaboration avec divers organismes utilisateurs.

iii. Rapports techniques. Ces rapports donneront des détails sur la méthode d'enquête, le plan d'échantillonnage, la méthode d'estimation, les procédures de collecte et de traitement des données, etc.

12.4 Confidentialité et contrôle de la divulgation

Le contrôle de la divulgation des résultats définitifs de l'enquête est une question importante. Il n'y aura pas de fichier de microdonnées à grande diffusion et le contrôle de la divulgation cible donc les tableaux diffusés et les autres mesures sommaires. (Le BSB pourrait cependant donner l'accès aux microdonnées à certains analystes autorisés qui auront prêté serment comme s'ils étaient employés du BSB. Toutes les pénalités pour divulgation non autorisée de renseignements confidentiels peuvent donc s'appliquer à ces personnes.)

Dans le cas des données publiées dans le principal rapport de l'enquête, l'équipe étudie des méthodes axées sur la restriction qui sont disponibles pour les données en tableau. La majorité des données de l'EGM sont discrètes, mais certaines variables sont continues et il est donc plus difficile de les traiter. Plusieurs méthodes sont à l'étude, mais l'équipe n'a pas encore tiré de conclusion, à savoir laquelle adopter. Étant donné cependant que l'un des principaux objectifs de la première édition de l'EGM est d'obtenir de l'information améliorée pour peaufiner l'EGM en cours, la stratégie générale adoptée doit être très conventionnelle. L'équipe est donc disposée à supprimer un nombre trop élevé de renseignements si nécessaire en faisant des évaluations détaillées de la qualité des données et du risque de divulgation pour l'application ultérieure de règles sur la divulgation qui maximiseront le nombre de renseignements qui peuvent être diffusés à l'avenir.

Voilà pourquoi l'équipe a déterminé une règle limite préliminaire précisant qu'il doit y avoir au moins dix répondants dans une case d'un tableau avant de le diffuser. L'équipe considère que cette mesure et la suppression résiduelle connexe sont suffisantes dans la plupart des cas et probablement très conventionnelles dans l'ensemble. L'équipe préfère regrouper les cases à caractère délicat avec les cases voisines lorsque c'est possible, au lieu de simplement les supprimer, parce que cette mesure aidera à minimiser les nombres autrement élevés de suppression de cases complémentaires.

Questions de récapitulation :

Le principal rapport de l'enquête devrait-il être mis à la disposition du grand public et être ainsi ajouté aux séries régulières de publications du BSB ou le rapport de la première édition devrait-il être restreint au personnel du BSB et aux utilisateurs de données connus dans d'autres ministères?

Recommandez d'autres méthodes de contrôle de la divulgation possible qui pourrait être considérées pour l'EGM.

Chapitre 13 - Planification et gestion de l'enquête

Enquête générale sur les ménages (EGM) de Belleterre
Évaluation personnelle des pratiques de planification et de gestion du chargé de projet

13.0 Introduction

L'objectif de ce rapport est de décrire et d'évaluer les pratiques de planification et de gestion appliquées à l'Enquête générale sur les ménages (EGM) de Belleterre, afin de tirer des leçons de l'expérience. Il s'agit de notes personnelles qui ne seront pas distribuées à l'équipe du projet ou au Comité directeur. À titre de chargé de projet, j'ajouterai cependant certains points considérés dans ce rapport au rapport officiel d'évaluation de l'enquête aux fins de l'amélioration du questionnaire et des procédures pour la prochaine édition de l'EGM.

13.1 Contexte

Pleinement conscient du besoin de plus en plus urgent d'information à jour sur l'état de l'économie et de la situation socioéconomique de la population, le Bureau de la statistique de Belleterre (BSB) a décidé d'améliorer son programme statistique. Le BSB convient en particulier de la nécessité d'obtenir des données pertinentes sur la situation des ménages en milieu urbain et rural au pays.

L'équipe du projet, avec le soutien du gouvernement national, a donné suite à une proposition d'enquête sur les conditions socioéconomiques des ménages en milieu urbain et rural au pays. Les grands sujets considérés dans l'enquête étaient les caractéristiques sociodémographiques, le marché du travail, les traits communs des revenus et dépenses, ainsi que les indicateurs des conditions de vie. Un certain nombre d'importants ministères nationaux ont demandé à l'équipe de l'EGM d'obtenir de l'information supplémentaire sur l'état de santé de la population, les activités agricoles des ménages en milieu rural et urbain et les petites entreprises.

Dans le contexte de ces objectifs, un Comité directeur a été nommé pour surveiller le déroulement de l'enquête. Le Comité directeur comprenait des directeurs représentant les domaines de la statistique sur la main-d'œuvre et les ménages, de la méthodologie, de l'informatique et de la collecte des données. Le but du Comité directeur était de donner suite aux besoins d'information du BSB et d'approuver d'importantes décisions, notamment, l'énoncé des objectifs, le budget, l'échéancier, etc.

