

Notes de cours

Échantillonnage STT2000

David Haziza
Département de mathématiques et de statistique
Université de Montréal

Automne 2011

PRÉFACE

Ces notes de cours ont été rédigées pour le cours STT-2000 (Échantillonnage) donné à l'Université de Montréal. Elles constituent une première introduction à la théorie de l'échantillonnage. Les sujets abordés ont été choisis avec un regard sur les aspects pratiques auxquels fera face un statisticien d'enquête, compte tenu des outils mathématiques et statistiques dont on dispose en début de deuxième année. Le traitement du sujet est similaire à celui de l'excellent ouvrage « Sampling : Design and Analysis » de Sharon L. Lohr (2010). Plusieurs exemples utilisés sont d'ailleurs empruntés à ce livre.

Afin de permettre aux étudiants de ne pas s'assoupir pendant les cours, j'ai omis de taper les preuves de résultats. Les étudiants devront donc 'remplir les trous' au stylo.

Je tiens à remercier Caroline Pelletier de Statistique Canada pour ses commentaires et suggestions qui ont grandement contribué à améliorer le contenu de ces notes.

David Haziza
Montréal, le 5 Août 2011

L'univers des enquêtes

Chapitre 1

1.1 Introduction

La place des enquêtes (ou sondages) dans notre société est sans cesse croissante. La grande majorité des citoyens est familière avec le concept de sondage. En effet, qui n'a pas été sélectionné, au moins une fois, pour participer à un sondage d'opinion durant une campagne électorale? L'emploi des sondages n'est toutefois pas limité au cas des enquêtes d'opinion. L'utilisation des enquêtes est répandue dans divers domaines : par exemple, la recherche de gisement pétrolier et la recherche minière donnent lieu à sondage; les contrôles fiscaux des contribuables se font généralement par sondage, etc. Nous commençons par définir des concepts de base qui nous serviront de point de départ dans ce cours. Les sections 1.1 à 1.3 sont largement inspirées de l'excellent ouvrage « Méthodes et pratiques d'enquête » publié par Statistique Canada.

Définition 1.1 : 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.

Définition 1.2 : Un *élément* est un objet sur lequel on va mesurer les caractéristiques d'intérêt.

Définition 1.3 : Une *population* est un ensemble d'éléments.

Définition 1.4 : Une *population cible* est une population pour laquelle l'information est requise.

Définition 1.5 : Les *unités d'échantillonnage* sont des entités disjointes dont l'union est égale à la population.

Définition 1.6 : La *base de sondage* donne les moyens d'identifier les unités d'échantillonnage de la population cible et de communiquer avec elles. C'est à partir de la base de sondage que l'on va sélectionner les unités d'échantillonnage qui vont être enquêtées.

Définition 1.7 : Un *échantillon* est un ensemble d'unités d'échantillonnage sélectionné à partir de la base de sondage.

Pour mieux saisir la distinction entre la notion d'élément et celle d'unité d'échantillonnage, considérons les exemples suivants : (1) on désire mener une enquête auprès des entreprises de restauration canadiennes afin d'estimer leur revenu total pour

une année donnée. On dispose d'une base de sondage comprenant toutes les entreprises de restauration au Canada. On décide donc de tirer un échantillon aléatoire d'entreprises et de les contacter par téléphone afin de recueillir les caractéristiques d'intérêt. Dans cet exemple, l'unité d'échantillonnage et l'élément référent à la même entité : l'entreprise.

(2) On désire mener une enquête auprès des infirmières travaillant en milieu hospitalier dans la province de Québec. Malheureusement, on ne dispose pas d'une base de sondage d'infirmières et créer une telle base prendrait beaucoup trop de temps. Par contre, une base de sondage comprenant tous les hôpitaux dans la province de Québec est à notre disposition. On décide donc de tirer un échantillon aléatoire d'hôpitaux et dans chaque hôpital sélectionné, d'interviewer toutes les infirmières y travaillant. Dans cet exemple, l'unité d'échantillonnage est un hôpital et l'élément est une infirmière.

On distingue deux types d'enquête : l'enquête-échantillon et le recensement. Le recensement cible la collecte de renseignements pour toutes les unités de la population alors que l'enquête-échantillon retient à cette fin une partie seulement (habituellement petite) des unités de la population.

1.2 Les étapes d'une enquête

Une enquête comporte plusieurs étapes. Dans ce qui suit, nous décrivons les plus importantes.

1.2.1 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. Un énoncé clair des objectifs oriente toutes les étapes ultérieures de l'enquête.

Par exemple, il faut définir le plus précisément possible la population cible. Il faut également préciser la couverture de la population : quels secteurs géographiques nous intéressent? La période de référence est également importante. Habituellement, cette dernière peut faire référence à la semaine dernière, le mois dernier ou l'année fiscale dernière. De plus, il faut préciser quels seront les principaux utilisateurs des données parce que leur rétroaction est très importante pendant la phase de planification.

1.2.2 Sélection de la base de sondage

La base de sondage donne les moyens d'identifier les unités de la population d'enquête et de communiquer avec elles. Il s'agit donc de choisir, si possible, une base de sondage qui est la plus similaire à la population cible. La base de sondage devrait comprendre les renseignements suivants, en tout ou en partie :

- (i) *Données d'identification* : ce sont les renseignements de la base de sondage qui identifient sans ambiguïté chaque unité de l'échantillon, par exemple, le nom et un numéro d'identification unique.
- (ii) *Données de communication* : ce 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* : ce sont les données servant à l'échantillonnage et, éventuellement, à l'estimation. Les données de classification comprennent une classification géographique (par exemple, province, division de recensement), classification type des professions ou des industries (par exemple, le Système de Classification des Industries de l'Amérique du Nord, SCIAN).

On distingue deux types de base de sondage : les bases listes et les base aréolaires.

- (i) *Base liste* : Une base liste peut être définie comme une liste conceptuelle ou physique de toutes les unités de la population d'enquête. Voici des exemples de bases listes :

- 1.2.3 registre des statistiques de l'état civil (par exemple, une liste de toutes les naissances ou de tous les décès dans la population);
- 1.2.4 registre des entreprises (liste de toutes les entreprises en opération) ;
- 1.2.5 registre des adresses (liste de logements avec adresses civiques);
- 1.2.6 annuaire téléphonique.

- (ii) *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 sont habituellement composées d'une hiérarchie d'unités géographiques. Des unités de la base de sondage à un niveau donné 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 être également divisés en plus petits secteur, par exemple, les îlots d'une ville.

Si chaque élément de la base de sondage correspond à un et un seul élément dans la population cible, on dira que la base de sondage est parfaite. Autrement, la base est défectueuse, ce qui aura vraisemblablement un impact sur la qualité des estimations. En pratique, une base parfaite n'existe pas. Il y a toujours des imperfections dont nous décrivons maintenant la nature :

- (i) *Sous-couverture* : La sous-couverture (ou sous-dénombrement) est le résultat de l'exclusion de la base de sondage de certaines unités qui font partie de la population cible. Par exemple, les nouvelles entreprises font partie de la population cible mais sont absentes de la base de sondage (registre des entreprises) car celle-ci n'a pas été mise à jour depuis la création de ces

entreprises. Un autre exemple survient lorsque l'annuaire téléphonique est utilisé comme base de sondage. Les ménages ayant un numéro de téléphone confidentiel font partie de la population cible mais sont absents de la base de sondage. (Voir Figure 1.1).

- (ii) *Sur-couverture* : La sur-couverture (ou sur-dénombrement) est le résultat de l'ajout à la base de sondage de certaines unités qui ne font pas partie de la population cible. Par exemple, les entreprises récemment « mortes » sont présentes sur la base de sondage mais ne font plus partie de la population cible. (Voir Figure 1.1).
- (iii) *Répétition* : Il y a répétition lorsque la même unité paraît plus d'une fois dans la base de sondage. Dans le registre des entreprises, par exemple, la même entreprise peut être énumérée une fois sous sa raison sociale et une fois sous son nom commercial.
- (iv) *Classification erronée* : Les erreurs de classification sont des valeurs inexactes attribuées à des variables de la base de sondage. Par exemple, un homme est inscrit par erreur à la catégorie femme ou encore une entreprise de détail est classée grossiste.

1.2.3 Sélection du plan de sondage

Il existe deux types d'échantillonnage: l'échantillonnage probabiliste et l'échantillonnage non-probabiliste. Dans le cas de l'échantillonnage non-probabiliste, une méthode subjective de sélection est appliquée à une population. C'est un moyen rapide, facile et bon marché de sélectionner un échantillon. Dans le cas de l'échantillonnage probabiliste, la sélection des unités est aléatoire. La probabilité d'inclusion dans l'échantillon pour chaque unité est donc connue.

Il existe plusieurs types d'échantillonnage non-probabiliste. En voici quelques-uns :

- (i) *Échantillonnage à participation volontaire* : Cette méthode fait appel à des répondants volontaires. Par exemple, 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 leur opinion. Seuls ceux que le sujet intéresse vraiment d'une façon ou d'une autre ont tendance à répondre, ce qui générera habituellement un *biais de sélection* marqué. Il y a présence de biais de sélection lorsque les unités sélectionnées dans l'échantillon possèdent des caractéristiques différentes des unités qui n'ont pas été sélectionnées.
- (ii) *Échantillonnage au jugé* : À l'aide de cette méthode, l'échantillonnage est effectué 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 décide de ce qui est

considéré un échantillon représentatif. Le biais de l'expert peut entacher les estimations si ses idées préconçues sont inexactes.

- (iii) *Échantillonnage par quotas* : C'est une des méthodes d'échantillonnage non-probabiliste les plus communes. L'échantillonnage est effectué jusqu'à ce qu'un nombre déterminé d'unités (quotas) soient sélectionnées dans diverses sous-populations. Par exemple, si la population comprend 100 hommes et 100 femmes et qu'il faille tirer un échantillon de taille 20, 10 hommes et 10 femmes seront interviewés. L'échantillonnage par quotas peut être considéré préférable à d'autres formes d'échantillonnage non-probabiliste parce qu'il faut inclure des membres de sous-population différentes.

Dans ce cours, seul l'échantillonnage probabiliste sera considéré. La grande majorité des grands organismes de statistique tels que Statistique Canada, utilisent l'échantillonnage probabiliste. Il existe plusieurs types d'échantillonnage probabiliste : l'échantillonnage aléatoire simple avec/sans remise, l'échantillonnage stratifié, l'échantillonnage proportionnel à la taille, l'échantillonnage systématique, l'échantillonnage à plusieurs degrés, etc. Ces méthodes feront l'objet de plusieurs chapitres du cours.

1.2.4 Conception du questionnaire

Un questionnaire est un groupe ou une séquence de questions conçues afin d'obtenir d'un répondant de l'information sur un sujet. Les questionnaires ont un rôle central dans les enquêtes car ils ont des répercussions importantes sur la qualité 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. Le processus de conception du questionnaire commence par la formulation des objectifs de l'enquête et des besoins d'information et continue avec les étapes suivantes :

- (i) consultation avec les utilisateurs des données et les répondants;
- (ii) examen des questionnaires précédents;
- (iii) élaboration d'une version provisoire du questionnaire;
- (iv) examen et révision du questionnaire;
- (v) mise à l'essai et révision du questionnaire;

(vi) touche finale apportée au questionnaire.

Un questionnaire bien conçu devrait :

- (i) permettre la collecte des données avec efficacité et le résultat devrait comprendre un nombre minimal d'erreurs et de données incohérentes;
- (ii) être convivial pour l'intervieweur et le répondant;
- (iii) diminuer dans l'ensemble le coût et le temps de la collecte des données.

1.2.5 Collecte des données

La collecte des données est le processus qui permet d'obtenir l'information nécessaire pour chaque unité sélectionnée de l'enquête. Les méthodes élémentaires de collecte des données sont présentées ci-dessous :

- (i) *Interview sur place* : Un intervieweur aide le répondant à remplir le questionnaire. 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. La méthode sur support papier est appelée interview papier et crayon (IPC) et la méthode assistée par ordinateur est appelée interview sur place assistée par ordinateur (IPAO).
- (ii) *Interview téléphonique* : Un intervieweur aide le répondant à remplir le questionnaire au téléphone. La méthode sur support papier est appelée interview papier et crayon (IPC) et la méthode assistée par ordinateur est appelée interview téléphonique assistée par ordinateur (ITAO).
- (iii) *Autodénombrement* : Le répondant remplit le questionnaire d'enquête par autodénombrement sans l'aide d'un intervieweur (par exemple, le recensement canadien). Divers moyens peuvent servir à envoyer le questionnaire au répondant et à le retourner à l'expéditeur : le service postal, internet, etc. La méthode sur support papier est appelée interview papier et crayon (IPC) et la méthode électronique est appelée auto-interview assistée par ordinateur (AIAO).
- (iv) *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 (par exemple, des données médicales ou étude sur des populations animales).

Le tableau 1.1 donne une comparaison des différentes méthodes de collecte en termes de coûts, de temps requis pour remplir le questionnaire et du taux de réponse habituellement observé.

| | Autodénombrement | Interview | |
|------------------------|-------------------------|------------------|---------------------|
| | | Sur place | Téléphonique |
| Coût | Faible | Élevé | Moyen |
| Temps | Plus long | Moyen | Moins long |
| Taux de réponse | Faible | Élevé | Moyen-élevé |

Tableau 1.1 : Caractéristiques des méthodes de collecte

1.2.6 Saisie des données et codage

Si les données n'ont pas été recueillies 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.

1.2.7 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. Certaines lacunes 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, auquel cas l'imputation est souvent utilisée pour régler les autres cas. L'imputation est un processus qui consiste à déterminer et à attribuer des valeurs de remplacement afin de résoudre les problèmes de données manquantes, non valables ou incohérentes.

1.2.8 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'obtenir des valeurs de la population d'intérêt et tirer des conclusions à partir de l'information obtenue d'un échantillon. Cet aspect sera traité en détail dans le cours.

1.2.9 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 pouvoir faire le lien entre les résultats de l'enquête et les questions et problèmes mentionnés dans l'énoncé des objectifs (section 1.2.1). L'analyse 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 d'analyses de régression, d'analyses de variance ou tests du khi-deux.

1.2.10 Diffusion des données

La diffusion des données est la distribution des données de l'enquête aux utilisateurs par l'intermédiaire de divers medias, par exemple, un communiqué de presse, une interview radio ou télédiffusée, etc.

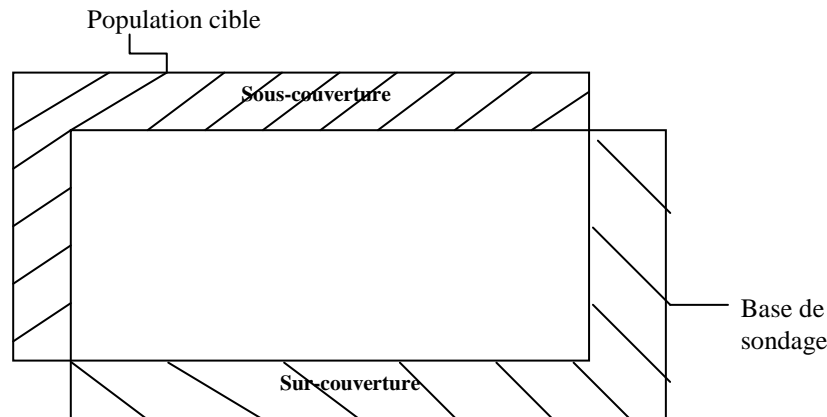


Figure 1.1 : Erreurs de couverture

1.3 Sources d'erreurs dans les enquêtes

Les estimations produites lors d'une enquête sont sujettes à de nombreuses erreurs : les erreurs dues à l'échantillonnage et les erreurs non dues à l'échantillonnage. Bien qu'il soit impossible d'éliminer toutes les erreurs, il est important d'allouer des ressources humaines et financières suffisantes afin de minimiser leur impact sur la qualité des estimations. Par exemple, dans le cas d'enquêtes avec interview (sur place ou téléphonique), la formation et la supervision des intervieweurs est importante afin de réduire certaines erreurs non dues à l'échantillonnage. La conception du questionnaire est un autre aspect exigeant des ressources importantes ce qui permettra d'éviter des réponses incohérentes, non valables et des erreurs de mesure (voir section 1.3.2). Il est important de tester le questionnaire à l'aide d'une enquête pilote afin d'identifier tout problème. Encore une fois, cela exige des ressources financières importantes puisque le coût d'une enquête pilote est non négligeable. Dans ce qui suit, nous présentons les principales sources d'erreurs inhérentes aux enquêtes.

1.3.1 Erreurs dues à l'échantillonnage

L'erreur due à l'échantillonnage est propre à toute enquête-échantillon. Il y a erreur d'échantillonnage lorsque l'on estime une caractéristique d'intérêt en mesurant seulement une partie de la population au lieu de la population au complet. L'ampleur de l'erreur due à l'échantillonnage peut être limitée par la taille de l'échantillon, le plan de sondage et la méthode d'estimation. Dans un recensement, il n'y a pas d'erreurs dues à l'échantillonnage car tous les membres de la population sont dénombrés.

1.3.2 Erreurs non dues à l'échantillonnage

On distingue essentiellement quatre types d'erreurs non dues à l'échantillonnage.

- (i) **Erreurs de couverture** : Ce type d'erreur survient lorsque la base de sondage et la population cible ne coïncident pas parfaitement. Il y a *sous-couverture* lorsque certaines unités de la population cible ne sont pas sur la base de sondage et ne peuvent donc pas être sélectionnées. Il y a *sur-couverture* lorsque certaines unités sur la base de sondage ne sont pas dans la population cible.
- (ii) **Erreurs de non-réponse** : Malgré tous les efforts mis de l'avant pour maximiser la quantité d'information recueillie, il y aura inévitablement de la non-réponse. On distingue deux types de non-réponse :
 - La *non-réponse totale* : elle survient lorsque l'unité échantillonnée ne répond à aucune des questions posées; elle peut être causée par un refus de subir l'interview ou parce que l'organisme en charge de l'enquête a été incapable de contacter l'unité échantillonnée. La non-réponse totale est généralement traitée par l'utilisation de méthodes de re-pondération qui consistent à hausser le poids de sondage des répondants (qui est le nombre d'individus que le répondant dans l'échantillon représente dans la population) pour compenser pour l'absence des non-répondants.
 - La *non-réponse partielle* : elle survient lorsque l'unité échantillonnée a répondu à certaines questions, mais pas toutes; elle peut être causée parce que la question est perçue comme indiscrete. La non-réponse partielle est généralement traitée par l'imputation qui consiste à déterminer une valeur de remplacement pour la valeur manquante.
- (iii) **Erreurs 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 et la méthode de collecte peuvent susciter ce type d'erreur.
- (iv) **Erreurs de traitement** : Le traitement transforme les réponses de l'enquête obtenues pendant la collecte en une mise en forme qui convient à l'analyse des

données. Il s'agit d'un ensemble d'activités manuelles et automatisées qui demandent beaucoup de temps et de ressources et ce volet peut donc être une source éventuelle d'erreurs. Ces activités comprennent le codage et la saisie des données décrits à la section 1.2.6.

1.4 Recensement vs. enquête-échantillon : comment choisir?

Tel que mentionné précédemment, on distingue deux types d'enquête : l'enquête-échantillon et le recensement. Ci-dessous, nous présentons les éléments les plus importants à considérer avant de choisir un recensement ou une enquête-échantillon.

1.4.1 Erreurs d'enquête

Comme nous l'avons constaté à la section 1.3, les estimations produites lors d'une enquête sont sujettes de nombreuses erreurs : les erreurs dues à l'échantillonnage et celles non dues à l'échantillonnage. Intuitivement, il serait tentant de choisir le recensement au lieu d'une enquête-échantillon car la qualité des estimations serait plus grande étant donné que toute la population est échantillonnée. Il est vrai que dans le cas d'un recensement, il n'y a pas d'erreur due à l'échantillonnage. Il ne faut cependant pas perdre de vue que l'ampleur des erreurs non dues à l'échantillonnage peut être potentiellement plus importante que celle des erreurs due à l'échantillonnage. Comme nous l'avons souligné à la section 1.3, les erreurs non dues à l'échantillonnage peuvent être dues à un manque de ressources financières, à une mauvaise formation des intervieweurs, etc. Dans le cas d'une enquête-échantillon, il est généralement plus simple de conserver un certain degré de contrôle qui garantira, par exemple, que les intervieweurs sont bien formés et qu'ils sont supervisés de manière adéquate, compte tenu de la taille (habituellement petite) de l'échantillon. Dans le cas d'un recensement, l'impact des erreurs non dues à l'échantillonnage sur la qualité des estimations peut être substantiel car les recensements exigent des moyens financiers et humains exorbitants, ce qui est rarement réalisé en pratique.

1.4.2 Coût

Étant donné que tous les membres de la population font l'objet de l'enquête, le recensement coûte beaucoup plus cher que l'enquête-échantillon.

1.4.3 Rapidité d'exécution

Il faut souvent obtenir et traiter les données, puis diffuser les résultats, au cours d'une période relativement courte. 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. Par exemple, dans le cas du recensement canadien, il s'écoule plusieurs années avant que les estimations finales soient disponibles.

1.4.4 Taille de la population

Lorsque la taille de la population est petite, le recensement est généralement préférable. En effet, pour obtenir des estimations ayant une petite erreur due à l'échantillonnage, il peut être nécessaire de tirer un échantillon de grande taille. Dans ce cas, et pour des frais supplémentaires minimales, les données peuvent être disponibles pour toute la population, au lieu d'une fraction seulement. Par exemple, les enquêtes portant sur la culture (cinémas, théâtre) menées à Statistique Canada sont des recensements.

1.4.5 Estimation pour petits domaines

Le recensement peut être préférable lorsque des estimations sont requises pour des secteurs géographiques restreints.

1.5 Définitions et notation

Dans l'univers des enquêtes, on fait face à des populations finies. C'est d'ailleurs un des facteurs principaux qui distingue la théorie de l'échantillonnage de la théorie statistique « classique » (voir section 1.6). Soit U une population finie de taille N . On écrit

$$U = \{1, \dots, i, \dots, N\}.$$

Dans une enquête, on distingue deux types de variables :

- (i) Les variables d'intérêt, y_1, \dots, y_p ; ce sont les variables que l'on tente de mesurer à l'aide de l'enquête.

- (ii) Les variables auxiliaires, x_1, \dots, x_q ; ce sont des variables disponibles
- (a) pour toutes les unités dans la population;
- ou
- (b) pour toutes les unités échantillonnées seulement mais leur total (ou leur moyenne) dans la population est connue.

Les variables auxiliaires de type (a) sont généralement disponibles sur la base de sondage. Par exemple, le registre des entreprises canadiennes contient, pour toutes les entreprises au Canada, des variables telles que la province, le code SCIAN ou le nombre d'employés. Les variables auxiliaires de type (b) proviennent généralement de sources telles que le recensement. Par exemple, le recensement canadien produit les effectifs dans la population par sexe et par groupe d'âge au niveau national mais également au niveau provincial et municipal. Plusieurs enquêtes de Statistique Canada utilisent les variables sexe et groupe d'âge comme variables auxiliaires à l'étape de l'estimation. Comme nous le verrons, les variables auxiliaires permettent d'utiliser des plans de sondage et des estimateurs plus efficaces. Elles permettent également de traiter la non-réponse totale et la non-réponse partielle.

Dans l'univers des enquêtes, on est généralement intéressé à estimer des *paramètres de la population finie* qui sont des paramètres descriptifs. Voici quelques exemples de paramètres :

- (i) Le total dans la population : $t_y = \sum_{i \in U} y_i$, où y_i désigne la valeur prise par la variable d'intérêt y pour l'unité i .
- (ii) La moyenne dans la population : $\bar{y}_U = t_y / N$.
- (iii) Le ratio de deux totaux : $R = \frac{t_{y_1}}{t_{y_2}}$.
- (iv) La dispersion dans la population de la variable d'intérêt y :

$$S_y^2 = \frac{1}{N-1} \sum_{i \in U} (y_i - \bar{y}_U)^2.$$
- (v) La covariance dans la population entre deux variables :

$$S_{y_1, y_2} = \frac{1}{N-1} \sum_{i \in U} (y_{1i} - \bar{y}_{1U})(y_{2i} - \bar{y}_{2U}).$$

On insistera sur le fait que les paramètres (i)-(v) sont des paramètres de population finie en ce sens qu'ils décrivent un aspect de la population finie sous étude, (moyenne, dispersion, etc.). Afin d'estimer ces paramètres, on sélectionne un échantillon aléatoire (que l'on désigne par s) selon une procédure appelée *plan de sondage*. Le plan de sondage joue un rôle important dans l'étude des propriétés statistiques d'un estimateur (par exemple, le biais et la variance).

Désignons par θ un paramètre d'intérêt (total, moyenne, etc.) et soit $\hat{\theta}$ un estimateur de θ . On définit maintenant des concepts de base tels que le biais et la variance d'un estimateur.

Définition 1.8 : Le biais de $\hat{\theta}$ est défini selon $\text{Biais}(\hat{\theta}) = E(\hat{\theta}) - \theta$, où $E(\cdot)$ désigne l'espérance mathématique d'une variable aléatoire. Un estimateur $\hat{\theta}$ est sans biais pour θ si $\text{Biais}(\hat{\theta}) = 0$.

Définition 1.9 : La variance de $\hat{\theta}$ est définie selon $V(\hat{\theta}) = E[\hat{\theta} - E(\hat{\theta})]^2$.

Définition 1.10 : L'erreur quadratique moyenne (EQM) de $\hat{\theta}$ est définie selon $EQM(\hat{\theta}) = E(\hat{\theta} - \theta)^2$.

Définition 1.11 : Le coefficient de variation $\hat{\theta}$ est défini selon $CV(\hat{\theta}) = \frac{\sqrt{V(\hat{\theta})}}{E(\hat{\theta})}$.

Définition 1.12 : Le ratio du biais de $\hat{\theta}$ est défini selon $RB(\hat{\theta}) = \frac{\text{Biais}(\hat{\theta})}{\sqrt{V(\hat{\theta})}}$.

Remarques :

(1) L'interprétation de l'espérance mathématique d'une variable aléatoire est habituellement présentée dans un cours de base de probabilité/statistique. Rappelons que l'espérance représente la moyenne des résultats obtenus lorsque l'on répète une expérience aléatoire une infinité de fois. En théorie de l'échantillonnage, on verra au Chapitre 2 que l'expérience aléatoire sous-jacente est la sélection de l'échantillon.

(2) Il n'est pas difficile de montrer que l'EQM d'un estimateur (Définition 1.10) peut s'exprimer en fonction de sa variance et de son biais au carré. En effet, on a

$$EQM(\hat{\theta}) = V(\hat{\theta}) + \text{Biais}(\hat{\theta})^2. \text{ Si } \hat{\theta} \text{ est un estimateur sans biais de } \theta, \text{ on a alors } EQM(\hat{\theta}) = V(\hat{\theta}).$$

- (3) Le coefficient de variation d'un estimateur (Définition 1.11) est fréquemment utilisé par les organismes statistiques pour indiquer le degré de précision d'un estimateur. C'est un indicateur de qualité très intéressant car il n'a pas d'unité, ce qui offre la possibilité de comparer différents estimateurs entre eux. Habituellement, on juge qu'un estimateur est précis lorsque son coefficient de variation est 5% ou moins. Ce seuil est toutefois subjectif.

1.6 Théorie de l'échantillonnage vs. statistique classique

En statistique classique, on suppose habituellement que les données sont générées indépendamment à partir d'une loi de probabilité connue. En effet, un échantillon est une suite de n observations indépendantes, y_1, \dots, y_n , qui sont des réalisations d'une variable aléatoire satisfaisant une distribution (ou modèle) connue. Par exemple, un modèle fréquemment étudié dans un cours de statistique de base est donné par

$$m: y_i = \mu + \varepsilon_i, \quad (1.1)$$

où les erreurs ε_i sont distribuées selon une loi normale de moyenne 0 et de variance σ^2 ; i.e., $E_m(\varepsilon_i) = 0$ et $V_m(\varepsilon_i) = \sigma^2$, où $E_m(\cdot)$ et $V_m(\cdot)$ désignent l'espérance et la variance par rapport au modèle (1.1). On peut réécrire (1.1) de manière plus compacte en écrivant $y_i \stackrel{i.i.d}{\sim} N(\mu, \sigma^2)$. Le modèle (1.1) peut s'interpréter de la façon suivante : les valeurs y_i sont générées à partir d'une population hypothétique de taille infinie. Si on avait toutes les valeurs y_i de la population à notre disposition, un graphique montrerait que les données sont en forme de « cloche » (allure d'une distribution normale). En statistique classique, on est intéressé à mener des inférences sur des paramètres μ et σ^2 , qui désignent respectivement la moyenne et la variance de la population conceptuelle. Cette dernière étant de taille infinie, il est clair qu'il n'est jamais possible de connaître la vraie valeur de μ . Il est bien connu que la moyenne échantillonnale, $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$, est un estimateur sans biais de la moyenne de la population μ et que la dispersion $s^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2$ est un estimateur sans biais de la variance σ^2 ; i.e., $E_m(\bar{y}) = \mu$ et $E_m(s^2) = \sigma^2$.

Dans l'univers des enquêtes, contrairement à l'univers de la statistique classique, on est en présence de populations finies (par exemple, la population des ménages au Canada). C'est essentiellement ce qui distingue la théorie de l'échantillonnage de celle de la statistique classique. De plus, dans les enquêtes, on est le plus souvent intéressé à des paramètres de la population finie tels que la moyenne dans la population, \bar{Y} . Dans le cas d'un recensement et en supposant que les erreurs non dues à l'échantillonnage sont

négligeables, on obtiendrait la vraie valeur de \bar{Y} . Rappelons que ceci n'est pas possible lorsque l'on cherche à estimer un paramètre de population infinie comme c'est le cas en statistique classique.

Dans les enquêtes, on tire un échantillon aléatoire selon un plan de sondage donné. Supposons que l'on cherche à estimer un paramètre θ et soit $\hat{\theta}$ un estimateur de θ . L'approche traditionnelle pour l'inférence est *l'approche basée sur le plan de sondage* (en anglais, *design-based approach*) selon laquelle le vecteur $\mathbf{y} = (y_1, \dots, y_N)'$ pour une variable d'intérêt donnée est traité comme fixe. Autrement dit, le vecteur \mathbf{y} est considéré non-aléatoire (contrairement à la statistique classique). À ce stade, on pose la question suivante : si l'on cherche à étudier les propriétés de $\hat{\theta}$ (par exemple, biais et variance) et compte tenu que le vecteur \mathbf{y} est non-aléatoire, quel est le mécanisme aléatoire sous-jacent qui entre en jeu? Pour répondre à cette question, on introduit une variable aléatoire importante en théorie de l'échantillonnage : la *variable indicatrice de sélection* définie selon

$$Z_i = \begin{cases} 1 & \text{si l'unité } i \text{ est sélectionné dans l'échantillon} \\ 0 & \text{sinon} \end{cases}$$

Cette variable est traitée comme étant aléatoire lors de l'étude des propriétés d'un estimateur. Par conséquent, les notions de biais et de variance font tous référence à l'échantillonnage répété, comme nous le verrons au Chapitre 2.

Plans de sondage simples

Chapitre 2

2.1 Introduction

Dans ce chapitre, nous commençons par l'étude du cadre de travail traditionnel utilisé en théorie de l'échantillonnage afin de mener une inférence. Dans la section 2.2, nous introduisons le concept de probabilité d'inclusion et nous définissons les concepts de biais, de variance et d'erreur quadratique moyenne. L'échantillonnage aléatoire simple sans remise sera considéré à la section 2.3. À la section 2.4, la construction des intervalles de confiance est considérée. Le calcul de la taille de l'échantillon fait l'objet de la section 2.5. Les sections 2.6 et 2.7 présentent deux plans de sondage relativement simples : le plan Bernoulli et le plan systématique. Finalement, nous introduirons l'estimateur de Horvitz-Thompson à la section 2.8 qui peut être utilisé pour estimer un total dans le cas de tous les plans de sondage considérés dans ce chapitre.

2.2 Cadre de travail pour l'échantillonnage probabiliste

Soit $U = \{1, 2, \dots, N\}$ une population finie. Un échantillon s est n'importe quel sous-ensemble de U . Soit Ω l'ensemble de tous les échantillons, s , possibles. Soit $p(s)$ la probabilité de sélectionner l'échantillon s , $s \in \Omega$. Un *plan de sondage* est une loi de probabilités sur Ω telle que

- (i) $p(s) \geq 0 \quad \forall s \in \Omega$;
- (ii) $\sum_{s \in \Omega} p(s) = 1$.

Autrement dit, le plan de sondage est une fonction qui associe une probabilité de sélection à chaque échantillon dans l'ensemble des échantillons possibles, Ω .

Exemple 2.1 : Supposons que la population est composée de 4 unités : $U = \{1, 2, 3, 4\}$. Considérons tous les échantillons sans remise de taille $n = 2$ tirés de U . On désigne par s_1, \dots, s_6 , les 6 échantillons possibles de taille 2. Autrement dit, l'ensemble de tous les échantillons possibles Ω contient 6 éléments, s_1, \dots, s_6 . Supposons que le plan de sondage utilisé est tel que : $p(s_1) = \frac{1}{3}, p(s_2) = \frac{1}{6}, p(s_6) = \frac{1}{2}$ et $p(s_3) = p(s_4) = p(s_5) = 0$. Le tableau suivant résume la situation :

| Échantillon | $s_1 = \{1, 2\}$ | $s_2 = \{1, 3\}$ | $s_3 = \{1, 4\}$ | $s_4 = \{2, 3\}$ | $s_5 = \{2, 4\}$ | $s_6 = \{3, 4\}$ |
|-------------|------------------|------------------|------------------|------------------|------------------|------------------|
| $p(s)$ | 1/3 | 1/6 | 0 | 0 | 0 | 1/2 |

Dans le cas de l'échantillonnage probabiliste, on affecte une probabilité de sélection à chaque échantillon possible. Chaque unité a donc une probabilité connue d'appartenir à un échantillon. Cette probabilité est appelée *probabilité d'inclusion*. La probabilité d'inclusion de l'unité i , que l'on désigne par π_i , est définie selon

$$\pi_i = P(i \in s) = P(Z_i = 1) = \sum_{\substack{s \in \Omega \\ s \ni i}} p(s),$$

où

$$Z_i = \begin{cases} 1 & \text{si l'unité } i \text{ est tirée dans l'échantillon} \\ 0 & \text{sinon} \end{cases}$$

La variable dichotomique Z_i est donc une variable indicatrice de sélection dans l'échantillon. Notons que, dans le cas de l'échantillonnage probabiliste, les probabilités d'inclusion π_i sont connues avant la sélection de l'échantillon. De plus, nous supposons que le plan de sondage est tel que $\pi_i > 0$ pour toutes les unités dans la population.

Exemple 2.1 (suite) : On a

$$\begin{aligned} \pi_1 &= p(s_1) + p(s_2) + p(s_3) = 1/2, \\ \pi_2 &= p(s_1) + p(s_4) + p(s_5) = 1/3, \\ \pi_3 &= p(s_2) + p(s_4) + p(s_6) = 2/3, \\ \pi_4 &= p(s_3) + p(s_5) + p(s_6) = 1/2. \end{aligned}$$

Notons que $\pi_1 + \pi_2 + \pi_3 + \pi_4 = 2 = n$. De manière générale, on montre que, pour un plan de sondage à taille fixe, on a $\sum_{i \in U} \pi_i = n$ (la preuve est laissée en exercice).

De manière similaire, on définit la *probabilité d'inclusion jointe* pour les unités i et j , π_{ij} , selon

$$\pi_{ij} = P(i \in s \& j \in s) = P(Z_i = 1, Z_j = 1) = \sum_{\substack{s \in \Omega \\ s \ni (i,j)}} p(s).$$

Notons que $\pi_{ii} = \pi_i$ et que $\pi_{ij} = \pi_{ji}$.

