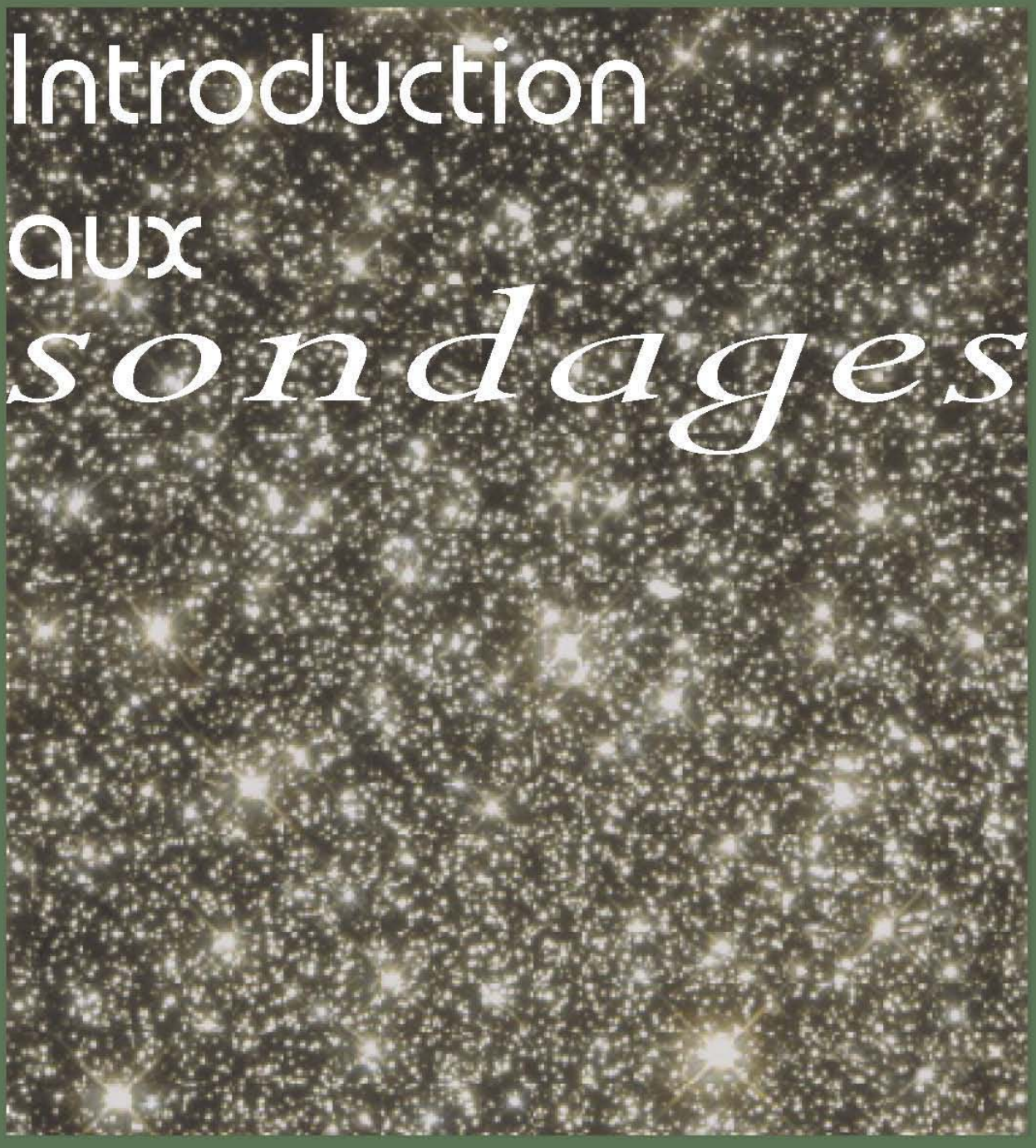
Serge Alalouf

Introduction aux *sondages*



Serge Alalouf

Introduction aux *sondages*

Alalouf, Serge

Introduction aux sondages

ISBN 978-2-923565-97-2

1. Échantillonnage (Statistique). I. Titre.

Copyright © 2015 Loze-Dion éditeur inc.
Courriel: commandes@lozedion.com
Web: www.lozedion.com

Tous droits réservés. On ne peut reproduire, enregistrer, ni diffuser aucune partie du présent ouvrage sous quelque forme ou par quelque procédé que ce soit, sans une autorisation écrite de l'éditeur.

Dépôt légal quatrième trimestre 2015
Bibliothèque et Archives nationales du Québec, 2015
Bibliothèque et Archives Canada, 2015

Imprimé au Canada

PRÉFACE

Une lente évolution amorcée dans la quatrième décennie du siècle dernier s'est accélérée au cours des vingt dernières années: vers le milieu des années cinquante, l'échantillonnage probabiliste a commencé à remplacer l'énumération complète. Il aura fallu, pour cela, rompre avec une longue tradition qui voulait que seul un recensement complet soit crédible. On ne doute plus, aujourd'hui, de la possibilité d'obtenir une information valable sur un grand ensemble à partir d'une partie de l'ensemble. Le développement et la diffusion de la théorie des sondages est sans doute le principal déclencheur de ce profond changement de mentalité.

À son tour, la pratique d'échantillonnage a suscité une activité fébrile de recherche théorique, en réponse à des besoins concrets de plus en plus variés et complexes. Le résultat, quelques décennies plus tard, est un corpus de connaissances extrêmement riche. Il va sans dire qu'un livre d'introduction aux sondages, qui ne peut tout couvrir, ou même effleurer, se doit avant tout de délimiter clairement son objet. Il se trouve que l'objectif de ce livre rend la chose relativement aisée.

OBJECTIF DU LIVRE

Mon intention est de présenter les outils nécessaires pour effectuer des sondages dans un contexte particulier: les sondages internes comme il s'en fait régulièrement dans les entreprises, les firmes conseil, les cabinets de comptables, et les agences gouvernementales.

Il m'est apparu évident, au cours de mes activités de consultation, qu'il existe un réel besoin de compétences en sondages dans ces milieux. En vérification comptable, par exemple, où l'échantillonnage fait depuis longtemps partie de la pratique, on observe une tendance à remplacer les échantillons dits «de jugement» par des échantillons probabilistes effectués dans les règles de l'art. En particulier, dans une situation de litige, la nécessité de justifier une conclusion de façon objective et d'évaluer les risques d'erreur impose un échantillonnage probabiliste.

La théorie classique des sondages repose sur un cadre clair: une population bien définie; et un mode de tirage qui permet de calculer la probabilité qu'une unité donnée soit incluse dans l'échantillon. Or ces conditions se trouvent réunies dans le contexte d'un échantillonnage interne: une population accessible (généralement sous la forme d'un fichier électronique) et bien identifiée; une base de sondage souvent riche et féconde; un mécanisme de sélection aisément applicable. Les grandes difficultés pour lesquelles les techniques les plus avancées ont été créées ne se présentent pas, ou rarement. Le paradigme de base est donc applicable et c'est à cela que ce livre s'en tient.

Cette limitation, face à l'énorme masse de connaissances qui existent en sondages, peut sembler radicale, mais pour le lecteur à qui s'adresse se livre, elle n'est presque pas contraignante.

À QUI S'ADRESSE CE LIVRE?

Ce livre est destiné à deux clientèles distinctes: les analystes qui sont ou seront appelés à effectuer des sondages; et les étudiants d'un premier cycle universitaire, aussi bien en mathématiques, actuariat ou statistique, qu'en sciences de la gestion. Il est possible d'accommoder ces diverses classes de lecteurs dans la mesure où chaque chapitre traite la théorie et les applications séparément. Chaque chapitre commence par décrire, discuter et illustrer les procédures avec un minimum de mathématiques, et passe ensuite à la théorie mathématique qui les justifie. Ceux qui s'intéressent essentiellement aux applications n'auront qu'à omettre la deuxième partie. Je suis convaincu que les étudiants des disciplines mathématiques, eux aussi, bénéficieront de cette approche. Les applications motivent et concrétisent la théorie.

LES CALCULS

Dans la mesure où ce livre se propose d'aider non seulement à comprendre des concepts mais également à développer une compétence dans ses applications, il est impossible d'éviter les calculs qu'impose l'analyse des résultats d'un sondage. Les données sont parfois imposantes et les calculs ardus. Pour certains exercices, des calculs préliminaires sont présentés afin d'alléger le labeur, mais ce qui reste à faire n'est pas toujours une mince besogne. Je tiens pour acquis que les lecteurs disposeront de leur propre logiciel de calcul, au-delà d'une simple calculatrice, et je n'en propose pas un en particulier. Mais certaines procédures, entre autres celles qui impliquent un algorithme itératif, ne peuvent être décrites simplement que dans le langage d'un logiciel possédant les outils nécessaires. J'ai choisi Excel pour cela, ce logiciel étant le plus universellement connu.

LES PRÉREQUIS

La lecture de ce livre exige comme préalable au moins un cours de probabilités et/ou de statistique générale. Ce qui ne veut pas dire qu'un tel cours met au même niveau tous les lecteurs ou tous les élèves d'une classe: le cours préalable peut avoir été théorique ou appliqué; axé sur les probabilités ou sur la statistique; rigoureusement mathématique ou vulgarisé. C'est pour cela que j'ai cru bon, dans un premier chapitre, assez long, de présenter toute information essentielle qui risquerait de manquer au lecteur. Y compris le lecteur équipé d'un cours de probabilités costaud, car ce chapitre traite de sujets qui ne sont pas toujours inclus dans un premier cours. Les exercices du chapitre aussi ont leur importance. Ils soulèvent des questions qui anticipent les problèmes de sondages qui seront traités par la suite. Présentés de façon épurée comme simples problèmes de probabilités, ils ont l'avantage de concentrer la réflexion sur le seul aspect probabiliste, sans l'encombrement d'un contexte plus chargé.

Ce contenu aurait pu paraître en annexe, tel un recueil d'informations à consulter seulement au besoin. Mais mon expérience me fait croire que ce n'est pas la meilleure façon de procéder. La lecture systématique de ce chapitre me semble être un passage obligé. Si certains le lisent rapidement et n'y trouvent rien de nouveau, tant mieux. Mais pour d'autres, probablement la majorité, l'investissement d'un temps — long ou court — à une

lecture sereine de ces notions préalables devrait se révéler rentable en temps et frustrations épargnées.

LES EXEMPLES

Certains des échantillons qui servent d'exemples sont tirés de données diffusées dans des documents officiels (Statistique Canada, ONU). Ils servent à illustrer les calculs ou à éclaircir une notion théorique. D'autres exemples servent à illustrer une situation pratique. Il est bien sûr souhaitable que ces exemples soient réels, et pour la plupart, ils le sont. Ils sont tirés de situations rencontrées au cours de mes consultations. Cependant, étant donné que je n'ai pas la liberté de divulguer les données, je les ai créées artificiellement. Mais le contexte est réel.

LE TITRE

Le terme *sondage* évoque normalement un sondage politique ou un sondage d'opinion auprès d'une population humaine. Le terme *échantillonnage* décrit mieux ce dont il est question. Malheureusement son sens est trop large et pourrait éventuellement se référer à la statistique tout entière puisque la notion d'échantillonnage y intervient partout. J'utiliserai donc le terme *sondages* pour désigner la préoccupation principale du livre, mais je continuerai, dans le texte, à employer le terme *échantillonnage* dans son acception statistique habituelle.

Toutes les données des exercices et exemples sont disponibles en formats texte et Excel à l'adresse internet suivante:

www.lozedion.com/introduction-aux-sondages-donnees/

À *Éva, Diane, Carole, Alfred, Margaret, Laïla et Jacob*

Table des matière

CHAPITRE 1 NOTIONS DE PROBABILITÉ ***7***

1.1 Espace échantillon, événements et probabilité 7
Combinaisons d'événements 9
1.2 Probabilité conditionnelle et indépendance 10
Indépendance d'événements 10
1.3 Analyse combinatoire 11
Arrangements 12
Permutations 13
Combinaisons 13
Répartition de *N* objets en *k* groupes 14
1.4 Variables aléatoires 15
Vecteur aléatoire : distributions conjointe, marginale et conditionnelle 17
Covariance 22
Covariance et indépendance 23
Espérance et variance conditionnelles 25
Coefficient de corrélation 27
1.5 Lois de probabilité 28
Loi binomiale 28
Loi binomiale négative 29
Loi hypergéométrique 29
Loi multinomiale 29
Loi normale 30
Variables normales indépendantes 30
RÉSUMÉ 31
EXERCICES 32

CHAPITRE 2 NOTIONS DE STATISTIQUE ***41***

Population 41
2.1 Statistiques et estimateurs 42
Estimateurs sans biais 42
Variance d'un estimateur 42
Erreur quadratique moyenne 43
Coefficient de variation 43
2.2 Intervalles de confiance 43
2.3 La technique de linéarisation 44
2.4 Propriétés de quelques estimateurs 45
Échantillonnage et sondages 46
2.5 Classes de sondages 47
2.6 Population et paramètres 49

2.7 Base de sondage 50
2.8 Modes d'échantillonnage 50
Échantillonnage aléatoire simple 51
Échantillonnage stratifié 51
Échantillonnage systématique 51
Échantillonnage par grappes 52
2.9 Estimateurs 53
RÉSUMÉ 54
EXERCICES 54

CHAPITRE 3 ÉCHANTILLONNAGE ALÉATOIRE SIMPLE ET ESTIMATION D'UNE MOYENNE *59*
3.1 Échantillon aléatoire simple 59
3.2 Paramètres de la population 60
3.3 Estimation d'une moyenne 61
3.4 Estimation d'un total 65
La notion de poids 67
3.5 Différence de moyennes 67
3.6 Détermination de la taille de l'échantillon 69
3.7 Démonstrations 71
Première approche 71
Deuxième approche 72
Troisième approche 73
3.8 Quelques simulations 75
Population asymétrique mais pas trop 75
Population fortement asymétrique 76
RÉSUMÉ 78
EXERCICES 78

CHAPITRE 4 ESTIMATION D'UNE PROPORTION *99*
4.1 Grand échantillon 100
Estimation de p 100
Estimation de N_c 101
4.2 Différence de proportions 102
4.3 Détermination de la taille de l'échantillon 105
4.4 Petit échantillon 109
Procédure de construction d'un intervalle de confiance pour N_c. 109
4.5 Population infinie 112
Calculs à l'aide d'Excel 115
4.6 Intervalle de confiance pour une médiane 117
Procédure – grands échantillons 118
Procédure – petits échantillons 118
4.7 Démonstrations 120

RÉSUMÉ 121
EXERCICES 121

CHAPITRE 5 ESTIMATION D'UN QUOTIENT 129

5.1 Estimateur d'un quotient et ses propriétés 130
Prévoir la taille de l'échantillon 134
5.2 Estimation des paramètres d'un domaine U_d 135
Estimation de la moyenne μ_d d'un domaine 136
Estimation du total τ_d d'un domaine 138
5.3 Justifications et discussions 140
La moyenne d'un domaine considérée comme quotient 140
Propriétés conditionnelles de l'estimateur de la moyenne d'un domaine 140
5.4 Différence de deux moyennes 143
5.5 Données dichotomiques 144
5.6 Démonstrations 146
RÉSUMÉ 149
EXERCICES 150

CHAPITRE 6 ESTIMATION PAR LA DIFFÉRENCE, PAR LE QUOTIENT ET PAR LA RÉGRESSION 161

6.1 Estimateur par la différence 164
Discussion sur l'estimateur par la différence 165
6.2 Estimateur par le quotient 166
6.3 Estimateur par la régression 168
6.4 Comparaison des estimateurs 172
6.5 Simulations 174
RÉSUMÉ 179
EXERCICES 180

CHAPITRE 7 STRATIFICATION 185

7.1 Estimation d'une moyenne ou d'un total 186
Décomposition de la dispersion totale 192
7.2 Allocation des observations 192
Allocation proportionnelle 193
Allocation optimale 193
Observations auto-pondérés 195
Strates recensées 195
Efficacité de la stratification 196
Formules des variances 197
Comparaison des variances 198
Détermination de la taille de l'échantillon 199
Optimisation du coût d'échantillonnage 200

7.3 Estimation d'une proportion 202
Allocations proportionnelle et optimale 204
7.4 Stratification après sélection 205
Non-réponses 206
Discussion 207
7.5 Échantillonnage double 209
7.6 Démonstrations 210
RÉSUMÉ 215
EXERCICES 216

CHAPITRE 8 STRATIFICATION—AUTRES ESTIMATEURS *225*
8.1 Estimateurs séparés 225
8.2 Estimateurs combinés 229
Estimateur par le quotient combiné 230
Estimateur par la régression combiné 231
8.3 Démonstrations 232
RÉSUMÉ 233
EXERCICES 234

CHAPITRE 9 ÉCHANTILLONNAGE PAR GRAPPES À UN DEGRÉ ET PROBABILITÉS ÉGALES *239*
9.1 Introduction, définitions et notation 239
9.2 Estimation par la moyenne 242
9.3 Estimation par le quotient 243
9.4 Estimation par la régression 244
Comparaison des variances 245
9.5 Tirage systématique 246
RÉSUMÉ 250
EXERCICES 252

CHAPITRE 10 ÉCHANTILLONNAGE PAR GRAPPES À UN DEGRÉ ET PROBABILITÉS INÉGALES *263*
10.1 Tirage avec remise et probabilités proportionnelles à la taille 264
Comment générer l'échantillon 268
10.2 Tirage sans remise et probabilités inégales — l'estimateur de Horvitz-Thompson 269
L'estimateur de Horvitz-Thompson 270
La méthode de Lahiri 273
La méthode d'échantillonnage systématique 276
10.3 Les bornes de Stringer 278
10.4 Démonstrations 281
EXERCICES 286

CHAPITRE 11 TIRAGE PAR GRAPPES À DEUX DEGRÉS 289

11.1 Premier degré : Tirage avec remise et probabilités proportionnelles à la taille 289
11.2 Un théorème fondamental 291
11.3 Premier degré : échantillonnage aléatoire simple 292
11.4 Estimateur de Horvitz-Thompson au premier degré 295
11.5 Démonstrations 297
EXERCICES 298

ANNEXE A 305

Jeux de données 305
Tableau A01 305
Tableau A01.1 306
Tableau A01.2 306
Tableau A01.3 306
Tableau A02 307
Tableau A02.1 308
Tableau A03 308
Tableau A03.1 309
Tableau A03.2 309
Tableau A04 309
Tableau A05 310
Tableau A06 312
Tableau A07 313
Tableau A07.1 314
Tableau A08 314
Tableau A08.1 315
Tableau A09 316
Tableau A09.1 317
Tableau A10 317
Tableau A10.1 318

TRAVAUX CITÉS 319

1 Notions de probabilité

Ce chapitre présente un rappel sommaire de quelques notions de probabilités. La théorie qui y est exposée ne se veut pas complète: plusieurs notions importantes sont omises. Elle sera développée dans le cadre restreint d'espaces échantillons finis. Cette restriction simplifie considérablement la théorie sans trop limiter sa portée. La plupart des théorèmes énoncés dans la suite pourront être démontrés à partir des quelques bases présentées dans ce chapitre.

Ce chapitre essentiellement théorique est destiné à ceux qui ont l'intention de suivre le développement mathématique qui sous-tend les techniques de sondages. Ceux qui s'intéressent principalement aux applications pratiques peuvent sauter ce chapitre. Les quelques notions indispensables (espérance mathématique, variance, écart-type, covariance) seront reprises au moment où elles seront appliquées une première fois.

1.1 ESPACE ÉCHANTILLON, ÉVÉNEMENTS ET PROBABILITÉ

La première étape dans le développement d'un modèle probabiliste consiste à identifier l'ensemble des résultats possibles de l'expérience. Cet ensemble est appelé *espace échantillon* ou *ensemble fondamental*.

DÉFINITION 1.1 ESPACE ÉCHANTILLON

L'ensemble des résultats possibles d'une expérience aléatoire est appelé *espace échantillon*. Nous le désignons par Ω.

Tableau 1.1
Une population de 10 personnes

Unité	Nom	Nom abrégé	Sexe	Âge
1	Anne	A	F	22
2	Betty	B	F	21
3	Carole	C	F	24
4	Diane	D	F	21
5	Émilie	E	F	20
6	Fanny	F	F	19
7	George	G	H	22
8	Henri	H	H	24
9	Isidore	I	H	23
10	Jean	J	H	22

Dans la plupart des cas, on se limitera à des espaces échantillons finis ou dénombrables[1]

Exemple 1.1 *Tirage de deux unités*

On tire au hasard et sans remise deux personnes dans la population présentée au Tableau 1.1. Déterminer un espace échantillon approprié.

Solution

L'espace échantillon est constitué de *paires* d'unités :

$$\Omega = \begin{vmatrix} AB & AC & AD & AE & AF & AG & AH & AI & AJ \\ BC & BD & BE & BF & BG & BH & BI & BJ & CD \\ CE & CF & CG & CH & CI & CJ & DE & DF & DG \\ DH & DI & DJ & EF & EG & EH & EI & EJ & FG \\ FH & FI & FJ & GH & GI & GJ & HI & HJ & IJ \end{vmatrix}$$

Exemple 1.2 *Échantillons ordonnés*

Dans l'Exemple 1.1, les résultats sont énumérés sans tenir compte de l'ordre : on y trouve AB. mais pas BA. Si on avait précisé qu'il s'agit bien de deux tirages successifs, on aurait pu distinguer chaque paire selon l'ordre du tirage, et inclure, par exemple, AB et BA dans Ω. Le nombre d'éléments de Ω se verrait alors doublé.

DÉFINITION 1.2 ÉVÉNEMENT

Tout sous-ensemble $E \subset \Omega$ est appelé *événement*.

[1] Un ensemble est *dénombrable* s'il peut être mis en correspondance avec les entiers naturels 1, 2, …

Exemple 1.3 *Événement dans un double tirage*

Dans l'Exemple 1.1, l'événement E : « Les deux personnes choisies sont des hommes », est représenté par le sous-ensemble E = {GH ; GI ; GJ ; HI ; HJ ; IJ}.

Exemple 1.4 *Exemples d'événements*

Dans l'Exemple 1.1, chacun des événements suivants est associé à un sous-ensemble de Ω. Le lecteur pourra facilement les identifier.

K : « Les deux personnes sont de même sexe »;

L : « Les deux personnes sont de sexes opposés »;

M : « Exactement une des deux personnes a plus de 22 ans »;

N : « Au moins l'une des deux personnes à plus de 22 ans »;

Q : « Les deux personnes ont 22 ans ou moins »;

R : « Les deux personnes sont des femmes ».

DÉFINITIONS 1.3 ÉVÉNEMENT IMPOSSIBLE ET ÉVÉNEMENT CERTAIN.

L'ensemble vide ϕ représente l'*événement impossible*, un événement qui ne peut pas se produire; et l'ensemble Ω lui-même est l'*événement certain*, un événement qui doit forcément se produire.

COMBINAISONS D'ÉVÉNEMENTS

Les opérations d'ensembles usuelles — la complémentation, l'intersection, l'union et la différence — appliquées à deux événements E et F engendrent de nouveaux ensembles, donc de nouveaux événements. Voici le sens de ces opérations en termes d'événements:

Opération	Sens en termes d'ensembles	Sens en termes d'événements
Complémentation E^c	L'ensemble des éléments de Ω qui n'appartiennent pas à E	« E ne s'est pas produit »
Intersection $E \cap F$	L'ensemble des éléments de Ω qui appartiennent à E et à F	« E et F se sont produits ».
Réunion $E \cup F$	L'ensemble des éléments de Ω qui appartiennent à E ou à F (ou aux deux)	« E ou F s'est produit ».
Différence E - F (ou E\F)	L'ensemble des éléments de Ω qui appartiennent à E mais pas à F	« E s'est produit mais pas F ». Noter que $E\backslash F = E \cap F^c$.

DÉFINITION 1.4 ÉVÉNEMENTS INCOMPATIBLES OU MUTUELLEMENT EXCLUSIFS

Deux événements E et F sont dits incompatibles ou mutuellement exclusifs, si

$$E \cap F = \phi.$$

Probabilité d'un résultat

Dans un espace échantillon fini ou dénombrable, il est possible d'attribuer une probabilité $p(\omega)$ à chaque résultat $\omega \in \Omega$. Les valeurs de p satisfont aux conditions suivantes:

(1.1) $$0 \leq p(\omega) \leq 1 \text{ pour tout } \omega \in \Omega \text{ et } \sum_{\omega \in \Omega} p(\omega) = 1 .$$

La probabilité d'un *événement* E peut ensuite être déterminée à partir des éléments $p(\omega)$; c'est la somme des probabilités des résultats qu'il contient:

(1.2) $$P(E) = \sum_{\omega \in E} p(\omega) .$$

1.2 PROBABILITÉ CONDITIONNELLE ET INDÉPENDANCE

DÉFINITION 1.5 PROBABILITÉ CONDITIONNELLE

La probabilité conditionnelle d'un événement F étant donné un événement E de probabilité non nulle, désignée par P(F | E), est définie par

$$P(F \mid E) = \frac{P(E \cap F)}{P(E)} .$$

INDÉPENDANCE D'ÉVÉNEMENTS

DÉFINITION 1.6 ÉVÉNEMENTS INDÉPENDANTS

Deux événements E et F sont dits *indépendants* si et seulement si

(1.3) $$P(E \cap F) = P(E)\, P(F).$$

Les énoncés suivants sont tous équivalents à (1.3):

$$P(E \mid F) = P(E) \qquad P(F \mid E) = P(F) \qquad P(E \mid F) = P(E \mid F^c) \qquad P(F \mid E) = P(F \mid E^c).$$

Indépendance mutuelle de plusieurs événements

La notion d'indépendance de deux événements se généralise à n événements E_1; ... ; E_n. Mais il ne suffit pas que

$$P(E_1 \cap E_2 \cap \ldots \cap E_n) = P(E_1) \times \ldots \times P(E_n).$$

Il faut en plus qu'une propriété analogue soit vérifiée pour tout sous-ensemble de l'ensemble $\{E_1; E_2; \ldots; E_n\}$.

DÉFINITION 1.7 ÉVÉNEMENTS MUTUELLEMENT INDÉPENDANTS

Les événements E_1, E_2, ... , E_n sont dits *mutuellement indépendants* si pour $k = 2, 3, \ldots . n$, la condition

$$P(E_{i_1} \cap E_{i_2} \cap \ldots \cap E_{i_k}) = P(E_{i_1}) \times P(E_{i_2}) \times \ldots \times P(E_{i_k})$$

est vérifiée pour tout choix i_1. ... . i_k de k entiers parmi les entiers 2, 3, ... ,n.

En sondages, la notion d'indépendance mutuelle intervient lorsqu'on parle de tirages successifs d'objets dans un ensemble d'objets. Lorsque les tirages se font *avec* remise (chaque objet tiré est remis dans la population avant le prochain tirage), on dit que les *tirages* sont indépendants, ce qui signifie ceci: si E_1, E_2,...,E_n sont des événements se rapportant, respectivement aux premier, deuxième,..., et n^e tirages, alors les événements E_1; ... ; E_n sont *mutuellement indépendants.*

1.3 ANALYSE COMBINATOIRE

Les probabilités $p(\omega)$ attribuées aux éléments d'un espace échantillon Ω dépendent de la nature de l'expérience aléatoire. Une hypothèse particulièrement simple mais néanmoins réaliste dans certains contextes est l'hypothèse d'équiprobabilité.

Hypothèse d'équiprobabilité

L'hypothèse d'équiprobabilité est l'hypothèse selon laquelle tous les éléments ω de Ω ont même probabilité. On en déduit que

$$p(\omega) = \frac{1}{Card(\Omega)},$$

où $Card(\Omega)$ désigne la cardinalité de Ω, le nombre d'éléments qu'il contient.

La probabilité d'un événement E est alors donnée par

$$P(E) = \frac{Card(E)}{Card(\omega)}. \tag{1.4}$$

Les ensembles E et Ω sont souvent très grands, et l'application de la formule (1.4) serait onéreuse s'il fallait énumérer les éléments de E et de Ω pour les compter. Quelques techniques élémentaires d'analyse combinatoire permettent de dénombrer les ensembles E et Ω sans avoir à en énumérer les éléments. Plusieurs problèmes de dénombrement peuvent être illustrés par un même paradigme, particulièrement pertinent dans le cadre des sondages : des tirages successifs dans une population. Chacune des quelques formules qui suivent correspond à un choix parmi les possibilités suivantes :

- Les tirages sont effectués *avec* remise ou *sans* remise;
- Dans l'énumération des résultats possibles, on tient compte de l'ordre ou pas.

Tirages avec remise

THÉORÈME 1.1 **TIRAGES AVEC REMISE, TENANT COMPTE DE L'ORDRE**

Le nombre de façons de tirer n objets parmi N, avec remise et en tenant compte de l'ordre est

$$N^n.$$

Exemple 1.5 *Nombre de « mots » de n lettres*

Combien peut-on former de mots de 5 lettres à partir d'un alphabet de 26 lettres (toute suite de 5 lettres est un « mot »)?

Solution

On compte ici le nombre de façons de placer une lettre dans chacune des 5 cases, comme, par exemple : | W | Z | W | U | Z |. Chacune peut être occupée de 26 façons. Le nombre total de mots est donc 26^5. Nous avons tenu compte de l'ordre dans le sens que le mot illustré ci-dessus est considéré comme différent de celui-ci : | U | Z | W | Z | W |.

Exemple 1.6 *Tirage avec remise de n objets parmi N*

Quinze finalistes participent à un triathlon : épreuve à vélo, à pied, et à la nage. Le gagnant de chaque épreuve recevra un prix (le prix est spécifique : un pour le vélo, un pour la course à pied, et un pour la nage). Combien y a-t-il de façons d'attribuer les trois prix ?

Solution

Si on désigne les 15 participants par A_1, A_2,…, A_{15}, chacun des résultats possibles peut être représenté par un triplé de lettres parmi ces quinze. Le triplet $A_3A_8A_{12}$, par exemple, identifierait le résultat : «les participants A_3, A_8 et A_{12} ont gagné aux épreuves à vélo, à pied et à la nage, respectivement». L'ordre compte parce que $A_3A_8A_{12}$ est un résultat distinct de $A_8A_{12}A_3$. Les tirages sont dits avec remise dans le sens que des résultats comme $A_8A_{12}A_8$ ou $A_8A_8A_8$ sont possibles. Il y a donc $15\times15\times15 = 3\,375$ résultats possibles.

ARRANGEMENTS

Combien y a-t-il de façons de choisir n objets parmi N? Commençons par préciser la question.

Exemple 1.7 *Nombre de comités de* 3 *personnes choisies parmi* 5

De combien de façons peut-on former un comité constitué d'un président, un trésorier et un secrétaire, lorsque ces trois personnes peuvent être choisies parmi 5 personnes (et une même personne ne peut cumuler plus d'un poste) ?

Solution

Si on désigne les 5 personnes par A, B, C, D et E, chaque résultat peut être représenté par une suite de 3 de ces lettres, comme par exemple celui-ci : | C | A | B |. Cet ensemble de 3 lettres désigne l'événement suivant : C est président, A trésorier, et B secrétaire. L'ordre compte, dans le sens que | C | A | B | et | B | A | C | sont deux résultats distincts. Le nombre total de comités est donc $5 \times 4 \times 3 = 60$.

Ce problème est un problème dit *d'arrangement* de n objets choisis parmi N, c'est-à-dire, le nombre de façons de choisir n objets parmi N et *de les placer*. On désigne ce nombre d'arrangements par A_n^N .

THÉORÈME 1.2 **ARRANGEMENTS**

Le nombre d'arrangements de n objets choisis parmi N est donné par

$$\mathrm{A}_n^N = \frac{N!}{(N-n)!}\,. \tag{1.5}$$

PERMUTATIONS

Le cas où $n = N$ a une signification particulière, car il représente le nombre de *permutations* de N objets, c'est-à-dire, le nombre de façons *d'ordonner* N objets.

COROLLAIRE **PERMUTATIONS**

Le nombre d'arrangements de N objets choisis parmi N est donné par

$$\mathrm{A}_n^N$$

Exemple 1.8 *Permutations de N objets*

Combien y a-t-il de façons de placer 25 volumes sur une tablette ?

Solution

25! = 15 511 210 043 330 985 984 000 000.

COMBINAISONS

La formule de A_n^N du nombre d'arrangements de n objets choisi parmi N tient compte de l'ordre. Le nombre de façons de choisir n objets parmi N *sans tenir compte de l'ordre*, désigné par $\binom{N}{n}$, est appelé le nombre de *combinaisons* de n objets choisis parmi N.

Exemple 1.9 *Nombre de comités de* 3 *personnes sans attribution de poste*

Dans le cadre de l'Exemple 1.7, supposons que les postes des trois membres du comité ne sont pas différenciés. Combien y a-t-il de comités possibles?

Solution

Le fait que les rôles sont identiques signifie que l'ordre des tirages ne compte pas: les choix | C | A | B | et | B | C | A | ne comptent que pour un. Il y a 3! permutations des lettres A, B, et C. Dans le calcul A_3^5 toute combinaison de 3 personnes est compté 3! fois. Le nombre de combinaisons est donc

$$\binom{5}{3} = \frac{\mathrm{A}_3^5}{3!} = \frac{5!/(5-3)!}{3!} = \frac{5!}{3!(5-3)!} = 10.$$

Voici la formule générale du nombre de combinaisons de n objets choisis parmi N.

THÉORÈME 1.3 **COMBINAISONS**

Le nombre de façons de choisir n unités parmi N, sans répétitions et sans tenir compte de l'ordre, appelé *nombre de combinaisons* de n objets choisis parmi N, est donné par

(1.6) $$\binom{N}{n} = \frac{N!}{n!(N-n)!}.$$

RÉPARTITION DE N *OBJETS EN* K *GROUPES*

Le nombre de combinaisons de n objets parmi N peut être vu comme le nombre de façons de répartir N objets en deux groupes, de n et N - n objets. On peut généraliser le problème et considérer le nombre de façons de répartir N objets en k groupes, où k peut être supérieur à 2.

Exemple 1.10 *Répartition d'objets dans des groupes*

Combien y a-t-il de façons de répartir 12 personnes en trois comités, de 3, 4, et 5 personnes?

Solution

Il y a $\binom{12}{3}$ façons de choisir les 3 personnes qui appartiendront au premier comité; ensuite il y a $\binom{9}{4}$ façons de choisir celles qui appartiendront au deuxième comité; et enfin $\binom{5}{5} = 1$ façon de placer les 5 derniers au troisième groupe. La réponse est donc $\binom{12}{3}\binom{9}{4}\binom{5}{5} = \frac{12!}{3!9!}\frac{9!}{4!5!}\frac{5!}{5!0!} = \frac{12!}{3!4!5!} = 27\,720$. Ce nombre est désigné par $\binom{2}{3;4;5}$.

THÉORÈME 1.4 **RÉPARTITION DE *N* OBJETS EN *K* GROUPES**

Le nombre de façons de répartir N objets en k groupes de tailles N_1, N_2 , ... , N_k. désigné par $\binom{N}{N_1 ; N_2 ; \ldots ; N_k}$, est donné par

(1.7) $$\binom{N}{N_1 ; N_2 ; \ldots ; N_k} = \frac{N!}{N_1!N_2!\ldots N_k!}.$$

La formule (1.7) représente aussi le nombre de permutations d'objets dont certains sont identiques (non discernables).

Exemple 1.11 *Permutations d'objets non discernables*

Combien y a-t-il de permutations discernables des 10 lettres AABBBCCCCC?

Solution

Chaque permutation de ces objets correspond à une et une seule façon de placer 10 objets dans 3 groupes, de 2, 3, et 5 objets. Par exemple, la permutation CABACBCCCB peut s'interpréter comme l'événement suivant: la première personne est dans le groupe C, la 2e dans le groupe A, la 3e dans le groupe B, …. , la 10e dans le groupe B. Donc chaque répartition des 10 objets en 3 groupes correspond à une et une seule permutation des 10 objets. Le nombre de permutations est donc le même, soit $\frac{10!}{2!3!5!} = 2\,520$.

1.4 VARIABLES ALÉATOIRES

Une variable aléatoire, généralement désignée par une lettre majuscule comme X, Y, Z, est une caractéristique numérique associée aux résultats d'une expérience aléatoire. En voici une définition formelle :

DÉFINITION 1.8 VARIABLE ALÉATOIRE

Une variable aléatoire est une fonction X qui fait correspondre à chaque élément $\omega \in \Omega$ un nombre réel $X(\omega)$.

Fonction de probabilité

On désigne par $X(\Omega)$ l'ensemble des valeurs que peut prendre X:

$$X(\Omega) = \{x \mid x = X(\omega) \text{ pour au moins un } \omega \in \Omega\}.$$

Pour tout $x \in X(\Omega)$, la probabilité que X prenne la valeur x, désignée par $P(X = x)$ ou $p(x)$, est la probabilité de l'ensemble des éléments $\omega \in \Omega$ pour lesquels $X(\omega) = x$. Cet ensemble est désigné par $X^{-1}(x)$:

$$X^{-1}(x) = \{\omega \in \Omega \mid X(\omega) = x\}.$$

La correspondance entre les éléments de $X(\Omega)$ et leur probabilité est appelée *fonction de probabilité* ou *fonction de masse.*

DÉFINITION 1.9 FONCTION DE PROBABILITÉ

La fonction de probabilité $p(x)$ d'une variable aléatoire discrète X fait correspondre à chaque valeur $x \in X(\Omega)$ la probabilité que X prenne la valeur x :

$$p(x) = P(X = x) = P[X^{-1}(x)].$$

Considérons l'espace échantillon défini à l'Exemple 1.1, et la variable aléatoire X : « Nombre de femmes dans l'échantillon ». Cette variable fait correspondre l'une des valeurs 0, 1, 2 à

chacun des éléments de l'espace échantillon. Les valeurs de $p(x)$ sont calculées à partir de l'espace échantillon constitué des 45 paires d'unités. Par exemple, $p(0) = P[X^{-1}(0)] = P(\{\omega \mid X(\omega) = 0\}) = P(\{GH ; GI ; GJ ; HI ; HJ ; IJ\})$. Si on admet que l'hypothèse d'équiprobabilité, soit que chaque choix de 2 unités a même probabilité, alors chaque résultat est de probabilité 1/45.

On peut ainsi déterminer la fonction de probabilité de X :

x	0	1	2	
$p(x)$	6/45	24/45	15/45	1

Une fonction de probabilité $p(x)$ satisfait aux conditions suivantes :

(1.8) $$p(x) \geq 0 \text{ et } \sum_x p(x) = 1 .$$

Espérance d'une variable aléatoire

DÉFINITION 1.10 ESPÉRANCE MATHÉMATIQUE

Soit X une variable de fonction de probabilité p. Alors *l'espérance mathématique* de X est définie par

(1.9) $$E(X) = \sum_x xp(x) .$$

Plus généralement, l'espérance d'une fonction $\varphi(X)$ est définie par:

(1.10) $$E[\varphi(X)] = \sum_x \varphi(x)p(x) .$$

Variance d'une variable aléatoire

DÉFINITION 1.11 VARIANCE D'UNE VARIABLE ALÉATOIRE

Soit X une variable de fonction de probabilité p et de moyenne μ. Alors la *variance* de X, désignée par $V(X)$, est définie par

(1.11) $$V(X) = \sum_x (x - \mu)^2 p(x) .$$

En vertu de (1.10), on peut aussi définir la variance comme $E[\varphi(X)]$, où $\varphi(X) = (X - \mu)^2$.

THÉORÈME 1.5 **AUTRE FORMULE DE LA VARIANCE**

Soit X une variable aléatoire de moyenne μ. Alors sa variance peut être calculée de la façon suivante :

(1.12) $$V(X) = \sum_x x^2 p(x) - \mu^2 = E(X^2) - [E(X)]^2.$$

DÉFINITION 1.12 ÉCART-TYPE

L'écart-type d'une variable aléatoire X, désignée par $ET(X)$, est la racine carrée de sa variance :

(1.13) $$ET(X) = \sqrt{V(X)} .$$

L'écart-type d'une variable aléatoire X est aussi désignée par σ ou σ_X.

Fonctions affines

L'espérance d'une fonction de $\varphi(X)$ se calcule aisément à partir de celle de X lorsque φ est une fonction affine, c'est-à-dire, une fonction de la forme

$$Y = a + bX,$$

où a et b sont des constantes.

THÉORÈME 1.6 **ESPÉRANCE ET VARIANCE D'UNE FONCTION AFFINE**

Soit X est une variable aléatoire et $Y = a + bX$, où a et b sont des constantes. Alors

(1.14) $$E(Y) = E(a + bX) = a + bE(X) \text{ et } V(Y) = V(a + bX) = b^2 V(X).$$

Exemple 1.12 *Espérance, variance et écart-type d'une fonction affine*

Vous donnez ordre à votre courtier de vous acheter 120 actions de la compagnie ABC au prix du marché X. Supposons que $E(X) = 27$ et $ET(X) = 3$. Vous recevrez une facture dont le montant Y est la valeur de vos actions, plus une commission forfaitaire de 50 $. Déterminer l'espérance et l'écart-type de Y.

Solution

$Y = 50 + 120X$. Alors $E(Y) = 50 + 120E(X) = 3\,290$ $. $ET(Y) = |120|ET(X) = 360$ $.

VECTEUR ALÉATOIRE : DISTRIBUTIONS CONJOINTE, MARGINALE ET CONDITIONNELLE

Un même contexte expérimental peut donner lieu à plusieurs variables aléatoires. Un sondage auprès d'une population de ménages, par exemple, fournira normalement plusieurs données sur un même ménage: le revenu X_1, le nombre de personnes X_2, le nombre d'enfants X_3, etc. Les fonctions de probabilité de ces variables ne recèlent pas d'information sur les *relations* entre elles. Pour cela, il nous faut la *fonction de probabilité conjointe* du triplé $\boldsymbol{X} = [X_1 ; X_2 ; X_3]$ — appelé *vecteur* aléatoire.

DÉFINITION 1.13 FONCTION DE PROBABILITÉ CONJOINTE

La fonction de probabilité conjointe d'un vecteur aléatoire $\boldsymbol{X} = [X_1 ; X_2 ; \ldots ; X_n]$ est une fonction $p_{\boldsymbol{X}}(x_1 ; x_2 ; \ldots ; x_n)$ qui fait correspondre à chaque vecteur de nombres $\boldsymbol{x} = [x_1 ; x_2 ; \ldots ; x_n]$ la probabilité $P(\boldsymbol{X} = \boldsymbol{x})$:

$$p_{\boldsymbol{X}}(x_1 ; x_2 ; \ldots ; x_n) = P(X_1 = x_1 \text{ et } X_2 = x_2 \text{ et } \ldots \text{ et } X_n = x_n).$$

Exemple 1.13 *Échantillon de taille* 2

Considérons une population de 4 ménages dont les revenus sont (en milliers de dollars) les suivants : {40 ; 60 ; 60 ; 120}. On tire deux ménages, l'un après l'autre, *sans* remise. Soit X le revenu du premier ménage et Y celui du deuxième. Déterminer la fonction de probabilité conjointe $p_{X;Y}(x ; y)$ de $[X ; Y]$.

Solution

Voici l'ensemble de toutes les valeurs possibles de $[X\,;\,Y]$ pour les 12 échantillons possibles. et leur probabilité conjointe :

		y			
		40	60	120	$p(x)$
x	40	0	1/6	1/12	1/4
	60	1/6	1/6	1/6	1/2
	120	1/12	1/6	0	1/4
	$p(y)$	1/4	1/2	1/4	1

Distributions marginales

De la distribution conjointe d'un vecteur aléatoire découlent les distributions de ses composantes.

DÉFINITION 1.14 FONCTION DE PROBABILITÉ MARGINALE

Lorsque la fonction de probabilité de X_i est déterminée à partir de la fonction de probabilité conjointe $p_{\boldsymbol{X}}(x_1\,;\,x_2\,;\,\ldots\,;\,x_n)$ d'un vecteur aléatoire $\boldsymbol{X} = [X_1\,;\,X_2\,;\,\ldots\,;\,X_n]$, elle est appelée *fonction de probabilité marginale*.

Le cas $n = 2$, qui est celui qui nous concernera le plus souvent dans ce livre, s'énonce facilement. La fonction de probabilité marginale $p_X(x)$ de X est la somme de $p_{X,Y}(x\,;\,y)$ sur toutes les valeurs de y. La fonction de probabilité $p_Y(y)$ est définie de façon analogue:

$$p_X(x) = \sum\nolimits_y p_{X;Y}(x;y) \qquad p_Y(y) = \sum\nolimits_x p_{X;Y}(x;y)\,.$$

Exemple 1.14 *Fonction de densité marginale*

Dans l'Exemple 1.13, déterminer la fonction de probabilité marginale de X.

Solution

Les fonctions de probabilité de X et de Y se trouvent dans les marges du tableau. Voici donc la fonction de probabilité de X :

x	40	60	120	
$p_X(x)$	1/4	1/2	1/4	1

Celle de Y est identique à celle-là.

Distributions conditionnelles

La notion de probabilité conditionnelle définie pour des événements s'applique également aux variables aléatoires. Rappelons que si E et F sont deux événements, alors la probabilité conditionnelle de F étant donné E est définie par $P(F \mid E) = P(E \cap F)/P(E)$. Si E est l'événement $\{X = x\}$ et F est l'événement $\{Y = y\}$, alors $P(F \mid E) = P(Y = y \mid X = x) = P(X = x\,;\,Y = y)/P(X = x)$. C'est ce qui motive la définition suivante :

DÉFINITION 1.15 FONCTION DE PROBABILITÉ CONDITIONNELLE

La fonction de probabilité conditionnelle de Y étant donné $X = x$ est définie par

$$p_{Y|x}(y) = \frac{p_{X,Y}(x\,;\,y)}{p_X(x)}\,.$$

Indépendance de variables aléatoires

La notion d'indépendance, définie plus haut pour des événements, est définie de façon analogue pour des variables aléatoires.

DÉFINITION 1.16 VARIABLES ALÉATOIRES INDÉPENDANTES

Deux variables aléatoires X et Y sont dites indépendantes si pour tout $(x\,;\,y)$.

$$p_{X.Y}(x\,;\,y) = p_X(x) \times p_Y(y).$$

C'est encore la règle du produit, mais appliquée à toute une famille d'événements, les événements de la forme $\{X = x\}$ et $\{Y = y\}$.

Remarquez que la condition est équivalente à $p_{Y|x}(y) = p_Y(y)$ (lorsque $p_X(x) > 0$).

Exemple 1.15 *Indépendance de variables aléatoires dans des tirages successifs*

Si X et Y sont les revenus de deux familles tirées *avec remise* d'une population, alors X et Y sont *indépendantes*. Voici la distribution conjointe de X et Y de l'exemple 1.13 :

		y			
		40	60	120	$p_X(x)$
x	40	1/16	1/8	1/16	1/4
	60	1/8	1/4	1/8	1/2
	120	1/16	1/8	1/16	1/4
	$p_Y(y)$	1/4	1/2	1/4	1

On peut vérifier que $p_{X.Y}(x\,;\,y) = p_X(x) \times p_Y(y)$ pour toute paire $[x\,;\,y]$. Ce qui n'était pas le cas dans l'Exemple 1.13, là où les tirages se faisaient *sans* remise.

Remarque *Indépendance de variables aléatoires*

L'indépendance illustrée dans le dernier exemple est facile à reconnaître en pratique. Mais il y a des cas où l'on se doute bien qu'une dépendance existe sans pour autant pouvoir le démontrer autrement qu'en prélevant des statistiques.

Si X et Y représentent, respectivement, la taille et le poids d'une même personne tirée au hasard dans une population, alors on s'attend à ce que X et Y soient dépendantes : si on sait que la personne est grande de taille, les chances que son poids soit élevé sont plus fortes.

Par contre, si X et Y représentent, respectivement, la taille et le quotient intellectuel, alors X et Y sont vraisemblablement indépendantes. « Les grands de taille ne sont ni plus

intelligents ni moins intelligents que les petits » : c'est ce qu'on suppose lorsqu'on affirme que X et Y sont indépendantes.

Fonctions scalaires de variables aléatoires

Si $Z = \varphi(X\,;\,Y)$ est une fonction scalaire de $[X\,;\,Y]$, l'espérance de Z est calculée par :

$$E[\varphi(X;Y)] = \sum_x \sum_y \varphi(x;y)\, p_{X,Y}(x;y)\ .$$

Exemple 1.16 *Espérance d'une fonction non linéaire*

On tire au hasard un ménage dans une population de 10 ménages. Soit X le nombre de personnes et Y le nombre de téléviseurs dans le ménage tiré. Voici la liste des ménages de la population, avec le nombre de personnes et le nombre de téléviseurs dans chacun :

Ménage #	Nombre de personnes	Nombre de téléviseurs	Ménage #	Nombre de personnes	Nombre de téléviseurs
1	1	1	6	2	3
2	1	2	7	3	1
3	2	1	8	3	1
4	2	2	9	3	2
5	2	3	10	3	3

Voici la distribution conjointe de $[X\,;\,Y]$:

		y (Nombre de téléviseurs)			
		1	2	3	$p_X(x)$
x (Nombre de personnes)	1	0,1	0,1	0	0,2
	2	0,1	0,1	0,2	0,4
	3	0,2	0,1	0,1	0,4
	$p_Y(y)$	0,4	0,3	0,3	1

Soit Z le nombre de téléviseurs par personne dans le ménage. Z est une fonction scalaire de $[X\,;\,Y]$: $Z = \varphi(X\,;\,Y) = Y/X$. Voici la valeur de $\varphi(x\,;\,y)$ correspondant à chaque paire $(x\,;\,y)$:

		y (Nombre de téléviseurs)		
		1	2	3
x (Nombre de personnes)	1	1	2	3
	2	1/2	1	3/2
	3	1/3	2/3	1

Alors

$$E(Z) = E[\varphi(X\,;\,Y)] = 1(0,1) + 2(0,1) + 3(0) + \frac{1}{2}(0,1) + 1(0,1) + \frac{3}{2}(0,2) + \frac{1}{3}(0,2) + \frac{2}{3}(0,1) + 1(01)$$

$$= 59/60.$$

Remarque

Au dernier exemple, on aurait pu procéder à partir des définitions de base. Z est une variable aléatoire dont l'espérance, selon la définition, est $E(Z) = \Sigma_z z p_Z(z)$. La fonction de probabilité $p_Z(z)$ peut être déterminée à partir du dernier tableau :

z	1/3	1/2	2/3	1	3/2	2	3	
$p_Z(z)$	0,2	0,1	0,1	0,3	0,2	0,1	0	1

Alors $E(Z) = \left(\frac{1}{3}\right)(0,2) + \left(\frac{1}{2}\right)(0,1) + \left(\frac{2}{3}\right)(0,1) + 1(0,3) + \left(\frac{3}{2}\right)(0,2) + 2(0,1) = \frac{59}{60}$.

Fonctions linéaires

Lorsque $Y = \varphi(X_1; X_2; \ldots ; X_n)$ est une combinaison linéaire de variables aléatoires, son espérance s'exprime simplement en termes des espérances des X_i :

THÉORÈME 1.7 **ESPÉRANCE D'UNE COMBINAISON LINÉAIRE DE VARIABLES ALÉATOIRES**

Soit $X_1, X_2, \ldots, X_n$ n variables aléatoires et $a_1, a_2, \ldots, a_n$ des constantes. Alors

$$E(a_1X_1 + a_2X_2 + \ldots + a_nX_n) = a_1E(X_1) + a_2E(X_2) + \ldots + a_nE(X_n).$$

Exemple 1.17 *Espérance d'une somme et d'une différence*

On interroge un ménage tiré au hasard dans un quartier. Soit X le revenu du père de famille et Y celui de la mère (on exclut de la population les familles monoparentales). Supposons que $E(X)$ = 58 000 $ et $E(Y)$ = 47 000 $. Quelle est l'espérance du revenu total du couple ?

Solution

Si $Z = X + Y$ est le revenu du couple, alors son espérance est $E(Z) = E(X + Y) = E(X) + E(Y)$ = 58 000 + 47 000 = 105 000 $.

Exemple 1.18 *Espérance d'une combinaison linéaire*

Vous donnez ordre à votre courtier d'acheter 100 actions de la compagnie A et 200 actions de la compagnie B, au prix du marché. Donc X et Y, les prix payés pour les actions de A et de B, respectivement, sont aléatoires. Supposons que $E(X)$ = 40 $ et $E(Y)$ = 60 $. Quelle est l'espérance de la valeur totale de vos actions ?

Solution

La valeur totale de vos actions est $100X + 200Y$, et son espérance est $E(100X + 200Y) = E(100X) + E(200Y) = 100E(X) + 200E(Y) = 100(40) + 200(60) =$ 16 000 $.

COVARIANCE

DÉFINITION 1.17 COVARIANCE

Soit X et Y deux variables aléatoires de moyennes $E(X) = \mu_x$ et $E(Y) = \mu_y$, et de fonction de probabilité conjointe $p_{X;Y}(x\,;\,y)$. La *covariance* entre X et Y, désignée par $C(X\,;\,Y)$, est définie par

$$C(X;Y) = E[(X-\mu_x)(Y-\mu_y)] = \sum_x \sum_y [(x-\mu_x)(y-\mu_y)]p_{X;Y}(x;y)\ . \tag{1.15}$$

Nous avons, pour la covariance comme pour la variance, une formule alternative :

$$C(X;Y) = E(XY) - E(X)E(Y) = \sum_x \sum_y xyp_{X;Y}(x;y) - [\sum_x xp(x)][\sum_y yp(y)]\ . \tag{1.16}$$

THÉORÈME 1.8 **PROPRIÉTÉS D'UNE COVARIANCE**

Soit X et Y deux variables aléatoires, et a, b et c des constantes. Alors

(1.17)

(i) $C(X\,;\,b) = 0$

(ii) $C(X + Y\,;\,Z) = C(X\,;\,Z) + C(Y\,;\,Z)$

(iii) $C(aX + c;\,bY) = ab\ C(X\,;\,Y)$

(iv) $C(X\,;\,X) = V(X)$

La variance d'une combinaison linéaire de variables aléatoires peut s'exprimer en fonction des variances et des covariances.

THÉORÈME 1.9 **VARIANCE D'UNE COMBINAISON LINÉAIRE**

Soit X et Y deux variables aléatoires indépendantes et a_1, a_2, … , a_n des constantes. Alors

$$V\left(a_1X_1 + a_2X_2 + \ \ldots\ + a_kX_k\right) = \sum_{i=1}^{k} a_i^2V(X_i) + \sum_i \sum_{j\neq i} a_ia_jC(X_i;X_j)\ . \tag{1.18}$$

Par la symétrie de la covariance, le deuxième terme peut aussi s'écrire comme

$$\sum_i \sum_{j\neq i} a_ia_jC(X_i;X_j)\ = 2\sum_i \sum_{j>i}^{n} a_ia_jC(X_i;X_j)\ .$$

Exemple 1.19 *Combinaison linéaire de trois variables*

Exprimer $V(a_1X_1 + a_2X_2 + a_3X_3)$ en termes des variances et des covariances des trois variables X_1, X_2, et X_3.

Solution

$V(a_1X_1+a_2X_2+a_3X_3)$ =

$$a_1^2V(X_1) + a_2^2V(X_2) + a_3^2V(X_3)\ +\ 2a_1a_2C(X_1;X_2) + 2a_1a_3C(X_1;X_3) + 2a_2a_3C(X_2;X_3)\ .$$

Cas particuliers

Soit X et Y deux variables aléatoires, et a et b deux constantes. Alors

$$V(aX + bY) = a^2V(X) + b^2V(Y) + 2abC(X\,;\,Y);$$

(1.19) $$V(X + Y) = V(X) + V(Y) + 2C(X\,;\,Y);$$

$$V(X - Y) = V(X) + V(Y) - 2C(X\,;\,Y).$$

COVARIANCE ET INDÉPENDANCE

THÉORÈME 1.10 **COVARIANCE DE VARIABLES INDÉPENDANTES**

Soit X et Y deux variables aléatoires indépendantes. Alors

(i) $E(XY) = E(X)\,E(Y)$; (ii) $C(X\,;\,Y) = 0$.

Exemple 1.20 *Calcul d'une covariance*

Dans l'Exemple 1.16, on a $E(X) = 2{,}2$, $E(Y) = 1{,}9$ et donc

$$\begin{aligned}C(X\,;\,Y) = {} & (1 - 2{,}2)(1 - 1{,}9)(0{,}1) + (1 - 2{,}2)(2 - 1{,}9)(0{,}1) + (1 - 2{,}2)(3 - 1{,}9)(0) \\ & + (2 - 2{,}2)(1 - 1{,}9)(0{,}1) + (2 - 2{,}2)(2 - 1{,}9)(0{,}1) + (2 - 2{,}2)(3 - 1{,}9)(0{,}2) \\ & + (3 - 2{,}2)(1 - 1{,}9)(0{,}2) + (3 - 2{,}2)(2 - 1{,}9)(0{,}1) + (3 - 2{,}2)(3 - 1{,}9)(0{,}1) \\ & = 0{,}02.\end{aligned}$$

THÉORÈME 1.11 **VARIANCE D'UNE FONCTION LINÉAIRE DE VARIABLES ALÉATOIRES INDÉPENDANTES**

Soit $X_1, X_2, \ldots, X_n$ n variables aléatoires indépendantes et $a_1, a_2, \ldots, a_n$ n constantes. Alors

$$V(a_1X_1 + a_2X_2 + \ldots + a_kX_k) = a_1^2V(X_1) + a_2^2V(X_2) + \ldots + a_k^2V(X_k)\,.$$

COROLLAIRE **VARIANCE D'UNE SOMME ET D'UNE DIFFÉRENCE DE VARIABLES INDÉPENDANTES**

Soit X et Y deux variables aléatoires indépendantes. Alors

(1.20) $$V(X - Y) = V(X + Y) = V(X) + V(Y).$$

Exemple 1.21 *Variance d'une somme de variables positivement dépendantes*

Considérons une population de 4 ménages avec les caractéristiques suivantes :

		Ménage			
		1	2	3	4
Revenu (en milliers de dollars)	Époux	25	55	80	150
	Épouse	45	45	60	100

On tire un ménage au hasard. Soit X le revenu de l'époux et Y celui de l'épouse. Comparer la variance de la somme, $V(X + Y)$, à la somme des variances, $V(X) + V(Y)$.

Solution

X prend les valeurs 25, 55, 80 et 150 avec probabilité 1/4; et Y prend les valeurs 45, 60 et 100 avec probabilité 1/2, 1/4, et 1/4. On trouve alors $V(X) = 2131{,}25$; $V(Y) = 506{,}25$; et donc $V(X) + V(Y) = 2\,637{,}5$.

La variable $X + Y$ prend les valeurs 70, 100, 140, 250, chacune avec probabilité 1/4. Alors $V(X + Y) = 4\,650$. On voit donc que $V(X + Y) >> V(X) + V(Y)$, conséquence de la dépendance positive entre X et Y. Cette dépendance est clairement visible dans le tableau, où les petites valeurs de X sont accompagnées de petites valeurs de Y.

Exemple 1.22 *Variance d'une somme de variables négativement dépendantes*

Considérons une population de 4 ménages dont les revenus sont (en milliers de dollars) les suivants : {40, 50, 60, 120}.

On tire deux ménages, l'un après l'autre, sans remise. Soit X et Y les revenus des premier et deuxième ménages, respectivement.

a) Déterminer $C(X\,;\,Y)$.

b) Déterminer $V(X + Y)$ et $V(X - Y)$.

c) Soit U et W les revenus des premier et deuxième ménages, respectivement, lorsque les tirages se font avec remise. Vérifier que

i) $C(U\,;\,W) = 0$; ii) $V(U + W) > V(X + Y)$ iii) $V(U - W) < V(X - Y)$.

Solution

a) Voici l'ensemble de toutes les valeurs possible de $[X\,;\,Y]$.

x	y	x	y	x	y
40	50	50	60	60	120
40	60	50	120	120	40
40	120	60	40	120	50
50	40	60	50	120	60

Chacun de ces couples est de probabilité 1/12. Leur covariance est -322,9167.

b) Voici l'ensemble de toutes les valeurs possibles de $X + Y$ pour les 12 échantillons possibles :

		y			
		40	50	60	120
x	40	*	$X+Y=90$	$X+Y=100$	$X+Y=160$
	50	$X+Y=90$	*	$X+Y=110$	$X+Y=170$
	60	$X+Y=100$	$X+Y=110$	*	$X+Y=180$
	120	$X+Y=160$	$X+Y=170$	$X+Y=180$	*

* Impossible dans un tirage sans remise

Il y a 12 résultats possibles, chacun de probabilité 1/12 si on admet l'hypothèse d'équiprobabilité. X prend la valeur 40 dans 3 échantillons sur 12, et donc P(X = 40) = 3/12. On vérifie que chacune des 4 valeurs possibles de X est de probabilité 1/4, et par conséquent, $V(X)$ = 968,75. De même, $V(Y)$ = 968,75, et donc $V(X) + V(Y)$ = 1937,5. La distribution de la variable $Z = X+Y$ est

z	90	100	110	160	170	180
$p(z)$	1/6	1/6	1/6	1/6	1/6	1/6

On obtient alors $V(X + Y) = V(Z)$ = 1291,67 < $V(X) + V(Y)$.

On peut cependant obtenir ce résultat plus simplement par (1.19): $V(X + Y) = V(X) + V(Y) + 2C(X ; Y)$ = 968,75 + 968,75 +2(-322,9167) = 1291,67. On trouve aussi $V(X - Y) = V(X) + V(Y) - 2C(X ; Y)$ = 978,75 + 978,75 - 2(-322,9167) = 2583,333.

c) U et W sont indépendantes. Donc i) $C(U ; W) = 0$; ii) $V(U + W) = V(U) + V(W)$ = 1937,5 > 1291,67 = $V(X + Y)$; $V(U - W) = V(U) + V(W)$ = 1937,5 < 2583,333 = $V(X - Y)$.

Remarque *Covariance entre deux valeurs d'un échantillon*

La dépendance négative entre X et Y dans le dernier exemple s'explique par le fait qu'une valeur prise par X ne peut pas être prise par Y. Donc si X = 120, la plus grande valeur, Y ne peut prendre que les valeurs 40, 50 ou 60 — plutôt faibles. Inversement, si X = 40, Y aura tendance à prendre des valeurs plus élevées.

ESPÉRANCE ET VARIANCE CONDITIONNELLES

Le calcul d'espérances et de variances, qui joue un rôle important en sondages, se révèle parfois assez complexe. Soit X et Y deux variables aléatoires de fonction de probabilité conjointe $p_{X;Y}(x ; y)$ et fonction de probabilité marginale $p_{Y|x}(y)$. Alors l'espérance et la variance conditionnelles sont définies par

$$E(Y \mid x) = \sum_y y p_{Y|x}(y) \quad \text{et} \quad V(Y \mid x) = \sum_y [y - E(Y \mid x)]^2 p_{Y|x}(y) .$$

Parfois le calcul de l'espérance et de la variance d'une variable Y doit se faire en conditionnant par rapport à une deuxième variable X, c'est-à-dire de commencer par calculer $E(Y \mid X)$ et $V(Y \mid X)$. Le théorème suivant montre ce qu'on peut faire ensuite.

THÉORÈME 1.12 **CALCUL D'UNE ESPÉRANCE ET D'UNE VARIANCE PAR CONDITIONNEMENT**

(1.21) $$E(Y) = E\left[E(Y \mid X)\right].$$

(1.22) $$V(Y) = V\left[E(Y \mid X)\right] + E\left[V(Y \mid X)\right].$$

Exemple 1.23 *Calcul par conditionnement*

Voici, dans une population de 900 ménages, la distribution conjointe des deux variables suivantes :

X = nombre de chambres à coucher; Y = nombre de personnes.

Effectifs

		Nombre de personnes (y)					
		1	2	3	4	5	Total
Nombre de chambres à coucher (x)	1	50	30	20	0	0	100
	2	80	100	170	90	60	500
	3	10	30	50	90	20	200
	4	0	0	20	30	50	100
	Total	140	160	260	210	130	900

La distribution marginale de X est :

x	1	2	3	4	
$p_X(x)$	1/9	5/9	2/9	1/9	1

Le tableau suivant donne les distributions, les espérances et les variances conditionnelles.

		y						
		1	2	3	4	5	$E(Y\|x)$	$V(Y\|x)$
x	1	0,50	0,30	0,20	0	0	1,7	0,61
	2	0,16	0,20	0,34	0,18	0,12	2,9	1,49
	3	0,05	0,15	0,25	0,45	0,1	3,4	1,04
	4	0	0	0,2	0,3	0,5	4,3	0,61

Par exemple,

$E(Y \mid x = 1) = 1(0,5)+2(0,3)+3(0,2) +4(0) + 5(0) = 1,7.$

$V(Y \mid x = 1) = (1\text{-}1,7)^2(0,5) + (2\text{-}1,7)^2(0,3) + (3\text{-}1,7)^2(0,2) +(4\text{-}1,7)^2(0) + (5\text{-}1,7)^2(0) = 0,61.$

Donc

$E(Y) = E[E(Y \mid X)]$

$=E(Y \mid X = 1)p_X(1) + E(Y \mid X = 2)p_X(2) + E(Y \mid X = 3)p_X(3) + E(Y \mid X = 4)p_X(4)$

$$=1,7\frac{1}{9}+2,9\frac{5}{9}+3,4\frac{2}{9}+4,3\frac{1}{9}=\frac{91}{30}.$$

$E[V(Y \mid x)] = V(Y \mid X = 1)p_X(1) + V(Y \mid X = 2)p_X(2) + V(Y \mid X = 3)p_X(3) + V(Y \mid X = 4)p_X(4)$

$$=0,61\frac{1}{9}+1,49\frac{5}{9}+1,04\frac{2}{9}+0,61\frac{1}{9}=\frac{43}{36}.$$

$V[E(Y \mid x)]=[E(Y \mid X = 1) - 91/30]^2 p_X(1)+[E(Y \mid X = 2) - 91/30]^2 p_X(2)$

$+[E(Y \mid X = 3) - 91/30]^2 p_X(3) + [E(Y \mid X = 4) - 91/30]^2 p_X(4)$

$=187/450 = 0{,}415555.$

Alors, $V(Y) = E[V(Y \mid X)] + V[E(Y \mid X)] = 43/36 + 187/450 = 1{,}61.$

COEFFICIENT DE CORRÉLATION

Le *coefficient* de *corrélation* $\rho(X ; Y)$ entre X et Y, également désigné par ρ ou par ρ_{XY}, est défini par

$$\rho = \frac{C(X;Y)}{\sqrt{V(X)}\sqrt{V(Y)}} .$$

Exemple 1.24 *Calcul d'une covariance et d'un coefficient de corrélation*

Considérer la distribution conjointe suivante :

		y			
		0	1	2	$p_X(x)$
x	0	1/8	0	0	1/8
	1	1/8	2/8	0	3/8
	2	0	2/8	1/8	3/8
	3	0	0	1/8	1/8
	$p_Y(y)$	1/4	1/2	1/4	1

Calculer a) l'espérance de X et l'espérance de Y ; *b*) la covariance entre X et Y; c) la variance de X et la variance de Y; d) le coefficient de corrélation entre X et Y.

Solution

a) Les espérances se calculent à partir des distributions marginales :

$$E(X) = 0\frac{1}{8} + 1\frac{3}{8} + 2\frac{3}{8} + 3\frac{1}{8} = \frac{3}{2}; \; E(Y) = 0\frac{1}{4} + 1\frac{1}{2} + 2\frac{1}{4} = 1$$

b) $C(X ; Y)= E(XY) - E(X)E(Y).$

$E(XY)=0 \times 0 \times (1/8) + 0 \times 1 \times (0) + 0 \times 2 \times (0)$

$+1 \times 0 \times (1/8) + 1 \times 1 \times (2/8) + 1 \times 2 \times (0)$

$+ 2 \times 0 \times (0) + 2 \times 1 \times (2/8) + 2 \times 2 \times (1/8)$

$+3 \times 0 \times (0) + 3 \times 1 \times (0) + 3 \times 2 \times (1/8) = 2$

Donc $C(X ; Y) = E(XY) - E(X)E(Y) = 2 - (1{,}5)(1) = 0{,}5.$

c) $V(X) = E(X^2) - [E(X)]^2 = 0^2 \times (1/8) + 1^2 \times (3/8) + 2^2 \times (3/8) + 3^2 \times (1/8) - (1{,}5)^2 = 0{,}75;$

$V(Y) = E(Y^2) - [E(Y)]^2 = 0^2 \times (1/4) + 1^2 \times (1/2) + 2^2 \times (1/4) - (1)^2 = 0{,}5.$

d) $\rho = \dfrac{C(X;Y)}{\sqrt{V(X)V(Y)}} = \dfrac{0,5}{\sqrt{0,75}\sqrt{0,5}} = 0,816.$

Exemple 1.25 *Variables aléatoires indépendantes*

Considérons une population U de factures dont les montants sont distribués selon le tableau suivant :

Montant ($)	100	120	200	250	300	400	600
Fréquence	0,05	0,15	0,20	0,30	0,15	0,10	0,05

On tire avec remise 6 factures de cette population. Soit X_1, ... , X_6 les montants de ces factures.

a) Quelle est la fonction de probabilité de X_1 ?

b) Quelle est la fonction de probabilité conjointe de [X_1 ; X_2] ?

c) Quelle est la fonction de probabilité conjointe de [X_1 ; ... ; X_6] ?

Solution

a) La fonction de probabilité de X_1 est donnée précisément dans le tableau ci-dessus.

Ainsi, $p(100) = 0,05$; $p(120) = 0,15$; ... ; $p(600) = 0,05$;

b) Les variables X_1, X_2 sont indépendantes et donc $p(x_1 ; x_2) = p(x_1)p(x_2)$. Avec $x_1 = 250$ et $x_2 = 400$, par exemple, on a $p(250 ; 400) = (0,30)(0,10) = 0,03$.

c) Par la règle du produit $p(x_1 ; x_2 ; x_3 ; x_4 ; x_5 ; x_6) = p(x_1)p(x_2)p(x_3)p(x_4)p(x_5)p(x_6)$ on peut calculer la valeur de $p(x_1 ; x_2 ; x_3 ; x_4 ; x_5 ; x_6)$ pour tout six-tuplé $[x_1 ; x_2 ; x_3 ; x_4 ; x_5 ; x_6]$.

Par exemple.

$p(600 ; 200 ; 250 ; 100 ; 400 ; 200) = 0,05 \times 0,20 \times 0,30 \times 0,05 \times 0,10 \times 0,20 = 0,000\,003.$

1.5 LOIS DE PROBABILITÉ

Rappelons quelques lois classiques et leurs propriétés.

LOI BINOMIALE

Soit X le *nombre de succès* en une expérience aléatoire constituée de n épreuves indépendantes, chacune desquelles peut présenter un « succès » avec probabilité p ou un « échec » avec probabilité 1 - p.

On dit alors que X est de *loi binomiale de paramètres n* et p, ce qui s'écrit : $X \sim \mathcal{B}(n ; p)$.

THÉORÈME 1.13 **LOI BINOMIALE**

Si $X \sim \mathcal{B}(n; p)$, alors sa fonction de probabilité est donnée par

$$p(x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0, 1, \ldots, n, \text{ et}$$
$$E(X) = np, \quad V(X) = np(1-p).$$

LOI BINOMIALE NÉGATIVE

Supposons que les épreuves qui caractérisent la loi binomiale se poursuivent jusqu'à l'obtention d'un r^e succès. Si X est le nombre d'épreuves effectuées au moment où ce r^e succès est réalisé, alors X est dit de loi *binomiale négative* de paramètres r et p : $X \sim \mathcal{BN}(r\,;\,p)$.

THÉORÈME 1.14 **LOI BINOMIALE NÉGATIVE**

Si $X \sim \mathcal{BN}(r\,;\,p)$, sa fonction de probabilité est donnée par

$$p(x) = \binom{x-1}{r-1} p^r (1-p)^{x-r}, x = r, r+1, \ldots, \text{ et}$$
$$E(X) = \frac{r}{p}; \quad V(X) = \frac{r(1-p)}{p^2}.$$

Remarque

Si $r = 1$, la loi binomiale négative porte le nom de loi géométrique.

LOI HYPERGÉOMÉTRIQUE

Supposons qu'on tire *sans remise* un échantillon de taille n d'une population U de taille N et soit $U_1 \subset U$ un sous-ensemble de cardinalité N_1. Soit X le nombre d'unités de l'échantillon appartenant à U_1. Alors X est dite de loi *hypergéométrique* de paramètres n, N_1 et $N_2 = N - N_1$; on écrit $X \sim \mathcal{H}(n\,;\,N_1;\,N_2)$.

THÉORÈME 1.15 **LOI HYPERGÉOMÉTRIQUE**

Si $X \sim \mathcal{H}(n\,;\,N_1\,;\,N_2)$, alors sa fonction de probabilité est donnée par

$$p(x) = \frac{\binom{N_1}{x}\binom{N_2}{n-x}}{\binom{N}{n}} \quad 0 \le x \le N_1,\ 0 \le n-x \le N_2\ , \text{ et}$$
$$E(X) = np\,;\ V(X) = npq\frac{N-n}{N-1},\ p = \frac{N_1}{N}, q = 1-p.$$

LOI MULTINOMIALE

La loi multinomiale est une généralisation de la loi binomiale: n épreuves indépendantes donnent chacune lieu à l'un de k résultats R_1, R_2, …, R_k avec probabilités p_1, p_2, … . p_k, ($\sum_{i=1}^{k} p_i = 1$). Soit $\boldsymbol{X} = [X_1\,;\,X_2\,;\,\ldots\,;\,X_k]$ le vecteur des résultats, X_i étant le nombre de fois que le résultat R_i est apparu au cours des n essais.

THÉORÈME 1.16 **LOI MULTINOMIALE**

a) La fonction de probabilité conjointe de $\boldsymbol{X}$ est donnée par

$$p(x_1 ; x_2 ; \dots ; x_k) = P(X_1 = x_1 ; X_2 = x_2 ; \dots ; X_k = x_k) = \frac{1}{\binom{n}{x_1 ; \cdots ; x_k}} p_1^{x_1} \dots p_k^{x_k} \quad , \quad \sum_{i=1}^{k} x_i = n.$$

b) Chaque X_i est de loi binomiale de paramètres n et p_i;

c) Pour tout sous-ensemble $[X_{i_1} ; X_{i_2} ; \cdots ; X_{i_\ell}]$ $(\ell < k)$, la distribution conditionnelle du vecteur $[X_{i_1} ; X_{i_2} ; \cdots ; X_{i_\ell}]$ étant donné l'événement $\{X_{i_1} + X_{i_2} + \cdots + X_{i_\ell} = m\}$ est multinomiale de paramètres m et r_1, r_2, …, r_ℓ, où $r_j = \dfrac{p_{i_j}}{\sum_{j=1}^{\ell} p_{i_j}}$.

d) $C(X_i ; X_j) = -np_ip_j$.