L'une des premières étapes du processus de planification a été l'identification des domaines à propos desquels peu ou pas d'information existait et la préparation d'une proposition d'enquête. Un chargé de projet intérimaire a été nommé et celui-ci a trouvé de l'information initiale sur les solutions de rechange à une enquête (p. ex., les données disponibles d'une autre enquête ou source administrative), ainsi que sur le coût et la faisabilité d'une enquête. Cet examen préliminaire était axée sur la consultation d'experts de diverses disciplines qui pourraient être engagés dans l'équipe de l'enquête (bien que ces experts n'aient pas, en définitive, été nommés membres de l'équipe de l'enquête). Le Comité directeur a considéré la proposition d'enquête et décidé de procéder à une élaboration plus substantielle en vue de formuler un énoncé des objectifs et de tracer un plan d'enquête.

13.2 Planification de l'EGM

Le Comité directeur a ensuite décidé de m'affecter à titre de chargé du projet de l'EGM. J'ai obtenu les grands objectifs de l'enquête et le nom des personnes-ressources d'organismes à l'externe qui s'intéressaient à l'information (p. ex., le ministère de la Santé). On m'a demandé par la suite de former une équipe de projet, d'élaborer rapidement un énoncé approximatif des objectifs et de tracer un plan d'enquête.

La composition de l'équipe du projet a été une étape importante. Mon but, comme tout chargé de projet, était de trouver des gens d'expérience avec qui j'avais collaboré avec succès à certaines occasions. J'ai communiqué avec le directeur de la Division de la méthodologie pour demander les services de M. X qui avait servi à titre de méthodologiste d'une équipe de projet que j'avais dirigée auparavant. J'ai été informé que M. X avait quitté le BSB et, compte tenu des conditions changeantes du marché, que la dotation était non seulement insuffisante à la division, mais qu'elle manquait aussi en particulier de personnel d'expérience. J'ai obtenu un méthodologiste (M. M) qui avait seulement quelques années d'expérience. J'ai été informé que le surveillant du méthodologiste examinerait toutes ses décisions. J'ai eu plus de succès lors du recrutement des autres membres de l'équipe :

- une coordonnatrice de la matière (M^{me} S) et une équipe d'experts en la matière dans les domaines de la statistique sur la main-d'œuvre, l'économie, etc.
- un expert des systèmes informatiques (M. P), qu'on m'a présenté comme un programmeur chevronné,
- un agent des opérations et de la collecte des données (M^{me} D) avec qui j'ai travaillé auparavant.

L'équipe de l'enquête maintenant formée a commencé à élaborer l'énoncé des objectifs et à confirmer la faisabilité de l'enquête. La première réunion de l'équipe du projet a été convoquée, le but étant de présenter les membres de l'équipe du projet l'un à l'autre, d'expliquer leurs rôles et la structure de l'obligation de faire rapport, et de discuter des buts du projet. J'ai précisé, à l'intention des novices de l'équipe du projet, que les surveillants hiérarchiques examineraient leur travail.

Le but de la deuxième réunion de l'équipe du projet était de commencer le travail sur l'énoncé des objectifs. Il est devenu évident, après une certaine discussion, que nous ne pourrions obtenir de l'information détaillée sur de nombreux sujets divers en une seule enquête. Il faudrait à cette fin imposer un fardeau trop lourd aux répondants et compromette éventuellement la qualité des données. De retour au Comité directeur, j'ai suggéré qu'il considère la possibilité de diminuer la portée de l'enquête ou de procéder à une étude officielle de faisabilité pour mettre à l'essai les procédures de collecte avant de procéder à une enquête complète. Le Comité a répondu qu'il étudierait mes suggestions.

L'équipe a entrepris simultanément l'identification des besoins particuliers d'information des divers utilisateurs. J'ai communiqué avec les ministères de la Santé et de l'Agriculture et les ministères à vocation économique qui avaient demandé des renseignements pour définir leurs besoins d'information. J'ai aussi commencé à peaufiner l'ébauche du budget et de l'échéancier préparé pour la proposition d'enquête. À l'aide de l'expérience acquise et après consultation avec divers membres de l'équipe du projet, j'ai estimé les ressources nécessaires pour planifier, élaborer, mettre en œuvre et évaluer l'enquête. Les ressources nécessaires estimées étaient très élevées étant donné la portée de l'enquête. Celle-ci avait cependant un caractère permanent et l'on pourrait considérer que les coûts de planification et d'élaboration pourraient être amortis sur plusieurs années (même si les ressources étaient nécessaires immédiatement). Seuls les coûts de la mise en œuvre et de l'évaluation de l'enquête (et des modifications occasionnelles du plan) seraient réguliers. J'ai rencontré de nouveau les représentants des ministères intéressés pour les informer des coûts estimés et leur enthousiasme envers le projet a considérablement diminué.