Exemple 2.1 (suite) : On a

$$\begin{aligned}\pi_{12} &= p(s_1) = 1/3, \\ \pi_{13} &= p(s_2) = 1/6, \\ \pi_{14} &= p(s_3) = 0, \\ \pi_{23} &= p(s_4) = 0, \\ \pi_{24} &= p(s_5) = 0, \\ \pi_{34} &= p(s_6) = 1/2.\end{aligned}$$

Comme nous l'avons brièvement évoqué au à la section 1.6, la majorité des résultats en échantillonnage repose sur la distribution d'une statistique (par exemple, un estimateur) qui est la distribution des valeurs prises par cette statistique obtenues lorsque l'on a tiré tous les échantillons possibles dans une population donnée. Autrement dit, les propriétés (par exemple, le biais et la variance) d'un estimateur sont étudiées dans un cadre particulier qui consiste à tirer tous les échantillons possibles.

Exemple 2.2 : Considérons la population suivante : $U = \{1, 2, 3, 4, 5, 6, 7, 8\}$. Les données sont exhibées dans le tableau suivant :

| | | | | | | | | |
|-------|---|---|---|---|---|---|---|---|
| i | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| y_i | 1 | 2 | 4 | 4 | 7 | 7 | 7 | 8 |

Considérons les échantillons sans remise de taille $n=4$. L'ensemble Ω contient $\binom{8}{4} = 70$ échantillons possibles. Supposons que $p(s_k) = \frac{1}{70}, k=1, \dots, 70$. Comme nous le verrons à la section 2.3, ce plan de sondage correspond à l'échantillonnage aléatoire simple sans remise. Dans ce cas, la probabilité d'inclusion de l'unité i est $\pi_i = \frac{1}{2}, i=1, \dots, 8$. Supposons que l'on veuille estimer le total dans la population $t_y = \sum_{i \in U} y_i$ et que nous décidions d'utiliser l'estimateur $\hat{t}_y = N\bar{y}_s$, où $\bar{y}_s = \frac{1}{4} \sum_{i=1}^4 y_i$ désigne la moyenne des observations dans l'échantillon. La distribution de l'estimateur \hat{t}_y est exhibée dans le tableau suivant.

| | | | | | | | | | | | | | | | |
|--------------------|------|------|------|------|------|------|------|-------|------|------|------|------|------|------|------|
| $\hat{t}_y = t$ | 22 | 28 | 30 | 32 | 34 | 36 | 38 | 40 | 42 | 44 | 46 | 48 | 50 | 52 | 58 |
| $P(\hat{t}_y = t)$ | 1/70 | 6/70 | 2/70 | 3/70 | 7/70 | 4/70 | 6/70 | 12/70 | 6/70 | 4/70 | 7/70 | 3/70 | 2/70 | 6/70 | 1/70 |

Définition 2.1 : L'espérance de \hat{t}_y est définie selon

$$E(\hat{t}_y) = \sum_{s \in \Omega} \hat{t}_y p(s).$$

L'espérance de l'estimateur \hat{t}_y est donc la moyenne pondérée de toutes les valeurs possibles prises par cet estimateur. Le poids $p(s)$ correspond à la probabilité d'occurrence d'une valeur particulière prise par l'estimateur, t . Il est donc important de comprendre, que dans le contexte de l'échantillonnage, l'espérance d'un estimateur est une moyenne pondérée des valeurs prises par cet estimateur lorsque l'on répète le processus d'échantillonnage. Le biais, défini selon $\text{Biais}(\hat{t}_y) = E(\hat{t}_y) - t_y = E(\hat{t}_y - t_y)$, est donc est la moyenne pondérée des valeurs prises par l'erreur due à l'échantillonnage, $\hat{t}_y - t_y$, lorsque l'on répète le processus d'échantillonnage.

Exemple 2.2 (suite) : D'abord, notons que $t_y = 40$. On a

$$E(\hat{t}_y) = \frac{1}{70}(22) + \frac{6}{70}(28) + \dots + \frac{1}{70}(58) = 40.$$

Puisque $E(\hat{t}_y) = t_y$, il s'ensuit que l'estimateur \hat{t}_y est sans biais pour t_y .

Définition 2.2 : La variance de \hat{t}_y est définie selon

$$\begin{aligned} V(\hat{t}_y) &= E\left[\hat{t}_y - E(\hat{t}_y)\right]^2 \\ &= \sum_{s \in \Omega} \left[\hat{t}_y - E(\hat{t}_y)\right]^2 p(s). \end{aligned}$$

Remarque : Si \hat{t}_y est sans biais pour t_y , (i.e., $E(\hat{t}_y) = t_y$), alors la variance de \hat{t}_y s'écrit simplement comme

$$V(\hat{t}_y) = \sum_{s \in \Omega} (\hat{t}_y - t_y)^2 p(s).$$

Exemple 2.2 (suite) : On a

$$V(\hat{t}_y) = \frac{1}{70}(22 - 40)^2 + \dots + \frac{1}{70}(58 - 40)^2 = 54,86.$$

Définition 2.3 : L'erreur quadratique moyenne de \hat{t}_y est définie selon

$$\begin{aligned}EQM(\hat{t}_y) &= E[\hat{t}_y - t_y]^2 \\ &= \sum_{s \in \Omega} (\hat{t}_y - t_y)^2 p(s).\end{aligned}$$

2.3 Échantillonnage aléatoire simple sans remise

L'échantillonnage aléatoire simple est l'un des plans de sondage les plus simples. On distingue l'échantillonnage aléatoire simple sans remise (EASSR) de l'échantillonnage aléatoire simple avec remise (EASAR). Dans ce chapitre, nous considérons l'EASSR seulement. L'étude de l'EASAR est laissée en exercice.

2.3.1 Définition

Considérons une population U de taille N . Dans le cas de l'EASSR, n'importe quel sous ensemble de n unités distinctes a la même probabilité d'être tiré. Puisqu'il existe $\binom{N}{n}$ échantillons possibles, la probabilité de tirer un échantillon s de taille n est égale à

$$p(s) = \frac{1}{\binom{N}{n}}.$$

L'EASSR est un plan de sondage à *taille fixe* puisque la taille n est fixée avant la sélection de l'échantillon.

Proposition 2.1 : Considérons un EASSR de taille n tiré d'une population de taille N . Alors,

(i) $\pi_i = \frac{n}{N}$ pour tout i .

(ii) $\pi_{ij} = \frac{n(n-1)}{N(N-1)}$ pour tout $(i, j), i \neq j$.

Démonstration :

Donc, dans le cas de l'EASSR, chaque unité a la même probabilité d'être incluse dans l'échantillon. On dit alors que l'EASSR est un plan de sondage *auto-pondéré* (i.e., un plan de sondage avec probabilité d'inclusion constante).

2.3.2 Quelques algorithmes de tirage

Comment tirer un EASSR en pratique? Plusieurs algorithmes ont été proposés dans la littérature. Dans ce qui suit nous décrivons deux algorithmes simples.

Algorithme 1 : On procède selon les étapes suivantes :

Étape 1 : On tire une première unité avec probabilité $\frac{1}{N}$;

Étape 2 : On tire une deuxième unité avec probabilité $\frac{1}{N-1}$;

· · ·
· · ·
· · ·

Étape n : On tire la dernière unité avec probabilité $\frac{1}{N-n+1}$.

Cette procédure mène à un EASSR puisque

$$p(s) = \left\{ \frac{1}{N} \times \frac{1}{N-1} \times \dots \times \frac{1}{N-n+1} \right\} \times n! = \frac{(N-n)!n!}{N!} = \frac{1}{\binom{N}{n}},$$

où le terme entre accolades $\{ \}$ représente la probabilité d'une permutation donnée de $(1, \dots, i, \dots, n)$ et $n!$ représente le nombre de permutations possibles.

Algorithme 2 : Cette méthode consiste à d'abord permuter les N unités aléatoirement. Ensuite, on sélectionne les n premières unités dans la liste ordonnée. En pratique, on peut procéder comme suit : on génère des variables aléatoires indépendantes $\varepsilon_1, \dots, \varepsilon_N$ correspondant aux unités $1, \dots, N$, respectivement, à partir d'une distribution uniforme de paramètres 0 et 1. Ensuite, on ordonne les unités en ordre croissant (ou décroissant) selon les valeurs générées ε_i . L'échantillon s est composé des n premières unités.

On peut montrer (mais nous ne le ferons pas!) que cette procédure mène à un EASSR, i.e.,

$$p(s) = \frac{1}{\binom{N}{n}}.$$

2.3.3 Estimation d'une moyenne

Une fois l'échantillon tiré et la collecte des données effectuée, on dispose des observations pour les n unités tirées. Supposons que l'on veuille estimer la moyenne dans la population d'une variable d'intérêt y , $\bar{y}_U = \frac{1}{N} \sum_{i \in U} y_i$. Quel estimateur raisonnable devrait-on utiliser?

Avant de répondre à cette question, nous énonçons le lemme suivant qui servira dans la démonstration de résultats importants.

Lemme 2.1 : Soit Z_i la variable indicatrice de sélection telle que $Z_i = 1$ si $i \in s$ et $Z_i = 0$ si $i \notin s$. Alors,

(i) $E(Z_i) = \frac{n}{N}$.

(ii) $V(Z_i) = \frac{n}{N} \left(1 - \frac{n}{N}\right)$.

(iii) $Cov(Z_i, Z_j) = -\frac{1}{N-1} \left(1 - \frac{n}{N}\right) \frac{n}{N}$ si $i \neq j$.

Démonstration :

Proposition 2.2 : Supposons que s désigne un EASSR de taille n tiré d'une population U de taille N . Alors, la moyenne dans l'échantillon $\bar{y}_s = \frac{1}{n} \sum_{i \in s} y_i$ est sans biais pour la moyenne de la population \bar{y}_U ; i.e., $E(\bar{y}_s) = \bar{y}_U$.

Démonstration :

Proposition 2.3 : Dans le cas de l'EASSR, la variance de l'estimateur \bar{y}_s est donnée par

$$V(\bar{y}_s) = \left(1 - \frac{n}{N}\right) \frac{S_y^2}{n}, \quad (2.1)$$

où $S_y^2 = \frac{1}{N-1} \sum_{i \in U} (y_i - \bar{y}_U)^2$ désigne la dispersion de la variable y dans l'ensemble de la population.

Démonstration :

Remarques :

(1) En vertu de l'expression (2.1), il est clair que l'estimateur \bar{y}_s a une petite variance si :

- (i) la taille de l'échantillon n est grande.
- (ii) la fraction de sondage n/N est grande.
- (iii) la dispersion de la variable y , S_y^2 , est petite.

(2) Le facteur $\left(1 - \frac{n}{N}\right)$ en (2.1) est *un facteur de correction pour population finie*.

Notons que, pour une taille de population N donnée, $\left(1 - \frac{n}{N}\right) \rightarrow 0$ quand n augmente. En pratique, il n'est pas rare que la fraction de sondage $\frac{n}{N}$ soit petite (et même négligeable).

Dans ce cas, on a $\left(1 - \frac{n}{N}\right) \approx 1$.

(3) La variance de \bar{y}_s , $V(\bar{y}_s)$, en (2.1) ne peut être calculée au moyen des observations dans l'échantillon car S_y^2 est une fonction des y pour toutes les unités dans la population. Puisque S_y^2 est inconnu, il faudra l'estimer. Un estimateur de la variance $V(\bar{y}_s)$ est donné par la Proposition 2.4. Mais d'abord, nous énonçons le lemme suivant :

Lemme 2.2 : La dispersion de la variable y dans l'échantillon, $s_y^2 = \frac{1}{n-1} \sum_{i \in S} (y_i - \bar{y}_s)^2$, est un estimateur sans biais de S_y^2 ; i.e., $E(s_y^2) = S_y^2$.

Démonstration :

Proposition 2.4 : Un estimateur sans biais de $V(\bar{y}_s)$ est donné par

$$\hat{V}(\bar{y}_s) = \left(1 - \frac{n}{N}\right) \frac{s_y^2}{n}. \quad (2.2)$$

On a donc $E[\hat{V}(\bar{y}_s)] = V(\bar{y}_s)$.

Démonstration : Le résultat découle du Lemme 2.2.

Exemple 2.3 : D'une population de $N = 8\,427$ comptes à recevoir, on prélève un EASSR de taille $n = 30$ afin d'estimer la valeur moyenne des comptes. Voici les résultats, en dollars :

| | | | | | | | | | |
|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|
| 240,82 | 232,50 | 740,81 | 860,32 | 224,10 | 7,15 | 324,12 | 240,12 | 190,08 | 182,75 |
| 160,21 | 148,22 | 132,19 | 119,25 | 113,85 | 108,30 | 107,10 | 101,19 | 99,21 | 93,12 |
| 88,13 | 80,15 | 78,13 | 72,15 | 67,13 | 65,14 | 41,10 | 32,17 | 10,02 | 9,15 |

- Estimer la moyenne \bar{y}_U de la population
- Estimer la variance de l'estimateur utilisé en (a).

Solution :

(a) Nous estimons \bar{y}_U par $\bar{y}_s = \frac{1}{n} \sum_{i \in s} y_i = \frac{4\,968,68}{30} = 165,62$.

(b) On a $\hat{V}(\bar{y}_s) = \left(1 - \frac{n}{N}\right) \frac{s_y^2}{n} = \left(1 - \frac{30}{8427}\right) \frac{35930,7}{30} = 1193,83$.

Exemple 2.4 : Afin de mieux saisir la notion d'espérance et de variance dans le contexte de populations finies, considérons l'expérience théorique suivante. J'ai généré une population de taille $N=1000$ comprenant une variable d'intérêt y de moyenne $\bar{y}_U = 100$. De la population ainsi générée, j'ai tiré 100 EASSR de taille $n=50$ et $n=200$. Dans chaque échantillon, j'ai calculé l'erreur relative (en %) $ER(\bar{y}_s) = 100 \times \left(\frac{\bar{y}_s - \bar{y}_U}{\bar{y}_U} \right)$.

Notons qu'en pratique, il n'est généralement pas possible d'obtenir la valeur de $ER(\bar{y}_s)$ puisque la moyenne de la population \bar{y}_U est inconnue (sinon, nous n'aurions pas besoin de mener une enquête!). Les figures 2.1 et 2.2 exhibent l'erreur relative en fonction de l'échantillon. L'examen de ces graphiques nous montre bien que l'erreur relative prend tantôt des valeurs négatives, tantôt des valeurs positives et qu'en moyenne l'erreur relative tourne autour de la valeur 0, ce qui n'est pas surprenant puisque l'estimateur \bar{y}_s est un estimateur sans biais de la moyenne \bar{y}_U (voir Proposition 2.2). Il est également intéressant de noter que lorsque $n=50$, l'erreur relative a tendance à prendre des valeurs plus grandes pour certains échantillons (le maximum en valeur absolue étant égale à environ 8%). Lorsque $n=200$, l'erreur relative prend des valeurs toutes comprises entre -5% et 5%. En fait pour la grande majorité des échantillons, l'erreur relative se situe entre -2.5% et 2.5%. Donc, lorsque la taille de l'échantillon augmente, l'erreur relative tourne autour de 0 tout en étant de plus en plus concentrée autour de 0. Autrement dit, lorsque la taille de l'échantillon augmente, la variance de l'estimateur diminue, ce qui n'est pas surprenant en vertu de la Proposition 2.3.

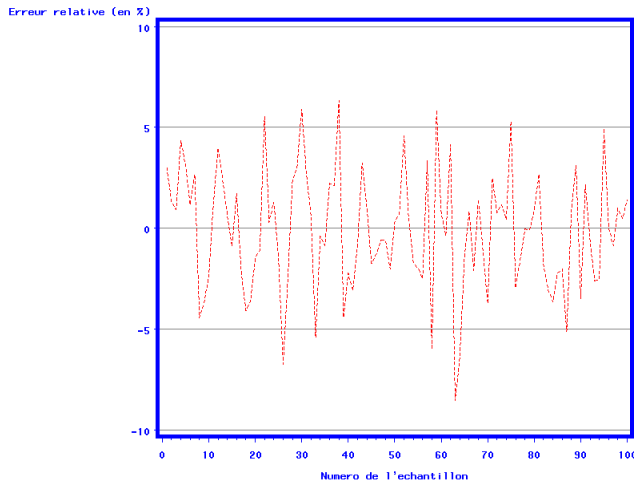


Figure 2.1 : Erreur relative avec $n = 50$.

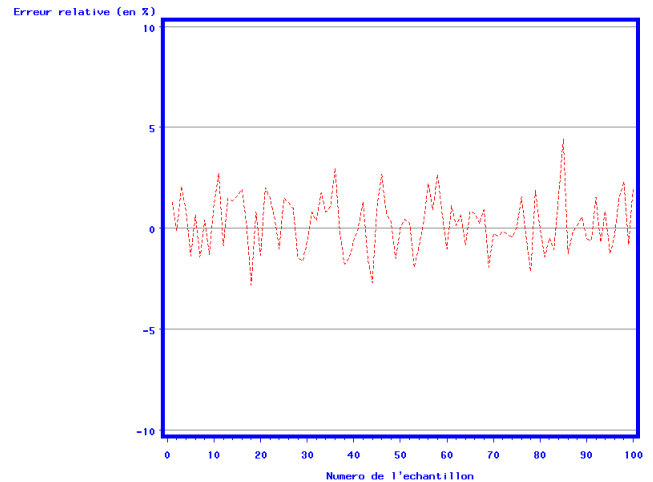


Figure 2.2 : Erreur relative avec $n = 200$.

2.3.4 Estimation d'un total

À la section 2.3.3, nous avons considéré le problème de l'estimation d'une moyenne. En pratique, des estimations pour des totaux sont fréquemment requises. Le total dans la population pour une variable d'intérêt y est donné par $t_y = \sum_{i \in U} y_i = N\bar{y}_U$. Supposons que la taille de la population N soit connue. Par la Proposition 2.2, un estimateur de t_y est donné par $\hat{t}_y = N\bar{y}_s$. À l'aide la Proposition 2.3, on détermine aisément sa variance qui est donnée par

$$V(\hat{t}_y) = V(N\bar{y}_s) = N^2V(\bar{y}_s) = N^2\left(1 - \frac{n}{N}\right)\frac{S_y^2}{n}.$$

Finalement, on déduit de (2.2) qu'un estimateur de la variance $V(\hat{t}_y)$ est donné par

$$\hat{V}(\hat{t}_y) = N^2\left(1 - \frac{n}{N}\right)\frac{s_y^2}{n}.$$

2.3.5 Estimation d'une proportion

Dans cette section, nous considérons le problème de l'estimation d'une proportion. Soit p_y la proportion d'unités dans la population possédant une certaine caractéristique y . Par exemple, la proportion d'hommes de moins de 25 ans dans la population Canadienne ou la proportion de ménages dont le revenu est de moins de 30 000\$ au Canada ou encore la proportion d'individus étant atteinte d'un certain type de cancer.

Soit

$$y_i = \begin{cases} 1 & \text{si l'unité } i \text{ possède la caractéristique} \\ 0 & \text{si l'unité } i \text{ ne possède pas la caractéristique} \end{cases}$$

La proportion p_y peut s'écrire comme $p_y = \frac{1}{N} \sum_{i \in U} y_i$. Une proportion peut donc être vue comme un cas particulier de la moyenne dans la population \bar{y}_U lorsque la variable d'intérêt y prend les valeurs 0 ou 1 seulement. On dira que la variable y est *dichotomique*. Par la Proposition 2.2, il s'ensuit qu'un estimateur sans biais de p_y est donné par $\hat{p}_y = \bar{y}_s$, qui représente la proportion d'unités dans l'échantillon possédant la caractéristique d'intérêt. Avant de donner une expression de la variance \hat{p}_y , notons que dans le cas d'une variable dichotomique y , on a

$$\begin{aligned} S_y^2 &= \frac{1}{N-1} \sum_{i \in U} (y_i - p_y)^2 = \frac{1}{N-1} \left[\sum_{i \in U} y_i^2 - 2p_y \sum_{i \in U} y_i + Np_y^2 \right] \\ &= \frac{N}{N-1} p_y (1 - p_y), \end{aligned}$$

en notant que $y_i^2 = y_i$. Il s'ensuit (par la Proposition 2.3) que la variance de \hat{p}_y est donnée par

$$V(\hat{p}_y) = \frac{N}{N-1} \left(1 - \frac{n}{N}\right) \frac{p_y (1 - p_y)}{n}.$$

De manière similaire, notons que

$$s_y^2 = \frac{1}{n-1} \sum_{i \in s} (y_i - \hat{p}_y)^2 = \frac{n}{n-1} \hat{p}_y (1 - \hat{p}_y). \quad (2.3)$$

De (2.2) et (2.3), on déduit qu'un estimateur sans biais de $V(\hat{p}_y)$ est donné par

$$\hat{V}(\hat{p}_y) = \left(1 - \frac{n}{N}\right) \frac{\hat{p}_y (1 - \hat{p}_y)}{n-1}.$$

Exemple 2.4 : Lors d'un sondage mené auprès de 138 étudiants de l'Université de Montréal choisis parmi les étudiants d'un programme dont l'effectif est $N = 988$. On leur a demandé, entre autre, de s'exprimer sur l'énoncé « Le singe et l'homme ont un ancêtre commun ». Il y en 53 parmi les 138 qui se sont déclarés d'accord, les autres exprimant leur désaccord. Soit p_y la proportion d'étudiants du programme qui sont d'accord. L'estimation \hat{p}_y de p_y est

$$\hat{p}_y = \frac{53}{138} = 0,3841.$$

L'estimation de la variance de \hat{p}_y est

$$\hat{V}(\hat{p}_y) = \left(1 - \frac{n}{N}\right) \frac{\hat{p}_y(1 - \hat{p}_y)}{n-1} = \left(1 - \frac{138}{988}\right) \frac{53}{138} \left(1 - \frac{53}{138}\right) = 0,001485.$$

2.4 Intervalles de confiance

L'estimation des paramètres d'intérêt effectuée, il importe de déterminer une marge d'erreur afin de savoir à quel point l'estimation est erronée. En pratique, on construit fréquemment des intervalles de confiances (IC), des intervalles dont on peut dire avec un certain degré de confiance qu'ils contiennent la valeur réelle du paramètre.

2.4.1 Construction d'un intervalle

Dans cette section, on montre comment construire un IC. Supposons que l'on veuille estimer la moyenne dans la population, $\bar{y}_U = \frac{1}{N} \sum_{i \in U} y_i$, d'une variable d'intérêt y . On tire un EASSR de taille n . Comme nous l'avons vu à la section 2.3.3, la moyenne échantillonnale \bar{y}_s est un estimateur sans biais de \bar{y}_U . De plus, on a $V(\bar{y}_s) = \left(1 - \frac{n}{N}\right) \frac{S_y^2}{n}$.

Hajek (1960) a montré que, sous certaines conditions et si les tailles n , N , and $N - n$ sont « suffisamment grandes », alors

$$\frac{\bar{y}_s - \bar{y}_U}{\sqrt{\left(1 - \frac{n}{N}\right) \frac{S_y^2}{n}}} \sim N(0,1). \quad (2.4)$$

Le résultat établi en (2.4) est connu sous le nom de *Théorème Limite Central* dans le contexte de populations finies. Partant de (2.4), on peut écrire

$$P \left(-z_{\alpha/2} \leq \frac{\bar{y}_U - \bar{y}_s}{\sqrt{\left(1 - \frac{n}{N}\right) \frac{S_y^2}{n}}} \leq z_{\alpha/2} \right) = 1 - \alpha,$$

où $z_{\alpha/2}$ est le quantile d'ordre $(1 - \alpha/2)$ d'une variable aléatoire normale centrée réduite; i.e., $z_{\alpha/2}$ est tel que $P(N(0,1) > z_{\alpha/2}) = \alpha/2$. La probabilité $1 - \alpha$ est appelée *niveau de confiance* de l'intervalle. Un IC de niveau $1 - \alpha$ pour la moyenne de la population \bar{y}_U est donné par

$$\left[\bar{y}_s - z_{\alpha/2} \sqrt{\left(1 - \frac{n}{N}\right) \frac{S_y^2}{n}}, \bar{y}_s + z_{\alpha/2} \sqrt{\left(1 - \frac{n}{N}\right) \frac{S_y^2}{n}} \right]. \quad (2.5)$$

Malheureusement, la formule (2.5) n'est pas utilisable en pratique car la dispersion de la variable y dans l'ensemble de la population, S_y^2 , est inconnue. On estime donc S_y^2 , par la dispersion de la variable y dans l'échantillon, s_y^2 , ce qui mène à

$$\left[\bar{y}_s - z_{\alpha/2} \sqrt{\left(1 - \frac{n}{N}\right) \frac{s_y^2}{n}}, \bar{y}_s + z_{\alpha/2} \sqrt{\left(1 - \frac{n}{N}\right) \frac{s_y^2}{n}} \right]. \quad (2.6)$$

Remarques :

(1) Les bornes inférieures et supérieures de l'intervalle de confiance (2.6) sont aléatoires puisqu'elles sont fonction de l'estimateur et de sa variance estimée qui sont eux-mêmes des variables aléatoires. On ne peut donc pas être sûr que \bar{y}_U sera contenu dans l'intervalle puisque ce dernier est aléatoire.

(2) Des intervalles de confiance pour un total t_y ou pour une proportion p_y peuvent être construits de manière similaire.

Exemple 2.5 : Une grande firme de construction possède 120 maisons à différents stades de construction. La firme souhaite estimer la valeur totale des coûts encourus pour la construction des maisons. Pour cela, un échantillon aléatoire simple sans remise de 12 maisons est sélectionné et le coût pour chacune des 12 maisons est déterminé. Voici les résultats obtenus :

| | | |
|--------|--------|--------|
| 35 500 | 30 200 | 28 900 |
| 36 400 | 29 800 | 34 100 |
| 32 600 | 26 400 | 38 000 |
| 38 200 | 32 200 | 27 500 |

- (a) Estimer le coût total pour les 120 maisons.
 (b) Construire un intervalle de confiance de niveau 95% pour le coût total.

Solution :

- (a) Soit y le coût d'une maison. On cherche à estimer $t_y = \sum_{i=1}^{120} y_i$. L'estimation du coût total est donné par

$$\hat{t}_y = N\bar{y}_s = 120 \left(\frac{389\,800}{12} \right) = 3\,898\,000.$$

- (b) Un intervalle de confiance de niveau 95% pour t_y est donné par

$$\hat{t}_y \pm 1,96 \sqrt{N^2 \left(1 - \frac{n}{N} \right) \frac{s_y^2}{n}},$$

où $s_y^2 = \frac{1}{n-1} \sum_{i \in s} (y_i - \bar{y}_s)^2 = 16\,123\,334,34$. L'intervalle de confiance recherché est donc donné par $3\,898\,000 \pm 263\,918,17$ ou encore $(3\,634\,081,83, 4\,161\,918,17)$.

2.4.2 Interprétation

L'interprétation d'un intervalle de confiance pour des populations finies est légèrement différente de celle d'un intervalle en statistique classique.

En statistique classique (cas de populations infinies) on interprète un intervalle de niveau 95% comme suit : Si l'on tire un grand nombre d'échantillons de taille n et que pour chacun de ceux-ci, l'on construit un intervalle de confiance pour \bar{y}_U , on s'attend à ce que 95% des IC ainsi construits contiennent la vraie valeur du paramètre \bar{y}_U . Rappelons que, dans le contexte de populations infinies, il existe une infinité d'échantillons possibles de taille n .

Dans le cas de population finie, il existe un nombre fini d'échantillons possibles. Par exemple, dans le cas de l'EASSR, on a $\binom{N}{n}$ échantillons possibles. S'il était possible de générer tous les échantillons possibles, on pourrait alors construire un IC pour chacun de ceux-ci. Dans ce cas, on pourrait, pour chaque échantillon, déterminer si celui-ci contient la vraie valeur du paramètre \bar{y}_U (en supposant que l'on connaisse cette valeur). On pourrait alors déterminer le niveau de confiance exact de l'IC, qui est défini par la proportion des échantillons qui contiennent la vraie valeur du paramètre \bar{y}_U parmi les $\binom{N}{n}$ échantillons possibles. En pratique, on n'a qu'un seul échantillon à notre disposition et la valeur du paramètre \bar{y}_U n'est pas connue. Il est donc impossible de connaître exactement la proportion d'échantillons qui couvre la vraie valeur du paramètre. Nous n'avons alors d'autre choix que de faire appel au théorème limite central (qui suppose que la taille de l'échantillon est suffisamment grande), ce qui nous assure que le niveau de confiance est approximativement égal à 95%.

2.4.3 L'hypothèse de normalité

La construction d'IC valides repose sur l'hypothèse de normalité (2.4) qui nécessite une grande taille d'échantillon n . Qu'en est-il si la taille de l'échantillon n'est pas « suffisamment grande »? En effet, que peut-on dire lorsque $n=5$ ou $n=10$? La réponse à cette question n'est pas simple car la réponse dépend de la population. Nous illustrons ce point à l'aide des deux exemples suivants :

Exemple 2.6 : J'ai généré une population de taille $N=1000$ comprenant une variable d'intérêt y de moyenne $\bar{y}_U=100$. La figure 2.3 exhibe la distribution de la variable y dans la population. Il est clair que cette distribution est fortement asymétrique. De la population, j'ai tiré 50 000 EASSR de taille $n=5; 10; 30; 100$. Dans chaque échantillon, j'ai calculé la moyenne échantillonnale \bar{y}_s . Les figures 2.4-2.7 exhibent la distribution de \bar{y}_s . La distribution de \bar{y}_s est clairement asymétrique lorsque $n=5$ et légèrement asymétrique lorsque $n=10$. Avec $n=30$ et $n=100$, la distribution de \bar{y}_s est presque normale. Il est donc risqué de construire des intervalles de confiance reposant sur l'hypothèse de normalité de \bar{y}_s lorsque $n=5$. Dans ce cas, le taux de couverture de l'intervalle de confiance pourrait être considérablement différent de 95%. En effet, parmi les 50 000 échantillons tirés, seulement 81.5% des intervalles de confiance de niveau 95%, $\left[\bar{y}_s - 1,96 \sqrt{\left(1 - \frac{n}{N}\right) \frac{s_y^2}{n}}, \bar{y}_s + 1,96 \sqrt{\left(1 - \frac{n}{N}\right) \frac{s_y^2}{n}} \right]$ contenaient la valeur réelle de la moyenne $\bar{y}_U=100$. Avec $n=10$, les résultats sont plus légèrement meilleurs puisque 87,1% des intervalles contenaient la valeur réelle de la moyenne. Finalement avec $n=30$

et $n = 100$, les résultats sont plus encourageants puisque 92,2% et 94,1% des intervalles contenaient la valeur réelle de la moyenne, respectivement.

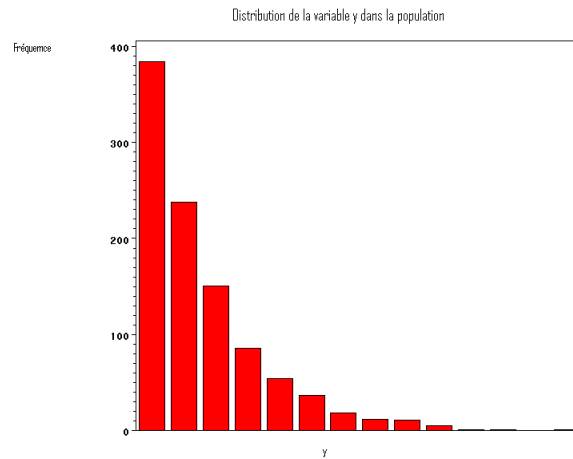


Figure 2.3

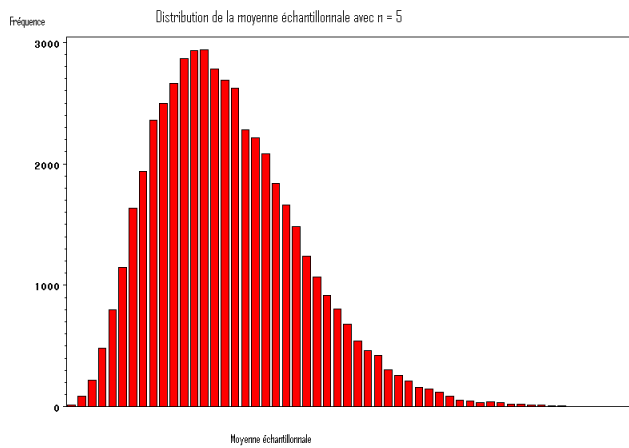


Figure 2.4

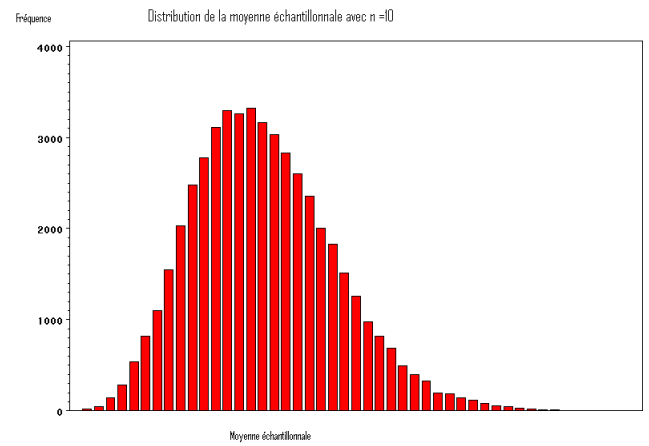


Figure 2.5

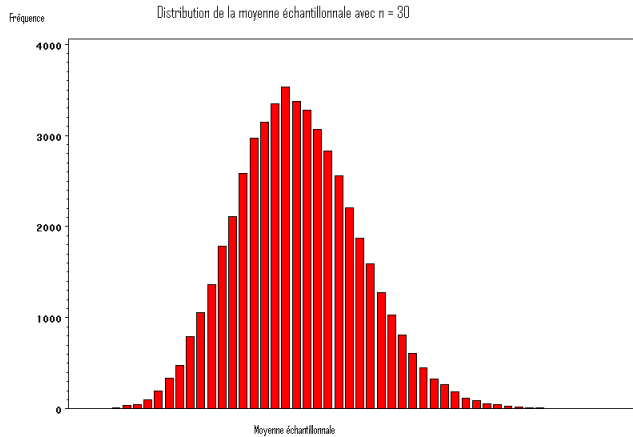


Figure 2.6

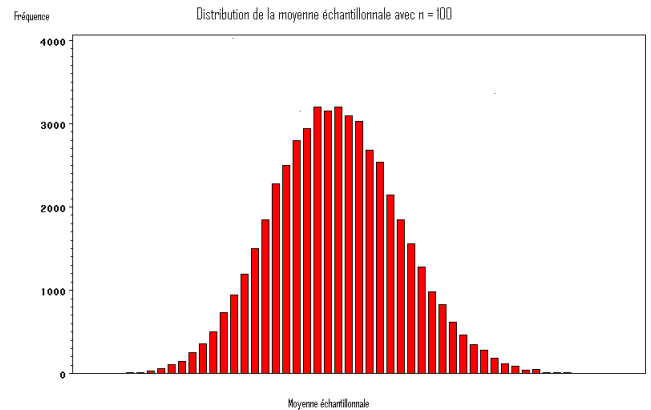


Figure 2.7

Exemple 2.7 : Comme dans l'exemple 2.4, j'ai généré une deuxième population de taille $N=1000$ comprenant une variable d'intérêt y de moyenne $\bar{y}_U=100$. La figure 2.8 exhibe la distribution de la variable y dans la population. Cette distribution semble être presque symétrique. De la population, j'ai tiré 50 000 EASSR de taille $n=5$ et $n=10$. Dans chaque échantillon, j'ai calculé la moyenne échantillonnale \bar{y}_s . Les figures 2.9 et 2.10 exhibent la distribution de \bar{y}_s . Déjà, avec $n=5$, la distribution de \bar{y}_s semble presque normale. Parmi les 50 000 échantillons tirés, seulement 88,2% des intervalles de confiance de contenaient la valeur réelle de la moyenne $\bar{y}_U=100$. Avec $n=10$, la distribution de \bar{y}_s semble, encore une fois, presque normale. Dans ce cas, 92,2% des intervalles de confiance contenaient la valeur réelle de la moyenne $\bar{y}_U=100$.