LOI NORMALE

La loi normale est une loi continue caractérisée par deux paramètres, la moyenne μ et la variance σ^2. On écrit $X \sim N(\mu ; \sigma^2)$ pour signifier que X est de loi normale de moyenne μ et de variance σ^2.

On désigne par $\Phi(z)$ sa fonction de répartition :

$$\Phi(z) = P(Z \le z) .$$

THÉORÈME 1.17 **FONCTION AFFINE D'UNE NORMALE**

Si $X \sim N(\mu; \sigma^2)$ et $Y = a + bX$ alors $Y \sim N(a + b\mu ; b^2\sigma^2)$.

NOTATION ON DÉSIGNE PAR z_β LE POINT TEL QUE, SI $Z \sim N(0 ; 1)$, ALORS

(1.23) $$P(Z > z_\beta) = \beta \Leftrightarrow \beta = 1 - \Phi(z_\beta)$$

VARIABLES NORMALES INDÉPENDANTES

Une propriété importante de la loi normale est le fait que toute combinaison linéaire de variables aléatoires normales indépendantes, est normale :

THÉORÈME 1.18 **COMBINAISON LINÉAIRE DE VARIABLES NORMALES INDÉPENDANTES**

Si X_1, X_2, … , X_n sont des variables aléatoires indépendantes de loi normale, $X_i \sim N(\mu_i;\sigma_i^2)$ et a_1, a_2, … , a_n des constantes, alors

$$Y = \sum_{i=1}^{n} a_i X_i \sim N(\mu;\sigma^2) \text{ où } \mu = \sum_{i=1}^{n} a_i\mu_i \text{ et } V(Y) = \sum_{i=1}^{n} a_i^2\sigma_i^2 .$$

COROLLAIRE **SOMME DE VARIABLES NORMALES INDÉPENDANTES DE LOI NORMALE IDENTIQUE**

Soit X_1, X_2, … , X_n des variables aléatoires indépendantes de loi normale de même moyenne μ et de même variance σ^2, alors $\overline{X} \sim \mathcal{N}\left(\mu ; \frac{\sigma^2}{n}\right)$.

En échantillonnage cela veut dire que si on tire avec remise un échantillon aléatoire d'une population normale, alors la moyenne $\overline{X}$ de l'échantillon est de loi normale.

Évidemment, on ne peut pas supposer a priori qu'une population inconnue est normale. Mais dans les applications en échantillonnage, c'est la distribution de la moyenne de l'échantillon $\overline{X}$ qui compte, et celle-ci peut être supposée à peu près normale en vertu du *Théorème limite central*, dont l'un des énoncés suit.

THÉORÈME 1.19 **THÉORÈME LIMITE CENTRAL**

Soit X_1, X_2, …, X_n une suite de variables aléatoires indépendantes de même loi, de moyenne μ et de variance σ^2. Alors la variable $Z_{\overline{X}} = \frac{\overline{X} - \mu}{\sigma / \sqrt{n}}$ tend en loi vers une loi $\mathcal{N}(0 ; 1)$.

En sondages, les observations d'un échantillon ne sont pas nécessairement indépendantes, car les tirages se font généralement sans remise dans une population finie. Donc ce théorème n'est pas applicable tel quel. Mais une variante de ce théorème mène à la même conclusion.

RÉSUMÉ

- La probabilité *conditionnelle* de F étant donné E est définie par $\mathrm{P}(\mathrm{F} \mid \mathrm{E}) = \frac{\mathrm{P}(\mathrm{E} \cap \mathrm{F})}{\mathrm{P}(\mathrm{E})}$.
- Deux événements E et F sont dits *indépendants* si et seulement si $\mathrm{P}(\mathrm{E} \cap \mathrm{F}) = \mathrm{P}(\mathrm{E})\mathrm{P}(\mathrm{F})$ [voir des conditions équivalentes au-dessous de (1.3) ainsi que la Définition 1.7 pour la généralisation].
- Sous l'hypothèse d'équiprobabilité, la probabilité d'un événement E est donnée par $\mathrm{P}(\mathrm{E}) = \frac{Card(\mathrm{E})}{Card(\omega)}$.
- Le nombre de façons de tirer n objets parmi N est:
  - N^n si on tire avec remise et on tient compte de l'ordre;
  - $\mathrm{A}_n^N = \frac{N!}{(N-n)!}$ si on tire *sans* remise, et on tient compte de l'ordre;
  - $\binom{N}{n} = \frac{N!}{n!(N-n)!}$ si on tire sans remise, et on ne tient pas compte de l'ordre.

- Le nombre de façons de répartir N objets en k groupes de tailles $N_1, N_2, ..., N_k$: $\binom{N}{N_1 ; N_2 ; \ldots ; N_k} = \frac{N!}{N_1!N_2!\ldots N_k!}$.
- Espérance mathématique μ d'une variable aléatoire : $\mu = E(X) = \sum_x xp(x)$.
- Variance σ^2 d'une variable aléatoire : $\sigma^2 = V(X) = \sum_x x^2 p(x) - \mu^2 = E[(X - \mu)^2] = E(X^2) - [E(X)]^2$.
- Fonction de probabilité conjointe: $P(\boldsymbol{X} = \boldsymbol{x}) : p_{\boldsymbol{X}}(x_1; x_2; \ldots; x_n) = P(X_1 = x_1; X_2 = x_2 ; \ldots ; X_n = x_n)$.
- Fonction de probabilité *conditionnelle* de Y étant donné $X = x$: $p_{Y|x}(y) = \frac{p_{X,Y}(x ; y)}{p_X(x)}$.
- Covariance entre X et Y : $C(X ; Y) = E(XY) - E(X)E(Y)$.
- Variance d'une combinaison linéaire:

$$V\left(a_1X_1 + a_2X_2 + \ \ldots \ + a_kX_k\right) \ = \ \sum_{i=1}^{k} a_i^2 V(X_i) + \ 2\sum_i \sum_{j>i} a_i a_j C(X_i; X_j) \ .$$

Calcul d'espérances et de variance par conditionnement:

$$E(Y) \ = E\left[E\left(Y \mid X\right)\right]; \ \ V\left(Y\right) = V\left[E\left(Y \mid X\right)\right] \ + E\left[V\left(Y \mid X\right)\right] .$$

Lois discrètes [voir les Théorèmes 1.13 à 1.16].

EXERCICES

1.1 On forme au hasard un comité de 12 personnes tirées successivement dans une population de 26 employés nommés A, B, …, Z. Les personnes A, B, …, O sont des femmes, les 11 autres des hommes.

a) Supposons que le comité comprend 12 postes distincts (on les numérotera 1, 2, …, 12) et que les tirages se font avec remise (une même personne peut occuper plus d'un poste). Déterminer les probabilités des événements suivants:

i) Les postes 1, 2, …, 12 sont occupés par les employés A, B, …, L, respectivement;

ii) Le comité est constitué des personnes A, B, …, et L;

iii) A est membre du comité (et occupe possiblement plus d'un poste);

iv) A est dans le comité et n'occupe qu'un seul poste;

v) A et B occupent chacun un et un seul poste;

vi) Tous les postes sont occupés par A.

b) Les tirages se font sans remise, et une fois les membres choisis, on leur attribue les postes au hasard (un seul poste par personne). Déterminer la probabilité de chacun des événements suivants:

i) Le comité est constitué des membres A, B, C, … , et L;

ii) Les postes 1, 2, …, 12 sont occupés par les employés A, B, …, L, respectivement;

iii) Z est membre du comité;

iv) A et B sont tous deux membres du comité.

c) On tire sans remise. Déterminer la probabilité (conditionnelle) que B soit dans le comité sachant que A y est.

d) On tire avec remise.

i) Déterminer la probabilité que les postes 1 à 7 soient occupés par des femmes et les postes 8 à 12 par des hommes.

ii) Déterminer la probabilité que le comité soit composé de 3 femmes et 9 hommes.

iii) Déterminer la probabilité qu'il y ait plus d'une femme dans le comité sachant qu'il y en a déjà au moins une.

e) On tire sans remise.

i) Déterminer la probabilité que le comité soit composé de 3 femmes et 9 hommes.

ii) Déterminer la probabilité que les postes 1 à 7 soient occupés par des femmes et les postes 8 à 12 par des hommes.

iii) Déterminer la probabilité qu'il y ait plus de deux femmes dans le comité sachant qu'il y en a déjà au moins deux.

f) On tire sans remise un échantillon de n éléments d'un ensemble de N éléments.

i) Montrer que la probabilité que l'élément k (k = 1, …, N) fasse partie de l'échantillon est n/N.

ii) Montrer que la probabilité que les éléments k et ℓ (k , ℓ = 1, …, N; $k \neq \ell$) fassent partie de l'échantillon est $\frac{n(n-1)}{N(N-1)}$.

1.2 Soit X et Y deux variables aléatoires dont voici les distributions:

x	1	2	3	4	
$p(x)$	0,2	0,4	0,3	0,1	1

y	1	2	3	
$p(y)$	0,5	0,4	0,1	1

a) Déterminer $E(X)$, $E(Y)$, $V(X)$, et $V(Y)$.

b) Déterminer

i) $E(2X)$ ii) $V(3Y)$ iii) $E(X+Y)$

iv) $E(X-Y)$ v) $E(2X+3Y)$ vi) $E(2X-3Y)$.

c) Supposons que X et Y sont indépendantes. Déterminer

i) $V(X+Y)$ ii) $V(X-Y)$ iii) $V(2X-3Y)$ iv) $V(2X+3Y)$.

d) Calculer les variances demandées en c) sous l'hypothèse que le coefficient de corrélation entre X et Y est $\rho = 0{,}4$.

e) Calculer $V(X+Y)$ et $V(X-Y)$ sous l'hypothèse que $\rho = -0{,}4$. Comparer avec les résultats en c).

f) Supposons que X et Y représentent respectivement le nombre de chambres à coucher et le nombre de salles de bains dans un logement tiré au hasard dans un quartier. L'hypothèse d'indépendance est-elle vraisemblable?

g) Supposons que X est le nombre de chambres à coucher dans un logement tiré au hasard dans un quartier; et Y est le nombre de chambres à coucher dans un logement tiré au hasard dans un *autre* quartier. L'hypothèse d'indépendance est-elle vraisemblable?

1.3 On prélève un échantillon de 3 personnes (sans remise) d'une population de 8 fumeurs. On s'intéresse au nombre de cigarettes fumées par jour. Les valeurs de cette variable pour les 8 personnes de la population sont:

$$U = \{3\ ;\ 6\ ;\ 9\ ;\ 12\ ;\ 15\ ;\ 18\ ;\ 21\ ;\ 24\}.$$

La moyenne de cette population est $\mu = 13{,}5$, et sa variance est $\sigma^2 = 47{,}25$. Soit T la somme des trois observations de l'échantillon. Le tableau suivant présente la fonction de probabilité de T, qu'on peut obtenir en énumérant l'ensemble de tous les échantillons possibles (il y en a 56, et vous pourriez faire cette énumération, ou du moins une ébauche).

t	18	21	24	27	30	33	36	39	42	45	48	51	54	57	60	63
56 p(t)	1	1	2	3	4	5	6	6	6	6	5	4	3	2	1	1

a) Déterminer $P(T \leq 33)$, $P(T > 32)$, $P(T < 51)$ et $P(T \geq 54)$.

b) Déterminer l'espérance μ, la variance σ^2, et l'écart-type σ de T.

c) Déterminer l'espérance, la variance et l'écart-type de la moyenne échantillonnale $\overline{X} = T/3$.

d) Évaluer les probabilités calculées en a) sous l'hypothèse que $\overline{X}$ est de loi normale. L'approximation normale vous semble-t-elle adéquate ici?

d) Quelle est la probabilité qu'il y ait plus de 2 points d'écart entre $\overline{X}$ et μ?

e) Déterminer $P\left(\bar{X} - \mu > 1{,}96\sqrt{V(\bar{X})}\right)$ et $P\left(|\bar{X} - \mu| > 1{,}96\sqrt{V(\bar{X})}\right)$ (sous l'hypothèse que $\bar{X}$ est de loi normale ces probabilités sont, respectivement, de 0,025 et 0,05).

1.4 D'une population de citoyens dont 50 % sont en faveur d'un certain projet de loi, on tire un échantillon de taille 8. Soit X le nombre de personnes dans l'échantillon qui sont en faveur du projet. La fonction de probabilité de X est donnée dans le tableau suivant :

x	0	1	2	3	4	5	6	7	8	
$256p(x)$	1	8	28	56	70	56	28	8	1	1

a) Vérifier les valeurs de la fonction de probabilité (X est de loi binomiale).

b) Déterminer l'espérance, la *variance*, et l'écart-type de X.

c) Déterminer l'espérance, la variance et l'écart-type de la *proportion* échantillonnale $Y = X/8$.

1.5 Les revenus annuels (en dizaines de milliers de dollars) de quatre personnes, Agnès, Barbie, Carole et Diane sont, respectivement, 3, 4, 6 et 7. On choisit deux personnes au hasard parmi ces 4. Soit X le revenu moyen des 2 personnes tirées. Déterminer, à l'aide d'une énumération complète de tous les échantillons possibles, la fonction de probabilité, l'espérance et la variance de X.

1.6 On tire avec remise 2 appartements parmi 4. Le nombre d'occupants dans les appartements est $\{1, 2, 4, 5\}$. Soit X le nombre de personnes dans le premier appartement tiré et Y le nombre de personnes dans le deuxième.

a) Soit Z le nombre total de personnes dans les deux appartements. Déterminer la fonction de probabilité de Z.

b) À partir de la fonction de probabilité trouvée en a), déterminer $E(Z)$ et $V(Z)$.

c) Calculer $E(X)$ et $E(Y)$, et vérifiez que leur somme est égale à $E(Z)$.

d) Calculer $V(X)$ et $V(Y)$, et vérifiez que leur somme est égale à $V(Z)$.

e) Déterminer la fonction de probabilité de $U = XY$, et calculer $E(XY)$. Est-ce que $E(XY) = E(X) \times E(Y)$?

f) Déterminer la fonction de probabilité de $D = X - Y$. Vérifiez que $E(X - Y) = E(X) - E(Y)$ et que $V(X - Y) = V(X) + V(Y)$.

1.7 Supposez, au numéro précédent, que vous tirez les deux appartements *sans* remise.

a) Déterminer la fonction de probabilité de Z, le nombre total d'occupants dans les deux appartements.

b) Calculer la variance de $Z = X + Y$. Pouvez-vous expliquer pourquoi $V(Z)$ ici est inférieure à $V(Z)$ au numéro précédent?

c) Calculer la variance de $D = Y - X$. Pouvez-vous expliquer pourquoi $V(D)$ ici est supérieure à $V(Z)$ au numéro précédent ?

d) Vérifiez que $E(XY) \neq E(X) \times E(Y)$.

1.8 Le prix X d'une action de A et le prix Y d'une action de B ont chacun la même espérance mathématique et la même variance. Les prix des deux actions sont des variables indépendantes. Qu'est-ce qui est le plus risqué? Acheter **100** actions de A ou acheter **50** actions de A et **50** de B. Déterminer la réponse sous chacune des hypothèses suivantes:

a) X et Y sont indépendantes;

b) Le coefficient de corrélation entre X et Y est $\rho = 0{,}4$ [$C(X\,;\,Y) = \rho\sqrt{V(X)}\sqrt{V(Y)}$];

c) Le coefficient de corrélation entre X et Y est $\rho = -0.3$.

1.9 Considérez la population suivante : $U = \{3\,;\ 5\,;\ 6\,;\ 8\,;\ 10\}$. On tire un échantillon de deux éléments de cette population, *sans* remise. Soit X le résultat du premier tirage et Y le résultat du deuxième. Voici l'ensemble des valeurs de $(X\,;\,Y)$, ainsi que les valeurs possibles des variable $U = XY$ et $W = (X + Y)/2$:

$(x\,;\,y)$	u	w	$(x\,;\,y)$	u	w	$(x\,;\,y)$	u	w
3 ; 5	15	4	5 ; 10	50	7,5	8 ; 6	48	7
3 ; 6	18	4,5	6 ; 3	18	4,5	8 ; 10	80	9
3 ; 8	24	5,5	6 ; 5	30	5,5	10 ; 3	30	6,5
3 ; 10	30	6,5	6 ; 8	48	7	10 ; 5	50	7,5
5 ; 3	15	4	6 ; 10	60	8	10 ; 6	60	8
5 ; 6	30	5,5	8 ; 3	24	5,5	10 ; 8	80	9
5 ; 8	40	6,5	8 ; 5	40	6,5			

a) Déterminer les fonctions de probabilité de $U = XY$ et de W.

b) Calculer $E(U) = E(XY)$. Est-ce que $E(XY) = E(X)\,E(Y)$?

c) Calculer $V(W)$. Quelle aurait été cette variance si les tirages avaient été effectués *avec* remise ?

Les numéros 1.10 à 1.17 portent sur une population constituée de 12 logements réunis en trois quartiers, A, B et C. Voici les données concernant le nombre de logements (X) dans les quartiers et le nombre de personnes (Y) dans les logements :

Quartier	X	Désignation des unités	Valeurs de Y
A	4	a_1 a_2 a_3 a_4	1 2 4 5
B	5	b_1 b_2 b_3 b_4 b_5	2 4 5 6 8
C	3	c_1 c_2 c_3	1 4 7

1.10 On tire au hasard un logement dans la population entière. Soit Y le nombre de personnes dans le logement sélectionné.

a) Déterminer $E(Y)$ et $V(Y)$ à partir de la distribution ci-dessus;

b) Déterminer $E(Y \mid X)$ pour chaque valeur de X;

c) Déterminer $V(Y \mid X)$ pour chaque valeur de X;

d) Déterminer $V[E(Y \mid X)]$;

e) Déterminer $E[V(Y \mid X)]$;

f) Déterminer $E(Y)$ à l'aide de la formule (1.9);

g) Déterminer $V(Y)$ à l'aide de la formule (1.22).

1.11 Au numéro 1.10, supposons qu'on procède différemment: on commence par choisir un quartier au hasard, puis on tire un logement dans le quartier sélectionné. Répondre aux questions b) à g) du numéro 1.10.

1.12 On tire au hasard un quartier (avec probabilités égales), puis un logement dans le quartier sélectionné. Déterminer :

a) $P(Y = 2 \mid X = 4)$ et $P(Y = 2)$.

b) $E(Y)$ et $V(Y)$. Vérifier que $E(Y) \neq \mu_y$, le nombre moyen de personnes par logement dans la population.

c) Considérer la variable Z, définie en fonction de Y comme ceci:

$$Z = \begin{cases} 4Y/4 & \text{si le logement est dans le quartier A} \\ 5Y/4 & \text{si le logement est dans le quartier B} \\ 3Y/4 & \text{si le logement est dans le quartier C} \end{cases}$$

Déterminer $E(Z)$ et $V(Z)$. Vérifier que $E(Z) = \mu_y$, le nombre moyen de personnes par logement dans la population.

1.13 On tire au hasard un quartier avec probabilités proportionnelles au nombre de logements: donc on sélectionne A avec probabilité 4/12, B avec probabilité 5/12, et C avec probabilité 3/12. Puis on tire au hasard un logement dans le quartier sélectionné. Soit W le nombre de personnes dans le logement choisi. Déterminer $E(W)$ et $V(W)$. Vérifier que $E(W) = \mu_y$. le nombre moyen de personnes par logement dans la population.

1.14 On tire au hasard deux quartiers, sans remise, puis 2 logements dans chacun des quartiers sélectionnés.

a) Déterminer la probabilité que l'échantillon soit constitué des unités a_1, a_2, b_1 et b_2.

b) Comparer cette probabilité à la probabilité du même événement sous l'hypothèse que les 4 logements ont été sélectionnés sans remise dans l'ensemble de 12 logements.

1.15 On tire au hasard un logement dans chaque quartier. Soit Y_1, Y_2 et Y_3 les valeurs de Y obtenues dans les quartiers A, B, et C, respectivement.

a) Soit $\bar{Y} = \frac{1}{3}(Y_1 + Y_2 + Y_3)$. Déterminer $E(\bar{Y})$ et $V(\bar{Y})$.

b) Soit $\bar{Y}^* = \frac{4}{12}Y_1 + \frac{5}{12}Y_2 + \frac{3}{12}Y_3$. Déterminer $E(\bar{Y}^*)$ et $V(\bar{Y}^*)$. Vérifier que $E(\bar{Y}^*) = \mu_y$, le nombre moyen de personnes par logement dans la population.

1.16 On tire au hasard un échantillon de 6 logements parmi les 12 logements (avec probabilités égales).

a) Soit n_1, n_2 et n_3 le nombre de logements dans l'échantillon appartenant aux quartiers A, B et C, respectivement ($n_1 + n_2 + n_3 = 6$). Déterminer la probabilité $P(n_1 = n_2 = n_3)$

i) lorsque les tirages se font avec remise;

ii) lorsque les tirages se font sans remise.

b) Si les tirages se font sans remise, déterminer la probabilité que l'échantillon soit constitué des unités a_1, a_2, b_1, b_2, c_1, et c_2.

c) Si les tirages se font sans remise, déterminer la probabilité conditionnelle que l'échantillon soit constitué des unités a_1, a_2, b_1, b_2, c_1, et c_2 étant donné que $n_1 = n_2 = n_3$.

1.17 Supposons que pour obtenir un échantillon de taille 6, on tire au hasard deux unités (sans remise) dans chacun des trois quartiers (donc $n_1 = n_2 = n_3$). Déterminer la probabilité que l'échantillon soit constitué des unités a_1, a_2, b_1, b_2, c_1, et c_2. Constatez qu'elle est égale à la probabilité calculée au numéro 1.16-c).

1.18 Soit X une variable de loi hypergéométrique de paramètres n, N_1 et N_2. Montrer que lorsque $N_1 \to \infty, N_2 \to \infty$ de telle sorte que $N_1/N = p$ ($N = N_1 + N_2$), alors la fonction de probabilité de X tend vers celle d'une variable $\mathcal{B}(n; p)$. Ce résultat confirme que si N est grand, il importe peu que les tirages soient effectués avec ou sans remise.

[*Suggestion*: Montrer que la fonction de probabilité d'une variable de loi hypergéométrique peut s'écrire comme: $\frac{\binom{N_1}{x}\binom{N_2}{n-x}}{\binom{N}{n}} =$

$$\frac{\binom{n}{x}}{(1-1/N)\ldots(1-(n-1)/N)} \times \frac{N_1}{N}\left(\frac{N_1}{N} - \frac{1}{N}\right)\ldots\left(\frac{N_1}{N} - \frac{x-1}{N}\right) \quad .]$$

1.19 Soit $\boldsymbol{X} = [X_1 ; X_2 ; \ldots ; X_n]$ un vecteur aléatoire de loi multinomiale de paramètres n et $[p_1 ; p_2 ; \ldots ; p_n]$. Montrer que la loi conditionnelle de $\boldsymbol{X}_2 = [X_2 ; X_3 ; \ldots ; X_n]$ étant donné $X_1 = x_1$ est multinomiale de paramètres $n - x_1$ et $\left[\frac{p_2}{1-p_1}; \frac{p_3}{1-p_1}; \cdots; \frac{p_n}{1-p_1}\right]$.

1.20 Dans chacun des numéros suivants, on décrit une expérience et deux variables aléatoires, X et Y. Dites si la covariance $C(X ; Y)$ est positive, négative ou nulle. Justifiez votre réponse.

a) On tire au hasard un ménage dans une population de ménages.

X : Le revenu familial;

Y : La superficie du logement.

b) Une ville est divisée en deux secteurs, Est et Ouest. On tire au hasard un logement dans chaque secteur.

X: La valeur de la propriété tirée dans le secteur Est;

Y: La valeur de la propriété tirée dans le secteur Ouest.

c) On tire un échantillon de 200 personnes d'une grande population.

X: Les frais médicaux encourus par les femmes de l'échantillon au courant de l'année écoulée;

Y: Les frais médicaux encourus par les hommes durant la même période.

d) On interroge au hasard un échantillon de 50 citoyens (dans une très grande population).

X : Le nombre de répondants favorables au plan d'austérité du gouvernement;

Y : Le nombre de répondants qui s'y objectent.

e) On tire au hasard et avec remise deux ménages dans une très grande population :

X : Le revenu familial du premier ménage;

Y : La *moyenne* des revenus des deux ménages.

f) On tire au hasard un quartier dans une ville, puis deux maisons, avec remise, dans le quartier sélectionné.

X : La valeur de la première maison tirée;

Y : La valeur de la deuxième maison tirée.

1.21 Pour chacune des paires de variables aléatoires X et Y, dites si d'après vous $V(X) > V(Y)$ ou si $V(X) < V(Y)$.

a) X : La valeur d'une habitation tirée au hasard dans un quartier huppé;

Y : La valeur d'une habitation tirée au hasard dans la ville entière.

b) X : La proportion d'objets défectueux dans un échantillon de 10 objets tirés d'une certaine population;

Y : La proportion d'objets défectueux dans un échantillon de 100 objets tirés de la même population.

c) X : Le nombre d'objets défectueux dans un échantillon de 10 objets tirés sans remise d'une certaine population;

Y : Le nombre d'objets défectueux dans un échantillon de 10 objets tirés avec remise de la même population;

d) X : Le revenu moyen de 10 familles choisies au hasard dans une population;

Y : Le revenu moyen de 100 familles choisies au hasard dans la même population.

e) X : La taille moyenne de deux personnes choisies indépendamment dans une population;

Y : La taille moyenne de deux jumeaux identiques choisis dans une population de jumeaux.

f) R_1 et R_2 sont les revenus de deux personnes tirées au hasard dans une grande population.

$X = R_1 + R_2$ lorsque les deux personnes sont tirées au hasard dans la population entière;

$Y = R_1 + R_2$ lorsque les deux personnes sont tirées au hasard dans un même quartier, ce quartier ayant été tiré préalablement au hasard (on suppose que les quartiers sont très grands).

g) On tire sans remise un échantillon de **5** personnes dans une salle contenant N personnes dont une certaine proportion p sont des fumeurs

X : Nombre de fumeurs dans l'échantillon lorsque $N = 10$;

Y : Nombre de fumeurs dans l'échantillon lorsque $N = 30$.

2 Notions de statistique

Nous rappelons ici certaines notions de la statistique classique — essentiellement celles liées à l'estimation de paramètres — dont la théorie des sondages s'inspire. Quelques ajustements seront nécessaires afin de les adapter au domaine des sondages, mais nous commençons par les présenter sous leur forme classique, car elles seront malgré tout omniprésentes dans la suite.

POPULATION

Traditionnellement, le terme « population » est défini comme une fonction de répartition $F(x \mid \boldsymbol{\theta})$ qui dépend d'un *paramètre* θ ou d'un vecteur $\boldsymbol{\theta} = [\theta_1 ; \theta_2 ; \ldots ; \theta_k]$ de paramètres. Par exemple les revenus des ménages d'une ville sont généralement résumés par un vecteur $\boldsymbol{\theta} = [\theta_1 ; \theta_2] = [\mu ; \sigma^2]$, où μ est la moyenne et σ^2 la variance; et F est la fonction de répartition des revenus (une loi normale, par exemple). Cette définition s'applique également en sondages, sauf qu'en sondages on ne fait pas nécessairement d'hypothèse concernant la loi de la variable: aucune forme particulière F n'est postulée.

Échantillon aléatoire simple

En statistique classique, un *échantillon aléatoire simple* est défini comme une suite de n variables aléatoires $X_1, \ldots, X_n$ indépendantes et identiquement distribuées.

Remarque

Ces variables représentent les valeurs observées d'une variable X lors de n tirages successifs dans une population. Les X_i sont de même loi parce que les tirages sont tous effectués dans la même population : si X_i est le revenu du i^e ménage choisi, alors, pour tout i, la distribution de X_i est précisément la distribution des valeurs de la population. L'hypothèse d'indépendance découle du fait que les tirages sont effectués dans des conditions qui ne changent pas selon les résultats des tirages précédents. L'hypothèse d'indépendance est vérifiée lorsque les tirages sont effectués avec remise. Nous ne pourrons maintenir cette hypothèse, car s'il s'agit d'une population finie, et si les tirages se font *sans* remise, alors les X_i ne sont plus indépendantes.

2.1 STATISTIQUES ET ESTIMATEURS

Soit X_1 , … , X_n un échantillon aléatoire simple provenant d'une population $F(x \mid \theta)$. Toute variable aléatoire $T(X_1; \ldots ; X_n)$ fonction de $X_1, \ldots, X_n$, est appelée une *statistique*. Toute statistique $\hat{\theta} = \hat{\theta}(X_1 ; X_2 ; \ldots ; X_n)$ destinée à estimer un paramètre θ est appelée *estimateur* de θ.

ESTIMATEURS SANS BIAIS

L'une des qualités généralement souhaitées d'un estimateur est celui d'être sans biais :

DÉFINITION 2.1 ESTIMATEUR SANS BIAIS

Un estimateur $\hat{\theta}$ est dit sans biais pour θ si

$$E(\hat{\theta}) = \theta ,$$

quelle que soit la valeur de θ.

DÉFINITION 2.2 LE BIAIS D'UN ESTIMATEUR

Le *biais* d'un estimateur $\hat{\theta}$ est défini par

$$\mathrm{B}(\hat{\theta}) = E(\hat{\theta}) - \theta .$$

Un estimateur biaisé peut être acceptable si le biais tend à disparaître lorsque l'échantillon est grand.

DÉFINITION 2.3 ESTIMATEUR ASYMPTOTIQUEMENT SANS BIAIS

Un estimateur $\hat{\theta}$ est dit *asymptotiquement sans biais* pour θ si

$$\lim_{n \to \infty} \mathrm{B}(\hat{\theta}) = 0.$$

VARIANCE D'UN ESTIMATEUR

Un estimateur sans biais n'est pas nécessairement bon. La moyenne $\overline{X}$ d'un échantillon aléatoire simple X_1, …, X_n est un estimateur sans biais de la moyenne μ, ce qui est une bonne chose. Mais ce seul critère ne suffit pas à le distinguer d'autres estimateurs, également sans biais mais nettement moins bons, comme, par exemple, la moyenne $(X_1 + X_2)/2$ des deux premières observations.

L'absence de biais signifie que l'estimateur a tendance à prendre la valeur juste, *en moyenne*. Mais cela ne l'empêche pas de s'en éloigner. Ce qui compte surtout, c'est que l'estimateur ait tendance à rester proche de μ: qu'il ne s'en éloigne pas trop et pas souvent. Une façon de mesurer cette tendance, c'est la variance de l'estimateur, $V(\hat{\theta})$: on la souhaite *faible*.

ERREUR QUADRATIQUE MOYENNE

Lorsque l'estimateur $\hat{\theta}$ n'est pas sans biais, l'écart entre $\hat{\theta}$ et θ est mesuré non pas par $V(\hat{\theta})$ mais par l'*erreur quadratique moyenne*, désignée par $EQM(\hat{\theta})$.

DÉFINITION 2.4 ERREUR QUADRATIQUE MOYENNE

L'erreur quadratique moyenne $EQM(\hat{\theta})$ d'un estimateur $\hat{\theta}$ est définie par

$$EQM(\hat{\theta}) = E[(\hat{\theta} - \theta)^2] .$$

Le théorème suivant identifie deux composantes de l'erreur quadratique moyenne : la variance et le carré du biais.

THÉORÈME 2.1 **DÉCOMPOSITION DE L'ERREUR QUADRATIQUE MOYENNE**

L'erreur quadratique moyenne se décompose en deux parties, comme ceci:

$$EQM(\hat{\theta}) = V(\hat{\theta}) + [B(\hat{\theta})]^2 .$$

Normalement, un estimateur biaisé est acceptable dans la mesure où la deuxième composante est faible par rapport à la première. Plus précisément, il faudrait que le rapport $\frac{B(\hat{\theta})}{\sqrt{V(\hat{\theta})}}$ tende vers 0 lorsque *n* tend vers l'infini.

COEFFICIENT DE VARIATION

Le coefficient de variation d'une variable est le rapport de l'écart-type sur sa moyenne. Quand il s'agit d'un estimateur $\hat{\theta}$ sans biais, le coefficient de variation $cv(\hat{\theta}) = \frac{\sqrt{V(\hat{\theta})}}{E(\hat{\theta})}$ relativise l'écart-type, ce qui le rend préférable à l'écart-type comme mesure descriptive. Le coefficient de variation a l'avantage aussi d'être indépendant de l'unité de mesure de $\hat{\theta}$.

2.2 INTERVALLES DE CONFIANCE

Un *intervalle de confiance* à 100(1 - α) % pour un paramètre θ est un intervalle aléatoire, $I(\omega) = [LI(\omega) ; LS(\omega)]$, fonction de l'échantillon ω, tel que $P[I(\omega) \ni \theta] \geq 1 - \alpha$. L'exemple le plus simple concerne l'estimation de la moyenne μ d'une population normale de moyenne μ et de variance σ^2, σ connu. La statistique $Z = \frac{\bar{X} - \mu}{\sqrt{V(\bar{X})}}$ est alors de loi $N(0 ; 1)$, où $V(\bar{X}) = \frac{\sigma^2}{n}$. Alors il existe un nombre $z_{\alpha/2}$ tel que $P\left(-z_{\alpha/2} \leq \frac{\bar{X} - \mu}{\sqrt{V(\bar{X})}} \leq z_{\alpha/2} \right) = 1 - \alpha$, où $z_{\alpha/2}$ est défini par (1.23).

L'événement $\left\{-z_{\alpha/2} \le \frac{\bar{X}-\mu}{\sqrt{V(\bar{X})}} \le z_{\alpha/2}\right\}$ est équivalent à $\left\{\bar{X} - z_{\alpha/2}\sqrt{V(\bar{X})} \le \mu \le \bar{X} + z_{\alpha/2}\sqrt{V(\bar{X})}\right\}$, et donc l'intervalle $\left[\bar{X} - z_{\alpha/2}\sqrt{V(\bar{X})}\ ;\ \bar{X} + z_{\alpha/2}\sqrt{V(\bar{X})}\right]$ contient μ avec probabilité 1 - α. C'est ce qui en fait un intervalle de confiance à 100(1-α) %, appelé le *niveau* de l'intervalle de confiance. Cet intervalle ne peut pas être calculé car en général la variance $V(\bar{X})$ n'est pas connue. On remplace donc $V(\bar{X})$ par une estimation $\hat{V}(\bar{X})$, et la statistique Z par $\hat{Z} = \frac{\bar{X}-\theta}{\sqrt{\hat{V}(\bar{X})}}$. $\hat{Z}$ ne suit pas une loi normale, mais sous l'hypothèse que la population est normale, elle suit une loi connue, la loi de Student, et cela permet de déterminer un intervalle de confiance.

Dans cet ouvrage, on ne fera pas d'hypothèse sur la nature de la population. On se fiera plutôt au théorème limite central. Dans la plupart des cas, le point de départ est une statistique de la forme $\hat{Z} = \frac{\hat{\theta}-\theta}{\sqrt{\hat{V}(\hat{\theta})}}$, où $\hat{V}(\hat{\theta})$ est un estimateur de $V(\hat{\theta})$. Nous déterminerons un intervalle de confiance par la formule

$$\left[\hat{\theta} - z_{\alpha/2}\sqrt{\hat{V}(\hat{\theta})}\ ;\ \hat{\theta} + z_{\alpha/2}\sqrt{\hat{V}(\hat{\theta})}\right],$$

basée sur l'hypothèse que $\hat{Z}$ est à peu près de loi normale centrée-réduite, ce qui est justifié si n est grand.

2.3 LA TECHNIQUE DE LINÉARISATION

Le Théorème 1.9 montre comment calculer la variance d'une fonction linéaire $\sum_{i=1}^{n} a_i X_i$ de variables aléatoires : $V(\sum_{i=1}^{k} a_i X_i) = \sum_{i=1}^{k} a_i^2 V(X_i) + 2\sum_i \sum_{j>i} a_i a_j C(X_i; X_j)$. Pour une fonction *non linéaire* $Y = f(X_1; X_2 ; \ldots ; X_k)$, il n'existe pas de formule simple pour calculer $V(Y)$. Cependant, lorsqu'un estimateur $\hat{\theta}$ est une fonction $\hat{\theta} = f(\hat{\theta}_1; \hat{\theta}_2; \cdots; \hat{\theta}_k)$ de statistiques convergentes[2], il existe une technique dite de *linéarisation* qui consiste à approcher f par une fonction linéaire, soit

$$\hat{\theta} \approx \tilde{\theta} = f(\theta_1; \theta_2; \ldots; \theta_k) + \sum_{i=1}^{k} a_i(\hat{\theta}_i - \theta_i),$$

où $a_i = \frac{\partial f(\theta_1; \theta_2; \ldots; \theta_k)}{\partial \theta_i}$.

[2] L'estimateur $\hat{\theta}$ d'un paramètre θ est dit *convergent* si pour tout $a > 0$, $P(|\hat{\theta} - \theta| > a)$ tend vers 0 lors que n tend vers l'infini.

On a alors

(2.1) $$E(\hat{\theta}) \approx E(\tilde{\theta}) = f(\theta_1; \theta_2; \ldots; \theta_k) + \sum_{i=1}^{k} a_i E(\hat{\theta}_i - \theta_i) = f(\theta_1; \theta_2; \ldots; \theta_k),$$

et

(2.2) $$V(\hat{\theta}) \approx V(\tilde{\theta}) = \sum_{i=1}^{k} a_i^2 V(\hat{\theta}_i) + 2\sum_i \sum_{j>i} a_i a_j C(\hat{\theta}_i ; \hat{\theta}_j).$$

On estimera $V(\hat{\theta})$ en remplaçant chaque paramètre inconnu par son estimation: θ_i par $\hat{\theta}_i$, a_i par $\hat{a}_i = \dfrac{\partial f(\hat{\theta}_1; \hat{\theta}_2; \ldots; \hat{\theta}_k)}{\partial \hat{\theta}_i}$; et on estimera la variance de $\hat{\theta}$ par

$$\hat{V}(\hat{\theta}) = \sum_{i=1}^{k} \hat{a}_i^2 \hat{V}(\hat{\theta}_i) + 2\sum_i \sum_{j>i} a_i a_j \hat{C}(\hat{\theta}_i; \hat{\theta}_j),$$

où $\hat{V}(\hat{\theta}_i)$ et $\hat{C}(\hat{\theta}_i ; \hat{\theta}_j)$ sont les estimateurs des variances $V(\hat{\theta}_i)$ et des covariances $C(\hat{\theta}_i ; \hat{\theta}_j)$.

Bien que la qualité de ces approximations ne soit pas rigoureusement assurée par les théorèmes limites connus, les études empiriques tendent à montrer que si l'échantillon est assez grand, les approximations sont satisfaisantes.

2.4 PROPRIÉTÉS DE QUELQUES ESTIMATEURS

Des sondages ont été conçus afin d'estimer toute une gamme de paramètres. Mais trois paramètres ont une importance particulière: une moyenne μ; une proportion p; et une variance σ^2. Les théorèmes suivants présentent des estimateurs sans biais et déterminent leur variance, toujours sous l'hypothèse que les observations X_1, X_2, …, X_n sont indépendantes.

Estimateur sans biais d'une moyenne μ

THÉORÈME 2.2 **ESTIMATION DE *M***

Soit X_1, …, X_n un échantillon aléatoire simple provenant d'une population de moyenne μ. Alors $\bar{X} = \dfrac{1}{n}\sum_{i=1}^{n} X_i$ est un estimateur sans bais de μ et $V(\bar{X}) = \dfrac{\sigma^2}{n}$.

Estimateur sans biais d'une proportion p

Un cas particulier important est celui où la variable X est *dichotomique*, c'est-à-dire, une variable dont les seules valeurs sont 0 ou 1. Une proportion p est la moyenne d'une variable dichotomique. Supposons qu'on tire, avec remise, un échantillon de taille n d'une grande population afin d'estimer la proportion p des unités de la population qui appartiennent à une certaine classe C. La situation peut être ramenée à celle du Théorème 2.2 en définissant X_i comme une variable dichotomique qui prend la valeur 1 si la i^e unité tirée appartient à la classe C, 0 sinon. La moyenne de la population est alors la proportion p, et son estimateur

$\overline{X}$ est la proportion $\hat{p}$ d'unités de l'*échantillon* qui appartiennent à la classe *C*. Le théorème suivant énonce les propriétés de cet estimateur

THÉORÈME 2.3 **ESTIMATION D'UNE PROPORTION**

Soit p la proportion des unités d'une population qui appartiennent à une certaine classe *C*, et $\hat{p}$ la proportion d'unités dans l'échantillon qui appartiennent à la classe *C*. Alors

$$\hat{p} \text{ est un estimateur sans bais de } p \text{ et } V(\hat{p}) = \frac{p(1-p)}{n} .$$

Ces conclusions découlent du Théorème 1.13.

Estimateur sans biais d'une variance

THÉORÈME 2.4 **ESTIMATION D'UNE VARIANCE**

Soit X_1, X_2, … , X_n un échantillon aléatoire provenant d'une population de variance σ^2. Alors

$$s^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1} \tag{2.3}$$

est un estimateur sans bais de σ^2.

On remarque que l'estimateur intuitif $\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n}$ n'est pas sans biais. Son biais est $B(\hat{\sigma}^2) = B\left[\frac{n-1}{n}s^2\right] = E\left[\frac{n-1}{n}s^2 - \sigma^2\right] = \frac{n-1}{n}\sigma^2 - \sigma^2 = -\frac{\sigma^2}{n}$. Le bais est donc négligeable lorsque n est grand et tend vers 0 lorsque n tend vers l'infini.

ÉCHANTILLONNAGE ET SONDAGES

L'objet de ce livre est l'échantillonnage. Or le terme « échantillonnage » pourrait désigner presque toute la science statistique, puisque ses préoccupations — une recherche scientifique en laboratoire, un sondage d'opinion, une étude des relations entre variables, l'analyse d'une série chronologique — font presque toujours intervenir la notion d'échantillonnage. Nous allons donc désigner par le terme « sondage » le type d'échantillonnage particulier dont il sera question ici.

Qu'est-ce qui distingue notre objet du reste de la statistique?

La théorie des sondages s'inspire largement des résultats de la théorie statistique classique résumée ci-dessus, mais elle se distingue de celle-là sur certains points importants.

Il y a d'abord le fait qu'un échantillon ne peut pas ici être défini comme *une suite de variables indépendantes*. Lorsqu'un échantillon est tiré sans remise d'une population finie, le résultat d'un tirage dépend de ceux qui l'ont précédé. De ce fait, les théorèmes énoncés ci-dessus devront être modifiés.

Une différence plus fondamentale, c'est le fait qu'en sondages une bonne partie du développement théorique — et des applications qui en découlent — peut se faire sans l'appui d'un modèle. Alors que dans la plupart des branches de la statistique classique, les modèles, au contraire, jouent un rôle fondamental. Pour illustrer: en régression linéaire simple, les observations $Y_1, Y_2, \ldots, Y_n$, sont supposées indépendantes, de loi normale, de même variance σ^2 et de moyennes $E(Y_i) = \beta_o + \beta x_i$. $i = 1 \ldots\ldots n$. D'importants théorèmes concernant les estimateurs des paramètres β_o, β_1 et σ^2 découlent de ces hypothèses et ne pourraient être démontrées sans elles.

En sondages, par contre, plusieurs techniques utiles peuvent être développées et appliquées à partir d'un ensemble minimal de postulats. Le point de départ: une population constituée de N unités, et un échantillon. On peut imposer une structure à la population : la répartir en sous-populations; ou rassembler ses unités élémentaires (des personnes, par exemple) en groupes (des ménages). Ces seuls éléments suffisent à déterminer des estimateurs et établir certaines de leurs propriétés. On exige seulement que le tirage des unités se fasse selon un mécanisme aléatoire bien défini.

Il est vrai que l'absence de modèle a un prix. Un sérieux handicap est la difficulté à connaître la distribution d'un estimateur $\hat{\theta}$: À moins de connaître la loi de $\hat{\theta}$, ou de pouvoir l'approcher, on ne peut construire d'intervalle de confiance. En statistique classique, lorsqu'on suppose que la population est normale, on peut conclure que certaines statistiques suivent une loi de Student, d'autres une loi khi-deux, d'autres encore une loi de Fisher. En sondages, à défaut d'une hypothèse concernant la forme de la population, on recourt à des théorèmes limites, en particulier un théorème limite central applicable à un échantillon tiré sans remise d'une population finie (Hajek, 1960). Son énoncé exact est complexe, mais ses implications en sondages se résument à peu près à ceci : si n et N sont grands et n/N n'est pas trop proche de 1, alors la moyenne d'un échantillon simple est à peu près de loi normale. Notre estimateur sera généralement une moyenne ou une fonction de plusieurs moyennes, ce qui nous permettra d'agir comme s'il était de loi normale et éventuellement de construire un intervalle de confiance, à condition que n soit grand. Il n'est pas facile de justifier cette procédure à partir des théorèmes asymptotiques, puisqu'il n'est jamais évident qu'un échantillon est assez « grand ». Nous reviendrons à ces questions dans le cadre de problèmes spécifiques. Pour l'instant, on note seulement que les expériences empiriques — des simulations par ordinateur — faites à date donnent des résultats plutôt encourageants.

L'utilisation de modèles en sondages n'est pas exclue. Au contraire, l'appel aux modèles est de plus en plus courant dans les recherches actuelles. Mais l'approche sans modèle est amplement suffisante pour les besoins des problèmes traités dans ce livre.

2.5 CLASSES DE SONDAGES

Les facultés universitaires qui portent un intérêt aux sondages — sociologie, sciences politiques, gestion, comptabilité — classent généralement les sondages en deux catégories: les sondages *empiriques* et les sondages *probabilistes*. Les firmes de sondages aussi utilisent cette classification dans leur offre de services.

Sondages empiriques

Les sondages empiriques choisissent les unités de façon « raisonnée », généralement motivés par souci d'économie ou pour des raisons de facilité. Ils comprennent certains sondages aux objectifs vastes et vagues, des sondages exploratoires auprès de populations dont on connaît peu les caractéristiques. Supposons qu'on tire un échantillon des comptes de dépenses soumis par les représentants d'une compagnie afin de « se faire une idée de la pertinence et du caractère raisonnable » des réclamations. On peut difficilement préciser ce qu'on cherche: on cherche des dérogations (aux règles écrites comme au bon sens), certes, mais on ne les connaît pas toutes, on ne les a pas préalablement répertoriées. On peut noter des abus imprévus; se poser de nouvelles questions suscitées par les résultats; ou faire des constatations qualitatives, comme « M. Untel pourrait choisir des restaurants moins chers ». Le choix des comptes peut avoir été motivé par la facilité: les documents à portée de la main; les comptes du mois courant; une poignée de documents tirés ici et là dans un tiroir. Ou encore, on peut avoir volontairement favorisé certaines classes de comptes: ceux des derniers représentants engagés; des représentants des régions éloignées; de ceux soupçonnés de dépenses abusives, etc.

Certaines approches, mêmes structurées, souffrent du même inconvénient. C'est le cas des sondages d'opinions dans lesquels les sujets sont invités à répondre à des questions posées aux internautes ou aux lecteurs d'un magazine. L'échantillonnage dit « de jugement », qui donne à l'enquêteur une certaine latitude dans le choix des répondants — en principe afin d'obtenir un échantillon « représentatif » — ne peut non plus être traité rigoureusement. L'échantillonnage dit « par quota » est également problématique : il consiste à poursuivre les tirages jusqu'à ce qu'on obtienne un certain « quota » de personnes appartenant à certaines classes (des femmes, des citadins, des professionnels, des immigrants, etc.).

Finalement, il y les panels ou *focus groups*, de plus en plus populaires et offerts dans la plupart des firmes de sondages.

Échantillonnage probabiliste

L'échantillonnage empirique peut être parfaitement acceptable pour certaines fins, mais il ne peut pas mener à des conclusions quantitatives et objectivement défendables. La validité des conclusions quantitatives (« 20 % des montants réclamés sont inadmissibles »), en particulier, est très douteuse. Seul un sondage probabiliste permet de déterminer une marge d'erreur et de justifier une conclusion de façon objective.

La possibilité de déterminer une marge d'erreur est particulièrement importante en cas de litige. Par exemple, un échantillon tiré d'un lot de pièces fabriquées afin d'estimer le taux de défectuosité doit être crédible aux deux partis: le fournisseur craint le risque d'une surestimation, et le client le risque d'une sous-estimation. Les partis peuvent alors s'entendre sur une marge d'erreur raisonnable, à condition que celle-ci soit calculable. Même problématique en vérification quand une liste de réclamations soumise par une compagnie (pour remboursement d'une taxe, par exemple) doit être vérifiée par le ministère du Revenu.

Dans un sondage probabiliste, les unités de l'échantillon sont sélectionnées selon un mécanisme aléatoire clairement défini. Le mécanisme peut varier, mais quel qu'il soit, il doit permettre de calculer la probabilité qu'une unité donnée soit sélectionnée, ainsi que la probabilité qu'une paire d'unités soit sélectionnée. Dans ce cas seulement est-il possible de déterminer une marge d'erreur. Il y a plusieurs modes de tirage possibles, et la façon de procéder a un effet crucial sur la suite — sur la façon d'estimer les paramètres, sur la qualité des estimations et sur la façon d'en évaluer la précision. Nous allons donc, dans cet ouvrage, développer en détail plusieurs des modes d'échantillonnage les plus courants — mais toujours probabilistes.

2.6 POPULATION ET PARAMÈTRES

En gros, un projet de sondage comporte les étapes suivantes:

- Identifier clairement la population;
- Identifier le ou les paramètres à estimer;
- Décider d'une façon de tirer l'échantillon et de sa taille;
- Décider comment estimer le ou les paramètres;
- Analyser les résultats: estimer les paramètres et en évaluer la fiabilité.

Une population, dans ce manuel, représentera généralement un ensemble fini et concret: l'ensemble de comptes soumis pour remboursement de taxes; l'ensemble des employés d'une compagnie; l'ensemble des comptes de dépenses soumis par les représentants d'une compagnie; l'ensemble des ménages d'une ville; l'ensemble des entreprises agricoles d'un territoire, etc.

Certaines des caractéristiques d'une population se résument par un *paramètre*. Les paramètres suivants seront traités dans ce livre:

- La *moyenne* d'une population : par exemple, le revenu moyen des ménages d'un quartier ;
- Le *total* d'une population : par exemple, la *production totale* de blé dans les fermes d'une région;
- Une *proportion* : par exemple, la proportion des employés d'une compagnie qui seraient favorables à un plan de soins dentaires ;
- Un *effectif* : par exemple, le *nombre* d'employés favorables à un plan de soins dentaires ;
- Un *quotient* : par exemple, la production moyenne de blé par acre d'une population de fermes.
- Un *écart-type* : par exemple, la dispersion des salaires des médecins.

Bien que peu nombreux, ces paramètres font l'objet de la très grande majorité des études conduites dans l'industrie, le commerce, ou les gouvernements.

Une fois l'objectif précisé, il faut élaborer un *plan d'échantillonnage*, c'est-à-dire, déterminer les étapes principales du projet, depuis la quantification des objectifs jusqu'à l'énoncé des conclusions. Une des premières décisions à prendre consiste à fixer un *mode de sélection* ou *mode d'échantillonnage*. Une autre consiste à décider de la façon d'estimer les paramètres — car un même paramètre peut être estimé de plusieurs façons.

2.7 BASE DE SONDAGE

Le mode d'échantillonnage est choisi en fonction des informations disponibles sur la population ciblée. Ces informations constituent ce qu'on appelle une *base de sondage*. Minimalement, une base de sondage comprend une liste des unités de la population. Par exemple, un vérificateur dont le mandat porte sur une population de transactions doit disposer de la liste des transactions, et doit s'assurer que les documents à l'origine des comptes sélectionnés seront accessibles. Il existe aussi des bases de sondage dites *aréolaires*, constituées d'unités géographiques. Pour effectuer un sondage auprès des ménages d'une ville, on peut commencer par dresser une liste des quartiers de la ville, délimités sur une carte.

Outre cette liste qui permet d'identifier et éventuellement d'accéder aux unités de la population, une base de sondages peut comprendre plusieurs autres éléments d'information utiles à l'élaboration d'un plan de sondages. En général, la population visée est rarement totalement inconnue. Certaines données peuvent être même connues pour la population entière. Par exemple, si une liste est constituée de régions géographiques urbaines, certaines informations sur ces régions peuvent avoir été recueillies lors d'un recensement pour toutes les régions de la population: le nombre de ménages, la superficie de la région, etc. Une telle information est précieuse, et un plan de sondage qui les met à contribution peut se révéler particulièrement efficace. Toute information issue d'un sondage antérieur (de la population ciblée elle-même, ou d'une autre lui ressemblant) peut aussi servir à cette fin. Elle peut aider à planifier un sondage futur: à choisir un mode d'échantillonnage, à suggérer des estimateurs, et à déterminer la taille de l'échantillon qui assurera un niveau de précision acceptable.

2.8 MODES D'ÉCHANTILLONNAGE

Le mode de sondage choisi a un effet sur les propriétés d'un estimateur, lequel peut être bon sous un mode d'échantillonnage et mauvais sous un autre. Chaque mode d'échantillonnage entraîne un choix distinct d'estimateurs.

Mathématiquement, ce qui caractérise un mode d'échantillonnage, c'est la distribution des probabilités $p(\omega)$ des éléments de l'espace échantillon Ω. C'est à partir de ces probabilités, qui doivent être calculables, que sont développées les propriétés des estimateurs.

Dans ce livre, nous examinerons les modes d'échantillonnage les plus courants, soit :

- L'échantillonnage aléatoire simple;
- L'échantillonnage stratifié;
- L'échantillonnage systématique;
- L'échantillonnage par grappes.

Un projet d'échantillonnage ne se limite pas nécessairement à un seul de ces quatre modes d'échantillonnage. Il peut s'appuyer sur une combinaison de ces méthodes.

ÉCHANTILLONNAGE ALÉATOIRE SIMPLE

L'échantillonnage aléatoire simple consiste à sélectionner un échantillon de n éléments de telle sorte que tout échantillon possible soit aussi probable que tout autre.

Exemple 2.1 *Échantillonnage aléatoire simple*

À partir d'une liste des N employés d'une compagnie, on prélève un échantillon de taille n en tirant successivement et sans remise n employés. Le mécanisme de sélection doit être tel que, au premier tirage, chaque employé a une chance sur N d'être sélectionné; au deuxième, chacun des N-1 qui restent a une chance sur (N-1) d'être sélectionné; ainsi de suite. L'échantillon qui en résulte est un échantillon aléatoire simple.

L'échantillonnage aléatoire simple est un mode élémentaire, employé parfois comme tel, mais plus souvent comme partie d'un mode d'échantillonnage plus complexe. En particulier, il fait souvent partie d'un échantillon stratifié.

ÉCHANTILLONNAGE STRATIFIÉ

Il est naturel, parfois, de partitionner une population avant de l'échantillonner. Par exemple, un sondage auprès des étudiants d'une université peut procéder par faculté (appelées, dans le langage technique, des *strates*): tous les étudiants appartiennent à une et une seule faculté. On tire alors un échantillon dans chaque faculté — aléatoire simple, si l'on veut, mais pas nécessairement. Ensuite, on réunit les résultats obtenus dans les facultés pour obtenir une estimation globale des paramètres de la population entière. Ce mode d'échantillonnage, appelé *échantillonnage par strates*, est recommandé parfois pour des raisons purement administratives. Dans une entreprise avec une structure régionale, les autorités locales sont mieux équipées pour prélever des données au niveau de leur région. L'administration d'un sondage est donc gérée localement d'abord: les responsables locaux tirent un échantillon dans leur région, puis le siège social se charge de les amalgamer en un seul échantillon.

La simple réunion des échantillons n'est plus un échantillon aléatoire simple et un estimateur qui est bon pour un échantillon aléatoire simple ne convient plus. Il faut donc développer des estimateurs propres à ce mode d'échantillonnage.

Une autre raison peut motiver le recourt à un échantillon stratifié: la précision des estimateurs. Dépendant de la façon dont la population est stratifiée et des tailles des échantillons tirés dans les strates, les estimateurs qui en résultent peuvent être plus efficaces qu'un échantillon aléatoire simple: on peut atteindre une plus grande précision avec les mêmes ressources financières.

ÉCHANTILLONNAGE SYSTÉMATIQUE

L'échantillonnage systématique est un mode d'échantillonnage intuitivement attrayant et dans certains cas particulièrement facile à exécuter. Il est donc très répandu. Il s'applique lorsque les unités de la population sont rangées dans un certain ordre : on identifie une 1re

unité, une 2e, …, une Ne . Par exemple, les spectateurs d'un film défilant à la sortie; les livres d'une bibliothèque rangés sur des tablettes; les fiches remplies par les passagers d'un avion à l'arrivée à l'aéroport; les plants de tomates dans une rangée.

Un *échantillonnage systématique* consiste à prélever les unités à des intervalles réguliers, ceci à partir d'un point de départ aléatoire.

Exemple 2.2 *Échantillonnage systématique*

Considérons une population de 1000 patients dont les dossiers sont rangés dans des tiroirs en ordre alphabétique. Pour prélever un échantillon de 20 dossiers, on tire systématiquement chaque 50e dossier dans les tiroirs à partir d'un dossier choisi au hasard parmi les 50 premiers dossiers. Si on tombe sur le 31e dossier, par exemple, l'échantillon sera constitué des 31e, 81e, 131e,…, et 981e dossier.

Un attrait de l'échantillonnage systématique, outre une certaine facilité d'exécution, c'est qu'il semble intuitivement prometteur : il donne l'assurance que chaque partie de la population sera représentée. Son intérêt est clair s'il s'agit, par exemple, d'échantillonner des plants de tomates alignées en rangées : un échantillon systématique comprendra nécessairement des unités provenant des différentes parties du terrain. C'est l'un des attraits de la méthode.

La méthode souffre, cependant, d'un inconvénient majeur : un échantillon systématique ne peut fournir d'information permettant d'évaluer la précision d'un estimateur. Donc pas de marge d'erreur, et pas d'intervalle de confiance.

ÉCHANTILLONNAGE PAR GRAPPES

Les unités d'une population appartiennent souvent à des groupes naturels : chaque ménage se trouve dans un quartier; chaque cégépien est inscrit à un cégep; chaque électeur est inscrit dans une circonscription. Ces groupements fournissent un moyen de sélectionner des unités lorsqu'un échantillonnage aléatoire simple se révèle impossible ou trop coûteux. Dans le cas des ménages, un tirage aléatoire simple n'est possible que si on dispose d'une liste de tous les ménages de la population. Il peut aussi être coûteux : si les ménages sélectionnés doivent être rencontrés personnellement, on a intérêt, une fois dans un quartier, à sélectionner plusieurs ménages du quartier afin de minimiser les coûts de déplacement.

Pour ces raisons, purement pratiques, on peut juger préférable de prélever un échantillon de *quartiers* (ou cégeps, ou circonscriptions) plutôt qu'un échantillon de *ménages*. Les unités sélectionnées sont les quartiers, appelées *grappes* ou *unités primaires*. Les ménages sont alors les *sous-unités* ou *unités secondaires*.

C'est dans le contexte d'un tirage par grappes que certaines nouvelles questions surgissent.

Probabilités d'inclusion égales ou inégales? Étant donné que les grappes peuvent être de tailles très variables, il est naturel de vouloir accorder plus d'importance aux plus grandes grappes, par exemple en donnant aux grandes grappes une plus forte probabilité d'inclusion (dans l'échantillon).

Tirage avec remise ou sans remise? Bien qu'un tirage avec remise soit forcément moins efficace qu'un tirage sans remise, il n'est pas pour autant exclu d'emblée. Car il peut être beaucoup plus facile à exécuter et les estimateurs qui en découlent (ainsi que tous les calculs subséquents) plus simples à effectuer. Lorsque le nombre de grappes dans la population est grand, le fait de tirer avec remise ne cause pas une grande perte d'efficacité.

Échantillonnage au second degré? Supposons que les unités primaires sont des quartiers et les ménages sont les sous-unités. Dans un quartier sélectionné, est-ce qu'on interroge *tous* les ménages qui s'y trouvent? Ou est-ce qu'on se contente d'un *échantillon* de ménages dans le quartier? Dans ce deuxième cas, on parle d'un échantillonnage *à deux degrés*.

On voit bien que le nombre de possibilités est quasiment illimité : les ménages peuvent être sélectionnés de plusieurs façons: par un échantillon aléatoire simple, stratifié, par grappes encore. Tout cela est possible.

Taille fixe ou aléatoire? La plupart des modes de sélection discutés dans ce livre supposent une taille d'échantillon fixe. Mais certaines techniques d'échantillonnage mènent à un échantillon de taille aléatoire. Nous les décrirons brièvement.

2.9 ESTIMATEURS

Étant donné un mode d'échantillonnage, on cherchera d'abord un estimateur sans biais. S'il n'y en pas, on se contentera d'un estimateur asymptotiquement sans biais. S'il en existe plus d'un (ce qui est toujours possible) on tentera de choisir « le meilleur », celui dont la variance (ou l'erreur quadratique moyenne, pour un estimateur biaisé) est la plus faible. La tâche n'est pas facile. La formule de la variance peut être déterminée, mais en général elle dépend d'un ou de plusieurs paramètres et ceux-là sont inconnus. Néanmoins, les formules de variance permettent de voir quels facteurs contribuent à la variabilité de l'estimateur, et ces facteurs peuvent être évalués intuitivement ou à l'aide d'informations provenant d'autres sondages.

Une difficulté en sondage tient justement à la mesure de la qualité d'un estimateur, fondée essentiellement sur la variance: un bon estimateur est un estimateur sans biais avec une petite variance. Mais fondamentalement, si on s'intéresse à la variance, c'est parce qu'elle reflète (ou est censée refléter) ce qui nous intéresse réellement: de savoir si l'erreur d'estimation a de bonnes chances de rester petite. Ce lien entre la variance et la probabilité d'erreur permet de déterminer un intervalle de confiance, et d'affirmer que l'erreur est contenue dans un certain intervalle (la marge d'erreur) avec probabilité de 95 % (disons). Le problème, c'est que ce lien est déterminé sous l'hypothèse que l'estimateur est de loi normale. Ce qui est vrai (à peu près) quand l'échantillon est assez grand. Mais le sens de « assez grand » varie selon la forme de la distribution des données de la population. Si cette distribution est à peu près symétrique, un échantillon de taille modeste est « assez grand ». Mais dans certains domaines, entre autres en vérification comptable et en finance, la population est souvent fortement asymétrique et dans ce cas la distribution de l'estimateur le sera aussi — à moins que l'échantillon soit extrêmement grand. Des solutions (ou

palliatifs) existent, mais elles varient selon le contexte. On les indiquera dans les situations qui s'y prêtent.

Pour chaque paramètre et chaque mode d'échantillonnage, nous suivrons à peu près la même démarche : nous proposerons des estimateurs raisonnables, nous nous assurerons qu'ils sont sans biais (ou du moins, *approximativement* sans biais); nous développerons la formule (parfois approximative) de sa variance ou de son erreur quadratique moyenne (EQM); et enfin, nous proposerons un estimateur de l'EQM.

RÉSUMÉ

- Le *biais* d'un estimateur $\hat{\theta}$ est défini par $\mathrm{B}(\hat{\theta}) = E(\hat{\theta}) - \theta$.
- Un estimateur $\hat{\theta}$ est dit *sans biais* pour θ si $\mathrm{B}(\hat{\theta}) = 0$ et *asymptotiquement sans biais* si $\lim_{n\to\infty} \mathrm{B}(\hat{\theta}) = 0$.
- L'erreur quadratique moyenne $EQM(\hat{\theta})$ d'un estimateur $\hat{\theta}$ est définie par $EQM(\hat{\theta}) = E[(\hat{\theta} - \theta)^2]$
- L'erreur quadratique moyenne admet la décomposition suivante: $EQM(\hat{\theta}) = V(\hat{\theta}) + [B(\hat{\theta})]^2$, et donc se confond avec la variance si $\hat{\theta}$ est sans biais.
- S'il est raisonnable de supposer que $\hat{Z} = \dfrac{\hat{\theta} - \theta}{\sqrt{\hat{V}(\hat{\theta})}}$ est à peu près de loi $\mathcal{N}(0 ; 1)$, un intervalle de confiance approximatif de niveau 100(1-α)% est $\left[\hat{\theta} - z_{\alpha/2}\sqrt{\hat{V}(\hat{\theta})} \ ; \ \hat{\theta} + z_{\alpha/2}\sqrt{\hat{V}(\hat{\theta})}\right]$.
- Soit $X_1, \ldots, X_n$ n variables aléatoires indépendantes de même loi, de moyenne μ et de variance σ^2. Alors
  - $\bar{X} = \dfrac{1}{n}\sum_{i=1}^{n} X_i$ est un estimateur sans bais de μ et $V(\bar{X}) = \dfrac{\sigma^2}{n}$, et
  - $s^2 = \dfrac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$ est un estimateur sans bais de σ^2.
- La proportion échantillonnale $\hat{p}$ est un estimateur sans bais de la proportion p de la population et sa variance est $V(\hat{p}) = \dfrac{p(1-p)}{n}$.

EXERCICES

2.1 Considérons une population de 9 appartements divisée en deux quartiers. Le nombre Y d'occupants dans les appartements est comme suit:

Quartier 1: {1 ; 2 ; 2 ; 3}

Quartier 2: {2 ; 4 ; 4 ; 4 ; 6}

On tire au hasard deux unités dans chaque quartier (sans remise). Soit $\bar{y}_1$ et $\bar{y}_2$ les moyennes ; et s_1^2 et s_2^2 les variances des échantillons (équation (2.3)).

a) Voici la liste des valeurs possibles des quartiers 1 et 2. Vérifiez les résultats obtenus.

Quartier 1				Quartier 2			
Valeurs	$\overline{y}_1$	s_1^2	Prob.	Valeurs	$\overline{y}_2$	s_2^2	Prob.
1 ; 2	1,5	0,5	4/12	2 ; 4	3	2	6/20
1 ; 3	2	2	2/12	2 ; 6	4	8	2/20
2 ; 2	2	0	2/12	4 ; 4	4	0	6/20
2 ; 3	2,5	0,5	4/12	4 ; 6	5	2	6/20

b) À partir du tableau ci-dessus, vérifier les distributions suivantes:

Quartier 1

$\overline{y}_1$	1,5	2	2,5	s_1^2	0	0,5	2
$p(\overline{y}_1)$	4/12	4/12	4/12	$p(s_1^2)$	2/12	8/12	2/12

Quartier 2

$\overline{y}_2$	3	4	5	s_2^2	0	2	8
$p(\overline{y}_2)$	6/20	8/20	6/20	$p(s_2^2)$	6/20	12/20	2/20

c) Déterminer la distribution de l'estimateur $\hat{\mu} = \frac{\overline{y}_1 + \overline{y}_2}{2}$ de la moyenne μ de la population.

d) Déterminer le biais de $\hat{\mu}$.

e) Déterminer l'erreur quadratique moyenne de $\hat{\mu}$,

f) Déterminer la distribution de la moyenne pondérée $\hat{\mu}_p = \frac{4\overline{y}_1 + 5\overline{y}_2}{9}$.

g) Montrer que $E(\hat{\mu}_p) = \mu$. [Noter que si le résultat en c) démontre que $\hat{\mu}$ n'est pas sans biais, le cas particulier traité ici ne démontre pas que $\hat{\mu}_p$ est sans biais. Pour cela il faudrait montrer que l'égalité $E(\hat{\mu}_p) = \mu$ est vérifiée pour toutes les valeurs possibles de μ.]

h) Déterminer la variance de $\hat{\mu}_p$.

i) Vérifier que $E\left(0,25s_1^2\right) = V(\overline{y}_1)$ et que $E(0,3s_2^2) = V(\overline{y}_2)$.

j) Utiliser résultat i) pour déterminer un estimateur sans biais de la variance de $\hat{\mu}_p$.

k) Montrer que $\frac{1}{18}(\overline{y}_1 - \overline{y}_2)$ est un estimateur sans biais du biais $\mathrm{B}(\hat{\mu})$.

2.2 Au numéro précédent, supposons qu'on prélève un échantillon de la façon suivante: On tire au hasard l'un des deux quartiers (avec probabilités ½ chacun). Ensuite on tire au hasard avec remise 4 appartements dans le quartier sélectionné. Soit Z une variable aléatoire qui prend la valeur 1 ou 2 selon que le quartier sélectionné est le quartier 1 ou 2. Soit $\bar{y}$ la moyenne des 4 appartements sélectionnés.

a) On propose $\hat{\mu} = \bar{y}$ comme estimateur de μ, la moyenne de la population. Déterminer $E(\hat{\mu} \mid Z = 1)$ et $E(\hat{\mu} \mid Z = 2)$.

b) Déterminer $E(\hat{\mu})$ et le biais de $\hat{\mu}$.

c) Déterminer $V(\hat{\mu} \mid Z = 1)$ et $V(\hat{\mu} \mid Z = 2)$.

d) Déterminer $V(\hat{\mu})$.

e) Considérer un estimateur qui pondère $\bar{y}$ selon le quartier: $\hat{\mu}_{sb} = \begin{cases} a\bar{y} & \text{si Z=1} \\ b\bar{y} & \text{si Z=2} \end{cases}$.
Déterminer a et b de telle sorte que $\hat{\mu}$ soit sans biais. [a = 8/9, b = 10/9].

f) Déterminer $V(\hat{\mu}_{sb})$.

2.3 Au numéro précédent, supposons que le tirage du quartier est effectué avec probabilités proportionnelles au nombre d'appartements, c'est-à-dire, les quartiers 1 et 2 sont sélectionnés avec probabilités 4/9 et 5/9, respectivement.

a) Montrer qu'alors $\hat{\mu} = \bar{y}$ est sans biais ($\bar{y}$ est la moyenne de l'échantillon tiré dans le quartier sélectionné).

b) Déterminer $V(\hat{\mu})$.

2.4 Soit x_1, x_2, …., x_n n valeurs d'une variable dichotomique et soit p la proportion de ces nombres qui égalent 1. Montrer que la variance $s^2 = \dfrac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$ peut s'écrire comme $\dfrac{n}{n-1}p(1-p)$.

2.5 Soit x_1, x_2, …, x_n et y_1, y_2, …, y_n les valeurs de deux variables dichotomiques. Montrer que la *covariance* $s_{xy} = \dfrac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$ est égale à $\dfrac{n}{n-1}(p_{XY} - p_X p_Y)$, où p_X est la proportion des x_i qui également 1, p_Y la proportion des y_i qui également 1, et p_{xy} la proportion des produits $x_i\, y_i$ qui égalent 1.

2.6 Soit X une variable aléatoire dichotomique. Soit p = P(X = 1). Montrer que $E(X) = p$ et $V(X) = p(1\text{-}p)$.

2.7 Admettons que $Z = \dfrac{\hat{\theta} - \theta}{\sqrt{\hat{V}(\hat{\theta})}}$ est à peu près de loi $N(0\,;\,1)$. Montrer que les deux intervalles suivants sont des intervalles à 90 % (à peu près):

$$\left[\hat{\theta} - 2{,}05\sqrt{\hat{V}(\hat{\theta})}\ ;\ \hat{\theta} + 1{,}41\sqrt{\hat{V}(\hat{\theta})}\right];\ \left[\hat{\theta} - 1{,}64\sqrt{\hat{V}(\hat{\theta})}\ ;\ \hat{\theta} + 1{,}64\sqrt{\hat{V}(\hat{\theta})}\right].$$

Quel critère ferait préférer l'un plutôt que l'autre?

2.8 Démontrer le Théorème 2.1.

2.9 Démontrer le Théorème 2.2.

2.10 Démontrer le Théorème 2.3.

2.11 Démontrer le Théorème 2.4.

2.12 Soit $\hat{\theta}^2$ un estimateur sans biais de θ^2. Montrer que $\hat{\theta}$ ne peut pas être sans biais pour θ (à moins qu'il soit de variance nulle) et que son biais est négatif. On peut donc en déduire que s sous-estime σ.

2.13 Soit $X_1, X_2, \ldots, X_n$ un échantillon aléatoire simple (les X_i supposées indépendantes) d'une population de moyenne μ. Les techniques pour estimer μ, estimer sa variance et déterminer un intervalle de confiance ont été exposées dans ce chapitre — tout cela sans postuler de modèle quelconque. Mais un modèle, s'il est crédible, peut améliorer la qualité des estimations. Supposons, par exemple, que chaque X_i est de loi de Poisson de paramètre μ [la fonction de probabilité d'une variable de loi de Poisson est $p(x) = \frac{e^{-\mu}\mu^x}{x!}$, $x = 0, 1, 2, \ldots$; $E(X) = V(X) = \mu$].

a) Montrer que $\bar{X}$ et S^2 sont tous deux des estimateurs sans biais de μ.

b) Montrer que la loi conditionnelle de X_i étant donné $\Sigma X_i = k$ est une loi binomiale de paramètres k et $1/n$.

c) Montrer que $E(X_i \mid \bar{X}) = \bar{X}$, $V(X_i \mid \bar{X}) = \frac{(n-1)\bar{X}}{n}$, et que $E(X_i^2 \mid \bar{X}) = \frac{(n-1)\bar{X}}{n} + \bar{X}^2$.

d) Montrer que $E(S^2 \mid \bar{X}) = \bar{X}$.

e) Montrer que $V(S^2) \geq V(\bar{X})$ [*Suggestion* : Calculer $V(S^2)$ en conditionnant par rapport à $\bar{X}$ selon (1.22)].

3 Échantillonnage aléatoire simple et estimation d'une moyenne

Ce chapitre est fondamental, bien que son objet puisse sembler, à première vue, plutôt restreint — l'estimation d'une moyenne, un paramètre particulier; dans un échantillon aléatoire simple, un mode de tirage particulier. Mais sa portée est en fait beaucoup plus large. C'est de toute la problématique d'estimation qu'il sera question ici. Certes, la moyenne d'une population n'est qu'un paramètre parmi d'autres, mais les notions fondamentales sur lesquelles son estimation repose sont communes à tout problème d'estimation. Elles seront revisitées à plusieurs reprises dans les chapitres suivants. Le mode de tirage aussi — l'échantillon aléatoire simple — est un mode de tirage parmi d'autres; mais il constitue la substruction des méthodes qui seront discutées aux trois prochains chapitres. On y reviendra donc régulièrement.

C'est ici aussi que nous fixerons le vocabulaire et définirons la notation particulière qui sera utilisée tout le long de cet ouvrage.

3.1 ÉCHANTILLON ALÉATOIRE SIMPLE

Considérons une base de sondage rudimentaire constituée d'une simple liste des N unités d'une population U:

$$U = \{u_1 ; u_2 ; \ldots ; u_N\}.$$

On tire un échantillon ω de taille n à l'aide d'un mécanisme de sélection au hasard. Soit

Ω = l'ensemble de tous les échantillons possibles.

Supposons que le mécanisme utilisé est tel que tout sous-ensemble de U de taille n est susceptible d'être sélectionné. Donc la cardinalité de Ω est

$$Card(\Omega) = \binom{N}{n}.$$

Le mode d'échantillonnage traité exclusivement dans ce chapitre est l'échantillonnage aléatoire simple, ou *éas*. En voici la définition formelle.