De retour au Comité directeur au nom de l'équipe du projet, j'ai présenté l'ébauche du budget, de l'échéancier (Tableau 13.1) et de l'énoncé des objectifs. Le Comité directeur a décidé que la planification devrait continuer seulement pour les quatre principaux thèmes et qu'on pourrait communiquer de nouveau avec les autres ministères intéressés après le premier cycle de l'enquête. Le budget, l'échéancier et l'énoncé des objectifs en étaient seulement à l'étape d'une ébauche approximative, mais le Comité directeur a approuvé le budget et approuvé officiellement l'élaboration et le plan de l'enquête.

Ayant vécu l'expérience du processus de planification à de nombreuses occasions, je savais que l'énoncé des objectifs demanderait encore beaucoup de travail (ainsi que le budget et l'échéancier) avant d'entreprendre le plan d'enquête en soi. L'équipe du projet a donc continué d'élaborer l'énoncé des objectifs. J'ai rencontré les ministères intéressés pour les informer que l'EGM serait plus étroitement ciblée au cours du premier cycle et qu'ils seraient invités, au cours des années ultérieures, à proposer des ajouts à l'enquête.

Le Comité directeur a informé peu après l'équipe du projet que le budget était réduit de moitié et qu'elle devrait concentrer ses efforts sur l'un des principaux thèmes, tout en préparant une infrastructure d'enquête qui pourrait immédiatement prendre de l'expansion au cours de la deuxième année. Le Comité directeur et l'équipe du projet ont considéré les leçons apprises jusqu'à maintenant, décidé de cibler la composante du marché du travail et de reporter à l'an prochain les autres sujets de la liste. La situation était décevante pour les membres de l'équipe du projet, en particulier ceux qui avaient travaillé sur les thèmes retranchés), mais ils ont vite réalisé que la décision était la meilleure à long terme. Les membres auraient maintenant davantage de temps pour concentrer leurs efforts sur la production d'un bon produit aux fins de la collecte de données de qualité. La touche finale a rapidement été apportée à l'énoncé des objectifs qui ont été présentés au Comité directeur pour approbation.

L'équipe du projet a ensuite entrepris la rédaction d'une ébauche de rapport de planification, afin de présenter les options de base de sondage, de plan d'échantillonnage, de collecte des données, de traitement, d'analyse, de diffusion, etc., et d'en discuter. Ce rapport de planification a été présenté au Comité directeur et l'équipe du projet a demandé des conseils sur plusieurs questions, y compris la définition des populations cible et observée, la base d'échantillonnage à utiliser, les méthodes de collecte des données, etc. Le Comité directeur a donné peu de conseils d'importance et, en majeure partie, a demandé à l'équipe du projet de prendre ces décisions.

13.3 Conception et élaboration

Lorsque les principales décisions ont été prises, chaque membre de l'équipe a commencé à préparer des plans de composante pertinents à sa responsabilité dans l'équipe.

i. Contenu

M^{me} S a entrepris l'élaboration du questionnaire après avoir formulé les concepts et définitions pour l'énoncé des objectifs. Elle surveillait de nouvelles recrues et voulait bien entendu que le groupe acquière une expérience en milieu d'équipe de projet. Les experts en la matière de son équipe n'avaient aucune expérience précédente de l'enquête, mais ils étaient enthousiastes, même si leurs attentes étaient parfois peu réalistes. Il y a eu de nombreuses réunions de l'équipe du projet (et beaucoup de temps perdu) pour discuter des propositions des membres subalternes de l'équipe des spécialistes du contenu. Il aurait été plus efficace que ces experts discutent entre eux et, après avoir établi le consensus sur l'option préférée, qu'ils l'aient présentée à l'équipe du projet. J'en suis arrivé à suggérer cette procédure et la conception du questionnaire, ainsi que les étapes ultérieures, se sont déroulées beaucoup plus facilement et rapidement.

ii. Méthodologie

Le méthodologiste d'enquête n'avait pas d'expérience, mais il a achevé son travail rapidement et efficacement. Il était aussi chargé de la conception des systèmes de contrôle qualitatif pour la saisie et le codage des données. M. M. n'était pas disposé au départ à faire des compromis sur le nombre d'inspections de contrôle qualitatif nécessaires. Il soutenait que les taux d'erreur des sorties pour la saisie et le codage des données devraient être près de 0 %. J'ai pu le convaincre après plusieurs discussions de la nécessité d'un compromis entre le coût et la qualité.

iii. Programmation

L'analyste des systèmes a obtenu toutes les spécifications à temps, mais les systèmes n'étaient pas prêts à la mise à l'essai au moment opportun. Je ne sais toujours pas quel a été le problème exactement. Le programmeur avait une autre charge de travail trop lourde, à mon avis, et il n'a pu répartir son temps avec efficacité. La situation a eu des répercussions importantes sur l'étape de la mise en œuvre (considérée à la section suivante).

iv. Opérations et collecte des données

L'agent des opérations et de la collecte des données a efficacement pris en charge le recrutement, la formation, la surveillance et le contrôle du personnel de la collecte des données, notamment les intervieweurs et les surveillants, ainsi que les opérateurs de la saisie et les codeurs des données. Tous les manuels ont été produits à temps et ils étaient d'excellente qualité. L'agent des opérations et de la collecte des données a visité plusieurs bureaux de la statistique régionale pour observer le recrutement et la formation.