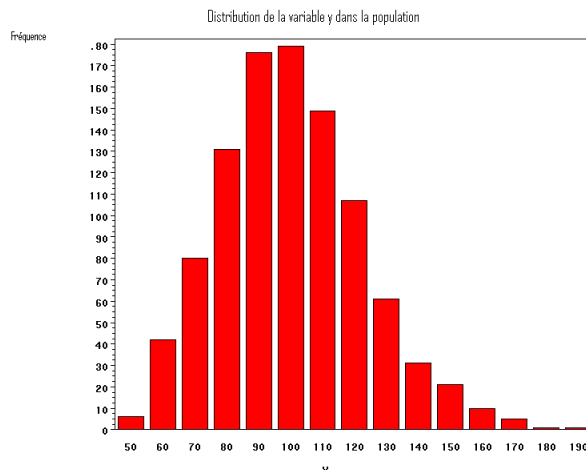


Figure 2.8

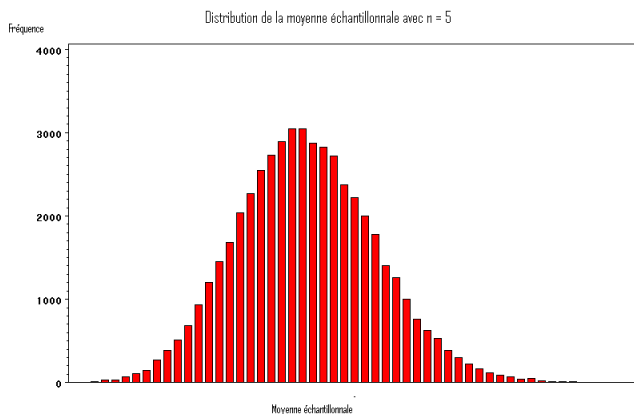


Figure 2.9

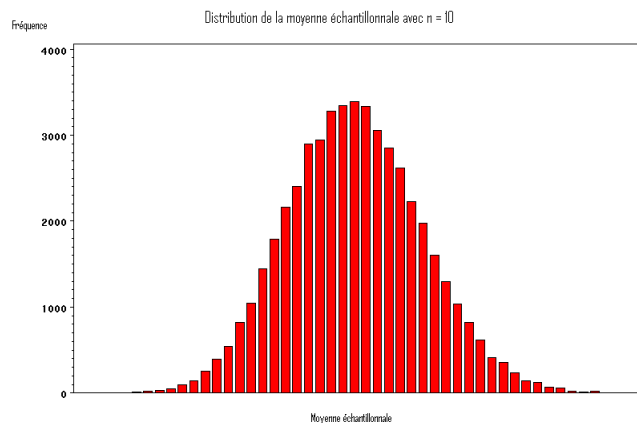


Figure 2.10

Les deux exemples précédents illustrent bien la problématique lorsqu'il s'agit de répondre à la question : À partir de quelle valeur juge-t-on que la taille de l'échantillon est « suffisamment grande »? Plus la distribution de la variable d'intérêt dans la population est différente d'une distribution normale, plus la taille de l'échantillon devra être grande afin de garantir que la distribution de la moyenne échantillonnale \bar{y}_s est presque normale. Il n'y a donc pas de « recette miracle ». Dans un premier cours de statistique, il est coutume d'utiliser la taille $n = 30$ comme seuil garantissant la normalité de la distribution de la moyenne échantillonnale, ce qui est réaliste la plupart du temps. Il ne faut cependant pas perdre de vue que, dans certains cas, cette taille n n'est pas suffisante pour garantir que la distribution de \bar{y}_s est normale.

2.5 Calcul de la taille de l'échantillon

Une question importante qui surgit en pratique est celle de la taille de l'échantillon. Pour satisfaire un niveau de précision donné, quelle taille n doit-on utiliser? Il est évident que plus la taille de l'échantillon est grande, plus le sondage est précis. Cependant, nous ne disposons pas d'un budget illimité. Il s'agira donc de préalablement déterminer notre niveau de tolérance à l'erreur. Soit e la marge d'erreur tolérée. En pratique, la marge d'erreur est fréquemment une valeur entre 1% et 5%.

2.5.1 Cas d'une moyenne

Le problème se pose comme suit : Il faut déterminer la taille n de l'échantillon telle que la moyenne échantillonnale \bar{y}_s se situe à moins de e de la moyenne de la population \bar{y}_U avec une probabilité de $1 - \alpha$.

$$P(|\bar{y}_s - \bar{y}_U| \leq e) = 1 - \alpha. \quad (2.7)$$

Habituellement, la valeur $\alpha = 0,05$ est utilisée auquel cas le niveau de confiance est égal à 95%. Il reste à déterminer la taille n qui satisfait (2.7). Après quelques manipulations algébriques, on déduit de (2.7)

$$P(-e \leq \bar{y}_s - \bar{y}_U \leq e) = 1 - \alpha,$$

ou, de façon équivalente

$$P\left(\frac{-e}{\sqrt{V(\bar{y}_s)}} \leq N(0,1) \leq \frac{e}{\sqrt{V(\bar{y}_s)}}\right) = 1 - \alpha.$$

On obtient alors :

$$\frac{e}{\sqrt{V(\bar{y}_s)}} = z_{\alpha/2}$$

ou

$$\begin{aligned} e &= z_{\alpha/2} \sqrt{V(\bar{y}_s)} \\ &= z_{\alpha/2} \sqrt{\left(1 - \frac{n}{N}\right) \frac{S_y^2}{n}}. \end{aligned} \quad (2.8)$$

On détermine aisément la taille n , en résolvant l'équation (2.8), ce qui mène à

$$n = \frac{z_{\alpha/2}^2 S_y^2}{e^2 + \frac{z_{\alpha/2}^2 S_y^2}{N}}. \quad (2.9)$$

Malheureusement, la dispersion de la variable y , S_y^2 , en (2.9) est inconnue. Il s'agit donc de donner une estimation a priori de S_y^2 . Pour cela, on peut faire appel à une des méthodes suivantes :

- (i) Utiliser une estimation obtenue dans le cadre de l'enquête pilote.
- (ii) Utiliser l'estimation obtenue dans des enquêtes similaires effectuées par d'autres organismes statistiques.
- (iii) Si une variable auxiliaire x est disponible pour toutes les unités dans la population, on connaît alors la dispersion de celle-ci dans l'ensemble de la population, S_x^2 . De plus, si l'on sait que la variable d'intérêt y est proportionnelle à x (i.e., $y_i = cx_i$), on aura $S_y^2 = c^2 S_x^2$, donnant une valeur de S_y^2 qui pourra être utilisée. En pratique, il est irréaliste de supposer l'existence d'une relation exacte entre les variables y et x . Cependant, si les variables y et x sont fortement corrélées, alors l'estimation de S_y^2 sera vraisemblablement de bonne qualité.

2.5.2 Cas d'une proportion

Rappelons que dans le cas d'une proportion, on a $p_y = \bar{y}_U$, où y désigne une variable dichotomique. Dans ce cas, l'expression (2.9) s'écrit comme

$$n = \frac{z_{\alpha/2}^2 \frac{N}{N-1} p_y (1-p_y)}{e^2 + z_{\alpha/2}^2 \frac{p_y (1-p_y)}{N-1}}, \quad (2.10)$$

en se rappelant que $S_y^2 = \frac{N}{N-1} p_y (1-p_y)$. Encore une fois, on fait face au fait que p_y est inconnue. On peut alors faire appel à l'une des méthodes décrites pour une moyenne à la section 2.5.1. Si aucune estimation a priori raisonnable de p_y n'est disponible (ce qui est relativement rare en pratique), il est coutume d'utiliser $p_y = \frac{1}{2}$ comme estimation a

priori. En effet, la fonction $p_y(1-p_y)$ possède un maximum $p_y = \frac{1}{2}$ (voir Figure 2.11). La taille n obtenue est donc conservatrice.

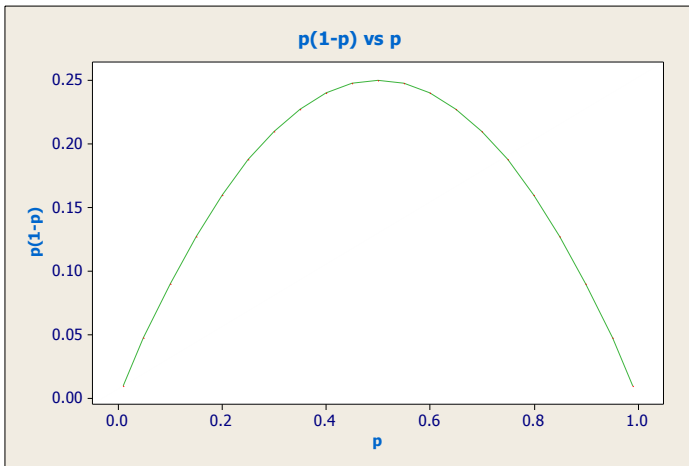


Figure 2.11 : Graphique de $p_y(1-p_y)$ en fonction de p_y .

Exemple 2.8 : Le responsable de marketing dans une compagnie de boissons gazeuses cherche à estimer la proportion p_y de consommateurs qui préfère un nouveau produit destiné à remplacer l'ancien. Quelle est la taille de l'échantillon qu'il doit prélever s'il veut que la marge d'erreur de son estimateur ne soit pas supérieure à 0,04? En considérant que la population est à toute fin pratique infinie, le numérateur dans (2.10) est approximativement égal à e^2 . N'ayant pas d'idées préalables à propos de la proportion p_y , on utilise $p_y = 1/2$ dans (2.10), ce qui mène à

$$n = \frac{(1,96)^2 \frac{1}{2} \left(1 - \frac{1}{2}\right)}{0,04^2} = 600,25 \approx 601.$$

L'estimation que nous avons faite dans le dernier exemple est correcte si de fait $p_y = 1/2$: la marge d'erreur serait alors égale à 0,04. Lorsque $p_y \leq 1/2$, l'estimation est pessimiste : elle donne une valeur de n plus grande que nécessaire. Le tableau suivant donne la taille de l'échantillon nécessaire sous plusieurs hypothèses concernant p_y . On constate qu'on ne commet pas une erreur très importante lorsque p_y n'est pas trop éloignée de 0,5. Mais lorsque p_y est petite, on surestime n considérablement.

| | | | | | | | |
|-------|------|------|------|------|------|------|------|
| p_y | 0,01 | 0,05 | 0,10 | 0,20 | 0,30 | 0,40 | 0,50 |
| n | 24 | 115 | 217 | 385 | 505 | 577 | 601 |

2.6 L'échantillonnage de Bernoulli

L'échantillonnage de Bernoulli (BE) est un plan de sondage extrêmement simple. Ce plan de sondage peut être décrit comme suit : Soit n_s la taille de l'échantillon s et soit π la probabilité d'inclusion de l'unité i , $i = 1, \dots, N$; i.e., $\pi_i = \pi, i = 1, \dots, N$. Notons que la probabilité d'inclusion est constante pour tout i .

2.6.1 Définition

Pour tirer un échantillon selon un plan BE, on procède selon les étapes suivantes :

- (i) On génère des variables aléatoires indépendantes et identiquement distribuées, $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_N$, à partir d'une distribution uniforme de paramètre 0 et 1.
- (ii) Si $\varepsilon_i < \pi$, l'unité i est sélectionnée; sinon l'unité i n'est pas sélectionnée.

La probabilité de tirer un échantillon s est alors

$$p(s) = \pi^{n_s} (1 - \pi)^{N - n_s}.$$

Remarques :

(1) Dans un plan BE, le nombre d'échantillons possibles est 2^N .

(2) La taille de l'échantillon n_s est aléatoire. Nous utilisons la notation n_s au lieu de n (comme c'est le cas pour l'EASSR qui est un plan à taille fixe) pour bien indiquer que la taille de l'échantillon varie d'un échantillon à un autre. Le plan de BE est donc un *plan à taille aléatoire*. En fait, on a $n_s \sim B(N, \pi)$. On en déduit les identités suivantes :

$$P(n_s = n) = \binom{N}{n} \pi^n (1 - \pi)^{N - n},$$

$$E(n_s) = N\pi$$

et

$$V(n_s) = N\pi(1 - \pi).$$

Dans le cas d'un plan BE, il n'est pas impossible d'observer $n_s = 0$ ou $n_s = N$ avec probabilités

$P(n_s = 0) = (1 - \pi)^N$ ou $P(n_s = N) = \pi^N$, respectivement, auxquels cas on a $s = \emptyset$ or $s = U$.

(3) L'étendue des valeurs probables n_s peut être évaluée au moyen d'un intervalle. Lorsque la taille N de la population est suffisamment grande, on peut utiliser une approximation par une loi normale, ce qui donne un intervalle pour n_s de la forme

$$\left[N\pi - z_{\alpha/2}\sqrt{N\pi(1-\pi)}; N\pi + z_{\alpha/2}\sqrt{N\pi(1-\pi)} \right]$$

Par exemple, lorsque $N = 10000$ et $\pi = 0,2$ un intervalle de niveau 99% est donné par

$$\left[2000 - 2,58\sqrt{1600}; 2000 + 2,58\sqrt{1600} \right] = [1897; 2103].$$

(4) Dans le cas d'un plan BE, notons que les variables indicatrices de sélection Z_i et Z_j pour les unités i et j , sont indépendantes. On a alors

$$\pi_{ij} = P(i \in s \text{ et } j \in s) = P(Z_i = 1 \text{ et } Z_j = 1) = P(Z_i = 1)P(Z_j = 1) = \pi_i\pi_j = \pi^2.$$

2.6.2 Estimation

Supposons que l'on veuille estimer le total dans la population, $t_y = \sum_{i \in U} y_i$, d'une variable d'intérêt y . Quel estimateur utiliser? La proposition suivante répond à cette question.

Proposition 2.5 : Dans le cas d'un plan BE, un estimateur sans biais d'un total t_y est donné par

$$\hat{t}_y = \frac{1}{\pi} \sum_{i \in s} y_i. \quad (2.11)$$

Démonstration : laissée en exercice.

Proposition 2.6 : Dans le cas d'un plan BE, la variance de \hat{t}_y est donnée par

$$V(\hat{t}_y) = \left(\frac{1}{\pi} - 1 \right) \sum_{i \in U} y_i^2. \quad (2.12)$$

Démonstration : laissée en exercice.

Bien sûr, la variance en (2.12) ne peut être calculée au moyen des unités échantillonnées seulement car elle dépend des valeurs de la variable y pour l'ensemble de la population. La proposition suivante propose un estimateur sans biais de $V(\hat{t}_y)$.

Proposition 2.7 : Un estimateur sans biais de $V(\hat{t}_y)$ en (2.12) est donné par

$$\hat{V}(\hat{t}_y) = \frac{1}{\pi} \left(\frac{1}{\pi} - 1 \right) \sum_{i \in S} y_i^2; \quad (2.13)$$

i.e., $E(\hat{V}(\hat{t}_y)) = V(\hat{t}_y)$.

Démonstration : laissée en exercice.

Remarque : Un estimateur de la moyenne \bar{y}_U découle aisément de la Proposition 2.5 en notant que $\bar{y}_U = \frac{t_y}{N}$. Si la taille de la population N est connue, on utilisera l'estimateur

$\hat{y}_U = \frac{\hat{t}_y}{N}$. Sa variance est égale à $V(\hat{y}_U) = \frac{1}{N^2} V(\hat{t}_y)$ et un estimateur sans biais de $V(\hat{y}_U)$ est donné par $\hat{V}(\hat{y}_U) = \frac{1}{N^2} \hat{V}(\hat{t}_y)$.

Exemple 2.9 : Un professeur d'université doit corriger 600 copies d'examen. Pour se faire une idée préliminaire du taux de réussite, il utilise le plan suivant. Pour chaque copie, il lance un dé ordinaire. Si le résultat observé est 6, il corrige l'examen. Sinon, il procède en passant à la prochaine copie. À la fin du processus, il trouve que 60, parmi 90 étudiants corrigés, ont obtenu la note de passage. À l'aide d'un intervalle de confiance, au niveau 95%, estimer le nombre d'étudiants qui ont réussi l'examen.

Solution : Notons d'abord que le plan utilise par le professeur est le plan BE avec $\pi = \frac{1}{6}$.

Soit y une variable dichotomique telle que $y_i = 1$ si l'étudiant i a réussi l'examen et $y_i = 0$ si l'étudiant i a échoué. On cherche donc à déterminer un intervalle de confiance pour $t_y = \sum_{i \in U} y_i$. On a donc $\hat{t}_y = 6 \sum_{i \in S} y_i = 6 \times 60 = 360$. Maintenant, un estimateur de la variance

de $V(\hat{t}_y)$ est égal à $\hat{V}(\hat{t}_y) = \frac{1}{\pi} \left(\frac{1}{\pi} - 1 \right) \sum_{i \in S} y_i^2 = 6 \times 5 \times 60 = 1\,800$ puisque

$\sum_{i \in S} y_i = \sum_{i \in S} y_i^2 = 60$. Un intervalle de confiance de niveau 95% est donné par

$$\left[360 - 1,96\sqrt{1800}; 360 + 1,96\sqrt{1800} \right] = [277; 443].$$

2.7 L'échantillonnage systématique

L'échantillonnage systématique (SY) est une procédure d'échantillonnage qui offre plusieurs avantages pratiques dont sa simplicité d'exécution sur le terrain. C'est pourquoi, l'échantillonnage SY est fréquemment utilisé dans les enquêtes, particulièrement dans les enquêtes auprès des ménages comme l'Enquête sur la Population Active Canadienne. Pour introduire le plan SY on considère l'exemple suivant:

Un médecin souhaite avoir une estimation de l'âge moyen de ses patients. Pour chaque patient, le médecin dispose d'une fiche comprenant plusieurs variables dont l'âge, le sexe, le poids et la taille. Toutes les fiches sont rangées dans le tiroir d'un classeur en ordre alphabétique. Supposons que le médecin a 1000 patients; i.e., $N = 1000$. Le médecin veut obtenir une estimation à partir d'un échantillon de $n = 100$ patients. Bien sûr, le médecin pourrait toujours tirer un EASSR dans la population mais ce n'est pas une méthode très pratique dans ce contexte. Il décide plutôt d'opter pour une procédure qui consiste à d'abord choisir un intervalle approprié. Ici, un intervalle approprié est de longueur 10. Ensuite, il choisit un nombre aléatoirement entre 1 et 10. Supposons qu'il ait choisi le nombre cinq. Alors, le premier élément sélectionné dans l'échantillon est la cinquième fiche, le deuxième est la quinzième fiche, le troisième élément est la vingt cinquième fiche, etc. Le médecin va donc tirer systématiquement chaque dixième fiche dans le tiroir. Le seul aspect aléatoire du tirage est le choix du premier élément tiré. Ce mode de tirage est appelé *échantillonnage systématique*.

2.7.1 Définition

Un échantillon qui est obtenu en tirant au hasard le premier élément parmi les k premiers éléments dans la base de sondage et de choisir systématiquement chaque k^e élément systématiquement est un échantillon systématique 1-dans- k .

L'échantillonnage SY est une alternative intéressante à l'EASSR pour les raisons suivantes :

(i) L'échantillonnage SY présente une facilité d'exécution sur le terrain et facilite le travail des interviewers qui vont systématiquement choisir chaque k^e logement dans une enquête auprès des ménages. Par exemple, dans le cadre du recensement canadien, un questionnaire long est distribué à chaque k^e logement.

(ii) Si les unités de sondages sont ordonnées dans la base de sondage selon un certains critère, l'échantillonnage SY permettra d'obtenir des unités pour chaque partie de la population. Par exemple, supposons que dans une enquête auprès des entreprises, celles-ci soient ordonnées en ordre croissant du nombre d'employés. Alors, l'échantillonnage SY permet de parcourir la liste et ce faisant, permettra de choisir des entreprises avec un

petit nombre d'employés, des entreprises avec un nombre d'employés moyen et des entreprises avec un grand nombre d'employés, ce qui n'est pas garanti avec l'EASSR.

2.7.2 Sélection de l'échantillon

Comment tire-t-on un échantillon SY? Si l'on cherche à tirer un échantillon SY 1-dans- k , on procèdera selon les étapes suivantes :

- (i) On tire au hasard une unité parmi les k premières. La première unité tirée est le *départ aléatoire*.
- (ii) Après sélection de la première unité, on choisit systématiquement chaque k^{e} unité et on parcourt ainsi toute la population. Le nombre k est le *pas d'échantillonnage*.

Comment choisit-on k ? Lorsque la fraction $\frac{N}{n}$ est un nombre entier, il est coutume de poser $k = \frac{N}{n}$, ce qui nous assurera qu'exactly n unités ont été tirées dans l'échantillon. Par exemple, si $N = 15\ 000$, et l'on désire une taille d'échantillon $n = 100$, on posera $k = \frac{15\ 000}{100} = 150$; i.e., l'échantillon sera un échantillon systématique 1-dans-150. Dans le cas où $\frac{N}{n}$ n'est pas un nombre entier, on peut utiliser (mais il ne sera pas considéré ici) *l'échantillonnage systématique circulaire*.

Il nous reste à déterminer la probabilité $p(s)$ d'un échantillon s donné. Pour cela, considérons l'exemple suivant.

Exemple 2.10 : On dispose d'une population de taille $N = 20$. On veut tirer un échantillon selon plan SY de taille $n = 4$. Voici la population :

4 8 1 7 6 3 10 9 4 7 5 12 2 8 4 1 11 3 6 5

D'abord, nous posons $k = \frac{N}{n} = \frac{20}{4} = 5$. On sélectionne donc un échantillon SY 1-dans-5.

Pour cela, on tire au hasard une unité parmi les cinq premières unités et chaque cinquième unité est systématiquement sélectionnée. La liste des échantillons possibles est exhibée ci-dessous :

$$s_1 = \{4, 3, 5, 1\}, s_2 = \{8, 10, 12, 11\}, s_3 = \{1, 9, 2, 3\}, s_4 = \{7, 4, 8, 6\}, s_5 = \{6, 7, 4, 5\}$$

Il y a donc cinq échantillons possibles de taille 4. La probabilité de tirer n'importe lequel de ces échantillons est donc égale à $\frac{1}{5}$.

En général, lorsque l'on tire un échantillon SY 1-dans- k , il y a k échantillons équiprobables possibles, chacun pouvant être tiré avec probabilité $\frac{1}{k}$. Autrement dit, on a

$$\Omega = \{s_1, s_2, \dots, s_k\} \text{ et } p(s) = \frac{1}{k} \text{ pour tout } s \in \Omega.$$

Remarques :

(1) La probabilité d'inclusion π_i est égale à $\pi_i = \frac{1}{k}$, pour tout $i \in U$ puisque chaque unité i appartient à un et un seul échantillon parmi les k échantillons possibles. De plus, pour tout i et j tel que $i \neq j$, on a

$$\pi_{ij} = \begin{cases} 1/k & \text{si } i \text{ et } j \text{ appartiennent au même échantillon } s \\ 0 & \text{sinon} \end{cases}$$

Donc, dans le cas de l'échantillonnage SY, certains couples (i, j) ont une probabilité d'inclusion jointe égale à 0, ce qui posera un problème lorsqu'il s'agira d'estimer la variance des estimateurs.

(2) Les échantillons s_1, s_2, \dots, s_k forment une partition de la population; i.e., $U = \bigcup_{g=1}^k s_g$ et $s_g \cap s_{g'} = \emptyset$ si $g \neq g'$. Il s'ensuit que le total dans la population d'une variable d'intérêt y , t_y , peut s'écrire comme $t_y = \sum_{g=1}^k t_g$, où $t_g = \sum_{i \in s_g} y_i$.

(3) Le nombre d'échantillons possibles dans un plan SY, k , est considérablement plus petit que le nombre d'échantillons possibles dans le cas de l'EASSR qui est égal à $\binom{N}{n}$, ou que le nombre d'échantillons possibles dans le cas d'un plan BE qui est égal à 2^N .

2.7.3 Estimation

Dans cette section, on considère le problème de l'estimation d'un total t_y . Quel estimateur utiliser? La proposition suivante répond à cette question.

Proposition 2.8 : Un estimateur sans biais de t_y est donné par

$$\hat{t}_y = kt_s, \quad (2.14)$$

où $t_s = \sum_{i \in s} y_i$ désigne le total de la variable y dans l'échantillon et s est membre de l'ensemble de tous les échantillons possibles $\Omega = \{s_1, \dots, s_g, \dots, s_k\}$.

Démonstration :

Proposition 2.9 : La variance de \hat{t}_y en (2.14) est donnée par

$$V(\hat{t}_y) = k(k-1)S_t^2, \quad (2.15)$$

où $S_t^2 = \frac{1}{k-1} \sum_{g=1}^k (t_g - \bar{t})^2$ avec $\bar{t} = \frac{1}{k} \sum_{g=1}^k t_g$.

Démonstration :

L'expression (2.15) montre que la variance de l'estimateur \hat{t}_y est grande lorsque les totaux $t_1, \dots, t_g, \dots, t_k$ sont grandement dispersés. La variance (2.15) sera étudiée en détail à la section 2.7.4. Comme nous l'avons noté à la section 2.7.2, certains couples (i, j) ont une probabilité d'inclusion jointe égale à 0. Dans ce cas, il n'existe pas d'estimateur sans biais de la variance (2.15). Si l'ordre des unités sur la base de sondage est arbitraire, il est coutume d'utiliser l'estimateur de la variance que l'on aurait obtenu si l'échantillon avait été tiré selon un plan EASSR, ce qui mène à l'estimateur de variance

$$\hat{V}(\hat{t}_y) = N^2 \left(1 - \frac{n}{N}\right) \frac{s_y^2}{n}.$$

Ce point sera éclairci à la section 2.7.4.

Remarque : Si l'on cherche à estimer la moyenne dans la population $\bar{y}_U = \frac{t_y}{N}$, on utilisera l'estimateur $\hat{\bar{y}}_U = \frac{\hat{t}_y}{N}$ dont la variance est égale à $V(\hat{\bar{y}}_U) = \frac{1}{N^2} V(\hat{t}_y)$.

Exemple 2.11 : Afin de connaître le niveau de scolarité de l'auditoire d'une pièce de théâtre, le metteur en scène décide de tirer un échantillon aléatoire de spectateurs. Il s'installe à la porte du théâtre et interroge, disons, chaque dixième spectateur à son arrivée en salle. Remarquons que nous n'avons pas besoin de connaître la taille de la population pour faire ce type d'échantillonnage, et c'est là un avantage de l'échantillonnage systématique. Par contre, dans ce cas, la taille de l'échantillon est aléatoire puisqu'on ne connaît pas, à l'avance, le nombre de spectateurs qui vont assister à la représentation.

Exemple 2.12 : Une compagnie de pâtes et papier est intéressée à estimer le volume de bois moyen par terrain d'un demi-hectare. On compte 525 terrains. Un échantillon systématique 1-dans-25 de terrains est tiré. Les données sont présentées dans le tableau

ci-dessous. Estimer le volume moyen de bois par terrain, \bar{y}_U , et construire un intervalle de confiance de niveau 95% pour \bar{y}_U .

| Terrain tiré | Volume (en pieds cube) | Terrain tiré | Volume (en pieds cube) |
|--------------|---------------------------|--------------|---------------------------|
| 4 | 7 030 | 279 | 7 540 |
| 29 | 6 720 | 304 | 6 720 |
| 54 | 6 850 | 329 | 6 900 |
| 79 | 7 210 | 354 | 7 200 |
| 104 | 7 150 | 379 | 7 100 |
| 129 | 7 370 | 404 | 6 860 |
| 154 | 7 000 | 429 | 6 800 |
| 179 | 6 390 | 454 | 7 050 |
| 204 | 6 570 | 479 | 7 420 |
| 229 | 6 910 | 504 | 7 090 |
| 254 | 7 380 | | |

On a $n = 21$, $\sum_{i \in S} y_i = 147\,260$, $\sum_{i \in S} y_i^2 = 1\,034\,331\,400$, $s_y^2 = 84\,408,84$. Une estimation de

\bar{y}_U est donnée par $\hat{\bar{y}}_U = \frac{1}{n} \sum_{i \in S} y_i = \frac{1}{21} 147\,260 = 7012,38$. Un intervalle de confiance pour

\bar{y}_U est donné par $\hat{\bar{y}}_U \pm 1,96 \sqrt{\hat{V}(\hat{\bar{y}}_U)}$, où $\hat{V}(\hat{\bar{y}}_U) = \left(1 - \frac{n}{N}\right) \frac{s_y^2}{n} = 3858,68$. L'intervalle de

confiance recherché est donc donné par $7012,38 \pm 1,96 \sqrt{3858,68}$. Notons que nous avons utilisé l'estimateur de la variance qu'on aurait obtenu si l'échantillon avait été tiré selon l'EASSR.

2.7.4 Comparaison de l'échantillonnage systématique et d l'échantillonnage aléatoire simple sans remise

Dans cette section, on compare l'échantillonnage systématique et l'échantillonnage aléatoire simple sans remise en termes d'efficacité. La variance en (2.15) est petite lorsque les totaux t_g sont approximativement égaux. Autrement dit, si l'ordre des unités qui prévaut dans la population est tel que tous les échantillons possibles ont approximativement le même total (des valeurs de y), alors la variance des estimateurs sera petite. Afin de mieux comprendre ce phénomène, nous considérons maintenant une décomposition de la variance similaire à celle que l'on retrouve dans le contexte de l'analyse de variance.

Prenons le cas $k = \frac{N}{n}$. L'estimateur \hat{t}_y en (2.14) est alors donné par $\hat{t}_y = N\bar{y}_s$. Afin

d'analyser plus en profondeur la variance (2.15), notons d'abord que la dispersion totale de la variable y dans la population peut se décomposer comme suit :

$$(N-1)S_y^2 \equiv SST = \sum_{i \in U} (y_i - \bar{y}_U)^2 = \sum_{g=1}^k \sum_{i \in s} (y_i - \bar{y}_{s_g})^2 + n \sum_{g=1}^k (\bar{y}_{s_g} - \bar{y}_U)^2$$

$$\equiv SSW + SSB, \quad (2.16)$$

où $\bar{y}_{s_g} = \frac{t_g}{n}$ et SST , SSW et SSB désignent respectivement, la somme des carrés « totale », la somme des carrés « intra » et la somme des carrés « inter ». Autrement dit, la variabilité totale dans l'ensemble de la population de la variable d'intérêt y peut être décomposée en deux termes : la variabilité des moyennes entre les échantillons et la somme des variabilités de la variable y à l'intérieur de chaque échantillon. Notons que, pour une population donnée, la dispersion S_y^2 est un nombre fixe. Par conséquent, un accroissement de SSW se traduit par la diminution correspondante de SSB . De plus, notons que la variance (2.15) peut s'écrire comme

$$V(\hat{t}_y) = N \cdot SSB.$$

Autrement dit, dans le cas d'un plan SY, la variance de \hat{t}_y sera petite lorsque les groupes ont approximativement la même moyenne. On espère donc que SSW sera la plus grande possible, ce qui entraînera que SSB sera petite. Cela veut donc dire que l'on cherche à avoir le moins d'homogénéité à l'intérieur des échantillons.

Comment mesure t-on l'homogénéité à l'intérieur d'un échantillon? On définit maintenant le coefficient de corrélation intra-classe (ICC) qui est une mesure de l'homogénéité à l'intérieur d'un échantillon et qui nous informe sur la similarité ou la dissimilarité des éléments à l'intérieur d'un groupe. Le coefficient de corrélation intra-classe est défini selon

$$ICC = 1 - \frac{n}{n-1} \frac{SSW}{SST} \quad (2.17)$$

Puisque $0 \leq \frac{SSW}{SST} \leq 1$, on déduit de (2.17) que

$$-\frac{1}{n-1} \leq ICC \leq 1.$$

D'une part, si les échantillons sont parfaitement homogènes, on a $SSW = 0$ et alors $ICC = 1$. D'autre part, si les échantillons ont exactement la même moyenne, on

$$SSB = 0 \text{ et alors } ICC = -\frac{1}{n-1}.$$

Comment l'échantillonnage systématique se compare-t-il à l'échantillonnage aléatoire simple sans remise en terme de variance? Nous répondons maintenant à cette question. Désignons par $V_{SY}(\hat{t}_y)$ la variance de l'estimateur \hat{t}_y lorsque le plan SY a été utilisé pour la sélection de l'échantillon et par $V_{EASSR}(\hat{t}_y)$ la variance de l'estimateur \hat{t}_y lorsque le plan EASSR a été utilisé pour la sélection de l'échantillon. Afin de comparer les deux plans, on définit une mesure utile appelée *effet de plan* (*design effect* en anglais).

Définition 2.4 : L'effet de plan, que l'on désigne par *deff*, est le rapport entre la variance d'un estimateur, selon un plan de sondage donné, et la variance du même estimateur d'un EASSR de même taille.

Dans le cas du plan SY, on peut montrer (la preuve est laissée en exercice) que l'effet de plan est donné par

$$\begin{aligned} deff &= \frac{V_{SY}(\hat{t}_y)}{V_{EASSR}(\hat{t}_y)} = \left(\frac{N-1}{N}\right) \left(1 - \frac{n}{N}\right)^{-1} [1 + (n-1)ICC] \\ &\approx [1 + (n-1)ICC] \end{aligned} \quad (2.18)$$

si la fraction de sondage, n/N , est négligeable (i.e., $n/N \approx 0$) et si pose $N-1 = N$. L'expression (2.18) suggère que lorsque $ICC \approx 1$, alors $deff \approx n$, auquel cas l'échantillonnage systématique s'avère très inefficace par rapport à l'échantillonnage aléatoire simple sans remise, particulièrement si la taille n de l'échantillon est grande. Par contre lorsque $ICC \approx 0$, alors $deff \approx 1$. Dans ce cas, les deux plans ont approximativement la même efficacité. Finalement, lorsque $ICC \leq 0$, on a $deff \leq 1$, auquel cas l'échantillonnage systématique est plus efficace que l'échantillonnage aléatoire simple sans remise.

Remarques :

(1) Si l'on permute aléatoirement les unités avant de tirer l'échantillon, il est raisonnable de penser (si le hasard a bien fait les choses!) que $ICC \approx 0$. Dans ce cas, on aura $deff \approx 1$, ce qui explique pourquoi il est raisonnable d'utiliser l'estimateur de variance (2.2) obtenu dans le cas de l'EASSR.

(2) Supposons que nous soyons intéressés à estimer le volume de transactions hebdomadaire dans une grande chaîne d'épicerie. Le volume de transactions est cependant extrêmement différent selon le jour de la semaine. On s'attend à ce que le volume de transactions soit très élevé du jeudi au samedi alors qu'il sera beaucoup moins important du dimanche au mercredi. Dans ce cas, il y a donc un problème de périodicité. Supposons que le plan SY est utilisé; un jour de la semaine, qui servira de départ aléatoire, est tiré au hasard parmi les sept jours de la semaine. Ensuite, chaque septième jour est systématiquement sélectionné. Si le jour de la semaine tiré est le mardi, on sous-

estimera vraisemblablement le volume de transactions hebdomadaire alors que si le jour tiré est le vendredi, on surestimera vraisemblablement le volume de transactions hebdomadaire. Pour contrer ce problème qui est inhérent aux populations périodiques, il est coutume de changer de départ aléatoire à plusieurs reprises. Cette procédure permettra de réduire la probabilité de tirer des observations à la même position dans le cycle.

2.8 Les poids de sondage et l'estimateur de Horvitz-Thompson

Jusqu'à maintenant, nous avons vu que, pour estimer un total $t_y = \sum_{i \in U} y_i$, un estimateur sans biais est donné par

- (i) $\hat{t}_y = N\bar{y}_s = \frac{N}{n} \sum_{i \in s} y_i$ pour un plan EASSR (section 2.3.4).
- (ii) $\hat{t}_y = \sum_{i \in s} \frac{y_i}{\pi}$ pour un plan BE (section 2.6.2).
- (iii) $\hat{t}_y = kt_s = k \sum_{i \in s} y_i$ pour un plan SY (section 2.7.3).