DÉFINITION 3.1 ÉCHANTILLON ALÉATOIRE SIMPLE

Un *échantillon aléatoire simple* (*éas*) de taille n issu d'une population de N unités est un échantillon tiré de telle sorte que tout ensemble de n unités a la même probabilité de constituer l'échantillon.

Donc pour tout $\omega \in \Omega$.

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

Remarques

Un mécanisme de sélection

La définition formelle d'un *éas* énonce les propriétés de l'échantillon mais ne dicte pas de mécanisme de sélection, c'est-à-dire, elle ne prescrit pas une procédure concrète pour le réaliser. Il en existe plusieurs, dont en voici une : on numérote de 1 à N les N unités de la population, ainsi que N boules qu'on place dans une urne. Ensuite on tire au hasard, successivement et *sans remise*, n boules dans l'urne. L'échantillon est l'ensemble des unités portant les numéros des boules tirées.

Évidemment, ce n'est pas la façon dont on procède en pratique: des logiciels informatiques peuvent être programmés pour simuler cette procédure.

Une définition opérationnelle d'un éas

On peut toujours définir un *éas* en fonction d'une procédure de sélection particulière, comme celle que nous venons de décrire: un *éas* est la suite des résultats de n tirages effectuées sans remise d'une population finie. Toute procédure qui équivaudrait à celle-là serait également considérée comme un *éas*. Cette définition opérationnelle approche la définition classique, selon laquelle un échantillon aléatoire simple est une suite de n variables aléatoires indépendantes et identiquement distribuées. La différence ici est que les tirages sont présumés effectués sans remise, donc dépendants.

3.2 PARAMÈTRES DE LA POPULATION

Soit Y une variable quantitative définie sur U: elle fait correspondre une valeur numérique y_k à l'unité k, $k = 1, \ldots, N$. Deux mesures statistiques seront évoquées régulièrement dans la suite: la moyenne μ et l'écart-type S. Lié de près à la moyenne est le *total* τ de la population. Voici leur définition formelle.

DÉFINITION 3.2 TOTAL, MOYENNE, ÉCART-TYPE DE LA POPULATION

Le total τ, la moyenne μ et l'écart-type S des valeurs $y_1, y_2, \ldots, y_N$ d'une population sont définies par

(3.1) $$\tau = \sum_{k=1}^{N} y_k \ , \ \mu = \frac{1}{N}\sum_{k=1}^{N} y_k \text{ et } S = \sqrt{\frac{\sum_{k=1}^{N}(y_k - \mu)^2}{N-1}} .$$

Remarque

Ces paramètres sont définis sur un ensemble de nombres. On se permettra néanmoins de parler de la moyenne de la population ou de l'écart-type de la population. On comprendra qu'il s'agit de la moyenne et de l'écart-type des valeurs d'une variable associée à la population.

Des quantités analogues sont définies pour les données de l'échantillon.

DÉFINITION 3.3 MOYENNE ET ÉCART-TYPE DE L'ÉCHANTILLON

Soit ω un *éas* tiré dans une population *U*. La moyenne $\bar{y}$ et l'écart-type *s* de l'échantillon sont définies par:

(3.2) $$\bar{y} = \frac{1}{n}\sum_{i\in\omega} y_i \text{ et } s = \sqrt{\frac{\sum_{i\in\omega}(y_i - \bar{y})^2}{n-1}} .$$

Remarques *Termes et notation*

La notation « $\Sigma_{i\in A}$ » désigne une somme prise sur toutes les données d'un ensemble A.

Pour simplifier la notation, nous identifions l'unité u_k par son indice, *k*. Ainsi donc, $k \in \omega$ est une façon abrégée d'écrire $u_k \in \omega$.

À l'occasion, il sera utile d'accoler le suffixe « ω » et désigner la moyenne de l'échantillon par $\bar{y}(\omega)$, pour souligner le fait qu'il s'agit de la moyenne des unités de ω. Mais c'est la notation allégée $\bar{y}$ qui sera privilégiée, tant qu'elle ne risque pas de causer d'ambiguïté.

Le terme « variance » est ambigu car il désigne aussi bien une mesure concrète comme S^2 que la variance $V(X)$ d'une variable aléatoire *X*. En général, le contexte rendra le sens clair.

3.3 ESTIMATION D'UNE MOYENNE

L'estimateur naturel de la moyenne μ de la population est la moyenne $\bar{y}$ de l'échantillon. Le prochain théorème affirme que $\bar{y}$ est sans biais et établit sa variance.

THÉORÈME 3.1 **ESPÉRANCE ET VARIANCE DE** $\bar{y}$

Soit $\bar{y}$ la moyenne d'un *éas* d'une population de taille *N*, de moyenne μ et d'écart-type S. Alors

$$E(\bar{y}) = \mu \text{ et } V(\bar{y}) = (1-f)\frac{S^2}{n} .$$

DÉFINITION 3.4 FRACTION D'ÉCHANTILLONNAGE ET FACTEUR DE CORRECTION

Le rapport $f = \frac{n}{N}$ est appelé fraction d'échantillonnage et le facteur (1-*f*) est appelé facteur de correction pour population finie ou facteur de correction tout court.

Remarques

Une inspection de la formule $V(\bar{y}) = (1-f)\frac{S^2}{n}$ permet d'identifier les trois facteurs qui contribuent à la variance de l'estimateur.

La taille de l'échantillon

Du fait que *n* est au dénominateur, la variance est d'autant plus petite que *n* est grand, ce qui ne fait que quantifier une évidence: plus l'échantillon est grand, mieux c'est.

L'écart-type S de la population

S est un autre facteur déterminant : l'estimateur est d'autant meilleur que *S* est petit. Ce qui se comprend aussi : si les données de la population sont très rapprochées de leur moyenne (*S* petit), la moyenne de l'échantillon tend à l'être aussi.

Le facteur de correction

Finalement, le facteur de correction joue un rôle qui reflète la taille de la population par rapport à celle de l'échantillon. Son effet est, la plupart du temps, relativement mineur, la fraction d'échantillonnage $f = n/N$ étant généralement faible.

Tirages avec remise

Si les tirages sont effectués avec remise, le facteur de correction est égal à 1, et *S* est remplacée par σ, où $\sigma = \sqrt{\frac{\sum_{k=1}^{N}(y_i-\mu)^2}{N}} = \sqrt{\frac{N-1}{N}}S$. La variance de $\bar{y}$ est $\frac{\sigma^2}{n} = \left(1-\frac{1}{N}\right)\frac{S^2}{n}$, comparée à $\left(1-\frac{n}{N}\right)\frac{S^2}{n}$ pour un tirage sans remise.

Estimation de la variance

Pour estimer sans biais la variance $V(\bar{y}) = (1-f)\frac{S^2}{n}$, il suffit de remplacer S^2 par une estimation sans biais. Un estimateur sans biais de S^2 est donné dans le théorème suivant :

THÉORÈME 3.2 **ESTIMATEUR DE S^2**

La statistique

$$s^2 = \frac{\sum_{i\in\omega}(y_i - \bar{y})^2}{n-1}$$

est un estimateur sans biais de S^2.

COROLLAIRE **ESTIMATEUR DE LA VARIANCE DE $\bar{y}$**

Un estimateur sans biais de la variance $V(\bar{y})$ est donné par

$$\hat{V}(\bar{y}) = (1-f)\frac{s^2}{n} \ .$$

Exemple 3.1 *Estimation d'une moyenne*

Un vérificateur tire un échantillon de 40 factures parmi les 1 200 factures émises par un restaurant au courant de l'année. Le but est d'estimer le montant moyen de la facture. Voici les montants des 40 factures de l'échantillon :

20,90	24,46	34,22	45,46	52,18	69,76	80,98	90,72	113,04	144,23
21,72	25,75	38,85	48,86	62,57	70,83	83,66	95,77	129,81	144,86
21,82	31,70	39,55	51,29	65,81	78,89	83,97	98,66	136,78	148,52
23,80	31,99	44,86	51,43	67,37	79,52	89,65	100,40	143,50	148,79

Estimer la moyenne de la population et l'écart-type de l'estimateur.

Solution

L'estimateur de la moyenne μ est la moyenne échantillonnale $\bar{y}$ = 73,42325. L'écart-type de l'échantillon est s = 40,60443. On estime donc l'écart-type de $\bar{y}$ par

$$\sqrt{\hat{V}(\bar{y})} = \sqrt{1-f}\,\frac{s}{\sqrt{n}} = \sqrt{1-0,0333}\,\frac{40,60443}{\sqrt{40}} = 6,312215 \ .$$

Intervalle de confiance

Un intervalle de confiance approximatif à 100(1-α)% est donné par

$$\bar{y} - z_{\alpha/2}\sqrt{\hat{V}(\bar{y})} \le \mu \le \bar{y} + z_{\alpha/2}\sqrt{\hat{V}(\bar{y})} \ , \tag{3.3}$$

où $z_{\alpha/2}$ est le point tel que

$$P(Z > z_{\alpha/2}) = \alpha/2 \quad \text{si } Z \text{ est de loi } N(0\ ;\ 1). \tag{3.4}$$

Exemple 3.2 *Intervalle de confiance pour une moyenne*

Déterminer un intervalle de confiance à 95 % pour la moyenne de la population définie à l'Exemple 3.1.

Solution

Le niveau de 1 - α = 0,95 signifie que α = 0,05 et $z_{\alpha/2}$ = 1,96. L'intervalle de confiance est donc $[\bar{y} - z_{\alpha/2}\sqrt{\hat{V}(\bar{y})}\,;\, \bar{y} + z_{\alpha/2}\sqrt{\hat{V}(\bar{y})}] = [73{,}42325 - 1{,}96(6{,}312215)\,;\, 73{,}42325 + 1{,}96(6{,}312215)]$

= [61,05 ; 85,79].

L'intervalle s'exprime parfois sous la forme $\bar{y} \pm z_{\alpha/2}\sqrt{\hat{V}(\bar{y})}$ et $z_{\alpha/2}\sqrt{\hat{V}(\bar{y})}$ est appelée *marge d'erreur*. Dans le dernier exemple, la marge d'erreur est 1,96(6,312215) = 12,37. Elle constitue une mesure de la précision de l'estimateur.

DÉFINITION 3.5 MARGE D'ERREUR

La *marge d'erreur à* 100(1 - α) % (ou *de niveau* 1 - α) de l'estimateur $\hat{\theta}$ d'un paramètre θ est la demi-largeur de l'intervalle de confiance à 100(1 - α) %. Lorsque l'estimateur est supposé de loi normale, la marge d'erreur est $z_{\alpha/2}\sqrt{\hat{V}(\bar{y})}$, où $z_{\alpha/2}$ est donné par (3.4).

La marge d'erreur est aussi interprétée comme ceci: la probabilité est de 1 - α pour que l'erreur d'estimation soit inférieure ou égale à la marge d'erreur.

Justification de la formule d'un intervalle de confiance

La formule (3.3) est basée sur le théorème limite de Hajek (voir la section 2.4) dont une conséquence est que la variable $Z = \dfrac{\bar{y} - \mu}{\sqrt{V(\bar{y})}}$ est de loi à peu près normale. De là, on conclut aussi que $\hat{Z} = \dfrac{\bar{y} - \mu_y}{\sqrt{\hat{V}(\bar{y})}}$ ~ $N(0\ ;\ 1)$, et c'est sur cette hypothèse que nous comptons pour légitimer la formule (3.3). Ce qui se défend en vertu de certains théorèmes limites, à condition, encore, que *n* soit grand. Mais la question est : quand est-ce qu'un échantillon est « grand »? Il n'y a pas de réponse simple et claire, et il ne pourrait y en avoir une, car cela dépend de la forme de la population: plus elle est asymétrique, pour l'échantillon doit être grand. Aucun théorème ne peut exprimer simplement et utilement le lien entre la forme de la population et l'approche à la normalité.

Mais il existe des études empiriques — non pas des théorèmes, mais des expériences, simulées à l'ordinateur à partir de populations réelles ou artificielles (voir la section 3.8). Ces études permettent d'estimer le niveau *réel* d'un intervalle de confiance théoriquement « à 100(1-α) % ». Ce qu'elles révèlent, généralement, c'est que si le niveau réel n'est jamais exactement égal à 1 - α, il lui est souvent assez proche. Et cela suffit, puisque, après tout, le choix du niveau — traditionnellement de 95 % — n'a rien d'incontournable. S'il se trouve que le niveau réel d'un « intervalle de confiance à 95 % » n'est en fait que de 92 %, ce n'est pas catastrophique; et s'il se trouve qu'il est de 97%, c'est tant mieux.

3.4 ESTIMATION D'UN TOTAL

Le total $\tau = \sum_{k=1}^{N} y_k$ d'une variable est un autre paramètre d'intérêt. Il est directement lié à la moyenne μ, puisque $\tau = N\mu$. Les résultats suivants sont donc immédiats :

L'estimateur

(3.5) $$\hat{\tau} = N\bar{y}$$

est sans biais pour τ. L'écart-type de $\hat{\tau}$ est

(3.6) $$\sqrt{V(\hat{\tau})} = N\sqrt{V(\bar{y})} = N(1-f)\frac{S}{\sqrt{n}}.$$

La variance de $\hat{\tau}$ est estimée sans biais par

(3.7) $$\hat{V}(\hat{\tau}) = N^2\hat{V}(\bar{y}) = N^2(1-f)\frac{s^2}{n}.$$

Un intervalle de confiance approximatif à 100(1-α)% est donné par

(3.8) $$\hat{\tau} - z_{\alpha/2}\sqrt{\hat{V}(\hat{\tau})} \le \tau \le \hat{\tau} + z_{\alpha/2}\sqrt{\hat{V}(\hat{\tau})}.$$

Exemple 3.3 *Estimation d'un total*

Dans l'Exemple 3.1, estimer les recettes totales τ de l'année et déterminer un intervalle de confiance à 95 % pour τ.

Solution

L'estimation du total est $\hat{\tau} = N\bar{y} = 1200(73{,}42325) = 88\,107{,}9$.

L'écart-type de $\hat{\tau}$ est estimé par $\hat{V}(\hat{\tau}) = N\sqrt{\hat{V}(\bar{y})}$ =1 200(6,312215) = 7 574,658. Un intervalle de confiance à 95 % pour τ est donc

$$[\hat{\tau} - 1{,}96\sqrt{\hat{V}(\hat{\tau})}\,;\,\hat{\tau} + 1{,}96\sqrt{\hat{V}(\hat{\tau})}] = [73\,262\,;\,102\,954].$$

On peut donc conclure avec (à peu près) 95 % de confiance que les recettes totales de l'année se situent entre 73 262 $ et 102 954 $.

Dans une population fortement asymétrique, les approximations normales peuvent être assez grossières à moins que l'échantillon soit excessivement grand. Cette difficulté n'est pas banale, en particulier dans certains domaines — dont la finance — où, typiquement, les distributions sont fortement asymétriques. Il n'existe pas *une* solution au problème, mais plusieurs palliatifs qui, choisis selon le contexte, peuvent améliorer les choses. Le prochain exemple illustre une approche relativement simple. D'autres solutions seront illustrées aux chapitres 6, 7 et 8.

Exemple 3.4 *Effets d'une asymétrie de la populario*n

Le tableau A02 présente une population de 1183 subdivisions de recensement québécoises. Considérons la variable Y : Nombre de logements privés. Pour un petit nombre d'unités (les grands centres urbains), Y prend des valeurs extrêmement élevées. La moyenne de la population est de 3 115, alors que 88,5 % des valeurs lui sont inférieures. Les quelque 11 valeurs supérieures à 40 000 (et les 5 supérieures à 100 000) ont pour effet de tirer la moyenne vers le haut.

Dans une situation comme celle-là, les propriétés établies dans ce chapitre — l'absence de biais de $\bar{y}$ et de s^2, si elles demeurent vraies, ne sont pas particulièrement réconfortantes. Car dans la plupart des cas $\bar{y}$ sous-estimera la moyenne et s^2 sous-estimera la variance. Pourquoi? Parce que les données extrêmes, étant peu nombreuses, ne se retrouveront que rarement dans l'échantillon, lequel sera presque toujours constitué entièrement de petites valeurs. La moyenne $\bar{y}$ sera inférieure à μ et la variance s^2 inférieure à S^2. L'estimateur $\bar{y}$ est sans biais dans le sens que les nombreuses *sous-estimations* sont compensées par quelques extrêmes *surestimations*, ce qui rend les estimateurs sans biais mais qui n'a rien de réjouissant.

Voici une solution possible: on isole ces très grandes unités qui causent tant d'ennuis, et on les traite séparément. Dans le cas présent, on peut certainement identifier les plus grandes unités (on sait que Montréal et Québec, par exemple, sont grands). Supposons, par exemple, qu'on identifie les 10 plus grosses unités et qu'on les recense au complet. Ensuite on tire un échantillon de taille 40 de la population U_1 constituée des N_1 = 1 173 unités qui restent. La moyenne μ_1 de U_1 sera estimée par la moyenne $\bar{y}_1$ de l'échantillon tiré dans U_1. La moyenne μ_2 de $U_2 = U \setminus U_1$ elle, sera connue. L'écart-type S_1 de U_1 sera vraisemblablement inférieur à S, l'écart-type de U, ce qui donnera une meilleure estimation de μ_1. On trouve en effet que S_1 = 3 637, nettement inférieur à S = 26 100. La moyenne de la population, $\mu = \frac{\tau}{N} = \frac{\tau_1 + \tau_2}{N}$, sera estimée par $\hat{\mu} = \frac{\hat{\tau}_1 + \tau_2}{N}$. L'écart-type de $\hat{\mu}$ est $\sqrt{V(\hat{\mu})} = \frac{1}{N}\sqrt{V(\hat{\tau}_1 + \tau_2)} = \frac{1}{N}\sqrt{V(\hat{\tau}_1)}$ car τ_2, le total de U_2, est fixe. L'écart-type estimé de $\hat{\tau}_1$ est $\sqrt{V(\hat{\tau}_1)} = 1173\sqrt{1 - \frac{40}{1173}}\frac{3637}{\sqrt{40}}$, et donc l'écart-type de $\hat{\mu}$ est $\sqrt{\hat{V}(\hat{\mu})} = \frac{1173}{1183}\sqrt{1 - \frac{40}{1173}}\frac{3637}{\sqrt{40}}$ = 560, un réel progrès par rapport à $\bar{y}$, dont l'écart-type est 3 612.

Par ailleurs, nous avons montré par des méthodes empiriques (voir la section 3.8) que les sous-estimations sont nettement moins nombreuses dans les échantillons tirés dans U_1. Par exemple, μ_1 est sous-estimée 56 % des fois alors que μ est sous-estimée 75 % des fois. De même, S_1^2 est sous-estimée 66 % des fois contre 91 % pour S^2.

Remarque *De meilleurs partages sont possibles*

Dans le dernier exemple, le choix du partage 1 173 + 10 était plutôt arbitraire. On aurait certainement pu faire d'autres choix, et il y en a de meilleurs. Nos calculs montrent qu'avec la répartition 1 157 + 26 (toujours avec un échantillon total de taille 50), l'écart-type de l'estimateur est réduit à 431. Cette question, et d'autres associées, seront abordées au chapitre 7.

Il faut reconnaître, cependant, qu'en pratique la solution proposée ici est un peu simpliste, car elle ne tient pas compte des coûts. Le coût de prélèvement d'une unité peut dépendre de sa taille. Un échantillon comprenant obligatoirement 26 grandes unités (et 24 petites) est plus efficace qu'un échantillon de 10 grandes et 40 petites, mais il est peut-être aussi beaucoup plus cher.

LA NOTION DE POIDS

L'estimateur $\hat{\tau}$ est une moyenne pondérée des observations, soit $\hat{\tau} = \sum_{i \in \omega} \alpha_i y_i$. Dans un *éas*, les α_i sont égaux, soit $\alpha_i = \frac{N}{n} = \frac{1}{n/N}$. Dépendamment du mode d'échantillonnage, les α_i peuvent varier, mais ils ont tous cette particularité: la réciproque de α_i (n/N ici) est la probabilité d'inclusion de y_i. Les agences de collection et d'analyse statistiques habituellement publient α_i en même temps que les valeurs observées et les moyennes, de façon à permettre à un utilisateur de pondérer ces quantités dans des estimations ultérieures.

3.5 DIFFÉRENCE DE MOYENNES

Considérons deux populations, de moyennes μ_1 et μ_2 et d'écarts-types S_1 et S_2; et un *éas* tiré de chacune des populations dans le but d'estimer la différence $\mu_d = \mu_1 - \mu_2$. Soit n_1, n_2 les tailles des échantillons; $\overline{y}_1$, $\overline{y}_2$ leurs moyennes; et s_1, s_2 leurs écarts-types. L'estimateur naturel de $\mu_1 - \mu_2$ est $\overline{y}_1 - \overline{y}_2$. Les échantillons sont indépendants, et donc la variance de l'estimateur est égale à la somme des variances, $V(\overline{y}_1 - \overline{y}_2) = V(\overline{y}_1) + V(\overline{y}_2)$, et sera estimée par $\hat{V}(\overline{y}_1 - \overline{y}_2) = \hat{V}(\overline{y}_1) + \hat{V}(\overline{y}_2)$. Si les échantillons sont assez grands, la statistique $Z = \frac{(\overline{y}_1 - \overline{y}_2) - (\mu_1 - \mu_2)}{\sqrt{\hat{V}(\overline{y}_1 - \overline{y}_2)}}$ approche la loi normale centrée-réduite et un intervalle de confiance à 100(1-α) % est donné approximativement par

$$(\overline{y}_1 - \overline{y}_2) - z_{\alpha/2}\sqrt{\hat{V}(\overline{y}_1 - \overline{y}_2)} \le \mu_1 - \mu_2 \le (\overline{y}_1 - \overline{y}_2) + z_{\alpha/2}\sqrt{\hat{V}(\overline{y}_1 - \overline{y}_2)} \ .$$

Exemple 3.5 *Différence de deux moyennes*

Le tableau suivant présente les prix d'un échantillon de maisons vendues dans deux quartiers de la région de Montréal.

Quartier 1 (N_1 = 433)	Quartier 2 (N_2 = 392)
717 499 579 505 459 510 471 456 459 438	501 511 439 464 385 399 402 434
422 431 397 405 440 426 368 389 365 370	425 419 432 376 389 375 391 372
355 345 335 303 320 328 306 320 277 295	362 340 318 347 283 255 224 196
312 262 215 235 202 171 163 171 145 1052	182 194 185 181

Quelques calculs :

Quartier 1	Quartier 2
n_1 = 40 $\bar{y}_1$ = 380,45 s_1 = 162,04272	n_2 = 28 $\bar{y}_2$ = 349,3214 s_2 = 99,59884

On a donc $\bar{y}_1 - \bar{y}_2$ = 31,12857, et $V(\bar{y}_1 - \bar{y}_2) = V(\bar{y}_1) + V(\bar{y}_2)$ $= \sqrt{595{,}8044 + 328{,}9773}$ = 30,41022. Un intervalle de confiance approximatif est donné par [31,12857-1,96(30,41022) ; 31,12857-1,96(30,41022). On affirme donc que la différence μ_1 - μ_2 se situe dans l'intervalle [-28,48 ; 90,73].

Remarque *Hypothèse d'égalité*

En général, lorsqu'on s'intéresse à la différence μ_1 - μ_2, c'est parce qu'on veut savoir si ces moyennes sont égales. L'objectif serait de tester l'hypothèse d'égalité, μ_1 - μ_2 = 0. Un intervalle de confiance pour la différence permet de conclure le test: si l'intervalle ne contient pas la valeur 0, on peut rejeter l'hypothèse. Dans le dernier exemple, puisqu'on affirme que μ_1 - μ_2 se situe dans l'intervalle [-28,48 ; 90,73], on ne peut exclure la possibilité que les moyennes soient égales : l'hypothèse ne peut pas être rejetée.

Facteur de correction dans un test d'hypothèse

Quand le but d'un intervalle de confiance est de tester une hypothèse, il est raisonnable, parfois, d'omettre le facteur de correction. Pourquoi? Parce que l'hypothèse porte souvent sur un paramètre dont le sens n'est pas celui qu'on lui donne ici. Dans le dernier exemple, μ_1 et μ_2 sont définies comme les moyennes concrètes des prix des maisons dans les populations échantillonnées de 433 et 392 maisons, l'hypothèse que μ_1 = μ_2 a été testée. Mais est-ce bien l'hypothèse qu'on veut tester ? Si la question posée est « Est-ce que les maisons ont plus de valeur dans le quartier 1 que dans le quartier 2 ? », alors il ne s'agit pas de μ_1 et μ_2, car ces moyennes sont elles-mêmes aléatoires : les prix auxquels les maisons se sont vendues ont subi des influences aléatoires et c'est plutôt sur leur *espérance* que porte la question. Il faut donc envisager chaque population non pas comme l'ensemble fini des prix auxquels les maisons se sont vendues, mais plutôt comme l'ensemble des prix auxquelles elles *auraient pu* se vendre. On aurait affaire à deux populations infinies, et le facteur de correction n'aurait pas sa place.

3.6 DÉTERMINATION DE LA TAILLE DE L'ÉCHANTILLON

Une question qui surgit normalement dès qu'on se propose de prélever un échantillon est, évidemment, «quelle doit être sa taille?» Si la taille n est déterminée par intuition, ou à partir de considérations économiques, ou par tradition, c'est après coup qu'on découvre si le choix a été bon ou mauvais. Un n trop petit ne donne pas la précision souhaitée; un n trop grand coûte inutilement cher. Il est souhaitable d'avoir une idée de ce qu'est la bonne taille avant de tirer l'échantillon.

Il faut d'abord noter qu'il n'existe pas de taille n qui suffise à tout besoin : un même échantillon peut assez bien estimer un paramètre mais pas un autre. Il peut être assez grand pour estimer, par exemple, le pourcentage de ménages connectés à l'Internet, mais pas pour estimer adéquatement le revenu moyen des ménages. On ne peut donc pas déterminer n avant d'avoir précisé les objectifs : définir les paramètres qui doivent être estimés; et décider du niveau de précision exigé.

Le niveau de précision s'exprime par la marge d'erreur qui, dans le cas d'une moyenne, est donnée par $z_{\alpha/2}\sqrt{V(\bar{y})}$. Si cette marge est fixée à une certaine valeur E, le problème consiste à choisir n de telle sorte que

$$\textit{Exigé} : \quad z_{\alpha/2}\sqrt{V(\bar{y})} = E.$$

Puisque $\sqrt{V(\bar{y})} = \sqrt{1-f}\,\frac{S}{\sqrt{n}}$ ($f = \frac{n}{N}$), nous devons déterminer la valeur de n qui satisfait l'équation

$$z_{\alpha/2}\sqrt{1-f}\,\frac{S}{\sqrt{n}} = E.$$

Si on ignore pour l'instant le facteur de correction $\sqrt{1-f}$ (qui de toute façon est souvent négligeable), l'équation à résoudre devient

$$\frac{z_{\alpha/2}S}{\sqrt{n}} = E$$

ce qui donne la solution provisoire

$$n_o = \left(\frac{z_{\alpha/2}S}{E}\right)^2 . \tag{3.9}$$

Cette solution correspond au cas où $f = 0$, c'est-à-dire, lorsque la population est infinie. Elle est parfaitement acceptable lorsque la population est grande par rapport à n. Mais si, comme dans l'exemple traité dans cette section, N n'est pas particulièrement grand, il suffit d'un simple ajustement pour obtenir la solution exacte. L'ajustement est le suivant :

$$n = \frac{n_o}{1+\frac{n_o}{N}} . \tag{3.10}$$

Tout cela, évidemment, exige que la variance S^2 soit connue ou puisse être estimée, avant même de tirer l'échantillon. S'il n'existe aucune information indépendante sur la population, on peut toujours estimer S^2 à partir d'un petit échantillon pilote.

Exemple 3.6 *Détermination de n*

Supposons que, dans le contexte de l'Exemple 3.1, on avait souhaité que la marge d'erreur soit de 12 $; et qu'afin de déterminer la taille de l'échantillon nécessaire, on avait procédé à un échantillonnage pilote de 15 unités pour estimer S. Supposons que les résultats obtenus avaient été les suivants :

20,14	39,48	47,46	100,14	124,69
33,89	40,68	66,57	114,95	140,86
35,50	41,24	71,71	122,42	147,28

À quoi aurait-on estimé n ?

Solution

On aurait utilisé l'écart-type de l'échantillon pilote pour estimer S, soit s = 44,04132. Une première approximation aurait donné :

$$n_o = \frac{[1,96(44,04132)]^2}{12^2} = 51,74523.$$

On ajuste :

$$n = \frac{51,74523}{1+\frac{51,74523}{1200}} \approx 50.$$

Selon l'échantillon pilote, donc, on aurait choisi un échantillon de taille 50. Remarquez, cependant, que cette taille est probablement surestimée, car à l'Exemple 3.1, on voit qu'un échantillon de taille 40 suffit: la marge d'erreur estimée là est d'environ 12 $, atteinte avec un échantillon de taille 40. La surestimation ici est due à une surestimation de S obtenue de l'échantillon pilote: 44,04, tandis qu'a l'Exemple 3.1 l'estimation est 40,60 — une estimation plus crédible étant donné la taille de l'échantillon.

Il n'est pas toujours possible ou nécessaire de tirer un échantillon pilote pour estimer S. Dans une situation particulière, il existe une variété d'informations contextuelles qui donnent une idée de sa valeur. Mais il est difficile de proposer des solutions générales. Ce que nous montrons ici, donc, ce sont des pistes de solution. En pratique, on complétera avec les moyens du bord.

Remarque *La marge d'erreur*

La marge d'erreur posée comme exigence est en principe déterminée en fonction de besoins pratiques. Mais en fait elle est souvent subjective et flexible. Si on projette

d'estimer le revenu moyen des ménages d'une population, on peut bien, a priori, souhaiter l'estimer avec une marge d'erreur de 5 000 $, mais cela n'a rien d'impératif. S'il se trouve que le coût d'une telle précision est prohibitif, on se contentera d'une précision moindre. Ainsi, le choix de n correspond toujours à un compromis — entre la précision souhaitée et son coût.

3.7 DÉMONSTRATIONS

Nous démontrons les théorèmes 3.1 et 3.2 trois fois afin d'illustrer trois approches différentes, chacune pouvant éventuellement se révéler utile.

On se référera aux résultats suivants (voir l'exercice 1.1-f).

- Chaque unité k est contenue dans $\binom{N-1}{n-1}$ échantillons et donc $P(k \in \omega) = \dfrac{\binom{N-1}{n-1}}{\binom{N}{n}} = \dfrac{n}{N}$.
- Chaque paire k, ℓ d'unités distinctes est contenue dans $\binom{N-2}{n-2}$ échantillons, et donc $P(k \in \omega \text{ et } \ell \in \omega) = \dfrac{n(n-1)}{N(N-1)}$.

On utilise aussi le fait que $[\sum_{k=1}^{N} y_k]^2 = \sum_{k=1}^{N} y_k^2 + \sum_k \sum_{\ell \neq k} y_k y_\ell = \sum_{k=1}^{N} y_k^2 + 2\sum_k \sum_{\ell > k} y_k y_\ell$.

PREMIÈRE APPROCHE

i) $E(\bar{y}) = \mu$

Démonstration

Nous désignerons $\bar{y}$ par $\bar{y}(\omega)$ quand nous jugerons nécessaire de rappeler qu'il s'agit de la moyenne de l'échantillon ω.

Par définition, $E(\bar{y}) = \sum_{\omega \in \Omega} \bar{y}(\omega) p(\omega) = \sum_{\omega \in \Omega} \bar{y}(\omega) \dfrac{1}{\binom{N}{n}} = \dfrac{1}{n} \sum_{\omega \in \Omega} (\sum_{i \in \omega} y_i) \dfrac{1}{\binom{N}{n}}$.

La dernière double somme peut être effectuée dans l'ordre inverse :

$$\frac{1}{n} \sum_{\omega \in \Omega} (\sum_{i \in \omega} y_i) \frac{1}{\binom{N}{n}} = \frac{1}{n} \sum_{k=1}^{N} y_k \sum_{\omega \ni k} \frac{1}{\binom{N}{n}} = \frac{1}{n\binom{N}{n}} \sum_{k=1}^{N} y_k \sum_{\omega \ni k} 1 =$$

$$\frac{1}{n\binom{N}{n}} \sum_{k=1}^{N} y_k \times (\text{Nombre d'échantillons contenant } k) = \frac{1}{n\binom{N}{n}} \sum_{k=1}^{N} y_k \binom{N-1}{n-1} = \frac{N\binom{N-1}{n-1}}{n\binom{N}{n}} \frac{\sum_{k=1}^{N} y_k}{N} = \mu .$$

ii) $V(\bar{y}) = (1-f)\dfrac{S^2}{n}$.

Démonstration

$$V(\bar{y}) = E[(\bar{y}-\mu)^2] = \frac{1}{n^2}\sum_{\omega}\sum_{i\in\omega}(y_i-\mu)^2 p(\omega) + \frac{1}{n^2}\sum_{\omega}[\sum_{i\in\omega}\sum_{j\neq i}(y_i-\mu)(y_j-\mu)]p(\omega)\ .$$

Le premier terme est

$$\frac{1}{n}\sum_{\omega}\left[\frac{\sum_{i\in\omega}(y_i-\mu)^2}{n}\right]p(\omega) = \frac{1}{n}\frac{\sum_{k=1}^{N}(y_k-\mu)^2}{N}\ . \tag{3.11}$$

On montre que le deuxième terme est

$$\frac{1}{n^2}\sum_{\omega}[\sum_{i\in\omega}\sum_{j\neq i}(y_i-\mu)(y_j-\mu)]p(\omega) = \frac{n-1}{nN(N-1)}\sum_{k=1}^{N}\sum_{\ell\neq k}(y_k-\mu)(y_\ell-\mu) \tag{3.12}$$

$$\frac{1}{n^2}\sum_{\omega}[\sum_{i\in\omega}\sum_{j\neq i}(y_i-\mu)(y_j-\mu)]p(\omega) = \frac{1}{n^2}\sum_{k=1}^{N}\sum_{\ell\neq k}\sum_{\substack{\omega\ni k\\ \omega\ni\ell}}(y_k-\mu)(y_\ell-\mu)p(\omega)$$

$$= \frac{1}{n^2}\sum_{k=1}^{N}\sum_{\ell\neq k}(y_k-\mu)(y_\ell-\mu)\sum_{\substack{\omega\ni k\\ \omega\ni\ell}}p(\omega) = \frac{1}{n^2}\sum_{k=1}^{N}\sum_{\ell\neq k}(y_k-\mu)(y_\ell-\mu)\binom{N-2}{n-2}p(\omega)$$

$$= \frac{1}{n^2}\sum_{k=1}^{N}\sum_{\ell\neq k}(y_k-\mu)(y_\ell-\mu)\frac{\binom{N-2}{n-2}}{\binom{N}{n}} = \frac{1}{n^2}\sum_{k=1}^{N}\sum_{\ell\neq k}(y_k-\mu)(y_\ell-\mu)\frac{n(n-1)}{N(N-1)}$$

$$= \frac{n-1}{nN(N-1)}\sum_{k=1}^{N}\sum_{\ell\neq k}(y_k-\mu)(y_\ell-\mu)\ .$$

Donc, réunissant (3.11) et (3.12), on a

$$V(\bar{y}) = \frac{1}{n}\frac{\sum_{k=1}^{N}(y_k-\mu)^2}{N} + \frac{n-1}{nN(N-1)}\sum_{k=1}^{N}\sum_{\ell\neq k}(y_k-\mu)(y_\ell-\mu)$$

$$= \frac{1}{n}\frac{\sum_{k=1}^{N}(y_k-\mu)^2}{N} + \frac{n-1}{nN(N-1)}\left\{\left[\sum_{k=1}^{N}(y_k-\mu)\right]^2 - \sum_{k=1}^{N}(y_k-\mu)^2\right\}$$

$$= \frac{1}{n}\frac{\sum_{k=1}^{N}(y_k-\mu)^2}{N} - \frac{n-1}{nN(N-1)}\sum_{k=1}^{N}(y_k-\mu)^2 = \frac{N-1}{nN}S^2 - \frac{n-1}{nN}S^2 = \left(1-\frac{n}{N}\right)\frac{S^2}{n}\ .$$

DEUXIÈME APPROCHE

Une deuxième approche part d'un mode de tirage particulier : *n* tirages successifs, *sans remise* : le premier est un choix au hasard de l'une des *N* unités de la population; le deuxième est un choix au hasard des *N*-1 qui restent, ainsi de suite. On voit bien que cette procédure résulte en un échantillon simple dans le sens de la Définition 3.1.

On définit alors *n* variables aléatoires X_1, X_2, … , X_n, où X_i est *le résultat du i^e tirage*. La moyenne des X_i est $\bar{X} = \bar{y}$. Nous montrerons que

i) $E(X_i) = \mu$; ii) $E(\bar{X}) = \mu$; iii) $V(X_i) = \dfrac{(N-1)S^2}{N}$;

iv) $C(X_i ; X_j) = -\dfrac{\sigma^2}{N-1} = -\dfrac{S^2}{N}$ $(i \neq j)$; v) $V(\bar{X}) = \left(1 - \dfrac{n}{N}\right)\dfrac{S^2}{n}$.

Démonstrations

i) La variable X_i prend les valeurs y_1, y_2, … , y_N, chacune avec probabilité $1/N$. Donc $E(X_i) = \sum_{k=1}^{N} y_k \dfrac{1}{N} = \mu$.

ii) $E(\bar{X}) = E\left[\dfrac{1}{n}\sum_{i=1}^{n} X_i\right] = \dfrac{1}{n}\sum_{i=1}^{n} E(X_i) = \dfrac{1}{n}\sum_{i=1}^{n} \mu = \mu.$

iii) Par définition, $V(X_i) = \sum_{k=1}^{N} (y_k - \mu)^2 \dfrac{1}{N} = \sigma^2 = \dfrac{(N-1)S^2}{N}$.

iv) $C(X_i ; X_j) = E[(X_i - \mu)(X_j - \mu)] =$

$$\frac{1}{N(N-1)}\sum_{k=1}^{N}\sum_{\ell \neq k}(y_k - \mu)(y_\ell - \mu) = \frac{1}{N(N-1)}[\sum_{k=1}^{N}(y_k - \mu)]^2 - \frac{1}{N(N-1)}\sum_{k=1}^{N}(y_k - \mu)^2 =$$

$$0 - \frac{1}{N(N-1)}\sum_{k=1}^{N}(y_k - \mu)^2 = -\frac{S^2}{N} .$$

v) $V(\bar{X}) = \dfrac{1}{n^2} V\left(\sum_{i=1}^{n} X_i\right) = \dfrac{1}{n^2}\sum_{i=1}^{n} V(X_i) + \dfrac{1}{n^2}\sum_{i=1}^{n}\sum_{j \neq i} C(X_i ; X_j) =$

$$\frac{1}{n^2}\sum_{i=1}^{n}\frac{(N-1)S^2}{N} + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j \neq i}\frac{-S^2}{N} = \frac{(N-1)S^2}{nN} - \frac{(n-1)S^2}{nN} = \left[\frac{N-1}{N} - \frac{n-1}{N}\right]\frac{S^2}{n} = \left(1 - \frac{n}{N}\right)\frac{S^2}{n} .$$

TROISIÈME APPROCHE

Une troisième approche se révélera particulièrement utile lorsqu'on traitera de plans d'échantillonnage plus complexes. Ici on définit N variables aléatoires ψ_1, ψ_2, …, ψ_N comme ceci :

$$\psi_k = \begin{cases} 1 & \text{si } k \in \omega \\ 0 & \text{sinon} \end{cases}, \quad k = 1, 2, \ldots, N.$$

L'intérêt de cette approche, c'est qu'on peut maintenant exprimer la moyenne de l'échantillon comme une combinaison linéaire des y_k :

$$\bar{y} = \frac{1}{n}\sum_{k=1}^{N} y_k \psi_k .$$

Les coefficients y_k sont des constantes alors que les ψ_k sont aléatoires.

Propriétés

i) $E(\psi_k) = P(\psi_k = 1) = \frac{n}{N}$; ii) $V(\psi_k) = f(1-f)$, $f = \frac{n}{N}$; iii) $C(\psi_k ; \psi_\ell) = -\frac{f(1-f)}{N-1}$.

Démonstration

ψ_i est une variable bernoullienne (binomiale avec $n = 1$) de probabilité f. Donc son espérance est f et sa variance est $f(1-f)$.

$C(\psi_\ell ; \psi_j) = E(\psi_k\psi_\ell) - E(\psi_k)E(\psi_\ell) = E(\psi_k\psi_\ell) - f^2$. Le produit $\psi_k\psi_\ell$ est aussi une variable bernoullienne de probabilité $P(\psi_k = 1 \text{ et } \psi_\ell = 1) = P(k \in \omega \text{ et } \ell \in \omega) = \frac{n(n-1)}{N(N-1)}$. Donc $C(\psi_k ; \psi_\ell)$

$$= \frac{n(n-1)}{N(N-1)} - \frac{n}{N}\frac{n}{N} = \frac{n}{N}\left(\frac{n-1}{N-1} - \frac{n}{N}\right) = -\frac{n}{N}\frac{N-n}{N(N-1)} = -\frac{f(1-f)}{N-1} .$$

On démontre alors

i) $E(\bar{y}) = \mu$

$$E(\bar{y}) = \frac{1}{n}\sum_{k=1}^{N} y_k E(\psi_k) = \frac{1}{n}\sum_{k=1}^{N} y_k \frac{n}{N} = \frac{1}{N}\sum_{k=1}^{N} y_k = \mu .$$

ii) $V(\bar{y}) = (1-f)\frac{S^2}{n}$

$$V(\bar{y}) = \frac{1}{n^2}\left[\sum_{k=1}^{N} y_k^2 V(\psi_i) + \sum_{k=1}^{N}\sum_{\ell\neq k} y_k y_\ell C(\psi_k;\psi_\ell)\right] = \frac{1}{n^2}\left[\sum_{k=1}^{N} y_k f(1-f) - \sum_{k=1}^{N}\sum_{\ell\neq k} y_i y_j \frac{f(1-f)}{N-1}\right]$$

$$= \frac{f(1-f)}{n^2}\left[\sum_{k=1}^{N} y_k^2 - \frac{\sum_{k=1}^{N}\sum_{\ell\neq k} y_k y_\ell}{N-1}\right] = \frac{f(1-f)}{n^2}\left[\frac{(N-1)\sum_{k=1}^{N} y_k^2 - \sum_{k=1}^{N}\sum_{\ell\neq k} y_k y_\ell}{(N-1)}\right]$$

$$= \frac{f(1-f)}{n^2}\left[\frac{N\sum_{k=1}^{N} y_k^2 - (\sum_{k=1}^{N} y_k^2 + \sum_{k=1}^{N}\sum_{\ell\neq k} y_k y_\ell)}{(N-1)}\right] = \frac{f(1-f)}{n^2}\left[\frac{N\sum_{k=1}^{N} y_k^2 - (\sum_{k=1}^{N} y_k)^2}{(N-1)}\right]$$

$$= \frac{f(1-f)N}{n^2}\left[\frac{\sum_{k=1}^{N} y_k^2 - (\sum_{k=1}^{N} y_k)^2/N}{(N-1)}\right] = \frac{f(1-f)N}{n^2}S^2 = (1-f)\frac{S^2}{n} .$$

iii) $E(s^2) = S^2$

$$E[(n-1)s^2] = E[\sum_{i\in\omega} y_i^2] - n[E(\bar{y}^2)]^2 =$$

$$nE\left[\frac{\sum_{i\in\omega} y_i^2}{n}\right] - n[V(\bar{y}) + \mu^2] = n\frac{\sum_{k=1}^{N} y_k^2}{N} - n\frac{(N-n)}{nN}S^2 - n\mu^2 = n\left[\frac{\sum_{k=1}^{N} y_k^2}{N} - \mu^2\right] - \frac{(N-n)}{N}S^2 =$$

$$\frac{n}{N}\left[\sum_{k=1}^{N} y_k^2 - N\mu^2\right] - \frac{(N-n)}{N}S^2 = \frac{n(N-1)}{N}S^2 - \frac{(N-n)}{N}S^2 = (n-1)S^2 \Longrightarrow E(s^2) = S^2.$$

3.8 QUELQUES SIMULATIONS

La construction d'un intervalle de confiance est basée sur de nombreuses approximations dont on sait qu'elles s'améliorent à mesure que n augmente. Pour évaluer la qualité de ces approximations, on recourt à des études dites *empiriques*. Ces études n'ont pas la généralité d'un théorème, mais elles permettent au moins de se faire une idée de la confiance qu'on peut — ou pas — accorder aux intervalles de confiance dans certaines situations particulière. Après quoi on peut extrapoler à d'autres situations, voisines de celles étudiées.

L'étude consiste à construire une population puis d'y tirer un grand nombre d'échantillons. L'ensemble des échantillons permet de répondre à un certain nombre de questions : Est-ce que $\bar{y}$ suit une loi normale ? À quel point la variable $\hat{Z} = \dfrac{\bar{y} - \mu}{\sqrt{\hat{V}(\bar{y})}}$ est-elle normale ? Quel est le niveau *réel* d'un intervalle de confiance « à 95 % » ?

Nous savons que les réponses dépendent fortement du degré de symétrie de la population. Nous considérons donc deux populations : une population de factures de restaurant; et une population de factures émises par une entreprise de service public à une clientèle qui peut être commerciale ou résidentielle. Elles sont caractéristiques des données financières : une forte concentration de petites valeurs, puis des valeurs de plus en plus élevées à des fréquences de plus en plus faibles. Aucune des deux n'est symétrique, mais le degré d'asymétrie est modéré dans la première, et très prononcé dans la deuxième. Dans chacune, nous avons simulé 10 000 fois le tirage d'un échantillon de taille n, pour certaines valeurs de n. Ensuite

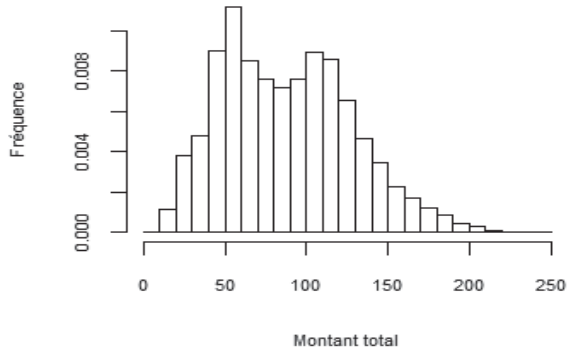
- nous examinons graphiquement la distribution de la moyenne échantillonnale $\bar{y}$ afin d'évaluer le degré de « normalité » de sa distribution; et
- nous estimons la probabilité qu'un intervalle de confiance, nominalement de niveau 95 %, contienne la moyenne de la population — une probabilité qui doit théoriquement approcher 0,95.

POPULATION ASYMÉTRIQUE MAIS PAS TROP

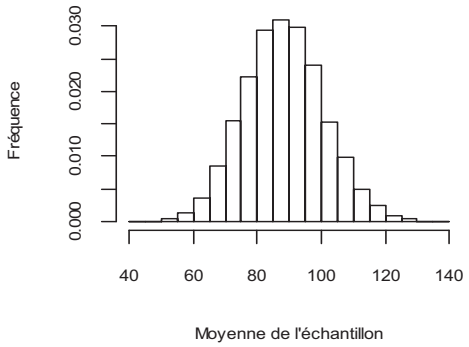
De cette première population, asymétrique mais pas trop, nous avons tiré des échantillons de tailles n = 10, 30 et 90. Les résultats sont présentés au Tableau 3.1. La distribution des moyennes $\bar{y}$ devrait s'approcher d'une loi normale à mesure que n augmente, et c'est bien ce qu'on observe. En fait la symétrie se manifeste déjà avec un échantillon de taille 10. Nous avons également compté la proportion des fois où l'intervalle contient le paramètre μ. Cette proportion estime le niveau *réel* de l'intervalle, par opposition au niveau nominal 0,95.

Tableau 3.1
Échantillons d'une population de N = 8 654 factures
μ = 88,2 ; S = 39,66

Population de factures

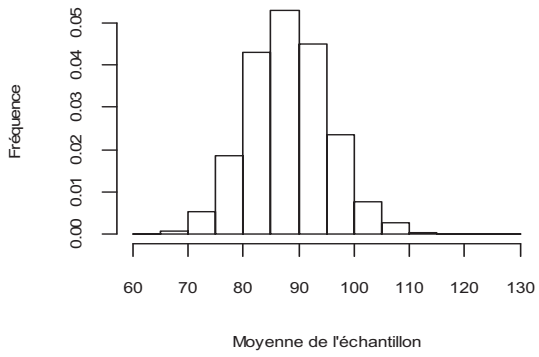


Échantillons de taille 10



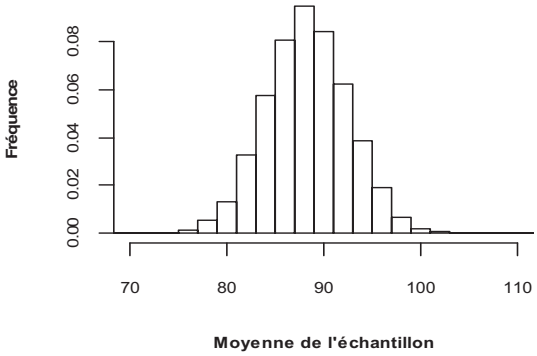
Niveau de confiance: 91,6 %

Échantillons de taille 30



Niveau de confiance: 93,8 %

Échantillons de taille 90



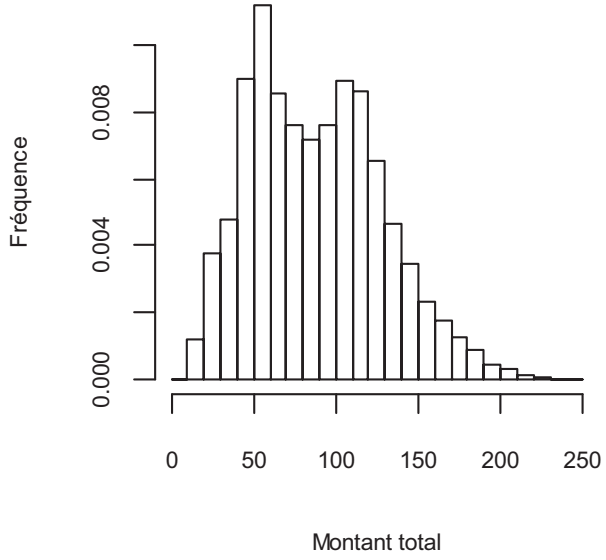
Niveau de confiance: 94,6 %

POPULATION FORTEMENT ASYMÉTRIQUE

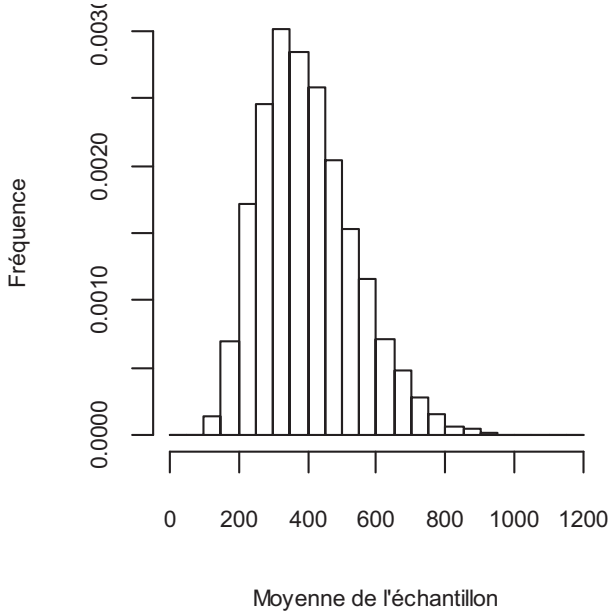
La population suivante est fortement asymétrique, plus encore qu'elle ne paraît dans l'histogramme (Tableau 3.2), car la population a été tronquée pour les besoins de la présentation graphique). On constate que la convergence vers la normalité est très lente. Un échantillon de taille 50 ne suffit pas: la distribution des $\bar{y}$ est encore nettement asymétrique et le niveau réel d'un intervalle de confiance, à 84,7 %, reste assez loin du niveau nominal de 95 %.

Tableau 3.2
Échantillons d'une population de $N = 53\,339$ factures
$\mu = 400{,}19$; $S = 980$

Population

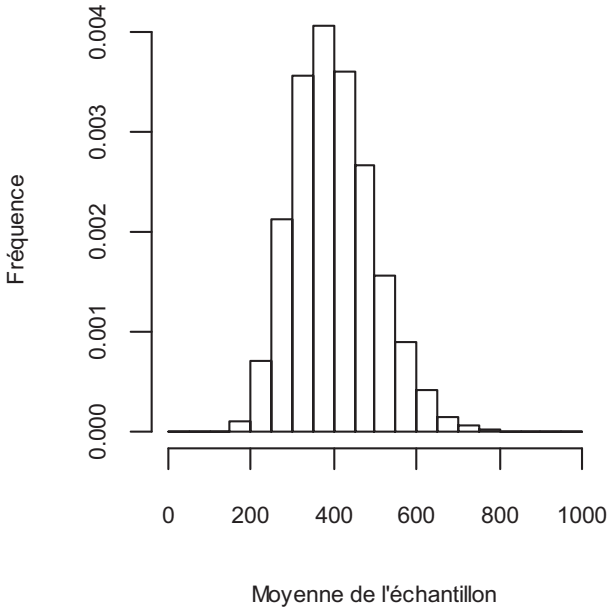


Échantillons de taille 50



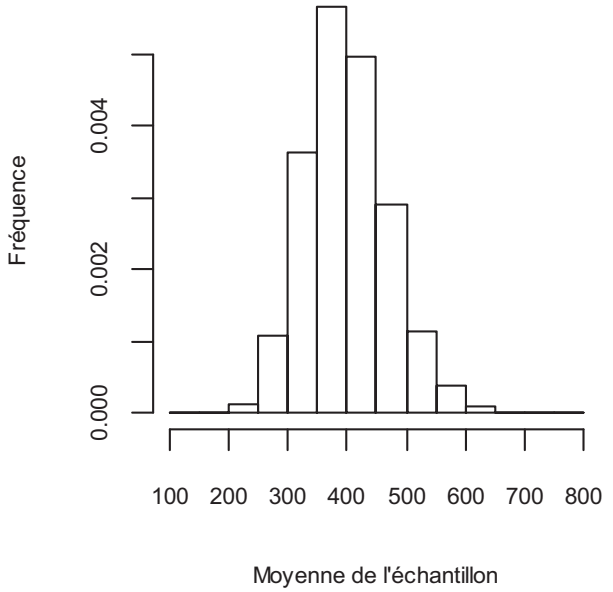
Niveau de confiance: 84.7 %

Échantillons de taille 100



Niveau de confiance: 89.3 %

Échantillons de taille 200



Niveau de confiance: 92.05 %

RÉSUMÉ

- Un échantillon aléatoire simple est un échantillon tiré de telle sorte que pour tout $\omega \in \Omega$, $p(\omega) = \frac{1}{\binom{N}{n}}$.

- Le total, la moyenne et l'écart-type d'une population sont désignés par $\tau = \sum_{k=1}^{N} y_k$, $\mu = \frac{1}{N}\sum_{k=1}^{N} y_k$ et $S = \sqrt{\frac{\sum_{k-1}^{N}(y_k - \mu^2}{N-1}}$.

- La moyenne et l'écart-type de l'échantillon sont désignés par $\bar{y} = \frac{1}{n}\sum_{i\in\omega} y_i$ et $s = \sqrt{\frac{\sum_{i\in\omega}(y_i - \bar{y})^2}{n-1}}$.

- Dans un échantillon aléatoire simple, la moyenne échantillonnale $\bar{y}$ est sans biais pour la moyenne de la population μ et sa variance est donnée par $V(\bar{y}) = (1-f)\frac{S^2}{n}$, où $f = \frac{n}{N}$ (la *fraction d'échantillonnage*). Le facteur (1-*f*) est appelé *facteur de correction*.

- La statistique $s^2 = \frac{\sum_{i\in\omega}(y_i - \bar{y})^2}{n-1}$ est un estimateur sans biais de S^2 et donc $\hat{V}(\bar{y}) = (1-f)\frac{s^2}{n}$ est un estimateur sans biais de $V(\bar{y})$.

- Le total $\tau = N\mu$ de la population est estimé sans biais par $\hat{\tau} = N\bar{y}$, sa variance $N^2V(\bar{y})$ est estimée par $N^2\hat{V}(\bar{y})$.

- La taille d'un échantillon nécessaire pour estimer μ avec une marge d'erreur E est $n = \frac{n_0}{1+\frac{n_0}{N}}$, où $n_o = \left(\frac{z_{\alpha/2}S}{E}\right)^2$.

EXERCICES

3.1 Une vérificatrice doit examiner un ensemble de 1 350 comptes de dépenses soumis par les représentants d'une grande compagnie. Elle prélève un échantillon de 20 comptes, et les corrige s'il y a lieu. Son objectif est d'estimer le montant réel des dépenses (par opposition au montant réclamé). Voici les montants réels (*Y*) :

23,90	110,52	79,95	146,65	19,51	26,62	27,67	65,79	12,38	12,44
72,14	57,37	62,93	135,88	15,22	36,35	31,98	7,39	33,05	46,96

a) Estimer la moyenne μ et la variance S^2 de *Y*.

b) Estimer l'écart-type de l'estimateur.

c) Déterminer un intervalle de confiance à 95 % pour μ.

d) Déterminer un intervalle de confiance (à 95 %) pour le montant total réel τ.

3.2 Mêmes données qu'au numéro précédent. Pour des raisons d'équité, la compagnie estime qu'elle ne peut pas contrôler les comptes des uns sans contrôler ceux des autres: elle doit soit les vérifier tous, soit n'en vérifier aucun. La décision dépend du montant total τ susceptible d'être épargné par la véfication. La somme des montants réclamés est de 80 341,65 $.

a) Déterminer un intervalle de confiance à 95 % unilatéral de la forme [0 ; b] pour le montant total τ que la compagnie paierait *en trop* si elle remboursait les montants réclamés sans vérification. Déterminer un intervalle de confiance pour τ. (Il suffit de déterminer un intervalle de confiance à 90 % et de se préoccuper seulement de la borne supérieure.)

b) On décide qu'on renoncerait à la vérification si on avait l'assurance (à 90 %) qu'elle ne permettrait pas de récupérer plus de 30 000 $. A-t-on cette assurance?

c) Supposons qu'au lieu de retenir le montant corrigé, on retient plutôt le montant de la correction, c'est-à-dire, la différence entre le montant réclamé et le montant admissible (cette différence est toujours positive ou nulle).

8,10	0	27,07	0	0	9,03	0	22,26	4,18	4,19
0	19,38	21,25	0	5,14	12,26	0	0	11,19	0

Répondre à la question c). Pouvez-vous expliquer la contradiction entre cette réponse et celle en b)?

3.3 On prélève un échantillon de 100 employés parmi les 1300 employés d'une compagnie afin d'estimer le nombre moyen μ de jours d'absence sur une période de 90 jours. On note Y le nombre de jours d'absence de chacun :

	0	1	2	3	4	5	Total
Effectifs (nombre d'employés)	40	24	19	11	4	2	100

a) Déterminez un intervalle de confiance à 95 % pour μ.

b) On évalue à 356 $ la perte occasionnée par l'absence d'un jour d'un employé. Déterminer un intervalle de confiance pour la perte totale due à toutes les absences durant la période considérée.

c) Durant la même période l'année dernière, le nombre moyen de jours d'absence a été de 0,98. Déterminer un intervalle de confiance pour l'accroissement moyen du nombre de jours d'absence depuis l'an dernier. Peut-on affirmer avec 95 % de confiance que cette moyenne a augmenté depuis l'année dernière?

d) Un analyste découvre que la moyenne de 0,98 citée en c) n'est pas la moyenne réelle: c'est la moyenne estimée à partir d'un échantillon de 100 employés tiré l'an dernier. Voici la distribution échantillonnale du nombre Z d'absences :

Nombre de jours d'absence (Z)	0	1	2	3	4	Total
Effectifs (nombre d'employés)	46	25	17	9	3	l00

Déterminer un intervalle de confiance pour l'accroissement moyen du nombre de jours d'absence depuis l'an dernier. Peut-on affirmer avec 95 % de confiance que cette moyenne a augmenté depuis l'année dernière? Expliquer pourquoi l'intervalle ici est plus large qu'en c).

e) Une deuxième analyste découvre que l'échantillon tiré l'année dernière était constitué des mêmes individus, sélectionnés l'année dernière. Pour chaque individu elle prend la différence $d = Y - Z$ entre le nombre d'absences cette année et le nombre d'absence l'année dernière. Voici les données:

Différence (d)	-1	0	1	2	Total
Effectifs (nombre d'employés)	2	80	11	7	100

Déterminer un intervalle de confiance pour l'accroissement moyen du nombre de jours d'absence depuis l'an dernier. Peut-on affirmer avec 95 % de confiance que cette moyenne a augmenté depuis l'année dernière? Expliquer pourquoi l'intervalle ici est plus étroit qu'en c) et qu'en d).

3.4 Un vérificateur tire un échantillon de 20 comptes de dépenses parmi les 780 soumis par une compagnie afin d'estimer la proportion des frais donnant droit à une exemption de la TVQ. Les montants réclamés sont tous connus. Leur moyenne est μ = 580,75 \$, alors que la moyenne de l'échantillon est de $\bar{y}$ = 765,91 \$. Selon une analyste, cet écart est excessif et ne peut être attribué au hasard de l'échantillonnage. Elle soupçonne plutôt que le vérificateur a délibérément choisi les plus gros comptes. Voici les montants réclamés dans les comptes de l'échantillon :

4,97	441,34	562,69	603,68	621,24
112,27	500,41	562,93	614,05	777,17
806,22	860,40	931,10	1055,88	1345,58
844,33	865,74	1007,15	1238,10	1562,99

Les doutes de l'analyste sont-ils justifiés?

3.5 La valeur aux livres des 1 230 comptes à recevoir d'une compagnie est de 136 220 \$. Pour vérifier l'exactitude de ce montant, on prélève un échantillon de 25 comptes. On observe les montants suivants.

112,34	98,25	17,32	21,05	23,45	7,38	32,06
18,25	12,35	234,54	19,87	12,34	240,23	12,50
230,18	332,12	23,45	15,34	14,32	12,36	10,23
112,50	124,99	167,35	18,45			

a) Déterminer un intervalle de confiance à 95 % pour le montant total réel τ.

b) Déterminez un intervalle de confiance à 95 % pour la différence (valeur aux livres - total réel). Peut-on (avec 95 % de confiance) affirmer que la valeur aux livres totale est supérieure à la valeur réelle ?

3.6 On tire un échantillon de taille n = 100 d'une population U de N = 1 584 déclarations d'impôt, afin d'estimer le montant total de revenus de dividendes déclarés.

a) Supposons que S = 244. Déterminer l'écart-type de l'estimateur du total.

b) Supposons maintenant qu'il est possible avec peu d'effort de parcourir la population entière afin d'exclure les déclarations qui ne comprennent aucun revenu de dividendes. Et qu'on a ainsi réduit la population à un sous-ensemble U_1 de N_1 = 620 déclarations, toutes comprenant un revenu de dividendes positif. Supposons que l'écart-type de cette sous-population est S_1 = 200. Déterminer l'écart-type de l'estimateur du total dans un échantillon de taille 100 tiré dans U_1.

c) Dans un échantillon de taille 100, on s'attend (étant donné que N_1 = 620) à trouver 39 données non nulles, et ce sont sur ces données que le total est estimé. Si on profitait de la taille réduite de la population (et de son écart-type) pour tirer un échantillon de taille 39 de U_1, quel serait l'écart-type de l'estimateur du total ?

3.7 Les données du tableau ci-dessous portent sur un échantillon de 45 comptes de dépenses soumis par les représentants d'une compagnie. La population est un ensemble de 2 500 comptes. Ces comptes ont été vérifiés et les réclamations incorrectes corrigées. Les variables observées sont donc

x : Montant réclamé; y : Montant corrigé.

x	y	x	y	x	y	x	y	x	y
270,18	215,72	19,37	14,42	23,88	20,29	260,69	260,69	73,83	73,83
23,54	23,54	16,84	16,84	237,42	175,45	187,92	187,92	165,54	165,54
93,48	93,48	36,13	36,13	66,37	66,37	30,9	26,00	15,77	15,77
31,51	31,51	404,82	404,82	231	231	33,98	33,98	64,76	64,76
11,66	11,66	314,26	314,26	109,3	109,3	6,57	6,57	43,05	43,05
99,66	99,66	329,27	329,27	68,86	68,86	30,43	30,43	72,3	65,86
57,52	57,52	1865,83	1865,83	48,42	48,42	518,77	518,77	48,07	48,07
6,9	6,9	49,47	49,47	190,62	190,62	381,73	381,73	93,4	93,4
13,39	13,39	61,16	61,16	76,67	76,67	249,98	249,98	385,33	385,33

$\sum x_i = 7420{,}5$; $\sum y_i = 7284{,}24$; $\sum x_i^2 = 4\,920\,941{,}52$;
$\sum y_i^2 = 4\,867\,399{,}46$; $\sum x_i y_i = 4\,890\,715{,}97$.

a) Déterminer un intervalle de confiance à 95% pour le montant réclamé moyen μ_x. Estimer le coefficient de variation de l'estimateur.

b) Soit $Z = X - Y$. Déterminer un intervalle de confiance à 95% pour $\mu_z = \mu_x - \mu_y$, l'erreur moyenne. (Cette erreur peut être estimée par $\bar{z}$, la moyenne des différences $z_i = x_i - y_i$). Estimer le coefficient de variation de l'estimateur.

c) Peut-on conclure avec 95 % de confiance que le montant total de l'erreur est inférieur à 20 000 $?

d) Estimer μ_z par $\bar{z} = \bar{x} - \bar{y}$ et déterminer un intervalle de confiance en estimant la variance de $\bar{z}$ par $\hat{V}(\bar{z}) = \hat{V}(\bar{x}) + \hat{V}(\bar{y})$. Pourquoi cette approche est-elle erronée? Pourquoi cet intervalle est tellement plus large que celui calculé en c) ?

e) Supposons que pour tester l'hypothèse que $\mu_x = \mu_y$, on procède de la façon suivante. On détermine deux intervalles de confiance, l'un pour μ_x, l'autre pour μ_y, et on rejette l'hypothèse si les deux intervalles sont disjoints. Selon les résultats que vous avez obtenus en a) et en b), que concluez-vous?

[Pour plusieurs raisons, la solution proposée en d) n'est pas correcte. La réponse à la question découle de l'intervalle construit en c).

3.8 [*Même contexte qu'au numéro* 3.7 *ci-dessus*]. Les montants réclamés peuvent aisément être déterminés pour la population entière, de sorte que leur moyenne est connue : $\mu_x = 207{,}72$. On peut donc estimer μ_z par $\hat{\mu}_z = \mu_x - \bar{y}$. Cet estimateur est-il meilleur ou moins bon que celui proposé au numéro 3.7-b)? Votre conclusion était-elle prévisible?

3.9 [*Même contexte qu'au numéro* 3.7 *ci-dessus*]. Un an plus tard, on planifie l'opération annuelle de vérification des comptes. Le nombre de comptes est maintenant 1 800. Autrement, on suppose que les écarts-types n'ont pas changé sensiblement et peuvent donc être estimés à partir de l'échantillon déjà tiré.

a) On souhaite estimer la moyenne μ_y avec une marge d'erreur relative (à 95 %) d'au plus 25 %. Quelle devrait être la taille de l'échantillon?

b) Supposons qu'on veuille estimer le montant moyen de la correction μ_{x-y} par $\bar{x} - \bar{y}$ avec une marge d'erreur relative (à 95 %) d'au plus 25 %. Quelle devrait être la taille de l'échantillon?

3.10 Le tableau A03 présente un ensemble de 8 654 factures émises par un restaurant au cours d'une année de 364 jours. Afin d'estimer le montant total des recettes, on décide de tirer un échantillon de 3 jours parmi les 364.

a) Les recettes totales de 3 jours sont : 1 248,90 $, 1 751,45 $, et 1 804,35 $. Déterminer $\hat{\tau}$, l'estimation du montant total τ des recettes de l'année et estimer l'écart-type de $\hat{\tau}$.

b) Les trois jours choisis représentent un total de 53 factures dont les valeurs ont une moyenne de 90,65472 et un écart-type de 45,98437. Si on avait traité (erronément) les données comme si elles constituaient un *éas* de taille 53 tiré d'une population de taille 8 654, quelle aurait été l'estimation du total?

c) Déterminer l'écart-type réel de l'estimateur $\hat{\tau}$ décrit en a) à partir du fait que l'écart-type de 364 totaux de la population est 670,5357.

d) Déterminer l'écart-type réel de l'estimateur de $\hat{\tau}$ basé sur un *éas* de 53 factures. L'écart-type des valeurs des 8 654 factures de la population est 39,66464.

3.11 [*Même contexte qu'au numéro* 3.7 *ci-dessus*]. Supposons que les valeurs de X (montant réclamé) sont connues pour la population entière, et qu'on en profite pour isoler les 10 plus gros comptes (selon la valeur de X) afin de les recenser entièrement; et de tirer ensuite un échantillon de taille 35 parmi les $N_1 = 2490$ comptes « moins gros ». L'échantillon donne les valeurs suivantes de X (montant réclamé) et Y (montant réel):

x	y	x	y	x	y	x	y	x	y
52,15	52,15	70,93	70,93	55,34	55,34	173,03	173,03	146,21	146,21
75,69	75,69	93,4	93,4	216,18	216,18	221,39	178,17	198,54	198,54
7,56	7,56	111,27	111,27	236,72	236,72	43,32	42,92	54,73	54,73
150,04	129,63	25,98	25,98	5962,23	5962,23	11,56	11,56	791,45	791,45
296,81	296,81	37,77	37,77	49,27	36,63	188,49	188,49	47,25	47,25
58,4	58,4	81,59	81,59	422,35	422,35	45,48	45,48	115,04	115,04
676,1	676,1	120,25	120,25	3468,54	3468,54	30,72	30,72	51,79	43,18

$\sum x_i = 14\,387{,}57$; $\sum y_i = 14\,302{,}29$; $\sum x_i^2 = 49\,322\,423$; $\sum y_i^2 = 49\,297\,508$; $\sum x_i y_i = 49\,308\,706$.

a) Estimer le montant corrigé moyen μ_y. Estimer l'écart-type et le coefficient de variation de votre estimateur. Les y correspondant aux 10 plus grandes valeurs de X sont sans erreur et leur moyenne est 15 165,52.

b) Estimer le montant total de la correction (τ_{x-y}). Estimer l'écart-type et le coefficient de variation de votre estimateur. Comparer la qualité de l'estimateur à celui du numéro 3.7-b).

3.12 Voici le nombre d'hectares d'orge produits dans un échantillon de 50 fermes tirées d'un échantillon de 3 343 fermes.

0	105	301	0	0	182	0	0	18	161
257	0	176	0	0	116	297	0	0	0
139	320	172	0	71	0	0	111	0	157
0	0	192	0	253	159	0	105	207	85
0	0	0	0	0	129	72	123	0	204

$\sum y_i = 4112$; $\sum y_i^2 = 818824$

a) Estimer la production moyenne d'orge dans les fermes de la population ainsi que l'écart-type de l'estimateur.

b) Estimer la production totale d'orge dans les fermes de la population ainsi que l'écart-type de l'estimateur.

c) Déterminer un intervalle de confiance à 95 % pour la production totale d'orge dans la population.

d) Soit U_1 l'ensemble des fermes de la population cultivant de l'orge (il y en a 1 579). Supposons qu'il est possible d'identifier les unités appartenant à U_1, et qu'on estime la production totale à partir d'un *éas* de taille 50 tiré dans U_1. Estimer l'écart-type de l'estimateur. Pour estimer S_1 (l'écart-type des données de U_1),

utilisez les 25 données non nulles de l'échantillon comme s'il s'agissait d'un *éas* tiré dans U_1 (une approche qui sera justifiée au chapitre 5).

e) Avec l'approche décrite en d), quelle taille d'échantillon aurait permis de réaliser la même précision que celle estimée en b)?

3.13 Considérons une population de taille *N*, de moyenne μ et d'écart-type *S*. Soit $\bar{y}_1$ et s_1 la moyenne et l'écart-type d'un échantillon de taille 100 tiré de cette population ; $\bar{y}_2$ et s_2 la moyenne et l'écart-type d'un échantillon de taille 200 tiré de cette même population. Inscrivez « < », « > » , « = » ou « ≈ » dans les espaces. Marquez « ? » si aucun lien ne peut être établi entre les deux quantités. Vous supposerez que $\bar{y}_1$ et $\bar{y}_2$ sont toutes deux de loi normale. Justifier vos réponses.

a) $E(\bar{y}_1)$ _____ $E(\bar{y}_2)$	h) $P(\bar{y}_1 > \mu+1{,}96)$ _____ $P(\bar{y}_2 > \mu+1{,}96)$
b) $E(\bar{y}_1)$ _____ μ	i) $P(\bar{y}_1 > \mu+1{,}96\sqrt{V(\bar{y}_1)})$ _____ $P(\bar{y}_2 > \mu+1{,}96\sqrt{V(\bar{y}_2)})$
c) s_1 _____ S	j) $P(\lvert \bar{y}_1 - \mu \rvert < 1{,}96)$ _____ $P(\lvert \bar{y}_2 - \mu \rvert < 1{,}96)$
d) $\sqrt{V(\bar{y}_1)}$ _____ S	k) $P(\lvert \bar{y}_1 - \mu \rvert > 1{,}96\sqrt{V(\bar{y}_1)})$ _____ 0,05
e) $E(s_1^2)$ _____ $E(s_2^2)$	l) $P(\bar{y}_1 > \mu + 1{,}96\sqrt{V(\bar{y}_1)})$ _____ 0,025
f) $\sqrt{V(\bar{y}_1)}$ _____ $\sqrt{V(\bar{y}_2)}$	m) $E[\hat{V}(\bar{y}_1)]$ _____ $E[\hat{V}(\bar{y}_2)]$
g) $E(s_1^2)$ _____ $E(s_2^2)$	n) $\sqrt{\hat{V}(\bar{y}_1)}$ _____ $\sqrt{\hat{V}(\bar{y}_1)}$

3.14 D'une population *U* de 4 337 factures de restaurant, on prélève un échantillon de 36 factures afin d'estimer la valeur totale τ des boissons alcooliques. Voici les montants dépensés en boisson alcooliques :

0	0	0	0	0	0	0	0	0
0	0	10,1	10,1	10,1	16,16	16,16	16,16	18,18
20,2	20,2	20,2	20,2	20,2	24,24	28,28	30,3	30,3
30,3	30,3	36,36	40,4	50,5	60,6	60,6	60,6	80,8

$\sum y_i = 761{,}54 \; ; \; \sum y_i^2 = 31570{,}05$

a) Estimer τ et estimer le coefficient de variation de votre estimateur

b) Supposons qu'il a été possible d'exclure de la population les factures qui ne comprenaient aucune composante de boisson alcoolique, et qu'après leur exclusion on se retrouve avec une sous-population U_1 de 3 409 factures dont toutes comprennent une composante alcool non nulle. Considérez les données non nulles comme un *éas* de taille 25 afin d'obtenir une nouvelle estimation de τ (nous montrerons au chapitre 6 que cette approche est légitime). Estimer le coefficient de variation de l'estimateur.