Au cours de l'étape de la conception et de l'élaboration de l'enquête, les réunions de l'équipe du projet ont été moins fréquentes pour donner davantage de temps aux membres de travailler à leurs plans de composante et pour examiner les plans des autres membres de l'équipe. Le Comité directeur a été informé de toutes les décisions de l'équipe du projet, mais la rétroaction a été rare.

13.4 Mise en œuvre

Au cours de la mise en œuvre, la cible est passée de la prise de décisions à l'action. Le questionnaire a été imprimé, l'échantillon a été sélectionné, la formation s'est déroulée dans les régions, les grappes ont été listées et les interviews ont eu lieu. Au retour des données aux bureaux, elles ont été saisies, codées, vérifiées et imputées. L'estimation et l'analyse ont suivi. Les réunions de l'équipe du projet ont été plus fréquentes pour considérer tous les problèmes remarqués pendant la mise en œuvre.

Il y a eu plusieurs problèmes pendant l'étape de la mise en œuvre.

i. Le taux de réponse obtenu était inférieur à celui que nous avions prévu. Le taux de réponse prévu était de 80 %, mais le taux réel a atteint 68 % seulement. Ce résultat a suscité deux préoccupations.

   a. Premièrement, le taux élevé de non-réponse pourrait ajouter un biais aux résultats de l'enquête. Une étude de suivi des non-réponses est prévue pour examiner la question et les caractéristiques des non-répondants, comparativement à celles des répondants. Cette étude servira à planifier les procédures de suivi des non-réponses pour les enquêtes ultérieures.

b. Deuxièmement, la précision des estimations régionales était inférieure à la cible établie à l'étape du choix de la taille de l'échantillon. Il a donc fallu calculer de nouveau la précision atteignable, compte tenu du taux de réponse, et présenter le résultat au Comité directeur et aux utilisateurs. Il a été décidé d'affecter davantage de ressources au suivi des non-réponses et à la conversion des refus en réponses pour essayer d'augmenter le taux de réponse et, si nécessaire, des données plus agrégées seraient diffusées.

ii. Étant donné la livraison tardive des applications informatiques (mentionnées ci-dessus), en particulier pour la saisie des données, un groupe de commis n'avait à rien faire pendant que l'équipe du projet faisait l'essai des programmes informatiques. Conscient que nous avions des commis assis à ne rien faire (mais quand même rémunérés), je dois admettre que j'ai exercé des pressions sur l'équipe du projet pour qu'elle accélère la mise à l'essai des applications. Il y avait donc malheureusement toujours des erreurs lorsqu'elles ont finalement été mises en œuvre. Les délais et les interruptions qui ont découlé des erreurs critiques du programme ont probablement coûté beaucoup plus de temps et d'argent que si nous avions réservé suffisamment de temps à la mise à l'essai du programme avant son application.

iii. Il est devenu évident à l'examen des commentaires à la section « Autre (veuillez préciser) » de la Question E3 que les répondants ou les intervieweurs n'ont pas compris les catégories de réponse. Ces commentaires ayant fait l'objet d'un examen pendant la collecte, il a été possible d'envoyer des instructions aux intervieweurs pour préciser les questions et les catégories de réponse, afin d'obtenir des données exactes par la suite. Dans le cas des questionnaires déjà reçus, les experts en la matière ont examiné les données de la question et changé les codes au besoin.

13.5 Évaluation

J'ai plusieurs observations à faire sur l'évaluation de la planification et de la gestion de l'EGM :

i. Il est assez rapidement devenu évident que le Comité directeur n'était pas suffisamment engagé dans le processus de prise de décisions. Il semblait satisfait de laisser l'équipe du projet prendre toutes les décisions, même celles qui avaient de grandes conséquences. J'aurais dû exprimer ma préoccupation à ce sujet au Comité directeur. J'en ai plutôt discuté avec l'équipe du projet, mais certains membres n'avaient pas suffisamment d'expérience pour faire des commentaires utiles.

ii. En rétrospective, j'aurais pu intervenir différemment pour régler le problème de la livraison tardive du système de saisie des données. J'aurais dû surveiller de plus près l'échéancier pour reconnaître le problème question plus tôt. Après en avoir constaté l'existence, j'aurais pu intervenir immédiatement pour remplacer le programmeur ou obtenir des ressources de programmation supplémentaires. J'ai hésité à ce moment-là, étant d'avis qu'il y aurait de nouveau trop de temps perdu pour apprendre aux nouveaux programmeurs les spécifications et les programmes en place. J'aurais aussi dû donner suffisamment de temps pour la mise à l'essai appropriée.