Ces trois estimateurs sont des cas particuliers de *l'estimateur de Horvitz-Thompson* (Horvitz et Thompson, 1952)

$$\hat{t}_y = \sum_{i \in s} \frac{y_i}{\pi_i}, \quad (2.19)$$

puisque $\pi_i = \frac{n}{N}$ dans le cas de l'EASSR, $\pi_i = \pi$ pour un plan BE et $\pi_i = \frac{1}{k}$ pour un plan SY. L'estimateur (2.19) est aussi appelé *estimateur par dilatation*.

Définition 2.5 : Le poids de sondage de l'unité i , que l'on désigne par w_i , est défini comme l'inverse de la probabilité d'inclusion π_i ; i.e., $w_i = \frac{1}{\pi_i}$.

Chaque unité échantillonnée représente un certain nombre d'unités dans la population. Par exemple, si le poids de sondage d'une unité est 5, alors cette dernière représente 5 unités dans la population (elle-même et quatre autres unités non-échantillonnées). L'estimateur de Horvitz-Thompson peut donc s'écrire comme la somme pondérée des valeurs de la variables y pour les unités échantillonnées, ce qui mène à

$$\hat{t}_y = \sum_{i \in s} w_i y_i. \quad (2.20)$$

Définition 2.6 : Un plan de sondage est dit *auto-pondéré* si le poids de sondage des unités est constant (i.e., le poids ne varie pas d'une unité à l'autre).

Les plans EASSR, BE et SY sont trois exemples de plans de sondage auto-pondérés. Dans ce qui suit, nous montrons que l'estimateur de Horvitz-Thompson est sans biais et nous donnons une expression de sa variance ainsi qu'un estimateur de celle-ci. Les propositions suivantes supposent que le plan de sondage $p(s)$ utilisé est tel que $\pi_i > 0$ pour tout i et que $\pi_{ij} > 0$ pour tout (i, j) .

Proposition 2.10 : L'estimateur de Horvitz-Thompson en (2.19) est sans biais pour t_y .

Démonstration : laissée en exercice.

Proposition 2.11 : La variance de l'estimateur de Horvitz-Thompson (2.20) est donnée par

$$V(\hat{t}_y) = \sum_{i \in U} \sum_{j \in U} (\pi_{ij} - \pi_i \pi_j) \frac{y_i}{\pi_i} \frac{y_j}{\pi_j}. \quad (2.21)$$

Démonstration :

Remarque : On vérifiera, que pour le plan EASSR, l'expression (2.21) se simplifie pour donner (2.1). De manière similaire, on peut montrer que dans le cas des plans BE et SY, l'expression (2.21) se réduit à (2.12) et (2.15), respectivement.

Bien sûr, la variance en (2.21) ne peut être obtenue puisque elle dépend des valeurs de y pour l'ensemble de la population. La proposition suivante exhibe un estimateur sans biais de la variance (2.21).

Proposition 2.12 : Un estimateur de la variance $V(\hat{t}_y)$, est donné par

$$\hat{V}(\hat{t}_y) = \sum_{i \in s} \sum_{j \in s} \frac{(\pi_{ij} - \pi_i \pi_j)}{\pi_{ij}} \frac{y_i}{\pi_i} \frac{y_j}{\pi_j}. \quad (2.22)$$

Démonstration :

Remarques :

(1) On vérifiera, que pour le plan EASSR, l'expression (2.22) se simplifie pour donner (2.2). De manière similaire, on peut montrer que dans le cas du plan BE, l'expression (2.22) se réduit à (2.13).

(2) Les fichiers de données produits par les organismes statistiques tels que Statistique Canada contiennent les valeurs des p variables d'intérêt ainsi que celles des q variables auxiliaires pour les unités échantillonnées ainsi qu'un poids de sondage. En pratique, le poids fourni par les organismes, souvent appelé poids final, a une structure assez complexe puisqu'il est le résultat de plusieurs ajustements dont un ajustement pour la non-réponse totale. Si aucun ajustement n'est effectué, le poids $w_i = 1/\pi_i$ sera utilisé (voir le Tableau 2.1). Il est alors facile pour les utilisateurs d'obtenir les estimations de totaux (pour les variables d'intérêt ou les variables auxiliaires) puisqu'ils n'ont qu'à calculer des sommes pondérées (par les poids de sondage) pour obtenir des estimations de la forme (2.20). La Proposition 2.10 nous garantit que, quelle que soit la variable pour laquelle on cherche à estimer son total (variable d'intérêt ou auxiliaire), l'utilisation d'un estimateur de la forme (2.20) sera sans biais par rapport au plan de sondage.

Tableau 2.1 : Fichier de donnée typique fourni par les organismes statistiques

| unité | y_1 | ... | y_p | x_1 | ... | x_q | w_i |
|-------|----------|-----|----------|----------|-----|----------|-------|
| 1 | y_{11} | ... | y_{p1} | x_{11} | ... | x_{q1} | w_1 |
| 2 | y_{12} | ... | y_{p2} | x_{12} | ... | x_{q2} | w_2 |
| . | . | . | . | . | . | . | . |
| . | . | . | . | . | . | . | . |
| . | . | . | . | . | . | . | . |
| n | y_{1n} | ... | y_{pn} | x_{1n} | ... | x_{qn} | w_n |

L'échantillonnage stratifié

Chapitre 3

3.1 Introduction

Dans le cas de l'échantillonnage aléatoire simple sans remise, l'échantillon est tiré à partir de la base de sondage sans l'aide de variables auxiliaires qui pourraient être présentes sur celle-ci. L'échantillonnage stratifié est une procédure qui permet l'utilisation d'une telle information auxiliaire, ce qui permet, entre autres, d'améliorer la qualité des estimateurs. À la section 3.2, l'estimation ponctuelle et l'estimation de la variance dans le cas de l'échantillonnage stratifié aléatoire simple seront traitées. À la section 3.3, on montre que les estimateurs ponctuels considérés à la section 3.2 peuvent s'écrire en fonction des poids de sondage. Le problème important de la répartition des observations dans les strates fera l'objet de la section 3.4. À la section 3.5, on compare l'échantillonnage stratifié aléatoire simple et l'échantillonnage aléatoire simple sans remise en termes d'efficacité. Finalement, la construction des strates sera brièvement abordée à la section 3.6.

Dans ce chapitre, on adoptera la notation suivante : Soit U une population de taille N , qui est divisée en H strates U_1, \dots, U_H , de tailles N_1, \dots, N_H , respectivement. Les strates forment une partition de la population; i.e., $\bigcup_{h=1}^H U_h = U$ et $U_h \cap U_l = \emptyset$ si $h \neq l$. On a donc $N = \sum_{h=1}^H N_h$. Dans ce chapitre, nous traitons exclusivement de l'échantillonnage stratifié aléatoire simple.

Définition 3.1 : Un échantillon stratifié aléatoire simple est obtenu en tirant un échantillon aléatoire simple sans remise dans chaque strate tel que la sélection dans une strate est indépendante de la sélection dans n'importe quelle autre strate.

Dans la strate U_h , on tire un échantillon aléatoire simple sans remise, s_h , de taille n_h , $h = 1, \dots, H$. Soit $s = \bigcup_{h=1}^H s_h$ l'échantillon « total » de taille $n = \sum_{h=1}^H n_h$. La situation est décrite par la Figure 3.1.

L'indépendance de sélection d'une strate à l'autre permet d'obtenir aisément la probabilité $p(s)$ donnée par

$$p(s) = \prod_{h=1}^H p(s_h) = \prod_{h=1}^H \frac{1}{\binom{N_h}{n_h}}.$$

Dans le cas de l'échantillonnage stratifié aléatoire simple, la probabilité d'inclusion dans l'échantillon de l'unité i appartenant à la strate h , π_{hi} , est égale à n_h/N_h . De plus, la probabilité d'inclusion jointe des unités i et j dans l'échantillon est donnée par

$$\pi_{ij} = \begin{cases} \frac{n_h}{N_h} \frac{n_l}{N_l} & \text{si } i \in U_h \text{ et } j \in U_l, h \neq l \\ \frac{n_h(n_h-1)}{N_h(N_h-1)} & \text{si } i \in U_h \text{ et } j \in U_h \end{cases}$$

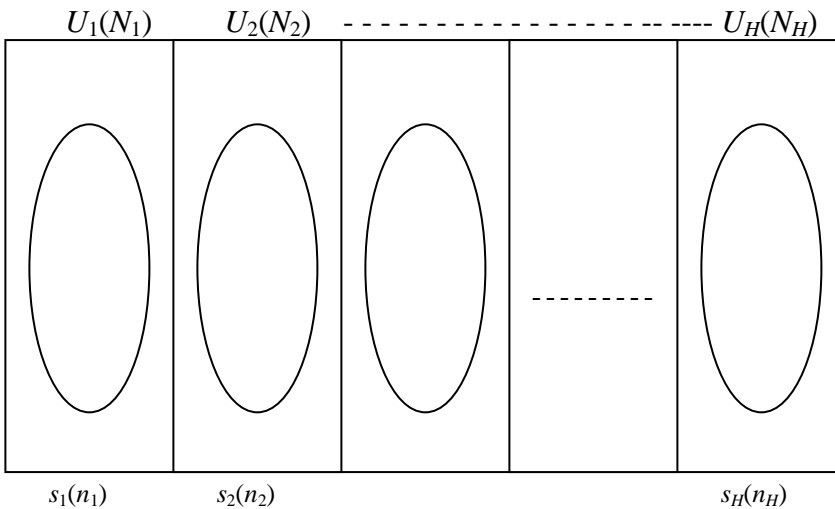


Figure 3.1 : Échantillonnage stratifié

Remarque : Bien que l'on ne considère que l'échantillonnage stratifié aléatoire simple dans ce chapitre, il existe d'autres types de plans stratifié. Par exemple, au lieu de tirer un échantillon aléatoire simple sans remise dans chaque strate, on pourrait utiliser l'échantillonnage systématique ou l'échantillonnage de Bernoulli dans chacune des strates. Le traitement de ces cas est très similaire à celui de l'échantillonnage stratifié aléatoire simple.

À ce stade, il est légitime de se demander pourquoi utiliserait-on l'échantillonnage stratifié aléatoire simple plutôt que l'échantillonnage aléatoire simple.

- (1) L'échantillonnage stratifié permet d'obtenir des estimations pour chacune des strates, ce qui est souvent requis en pratique. Par exemple, on peut stratifier la population canadienne par province si des estimations sont requises à ce niveau.
- (2) L'échantillonnage stratifié est plus pratique du point de vue opérationnel et administratif. En effet, il permet à l'organisme statistique de contrôler la répartition du travail sur le terrain entre ses bureaux régionaux. De plus, il est possible d'utiliser une stratégie différente dans chaque strate. Par exemple, dans le cas d'une enquête auprès des entreprises, un questionnaire peut être envoyé aux grandes entreprises par courrier postal alors qu'une interview (téléphonique ou sur place) peut être utilisée pour les petites entreprises. D'autres types d'enquête peuvent utiliser des méthodes d'échantillonnage différentes pour les strates urbaines et rurales.
- (3) L'échantillonnage stratifié permet d'éviter de tirer un « mauvais » échantillon. Par exemple, si l'on tire un échantillon aléatoire simple sans remise de taille 100 dans une population composée de 1000 hommes et 1000 femmes, il est possible que l'échantillon tiré contienne peu ou pas d'hommes, bien que l'occurrence d'un tel échantillon est très petite. Si les hommes et les femmes répondent différemment aux questions posées, les estimations pourraient être trop grandes ou trop petites. Si la variable sexe est disponible dans la base de sondage, il serait préférable de créer deux strates correspondant aux hommes et aux femmes et de tirer un échantillon aléatoire simple sans remise de taille 50 dans chaque strate. Une telle stratégie éliminerait la possibilité d'obtenir un mauvais échantillon (i.e., un échantillon non-représentatif).
- (4) L'échantillonnage stratifié permet généralement d'obtenir une stratégie d'échantillonnage plus efficace que l'échantillonnage aléatoire simple sans remise lorsque les strates sont telles que les unités dans chaque strate sont homogènes par rapport aux variables d'intérêt. Par conséquent, les estimateurs résultants ont généralement une plus petite variance que ceux obtenus dans le cas de l'échantillonnage aléatoire simple sans remise.

Pour toutes ces raisons, la stratification est une technique très fréquemment utilisée en pratique. Voici quelques exemples d'enquête utilisant l'échantillonnage stratifié :

- (1) À Statistique Canada, le plan de sondage typique dans les enquêtes auprès des entreprises est l'échantillonnage stratifié aléatoire simple. Les strates sont habituellement formées en croisant les variables auxiliaires *province*, *code SCIAN* et une variable de taille (par exemple, le *nombre d'employés*).
- (2) Considérons une enquête sur les fumeurs. L'objectif de l'enquête est de déterminer la proportion de fumeurs dans la population. Sachant que la proportion de fumeurs

varie considérablement selon l'âge, le sexe et le type d'emploi, il serait judicieux (si ces variables sont disponibles sur la base de sondage) de tirer un échantillon qui contient des membres de chacun de ces groupes. Par exemple, les strates pourraient être formées en croisant les variables âge, sexe et type d'emploi.

- (3) L'Agence du Revenu du Canada vérifie une petite fraction des rapports d'impôts soumis chaque année par les millions de contribuables. Le nombre de vérifications dépend habituellement du personnel disponible. En général, environ 0,25% des rapports d'impôt sont vérifiés chaque année. Les rapports à vérifier sont tirés aléatoirement selon un plan stratifié. Les rapports d'impôt sont d'abord stratifiés selon un ensemble de critères. Le revenu du contribuable, son type d'emploi (docteur, avocat, artiste, etc.), le type de déductions réclamées par le contribuable (voyage, abris fiscaux, etc.) sont des exemples de critères.

3.2 Estimation

Supposons que l'on cherche à estimer le total $t_y = \sum_{i \in U} y_i$, d'une variable d'intérêt y . On peut écrire le total t_y comme suit :

$$t_y = \sum_{h=1}^H \sum_{i \in U_h} y_{hi} = \sum_{h=1}^H t_h,$$

où $t_h = \sum_{i \in U_h} y_{hi}$ est le total dans la population pour la strate h et y_{hi} désigne la valeur de la variable y pour l'unité i appartenant à la strate h , $h=1, \dots, H$; $i=1, \dots, N_h$. Estimer t_y revient donc à estimer chaque t_h , ce que l'on sait faire (voir Chapitre 2, section 3.4). Un estimateur de t_y , que l'on désigne par \hat{t}_{st} , est donné par

$$\hat{t}_{st} = \sum_{h=1}^H \hat{t}_h, \quad (3.1)$$

où $\hat{t}_h = N_h \bar{y}_h$ et $\bar{y}_h = \frac{1}{n_h} \sum_{i \in s_h} y_{hi}$. L'estimateur (3.1) peut donc être réécrit comme suit :

$$\hat{t}_{st} = \sum_{h=1}^H N_h \bar{y}_h, \quad (3.2)$$

qui est une somme pondérée des moyennes échantillonnales \bar{y}_h . Maintenant, supposons que l'on cherche à estimer la moyenne dans la population \bar{y}_U . En notant que $\bar{y}_U = \frac{t_y}{N}$, un estimateur de \bar{y}_U , que l'on désigne par \bar{y}_{st} , est donné par

$$\bar{y}_{st} = \frac{\hat{t}_{st}}{N} = \sum_{h=1}^H \frac{N_h}{N} \bar{y}_h, \quad (3.3)$$

qui est une somme pondérée des moyennes échantillonnales \bar{y}_h . Le poids, $\frac{N_h}{N}$, représentent la taille relative de la strate h .

Proposition 3.1 : L'estimateur \hat{t}_{st} en (3.1) est sans biais pour t_y ; i.e., $E(\hat{t}_{st}) = t_y$.

Démonstration :

Remarque : Puisque $\bar{y}_{st} = \frac{\hat{t}_{st}}{N}$, il découle de la Proposition 3.1 que

$E(\bar{y}_{st}) = \frac{E(\hat{t}_{st})}{N} = \frac{t_y}{N} = \bar{y}_U$. Donc, \bar{y}_{st} est un estimateur sans biais de \bar{y}_U .

Proposition 3.2 : La variance de \hat{t}_{st} est donnée par

$$V(\hat{t}_{st}) = \sum_{h=1}^H N_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{S_h^2}{n_h}, \quad (3.4)$$

où $S_h^2 = \frac{1}{N_h - 1} \sum_{i \in U_h} (y_{hi} - \bar{y}_{Uh})^2$ et $\bar{y}_{Uh} = \frac{1}{N_h} \sum_{i \in U_h} y_{hi}$.

Démonstration :

Remarques :

(1) Il découle de la Proposition 3.2 que la variance de \bar{y}_{st} en (3.3) est donnée par

$$V(\bar{y}_{st}) = \sum_{h=1}^H \left(\frac{N_h}{N} \right)^2 \left(1 - \frac{n_h}{N_h} \right) \frac{S_h^2}{n_h}. \quad (3.5)$$

(2) Les expressions (3.4) et (3.5) suggèrent que la variance des estimateurs sera faible lorsque

(a) les tailles n_h sont grandes.

(b) les fractions de sondage n_h/N_h sont grandes.

(c) les dispersions de la variable y , S_h^2 , sont petites. Autrement dit, l'échantillonnage stratifié est une stratégie efficace lorsque les strates sont homogènes par rapport à la variable y . Il est donc préférable, compte tenu des contraintes pratiques, de construire des strates homogènes. La construction des strates sera brièvement traitée à la section 3.6.

La variance (3.4) ne pouvant être calculée au moyen des unités échantillonnées, on acceptera de l'estimer.

Proposition 3.3 : Un estimateur sans biais de $V(\hat{t}_{st})$ en (3.4) est donnée par

$$\hat{V}(\hat{t}_{st}) = \sum_{h=1}^H N_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{s_h^2}{n_h},$$

où $s_h^2 = \frac{1}{n_h - 1} \sum_{i \in s_h} (y_{hi} - \bar{y}_h)^2$.

Remarque : Il découle de la Proposition 3.3 qu'un estimateur sans biais de la variance de \bar{y}_{st} en (3.5) est donné par

$$\hat{V}(\bar{y}_{st}) = \sum_{h=1}^H \left(\frac{N_h}{N}\right)^2 \left(1 - \frac{n_h}{N_h}\right) \frac{s_h^2}{n_h}.$$

Nous sommes maintenant en mesure de construire des intervalles de confiance dans le contexte de l'échantillonnage stratifié. Si la taille d'échantillon est suffisamment grande dans chaque strate ou que le nombre de strates est suffisamment grand, on peut appliquer le théorème limite central pour obtenir un intervalle de confiance pour la moyenne \bar{y}_U :

$$\bar{y}_{st} \pm z_{\alpha/2} \sqrt{\hat{V}(\bar{y}_{st})}.$$

Exemple 3.1 : (Lohr, 2010, exercice no. 5 p. 102) Le American Council of Learned Societies (ACLS) a tiré un échantillon de sociétés dans sept disciplines faisant parties de l'ACLS selon un plan stratifié aléatoire simple. Le but était d'étudier les habitudes de publication et l'utilisation des bibliothèques et du matériel informatique parmi les membres (étudiants, chercheurs et professeurs) des différents départements appartenant à une société de l'ACLS. On a demandé si le répondant est d'accord ou en désaccord avec la question suivante : « Lorsque je feuillette les pages du nouveau numéro de la revue la plus importante dans mon champs académique, je trouve rarement un article qui m'intéresse ». Les résultats par champs sont exhibés ci-dessous. Estimez la proportion d'individus dans la population qui sont en accord avec la question posée et estimez la variance de votre estimateur.

| Champ | Taille du département N_h | Taille de l'échantillon n_h | En accord (%) |
|---------------------|--------------------------------|----------------------------------|---------------|
| Littérature | 9100 | 915 | 37 |
| Études classiques | 1950 | 633 | 23 |
| Philosophie | 5500 | 658 | 23 |
| Histoire | 10850 | 855 | 29 |
| Langues | 2100 | 667 | 19 |
| Sciences politiques | 5500 | 833 | 43 |
| Sociologie | 9000 | 824 | 41 |

Solution :

Rappelons qu'une proportion est un cas particulier d'une moyenne pour laquelle la variable d'intérêt est dichotomique. On peut donc appliquer les résultats de la section 3.2 pour une moyenne en se rappelant également des résultats obtenus à la section 2.4 du Chapitre 2. On a donc

$$\hat{p}_{str} = \sum_{h=1}^H \frac{N_h}{N} \hat{p}_h = \left(\frac{9100}{44000}\right)(0,37) + \left(\frac{1950}{44000}\right)(0,23) + \dots + \left(\frac{9000}{44000}\right)(0,41) = 0,334.$$

De plus, on a

$$\begin{aligned} \hat{V}(\hat{p}_{str}) &= \sum_{h=1}^H \left(\frac{N_h}{N}\right)^2 \left(1 - \frac{n_h}{N_h}\right) \frac{\hat{p}_h(1 - \hat{p}_h)}{n_h - 1} = 1,46 \times 10^{-5} + 5,94 \times 10^{-7} + \dots + 1,61 \times 10^{-5} \\ &= 6,241 \times 10^{-5}. \end{aligned}$$

3.3 Écriture en fonction des poids de sondage

Rappelons que dans le cas de l'échantillonnage stratifié aléatoire simple, la probabilité d'inclusion de l'unité i dans la strate h est égale à $\pi_{hi} = \frac{n_h}{N_h}$. Le poids de sondage de l'unité i dans la strate h , défini selon l'inverse de la probabilité d'inclusion, est donc donné par $w_{hi} = \frac{N_h}{n_h}$. L'estimateur \hat{t}_{st} en (3.1) peut donc s'écrire comme

$$\hat{t}_{st} = \sum_{h=1}^H \sum_{i \in s_h} w_{hi} y_{hi}, \quad (3.6)$$

et l'estimateur \bar{y}_{st} en (3.3) peut s'écrire comme

$$\bar{y}_{st} = \frac{\sum_{h=1}^H \sum_{i \in s_h} w_{hi} y_{hi}}{\sum_{h=1}^H \sum_{i \in s_h} w_{hi}}. \quad (3.7)$$

Notons que $\sum_{h=1}^H \sum_{i \in s_h} w_{hi} = N$, ce signifie que l'échantillon « représente » bien la population.

3.4 Répartition de la taille de l'échantillon

Une fois que les strates sont construites, il s'agit de savoir comment répartir la taille d'échantillon globale n entre les différentes strates. Autrement dit, compte tenu que l'on dispose d'un budget donné, comment détermine-t-on les tailles n_1, n_2, \dots, n_H ?

À ce stade, on introduit la notion de fonction de coût. Supposons que le coût (ou budget) total, C , de l'enquête est donné par

$$C = c_0 + \sum_{h=1}^H n_h c_h, \quad (3.8)$$

où c_0 représente un coût de base et $c_h > 0$ représente le coût d'enquêter une unité dans la strate $h, h=1, \dots, H$. On cherche à répartir les unités dans les strates de manière à minimiser $V(\bar{y}_{st})$ étant donné un budget égal à C , ou de manière équivalente, minimiser C pour une variance donnée $V(\bar{y}_{st})$. On supposera que les coûts c_1, \dots, c_H sont fixes (non-aléatoires). L'objectif visé consiste donc à obtenir les estimations les plus précises pour un budget donné.

Proposition 3.4 : Minimiser la variance $V(\bar{y}_{st})$ en (3.5) étant donné le coût (3.8) mène à

$$n_h = n \left(\frac{\frac{N_h S_h}{\sqrt{c_h}}}{\sum_{l=1}^H \frac{N_l S_l}{\sqrt{c_l}}} \right). \quad (3.9)$$

La répartition en (3.9) est appelée *répartition optimale*.

Démonstration :

L'expression (3.9) suggère qu'il faille tirer un échantillon de grande taille dans la strate h si

- (i) La taille relative de la strate, $\frac{N_h}{N}$, est importante.
- (ii) La dispersion de la variable y dans la population, S_h^2 , est grande; dans ce cas, on accroît la taille de l'échantillon afin de compenser pour l'hétérogénéité correspondant à la variable y .
- (iii) Le coût d'enquêter une unité, c_h , est faible.

Si les coûts c_h sont égaux dans toutes les strates, l'expression (3.9) se simplifie pour donner

$$n_h = n \frac{N_h S_h}{\sum_{l=1}^H N_l S_l}. \quad (3.10)$$

La répartition (3.10) est appelée *répartition de Neyman*. Cette répartition est optimale si on a bien des coûts égaux. Par contre, elle est sous-optimale si les coûts varient d'une strate à l'autre. Si, en plus des coûts, la dispersion de la variable y est la même dans toutes les strates, i.e., $S_h = S$, alors l'expression (3.10) se simplifie pour donner

$$n_h = n \frac{N_h}{N}. \quad (3.11)$$

La répartition (3.11) est appelée *répartition proportionnelle*. Cette répartition est optimale si on a bien des coûts égaux et une même dispersion dans toutes les strates. Par contre, elle est sous-optimale si les coûts et/ou les dispersions varient d'une strate à l'autre.

Remarques :

- (1) Dans les cas des répartitions (3.9) et (3.10), le calcul des n_h dépend de la dispersion de la variable y dans la population, S_h , qui est inconnue. Dans des enquêtes répétées, il est possible d'utiliser les estimations des S_h obtenues antérieurement.
- (2) Dans le cas d'une répartition proportionnelle, on a $\pi_{hi} = \frac{n_h}{N_h} = \frac{n}{N}$, auquel cas $w_{hi} = N/n$. La répartition proportionnelle mène donc à un plan de sondage auto-pondéré.
- (3) Au lieu d'une répartition du type (3.9), (3.10) ou (3.11), il est possible d'effectuer une répartition manuelle pour lequel le choix des n_h est dicté par d'autres critères que la variance. Il ne faut cependant pas perdre de vue que plus la répartition

choisie s'éloigne d'une répartition optimale et plus les estimateurs risquent d'être forts inefficaces.

Exemple 3.2 : (Lohr, 2010, exercice no. 8 p. 104) Un chercheur dispose d'un budget de 20 000\$ pour effectuer une enquête. Il sait que, dans la population, 90% des ménages ont un numéro de téléphone. Les interviews téléphoniques coûtent 10\$ par ménage; les interviews sur place coûtent, quant à elles, 30\$ chacune si toutes les interviews sont effectuées sur place et coûtent 40\$ si elles ne sont utilisées que pour les ménages ne possédant pas un numéro de téléphone (car dans ce cas, les coûts de déplacement sont plus élevés). Supposons que la dispersion dans les strates des ménages ayant un téléphone est égale à celle dans la strate des ménages n'ayant pas de téléphone. De plus, supposons que le coût de base $c_0 = 5000\$$. Combien de ménages doit-on interviewer dans chaque strate si

- (a) tous les ménages bénéficient d'une interview sur place.
- (b) les ménages ayant un téléphone sont interviewés par téléphone et les ménages n'ayant pas de téléphone sont interviewés sur place.

Solution :

- (a) Puisque le budget pour les interviews est de 15 000\$, le nombre de ménages que l'on peut interviewer sur place, compte tenu du budget disponible, est égal à $15\ 000/30 = 500$. Les dispersions sont supposées égales dans les deux strates. La répartition proportionnelle est donc optimale (puisque l'on est en présence de coûts et de variances égaux). Donc, 450 ménages ayant un téléphone et 50 ménages n'ayant pas de téléphone seront interviewés.
- (b) Les dispersions dans les deux strates sont supposées égales. La répartition optimale devient donc

$$n_h = n \frac{N_h / \sqrt{c_h}}{\sum_{l=1}^H N_l / \sqrt{c_l}}$$

| Strate | c_h | N_h/N | $N_h / (N \sqrt{c_h})$ |
|----------------|-------|---------|------------------------|
| Avec téléphone | 10 | 0,9 | 0,284605 |
| Sans téléphone | 40 | 0,1 | 0,015811 |
| Total | | 1,0 | 0,300416 |

Les calculs dans le tableau ci-dessus impliquent que le nombre de ménages à tirer dans la strate des ménages ayant un téléphone, n_{tel} , est égal à

$$n_{\text{tel}} = \frac{0,284605}{0,300416} n.$$

Les contraintes de coût mènent à

$$10n_{\text{tel}} + 40n_{\text{nontel}} = 10n_{\text{tel}} + 40(n - n_{\text{tel}}) = 15\,000,$$

où n_{nontel} représente le nombre de ménages à tirer dans la strate des ménages n'ayant pas de téléphone. On résout finalement l'équation précédente, ce qui donne

$$n_{\text{tel}} = 1\,227, n_{\text{nontel}} = 68, \text{ et } n = 1\,295.$$

On constate que le nombre d'interviews à effectuer dans chaque strate est plus élevé que celui obtenu en (a), ce qui s'explique par la présence d'interviews téléphoniques qui coûtent considérablement moins cher que les interviews sur place.

3.5 Comparaison de l'échantillonnage stratifié aléatoire simple et de l'échantillonnage aléatoire simple sans remise

Tel que mentionné à la section 3.1, l'échantillonnage stratifié aléatoire simple est, en général, plus efficace que l'échantillonnage aléatoire simple sans remise. Afin de comparer ces deux stratégies, on se restreint au cas de la répartition proportionnelle (3.11), auquel cas la variance de \bar{y}_{st} , que l'on désigne par $V_{prop}(\bar{y}_{st})$, s'écrit comme

$$V_{prop}(\bar{y}_{st}) = \sum_{h=1}^H \left(\frac{N_h}{N} \right)^2 \left(1 - \frac{n_h}{N_h} \right) \frac{S_h^2}{n_h} = \left(1 - \frac{n}{N} \right) \frac{1}{Nn} \sum_{h=1}^H N_h S_h^2. \quad (3.12)$$

Nous décomposons maintenant la dispersion totale, $(N-1)S_y^2 = \sum_{i \in U} (y_i - \bar{y}_U)^2$, selon

$$\begin{aligned}
(N-1)S_y^2 &= SST \\
&= \sum_{i \in U} (y_i - \bar{y}_U)^2 \\
&= \sum_{h=1}^H \sum_{i \in U_h} (y_i - \bar{y}_U)^2 \\
&= \sum_{h=1}^H \sum_{i \in U_h} (y_{hi} - \bar{y}_{Uh} + \bar{y}_{Uh} - \bar{y}_U)^2 \\
&= \sum_{h=1}^H \sum_{i \in U_h} (y_{hi} - \bar{y}_{Uh})^2 + \sum_{h=1}^H N_h (\bar{y}_{Uh} - \bar{y}_U)^2 \\
&= \sum_{h=1}^H (N_h - 1)S_h^2 + \sum_{h=1}^H N_h (\bar{y}_{Uh} - \bar{y}_U)^2 \\
&\equiv SSW + SSB.
\end{aligned} \tag{3.13}$$

Le terme *SSW* en (3.13) désigne la dispersion intra-strate alors que *SSB* désigne la dispersion inter-strate. Autrement dit, la variabilité totale de la variable d'intérêt y s'exprime comme la somme de la dispersion intra-strate et de la dispersion inter-strate. En vertu de la décomposition (3.12) la variance en (3.13) est donnée

$$V_{prop}(\bar{y}_{st}) = \left(1 - \frac{n}{N}\right) \frac{1}{Nn} \left[SSW + \sum_{h=1}^H S_h^2 \right]. \tag{3.14}$$

Désignons par $V_{EASSR}(\bar{y}_s)$ la variance de \bar{y}_s que l'on aurait obtenu, si au lieu de tirer un échantillon stratifié aléatoire simple, on avait tiré un échantillon aléatoire simple sans remise. On a vu au Chapitre 2 que

$$\begin{aligned}
V_{EASSR}(\bar{y}_s) &= \left(1 - \frac{n}{N}\right) \frac{S_y^2}{n} \\
&= \left(1 - \frac{n}{N}\right) \frac{1}{n} \frac{SST}{N-1} \\
&= \left(1 - \frac{n}{N}\right) \frac{1}{n} \frac{(SSW + SSB)}{N-1}.
\end{aligned} \tag{3.15}$$

Comparant (3.14) et (3.15), on obtient

$$V_{EASSR}(\bar{y}_s) - V_{prop}(\bar{y}_{st}) = \left(1 - \frac{n}{N}\right) \frac{1}{nN} \frac{1}{(N-1)} \left[N(SSB) - \sum_{h=1}^H (N - N_h) S_h^2 \right].$$

On a donc

$$V_{EASSR}(\bar{y}_s) - V_{prop}(\bar{y}_{st}) \geq 0$$

si et seulement si

$$SSB \geq \sum_{h=1}^H \left(1 - \frac{N_h}{N}\right) S_h^2.$$

L'échantillonnage stratifié aléatoire simple avec répartition proportionnelle est donc une stratégie plus efficace que l'échantillonnage aléatoire simple sans remise à moins que

$$SSB < \sum_{h=1}^H \left(1 - \frac{N_h}{N}\right) S_h^2. \quad (3.16)$$

Plus les moyennes \bar{y}_{Uh} sont différentes, plus le gain de l'échantillonnage proportionnel est important. L'inégalité (3.16) est rarement satisfaite en pratique lorsque les tailles des strates N_h sont grandes puisque, dans ce cas, on s'attend à avoir $N_h (\bar{y}_{Uh} - \bar{y}_U)^2 > S_h^2$. En conclusion, l'échantillonnage stratifié aléatoire simple est habituellement une stratégie plus efficace que l'échantillonnage aléatoire simple sans remise.

3.6 Construction des strates

Comme nous l'avons vu à la section 3.5, la stratification est une stratégie judicieuse lorsque les strates sont homogènes. Dans ce cas, la variabilité inter-strate est grande alors que la variabilité intra-strate est petite. Par conséquent, lors de la construction des strates, il s'agira de s'assurer que celles-ci soient le plus homogène possible. Dans un monde idéal, on stratifierait par rapport aux valeurs de la variable d'intérêt y . Par exemple, si l'on désire estimer le montant dépensé par une certaine catégorie d'entreprises pour la publicité, on aimerait bien pouvoir regrouper les entreprises qui dépensent le plus dans une même strate (disons la strate 1), celles qui dépensent un peu moins dans une autre strate (disons la strate 2), etc. Dans ce cas, les strates seront homogènes par rapport à la variable « montant alloué à la publicité ». Le problème est que les strates étant formées avant la sélection de l'échantillon, on ne dispose pas des valeurs de la variable y (sinon, nous n'aurions pas besoin d'une enquête!). Pour contourner le problème, supposons que l'on dispose d'une variable auxiliaire pour toutes les unités de la population fortement

liée à la variable d'intérêt y . Par exemple, on pourrait stratifier à l'aide la variable 'nombre d'employés' dans l'exemple ci-dessus. L'hypothèse sous-jacente est que plus le nombre d'employés dans une entreprise est grand, plus les montants alloués à la publicité sont importants. Nous présentons maintenant une méthode appelée *méthode de la racine carrée de la fréquence cumulée* (en anglais, *cumulative square root of the frequency method*) qui fonctionne relativement bien pour construire les strates.