3.15 On doit prélever un échantillon de maisons unifamiliales vendues au courant de l'année dans une ville afin d'en estimer le prix moyen. La population est répartie en 21

quartiers. Voici, pour chaque quartier, le nombre M_i de maisons, le prix total t_i de toutes les maisons du quartier et l'écart-type s_i des prix des maisons du quartier :

Quartier	M_i	t_i	S_i	*Quartier*	M_i	t_i	S_i
1	9	3555	52,32	12	19	7904	82,27
2	12	6752	255,55	13	22	8060	55,12
3	13	13537	366,58	14	13	4966	110,49
4	5	2792	136,77	15	11	4043	41,63
5	15	6299	55,32	16	14	8030	216,77
6	11	3741	50,07	17	19	20072	400,72
7	12	4236	83,27	18	16	8296	197,55
8	6	2233	57,79	19	15	6361	63,12
9	19	10116	182,65	20	13	4993	69,25
10	16	16549	355,29	21	19	7689	96,58
11	11	6212	144,59				

a) On tire un échantillon de 3 quartiers. On tombe sur les quartiers 4, 7 et 10. Estimer la valeur moyenne de toutes les maisons vendues. Estimer l'écart-type de l'estimateur.

b) Supposons qu'on ait décidé plutôt de tirer un *éas* de 33 maisons tirées dans la liste des 290 maisons de la population et qu'on ait obtenu les 33 valeurs suivantes :

497	356	340	171	361	352	373	305	285
391	505	476	669	343	245	723	441	481
489	428	413	1162	905	297	377	458	537
492	296	285	876	1323	963			

$\sum_{i\in\omega} y_i = 16\,615$; $s_y = 266,6734$

Estimer la valeur moyenne de toutes les maisons vendues. Estimer l'écart-type de l'estimateur.

c) La procédure employée en a) est peu efficace, mais elle est correcte. Ce qui aurait été incorrect, ç'aurait été d'agir comme si on disposait d'un *éas* de 33 maisons. Quelle aurait été l'estimation de la moyenne dans ce cas et à quoi aurait-on évalué $\sqrt{\hat{V}(\hat{\tau})}$? [L'écart-type des valeurs des 33 maisons est s = 409,688.]

d) Utiliser maintenant les données ci-dessus sur la population entière afin de déterminer l'écart-type de l'estimateur de la moyenne dans les situations suivantes:

i) On tire un échantillon de trois quartiers.

ii) On tire un *éas* de 41 maisons (le nombre moyen de maisons dans un échantillon de 3 quartiers) parmi les 290 maisons de la population. L'écart-type des prix des maisons dans la population est S = 301,0797.

Remarque Il est clair qu'il est préférable de tirer 41 maisons dans l'ensemble des maisons plutôt que 3 quartiers parmi les 21 quartiers. On peut expliquer ce phénomène en comparant l'écart-type de la population, S = 301,0797 avec les écarts-types des quartiers S_i, généralement inférieurs (leur moyenne est de 146). Les quartiers regroupent des maisons semblables, et par conséquent les écarts entre les quartiers sont importants.

3.16 Le but de cet exercice est d'évaluer l'importance de la fraction d'échantillonnage $f = n/N$ dans la qualité de l'estimation d'une moyenne μ. Soit $\bar{y}$ la moyenne d'un *éas* de taille n tiré d'une population de moyenne μ et d'écart-type S = 30. Le tableau suivant présente la probabilité $P(|\bar{y} - \mu| > 4)$ sous chaque combinaison de trois hypothèses concernant n et cinq hypothèses concernant f (on suppose que $\bar{y}$ est de loi normale):

		n/N				
		0,2	0,1	0,05	0,01	0,005
	20	0,505	0,530	0,541	0,549	0,550
n	25	0,456	0,482	0,494	0,503	0,504
	50	0,292	0,320	0,333	0,343	0,345

Confirmer les résultats des calculs ci-dessus et commentez sur l'effet de n et l'effet de la fraction d'échantillonnage.

3.17 D'une population de 1 640 étudiants, on veut prélever un échantillon pour estimer certains paramètres concernant la variable y : montant dépensé mensuellement en transport. Un échantillonnage préliminaire a donné une estimation de μ et de S : $\mu \approx 140$, $S \approx 25$. Déterminer la taille de l'échantillon qu'on doit prélever s'il faut que

a) la marge d'erreur de l'estimateur de la moyenne soit de 2 $;

b) la marge d'erreur de l'estimateur de la moyenne soit de 4 %;

c) la marge d'erreur de l'estimateur du total soit de 2 000 $;

d) la marge d'erreur de l'estimateur du total soit de 4 %.

3.18 D'une population de taille 678 on doit prélever un échantillon afin d'estimer la moyenne de la population avec une marge d'erreur relative (à 95 %) de 10 %. On prélève un échantillon préliminaire de taille 12 et obtient les résultats suivants :

5,17	42,21	81,26	52,34	64,15	3,06
30,36	26,96	24,24	4,54	120,24	48,68

Estimez la taille de l'échantillon qu'il faut prélever.

3.19 On doit tirer un échantillon d'une population de N = 5 000 factures de restaurant. Soit Y le montant total de la facture et X la composante «boisson alcoolique». À partir d'un échantillon pilote, on obtient les estimations suivantes : μ_y = 93,79; μ_x = 17,8; S_y = 54,51; S_x = 18,12; S_{xy} = 745,16.

a) Déterminer la taille de l'échantillon nécessaire pour estimer μ_y avec une marge d'erreur relative (à 95 %) de 25 %.

b) Déterminer la taille de l'échantillon nécessaire pour estimer μ_x avec une marge d'erreur relative (à 95 %) de 20 %.

c) Déterminer la taille de l'échantillon nécessaire pour estimer $\mu_{y\text{-}x}$ (le montant moyen de nourriture) avec une marge d'erreur relative (à 95 %) de 15 %.

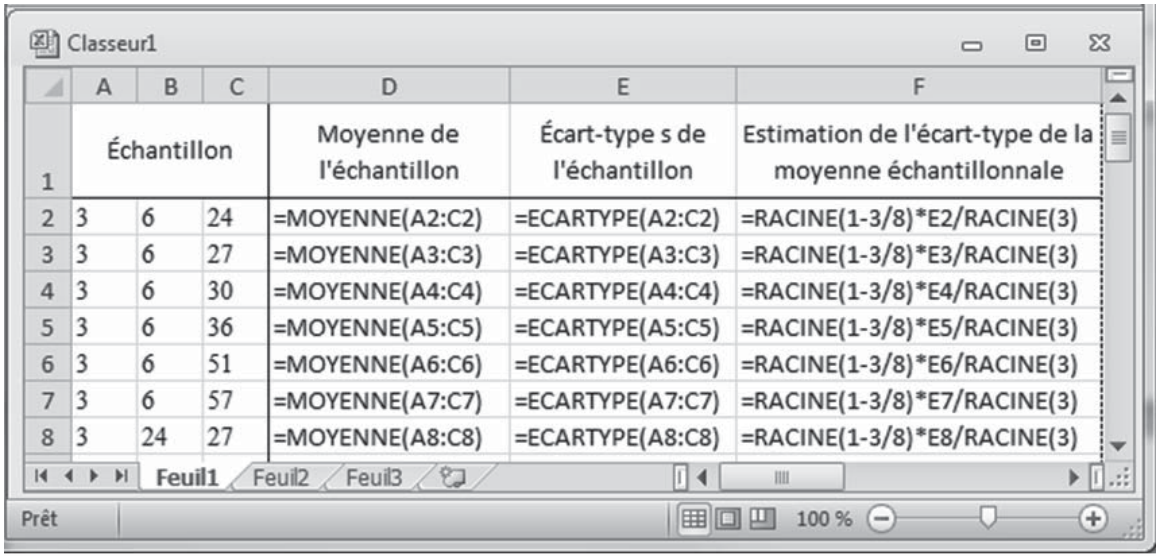
3.20 Considérez la population composée des 8 unités suivantes :

3 6 24 27 30 36 51 57

Supposez qu'on tire un échantillon de taille 3 de cette population. Voici l'ensemble des 56 échantillons possibles :

#	Échantillon			#	Échantillon			#	Échantillon			#	Échantillon		
1	3	6	24	15	3	27	57	29	6	27	51	43	24	30	57
2	3	6	27	16	3	30	36	30	6	27	57	44	24	36	51
3	3	6	30	16	3	30	51	31	6	30	36	45	24	36	57
4	3	6	36	18	3	30	57	32	6	30	51	46	24	51	57
5	3	6	51	19	3	36	51	33	6	30	57	47	27	30	36
6	3	6	57	20	3	36	57	34	6	36	51	48	27	30	51
7	3	24	27	21	3	51	57	35	6	36	57	49	27	30	57
8	3	24	30	22	6	24	27	36	6	51	57	50	27	36	51
9	3	24	36	23	6	24	30	37	24	27	30	51	27	36	57
10	3	24	51	24	6	24	36	38	24	27	36	52	27	51	57
11	3	24	57	25	6	24	51	39	24	27	51	53	30	36	51
12	3	27	30	26	6	24	57	40	24	27	57	54	30	36	57
13	3	27	36	27	6	27	30	41	24	30	36	55	30	51	57
14	3	27	51	28	6	27	36	42	24	30	51	56	36	51	57

Le tableau Excel suivant présente les quelques premiers échantillons et montre comment calculer certaines statistiques échantillonnales :



a) Calculez μ et S pour la population.

b) Confirmer numériquement que $E(\bar{y}) = \mu$ et que $V(\bar{y}) = (1-f)\frac{S}{\sqrt{n}}$.

c) Confirmer numériquement que s^2 est sans biais pour S^2.

d) Montrer que s n'est pas un estimateur sans biais de S.

e) Déterminez la probabilité de commettre une erreur

(i) de plus de 2 unités ; (ii) de plus de 5 unités ; (iii) de plus de 25 %

dans l'estimation de la moyenne.

f) Quelle est la probabilité de se tromper de plus de 20 % dans l'estimation de l'écart-type S de la population ?

g) Quel est le niveau réel de confiance de l'intervalle de confiance déterminé par la formule.

L'intervalle de confiance $\bar{y} - 1{,}96\sqrt{\hat{V}(\bar{y})} \leq \mu \leq \bar{y} + 1{,}96\sqrt{\hat{V}(\bar{y})}$ est censé être de niveau 95 % à peu près. Déterminer son niveau exact.

h) Si on connaissait la variance S^2 de la population. on aurait utilisé $\sqrt{V(\bar{y})}$ plutôt que $\sqrt{\hat{V}(\bar{y})}$. Quel est le niveau de confiance exact de l'intervalle de confiance $\bar{y} - 1{,}96\sqrt{\hat{V}(\bar{y})} \leq \mu \leq \bar{y} + 1{,}96\sqrt{\hat{V}(\bar{y})}$?

i) Déterminer le niveau de confiance exact de l'intervalle de confiance donné par la formule $\bar{y} - 3\sqrt{\hat{V}(\bar{y})} \leq \mu \leq \bar{y} + 3\sqrt{\hat{V}(\bar{y})}$.

3.21 Considérez une population de personnes dont les pointures des chaussures (Y) sont distribuées comme suit.

Prélevez (à l'aide d'Excel ou autre logiciel) 500 échantillons de taille 5 et tâchez de répondre aux questions suivantes du mieux que possible, par simulation. La population est considérée infinie, ce qui veut dire que σ et S sont confondus, et que $f \approx 0$.

y	Fréquence
5,0	0,02
5,5	0,03
6,0	0,04
6,5	0,06
7,0	0,10
7,5	0,12
8,0	0,13
8,5	0,10
9,0	0,07
9,5	0,05
10,0	0,05

y	Fréquence
10,5	0,04
11,0	0,04
11,5	0,03
12,0	0,03
12,5	0,02
13,0	0,02
13,5	0,01
14,0	0,01
14,5	0,01
15,0	0,01
15,5	0,01

a) Calculer la moyenne $\bar{y}$ de chaque échantillon; puis calculer la moyenne $\bar{\bar{y}}$ de ces moyennes, $\bar{\bar{y}}$ est une estimation de $E(\bar{y}) = \mu$ ($\bar{\bar{y}}$ aurait été *exactement* égale à μ si on avait tiré *tous* les échantillons possibles.)

b) Calculez la variance des $\bar{y}$, soit $s^2_{\bar{y}} = \dfrac{\sum_{i=1}^{500}(\bar{y}_i - \bar{\bar{y}})^2}{500-1}$, où $\bar{y}_i$ est la moyenne du i^e échantillon. Quelle est l'espérance de $s^2_{\bar{y}}$? Énoncez le théorème qui justifie votre réponse.

c) Lorsque n est grand, la distribution de $\bar{y}$ est à peu près symétrique et d'allure assez proche d'une distribution normale. Est-ce le cas lorsque $n = 10$? (Faites un histogramme).

d) Soit $Z = |\bar{y} - \mu|$ l'erreur commise dans l'estimation de μ. Estimer les probabilités suivantes à l'aide des données générées.

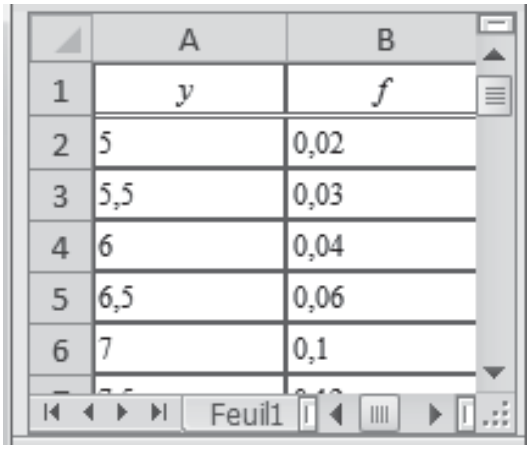
i) P(|Z| > 1,17); ii) P(|Z| > 1,40); iii) P(|Z|> 1,84)

[Théoriquement, ces probabilités sont de 0,9; de 0,95 et de 0,99].

e) Le niveau de l'intervalle de confiance calculé à partir de la formule $\bar{y} - \dfrac{1{,}96s}{\sqrt{9}} \le \mu \le \bar{y} + \dfrac{1{,}96s}{\sqrt{9}}$ est présumé être d'environ 95%. Est-ce que vos données sont conformes à cette supposition ?

f) Si on se base sur la supposition que la population est normale. l'intervalle $\{0{,}47s^2 \le \sigma^2 \le 3{,}33s^2\}$ est un intervalle de confiance à 95 %. Selon votre simulation, cette formule est-elle adéquate?

Instructions pour les calculs à l'aide d'Excel. Supposons que les valeurs de y et f sont placées dans les colonnes A et B. respectivement :



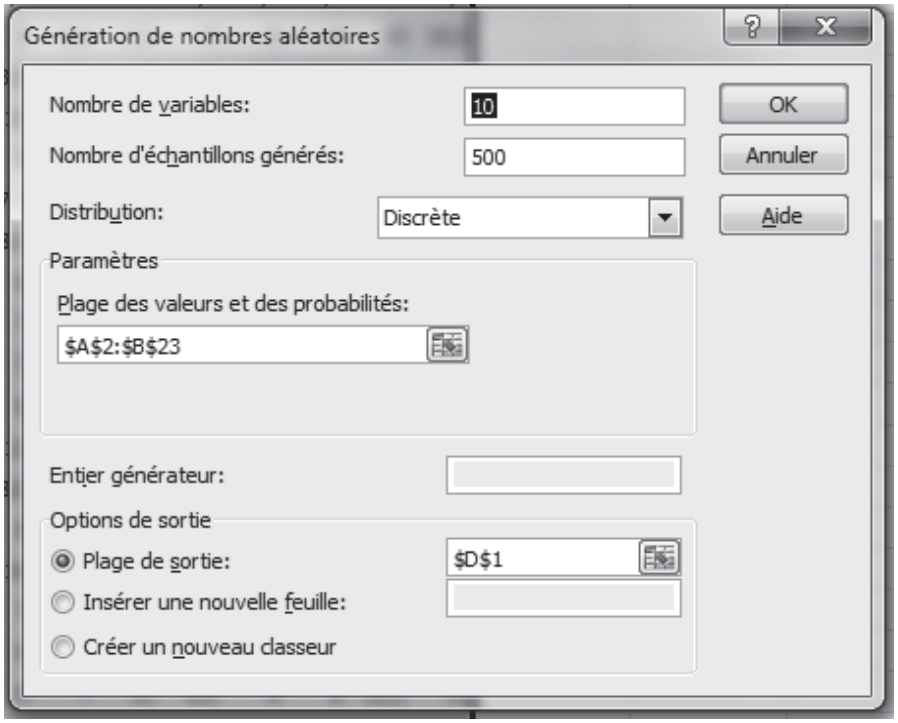
	A	B
1	*y*	*f*
2	5	0,02
3	5,5	0,03
4	6	0,04
5	6,5	0,06
6	7	0,1

Feuil1

On utilise le générateur de nombres aléatoires :

Données>Utilitaire d'analyse>Génération de nombres aléatoires.

Dans la fenêtre qui suit. on indique la position des valeurs de y et f (`Plage des valeurs et des probabilités`), la taille de l'échantillon (`Nombre de variables`), le nombre d'échantillons à générer (`Nombre d'échantillons générés`), et l'emplacement des données générées (`Plage de sortie`). Dans la fenêtre ci-dessous, on génère 500 échantillons de taille 10 qu'on trouvera dans la plage (D1 à M500).



3.22 Soit ω un *éas* de taille n tiré d'une population $U = \{1; 2; \ldots; N\}$, et $\omega_0 \subset U$ un sous-ensemble de U de cardinalité n contenant l'unité k. Montrez que $P(\omega = \omega_0 \mid k \in \omega) = \frac{1}{\binom{N-1}{n-1}}$.

3.23 Supposons que $N = nk$, où k est un entier naturel. Considérez le schème suivant : on dispose au hasard les N unités d'une population dans un tableau de n lignes et k colonnes (la probabilité de chacun des $N!$ tableaux étant de $1/N!$); on prend pour échantillon les n unités d'une colonne choisie au hasard. Montrez que $p(\omega) = \frac{1}{\binom{N}{n}}$ pour tout choix ω de n unités distinctes. [On peut trouver ce résultat évident, mais en faire une démonstration formelle est un bon exercice.]

3.24 On tire *sans* remise un échantillon ω de taille n d'une population de taille N. Puis dans l'échantillon tiré, on tire un échantillon de taille m *avec* remise. Soit $\bar{y}$ la moyenne du premier échantillon et s son écart-type. Soit $\bar{y}^*$ la moyenne du deuxième échantillon (tiré avec remise).

a) Déterminer si $\bar{y}^*$ est sans biais pour la moyenne de la population.

b) Déterminer $V(\bar{y}^* \mid \omega)$, $E[V(\bar{y}^* \mid \omega)]$ et $V[E(\bar{y}^* \mid \omega)]$.

c) Déterminer $V(\bar{y}^*)$ et montrer que $V(\bar{y}^*) \geq V(\bar{y})$.

3.25 On tire un *éas* ω de taille n d'une population U de taille N. Soit U_1 un sous-ensemble de U de taille N_1; Y une variable définie sur U; $\tau_1 = \sum_{k \in U_1} y_k$, et $t_1 = \sum_{i \in \omega_1} y_i$; où n_1 le nombre d'éléments dans $\omega_1 = \omega \cap U_1$. Montrer que pour tout ensemble ω_{1o} de taille n_1, $P(\omega_1 = \omega_{1o} \mid n_1) = \frac{1}{\binom{N_1}{m}}$. En déduire que Nt_1/n est un estimateur sans biais de τ_1.

3.26 (Cochran, 1977, p. 48) Soit ω un *éas* de taille n, et ω_1 un sous-échantillon aléatoire simple de taille n_1 tiré de l'échantillon ω. Soit $\bar{y}$ la moyenne de ω, $\bar{y}_1$ la moyenne de ω_1, et $\bar{y}_2$ la moyenne des $n_2 = n - n_1$ unités de $\omega \setminus \omega_1$. Montrez que :

a) $V(\bar{y}_1 - \bar{y}_2) = \frac{nS^2}{n_1 n_2}$;

b) $C(\bar{y}; \bar{y}_1 - \bar{y}) = 0$.

3.27 (Raj, 1958, p. 227) Supposons qu'on effectue des tirages aléatoires avec remise dans une population de taille N jusqu'à ce qu'on obtienne n unités distinctes. Soit η le nombre de tirages effectués; X_i la i^e valeur observée, $i = 1, 2, \ldots, \eta$; et ω l'ensemble des n unités distinctes. Deux estimateurs possibles sont $\bar{y} = \frac{1}{n}\sum_{i \in \omega} y_i$, la moyenne des unités distinctes, et $\bar{X} = \frac{1}{\eta}\sum_{i=1}^{\eta} X_i$. On admettra que $E(\bar{X} \mid \omega) = \bar{y}$, ce qui peut être démontré par un argument de symétrie. Montrer que

a) $\bar{y}$ et $\bar{X}$ sont sans biais ;

b) $E(\eta) = N\left(\frac{1}{N} + \frac{1}{N-1} + \ldots + \frac{1}{N-n+1}\right)$;

c) Montrer que $V(\bar{X}) = V(\bar{y}) + \sum_{\omega\in\Omega} V(\bar{X} \mid \omega) p(\omega)$ et conclure que $V(\bar{X}) > V(\bar{y})$.

3.28 (Cochran, 1977, p. 48) L'espace échantillon Ω correspondant à un tirage aléatoire simple est l'ensemble de tous les échantillons possibles, au nombre de $\binom{N}{n}$, et la probabilité attribuée à chaque élément de Ω est $1/\binom{N}{n}$. Mais lorsqu'on estime une moyenne ou un total, les propriétés essentielles, c'est-à-dire, l'espérance et la variance de l'estimateur, ne dépendent que des probabilités d'inclusion : la probabilité qu'une unité donnée soit sélectionnée et la probabilité que deux unités données le soient. Ces probabilités sont, respectivement, n/N et $n(n\text{-}1)/[N(N\text{-}1)]$. Il est possible, cependant, de réduire la cardinalité de Ω tout en maintenant ces probabilités d'inclusion. Certains espaces échantillons peuvent être construits selon un modèle de plan d'expérience appelé *plans équilibré en blocs incomplets* (*BIBD*: *Balanced incomplete Bloc design*). Il s'agit d'un espace échantillon qui satisfait les conditions suivantes :

- chaque échantillon est de taille n;
- chaque unité figure dans le même nombre d'échantillons, r;
- chaque paire d'unités figure dans le même échantillon λ fois;

L'espace échantillon est de cardinalité $b \leq \binom{N}{n}$.

Le tableau suivant illustre un tel espace échantillon avec les paramètres suivants : $N = 7$; $n = 3$; $r = 3$; $\lambda = 1$. Il est de cardinalité 7 alors qu'il serait normalement de cardinalité $\binom{7}{3} = 35$.

		Unité						
		1	2	3	4	5	6	7
Échantillon	s_1	*	*		*			
	s_2	*		*				*
	s_3	*				*	*	
	s_4		*	*		*		
	s_5		*				*	*
	s_6			*	*		*	
	s_7				*	*		*

Un tel espace échantillon n'est possible que lorsque N, n, r et λ satisfont certaines contraintes.

a) Montrez que $bn = Nr$ [*Suggestion* : Comptez les étoiles dans un tableau comme celui présenté en exemple. Dans un tableau général (b lignes et N colonnes), le nombre d'étoiles dans chaque ligne multiplié par le nombre de lignes est égal au nombre d'étoiles dans chaque colonne multiplié par le nombre de colonnes].

b) Montrez que $r(n\text{-}1) = \lambda(N\text{-}1)$ [*Suggestion* : Considérez le nombre d'étoiles correspondant (dans l'exemple) aux colonnes 2 à 7, lignes 1 à 3].

c) Montrez que les probabilités d'inclusion sont les mêmes que dans un *éas*.

d) Dans l'exemple, soit {1, 4, 5, 9, 20, 32, 50} l'ensemble des valeurs de Y dans la population. Vérifier numériquement que $\bar{y}$ est sans biais pour la moyenne de la population et que s^2 est sans biais pour S^2.

e) Vous remarquerez que les résultats dépendent d'un choix arbitraire, l'indice des valeurs. Par exemple, les unités de la population auraient pu être énumérées et indicées de la façon suivante : {50, 32, 20, 9, 5, 4, 1}. Vérifiez numériquement que, bien que les moyennes échantillonnales seront différentes, leurs propriétés essentielles ne changent pas.

3.29 Voici un autre exemple du schéma présenté au numéro précédent. La population est de taille $N = 13$, l'échantillon est de taille $n = 4$. Les valeurs d'une variable Y dans la population sont : {7 , 9 , 10 , 19 , 25 , 27 , 39 , 45 , 59 , 60 , 68 , 72 , 74}.

a) Vérifier que les valeurs $r = 4$, $n = 4$, $\lambda = 1$, et $b = 13$ satisfont les contraintes a) et b).

b) Le tableau ci-dessous présente un plan d'échantillonnage *bibd*. Vérifier numériquement que $\bar{y}$ est un estimateur sans biais de μ.

c) Vérifier numériquement que s^2 est un estimateur sans biais de S^2.

Échantillon	Unité 1	2	3	4	5	6	7	8	9	10	11	12	13	Somme	Somme des carrés
ω_1		*		*				*					*	147	7943
ω_2		*	*		*	*								71	1535
ω_3						*		*	*	*				191	9835
ω_4	*				*			*			*			145	7323
ω_5				*	*					*		*		176	9770
ω_6	*	*					*			*				115	5251
ω_7	*					*						*	*	180	11438
ω_8			*				*	*				*		166	8830
ω_9					*		*		*				*	197	11103
ω_{10}			*							*	*		*	212	13800
ω_{11}	*		*	*					*					95	3991
ω_{12}		*							*		*	*		208	13370
ω_{13}				*		*	*				*			153	7235

3.30 (Cochran, 1977, p. 48) Soit ω un *éas* de taille n, ω_1 un sous-échantillon aléatoire simple de taille n_1 tiré dans ω,, et ω^* la réunion de ω et ω_1 ($n + n_1$ observations). Soit $\bar{y}$ la moyenne des données de ω et $\bar{y}^*$ la moyenne des données de ω^*. Démontrer les énoncés suivants:

a) $\bar{y}^*$ est sans biais pour la moyenne de la population;

b) $V(\bar{y}^*) = V(\bar{y}) + \dfrac{n_1(n-n_1)}{(n_1+n)^2}\dfrac{S^2}{n}$.

3.31 Soit $\bar{y}$ la moyenne d'un échantillon tiré avec remise d'une population de moyenne μ et de variance σ^2. Un estimateur de la forme $\hat{\mu} = nw\bar{y}$ ($w > 0$) est biaisé mais pourrait bien avoir une erreur quadratique moyenne inférieure à $V(\bar{y})$.

a) Montrer que l'erreur quadratique moyenne de $\hat{\mu}$ est $EQM(\hat{\mu}) = E[(\hat{\mu}-\mu)^2] = nw^2\sigma^2 + \mu^2(1-nw)^2$.

b) Montrer que la valeur de w qui minimise $EQM(\hat{\mu})$ est $w_0 = \dfrac{1}{n+k^2}$, où $k = \sigma/\mu$.

c) Supposons que le coefficient de variation de la population peut être assez bien estimé à partir de données indépendantes de l'échantillon. Montrer que l'erreur quadratique moyenne de $\hat{\mu}_0 = nw_0\bar{y}$ est $EQM(\hat{\mu}) = \dfrac{\sigma^2}{n+k^2} < V(\bar{y})$.

3.32 Soit U une population de taille N et $U_1 = U\backslash\{k\}$ (la population U de laquelle l'unité k a été soustraite). Soit S l'écart-type de la population et S_1 l'écart-type de U_1.

a) Montrer que $S^2 = \dfrac{N-2}{N-1}S_1^2 + \dfrac{N}{(N-1)^2}(y_k-\mu)^2$ où μ est la moyenne de U.

b) Si l'unité k est obligatoirement incluse dans l'échantillon, et $\bar{y}_1$ est la moyenne d'un échantillon de taille n-1 tiré dans U_1, montrer que $\hat{\mu} = \dfrac{y_k + (N-1)\bar{y}_1}{N}$ est un estimateur sans biais de la moyenne de la population.

3.33 Montrez que la procédure suivante donne un *éas* :

- On tire un échantillon de taille n avec remise;
- Si les n unités sont distinctes, on garde l'échantillon;
- Si l'échantillon contient des répétitions, on recommence à l'étape 1.

Déterminez la variance et l'espérance du nombre d'essais qu'il faudra effectuer pour enfin obtenir un échantillon sans répétitions. Montrez que l'espérance tend vers 1 et la variance vers 0 lorsque $N \to \infty$.

3.34 Supposons que pour tirer un *éas* de taille n d'une population de taille N, on prélève d'abord un *éas* ω^* de taille $n^* > n$, puis un *éas* ω de taille n dans ω^*. Soit $\bar{y}$ la moyenne de l'échantillon ω.

a) Montrer que $E(\bar{y}) = \mu$.

b) Montrer que $V(\bar{y}) = \frac{S^2}{n}(1-f)$ [*Suggestion* : Commencer par calculer l'espérance et la variance conditionnelles de $\bar{y}$ étant donné ω^*, puis utiliser le Théorème 1.12].

c) Démontrer le résultat plus général suivant : pour tout ensemble ω_0 de n unités, $P(\omega = \omega_0) = \frac{1}{\binom{N}{n}}$.

3.35 Certains modes de tirage donnent un échantillon dont la taille est aléatoire. Par exemple, on peut procéder de la façon suivante : pour chaque unité de la population, on décide à l'aide d'un mécanisme aléatoire indépendant de l'inclure ou non dans l'échantillon. Soit π_k la probabilité que l'unité k soit sélectionnée, $k = 1, 2, \ldots, N$. Soit η la taille (aléatoire) de l'échantillon.

a) Montrer que $\hat{\tau} = \sum_{i\in\omega} \frac{y_i}{\pi_i}$ est sans biais pour le total τ de la population. [*Suggestion* : définir les variables $Z_k = \begin{cases} 1 \text{ si } k \in \omega \\ 0 \text{ sinon} \end{cases}$ et exprimer l'estimateur comme $\hat{\tau} = \sum_{k=1}^{N} \frac{y_k}{\pi_k} z_k$].

b) Montrer que la variance de $\hat{\tau}$ est $V(\hat{\tau}) = \sum_{k=1}^{N} \frac{1-\pi_k}{\pi_k} y_k^2$.

c) Montrer que $\hat{V}(\hat{\tau}) = \sum_{i\in\omega} \frac{1-\pi_i}{\pi_i^2} y_i^2$ est un estimateur sans biais de $V(\hat{\tau})$.

d) Supposons que $\pi_k = \pi$ pour tout k.

i) Montrer que, conditionnellement au nombre η d'unités, ω a les mêmes propriétés qu'un *éas*. En d'autres termes, que pour tout ensemble ω_0 de n unités, $P(\omega = \omega_0 \mid \eta = n) = 1/\binom{N}{n}$.

ii) Montrer que $V[E(\hat{\tau} \mid \eta)] = \frac{\tau^2(1-\pi)}{N\pi}$.

iii) Montrer que $E[V(\hat{\tau} \mid \eta)] = \frac{1-\pi}{\pi}(N-1)S^2$.

iv) Développer la formule de la variance de $\hat{\tau}$ à partir de ii) et iii).

v) Montrer que $\hat{S}^2 = \frac{1}{N}\sum_{i\in\omega}\frac{y_i^2}{\pi} - \frac{1}{N(N-1)}\sum_{i\in\omega}\sum_{j\in\omega, j\neq i}\frac{y_i y_j}{\pi^2}$ est un estimateur sans biais de S^2.

3.36 [(Särndal, 1972)] Considérez un *éas* tiré d'une population U de taille N dont l'une des valeurs, y_1, est très petite et une autre, y_N, est très grande. Considérer l'estimateur suivant de la moyenne μ de la population :

$$\hat{\mu} = \begin{cases} \bar{y} + c & \text{si } y_1 \in \omega \text{ et } y_N \notin \omega \\ \bar{y} - c & \text{si } y_N \in \omega \text{ et } y_1 \notin \omega \\ \bar{y} & \text{dans tous les autres cas} \end{cases}$$

où c est une constante positive. Soit A l'événement « $y_1 \in \omega$ » et B l'événement « $y_N \in \omega$ ».

a) Démontrer les résultats suivants.

i) $P(A\backslash B) = P(B\backslash A) = \frac{n(N-n)}{N(N-1)}$;

ii) $P(A\cap B) = \frac{n(n-1)}{N(N-1)}$;

iii) $P(A^c\cap B^c) = \frac{(N-n)(N-n-1)}{N(N-1)}$.

b) Démontrer les résultats suivants.

i) $E(\hat{\mu} \mid A\backslash B) = \frac{y_1}{n} + \frac{n-1}{n}\mu_1 + c$, où $\mu_1 = \frac{\sum_{i=2}^{N-1} y_i}{N-2}$;

ii) $E(\hat{\mu} \mid B\backslash A) = \frac{y_N}{n} + \frac{n-1}{n}\mu_1 - c$;

iii) $E(\hat{\mu} \mid A \cap B) = \frac{y_1 + y_N}{n} + \frac{n-2}{n}\mu_1$;

iv) $E(\hat{\mu} \mid A^c \cap B^c) = \mu_1$.

c) Utiliser les résultats ci-dessus pour montrer que $\hat{\mu}$ est sans biais pour μ.

d) Démontrer les résultats suivants :

i) $V(\hat{\mu} \mid A\backslash B) = V(\hat{\mu} \mid B\backslash A) = \frac{(n-1)S_1^2}{n^2}\left(1 - \frac{n-1}{N-2}\right)$, où S_1^2 est la variance de la population réduite $U_1 = U \setminus \{y_1 ; y_N\}$;

ii) $V(\hat{\mu} \mid A \cap B) = \frac{(n-2)S_1^2}{n^2}\left(1 - \frac{n-2}{N-2}\right)$;

iii) $V(\hat{\mu} \mid \mathrm{A}^c \cap \mathrm{B}^c) = \frac{S_1^2}{n}\left(1 - \frac{n}{N-2}\right)$.

Montrer que l'espérance de la variance conditionnelle est $\frac{(N-3)(N-n)}{nN(N-1)} S_1^2$.

e) Considérer les données suivantes sur une variable Y définie sur une population (tableau A02, Y : Superficie en 2011).

N = 1 183, μ = 570,017 et S = 8 850,919. La plus petite donnée est 0,0358 et la plus grande est 297 332,8. La population amputée de ces deux valeurs est de moyenne μ_1 = 319,2187 et d'écart-type S_1 = 1 942,749.

Considérer l'estimateur $\hat{\mu}$ proposé dans ce numéro avec n = 45 et c = 1 000. Vérifiez que

i) $E[V(\hat{\mu}]$ = 80 545,79;

ii) $V[E(\hat{\mu})]$ = 1 183 377;

iii) $V(\hat{\mu})$ = 1 263 923.

Comparer avec la variance de l'estimateur habituel $\bar{y}$.

4 Estimation d'une proportion

Soit U une population de taille N et $C \subset U$ une classe d'unités possédant une caractéristique particulière. Par exemple, C pourrait être l'ensemble des créances *en souffrance* dans une population de comptes à recevoir; les pièces *défectueuses* dans une livraison de N pièces; ou les jeunes de moins de 20 ans parmi les clients d'une compagnie d'assurance auto. On peut s'intéresser à estimer le nombre N_c d'éléments dans C ou, plus souvent, à la *proportion* $p = \frac{N_c}{N}$ d'unités appartenant à C.

Nous définirons les estimateurs $\hat{p}$ et $\hat{N}_c$ de p et de N_c, respectivement, et nous développerons leurs propriétés essentielles: on montrera qu'ils sont sans biais, on développera les formules de $V(\hat{p})$ et $V(\hat{N}_c)$ et de leurs estimateurs $\hat{V}(\hat{p})$ et $\hat{V}(\hat{N}_c)$.

Variable dichotomique

Strictement parlant, ces propriétés ont déjà été établies, car p et N_c ne sont que des cas particuliers de moyenne μ et de total τ — des paramètres déjà traités au Chapitre 3. Ce cas est particulier dans le sens qu'il se limite à une variable *dichotomique*: une variable qui ne prend que les valeurs 0 et 1. Soit Y une variable définie comme ceci:

$$y_k = \begin{cases} 1 & \text{si } k \in C \\ 0 & \text{sinon} \end{cases}, \; k = 1, 2, \ldots, N.$$

Alors p et N_c ne sont que la moyenne μ et le total τ de Y :

$$\mu = \frac{\sum_{k=1}^{N} y_k}{N} = p \; ; \; \tau = \sum_{k=1}^{N} y_k = N_c.$$

Donc tous les théorèmes concernant p et N_c découlent des théorèmes concernant μ et τ, de sorte qu'on pourrait, à la limite, se fier aux formules du Chapitre 3 pour estimer les paramètres et leur variance. On a intérêt, cependant, à traiter ce cas particulier séparément. D'abord parce que le caractère dichotomique de Y permet d'exprimer les variances et leur estimateur plus simplement. Ensuite, parce dans le cas d'une variable dichotomique, la distribution des estimateurs est connue: à une constante multiplicative près, les estimateurs de p et de N_c sont de loi hypergéométrique. Par conséquent, un

intervalle de confiance, jusqu'ici basé sur des lois asymptotiques (donc approximatives) pourra, lorsque nécessaire, être déterminé sans recours à des approximations.

Il est utile d'établir le lien entre les quantités définies au Chapitre 3 et leur expression en termes de n, N et p. Outre la proportion p qui prend la place de $\bar{y}$ et N_c, qui prend la place de τ, il y a la variance S^2, qui s'exprime comme ceci:

(4.1) $$S^2 = \frac{Np(1-p)}{N-1} .$$

4.1 GRAND ÉCHANTILLON

Lorsque n est grand, on établit les propriétés des estimateurs de p et de N_c simplement en ré-exprimant les théorèmes du dernier chapitre dans les termes d'une variable dichotomique

ESTIMATION DE P

Supposons que dans un échantillon simple de taille n on trouve n_c unités appartenant à C. Naturellement, on estimera p par son analogue échantillonnal, soit

$$\hat{p} = \frac{n_c}{n} .$$

Par le Théorème 3.1, $\hat{p}$ est sans biais, et, substituant l'expression (4.1) à S^2 dans la formule de $V(\bar{y})$, on obtient la variance de $\hat{p}$:

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

Par la version échantillonnale de l'équation (4.1), soit $s^2 = \frac{n\hat{p}(1-\hat{p})}{n-1}$, et le Théorème 3.2 et son corollaire, un estimateur sans biais de $V(\hat{p})$ est

(4.3) $$\hat{V}(\hat{p}) = (1-f)\frac{\hat{p}(1-\hat{p})}{n-1} ,$$

où $f = n/N$.

Finalement, l'équation (3.3), adaptée au contexte présent donne la formule d'un intervalle de confiance à 100(1-α) % pour p:

(4.4) $$\hat{p} - z_{\alpha/2}\sqrt{\hat{V}(\hat{p})} \le p \le \hat{p} + z_{\alpha/2}\sqrt{\hat{V}(\hat{p})} .$$

Exemple 4.1 *Intervalle de confiance pour p*

On tire un *éas* de 90 étudiants choisis parmi les étudiants de la faculté des sciences d'une université dont l'effectif est N = 988. On leur demande, entre autres, s'ils croient à la théorie de l'évolution. Parmi les 90 répondants, 33 se sont déclarés d'accord. Soit p la proportion

d'étudiants du programme qui sont d'accord. Déterminer un intervalle de confiance à 95 % pour p.

Solution

L'estimation $\hat{p}$ de p est

$$\hat{p} = \frac{33}{90} = 0{,}3667.$$

L'estimation de l'écart-type de $\hat{p}$ est

$$\sqrt{\hat{V}(\hat{p})} = \sqrt{1 - \frac{90}{988}} \sqrt{\frac{\left(\frac{33}{90}\right)\left(1 - \frac{33}{90}\right)}{90 - 1}} = 0{,}04870.$$

Voici, finalement, un intervalle de confiance à 95% pour p :

$$0{,}3667 \pm 1{,}96(0{,}04870) = 0{,}3667 \pm 0{,}09545.$$

On estime donc que la proportion de personnes dans le programme qui croient à l'évolution est de 36,67 %, avec une marge d'erreur de 9,545 %. Ou encore, que 27,1 % $\leq p \leq$ 46,2 %.

Remarque *Le facteur de correction*

Dans certaines situations, dont celle du dernier exemple, le facteur de correction $\sqrt{1 - \frac{90}{988}}$ peut raisonnablement être omis. Le paramètre p qui a été estimé dans l'exemple a été présenté comme la proportion des étudiants du programme de la faculté des sciences qui sont d'accord avec la proposition. Il est rare que ce soit vraiment cela qu'on tente d'estimer. Une telle recherche aurait vraisemblablement été motivée par la question suivante : « Quelle est la proportion des étudiants de science qui croient à la théorie de l'évolution? » La population visée est donc plus vaste que l'ensemble des étudiants du programme. Elle est en fait infinie et p est une probabilité et non une proportion.

Cette extrapolation — depuis la population des étudiants de la faculté en question à la population des « étudiants de science » — est raisonnable et peut être défendue par des arguments de bon sens, mais elle n'est pas sanctionnée par la statistique.

ESTIMATION DE N_C

L'estimateur de N_c et ses propriétés découlent immédiatement de ce qui précède. Puisque $N_c = Np$, les résultats suivants sont immédiats :

- L'estimateur $\hat{N}_c = N\hat{p}$ est sans biais pour N_c;
- La variance de $\hat{N}_c$ est $V(\hat{N}_c) = N^2 V(\hat{p}) = N^2 \frac{N-n}{N-1} \frac{p(1-p)}{n}$;

- Un estimateur sans biais de la variance de $\hat{N}_c$ est $\hat{V}(\hat{N}_c) = N^2(1-f)\dfrac{\hat{p}(1-\hat{p})}{n-1}$.

Un intervalle de confiance à 100(1-α) % pour N_c est donné approximativement par

$$\hat{N}_c - z_{\alpha/2}\sqrt{\hat{V}(\hat{N}_c)} \le N_c \le \hat{N}_c + z_{\alpha/2}\sqrt{\hat{V}(\hat{N}_c)} .$$

Une borne pour la variance

La variance de $\bar{y}$ dépend d'une variance S^2 qui, pour une variable quelconque, n'a pas de limite supérieure évidente. Mais lorsqu'il s'agit d'une variable dichotomique, la variance de $\hat{p}$ est bornée. Le produit $p(1-p)$, pour p dans l'intervalle [0 ; 1], atteint son maximum lorsque $p = 1/2$. Donc $p(1-p) \le 1/4$ et alors

$$V(\hat{p}) = \frac{N-n}{N-1}\frac{p(1-p)}{n} \le \frac{N-n}{N-1}\frac{1/4}{n} \le \frac{1}{4n} .$$

La loi de $\hat{p}$

Notons le fait que la distribution de $\hat{p}$ est connue : son numérateur n_c est la valeur d'une variable aléatoire X de loi hypergéométrique de paramètres n, N_c et $N - N_c$. Sa variance est donc, selon le Théorème 1.15, $V(X) = n\dfrac{N-n}{N-1}p(1-p)$, et l'équation (4.2) en découle. Si la population est grande au point d'être, à toute fin pratique, infinie, alors X est de loi *binomiale* de paramètres n et p et sa variance est $np(1-p)$ (voir le Théorème 1.13).

4.2 DIFFÉRENCE DE PROPORTIONS

Soit p_1 et p_2 deux proportions dont on voudrait estimer la différence $p_2 - p_1$. Nous allons illustrer deux situations. Dans les deux cas, il existe des estimateurs sans biais $\hat{p}_1$ et $\hat{p}_2$ et donc aussi un estimateur sans biais de la différence $p_2 - p_1$, soit $\hat{p}_2 - \hat{p}_1$. Mais les deux situations diffèrent quant à la variance de $\hat{p}_2 - \hat{p}_1$. Dans le premier cas, $\hat{p}_1$ et $\hat{p}_2$ sont calculés à partir d'échantillons distincts et sont donc indépendants; la variance de la différence $\hat{p}_2 - \hat{p}_1$ est alors égale à la somme des variances: $V(\hat{p}_2 - \hat{p}_1) = V(\hat{p}_2) + V(\hat{p}_1)$. Dans le deuxième cas, les proportions $\hat{p}_1$ et $\hat{p}_2$ réfèrent à deux classes différentes, C_1 et C_2, mais sont calculées à partir des mêmes unités. Ce qui fait qu'elles ne sont pas indépendantes et que la variance de leur différence n'est plus la somme des variances. La variance peut néanmoins être déterminée dans le contexte que nous décrirons plus bas.

Échantillon indépendants

Soit p_1 et p_2 des proportions associées à deux populations U_1 et U_2 et $C_1 \subset U_1$ et $C_2 \subset U_2$ deux classes particulières. Par exemple, U_1 est l'ensemble des logements habités par leur propriétaire et U_2 l'ensemble des logements habités par des locataires; C_1 et C_2 sont les logements jugés insalubres; p_1 et p_2 sont les proportions de logements jugés insalubres.

On tire un échantillon dans chacune des deux populations. Soit

- N_1 et N_2 les tailles des populations U_1 et U_2;
- p_1 et p_2 les proportions d'unités appartenant à C_1 et C_2, respectivement;
- n_1 et n_2 les tailles des échantillons tirés dans U_1 et U_2, respectivement;
- n_{1c} et n_{2c} le nombre d'unités dans les échantillons appartenant à C_1 et C_2, respectivement.

Soit $\hat{p}_1 = \frac{n_{1c}}{n_1}$ et $\hat{p}_2 = \frac{n_{2c}}{n_2}$ les estimateurs usuels de p_1 et p_2. La différence p_2 - p_1 est estimée sans biais par $\hat{p}_2 - \hat{p}_1$. Étant donné que les échantillons sont tirés indépendamment l'un de l'autre, les variables $\hat{p}_1$ et $\hat{p}_2$ sont indépendantes et, par conséquent, $V(\hat{p}_2 - \hat{p}_1) = V(\hat{p}_2) + V(\hat{p}_1) = \frac{N_2 - n_2}{N_2 - 1}\frac{p_2(1-p_2)}{n_2} + \frac{N_1 - n_1}{N_1 - 1}\frac{p_1(1-p_1)}{n_1}$. On estime $V(\hat{p}_2 - \hat{p}_1)$ par

$$\hat{V}(\hat{p}_2 - \hat{p}_1) = \hat{V}(\hat{p}_2) + \hat{V}(\hat{p}_1) = (1 - f_2)\frac{\hat{p}_2(1-\hat{p}_2)}{n_2 - 1} + (1 - f_1)\frac{\hat{p}_1(1-\hat{p}_1)}{n_1 - 1},$$

où $f_1 = \frac{n_1}{N_1}$ et $f_2 = \frac{n_2}{N_2}$.

Si les échantillons sont assez grands, la variable

$$Z = \frac{(\hat{p}_2 - \hat{p}_1) - (p_2 - p_1)}{\sqrt{\hat{V}(\hat{p}_2 - \hat{p}_1)}}$$

est de loi à peu près normale centrée-réduite. Un intervalle de confiance approximatif à 100(1 - α) % est donné par

$$(\hat{p}_2 - \hat{p}_1) - z_{\alpha/2}\sqrt{\hat{V}(\hat{p}_2 - \hat{p}_1)} \le p_2 - p_1 \le (\hat{p}_2 - \hat{p}_1) + z_{\alpha/2}\sqrt{\hat{V}(\hat{p}_2 - \hat{p}_1)}.$$

Exemple 4.2 *Différence de deux proportions — échantillons indépendants*

Parmi les clients d'une compagnie d'assurance auto se trouvent 6 175 étudiants en première année universitaire. Quatre ans plus tard, on sépare les dossiers en deux groupes. Le groupe 1 comprend les 4 855 clients qui n'ont fait aucune réclamation; le groupe 2 comprend les 1 320 qui en ont fait au moins une. On prélève donc un échantillon de 400 clients du premier groupe et un échantillon de 200 étudiants du deuxième. On constate que 196 (49 %) des étudiants ont abandonné leurs études avant d'obtenir leur diplôme, comparé à 104 (52 %) dans le deuxième groupe. Soit p_1 et p_2 les proportions correspondantes dans la population.

a) Déterminer un intervalle de confiance à 95 % pour la différence p_2 - p_1.

b) Peut-on conclure qu'il y a une réelle différence entre les deux groupes?

Solution

a) $\hat{p}_1$ = 0,49, $\hat{p}_2$ = 0,52, $\hat{p}_2 - \hat{p}_1$ = 0,03. $\sqrt{\hat{V}(\hat{p}_2 - \hat{p}_1)} = \sqrt{(0,02397319)^2 + (0,03262254)^2}$ = 0,04048. L'intervalle de confiance à 95 % (à peu près) est donc [0,03-1.96(0,04048) ; 0,03+1,96(0,04048) = [-0,049 ; 0,1093].

b) Selon l'intervalle de confiance, on ne peut pas affirmer (à 95 % de confiance) qu'il y a une réelle différence entre les deux groupes.

Données appariées

Les estimateurs $\hat{p}_1$ et $\hat{p}_2$ ne sont pas indépendants s'ils sont calculés à partir des mêmes unités échantillonnales. Par exemple, on tire un échantillon de citoyens afin d'estimer le pourcentage de personnes en faveur de la gratuité scolaire. Un an plus tard, on pose la même question aux mêmes personnes. Les paramètres p_1 et p_2 sont les proportions en faveur de la gratuité scolaire lors du premier et du deuxième sondage, respectivement. Les estimateurs de p_1 ,- p_2 et p_2 - p_1 sont encore $\hat{p}_1$,- $\hat{p}_2$ et $\hat{p}_2$ - $\hat{p}_1$: ils demeurent sans biais. Ici, cependant, les estimateurs, ayant été calculés à partir des mêmes répondants, ne sont plus indépendants et donc $V(\hat{p}_2 - \hat{p}_1) \neq V(\hat{p}_1) + V(\hat{p}_2)$. Pour déterminer et estimer la variance, on réduit les deux échantillons à échantillon unique. Si Y_1 et Y_2 sont les deux variables dichotomiques de moyennes p_1 et p_2, alors la variable $Y = Y_2 - Y_1$ est une variable de moyenne $p_2 - p_1$. Le problème revient donc à estimer la moyenne d'une seule variable Y, ce qui peut se faire à l'aide des techniques de la section 3.5.

Exemple 4.3 *Différence de deux proportions — données appariées*

On tire un échantillon de taille 160 parmi les 4 235 membres d'un syndicat. On leur pose la question suivante: « Appuyez-vous les demandes de vos négociateurs syndicaux? ». On trouve que 70 répondants ont dit « Oui ». Quelque temps plus tard, à la suite de certains affrontements, on pose la même question aux mêmes individus; 60 disent « Oui ».

a) Estimer $p_2 - p_1$ où p_1 et p_2 sont les proportions de « Oui » avant et après.

b) Estimer l'écart-type de l'estimateur.

Solution

a) $\hat{p}_1$ = 70/160; $\hat{p}_2$ = 60/160; $\hat{p}_2$ - $\hat{p}_1$ = -0,0625.

b) On ne peut pas estimer l'écart-type de $\hat{p}_2$ - $\hat{p}_1$ sans information supplémentaire. Celle-ci est contenue dans le tableau suivant:

		Deuxième sondage		
		Oui	Non	
Premier sondage	Oui	40	30	70
	Non	20	70	90
		60	100	160

Soit Y_1 une variable qui prend la valeur 1 si la réponse est « Oui » au 1er sondage et 0 sinon; et soit Y_2 une variable dichotomique définie de façon analogue pour le 2e sondage. La différence $Y = Y_2 - Y_1$ prend les valeurs -1, 0 et 1 avec effectifs 30, 110, et 20, respectivement. On a alors les résultats suivants: $\bar{y}$ = -0,0625, s_y = 0,5572563, $\sqrt{\hat{V}(\bar{y})}$ = 0,04321476. L'intervalle de confiance à 95 % (à peu près) est [-0,147 ; 0,022]. Ce résultat ne permet pas de conclure que l'appui des syndiqués a changé (en moyenne: certains ont manifestement changé d'opinion, dans les deux sens, mais la différence moyenne peut être nulle si l'effet des uns annule celui des autres).

Remarque

Dans le dernier exemple, si le seul objectif avait été de savoir si $p_1 = p_2$, on aurait pu procéder autrement. Si on se limite aux n = 50 personnes qui ont changé d'avis, on constate qu'il y en a 30 qui ont passé de oui à non et 20 qui ont passé de non à oui. Si la proportion globale d'appui aux négociateurs n'a pas changé, la proportion p (parmi ceux qui ont changé d'avis) qui ont passé de non à oui devrait égaler ½. Un intervalle de confiance à 95 % pour p, étant donné que n = 50, est, ignorant le facteur de correction, [0,263 ; 0,537]. Puisque la proportion $p = 1/2$ n'est pas exclue de l'intervalle on ne peut encore pas conclure qu'il y a eu, globalement, un changement dans les appuis aux négociateurs.

4.3 DÉTERMINATION DE LA TAILLE DE L'ÉCHANTILLON

Comment déterminer, avant de tirer un échantillon, la taille nécessaire pour que p soit estimée avec une marge d'erreur donnée? Le problème a été traité au Chapitre 3 pour une moyenne et peut l'être ici de la même manière puisque une proportion est une moyenne. Nous avons montré dans la section 3.6 que pour estimer une moyenne avec une marge d'erreur égale à E, la taille doit être approximativement égale à $n_o = \left(\frac{z_{\alpha/2}S}{E}\right)^2$ [(3.9), ajustée ensuite par la transformation $n = \frac{n_o}{1+\frac{n_o}{N}}$].

Donc, utilisant l'expression (4.1) de S^2, nous trouvons que pour que la marge d'erreur soit égale à E il faut que $n_0 = \frac{z^2_{\alpha/2}}{E^2}S^2 = \frac{z^2_{\alpha/2}}{E^2}\frac{N}{N-1}p(1-p)$. Omettant le facteur $\frac{N}{N-1}$, on obtient

(4.5) $$n_o = \frac{z^2_{\alpha/2}p(1-p)}{E^2}.$$

Ici aussi la formule fait intervenir un paramètre inconnu, p. Notons, cependant, que c'est le produit $p(1-p)$ qu'il faut connaître et non p. À défaut de connaître ou de pouvoir estimer ce produit, on peut considérer l'hypothèse pessimiste qui consiste à remplacer $p(1-p)$ par sa

valeur maximale, soit ¼, atteinte lorsque $p = 1/2$. Donc l'estimation la plus pessimiste qu'on puisse obtenir de n_0 est celle où l'on remplace p par 1/2. On a alors :

$$n_o = \frac{z^2_{\alpha/2}}{4E^2}\,.$$

Exemple 4.4 *Taille d'échantillon maximale pour une marge d'erreur absolue donnée*

Le responsable de marketing dans une compagnie de boissons gazeuses veut estimer la proportion p de consommateurs qui préfèrent 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 (à 95 %) de son estimateur ne soit pas supérieure à 0,04 ?

Solution

La valeur de n_o est $\frac{(1,96)^2}{4(0,04)^2} = 600$. Nous ne ferons pas d'ajustement, puisque la population est essentiellement infinie. Donc si on prélève un échantillon de taille 600, l'écart-type de l'estimateur $\hat{p}$ ne sera pas supérieur à 0,02.

La valeur $n = 600$ est presque certainement une surestimation. À moins que p soit exactement égale à ½, la marge d'erreur est inférieure à 0,04 [on dira aussi 4 *points de pourcentage*]. Le tableau suivant donne la taille de l'échantillon nécessaire sous plusieurs hypothèses concernant p.

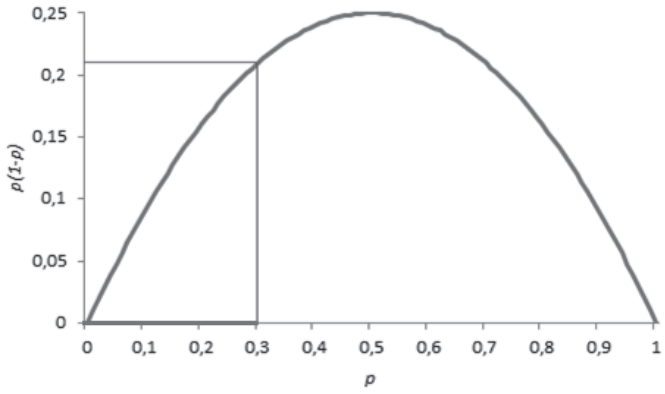
p	0,01	0,05	0,10	0,20	0,30	0,40	0,50
n_o	24	114	216	384	504	576	600

On constate que l'erreur de surestimation n'est pas grave lorsque p n'est pas trop éloignée de 0,5. Mais lorsque p est petit, un échantillon de taille 600 est beaucoup plus grand que nécessaire. Par exemple, si $p = 0,2$, un échantillon de taille 384 suffit; un échantillon de 600 est inutilement coûteux.

Il est rare en pratique qu'on soit démuni au point de ne pouvoir estimer p qu'en le remplaçant par 1/2. On a souvent une idée de la valeur de p, assez bonne pour au moins fixer une borne à sa valeur, soit supérieure, soit inférieure. Par exemple, un service de vérification interne qui se propose de tirer un échantillon afin d'estimer la proportion de factures erronées dans l'année possède une assez bonne connaissance du contexte, basée sur des données historiques, pour être sûr que ce pourcentage n'approchera pas 50 %. Il pourrait, par exemple, prévoir avec assurance que le pourcentage est inférieur à, disons, 30 %. Comme le montre la Figure 4.1, le produit $p(1-p)$ sera alors inférieur à (0,30)(0,70) = 0,21.

La fonction $p(1-p)$ étant croissant l'intervalle [0 ; 1/2], dans l'intervalle [0 ; 0,3[le produit $p(1-p)$ est inférieur à (0,30)(0,70) = 0,21. En général, si p est inférieure à un certain nombre $p_o < 1/2$, ou si p est supérieure à un nombre $p_o > 1/2$ alors $p(1-p) < p_o(1-p_o)$.

Figure 4.1
La fonction *p*(1-*p*)



Erreur relative

La marge d'erreur s'exprime normalement en termes relatifs, c'est-à-dire, comme pourcentage de *p*. Dans l'Exemple 4.4, la marge d'erreur souhaitée n'aurait pas été fixée à 0,04 si on estimait que *p* est aux alentours de 0,02. Lorsque *p* est petit, la marge exigée est étroite, possiblement proportionnelle à *p*. En d'autres termes, l'exigence est de la forme

$$z_{\alpha/2}\sqrt{V(\hat{p})} = pR\ .$$

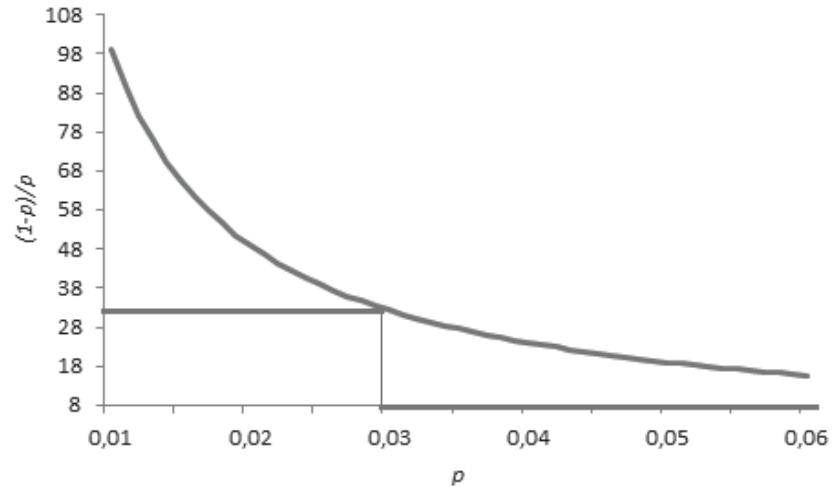
Donc *pR* remplace *E* et on a alors ceci :

$$n_o \approx \frac{z_{\alpha/2}^2 p(1-p)}{(pR)^2} = \left(\frac{z_{\alpha/2}}{R}\right)^2 \frac{(1-p)}{p}\ .$$

L'ajustement pour une population finie est $n = \dfrac{n_o}{1+\dfrac{n_o}{N}}$. Si la valeur de *p* ne peut pas être estimée, il faut trouver un moyen de borner supérieurement la quantité $\dfrac{1-p}{p}$. Or celle-ci est une fonction décroissante de *p* (voir la Figure 4.2), ce qui veut dire que $\dfrac{1-p}{p}$ est à son maximum lorsque *p* est à son minimum réaliste.

Figure 4.2

La fonction (1-*p*)/*p*



Exemple 4.5 *Taille d'échantillon – marge absolue*

On voudrait estimer le pourcentage de pièces défectueuses dans un lot de 5 000 pièces. Quelle est la taille de l'échantillon qu'il faut prélever pour une marge d'erreur (à 95 %) de 0,04 ?

On sait que, normalement, le taux de défectuosité est de l'ordre de 5 %, et on estime qu'il est encore de cet ordre-là. Pour être sûrs, on la pose à 10 %, bien au-dessus de ce qu'on croit être sa valeur.

Solution

Nous avons $n_o = \left(\frac{z_{\alpha/2}}{0,04}\right)^2 p(1-p)$, et puisque le numérateur prend au maximum la valeur (0,1)(0,9) = 0,09, on a $n_o = \left(\frac{1,96}{0,04}\right)^2 p(1-p) = 216$. L'ajustement donne $n = 216/(1+216/5\,000) \approx 207$.

Remarquez que si on avait pris la valeur 1/2 pour p, nous aurions obtenu $n_o = 600,25$ et $n = 536$, beaucoup plus grand que nécessaire.

Exemple 4.6 *Taille d'échantillon – marge relative*

Dans le même contexte que dans l'Exemple 4.5, déterminer la taille de l'échantillon qu'il faut prélever pour une marge *relative* de 25 %. On estime encore que $p \approx 0,05$, mais pour être sûrs, cette fois-ci, on la pose à 0,03.

Solution

Ce que vous voulez, c'est que $z_{\alpha/2}\sqrt{V(\hat{p})} = 0{,}25p \Rightarrow n_o = \left(\frac{1{,}96}{0{,}25}\right)^2 \frac{1-p}{p}$. La plus grande valeur de $\frac{1-p}{p}$ est atteinte lorsque p est à son minimum, comme on le voit dans la Figure 4.2. Ce minimum est 0,03. On prendra donc $n_o = \left(\frac{1{,}96}{0{,}25}\right)^2 \frac{1-0{,}03}{0{,}03} = 1\,987$ et $n = 1422$.

4.4 PETIT ÉCHANTILLON

L'approximation normale est justifiée du fait de n_c est la valeur observée d'une variable de loi hypergéométrique, que lorsque N est assez grand (ce qui est souvent le cas), la loi hypergéométrique peut être approchée par la loi binomiale, et celle-ci à son tour peut être approchée par la loi normale. Tout cela à condition que n soit grand et p pas trop éloignée de ½. La plupart du temps, c'est la deuxième condition qui fait défaut : une probabilité p excessivement petite. C'est le cas en vérification, où p représente la proportion de transactions erronées; ou en contrôle de production, où p est la proportion de pièces défectueuses; ou la proportion de réclamations de frais médicaux de plus de 100 000 $. Lorsque p est très petit, l'approximation normale est inadéquate. Il est préférable dans ce cas de déterminer l'intervalle de confiance pour p en faisant appel à la loi exacte : la loi hypergéométrique (voir le Théorème 1.15).

Nous décrivons la procédure pour N_c, laquelle évidemment mène directement à un intervalle de confiance pour p : l'intervalle de confiance $a \le N_c \le b$ est équivalent à $\frac{a}{N} \le p \le \frac{b}{N}$.

PROCÉDURE DE CONSTRUCTION D'UN INTERVALLE DE CONFIANCE POUR N_c.

Nous commençons par construire un intervalle de confiance *unilatéral* à gauche, c'est-à-dire, un intervalle de la forme

$$N_c \le b \text{ ou } p \le b/N.$$

Ce n'est pas seulement pour des raisons pédagogiques que nous commençons là. C'est aussi parce que la borne supérieure est parfois la seule qui compte : si p est un taux d'erreur, ou de défectuosité, ou de dérogations diverses, on veut s'assurer qu'il *n'excède pas* une limite *supérieure* tolérable. On ne craint pas que p soit trop petit.

La procédure

Soit X une variable aléatoire de loi hypergéométrique de paramètres n, N_c, et N - N_c. Supposons que dans un échantillon de taille n on observe $X = n_c$ unités appartenant à la classe C. L'intervalle de confiance est l'ensemble suivant :

$$\mathcal{I}(n_c) = \left\{ N_c : \mathrm{P}(X \le n_c \mid N_c) \ge \alpha \right\} .$$

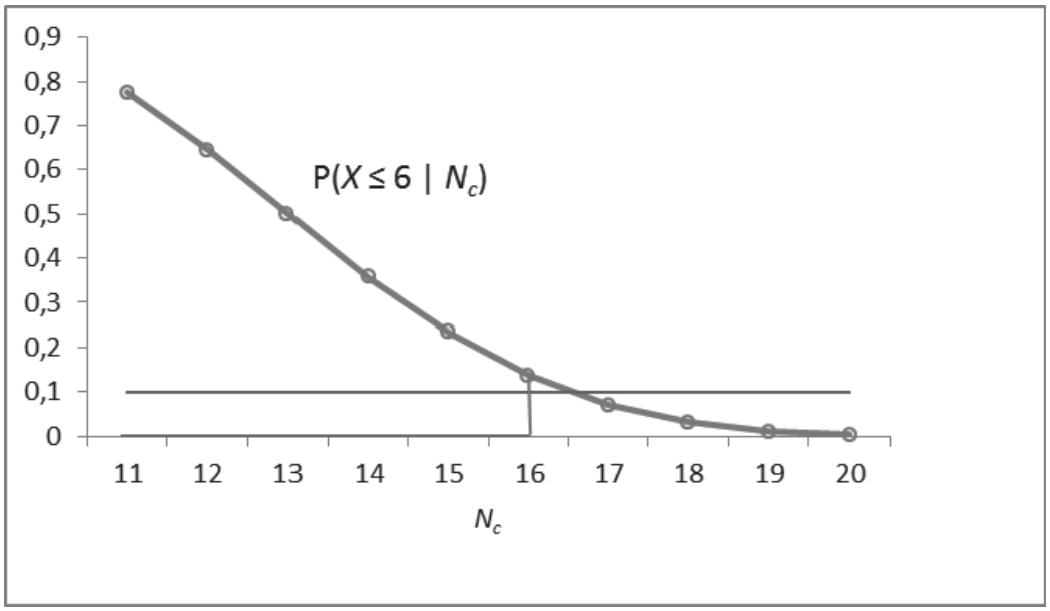
Exemple 4.7 *Intervalle de confiance pour N_c*

Dans un échantillon de taille $n = 15$ d'une population de taille $N = 30$, on observe $n_c = 6$. On calcule alors $\mathrm{P}(X \le 6 \mid N_c)$ pour plusieurs valeurs de N_c afin d'identifier celles pour lesquelles la probabilité $\mathrm{P}(X \le 6 \mid N_c)$ est supérieure ou égale à α. Le tableau suivant calcule ces probabilités. Si on fixe α = 0,1, l'intervalle de confiance est $\{N_c \le 16\}$.

C'est un intervalle de confiance de niveau 1 - α dans le sens que la probabilité qu'il contienne N_c est au moins égale à 1 - α.

N_c	$\mathrm{P}(X \le 6 \mid N_c)$	N_c	$\mathrm{P}(X \le 6 \mid N_c)$	N_c	$\mathrm{P}(X \le 6 \mid N_c)$
6	1,0000	11	0,7751	16	0,1362
7	0,9968	12	0,6448	17	0,0697
8	0,9824	13	0,5000	18	0,0302
9	0,9457	14	0,3576	19	0,0105
10	0,8775	15	0,2330	20	0,0026

Voici une représentation graphique de la procédure :



Les valeurs de N_c pour lesquelles la courbe est au-dessus de la ligne 0,10 sont toutes les valeurs inférieures ou égales à 16.

Pour déterminer un intervalle de confiance borné à gauche, donc de la forme

$$N_c \ge a \quad \text{ou } p \ge a/N,$$

on procède de façon analogue : On calcule $\mathrm{P}(X \ge 6 \mid N_c)$ pour plusieurs valeurs de N_c afin d'identifier celles pour lesquelles la probabilité $\mathrm{P}(X \ge 6 \mid N_c)$ est supérieure ou égale à α.

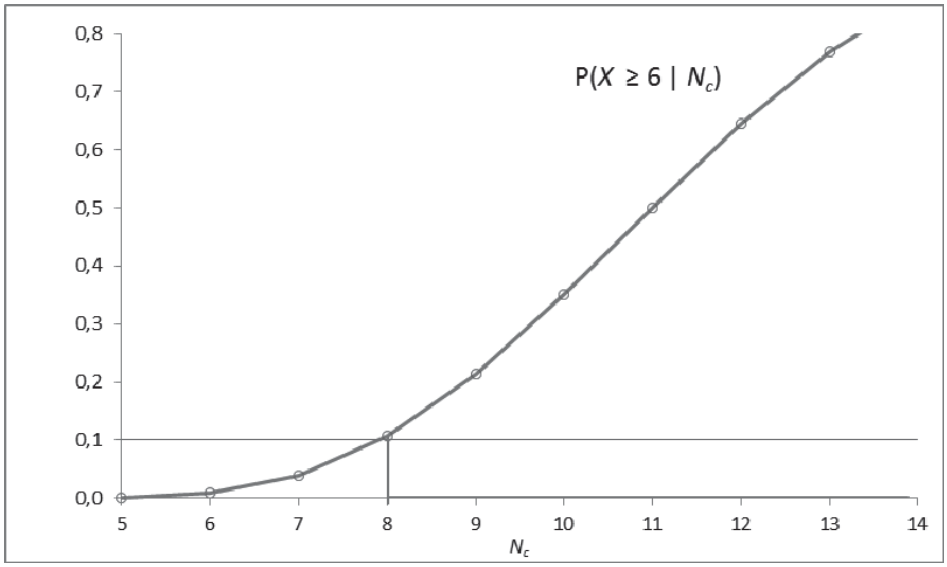
Voici les calculs :

N_c	$P(X \geq 6 \mid N_c)$	N_c	$P(X \geq 6 \mid N_c)$
5	0	10	0,3499
6	0,0084	11	0,5000
7	0,0400	12	0,6448
8	0,1074	13	0,7689
9	0,2135	14	0,8638

Si on fixe $\alpha = 0,1$, l'ensemble de confiance est $\{N_c \geq 8\}$ (le terme « intervalle » n'étant pas tout à fait approprié lorsqu'on se réfère à une liste d'entiers, on parlera plutôt d'« ensemble » de confiance).

On peut combiner les deux résultats et affirmer que $8 \leq N_c \leq 16$.

Mais alors il s'agit là d'un intervalle de confiance de niveau $2\alpha = 20$ %.



Procédure:

Un intervalle de confiance bilatéral à $100(1-\alpha)$ % est l'intersection d'un intervalle à gauche de niveau $100(1-\alpha_1)$ % et un intervalle à droite de niveau $100(1-\alpha_2)$ %, où $\alpha = \alpha_1 + \alpha_2$.

Dans le dernier exemple, nous avons pris $\alpha_1 = \alpha_2 = 0,1$.

Remarque

Le choix $\alpha_1 = \alpha_2 = \alpha/2$ est simple, mais il n'est pas optimal : il ne donne généralement pas l'intervalle le plus court. L'effort d'optimiser l'intervalle serait trop fastidieux, cependant, et le résultat n'en vaut probablement pas la peine.

4.5 POPULATION INFINIE

Quand la population échantillonnée est infinie, les calculs par la loi hypergéométrique peuvent être trop onéreux. Dans ce cas, cependant, la population peut, à toute fin pratique, être considérée infinie et alors la loi hypergéométrique peut être approchée par la loi binomiale.

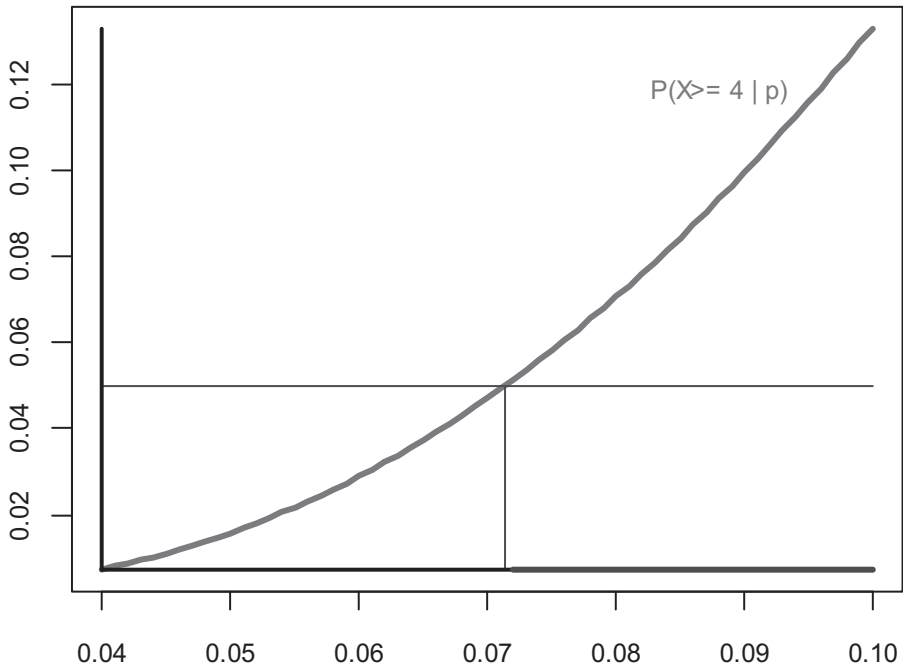
On observe donc une variable aléatoire X de loi $\mathcal{B}(n\ ;\ p)$ et l'objectif est de déterminer un intervalle de confiance de niveau 1 - α. Le principe est le même : si n_c est la valeur observée de X, nous devons déterminer les valeurs LI et LS telles que

(4.6) $$P(X \geq n_c \mid p = LI) = \alpha/2 \text{ et } P(X \leq n_c \mid p = LS) = \alpha/2.$$

Ce qui distingue cette situation de la précédente, c'est le fait que, contrairement à N_c, qui prend les valeurs discrètes 0, 1, 2, ..., N, p prend des valeurs continues. Donc on ne peut pas résoudre les inégalités (4.6) en calculant les probabilités (4.6) pour un nombre fini de valeurs de p. Les solutions de ces égalités peuvent être déterminées à l'aide de procédures numériques. Ces procédures sont longues et complexes — ce ne sont pas des formules, mais une série d'opérations qui permettent d'approcher la solution par étapes. Il se trouve, heureusement, que plusieurs logiciels, dont Excel, peuvent effectuer ces opérations. Nous montrons plus bas comment le faire avec Excel. Ici nous illustrons le principe.