iii. L'échéancier original était trop serré. Nous aurions dû attribuer davantage de temps entre les produits à livrer pour éviter les problèmes. Étant donné le manque d'expérience et la lourde charge de travail de plusieurs membres de l'équipe du projet, j'aurais dû établir un échéancier plus conventionnel. D'autre part, j'aurais pu maintenir l'échéancier constant et affecter davantage de personnes à certaines des principales tâches, notamment la programmation.

iv. Les non-réponses sont un problème de l'EGM qu'il faut régler. Il faudra évaluer la possibilité de biais de non-réponse (en particulier dans certains sous-groupes de la population) dans l'EGM actuel. Aux fins des éditions ultérieures de l'EGM, je recommanderais l'application de procédures améliorées de suivi des non-réponses pour obtenir un taux de réponse élevé. Nous avons actuellement des données très agrégées seulement à la disposition des utilisateurs. Elles seraient beaucoup plus utiles si elles pouvaient être plus détaillées.

v. Un grand nombre de migrants récents occupaient des logements temporaires. La situation peut causer certains problèmes si les mêmes logements sont inclus dans plus d'une phase de l'enquête.

Tableau 13.1 : Échéancier de l'EGM

Étapes	Responsable	Début	Fin
Énoncé des objectifs	Direction, matière	1er janvier	31 janvier
Plan d'enquête	Tous	1er février	28 février
Budget	Direction	1er janvier	31 janvier
Conception du questionnaire	Matière	1er mars	31 mars
Plan d'analyse	Matière	1er avril	30 avril
Spécifications – Plan d'échantillonnage	Méthodologie	1er mars	30 avril
Spécifications – Méthode de collecte des données	Collecte des données	1er mars	31 mars
Spécifications – Vérification et imputation	Méthodologie, matière	1er avril	30 avril
Spécifications – Contrôle qualitatif	Méthodologie	1er mai	31 mai
Spécifications – Saisie des données	Collecte des données	1er avril	30 avril
Spécifications – Estimation, variance	Méthodologie	1er juin	30 juin
Élaboration du système informatique	Programmation	1er juin	15 août
Mise à l'essai du système informatique	Tous	15 juillet	31 août
Manuels sur place	Collecte des données	1er juin	31 juillet
Relations publiques	Direction, Collecte des données	1er août	31 août
Formation des intervieweurs	Collecte des données	15 août	31 août
Listage	Collecte des données	1er septembre	12 septembre
Collecte des données	Collecte des données	15 septembre	30 septembre
Formation – Codage	Collecte des données	26 septembre	30 septembre
Codage	Collecte des données	1er octobre	14 octobre
Formation – Saisie des données	Collecte des données	10 octobre	14 octobre
Saisie des données	Collecte des données	15 octobre	31 octobre
Vérification et imputation	Collecte des données	1er novembre	14 novembre
Estimation	Méthodologie, matière	15 novembre	30 novembre
Analyse	Matière	1er décembre	14 décembre
Évaluation de la qualité des données	Méthodologie, matière	1er décembre	25 décembre
Contrôle de la divulgation	Méthodologie	15 décembre	25 décembre
Diffusion	Direction, matière	31 décembre	31 décembre
Documentation	Tous	1er janvier	31 décembre

PUBLICATIONS ÉLECTRONIQUES
DISPONIBLES À

www.statcan.gc.ca

Index

- **Aberrantes (données) :** 10.5

- **Analyse de données:** 1.1.9 , 11.0
  - **Mesures de récapitulation:** 11.3
    - **Données d'enquête simple**: 11.3.1
      - **dispersion (variance de population et d'échantillonnage, étendue, intervalle interquartile):** 11.3.1.3, 7.3.1.
      - **emplacement (moyenne, médiane, mode, quartiles, centiles)**: 11.3.1.2
      - **estimation et présentation des distribution de fréquences**: 11.3.1.1
        - **diagrammes et schémas:** 11.3.1.1.1, 12.2.2, 12.2.3 .
    - **Données d'enquête complexe**: 11.3.2
      - **médiane:** 11.3.2.1
      - **dispersion (variance de population et d'échantillonnage, étendue, intervalle interquartile):** 11.3.2.2
  - **Rapport d'analyse**: 12.3
  - **Tests d'hypothèse pour données continues**:11.4
    - **Données d'enquête complexe**: 11.4.3
      - **plusieurs moyennes (ANOVA et régression):** 11.4.3.2
      - **une moyenne**: 11.4.3.1
    - **Données d'enquête simple**: 11.4.2
      - **deux moyennes**: 11.4.2.2
      - **plusieurs moyennes**: 11.4.2.3
        - **analyse de variance (ANOVA):** 11.4.2.3.1
        - **régression linéaire**: 11.4.2.3, 11.4.2.3.2
      - **une moyenne:** 11.4.2.1
  - **Tests d'hypothèse pour données discrètes**: 11.5
    - **Données d'enquête complexe**: 11.5.2
      - **tests d'indépendance et d'homogénéité**: 11.5.2
    - **Données d'enquête simple:** 11.5.1
      - **modèles log-linéaires**: 11.5.1.3.
      - **test d'homogénéité**: 11.5.1.2
      - **test d'indépendance**: 11.5.1.1