Exemple 3.3 : On cherche à estimer le volume annuel de ventes pour 56 entreprises à partir d'un échantillon de taille $n = 15$. On dispose de la variable « Revenu » pour les 56 entreprises. Les données sont exhibées dans le tableau ci-dessous. Comment répartir les entreprises de la population si l'on veut utiliser $H = 3$ strates?

| Revenu (en milliers de dollars) | Fréquence | $\sqrt{\text{Fréquence}}$ | $\sqrt{\text{Fréquence cumulée}}$ |
|---------------------------------------|-----------|---------------------------|-----------------------------------|
| 100 – 150 | 11 | 3,32 | 3,32 |
| 150 – 200 | 14 | 3,74 | 7,06 |
| 200 – 250 | 9 | 3,00 | 10,06 |
| 250 – 300 | 4 | 2,00 | 12,06 |
| 300 – 350 | 5 | 2,24 | 14,30 |
| 350 – 400 | 8 | 2,83 | 17,13 |
| 400 – 450 | 3 | 1,73 | 18,86 |
| 450 – 500 | <u>2</u> | 1,41 | 20,27 |

56

Solution : Notons que dans le tableau précédent, nous avons ajouté deux colonnes aux deux premières : l'une représentant la racine carrée de la fréquence et l'autre la racine carrée de la fréquence cumulée. La méthode consiste à découper la dernière colonne en 3 intervalles de taille égale. Puisque $(20,27)/3 = 6,76$, la frontière de la première strate doit être aussi près que possible de 6,76 alors que la frontière de la deuxième strate doit être aussi près que possible de $2(6,76) = 13,52$. Dans le tableau, on constate que 7,06 est le nombre le plus près de 6,76, et 14,30 est le nombre le plus près de 13,52. Les 3 strates sont donc définies comme suit :

Strate 1 : Entreprises dont le revenu est entre 100 000 et 200 000

Strate 2 : Entreprises dont le revenu est entre 200 001 et 350 000

Strate 3 : Entreprises dont le revenu est entre 350 001 et 500 000

Estimation d'autres paramètres

Chapitre 4

4.1 Introduction

Jusqu'à maintenant, nous avons traité du problème de l'estimation d'un total ou d'une moyenne (qui inclut une proportion comme cas particulier). En pratique, il est fréquemment requis d'estimer des paramètres plus complexes. À la section 4.2, on considère l'estimation du total d'un domaine. Le ratio (ou quotient) de deux totaux fera l'objet de la section 4.3. Le cas de la moyenne d'un domaine, qui est un cas particulier d'un ratio, est traité à la section 4.4. Finalement, la technique de linéarisation par séries de Taylor utile pour approximer des fonctions lisses de totaux, fait l'objet de la section 4.5.

4.2 Estimation du total d'un domaine

Dans la grande majorité des enquêtes, les estimations sont requises, non seulement au niveau de toute la population mais également au niveau de sous-populations appelées domaines. Un *domaine* est donc un sous-ensemble de la population pour lequel une estimation est requise. Par exemple, dans l'Enquête Canadienne sur la Population Active (EPA), les estimations du taux de chômage sont requises non seulement aux niveaux national et provincial mais également par groupe d'âge ou par sexe.

Soit U une population de taille N . Soit $U_d \subseteq U$ un domaine de taille N_d , que l'on suppose inconnue. Le but est d'estimer le total d'une variable d'intérêt y au niveau du domaine

$$t_d = \sum_{i \in U_d} y_i. \quad (4.1)$$

On peut réécrire t_d en (4.1) comme

$$t_d = \sum_{i \in U} \delta_i y_i, \quad (4.2)$$

où δ_i est une variable indicatrice de domaine telle que $\delta_i = 1$ si l'unité i appartient au domaine U_d et $\delta_i = 0$, sinon. Supposons que l'on tire un EASSR, s , de taille n . Soit $s_d = s \cap U_d$ le sous-ensemble de s qui appartient au domaine U_d . Soit n_d le nombre d'unités dans s_d . Notons que n_d est une variable aléatoire. La figure 4.1 résume la notation utilisée dans cette section :

Population U de taille N

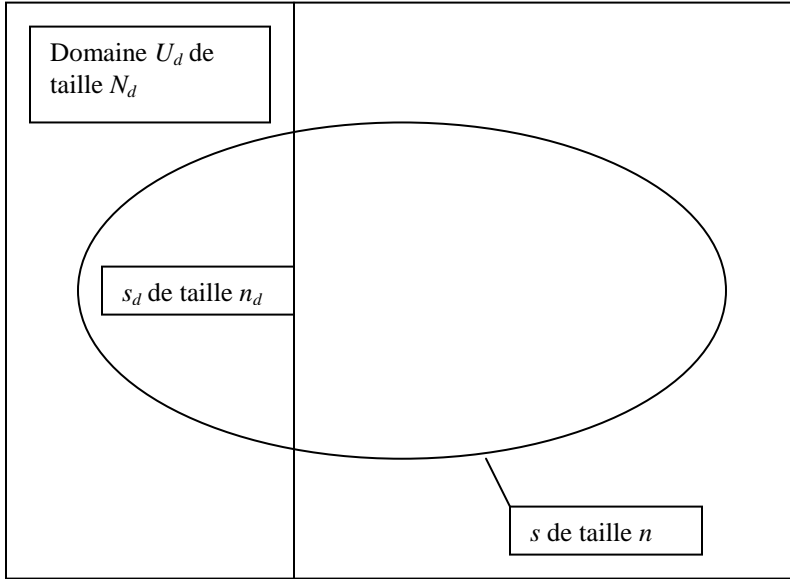


Figure 4.1 : Estimation par domaine

L'expression (4.2) suggère qu'estimer le total d'un domaine est équivalent à estimer le total $\sum_{i \in U} u_i$, où $u_i = \delta_i y_i$, ce que l'on sait faire (voir section 2.3.4). Autrement dit, les résultats suivants sont obtenus à partir de ceux présentés à la section 2.3.4 en remplaçant y_i par u_i . Un estimateur de t_d est donc donné par

$$\hat{t}_d = \frac{N}{n} \sum_{i \in s} u_i = \frac{N}{n} \sum_{i \in s} \delta_i y_i = \frac{N}{n} \sum_{i \in s_d} y_i = N \frac{n_d}{n} \bar{y}_d, \quad (4.3)$$

où $\bar{y}_d = \frac{1}{n_d} \sum_{i \in s_d} y_i$. Notons que le terme $N \frac{n_d}{n}$ représente une estimation de la taille N_d du domaine. Autrement dit, $\hat{N}_d = N \frac{n_d}{n}$. Par conséquent, l'estimateur de t_d peut s'écrire comme $\hat{t}_d = \hat{N}_d \bar{y}_d$, ce qui est tout à fait naturel. De manière similaire, on obtient la variance de \hat{t}_d :

$$V(\hat{t}_d) = N^2 \left(1 - \frac{n}{N}\right) \frac{S_u^2}{n}, \quad (4.4)$$

où $S_u^2 = \frac{1}{N-1} \sum_{i \in U} \left(u_i - \frac{\sum_{i \in U} u_i}{N} \right)^2$. Remarquons que

$$\begin{aligned}
S_u^2 &= \frac{1}{N-1} \sum_{i \in U} \left(\delta_i y_i - \frac{t_d}{N} \right)^2 \\
&= \frac{1}{N-1} \sum_{i \in U} \left(\delta_i y_i - \delta_i \bar{y}_{U_d} + \delta_i \bar{y}_{U_d} - \frac{t_d}{N} \right)^2 \\
&= \frac{1}{N-1} \left[\sum_{i \in U} \delta_i (y_i - \bar{y}_{U_d})^2 + N_d \left(1 - \frac{N_d}{N} \right) \bar{y}_{U_d}^2 \right],
\end{aligned}$$

où $\bar{y}_{U_d} = t_d/N_d$ désigne la moyenne de la variable d'intérêt y dans le domaine U_d . Notant que

$$\sum_{i \in U} \delta_i (y_i - \bar{y}_{U_d})^2 = \sum_{i \in U_d} (y_i - \bar{y}_{U_d})^2 = (N_d - 1) S_{yd}^2,$$

où $S_{yd}^2 = \frac{1}{(N_d - 1)} \sum_{i \in U_d} (y_i - \bar{y}_{U_d})^2$, S_u^2 s'écrit comme

$$S_u^2 = \frac{(N_d - 1)}{(N - 1)} S_{yd}^2 + \left(\frac{N_d}{N - 1} \right) \left(1 - \frac{N_d}{N} \right) \bar{y}_{U_d}^2$$

et la variance de \hat{t}_d en (4.4) peut s'écrire comme

$$V(\hat{t}_d) = N^2 \left(1 - \frac{n}{N} \right) \frac{1}{n} \left[\left(\frac{N_d - 1}{N - 1} \right) S_{yd}^2 + \left(\frac{N_d}{N - 1} \right) (1 - P_d) \bar{y}_{U_d}^2 \right], \quad (4.5)$$

où $P_d = \frac{N_d}{N}$ désigne la taille relative du domaine U_d . En supposant que $N_d - 1 \approx N_d$ et $N - 1 \approx N$, on peut approximer en (4.5) par

$$V(\hat{t}_d) \approx N^2 \left(1 - \frac{n}{N} \right) \frac{1}{n} P_d \left[S_{yd}^2 + (1 - P_d) \bar{y}_{U_d}^2 \right].$$

Un estimateur de $V(\hat{t}_d)$, que l'on désigne par $\hat{V}(\hat{t}_d)$, est obtenu en estimant les quantités inconnues dans (4.5), ce qui mène à

$$\hat{V}(\hat{t}_d) = N^2 \left(1 - \frac{n}{N} \right) \frac{1}{n} \left[\left(\frac{n_d - 1}{n - 1} \right) s_{yd}^2 + \left(\frac{n_d}{n - 1} \right) (1 - p_d) \bar{y}_d^2 \right], \quad (4.6)$$

où $p_d = \frac{n_d}{n}$ et $s_{yd}^2 = \frac{1}{n_d - 1} \sum_{i \in s_d} (y_i - \bar{y}_d)^2$. L'estimateur $\hat{V}(\hat{t}_d)$ en (4.6) est sans biais pour $V(\hat{t}_d)$ en (4.5). En supposant que $n_d - 1 \approx n_d$ et $n - 1 \approx n$, on peut approximer en (4.5) par

$$\hat{V}(\hat{t}_d) \approx N^2 \left(1 - \frac{n}{N}\right) \frac{1}{n} p_d [s_{yd}^2 + (1 - p_d) \bar{y}_d^2].$$

4.3 Estimation d'un ratio

Le ratio de deux totaux que l'on désigne par B , est défini selon

$$B = \frac{t_{y_1}}{t_{y_2}} = \frac{\bar{y}_{1U}}{\bar{y}_{2U}}, \quad (4.7)$$

où y_1 et y_2 désignent deux variables d'intérêt. Par exemple, on peut chercher à estimer la part des ventes (en dollars) annuelles de voitures pour le premier trimestre ou la part du revenu mensuel alloué à l'achat de nourriture dans les familles avec enfants ou encore le nombre d'enfants par ménage. Bien que l'on mette l'accent sur l'échantillonnage aléatoire simple sans remise, les résultats dans ce chapitre peuvent être aisément généralisés au cas d'un plan de sondage arbitraire.

Pour estimer B en (4.7), on tire un EASSR, s , de taille n , dans une population finie de taille N . Comment estimer B à partir des observations de l'échantillon? Un estimateur naturel de B consiste à estimer séparément le numérateur et le dénominateur en (4.7), ce que l'on sait faire. Un estimateur de B est donc donné par

$$\hat{B} = \frac{\hat{t}_{y_1}}{\hat{t}_{y_2}} = \frac{\bar{y}_{1s}}{\bar{y}_{2s}}, \quad (4.8)$$

où $\hat{t}_{y_1} = N\bar{y}_{1s}$, $\hat{t}_{y_2} = N\bar{y}_{2s}$ et $(\bar{y}_{1s}, \bar{y}_{2s}) = \frac{1}{n} \sum_{i \in s} (y_{1i}, y_{2i})$. Une question se pose naturellement:

Est-ce que l'estimateur \hat{B} en (4.8) est sans biais pour B ? Malheureusement, la réponse est non puisque

$$E(\hat{B}) = E\left(\frac{\hat{t}_{y_1}}{\hat{t}_{y_2}}\right) \neq \frac{E(\hat{t}_{y_1})}{E(\hat{t}_{y_2})} = \frac{t_{y_1}}{t_{y_2}} = B.$$

Autrement dit, l'espérance d'un ratio n'est pas égale au ratio des espérances. Peut-on alors, évaluer le biais de \hat{B} ? Comme nous le verrons, il n'est pas facile d'obtenir l'expression exacte du biais de \hat{B} . Dans un premier temps, on se contentera d'obtenir une borne supérieure pour le biais de \hat{B} . Dans un deuxième temps, on obtiendra une expression du biais approximatif à l'aide de l'approximation par série de Taylor.

Proposition 4.1 : Dans le cas d'un EASSR, une borne supérieure du ratio du biais de \hat{B} est donnée par

$$|BR(\hat{B})| = \frac{|Biais(\hat{B})|}{\sqrt{V(\hat{B})}} \leq CV(\hat{t}_{y_2}) = \frac{\sqrt{V(\hat{t}_{y_2})}}{t_{y_2}} = \sqrt{\left(1 - \frac{n}{N}\right)} \frac{S_{y_2}}{\sqrt{n}\bar{y}_{2U}} = \sqrt{\left(1 - \frac{n}{N}\right)} \frac{CV(y_2)}{\sqrt{n}},$$

où $CV(y_2) = \frac{S_{y_2}}{\bar{y}_{2U}}$ désigne le coefficient de variation de la variable y_2 .

Démonstration :

La Proposition 4.1 montre que le ratio du biais est petit lorsque

- (i) la taille de l'échantillon n est grande;
- (ii) la fraction de sondage n/N est grande;
- (iii) le coefficient de variation de la variable y_2 , $CV(y_2)$, est petit.

Cependant, lorsque la taille de l'échantillon est très petite et/ou $CV(y_2)$ est très grand, l'estimateur \hat{B} peut être considérablement biaisé. En général, on peut estimer que le biais sera petit lorsque la taille de l'échantillon sera plus grande que 30 ou lorsque $CV(\hat{t}_{y_2}) \leq 0,1$. La proposition ci-dessous donne une expression du biais approximatif de l'estimateur \hat{B} .

Proposition 4.2 : Dans le cas de l'EASSR, le biais approximatif de \hat{B} est donné par

$$E(\hat{B} - B) \approx \left(1 - \frac{n}{N}\right) \frac{1}{n\bar{y}_{2U}^2} (BS_{y_2}^2 - RS_{y_1}S_{y_2}),$$

où $R = \frac{\sum_{i \in U} (y_{1i} - \bar{y}_{1U})(y_{2i} - \bar{y}_{2U})}{(N-1)S_{y_1}S_{y_2}}$ désigne le coefficient de corrélation entre les variables y_1 et y_2 .

Démonstration : voir la section 4.5.

La Proposition 3.2 suggère que le biais approximatif de \hat{B} est petit lorsque

- (i) la taille de l'échantillon n est grande;
- (ii) la fraction de sondage n/N est grande;
- (iii) la moyenne \bar{y}_{2U} est grande;
- (iv) la dispersion dans la population de la variable y_2 , S_{y_2} , est petite;
- (v) le coefficient de corrélation R est près de 1.

À partir de maintenant, nous considérerons que le biais de \hat{B} est négligeable lorsque la taille de l'échantillon est suffisamment grande. Autrement dit, on acceptera d'écrire

$E(\hat{B}) \approx \frac{t_{y_1}}{t_{y_2}} = B$ pour n grand. Il s'ensuit que l'erreur quadratique moyenne (EQM) de \hat{B}

peut être approximée par sa variance; i.e.,

$$EQM(\hat{B}) = V(\hat{B}) + \text{Biais}(\hat{B})^2 \approx V(\hat{B}).$$

Déterminer explicitement la variance de \hat{B} est virtuellement impossible car \hat{B} est une fonction non-linéaire de deux totaux estimés \hat{t}_{y_1} et \hat{t}_{y_2} . On se contentera d'une expression de la variance approximative de \hat{B} .

Proposition 4.3 : Dans le cas de l'EASSR, la variance de \hat{B} est approximativement égale à

$$V(\hat{B}) \approx \left(1 - \frac{n}{N}\right) \frac{1}{\bar{y}_{2U}^2} \frac{S_e^2}{n}, \quad (4.9)$$

où $S_e^2 = \frac{1}{N-1} \sum_{i \in U} (y_{1i} - B y_{2i})^2$.

Démonstration : voir la section 4.5.

Corollaire 4.1 : Une expression alternative de la variance approximative en (4.9) est donnée par

$$V(\hat{B}) \approx \left(1 - \frac{n}{N}\right) \frac{1}{\bar{y}_{2U}^2} \frac{1}{n} \left[S_{y_1}^2 - 2BRS_{y_1} S_{y_2} + B^2 S_{y_2}^2 \right]. \quad (4.10)$$

Démonstration : laissée en exercice.

Les expressions (4.9) et (4.10) suggèrent que la variance approximative de \hat{B} est petite lorsque

- (i) la taille n de l'échantillon est grande;
- (ii) la fraction de sondage n/N est grande;
- (iii) les « résidus », $y_{1i} - By_{2i}$, sont petits;
- (iv) le coefficient de corrélation R est près de 1;
- (v) la moyenne \bar{y}_{2U} est grande.

Bien sûr, il n'est pas possible de calculer la variance de \hat{B} en (4.9) puisqu'elle dépend de B qui est inconnu. Il faudra donc l'estimer.

Proposition 4.4 : Un estimateur *approximativement* sans biais de la variance approximative de \hat{B} en (4.9) est donné par

$$\hat{V}(\hat{B}) = \left(1 - \frac{n}{N}\right) \frac{s_e^2}{n\bar{y}_{2s}^2},$$

(4.11)

où $s_e^2 = \frac{1}{n-1} \sum_{i \in s} (y_{1i} - \hat{B}y_{2i})^2$ et $\bar{y}_{2s} = \frac{1}{n} \sum_{i \in s} y_{2i}$. On a donc $E[\hat{V}(\hat{B})] \approx V(\hat{B})$.

Démonstration : omise.

Exemple 4.1 : Une enquête auprès des ménages d'une petite communauté a été effectuée afin de déterminer la part du revenu total alloué à l'achat de nourriture durant l'année. Soit y_1 montant (en dollars) alloué à l'achat de nourriture et y_2 le revenu total d'un ménage. Un EASSR de 14 ménages a été tiré d'une population de 150 ménages dans la communauté. Les données sont exhibées dans le tableau ci-dessous.

- (a) Estimer B , la part du revenu total alloué à l'achat de nourriture durant l'année.
- (b) Construire un intervalle de confiance de niveau 95% pour B .

| Ménage | y_{1i} | y_{2i} | $y_{1i} - \hat{B}y_{2i}$ |
|--------|----------|----------|--------------------------|
| 1 | 3 800 | 25 100 | 117,33 |
| 2 | 5 100 | 32 200 | 375,62 |
| 3 | 4 200 | 29 600 | -142,91 |
| 4 | 6 200 | 35 000 | 1064,80 |
| 5 | 5 800 | 34 400 | 752,83 |
| 6 | 4 100 | 26 500 | 211,92 |
| 7 | 3 900 | 28 700 | -310,86 |
| 8 | 3 600 | 28 200 | -537,50 |
| 9 | 3 800 | 34 600 | -1276,51 |
| 10 | 4 100 | 32 700 | -697,74 |
| 11 | 4 500 | 31 500 | -121,68 |
| 12 | 5 100 | 30 600 | 610,37 |
| 13 | 4 200 | 27 700 | 135,86 |
| 14 | 4 000 | 28 500 | -181,52 |

$$\sum y_{1i} = 62\,400 \quad \sum y_{2i} = 425\,300$$

Solution :

(a) L'estimation de B est donnée par $\hat{B} = \frac{\hat{t}_{y_1}}{\hat{t}_{y_2}} = \frac{\sum_{i \in S} y_{1i}}{\sum_{i \in S} y_{2i}} = \frac{62\,400}{425\,300} = 0,147$.

Approximativement 15% du revenu total des ménages est alloué à l'achat de nourriture dans la petite communauté.

(b) Un intervalle de confiance de niveau 95% est donné par $\hat{B} \pm 1,96\sqrt{\hat{V}(\hat{B})}$, où $\hat{V}(\hat{B})$ est donné par (4.11). On a $s_e^2 = \frac{1}{n-1} \sum_{i \in S} (y_{1i} - \hat{B}y_{2i})^2 = 373951,2$. Après calculs, on obtient $\hat{V}(\hat{B}) = 0,00002624$. L'intervalle de confiance est donc donné par $0,147 \pm 0,0101$.

Remarque : Dans le cas d'un plan de sondage arbitraire, un estimateur du ratio B en (4.7) est obtenu en estimant séparément le numérateur et le dénominateur par leur estimateur de Horvitz-Thompson respectif (voir section 2.8).

4.4 Un cas particulier d'un ratio : la moyenne d'un domaine

À la section 4.2, on a étudié le problème de l'estimation du total d'un domaine. Dans cette section, on se penche sur l'estimation de la moyenne d'un domaine. Nous adoptons la notation de la section 4.2. La moyenne du domaine est donnée par

$$\bar{y}_{U_d} = \frac{1}{N_d} \sum_{i \in U_d} y_i, \quad (4.12)$$

d'une variable d'intérêt y . On peut réécrire \bar{y}_{U_d} en (4.12) comme

$$\bar{y}_{U_d} = \frac{\sum_{i \in U} \delta_i y_i}{\sum_{i \in U} \delta_i}. \quad (4.13)$$

L'expression (4.13) suggère qu'estimer la moyenne d'un domaine est équivalent à estimer le ratio de deux totaux, $\sum_{i \in U} u_i$ et $\sum_{i \in U} \delta_i$, où $u_i = \delta_i y_i$. Autrement dit, les résultats qui suivent sont obtenus à partir de ceux obtenus à la section 4.3 en remplaçant y_{1i} par u_i et y_{2i} par δ_i . En vertu de (4.8), un estimateur de \bar{y}_{U_d} est donc donné par

$$\bar{y}_d = \frac{\sum_{i \in s} u_i}{\sum_{i \in s} \delta_i} = \frac{\sum_{i \in s} \delta_i y_i}{\sum_{i \in s} \delta_i} = \frac{1}{n_d} \sum_{i \in s_d} y_i. \quad (4.14)$$

La variance approximative de \bar{y}_d en (4.14) est donc obtenue de (4.9) en remplaçant y_{1i} par u_i et y_{2i} par δ_i , ce qui mène à

$$V(\bar{y}_d) \approx \left(1 - \frac{n}{N}\right) \frac{1}{n \bar{y}_{2U}^2} S_e^2, \quad (4.15)$$

où $S_e^2 = \frac{1}{N-1} \sum_{i \in U} (\delta_i y_i - B \delta_i)^2$ et $B = \bar{y}_{U_d}$. Notant que $\bar{y}_{2U} = \frac{N_d}{N}$ et que

$$\sum_{i \in U} \delta_i (y_i - \bar{y}_{U_d})^2 = \sum_{i \in U_d} (y_i - \bar{y}_{U_d})^2 = (N_d - 1) S_{yd}^2,$$

où $S_{yd}^2 = \frac{1}{(N_d - 1)} \sum_{i \in U_d} (y_i - \bar{y}_{U_d})^2$, l'expression (4.15) s'écrit comme

$$\begin{aligned}
V(\bar{y}_d) &\approx \left(1 - \frac{n}{N}\right) \frac{1}{n} \left(\frac{N}{N_d}\right)^2 \left(\frac{N_d - 1}{N - 1}\right) S_{y_d}^2 \\
&\approx \left(1 - \frac{n}{N}\right) \frac{S_{y_d}^2}{nP_d},
\end{aligned} \tag{4.16}$$

où $P_d = \frac{N_d}{N}$ désigne la taille relative du domaine U_d et en supposant que $\frac{N_d - 1}{N - 1} \approx \frac{N_d}{N}$.

Un estimateur de $V(\bar{y}_d)$ en (4.16), que l'on désigne par $\hat{V}(\bar{y}_d)$, est obtenu en estimant les quantités inconnues dans (4.16), ce qui mène à

$$\hat{V}(\bar{y}_d) = \left(1 - \frac{n}{N}\right) \frac{s_{y_d}^2}{n_d}. \tag{4.17}$$

Exemple 4.2 : Un économiste cherche à estimer le montant moyen dépensé pour l'achat de nourriture (y) par des familles avec enfants vivant dans une région affichant un important taux de pauvreté. Une liste complète des 250 familles dans la région est disponible, mais il est impossible d'identifier sur la base de sondage les familles avec enfants. L'économiste tire donc un EASSR de taille $n = 50$ famille et trouve que 42 familles dans l'échantillon ont au moins un enfant. Ces 42 familles sont interviewées et

l'on obtient $\sum_{i=1}^{42} y_i = 1\,720$ $\sum_{i=1}^{42} y_i^2 = 7\,2200$.

- (a) Estimer le montant moyen dépensé pour la nourriture par les familles avec enfants dans la région.
- (b) Construire un intervalle de confiance de niveau 95% pour le montant moyen dépensé pour la nourriture par les familles avec enfants dans la région.

Solution :

(a) Le domaine d'intérêt est la sous-population U_d des familles avec enfants. On veut estimer \bar{y}_{U_d} . Une estimation de \bar{y}_{U_d} est donnée par

$$\bar{y}_d = \frac{1}{n_d} \sum_{i \in s_d} y_i = \frac{1}{42} (1\,720) = 40.95$$

(b) Un intervalle de confiance de niveau 95% est donné par $\bar{y}_d \pm 1,96\sqrt{\hat{V}(\bar{y}_d)}$, où $\hat{V}(\bar{y}_d)$ est donné par (4.17). D'abord, on a

$$s_d^2 = \frac{1}{(n_d - 1)} \sum_{i \in s_d} (y_i - \bar{y}_d)^2 = \frac{1}{(n_d - 1)} \left[\sum_{i \in s_d} y_i^2 - n_d \bar{y}_d^2 \right] = \frac{1}{41} [72\,200 - (42)(40,95)^2] = \frac{1\,762}{41}$$

Après quelques calculs, on obtient $\hat{V}(\bar{y}_d) = \left(1 - \frac{50}{250}\right) \frac{1}{42} \frac{1\,762}{41} = 0,819$. L'intervalle de confiance recherché est donné par $40,95 \pm 1,96\sqrt{0,819}$.

4.5 Linéarisation par séries de Taylor

À la section 4.3, nous avons intuitivement décidé d'estimer le ratio de deux totaux,

$$B = \frac{t_{y_1}}{t_{y_2}},$$

en estimant séparément le numérateur et le dénominateur, ce qui a mené à

$$\hat{B} = \frac{\hat{t}_{y_1}}{\hat{t}_{y_2}}.$$

Les propositions 4.1 et 4.2 suggèrent que le biais devient négligeable lorsque la

taille de l'échantillon est suffisamment grande. L'estimateur \hat{B} est donc un ratio de deux variables aléatoires, \hat{t}_{y_1} et \hat{t}_{y_2} . Autrement dit, l'estimateur \hat{B} est une fonction non-linéaire de variables aléatoires. Déterminer une expression exacte de son biais et une expression exacte de sa variance s'avère donc virtuellement impossible. Une manière de contourner le problème est d'approximer la fonction non-linéaire par une fonction linéaire des observations. On dira que l'on « linéarise » la fonction non-linéaire. La linéarisation par séries de Taylor, que l'on considère dans cette section, est un outil utile permettant la linéarisation de fonctions non-linéaires.

Supposons que l'on cherche à estimer un paramètre θ . De plus, supposons que θ peut s'écrire comme une fonction « lisse » de totaux; i.e., $\theta = f(t_1, \dots, t_q)$, où $t_j = \sum_{i \in U} y_{ji}$, où

y_{ji} est la i^{e} valeur de la variable d'intérêt y_j , $j = 1, \dots, q$. Par exemple, le ratio B peut s'écrire comme $B = f(t_{y_1}, t_{y_2})$. De la population U de taille N , on tire un EASSR, s , de

taille n . Soit $\hat{\theta}$ un estimateur de θ . Comment estimer θ ?

Le principe est le suivant : dans la fonction $f(., \dots, .)$, on remplace chaque total inconnu t_j par son estimateur sans biais \hat{t}_j ; i.e., $E(\hat{t}_j) = t_j$. L'estimateur résultant de θ est donc

$\hat{\theta} = f(\hat{t}_1, \dots, \hat{t}_q)$. Par exemple, un estimateur de B est donné par $\hat{B} = f\left(\hat{t}_{y_1}, \hat{t}_{y_2}\right) = \frac{\hat{t}_{y_1}}{\hat{t}_{y_2}}$.

Quelles sont les propriétés d'un tel estimateur en termes de biais et de variance? La réponse à cette question nous amène à introduire le concept de linéarisation par séries de Taylor du premier ordre et du deuxième ordre.

4.5.1 Linéarisation par séries de Taylor du premier ordre

La technique présentée ici est similaire à celle étudiée dans un cours d'analyse de base. On acceptera sans démonstration que $\hat{\theta}$ peut être approximé par

$$\hat{\theta} = f(\hat{t}_1, \dots, \hat{t}_q) = f(t_1, \dots, t_q) + \sum_{j=1}^q \frac{\partial f(\hat{t}_1, \dots, \hat{t}_q)}{\partial \hat{t}_j} \Big|_{\hat{t}_1=t_1, \dots, \hat{t}_q=t_q} (\hat{t}_j - t_j) \quad (4.18)$$

+ Reste.

Le développement (4.18) est un développement par séries de Taylor du premier ordre. Il est possible de montrer que le reste est négligeable lorsque la taille de l'échantillon est suffisamment grande. On peut donc réécrire (4.18) comme

$$\hat{\theta} = f(\hat{t}_1, \dots, \hat{t}_q) \approx f(t_1, \dots, t_q) + \sum_{j=1}^q \frac{\partial f(\hat{t}_1, \dots, \hat{t}_q)}{\partial \hat{t}_j} \Big|_{\hat{t}_1=t_1, \dots, \hat{t}_q=t_q} (\hat{t}_j - t_j).$$

Exemple 4.3 : Soit $B = \frac{t_{y_1}}{t_{y_2}}$ et $\hat{B} = \frac{\hat{t}_{y_1}}{\hat{t}_{y_2}}$. On a bien $\hat{B} = f(\hat{t}_{y_1}, \hat{t}_{y_2})$. Maintenant, on évalue

les dérivées partielles $\frac{\partial f(\hat{t}_{y_1}, \hat{t}_{y_2})}{\partial \hat{t}_{y_1}}$ et $\frac{\partial f(\hat{t}_{y_1}, \hat{t}_{y_2})}{\partial \hat{t}_{y_2}}$, ce qui mène à

$$\frac{\partial f(\hat{t}_{y_1}, \hat{t}_{y_2})}{\partial \hat{t}_{y_1}} \Big|_{\hat{t}_{y_1}=t_{y_1}, \hat{t}_{y_2}=t_{y_2}} = \frac{1}{t_{y_2}}$$

et

$$\frac{\partial f(\hat{t}_{y_1}, \hat{t}_{y_2})}{\partial \hat{t}_{y_2}} \Big|_{\hat{t}_{y_1}=t_{y_1}, \hat{t}_{y_2}=t_{y_2}} = -\frac{t_{y_1}}{t_{y_2}^2}.$$

Par (4.18), on a

$$\begin{aligned}\hat{B} &= f(\hat{t}_{y_1}, \hat{t}_{y_2}) = f(t_{y_1}, t_{y_2}) + \frac{1}{t_{y_2}}(\hat{t}_{y_1} - t_{y_1}) - \frac{t_{y_1}}{t_{y_2}^2}(\hat{t}_{y_2} - t_{y_2}) + \text{Reste} \\ &\approx B + \frac{1}{t_{y_2}}(\hat{t}_{y_1} - t_{y_1}) - \frac{t_{y_1}}{t_{y_2}^2}(\hat{t}_{y_2} - t_{y_2}).\end{aligned}$$

On a donc

$$\hat{B} - B \approx \frac{1}{t_{y_2}}(\hat{t}_{y_1} - B\hat{t}_{y_2}), \quad (4.19)$$

ou encore

$$\hat{B} - B \approx \left(\frac{N}{n}\right) \sum_{i \in S} z_i, \quad (4.20)$$

où $z_i = \frac{1}{t_{y_2}}(y_{1i} - By_{2i})$. La variable z_i est appelée *variable linéarisée*. Il découle de (4.19)

que

$$E(\hat{B} - B) \approx \frac{1}{t_{y_2}} [E(\hat{t}_{y_1}) - BE(\hat{t}_{y_2})] = \frac{1}{t_{y_2}} [t_{y_1} - Bt_{y_2}] = 0.$$

Un développement par séries de Taylor du premier ordre suggère donc que \hat{B} est approximativement sans biais pour B . Autrement dit, le biais de \hat{B} est considéré négligeable pour une grande taille d'échantillon. Il est intéressant de noter de (4.19) et (4.20) que le développement par séries de Taylor a permis d'approximer la fonction non-linéaire, $\hat{B} - B$, par un total, \hat{t}_z qui est une fonction linéaire des observations. Ces expressions suggèrent également que la variance de \hat{B} peut être approximée par la variance du total \hat{t}_z . Autrement dit, il découle de (4.20) que

$$V(\hat{B} - B) = V(\hat{B}) \approx V(\hat{t}_z).$$

Obtenir la variance de \hat{t}_z , $V(\hat{t}_z)$, consiste alors à utiliser les expressions de variance obtenues au chapitre 2 pour l'estimateur d'un total en remplaçant la variable y_i par la linéarisée z_i . Dans le cas de l'EASSR (voir Proposition 2.3), il s'ensuit que la variance de \hat{B} peut être approximée par

$$V(\hat{B}) \approx V(\hat{t}_z) = N^2 \left(1 - \frac{n}{N}\right) \frac{S_z^2}{n}, \quad (4.21)$$

où

$$S_z^2 = \frac{1}{N-1} \sum_{i \in U} (z_i - \bar{z}_U)^2 = \frac{1}{N-1} \sum_{i \in U} z_i^2 = \frac{1}{t_{y_2}^2} \frac{1}{N-1} \sum_{i \in U} (y_{1i} - B y_{2i})^2,$$

en notant que $\bar{z}_U = \frac{1}{N} \sum_{i \in U} z_i = 0$. La démonstration de la Proposition 4.3 est donc complète

en posant $S_e^2 = \frac{1}{N-1} \sum_{i \in U} (y_{1i} - B y_{2i})^2$ et en notant que $\frac{t_{y_2}^2}{N^2} = \bar{y}_{2U}^2$.