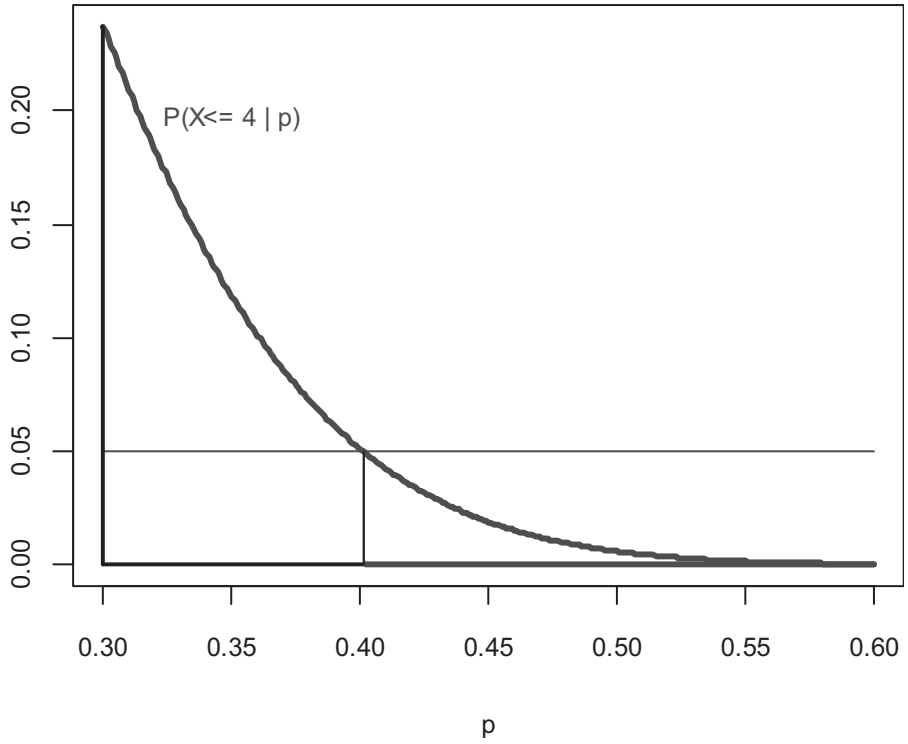
Exemple 4.8 *Intervalle de confiance exact pour p*

Dans un échantillon de taille n = 20, déterminons un intervalle de confiance à 90 % pour p ayant observé n_c = 4. La limite inférieure est la valeur LI de p pour laquelle $P(X \geq 4 \mid p = LI) = 0{,}05$. Le graphique suivant présente la probabilité $P(X \geq 4 \mid p)$ en fonction de p :



Nous excluons de l'intervalle de confiance toute valeur $p < 0{,}071$. La limite inférieure est donc LI = 0,071.

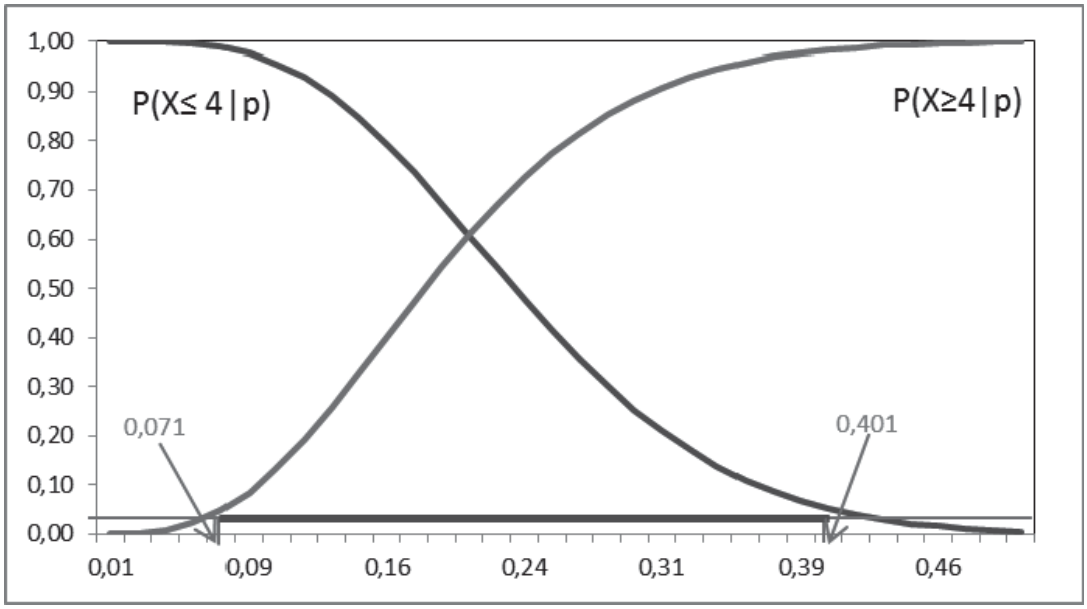
La limite supérieur est la valeur LS de p pour laquelle $P(X \leq 4 \mid p = LS) = 0{,}05$. Le graphique suivant présente la probabilité $P(X \leq 4 \mid p)$ en fonction de p :



Nous excluons de l'intervalle de confiance toute valeur $p > 0{,}401$. La limite supérieure est donc LS = 0,401.

Finalement, l'intervalle de confiance est $0{,}071 \leq p \leq 0{,}401$.

Le graphique suivant combine les deux précédents et montre plus clairement l'ensemble des valeurs incluses dans l'intervalle de confiance.



L'intervalle de confiance approximatif basé sur la loi normale est $0{,}049 \le p \le 0{,}351$.

Remarque *Quand est-ce que la méthode exacte est nécessaire?*

L'approximation normale est assez bonne lorsque la valeur de p n'est pas trop éloignée de 1/2. Mais la méthode exacte est généralement indispensable dans les cas plus extrêmes. Le lecteur peut vérifier que si on observe $n_c = 2$ dans un échantillon de taille 18, la méthode approximative donne l'intervalle (à 90 %) $-0{,}014 \le p \le 0{,}236$. L'intervalle calculé par la méthode exacte est plutôt $0{,}02 \le p \le 0{,}24$.

Dans un grand échantillon, par contre, la méthode approximative est acceptable, même lorsque p est relativement petit. Voici quelques intervalles de confiance (à 90 %) comparatifs avec $n = 18$ et $n = 100$, et quelques valeurs de $n_c = 11$:

	Méthode exacte	Méthode approximative
$n = 18,\ n_c = 2$	$0{,}020 \le p \le 0{,}310$	$0 \le p \le 0{,}236$
$n = 18,\ n_c = 7$	$0{,}199 \le p \le 0{,}608$	$0{,}194 \le p \le 0{,}583$
$n = 100,\ n_c = 11$	$0{,}063 \le p \le 0{,}176$	$0{,}058 \le p \le 0{,}162$
$n = 100,\ n_c = 2$	$0{,}00357 \le p \le 0{,}0616$	$0 \le p \le 0{,}0431$
$n = 100,\ n_c = 1$	$0{,}00051 \le p \le 0{,}0466$	$0 \le p \le 0{,}0264$
$n = 100,\ n_c = 0$	$0 \le p \le 0{,}0295$	Impossible

Cette procédure peut être effectuée à l'aide de l'application « Solveur » du complément « Utilitaire d'analyse » d'Excel.

CALCULS À L'AIDE D'EXCEL

Comment déterminer un intervalle de confiance à 90 % pour dans un échantillon de taille $n = 15$ lorsqu'on observe $n_c = 3$.

1. Il est utile de nommer les valeurs de n, n_c, et α, ici placées dans les cases B1, B2 et D1, respectivement.

2. On nomme également les limites inférieure (LimInf) et supérieure (LimSup) de l'intervalle de confiance. On leur donne une valeur initiale dans les cases D3 et D4, respectivement. La valeur provisoire 0,2 est inscrite ici dans les deux cases.

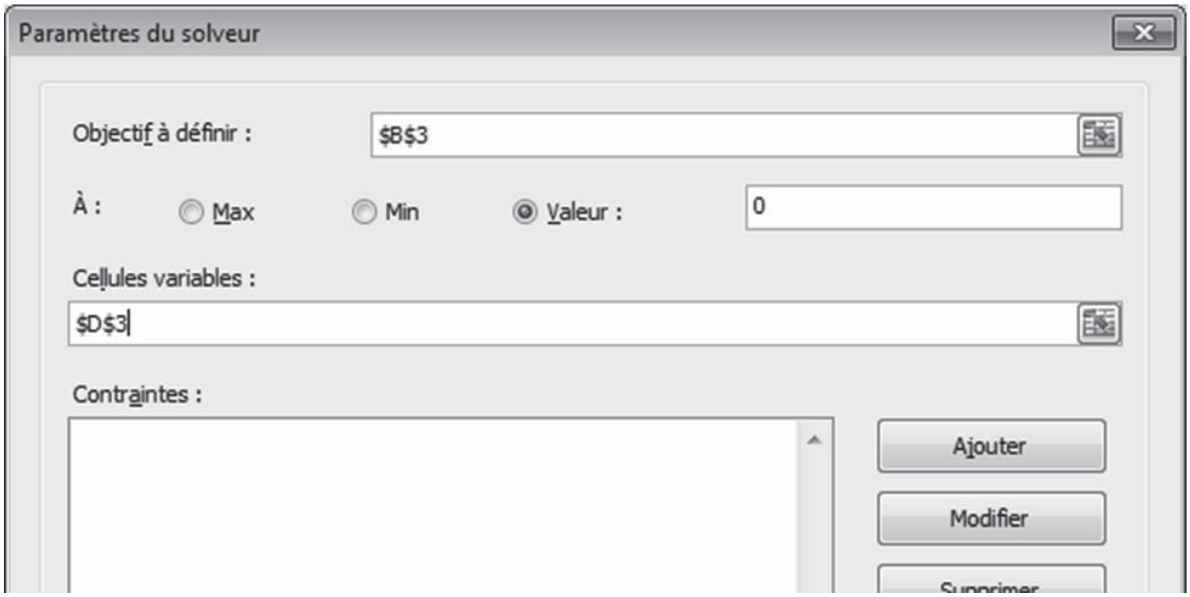
	A	B	C	D
1	n	15	alpha	0,1
2	x	3		
3	P(X>x)	=1-LOI.BINOMIALE.N(x-1;n;LimInf;1)-alpha/2	LimInf	0,2
4	P(X<=x)	=LOI.BINOMIALE.N(x;n;LimSup;1)-alpha/2	LimSup	0,2

Intervalle de confiance

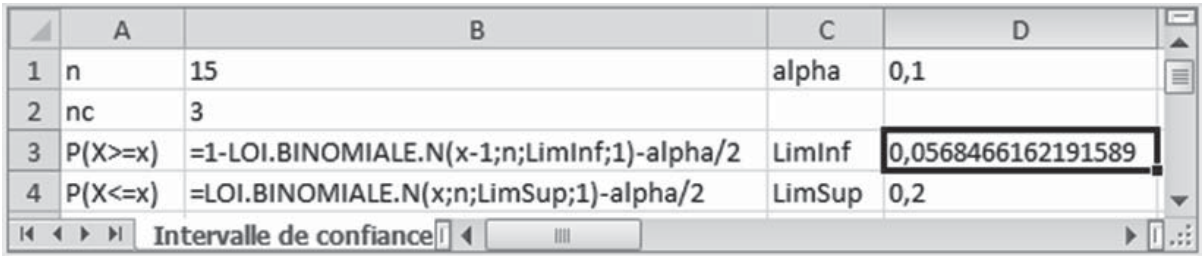
3. Dans B3, on inscrit la commande pour calculer $P(X \geq 3) - \alpha/2$ lorsque X est de loi $\mathcal{B}(n\,;\,p)$.

4. L'objectif est de déterminer la valeur LimInf de p pour laquelle la probabilité $P(X \geq 3) - \alpha/2 = 0$. Dans l'onglet **Données > Solveur**, on inscrit \$B\$3 dans « **Objectif à définir** », on clique sur « **Valeur** » et on inscrit dans la case la valeur « 0 » qui doit être atteinte dans B3 (il n'est pas nécessaire d'inscrire \$B\$3 manuellement : il suffit de placer le curseur dans la case et clique sur la case B3).

5. On inscrit dans « **Cellules variables** » la valeur LimInf à déterminer.



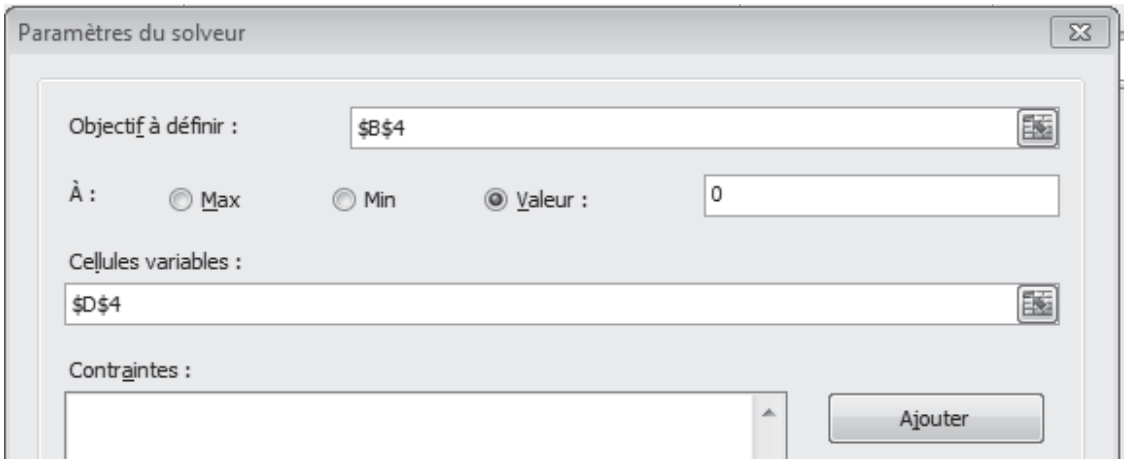
On obtient la limite LimInf, la limite inférieure de l'intervalle de confiance :



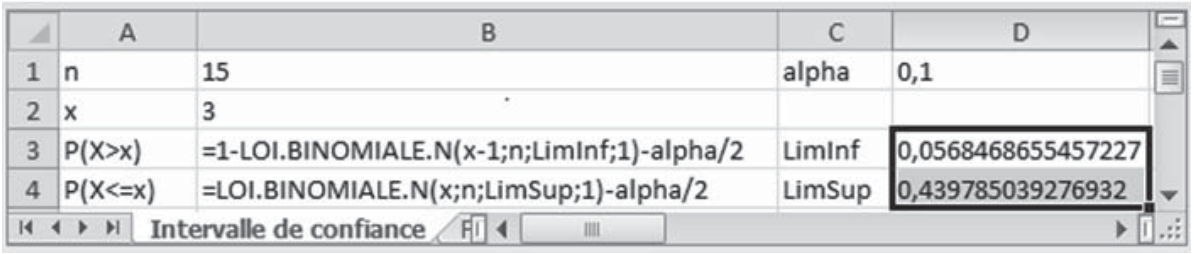
	A	B	C	D
1	n	15	alpha	0,1
2	nc	3		
3	P(X>=x)	=1-LOI.BINOMIALE.N(x-1;n;LimInf;1)-alpha/2	LimInf	0,0568466162191589
4	P(X<=x)	=LOI.BINOMIALE.N(x;n;LimSup;1)-alpha/2	LimSup	0,2

Intervalle de confiance

6. De la même façon, on détermine la limite supérieure LimSup.



7. On obtient finalement ceci :

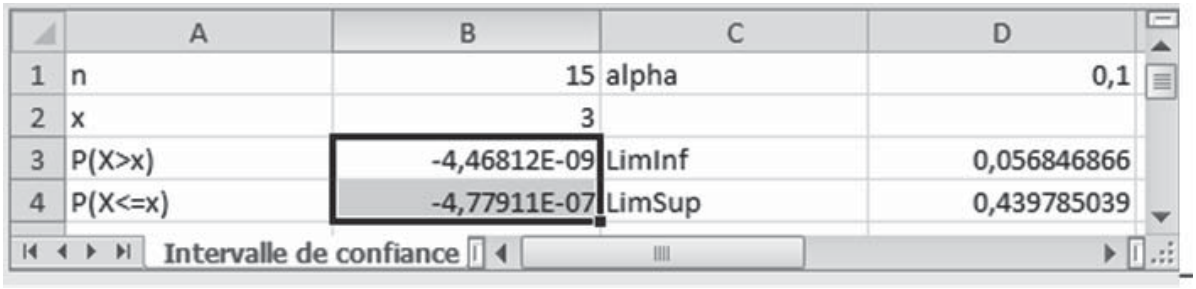


	A	B	C	D
1	n	15	alpha	0,1
2	x	3		
3	P(X>x)	=1-LOI.BINOMIALE.N(x-1;n;LimInf;1)-alpha/2	LimInf	0,0568468655457227
4	P(X<=x)	=LOI.BINOMIALE.N(x;n;LimSup;1)-alpha/2	LimSup	0,439785039276932

Intervalle de confiance

8. L'intervalle de confiance est donc $0{,}0568 \leq p \leq 0{,}440$. L'approche approximative par la loi normale aurait donné $0{,}024 \leq p \leq 0{,}376$.

9. Il est bon de s'assurer que les probabilités dans B3 et 4 sont proches de 0, comme il se doit :



	A	B	C	D
1	n	15	alpha	0,1
2	x	3		
3	P(X>x)	-4,46812E-09	LimInf	0,056846866
4	P(X<=x)	-4,77911E-07	LimSup	0,439785039

Intervalle de confiance

4.6 INTERVALLE DE CONFIANCE POUR UNE MÉDIANE

Nous avons vu qu'un intervalle de confiance pour une moyenne μ est problématique quand la population est fortement asymétrique. Nous avons examiné une solution qui consiste à soustraire les valeurs les plus extrêmes de la population, afin de leur accorder un traitement particulier (les recenser entièrement, par exemple). Les techniques des trois prochains chapitres elles aussi peuvent y remédier jusqu'à un certain point.

Mais la présence de quelques données extrêmes ne crée pas seulement des problèmes d'estimation. Dans une population comptant des données extrêmes, c'est la moyenne arithmétique elle-même, comme mesure de tendance centrale, qui est problématique — dans le sens qu'elle peut difficilement être considérée comme donnée « représentative » ou « centrale ». Par exemple, dans l'ensemble des municipalités québécoises (tableau A05), le nombre d'habitants est une variable de moyenne μ = 6 733, une valeur supposément « centrale ». Or une très forte proportion (88 %) des données lui sont inférieures. Dans des populations comme celles-là, c'est l'usage de la moyenne elle-même qui doit être mis en cause. D'autres mesures de tendance centrale doivent être considérées. La *médiane* en est une, probablement la plus appropriée. (C'est pour cette raison que, dans les documents de statistiques officiels, la moyenne des revenus des ménages, une variable notoirement asymétrique, est généralement accompagnée de la médiane.)

Nous allons donc montrer comment déterminer un intervalle de confiance pour une médiane. Nous commençons par décrire la procédure, d'abord dans de grands échantillons puis dans de petits échantillons. Ensuite, nous la justifierons. Nous nous en tiendrons aux cas où la population est assez grande pour qu'on puisse poser $N = \infty$.

La médiane est le nombre m qui sépare la population en deux parties égales: celle où $Y \leq m$ et celle où $Y \geq m$. Elle n'est pas nécessairement unique. Par exemple, la médiane des valeurs 2, 3, 4, 7, 8, 9 est n'importe quelle nombre dans l'intervalle]4 ; 7[. Il n'est pas nécessaire de s'attarder sur les diverses conventions qui existent pour éliminer cette ambiguïté. Lorsque la population est assez grande cette difficulté est sans importance.

PROCÉDURE – GRANDS ÉCHANTILLONS

Soit n la taille de l'échantillon, et $y_{(1)} \le y_{(2)} \le \ldots \le y_{(n)}$ les valeurs de y ordonnées de la plus petite à la plus grande. Soit X une variable de loi binomiale de paramètres n et $p = 1/2$. Voici les étapes à suivre pour déterminer un intervalle de confiance à 100(1-α) % pour la médiane :

On détermine x_1, le plus petit entier supérieur ou égal à $\frac{n - z_{\alpha/2}\sqrt{n}}{2}$ et $x_2 = n - x_1$.

L'intervalle de confiance pour la médiane m est donné par

$$y_{(x_1)} \le m \le y_{(x_2)}$$

où $y_{(i)}$ désigne la i^e plus petite donnée.

Exemple 4.9 *Intervalle de confiance pour une médiane*

Les données suivantes représentent les prix d'un échantillon de 40 maisons vendues dans la région de Québec. On les présente en ordre croissant.

163	191	202	206	222	224	301	320	329	335
349	372	376	**385**	396	396	399	403	419	442
445	464	499	523	601	**778**	871	875	932	938
1037	1052	1101	1139	1194	1228	1257	1324	1544	1628

Pour un intervalle de confiance à 95 %, $z_{\alpha/2} = z_{0,025} = 1{,}96$. On a $\frac{n - z_{\alpha/2}\sqrt{n}}{2} = 13{,}8$ et $\frac{n + z_{\alpha/2}\sqrt{n}}{2} = 26{,}2$. On arrondit le premier vers le haut et le deuxième vers le bas, ce qui donne $x_1 = 14$ et $x_2 = 26$. Les limites de l'intervalle de confiance pour la médiane sont alors $y_{(14)} = 385$ et $y_{(26)} = 778$: On affirme que la médiane des prix des maisons de la population se situe entre 385 000 $ et 778 000 $.

PROCÉDURE – PETITS ÉCHANTILLONS

Le nombre x_1 défini ci-dessus est une approximation. La définition exacte est celle-ci : x_1 est le plus petit des entiers x tels que $P(X \le x) \ge \alpha/2$, où X est une variable de loi $\mathcal{B}(n\,;\,½)$.

La détermination de x_1 est donc plutôt laborieuse, mais elle se fait bien à l'aide d'un logiciel comme Excel.

CALCULS D'UN INTERVALLE POUR LA MÉDIANE À L'AIDE D'EXCEL:

La fonction LOI.BINOMIALE.INVERSE() définit précisément le nombre x_1. Lorsque $n = 40$ et $\alpha/2 = 0{,}025$, la commande exacte est LOI.BINOMIALE.INVERSE(40;0,5;0,025). Voici l'image de la commande Excel:

	A	B	C
1	n	x1	x2
2	40	=LOI.BINOMIALE.INVERSE(A2;0,5;0,025)	=A2-B2

Voici les résultats:

	A	B	C
1	n	x1	x2
2	40	14	26

Le résultat obtenu est le même qu'avec l'approche approximative.

Remarque

La loi binomiale avec $p = ½$ étant très bien approchée par une loi normale, il sera rarement nécessaire de recourir à l'approche exacte.

Justification

Soit m la médiane d'une population et X le nombre d'observations inférieures ou égales à m. Alors X est de loi binomiale de paramètres n et $p = 1/2$. L'intervalle $\left[y_{(x_1)} ; y_{(x_2)} \right]$ recouvre m si et seulement si $x_1 \le X \le x_2$. Donc le niveau de l'intervalle est $P(x_1 \le X \le x_2)$. Si x_1 est le plus petit entier tel que $P(X \le x_1) \ge \alpha/2$ et $x_2 = n - x_1$, alors $P(x_1 \le X \le x_2) \ge 1 - \alpha$, ce qui montre que l'intervalle est de niveau $1 - \alpha$.

Exemple 4.10 *Une justification de la procédure*

Voici une autre façon de justifier la procédure. On exclut de l'intervalle toute valeur de m qui est peu vraisemblable au vu de l'échantillon. Cela inclut les valeurs trop grandes ainsi que les valeurs trop petites. Reprenons les données de l'exemple et considérons, par exemple, la valeur $m = 349$. Si c'est bien la médiane, il devrait y avoir en moyenne 20 observations inférieures ou égales à 349. Or il n'y en a eu que 11, ce qui semble un peu trop petit pour être attribué au hasard: sous l'hypothèse que $m = 349$, le nombre X de valeurs inférieures ou égales à la médiane est une variable aléatoire de loi binomiale de paramètres $n = 40$ et $p = ½$ et alors $P(X \le 11 \mid m = 349) = 0{,}0032$: il est peu probable que le nombre d'observations inférieures ou égales à la médiane soit inférieure ou égale à 11. À l'autre extrémité, une grande valeur m sera rejetée si Y, le nombre d'observations *supérieures* ou égales à m est trop petit. Par exemple, la valeur $m = 932$ est rejetée car il n'y a que 12 données qui lui sont supérieures ou égales, et $P(Y \ge 12 \mid m = 932) = 0{,}0032$. Quelques-uns de ces calculs sont illustrés au Tableau 4.1.

Tableau 4.1
Intervalle de confiance pour une médiane

m	x	$P(X \le x \mid m)$	m	y	$P(Y \ge y \mid m)$
380	13	0,01924	454	21	0,43731
390	14	0,04035	482	22	0,31791
396	16	0,13409	511	23	0,21480
398	16	0,13409	562	24	0,13409
401	17	0,21480	690	25	0,07693
411	18	0,31791	824	26	0,04035
430	19	0,43731	873	27	0,01924

Remarque

Plusieurs raffinements sont possibles, mais rarement nécessaires. Par exemple, l'approximation normale devrait tenir compte du facteur de correction, ce que nous n'avons pas fait ici. Et, strictement parlant, c'est la loi hypergéométrique que nous aurions dû appliquer et non la loi binomiale. À moins que N soit vraiment petit, ces ajustements laborieux n'en valent pas vraiment la peine.

4.7 DÉMONSTRATIONS

Intervalle de confiance pour N_c avec petit échantillon

L'observation n_c est la valeur observée d'une variables X une variable de loi hypergéométrique de paramètres n, N_c, et $N - N_c$. Pour une valeur $X = n_c$ donnée, l'intervalle de confiance est défini comme $\{N_c : P(X \le n_c \mid N_c) \ge \alpha/2$ et $P\{N_c : P(X \le n_c \mid N_c) \ge \alpha/2\}$. La probabilité que l'intervalle contienne N_c est donc $P[\{x : P(X \le x \mid N_c) \ge \alpha/2\} \cap \{x : P(X \ge x \mid N_c) \ge \alpha/2\}]$. Soit x_1 la plus petite valeur x pour laquelle $P(X \le x) \ge \alpha/2$ et x_2 la plus grande valeur x pour laquelle $P(X \ge x) \ge \alpha/2$. On a donc $P(X \le x_1) \ge \alpha/2$ et $P(X < x_1) < \alpha/2$; de même, $P(X \ge x_2) \ge \alpha/2$ et $P(X > x_2) < \alpha/2$. Donc la valeur du paramètre se trouvera dans l'intervalle de confiance si et seulement $x_1 \le X \le x_2$. La probabilité de cet événement est $P(x_1 \le X \le x_2) = 1 - P(X < x_1) - P(X > x_2) \ge 1 - \alpha$.

RÉSUMÉ

- Lorsqu'une variable Y est dichotomique, sa variance est $S^2 = \frac{Np(1-p)}{N-1}$, où p est la proportion des valeurs de Y qui égalent 1.
- La variance d'une proportion échantillonnale $\hat{p} = \frac{n_c}{n}$ est $V(\hat{p}) = \frac{N-n}{N-1}\frac{p(1-p)}{n}$; elle est estimée sans biais par $\hat{V}(\hat{p}) = (1-f)\frac{\hat{p}(1-\hat{p})}{n-1}$.
- La taille de l'échantillon nécessaire pour que la marge d'erreur de l'estimateur de p soit égale à E est $n = \frac{n_o}{1+\frac{n_o}{N}}$, où $n_o = \frac{z^2_{\alpha/2}p(1-p)}{E^2}$. Le produit $p(1\text{-}p)$ est borné supérieurement par 1/4.
- La taille de l'échantillon nécessaire pour que la marge d'erreur relative soit égale à R est $n = \frac{n_o}{1+\frac{n_o}{N}}$, où $n_o \approx \left(\frac{z_{\alpha/2}}{R}\right)^2 \frac{(1-p)}{p}$.

Intervalles de confiance exacts

- Soit $\hat{p} = \frac{n_c}{n}$. Un intervalle de confiance de niveau 1- α est donné par $\{N_c \mid \mathrm{P}(X \geq n_c \mid N_c) \geq \alpha/2$ et $\mathrm{P}(X \leq n_c \mid N_c) \geq \alpha/2\}$ où X est de loi hypergéométrique de paramètres n, N_c et N - N_c.
- Si N est grand, l'intervalle est approché par $\{p \mid \mathrm{P}(X \geq n_c \mid p) \geq \alpha/2$ et $\mathrm{P}(X \leq n_c \mid p) \geq \alpha/2\}$ où X est de loi binomiale de n et p.

Intervalle de confiance pour une médiane

- Un intervalle de confiance de niveau 1 - α pour la médiane de la population est donné par $\left[y_{(x_1)} ; y_{(x_2)}\right]$, où $y_{(i)}$ est la i^e plus petite donnée de l'échantillon, $x_2 = n - x_1$, et x_1 est le plus petit des entiers x pour lesquels que $\mathrm{P}(X \leq x_1) \geq \alpha/2$ lorsque X est une variable de loi $\mathcal{B}(n ; 1/2)$.

EXERCICES

Note: Dans les exercices qui suivent, à moins d'indication contraire, on supposera un niveau de confiance de 95 % dans le calcul d'une marge d'erreur ou d'un intervalle de confiance.

4.1 On tire un *éas* de 300 étudiants dans une population de 6 455 étudiants afin d'estimer la proportion p et le nombre N_c d'étudiants qui ont accès à l'Internet chez eux. On trouve que 250 étudiants dans l'échantillon ont accès à l'Internet.

a) Estimer p et N_c et estimer l'écart-type de vos estimateurs.

b) Déterminer un intervalle de confiance pour p et pour N_c.

4.2 Dans une université, on se propose d'estimer le *nombre* d'étudiants qui possèdent un ordinateur. Le nombre d'étudiants dans l'université est de 45 635. Sachant par expérience que la proportion d'étudiants qui possèdent un ordinateur se situe quelque part entre 40 % et 65 %, déterminer la taille de l'échantillon qu'on doit prélever dans chacun des cas suivants :

a) On tient à ce que la marge d'erreur soit d'environ 1 000 étudiants.

b) On tient à ce que la marge d'erreur soit de 15 %.

4.3 On s'apprête à tirer un échantillon dans une population de 28 635 étudiants afin d'estimer la proportion p d'étudiants (ou N_c, leur nombre) qui ont accès à l'Internet. Sachant par expérience que p se situe quelque part entre 15 % et 25 %, déterminer la taille de l'échantillon qu'il faut prélever pour atteindre les objectifs suivants :

a) Estimer N_c avec une marge d'erreur de 500 étudiants.

b) Estimer N_c avec une marge d'erreur relative de 10 %.

c) Estimer p avec une marge d'erreur d'environ 2 % (c'est-à-dire, 2 points de pourcentage).

d) Estimer p avec une marge d'erreur relative de 10 % (par rapport à la vraie proportion p).

4.4 Un *éas* de taille 835 est tiré d'une population afin d'estimer une proportion p. Un intervalle de confiance de niveau 95 % pour p donne l'intervalle suivant : [0,57 ; 0,64]. Pour chacune des propositions suivantes autres que d) et g), indiquer si elle est vraie, fausse, ou si on ne peut pas se prononcer.

a) La proportion p est dans l'intervalle [0,57 ; 0,64].

b) L'estimation ponctuelle de p est dans l'intervalle [0,57 ; 0,64].

c) La probabilité que p soit dans l'intervalle [0,57 ; 0,64] est de 0,85.

d) À partir de cet intervalle, déterminer un intervalle de confiance à 99 %.

e) La marge d'erreur de l'intervalle [0,57 ; 0,64] est de 7 (points de pourcentage).

f) Si on avait pris un échantillon de taille n = 1 225, on aurait sans doute obtenu un intervalle plus grand (plus large).

g) Dans le calcul de cet intervalle, on n'a pas tenu compte de la taille de la population. Corriger l'intervalle de confiance sachant qu'en fait N = 1100?

4.5 Une analyste doit vérifier une population des 1 450 comptes d'investissement d'une succursale bancaire. Elle doit estimer deux choses : la valeur totale des comptes ainsi que le nombre de comptes de 50 000 $ ou plus. Elle souhaite que sa marge d'erreur relative soit d'au plus 12 % pour les deux estimations. On dispose des données suivantes. Dans les années passées, le coefficient de variation des montants des

comptes se situait, bon an, mal an, entre 60 % et 80 %. Aussi, la proportion des comptes de 50 000 $ ou plus s'est toujours située entre 19 % et 23 %. Ce pourcentage, cependant, croît avec le temps et pourrait aller jusqu'à 25 % cette année. Quelle est la taille de l'échantillon qu'elle devrait prélever ?

4.6 Après avoir annoncé son produit lors d'une émission de télévision locale, le propriétaire d'une chaîne de magasins fait faire une petite enquête auprès d'un *éas* de 500 habitants de la région couverte par la station de télévision (17 324 habitants). Voici un résumé des résultats :

		Ont vu l'annonce		
		Oui	Non	Total
Ont vu l'émission	Oui	250	100	350
	Non	0	150	150
	Total	250	250	500

(Ignorez le facteur de correction dans cette question).

a) Déterminez un intervalle de confiance pour la proportion de la population qui ont vu l'annonce.

b) Déterminez un intervalle de confiance pour la proportion de la population qui ont vu l'émission.

c) Déterminez un intervalle de confiance pour la proportion de ceux qui ont vu l'annonce parmi ceux qui ont vu l'émission. Vous supposerez que le nombre (350) de personnes qui ont vu l'émission est un nombre fixe et vous estimerez la taille de la population (ceux qui ont vu l'émission) à partir de l'échantillon (voir l'exercice 4.14 pour une justification de cette pratique).

d) Déterminez un intervalle de confiance pour le nombre de personnes qui ont vu l'annonce.

e) Déterminez un intervalle de confiance pour le nombre de personnes qui ont vu l'annonce, en utilisant plutôt le fait, connu indépendamment, que 13 000 personnes dans la population ont vu l'émission l'échantillon (voir l'exercice 4.14).

4.7 On sait par expérience que le poids de certains sacs de pommes de terre est une variable de loi à peu près normale d'écart-type 0,45 kg. Un lot de 4 000 sacs passe par un processus de sélection automatique qui permet, sans les peser individuellement, d'exclure tous les sacs de moins de 9 kg.

a) On a trouvé que 182 des 4 000 sacs pesaient moins de 9 kg. Utiliser cette information pour obtenir une approximation du poids moyen des 4 000 sacs.

b) Supposons que les 4 000 sacs observés ne sont, en réalité, qu'un échantillon d'une population infinie et que la proportion observée $\hat{p}$ = 0,0455 est une estimation d'une proportion p inconnue. Déterminer un intervalle de confiance pour p et, par suite, un intervalle de confiance pour μ (μ étant cette fois-ci la moyenne non pas des 4 000 sacs mais de la population dont ceux-là font partie).

4.8 On tire un échantillon de 5 appartements dans un immeuble de 20 appartements afin d'estimer le nombre N_c de locataires qui n'ont pas toujours payé leur loyer à temps. On en trouve 2 parmi ceux des 5 appartements sélectionnés.

a) Déterminer un ensemble de confiance à 80 % pour N_c

i) par la méthode exacte; et

ii) par l'approximation normale.

b) Supposons que l'échantillon a été tiré par un acheteur potentiel qui voudrait s'assurer (avec 80 % de confiance) qu'il n'y a pas plus de 10 mauvais payeurs parmi les locataires. Déterminer un intervalle de confiance unilatéral qui donnerait (ou pas) cette assurance.

4.9 D'une très grande population de comptes, on tire un échantillon de 15 comptes. On y trouve 2 qui sont erronés. Déterminer par la méthode exacte un intervalle de confiance à 90 % pour la proportion p de comptes erronés dans la population. Comparez votre intervalle à celui que vous auriez obtenu par l'approximation normale.

4.10 Au numéro 4.9, supposons que dans un échantillon de taille 15 on n'ait observé aucun compte erroné. Déterminer un intervalle de confiance à 90 % pour p. Noter que l'intervalle est forcément unilatéral, c'est-à-dire, de la forme $p \leq \text{LS}$.

4.11 Au numéro 4.9, supposons que dans un échantillon de taille 15 on ait observé 15 comptes erronés. Déterminer un intervalle de confiance à 90 % pour p. Notez que l'intervalle est forcément de la forme $p \geq \text{LI}$.

4.12 Voici les montants d'un échantillon de 45 comptes de dépenses tirés d'une grande population :

7,12	8,16	10,53	14,80	23,06
23,78	24,88	32,35	32,48	37,28
38,35	45,52	46,79	46,88	48,38
60,45	62,12	71,47	72,14	77,22
89,13	94,58	99,34	120,46	121,58
124,36	127,81	132,73	133,93	152,92
166,20	170,20	237,31	263,20	295,79
299,11	373,86	417,37	451,66	595,21
640,22	686,14	699,23	862,58	9910,48

Déterminer un intervalle de confiance à 90 % pour la médiane de la population. Comparer cet intervalle à un intervalle de confiance pour la moyenne. [Une simulation à partir des données de la population a montré que le niveau réel d'un intervalle de confiance pour la médiane est très proche du niveau théorique de 90 %, alors que le niveau réel d'un intervalle de confiance pour la moyenne est de l'ordre de 38 %].

4.13 Démontrer que si Y est une variable dichotomique, alors $S^2 = \dfrac{Np(1-p)}{N-1}$.

4.14 Au numéro 4.6, montrer que les solutions proposées en c) et en d) sont légitimes dans le sens que les estimateurs de $\hat{V}(\hat{p})$ sont sans biais pour la variance *conditionnelle* $V(\hat{p} \mid \mathrm{A})$, où A est l'événement « le nombre de personnes ayant vu l'émission est 350 ». [Les 3 effectifs dont les valeurs observées sont 250, 100, et 250 est un vecteur de loi multinomiale. Le vecteur $\boldsymbol{X} = [X_1 ; X_2 ; X_3]$ est de loi multinomiale de paramètres 500 et p_1, p_2, et p_3. Alors (voir le Théorème 1.16), $P(X_1 = x_1 \mid X_1 + X_2 = m) = \binom{m}{x_1}\left(\frac{p_1}{p_1+p_2}\right)^{x_1}\left(\frac{p_2}{p_1+p_2}\right)^{m-x_1}$. Autrement dit, étant donné que $X_1 + X_2 = 250$, X_1 est conditionnellement de loi binomiale.]

4.15 On suppose que parmi les nouveaux arrivés dans le pays, une certaine proportion p a été admise avec de faux papiers. Afin de se faire une idée de la valeur de p, on tire des individus au hasard jusqu'à ce qu'on tombe sur une personne avec de faux papiers. On compte alors le nombre X d'unités qu'il a fallu tirer. On ne peut pas estimer p sans biais, mais on peut estimer sa réciproque $\theta = 1/p$, une information qui n'est pas dénuée de sens : θ - 1 est le nombre d'immigrants légitimes pour chaque immigrant illégal. Soit X le nombre d'essais effectués au moment où on tombe sur un premier immigrant illégal. Considérons l'estimateur $\hat{\theta} = X$.

a) Montrer que $\hat{\theta}$ est un estimateur sans biais de θ.

b) Montrer que $V(\hat{\theta}) = (1-p)/p^2$.

c) Montrer que $\hat{V}(\hat{\theta}) = \frac{X(X-1)}{2}$ est un estimateur sans biais de $V(\hat{\theta})$.

[*Suggestion*: Utiliser le fait que $\sum_{x=1}^{\infty} x(1-p)^{x-1} = 1/p^2$ et que $\sum_{x=1}^{\infty} x^2(1-p)^{x-1} = (2-p)/p^3$.]

4.16 L'expérience décrite au numéro 4.15 ne suffit pas à donner un estimateur précis: il est assez probable, même si p est petit, que le premier succès survienne assez vite, laissant croire que p est grand; ou, au contraire, que le premier succès peut tarder à se réaliser, alors que p est grand. Pour réduire ces risques, il faudrait que l'expérience soit répétée plusieurs fois. On décide donc d'avance d'effectuer des tirages jusqu'à l'obtention d'un r^e succès. X est encore le nombre d'essais effectués.

a) Montrer que $\hat{p} = \frac{r-1}{X-1}$ est un estimateur sans biais de p.

b) Montrer que si $r > 2$, $V(\hat{p}) \le \frac{p^2}{r-2}$, et que cette borne peut être estimée sans biais par $\frac{r-1}{(X-1)(X-2)}$. X suit une loi binomiale négative (voir le Théorème 1.14). Les sommes qui définissent l'espérance et la variance peuvent être simplifiées en utilisant le fait que $\sum_{x \ge r} p(x) = 1$ lorsque $p(x)$ est de la forme $p(x) = \binom{x-1}{r-1} p^r (1-p)^{x-r}$.

c) L'inconvénient de cette approche est que le nombre d'essais risque d'être trop élevé. Son espérance est r/p et son écart-type est $\frac{\sqrt{r(1-p)}}{p}$. Justifier par un argument verbal le fait que P($X > x$) = $\sum_{i=0}^{r-1}\binom{x}{i}p^i(1-p)^{x-i}$.

d) Afin d'estimer la proportion de documents erronés dans une très grande population, on prélève et vérifie des documents un à la fois. On s'arrête après avoir découvert 4 documents erronés. Supposons qu'on découvre le quatrième document erroné au 10e essai. Utiliser les techniques exactes pour montrer que l'intervalle [0,15002782 ; 0,65505895] est un intervalle de confiance à 90 % pour p.

4.17 Le lendemain de la diffusion d'une importante émission de télévision, la population est divisée en 3 catégories: N_1 personnes ont vu l'émission et l'ont aimée; N_2 l'ont vue mais ne l'ont pas aimée; et N_3 ne l'ont pas vue. On tire un *éas* de 300 personnes. Soit X_1, X_2 et X_3 le nombre de personnes dans l'échantillon appartenant aux catégories 1, 2 et 3, respectivement. Soit p la proportion de ceux qui ont aimé l'émission (parmi ceux qui l'ont vue). Considérer que la population est très grande.

a) Montrer que la distribution conditionnelle de X_1 étant donné $X_1+X_2 = m$ est une loi hypergéométrique de paramètres m, N_1 et N_2.

b) Soit $\hat{p} = \frac{X_1}{X_1 + X_2}$. Montrer que $E(\hat{p} \mid X_1 + X_2 = m) = p$ et conclure que $\hat{p}$ est un estimateur sans biais de p;

c) Montrer que $V(\hat{p} \mid X_1 + X_2 = m) = \frac{M-m}{M-1}\frac{p(1-p)}{m}$, où $M = N_1+N_2$;

d) Montrer que $\hat{V}(\hat{p} \mid X_1 + X_2 = m) = \left(1 - \frac{m}{M}\right)\frac{\hat{p}(1-\hat{p})}{m-1}$ est un estimateur conditionnellement sans biais de $V(\hat{p} \mid X_1 + X_2 = m)$;

e) Supposons que N = 6 419, X_1 =25, X_2 = 70 et X_3 = 205. Estimer p, ainsi que sa variance conditionnelle sachant que 95 de personnes ont vu l'émission). Pour estimer la fraction d'échantillonnage m/M, remplacer M par une estimation $\hat{M}$.

4.18 La municipalité d'une petite ville comprenant 4 824 logements lance une campagne visant à encourager l'installation de détecteurs de fumée. Afin d'estimer son efficacité, elle tire un échantillon de 400 logements avant le début de la campagne et une deuxième fois, après la fin de la campagne, indépendamment de la première. Voici un résumé des réponses obtenues des occupants:

	Détecteur de fumée?		
	Non	Oui	
Avant	172	228	400
Après	120	280	400

Soit N_1 le nombre de logements ayant un détecteur de fumée avant la campagne et N_2 le nombre de logements ayant un détecteur de fumée après la campagne.

a) Déterminer un intervalle de confiance à 95 % pour N_1 et un intervalle de confiance pour N_2.

b) Déterminer un intervalle de confiance à 95 % pour l'accroissement N_2 - N_1 du nombre de logements avec détecteur de fumée.

4.19 Même contexte qu'au numéro 4.18, avec la différence suivante: plutôt que d'effectuer un deuxième tirage indépendant, on interroge les mêmes 400 occupants sélectionnés lors du premier sondage. On obtient les réponses suivantes:

		Deuxième tirage		
	Détecteur de fumée?	Non	Oui	
Premier tirage	Non	115	57	172
	Oui	5	223	228
		120	280	400

a) Soit p la proportion, parmi les logements sans détecteur lors du premier sondage, qui en avaient un lors du second sondage. Déterminer un intervalle de confiance à 95 % pour p.

b) Déterminer un intervalle de confiance à 95 % pour l'accroissement du nombre de logements avec détecteur de fumée depuis le premier sondage.

c) Pourquoi l'intervalle ici est-il plus étroit qu'au numéro 4.18-b)?

4.20 Dans l'Exemple 4.3, quelle aurait été l'estimation de $V(\hat{p}_2 - \hat{p}_1)$ si on avait supposé que $\hat{p}_1$ et $\hat{p}_2$ étaient indépendants? On remarque que la variance est surestimée ici. Pouvez-vous expliquer pourquoi?

4.21 On compare deux quartiers quant à la proportion des maisons ayant un garage. Voici les données de deux échantillons indépendants tirés dans ces populations:

	Quartier 1	Quartier 2
Nombre de maisons dans la population	1000	500
Nombre de maisons échantillonnées	400	300
Nombre de maisons avec garage (dans l'échantillon)	189	120

L'objectif est de savoir s'il y a vraiment une différence entre les deux proportions. Une procédure informelle est la suivante: on détermine un intervalle de confiance pour chacune des deux proportions; on rejettera l'hypothèse que les deux proportions sont égales si les deux intervalles ne se recoupent pas.

a) Montrer qu'en suivant cette procédure, à un niveau de 95 %, on ne peut pas rejeter l'hypothèse d'égalité des proportions.

b) Montrer en fait que cette procédure est conservatrice: la probabilité de rejeter l'hypothèse est inférieure à celle qui résulterait de la procédure décrite dans la section 4.2.

c) Montrer que l'hypothèse acceptée en a) aurait été rejetée si on avait suivi plutôt la procédure décrite dans la section 4.2.

5 Estimation du quotient

Le quotient de deux moyennes (ou de deux totaux) est un autre paramètre d'intérêt. Quelques exemples: la production moyenne d'orge par hectare dans une population de fermes; la portion relative des frais de nourriture dans les budgets familiaux; la teneur en sucre d'un type d'agrume; la proportion d'aînés dans les familles d'une ville.

Formellement, un quotient est fonction de deux variables, X et Y, définies sur une même population U : à chaque unité k correspondent deux valeurs, x_k et y_k. Le quotient R est défini comme le rapport de la somme (ou de la moyenne) des y_k sur la somme (ou la moyenne) des x_k :

$$R = \frac{\tau_y}{\tau_x} = \frac{\mu_y}{\mu_x}, \text{ où } \tau_x = \sum_{k=1}^{N} x_k \text{ et } \tau_y = \sum_{k=1}^{N} y_k .$$

Nous maintenons la notation du chapitre 3, en y accolant un x ou un y pour identifier la variable dont il est question. Ainsi les moyennes de la population seront désignées par μ_x et μ_y, et leurs variances par S_x^2 et S_y^2. Les moyennes échantillonnales seront désignées par $\bar{x}$ et $\bar{y}$ et les variances par s_x^2 et s_y^2, etc.

Il est nécessaire maintenant de définir un paramètre qui mesure la dépendance entre les variables X et Y, la *covariance*.

DÉFINITION 5.1 COVARIANCE ENTRE DEUX SÉRIES DE DONNÉES

La covariance entre les valeurs de X et les valeurs de Y est définie par

$$S_{xy} = \frac{\sum_{k=1}^{N}(x_k - \mu_x)(y_k - \mu_y)}{N-1} \tag{5.1}$$

pour les données de la population et par

$$s_{xy} = \frac{\sum_{i\in\omega}(x_i - \bar{x})(y_i - \bar{y})}{n-1} \tag{5.2}$$

pour les données de l'échantillon.

THÉORÈME 5.1 **ESTIMATEUR D'UNE COVARIANCE**

La covariance échantillonnale est sans biais pour la covariance de la population : $E(s_{xy}) = S_{xy}$.

La formule de la covariance $C(\bar{x};\bar{y})$ entre deux moyennes échantillonnales est analogue à celle de la variance d'une moyenne.

THÉORÈME 5.2 **COVARIANCE ENTRE DEUX MOYENNES ÉCHANTILLONNALES**

La covariance entre $\bar{x}$ et $\bar{y}$ est donnée par

$$C(\bar{x};\bar{y}) = (1-f)\frac{S_{xy}}{n}, \quad f = \frac{n}{N}.$$

THÉORÈME 5.3 **ESTIMATION DE LA COVARIANCE ENTRE DEUX MOYENNES**

$\hat{C}(\bar{x};\bar{y}) = (1-f)\frac{s_{xy}}{n}$ est un estimateur sans biais de $C(\bar{x};\bar{y})$.

5.1 ESTIMATEUR D'UN QUOTIENT ET SES PROPRIÉTÉS

Supposons qu'on tire un échantillon simple de n fermes dans une population U afin d'estimer la production moyenne d'orge par hectare dans les fermes de la population. Si x_k est la grandeur de la ferme (en nombre d'hectares de terre) et y_k est la production d'orge (en tonnes), le quotient est le rapport

$$R = \frac{\text{Production totale de toutes les fermes de la population}}{\text{Nombre total d'acres dans toutes les fermes de la population}}.$$

Nous proposerons un estimateur $\hat{R}$ de R, et développerons ses propriétés. Contrairement aux paramètres discutés aux chapitres précédents, ces propriétés sont asymptotiques: elles ne sont vérifiées — approximativement — que lorsque n est grand. Règle générale, donc, si n n'est pas grand, on interprétera les résultats avec circonspection.

Estimateur de R

L'estimateur de R sera celui que l'intuition recommanderait, soit le quotient analogue à R calculée à partir des données de l'échantillon :

$$\text{Estimateur de } R: \ \hat{R} = \frac{\bar{y}}{\bar{x}}.$$

$\hat{R}$ n'est pas un estimateur sans biais de R, mais le biais tend vers 0 lorsque n tend vers l'infini. Il est négligeable lorsque n est assez grand.

Variance de $\hat{R}$

L'estimateur $\hat{R}$ n'étant pas sans biais, on devrait parler de *l'erreur quadratique moyenne* plutôt que de la variance. Mais si n est assez grand, l'erreur quadratique moyenne se

confondra presque avec la variance: $EQM(\hat{R}) \approx V(\hat{R})$. On se permettra donc de désigner l'erreur quadratique moyenne par $V(\hat{R})$.

THÉORÈME 5.4 **ERREUR QUADRATIQUE MOYENNE DE** $\hat{R}$

Lorsque n est grand, $\hat{R}$ est à peu près sans biais, et son erreur quadratique moyenne est donnée approximativement par

$$V(\hat{R}) \approx \frac{(1-f)}{n\mu_x^2}[S_y^2 + R^2 S_x^2 - 2RS_{xy}] \tag{5.3}$$
$$= \frac{1}{\mu_x^2}[V(\bar{y}) + R^2 V(\bar{x}) - 2R\,C(\bar{x};\bar{y})]$$

Une autre expression: **L'équation (5.3) peut également s'écrire comme ceci:**

$$EQM(\hat{R}) \approx \frac{(1-f)}{n\mu_x^2} \frac{\sum_{k=1}^{N}(y_k - Rx_k)^2}{N-1}.$$

Estimation de la variance $V(\hat{R})$

La variance $V(\hat{R})$ est fonction de plusieurs paramètre inconnus : S_y^2, S_x^2, S_{xy}, R et μ_x. Pour l'estimer, on remplacera chaque paramètre inconnu par son estimation: on remplacera S_y^2, S_x^2, S_{xy}, R et μ_x par s_y^2, s_x^2, s_{xy}, $\hat{R}$ et $\bar{x}$, respectivement. L'estimateur est donc :

$$\hat{V}(\hat{R}) = \frac{1-f}{n\bar{x}^2}[s_y^2 + \hat{R}^2 s_x^2 - 2\hat{R}s_{xy}] \tag{5.4}$$
$$= \frac{1}{\bar{x}^2}[\hat{V}(\bar{y}) + \hat{R}^2\hat{V}(\bar{x}) - 2\hat{R}\hat{C}(\bar{x};\bar{y})] \qquad f = \frac{n}{N}.$$

On peut écrire cette formule sous la forme suivante:

$$\hat{V}(\hat{R}) = \frac{1-f}{n\bar{x}^2} \frac{\sum_{i\in\omega}(y_i - \hat{R}x_i)^2}{n-1}. \tag{5.5}$$

Intervalle de confiance pour R

Comme pour les estimations du chapitre précédent, un intervalle de confiance approximatif à 100(1 - α) % est donné par

$$\hat{R} - z_{\alpha/2}\sqrt{\hat{V}(\hat{R})} \le R \le \hat{R} + z_{\alpha/2}\sqrt{\hat{V}(\hat{R})}.$$

Note Nous supposerons, pour tout problème concernant un quotient, que X ne prend que des valeurs positives.

Remarque *Imperfection des estimateurs*

Il est nécessaire de noter les imperfections de ces estimateurs.

1) Tout d'abord, la formule (5.3), même si elle est évaluée à partir des données de la population entière ne donne pas la variance exacte de $\hat{R}$.

2) La formule (5.4), étant calquée sur celle de $V(\hat{R})$, qui n'est pas exactement la variance, estime, au mieux, une *approximation* de la variance et non la variance elle-même.

Bien que ces observations invitent à la prudence, elles n'invalident pas les procédures que nous décrivons ici, qui sont quand même adéquates lorsque l'échantillon est assez grand. Le biais de $\hat{R}$, en tout cas, ne devrait pas causer trop d'inquiétude, comme le montre le théorème suivant :

THÉORÈME 5.5 **UNE BORNE AU BIAIS DE** $\hat{R}$

$$\frac{\left|Biais(\hat{R})\right|}{\sqrt{V(\hat{R})}} \leq \frac{\sqrt{V(\bar{x})}}{\mu_x}.$$

Le membre de gauche mesure l'importance relative du biais. Rappelons que l'erreur quadratique moyenne de $\hat{R}$ comprend deux parties, selon le Théorème 2.1: $EQM(\hat{R}) = V(\hat{R}) + [\mathrm{B}(\hat{R})]^2$. C'est la deuxième composante, $[\mathrm{B}(\hat{R})]^2$, qui doit être petite comparée à la variance $V(\hat{R})$. Le fait que $V(\bar{x})$ tend vers 0 lorsque n tend vers l'infini signifie que le biais relatif sera négligeable quand n est grand.

Exemple 5.1 *Estimation d'un quotient*

Considérons le Tableau 5.1, qui présente la production d'orge de 50 fermes tirées d'une population de 1 579 fermes qui cultivent de l'orge (ce tableau est extrait du tableau A01, qui comprend en outre des données sur 1 764 fermes qui ne cultivent pas l'orge). Les variables sont :

X : La surface cultivée (en hectares)

Y : La production d'orge (en tonnes)

Tableau 5.1
Surface cultivée (X) et production d'orge (Y) d'un échantillon de 50 fermes

x	y	x	y	x	y	x	y	x	y
29,4	92	37,8	122	63,1	182	27,8	87	47	125
81,8	242	66,6	188	36,2	101	41,8	123	23,1	71
46	148	58,4	180	71,7	223	45,1	139	45,9	137
42,3	125	49,9	160	24,8	79	52,4	162	11,1	35
29,1	77	25,5	79	53,2	165	92,1	272	66,5	214
18,8	58	16,6	52	36,1	117	80,9	249	67	210
65,5	190	26,6	79	42,6	129	68,4	187	24,3	67
46	156	60	181	99,8	289	36,4	104	29,9	93
26,2	73	100,9	314	62,2	205	46,9	133	57,3	187
80	250	71,7	225	36,9	120	47,4	141	66,9	197

$n = 50; N = 1579; \bar{x} = 49,678; \bar{y} = 150,68; s_x = 21,74509; s_y = 66,11043; s_{xy} = 1426,179$

L'objectif du sondage est d'estimer R, la production moyenne par hectare.

Solution

$$\hat{R} = \frac{150,68}{49,678} = 3,033133.$$

L'écart-type de $\hat{R}$ est estimé par :

$$\sqrt{\hat{V}(\hat{R})} = \sqrt{1 - \frac{50}{1579}} \frac{\sqrt{(66,11043)^2 + (3,033133)^2(21,74509)^2 - 2(3,033133)(1426,179)}}{\sqrt{50}\,(49,678)}$$

$$= 0,02329879.$$

Un intervalle de confiance à 95 % est donné par

$$\hat{R} - 1,96\sqrt{\hat{V}(\hat{R})} \le R \le \hat{R} + 1,96\sqrt{\hat{V}(\hat{R})} \Leftrightarrow 2,99 \le R \le 3,08.$$

On affirme donc que la production moyenne par hectare se situe entre 2,99 et 3,08 tonnes.

Remarque *Moyenne, proportion, quotient*

Remarquez que nous disons « production *moyenne* » en nous référant à *R*. C'est le terme employé communément dans ce contexte et nous l'utiliserons dans ce sens lorsque celui-ci est conforme à l'usage. Mais ailleurs, les termes « moyenne », « proportion » et « quotient » ne seront employés que dans leur acception technique définie formellement.

PRÉVOIR LA TAILLE DE L'ÉCHANTILLON

Il n'est pas facile de savoir d'avance quelle taille d'échantillon donnerait une estimation acceptable d'un quotient; trop de paramètres doivent être connus : S_x, S_y, S_{xy}, μ_x et/ou μ_y. Nous supposerons qu'ils ont été estimés, d'une façon ou d'une autre, parfois à l'aide d'un échantillon pilote tiré à cette fin, ou d'un échantillon tiré antérieurement dans la même population. Les objectifs de précision peuvent normalement s'exprimer comme ceci :

(5.6) Marge d'erreur *absolue* : $z_{\alpha/2}\sqrt{V(\hat{R})} = E \Leftrightarrow z_{\alpha/2}\sqrt{1-\frac{n}{N}}\frac{\sqrt{S_y^2 + R^2S_x^2 - 2RS_{xy}}}{\mu_x\sqrt{n}} = E$

(5.7) Marge d'erreur *relative* : $z_{\alpha/2}\frac{\sqrt{V(\hat{R})}}{R} = r \Leftrightarrow z_{\alpha/2}\sqrt{1-\frac{n}{N}}\frac{\sqrt{S_y^2 + R^2S_x^2 - 2RS_{xy}}}{\mu_x R\sqrt{n}} = r$.

Si on ignore pour l'instant le facteur de correction, l'équation (5.6) mène à

(5.8) $$n_0 = z_{\alpha/2}^2\frac{S_y^2 + R^2S_x^2 - 2RS_{xy}}{\mu_x^2E^2},$$

alors que (5.7) donne

(5.9) $$n_0 = z_{\alpha/2}^2\frac{S_y^2 + R^2S_x^2 - 2RS_{xy}}{\mu_y^2r^2}.$$

Ces valeurs ne tiennent pas compte du facteur de correction. Pour en tenir compte, on effectue l'ajustement suivant:

(5.10) $$n = \frac{n_0}{1 + n_0 / N}.$$

L'équation (5.9) peut s'écrire comme

(5.11) $$n_0 = z_{\alpha/2}^2\frac{cv^2(y) + cv^2(x) - 2\rho cv(x)cv(y)}{r^2}.$$

où $cv(.)$ désigne le coefficient de variation et ρ est le coefficient de corrélation entre X et Y. Cette formule est utile dans la mesure où un coefficient de variation peut parfois être mieux connu a priori qu'un écart-type.

Exemple 5.2 *Taille d'échantillon, estimation d'un quotient*

Dans le contexte de l'exemple 5.1, on planifie un nouveau sondage, cinq années plus tard, dans la même population afin d'estimer R et on voudrait encore estimer la taille de l'échantillon qu'il faut prélever. On sait bien que les paramètres ont changé de valeur, mais pour les besoins présents, les estimations obtenues à l'exemple peuvent être utilisées.

a) Quelle est la taille de l'échantillon qu'on doit prélever s'il faut que la marge d'erreur absolue (à 95 %) soit de 0,02 (tonnes/hectares).

b) Quelle est la taille de l'échantillon qu'on doit prélever s'il faut que la marge d'erreur relative (à 95 %) soit de 1 %.

Solution

Pour un niveau de confiance à 95 %, $z_{\alpha/2} = z_{0,025} = 1{,}96$.

a) On utilisera l'équation (5.8), mais on remplacera les paramètres par les estimateurs obtenus à l'Exemple 5.1. On a alors

$$s_y^2 + \hat{R}^2 s_x^2 - 2\hat{R}s_{xy} = 69{,}17349, \quad n_0 = z_{\alpha/2}^2 \frac{s_y^2 + \hat{R}^2 s_x^2 - 2\hat{R}s_{xy}}{\bar{x}^2 (0{,}02)^2} = 269; \quad n = 230.$$

Il est évident que les paramètres peuvent avoir changé de valeur depuis 5 ans, le risque de sous-estimer la taille est bien réel. La seule façon de se prémunir contre ce risque est de majorer l'effectif, dans la mesure où les ressources financières le permettent.

b) $cv^2(y) + cv^2(x) - 2\rho cv(x)cv(y) = 0{,}003046691;$

$$n_0 = z_{\alpha/2}^2 \frac{cv^2(y) + cv^2(x) - 2\rho cv(x)cv(y)}{(0{,}01)^2} = 117{,}0417 \; ; \; n = 109.$$

5.2 ESTIMATION DES PARAMÈTRES D'UN DOMAINE U_D

Il arrive qu'on se serve d'un échantillon pour estimer non seulement les paramètres de la population mais aussi ceux d'une *partie* de la population.

Quelques exemples:

- Au service de comptabilité d'une compagnie, on tire un échantillon d'une population de clients dont les comptes sont en souffrance. On constate, une fois l'échantillon tiré, qu'une partie de cette clientèle a déjà posé des problèmes de retards dans leurs paiements. On souhaite alors étudier plus à fond la sous-population des retardataires.
- On tire un *éas* des propriétés de Montréal afin d'estimer leur valeur moyenne. On s'intéresse ensuite à la valeur Y des propriétés d'un arrondissement particulier — Lachine, disons. L'échantillon comprend des maisons de Lachine, et on compte s'en servir afin d'estimer la moyenne de Y à Lachine.

- On tire un échantillon de détenteurs d'une assurance auto. On décide ensuite de se pencher de plus près sur ceux qui ont fait plus d'une réclamation en un an.

Une sous-population dont on cherche à estimer les paramètres est appelée *domaine*.

DÉFINITION 5.2 : DOMAINE

Un domaine est une sous-population dont on souhaite estimer certains paramètres à partir d'un échantillon tiré de la population entière.

Considérons un échantillon tiré d'une population U (de taille N), et supposons qu'on veuille estimer, en plus de certains paramètres relatifs à U, les paramètres d'un domaine $U_d \subset U$.

Notation Les quantités relatives à la population, à l'échantillon et au domaine seront distinguées par les notations suivantes:

	Population		Échantillon	
	Entière (U)	Domaine (U_d)	Entier (ω)	Domaine (ω_d)
Taille	N	N_d	n	n_d
Moyenne	μ	μ_d	$\bar{y}$	$\bar{y}_d$
Écart-type	S	S_d	s	s_d

ESTIMATION DE LA MOYENNE M_D D'UN DOMAINE

Soit Y une variables définie sur les unités de la population, et μ_d la moyenne de Y sur les unités du domaine:

$$\mu_d = \frac{\sum_{k \in U_d} y_k}{N_d}.$$

Soit ω l'échantillon et ω_d le sous-ensemble de ω appartenant à U_d. Il est naturel, pour estimer μ_d, d'utiliser les données de ω_d et de procéder comme si ω_d était un échantillon de taille n_d tiré d'une population U_d de taille N_d. Et lorsque N_d est connu, c'est exactement ce qu'on fait, et en ce qui concerne les applications, il n'y aura rien de nouveau. Mais il est nécessaire de justifier cette procédure théoriquement, car n_d est aléatoire alors que la théorie développée jusqu'ici tient la taille de l'échantillon pour fixe. Par ailleurs, lorsque N_d est inconnue, le fait de ne pas connaître la taille de la population force à modifier la procédure d'estimation, surtout lorsqu'on souhaite estimer le total du domaine.

Exemple 5.3 *Estimation de μ_d, N_d connu*

Le tableau suivant présente des données sur un échantillon de taille 50 tiré d'une population U de 8 654 factures de restaurant. La variable d'intérêt Y est le montant de la facture, et l'échantillon a été tiré afin d'estimer la moyenne μ. Mais on s'interroge également sur le cas des factures émises un week-end (samedi ou dimanche) — le domaine U_d. Voici les valeurs de Y dans l'échantillon, classées selon qu'elles appartiennent au domaine ou pas.

Unités dans U_d (week-end)	Unités à l'extérieur de U_d (jour de semaine)
25,30 31,05 52,90 54,05 56,35 63,25 81,65 92,00 98,90 100,05 123,05 124,20 133,40 155,25	21,85 31,05 35,65 39,10 42,55 47,15 49,45 49,45 49,45 52,90 58,65 59,80 62,10 64,40 64,40 65,55 66,70 69,00 70,15 82,80 86,25 86,25 87,40 90,85 94,30 96,60 100,05 106,95 106,95 110,40 124,20 124,20 131,10 133,40 142,60 151,80

Soit μ_d la moyenne de U_d. L'échantillon comprend n_d = 14 unités dans le domaine U_d. On les considérera comme un échantillon de 14 factures tirées du domaine U_d, traité comme une population. L'estimateur de μ_d est la moyenne des montants des 14 factures du domaine: $\bar{y}_d$ = 85,10.

Lorsque N_d est connue, la formule est évaluée telle quelle. Supposons que N_d = 3234, n_d = 14, et l'écart-type des données du domaine est s_d = 39,70928. Nous avons alors

$$\sqrt{\hat{V}(\bar{y}_d)} = \sqrt{1-f_d}\,\frac{s_d}{\sqrt{n_d}} = \sqrt{1-\frac{14}{3234}}\,\frac{39,70928}{\sqrt{14}} = 10,58976.$$

L'intervalle de confiance est donc

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

À un niveau de 95 %. cela donne 85,10 ± 1,96(10,58976) :

$$64,34 \le \mu_d \le 105,86.$$

Lorsque N_d n'est pas connu, le facteur de correction $f_d = \frac{n_d}{N_d}$ doit être remplacé par $\hat{f}_d = \frac{n_d}{\hat{N}_d}$.

où $\hat{N}_d = \frac{n_d}{n}N$ est un estimateur de N_d. Le facteur de correction est estimé par $1-\hat{f}_d$, où $\hat{f}_d = \frac{n_d}{\hat{N}_d} = \frac{n}{N}$. On estimera alors $\sqrt{V(\bar{y}_d)}$ par

$$\sqrt{\hat{V}(\bar{y}_d)} = \sqrt{1-\frac{n}{N}}\,\frac{s_d}{\sqrt{n_d}}\,.$$

Exemple 5.4 *Estimation de la moyenne, N_d inconnu*

Dans l'exemple 5.3, supposons que N_d n'est pas connu. Alors l'écart-type de $\bar{y}_d$ est estimé par

$$\sqrt{\hat{V}(\bar{y}_d)} = \sqrt{1-\frac{n}{N}}\,\frac{s_d}{\sqrt{n_d}} = \sqrt{1-\frac{50}{8654}}\,\frac{39,70928}{\sqrt{21}} = 10,58205$$

une estimation relativement proche de la première (10,58976). L'intervalle de confiance est

$$64,36 \le \mu_d \le 105,84.$$

ESTIMATION DU TOTAL T_D D'UN DOMAINE

L'estimateur du total τ_d du domaine est normalement $\hat{\tau}_d = N_d \overline{y}_d$. Ici, la démarche change radicalement selon que N_d est connue ou pas.

N_d connu

Si N_d est connu, il n'y a aucune difficulté nouvelle, nous suivons la démarche habituelle. Nous désignons l'estimateur du total par $\hat{\tau}_{d1}$ pour le distinguer du cas où N_d n'est pas connu :

$$\hat{\tau}_{d1} = N_d \times \overline{y}_d .$$

L'écart-type de $\hat{\tau}_{d1}$ est

$$\sqrt{V(\hat{\tau}_{d1})} = N_d \sqrt{V(\overline{y}_d)} .$$

Exemple 5.5 *Estimation de T_d, N_d connu*

Dans l'exemple 5.3, si N_d = 3 234, on a alors $\hat{\tau}_{d1}$ = 3 234 ×85,10 = 275 213,40. $\sqrt{V(\hat{\tau}_{d1})}$ = 34 247,27, et l'intervalle de confiance pour τ_d est

$$\hat{\tau}_{d1} - z_{\alpha/2}\sqrt{\hat{V}(\hat{\tau}_{d1})} \leq \tau_d \leq \hat{\tau}_{d1} + z_{\alpha/2}\sqrt{\hat{V}(\hat{\tau}_{d1})} .$$

À 95 %, cela donne

$$208\ 089 \leq \tau_d \leq 342\ 338.$$

N_d inconnu

Quand N_d est inconnu, il faut remplacer N_d par $\hat{N}_d = (n_d / n)N$ et estimer τ_d par

$$\hat{\tau}_{d2} = \hat{N}_d \overline{y}_d .$$

Comment estimer l'écart-type de $\hat{\tau}_{d2}$? On ne peut pas, comme on l'a fait précédemment, multiplier par $\hat{N}_d$ l'estimateur $\sqrt{\hat{V}(\overline{y}_d)}$, car $\hat{N}_d$ n'est pas une constante.

La solution, pourtant, n'est pas difficile, bien qu'à première vue elle puisse sembler artificielle. Elle consiste à supposer que les montants des factures à l'extérieur du domaine sont tous nuls. Pourquoi ? Parce qu'alors, le total du domaine est exactement égal au total de la population entière — et le problème est réduit à l'estimation d'un total, tout simplement. En pratique, donc, il suffit de remplacer par 0 toutes les valeurs correspondant à des unités extérieures au domaine et de procéder normalement à l'estimation du total.

Exemple 5.6 *Estimation de* τ_d, N_d *inconnu*

Supposons, dans l'exemple 5.3, que N_d est inconnu. On remplace par 0 tout montant correspondant à un jour de semaine. L'échantillon présenté se voit donc transformé en celui-ci :

25,30	31,05	52,9	54,05	56,35	63,25	81,65	92	98,9	100,05
123,05	124,20	133,4	155,25	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0

Cela équivaut à définir une nouvelle variable, Y', qui prend les mêmes valeurs que Y dans le domaine, et 0 ailleurs. À partir d'ici, on estime un total, tout simplement. On a n = 50, N = 8 654. La moyenne échantillonnale, qu'on désigne par $\bar{y}'$, est la moyenne des 50 données ci-dessus, incluant les zéros; de même, l'écart-type, s', est l'écart-type des 50 données, incluant les zéros.

On obtient les résultats

$$\bar{y}' = 23{,}828 \qquad s' = 43{,}68211.$$

On estime donc le total par

$$\hat{\tau}_{d2} = N\bar{y}' = 8654(23{,}828) = 206\ 207{,}50.$$

L'écart-type de l'estimateur est estimé par

$$\sqrt{V(\hat{\tau}_{d2})} = N\sqrt{1-\frac{n}{N}}\frac{s'}{\sqrt{n}} = (8654)\sqrt{1-\frac{50}{8654}}\frac{43{,}68211}{\sqrt{50}} = 53\ 306{,}14.$$

Un intervalle de confiance à 95 % pour le total et donc

$$\hat{\tau}_{d2} \pm z_{\alpha/2}\sqrt{\hat{V}(\hat{\tau}')} = 206\ 207{,}5 \pm 1{,}96(53\ 306{,}14).$$

$$101\ 7298 \leq \tau_d \leq 310\ 688.$$

Sans surprise, cet intervalle est beaucoup plus large que le précédent: le fait de ne pas connaître la taille du domaine a un coût. Ainsi dans l'exemple traité ici, l'écart-type de l'estimateur $\hat{\tau}_{d1} = N_d\bar{y}_d$ est estimé à 34 247,27 alors que celui de $\hat{\tau}_{d2}$ est estimé à 53 306,14, près de 1,6 fois supérieur. Il fallait s'y attendre : l'estimation de N_d dans $\hat{\tau}_{d2}$ introduit une source d'erreur supplémentaire qui n'est pas présente quand N_d est connue.

Contrairement aux cas précédent, la variance de l'estimateur $\hat{\tau}_{d2}$ n'est pas une variance conditionnelle.

5.3 JUSTIFICATIONS ET DISCUSSION

Les procédures décrites ci-dessus peuvent se justifier de deux façons. La première consiste à prendre le taureau par les cornes: puisque n_d est aléatoire, la moyenne $\bar{y}_d = \frac{\sum_{i\in\omega_d} y_i}{n_d}$ est en fait un quotient. Et on la traite comme tel. On verra que, étant donné les approximations faites dans l'estimation d'un quotient, cela revient en fin de compte à traiter n_d comme s'il était fixe. La deuxième justification, une approche conditionnelle, est intéressante d'un point de vue théorique car elle sert également à justifier d'autres procédures.

LA MOYENNE D'UN DOMAINE CONSIDÉRÉE COMME QUOTIENT

La moyenne μ_d est en fait un quotient, et l'estimateur proposé $\bar{y}_d$ l'est aussi. On définit deux variables, Y' et X:

$$y_k = \begin{cases} y_k & \text{si } k \in U_d \\ 0 & \text{sinon} \end{cases} \quad , \quad x_k = \begin{cases} 1 & \text{si } k \in U_d \\ 0 & \text{sinon} \end{cases} .$$

Alors $\mu_d = \frac{\sum_{k=1}^{N} y'_k}{\sum_{k=1}^{N} x_k}$ et $\bar{y}_d = \frac{\sum_{i\in\omega} y'_i}{\sum_{i\in\omega} x_i}$. Il suffit alors d'utiliser les techniques établies pour l'estimation d'un quotient. L'estimateur (5.5) de la variance donne $\hat{V}(\bar{y}_d) = (1-f)\frac{n(n_d-1)}{n_d(n-1)}\frac{s_d^2}{n_d}$, ce qui est assez proche de $\hat{V}(\bar{y}_d) = (1-f)\frac{s_d^2}{n_d}$, la formule qu'on utiliserait si on tenait n_d pour fixe (et N_d inconnu).