- **Analyse de variance (ANOVA):** 11.4.2.3.1

- **Autodénombrement:** 4.1.1

- **Base aréolaire:** 3.3.1.2

- **Base de liste:** 3.3.1.1

- **Base de sondage**: 3.3
  - **Base aréolaire:** 3.3.1.2, 9.3.3
  - **Base de liste:** 3.3.1.1
  - **Bases multiples**: 3.3.1.3

  - **Défauts d'une base de sondage:** 3.3.2
  - **Qualités d'une bonne base de sondage:** 3.3.3

- **Bases multiples:** 3.3.1.3

- **Biais**: 6.2.1, 7.3.1, 11.3.2.2.1

- **Bootstrap**: 7.3.4

- **Calibration et régression généralisée :** 7.1.4.3

- **CAPI/CASI/CATI:** 4.2.

- **Codage:** 10.1

- **Coefficient de variation (CV):** 7.3.2.1

- **Collecte assistée par interviewer:** 4.1.2

- **Collecte des données**: 1.1.5; 4.0
  - **Assistée par intervieweur**: 4.1, 4.1.2
    - **Par téléphone**: 4.1, 4.1.2.2
    - **Sur place**: 4.1, 4.1.2.1
  - **Assistée par ordinateur:** 4.1; détails 4.2
    - **Au téléphone (CATI):** 4.1
    - **Autodénombrement (CASI):** 4.1
    - **En personne (CAPI):** 4.1
  - **Autodénombrement:** 4.1, détails 4.1.1
  - **Autres méthodes de collecte**:
    - **Déclaration électronique de données (DED):** 4.3.2
    - **Données administratives:** 4.3.3, Annexe A
    - **Enquêtes omnibus et supplémentaires**: 4.3.5
    - **Méthodes combinées**: 4.3.4
    - **Observation directe**: 4.3.1
  - **Comparaison des méthodes de collecte**: tableau 2 du Chapitre 4.
  - **Papier**: 4.1
    - **Entrevue "papier crayon" (PAPI):** 4.1
  - **Opérations de collecte:** Chapitre 9
    - **Déroulement des interviews**: 9.4
      - **approfondissement:** 9.4.4
      - **conclusion de l'interview**: 9.4.5
      - **interview efficace**: 9.4.8
      - **préparation des interviews et établissement de l'horaire**: 9.4.1
      - **procuration / substitut**: 9.4.7
      - **refus et autres situations délicates**: 9.4.6
      - **techniques de présentation**: 9.4.2
      - **utilisation du questionnaire:** 9.4.3
    - **Organisation de la collecte**: 9.1
    - **Préparation des procédures de collecte**: 9.3

                - **dépistage:** 9.3.4
                - **embauche et formation des interviewers**: 9.3.2
                - **listage:** 9.3.3,3.3.1.2
                - **manuel des interviewers:** 9.3.1.1
                - **manuel du superviseur:** 9.3.1.2
            - **Relations avec les répondants**: 9.2.2
            - **Relations publiques**: 9.2
            - **Surveillance de la qualité et du rendement**: 9.5.1
                - **gestion des tâches d'intervieweurs**: 9.5.2
                - **surveillance des surveillants**: 9.5.3

- **Composition aléatoire de numéros (RDD):** 4.1.2.2.1

- **Confidentialité:** 5.1.3, 5.5.1, 9.2.2; 9.3.1.1; 9.4; 9.4.2; 12.5.2

- **Contrôle d'acceptation:** Annexe B : 2.4

- **Contrôle statistique du processus**: Annexe B

- **Contrôle statistique du produit** : Annexe B

- **Cycle de vie d'une enquêtes:** 1.2
    - **Conception**:1.2.2
    - **Évaluation**: 1.2.4
    - **Mise en oeuvre:**1.2.3
    - **Planification**: 1.2.1, 13.1

- **Dépistage:** 9.3.4

- **Diffusion des données**: 1.1.10 et 12.1 (détails au Chapitre 12)
    - **Divulgation:** 1.1.10 et 12.5.1
        - **Protection des fichiers à grande diffusion**: 12.5.2.2
            - **identification des enregistrements délicats**: 12.5.2.2
            - **traitement des enregistrements délicats (réduction / perturbation des données)**: 12.5.2.2
        - **Protection des tableaux**: 12.5.2.1
            - **identification des cases délicates (limite / règle (n,k) / règle p-pour cent):** 12.5.2.1
            - **traitement des cases délicates (réduction / perturbation des données)**: 12.5.2.1
    - **Rapport d'analyse des données**: 12.3
    - **Rapport principal:** 12.2
        - **Lignes directrices sur la rédaction**: 12.2.1
    - **Rapport sur la qualité des données**: 12.4
    - **Tableaux / graphiques:** 12.2.2, 12.2.3, 11.3.1.1.1