4.5.2 Linéarisation par séries de Taylor du deuxième ordre

Il est possible d'obtenir un développement par séries de Taylor du deuxième ordre en allant un pas plus loin, ce qui nous amène à évaluer les dérivées partielles du deuxième ordre. Il en résulte une approximation plus précise que celle obtenue à l'aide d'une approximation du premier ordre. On acceptera sans démonstration que $\hat{\theta}$ peut être approximé par

$$\hat{\theta} = f(\hat{t}_1, \dots, \hat{t}_q) = f(t_1, \dots, t_q) + \sum_{j=1}^q \frac{\partial f(\hat{t}_1, \dots, \hat{t}_q)}{\partial \hat{t}_j} \Big|_{\hat{t}_1=t_1, \dots, \hat{t}_q=t_q} + \frac{1}{2!} \sum_{j=1}^q \sum_{k=1}^q \frac{\partial^2 f(\hat{t}_1, \dots, \hat{t}_q)}{\partial \hat{t}_j \partial \hat{t}_k} \Big|_{\hat{t}_1=t_1, \dots, \hat{t}_q=t_q} (\hat{t}_j - t_j)(\hat{t}_k - t_k) + \text{Reste.}$$

Exemple 4.4 (suite de l'exemple 4.3) : On évalue les dérivées partielles du deuxième

ordre, $\frac{\partial^2 f(\hat{t}_{y_1}, \hat{t}_{y_2})}{\partial^2 \hat{t}_{y_2}}$, $\frac{\partial^2 f(\hat{t}_{y_1}, \hat{t}_{y_2})}{\partial^2 \hat{t}_{y_1}}$ et $\frac{\partial^2 f(\hat{t}_{y_1}, \hat{t}_{y_2})}{\partial^2 \hat{t}_{y_1} \hat{t}_{y_2}}$, ce qui mène à

$$\frac{\partial^2 f(\hat{t}_{y_1}, \hat{t}_{y_2})}{\partial^2 \hat{t}_{y_1}} \Big|_{\hat{t}_{y_1}=\hat{t}_{y_2}=\hat{t}_{y_2}} = 0;$$

$$\frac{\partial^2 f(\hat{t}_{y_1}, \hat{t}_{y_2})}{\partial^2 \hat{t}_{y_2}} \Big|_{\hat{t}_{y_1}=\hat{t}_{y_2}=\hat{t}_{y_2}} = \frac{2t_{y_1}}{t_{y_2}^3}$$

et

$$\frac{\partial^2 f(\hat{t}_{y_1}, \hat{t}_{y_2})}{\partial^2 \hat{t}_{y_1} \hat{t}_{y_2}} \Big|_{\hat{t}_{y_1}=t_{y_1}, \hat{t}_{y_2}=t_{y_2}} = -\frac{1}{t_{y_2}^2}.$$

On a alors

$$\begin{aligned} \hat{B} &= B + \frac{1}{t_{y_2}}(\hat{t}_{y_1} - t_{y_1}) - \frac{t_{y_1}}{t_{y_2}^2}(\hat{t}_{y_2} - t_{y_2}) + \frac{2t_{y_1}}{2!t_{y_2}^3}(\hat{t}_{y_2} - t_{y_2})^2 - \frac{2}{2!t_{y_2}^2}(\hat{t}_{y_1} - t_{y_1})(\hat{t}_{y_2} - t_{y_2}) + \text{Reste} \\ &\approx B + \frac{1}{t_{y_2}}(\hat{t}_{y_1} - t_{y_1}) - \frac{t_{y_1}}{t_{y_2}^2}(\hat{t}_{y_2} - t_{y_2}) + \frac{2t_{y_1}}{2!t_{y_2}^3}(\hat{t}_{y_2} - t_{y_2})^2 - \frac{2}{2!t_{y_2}^2}(\hat{t}_{y_1} - t_{y_1})(\hat{t}_{y_2} - t_{y_2}). \end{aligned}$$

Il s'ensuit que

$$\hat{B} - B \approx \frac{1}{t_{y_2}}(\hat{t}_{y_1} - t_{y_1}) - \frac{B}{t_{y_2}}(\hat{t}_{y_2} - t_{y_2}) + \frac{B}{t_{y_2}^2}(\hat{t}_{y_2} - t_{y_2})^2 - \frac{1}{t_{y_2}^2}(\hat{t}_{y_1} - t_{y_1})(\hat{t}_{y_2} - t_{y_2}). \quad (4.22)$$

Il découle de (4.22) que

$$\begin{aligned} E(\hat{B} - B) &\approx \frac{B}{t_{y_2}^2} E(\hat{t}_{y_2} - t_{y_2})^2 - \frac{1}{t_{y_2}^2} E[(\hat{t}_{y_1} - t_{y_1})(\hat{t}_{y_2} - t_{y_2})] \\ &= \frac{1}{t_{y_2}^2} [BV(\hat{t}_{y_2}) - Cov(\hat{t}_{y_1}, \hat{t}_{y_2})] \\ &= N^2 \left(1 - \frac{n}{N}\right) \frac{1}{nt_{y_2}^2} [BS_{y_2}^2 - S_{y_1 y_2}], \end{aligned}$$

où $S_{y_1 y_2} = \frac{1}{N-1} \sum_{i \in U} (y_{1i} - \bar{y}_{1U})(y_{2i} - \bar{y}_{2U})$. La démonstration de la Proposition 4.2 est

complète en notant que $S_{y_1 y_2} = RS_{y_1} S_{y_2}$ et $\frac{t_{y_2}^2}{N^2} = \bar{y}_{2U}^2$.

Estimation en présence d'information auxiliaire

Chapitre 5

5.1 Introduction

Jusqu'à maintenant, nous avons vu qu'il était possible d'utiliser de l'information auxiliaire à l'étape du plan de sondage dans le cas de l'échantillonnage stratifié. Par exemple, dans les enquêtes auprès des entreprises, l'échantillonnage stratifié aléatoire simple est habituellement utilisé, les strates étant obtenues en croisant des variables disponibles sur le registre des entreprises (la base de sondage) pour toutes les unités dans la population. Ces variables incluent habituellement la province, le code SCIAN et le nombre d'employés. Il arrive fréquemment qu'une certaine quantité d'information auxiliaire est disponible à l'étape de l'estimation. Cette information auxiliaire comprend les variables auxiliaires qui étaient disponibles à l'étape du plan de sondage mais également d'autres variables auxiliaires pouvant provenir, par exemple, de sources administratives. Dans ce chapitre, nous étudions quelques procédures d'estimation qui utilisent l'information auxiliaire disponible. Les estimateurs résultant sont généralement plus précis que les estimateurs n'utilisant pas d'information auxiliaires étudiés au chapitre 2. Dans ce chapitre, on considère les procédures d'estimation suivantes : la post-stratification, l'estimateur par le ratio, l'estimateur par la régression et l'estimateur par la différence. Encore une fois, par souci de simplicité, on met l'accent sur l'échantillonnage aléatoire simple sans remise dans ce chapitre mais la généralisation à des plans arbitraires est relativement aisée.

Rappelons qu'une variable est appelée variable auxiliaire si celle-ci est disponible

- (i) pour toutes les unités dans la population;
- (ii) pour les unités échantillonnées seulement mais son total au niveau de la population est connu.

Voici quelques exemples de variables auxiliaires rencontrées fréquemment en pratique :

1. Une liste peut contenir le nom des villes ainsi que le nombre d'habitants dans chaque ville à la date du dernier recensement.
2. Le registre des fermes Canadiennes contient la superficie de chaque ferme.
3. Le registre des entreprises Canadiennes contient des variables telles que le nombre d'employés et le type d'industrie.

5.2 La post-stratification

La post-stratification est une technique extrêmement importante car elle est utilisée dans la grande majorité des enquêtes. Supposons que l'on cherche à estimer la moyenne de la population d'une variable d'intérêt y , $\bar{y}_U = \frac{1}{N} \sum_{i \in U} y_i$. Pour cela, on tire un EASSR, s , de taille n . De plus, supposons qu'une variable auxiliaire catégorielle x est disponible à l'étape de l'estimation et qu'elle prend J valeurs distinctes. Par exemple, la variable x peut représenter le groupe d'âge, auquel cas on aurait $x_i = 1$ si l'unité i est âgée entre 15 et 19 ans, $x_i = 2$ si l'unité i est âgée entre 20 et 24 ans, etc. Comme dans le cas de la stratification, la post-stratification consiste à d'abord utiliser la variable x afin de diviser la population en J sous-groupes disjoints, U_1, \dots, U_J , que l'on appelle post-strates; i.e., $U = \bigcup_{j=1}^J U_j$ et $U_j \cap U_l = \emptyset, j \neq l$. Soit N_j la taille de U_j . On suppose que les tailles N_j sont connues, ce qui est une hypothèse inhérente à la post-stratification. On a donc $N = \sum_{j=1}^J N_j$. Notons que les tailles N_j sont souvent disponibles grâce aux effectifs produits par le recensement canadien de la population. Ces effectifs sont utilisés dans les enquêtes auprès des ménages menées par Statistique Canada. De plus, notons que, contrairement à la stratification pour laquelle les strates sont formées avant la sélection de l'échantillon, les post-strates sont formées après la sélection de l'échantillon.

La partition de la population mène à la partition correspondante de l'échantillon, s , en sous-groupes s_1, \dots, s_J , où $s_j = U_j \cap s$, de taille n_1, \dots, n_J , où n_j désigne la taille de s_j . Notons que les tailles n_j sont des variables aléatoires. La Figure 5.1 résume bien la situation.

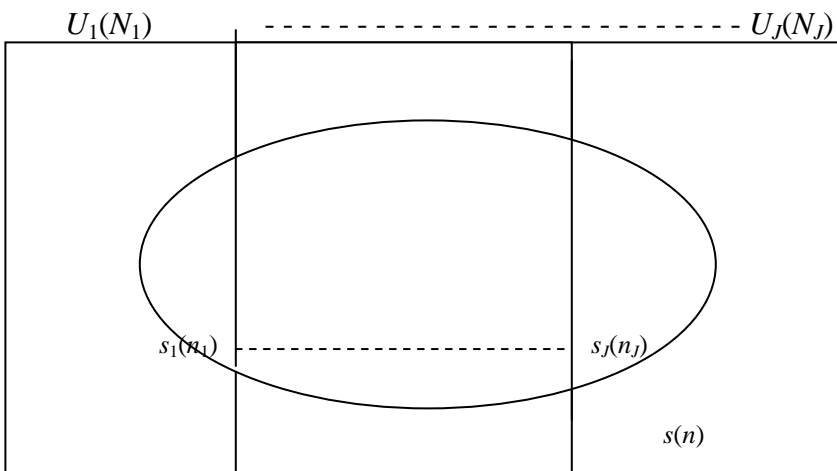


Figure 5.1 : Post-stratification

La moyenne de la population \bar{y}_U que l'on cherche à estimer peut s'écrire comme

$$\bar{y}_U = \sum_{j=1}^J \frac{N_j}{N} \bar{y}_{jU},$$

où $\bar{y}_{jU} = \frac{1}{N_j} \sum_{i \in U_j} y_i$ désigne la moyenne de la post-strate U_j . L'estimateur post-stratifié de \bar{y}_U est donné par

$$\bar{y}_{post} = \sum_{j=1}^J \frac{N_j}{N} \bar{y}_j, \quad (5.1)$$

où $\bar{y}_j = \frac{1}{n_j} \sum_{i \in s_j} y_i$. Notons que l'estimateur post-stratifié (5.1) peut également être exprimé comme suit :

$$\bar{y}_{post} = \frac{1}{N} \sum_{j=1}^J \sum_{i \in s_j} \frac{N}{n} \frac{N_j}{\hat{N}_j} y_i = \frac{1}{N} \sum_{j=1}^J \sum_{i \in s_j} \tilde{w}_i y_i, \quad (5.2)$$

où $\tilde{w}_i = \frac{N}{n} \frac{N_j}{\hat{N}_j}$ et $\hat{N}_j = \frac{N}{n} n_j$ est un estimateur la taille de U_j , N_j . L'expression (5.2)

suggère que la post-stratification a pour effet d'ajuster les poids de sondage $w_i = \frac{N}{n}$,

pour obtenir les poids $\tilde{w}_i = w_i \frac{N_j}{\hat{N}_j}$. Si $\hat{N}_j = N_j$, alors $\tilde{w}_i = w_i$ ce qui signifie qu'aucun

ajustement n'est nécessaire. Par contre, si $\frac{N_j}{\hat{N}_j} < 1$ alors $\tilde{w}_i < w_i$. De même, si $\frac{N_j}{\hat{N}_j} > 1$

alors $\tilde{w}_i > w_i$. Autrement dit, la post-stratification a comme effet de hausser la « visibilité » des unités qui sont sous-représentées dans l'échantillon et de diminuer celle des unités qui y sont surreprésentées de façon à ce que les unités de l'échantillon représentent adéquatement la population.

Remarques :

- (1) Puisque les tailles n_j sont des variables aléatoires, il n'est pas impossible que $n_j = 0$ pour un certain $j = 1, \dots, J$. Dans ce cas, la moyenne \bar{y}_j n'est pas définie. En pratique, lorsque cette situation survient, il est coutume de combiner des post-strates pour contourner cette difficulté. Dans ce qui suit, on supposera $n_j > 0$ pour tout $j = 1, \dots, J$.
- (2) Au chapitre 2, on a vu qu'un fichier typique produit par les organismes statistiques contient les valeurs des variables d'intérêt et des variables auxiliaires pour les unités ainsi

qu'un poids de sondage. Dans le contexte d'une post-stratification, on rapporte le poids \tilde{w}_i plutôt que le poids $w_i = 1/\pi_i$ (voir Tableau 5.1). Il est alors facile pour les utilisateurs d'obtenir les estimations de totaux (pour les variables d'intérêt ou les variables auxiliaires) puisqu'ils n'ont qu'à calculer des sommes pondérées (par les poids \tilde{w}_i) pour obtenir des estimations de la forme (5.2). Pour simplifier la discussion, supposons que l'on a post-stratifié en utilisant la variable *sexe*. Soit $x_{1i} = 1$ si l'unité i est une femme et $x_{1i} = 0$, sinon. De manière similaire, soit $x_{2i} = 1$ si l'unité i est un homme et $x_{2i} = 0$, sinon. Un utilisateur de données est intéressé à estimer le nombre total de femmes et celui des hommes dans la population. Autrement dit, on cherche à estimer $N_F = \sum_{i \in U} x_{1i}$ et $N_H = \sum_{i \in U} x_{2i}$. Deux estimateurs possibles sont donnés par $\hat{N}_F = \sum_{i \in s} w_i x_{1i} = \frac{N}{n} n_F$ et $\hat{N}_H = \sum_{i \in s} w_i x_{2i} = \frac{N}{n} n_H$, où n_F et n_H désignent respectivement le nombre de femmes et le nombre d'hommes dans l'échantillon. Notons que ces deux estimateurs ont été obtenus en calculant la somme pondérée des variables x_{1i} et x_{2i} , pondérée par le poids $w_i = 1/\pi_i$. Qu'obtiendrions-nous si on utilisait les poids \tilde{w}_i plutôt? Dans ce cas, les estimateurs de N_F et N_H sont respectivement donnés par

$$\hat{N}_{F(post)} = \sum_{i \in s_F} \tilde{w}_i x_{1i} = \sum_{i \in s_F} \frac{N}{n} \frac{N_F}{\hat{N}_F} x_{1i} = \frac{N_F}{\hat{N}_F} \hat{N}_F = N_F$$

et

$$\hat{N}_{H(post)} = \sum_{i \in s_H} \tilde{w}_i x_{2i} = \sum_{i \in s_H} \frac{N}{n} \frac{N_H}{\hat{N}_H} x_{2i} = \frac{N_H}{\hat{N}_H} \hat{N}_H = N_H,$$

où s_F et s_H désignent respectivement l'échantillon des femmes et celui des hommes.

Autrement dit, l'utilisation des poids \tilde{w}_i afin d'estimer la taille des post-strates nous ramène aux totaux connus des post-strates, ce qui est une propriété attrayante. Pour cette raison, les poids \tilde{w}_i sont souvent appelés *poids de calage*.

Tableau 5.1 : Fichier de donnée typique fourni par les organismes statistiques

| unité | y_1 | ... | y_p | x_1 | ... | x_q | \tilde{w}_i |
|-------|----------|-----|----------|----------|-----|----------|---------------|
| 1 | y_{11} | ... | y_{p1} | x_{11} | ... | x_{q1} | \tilde{w}_1 |
| 2 | y_{12} | ... | y_{p2} | x_{12} | ... | x_{q2} | \tilde{w}_2 |
| . | . | . | . | . | . | . | . |
| . | . | . | . | . | . | . | . |
| . | . | . | . | . | . | . | . |
| N | y_{1n} | ... | y_{pn} | x_{1n} | ... | x_{qn} | \tilde{w}_n |

Si (a) les tailles n_j sont suffisamment grandes (≥ 30) et (b) la taille de l'échantillon n est grande, alors on peut montrer que la variance de l'estimateur post-stratifié (5.2) peut être estimée par

$$\hat{V}(\bar{y}_{post}) = \left(1 - \frac{n}{N}\right) \sum_{j=1}^J \frac{N_j}{N} \frac{s_j^2}{n}$$

qui est identique à l'expression de l'estimateur de variance obtenu dans le cas de l'échantillonnage aléatoire simple avec une répartition proportionnelle (voir chapitre 3). Il est intéressant de noter que la variance estimée de l'estimateur post-stratifié sera petite lorsque les dispersions s_j^2 seront petites, ce qui surviendra lorsque la variable auxiliaire x utilisée pour créer les post-strates est fortement liée à la variable d'intérêt y . Autrement dit, les post-strates doivent être homogènes par rapport à la variable d'intérêt pour que la post-stratification exhibe un gain substantiel en termes de précision par rapport à l'estimateur usuel \bar{y}_s .

Exemple 5.1 : Un EASSR, s , de taille $n = 100$ individus a été tiré dans une population contenant le même nombre d'hommes et de femmes. La variable d'intérêt est le poids du répondant et l'objectif est d'estimer la moyenne du poids des individus (hommes et femmes) dans la population. On obtient les données suivantes :

| Hommes | Femmes |
|--------------------------|--------------------------|
| $n_1 = 20$ | $n_2 = 80$ |
| $\bar{y}_H = 180$ livres | $\bar{y}_F = 110$ livres |
| $\bar{y}_s = 124$ | |

Il est clair que les hommes sont sous-représentés dans l'échantillon. Par conséquent, l'estimateur usuel qui n'utilise pas d'information auxiliaire mène à $\bar{y}_s = 124$ qui semble beaucoup trop petit. Si on post-stratifie l'échantillon en utilisant le sexe du répondant, on obtient

$$\bar{y}_{post} = \left(\frac{N_1}{N}\right)\bar{y}_1 + \left(\frac{N_2}{N}\right)\bar{y}_2 = 0,5(180) + 0,5(110) = 145.$$

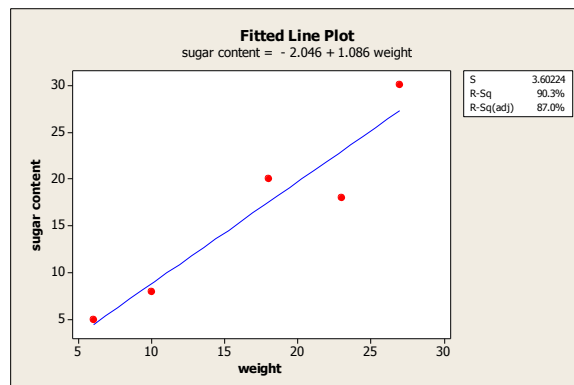
Cette estimation semble beaucoup plus réaliste que la précédente. Notons que dans cet exemple, bien que l'on ne connaisse pas les valeurs de N_1 et N_2 , on sait que $N_j/N \approx 50\%$, $j = 1, 2$.

5.3 L'estimateur par le ratio

Dans cette section, on étudie l'estimateur par le ratio. On suppose qu'une variable auxiliaire *continue* x est disponible. Considérons l'exemple suivant :

Le prix de gros des oranges dans les grosses cargaisons est fonction de la quantité de sucre dans la cargaison. La quantité totale exacte de sucre ne peut bien sûr pas être déterminée avant l'achat et l'extraction du jus pour toute la cargaison. En revanche, on peut l'estimer. Soit t_y la quantité totale de sucre dans la cargaison. On peut estimer t_y en tirant un EASSR de taille n oranges et en utilisant l'estimateur usuel $\hat{t}_y = N\bar{y}_s$ (où \bar{y}_s est un estimateur sans biais de \bar{y}_U , la vraie moyenne de la quantité de sucre par orange). Malheureusement, cet estimateur ne peut être utilisé puisque la taille de la population (i.e., le nombre d'oranges dans la cargaison) est inconnue (le comptage des oranges serait beaucoup trop long). Comment peut-on estimer t_y si la taille de la population N est inconnue? Pour répondre à cette question, notons d'abord les faits suivants :

1. La quantité de sucre dans une orange y est fortement liée à son poids x . Il est raisonnable d'imaginer que la relation entre x et y est de la forme exhibée ci-dessous. Autrement, il est raisonnable de penser que la quantité de sucre dans une orange est approximativement proportionnelle à son poids.



2. Soit t_x le poids total du chargement que l'on suppose connu. On peut alors écrire

$$\frac{\bar{y}_U}{\bar{x}_U} = \frac{N\bar{y}_U}{N\bar{x}_U} = \frac{t_y}{t_x}$$

d'où

$$t_y = \frac{\bar{y}_U}{\bar{x}_U} t_x.$$

Puisque $t_x = \sum_{i \in U} x_i$ est connu, un estimateur de t_y est donné par

$$\hat{t}_{yr} = \hat{B}t_x, \quad (5.3)$$

où $\hat{B} = \frac{\hat{t}_y}{\hat{t}_x}$. L'estimateur (5.3) est appelé *estimateur par le ratio*.

Remarques :

(1) Si l'on cherche à estimer la moyenne dans la population \bar{y}_U , on utilisera

$$\hat{y}_r = \hat{B}\bar{x}_U. \quad (5.4)$$

Dans ce cas, il faudra cependant connaître la moyenne dans la population de la variable auxiliaire, \bar{x}_U .

(2) Notons que les estimateurs par le ratio (5.3) et (5.4) ne requièrent pas la connaissance des valeurs de x pour toutes les unités de la population. Seules les valeurs de x pour les unités échantillonnées ainsi que son total (ou la moyenne) dans la population sont requises.

Pourquoi utiliser l'estimateur par le ratio au lieu de l'estimateur usuel?

1. Comme nous l'avons vu dans l'exemple précédent, l'estimateur par le ratio peut être utilisé lors de l'estimation d'un total lorsque la taille de la population N est inconnue.
2. L'estimateur par le ratio est fréquemment utilisé afin d'améliorer la précision des estimateurs. En effet, lorsque la variable y est approximativement proportionnelle à la variable x , l'estimateur par le ratio \hat{t}_{yr} peut être considérablement plus précis que l'estimateur usuel, \hat{t}_y . En fait, si $y_i = cx_i$, pour une constante donnée c , l'estimateur par le ratio est égal à $\hat{t}_{yr} = ct_x$, qui est une constante (i.e., non-aléatoire). Dans ce

cas, la variance de l'estimateur par le ratio est égale à 0. Bien sûr, il n'est pas réaliste de supposer que la variable y est parfaitement proportionnelle à la variable x . Il faut cependant s'attendre à ce que l'estimateur par le ratio soit très précis lorsque la variable y est approximativement proportionnelle à x .

3. Supposons que l'on cherche à estimer le total du revenu y au temps t pour les entreprises dans une région donnée. Pour cela, on tire un EASSR de taille n dans la population des entreprises de la région. On se rend compte après la sélection de l'échantillon que l'on fait face à un mauvais échantillon car ce dernier contient les n plus petites entreprises de la région. Dans ce cas, l'utilisation de l'estimateur usuel

$$\hat{t}_y = \frac{N}{n} \sum_{i \in s} y_i$$

va clairement sous-estimer le vrai total dans la population. Supposons que le revenu x des entreprises au temps $t-1$ est disponible pour toutes les unités échantillonnées et que le total du revenu au temps $t-1$ est également disponible. Il y a fort à parier que la relation entre le revenu au temps $t-1$ et celui au temps t est linéaire et qu'elle passe par l'origine. De plus, on peut supposer sans trop de risques que cette relation est forte. Dans ce cas, on s'attend à ce que le ratio $\frac{t_x}{\hat{t}_x}$ soit

significativement supérieur à 1 puisqu'une petite valeur de \hat{t}_y entrainera une petite valeur de \hat{t}_x . L'utilisation de l'estimateur par le ratio, $\hat{t}_{yr} = \frac{t_x}{\hat{t}_x} \hat{t}_y$, aura comme effet

de redresser à la hausse l'estimation (trop petite) obtenue à l'aide de l'estimateur usuel. De façon similaire, si les entreprises sélectionnées sont les n plus grandes, l'estimateur usuel \hat{t}_y donnera une valeur beaucoup trop grande. Dans ce cas, on

s'attend à ce que le ratio $\frac{t_x}{\hat{t}_x}$ soit significativement inférieur à 1 et l'utilisation de

l'estimateur par le ratio aura comme effet de redresser à la baisse l'estimation (trop grande) obtenue à l'aide de l'estimateur usuel. L'estimateur usuel \hat{t}_y et l'estimateur

par le ratio \hat{t}_{yr} coïncident lorsque $\frac{t_x}{\hat{t}_x} = 1$. Le ratio $\frac{t_x}{\hat{t}_x}$ peut donc être vu comme un

ajustement que l'on applique lorsque l'échantillon n'est pas équilibré par rapport à la variable x .

4. L'estimateur par le ratio \hat{t}_{yr} peut également s'écrire comme $\hat{t}_{yr} = \sum_{i \in s} \tilde{w}_i y_i$, où

$$\tilde{w}_i = \frac{N}{n} \frac{t_x}{\hat{t}_x}$$

Autrement dit, l'estimateur par le ratio s'exprime comme une somme pondérée des valeurs de y , pondérée par les poids \tilde{w}_i . Notons que $\hat{t}_{xr} = \sum_{i \in s} \tilde{w}_i x_i = t_x$.

Appliquer les poids \tilde{w}_i à la variable auxiliaire x nous ramène donc au total connu dans la population t_x . Pour cette raison, les poids \tilde{w}_i sont appelés *poids de calage*.

L'estimateur $\hat{t}_{yr} = \hat{B}t_x$ n'est pas sans biais pour t_y car il s'écrit comme une fonction non-linéaire de totaux estimés. Puisque t_x est une constante, on peut utiliser la Proposition 4.2 pour obtenir une expression du biais approximatif de \hat{t}_{yr} .

Proposition 5.1 : Dans le cas de l'EASSR, le biais approximatif de \hat{t}_{yr} est donné par

$$E(\hat{t}_{yr} - t_y) \approx N \left(1 - \frac{n}{N}\right) \frac{1}{n\bar{x}_U} (BS_x^2 - RS_x S_y),$$

où $R = \frac{\sum_{i \in U} (x_i - \bar{x}_U)(y_i - \bar{y}_U)}{(N-1)S_x S_y}$.

Démonstration : Elle découle directement de la Proposition 4.2 en remplaçant y_{li} par y_i et y_{2i} par x_i et en utilisant le fait que t_x est une constante.

Proposition 5.2 : Dans le cas de l'EASSR, la variance de \hat{t}_{yr} peut être approximée par

$$V(\hat{t}_{yr}) \approx N^2 \left(1 - \frac{n}{N}\right) \frac{S_e^2}{n}, \quad (5.5)$$

où $S_e^2 = \frac{1}{N-1} \sum_{i \in U} (y_i - Bx_i)^2$.

Démonstration : Elle découle directement de la Proposition 4.3 en remplaçant y_{li} par y_i et y_{2i} par x_i et en utilisant le fait que t_x est une constante.

Corollaire 5.1 : Une expression alternative de la variance de \hat{t}_{yr} est donnée par

$$V(\hat{t}_{yr}) \approx N^2 \left(1 - \frac{n}{N}\right) \frac{1}{n} [S_y^2 - 2BRS_x S_y + B^2 S_x^2]. \quad (5.6)$$

Démonstration : Elle découle directement du Corollaire 4.1 en remplaçant y_{li} par y_i et y_{2i} par x_i et en utilisant le fait que t_x est une constante.

Les expressions (5.5) et (5.6) suggèrent que la variance de l'estimateur par le ratio est petite lorsque

- (i) la taille n de l'échantillon est grande;
- (ii) la fraction de sondage n/N est grande;
- (vi) les « résidus », $y_i - Bx_i$, sont petits;
- (vii) le coefficient de corrélation R est près de 1.

Bien sûr, la variance approximative de \hat{t}_{yr} ne peut être calculée puisqu'elle dépend de paramètres inconnus. Il faudra donc l'estimer.

Proposition 5.3 : Un estimateur approximativement sans biais de la variance de \hat{t}_{yr} en (5.5) est donné par

$$\hat{V}(\hat{t}_{yr}) = N^2 \left(1 - \frac{n}{N}\right) \frac{s_e^2}{n}, \quad (5.7)$$

où $s_e^2 = \frac{1}{n-1} \sum_{i \in S} (y_i - \hat{B}x_i)^2$.

Démonstration : Elle découle directement de la Proposition 4.4 en remplaçant y_{1i} par y_i et y_{2i} par x_i et en utilisant le fait que t_x est une constante.

Remarque : La variance de l'estimateur par le ratio d'une moyenne (5.4) est obtenue simplement en divisant la variance (5.5) ou (5.6) par N^2 . De manière similaire, un estimateur de la variance de l'estimateur (5.4) est obtenu en divisant (5.7) par N^2 .

À ce stade, une question légitime vient à l'esprit : l'estimateur par le ratio est-il toujours plus précis que l'estimateur usuel? La réponse est malheureusement non comme le montre la prochaine proposition.

Proposition 5.4 : Supposons que la taille d'échantillon n soit suffisamment grande pour qu'on puisse négliger le biais de \hat{t}_{yr} . Alors, $EQM(\hat{t}_{yr}) \leq EQM(\hat{t}_y)$ si et seulement si

$$R \geq \frac{1}{2} \frac{CV(x)}{CV(y)}, \quad (5.8)$$

où $CV(x)$ et $CV(y)$ désignent les coefficients de variation pour les variables x et y , donnés par S_x/\bar{x}_U et S_y/\bar{y}_U , respectivement.

Démonstration :

La Proposition 5.4 suggère que lorsque le ratio $CV(x)/CV(y)$ est près de 1, l'estimateur par le ratio est plus précis que l'estimateur usuel si la corrélation entre les variables x et y est supérieure à $1/2$. L'estimateur par le ratio est approprié lorsque la relation entre x et y ne passe pas « trop loin » de l'origine. Si l'ordonnée à l'origine est fortement significative, il est fort probable que, dans ce cas, le ratio $CV(x)/CV(y)$ soit supérieur à 1.

Exemple 5.2 : (Lohr, 2010, exercice no. 3 p. 156) Des chercheurs veulent estimer l'âge des arbres dans une petite forêt. Cependant, déterminer l'âge d'un arbre n'est pas aisé car il faut, pour cela, compter le nombre d'anneaux au niveau du tronc. Par contre, on sait que plus un arbre est vieux, plus le diamètre de celui-ci est grand. Le diamètre étant relativement facile à obtenir, les chercheurs décident donc de mesurer celui des 1132 arbres de la forêt et déterminent que le diamètre moyen dans la population est égal à 10,3 pieds. Les chercheurs tirent alors un EASSR de taille $n = 20$ arbres.

(a) Estimez \bar{y}_U .

(b) Construisez un intervalle de confiance de niveau 95% pour \bar{y}_U .

| Arbre no. | Diamètre, x | Âge, y | Arbre no. | Diamètre, x | Âge, y |
|-----------|---------------|----------|-----------|---------------|----------|
| 1 | 12,0 | 125 | 11 | 5,7 | 61 |
| 2 | 11,4 | 119 | 12 | 8,0 | 80 |
| 3 | 7,9 | 83 | 13 | 10,3 | 114 |
| 4 | 9,0 | 85 | 14 | 12,0 | 147 |
| 5 | 10,5 | 99 | 15 | 9,2 | 122 |
| 6 | 7,9 | 117 | 16 | 8,5 | 106 |
| 7 | 7,3 | 69 | 17 | 7,0 | 82 |
| 8 | 10,2 | 133 | 18 | 10,7 | 88 |
| 9 | 11,7 | 154 | 19 | 9,3 | 97 |
| 10 | 11,3 | 168 | 20 | 8,2 | 99 |

D'abord, effectuons une analyse de régression afin de déterminer si l'estimateur par le ratio est approprié dans ce cas. La sortie 1 montre que l'ordonnée à l'origine n'est pas significative (p -value = 0,74) ce qui permet de penser que la relation entre l'âge et le diamètre passe par l'origine. De plus, un simple coup d'œil à la Figure 5.2 semble montrer que la relation est bien de nature linéaire. Dans ce cas, l'estimateur par le ratio semble approprié.

Sortie 1 : Analyse de régression - Âge vs. Diamètre

The regression equation is
Age = - 7.5 + 12.2 Diameter

| Predictor | Coef | SE Coef | T | P |
|-----------|--------|---------|-------|-------|
| Constant | -7.52 | 22.32 | -0.34 | 0.740 |
| Diameter | 12.209 | 2.331 | 5.24 | 0.000 |

S = 18.5856 R-Sq = 60.4% R-Sq(adj) = 58.2%

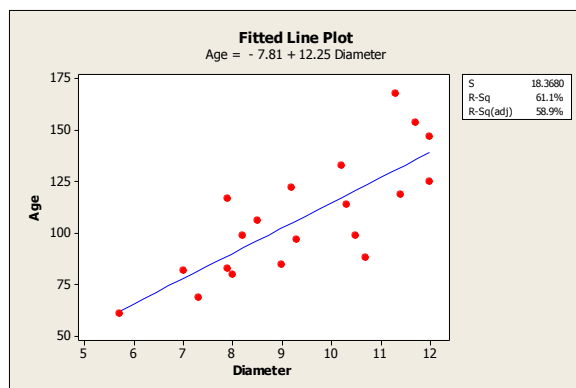


Figure 5.2 : Âge vs. diamètre

- (a) On a $\bar{x}_U = 10,3$. L'estimateur par le ratio mène à $\hat{y}_r = \hat{B}\bar{x}_U = 11,41 \times 10,3 = 117,6$.
- (b) L'estimateur de la variance de \hat{y}_r est donné par $\hat{V}(\hat{y}_r) = \left(1 - \frac{n}{N}\right) \frac{s_e^2}{n}$. Après calculs, on obtient $\hat{V}(\hat{y}_r) = 40,32$. Un intervalle de confiance de niveau 95% est donné par $117,6 \pm 1,96(6,35)$.

5.4 L'estimateur par la régression

À la section 5.3, on a vu que l'estimateur par le ratio est approprié lorsque la relation qui lie les variables x et y est linéaire et que la droite passe par l'origine. Souvent, la relation liant x et y est bien linéaire mais la droite ne passe pas par l'origine. En effet, la relation est souvent de la forme

$$y = B_0 + B_1x. \quad (5.9)$$

On suppose que la moyenne dans la population, \bar{x}_U , de la variable auxiliaire x est connue. On cherche à estimer la moyenne de la population \bar{y}_U . Dans la population, on tire un EASSR, s , de taille n . En utilisant (5.9), la moyenne \bar{y}_U peut s'écrire comme

$$\bar{y}_U = \frac{1}{N} \sum_{i \in U} y_i = \frac{1}{N} \sum_{i \in U} (B_0 + B_1x_i) = B_0 + B_1\bar{x}_U. \quad (5.10)$$

Puisque la moyenne \bar{x}_U est connue, un estimateur de \bar{y}_U , que nous appellerons *l'estimateur par la régression*, est obtenu en estimant B_0 et B_1 en (5.10), ce qui mène

$$\hat{y}_{reg} = \hat{B}_0 + \hat{B}_1\bar{x}_U, \quad (5.11)$$

où $\hat{B}_1 = \frac{\sum_{i \in s} (x_i - \bar{x}_s)(y_i - \bar{y}_s)}{\sum_{i \in s} (x_i - \bar{x}_s)^2}$ et $\hat{B}_0 = \bar{y}_s - \hat{B}_1\bar{x}_s$. L'estimateur par la régression peut également s'écrire comme

$$\hat{y}_{reg} = \bar{y}_s + \hat{B}_1(\bar{x}_U - \bar{x}_s). \quad (5.12)$$

L'expression (5.12) suggère que l'estimateur par la régression s'exprime comme la somme de la moyenne échantillonnale et un terme d'ajustement, $\hat{B}_1(\bar{x}_U - \bar{x}_s)$. Notons que $\hat{y}_{reg} = \bar{y}_s$ si $\hat{B}_1 = 0$ (i.e., il n'y a pas de relation entre x et y) ou si $\bar{x}_s = \bar{x}_U$.

L'estimateur par la régression (à l'instar de l'estimateur par le ratio) est généralement biaisé puisque le coefficient \hat{B}_1 est une fonction non-linéaire de totaux estimés. Le biais de \hat{y}_{reg} est donné par

$$\begin{aligned} \text{Biais}(\hat{y}_{reg}) &= E(\hat{y}_{reg}) - \bar{y}_U \\ &= E(\bar{y}_s) + E[\hat{B}_1(\bar{x}_U - \bar{x}_s)] - \bar{y}_U \\ &= E[\hat{B}_1(\bar{x}_U - \bar{x}_s)] \\ &= -\text{Cov}(\hat{B}_1, \bar{x}_s). \end{aligned}$$

Le biais de \hat{y}_{reg} est donc égal à zéro si la covariance entre \hat{B}_1 et \bar{x}_s est nulle, ce qui survient, par exemple, lorsque la droite passe par tous les points (x_i, y_i) dans la population.

En utilisant un développement par séries de Taylor, on peut montrer que l'estimateur par la régression est approximativement sans biais pour \bar{y}_U ; i.e.,

$$E(\hat{y}_{reg}) \approx \bar{y}_U.$$

Autrement dit, on considère que le biais de \hat{y}_{reg} est négligeable lorsque la taille de l'échantillon est suffisamment grande. On a donc

$$EQM(\hat{y}_{reg}) = V(\hat{y}_{reg}) + \text{Biais}(\hat{y}_{reg})^2 \approx V(\hat{y}_{reg}).$$

Proposition 5.5 : Dans le cas de l'EASSR, la variance de l'estimateur par la régression \hat{y}_{reg} peut être approximée par

$$V(\hat{y}_{reg}) \approx \left(1 - \frac{n}{N}\right) \frac{S_d^2}{n}, \quad (5.13)$$

où $S_d^2 = \frac{1}{N-1} \sum_{i \in U} (y_i - B_0 - B_1 x_i)^2$.