PROPRIÉTÉS CONDITIONNELLES DE L'ESTIMATEUR DE LA MOYENNE D'UN DOMAINE

On peut considérer n_d fixe à condition de reconnaître que les propriétés qui en découlent sont des propriétés *conditionnelles*: l'espérance *conditionnelle* de l'estimateur et la variance *conditionnelle* de l'estimateur étant donné la valeur observée n_d. Le théorème pertinent est le suivant.

THÉORÈME 5.6 **PROBABILITÉ CONDITIONNELLE SOUS-ÉCHANTILLON**

Soit $U = U_1 \cup U_2$ une partition de la population U en deux sous-populations de tailles N_1 et N-N_1, respectivement. Soit ω un *éas* de taille n tiré de cette population, $\omega_1 = \omega \cap U_1$ et η la cardinalité de ω_1. Alors, pour toute paire de sous-ensembles (ω_{1o} ; ω_{20}), $\omega_{10} \subset U_1$ et $\omega_{20} \subset U_2$ de cardinalité n_1 et $n_2 = n - n_1$, respectivement,

- $P(\omega_1 = \omega_{10} | \eta_1 = n_1) = \dfrac{1}{\binom{N_1}{n_1}}$,
- $P(\omega \backslash \omega_1 = \omega_{20} | \eta_1 = n_1) = \dfrac{1}{\binom{N-N_1}{n-n_1}}$, et
- $P(\omega_1 = \omega_{10} \text{ et } \omega \backslash \omega_1 = \omega_{20} | \eta_1 = n_1) = \dfrac{1}{\binom{N_1}{n_1}} \dfrac{1}{\binom{N_2}{n_2}}$.

L'implication de ce théorème est que si on restreint l'espace échantillon à l'ensemble des échantillons pour lesquels ω_1 est de taille n_1, alors les propriétés de ω_1 et de $\omega \backslash \omega_1$ sont exactement celles de deux *éas* indépendants, tirés de U_1 et de U_2, respectivement.

Dans le cas de l'estimation de la moyenne d'un domaine cela veut dire que si on restreint l'espace échantillon à l'ensemble des échantillons pour lesquels ω_d est de taille n_d, alors les propriétés de ω_d sont celles d'un échantillon aléatoires simple de taille n_d tiré de U_d. On en déduit immédiatement que $E(\bar{y}_d \mid \eta_d) = \mu_d$, et que $V(\bar{y}_d \mid \eta_d) = (1 - f_d)\dfrac{S_d^2}{\eta_d}$, où $f_d = \dfrac{\eta_d}{N_d}$.

Remarque $\bar{y}_d$ *inconditionnellement sans biais*

Le fait que l'espérance $E(\bar{y}_d \mid \eta_d) = \mu_d$ ne dépend pas de η_d signifie que $\bar{y}_d$ n'est pas seulement *conditionnellement* sans biais : il est sans biais tout court.

La variance conditionnelle, par contre, dépend carrément de η_d. Peut-on s'en contenter? Ou est-il nécessaire de déterminer la variance inconditionnelle? Avant d'aborder cette question, voyons comment on calculerait la variance inconditionnelle.

Variance inconditionnelle

On peut déterminer une expression approximative pour la variance inconditionnelle en utilisant la formule $V(\bar{y}_d) = V[E(\bar{y}_d \mid \eta_d)] + E[V(\bar{y}_d \mid \eta_d)]$ (équation (1.22)). Puisque $E(\bar{y}_d \mid \eta_d) = \mu_d$ est un nombre fixe, $V[E(\bar{y}_d \mid \eta_d)] = 0$ et donc $V(\bar{y}_d) = E[V(\bar{y}_d \mid \eta_d)] = E\left[(1 - f_d)\dfrac{S_d^2}{\eta_d}\right] = S_d^2 E\left(\dfrac{1}{\eta_d}\right) - \dfrac{S_d^2}{N_d}$. L'espérance $E\left(\dfrac{1}{\eta_d}\right)$ peut être déterminée à l'aide de la loi hypergéométrique en fonction des paramètres n, N_d et $N - N_d$. (Il faudrait quand même imposer une condition, soit $\eta_d \neq 0$, sans quoi l'espérance ne serait même pas définie).

On pourrait également appliquer la méthode de linéarisation (section 2.3) pour approcher $E\left(\dfrac{1}{\eta_d}\right)$. On a alors $\dfrac{1}{\eta_d} \approx \dfrac{1}{np_d} - \dfrac{\eta_h - np_d}{(np_d)^2} + \dfrac{(\eta_h - np_d)^2}{(np_d)^3}$, où $p_d = N_d / N$. Si on s'en tient au premier terme, $E\left(\dfrac{1}{\eta_d}\right) \approx \dfrac{1}{np_d}$, et si on estime p_d par n_d / n, cela revient à remplacer $E\left(\dfrac{1}{\eta_d}\right)$ par $\dfrac{1}{n_d}$ et à

déclarer que la formule de la variance conditionnelle est une approximation acceptable de la variance inconditionnelle.

On pourrait améliorer légèrement l'approximation en conservant les trois premiers termes du développement de $\frac{1}{\eta_d}$. Dans ce cas, on remplacerait $\frac{1}{n_d}$ dans la formule par $E\left(\frac{1}{\eta_d}\right) \approx \frac{1}{np_d} + \frac{(1-n_d/n)^2}{n_d^2}$.

Remarque *L'estimateur de la variance est inconditionnellement sans biais*

Il est intéressant de noter que si $V(\bar{y}_d) = (1-f_d)S_d^2/\eta_d = \frac{S_d^2}{\eta_d} - \frac{S_d^2}{N_d}$ est la variance *conditionnelle* de $\bar{y}_d$, l'estimateur $\hat{V}(\bar{y}_d) = (1-f_d)s_d^2/\eta_d$ est sans biais pour la variance *inconditionnelle* : $E[\hat{V}(\bar{y}_d)] = E(E[\hat{V}(\bar{y}_d) \mid \eta_d]) = E\left\{E\left[\frac{s_d^2}{\eta_d} - \frac{s_d^2}{N_d} \middle| \eta_d\right]\right\} = S_d^2 E\left(\frac{1}{\eta_d}\right) - \frac{S_d^2}{N_d} = S_d^2 E\left(\frac{1}{\eta_d}\right) - \frac{S_d^2}{N_d} = V(\bar{y}_d)$.

Variance conditionnelle ou inconditionnelle?

La variance $\frac{S_d^2}{n_d} - \frac{S_d^2}{N_d}$ est une variance conditionnelle, alors que la variance inconditionnelle est $S_d^2 E\left(\frac{1}{\eta_d}\right) - \frac{S_d^2}{N_d}$. Dans l'exemple traité ici, S_d = 69,9566, N_d = 3234, et n_d = 14. La variance conditionnelle étant donné η_d = 14 est donc $\frac{(39,9566)^2}{14} - \frac{(39,9566)^2}{3234}$ = 113,54. C'est la variance de $\bar{y}_d$ calculée sur tous les échantillons de taille 50 dans lesquels le nombre d'unités appartenant au domaine est 14. La variance *inconditionnelle* est calculée sur l'ensemble de tous les échantillons, dans lesquels n_d prend toutes les valeurs possibles (à l'exception de n_d = 0). Dans ce cas, on obtient une variance inconditionnelle de 87.

Laquelle de ces deux variances est plus pertinente comme mesure de la qualité de l'estimateur? Rappelons pourquoi on s'intéresse à la variance d'un estimateur (quand il est sans biais). Si la variance est petite, cela veut dire que si on tire un grand nombre d'échantillons (en fait *tous* les échantillons possibles), l'estimation sera souvent proche du paramètre. C'est ce qui nous conforte, une fois l'échantillon tiré, dans l'idée que l'estimation calculée n'est pas trop éloignée du paramètre. On ne peut pas savoir si l'estimateur qu'on vient de calculer est proche ou non du paramètre. On porte un jugement sur ce qui est arrivé en considérant l'ensemble de tout ce *qui aurait pu arriver*. Doit-on inclure, parmi ce qui aurait pu arriver, les échantillons dans lesquels η_d = 1, 2, …, 50? Doit-on, par exemple, s'inquiéter de ce que η_d *aurait pu* être égale à 1 (et aurait donné une variance de 1 576) ? Nous n'allons pas poursuivre cette discussion plus longtemps car la réponse dépend du

contexte. On peut dire quand même que dans le problème traité ici il est parfaitement raisonnable de se fier à la variance conditionnelle.

5.4 DIFFÉRENCE DE DEUX MOYENNES

Les implications du Théorème 5.6 s'étendent à d'autres problèmes, entre autres celui qui concerne la différence de deux moyennes et traité à la section 3.5. Mais là il s'agissait de deux échantillons *indépendants* tirés dans deux populations distinctes, U_1 et U_2. Ce n'est pas toujours le cas: les deux échantillons sont parfois issus d'un même échantillon ω tiré d'une population U partitionnée en deux sous-populations, $U = U_1 \cup U_2$. L'échantillon ω est ensuite réparti en deux, conformément à la partition de U: $\omega = \omega_1 \cup \omega_2$, $\omega_1 \subset U_1$, $\omega_2 \subset U_2$. Supposons, par exemple, qu'on tire un échantillon dans la liste des personnes inscrites à l'ordre des ingénieurs, et qu'on veuille estimer la différence de salaire entre les hommes et les femmes. L'échantillon est constitué d'hommes et de femmes, mais ces deux groupes ne sont pas des échantillons indépendants. On devine, cependant, et le Théorème 5.6 le confirme, que la dépendance résulte uniquement de la variabilité des *tailles* des deux groupes, et donc qu'une fois ces tailles fixées, les observations des deux groupes sont indépendantes. On développera donc les propriétés conditionnelles des deux échantillons. Ainsi les variances calculées et les intervalles de confiance sont alors conditionnels.

Soit $\omega = \omega_1 \cup \omega_2$ un *éas* d'une population partitionnée $U = U_1 \cup U_2$, où $\omega_1 = \omega \cap U_1$ et $\omega_2 = \omega \cap U_2$. Soit N_1 et N_2 les tailles de U_1 et U_2, respectivement. Soit Y une variable définie sur U, μ_1 et μ_2 les moyennes de Y dans U_1 et U_2, $\bar{y}_1$ et $\bar{y}_2$ les moyennes de Y dans ω_1 et ω_2. Soit η_1 la taille de ω_1. Alors $E(\bar{y}_1 - \bar{y}_2 \mid \eta_1) = \mu_1 - \mu_2$ et

$$V(\bar{y}_1 - \bar{y}_2 \mid \eta_1) = V(\bar{y}_1 \mid \eta_1) + V(\bar{y}_2 \mid \eta_1) = (1 - f_1) S_1^2 / \eta_1 + (1 - f_2) S_2^2 / \eta_2 ,$$

où S_1^2 et S_2^2 sont les variances des deux sous-populations et $\eta_2 = n - \eta_1$. La variance conditionnelle de $\bar{y}_1 - \bar{y}_2$ est estimée par

$$\hat{V}(\bar{y}_1 - \bar{y}_2 \mid \eta = n_1) = \hat{V}(\bar{y}_1 \mid \eta_1 = n_1) + \hat{V}(\bar{y}_2 \mid \eta_1 = n_1) = (1 - f_1) s_1^2 / n_1 + (1 - f_2) s_2^2 / n_2 ,$$

où s_1^2 et s_2^2 sont les variances de ω_1 et ω_2. Un intervalle de confiance approximatif de niveau 1-α est donné par :

$$(\bar{y}_1 - \bar{y}_2) - z_{\alpha/2} \sqrt{\hat{V}(\bar{y}_1) + \hat{V}(\bar{y}_2)} \le \mu_1 - \mu_2 \le (\bar{y}_1 - \bar{y}_2) + z_{\alpha/2} \sqrt{\hat{V}(\bar{y}_1) + \hat{V}(\bar{y}_2)} .$$

Exemple 5.7 *Estimation de la différence de deux moyennes*

Utilisons les données de l'Exemple 5.3 pour estimer la différence entre la facture moyenne le week-end et la facture moyenne en semaine. Le nombre de factures dans la population est de 3 234 le week-end et de 5 420 en semaine.

Solution

Les données de base sont : n_1 = 14, n_2 = 36, $\bar{y}_1$ = 85,1, $\bar{y}_2$ = 79,31806, s_1 = 39,70928, s_2 = 33,659. Les écarts-types conditionnels sont:

$$\sqrt{\hat{V}(\bar{y}_1)} = 10{,}58976,\ \sqrt{\hat{V}(\bar{y}_2)} = 5{,}591171,\ \sqrt{\hat{V}(\bar{y}_1 - \bar{y}_2)} = 11{,}97515.$$

Un intervalle de confiance à 95 % est donné par

$$5{,}7819 - 1{,}96(11{,}97515) \le \mu_1 - \mu_2 \le 5{,}7819 + 1{,}96(11{,}97515)$$

$$-17{,}69 \le \mu_1 - \mu_2 \le 29{,}25.$$

Remarque $\hat{V}(\bar{y}_1 - \bar{y}_2)$ *est l'estimateur de la variance inconditionnelle*

Bien que toutes ces estimations soient basées sur la distribution conditionnelle de $\bar{y}_1$ et $\bar{y}_2$, et que $(1-f_1)\frac{S_1^2}{n_1} + (1-f_2)\frac{S_2^2}{n_2}$ soit la variance conditionnelle, la statistique $\hat{V}(\bar{y}_1 - \bar{y}_2)$ $= (1-f_1)s_1^2/n_1 + (1-f_2)s_2^2/n_2$ est un estimateur sans biais de la variance *inconditionnelle*:

$$E[\hat{V}(\bar{y}_1 - \bar{y}_2)] = E\left\{E[\hat{V}(\bar{y}_1 - \bar{y}_2) \mid n_1]\right\} = E\left\{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_1} - \frac{S_1^2}{N_1} - \frac{S_2^2}{N_2}\right\} = S_1^2 E\left(\frac{1}{n_1}\right) + S_2^2 E\left(\frac{1}{n_2}\right) -$$

$$S_1^2 E\left(\frac{1}{N_1}\right) - S_2^2 E\left(\frac{1}{N_2}\right) = V(\bar{y}_1 - \bar{y}_2) \ .$$

En ce qui concerne l'intervalle de confiance, il est également non conditionnel. On peut affirmer que la probabilité conditionnelle que l'intervalle contienne le paramètre (μ_1 - μ_2) est d'à peu près 1 - α (sous réserve d'une approximation normale adéquate). Or, puisque cette probabilité ne dépend pas de n_1, c'est également la probabilité tout court.

5.5 DONNÉES DICHOTOMIQUES

Les variables Y et X peuvent bien être dichotomiques sans que la théorie doive être modifiée. Mais le langage change : on parlera de *proportions* plutôt que de *moyennes*, et les écarts-types s, s_1 et s' prennent une forme relativement simple.

Soit U une population dont certaines unités appartiennent à une classe particulière C, ω un *éas* de taille n; U_d un domaine de taille N_d. Soit n_d le nombre d'unités dans $\omega_d = \omega \cap U_d$, x le nombre d'unités dans $\omega \cap C$, et x_d le nombre d'unités dans $\omega_d \cap C$. Le tableau suivant présente la correspondance entre les formules générales et celles spécifiques aux données dichotomiques.

Formules générales	Formules propres aux variables dichotomiques	Formules générales	Formules propres aux variables dichotomiques
$\overline{y}$	$\hat{p} = \frac{x}{n}$	s	$\sqrt{\frac{n\hat{p}(1-\hat{p})}{n-1}}$
$\overline{y}_1$	$\hat{p}_d = \frac{x_d}{n_d}$	s_d	$\sqrt{\frac{n_d\hat{p}_d(1-\hat{p}_d)}{n_d-1}}$
$\overline{y}'$	$\hat{p}' = \frac{x_d}{n}$	s'	$\sqrt{\frac{n\hat{p}'(1-\hat{p}')}{n-1}}$

Exemple 5.8 *Estimation de proportions*

On tire un échantillon de 165 logements dans un quartier comprenant 1 200 logements. L'objectif est d'estimer la proportion p de logements sans détecteur de fumée.

a) Dans l'échantillon, on trouve 50 maisons sans détecteurs de fumée. Estimer p ainsi que le nombre de logements sans détecteur de fumée. Estimer le coefficient de variation de l'estimateur.

b) On décide après coup d'estimer la proportion p_d de logements sans détecteur de fumée parmi les *logements unifamiliaux*. Dans l'échantillon, 45 logements sont unifamiliaux, et 10 de ceux-là n'ont pas de détecteur de fumée. Estimer p_d ainsi que l'écart-type de l'estimateur.

c) Estimer N_{cd}, le nombre total de logements unifamiliaux qui n'ont pas de détecteur de fumée et estimer l'écart-type de l'estimateur.

d) Supposons que les habitations ont été répertoriées, de sorte qu'on sait que le nombre de logements unifamiliaux dans la population est 300. Utiliser cette information pour estimer N_{cd} et estimer l'écart-type de l'estimateur.

Solution

a) $\hat{p} = 0,3030;\ s = \sqrt{\frac{n\hat{p}(1-\hat{p})}{n-1}} = 0,46097\ ;\ \sqrt{\hat{V}(\hat{p})} = \sqrt{1-\frac{n}{N}}\frac{s}{\sqrt{n}} = 0,0333$. Coefficient de variation estimé: 10,9 %.

b) $n_d = 45;\ x_d = 10;\ \hat{p}_d = \frac{x_d}{n_d} = 0,2222;\ s_d = \sqrt{\frac{n_d\hat{p}_d(1-\hat{p}_d)}{n_d-1}} = 0,42044$;

$$\sqrt{\hat{V}(\hat{p}_d)} = \sqrt{1-\frac{n}{N}}\frac{s_d}{\sqrt{n_d}} = 0,058207\ .$$

c) $\hat{p}' = \frac{x_d}{n} = 0,06061;\ s' = \sqrt{\frac{n\hat{p}'(1-\hat{p}')}{n-1}} = 0,2393;\ \hat{N}_{cd} = N\hat{p}' = 72,72727$;

$$\sqrt{\hat{V}(\hat{N}_{cd})} = N\sqrt{\hat{V}(\hat{p}')} = N\sqrt{1-\frac{n}{N}}\frac{s'}{\sqrt{n}} = 20,76447.$$

d) $n_d = 45; N_d = 300;\ N_d\hat{p}_d = 66,6667\ ; s_d = \sqrt{\dfrac{n_d\hat{p}_d(1-\hat{p}_d)}{n_d - 1}} = 0,42044\ ;$

$$N_d\sqrt{\hat{V}(\hat{p}_d)} = 17,335.$$

5.6 DÉMONSTRATIONS

THÉORÈME 5.1 **ESTIMATEUR D'UNE COVARIANCE**

La covariance échantillonnale est sans biais pour la covariance de la population :

$$E(s_{xy}) = S_{xy}.$$

Démonstration

Définissons la variable $z_k = x_k + y_k$. Par le Théorème 3.2. $E(s_z^2) = S_z^2$. Par ailleurs, on établit aisément la décomposition $s_z^2 = s_x^2 + s_y^2 + 2s_{xy}$ ainsi que la décomposition analogue de S_z^2 . Alors

$E(s_z^2) = S_z^2 \Rightarrow E(s_x^2) + E(s_y^2) + 2E(s_{xy}) = S_x^2 + S_y^2 + 2S_{xy}$. Étant donné que $E(s_x^2) = S_x^2$ et $E(s_y^2) = S_y^2$, cela entraîne $2E(s_{xy}) = 2S_{xy} \Rightarrow E(s_{xy}) = S_{xy}$.

THÉORÈME 5.2 **COVARIANCE ENTRE DEUX MOYENNES**

La covariance entre $\bar{x}$ et $\bar{y}$ est donnée par

$$C(\bar{x};\bar{y}) = (1-f)\frac{S_{xy}}{n}\ ,\quad f = \frac{n}{N}\ .$$

Démonstration

La démonstration suit celle du théorème précédent: $V(\bar{z}) = V(\bar{x}) + V(\bar{y}) + 2C(\bar{x};\bar{y})$. Par ailleurs, $V(\bar{z}) = (1-f)\dfrac{S_z^2}{n} = (1-f)\dfrac{S_x^2 + S_y^2 + 2S_{xy}}{n} = (1-f)\dfrac{S_x^2}{n} + (1-f)\dfrac{S_y^2}{n} + 2(1-f)\dfrac{S_{xy}}{n}$

$$= V(\bar{x}) + V(\bar{y}) + 2(1-f)\frac{S_{xy}}{n}\ .$$

Donc $V(\bar{x}) + V(\bar{y}) + 2C(\bar{x};\bar{y}) = V(\bar{x}) + V(\bar{y}) + 2(1-f)\dfrac{S_{xy}}{n} \Rightarrow C(\bar{x};\bar{y}) = (1-f)\dfrac{S_{xy}}{n}$.

THÉORÈME 5.3 **ESTIMATION DE LA COVARIANCE ENTRE DEUX MOYENNES**

$\hat{C}(\bar{x};\bar{y}) = (1-f)\dfrac{s_{xy}}{n}$ est un estimateur sans biais de $C(\bar{x};\bar{y})$

Démonstration

Découle directement du Théorème 5.1 et du Théorème 5.2.

THÉORÈME 5.4 **ERREUR QUADRATIQUE MOYENNE DE** $\hat{R}$

Lorsque n est grand, $\hat{R}$ est à peu près sans biais, et son erreur quadratique moyenne est donnée approximativement par $EQM(R) \approx V(\hat{R}) \approx (1-f)\dfrac{S_y^2 + R^2 S_x^2 - 2RS_{xy}}{n\mu_x^2}$.

Démonstration

Nous utilisons maintenant la technique de linéarisation définie à la section 2.3.

$$\hat{R} = \frac{\overline{y}}{\overline{x}} = f(\overline{y};\overline{x}) \approx f(\mu_y;\mu_x) + \frac{\partial f(\mu_y;\mu_x)}{\partial \mu_y}(\overline{y}-\mu_y) + \frac{\partial f(\mu_y;\mu_x)}{\partial \mu_x}(\overline{x}-\mu_x)$$

$$= \frac{\mu_y}{\mu_x} + \frac{1}{\mu_x}(\overline{y}-\mu_y) - \frac{\mu_y}{\mu_x^2}(\overline{x}-\mu_x) = \frac{1}{\mu_x}\left[\mu_y + (\overline{y}-\mu_y) - R(\overline{x}-\mu_x)\right]$$

Alors $E(\hat{R}) \approx \dfrac{1}{\mu_x}\left[\mu_y + E(\overline{y}-\mu_y) - R\,E(\overline{x}-\mu_x)\right] = \dfrac{1}{\mu_x}\left[\mu_y + 0 - R\,(0)\right] = R.$

On en conclut que $\hat{R}$ est à peu près sans biais et donc que $V(\hat{R}) \approx EQM(\hat{R})$.

$$V(\hat{R}) \approx \frac{1}{\mu_x^2} V\left[\mu_y + (\overline{y}-\mu_y) - R(\overline{x}-\mu_x)\right]$$

$$= \frac{V(\overline{y}-\mu_y) + R^2 V(\overline{x}-\mu_x) - 2R \times C[(\overline{x}-\mu_x);(\overline{y}-\mu_y)]}{\mu_x^2} = \frac{V(\overline{y}) + R^2 V(\overline{x}) - 2R \times C(\overline{x};\overline{y})}{\mu_x^2}$$

$$= \frac{(1-f)S_y^2 + R^2(1-f)S_x^2 - 2(1-f)RS_{xy}}{n\mu_x^2} = \frac{(1-f)}{n\mu_x^2}[S_y^2 + R^2 S_x^2 - 2RS_{xy}].$$

THÉORÈME 5.5 **UNE BORNE AU BIAIS DE** $\hat{R}$

$$\frac{\left|Biais(\hat{R})\right|}{\sqrt{V(\hat{R})}} \leq \frac{\sqrt{V(\overline{x})}}{\mu_x}.$$

Démonstration

Considérons la covariance entre $\bar{x}$ et $\hat{R}$: $C(\bar{x};\hat{R}) = E(\bar{x}\hat{R}) - E(\bar{x})E(\hat{R}) = E(\bar{y}) - E(\bar{x})E(\hat{R}) = \mu_y - \mu_x E(\hat{R}) = \mu_x(R - E(\hat{R}))$. Donc $|E(\hat{R}) - R| = \frac{|C(\bar{x};\hat{R})|}{\mu_x} = \frac{|\rho\sqrt{V(\bar{x})}\sqrt{V(\hat{R})}|}{\mu_x} \leq \frac{\sqrt{V(\bar{x})}\sqrt{V(\hat{R})}}{\mu_x}$

$\Rightarrow \frac{|E(\hat{R}) - R|}{\sqrt{V(\hat{R})}} \leq \frac{\sqrt{V(\bar{x})}}{\mu_x}$.

THÉORÈME 5.6 **PROBABILITÉ CONDITIONNELLE**

Soit $U = U_1 \cup U_2$ une partition d'une population U de taille N en deux sous-populations de tailles N_1 et N_2, respectivement. Soit ω un *éas* de taille n tiré de cette population; $\omega_1 = \omega \cap U_1$ et $\omega_2 = \omega \cap U_2$; η_1 la taille de ω_1. Soit $\omega_{10} \subset U_1$ et $\omega_{20} \subset U_2$ deux sous-ensembles de tailles n_1 et n_2, respectivement, $n_1 + n_2 = n$. Alors $P(\omega_1 = \omega_{10} \mid \eta_1 = n_1) = \frac{1}{\binom{N_1}{n_1}}$, $P(\omega_2 = \omega_{20} \mid \eta_1 = n_1) = \frac{1}{\binom{N_2}{n_2}}$, et

$P(\omega_1 = \omega_{10} \text{ et } \omega_2 = \omega_{20} \mid \eta_1 = n_1) = \frac{1}{\binom{N_1}{n_1}} \frac{1}{\binom{N_2}{n_2}}$.

Démonstration

$P(\omega_1 = \omega_{10} \mid \eta_1 = n_1) = P(\omega_1 = \omega_{10} \text{ et } Card(\omega_1) = n_1)/P(Card(\omega_1) = n_1) = \frac{P(\omega_1 = \omega_{10})}{P(Card(\omega_1) = n_1)}$. Le nombre d'échantillons dont la partie ω_1 est ω_{10} est le nombre de choix possibles de ω_2, soit $\binom{N_2}{n_2}$. Donc $P(\omega_1 = \omega_{10}) = \frac{\binom{N_2}{n_2}}{\binom{N}{n}}$. Le nombre d'échantillons dans lesquels la cardinalité de ω_1 est n_1 est $\binom{N_1}{n_1}\binom{N_2}{n_2}$. Donc $P(Card(\omega_1) = n_1) = \frac{\binom{N_1}{n_1}\binom{N_2}{n_2}}{\binom{N}{n}}$. Donc $P(\omega_1 = \omega_{10} \mid n_1)$

$= \frac{\binom{N_2}{n_2}\Big/\binom{N}{n}}{\binom{N_1}{n_1}\binom{N_2}{n_2}\Big/\binom{N}{n}} = \frac{1}{\binom{N_1}{n_1}}$. De même $P(\omega_2 = \omega_{20} \mid \eta_1 = n_1) = \frac{1}{\binom{N_2}{n_2}} \cdot \frac{\binom{N_2}{n_2}}{\binom{N}{n}}$,

$P(\omega_1 = \omega_{10} \text{ et } \omega_2 = \omega_{20} \mid \eta_1 = n_1) = \frac{P[\omega_1 = \omega_{10}, \omega_2 = \omega_{20}, \text{et } \eta_1 = n_1]}{P(Card(\omega_1) = n_1)} = \frac{1\Big/\binom{N}{n}}{\binom{N_1}{n_1}\binom{N_2}{n_2}\Big/\binom{N}{n}} = \frac{1}{\binom{N_1}{n_1}\binom{N_2}{n_2}} \frac{1}{\binom{N_1}{n_1}\binom{N_2}{n_2}}$.

RÉSUMÉ

- Pour toute paire de variables X, Y, $C(\overline{x}\,;\overline{y}) = (1-f)S_{xy}/n$, $f = n/N$. Ces paramètres sont estimés par leur équivalent échantillonnal, $\hat{V}(\overline{y}) = (1-f)s^2/n$ et $\hat{C}(\overline{x}\,;\overline{y}) = (1-f)s_{xy}/n$.

- $\hat{R} = \dfrac{\overline{y}}{\overline{x}}$ est un estimateur à peu près sans biais pour $R = \dfrac{\mu_y}{\mu_x}$. Son erreur quadratique moyenne est à peu près égale à $V(\hat{R}) \approx \dfrac{(1-f)}{n\mu_x^2}[S_y^2 + R^2S_x^2 - 2RS_{xy}] = \dfrac{(1-f)}{n\mu_x^2}\dfrac{\sum_{k=1}^{N}(y_k - Rx_k)^2}{N-1}$. L'erreur quadratique moyenne est estimée par $\hat{V}(\hat{R}) = \dfrac{1-f}{n\overline{x}^2}[s_y^2 + \hat{R}^2s_x^2 - 2\hat{R}s_{xy}] = \dfrac{1}{\overline{x}^2}[\hat{V}(\overline{y}) + \hat{R}^2\hat{V}(\overline{x}) - 2\hat{R}\hat{C}(\overline{x};\overline{y})] =$ $\dfrac{1-f}{n\overline{x}^2}\dfrac{\sum_{i\in w}(y_i - \hat{R}x_i)^2}{n-1}$, $f = \dfrac{n}{N}$.

- La taille de l'échantillon nécessaire pour estimer R est $n = \dfrac{n_0}{1 + n_0/N}$, où $n_0 = z_{\alpha/2}^2\dfrac{S_y^2 + R^2S_x^2 - 2RS_{xy}}{\mu_x^2E^2}$ pour une marge d'erreur absolue E; et $n_0 = z_{\alpha/2}^2\dfrac{S_y^2 + R^2S_x^2 - 2RS_{xy}}{\mu_y^2r^2}$ pour une marge d'erreur relative r.

Estimation de la moyenne et du total d'un domaine.

		Moyenne	Total
N_d connu	Estimateur	$\overline{y}_d$	$\overline{y}_dN_d$
	Variance estimée	$(1-n_d/N_d)\dfrac{s_d^2}{n_d^2}$	$N_d^2(1-n_d/N_d)\dfrac{s_d^2}{n_d^2}$ =
N_d Inconnu	Estimateur	$\overline{y}_d$	$N\overline{y}'$
	Variance estimée	$(1-n/N)\dfrac{s_d^2}{n_d^2}$	$N^2(1-n/N)(s'/n)^2$

- N_d et n_d sont la taille du domaine dans la population et dans l'échantillon, respectivement; $\overline{y}_d$ et s_d sont la moyenne et l'écart-type des données de l'échantillon appartenant au domaine; $\overline{y}'$ et s' sont la moyenne et l'écart-type d'une variable qui prend les valeurs de y dans le domaine et 0 à l'extérieur du domaine.

- *Données dichotomiques* Voir la section 5.5.

EXERCICES

5.1 Le tableau A02 présente des données sur une population de 1183 municipalités du Québec. Le Tableau 5.2 présente le nombre d'habitants en 2011 (pop11) et en 2006 (pop06), ainsi que le nombre de logements en 2011 (loge11) dans un échantillon de 45 municipalités.

a) Estimer le pourcentage d'accroissement du nombre d'habitants de 2006 à 2011.

b) Estimez le coefficient de variation de votre estimateur.

Tableau 5.2
Nombre d'habitants en 2011 (pop11) et en 2006 (pop06) et nombre de logements en 2011 (loge11 de 45 municipalités québécoises

pop11	pop06	loge11	pop11	pop06	loge11	pop11	pop06	loge11
410	377	177	860	859	394	3418	3150	1484
522	403	673	938	873	1003	3274	3260	1480
403	416	175	1109	1096	595	3198	3309	1470
484	458	307	1440	1353	1958	3275	3336	1616
485	458	228	1563	1430	754	3456	3409	1566
485	497	321	1606	1671	740	4061	4045	2065
542	508	242	1939	1853	961	4966	4237	2133
572	549	347	1971	1985	829	5692	5243	2406
536	596	320	2375	2132	925	5934	5687	3592
674	608	734	2624	2288	913	6933	6772	3323
743	673	354	2229	2351	1007	9485	8383	3965
736	682	282	2884	2637	1067	12171	10132	5015
771	692	349	2201	2713	1774	18200	15720	7406
786	718	707	2679	2800	1598	26816	26468	10337
674	727	460	3051	2895	1107	42491	33764	16290

Sommes et variances

	pop11	pop06	loge11
Sommes	191 662	174 213	85 449
Variances s^2	58 179 418	41 997 158	8 575 290

Covariances

C(pop11; pop06) = 49 149 799; C(pop11; loge11) = 22 200 704; C(pop06; loge11) = 18 785 924.

5.2 [*Suite du numéro* 5.1] Utiliser les données du Tableau 5.2 pour estimer le nombre moyen de personnes par logement en 2011. Estimer l'écart-type de l'estimateur.

5.3 D'une population de 2 500 comptes de dépenses soumis par des représentants d'une compagnie, on tire un échantillon de 45 comptes. Ceux-ci sont analysés de près et le montant réclamé est corrigé s'il y a lieu. Le Tableau 5.3 présente les résultats.

a) Estimer le rapport du montant corrigé sur le montant réclamé (μ_y/μ_x).

b) Estimer l'écart-type et le coefficient de variation de l'estimateur.

c) On considère cesser de vérifier les comptes et donc de payer les montants réclamés tels quels. On s'attend donc à ce qu'un certain pourcentage p des montants payés sera payé en trop. Estimer et déterminer un intervalle de confiance pour ce pourcentage.

d) Supposons que le montant moyen réclamé est connu: $\mu_x = 207{,}72$. On peut donc estimer le rapport μ_y/μ_x par $\frac{\bar{y}}{\mu_x}$. Déterminer l'écart-type de cet estimateur.

Tableau 5.3
Montants réclamés (x) et ajustés (y) de 45 comptes de dépenses

x	y	x	y	x	y
270,18	215,72	1865,83	1865,83	33,98	33,98
23,54	23,54	49,47	49,47	6,57	6,57
93,48	93,48	61,16	61,16	30,43	30,43
31,51	31,51	23,88	20,29	518,77	518,77
11,66	11,66	237,42	175,45	381,73	381,73
99,66	99,66	66,37	66,37	249,98	249,98
57,52	57,52	231	231	73,83	73,83
6,9	6,9	109,3	109,3	165,54	165,54
13,39	13,39	68,86	68,86	15,77	15,77
19,37	14,42	48,42	48,42	64,76	64,76
16,84	16,84	190,62	190,62	43,05	43,05
36,13	36,13	76,67	76,67	72,3	65,86
404,82	404,82	260,69	260,69	48,07	48,07
314,26	314,26	187,92	187,92	93,4	93,4
329,27	329,27	30,9	26	385,33	385,33

$\sum x_i = 7420{,}55$; $\sum y_i = 7284{,}24$; $s_x = 289{,}8779$; $s_y = 289{,}5249$; $s_{xy} = 83853{,}11$

5.4 Le Tableau 5.4 présente les données d'un échantillon de taille 30 tiré d'une population de 181 pays (tableau A08).

a) Estimer le rapport R de l'espérance de vie (à 60 ans) des femmes sur l'espérance de vie (à 60 ans) des hommes. Estimer le coefficient de variation de l'estimateur.

b) On peut s'interroger sur la pertinence du quotient R estimé en a). Certains pourraient préconiser plutôt la moyenne de quotients (plutôt que le quotient des moyennes R), c'est-à-dire, la moyenne μ_z de la variable Z dont les valeurs sont $z_k = \frac{y_k}{x_k}$. Estimer μ_z ainsi que l'écart-type de l'estimateur à partir des données suivantes: $\sum z_i = 35{,}10674$; $\sum z_i^2 = 41{,}33144$.

[*Remarque* : R et μ_z peuvent tous deux être contestés, dépendant de l'usage qu'on en fait. Si l'intention est de comparer la longévité des femmes à celle des hommes dans le monde, les deux indices sont inadéquats dans la mesure où ils ne pondèrent pas les moyennes ou les taux selon l'importance du pays.]

Tableau 5.4
Espérance de vie à 60 ans des hommes (*x*) et des femmes (*y*) dans un échantillon de 30 pays tirés d'un ensemble de 181 pays

x	y	x	y	x	y	x	y	x	y	x	y
18	22	22	26	13	17	18	21	21	25	15	17
18	23	21	25	16	18	15	17	15	16	19	21
14	21	20	22	20	22	22	25	17	20	22	25
21	25	22	24	14	19	17	18	22	26	17	20
18	19	21	23	17	20	20	23	23	27	20	23

$\sum x_i = 558$; $\sum y_i = 650$; $\sum x_i^2 = 10618$; $\sum y_i^2 = 14366$; $\sum x_i y_i = 12325$

5.5 Même contexte qu'au numéro 3.3. Voici la distribution conjointe de (Z ; Y)

		y						
		0	1	2	3	4	5	Total
	0	38	6	2	0	0	0	46
	1	2	18	2	3	0	0	25
z	2	0	0	15	1	1	0	17
	3	0	0	0	7	1	1	9
	4	0	0	0	0	2	1	3
	Total	40	24	19	11	4	2	100

Estimer le rapport du nombre moyen d'absences cette année (μ_y) sur la moyenne l'année dernière (μ_z) et estimer l'écart-type de l'estimateur.

5.6 Dans l'exemple 5.1, supposons qu'on réussisse à connaître la superficie moyenne de toutes les fermes de la population, soit μ_x = 52,42242. Alors $\hat{R}^* = \frac{\overline{y}}{\mu_x}$ est un estimateur sans biais du quotient $R = \frac{\mu_y}{\mu_x}$. Évaluer $\hat{R}^*$ et estimer son écart-type. Comparer avec vos résultats à ceux de l'Exemple 5.1. $\hat{R}^*$ semble moins bon que $\hat{R}$. Était-ce prévisible?

5.7 Dans une étude sur la santé, on prélève un échantillon de 40 ménages afin d'estimer la proportion p des gens qui font de l'exercice régulièrement. Les résultats sont présentés dans le Tableau 5.5. La population est très grande.

a) Estimer p en considérant que l'échantillon est constitué de n = 40 ménages tirés dans une population de ménages (dans ce cas, p est en fait un *quotient*.) Estimer l'écart-type de l'estimateur.

b) Estimer p en supposant que l'échantillon est constitué de n = 160 *personnes* tirées au hasard dans l'ensemble des personnes (dans ce cas, p est une *proportion*.) Estimez l'écart-type de l'estimateur.

c) Pourquoi l'approche en b) est-elle erronée? Justifier votre réponse. [Examiner les valeurs de $p_i = y_i/x_i$.]

Tableau 5.5
Enquête sur la santé

x	y	x	y	x	y	x	y
4	0	6	6	5	1	3	0
4	4	2	0	6	0	3	3
3	3	3	3	4	3	5	1
4	0	2	2	2	2	6	0
3	0	4	4	3	1	8	0
5	4	4	3	3	0	1	0
5	0	4	1	4	4	3	2
3	3	4	0	5	5	3	3
3	0	4	0	6	1	4	1
4	2	4	4	7	0	4	2

x : Nombre de personnes dans le ménage
y : Nombre de personnes qui font de l'exercice régulièrement
Calculs
$\overline{x} = 4$; $\overline{y} = 1,7$
$s_x = \sqrt{2}$; $s_y = \sqrt{2,9333}$
$s_{xy} = -0,025641$

5.8 D'une population de 880 comptes à payer, on prélève un échantillon de 30 comptes. On s'intéresse particulièrement au domaine U_d des comptes qui sont dus depuis 6 mois ou plus. Dans l'échantillon, on trouve 17 unités appartenant à U_d. Les montants de ces 17 comptes sont les suivants :

141,22	34,97	12,50	99,95	49,44	109,7	7,04	229,95	50,55
55,41	68,86	110,75	27,88	15,04	4,01	25,25	18,97	

a) Estimer le montant total τ_d de tous les comptes de U_d, et estimer l'écart-type de l'estimateur.

b) Supposons qu'on ait déterminé auparavant que N_d, la taille de U_d, est égale à 460. Estimer τ_d en utilisant cette information et estimer l'écart-type de l'estimateur.

c) [*Exige un logiciel*] Supposons que l'écart-type de Y dans U_d est $S_d = 65$. Déterminer la variance conditionnelle de $\bar{y}_d$ étant donné que $n_d = 17$, puis déterminer la variance inconditionnelle (mais conditionnelle quand même à $n_d \neq 0$]. On suppose encore que $N_d = 460$ et $n = 30$.

Tableau 5.6
Nombre d'adultes et d'enfants
selon le type d'habitation

Appartements

Adultes	5	2	3	3	1	4	1	5	1	2	2	3	3	2	4	1	3	1
Enfants	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	2	2

Sommes: adultes:46; enfants:6;

Maisons unifamiliales

Adultes	5	2	3	3	1	4	1	5	1	2	2	3	3	2	4	1	3
Enfants	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	2
Adultes	6	2	1	1	1	3	1	6	2	2	4	5	2	3	1	2	1
Enfants	1	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2	2
Adultes	1	4	2	2	2	1	3	4	1	4	1	4	4	2	2	2	
Enfants	2	0	0	0	0	0	0	1	2	2	2	2	2	3	4	4	

	Appartements			Maisons		
	Somme	s^2	Cov.	Somme	s^2	Cov.
Adultes	46	1,7908497	-0,137255	127	2,008571	-0,08327
Enfants	6	0,4705882		52	1,141225	

5.9 D'une population de 1 500 logements abritant 5 370 personnes, on tire un *éas* de 68 logements. Le Tableau 5.6 présente les données. Les variables observées sont :

- Adultes Le nombre d'adultes vivant dans le logement ;
- Enfants Le nombre d'enfants vivant dans le logement ;
- Habitation Le type d'habitation : « Maison unifamiliale » et « Appartement ».

[Estimer l'écart-type des estimateurs calculés.]

a) Estimer la proportion d'adultes dans la population.

b) En a), le paramètre à estimer est un quotient. Supposons qu'on l'avait traité comme une proportion p et considéré les données comme un *éas* de 231 personnes tiré dans une population de personnes. Estimer l'écart-type de $\hat{p}$ sous ce modèle. Pourquoi cette approche est-elle erronée?

c) Estimer la proportion d'adultes dans les maisons unifamiliales.

d) Estimer la proportion de maisons unifamiliales où il n'y a pas d'enfants.

e) Dans le domaine des maisons unifamiliales, estimer la proportion d'habitations qui n'ont pas d'enfants.

5.10 [*Ce numéro est un exercice récapitulatif; il peut comprendre des problèmes relevant de chapitres antérieurs*] Le Tableau 5.7 présente les données d'un échantillon de taille 50 tiré d'une population de 8 654 factures de restaurant (tableau A03) selon qu'elles ont été émises un jour de semaine ou le week-end. X est la valeur du vin consommé, Y le montant total de la facture. Dans chaque cas, estimer l'écart-type de l'estimateur.

a) Estimer la proportion des recettes réalisées en fin de semaine (week-end).

b) Estimer le montant total des recettes provenant du service de vin.

c) Estimer le montant total des recettes provenant du service de vin en utilisant le fait que 6 833 factures comprenaient du vin.

d) Estimer la proportion des recettes provenant du service de vin.

e) Soit U_d l'ensemble des repas ayant inclus du vin. Estimer la proportion des recettes provenant du service de vin dans le domaine U_d.

f) Estimer la proportion des repas servis sans vin.

g) Estimer le nombre de repas servis sans vin.

h) Estimer le nombre de repas servis sans vin le week-end.

i) Estimer le nombre de repas servis sans vin le week-end en utilisant le fait que 3 234 repas ont été servis le week-end.

j) Estimer le montant total des recettes provenant du vin servi le week-end (sans utiliser l'information ajoutée en i).

k) Estimer le montant total des recettes provenant du vin servi le week-end en utilisant le fait que 3 234 repas ont été servis un week-end;

l) Estimer le montant total des recettes provenant du vin servi le week-end en utilisant le fait que 2 758 repas comprenant du vin ont été servis un week-end.

Statistiques descriptives

	En semaine				En week-end			
Repas	*n*	*N*	$\sum y_i$	$\sum y_i^2$	*n*	*N*	$\sum y_i$	$\sum y_i^2$
Sans vin	10	1345	488,75	26195	4	476	238,05	18281
Avec vin	26	4075	2366,70	239947	10	2758	953,35	103606

Tableau 5.7
Échantillon de 50 factures de restaurant

Factures émises un jour de semaine

x	*y*	*x*	*y*	*x*	*y*	*x*	*y*
0	49,45	0,00	35,65	9,20	64,40	18,40	100,05
0	65,55	9,20	47,15	9,20	86,25	18,40	110,40
0	64,40	9,20	86,25	13,80	39,10	18,40	124,20
0	49,45	9,20	62,10	13,80	82,80	23,00	106,95
0	59,80	9,20	151,80	13,80	106,95	24,15	133,40
0	69,00	9,20	66,70	13,80	70,15	27,60	94,30
0	42,55	9,20	131,10	13,80	49,45	27,60	87,40
0	31,05	9,20	96,60	14,95	58,65	32,20	142,60
0	21,85	9,20	52,90	14,95	90,85	32,20	124,20

$\sum x_i = 412{,}85;\ \sum y_i = 2855{,}45;\ \sum x_i^2 = 7970{,}707;\ \sum y_i^2 = 266\ 141{,}2;\ \sum x_i y_i = 40\ 264{,}83.$

Factures émises le week-end

Vin	total	Vin	total	Vin	total	Vin	total
0	100,05	41,4	155,25	13,8	63,25	18,4	54,05
0	81,65	32,2	123,05	18,4	124,2	13,8	56,35
0	31,05	9,2	52,9	18,4	133,4	13,8	92
0	25,30	32,2	98,9				

$\sum x_i = 211{,}6;\ \sum y_i = 1191{,}4;\ \sum x_i^2 = 5459{,}28;\ \sum y_i^2 = 121\ 886{,}89;\ \sum x_i y_i = 22\ 715{,}26.$

5.11 Au numéro 5.3, on estime l'écart-type de $\hat{R}$ à 0,012. Estimer la taille de l'échantillon qu'il aurait fallu tirer pour estimer R avec une marge d'erreur absolue (à 95 %) de 0,015.

5.12 Considérons un *éas* de taille 3 tiré d'une population de N = 8 unités dont les valeurs de deux variables, x et y, sont les suivantes :

y	8	10	67	44	66	56	89	99
x	3	6	24	27	30	36	51	57

À l'aide d'un logiciel (tel Excel) énumérer tous les échantillons possibles de taille 3 et répondre aux questions suivantes.

a) Vérifiez que l'estimateur $\hat{R}$ est biaisé, comme on le sait, et exprimez une opinion sur l'importance du biais dans ce cas.

b) Calculez la variance (la *vraie* variance) de $\hat{R}$, et comparez avec la variance telle que calculée par la formule approximative $V(\hat{R})$ (qui ne donne qu'une approximation et ne donne pas la vraie variance).

c) Quelle est la probabilité de se tromper de plus de 20% dans l'estimation du quotient ?

d) Comparez $V(\hat{R})$ à la quantité obtenue par la formule usuelle, $F = \frac{1-n/N}{n\mu_x^2}(s_y^2 + R^2 s_x^2 - 2Rs_{xy})$. Est-ce que la formule habituelle $\hat{V}(\hat{R})$ serait un estimateur raisonnable de F plutôt que de la vraie variance $V(\hat{R})$? Dites ce que vous en pensez à la lumière de vos calculs.

e) D'après ces données, est-ce que s_{xy} est un estimateur sans biais de S_{xy} ?

5.13 L'échantillon présenté au Tableau 5.1 est tiré d'une population de 1 579 fermes qui produisent de l'orge. On classe les fermes de la population en deux catégories : la classe 1 comprend les 756 fermes qui cultivent 50 hectares ou moins d'orge; la classe 2 comprend les 823 autres.

a) Estimer R_1, la production moyenne par hectare des fermes de la classe 1 ainsi que l'écart-type de l'estimateur (considérer les données appartenant à la classe 1 comme un échantillon de 29 fermes tirés dans une population de taille 75).

b) Répéter l'exercice a) pour les fermes de la classe 2, puis déterminer un intervalle de confiance pour la différence de quotients $R_2 - R_1$. Peut-on conclure avec 95 % de confiance qu'il y a une différence de moyennes par hectare entre les deux classes de fermes?

c) En a), R_1 aurait pu être défini comme le quotient $\mu_{y'}/\mu_{x'}$, où la variable y' est définie par $y_k' = \begin{cases} y_k & \text{si } k \text{ appartient à la classe 1} \\ 0 & \text{sinon} \end{cases}$, et x' définie de façon analogue. Estimer l'écart-type de l'estimateur $\hat{R}_1' = \frac{\sum_{i\in\omega} y_i'}{\sum_{i\in\omega} x_i'}$.

Quelques données:

Classe 1 :

$n_1 = 29; \sum x_i = 997{,}5; \sum y_i = 3022; \sum x_i^2 = 37\,512{,}17 ; \sum y_i^2 = 345\,996 ; \sum x_i y_i = 113\,693{,}5$.

Classe 2 :

$n_1 = 21; \sum x_i = 1486{,}4; \sum y_i = 4512; \sum x_i^2 = 109052{,}6$;

$\sum y_i^2 = 1\,003\,386$; $\sum x_i y_i = 330\,463{,}3$.

5.14 On doit prélever un *éas* d'une population de 1 160 municipalités afin d'estimer certains paramètres, tous fonctions des variables suivantes: m le nombre de familles; h le nombre d'habitants; f le nombre d'unilingues francophones; p le nombre de familles avec un enfant ou plus; e le nombre de familles avec deux enfants. Les estimations suivantes (moyennes. écarts-types et covariances) ont été obtenues à partir de données historiques: μ_m= 1898; μ_h = 6733; S_m = 13818; S_h = 53542; S_{mh}= 739 032 485; μ_f = 5260; S_f = 31539; S_{hf} = 1 632 769 884; μ_p = 1097; μ_e = 420; S_p = 8468; S_e = 3047; S_{pe} = 25 766 503

Déterminer la taille de l'échantillon nécessaire afin d'estimer les paramètres suivants avec les marges d'erreur (à 95 %) absolues et relatives indiquées.

a) La taille moyenne d'une famille. Marges d'erreur: absolue: 0,5; relative: 10 %.

b) La proportion d'unilingues francophones. Marges d'erreur: absolue: 0,5; relative: 10 %.

c) La proportion de ménages avec 2 enfants parmi les ménages avec au moins un enfant. Marges d'erreur: absolue: 0,05; relative: 4 %.

5.15 Au numéro 5.14, soit a le nombre d'unilingues anglophones et h le nombre d'habitants. Supposons qu'on connaît les valeurs des paramètres suivants: μ_a = 514,1034; S_a = 6294,618; S_{ah} = 315 498 400, S_{af}= 163 341 731. Soit R_f et R_a la proportion d'unilingues francophones et anglophones, respectivement.

a) Déterminer l'écart-type et le coefficient de variation de l'estimateur de la différence $R_f - R_a$ dans un échantillon de taille 300. Attention, les estimateurs $\hat{R}_f$ et $\hat{R}_a$ ne sont pas indépendants. Le Théorème 1.8 peut être utile ici.

b) Déterminer la taille de l'échantillon qu'il faudrait tirer pour que le coefficient de variation soit de 10 %.

c) Montrer que si $\hat{R}_f$ et $\hat{R}_a$ étaient indépendantes, le coefficient de variation de l'estimateur $\hat{R}_f - \hat{R}_a$ aurait été évalué (erronément) à 14,9 %.

5.16 Considérons une population sur laquelle sont définies deux variables x et y, dont les valeurs sont $(x_1 ; y_1),\ldots, (x_N ; y_N)$. On sait que l'estimateur $\hat{R}$ est biaisé lorsque le

tirage est aléatoire simple. Si les valeurs de X sont connues pour la population entière, il existe un mode de tirage, la méthode de Lahiri (Lahiri, 1951), pour laquelle $\hat{R}$ est sans biais. Elle consiste à tirer les échantillons avec probabilités proportionnelles aux $\bar{x}$: $p(\omega) = k\bar{x}(\omega)$.

a) Montrer que $k = \dfrac{1}{\mu_x \binom{N}{n}}$.

b) Montrer que $\hat{R}$ est sans biais: $E_L(\hat{R}) = R$, où E_L désigne l'espérance sous le tirage de Lahiri.

c) Montrer que $V_L(\hat{R}) = E_a\left(\bar{y}^2/\bar{x}\mu_x\right) - R^2$, où V_L désigne la variance sous un tirage de Lahiri et E_a désigne l'espérance sous un échantillonnage aléatoire simple.

d) Montrer qu'une façon d'effectuer un tirage de Lahiri est la suivante: on tire une première unité avec probabilités proportionnelles aux x_i; puis on tire un *éas* aléatoire simple de n -1 unités parmi les N - 1 qui restent (Midzuno, 1951).

5.17 [*Estimation d'un produit*] Montrer que $\hat{\pi} = \bar{x}\bar{y} - (1-f)\dfrac{s_{xy}}{n}$, $f = n/N$, est un estimateur sans biais d'un produit $\pi = \mu_x\mu_y$.

5.18 Supposons que μ_x est connue. Montrer que $\hat{R}^* = \dfrac{1}{n}\sum_{i\in\omega} r_i + \dfrac{(N-1)s_{rx}}{N\mu_x}$ est un estimateur sans biais de R, où $r_i = y_i/x_i$.

5.19 Considérer un tirage selon la méthode de Midzuno (numéro 5.16-d).

a) Montrer que l'estimateur $\hat{R}^* = \dfrac{\bar{y}}{x_\ell}$, où x_ℓ est la valeur de x correspondant à la première unité tirée, est sans biais pour R. [$\hat{R}$ est défini sur l'ensemble ω des échantillons distincts ω alors que $\hat{R}^*$ est défini sur l'ensemble Ω^* des échantillons ω^* qui distinguent les échantillons selon le résultat du premier tirage. Chaque ω est compris dans n échantillons ω^*.

b) Soit $p(i \mid \omega)$ la probabilité conditionnelle que i soit la première unité sélectionnée étant donné un échantillon $\omega\in\Omega$. Montrer que $p(i \mid \omega) = \dfrac{x_i}{n\bar{x}}$.

c) Montrer que $E(\hat{R}^* \mid \omega) = \hat{R}$.

d) Utiliser le Théorème 1.12 pour démontrer que $V(\hat{R}^*) \geq V(\hat{R})$.

e) Montrer que $V(\hat{R}^* \mid \omega) = \dfrac{\bar{y}^2}{n\bar{x}}\sum_{i\in\omega}\dfrac{1}{x_i} - \hat{R}^2$.

5.20 Considérer la population présentée au numéro 5.12.

a) Vérifiez numériquement que dans un tirage de Lahiri, $\hat{R}$ est sans biais.

b) Vérifiez que $\hat{R}$ est une meilleure estimation du quotient dans un tirage de Lahiri que dans un tirage aléatoire simple en comparant la variance de l'estimateur sous un tirage de Lahiri à l'erreur quadratique moyenne de l'estimateur sous un tirage aléatoire simple.

6 Estimation par la différence, le quotient et la régression

Jusqu'ici, le seul estimateur de μ_y proposé a été la moyenne échantillonnale $\bar{y}$; et le seul proposé pour le total τ a été $\hat{\tau} = N\bar{y}$. Mais s'ils sont souvent utilisés, c'est en partie parce que l'information préalable qu'ils exigent — la *base de sondage* — est minimale : une liste des unités de la population suffit. Si on n'a rien de plus, on peut difficilement faire mieux que d'estimer μ par $\bar{y}$.

Mais une population qu'on projette d'échantillonner n'est pas toujours totalement inconnue. La base de sondage peut contenir, en plus d'une liste, des éléments d'information qui peuvent aider à améliorer la précision d'un estimateur. Ces informations prennent souvent la forme d'une deuxième variable X, appelée *variable auxiliaire*, dont les valeurs sont connues *pour la population entière*.

Par exemple :

- Dans les applications en vérification, la valeur corrigée Y d'un compte n'est connue qu'après vérification d'un échantillon de comptes; mais X, la valeur aux livres, est généralement connue pour la population entière.
- Pour estimer le revenu moyen μ_y des ménages d'une ville, il peut être utile de connaître le montant X de leur taxe foncière, une information dont la municipalité dispose pour la population entière.
- Pour estimer le nombre moyen μ_y de naissances dans les villes du Québec en 2015, il est utile de connaître le nombre X de naissances lors du dernier recensement.
- Pour estimer la production moyenne d'une céréale dans une population de fermes, un estimateur qui tient compte de la superficie cultivée, si celle-ci est connue pour l'ensemble de la population, est sûrement meilleur.

Dans ce chapitre, nous introduisons trois nouveaux estimateurs qui mettent à profit l'information contenue dans une variable auxiliaire X. Ce sont l'*estimateur par le quotient*, l'*estimateur par la différence*, et *l'estimateur par la régression*. Le prochain exemple illustre deux des trois estimateurs mentionnés ci-dessus et les compare à $\bar{y}$. L'estimateur par la régression, moins évident, sera traité plus loin.

Exemple 6.1 *Utilisation d'une variable auxiliaire*

Le Tableau 5.2 présente des données sur un *éas* de taille 45 tiré d'un ensemble de 1 183 municipalités du Québec (tableau A02). Les variables observées sont :

X : le nombre d'habitants en 2006

Y : le nombre d'habitants en 2011

L'objectif est d'estimer le nombre moyen d'habitants μ par ville en 2011. Nous pouvons toujours estimer μ par la moyenne échantillonnale $\bar{y}$. Mais supposons que les valeurs de X sont connues non seulement pour l'échantillon, mais aussi — par recensement — pour la population entière. Dans ce cas, d'autres possibilités se présentent. Voici trois possibilités.

a) *Estimation par la moyenne* La moyenne de l'échantillon, $\bar{y}$ = 4 259,156, est la seule estimation raisonnable si on ne dispose d'aucune autre information pertinente sur la population. Elle sera appelée *estimation par la moyenne*.

b) *Estimation par la différence* Si on connaît la moyenne μ_x de la variable X = nombre d'habitants en 2006, soit μ_x = 6 378,709, on peut raisonner comme ceci : d'après l'échantillon, le nombre d'habitants s'est accru de 387,7556 personnes en moyenne [$\bar{y} - \bar{x}$ = 4259,156 – 3871,4 = 387,7556]. Puisqu'on *sait* qu'il était de 6 378,709 en 2006, on l'estime maintenant à 6 378,709 + 387,7556 = 6 766,465. Ainsi l'échantillon a été utilisé ici pour estimer *l'accroissement* du nombre d'habitants, un accroissement qu'on ajoute à la moyenne connue de 2006. C'est l'*estimateur par la différence*.

c) *Estimation par le quotient* Une deuxième approche consiste à estimer non pas l'accroissement absolu mais plutôt l'accroissement *relatif*. D'après l'échantillon, le rapport du nombre d'habitants en 2011 sur le nombre d'habitants en 2006 est $\hat{R} = \frac{\bar{y}}{\bar{x}}$ = 1,100159. Puisque nous *savons* que le nombre moyen d'habitants était de 6 378,709 en 2006, il suffira de majorer ce nombre par le facteur 1,100159, ce qui donne 6 378,709(1,100159) = 7 017,594. C'est *l'estimateur par le quotient*.

Les trois méthodes illustrées dans le dernier exemple donnent trois estimations différentes :

- Estimation par la moyenne : 4 259,156 ;
- Estimation par la différence : 6 766,465 ;
- Estimation par le quotient : 7 017,594.

Étant donné qu'il n'est pas question d'utiliser trois estimateurs pour un même paramètre, il faudra faire un choix. Un examen des propriétés générales de ces estimateurs permettra d'entrevoir lequel des estimateurs est susceptible de donner les meilleurs résultats dans un contexte donné. On sait déjà, dans le cas de cet exemple, que les deux derniers estimateurs sont préférables au premier : il n'est pas nécessaire de calculer quoi que ce soit pour savoir que la corrélation entre les variables X et Y doit être très forte, et donc qu'on a intérêt à l'exploiter.

Un domaine où l'apport d'une variable auxiliaire peut se révéler appréciable est la vérification comptable. Une situation typique: la population est un ensemble, souvent très important, de transactions ou de comptes; x_k est la valeur nominale (ou valeur aux livres); et y_k est la valeur corrigée, suite à une vérification. Les valeurs x_k et y_k sont donc *égales* pour une grande partie de la population, et la corrélation entre X et Y est probablement forte. Et on découvre en effet que l'estimation par le quotient (ou par la différence) est souvent beaucoup plus précise que l'estimation par la moyenne.

D'autres cas sont moins évidents. Supposons que Y est la production de maïs dans les fermes d'un territoire et X est le nombre d'hectares cultivés dans les fermes. Si les valeurs de X sont répertoriées pour toute la population, on peut s'en servir pour estimer μ_x. Devrait-on le faire ? Peut-être bien, mais ce n'est pas évident. Il est vrai que la corrélation entre X et Y est positive : les grandes fermes produisent plus de maïs que les petites. Mais il y a des fermes, des grandes et des petites, qui ne cultivent pas de maïs du tout, et ceci affaiblit considérablement la corrélation. Jusqu'à quel point? Parfois on peut le deviner, parfois pas, dépendant de ce qu'on sait de la population. Par ailleurs, s'il est possible d'identifier a priori les fermes où l'on cultive le maïs, de sorte qu'on peut restreindre la population à celles-là, alors dans ce cas la corrélation est probablement plus forte. L'est-elle assez? Cela dépend. Si les fermes qui cultivent le maïs ne cultivent *que* le maïs, alors oui, X est probablement une bonne variable à exploiter. Autrement non. C'est là qu'une connaissance du contexte est indispensable : ceux qui œuvrent dans le domaine, et qui collectent régulièrement des données sur la production des fermes, finissent par développer un corpus d'information qui les aide dans leurs choix.

S'il n'est généralement par trop difficile de décider d'utiliser ou non une variable auxiliaires, le choix entre les estimateurs les utilisant n'est pas toujours facile à faire. Il se fera sur le terrain, sur la base de ce qu'on sait du contexte et des propriétés des estimateurs.

Notation Les observations de la population sont désignées par

$$x_1, x_2, \ldots, x_N$$

$$y_1, y_2. \ldots, y_N$$

Note : Nous supposerons dans ce chapitre que la variable auxiliaire X ne prend que des valeurs positives ou nulles.

Rappelons la définition du coefficient de corrélation, que nous désignerons par ρ pour la population et par r pour l'échantillon :

$$\rho = \frac{S_{xy}}{S_x S_y} \quad , \quad r = \frac{s_{xy}}{s_x s_y} \, . \tag{6.1}$$

C'est principalement le paramètre ρ qui détermine si la variable auxiliaire vaut la peine d'être utilisée dans l'estimation de μ_y (ou de τ_y).

6.1 ESTIMATEUR PAR LA DIFFÉRENCE

L'estimateur par la différence est

(6.2) $$\hat{\mu}_{ydiff} = \mu_x + (\overline{y} - \overline{x})$$

Propriétés de $\hat{\mu}_{ydiff}$

Nous posons les trois questions habituelles concernant un estimateur: $\hat{\mu}_{ydiff}$ est-il sans biais ? Quelle est son écart-type ? Comment estimer son écart-type ? Le théorème suivant donne les réponses.

THÉORÈME 6.1 **ESTIMATEUR PAR LA DIFFÉRENCE**

1. L'estimateur par la différence est sans biais pour la moyenne μ_y:

$$E(\hat{\mu}_{diff}) = \mu_y.$$

2. L'écart-type de $\hat{\mu}_{ydiff}$ est donné par

(6.3) $$\sqrt{V(\hat{\mu}_{ydiff})} = \sqrt{1-f}\,\frac{\sqrt{S_y^2 + S_x^2 - 2S_{xy}}}{\sqrt{n}} \quad \text{où } f = \frac{n}{N}.$$

3. L'estimateur de $\sqrt{V(\hat{\mu}_{ydiff})}$ est donné par

(6.4) $$\sqrt{\hat{V}(\hat{\mu}_{ydiff})} = \sqrt{1-f}\,\frac{\sqrt{s_y^2 + s_x^2 - 2s_{xy}}}{\sqrt{n}}.$$

4. L'estimateur $\sqrt{\hat{V}(\hat{\mu}_{ydiff})}$ est sans biais pour $V(\hat{\mu}_{ydiff})$.

5. Un intervalle de confiance approximatif est donné par

(6.5) $$\hat{\mu}_{yd} - z_{\alpha/2}\sqrt{\hat{V}(\hat{\mu}_{ydiff})} \le \mu_d \le \hat{\mu}_{yd} + z_{\alpha/2}\sqrt{\hat{V}(\hat{\mu}_{ydiff})}$$

Exemple 6.2 *Estimation par la différence*

Considérons encore l'échantillon de 45 municipalités québécoises au Tableau 5.2 tiré de la population de taille 1183 présentée au tableau A02.

a) Estimer par la différence le nombre moyen d'habitants μ_y en 2011, en utilisant le nombre d'habitants en 2006 comme variable auxiliaire, sachant que le nombre moyen d'habitants en 2006 était 6 378,709.

b) Déterminer un intervalle de confiance pour μ_y basé sur l'estimation par la différence.

Solution

Quelques calculs de base :

$\bar{y}$ = 4 259,156 ; $\bar{x}$ = 3 871,4 ; s_y = 7 627,543 ; s_x = 6 480,521; et s_{xy} = 49 149 799.

a) $\hat{\mu}_{ydiff} = \mu_x + (\bar{y} - \bar{x})$ = 6 378,709 + (4 259,156 - 3871,4) = 6 378,709 + 387,7556 = 6 766,465.

b) Estimons maintenant l'écart-type de $\hat{\mu}_{ydiff}$:

$$\hat{V}(\hat{\mu}_{ydiff}) = \sqrt{1-f}\,\frac{\sqrt{s_y^2 + s_x^2 - 2s_{xy}}}{\sqrt{n}} = \sqrt{1-45/1183}\,\frac{\sqrt{(7627,543)^2 + (6480,521)^2 - 2(49149799)}}{\sqrt{45}} = 200,3098$$

L'intervalle de confiance est donc 6 766,465 ± $z_{\alpha/2}$(200,3098). À un niveau de 95 % cela donne 6 373,86 ≤ μ_y ≤ 7 159,07.

À titre de comparaison, considérons l'estimateur $\bar{y}$. Son écart-type est estimé à

$$\sqrt{\hat{V}(\bar{y})} = \sqrt{1-f}\,\frac{s_y}{\sqrt{n}} = \sqrt{1-\frac{45}{1183}}\,\frac{7627,543}{\sqrt{45}} = 1\ 115,211.$$

plus de 5 fois l'écart-type de $\hat{\mu}_{ydiff}$.

Remarque *$\hat{\mu}_{ydiff}$ est rarement à ce point précis*

Il faut admettre que la supériorité spectaculaire de l'estimateur pas la différence dans le dernier exemple est un cas exceptionnel, dû à la très forte corrélation entre *X* et *Y*. L'estimateur par la différence est souvent plus efficace que la moyenne, mais rarement à ce point-là.

DISCUSSION SUR L'ESTIMATEUR PAR LA DIFFÉRENCE

Il y a deux façons de justifier intuitivement cet estimateur.

Première justification

On peut décomposer μ_y en deux parties, l'une connue, l'autre pas :

$$\mu_y = \mu_x + (\mu_y - \mu_x) .$$

La première composante, μ_x, est connue, et n'a pas à être estimée. Seule la deuxième composante, μ_y - μ_x, doit être estimée. On l'estime, naturellement, par la différence entre les deux moyennes échantillonnales, $\bar{y} - \bar{x}$. Donc les propriétés de l'estimateur par la différence sont essentiellement celles de $\bar{y} - \bar{x}$. La variance de cette différence est $V(\bar{x} - \bar{y}) = V(\bar{x}) + V(\bar{y}) - 2C(\bar{x};\bar{y})\ V(\bar{x}) + V(\bar{y}) - 2\rho\sqrt{V(\bar{x})}\sqrt{V(\bar{y})}$, où ρ est le coefficient de corrélation entre *X* et *Y*. Cette variance est petite lorsque ρ est positif et fort.

Remarque *Un défaut de* $\hat{\mu}_{ydiff}$

Dans la meilleure des hypothèses (ρ = 1), $V(\bar{x}-\bar{y}) = [\sqrt{V(\bar{x})}-\sqrt{V(\bar{y})}]^2$. Ceci révèle un défaut de l'estimateur, car si ρ = 1, la valeur de Y est entièrement prévisible et on devrait s'attendre à ce que la variance de l'estimateur soit nulle. Ce n'est pas le cas, à moins que $S_x = S_y$.

Deuxième justification

L'estimateur naturel de μ_y est $\bar{y}$ et a priori c'est l'estimateur privilégié. L'estimateur par la différence, écrit comme

$$\bar{y} + (\mu_x - \bar{x}),$$

ajuste l'estimateur $\bar{y}$ par l'ajout du terme $(\mu_x - \bar{x})$. Grâce à l'information disponible sur la variable *X,* on peut deviner si, en l'occurrence, $\bar{y}$ surestime ou sous-estime μ_y. Reprenons l'exemple. Nous *savons* que le nombre moyen d'habitants en 2006 était de 6 379 — ce n'est pas une estimation, c'est la vraie moyenne. Or *l'échantillon* a donné une moyenne de 3 871 en 2006. On en déduit que l'échantillon comprend des villes plus petites que la moyenne, et donc que $\bar{y}$ = 4 259 (nombre moyen d'habitants en 2011) est probablement une sous-estimation de μ_y. Il serait bon alors de la majorer. C'est ce qu'on fait en ajoutant l'écart de 6 379 - 3 871 = 2 508 habitants entre les moyennes de l'échantillon et de la population en 2006. L'estimation est donc $\bar{y}$ + 2 508 = 6 767.

Mais s'il est raisonnable de vouloir majorer $\bar{y}$, il n'est pas évident que le simple ajout de la différence $\mu_x - \bar{x}$ soit la meilleure façon de le faire. Dans le cas présent, cette approche est défendable, étant donné qu'elle repose sur une hypothèse raisonnable : si l'échantillon sous-estime μ_x par 2 508 habitants, il est probable qu'il sous-estime μ_y par à peu près ce nombre.

6.2 ESTIMATEUR PAR LE QUOTIENT

L'estimateur par le quotient est défini par

(6.6) $$\hat{\mu}_{yq} = \hat{R}\,\mu_x$$

où $\hat{R} = \dfrac{\sum_{i\in\omega} y_i}{\sum_{i\in\omega} x_i} = \dfrac{\bar{y}}{\bar{x}}$. Les propriétés de $\hat{\mu}_{yq}$ découlent trivialement de celles de $\hat{R}$ comme estimateur du quotient $R = \dfrac{\sum_{i=1}^{N} y_i}{\sum_{i=1}^{N} x_i} = \dfrac{\mu_y}{\mu_x}$, développées au Chapitre 5. Puisque $\hat{R}$ n'est pas sans biais, $\hat{\mu}_{yq}$ ne l'est pas non plus, mais le biais est négligeable lorsque n est grand. Le théorème suivant résume ses propriétés.

THÉORÈME 6.2 **ESTIMATEUR PAR LE QUOTIENT**

1. L'estimateur $\hat{\mu}_{yq}$ est à peu près sans biais

$$E(\hat{\mu}_{yq}) = \mu_x E(\hat{R}) \approx \mu_x R = \mu_y.$$

2. L'erreur quadratique moyenne de $\hat{\mu}_{yq}$ (déduite du Théorème 5.4), de $\hat{\mu}_{yq}$ est à peu près égale à

(6.7) $$V(\hat{\mu}_{yq}) = \mu_x^2 V(\hat{R}) \approx (1-f)\frac{S_y^2 + R^2 S_x^2 - 2RS_{xy}}{n} = V(\bar{y}) + R^2 V(\bar{x}) - 2RC(\bar{x};\bar{y}) \ .$$

3. Un estimateur de $V(\hat{\mu}_{yq})$ découle de (5.4):

(6.8) $$\hat{V}(\hat{\mu}_{yq}) = \mu_x^2 \hat{V}(\hat{R}) \approx \mu_x^2 \frac{[\hat{V}(\bar{y}) + \hat{R}^2 \hat{V}(\bar{x}) - 2\hat{R}\hat{C}(\bar{x};\bar{y})]}{\bar{x}^2} = \frac{\mu_x^2}{\bar{x}^2}[\hat{V}(\bar{y}) + \hat{R}^2 \hat{V}(\bar{x}) - 2\hat{R}\hat{C}(\bar{x};\bar{y})] \ .$$

4. Un intervalle de confiance approximatif à 100(1-α) % est donné par

(6.9) $$\hat{\mu}_{yq} - z_{\alpha/2}\sqrt{\hat{V}(\hat{\mu}_{yq})} \le \mu_y \le \hat{\mu}_{yq} + z_{\alpha/2}\sqrt{\hat{V}(\hat{\mu}_{yq})} \ .$$

Le facteur $\frac{\mu_x^2}{\bar{x}^2}$ est parfois omis, ce qui se justifie par l'équation (6.7), ce qui donne l'estimateur suivant:

(6.10) $$\hat{V}(\hat{\mu}_{yq}) = [\hat{V}(\bar{y}) + \hat{R}^2 \hat{V}(\bar{x}) - 2\hat{R}\hat{C}(\bar{x};\bar{y})] \ .$$

C'est l'estimateur (6.8) que nous privilégierons dans la suite.

Exemple 6.3 *Estimation par le quotient*

Considérons l'estimateur par le quotient dans le problème traité à l'Exemple 6.1 (données du Tableau 5.2). On a

$$\hat{R} = \frac{\bar{y}}{\bar{x}} = \frac{4259{,}156}{3871{,}4} = 1{,}00159, \text{ et}$$

$$\hat{\mu}_{yq} = 1{,}00159(6378{,}709) = 7017{,}594.$$

L'estimation de l'écart-type est

$$\sqrt{\hat{V}(\hat{\mu}_{yq})} = \frac{6378{,}709}{3871{,}4}\sqrt{1-\frac{45}{1183}}\,\frac{\sqrt{(7627{,}543)^2 + (1{,}100159)^2(6480{,}521)^2 - 2(1{,}100159)(49149799)}}{\sqrt{45}}$$

$$= 224{,}1128.$$

Il y a peu de différence entre cet estimateur et l'estimateur par la différence, dont l'écart-type est estimé à 200,3098.