- **Divulgation (contrôle):** 12.5.1

- **Documentation:** 1.1.11, 12.2

- **Données administratives**: 4.3.3 et Annexe A;
  - **Sources de données administratives**: Annexe A
  - **Utilisation des données administratives**: Annexe A
  - **Utilité des données administratives**: Annexe A

- **Échantillonnage:** 6.0
-
- **Échantillonnage aléatoire simple :** 6.2.2

- **Échantillonnage à plusieurs degrés:** 6.2.7

- **Échantillonnage à plusieurs phases:** 6.2.8

- **Échantillonnage avec probabilité proportionnelle à la taille (PPT):** 6.2.4

- **Échantillonnage d'acceptation:** Annexe B : 2.1.1

- **Échantillonnage double / Échantillonnage à deux phases**: 6.2.8

- **Échantillonnage non probabiliste:** 6.1

- **Échantillonnage par grappes:** 6.2.5

- **Échantillonnage par quota** 6.1.4.

- **Échantillonnage par répliques:** 6.2.9

- **Échantillonnage probabiliste:** 6.2

- **Échantillonnage stratifié:** 6.2.6

- **Échantillonnage systématique :** 6.2.3

- **Effet de plan (*deff*):** 7.3.3

- **Enquêtes longitudinales:** 6.3.1

- **Erreurs dans une enquête:**
  - **Erreur d'échantillonnage**: 3.1, 7.3
  - **Erreurs non dues à l'échantillonnage**: 3.1
    - **Couverture:** 3.4.2.1
    - **Erreurs systématiques / aléatoires**: 3.4.2
    - **Mesure / réponse:** 3.4.2.2; 5.4
      - **sources de l'erreur de réponse**: 5.4.1
      - **techniques de réduction des erreurs de réponse**: 5.4.2
    - **Non-réponse**: 3.4.2.3
    - **Traitement:** 3.4.2.4

- **Erreur d'échantillonnage :** 4.1, 7.3

- **Erreurs non dues à l'échantillonnage:** 3.1

- **Erreur quadratique moyenne (EQM, MSE)**: 11.3.2.2.1.

- **Estimateur:** 6.2.1,7 et 11.1
  - **Distribution d'échantillonnage d'un estimateur**: 6.2.1 et 11.1.
  - **Estimateur exact:** 6.2.1
  - **Estimateur précis**: 6.2.1, 7.3
  - **Estimateur robuste :** 10.5.2
  - **Estimateur sans biais**: 6.2.1

- **Estimation:** 1.1.8; 7.0
  - **Estimation d'un total:** 7.2.1
  - **Estimation d'une moyenne:** 7.2.1
  - **Estimation d'une proportion**: 7.2.1
  - **Estimation de la variance de la population:** 7.3.1, 7.3.1.1 et 7.3.2.3
  - **Estimation de la variance d'échantillonnage**: 3.4.1; 7.3.1, 8.1.3, 10.4.4, 11.3.1.3, 11.3.2.2
    - **estimation d'un coefficient of variation (*CV*):** 7.3.2.1
    - **estimation d'intervalles de confiance:** 7.3.2.2
      - en présence de biais: 11.3.2.2.1
    - **estimation d'un effet de plan (*deff*):** 7.3.3 , Chapitre 8 et 11.5.2
    - **estimation de la marge d'erreur**: 7.3.2.2.
    - **estimation de l'erreur type:** 7.3.2.1; 7.3.2.4
    - **estimation de la variance par répliques / jackknife / bootstrap:** 7.3.4;
    - **facteurs affectant la précision**: 3.4.1 et 8.1.2
  - **Estimation par le quotient / ratio:** 7.1.4.2
  - **Estimation pour petits domaines**: 7.2.3.1

- **Étapes d'une enquêtes:** 1.1

- **Exactitude**: 6.2.1

- **Famille:** 2.1.3

- **Gestion d'une enquête** 1.2.1, Chapitre 13

- **Graphique de contrôle:** Annexe B : 2.2.1

- **Groupe de discussion (focus group):** 5.1.5.3

- **Khi carré** : 11.4.1, 11.5.1.1, 11.5.1.2, 11.5.1.3, 11.5.2.

- **Imputation:** 10.4

- **Inférence:** 11.4.

- **Interview:** 4.1.2 **,** 9.4

- **Jackknife:** 7.3.4

- **Listage:** 9.3.3

- **Logement:** 2.1.3

- **Ménage:** 2.1.3

- **Modèles log-linéaires**: 11.5.1.3.
- **Objectifs d'une enquête :** Chapitre 2

- **Paramètre :** .2.1 et 11.1

- **Plans d'échantillonnage**: 1.1.3; Chapitre 6
    - **Échantillonnage non probabiliste**: 6.1
        - **À l'aveuglette:** 6.1.1
        - **Au jugé**: 6.1.3
        - **Boule de neige / réseau**: 6.3.3
        - **Probabiliste modifié**: 6.1.5
        - **Quota**: 6.1.4
        - **Volontaires**: 6.1.2
    - **Échantillonnage probabiliste**: 6.2
        - **Aléatoire simple:** 6.2.2
        - **À plusieurs degrés**: 6.2.7
        - **À plusieurs phases:** 6.2.8
        - **Avec probabilité proportionnelle à la taille:** 6.2.4
        - **Efficacité comparée**: 6.2.1
        - **En grappes**: 6.2.5
        - **Entrée / sortie:** 6.3.2
        - **Par répliques:** 6.2.9.
        - **Répété / longitudinal:** 6.3.1
        - **Stratifié**: 6.2.6
        - **Systématique:** 6.2.3

- **Planification d'une enquête** 1.2.1, Chapitre 13.