Démonstration : omise.

Corollaire 5.2 : La variance approximative (5.13) peut également s'écrire comme

$$V(\hat{y}_{reg}) \approx \left(1 - \frac{n}{N}\right) \frac{S_y^2}{n} (1 - R^2), \quad (5.14)$$

où $R = \frac{\sum_{i \in U} (x_i - \bar{x}_U)(y_i - \bar{y}_U)}{(N-1)S_x S_y}$.

Démonstration :

La variance approximative (5.14) est petite lorsque

- (i) la taille n de l'échantillon est grande;
- (ii) la fraction de sondage n/N est grande;
- (iii) la dispersion de la variable y dans la population, S_y , est petite;
- (iv) la corrélation R est près de 1 ou de -1.

La variance (5.13) ou (5.14) ne peut être calculée car elle dépend de paramètres de population. Il faudra donc l'estimer.

Proposition 5.6 : Un estimateur approximativement sans biais de $V(\hat{y}_{reg})$ en (5.14) est donné par

$$\hat{V}(\hat{y}_{reg}) = \left(1 - \frac{n}{N}\right) \frac{s_d^2}{n}, \quad (5.15)$$

où $s_d^2 = \frac{1}{n-1} \sum_{i \in S} (y_i - \hat{B}_0 - \hat{B}_1 x_i)^2$.

Démonstration : omise.

Dans ce qui suit, on compare l'estimateur par la régression à l'estimateur usuel et à l'estimateur par le ratio.

Proposition 5.7 : Supposons que la taille d'échantillon n soit suffisamment grande pour qu'on puisse négliger le biais de \hat{y}_{reg} . Alors, dans le cas de l'EASSR, on a $EQM(\hat{y}_{reg}) \leq EQM(\bar{y}_s)$ avec égalité si $R^2 = 0$.

Démonstration : laissée en exercice.

Proposition 5.8 : Supposons que la taille d'échantillon n soit suffisamment grande pour qu'on puisse négliger le biais de \hat{y}_r et celui de \hat{y}_{reg} . Dans le cas de l'EASSR, on a $EQM(\hat{y}_{reg}) \leq EQM(\hat{y}_r)$ avec égalité si et seulement si $B_1 = \frac{\bar{y}_U}{\bar{x}_U}$.

Démonstration : laissée en exercice.

Remarques :

(1) Si l'on cherche à estimer un total, $t_y = \sum_{i \in U} y_i$, plutôt qu'une moyenne, on utilisera l'estimateur par la régression, $\hat{t}_{yreg} = N\hat{y}_{reg}$. La variance de \hat{t}_{yreg} peut être approximée par $V(\hat{t}_{yreg}) = N^2V(\hat{y}_{reg})$ et un estimateur de cette variance est donné par $\hat{V}(\hat{t}_{yreg}) = N^2\hat{V}(\hat{y}_{reg})$.

(2) Les propositions 5.7 et 5.8 établissent que pour des grandes tailles d'échantillons, l'estimateur par la régression est toujours plus précis que l'estimateur usuel ou que l'estimateur par le ratio. Cette propriété n'est vraie que pour un plan EASSR. En général, on ne peut extrapoler ces résultats à des plans de sondage quelconques. De plus, il ne faut pas perdre de vue que, contrairement à l'estimateur usuel, l'estimateur par la régression peut être considérablement biaisé si la taille n de l'échantillon est très petite ($n < 10$).

(3) Le lecteur vérifiera que l'estimateur par la régression \hat{t}_{yreg} peut également s'écrire comme $\hat{t}_{yreg} = \sum_{i \in s} \tilde{w}_i y_i$, pour un certain choix de poids \tilde{w}_i .

Exemple 5.3 : Avant l'entrée dans un collège, 486 étudiants ont été soumis à un test de mathématiques. Un EASSR de taille $n = 10$ étudiants a été tiré afin d'étudier leur progrès et leur note finale dans le cours de calcul a été observé. Les notes finales du cours de calcul ainsi que les résultats du test de mathématique sont exhibés dans le tableau présenté ci-dessous. De plus, on sait que la moyenne des 486 étudiants pour le test de mathématiques, \bar{x}_U , est égale à 52.

- (a) Estimez la moyenne, \bar{y}_U , de la note finale de calcul dans la population composée des 486 étudiants.
- (b) Construisez un intervalle de confiance de niveau 95% pour la moyenne de la note finale de calcul.

| Étudiant | Résultat du test | |
|----------|-----------------------|----------------------------|
| | de mathématiques, x | Note finale en calcul, y |
| 1 | 39 | 65 |
| 2 | 43 | 78 |
| 3 | 21 | 52 |
| 4 | 64 | 82 |
| 5 | 57 | 92 |
| 6 | 47 | 89 |
| 7 | 28 | 73 |
| 8 | 75 | 98 |
| 9 | 34 | 55 |
| 10 | 52 | 75 |

D'abord, effectuons une analyse de régression afin de déterminer si l'estimateur par la régression est approprié dans ce cas. La sortie 2 montre que l'ordonnée à l'origine est significative (p -value = 0,002) ce qui élimine l'estimateur par le ratio comme candidat possible. De plus, un simple coup d'œil à la Figure 5.3 semble montrer que la relation est bien de nature linéaire. L'estimateur par la régression semble donc approprié.

Sortie 2 : Analyse de régression - Note finale en calcul vs. Note au test de mathématiques

The regression equation is
 Final calculus grade = 40.5 + 0.770 Achievement test

| Predictor | Coef | SE Coef | T | P |
|-----------------|--------|---------|------|-------|
| Constant | 40.461 | 8.663 | 4.67 | 0.002 |
| Achievment test | 0.7704 | 0.1782 | 4.32 | 0.003 |

S = 8.86351 R-Sq = 70.0% R-Sq(adj) = 66.3%

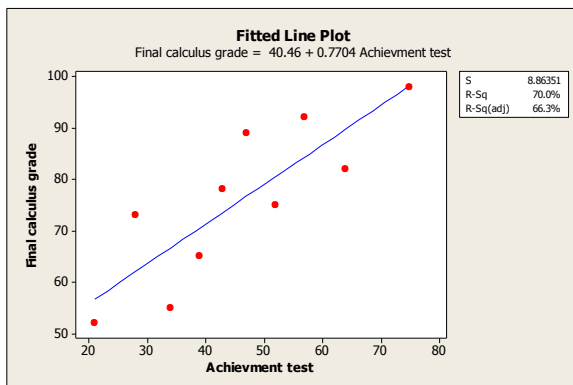


Figure 5.3 : Note finale dans le cours de calcul vs. Résultat du test de mathématiques

(a) On a $\hat{y}_{reg} = \bar{y}_s + \hat{B}_1(\bar{x}_U - \bar{x}_s) = 76 + (0,77)(52 - 46) = 80,6$.

(b) Un intervalle de confiance de niveau 95% est donné par $\hat{y}_{reg} \pm 1,96\sqrt{\hat{V}(\hat{y}_{reg})}$, où $\hat{V}(\hat{y}_{reg})$ est donnée par (5.15). Après calculs, on obtient

$$s_d^2 = \frac{1}{n-1} \sum_{i \in s} (y_i - \hat{B}_0 - \hat{B}_1 x_i)^2 = 69,83 \text{ et}$$

$$\hat{V}(\hat{y}_{reg}) = \left(1 - \frac{10}{486}\right) \frac{69,83}{10} = 6,83. \text{ L'intervalle de confiance recherché est donc donné par}$$

$$80,6 \pm 1,96\sqrt{6,83} \text{ ou } 80,6 \pm 5,12.$$

5.5 L'estimateur par la différence

L'estimateur par la différence, étudié dans cette section, peut être vu comme un cas particulier de l'estimateur par la régression. Il est utilisé lorsque que le statisticien « sait » que la pente B_1 est égale à 1. Cela survient fréquemment dans un contexte de vérification comptable (voir exemple 5.4). L'estimateur par la différence d'une moyenne \bar{y}_U est donc donné par

$$\hat{y}_{diff} = \bar{y}_s + (\bar{x}_U - \bar{x}_s). \quad (5.16)$$

L'estimateur \hat{y}_{diff} est sans biais pour \bar{y}_U puisque $E(\bar{y}_s) = \bar{y}_U$ et $E(\bar{x}_s) = \bar{x}_U$. Il est souvent employé dans la cadre de procédures comptables de vérification des rapports d'impôt (individuels et de compagnies).

La variance de l'estimateur par la différence est donnée par

$$V(\hat{y}_{diff}) = \left(1 - \frac{n}{N}\right) \frac{S_d^2}{n}, \quad (5.17)$$

où $S_d^2 = \frac{1}{N-1} \sum_{i \in U} (d_i - \bar{d}_U)^2$, avec $d_i = y_i - x_i$ et $\bar{d}_U = \frac{1}{N} \sum_{i \in U} d_i$. Un estimateur sans biais de la variance (5.17) est donné par

$$\hat{V}(\hat{y}_{diff}) = \left(1 - \frac{n}{N}\right) \frac{s_d^2}{n}, \quad (5.18)$$

où $s_d^2 = \frac{1}{n-1} \sum_{i \in s} (d_i - \bar{d}_s)^2$ et $\bar{d}_s = \frac{1}{n} \sum_{i \in s} d_i$.

Exemple 5.4 : Les vérificateurs comptables sont fréquemment intéressés à comparer la valeur vérifiée de certains items avec leur valeur comptable. Généralement, la valeur comptable est connue pour toutes les unités de la population (par exemple, la population des contribuables au Canada) et la valeur vérifiée est obtenue par les vérificateurs pour un échantillon seulement. La valeur comptable peut être vue comme une variable auxiliaire pour améliorer la qualité de l'estimation de la moyenne des valeurs vérifiées dans la population. Supposons que la population est composée de 180 items ayant une valeur comptable totale égale à 13 320\$. Soit x_i la valeur comptable pour l'item i et y_i sa valeur vérifiée. Un EASSR de taille $n = 10$ a permis de recueillir les données présentées dans le tableau ci-dessous.

- (a) Estimer la moyenne des valeurs vérifiées dans la population, \bar{y}_U .
 (b) Construire un intervalle de confiance de niveau 95% pour \bar{y}_U .

| Échantillon | Valeur vérifiée, y_i | Valeur comptable, x_i | $d_i = y_i - x_i$ |
|-------------|------------------------|-------------------------|-------------------|
| 1 | 9 | 10 | -1 |
| 2 | 14 | 12 | +2 |
| 3 | 7 | 8 | -1 |
| 4 | 29 | 26 | +3 |
| 5 | 45 | 47 | -2 |
| 6 | 109 | 112 | -3 |
| 7 | 40 | 36 | +4 |
| 8 | 238 | 240 | -2 |
| 9 | 60 | 59 | +1 |
| 10 | 170 | 167 | +3 |

D'abord, effectuons une analyse de régression afin de déterminer si l'estimateur par la différence est approprié dans ce cas. La sortie 3 montre que le coefficient $\hat{B}_1 = 0,99251 \approx 1$. De plus, un simple coup d'œil à la Figure 5.4 semble montrer que la relation est bien de nature linéaire. L'estimateur par la différence semble donc approprié.

Sortie 3 : Analyse de régression : Valeur comptable vs. Valeur vérifiée

The regression equation is
 Audit value = 0.94 + 0.993 Book value

| Predictor | Coef | SE Coef | T | P |
|------------|---------|---------|-------|-------|
| Constant | 0.937 | 1.138 | 0.82 | 0.434 |
| Book value | 0.99251 | 0.01105 | 89.81 | 0.000 |

S = 2.58212 R-Sq = 99.9% R-Sq(adj) = 99.9%

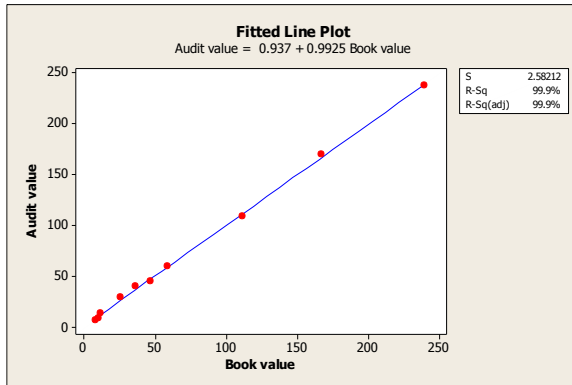


Figure 5.4 : Valeur comptable vs. valeur vérifiée

(a) On a $\hat{y}_{diff} = \bar{y}_s + (\bar{x}_U - \bar{x}_s) = 72,1 + (74 - 71,7) = 74,4$.

(b) Un intervalle de confiance de niveau 95% est donné par $\hat{y}_{diff} \pm 1,96\sqrt{\hat{V}(\hat{y}_{diff})}$, où $\hat{V}(\hat{y}_{diff})$ est donnée par (5.18). Après quelques calculs, on obtient

$s_d^2 = \frac{1}{n-1} \sum_{i \in S} (d_i - \bar{d}_s)^2 = 6,44$ et $\hat{V}(\hat{y}_{diff}) = \left(1 - \frac{10}{180}\right) \frac{6,44}{10} = 0,6086$. L'intervalle de confiance recherché est donc donné par

$$74,4 \pm 1,96\sqrt{0,6068} \text{ ou } 74,4 \pm 1,52.$$

Échantillonnage par grappes à probabilités égales

Chapitre 6

6.1 Introduction

Toutes les procédures d'échantillonnage étudiées jusqu'à maintenant supposaient qu'un échantillonnage direct des éléments étaient possibles. La base de sondage était composée des éléments. Autrement dit, l'unité d'échantillonnage et l'élément coïncidaient, ce qui n'est pas toujours le cas en pratique. Dans une bonne partie des enquêtes sociales et des enquêtes auprès des ménages, l'échantillonnage direct n'est pas utilisé pour l'une des deux raisons suivantes :

- (i) Une base de sondage identifiant chacun des éléments dans la population n'est pas disponible et le coût pour la construction d'une telle base serait exorbitant.
- (ii) La population d'intérêt est dispersée dans une aire géographique très vaste (par exemple, le Canada) et un échantillonnage direct résulterait vraisemblablement en un échantillon très dispersé. Dans ce cas, les coûts de voyages (si des interviews sur place sont utilisées) seraient potentiellement très élevés. De plus, la supervision du travail sur le terrain pourrait s'avérer difficile, ce qui pourrait résulter en de haut taux de non-réponse et erreurs de mesure.

Lorsque l'échantillonnage direct n'est pas possible, on a souvent recours à l'échantillonnage par grappes.

Exemples :

- (1) Supposons que l'on cherche à étudier les habitudes de consommation de drogues dans la population des étudiants du secondaire V au Québec. Il n'existe pas une liste d'étudiants du secondaire V au Québec. Par contre, une liste des écoles est disponible. Dans ce cas, il serait judicieux de tirer des écoles au hasard dans la base de sondage (liste de toutes les écoles au Québec). Dans chaque école sélectionnée, on pourrait alors enquêter tous les étudiants du secondaire V. On vient de donner un exemple de l'échantillonnage par grappe à un degré, où les *grappes* (ou *unités primaires d'échantillonnage (upe)*) sont les écoles et les éléments (ou *unités secondaires d'échantillonnage (use)*) sont les étudiants du secondaire V. Si au lieu d'enquêter tous les étudiants dans chaque école sélectionnée, on tire au hasard des classes du secondaire V et, dans chaque classe sélectionnée, on enquête tous les étudiants, alors, on est dans un cas d'échantillonnage à deux degrés. Si au lieu d'enquêter tous les étudiants dans chaque classe sélectionnée, on tire des étudiants au hasard dans chaque classe sélectionnée, alors on est dans un cas d'échantillonnage à trois degrés. Dans ce cas, les écoles sont les upe, les classes sont les use et les étudiants sont *les unités tertiaires (ute)*. On en déduit que le nombre de degrés correspond au nombre de mécanismes aléatoires utilisés afin d'avoir accès aux éléments que l'on va enquêter.

(2) Supposons que l'on cherche à estimer le revenu moyen des ménages dans une grande ville. Comment choisir un échantillon? Si l'on voulait tirer un EASSR ou un échantillon systématique de ménages, il faudrait que nous disposions d'une base de sondage des ménages de la ville. Une telle base n'est habituellement pas disponible. Pour contrer ce problème, on pourrait diviser la ville en régions (par exemple, en îlots) et tirer un échantillon aléatoire d'îlots. On pourrait alors mesurer le revenu de tous les ménages des îlots tirés. Dans cet exemple, on fait face, encore une fois, à l'échantillonnage par grappes à un degré dans lequel les îlots sont les upe et les ménages sont les use. Si au lieu d'enquêter tous les ménages dans un îlot sélectionné, on tire au hasard un échantillon de ménages, on est dans un cas d'échantillonnage à deux degrés.

La première tâche consiste à définir les upe de façon appropriée (par exemple, la taille des grappes). Les éléments à l'intérieur d'une grappe tendent à être près physiquement l'un de l'autre et donc à posséder des caractéristiques similaires. Autrement dit, les éléments à l'intérieur des grappes tendent à être fortement corrélés. Dans ce cas, il n'est pas nécessaire de tirer un grand nombre d'éléments à l'intérieur d'une grappe puisque ceux-ci présentent des caractéristiques similaires, compte tenu du fait que tirer des éléments supplémentaires mène potentiellement à des coûts substantiels. Dans certaines situations, il est possible que les éléments à l'intérieur d'une grappe soient très différents les uns des autres, auquel cas, un échantillon contenant quelques grandes grappes pourrait mener à des estimations de bonne qualité. Le problème consistant à définir des grappes appropriées est loin d'être trivial en pratique.

6.2 Notation pour l'échantillonnage par grappes

Considérons une population composée de N upe, U_1, \dots, U_N , de taille M_1, \dots, M_N , respectivement. De cette population d'upe, on tire un EASSR, s , de taille n upe. Soit s_i l'échantillon d'éléments (use) tiré dans la i^{e} upe. La notation ci-dessous sera utilisée dans le présent chapitre ainsi que dans le chapitre 7.

Soit y_{ij} la valeur d'une variable d'intérêt y pour le j^{e} élément dans la i^{e} grappe. Notons que dans les chapitres 6 et 7, N représente le nombre d'upe dans la population et non pas le nombre d'éléments!

Paramètres de population au niveau des upe (grappes)

N = nombre d'upe dans la population

M_i = nombre d'use dans la i^{e} upe

$K = \sum_{i \in U} M_i$ = nombre total d'use dans la population

$t_i = \sum_{j \in U_i} y_{ij}$ = total dans la i^{e} upe

$t_y = \sum_{i \in U} t_i = \sum_{i \in U} \sum_{j \in U_i} y_{ij}$ = total dans la population

$S_t^2 = \frac{1}{N-1} \sum_{i \in U} \left(t_i - \frac{t_y}{N} \right)^2$ = dispersion des totaux d'upe dans la population

Paramètres de population au niveau des use (éléments)

$\bar{y}_U = \frac{1}{K} \sum_{i \in U} \sum_{j \in U_i} y_{ij}$ = moyenne dans la population de la variable d'intérêt y

$\bar{y}_{iU} = \frac{1}{M_i} \sum_{j \in U_i} y_{ij} = \frac{t_i}{M_i}$ = moyenne dans la i^{e} upe

$S_y^2 = \frac{1}{(K-1)} \sum_{i \in U} \sum_{j \in U_i} (y_{ij} - \bar{y}_U)^2$ = dispersion de la variable y dans la population

$S_i^2 = \frac{1}{(M_i-1)} \sum_{j \in U_i} (y_{ij} - \bar{y}_{iU})^2$ = dispersion de la variable y dans la i^{e} upe

Quantités échantillonnables

n = nombre d'upe dans l'échantillon

m_i = nombre d'éléments (use) dans l'échantillon appartenant à la i^{e} upe, s_i .

$\bar{y}_i = \frac{1}{m_i} \sum_{j \in s_i} y_{ij}$ = moyenne estimée de la variable d'intérêt y dans la i^{e} upe

$\hat{t}_i = \frac{M_i}{m_i} \sum_{j \in s_i} y_{ij}$ = total estimé dans la i^{e} upe

$\hat{t}_y = \frac{N}{n} \sum_{i \in s} \hat{t}_i$ = estimateur du total pour une variable d'intérêt y

$s_t^2 = \frac{1}{n-1} \sum_{i \in s} \left(\hat{t}_i - \frac{\hat{t}_y}{N} \right)^2$ = dispersion des totaux d'upe dans l'échantillon

$s_i^2 = \frac{1}{m_i-1} \sum_{j \in s} (y_{ij} - \bar{y}_i)^2$ = dispersion dans l'échantillon de la variable y dans la i^{e} upe

6.3 Échantillonnage par grappes à un degré

6.3.1 Estimation

Considérons une population composée de N upe. De cette population, on tire un EASSR, s , de taille n upe. Dans le cas de l'échantillonnage par grappes à un degré, tous les éléments dans les upe sélectionnées sont tirés et alors on a $m_i = M_i$. Supposons que l'on cherche à estimer le total dans la population, $t_y = \sum_{i \in U} t_i$. Les totaux t_i étant connus pour les upe sélectionnées, un estimateur de t_y est donc défini selon

$$\hat{t}_y = \frac{N}{n} \sum_{i \in s} t_i. \quad (6.1)$$

L'estimateur \hat{t}_y en (6.1) est sans biais pour t_y puisque

$$E(\hat{t}_y) = \frac{N}{n} \sum_{i \in U} E(Z_i) t_i = \frac{N}{n} \sum_{i \in U} \frac{n}{N} t_i = \sum_{i \in U} t_i = t_y,$$

où

$$Z_i = \begin{cases} 1 & \text{si l'upe } i \text{ est tirée dans l'échantillon} \\ 0 & \text{sinon} \end{cases}$$

De plus, la variance de \hat{t}_y est donnée par

$$V(\hat{t}_y) = N^2 \left(1 - \frac{n}{N}\right) \frac{S_t^2}{n}, \quad (6.2)$$

où $S_t^2 = \frac{1}{N-1} \sum_{i \in U} \left(t_i - \frac{t_y}{N}\right)^2$. Un estimateur de la variance en (6.2) est donné par

$$\hat{V}(\hat{t}_y) = N^2 \left(1 - \frac{n}{N}\right) \frac{s_t^2}{n},$$

où $s_t^2 = \frac{1}{n-1} \sum_{i \in s} \left(t_i - \frac{\hat{t}_y}{N}\right)^2$.

Remarque : Pour estimer la moyenne dans la population, \bar{y}_U , on utilisera le fait que

$\bar{y}_U = \frac{t_y}{K}$, où $K = \sum_{i \in U} M_i$ désigne le nombre total d'éléments dans la population. Donc, un

estimateur de \bar{y}_U est donné par $\hat{\bar{y}}_U = \frac{\hat{t}_y}{K}$, sa variance est égale à $V(\hat{\bar{y}}_U) = \frac{N^2}{K^2} \left(1 - \frac{n}{N}\right) \frac{S_t^2}{n}$

et un estimateur de la variance est donné par $\hat{V}(\hat{\bar{y}}_U) = \frac{N^2}{K^2} \left(1 - \frac{n}{N}\right) \frac{s_t^2}{n}$.

Exemple 6.1 : Un étudiant cherche à estimer la moyenne académique des étudiants vivant dans les résidences universitaires. Ne pouvant obtenir une liste complète des étudiants vivants dans la résidence, il remarque que celle-ci est composée de 100 chambres, chacune contenant quatre étudiants. Il décide donc de tirer 5 chambres selon un plan aléatoire simple sans remise et, dans chaque chambre sélectionnée, il interroge les quatre étudiants à propos de leur moyenne académique. Les résultats sont présentés ci-dessous.

| Individu | Chambre | | | | |
|-----------------|---------|-------|------|-------|-------|
| | 1 | 2 | 3 | 4 | 5 |
| 1 | 3,08 | 2,36 | 2,00 | 3,00 | 2,68 |
| 2 | 2,60 | 3,04 | 2,56 | 2,88 | 1,92 |
| 3 | 3,44 | 3,28 | 2,52 | 3,44 | 3,28 |
| 4 | 3,04 | 2,68 | 1,88 | 3,64 | 3,20 |
| Total (t_i) | 12,16 | 11,36 | 8,96 | 12,96 | 11,08 |

Ici, les chambres représentent les upe. On a donc $N = 100$, $n = 5$, et $M_i = M = 4$. Une estimation de la moyenne académique dans la population des étudiants vivant dans la résidence est donnée par

$$\hat{\bar{y}}_U = \frac{\hat{t}_y}{K} = \frac{100}{5 \times 400} (12,16 + 11,36 + 8,96 + 12,96 + 11,08) = 2,82.$$

Notant que $s_t^2 = \frac{1}{5-1} [(12,16 - 11,304)^2 + \dots + (11,08 - 11,304)^2] = 2,256$, un estimateur

de la variance de $\hat{\bar{y}}_U$ est donné par $\hat{V}(\hat{\bar{y}}_U) = \frac{100^2}{400^2} \left(1 - \frac{5}{100}\right) \frac{2,256}{5} = 0,02679$.

6.3.2 Comparaison avec l'échantillonnage aléatoire simple sans remise

Dans cette section, on compare l'échantillonnage par grappes à un degré avec l'échantillonnage aléatoire simple sans remise en termes d'efficacité. Par souci de simplicité, on considère le cas $M_i = M$. Comme dans le cas de l'échantillonnage stratifié, on peut construire le tableau 6.1. Il découle de (6.2) que

$$\begin{aligned} V(\hat{t}_y) &= N^2 \left(1 - \frac{n}{N}\right) \frac{1}{n} \frac{1}{N-1} \sum_{i \in U} \left(t_i - \frac{t_y}{N}\right)^2 \\ &= N^2 \left(1 - \frac{n}{N}\right) \frac{1}{n} \frac{1}{N-1} \sum_{i \in U} M^2 (\bar{y}_{iU} - \bar{y}_U)^2 \\ &= N^2 \left(1 - \frac{n}{N}\right) \frac{M(MSB)}{n}. \end{aligned}$$

Par conséquent, la variance de \hat{t}_y est grande si MSB est grande. Il s'ensuit que l'échantillonnage par grappes à un degré est inefficace lorsque la dispersion inter-grappe est grande. Autrement dit, l'échantillonnage par grappes à un degré est inefficace lorsque les upe sont homogènes. En pratique, les éléments à l'intérieur des upe tendent naturellement à être similaires. L'échantillonnage par grappes à un degré est donc fréquemment moins efficace que l'EASSR dans les enquêtes mais – rappelons le – beaucoup plus pratique compte tenu des nombreuses contraintes opérationnelles.

Tableau 6.1 : Analyse de la variance

| Dispersion | Degrés de liberté | Somme de carrés | Moyenne de carrés |
|--------------|-------------------|--|------------------------------|
| Inter grappe | $N - 1$ | $SSB = \sum_{i=1}^N \sum_{j=1}^M (\bar{y}_{iU} - \bar{y}_U)^2$ | $MSB = \frac{SSB}{N - 1}$ |
| Intra-grappe | $N(M - 1)$ | $SSW = \sum_{i=1}^N \sum_{j=1}^M (y_{ij} - \bar{y}_{iU})^2$ | $MSW = \frac{SSW}{N(M - 1)}$ |
| Total | $NM - 1$ | $SST = \sum_{i=1}^N \sum_{j=1}^M (y_{ij} - \bar{y}_U)^2$ | S_y^2 |

Plus formellement, on définit le coefficient de corrélation intra-classe (ICC) afin de mesurer l'homogénéité des unités à l'intérieur des grappes. Celui-ci est défini selon

$$ICC = 1 - \frac{M}{M-1} \frac{SSW}{SST}. \quad (6.3)$$

Puisque $0 \leq \frac{SSW}{SST} \leq 1$, il découle de (6.3)

$$-\frac{1}{M-1} \leq ICC \leq 1.$$

Si les upe sont parfaitement homogènes, on a $SSW = 0$ et $ICC = 1$.

Comparons maintenant la variance de \hat{t}_y obtenue dans le cas de l'EASSR (que l'on désignera par \hat{t}_y^{EASSR}) avec celle de \hat{t}_y obtenue dans le cas de l'échantillonnage par grappes à un degré (que l'on désignera par \hat{t}_y^{Grappe}) avec la même taille d'échantillon nM . On peut montrer que l'effet de plan est donné par

$$\begin{aligned} deff &= \frac{V(\hat{t}_y^{Grappe})}{V(\hat{t}_y^{EASSR})} = \left(1 - \frac{n}{M}\right)^{-1} [1 + (M-1)ICC] \\ &\approx [1 + (M-1)ICC] \end{aligned} \quad (6.4)$$

si la fraction de sondage n/N est négligeable. L'expression (6.4) suggère que lorsque $ICC \approx 1$, on a $deff \approx M$ et l'échantillonnage par grappes s'avère très inefficace par rapport à l'EASSR. Si $ICC \approx 0$, alors on a $V(\hat{t}_y^{Grappe}) \approx V(\hat{t}_y^{EASSR})$. Finalement, lorsque $ICC \leq 0$, on a $V(\hat{t}_y^{Grappe}) \leq V(\hat{t}_y^{EASSR})$ et l'échantillonnage par grappes est plus efficace que l'EASSR.

Remarque : L'échantillonnage systématique considéré au Chapitre 2 est un cas particulier de l'échantillonnage par grappes à un degré avec $n = 1$.

Notons que le coefficient ICC n'est défini que lorsque $M_i = M$. Une mesure alternative de l'homogénéité qui peut être utilisée dans le cas de taille inégales est donnée par

$$R_a^2 = 1 - \frac{MSW}{S_y^2}.$$

Si les upe sont homogènes, alors les moyennes des grappes seront fortement dispersées et le coefficient R_a^2 sera grand.

6.3.3 Estimateur par le ratio

On a vu à la section 6.1 comment estimer le total t_y . Un autre estimateur de t_y du type ratio peut être utilisé en notant que le total de l'upe i , t_i , est généralement corrélé avec la taille M_i . On peut donc utiliser la taille des upe M_i comme variable auxiliaire si l'on suppose que le nombre total d'éléments dans la population, $K = \sum_{i \in U} M_i$, est connu. Un estimateur de type ratio est donc donné par

$$\hat{t}_r = \frac{\sum_{i \in s} t_i}{\sum_{i \in s} M_i} \sum_{i \in U} M_i = K \frac{\sum_{i \in s} t_i}{\sum_{i \in s} M_i} = K \hat{B},$$

où $\hat{B} = \frac{\sum_{i \in s} t_i}{\sum_{i \in s} M_i} \equiv \hat{y}_r$. En utilisant une approximation par séries de Taylor, on obtient un expression similaire à celle vue au Chapitre 5,

$$\hat{V}(\hat{t}_r) = N^2 \left(1 - \frac{n}{N}\right) \frac{1}{n} s_e^2,$$

$$\text{où } s_e^2 = \frac{1}{n-1} \sum_{i \in s} M_i^2 (\bar{y}_i - \hat{y}_r)^2.$$

6.4 Échantillonnage par grappes à deux degrés

L'échantillonnage par grappes à un degré consiste à sélectionner tous les éléments des upe tirées. Cependant, cette stratégie pourrait s'avérer très coûteuse si les tailles M_i des upe sont grandes. En pratique, il est coutume de tirer un sous-échantillon d'éléments dans les upe sélectionnées, auquel cas on parlera de l'échantillonnage par grappes à deux degrés. On procèdera selon les étapes suivantes :

- 1- On tire un échantillon, s , de taille n upe selon un plan aléatoire simple sans remise.
- 2- Dans chaque upe tirée à l'étape 1, on tire un échantillon, s_i , de taille m_i selon un plan aléatoire simple sans remise.

Proposition 6.1 : Supposons que l'on cherche à estimer le total dans la population, $t_y = \sum_{i \in U} t_i$. L'estimateur

$$\hat{t}_y = \frac{N}{n} \sum_{i \in s} \hat{t}_i = \frac{N}{n} \sum_{i \in s} M_i \bar{y}_i, \quad (6.5)$$

est sans biais pour t_y , où $\bar{y}_i = \frac{1}{m_i} \sum_{j \in s_i} y_{ij}$.

Démonstration :

Proposition 6.2 : La variance de l'estimateur \hat{t}_y en (6.5) est donné par

$$V(\hat{t}_y) = N^2 \left(1 - \frac{n}{N} \right) \frac{S_t^2}{n} + \frac{N}{n} \sum_{i \in U} \left(1 - \frac{m_i}{M_i} \right) M_i^2 \frac{S_i^2}{m_i}, \quad (6.6)$$

où $S_t^2 = \frac{1}{N-1} \sum_{i \in U} \left(t_i - \frac{t_y}{N} \right)^2$ et $S_i^2 = \frac{1}{M_i-1} \sum_{j \in U_i} (y_{ij} - \bar{y}_{iU})^2$.

Démonstration :

Remarque : Notons que la variance en (6.6) s'exprime comme la somme de deux composantes : (i) la variabilité inter-grappe (SSW) et (ii) la variabilité intra-grappe (SSB).

Proposition 6.3 : Un estimateur sans biais de la variance de \hat{t}_y en (6.6) est donné par

$$\hat{V}(\hat{t}_y) = N^2 \left(1 - \frac{n}{N} \right) \frac{s_t^2}{n} + \frac{N}{n} \sum_{i \in S} \left(1 - \frac{m_i}{M_i} \right) M_i^2 \frac{s_i^2}{m_i},$$

$$\text{où } s_t^2 = \frac{1}{n-1} \sum_{i \in S} \left(\hat{t}_i - \frac{\hat{t}_y}{N} \right)^2 \text{ et } s_i^2 = \frac{1}{m_i-1} \sum_{i \in S_i} (y_{ij} - \bar{y}_i)^2.$$

Démonstration : laissée en exercice.

Remarque : Si l'on cherche à estimer une moyenne $\bar{y}_U = \frac{t_y}{K}$, on utilisera $\hat{y}_U = \frac{\hat{t}_y}{K}$ (en supposant que K est connu) dont la variance est donnée par $V(\hat{y}_U) = \frac{V(\hat{t}_y)}{K^2}$ et un estimateur de la variance est donné par $\hat{V}(\hat{y}_U) = \frac{\hat{V}(\hat{t}_y)}{K^2}$.

Exemple 6.2 : Une université possède 10 locaux dans lesquels des ordinateurs sont à la disposition des étudiants qui veulent utiliser l'internet. Le responsable de l'informatique à l'université aimerait estimer le nombre moyen d'heures en panne par ordinateur dans les trente derniers jours. Les locaux étant dispersés dans la ville (les départements étant dispersés) il décide d'opter pour un échantillonnage par grappes. D'abord, il sélectionne certains locaux au hasard (en utilisant l'EASSR) et dans chaque local tiré, il décide de sélectionner des ordinateurs au hasard par EASSR. Les résultats suivants ont été obtenus avec $n = 3$ et $m_i = 0.2M_i$. On supposera que $\sum_{i=1}^{10} M_i = 250$. Estimer le nombre moyen d'heures en panne par ordinateur et construire un intervalle de confiance de niveau 95% pour ce paramètre.

| Local | M_i | m_i | Nombre d'heures en panne | \bar{y}_i | s_i^2 |
|-------|-------|-------|--------------------------|-------------|---------|
| 1 | 30 | 6 | 3; 4; 2; 4; 5; 6 | 4 | 2,0 |
| 2 | 25 | 5 | 4; 5; 2; 6; 3 | 4 | 2,5 |
| 3 | 20 | 4 | 5; 6; 4; 5 | 5 | 0,67 |

Une estimation du nombre moyen d'heures en panne par ordinateur est égale à

$$\hat{\bar{y}}_U = \frac{\hat{t}_y}{K} = \frac{N}{Kn} \sum_{i \in S} \hat{t}_i = \frac{N}{Kn} \sum_{i \in S} M_i \bar{y}_i = \frac{10}{3(250)} [30(4) + 25(4) + 20(5)] = \frac{10}{3(250)} (320) = 4,267 \text{ heures.}$$

Maintenant, on a

$$\begin{aligned} \hat{V}(\hat{\bar{y}}_U) &= \frac{N^2}{K^2} \left(1 - \frac{n}{N}\right) \frac{s_t^2}{n} + \frac{N}{nK^2} \sum_{i \in S} \left(1 - \frac{m_i}{M_i}\right) M_i^2 \frac{s_i^2}{m_i} \\ &= \frac{10^2 (0,7)}{3 \times 2 \times 250^2} \left[30^2 (4)^2 + 25^2 (4)^2 + 20^2 (5)^2 - \left(\frac{3 \ 200}{3}\right)^2 \frac{3}{100} \right] \\ &\quad + \frac{10(0,8)}{3 \times 250^2} \left[30^2 \frac{(2,0)}{6} + 25^2 \frac{(2,5)}{5} + 20^2 \frac{(0,67)}{4} \right] \\ &= \frac{4 \ 923,46}{250^2} = 0,078. \end{aligned}$$

Un intervalle de confiance de niveau 95% est donc donné par $4,267 \pm 1,96\sqrt{0,078}$ ou $4,267 \pm 0,561$.