Intervalle de confiance à 95 % :

$$7017{,}594 - 1{,}96(224{,}1128) \leq \mu_y \leq 7017{,}594 + 1{,}96(224{,}1128)$$

$$6578 \leq \mu_y \leq 7457.$$

6.3 ESTIMATEUR PAR LA RÉGRESSION

L'estimateur par la régression permet de corriger les défauts de l'estimateur par la différence. L'estimateur par la différence fonctionne relativement bien lorsque les variables X et Y s'expriment dans les mêmes unités et sont du même ordre de grandeur. Appliqué, comme à l'Exemple 6.2, au nombre d'habitants en 2011 et 2006, il donne de bons résultats. Mais dans d'autres contextes, l'estimateur par la différence — tel que présenté — est plus difficile à justifier et peut parfois sembler carrément insensé.

Dans l'Exemple 5.1, où Y est la production d'orge et X la superficie d'une population de fermes, la quantité $(\overline{y} - \overline{x})$ dans $\hat{\mu}_{ydiff} = \mu_x + (\overline{y} - \overline{x})$ est une différence entre un nombre d'hectares et un nombre de tonnes, ce qui n'a pas un sens concret. La différence $\mu_x - \overline{x}$ dans $\hat{\mu}_{ydiff} = \overline{y} + (\mu_x - \overline{x})$ s'exprime en hectares mais il s'ajoute à $\overline{y}$, un nombre de tonnes. Ce n'est pas le genre de chose qu'on devrait faire sans une certaine normalisation qui rend les valeurs de X — et par conséquent celles de l'écart $\mu_x - \overline{x}$ — commensurables avec $\overline{y}$. En d'autres termes, l'estimateur par la différence devrait être défini plus généralement par

$$\hat{\mu}_{ydiff} = \overline{y} + b(\mu_x - \overline{x}) , \tag{6.11}$$

où le facteur « b » modifie l'unité de mesure — de façon à rendre les unités de X comparables à celles de Y. C'est pourquoi le choix $b = 1$ est raisonnable quand un changement de l'unité de mesure n'est pas nécessaire, comme dans l'Exemple 6.2. Autrement, il faut sérieusement s'interroger sur ce qui constitue un choix approprié.

Exemple 6.4 *Effet de l'unité de mesure sur l'estimateur par la différence*

Dans le contexte de l'Exemple 5.1, considérons l'estimateur par la différence avec pour variable auxiliaire la superficie X des terres réservées à la culture de l'orge, dont la moyenne $\mu_x = 52{,}42242$ hectares est connue. Le tableau suivant présente l'écart-type (l'écart-type exact, calculé à partir des données de la population) d'un même estimateur, avec la même variable auxiliaire, mais exprimées en unités différentes, de plusieurs estimateurs basés sur un *éas* de taille $n = 50$. Le tableau suivant montre à quel point les qualités de $\hat{\mu}_{ydiff}$ dépendent de l'unité de mesure de la variable auxiliaire.

	Estimateur et variable auxiliaire	Écart-type de l'estimateur
1	X = Superficie en hectares	8,98
2	$X_2 = 2{,}4711X$ (Superficie en acres)	2,58
3	$X_3 = 2{,}9249X$ (Superficie en arpents)	1,09
4	$X_4 = 0{,}01X$ (Superficie en km carrés)	13,37
5	$X_5 = 1000X$ (Superficie en mètres carrés)	4440
6	$X_6 = 3{,}003\,X$ (Estimateur par la régression)	1,03

Comment choisir le facteur b par lequel normaliser X? L'estimateur *par la régression* correspond à un choix particulier de b, que nous désignons désormais par $\hat{B}$, soit

(6.12) $$\hat{B} = \frac{s_{xy}}{s_x^2} .$$

L'estimateur par la régression est donc

(6.13) $$\hat{\mu}_{yr} = \overline{y} + \hat{B}(\mu_x - \overline{x}) \ .$$

Remarque *La notation* $\hat{B}$

La notation « $\hat{B}$ » semble dire que $\hat{B}$ est l'estimateur d'un paramètre B. En effet : B est le rapport $B = \frac{S_{xy}}{S_x^2}$. Si B était connu, l'estimateur « idéal » parmi tous ceux de cette forme aurait été $\overline{y} + B(\mu_x - \overline{x})$, idéal dans le sens qu'il aurait la plus petite variance possible. Puisque B est inconnu, on l'estime par $\hat{B}$.

L'estimateur par la régression n'est pas strictement sans biais, mais son biais est négligeable lorsque l'échantillon est grand.

Le théorème suivant énonce ses propriétés.

THÉORÈME 6.3 **ESTIMATEUR PAR LA RÉGRESSION**

1. L'estimateur par la régression est à peu près sans biais:

$$E(\hat{\mu}_{yr}) \approx \mu_y \ .$$

2. Son erreur quadratique moyenne est à peu près égale à sa variance, laquelle est à peu près égale à

(6.14) $$V(\hat{\mu}_{yr}) \approx \frac{1-f}{n}[S_y^2 - BS_{xy}] = \frac{1-f}{n}S_y^2(1-\rho^2) \ = V(\overline{y})(1-\rho^2) \ ,$$

où $B = \frac{S_{xy}}{S_x^2}$ et $\rho = \frac{S_{xy}}{S_x S_y}$.

3. On estime $V(\hat{\mu}_{yr})$ par

(6.15) $$\hat{V}(\hat{\mu}_{yr}) = \frac{1-f}{n}[s_y^2 - \hat{B}s_{xy}] = \frac{1-f}{n}s_y^2(1-r^2) = \hat{V}(\bar{y})(1-r^2)\,,$$

où $r = \dfrac{s_{xy}}{s_x s_y}$ est le coefficient de corrélation échantillonnal.

Dans l'équation (6.15), on remplace parfois s_y^2 par $\dfrac{n-1}{n-2}s_y^2$, ce qui donne l'estimateur suivant:

(6.16) $$\hat{V}(\hat{\mu}_{yr}) = \frac{1-f}{n}[s_y^2 - \hat{B}s_{xy}] = \frac{1-f}{n}\frac{\sqrt{n-1}}{\sqrt{n-2}}s_y^2(1-r^2) = \hat{V}(\bar{y})(1-r^2)\,.$$

Cette modification — inspirée par un modèle théorique que nous n'invoquons pas ici — est préférable dans le sens qu'il risque un peu moins de sous-estimer la variance. Mais la différence entre (6.15) et (6.16) est négligeable, dans la mesure où l'estimateur par la régression n'est recommandé que lorsque n est grand.

Exemple 6.5 *Estimation par la régression*

Dans l'exemple 6.2, nous avons s_x = 6480,521, s_y = 7627,543, s_{xy} = 49 149 799 et donc

$\hat{B}$ = 1,170312. Donc $\hat{\mu}_{yr}$ = 4259,156 + 1,170312(6378,709 – 3871,4) = 7193,491.

Le coefficient de corrélation entre X et Y dans l'échantillon est r = 0,9943221. L'écart-type de l'estimateur est estimé par $\hat{V}(\hat{\mu}_{yr}) = \dfrac{\sqrt{1-45/1183}}{\sqrt{45}}\dfrac{\sqrt{45-1}}{\sqrt{45-2}}(7627{,}543)\sqrt{1-(0{,}9943221)^2}$ = 120,04.

Justification de l'estimateur par la régression

L'estimateur par la régression peut être considéré comme une amélioration, voire une *correction*, de l'estimateur par la différence. Défini largement comme $\bar{y} + b(\mu_x - \bar{x})$, l'estimateur par la différence soulève la question de savoir comment choisir b. Une valeur trop petite n'apporte pas les bénéfices escomptés alors qu'une valeur trop grande peut nuire considérablement à la qualité de l'estimateur. Mais entre inutile et nuisible, il existe une valeur *optimale*, celle qui minimise la variance de l'estimateur. Quelle est cette valeur? Le théorème suivant le dit.

THÉORÈME 6.1 **LA VALEUR DE B QUI MINIMISE LA VARIANCE**

La valeur de b qui minimise la variance

(6.17) $$V[\bar{y} + b(\mu_x - \bar{x})] = V(\bar{y}) + b^2V(\bar{x}) - 2bC(\bar{y};\bar{x})$$

de l'estimateur $\bar{y} + b(\mu_x - \bar{x})$ est

(6.18) $$B = \frac{C(\bar{x};\bar{y})}{V(\bar{x})} = \frac{(1-f)S_{xy}/n}{(1-f)S_x^2/n} = \frac{C(\bar{x};\bar{y})}{V(\bar{x})} = \frac{S_{xy}}{S_x^2}.$$

C'est ce qui motive l'estimateur par la régression. Si B était connu, « l'estimateur » $\bar{y} + B(\mu_x - \bar{x})$ serait le meilleur parmi tous les estimateurs de la forme $\bar{y} + b(\mu_x - \bar{x})$. Dans l'Exemple 6.5, la valeur réelle de B (calculée à partir de la population entière) est B = 1,026 et $\sqrt{V[\bar{y} + B(\mu_x - \bar{x})]}$ = 177,6. Aucun autre estimateur de la forme $\bar{y} + b(\mu_x - \bar{x})$ ne donne un écart-type aussi petit.

En pratique, B n'étant pas connu, on l'estimera par $\hat{B} = \frac{\hat{C}(\bar{x};\bar{y})}{\hat{V}(\bar{x})} = \frac{s_{xy}}{s_x^2}$. C'est ce qui donne l'estimateur par la régression, $\hat{\mu}_{yr} = \bar{y} + \hat{B}(\mu_x - \bar{x})$.

THÉORÈME 6.2 **LA VARIANCE MINIMALE**

L'espérance de $\bar{y} + B(\mu_x - \bar{x})$ est μ_y et sa variance est $\frac{1-f}{n}[S_y^2 - BS_{xy}] = \frac{1-f}{n}S_y^2(1-\rho^2)$.

Ces propriétés sont perdues lorsqu'on remplace B par $\hat{B}$. $\hat{\mu}_{yr}$ n'est pas sans biais et sa variance n'est pas si aisément calculée. On peut cependant montrer que lorsque n est grand, le biais est négligeable et la variance est à peu près égale à $\frac{1-f}{n}[S_y^2 - BS_{xy}] = \frac{1-f}{n}S_y^2(1-\rho^2)$. Pour estimer la variance, on remplace les paramètres dans cette expression par leur estimateur, soit S_y par s_y et ρ par r. Habituellement, cependant, on remplace s_y par $\sqrt{\frac{\sum_{i\in\omega}(y_i - \bar{y})^2}{n-2}} = \frac{\sqrt{n-1}s_y}{\sqrt{n-2}}$.

Deuxième justification de l'estimateur par la régression

Il existe une deuxième façon de motiver l'estimateur par la régression. Si les données de la population présentent une relation à peu près linéaire entre X et Y, la droite de régression (ou droite des moindres carrés) est donnée par

$$y = \mu_y + B(x - \mu_x).$$

Évaluée à $x = \mu_x$, la droite de régression donne $y = \mu_y + B(\mu_x - \mu_x) = \mu_y$. L'équation de la droite peut être estimée par $y = \bar{y} + \hat{B}(x - \bar{x})$ et on l'évalue à $x = \mu_x$, ce qui donne l'estimateur $\bar{y} + \hat{B}(\mu_x - \bar{x})$. C'est l'estimateur par la régression.

Remarque *L'estimateur est indépendant des unités*

L'estimateur par le quotient ne souffre pas de l'inconvénient illustré à l'Exemple 6.4, car il ne dépend pas de l'unité choisie pour *X*. Si *X* est multiplié par une constante *b*, μ_x et $\bar{x}$ se voient multipliés par *b* et l'estimateur $\frac{\bar{y}}{\bar{x}}\mu_x$ devient $\frac{\bar{y}}{b\bar{x}}b\mu_x = \frac{\bar{y}}{\bar{x}}\mu_x$: il ne change pas.

6.4 COMPARAISON DES ESTIMATEURS

Dans l'Exemple 6.1, comparons les quatre estimateurs. Les estimations elles-mêmes sont très différentes : de l'ordre de 4 200 pour l'estimateur par la moyenne et de 7 000 pour les deux autres estimateurs. Ceci met en évidence l'importance de choisir le bon estimateur. La conséquence n'est pas banale. Le tableau suivant, qui présente leur écart-type (réel mais approximatif dans le cas des deux derniers), montre à quel point leur qualité est variable.

		Écart-type
1	Estimateur par la moyenne	7852
2	Estimateur par la différence	268
3	Estimateur par le quotient	240
4	Estimateur par la régression	178

Remarque

Ces résultats montrent que l'estimateur par le quotient est supérieur à l'estimateur par la différence, et c'est ce à quoi on s'attendrait. Les deux estimateurs sont basés sur une estimation de l'accroissement du nombre d'habitants de 2006 à 2011 : l'accroissement *absolu* dans le cas de l'estimateur par la différence; et l'accroissement *relatif* dans le cas de l'estimateur par le quotient. Il y a de bonnes raisons de croire que l'accroissement *relatif* est plus stable comparé à l'accroissement absolu. L'échantillon, pourtant, a mené à la conclusion contraire : l'écart-type est estimé à 200 pour l'estimateur par la différence et à 224 pour l'estimateur par le quotient. Cette conclusion est accidentelle, comme le montrent les calculs ci-dessus, basés sur les données de la population.

On constate que les trois derniers estimateurs, qui utilisent la variable auxiliaire, en tirent un profit considérable.

Devant cet embarras du choix, il serait utile de se faire une idée des qualités relatives des estimateurs. Prenons pour base de comparaison l'estimateur par la moyenne. Pour chacun des autres estimateurs nous poserons la question suivante : dans quelles circonstances est-il supérieur à l'estimateur par la moyenne, le terme « supérieur » étant utilisé dans le sens habituel: $\hat{\theta}_1$ est supérieur à $\hat{\theta}_2$ si $EQM(\hat{\theta}_1) < EQM(\hat{\theta}_2)$.

C'est en fonction de *B*, la pente de la droite de régression de *Y* sur *X*, et de ρ, le coefficient de corrélation entre *X* et *Y*, que nous allons distinguer les estimateurs. Commençons par

énoncer formellement les conditions dans lesquelles les estimateurs par la différence, par le quotient et par la régression sont préférables à $\bar{y}$.

Estimateur par la différence

Considérons l'estimateur par la différence dans sa forme générale, $\hat{\mu}_{yd} = \bar{y} + b(\mu_x - \bar{x})$, où b est un nombre fixe. L'estimateur par la différence est supérieur à l'estimateur par la moyenne si et seulement si

$$B > \frac{b}{2}, \text{ ou encore, si } \rho > \frac{b}{2}\frac{S_x}{S_y}.$$

Comment savoir si ces conditions sont vérifiées? La quantité B est inconnue, mais son sens concret permet en général d'en avoir une bonne idée — et de choisir b en conséquence, c'est-à-dire, $b < 2B$. B est le taux d'accroissement de Y par rapport à X. Dans le contexte de l'Exemple 6.1, B répond à la question suivante: si l'écart entre deux villes était de 1000 habitants en 2006, que devrait-il être en 2011 ? A priori on soupçonne ceci: $B \approx 1000$ (très vraisemblable); ou B légèrement supérieure à 1000 (possible aussi) ; ou moins de 500 (fort peu probable). Il est très probable que B soit de l'ordre de 1, sinon un peu supérieur. On peut donc prévoir que l'estimateur par la différence avec $b < 2$ donnera de bons résultats (meilleurs que la moyenne).

Un fait à remarquer : il existe toujours un choix de b qui rend l'estimateur par la différence supérieur à l'estimateur par la moyenne: il suffit de choisir b assez petit. Si $b = 0$, l'estimateur par la différence coïncide avec la moyenne $\bar{y}$; il s'améliore à mesure que b augmente, tant que b demeure inférieur à $2B$. Après quoi, l'estimateur par la différence perd de son efficacité jusqu'à être nettement moins bon que la moyenne $\bar{y}$. Lorsque, comme à l'Exemple 6.1, X et Y s'expriment dans les mêmes unités et que $S_y \approx S_x$, la condition $\rho > \frac{b}{2}\frac{S_x}{S_y}$ devient $b < 2\rho$.

Estimateur par le quotient

L'estimateur par le quotient est plus précis que l'estimateur par la moyenne si et seulement si

$$B > \frac{R}{2}, \text{ ou encore, si } \rho > \frac{1}{2}\frac{cv(X)}{cv(Y)},$$

où $cv(X) = \frac{S_x}{\mu_x}$ et $cv(Y) = \frac{S_y}{\mu_y}$ sont les coefficients de variation de X et de Y. Normalement, les conditions qui recommandent l'estimation par la différence recommandent aussi l'estimation par le quotient.

Le choix entre l'estimateur par la différence ou l'estimateur par le quotient est plus difficile à faire. Formellement, on peut démontrer que l'estimateur par la différence est préférable à l'estimateur pas le quotient si $|b - B| < |R - B|$.

Donc l'estimateur par la différence battra l'estimateur par le quotient dans la mesure où le choix de b s'approche de la valeur optimale B. En revanche, l'estimateur par le quotient est meilleur si la droite des moindres carrés entre Y et X passe par l'origine. Dans ce cas $\mu_y - B\mu_y = 0 \Rightarrow R = B$ et donc forcément $|b - B| \geq |R - B|$.

Estimateur par la régression

L'estimateur par la régression $\hat{\mu}_{yr}$ est en principe supérieur aux trois autres. Le gain qu'il permet de réaliser se mesure par les différences entre leurs variances :

$$
\begin{aligned}
&V(\bar{y}) - V(\hat{\mu}_{yr}) = \rho^2 S_y^2; \\
&V(\hat{\mu}_{yq}) - V(\hat{\mu}_{yr}) = S_x^2 (R - B)^2; \\
&V(\hat{\mu}_{yd}) - V(\hat{\mu}_{yr}) = S_x^2 (b - B)^2.
\end{aligned}
\tag{6.19}
$$

Mise en garde Rappelons que ces conclusions sont basées sur des calculs qui, dans le cas des estimateurs par le quotient et par la régression, sont approximatifs. Elles ne sont justifiées que lorsque les échantillons sont grands.

6.5 SIMULATIONS

Nous avons effectué quelques simulations sur la population décrite au tableau A05, constituée de 1 160 municipalités québécoises. Notre objectif est d'examiner les propriétés de trois estimateurs de la moyenne μ_y de la variable

Y : Nombre de familles avec 3 enfants ou plus

à partir d'un échantillon de n = 100 municipalités.

Les trois estimateurs: estimateurs par la moyenne, par le quotient et par la régression.

Dans les deux derniers cas, la variable auxiliaire est

X : le nombre d'habitants

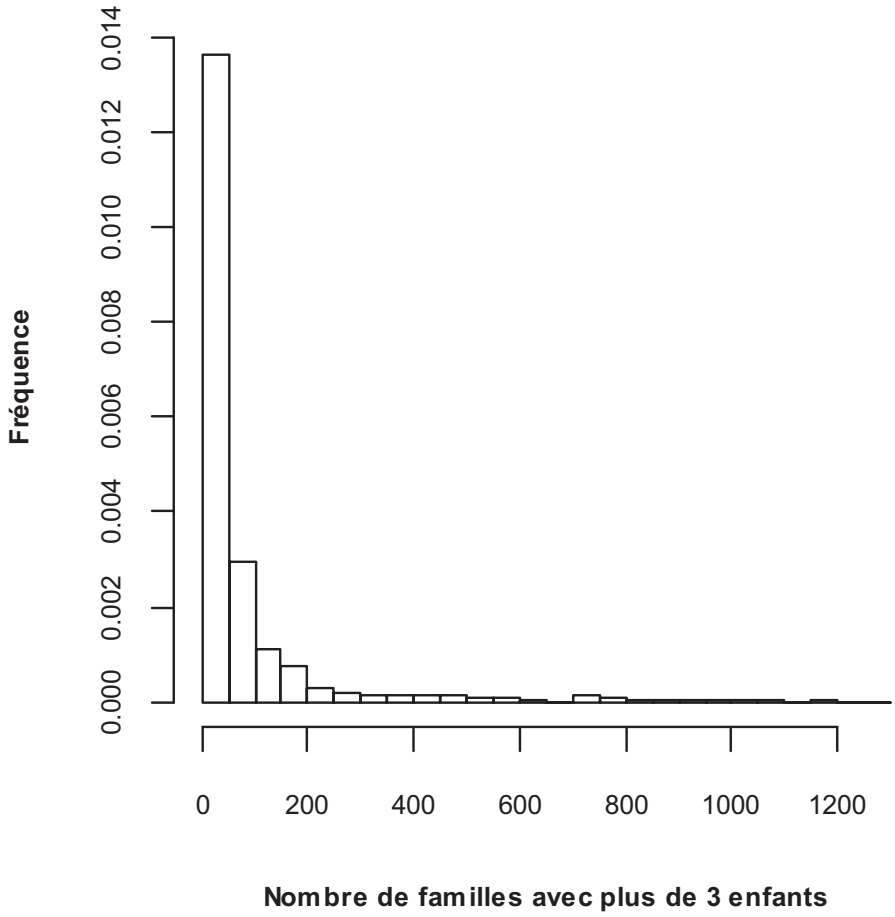
dont la moyenne, supposée connue, est μ_x = 6 733.

Nous avons généré 10 000 échantillons de taille 100 de façon à estimer les propriétés des trois estimateurs:

Y est la variable d'intérêt, X une variable auxiliaire. Les moyennes de ces variables sont: μ_y = 163,5474 et μ_x = 6 733,103.

La distribution de Y est fortement asymétrique, comme on le voit dans la Figure 6.1:

Figure 6.1
Distribution du nombre de familles avec plus de 3 enfants



(En réalité, l'asymétrie est encore plus prononcée qu'elle ne le paraît dans le graphique, car nous avons omis les données les plus extrêmes pour les besoins de cette présentation.)

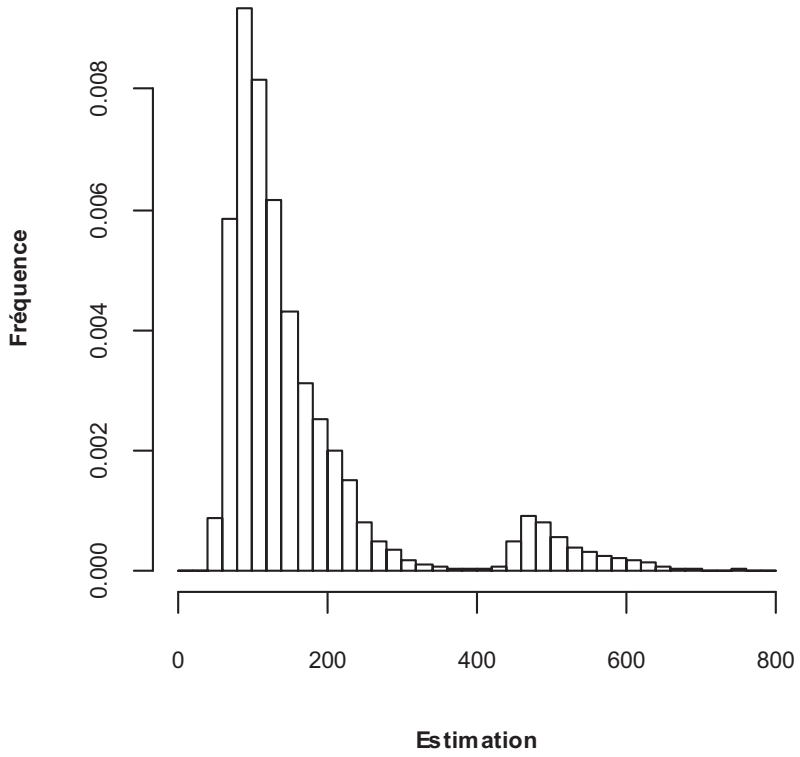
Un tel degré d'asymétrie ne manquera pas de présenter de sérieux problèmes. Examinons-en les effets sur les trois estimateurs.

Estimation par la moyenne.

Commençons par l'estimateur $\bar{y}$. On sait qu'il est sans biais et que l'estimateur de sa variance l'est aussi. Ce qui pose problème, c'est plutôt la *distribution* de $\bar{y}$, dont la normalité est douteuse. Par conséquent, le niveau réel d'un intervalle de confiance risque de s'éloigner du niveau stipulé. C'est ce que confirment nos simulations. La Figure 6.2 présente la distribution observée des $\bar{y}$.

Comme prévu, la loi de $\bar{y}$ est loin de la normalité. Son caractère bimodal était prévisible, étant donné les quelques villes beaucoup plus grandes que les autres: quelques échantillons les contiennent, d'autres pas. Dans les deux cas, l'intervalle de confiance a de fortes chances de ne pas inclure la moyenne. On peut donc s'attendre à ce qu'un intervalle de confiance basé sur la normalité soit peu fiable.

Figure 6.2
Distribution de la moyenne échantillonnale



On trouve en effet que les niveaux réels des intervalles de confiance sont inférieurs à leur niveau nominal. On constate, par exemple, que le niveau réel d'un intervalle de confiance à 95 % (théoriquement) est en fait un intervalle de confiance à 68 %[3].

Comment résoudre ce problème? L'utilisation d'une variable auxiliaire est certainement une piste à suivre. L'estimateur par la moyenne fait défaut à cause des quelques grandes villes (Montréal, Québec, Laval, Lachine) dont la valeur de Y s'écarte considérablement de la moyenne. Mais ces grandes villes ne se distinguent pas particulièrement des autres quant au *rapport* du nombre de familles avec trois enfants ou plus sur le nombre d'habitants.

Estimation par le quotient

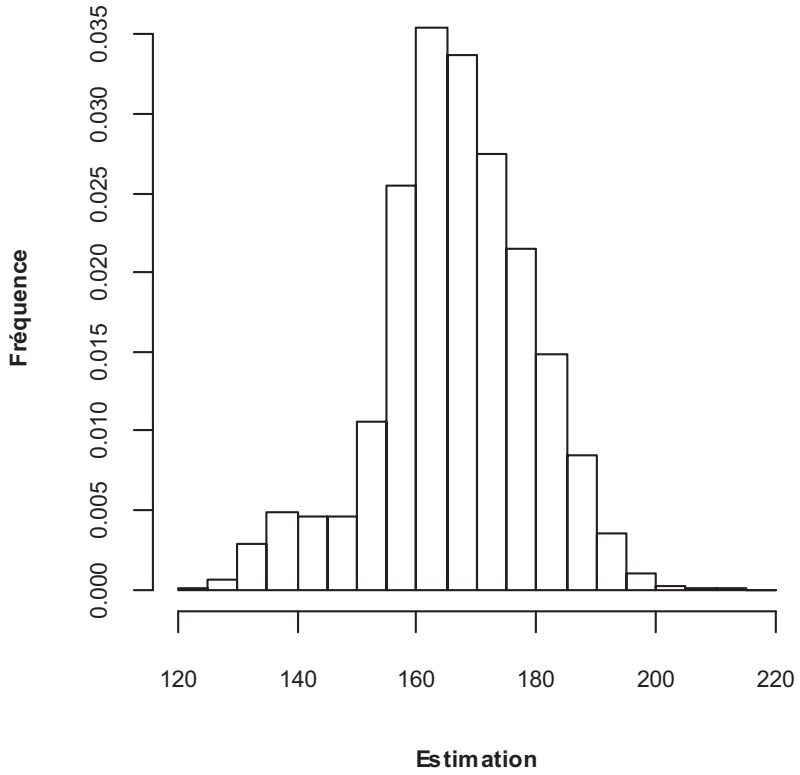
Considérons donc l'estimateur par le quotient, $\hat{\mu}_R$. La Figure 6.3 présente la distribution de l'estimateur par le quotient.

On voit donc que la bi-modalité observée dans le cas de $\bar{y}$ a disparu. Le biais de l'estimateur n'est pas inquiétant: la simulation l'estime à 1,8 %. Par contre, le biais de l'estimateur de la variance est de -47,5 %: une tendance à sous-estimer la variance. Cette

[3] Notez, cependant, que le niveau « réel » n'est pas tout à fait réel non plus. Il est, lui aussi, estimé. Mais étant donné le grand nombre d'échantillons (10 000), l'estimation est probablement assez fiable.

estimation optimiste se traduira forcément par un intervalle trop étroit. Néanmoins, on estime le niveau réel à 91 %, quand même pas très éloigné du niveau nominal.

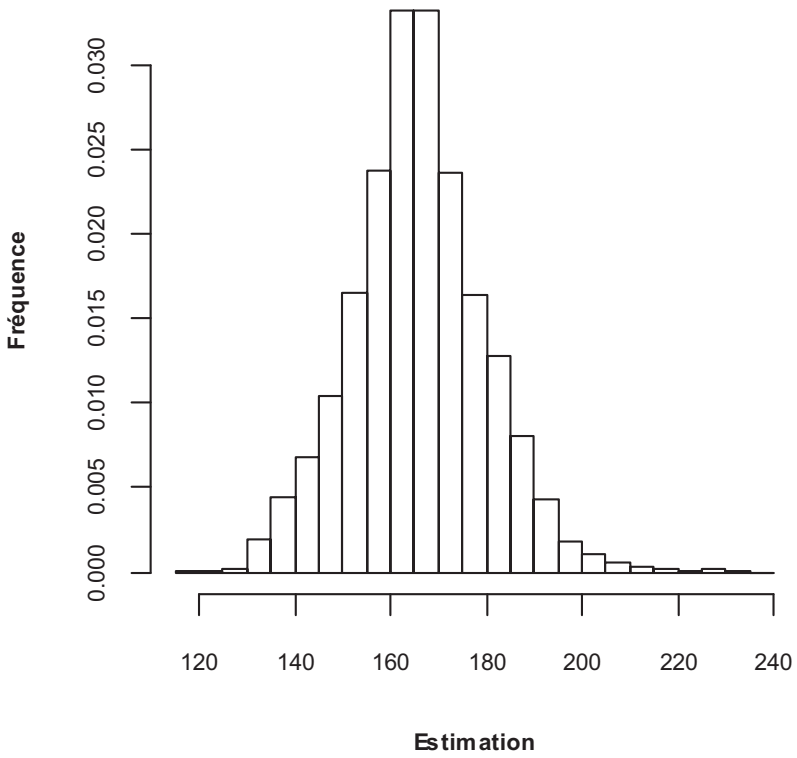
Figure 6.3
Distribution de l'estimateur par le quotient



Estimation par la régression

L'estimateur par la régression devrait, comme l'estimation par le quotient, contourner la difficulté présentée par $\bar{y}$. La Figure 6.4 présente la distribution de l'estimateur par la régression :

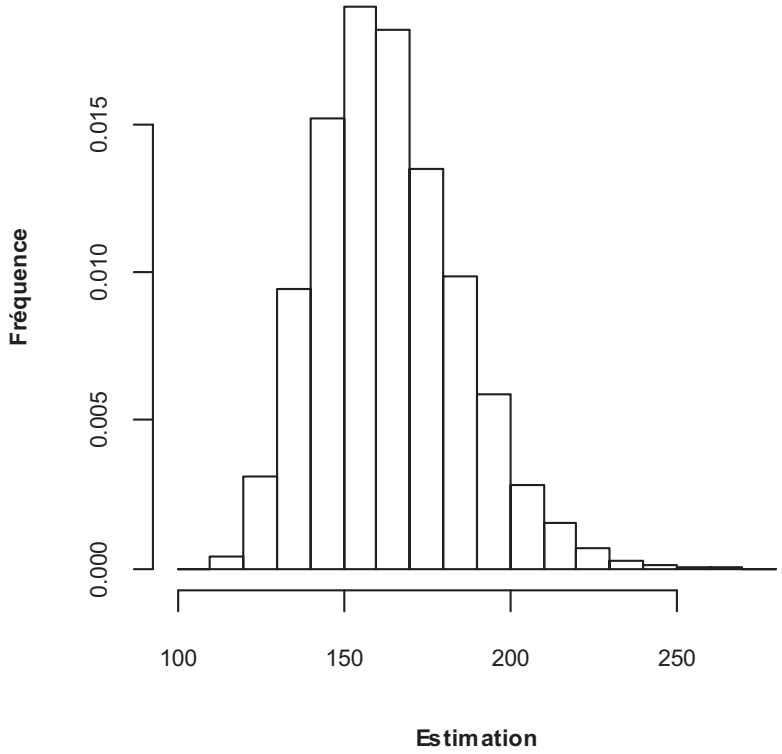
Figure 6.4
Distribution de l'estimateur par la régression



Le problème ici, comme avec l'estimateur par le quotient, est l'estimateur de la variance — et non la normalité de l'estimateur. On découvre, en effet, que l'estimateur de la variance présente un biais négatif d'environ 67 %, une estimation trop optimiste de la qualité de l'estimateur. Et dans ce cas-ci, cela affecte le niveau réel de l'intervalle de confiance, estimé à 70 %.

Comment se comparent ces estimateurs à une autre technique, déjà explorée, qui consiste à soustraire un certain nombre m des plus grandes données de la population? La Figure 6.5 présente la distribution de la moyenne échantillonnale lorsque les 10 plus grandes données sont exclues de la population. La technique est utile lorsqu'on utilise l'estimation par la moyenne car elle rend sa distribution plus symétrique. Elle l'est moins lorsqu'on utilise les estimateurs par le quotient et par la régression. Il est rarement utile d'appliquer les deux techniques simultanément, c'est-à-dire, exclure les m plus grandes unités et utiliser l'estimation par le quotient ou par la régression dans un échantillon de taille $n - m$.

Figure 6.5
Distribution de l'estimateur de la moyenne
après l'exclusion des 10 plus grandes données de la population



RÉSUMÉ

Notation

- Pour toute paire de variables *X*, *Y*, $C(\overline{x}\,;\overline{y}) = (1-f)S_{xy}/n$, $f = n/N$. Ce paramètre est estimé par son équivalent échantillonnal, $\hat{C}(\overline{x}\,;\overline{y}) = (1-f)s_{xy}/n$.

- Estimateur par la différence: $\hat{\mu}_{ydiff} = \mu_x + b(\overline{y}-\overline{x})$; $\sqrt{\hat{V}(\hat{\mu}_{ydiff})} = \sqrt{1-f}\,\dfrac{\sqrt{s_y^2 + b^2 s_x^2 - 2bs_{xy}}}{\sqrt{n}}$, où *b* est un nombre fixe choisi d'avance, souvent égal à 1.

- Estimateur par le quotient: $\hat{\mu}_{yq} = \hat{R}\mu_x$; $\hat{V}(\hat{R}) \approx \dfrac{\mu_x^2}{\overline{x}^2}[\hat{V}(\overline{y}) + \hat{R}^2\hat{V}(\overline{x}) - 2\hat{R}\hat{C}(\overline{x};\overline{y})]$.

- Estimateur par la régression $\hat{\mu}_{yr} = \overline{y} + \hat{B}(\mu_x - \overline{x})$; $\hat{B} = \dfrac{s_{xy}}{s_x^2}$; $\hat{V}(\hat{\mu}_{yr}) = \dfrac{1-f}{n}\dfrac{(n-1)s_y^2}{n-2}(1-r^2)$.

EXERCICES

6.1 Les 2 350 abonnés d'un club vidéo ont dépensé la somme de 60 $ en moyenne le mois dernier. Vous voulez estimer les recettes totales τ pour le mois qui vient de s'écouler. à partir d'un échantillon de taille 20. Pour chaque abonné de l'échantillon, vous prenez note de *X*, les dépenses le mois dernier, et de *Y*, les dépenses pour ce mois-ci. Les données sont présentées au Tableau 6.1:

Tableau 6.1

x	y	x	y	x	y	x	y
58,58	52,00	59,85	53,65	46,53	31,84	50,54	55,12
49,79	53,60	48,57	43,17	58,51	57,16	62,20	52,76
83,50	80,49	61,78	67,26	75,27	62,89	66,07	61,33
53,90	54,22	56,30	60,59	77,08	75,39	50,88	51,21
60,63	71,41	49,90	48,41	79,87	79,85	67,87	70,72

$\Sigma x_i = 1217,62;\ \sum y_i = 1183,07\ ;\ \sum x_i^2 = 76\,439,72\ ;\ \sum y_i^2 = 72\,878,21\ ;\ \sum x_i y_i = 74\,250,98$

a) Estimer τ par la méthode de la moyenne $\bar{y}$ et déterminer un intervalle de confiance à 95%.

b) Estimer τ par la différence et déterminer un intervalle de confiance à 95%.

c) Estimer τ par le quotient et déterminer un intervalle de confiance à 95%.

d) Estimer τ par la régression et déterminer un intervalle de confiance à 95%.

6.2 Vous prélevez un échantillon de 20 comptes d'une population de 4 890 comptes. Vous connaissez les valeurs aux livres des comptes de la population; leur moyenne est de 80 $. Les valeurs aux livres (*X*) et les valeurs corrigées (*Y*) sont présentées au Tableau 6.2:

Tableau 6.2

x	y	x	y	x	y	x	y
47,90	47,90	60,08	60,08	63,72	63,72	74,82	74,82
64,41	64,41	69,81	69,81	58,01	58,01	58,28	58,28
65,51	65,51	59,96	59,96	55,59	55,59	63,25	63,25
64,91	64,91	71,27	71,27	62,51	62,51	52,34	52,34
91,55	75,25	79,40	59,70	64,45	71,62	72,75	68,19

Σx =1300,52; s_x = 9,808355; Σy = 1267,13; s_y = 7,212526; s_{xy} = 56,474343.

a) Estimer τ par la moyenne $\bar{y}$ et déterminer un intervalle de confiance à 95%.

b) Estimer τ par la différence et déterminer un intervalle de confiance à 95%.

c) Estimer τ par le quotient et déterminer un intervalle de confiance à 95%.

6.3 [*Suite du numéro* 6.1] Estimer par la régression les dépenses moyennes par abonné et estimer l'écart-type de l'estimateur.

6.4 [*Suite du numéro* 6.2] Estimer par la régression les dépenses moyennes par abonné et estimer l'écart-type de l'estimateur.

6.5 Considérez un échantillon de taille 80 présenté à l'annexe A09.1, tiré de la population de 708 subdivisions de recensement. Le nombre moyen de familles dans la population est 9 654,254. Soit Y le nombre de couples mariés sans enfants et X le nombre de familles, une variable auxiliaire. Estimez μ_y par la méthode indiquée et estimer l'écart-type de l'estimateur. Dans chaque cas, estimer l'écart-type de l'estimateur ; pour b), c) et d) utiliser le nombre de familles comme variable auxiliaire.

a) Estimer μ_y par la moyenne.

b) Estimer μ_y par la différence avec X comme variable auxiliaire.

c) Estimer μ_y par le quotient.

d) Estimer μ_y par la régression.

$\sum x_i = 631\,700;\ \sum y_i = 223\,975\ ;\ \ s_x = \ 17197{,}15;\ s_y = 5676{,}089\ ;\ s_{xy} = 91370165$

6.6 Reprenez l'exercice 6.5 en prenant cette fois-ci pour variable auxiliaire le nombre X de couples mariés dans la population (Moyenne de la population: 7 793,044).

$\sum x_i = 456\,765;\ \sum y_i = 223\,975;\ \ s_x = \ 11\,523{,}49;\ s_y = 5676{,}089;\ s_{xy} = 91\,372\,291$

6.7 On remarque, dans l'exercice 6.5, que l'estimateur par la différence (avec pour variable auxiliaire le nombre total de familles) est nettement moins précis que ses concurrents. Ceci est dû au fait que l'ordre de grandeur de la variable auxiliaire X est supérieur à celui de la variable d'intérêt.

a) Montrer que l'estimateur par la différence s'améliore considérablement lorsqu'on prend $0{,}3X$ pour variable auxiliaire. Remarquer que l'effet de cette transformation a été de rendre les valeurs de la variable auxiliaires comparables à celles de Y.

b) Déterminer l'estimation par la régression (avec X pour variable auxiliaire). Vous remarquerez que l'écart-type est proche de celui calculé en a) et que $\hat{B}$ est proche du facteur 0,3 utilisé en a). Commenter.

6.8 [*Suite de l'exercice* 6.5] Contrairement à l'estimateur par la différence, l'estimateur par la régression ne dépend pas de l'unité de mesure. Vérifier ce fait numériquement en estimant par la régression le nombre moyen de couples mariés sans enfants en prenant pour variable auxiliaire

a) le nombre X de couples mariés dans la population;

b) la moitié $0{,}5X$ du nombre de couples mariés dans la population.

6.9 Lequel des trois estimateurs de la moyenne μ_y serait, d'après vous, le meilleur : l'estimateur par la moyenne, par la différence ou par le quotient ? La réponse sera nécessairement spéculative, certainement pas catégorique.

a) La population est l'ensemble des fermes d'une région ; μ_y est la récolte moyenne (en milliers de tonnes) de blé. Quel est le meilleur estimateur si

i) vous connaissez la superficie cultivée dans chacune des fermes de la population ?

ii) vous connaissez le nombre de personnes qui travaillent à la récolte dans chacune des fermes de la population ?

iii) vous connaissez la récolte des mêmes fermes l'année dernière ?

b) La population est l'ensemble des employés d'une grande compagnie ; μ_y est le revenu moyen des employés. Vous connaissez les salaires des employés de la population.

c) La population est l'ensemble des employés d'une grande compagnie ; μ_y est le revenu moyen des ménages auxquels ils appartiennent. Vous connaissez le salaire des employés de la population.

d) μ_y est la moyenne des recettes d'une population de petits commerces ; vous connaissez le nombre d'employés dans chaque commerce.

e) μ_y est la moyenne des dépenses d'une population de petits commerces ; vous connaissez les recettes de chaque commerce.

f) μ_y est le nombre moyen de téléviseurs dans les ménages d'une population. Vous connaissez le nombre de personnes dans chaque ménage de la population.

g) μ_y est le nombre moyen de personnes dans les ménages d'une population. Vous connaissez le nombre de personnes dans les mêmes ménages l'année dernière.

6.10 L'exercice 6.7 montre que l'estimateur par la différence peut être amélioré lorsqu'on transforme la variable auxiliaire X en une variable $X' = cX$, où c est choisi de telle sorte que les valeurs de X' sont du même ordre de grandeur que celles de Y.

a) On pourrait alors s'inspirer de l'échantillon pour choisir le nombre c qui aura cet effet, soit $c = \frac{\bar{y}}{\bar{x}}$. Montrer qu'en utilisant $X' = \frac{\bar{y}}{\bar{x}}X$ pour variable auxiliaire, l'estimateur par la différence n'est nul autre que l'estimateur par le quotient.

b) Supposons qu'on tente de multiplier les « améliorations » en utilisant un estimateur par la régression avec X' pour variable auxiliaire. Montrer qu'on n'a rien gagné: l'estimateur par la régression est le même, qu'on utilise X' ou X.

6.11 Dans le contexte de l'exercice 5.9 (données du Tableau 5.6), estimer les paramètres suivants ainsi que l'écart-type de l'estimateur. La population est de taille 1 500.

a) Estimer le nombre total d'adultes dans la population en utilisant le fait que le nombre total de personnes dans la population est de 5 370

   i) par le quotient.

   ii) par la régression.

b) Estimer par le quotient le nombre total d'adultes dans les maisons unifamiliales en utilisant le nombre de personnes comme variable auxiliaire et le fait que le nombre total de personnes dans la population est 5 370.

6.12 [*Exige un logiciel*] Considérons la population de $N = 8$ unités pour lesquels sont définies deux variables, X et Y, dont les valeurs sont les suivantes :

y	8	10	67	44	66	56	89	99
x	3	6	24	27	30	36	51	57

Le but de cet exercice est de comparer les estimateurs de μ_y, dans un échantillon de taille $n = 3$. Y est la variable d'intérêt, X une variable auxiliaire supposée connue pour toutes les unités de la population.

a) Calculer $\bar{y}$ et $\hat{\mu}_{yd}$ pour chacun des 56 échantillons possibles, confirmer numériquement que ces estimateurs sont sans biais, et vérifier numériquement les formules de l'écart-type des estimateurs.

b) Calculer $\hat{R}\mu_x$ pour chacun des 56 échantillons possibles et déterminer le biais de l'estimateur par le quotient.

c) Calculer $V(\hat{R}\mu_x)$, et comparer à ce que donnent les formules approximatives (6.8) et (6.10) .

d) Comparer maintenant les estimateurs par le quotient et par la moyenne sous un autre angle. Calculer, pour les estimateurs par la moyenne et par le quotient, la probabilité de se tromper

   i) de plus de 10 % dans l'estimation de la moyenne ;

   ii) de plus de 20 % dans l'estimation de la moyenne.

   Présentez les résultats dans un tableau et commentez.

e) Calculer $\hat{V}(\hat{\mu}_{yq})$ pour chaque échantillon et déterminer le biais de $\hat{V}(\hat{\mu}_{yq})$.

f) Déterminer le biais de l'estimateur par la régression $\hat{\mu}_{yr}$.

g) Déterminer le biais de $\hat{V}(\hat{\mu}_{yr})$.

6.13 [*Propriétés des estimateurs par le quotient et par la régression lorsque la variable auxiliaire X est dichotomique*] Une variable dichotomique X définit une partition de la population U en deux parties, U_1 comprenant les unités pour lesquelles $X = 1$, et son complément $U_2 = U \backslash U_1$. Soit Y une variable définie dans U, $\tau = \sum_{k \in U} t_k$, et $\tau_1 = \sum_{k \in U_1} t_k$.

a) Montrer que l'estimateur par le quotient est plus précis que l'estimateur par la moyenne si et seulement si $\frac{\tau_1}{\tau} \geq \frac{1}{2}\left(1+\frac{N_1}{N}\right)$, où N_1 est la taille de U_1. [Utiliser les formules approximatives pour les variances].

b) Montrer que l'estimateur par le quotient est $\hat{\mu}_{yq} = \bar{y}\frac{N_1/N}{n_1/n}$. Donc l'estimateur majore $\bar{y}$ dans la mesure où U_1 est sous représenté dans l'échantillon.

c) À première vue, l'estimateur par le quotient ne semble pas raisonnable, mais il peut s'expliquer par le sens que prend dans ce contexte le coefficient de corrélation ρ entre X et Y. Montrer que $\rho = 0$ si et seulement si $\mu_1=\mu_2$, où $\mu_1 = \frac{\tau_1}{N_1}, \mu_2 = \frac{\tau_2}{N_2}$.

d) Montrer que l'estimateur de la moyenne par la régression est la moyenne pondérée $\frac{N_1}{N}\bar{y}_1 + \frac{N_2}{N}\bar{y}_2$, où $N_2 = N - N_1$ et $\bar{y}_1$ et $\bar{y}_2$ sont les moyennes des unités échantillonnales appartenant à U_1 et U_2, respectivement.

e) Supposons que Y aussi est dichotomique. Soit p_X la proportion des unités pour lesquelles $X = 1$, p_Y la proportion des unités pour lesquelles $Y = 1$ et p_{XY} la proportion des unités pour lesquelles $X = 1$ *et* $Y = 1$. Montrer que l'estimateur par le quotient est plus précis que l'estimateur par la moyenne si et seulement si $p_{XY} - p_X p_Y \geq \frac{1}{2} p_Y(1-p_X)$ [L'estimateur par la moyenne est simplement la proportion des unités échantillonnales pour lesquelles $Y = 1$]. Le terme à gauche ($p_{XY} - p_X p_Y$) est un indice de la dépendance entre X et Y.

6.14 [*Estimateur par le produit*] Utiliser la méthode de linéarisation pour montrer que $V(\bar{x}\,\bar{y}) \approx \mu_x^2\mu_y^2\left[\frac{V(\bar{x})}{\mu_x^2} + \frac{V(\bar{y})}{\mu_y^2} + 2\frac{C(\bar{x};\bar{y})}{\mu_x\mu_y}\right]$. Utiliser cette formule approximative pour montrer que la variance de l'estimateur de μ_y par le produit, $\hat{\mu}_y = \frac{\bar{x}\,\bar{y}}{\mu_x}$ est inférieure à celle de $\bar{y}$ si et seulement si $\rho > -\frac{1}{2}\frac{cv(x)}{cv(y)}$, où ρ est le coefficient de corrélation entre X et Y et $cv(.)$ désigne le coefficient de variation.

7 Stratification

L'échantillonnage aléatoire simple — seul mode d'échantillonnage considéré jusqu'ici — n'en est qu'un parmi d'autres. En pratique, il est souvent préférable, voire nécessaire, de recourir à d'autres modes de tirage, généralement plus complexes. D'abord parce que le tirage aléatoire simple (*éas*) n'est pas toujours facile à exécuter. Malgré sa simplicité, la base de sondage qu'il exige — une simple liste des unités de la population — n'est pas toujours disponible.

Un *éas* est problématique aussi lorsqu'on souhaite estimer les paramètres d'un ou de plusieurs *domaines*, en plus de ceux de la population ciblée. L'estimation de la moyenne d'un domaine, telle que discutée à la section 5.2, est basée sur un échantillon dont la taille est aléatoire, une taille qui peut se révéler insuffisante. Supposons, par exemple, qu'on prélève un échantillon des réclamations soumises à une compagnie d'assurance automobile afin d'estimer une certaine moyenne ou une proportion. On s'assurera que l'échantillon soit assez grand pour donner des estimations assez précises. Mais si du même coup, on tient à estimer les moyennes ou proportions relatives à certains groupes d'assurés — les femmes, les jeunes, les accidentés fréquents — on doit s'assurer que ces groupes soient bien représentés dans l'échantillon. Leur nombre ne peut être laissé au hasard. Un mode d'échantillonnage qui contrôle le nombre d'observations dans ces catégories n'est plus un *éas*.

Finalement, le tirage aléatoire simple peut être écarté pour des raisons d'efficacité: comme nous le verrons dans ce chapitre, un *éas* peut être moins efficace que certains autres plans d'échantillonnage.

L'un de ces plans, appelé *échantillonnage par strates*, est un mode de tirage parfois matériellement plus facile à exécuter et peut en plus être plus efficace. Il consiste à répartir la population en sous-populations disjointes, puis de tirer un échantillon dans chacune des sous-populations. Cette approche vient spontanément à l'esprit quand la population est naturellement constituée de sous-populations disjointes (les facultés d'une université, les arrondissements d'une ville, les catégories d'employés, etc.), et d'un point de vue administratif, c'est une façon efficace de gérer le processus d'échantillonnage.

On considère donc une population partitionnée en sous-populations disjointes, appelées *strates*; et on tire indépendamment un échantillon dans chacune des strates. Cette façon de procéder entraîne de nouveaux estimateurs, ceux définis dans les chapitres précédents n'étant généralement plus valables. La moyenne échantillonnale $\bar{y}$, par exemple, proposée

comme estimateur de la moyenne μ dans un *éas*, peut être sérieusement biaisée dans un échantillon stratifié. Nous allons donc définir des estimateurs propres à l'échantillonnage stratifié, et nous établirons les formules appropriées de variance et d'intervalle de confiance.

En un premier temps, nous nous limiterons au cas où l'échantillon tiré dans chaque strate est un *éas*. Cette restriction n'est pas incontournable et nous l'éliminerons en un deuxième temps.

7.1 ESTIMATION D'UNE MOYENNE OU D'UN TOTAL

Nous introduisons d'abord une notation propre à l'échantillonnage par strates. Elle est calquée sur celle utilisée jusqu'ici.

Notation

Considérons une population répartie en L strates.

$$U = U_1 \cup U_2 \cup \ldots \cup U_L$$

et un échantillon

$$\omega = \omega_1 \cup \omega_2 \, \omega \ldots \cup \omega_L \, .$$

U_h est l'ensemble des unités de la strate h, et ω_h est l'échantillon tiré dans la strate h. Soit Y une variable définie sur U. Les paramètres relatifs à la population sont définis dans le tableau suivant :

N_h : Taille de U_h; $N = N_1 + \ldots + N_L$	$S_h = \sqrt{\frac{\sum_{k=1}^{N_h}(y_{hk} - \mu_h)^2}{N_h - 1}}$ = Écart-type de Y dans U_h
$W_h = N_h / N$	$\tau_h = N_h \mu_h = \sum_{h=1}^{L} \sum_{k=1}^{N_h} y_{hk}$ = le total de U_h
y_{hk} = La k^e valeur de Y dans la strate U_h	$\tau = N\mu = \sum_{h=1}^{L} \tau_h$: le total de la population
$\mu_h = \frac{1}{N} \sum_{k=1}^{N_h} y_{hk}$ = moyenne de Y dans U_h	$\mu = \Sigma_h W_h \mu_h$: moyenne de la population

On tire un *éas* dans chacune des strates. Le tableau suivant définit les statistiques échantillonnales et résume les propriétés de l'estimateur de la moyenne dans chacune des strates. Ces propriétés sont connues, puisque dans chaque strate le tirage est un *éas*. Seule la notation change ici : les statistiques et les paramètres seront affectés d'un indice (h) désignant la strate.

n_h = La taille de ω_h; $n = n_1 + \ldots + n_L$	*Variance de* $\bar{y}_h$ $V(\bar{y}_h) = (1-f_h)\frac{S_h^2}{n_h}$, $f_h = \frac{n_h}{N_h}$
$\bar{y}_h = \frac{1}{n_h}\sum_{j\in\omega_h} y_{hj}$ = Moyenne de Y dans ω_h	*Estimateur de* $V(\bar{y}_h)$ $V(\bar{y}_h) = (1-f_h)\frac{s_h^2}{n_h}$, $f_h = \frac{n_h}{N_h}$
$s_h = \sqrt{\frac{\sum_{i\in\omega_h}(y_{hi}-\bar{y}_h)^2}{n_h-1}}$ = Écart-type de Y dans ω_h	

L'objectif maintenant est de réunir l'information contenue dans les strates en une estimation de la moyenne globale μ et du total τ.

Estimateur de μ

La moyenne μ est une fonction linéaire des μ_h :

$$\mu = \frac{\tau}{N} = \frac{\sum_{h=1}^{L}\tau_h}{N} = \frac{\sum_{h=1}^{L}N_h\mu_h}{N} = \sum_{h=1}^{L}W_h\mu_h \ .$$

On estimera donc μ en remplaçant chaque μ_h par son estimateur $\bar{y}_h$. On désignera l'estimateur de μ par $\bar{y}_{st}$:

(7.1) $$\bar{y}_{st} = \sum_{h=1}^{L}W_h\bar{y}_h \ .$$

Variance de l'estimateur

Les estimateurs $\bar{y}_h$ sont des variables indépendantes : l'échantillon obtenu dans une strate donnée ne dépend pas des résultats obtenus dans les autres strates.

La variance de la somme $V(\sum_{h=1}^{L}W_h\bar{y}_h)$ est donc égale à la somme des variances : $V(\bar{y}_{st}) = V(\sum_{h=1}^{L}W_h\bar{y}_h) = \sum_{h=1}^{L}V(W_h\bar{y}_h) = \sum_{h=1}^{L}W_h^2V(\bar{y}_h)$. On a donc

(7.2) $$V(\bar{y}_{st}) = \sum_{h=1}^{L}W_h^2V(\bar{y}_h) = \sum_{h=1}^{L}W_h^2(1-f_h)\frac{S_h^2}{n_h} \ .$$

Estimation de la variance

On obtient un estimateur sans biais de $V(\bar{y}_{st})$ en remplaçant chaque S_h^2 par l'équivalent échantillonnal s_h^2 :

(7.3) $$V(\bar{y}_{st}) = \sum_{h=1}^{L}W_h^2(1-f_h)\frac{s_h^2}{n_h} \ .$$

Intervalle de confiance pour μ

On détermine un intervalle de confiance de niveau 100(1-α) % pour μ par la formule habituelle:

(7.4) $$\bar{y}_{st} - z_{\alpha/2}\sqrt{\hat{V}(\bar{y}_{st})} \leq \mu \leq \bar{y}_{st} + z_{\alpha/2}\sqrt{\hat{V}(\bar{y}_{st})} .$$

L'exemple suivant illustre tous les calculs.

Exemple 7.1 *Estimation et intervalle de confiance - échantillon stratifié*

D'une population répartie en trois strates de tailles 10, 25 et 50, on prélève un échantillon stratifié de 2 observations de la strate 1, 6 de la strate 2 et 12 de la strate 3. On obtient les données suivantes :

Strate 1	30215	95 684				
Strate 2	5656	6532	6521	6582	8457	9856
Strate 3	214	659	654	658	123	452
	325	445	323	765	325	235

Déterminer

a) une estimation de la moyenne de la population.

b) une estimation du total de la population.

c) une estimation de $\sqrt{V(\bar{y}_{st})}$.

d) une estimation de $\sqrt{V(\hat{\tau}_{st})}$.

e) un intervalle de confiance à 95% pour la moyenne de la population.

f) un intervalle de confiance à 95 % pour τ_y.

Solution.

[Les valeurs des calculs intermédiaires présentées sont moins précises que celles qui ont été retenues pour les calculs subséquents. On peut s'attendre à quelques différences dans les résultats finaux dues à des erreurs d'arrondi].

$\bar{y}_1$ = 62 949,50 s_1 = 46 293,57 ;

$\bar{y}_2$ = 7 267,33 s_2 = 1 567,18 ;

$\bar{y}_3$ = 431,50 s_3 = 208,88.

a) L'estimation de la moyenne est

$$\bar{y}_{st} = W_1\bar{y}_1 + W_2\bar{y}_2 + W_3\bar{y}_3 = \frac{10}{85}(62949{,}5) + \frac{25}{85}(7267{,}33) + \frac{50}{85}(431{,}50) = 9\,797{,}10.$$

b) Si on estime la moyenne à $\bar{y}_{st}$ = 9 797,10, alors une estimation du total est $\hat{\tau} = N\bar{y}_{st}$ = 85(9 797) = 832 753.

c) Les écarts-types estimés des trois moyennes échantillonnales sont

$$\sqrt{\hat{V}(\bar{y}_1)} = \sqrt{1-f_1}\,\frac{s_1}{\sqrt{n_1}} = \sqrt{1-\frac{2}{10}}\,\frac{46294}{\sqrt{2}} = 29279,$$

$$\sqrt{\hat{V}(\bar{y}_2)} = \sqrt{1-f_2}\,\frac{s_2}{\sqrt{n_2}} = \sqrt{1-\frac{6}{25}}\,\frac{1567{,}18}{\sqrt{6}} = 557{,}76, \text{ et}$$

$$\sqrt{\hat{V}(\bar{y}_3)} = \sqrt{1-f_3}\,\frac{s_3}{\sqrt{n_3}} = \sqrt{1-\frac{12}{50}}\,\frac{208{,}880}{\sqrt{12}} = 52{,}567.$$

La variance de l'estimateur de la moyenne est estimée à 11 892 754, et l'écart-type estimé par $\sqrt{\hat{V}(\bar{y}_{st})} = \sqrt{11\,892\,754} = 3\,448{,}59$.

d) $\sqrt{\hat{V}(\hat{\tau}_{st})} = N\sqrt{\hat{V}(\bar{y}_{st})} = 85(3448{,}59) = 293\,130.$

e) Un intervalle de confiance approximatif à 95% pour μ est donné par

$\bar{y}_{st} - 1{,}96\sqrt{\hat{V}(\bar{y}_{st})} \le \mu \le \bar{y}_{st} + 1{,}96\sqrt{\hat{V}(\bar{y}_{st})} = 9\,797{,}10 - 1{,}96\,(3\,448{,}59) \le \mu \le 9\,797{,}10 + 1{,}96\,(3\,448{,}59) \Rightarrow 9\,797{,}10 - 6\,759 \le \mu \le 9\,797{,}10 + 6\,759 \Rightarrow 3\,038 \le \mu \le 16\,556.$

f) L'intervalle de confiance à 95 % pour τ s'obtient directement en multipliant par N l'intervalle de confiance pour μ :

$$\hat{\tau} - 1{,}96\sqrt{\hat{V}(\hat{\tau}_{st})} \le \tau \le \hat{\tau} + 1{,}96\sqrt{\hat{V}(\hat{\tau}_{st})} \Rightarrow N[\bar{y}_{st} - 1{,}96\sqrt{\hat{V}(\bar{y}_{st})}] \le \tau \le N[\bar{y}_{st} + 1{,}96\sqrt{\hat{V}(\bar{y}_{st})}]$$

$\Rightarrow 85(3\,038) \le \tau \le 85(16\,556) \Rightarrow 25\,8229 \le \tau \le 1\,407\,277$

Un deuxième exemple présente des données qui seront utilisées fréquemment dans la suite.

Exemple 7.2 *Échantillon stratifié*

Le Tableau 7.1 présente des données d'un échantillon stratifié d'une population (tableau A05) de $N = 1\,160$ municipalités. Y est le nombre de familles ayant trois enfants ou plus.

La stratification de la population est basée sur le nombre d'habitants en 2011, connu par recensement pour la population entière. Les strates sont de tailles 468, 344, 232, et 116. Les échantillons sont de tailles 20, 15, 10, et 5, respectivement.

Tableau 7.1
Échantillon stratifié d'une population de 1160 municipalités québécoises

	Strate (h)			
	h = 1	h = 2	h = 3	h = 4
	5 5 5 5 5 10 10 10 10 10 15 15 20 20 20 20 25 25 30 30	10 15 25 25 30 35 35 40 40 45 45 50 55 65 85	70 70 75 80 90 105 110 145 155 170	145 280 495 500 870
Intervalle des valeurs de X	]50 ; 890]	]890 ; 2275]	]2275 ; 7645]	> 7645
Taille de la strate (N_h)	468	344	232	116
Taille relative de la strate ($W_h = \frac{N_h}{N}$)	0,4034	0,2966	0,2	0,1
Taille de l'échantillon (n_h)	20	15	10	5
Moyenne de l'échantillon ($\bar{y}_h$)	14,75	40,00	107,00	458,00
Écart-type de l'échantillon (s_h)	8,503095	19,179602	37,282704	275,013636
Écart-type estimé de l'estimateur : $\sqrt{\hat{V}(\bar{y}_h)} = \sqrt{1-f_h}\,\frac{s_h}{\sqrt{n_h}}$, $f_h = \frac{n_h}{N_h}$	1,860279	4,842980	11,532936	120,309998

La moyenne μ est estimée par $\bar{y}_{st}$: $\bar{y}_{st}$ = 85,01.

$$\hat{V}(\bar{y}_{st}) = 152{,}69;\ \sqrt{\hat{V}(\bar{y}_{st})} = 12{,}35683.$$

Intervalle de confiance à 95 % :

$$60{,}79 \le \mu \le 109{,}23.$$

Remarque *Effet de l'asymétrie de la population*

Il faut se rappeler que si les estimateurs s_h^2 et $\hat{V}(\bar{y}_{st})$ sont sans biais, leur valeur dans un cas particulier peut être assez éloignée de leur paramètre. Dans l'exemple actuel, étant donné que la population est connue, nous pouvons déterminer les vraies valeurs de ces paramètres.

	Strate 1	Strate 2	Strate 3	Strate 4
S_h	11,0	24,3	52,0	3791,2
s_h	8,5	19,2	37,3	275,0

$\sqrt{V(\bar{y}_{st})}$ =165,9.

On remarque que les écarts-types sont sous-estimées dans toutes les strates (gravement dans la quatrième). Ceci est dû à l'extrême asymétrie des données des strates, chacune ayant un petit nombre de valeurs extrêmes. Ces valeurs extrêmes font augmenter la dispersion, mais sont rarement présentes dans l'échantillon. La plupart des échantillons présentent donc une faible dispersion.

Il faut remarquer aussi que nous n'avons pas tiré grand profit de la stratification ici, puisque l'écart-type $\sqrt{V(\bar{y}_{st})}$ =165,9 n'est pas tellement plus petit que $\sqrt{V(\bar{y})}$ = 172,9, celui de la moyenne dans un *éas*. Il ne suffit donc pas de stratifier.

La manière dont les strates sont constituées joue un rôle important dans la précision de l'estimateur. Une stratification conçue en fonction des valeurs d'une variable auxiliaire *X* (le nombre d'habitants), est généralement avantageuse lorsque *X* est corrélée avec la variable d'intérêt *Y*. On peut alors, comme dans le dernier exemple, grouper dans une même strate les unités dont les valeurs de *X* sont rapprochées les unes des autres.

Mais le choix des valeurs de *X* qui délimitent les strates est également important. Il est difficile de proposer des techniques générales pour le faire, mais on peut explorer diverses possibilités et les évaluer selon leur efficacité dans l'« estimation » de μ_x, dont la valeur est connue.

Dans le prochain exemple, on propose une stratification inspirée de telles explorations.

Dans l'Exemple 7.2, la stratification aurait pu être mieux faite, puisque l'écart-type de la strate 4 est démesurément élevé. Le prochain exemple montre qu'on peut améliorer la précision de l'estimateur en réduisant la taille de la strate 4.

Exemple 7.3 *Effet d'une différente stratification*

Le tableau suivant montre les effets d'une nouvelle stratification pour la population décrite dans l'Exemple 7.2. Les données sont celles de la population elle-même.

	Strate (définie par les intervalles des valeurs de *X*)			
	1	2	3	4
	$0 < x \leq 2130$	$2130 < x \leq 8445$	$8445 < x \leq 25\,350$	$x > 25\,350$
N_h	798	251	70	41
n_h	20	15	10	5
μ_h	24,58	99,08	364,5	2919,88
S_h	20,67	54,58	168,09	6075,37
$\sqrt{V(\bar{y}_h)}$	4,56	13,66	49,21	2545,93

L'écart-type de l'estimateur est maintenant $\sqrt{V(\bar{y}_{st})}$ = 90,14, nettement inférieur à l'écart-type de 165,9 obtenu avec la stratification précédente, ceci en dépit du fait que la dispersion dans chacune des strates est supérieure ici. Qu'est-ce qui a permis une telle amélioration? La variance $V(\bar{y}_4)$ de $\bar{y}_4$, qui est énorme, contribue fortement à la variance de l'estimateur. Pour atténuer son effet, nous avons réduit son poids, W_4^2, en réduisant la taille de la strate.

On pourrait d'ailleurs pousser cette idée plus loin et réduire la strate 4 à une seule unité, Montréal (et majorant les tailles des autres échantillons de façon à maintenir une taille totale de 50). Le lecteur peut vérifier que si toutes les données de la strate 4 sauf Montréal sont absorbées par la strate 3 (et si les échantillons des 3 premières strates sont de tailles

22, 16 et 11), l'écart-type de l'estimateur est réduit encore plus, à 43,60. Il existe même une stratification qui donne $\sqrt{V(\bar{y}_{st})}$ = 12,4.

DÉCOMPOSITION DE LA DISPERSION TOTALE

Les propriétés d'un échantillon stratifié s'expliquent bien en termes de la dispersion des données de la population, que l'on peut décomposer en deux parties significatives. On peut démontrer ceci :

THÉORÈME 7.1 **DÉCOMPOSITIONS DE LA VARIANCE TOTALE**

(7.5) $$(N\text{-}1)S^2 = \sum_{h=1}^{L}(N_h - 1)S_h^2 + \sum_{h=1}^{L} N_h(\mu_h - \mu)^2 \ .$$

Cette décomposition porte le nom *d'analyse de variance*. Les trois sommes de carrés sont désignées par, respectivement, SCT, SCI et SCE :

- La première, SCT = $(N\text{-}1)S^2$, appelée *Somme des carrés totale*, représente la dispersion de toutes les données de la population par rapport à la moyenne de la population;
- La deuxième, SCI = $\sum_{h=1}^{L}(N_h - 1)S_h^2$, que nous appellerons *Somme des carrés interne*, est la dispersion des données à *l'intérieur* des strates, puisqu'elle est fonction des S_h.
- La troisième, SCE= $\sum_{h=1}^{L} N_h(\mu_h - \mu)^2$, la *Somme des carrés externe*, mesure les écarts *entre* les strates.

La décomposition s'écrit

$$\text{SCT} = \text{SCI} + \text{SCE}.$$

Dans un *éas*, la variance de l'estimateur est fonction de *S*, donc de SCT, alors que dans un échantillon stratifié, elle dépend des S_h. Par conséquent, pour que la stratification soit avantageuse, il faut que SCI soit faible, et SCE conséquemment élevée. En d'autres termes, il est souhaitable que les strates soient différentes les unes des autres et que les unités d'une même strate soient aussi semblables que possible.

7.2 ALLOCATION DES OBSERVATIONS

Supposons que la taille totale *n* d'un échantillon stratifié a été fixée. Comment répartir ces *n* observations entre les strates ? Il semble raisonnable de maintenir une certaine proportionnalité entre les tailles des échantillons et les tailles des strates: les grandes strates devant être plus fortement représentées dans l'échantillon. Cette répartition, appelée *allocation proportionnelle*, est très répandue et elle a du mérite. Nous la discutons en premier. Nous verrons ensuite qu'il est parfois possible de faire mieux, en fait de répartir les observations de la meilleure façon possible (une notion que nous prendrons soin de définir convenablement). Cette répartition est appelée *allocation optimale*. Commençons par l'allocation proportionnelle.

ALLOCATION PROPORTIONNELLE

L'*allocation proportionnelle* consiste à répartir l'échantillon de la même façon que la population. C'est-à-dire, les n_h sont tels que $n_h/n = N_h/N = W_h$, ou encore

(7.6) $$n_h = nW_h.$$

Dans l'Exemple 7.2 les tailles des 4 strates sont 464, 348, 232, 116; et les tailles relatives sont $W_1 = 0{,}4034$; $W_2 = 0{,}2966$; $W_3 = 0{,}2$; $W_4 = 0{,}1$. Supposons qu'on veuille prélever un échantillon de taille $n = 50$, avec allocation proportionnelle. On répartira les $n = 50$ unités comme suit :

$n_1 = 50\,W_1 = 20$; $n_2 = 50\,W_2 = 15$; $n_3 = 50\,W_3 = 10$; $n_4 = 50\,W_4 = 5$.

Formule de la variance de $\bar{y}_{st}$ — allocation proportionnelle.

THÉORÈME 7.2 **VARIANCE, ALLOCATION PROPORTIONNELLE**

La variance de $\bar{y}_{st}$ dans une allocation proportionnelle est à peu près

(7.7) $$V_{prop} = \frac{(1-f)}{n}\sum_{h=1}^{L} W_h S_h^2 \text{ , où } f = n/N.$$

ALLOCATION OPTIMALE

Bien qu'il soit naturel de répartir les n_h de façon proportionnelle aux N_h (ou aux W_h), ce n'est pas la meilleure allocation possible. L'allocation des effectifs a un effet important sur la précision d'un estimateur. Considérons, par exemple, la population présentée au tableau A05, une population de $N = 1\,160$ municipalités québécoises. Supposons que pour estimer μ_y, le nombre moyen de naissances, on projette de tirer un échantillon de taille $n = 50$. On établit des strates dont les propriétés sont les suivantes :

	Strate 1	Strate 2	Strate 3	Strate 4
N_h	420	563	134	43
S_h	9,8	34,8	168,5	5951

Nous allons considérer différentes allocations possibles et calculer l'écart-type de l'estimateur pour chaque répartition, en supposant encore, pour les besoins de ces calculs, que la population est entièrement connue. Les formules sont

$$V(\bar{y}_{st}) = \sum_{h=1}^{4} W_h^2 V(\bar{y}_h) \text{ , où } V(\bar{y}_h) = (1-f_h)\frac{S_h^2}{n_h}, f_h = \frac{n_h}{N_h}$$

Allocation	n_h	$\sqrt{V(\bar{y}_{st})}$
1. Proportionnelle	$n_1 = 18$ $n_2 = 24$ $n_3 = 6$ $n_4 = 2$	153
2. À peu près uniforme	$n_1 = 12$ $n_2 = 13$ $n_3 = 12$ $n_4 = 13$	52
3. Mauvais choix	$n_1 = 38$ $n_2 = 6$ $n_3 = 5$ $n_4 = 1$	218
4. Optimale	$n_1 = 1$ $n_2 = 3$ $n_3 = 4$ $n_4 = 42$	15
	éas, $n = 50$	173

On remarque que, dépendant de l'allocation, l'échantillonnage stratifié peut être plus efficace ou moins efficace qu'un *éas*. La stratification avec allocation proportionnelle (écart-type : 153) offre généralement de meilleurs résultats que le tirage aléatoire simple (173). Mais il est possible de faire mieux. La deuxième allocation dans le tableau, à peu près uniforme, est meilleure (52). La troisième (218) montre que la stratification ne garantit pas la qualité et peut donner de très mauvais résultats. Finalement, la quatrième, avec un écart-type de 15, a le plus petit écart-type. C'est l'allocation dite *optimale*.

DÉFINITION 7.1 ALLOCATION OPTIMALE

L'allocation optimale est l'allocation

i) qui minimise l'écart-type de l'estimateur pour une taille n fixée; ou, ce qui est équivalent,

ii) qui minimise la taille de l'échantillon pour un écart-type donné.

Dire que l'allocation $n_1 = 1$, $n_2 = 3$, $n_3 = 4$, $n_4 = 42$ est optimale, c'est dire qu'il n'y a aucune autre allocation (d'un échantillon de taille 50) qui donne un écart-type aussi petit que 15. C'est aussi dire qu'un écart-type de 15 ne peut être réalisé avec un échantillon plus petit que 50.

THÉORÈME 7.3 **ALLOCATION OPTIMALE**

Les valeurs de $n_1,\ldots, n_L$ qui minimisent la variance $V(\bar{y}_{st})$ pour une taille n donnée sont les valeurs proportionnelles aux $W_h S_h$, c'est-à-dire,

$$n_{opt} = \frac{N_h S_h}{\sum_{i=1}^{L} N_i S_i} n \,. \tag{7.8}$$

On peut voir dans cette formule ce qui fait qu'une strate exige un échantillon plus grand ou plus petit. La taille N_h joue évidemment un rôle : plus la strate est grande, plus il faut y puiser des observations. C'est le principe d'une allocation proportionnelle. Mais contrairement à celle-là, l'allocation optimale tient compte aussi de l'écart-type S_h : plus les données de la strate sont dispersées, plus il faut y prélever d'observations.

Exemple 7.4 *Calcul d'une allocation optimale*

C'est la formule que nous avons utilisée ci-dessus. Voici les calculs :

	Strate 1	Strate 2	Strate 3	Strate 4	Somme
N_h	420	563	134	43	1160
S_h	9,8259	34,7908	160,4636	5951,2908	
W_hS_h	3,5576	16,8855	18,5363	220,6082	259,5877
$\frac{W_hS_h}{\sum W_iS_i} \times 50$	0,69	3,25	3,57	42,49	50
n_h	1	3	4	42	50

On constate que l'allocation optimale est loin d'être proportionnelle : le plus grand échantillon est alloué à la plus petite strate, la strate 4, alors que sont alloués des échantillons minuscules aux strates 1, 2 et 3, pourtant plus grandes. C'est que la dispersion dans ces dernières est faible comparée à celle de la strate 4. C'est justement de là que la stratification tient son efficacité : un choix judicieux alloue des observations là où on en a le plus besoin. En pratique, on ne s'en tient pas nécessairement à l'allocation optimale telle que calculée. Les tailles des échantillons dans les trois premières strates sont beaucoup trop petites: on ne peut pas compter sur la normalité des moyennes avec des échantillons si petits. En plus, un échantillon de taille 1 (strate 1) n'est pas acceptable car l'écart-type s_1 n'est pas calculable.