- **Poids**
    - **Ajustement pour non-réponse**: 7.1.3
    - **Information auxiliaire**: 7.1.4
        - **Calibration et régression généralisée**: 7.1.4.3
        - **Estimation par le ratio:** 7.1.4.2
        - **Stratification *a posteriori***: 7.1.4.1
    - **Poids du plan**: 7.1
        - **Pour un plan avec probabilités inégales**: 7.1.2
        - **Pour un plan équiprobable**: 7.1.1

- **Population cible:** 3.2

- **Population d'enquête :** 3.2

- **Précision:** 6.2.1

- **Qualité:** Annexe B
    - **Assurance de qualité**: Annexe B : 3.0
    - **Contrôle de qualité**: Annexe B

        - **Contrôle statistique du produit**: Annexe B: 2.1; 2.3
            - **échantillonnage d'acceptation**: Annexe B 2.1.1
        - **Contrôle Statistique du processus**: Annexe B: 2.2; 2.3
            - **graphique de contrôle:** Annexe B 2.2.1
    - **Contrôle d'acceptation**: Annexe B 2.4

- **Questionnaire:** 1.1.4; 5.0
    - **Conception**: 5.1
    - **Erreur de réponse:** 5.4
    - **Formulation des questions**: 5.3
    - **Mise à l'essai**: 5.1.5
        - **Compte rendu d'interviewers**: 5.1.5.4
        - **Échantillons fractionnés**: 5.1.5.6
        - **Enquêtes pilotes :** 5.1.5.7
        - **Groupes de discussion**: 5.1.5.3
        - **Méthodes cognitives**: 5.1.5.2
        - **Pré-test:** 5.1.5.1
    - **Présentation**: 5.5
    - **Types de questions**
        - **Questions fermées**: 5.2, 10.1.2
            - **à choix multiples**: 5.2.2
            - **dichotomiques**: 5.2.1
            - **échelles:** 5.2.3, 5.2.4
        - **Questions ouvertes**: 5.2 , 10.1.1

- **Recensement**: 6.1

- **Régression linéaire**: 11.4.2.3

- **Règle (n,k)**: 12.5.2.1.

- **Règle p-pourcent:** 12.5.2.1

- **Répartition**
    - **À CV égaux** 8.2.1.2
    - **De Neyman**: 8.2.2.2.5
    - **Égale** 8.2.1.1
    - **En puissance**: 8.2.2.2.3
    - **Optimale** 8.2.2.2.4
    - **Proportionnelle (à N)** 8.2.2.1
    - **Proportionnelle à Y, à** $\sqrt{N}$ **, à** $\sqrt{Y}$ 8.2.2.2

- **Saisie des données**: 10.2

- **Stratification *a posteriori*** 7.1.4.1

- **Tableaux de contingence**: 11.5

- **Taille d'échantillon**: chapitre 8
    - **Exigences de précision** : 8.1.1

  - **Formules**: 8.1.3

- **Téléphone (RDD):** 4.1.2.2.1

- **Test d'hypothèse**: 11.4

- **Traitement:** 3.4.2.4 et 10.0;
  - **Codage:** 1.1.6 et 10.1; 3.4.2.4
    - **Questions fermées** 10.1.1
    - **Questions ouvertes** 10.1.3
  - **Données aberrantes**: 7.2.3.2 et 10.5
    - **Estimateurs robustes**: 10.5.2
    - **Identification** : 10.5.1
    - **Traitement** : 10.5.2
  - **Imputation:** 1.1.7 et 10.4 (détails in 10.4); 3.4.2.4
    - **Cold-deck**: 10.4.1.5
    - **Déterministe avec résidus aléatoires**: 10.4.1.7
    - **Estimation de la variance sous imputation**: 10.4.4
    - **Évaluation** 10.4.6
    - **Hot-deck**: 10.4.1.4
    - **Lignes directrices** 10.4.5
    - **Par déduction**: 10.4.1.1
    - **Par la moyenne**: 10.4.1.2
    - **Par le plus proche voisin**: 10.4.1.6
    - **Par ratio/régression**: 10.4.1.3
    - **Par donneur**: 10.4.3
  - **Saisie**: 1.1.5, 3.4.2.4 et 10.2
  - **Vérification**: 1.1.7 , 3.4.2.4 et 10.3

- **Unités:**
  - **Unité d'échantillonnage, de référence, d'enquête** 3.3

- **Vérification:** 10.3