6.5 Réécriture des estimateurs en fonction des poids de sondage

Rappelons que le poids de sondage d'un élément est défini par l'inverse de sa probabilité d'inclusion dans l'échantillon. Dans le cas de l'échantillonnage par grappes à un degré par EASSR, la probabilité d'inclusion de l'élément j appartenant à l'upe i est donnée par

$$\begin{aligned} P(\text{l'élément } j \text{ est tiré}) &= P(\text{l'upe } i \text{ est tirée}) \times P(\text{l'élément } j \text{ est tiré} | \text{l'upe } i \text{ est tirée}) \\ &= P(\text{l'upe } i \text{ est tirée}) = \frac{n}{N}, \end{aligned}$$

puisque $P(\text{l'élément } j \text{ est tiré} | \text{l'upe } i \text{ est tirée}) = 1$. Donc, dans le cas de l'échantillonnage par grappes à un degré par EASSR, le poids de l'élément j appartenant à l'upe i , que l'on désigne par w_{ij} , est donné par $w_{ij} = \frac{N}{n}$. Donc l'échantillonnage par grappe à un degré par EASSR est un exemple de plan de sondage auto-pondéré.

De manière similaire, dans le cas de l'échantillonnage par grappes à deux degrés avec EASSR à chaque degré, on a

$$P(\text{l'élément } j \text{ est tiré}) = P(\text{l'upe } i \text{ est tirée}) \times P(\text{l'élément } j \text{ est tiré} | \text{l'upe } i \text{ est tirée}) = \frac{n}{N} \cdot \frac{m_i}{M_i},$$

puisque $P(\text{l'élément } j \text{ est tiré} | \text{l'upe } i \text{ est tirée}) = \frac{m_i}{M_i}$. Le poids de sondage de l'élément j

appartenant à l'upe i est donc égale à $w_{ij} = \frac{N}{n} \frac{M_i}{m_i}$. Ici, notons que le poids varie d'un

élément à l'autre (lorsque les éléments n'appartiennent pas à la même upe). Donc, en général, l'échantillonnage par grappe à deux degrés ne mène pas à un plan de sondage auto-pondéré. Pour obtenir un plan de sondage auto-pondéré, il faut que le poids de sondage w_{ij} soit constant (i.e., indépendant de i et de j), ce qui surviendra lorsque le ratio

$\frac{m_i}{M_i}$ est constant (par exemple, $\frac{m_i}{M_i} = c$). Dans ce cas, on a $w_{ij} = \frac{N}{n} c$. Remarquons que,

pour que le ratio $\frac{m_i}{M_i}$ soit constant, il faut que $m_i = cM_i$. Autrement dit, le nombre

d'unités que l'on sélectionne dans l'upe i , m_i , doit être proportionnel à la taille de la grappe M_i .

Finalement, dans le cas de l'échantillonnage par grappes à un et deux degrés, les estimateurs (6.1) et (6.5) du total $t_y = \sum_{i \in U} t_i = \sum_{i \in U} \sum_{j \in U_i} y_{ij}$ peuvent s'écrire comme

$$\hat{t}_y = \sum_{i \in s} \sum_{j \in s_i} w_{ij} y_{ij}.$$

6.6 Élaboration d'un plan de sondage par grappes

Lors de l'élaboration d'un plan de sondage par grappes, il est important de considérer les points suivants :

1. Quel niveau de précision requiert-on?
2. Quelle devrait être la taille des upe?
3. Combien d'éléments doit-on tirer dans chaque upe?
4. Combien de upe doit-on tirer?

Dans ce qui suit, nous tentons de répondre (au moins partiellement) aux trois dernières questions.

Réponse à la question 2 : Les upe sont souvent définies de façon naturelle (par exemple, des aires géographiques). Cependant, dans certains cas, il est possible de faire face à un vaste choix quant aux tailles des upe. Dans un tel contexte, il n'est pas du tout trivial de déterminer les tailles des upe. Rappelons quand même que l'on s'attend à ce que l'homogénéité à l'intérieur des upe diminue à mesure que la taille de celles-ci augmente. Les coefficients ICC ou R_a^2 seront donc vraisemblablement plus petits dans une upe de grande taille. Dans ce cas, on s'attend à ce que l'échantillonnage par grappe soit relativement efficace. Par contre, rappelons, que si les upe sont trop grandes, les coûts (de voyage) peuvent augmenter de façon substantielle auquel cas, on aura perdu une des caractéristiques intéressantes de l'échantillonnage par grappes.

Réponse à la question 3 : Afin de déterminer les tailles m_i , considérons l'exemple particulier suivant. Supposons que l'on soit dans le cas de l'échantillonnage par grappes à deux degrés avec EASSR à chacun des degrés et tel que $M_i = M$ et $m_i = m$. Alors, la variance de \hat{y}_U , $V(\hat{y}_U)$, est donnée par

$$V(\hat{y}_U) = \left(1 - \frac{n}{N}\right) \frac{MSB}{nM} + \left(1 - \frac{m}{M}\right) \frac{MSW}{nm}. \quad (6.7)$$

Considérons la fonction de coût suivante :

$$C = c_1 n + c_2 n m, \quad (6.8)$$

où c_1 est le coût au niveau des upe c_2 est le coût encouru afin de mesurer chaque élément. On peut montrer (la démonstration est laissée en exercice) que minimiser la variance de \hat{y}_U en (6.7), étant donné le coût (6.8), mène à

$$n = \frac{c}{c_1 + c_2 m}$$

et

$$m = \sqrt{\frac{c_1 M (MSW)}{c_2 (MSB - MSW)}} = \sqrt{\frac{c_1 M (N-1) \left(\frac{1}{R_a^2} - 1 \right)}{c_2 (NM - 1)}}.$$

Réponse à la question 4 : Une fois que les tailles des upe sont déterminées et que les fractions de sondage au deuxième degré, m_i/M_i , sont fixées, on peut se questionner sur le nombre d'upe, n , à tirer. Encore une fois supposons que $M_i = M$ et $m_i = m$, auquel cas la variance de \hat{y}_U est donnée en (6.7). On a

$$V(\hat{y}_U) = \left(1 - \frac{n}{N}\right) \frac{MSB}{nM} + \left(1 - \frac{m}{M}\right) \frac{MSW}{nm} \leq \frac{1}{n} \left[\frac{MSB}{M} + \left(1 - \frac{m}{M}\right) \frac{MSW}{m} \right] = \frac{1}{n} \cdot c,$$

où $c = \frac{MSB}{M} + \left(1 - \frac{m}{M}\right) \frac{MSW}{m}$, si l'on ignore le facteur $\left(1 - \frac{n}{N}\right)$. La précision désirée peut s'exprimer comme suit :

$$P\left(|\hat{y}_U - \bar{y}_U| \leq e\right) = 1 - \alpha, \quad (6.9)$$

où e désigne la marge d'erreur. Il découle de (6.9) que

$$P\left(\frac{-e}{\sqrt{V(\hat{y}_U)}} \leq \frac{\hat{y}_U - \bar{y}_U}{\sqrt{V(\hat{y}_U)}} \leq \frac{e}{\sqrt{V(\hat{y}_U)}}\right) = 1 - \alpha,$$

et alors,

$$\frac{e}{\sqrt{V(\hat{y}_U)}} = z_{\alpha/2},$$

ce qui entraîne que

$$n = \frac{z_{\alpha/2}^2 c}{e^2}.$$

Bien sûr, il faut avoir une idée préalable de la valeur de c pour être en mesure de déterminer la valeur de n qui nous garantira la précision recherchée. On pourra, par exemple, utiliser l'information provenant d'enquêtes antérieures.

Échantillonnage par grappes à probabilités inégales

Chapitre 7

7.1 Introduction

Dans le chapitre 6, nous avons étudié l'échantillonnage par grappes à un et deux degrés avec sélection des grappes et des éléments selon un plan aléatoire simple sans remise. Afin d'améliorer la qualité des estimateurs, il est parfois judicieux de tirer les grappes et/ou les éléments avec probabilités inégales. Par exemple, supposons que l'on cherche à estimer le nombre total d'employés dans les entreprises de construction dans une ville donnée. La ville compte un grand nombre de petites entreprises de construction et un petit nombre de grandes entreprises. Imaginons que les petites entreprises emploient un petit nombre d'employés alors que les grandes entreprises de construction emploient un grand nombre d'employés. La contribution des grandes entreprises au nombre d'employés est donc très important. Par exemple, on peut imaginer, que bien qu'elles soient peu nombreuses, la contribution des grandes entreprises est environ égale à 90%. Si l'on utilise l'EASSR pour tirer les entreprises, la taille de celles-ci n'est pas prise en compte, auquel cas on s'attend, la plupart du temps, à obtenir un échantillon composé d'un grand nombre de petites entreprises. Dans ce cas, l'estimation du nombre total d'employés risque d'être bien en deçà de la vraie valeur. Par contre, si l'échantillon contient toutes les grandes entreprises, l'estimation du nombre total d'employés risque d'être assez proche de la vraie valeur puisque le nombre d'employés est fortement influencé par les grandes entreprises. Dépendamment de la composition de l'échantillon, l'estimation du nombre d'employés risque d'être volatile et la variance de l'estimateur grande. Imaginons le plan de sondage alternatif suivant : on tire un échantillon de telle manière à ce que les grandes entreprises ont une grande probabilité (éventuellement égale à 1) d'être tirées dans l'échantillon. Dans ce cas, on s'attend à ce que, quel que soit l'échantillon tiré (qui contiendra toujours les grandes entreprises), l'estimation du nombre d'employés varie très peu. Autrement dit, l'estimateur aura une petite variance. Une méthode permettant de tirer un tel échantillon s'appelle *l'échantillonnage proportionnel à la taille* (échantillonnage PPT) et fera l'objet de ce chapitre.

7.2 Échantillonnage d'une seule unité

Nous supposons qu'une *mesure de taille* x est disponible pour toutes les unités dans la population et soit $t_x = \sum_{i \in U} x_i$ le total de la variable de taille dans la population. Soit ψ_i la probabilité que l'unité i est sélectionnée au premier tirage. Supposons que l'on veuille tirer un échantillon de taille $n = 1$ grappe (ou élément). On veut tirer la grappe selon un plan de sondage avec probabilités proportionnelles à la taille; i.e., on veut sélectionner la grappe i telle que $\psi_i \propto x_i$. Afin d'illustrer ce type d'échantillonnage, considérons l'exemple suivant :

Une petite ville possède 4 supermarchés dont la superficie varie de 100 m² à 1000 m². On veut estimer le montant total des ventes dans les quatre supermarchés durant le mois dernier en choisissant un seul supermarché au hasard. On s'attend à ce que le montant des ventes pour un supermarché soit proportionnel à la superficie du supermarché. Autrement dit, on s'attend à ce qu'un petit supermarché fasse moins de ventes qu'un grand supermarché. On décide donc de tirer un supermarché au hasard avec probabilité proportionnelle à sa superficie. Supposons (pour les besoins de la cause) que le total des ventes du magasin i , t_i , est connu t_i (ce qui ne sera évidemment pas le cas en pratique!). On a les données suivantes :

| Supermarché | Taille (m ²) | ψ_i | t_i (en millier de dollars) |
|-------------|-----------------------------|----------|----------------------------------|
| A | 100 | 1/16 | 11 |
| B | 200 | 2/16 | 20 |
| C | 300 | 3/16 | 24 |
| D | 1 000 | 10/16 | 245 |
| | 1 600 | 1 | 300 |

Le but est de tirer un supermarché avec probabilité ψ_i et d'estimer le total des ventes,

$t_y = \sum_{i=1}^4 t_i$. Comment estimer t_y ? Nous accepterons pour l'instant d'utiliser l'estimateur

$$\hat{t}_\psi = \sum_{i \in S} \frac{t_i}{\psi_i}. \quad (7.1)$$

Dans le cas d'un échantillon de taille $n = 1$, l'estimateur (7.1) est égal à $\hat{t}_\psi = \frac{t_i}{\psi_i}$. Afin de nous convaincre que l'estimateur (7.1) est raisonnable, on montre maintenant qu'il est sans biais pour t_y . Pour cela, on détermine la distribution de \hat{t}_ψ :

| Échantillon | {A} | {B} | {C} | {D} |
|-------------|------|------|------|-------|
| $p(s)$ | 1/16 | 2/16 | 3/16 | 10/16 |

Rappelons (voir Chapitre 2) que

$$E(\hat{t}_\psi) = \sum_{s \in \Omega} \hat{t}_\psi p(s) = \frac{1}{16}(176) + \frac{2}{16}(160) + \frac{3}{16}(128) + \frac{10}{16}(392) = 300 = t_y.$$

Donc, l'estimateur \hat{t}_ψ est sans biais pour le total t_y . De plus, la variance de \hat{t}_ψ est donnée par

$$\begin{aligned}
V(\hat{t}_\psi) &= E\left[(\hat{t}_\psi - t_y)^2\right] = \sum_{s \in \Omega} (\hat{t}_\psi - t_y)^2 p(s) \\
&= \frac{1}{16}(15\,376) + \frac{2}{16}(19\,600) + \frac{3}{16}(29\,584) + \frac{10}{16}(8\,464) = 14\,248.
\end{aligned}$$

Au lieu de sélectionner l'échantillon selon un plan PPT, on décide de tirer un supermarché selon l'EASSR et d'utiliser l'estimateur usuel $\hat{t}_y = N\bar{y}_s$ (qui est un cas particulier de (7.1) avec $\psi_i = 1/N$). La distribution de \hat{t}_y est donnée par

| Échantillon | {A} | {B} | {C} | {D} |
|-------------|-----|-----|-----|-----|
| $p(s)$ | 1/4 | 1/4 | 1/4 | 1/4 |

On a alors, $E(\hat{t}_\psi) = \frac{1}{4}[44 + 80 + 96 + 980] = 300$ et \hat{t}_y est sans biais pour t_y , ce qui n'est pas surprenant compte tenu des résultats obtenus au Chapitre 2. La variance de \hat{t}_y est donnée par

$$V(\hat{t}_\psi) = \frac{1}{4}[65\,536 + 48\,400 + 41\,616 + 462\,400] = 154\,488.$$

On remarquera alors que la variance obtenue dans le cas de l'échantillonnage PPT (14 248) est plus de dix fois plus petite que celle obtenue dans le cas de l'EASSR (154 488).

7.3 Échantillonnage à un degré avec remise

Supposons que l'on veuille tirer un échantillon aléatoire selon un plan PPT avec remise de taille $n > 1$ grappes. Soit $\psi_i = P$ (sélectionner la grappe i au premier tirage). Puisque l'échantillonnage est effectué avec remise, ψ_i représente également la probabilité de sélectionner la grappe i à n'importe quel tirage (puisque les probabilités de sélection ne changent pas après que nous avons sélectionné la première grappe).

La probabilité d'inclusion π_i de la grappe i est la probabilité que la grappe i soit tirée au moins une fois dans l'échantillon. Il n'est pas difficile de montrer que

$$\pi_i = 1 - (1 - \psi_i)^n$$

La démonstration de ce résultat est similaire à celle obtenue dans le cas de l'échantillonnage aléatoire simple avec remise (EASAR). Notons que l'EASAR est un cas particulier de l'échantillonnage PPT avec remise avec $\psi_i = \frac{1}{N}$. De plus, si $n = 1$, on a $\psi_i = \pi_i$ alors que lorsque $n > 1$, on a $\psi_i \neq \pi_i$, en général.

7.3.1 Procédure de sélection

Comment tirer un échantillon selon un plan PPT avec remise de taille n ? Une méthode simple est appelée *méthode des tailles cumulées* (en anglais, *cumulative size method*). Pour appliquer cette méthode, on construit d'abord le tableau suivant :

| Unité | ψ_i | Probabilité cumulée | Intervalle |
|-------|--------------|--|---|
| 1 | ψ_1 | ψ_1 | $(0; \psi_1]$ |
| 2 | ψ_2 | $\psi_1 + \psi_2$ | $(\psi_1; \psi_1 + \psi_2]$ |
| 3 | ψ_3 | $\psi_1 + \psi_2 + \psi_3$ | $(\psi_1 + \psi_2; \psi_1 + \psi_2 + \psi_3]$ |
| ⋮ | ⋮ | ⋮ | ⋮ |
| ⋮ | ⋮ | ⋮ | ⋮ |
| ⋮ | ⋮ | ⋮ | ⋮ |
| $N-1$ | ψ_{N-1} | $\psi_1 + \psi_2 + \psi_3 + \dots + \psi_{N-1}$ | $(\psi_1 + \psi_2 + \dots + \psi_{N-2}; \psi_1 + \psi_2 + \psi_3 + \dots + \psi_{N-1}]$ |
| N | ψ_N | $\psi_1 + \psi_2 + \psi_3 + \dots + \psi_{N-1} + \psi_N = 1$ | $(\psi_1 + \dots + \psi_{N-1}; 1]$ |

Ensuite, on génère un nombre aléatoire u à partir d'une distribution $U(0,1)$. La grappe que l'on choisira est celle dont l'intervalle (voir tableau) contiendra le nombre généré u . Nous avons ainsi sélectionné une grappe. Pour en sélectionner n , on répétera cette procédure n fois.

Exemple 7.1 : Une école est composée de 8 classes de taille 30, 20, 25, 45, 35, 40, 50 et 55. Supposons que l'on veuille tirer un échantillon aléatoire de 3 classes avec remise et probabilité proportionnelle à la taille de la classe. On construit le tableau ci-dessous. Ensuite, on génère 3 nombres aléatoires à partir d'une distribution $U(0,1)$: 0,1881; 0,6452; 0,0804. Les classes sélectionnées sont donc les classes 3, 6 et 1.

| Classe | Taille x_i | $\psi_i = \frac{x_i}{t_x}$ | Probabilité cumulée | Intervalle |
|--------|--------------|----------------------------|---------------------|------------------|
| 1 | 30 | 0,1 | 0,1 | $(0; 0,1]$ |
| 2 | 20 | 0,0666 | 0,1666 | $(0,1; 0,1666]$ |
| 3 | 25 | 0,0833 | 0,25 | $(0,1666; 0,25]$ |
| 4 | 45 | 0,15 | 0,40 | $(0,25; 0,40]$ |
| 5 | 35 | 0,1166 | 0,5166 | $(0,40; 0,5166]$ |
| 6 | 40 | 0,1333 | 0,65 | $(0,5166; 0,65]$ |
| 7 | 50 | 0,1666 | 0,8166 | $(0,65; 0,8166]$ |
| 8 | 55 | 0,1833 | 1 | $(0,8166; 1]$ |
| 300 | | | | |

7.3.2 Estimation

Supposons que l'on veuille estimer le total dans la population $t_y = \sum_{i \in U} t_i$. Quel estimateur utiliser dans le cas de l'échantillonnage PPT avec remise? Un estimateur possible est donné par

$$\hat{t}_\psi = \frac{1}{n} \sum_{i \in U} Q_i \frac{t_i}{\psi_i}, \quad (7.2)$$

où Q_i désigne le nombre de fois que l'unité i est tirée dans l'échantillon. On a $\mathbf{Q} = (Q_1, \dots, Q_N) \sim MN(n, \psi_1, \dots, \psi_N)$. Il s'ensuit que $E(Q_i) = n\psi_i$ et $V(Q_i) = n\psi_i(1 - \psi_i)$. Il découle de ces résultats que \hat{t}_ψ est sans biais pour t_y (i.e., $E(\hat{t}_\psi) = t_y$) et que

$$V(\hat{t}_\psi) = \frac{1}{n} \sum_{i \in U} \psi_i \left(\frac{t_i}{\psi_i} - t_y \right)^2. \quad (7.3)$$

Les démonstrations des expressions (7.2) et (7.3) sont laissées en exercice.

Proposition 7.1 : Un estimateur sans biais de la variance de \hat{t}_ψ , $V(\hat{t}_\psi)$, en (7.3) est donné par

$$\hat{V}(\hat{t}_\psi) = \frac{1}{n(n-1)} \sum_{i \in U} Q_i \left(\frac{t_i}{\psi_i} - \hat{t}_\psi \right)^2.$$

Démonstration :

Remarque : Lorsque $t_i = c\psi_i$, on a $\hat{t}_\psi = \frac{1}{n} \sum_{i \in U} Q_i \frac{c\psi_i}{\psi_i} = \frac{c}{n} \sum_{i \in U} Q_i = c$ puisque $\sum_{i \in U} Q_i = n$.

Dans ce cas, on a $V(\hat{t}_\psi) = 0$. L'échantillonnage PPT avec remise est donc très efficace lorsque les totaux des grappes t_i sont fortement corrélés avec les probabilités de sélection ψ_i .

Dans un monde idéal, on aurait donc $\psi_i = \frac{t_i}{t_y}$. Le problème avec cette stratégie est

que l'on ne connaît pas les totaux t_i avant de tirer l'échantillon (sinon, on n'aurait pas besoin d'enquête !) et il ne sera donc pas possible de définir les probabilités ψ_i de cette manière. En revanche, si une variable auxiliaire x fortement corrélée avec y est

disponible pour toutes les unités dans la population, on pourra l'utiliser et poser $\psi_i = \frac{x_i}{t_x}$.

Par exemple, on sait que la taille M_i des upe est généralement corrélée au total t_i . On

pourrait donc utiliser $\psi_i = \frac{M_i}{K}$, où $K = \sum_{i \in U} M_i$.

7.4 Échantillonnage à un degré sans remise

L'échantillonnage PPT avec remise est généralement moins efficace que l'échantillonnage PPT sans remise. Cependant, la théorie dans le cas de l'échantillonnage PPT sans remise est généralement beaucoup plus complexe, et ce, même dans des situations relativement simples. Supposons qu'une variable de taille x est disponible pour toutes les unités de la population avant de tirer l'échantillon. On cherche à sélectionner les unités de manière à satisfaire

$$\pi_i = n \frac{x_i}{t_x} = n\psi_i \text{ pour tout } i. \quad (7.4)$$

Autrement dit, on veut que la probabilité d'inclusion de l'unité i dans l'échantillon soit proportionnelle à sa taille. Comme on le verra dans l'exemple suivant, déterminer une procédure d'échantillonnage sans remise qui satisfait (7.4) n'est pas triviale.

7.4.1 Procédure de sélection

Exemple 7.2 : Revenons à l'exemple des supermarchés considéré à la section 7.2. Supposons que l'on veuille tirer un échantillon aléatoire sans remise de taille $n = 2$ supermarchés selon un plan PPT. On veut donc que la procédure satisfasse $\pi_i = 2\psi_i$. Supposons qu'au premier tirage, le supermarché A est sélectionné. On a

$$\psi_A = P(\text{supermarché A est sélectionné au premier tirage}) = \frac{1}{16}.$$

Une fois que A est tiré, on doit sélectionner un autre supermarché. Supposons que l'on tire le supermarché B au deuxième tirage. On a

$$\begin{aligned} P(\text{supermarché B est sélectionné au deuxième tirage} \mid \text{supermarché A est sélectionné au premier tirage}) \\ &= \frac{2/16}{1-1/16} \\ &= \frac{2}{15} \\ &= \frac{\psi_B}{1-\psi_A}. \end{aligned}$$

En général, on a

$$P(\text{unité } j \text{ est sélectionnée au deuxième tirage} \mid \text{unité } i \text{ est sélectionnée au premier tirage}) = \frac{\psi_j}{1-\psi_i}.$$

Par conséquent, on a

$$\begin{aligned} &P(\text{unité } i \text{ est sélectionnée au premier tirage, unité } j \text{ est sélectionnée au deuxième tirage}) = \\ &P(\text{unité } i \text{ est sélectionnée au premier tirage}) \\ &\quad \times P(\text{unité } j \text{ est sélectionnée au deuxième tirage} \mid \text{unité } i \text{ est sélectionnée au premier tirage}) \\ &= \psi_i \frac{\psi_j}{1-\psi_i}. \end{aligned}$$

De manière similaire, on a

$$P(\text{unité } j \text{ est sélectionnée au premier tirage, unité } i \text{ est sélectionnée au deuxième tirage}) = \psi_j \frac{\psi_i}{1-\psi_j}.$$

Il s'ensuit que la probabilité d'inclusion jointe des unités i et j dans l'échantillon est donnée par

$$\begin{aligned}
\pi_{ij} &= P(\text{unité } i \text{ et unité } j \text{ tirées dans l'échantillon}) \\
&= P(\text{unité } j \text{ est sélectionnée au premier tirage, unité } i \text{ est sélectionnée au deuxième tirage}) \\
&\quad + P(\text{unité } i \text{ est sélectionnée au premier tirage, unité } j \text{ est sélectionnée au deuxième tirage}) \\
&= \psi_i \frac{\psi_j}{1-\psi_i} + \psi_j \frac{\psi_i}{1-\psi_j}.
\end{aligned}$$

Malheureusement, cette procédure d'échantillonnage n'est pas adéquate car elle ne mène pas à $\pi_i = 2\psi_i$. En effet, on a

$$\begin{aligned}
\pi_i &= P(i \in s) = P(i \text{ est sélectionnée d'abord et n'importe quelle unité est sélectionnée au deuxième tirage}) \\
&\quad + P(\text{n'importe quelle unité est sélectionnée au premier tirage et } i \text{ est sélectionnée au deuxième tirage}) \\
&= \psi_i + \sum_{j \neq i} \psi_j \frac{\psi_i}{1-\psi_j} \neq 2\psi_i,
\end{aligned}$$

en général. Plusieurs procédures de tirage PPT sans remise ont été proposées dans la littérature satisfaisant $\pi_i = n\psi_i$. Ces procédures sont généralement complexes et ne seront pas considérées ici. Dans ce qui suit, on supposera qu'une telle procédure a été utilisée pour sélectionner l'échantillon.

7.4.2 Estimation dans le cadre de l'échantillonnage par grappes à un degré

Dans cette section, on considère le cas de l'échantillonnage par grappes à un degré. Supposons que l'on tire n upe selon un plan de sondage à probabilités inégales (par exemple, l'échantillonnage PPT sans remise). On cherche à estimer le total $t_y = \sum_{i \in U} t_i$. Un estimateur sans biais de t_y est l'estimateur de type Horvitz-Thompson, donné par

$$\hat{t}_{HT} = \sum_{i \in s} \frac{t_i}{\pi_i}. \quad (7.5)$$

Remarque : Lorsque $t_i = c\pi_i$, on a $\hat{t}_{HT} = c$, auquel cas, $V(\hat{t}_{HT}) = 0$. Par conséquent, définir la probabilité d'inclusion π_i proportionnelle à t_i mènera à un estimateur ayant une petite variance. Cependant, cette stratégie ne peut être utilisée en pratique puisque les totaux t_i sont inconnus avant de tirer l'échantillon. Supposons qu'une variable auxiliaire x est disponible pour toutes les unités de la population et que la variable x_i est approximativement proportionnelle à t_i . Choisir π_i proportionnelle à x_i fera en sorte que

les ratios t_i/π_i seront approximativement constants. Il s'ensuit que la variance de l'estimateur (7.5) sera vraisemblablement petite.

Proposition 7.2 : L'estimateur (7.5) est sans biais pour t_y .

Démonstration : laissée en exercice.

Proposition 7.3 : La variance de l'estimateur (7.5) est donnée par

$$V(\hat{t}_{HT}) = \sum_{i \in U} \sum_{j \in U} (\pi_{ij} - \pi_i \pi_j) \frac{t_i}{\pi_i} \frac{t_j}{\pi_j}. \quad (7.6)$$

Démonstration : laissée en exercice.

Proposition 7.4 : Un estimateur sans biais de $V(\hat{t}_{HT})$ en (7.6) est donné par

$$\hat{V}(\hat{t}_{HT}) = \sum_{i \in s} \sum_{j \in s} \frac{(\pi_{ij} - \pi_i \pi_j)}{\pi_{ij}} \frac{t_i}{\pi_i} \frac{t_j}{\pi_j}. \quad (7.7)$$

Démonstration : laissée en exercice.

7.4.3 Estimation dans le cadre de l'échantillonnage par grappes à deux degrés

Dans cette section, on considère le cas de l'échantillonnage par grappes à deux degrés. Au premier degré, on tire un échantillon aléatoire de upe, s , de taille n à l'aide d'un plan de sondage à probabilités inégales (par exemple, l'échantillonnage PPT sans remise). Au deuxième degré, on tire des sous-échantillons aléatoires, s_i , de taille m_i , dans les upe sélectionnées au premier degré à l'aide d'un plan de sondage à probabilités inégales (par exemple, l'échantillonnage PPT sans remise). Supposons que l'on cherche à estimer le total dans la population, $t_y = \sum_{i \in U} t_i$. On utilisera l'estimateur de type Horvitz-Thompson

$$\hat{t}_{HT} = \sum_{i \in s} \frac{\hat{t}_i}{\pi_i}, \quad (7.8)$$

où $\hat{t}_i = \sum_{j \in s_i} \frac{y_{ij}}{\pi_{j|i}}$ et

$\pi_{j|i} = P(\text{élément } j \text{ est sélectionné dans } s_i | \text{upe } i \text{ est tirée au premier degré})$ désigne la probabilité conditionnelle de l'élément j étant donné que l'upe i a été tirée au premier degré.

Proposition 7.5 : L'estimateur (7.8) est sans biais pour t_y .

Démonstration :

Proposition 7.6 : La variance de l'estimateur (7.8) est donnée par

$$V(\hat{t}_{HT}) = \sum_{i \in U} \sum_{k \in U} (\pi_{ik} - \pi_i \pi_k) \frac{t_i}{\pi_i} \frac{t_k}{\pi_k} + \sum_{i \in U} \frac{V_i}{\pi_i}, \quad (7.9)$$

où $V_i = V_2(\hat{t}_i | s) = \sum_{j \in U_i} \sum_{j' \in U_i} (\pi_{jj'|i} - \pi_{j|i} \pi_{j'|i}) \frac{y_{ij}}{\pi_{j|i}} \frac{y_{ij'}}{\pi_{j'|i}}$.

Démonstration :

Proposition 7.7 : Un estimateur sans biais de la variance (7.9) est donné par

$$\hat{V}(\hat{t}_{HT}) = \sum_{i \in S} \sum_{k \in S} \frac{(\pi_{ik} - \pi_i \pi_k)}{\pi_{ik}} \frac{\hat{t}_i}{\pi_i} \frac{\hat{t}_k}{\pi_k} + \sum_{i \in S} \frac{\hat{V}_i}{\pi_i},$$

$$\text{où } \hat{V}_i = \sum_{j \in S_i} \sum_{j' \in S_i} \frac{(\pi_{jj'|i} - \pi_{j|i} \pi_{j'|i})}{\pi_{jj'|i}} \frac{y_{ij}}{\pi_{j|i}} \frac{y_{ij'}}{\pi_{j'|i}}.$$

Démonstration : Omise.

Annexe A

Dans cette annexe, on rappelle des notions de base de probabilités qui sont utiles aux chapitres 1 à 7.

A.1 Probabilités

Soit Ω l'ensemble de tous les résultats possibles d'une expérience aléatoire. Un événement A est un sous-ensemble de Ω .

Soit A et B deux événements.

- (a) A et B sont dits mutuellement exclusifs si $A \cap B = \emptyset$. Dans ce cas, $P(A \cap B) = 0$.
- (b) A et B sont dits indépendants si et seulement si $P(A \cap B) = P(A)P(B)$. Cette définition est équivalente à chacune des deux suivantes : $P(A|B) = P(A)$ pourvu que $P(B) \neq 0$ et $P(B|A) = P(B)$ pourvu que $P(A) \neq 0$.

A.2 Espérance d'une variable aléatoire

Soit X une variable aléatoire discrète. L'espérance de X , $E(X)$, est définie par

$$E(X) = \sum_x xp(x).$$

La variance de X , $V(X)$, est définie par

$$V(X) = E[X - E(X)]^2.$$

Soit X_1, X_2, \dots, X_n une suite de n variables aléatoires discrètes. Soit $Y = \sum_{i=1}^n a_i X_i$, où a_1, a_2, \dots, a_n sont des constantes données. L'espérance de Y , $E(Y)$, est donnée par

$$E(Y) = \sum_{i=1}^n a_i E(X_i).$$

Si les variables aléatoires X_1, X_2, \dots, X_n , sont indépendantes, alors la variance de Y , $V(Y)$, est donnée par

$$V(Y) = \sum_{i=1}^n a_i^2 V(X_i).$$

Par contre, si les variables aléatoires X_1, X_2, \dots, X_n , ne sont pas indépendantes, il faut tenir compte des covariances. Dans ce cas, la variance de Y , $V(Y)$, est donnée par

$$V(Y) = \sum_{i=1}^n a_i^2 V(X_i) + \sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n a_i a_j \text{Cov}(X_i, X_j),$$

où $\text{Cov}(X_i, X_j) = E(X_i X_j) - E(X_i)E(X_j)$ désigne la covariance entre X_i et X_j .

A.3 Espérance et variance conditionnelles

Soit X et Y deux variables aléatoires discrètes. L'espérance conditionnelle de Y étant donné X , $E(Y|X)$, est définie par

$$E(Y|X) = \sum_y y p(Y = y | X = x).$$

La variance de conditionnelle de Y étant donné X , $V(Y|X)$, est définie par

$$V(Y|X) = E(Y^2|X) - [E(Y|X)]^2.$$

On peut toujours exprimer l'espérance de Y comme :

$$E(Y) = E[E(Y|X)].$$

La variance de Y peut être exprimée comme

$$V(Y) = V[E(Y|X)] + E[V(Y|X)].$$

A.4 Loi binomiale

Considérons n essais indépendants d'une expérience aléatoire. Chaque essai donne lieu à deux résultats possibles : succès avec probabilité p et échec avec probabilité $1-p$. Soit X le nombre de succès en n essais. Alors, X suit une loi binomiale de paramètres n et p et on écrit : $X \sim B(n, p)$. On a

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0, 1, \dots, n.$$

De plus, $E(X) = np$ et $V(X) = np(1-p)$.

A.5 Loi multinomiale

Considérons une expérience aléatoire avec k résultats possibles, r_1, \dots, r_k . La probabilité du résultat r_j sera dénotée p_j , $j = 1, \dots, k$. On a donc $0 \leq p_j \leq 1$ et $\sum_{j=1}^k p_j = 1$. Considérons n répétitions indépendantes de cette expérience aléatoire. Soit Q_j le nombre de fois que l'on observe r_j et soit $\mathbf{Q} = (Q_1, \dots, Q_k)$. Alors, le vecteur \mathbf{Q} suit une loi multinomiale de paramètres n, p_1, \dots, p_k et on écrit $\mathbf{Q} \square MN(n, p_1, \dots, p_k)$. On a

$$P(Q_1 = n_1, \dots, Q_k = n_k) = \frac{n!}{n_1! n_2! \dots n_k!} p_1^{n_1} p_2^{n_2} \dots p_k^{n_k}, \quad \sum_{j=1}^k n_j = n.$$

Il est intéressant de noter que $Q_j \square B(n, p_j)$, $j = 1, \dots, k$, et $Cov(Q_i, Q_j) = -np_i p_j$, $i \neq j = 1, \dots, k$.