OBSERVATIONS AUTO-PONDÉRÉS

L'estimateur du total dans un échantillon stratifié est une moyenne pondérée de la forme $\hat{\tau} = \sum_{h=1}^{L} \sum_{j \in \omega_h} \alpha_{hj} y_{hj} = \sum_{h=1}^{L} \sum_{j \in \omega_j} \frac{y_{hj}}{n_h / N_h}$. Le coefficient de y_{hj} est, comme toujours, la réciproque de la probabilité de sélection de y_{hj} (la j^e observation de l'échantillon ω_h). Un échantillon est dit **auto-pondéré** si les coefficients α_{hj} sont égaux, donc si $n_h = nW_h$. Dans un échantillon stratifié, un échantillon auto-pondéré est un échantillon avec allocation proportionnelle.

STRATES RECENSÉES

Il peut arriver que la formule pour l'allocation optimale donne pour certaines strates une valeur de n_h supérieure à N_h— ce qui est impossible si les tirages se font sans remise. Dans ce cas, on prélève toutes les unités des strates en question, et on utilise l'allocation optimale pour les autres strates.

Exemple 7.5 *Strate recensée*

Soit une population de taille $N = 515$ unités réparties en trois strates de tailles 20, 95 et 400. Afin d'estimer la moyenne μ d'une variable Y, on décide de tirer un échantillon de taille 40. On estime les écarts-types des strates à 13 250, 800 et 150. Déterminer l'allocation optimale pour un échantillon de taille 40 et estimer l'écart-type de $\bar{y}_{st}$.

Solution

Les valeurs arrondies de $\frac{N_h S_h}{\sum_{i=1}^{L} N_i S_i}(40)$ sont 26 ; 8 et 6. On ne peut pas tirer un échantillon de 26 unités de la strate 1, qui n'en contient que 20. Donc on recense entièrement la strate 1 et on répartit les 20 autres, $\frac{N_2 S_2}{N_2 S_2 + N_3 S_3}$ = 11,18 et $\frac{N_3 S_3}{N_2 S_2 + N_3 S_3}$ = 8,82. Donc la répartition optimale est n_1 = 20; n_2 = 11 et n_3 = 9. Les écarts-types $V(\bar{y}_h)$ sont 0 ; 226,8148 ; et 49,4343. L'écart-type de $\bar{y}_{st}$ est alors estimé à 56,79.

Remarque *Difficulté : Paramètres inconnus*

Pour déterminer l'allocation proportionnelle, il suffit de connaître les W_h (s'ils ne sont pas connus, et si on ne peut pas les estimer avec précision, il n'est pas question de faire de l'échantillonnage stratifié : l'estimateur lui-même dépend des W_h). Pour déterminer l'allocation optimale, par contre, il faut aussi connaître les S_h. Or les S_h sont eux aussi inconnus, et ne peuvent être estimés qu'une fois l'échantillon tiré.

Cependant, il n'est pas nécessaire d'estimer les S_h de façon très précise pour appliquer les techniques discutées ici. Une approximation assez grossière devrait suffire et en pratique, cela est souvent possible. On peut obtenir une estimation des S_h soit par un échantillon pilote conçu à cette fin, soit en se basant sur des données semblables dans d'autres populations, soit encore en se fiant aux S_h, connus ou estimés, d'une *autre* variable, corrélée avec la variable d'intérêt.

EFFICACITÉ DE LA STRATIFICATION

Dans un *éas*, la variance de l'estimateur est fonction de S, l'écart-type de la population entière, alors que dans un échantillon stratifié elle est fonction des écarts-types S_h des strates. C'est ce qui fait qu'une stratification judicieuse peut donner des estimations plus précises qu'un *éas* de même taille. Une stratification est particulièrement efficace si les données d'une même strate sont homogènes (les S_h sont petits), ou encore, ce qui revient au même, si les moyennes μ_h des strates sont très différentes les unes des autres. L'exemple suivant montre l'effet d'une stratification dans un tirage avec allocation proportionnelle.

Exemple 7.6 *Efficacité de la stratification*

Voici les données d'une population de 45 unités :

20	45	55	65	90	100	125	135	145	170	180	205	215	225	250
25	40	60	70	85	105	120	140	150	165	185	200	220	230	245
30	35	50	75	80	110	115	130	155	160	190	195	210	235	240

La moyenne de cette population est μ = 135 et l'écart-type est S = 69,57664. Si on tire un *éas* de taille n = 12, l'écart-type de l'estimateur $\bar{y}$ est $\sqrt{V(\bar{y})}$ = 17,20.

Supposons qu'on répartit la population en trois strates de taille 15, dans l'ordre présenté, et qu'on tire un échantillon de taille 4 dans chacune des strates. Voici les données :

Strate (h)	Données	μ_h	S_h
1	20 45 55 65 90 100 125 135 145 170 180 205 215 225 250.	135	71,7
2	25 40 60 70 85 105 120 140 150 165 185 200 220 230 245.	136	71,1
3	30 35 50 75 80 110 115 130 155 160 190 195 210 235 240.	134	70,8

Le lecteur peut vérifier que la stratification n'apporte ici aucun gain de précision : l'écart-type de $\bar{y}_{st}$ est $\sqrt{V(\bar{y}_{st})}$ = 17,60, quasiment identique à celui de $\bar{y}$, $\sqrt{V(\bar{y})}$ = 17,60. La raison est que les moyennes et les écarts-types des strates sont à peu près égales.

Considérons maintenant une stratification qui rassemble les petites valeurs dans une même strate ; les valeurs moyennes dans une deuxième; et les valeurs élevées dans une troisième.

Strate (h)	Données	μ_h	S_h
1	20 25 30 35 40 45 50 55 60 65 70 75 80 85 90	55	22,36
2	100 105 110 115 120 125 130 135 140 145 150 155 160 165 170	135	22,36
3	180 185 190 195 200 205 210 215 220 225 230 235 240 245 250	215	22,36

L'avantage de la stratification est très net ici : l'écart-type de l'estimateur passe de $\sqrt{V(\bar{y}_{st})}$ = 17,60 à $\sqrt{V(\bar{y}_{st})}$ = 5,53.

La règle générale (qui découle du Théorème 7.1) est celle-ci : quand une allocation proportionnelle est prévue, la stratification est d'autant plus efficace que les *moyennes des strates* sont très différentes. Si on compte recourir à une allocation optimale; l'estimateur s'améliore encore, cette fois-ci en fonction des différences entre les écarts-types S_h des strates : plus les S_h sont différents les uns des autres, plus l'allocation optimale est avantageuse. Si les S_h sont égaux, l'allocation proportionnelle est optimale.

FORMULES DES VARIANCES

Formule de la variance de $\bar{y}_{st}$ — allocation optimale

THÉORÈME 7.4 **VARIANCE, ALLOCATION OPTIMALE**

La variance de $\bar{y}_{st}$ dans un échantillon stratifié avec allocation optimale est égale à

(7.9) $$V_{opt} = \frac{1-f}{n}\sum_{h-1}^{L} W_h S_h^2 - \frac{1}{n}\sum_{h=1}^{L} W_h (S_h - \bar{S})^2 ,$$

où

$$\bar{S} = \sum_{h=1}^{L} W_h S_h \quad \text{et } f = n/N.$$

THÉORÈME 7.5 **VARIANCE DE $\bar{y}$ EN FONCTION DES PARAMÈTRES DES STRATES**

La variance de $\bar{y}$ dans un *éas* peut s'écrire comme

$$V_a = \frac{1-f}{n}\left[\sum_{h=1}^{L}\frac{(N_h-1)}{N-1}S_h^2 + \sum_{h=1}^{L}\frac{N_h}{N-1}(\mu_h-\mu)^2\right]. \tag{7.10}$$

COMPARAISON DES VARIANCES

Nous comparons ici les variances sous trois approches: un *éas* de taille n; un échantillon stratifié de même taille avec allocation proportionnelle; et un échantillon stratifié de même taille avec allocation optimale. Pour ce faire, cependant, nous recourons à certaines approximations, valables lorsque les N_h sont grands: on remplacera N_h -1 par N_h et N-1 par N. Ainsi donc, les variances V_a, V_{prop} et V_{opt} correspondant aux trois approches sont à peu près celles-ci :

$$V_a \approx \frac{1-f}{n}\sum_{h=1}^{L} W_h S_h^2 + \frac{1-f}{n}\sum_{h=1}^{L} W_h(\mu_h-\mu)^2 \tag{7.11}$$

$$V_{prop} = \frac{(1-f)}{n}\sum_{h=1}^{L} W_h S_h^2 \tag{7.12}$$

$$V_{opt} = \frac{1-f}{n}\sum_{h-1}^{L} W_h S_h^2 - \frac{1}{n}\sum_{h=1}^{L} W_h(S_h-\bar{S})^2 . \tag{7.13}$$

Une comparaison de ces formules permet d'évaluer les gains réalisés par la stratification (et allocation proportionnelle) ainsi que les gains réalisés par l'allocation optimale par rapport à l'allocation proportionnelle.

$$V_a - V_{prop} \approx \frac{1-f}{n}\sum_{h=1}^{L} W_h(\mu_h-\mu)^2$$

$$V_{prop} - V_{opt} \approx \frac{1}{n}\sum_{h=1}^{L} W_h(S_h-\bar{S})^2 .$$

Donc la stratification avec allocation proportionnelle apporte des gains par rapport à l'échantillonnage simple dans la mesure où il y a de grands écarts entre les moyennes des strates. L'allocation optimale rapporte des gains supplémentaires dans la mesure où les écarts entre les *écarts-types* des strates sont importants.

Remarque *Les formules de V_{prop} et V_{opt} sont approximatives*

Les formules de V_{prop} et V_{opt} sont approximatives car elles sont développées en posant $n_h = nW_h$ dans le premier cas et $n_h = \dfrac{N_h S_h}{\sum_{i=1}^{L} N_i S_i}$ dans le deuxième, alors que les n_h doivent être arrondis à l'entier le plus proche.

DÉTERMINATION DE LA TAILLE DE L'ÉCHANTILLON

Quelle est la taille d'un échantillon qui assure une précision donnée? Supposons que les S_h sont à peu près connus. Comment déterminer la taille de l'échantillon nécessaire pour assurer une précision donnée? L'exigence de précision peut toujours s'exprimer en fonction de la variance de l'estimateur:

$$\text{Exigé : } V(\bar{y}_{st}) = V_0$$

où V_0 est un nombre donné.

La réponse dépend de l'allocation prévue.

Voyons concrètement quelles équations doivent être satisfaites sous les deux allocations considérées dans ce chapitre.

Allocation proportionnelle

La condition à satisfaire est: $\dfrac{(1-f)}{n}\sum_{h=1}^{L} W_h S_h^2 = V_0$. Si on néglige le facteur de correction, la solution approximative est

$$n_0 = \frac{\sum_{h=1}^{L} W_h S_h^2}{V_0}\ .$$

Si N n'est pas trop grand, le correctif usuel, $n = \dfrac{n_0}{1 + n_0 / N}$, permet de réduire quelque peu la taille de l'échantillon .

Allocation optimale

La condition à satisfaire est: $\dfrac{1}{n}\left(\sum_{h=1}^{L} W_h S_h\right)^2 - \dfrac{1}{N}\sum_{h=1}^{L} W_h S_h^2 = V_0$, ce qui entraîne la solution suivante:

$$n = \frac{\left(\sum_{h=1}^{L} W_h S_h\right)^2}{V_0 + \dfrac{1}{N}\sum_{h=1}^{L} W_h S_h^2}\ .$$

Exemple 7.7 *Taille d'échantillon. population stratifiée*

On projette de tirer un échantillon stratifié d'une population de 8 654 factures de restaurant afin d'estimer le montant moyen μ des factures, avec une marge d'erreur (à 95 %) de 5 $. La population est stratifiée selon le jour de la semaine. Afin de se faire une idée de la dispersion des données, on tire un *éas* de 36 factures, on les classe selon la strate, et on estime l'écart-type de chacune des strates. Voici les résultats:

	Strate 1 (Lundi à vendredi)	Strate 2 (Samedi)	Strate 3 (Dimanche
N_h	5 420	1 535	1 699
n_h	24	4	8
s_h	41,79462	33,86159	33,60080

a) Quelle devrait être la taille de l'échantillon si on se propose d'utiliser l'allocation proportionnelle?

b) Quelle devrait être la taille de l'échantillon si on tient à minimiser la taille de l'échantillon?

Solution

a) L'échantillon doit satisfaire l'égalité $1{,}96\sqrt{V(\bar{y}_{st})} = 5 \Rightarrow V(\bar{y}_{st}) = (5/1{,}96)^2$. Posons $V_0 = (5/1{,}96)^2$. La solution approximative est $n_0 = \dfrac{\sum_{h=1}^{L} W_h S_h^2}{V_0}$. Utilisant les s_h à la place des S_h. nous obtenons $n_0 = 233{,}4229$. L'ajustement donne $n = n_0/(1+n_0/N) \approx 227$, avec l'allocation 142, 40, 45.

b) $n = \dfrac{\left(\sum_{h=1}^{L} W_h S_h\right)^2}{\frac{1}{N}\sum_{h=1}^{L} W_h S_h^2 + V_0^2} = 225$. L'allocation optimale est 153, 35, 39.

Remarquer le peu de différence entre l'allocation optimale et l'allocation proportionnelle. Ceci est dû au fait que les S_h sont assez proches les uns des autres.

OPTIMISATION DU COÛT D'ÉCHANTILLONNAGE

L'allocation optimale présentée ci-dessus minimise n pour une variance donnée (ou minimise la variance pour une taille totale donnée n). Mais cette formulation ne tient pas compte du coût d'échantillonnage, qui peut différer d'une strate à l'autre. Une allocation optimale peut minimiser n mais à un coût excessif, si les strates qui requièrent un grand nombre d'observations sont celles où le coût d'échantillonnage ou d'analyse est élevé. Il est plus pertinent de minimiser le coût plutôt que n (ou minimiser la variance pour un coût donné). Supposons que le coût par unité dans la strate h est c_h et donc que le coût total est $B = \sum_{h=1}^{L} c_h n_h$, en plus, possiblement, d'un coût fixe c_0.

Le théorème suivant généralise le Théorème 7.3.

THÉORÈME 7.6 **ALLOCATION OPTIMALE AVEC COÛT**

Soit $c_0 + B$ le coût d'échantillonnage , où $B = \sum_{h=1}^{L} c_h n_h$. L'allocation qui minimise B pour une variance $V(\bar{y}_{st}) = V_0$ donnée , ou qui minimise la variance pour un coût $c_0 + B_0$ donné est donnée par

(7.14) $$n_h = k\frac{W_h S_h}{\sqrt{c_h}} \ .$$

En d'autres termes, les n_h doivent être proportionnelles aux $\frac{W_h S_h}{\sqrt{c_h}}$. Supposons qu'on fixe un budget: on dispose d'un montant $c_0 + B_0$, où B_0 est la partie qui dépend des n_h : $B_0 = \sum_{h=1}^{L} c_h n_h$. Les n_h sont à déterminer. De (7.14), on a $B_0 = \sum_{h=1}^{L} c_h n_h = k\sum_{h=1}^{L} W_h S_h \sqrt{c_h} \Rightarrow$ $k = \frac{B_0}{\sum_{h=1}^{L} W_h S_h \sqrt{c_h}}$, ce qui entraîne que

(7.15) $$n_h = k\frac{W_h S_h}{\sqrt{c_h}} = \frac{B_0}{\sum_{i=1}^{L} W_i S_i \sqrt{c_i}} \frac{W_h S_h}{\sqrt{c_h}} \ .$$

Inversement, si c'est la valeur V_0 de la variance de $\bar{y}_{st}$ qui est fixée,

$$V(\bar{y}_{st}) = \sum_{h=1}^{L} \frac{W_h^2 S_h^2}{n_h} - \sum_{h=1}^{L} \frac{W_h S_h^2}{N} = V_0 \ ,$$

et alors $\sum_{h=1}^{L} \frac{W_h^2 S_h^2}{n_h} - \sum_{h=1}^{L} \frac{W_h S_h^2}{N} = V_0 \Leftrightarrow \sum_{h=1}^{L} \frac{W_h^2 S_h^2 \sqrt{c_h}}{k W_h S_h} - \sum_{h=1}^{L} \frac{W_h S_h^2}{N} = V_0 \Leftrightarrow \frac{1}{k}\sum_{h=1}^{L} W_h S_h \sqrt{c_h} =$ $V_0 + \sum_{h=1}^{L} \frac{W_h S_h^2}{N} \Leftrightarrow k = \frac{\sum_{h=1}^{L} W_h S_h \sqrt{c_h}}{V_0 + \sum_{h=1}^{L} \frac{W_h S_h^2}{N}}$, ce qui entraîne que

(7.16) $$n_h = \frac{\sum_{h=1}^{L} W_h S_h \sqrt{c_h}}{V_0 + \sum_{h=1}^{L} \frac{W_h S_h^2}{N}} \frac{W_h S_h}{\sqrt{c_h}} \ .$$

Exemple 7.8 *Minimisation du coût ou de la variance pour un budget donné*

On projette de tirer un échantillon stratifié d'une population de 8654 factures de restaurant. On peut stratifier la population selon le nombre de convives compris dans la facture. Voici les particularités de la population (les S_h sont estimés):

	1-2 convives	3-4 convives	5+ convives
N_h	3 427	4085	1 142
S_h	15,6526	21,4568	25,2428
c_h	32 $	48 $	68 $
$W_h S_h$	6,198468	10,128397	3,331097
$W_h S_h / \sqrt{c_h}$	1,0957447	1,4619081	0,4039549
$W_h S_h \sqrt{c_h}$	35,06383	70,17159	27,46893

a) On dispose d'un budget de 1 800 $ dont un montant fixe de 300 $ et 1 500 $ affectés à la cueillette et l'analyse des factures. Quelles sont les tailles n_h qui minimisent la variance sous cette contrainte budgétaire? Quelle sera la variance de l'estimateur?

b) On tient à ce que la variance de l'estimateur $\bar{y}_{st}$ soit $V_0 = 7$. Quelles doivent être les n_h et quel sera le coût de l'opération?

Solution

a) B = 1 500. $\sum_{h=1}^{L} W_h S_h \sqrt{c_h}$ = 132,7044, donc $n_h = \dfrac{B}{\sum_{h=1}^{L} W_h S_h \sqrt{c_h}} \dfrac{W_h S_h}{\sqrt{c_h}}$. Les valeurs de n_h sont donc 12, 17 et 5. Avec ces effectifs, la variance est 11,4.

b) $n_h = \dfrac{\sum_{h=1}^{L} W_h S_h \sqrt{c_h}}{V_0 + \sum_{h=1}^{L} \dfrac{W_h S_h^2}{N}} \dfrac{W_h S_h}{\sqrt{c_h}}$. Les valeurs de n_h sont donc 21, 28 et 8. Le coût du projet s'élève à $300 + \sum_{h=1}^{L} c_h n_h$ = 2 860 $.

7.3 ESTIMATION D'UNE PROPORTION

Un échantillonnage par stratification peut également être employé pour estimer une proportion p. Aucune théorie nouvelle n'est réellement nécessaire, puisqu'une proportion est essentiellement une moyenne. La notation du chapitre 5 est adaptée ici aux échantillons stratifiés: la proportion d'unités appartenant à une classe C est désignée par p_h dans U_h et $\hat{p}_h$ dans ω_h. Nous appliquons l'équation (4.1) à chaque strate ce qui donne

(7.17) $$S_h^2 = \frac{N_h p_h (1 - p_h)}{N_h - 1} \quad \text{et} \quad s_h^2 = \frac{n_h \hat{p}_h (1 - \hat{p}_h)}{n_h - 1}.$$

L'estimateur de la proportion p est

$$\hat{p}_{st} = \sum_{h=1}^{L} W_h \hat{p}_h .$$

L'écart-type de $\hat{p}_{st}$ est

$$\sqrt{V(\hat{p}_{st})} = \sqrt{\sum_{h=1}^{L} W_h^2 V(\hat{p}_h)}$$

où, selon l'équation (4.2),

$$V(\hat{p}_{st}) = \left(\frac{N_h - n_h}{N_h - 1}\right) \frac{p_h(1-p_h)}{n_h}.$$

Un estimateur de $\sqrt{V(\hat{p}_{st})}$ est donné par

$$\sqrt{\hat{V}(\hat{p}_{st})} = \sqrt{\sum_{h=1}^{L} W_h^2 \hat{V}(\hat{p}_h)}$$

où, suivant l'équation (4.3),

$$\hat{V}(\hat{p}_h) = (1-f_h)\frac{\hat{p}_h(1-\hat{p}_h)}{n_h - 1}.$$

Un intervalle de confiance approximatif de niveau 100(1-α) % est donné par

$$\hat{p}_{st} - z_{\alpha/2}\sqrt{\hat{V}(\hat{p}_{st})} \le p \le \hat{p}_{st} + z_{\alpha/2}\sqrt{\hat{V}(\hat{p}_{st})}.$$

Exemple 7.9 *Estimation et intervalle de confiance - proportion et effectif*

Pour estimer le nombre N_c d'employés en faveur d'un plan de soins dentaires, on prélève un *éas* dans chacune des 4 divisions de la compagnie. Les effectifs des 4 divisions sont 4523, 3456, 1300 et 1124; et les tailles respectives des échantillons sont 44, 34, 12, et 10. Le nombre de personnes favorables est 20, 10, 6, et 6 dans les quatre échantillons, respectivement.

a) Estimez la proportion p et le nombre N_c d'employés en faveur d'un plan dentaire dans la compagnie ;

b) Estimez les écarts-types des estimateurs évalués en a) ;

c) Déterminez un intervalle de confiance à 95 % pour p et pour N_c.

Solution Les W_h sont $\frac{4523}{10403}$, $\frac{3456}{10403}$, $\frac{1300}{10403}$ et $\frac{1124}{10403}$. Les $\hat{p}_h$ sont 20/44, 10/34, 6/12 et 6/10. Résumé des calculs:

W_h	$\hat{p}_h$	$W_h\hat{p}_h$	$\sqrt{\hat{V}(\hat{p}_h)}$
0,43478	0,45455	0,19763	0,07556
0,33221	0,29412	0,09771	0,07893
0,12496	0,5	0,06248	0,15006
0,10805	0,6	0,06483	0,16257

a) La proportion p est estimée par $\hat{p}_{st} = \frac{4523}{10403} \times \frac{20}{44} + \frac{3456}{10403} \times \frac{10}{34} + \frac{1300}{10403} \times \frac{6}{12} + \frac{1124}{10403} \times \frac{6}{10}$ = 0,42264536.

On estime que 42,26% des employés de la compagnie sont en faveur du plan et donc que le nombre d'employés en faveur du plan est $\hat{N}_c = N\hat{p}_{st}$ = 10 403 × 0,42264536 = 4397.

b) La variance de $\hat{p}_{st}$ est $(0,43478)^2(0,07556)^2 + (0,33221)^2(0,07893)^2 + (0,12496)^2(0,15006)^2 + (0,10805)^2(0,16257)^2$ = 0,002427011, et donc $\sqrt{\hat{V}(\hat{p}_{st})} = \sqrt{0,002427011}$ = 0,0492647.

L'écart-type de $\hat{N}_c$ est $\sqrt{\hat{V}(\hat{N}_c)} = N\sqrt{\hat{V}(\hat{p}_{st})}$ = 10 403(0,0492647) = 512,5.

c) L'intervalle de confiance pour p est [0,4226 – 1,96 (0,0492647) ; 0,4226 + 1,96 (0,0492647)] = [0,326088 ; 0,519202].

On affirme donc avec environ 95 % de confiance que le pourcentage d'employés en faveur du plan est compris entre 32,6 % et 51,9 %. Le nombre d'employés en faveur du plan est donc compris entre (10 403)(0,326088) = 3 392 et (10 403)(0,326088) = 5 401.

ALLOCATIONS PROPORTIONNELLE ET OPTIMALE

L'allocation proportionnelle et l'allocation optimale peuvent aussi bien être définies pour l'estimation d'une proportion que pour l'estimation d'une moyenne. Mais il se trouve que la précision est peu affectée par l'allocation. La stratification avec allocation proportionnelle est avantageuse (par rapport à un *éas*) si les p_h sont différents les uns des autres. Mais l'allocation optimale rapporte rarement d'importants gains par rapport à l'allocation proportionnelle. Nous avons vu que l'allocation optimale est particulièrement efficace lorsque les S_h sont différents les uns des autres. Or les S_h sont fonctions des produits $p_h(1-p_h)$, et ces produits ne varient pas trop d'une strate à l'autre à moins que les écarts entre les p_h soient très importants.

Le prochain exemple illustre ce fait.

Exemple 7.10 *Répartitions proportionnelle et optimale en fonction des p_h*

Les 3 strates d'une population contiennent respectivement 175, 375 et 450 unités. On tire un échantillon de 100 unités afin d'estimer une proportion p. Le tableau suivant compare les coefficients de variation (en pourcentage) de l'estimateur de p sous trois modes d'échantillonnage: aléatoire simple, stratifié avec allocation proportionnelle, et stratifié avec allocation optimale.

	Mode d'échantillonnage		
	Aléatoire	Stratification. allocation…	
p_h	simple	proportionnelle	optimale
0,45 ; 0,50 ; 0,55	9,64	9,17	9,28
0,05 ; 0,15 ; 0,25	21,13	20,00	20,52
0,05 ; 0,30 ; 0,80	10,15	7,84	9,9

On constate que dans les deux premiers cas, la différence entre les trois modes d'échantillonnage est négligeable, ce qui s'explique par le fait que les p_h sont relativement proches les uns des autres. Dans le troisième cas, les différences entre les p_h sont importantes, et là, le mode d'échantillonnage importe. Mais là encore, la différence est entre un *éas* et un échantillon stratifié. Il n'y a, à toute fin pratique, pas de différence entre l'allocation proportionnelle et l'allocation optimale.

7.4 STRATIFICATION APRÈS SÉLECTION

La stratification décrite ci-dessus est une stratification planifiée : on commence par stratifier la population, puis on tire indépendamment un échantillon — de taille prédéterminé — dans chaque strate. Mais la stratification peut aussi être effectuée après la sélection d'un *éas*. Cette technique est particulièrement utile lorsque, au vu de l'échantillon, on soupçonne qu'un biais de sélection s'est produit. On peut, par exemple, constater que certaines parties de la population semblent être sous-représentées ou surreprésentées. La stratification après sélection tente de corriger le biais.

Exemple 7.11 *Stratification après sélection*

Le Tableau 7.2 présente les données d'un *éas* de 70 villes tirées de la population des villes présentée au tableau A05. La variable Y est le nombre de familles avec 3 enfants ou plus. Il existe une stratification en quatre régions géographiques, mais on n'en a pas tenu compte lors du tirage.

On obtient les résultats suivants : $\bar{y} = 209{,}79$; $s_y = 479{,}13$; $\sqrt{\hat{V}(\bar{y})} = 55{,}51$.

Maintenant, si on compare la distribution des strates dans l'échantillon à celle de la population, on découvre des écarts non négligeables. Par exemple, 20 données (29 % de l'échantillon) proviennent de la strate 1 alors que celle-ci comprend que 40 % de la population; 19 % de l'échantillon appartiennent à la strate 4 alors que celle-ci ne contient que 10 % de la population. Ces écarts pourraient être causés par un défaut dans la procédure de sélection et si c'est le cas on a intérêt à appliquer un correctif. La stratification de l'échantillon en est un. On classe les données selon la strate, puis on estime la moyenne comme si on avait tiré un échantillon stratifié. Le Tableau 7.2 montre les calculs. qui sont en fait identiques aux calculs effectués pour un échantillon proprement stratifié.

Tableau 7.2
Stratification après sélection

Échantillons	n_h	N_h	$\bar{y}_h$	s_h	$\sqrt{\hat{V}(\bar{y}_h)}$
0 0 0 5 5 5 10 10 10 10 10 10 15 15 15 15 15 20 20 20	20	468	10,5	6,468547	1,41517
15 15 15 25 30 30 35 40 40 40 45 45 45 55 65 65 70	17	344	39,70588	17,272683	4,08442
40 45 45 55 65 85 90 90 95 100 105 115 120 140 140 150 155 175 175 390	20	232	118,75	76,448727	16,3410
220 275 300 420 475 500 505 515 715 1000 1265 2235 3000	13	116	878,84615	839,253522	219,3366

Les résultats sont alors : $\bar{y}_{st}$ = 127,65; $\sqrt{\hat{V}(\bar{y}_{st})}$ = 22,22.

La moyenne $\bar{y}$ = 209,79 est une surestimation due à une surreprésentation de la strate 4 et une sous-représentation de la strate 1. L'écart-type estimé est supérieur à celui estimé en premier, mais le but de la stratification après la sélection n'est pas de réduire l'écart-type; c'est de corriger un biais potentiel. L'estimation $\bar{y}_{st}$ = 127,65 est plus crédible que $\bar{y}$ = 209,79.

NON-RÉPONSES

Un des problèmes les plus épineux en sondages est celui des non-réponses. Un questionnaire envoyé à 5 000 personnes qui ne récolte que 500 réponses n'est pas très crédible. Ce n'est pas que le nombre est insuffisant, sa taille peut bien être adéquate. Mais un tel échantillon n'est plus un *éas*, et ses propriétés sont inconnues, car on ne sait pas de quelle façon les répondants se sont sélectionnés. Plus le taux de réponses est faible (comme il l'est infailliblement dans un sondage par courrier ou par téléphone), plus le risque de biais est élevé. Par exemple, dans un *éas*, la probabilité d'appartenance des unités devraient être égales. Il est peu probable qu'elles le soient vraiment lorsque les répondants se sélectionnent eux-mêmes. Un sondage d'opinion politique, par exemple, suscitera plus de réponses de la part des plus politisés; ou moins de réponses de certains partis que d'autres.

Il n'existe pas de vraies « solutions » à ce problème, mais il y a quelques approches qui peuvent aider à réduire le biais. La stratification est l'une d'elles; la stratification a posteriori aussi. Dans un sondage politique, par exemple, on peut constater que le taux de réponses est nettement inférieur dans certains groupes d'âge. Donc dans un *éas*, certains groupes seront surreprésentés, d'autres sous-représentés. Une stratification selon le groupe d'âge atténue ce déséquilibre.

Une stratification a posteriori ne peut se faire que si les tailles relatives W_h des strates sont connues. Mais dans certaines situations, ces W_h peuvent être estimées. On l'illustre dans le prochain exemple.

Exemple 7.12 *Ajustement pour non-réponses*

Une succursale de banque envoie un questionnaire à ses 3 000 clients afin d'estimer, entre autres, leur revenu. Elle obtient 1 200 réponses. Sachant que la tendance à répondre dépend de l'âge — le taux de réponse est traditionnellement plus élevé parmi les retraités, par exemple — on stratifie l'échantillon en trois groupes. Voici leur répartition:

	≤ 30 ans	31 - 64 ans	≥ 65	Tous
Répondants (n_{hr})	200	540	460	1200
Non-répondants	300	1260	240	1800
n_h	500	1800	700	3000
Taux de réponse	40 %	30 %	66 %	

Les personnes de 65 ans ou plus sont surreprésentées alors que les 31-64 ans sont sous-représentés. La moyenne ordinaire $\bar{y}$, qui ne peut être calculée que pour les répondants, en serait sérieusement affectée. L'estimateur par stratification après sélection, $\sum_{h=1}^{3} W_h \bar{y}_{hr}$, où $\bar{y}_{hr}$ est la moyenne des répondants de la strate h, permet de pondérer les moyennes $\bar{y}_{1r}$, $\bar{y}_{2r}$ et $\bar{y}_{3r}$ de façon à refléter la population. Cette solution exige que l'âge des membres de la population puisse être déterminé.

Lorsque les poids W_h des strates ne sont pas connus, il faudrait au moins connaître l'âge de tous ceux à qui le questionnaire a été envoyé. Dans ce cas on peut estimer les W_h par n_h/n. La moyenne est donc estimée par $\sum_{h=1}^{L} \frac{n_h}{n} \bar{y}_{hr}$, soit $\frac{500}{3000}\bar{y}_{1r} + \frac{1800}{3000}\bar{y}_{2r} + \frac{700}{3000}\bar{y}_{3r}$ dans cet exemple. Notons, cependant, que cette situation (où l'on connaît l'âge des non-répondants de l'échantillon) est rare.

Remarque *Correction pour non-réponses*

L'estimateur décrit dans l'Exemple 7.12 corrige le biais créé par les différences entre les taux de non-réponse en donnant aux moyennes $\bar{y}_{hr}$ la pondération qu'il faut, mais il se fie sur l'hypothèse que $\bar{y}_{hr}$ est un estimateur sans biais de μ_h. Or ce que $\bar{y}_{hr}$ estime, c'est plutôt la moyenne de ceux dans la strate qui sont susceptibles de répondre, et cette moyenne-là est peut-être différente: à l'intérieur d'une même strate, les répondants peuvent constituer une classe distincte du reste de la strate.

DISCUSSION

La légitimité de la stratification après sélection ne découle pas directement de la théorie de l'échantillonnage par strates développée ci-dessus, car cette dernière part du principe que les tailles n_h sont des nombres fixes, choisis d'avance. Ce qui n'est pas le cas ici puisque le nombre d'observations qui tombent dans une strate donnée est aléatoire. On agit néanmoins comme s'ils étaient fixes. Cette pratique est légitimée pas le théorème suivant.

THÉORÈME 7.7 **DISTRIBUTION CONDITIONNELLE DES VALEURS DES STRATES**

Soit $U = U_1 \cup U_2 \cup \ldots \cup U_L$ une partition en strates d'une population U, et ω un *éas* de taille n. Soit $\omega = \omega_1 \cup \omega_2 \cup \ldots \cup \omega_L$ une partition de ω, ω_h étant l'ensemble des éléments de ω appartenant U_h. Soit n_h le nombre d'éléments dans ω_h. Alors

$$p(\omega \mid n_1; n_2; \ldots; n_L) = \frac{1}{\binom{N_1}{n_1}\binom{N_2}{n_2}\ldots\binom{N_L}{n_L}} .$$

Ceci montre que la distribution conditionnelle des échantillons est précisément celle d'un échantillon stratifié. Toutes les propriétés énoncées pour un échantillon stratifié sont valables et les statistiques de l'échantillon se comportent comme s'il s'agissait d'un échantillon stratifié proprement dit. Ce qui veut dire que l'espérance *conditionnelle* de $\bar{y}_{st}$ est égale à μ:

$$E(\bar{y}_{st} \mid n_1; n_2; \ldots; n_L) = \mu;$$

que l'expression (7.2) donne la variance *conditionnelle* de $\bar{y}_{st}$,

$$V(\bar{y}_{st} \mid n_1; n_2; \ldots; n_L) = \sum_{h=1}^{L} W_h^2 (1 - f_h) \frac{S_h^2}{n_h} ;$$

et que (7.3) est conditionnellement sans biais dans le sens que son espérance conditionnelle est égale à la variance conditionnelle de $\bar{y}_{st}$:

$$E[\hat{V}(\bar{y}_{st} \mid n_1; n_2; \ldots; n_L)] = V(\bar{y}_{st} \mid n_1; n_2; \ldots; n_L) .$$

Que signifie, concrètement, ce qualificatif, « conditionnelle »? L'espérance non conditionnelle de $\bar{y}_{st}$, par exemple, est la moyenne des $\bar{y}_{st}$ de l'ensemble de *tous les échantillons possibles*, y compris ceux dont les n_h diffèrent des n_h observées. L'espérance *conditionnelle* de $\bar{y}_{st}$ étant donné les n_h est la moyenne des $\bar{y}_{st}$ de tous les échantillons constitués de ces mêmes effectifs de strates.

En ce qui concerne l'espérance conditionnelle de $\bar{y}_{st}$, il se trouve qu'elle ne dépend pas des n_h et donc qu'elle est aussi l'espérance inconditionnelle: $\bar{y}_{st}$ est sans biais tout court.

Ce n'est pas le cas avec la variance conditionnelle, $\sum_{h=1}^{L} W_h^2 (1 - n_h / N_h) S_h^2 / n_h$ qui dépend bel et bien des n_h. Quelle est la variance *inconditionnelle* de $\bar{y}_{st}$? On peut la déterminer approximativement en utilisant la formule $V(\bar{y}_{st} \mid n_1; \ldots; n_L) = V[E(\bar{y}_{st} \mid n_1; \ldots; n_L)] + E[V(\bar{y}_{st} \mid n_1; \ldots; n_L)]$ Puisque $E(\bar{y}_{st} \mid n_1, \ldots, n_L) = \mu$ est un nombre fixe, $V[E(\bar{y}_{st} \mid n_1, \ldots, n_L)] = 0$ et donc

$$V(\bar{y}_{st}) = E[V(\bar{y}_{st} \mid n_1; \ldots; n_L)]$$

$$= E\left[\sum_{h=1}^{L} W_h^2 (1 - n_h / N_h) \frac{S_h^2}{n_h}\right] = E\left[\sum_{h=1}^{L} \frac{W_h^2 S_h^2}{n_h} - \sum_{h=1}^{L} \frac{W_h^2 S_h^2}{N_h}\right]$$

$$= \sum_{h=1}^{L} W_h^2 S_h^2 E\left(\frac{1}{n_h}\right) - \frac{1}{N}\sum_{h=1}^{L} W_h S_h^2 .$$

La méthode de linéarisation donne à peu près $E\left(\frac{1}{n_h}\right) = \frac{1}{nW_h}$ ce qui aboutit finalement à $V(\bar{y}_{st}) \approx \frac{1}{n}\sum_{h=1}^{L} W_h(1-n/N)S_h^2$. C'est la variance dans un échantillon stratifié avec allocation proportionnelle. Ce qui est naturel, puisque le hasard fait que les n_h ont tendance à être à peu près proportionnelles aux W_h.

7.5 ÉCHANTILLONNAGE DOUBLE

Un échantillonnage par grappes exige, naturellement, qu'on puisse identifier à l'avance la strate d'appartenance des unités de la population. On peut stratifier après sélection, mais même là il faut connaître les N_h. Sans cette information, impossible de stratifier, ni a priori ni a posteriori. Si on y tient quand même, croyant que la stratification promet d'importants gains de précision, on peut considérer ce qu'on appelle un *échantillonnage double*. L'échantillonnage double consiste à estimer, à l'aide d'un grand échantillon, l'information qui permettra de stratifier. On définit les strates dans ce grand échantillon, et puis on tire dans celui-ci un plus petit échantillon stratifié. Évidemment, pour que l'amélioration en vaille la peine, il faut que le premier échantillon, qui se doit d'être grand, puisse se faire à faible coût.

Supposons donc que la population est de taille M et les strates de tailles M_h, $h = 1, \ldots, L$. On tire un *éas* de taille N. Soit N_h le nombre d'unités parmi les N qui appartiennent à la strate h. Soit $W_h = \frac{M_h}{M}$, $\hat{W}_h = \frac{N_h}{N}$. On tire un échantillon stratifié dans le premier échantillon. Soit n_h la taille de l'échantillon tiré de la strate h, $\bar{y}_h$ sa moyenne et s_h son écart-type. Étant donné que les N_h sont inconnues, il n'est pas raisonnable de fixer les n_h d'avance. Nous supposerons cependant que les proportions $r_h = n_h/N_h$ sont fixées d'avance. Nous avons alors les résultats suivants:

THÉORÈME 7.8 **ÉCHANTILLONNAGE DOUBLE**

a) L'estimateur

(7.18) $$\hat{\bar{y}} = \sum_{h=1}^{L} \hat{W}_h \bar{y}_h$$

est sans biais pour la moyenne de la population μ;

b) La variance de $\hat{\bar{y}}$ est

(7.19) $$V(\hat{\bar{y}}) = \frac{1}{N}\sum_{h=1}^{L} W_h S_h^2\left(\frac{1}{r_h} - 1\right) + \frac{M-N}{MN} S^2$$

où S^2 est la variance de la population est S_h^2 est la variance de la strate h.

c) La variance de $\hat{\bar{y}}$ est estimée par

$$\hat{V}(\hat{\bar{y}}) = \frac{1}{N}\sum_{h=1}^{L} \hat{W}_h s_h^2 \left(\frac{1}{r_h} - 1\right) + \frac{M-N}{MN}\hat{S}^2(\omega_2)$$

où

$$(7.20) \qquad \hat{S}^2(\omega_2) = \frac{1}{N-1}\left\{\sum N_h(\bar{y}_h - \hat{\bar{y}}_{st})^2 + \sum N_h s_h^2 \left[1 + \frac{N_h}{Nn_h} - \frac{1}{N} - \frac{1}{n_h}\right]\right\}$$

est un estimateur sans biais de S^2.

Remarque *Le choix des r_h*

Le choix des r_h devrait normalement être motivé par des considérations de précision. Les gains potentiels de précision sont importants seulement si les S_h diffèrent considérablement d'une strate à l'autre. Dans ce cas, on choisit r_h grand là où S_h est grand et petit là où S_h est petit. En l'absence d'une motivation quelconque, on choisirait des r_h constants, donc $r_h = n/N$. Dans ce cas, il est peu probable que l'échantillonnage double soit plus efficace qu'une simple stratification a posteriori.

7.6 DÉMONSTRATIONS

THÉORÈME 7.1 **DÉCOMPOSITION D'UNE SOMME DE CARRÉS**

$$(N\text{-}1)S^2 = \sum_{h=1}^{L}(N_h - 1)S_h^2 + \sum_{h=1}^{L} N_h(\mu_h - \mu)^2 \ .$$

DÉMONSTRATION

Soit y_{hj} la j^e donnée de la strate h. Alors

$$(N-1)S^2 = \sum_{h=1}^{L}\sum_{j=1}^{N_h}(y_{hj} - \mu)^2 = \sum_{h=1}^{L}\sum_{j=1}^{N_h}[(y_{hj} - \mu_h) + (\mu_h - \mu)]^2$$

$$= \sum_{h=1}^{L}\sum_{j=1}^{N_h}(y_{hj} - \mu_h)^2 + \sum_{h=1}^{L}\sum_{j=1}^{N_h}(\mu_h - \mu)^2 + 2\sum_{h=1}^{L}\sum_{j=1}^{N_h}(y_{hj} - \mu_h)(\mu_h - \mu)$$

$$= \sum_{h=1}^{L}(N_h - 1)S_h^2 + \sum_{h=1}^{L} N_h(\mu_h - \mu)^2 \text{ , la somme}$$

$$2\sum_{h=1}^{L}\sum_{j=1}^{N_h}(y_{hj} - \mu_h)(\mu_h - \mu) = 2\sum_{h=1}^{L}(\mu_h - \mu)\left[\sum_{j=1}^{N_h}(y_{hj} - \mu_h)\right]$$

étant nulle.

THÉORÈME 7.2 **VARIANCE, ALLOCATION PROPORTIONNELLE**

DÉMONSTRATION

$$V(\bar{y}_{st}) = \sum_{h=1}^{L} W_h^2 \frac{S_h^2}{n_h} - \sum_{h=1}^{L} W_h^2 \frac{S_h^2}{N_h} = \sum_{h=1}^{L} W_h^2 \frac{S_h^2}{nW_h} - \sum_{h=1}^{L} W_h^2 \frac{S_h^2}{NW_h}$$

$$= \frac{1}{n}\sum_{h=1}^{L} W_h S_h^2 - \frac{1}{N}\sum_{h=1}^{L} W_h S_h^2 = \frac{(1-f)}{n}\sum_{h=1}^{L} W_h S_h^2 \text{ , où } f = n/N.$$

THÉORÈME 7.4 **VARIANCE, ALLOCATION OPTIMALE**

Deux expressions pour la variance optimale:

(7.21) $$V_{opt} = \frac{1}{n}(\sum_{h=1}^{L} W_h S_h)^2 - \frac{1}{N}\sum_{h=1}^{L} W_h S_h^2 \ ;$$

(7.22) $$V_{opt} = \frac{1-f}{n}\sum_{h-1}^{L} W_h S_h^2 - \frac{1}{n}\sum_{h=1}^{L} W_h (S_h - \bar{S})^2 ,$$

où

$$\bar{S} = \sum_{h=1}^{L} W_h S_h \ .$$

DÉMONSTRATION

$$V(\bar{y}_{st}) = \sum_{h=1}^{L} W_h^2 \left(1 - \frac{n_h}{N_h}\right)\frac{S_h^2}{n_h} = \sum_{h=1}^{L} \frac{W_h^2 S_h^2}{n_h} - \sum_{h=1}^{L} \frac{W_h^2 S_h^2}{N_h} = \sum_{h=1}^{L} \frac{W_h^2 S_h^2}{nW_h S_h / (\sum_{i=1}^{L} W_i S_i)} - \sum_{h=1}^{L} \frac{W_h^2 S_h^2}{NW_h} =$$

$\frac{1}{n}(\sum_{h=1}^{L} W_h S_h)^2 - \frac{1}{N}\sum_{h=1}^{L} W_h S_h^2$, ce qui démontre (7.21).

La forme (7.22) découle de la décomposition suivante (voir l'équation (1.1)):

(7.23) $$\sum_{h=1}^{L} W_h (S_h - \bar{S})^2 = \sum_{h=1}^{L} W_h S_h^2 - (\sum_{h=1}^{L} W_h S_h)^2 \ .$$

Il suffit de substituer au terme $(\sum_{h=1}^{L} W_h S_h)^2$ dans (7.21) l'expression $(\sum_{h=1}^{L} W_h S_h)^2 = \sum_{h=1}^{L} W_h S_h^2 - \sum_{h=1}^{L} W_h (S_h - \bar{S})^2$, ce qui donne (7.22) après simplification.

THÉORÈME 7.5 **VARIANCE DE $\bar{y}$ EN FONCTION DES PARAMÈTRES DES STRATES**

$$V_a = \frac{1-f}{n}\left[\sum_{h=1}^{L} \frac{(N_h - 1)}{N-1} S_h^2 + \sum_{h=1}^{L} \frac{N_h}{N-1}(\mu_h - \mu)^2\right] \ .$$

DÉMONSTRATION

Selon le Théorème 7.1, $S^2 = \sum_{h=1}^{L} \frac{(N_h - 1)}{N-1} S_h^2 + \sum_{h=1}^{L} \frac{N_h}{N-1} (\mu_h - \mu)^2$. Le résultat suit en substituant cette expression à S^2 dans $V_a = (1 - n/N)\frac{S^2}{n}$.

THÉORÈME 7.6 **OPTIMISATION DU COUT**

L'allocation qui minimise le coût $c_0 + \sum_{i=1}^{L} c_h n_h$ pour une variance donnée V_0; ou qui minimise la variance pour un coût donné $c_0 + \sum_{i=1}^{L} c_h n_h = c_0 + B_0$ est $n_h \propto \frac{W_h S_h}{\sqrt{c_h}}$.

DEMONSTRATION

Nous nous servons de l'inégalité de Cauchy-Schwarz. Voici le théorème sous une forme propre au contexte.

Inégalité de Cauchy-Schwarz. Soit $a_1, a_2, \ldots, a_L$ et $b_1, b_2, \ldots, b_L$ deux séries de nombres. Alors

(7.24) $$\left(\sum_{h=1}^{L} a_h b_h\right)^2 \le \left(\sum_{h=1}^{L} a_h^2\right)\left(\sum_{h=1}^{L} b_h^2\right),$$

et

(7.25) $$\left(\sum_{h=1}^{L} a_h b_h\right)^2 = \left(\sum_{h=1}^{L} a_h^2\right)\left(\sum_{h=1}^{L} b_h^2\right) \Leftrightarrow a_h \propto b_h.$$

L'inégalité est stricte à moins que, pour une constante k, $a_h = kb_h$.

La variance de $\bar{y}_{st}$ est une somme $V(\bar{y}_{st}) = \sum_{h=1}^{L} \frac{W_h^2 S_h^2}{n_h} - \frac{1}{N}\sum_{h=1}^{L} W_h S_h^2$ dont seul le premier terme dépend des n_h. Donc minimiser la variance équivaut à minimiser le premier terme $\sum_{h=1}^{L} \frac{W_h^2 S_h^2}{n_h}$. Par l'inégalité de Cauchy-Schwarz avec $a_h = \frac{W_h S_h}{\sqrt{n_h}}$ et $b_h = \sqrt{c_h n_h}$,

$\left(\sum_{h=1}^{L} W_h S_h \sqrt{c_h}\right)^2 = \left(\sum_{h=1}^{L} \frac{W_h S_h}{\sqrt{n_h}} \sqrt{c_h n_h}\right)^2 \le \left(\sum_{h=1}^{L} \frac{W_h^2 S_h^2}{n_h}\right)\left(\sum_{h=1}^{L} c_h n_h\right)$. Donc

$$\sum_{h=1}^{L} \frac{W_h^2 S_h^2}{n_h} \ge \frac{\left(\sum_{h=1}^{L} W_h S_h \sqrt{c_h}\right)^2}{\left(\sum_{h=1}^{L} c_h n_h\right)} \quad \text{ou} \quad \sum_{h=1}^{L} c_h n_h \ge \frac{\left(\sum_{h=1}^{L} W_h S_h \sqrt{c_h}\right)^2}{\left(\sum_{h=1}^{L} \frac{W_h^2 S_h^2}{n_h}\right)}.$$

Si on fixe le coût à $\sum_{h=1}^{L} c_h n_h = B_0$, $\left(\sum_{h=1}^{L} \frac{W_h^2 S_h^2}{n_h}\right)$ la première inégalité ci-dessus donne une borne inférieure pour la variance :

$$\left(\sum_{h=1}^{L} \frac{W_h^2 S_h^2}{n_h}\right) \geq \frac{\left(\sum_{h=1}^{L} W_h S_h \sqrt{c_h}\right)^2}{B_0} .$$

Si c'est la variance qui est fixée, $V(\bar{y}_{st}) = V_0 \Rightarrow \sum_{h=1}^{L} \frac{W_h^2 S_h^2}{n_h} = V_0 + \frac{1}{N}\sum_{h=1}^{L} W_h S_h^2$, alors la deuxième équation donne un borne inférieure pour le coût

$$\sum_{h=1}^{L} c_h n_h \geq \frac{\left(\sum_{h=1}^{L} W_h S_h \sqrt{c_h}\right)^2}{\left(V_0 + \frac{1}{N}\sum_{h=1}^{L} W_h S_h^2\right)} .$$

Dans les deux cas, selon (7.25), ces bornes sont atteintes si et seulement si $\frac{W_h S_h}{\sqrt{n_h}} \propto \sqrt{c_h n_h}$, ce qui est équivalent à $n_h \propto \frac{W_h S_h}{\sqrt{c_h}}$.

THÉORÈME 7.8 **ÉCHANTILLONNAGE DOUBLE**

Dans un échantillonnage double, (7.18) est un estimateur sans biais de la moyenne μ; (7.19) est sa variance; et (7.20) est un estimateur sans biais de la variance.

DEMONSTRATION

Soit U_h l'ensemble des unités de la strate h de la population; ω_{1h} l'ensemble des unités appartenant à la strate h parmi les unités du premier échantillon; et ω_{2h} l'échantillon de taille n_h tiré dans ω_{1h} . Soit M_h, N_h et n_h les effectifs de U_h, ω_{2h} et ω_{1h}, respectivement; et M, N, et n les tailles de la population, du premier échantillon et du deuxième, respectivement. Soit μ_h et S_h la moyenne et l'écart-type de U_h; $\hat{\mu}_h$ et $\hat{S}_h$ ceux de ω_{1h}; et $\bar{y}_h$ et s_h ceux de ω_{2h}.

On conditionne deux fois. E_1 et V_1 désignent l'espérance et la variance par rapport à ω_2 étant donné ω_1; E_2 et V_2 désignent l'espérance et la variance par rapport à ω_1 étant donné les N_h; et E_3 et V_3 désignent l'espérance et la variance de par rapport aux $\hat{W}_h$ (ou N_h). Notons que conditionnellement à ω_1, ω_2 est un échantillon stratifié d'une population stratifiée dont les strates sont de tailles N_h, de moyenne $\hat{\mu}_h$ et de variance $\hat{S}_h^2$. Conditionnellement aux N_h, ω_2 est un échantillon stratifié d'une population dont les strates sont de tailles M_h, de moyennes μ_h et de variances S_h^2 (voir le Théorème 7.7) .

a) $E(\hat{\bar{y}}_{st}) = E[E(\hat{\bar{y}}_{st} \mid \omega_1)]$

$$E(\hat{\bar{y}}_{st} \mid \omega_1) = E[\sum_{h=1}^{L} \hat{W}_h \bar{y}_h \mid \omega_1] = \sum_{h=1}^{L} E[E(\hat{W}_h \bar{y}_h \mid N_h) \mid \omega_1] = \sum_{h=1}^{L} E(\hat{W}_h \hat{\mu}_h \mid \omega_1) = \bar{y}(\omega_1) .$$

On a donc

(7.26) $$E(\hat{\bar{y}}_{st} \mid \omega_1) = \bar{y}(\omega_1) ,$$

où $\bar{y}(\omega_1)$ est la moyenne de toutes les observations de ω_1. Alors $E(\hat{\bar{y}}_{st}) = E[E(\hat{\bar{y}}_{st} \mid \omega_1)] = E[\bar{y}(\omega_1)] = \mu$.

b) $V(\hat{\bar{y}}_{st}) = V[E(\hat{\bar{y}}_{st} \mid \omega_1)] + E[V(\hat{\bar{y}}_{st} \mid \omega_1)]$. Le premier terme est, par (7.26), $V[E(\hat{\bar{y}}_{st} \mid \omega_1)] = V[\bar{y}(\omega_1)]$, et donc

(7.27) $$V[E(\hat{\bar{y}}_{st} \mid \omega_1)] = (1 - N / M)\frac{S^2}{N} .$$

Le deuxième terme est

$$V(\hat{\bar{y}}_{st} \mid \omega_1) = \sum_{h=1}^{L} \hat{W}_h^2 (1 - n_h / N_h)\frac{\hat{S}_h^2}{n_h} = \frac{1}{N^2}\sum_{h=1}^{L}\left(\frac{N_h^2}{r_h N_h} - N_h\right)\frac{\hat{S}_h^2}{n_h} = \frac{1}{N}\sum_{h=1}^{L}\left(\frac{1}{r_h} - 1\right)\hat{W}_h\hat{S}_h^2 .$$

On a donc

(7.28) $$V(\hat{\bar{y}}_{st} \mid \omega_1) = \frac{1}{N}\sum_{h=1}^{L}\left(\frac{1}{r_h} - 1\right)\hat{W}_h\hat{S}_h^2 ,$$

et

(7.29) $$E[V(\hat{\bar{y}}_{st} \mid \omega_1] = \frac{1}{N}\sum_{h=1}^{L}\left(\frac{1}{r_h} - 1\right)E(\hat{W}_h\hat{S}_h^2) .$$

Or $E(\hat{W}_h\hat{S}_h^2) = E[E(\hat{W}_h\hat{S}_h^2 \mid N_h)] = E[\hat{W}_h E(\hat{S}_h^2 \mid N_h)] = E(\hat{W}_h S_h^2) = E(\hat{W}_h)S_h^2 = W_h S_h^2$.

Donc

(7.30) $$E[V(\hat{\bar{y}}_{st} \mid \omega_1] = \frac{1}{N}\sum_{h=1}^{L}\left(\frac{1}{r_h} - 1\right)W_h S_h^2 .$$

Finalement,

$$V(\hat{\bar{y}}_{st}) = V[E(\hat{\bar{y}}_{st} \mid \omega_1)] + E[V(\hat{\bar{y}}_{st} \mid \omega_1)] = (1 - N / M)\frac{S^2}{N} + \frac{1}{N}\sum_{h=1}^{L}\left(\frac{1}{r_h} - 1\right)W_h S_h^2 .$$

c) Le premier terme de $V(\hat{\bar{y}}_{st})$, soit $Q_1 = \frac{1}{N}\sum_{h=1}^{L}\left(\frac{1}{r_h} - 1\right)W_h S_h^2$, peut être estimé par $\hat{Q}_1 = \frac{1}{N}\sum_{h=1}^{L}\left(\frac{1}{r_h} - 1\right)\hat{W}_h s_h^2$, car $E(\hat{Q}_1 \mid \omega_1) = \frac{1}{N}\sum_{h=1}^{L}\left(\frac{1}{r_h} - 1\right)E(\hat{W}_h s_h^2 \mid \omega_1) = \frac{1}{N}\sum_{h=1}^{L}\left(\frac{1}{r_h} - 1\right)\hat{W}_h\hat{S}_h^2$, et donc $E(\hat{Q}_1) = E[E(\hat{Q}_1 \mid \omega_1)] = E[V(\hat{\bar{y}}_{st} \mid \omega_1]$ selon l'équation (7.29): $E(\hat{Q}_1) = E[E(\hat{Q}_1 \mid \omega_1)] = \frac{1}{N}\sum_{h=1}^{L}\left(\frac{1}{r_h} - 1\right)E(\hat{W}_h\hat{S}_h^2)$.

Un estimateur sans biais de S^2 est donné par

$$\frac{\sum N_h(\bar{y}_h - \hat{\bar{y}}_{st})^2 + \sum N_h s_h^2\left[1 + \frac{N_h}{Nn_h} - \frac{1}{N} - \frac{1}{n_h}\right]}{N-1}. \tag{7.31}$$

Son espérance conditionnelle (étant donné ω_1) est $\hat{S}^2$ (voir l'exercice 7.15), et donc le terme $(1 - N/M)\frac{S^2}{N}$ est estimé par $\frac{M-N}{MN}\frac{\sum N_h(\bar{y}_h - \hat{\bar{y}}_{st})^2 + \sum N_h s_h^2\left[1 + \frac{N_h}{Nn_h} - \frac{1}{N} - \frac{1}{n_h}\right]}{N-1}$. Ce théorème a été démontré par, entre autres, (Rao, 1973).

RÉSUMÉ

- $\bar{y}_{st} = \sum_{h=1}^{L} W_h \bar{y}_h$ est un estimateur sans biais de la moyenne de la population μ.
- La variance de $\bar{y}_{st}$ est $V(\bar{y}_{st}) = \sum_{h=1}^{L} W_h^2 V(\bar{y}_h)$, où $V(\bar{y}_h) = (1 - f_h)\frac{S_h^2}{n_h}$.
- La variance $V(\bar{y}_{st})$ est estimée par $\hat{V}(\bar{y}_{st}) = \sum_{h=1}^{L} W_h^2 \hat{V}(\bar{y}_h)$ où $\hat{V}(\bar{y}_h) = (1 - f_h)\frac{s_h^2}{n_h}$ est un estimateur sans biais de $V(\bar{y}_h)$.

Allocation des observations.

- Allocation proportionnelle : $n_h = nW_h$, $W_h = N_h/N$.
- Allocation optimale : $n_h = \frac{W_h S_h}{\sum_{i=1}^{L} W_i S_i} n$.
- Variance sous trois modes d'échantillonnage :

Mode d'échantillonnage	Variance approximative	Comparaisons
Aléatoire simple	$V_a \approx \frac{1-f}{n}[\sum_{h=1}^{L} W_h S_h^2 + \sum_{h=1}^{L} W_h(\mu_h - \mu)^2]$	$V_a - V_{prop} \approx \frac{1-f}{n}\sum_{h=1}^{L} W_h(\mu_h - \mu)^2$
Stratifié — allocation proportionnelle	$V_{prop} \approx \frac{(1-f)}{n}\sum_{h=1}^{L} W_h S_h^2$	$V_{prop} - V_{opt} \approx \frac{1}{n}\sum_{h=1}^{L} W_h(S_h - \bar{S})^2$
Stratifié — allocation optimale	$V_{opt} \approx \frac{1-f}{n}\sum_{h-1}^{L} W_h S_h^2 - \frac{1}{n}\sum_{h=1}^{L} W_h(S_h - \bar{S})^2$.	

Estimation d'une proportion

- L'estimateur d'une proportion dans un échantillon stratifié est $\hat{p}_{st} = \sum_{h=1}^{L} W_h \hat{p}_h$. Son écart-type est estimé par $\sqrt{\hat{V}(\hat{p}_{st})} = \sqrt{\sum_{h=1}^{L} W_h^2 \hat{V}(\hat{p}_h)}$ où $\sqrt{\hat{V}(\hat{p}_h)} = \sqrt{(1-f_h)}\sqrt{\frac{\hat{p}_h(1-\hat{p}_h)}{n_h - 1}}$.

- La stratification a posteriori consiste à stratifier un échantillons aléatoire simple et à estimer μ par $\overline{y}_{st}$.

- Un échantillonnage double consiste à tirer un grand échantillon de taille *N* afin d'établir une stratification dans celui-ci; puis de tirer, dans ce premier échantillon, un plus petit échantillon stratifié de taille *n*. L'estimateur $\hat{\overline{y}}_{st} = \sum_{h=1}^{L} \hat{W}_h \overline{y}_h$ où $\hat{W}_h = N_h / N$ et N_h est le nombre d'unités appartenant à la strate *h* dans le premier échantillon, est sans biais pour μ. La variance et son estimateur sont donnés par le Théorème 7.8.

EXERCICES

7.1 Une population formée de 3 strates de 610, 1 670 et 915 unités. On prélève un échantillon stratifié. On obtient les résultats suivants :

Strate	Données	$n_h \overline{y}_h$	s_h
1	68 98 87 56 34 33 44 28	448	26,21341
2	2 3 4 2 3 4 3 2 6 5 3 4 2 5 6 2 3 5 4 2 3 8	81	1,615000
3	687 675 237 99 123 456 231 324 543 654 345 234	4608	213,5926

a) Estimer la moyenne de la population.

b) Estimer le total de la population.

c) Estimer l'écart-type de l'estimateur de la moyenne.

d) Estimer l'écart-type de l'estimateur du total.

e) Déterminer un intervalle de confiance à 95% pour la moyenne de la population.

f) Déterminer un intervalle de confiance à 95% pour le total de la population.

g) Utiliser les données de l'échantillon pour déterminer l'allocation optimale d'un échantillon de taille 42.

h) Estimer ce qu'aurait été l'écart-type de $\overline{y}_{st}$ si l'allocation optimale avait été utilisée.

7.2 Une population est formée de 5 strates comprenant 235, 432, 1 590, 2 300, et 4 321 unités. On estime que les écarts-types S_h sont 60, 36, 14, 12 et 10.

a) Déterminer l'allocation optimale d'un échantillon de taille 180.

b) Déterminer l'allocation proportionnelle pour un échantillon de taille 180 et comparer l'écart-type de $\bar{y}_{st}$ pour l'allocation proportionnelle et l'allocation optimale.

7.3 On projette de tirer un échantillon stratifié de taille 100 d'une population dont les 3 strates contiennent 30, 300 et 2 000 unités et les écarts-types sont 150, 38 et 14.

a) Déterminer l'allocation proportionnelle et l'écart-type de $\bar{y}_{st}$ sous une allocation proportionnelle.

b) Déterminer l'allocation optimale et l'écart-type de $\bar{y}_{st}$ sous une allocation optimale.

7.4 Une population de 85 comptes est répartie selon le type de client. Les montants de ces comptes sont les suivants :

Strate 1 : Clients industriels

50 212	30 215	12 564	36 598	36 598
36 527	96 532	95 684	69 854	68 594

Strate 2 : Grossistes

3 652	6 598	6 532	5 656	6 532
6 563	6 521	6 532	6 598	6 532
3 268	8 854	6 582	8 457	6 584
9 658	6 532	9 564	9 856	6 598
9 658	6 532	2 147	3 345	5 465

Strate 3 : Détaillants

325	695	658	423	214	659	854	632	632	654
985	658	745	698	365	256	985	654	965	965
985	658	321	123	365	965	965	856	452	325
445	323	765	139	239	432	871	347	138	325
762	769	126	247	246	235	345	345	345	298

Quelques calculs

	μ_h	S_h
Strate 1	53 337,80	28 312,66
Strate 2	6592,64	2031,519
Strate 3	535,58	275,6567

a) Quel est l'écart-type de l'estimateur de la moyenne basé sur un échantillon de taille n = 20, réparti de la façon suivante ?

i) n_1 = 6, n_2 = 11, n_3 = 3 ; ii) n_1 = 10, n_2 = 8, n_3 = 2 ; iii) n_1 = 1, n_2 = 4, n_3 = 15.

b) Déterminer la répartition proportionnelle d'un échantillon de taille 20 et déterminer l'écart-type de l'estimateur sous l'allocation proportionnelle.

c) Déterminer la répartition optimale d'un échantillon de taille 20 et déterminer l'écart-type de l'estimateur sous l'allocation optimale.

7.5 Les étudiants d'une université sont répartis en 4 facultés ayant 1 230, 3 000, 2 500, et 8 000 étudiants respectivement. On prélève un échantillon de 25, 61, 51, et 163 étudiants dans les 4 strates pour estimer la proportion p d'étudiants qui ont déjà utilisé la coop. On trouve que les nombres d'étudiants qui l'ont déjà utilisée dans les 4 échantillons sont 20, 43, 46, et 81 respectivement.

a) Estimer p ainsi que l'écart-type de l'estimateur.

b) Déterminer un intervalle de confiance à 95% pour p.

c) Estimer le nombre N_c d'étudiants dans la population qui ont déjà utilisé la coop.

d) Déterminer un intervalle de confiance à 95% pour N_c.

e) Utiliser l'estimation de p obtenue en *a*) pour estimer l'écart-type d'un estimateur basé sur un *éas* de taille 300.

f) Supposons qu'on a l'intention de tirer un nouvel échantillon stratifié, avec allocation optimale, afin d'estimer p avec une marge d'erreur de 3 points de pourcentage. Quelle devrait être la taille totale de l'échantillon, et comment devrait-on le répartir

i) si on compte allouer les échantillons de façon proportionnelles?

ii) si on compte allouer les échantillons de façon optimale?

7.6 Considérez une population de 1 800 factures d'un magasin de variétés dont on voudrait estimer le montant moyen μ_y. Une variable auxiliaire, le nombre X d'articles, est connue pour la population entière. La variable auxiliaire peut servir à estimer la moyenne par la différence, par le quotient ou par la régression dans un tirage aléatoire simple; ou servir comme base de stratification. Le but de cet exercice est d'évaluer et comparer ces procédures. Supposons donc qu'on tire un échantillon de taille 100. Déterminer l'écart-type de l'estimateur de μ_y dans chacune des circonstances suivantes. (Les données sur la population sont présentées plus bas.)

a) On tire un *éas* de taille 100 et

   i) On estime μ_y par la moyenne;

   ii) On estime μ_y par la différence;

   iii) On estime μ_y par le quotient;

   iv) On estime μ_y par la régression.

b) On tire un échantillon stratifié de taille 100 avec allocation proportionnelle.

c) On tire un échantillon stratifié de taille 100 avec allocation optimale.

[À remarquer: la stratification en fonction de X a presque le même effet que l'estimation par le quotient ou par la régression .

Les données suivent: μ_x = 1,905556; μ_y = 26,45444; S_x = 0,8546197; S_y = 13,31719; S_{xy} = 10,32477

La variable X prend les valeurs 1, 2, 3, 4, et la population est stratifiée selon la valeur de X.

Données par strate:

	Strate (Valeur de X)			
	1	2	3	4
N_h	670	710	340	80
μ_{yh}	13,73881	27,68592	41,85294	56,57500
S_{yh}	4,049634	5,856 651	6,795 072	8,343 534

7.7 Dans le cadre d'un projet de recherche sur les stratégies utilisées par les vendeurs lors de la mise en vente, une agence immobilière doit tirer un échantillon stratifié de maisons récemment vendues (tableau A06). L'objectif est d'estimer le montant moyen μ que le vendeur espérait obtenir pour sa maison. Afin de guider la planification, on tire un échantillon pilote de taille 30 stratifié selon le quartier. Les personnes choisies sont interviewées et révèlent le montant Y qu'ils espéraient obtenir. Voici les données (en milliers de dollars):

	N_h	n_h	$\bar{y}_h$	s_h
Ahuntsic-Cartierville	28	4	450,75	56,62964
Côte-des-Neiges/N.-D.-de-Grâce	36	5	374,4	50,3468
Ville-Mont-Royal	86	12	887,1667	532,62401
Villeray/Saint-Michel/Park-Ext.	40	5	314,8	113,09819
Autres	32	4	376	55,47973

On tolère une marge d'erreur (à 95 %) de 75 (mille dollars).

a) Si on suppose que les $\overline{y}_h$ du tableau sont d'assez bonnes estimations des moyennes des populations, peut-t-on prévoir que la stratification avec allocation proportionnelle sera efficace comparée à un *éas*?

b) Si on suppose que les s_h du tableau sont d'assez bonnes estimations des écarts-types des populations, peut-t-on prévoir que la stratification avec allocation optimale présentera d'importants gains de précision par rapport à une allocation proportionnelle?

c) Quelle devrait être la taille de l'échantillon si l'on prévoit allouer les observations proportionnellement?

d) Quelle devrait être la taille de l'échantillon si l'on prévoit utiliser l'allocation optimale? [En pratique, on évitera les échantillons trop petits. Ici, on devra majorer au moins la taille de l'échantillon dans la strate 1 et possiblement des deux dernières.]

7.8 On tire un échantillon stratifié de taille 30 dans la population décrite au numéro 7.7. La variable d'intérêt Y est le prix de la maison. Une variable auxiliaire, X, le prix auquel la maison est mise sur le marché, est connu pour la population entière. Voici les données de l'échantillon:

n_h	N_h	$\overline{x}_h$	$\overline{y}_h$	s_{xh}	s_{yh}	s_{xyh}
4	28	369,000	315,7500	193,3891	157,1949	30345,67
5	36	832,800	719,8000	590,7125	513,2574	302135,20
12	86	1076,583	981,0833	797,2557	718,1510	571577,77
5	40	544,000	475,0000	396,3559	338,6606	134064,75
4	32	586,500	513,0000	365,4883	357,4306	130479,00

a) Définir la variable $z_i = x_i - y_i$ et estimer la différence moyenne $\mu_z = \mu_x - \mu_y$ entre le prix demandé et le prix vendu. Déterminer l'écart-type de l'estimateur.

b) Est-ce que l'écart-type de l'estimateur par la différence aurait pu être estimée par $\sqrt{\hat{V}(\overline{x}_{st}) + \hat{V}(\overline{y}_{st})}$? Pourquoi ou pourquoi pas?

c) Sachant que μ_x = 719,1982 est connu, on peut également estimer μ_z par $\mu_x - \overline{y}_{st}$. Quel est le meilleur estimateur: celui-ci ou celui défini en a)? Répondre sans calculer, puis confirmer avec des calculs.

d) Déterminer un intervalle de confiance (à 95 %) pour la différence entre le prix moyen des maisons à Ville Mont-Royal (la troisième strate dans le tableau ci-dessus) et celui des autres maisons de la population.

7.9 Le tableau A06 présente des données sur une population de 222 maisons vendues dans la région de Montréal. On se propose de stratifier la population afin d'estimer le prix moyen μ_y des maisons de la population. Considérons une stratification basée sur le nombre de chambres à coucher, comme suit:

Strate	Nombre de chambres à coucher	N_h	μ_h	S_h
1	0 ou 1	27	276,6296	197,4421
2	2	62	410,1613	203,8387
3	3	68	520,3824	267,2459
4	4 ou plus	65	1158,6769	767,7695

Supposons qu'on tire un échantillon stratifié de taille 20 afin d'estimer μ_y.

a) Déterminer l'allocation proportionnelle et calculer l'écart-type de l'estimateur de μ_y dans un échantillon stratifié de taille 20 avec allocation proportionnelle;

b) Déterminer l'allocation optimale et calculer l'écart-type de l'estimateur de μ_y dans un échantillon stratifié de taille 20 avec allocation optimale;

c) L'écart-type de $\bar{y}$ dans un *éas* de taille 20 est 121. Expliquer le gain de précision en a) par rapport à un *éas*; ainsi que l'amélioration due à l'allocation optimale par rapport à l'allocation proportionnelle.

7.10 Le tableau A06 présente des données sur une population de 222 maisons vendues dans la région de Montréal. On se propose de stratifier la population afin d'estimer le prix moyen μ_y des maisons de la population. Considérons une stratification basée sur le nombre de salles de bains. comme suit:

Strate	Nombre de salles de bains	N_h	μ_h	S_h
1	1	113	388,0531	203,0322
2	2	75	694,7733	410,8046
3	3 ou plus	34	1401,2059	907,1333

Supposons qu'on tire un échantillon stratifié de taille 20 afin d'estimer μ_y.

a) Déterminer l'allocation proportionnelle et calculer l'écart-type de l'estimateur de μ_y dans échantillon stratifié de taille 20 avec allocation proportionnelle;

b) Déterminer l'allocation optimale et calculer l'écart-type de l'estimateur de μ_y dans un échantillon stratifié de taille 20 avec allocation optimale;

7.11 On tire un *éas* dans une population de 222 maisons (tableau A06) afin d'estimer le prix moyen μ_y. Le tableau ci-dessous présente les données de l'échantillon. La strate est définie par le nombre de salles de bains, la strate 3 comprenant les maisons avec 3 salles de bains ou plus. Les tailles des strates sont 113, 75 et 34.

y	strate	y	strate	y	strate	y	strate	y	strate
399	1	376	2	1264	3	808	1	426	1
389	2	129	1	1081	2	523	2	1052	3
340	1	2460	3	1010	3	447	3	397	1
283	1	1201	3	888	2	303	1	335	2
182	1	1244	3	853	2	160	2	171	1

a) Estimer μ_y ainsi que l'écart-type de l'estimateur

b) Estimer μ_y par stratification a posteriori ainsi que l'écart-type de l'estimateur.

c) Remarquer que la strate 2 est surreprésentée dans l'échantillon. Expliquer pourquoi cela résulte en un écart-type (estimé) plus petit en b) qu'en c).

7.12 On tire un *éas* de 400 logements d'une petite ville comprenant 4575 logements. L'objectif est d'estimer la proportion p de logements équipés d'un détecteur de fumée.

a) Dans l'échantillon, on trouve que 127 des 400 logements sont équipés d'un détecteur de fumée. Estimer p et estimer l'écart-type de l'estimateur.

b) La ville est répartie en 4 secteurs. Voici les données réparties selon le secteur (z_h est le nombre de logements équipés dans les unités échantillonnales de la strate h):

	Secteur				
	1	2	3	4	Somme
N_h	2345	1530	500	200	4575
n_h	170	113	84	33	400
z_h	17	34	50	26	127

Estimer p par stratification (selon le secteur) a posteriori, et estimer l'écart-type de l'estimateur.

c) Qu'est-ce qui explique que l'estimation en a) est plus élevée qu'en b)?

d) Il y a des raisons de croire que l'échantillon n'a pas vraiment été tiré de façon parfaitement aléatoire, certaines strates étant fortement surreprésentées, d'autres sous-représentées dans l'échantillon. Montrer, par exemple, que le nombre d'observations dans la strate 4 est trop élevé pour être attribué au hasard. Est-ce que cette conclusion conforte le choix de procéder par stratification a posteriori, ou non?

e) Les données ci-dessus ont été recueillies en 2013, alors que les autorités municipales amorçaient une campagne visant à encourager les citoyens à installer des détecteurs de fumée. Un an plus tard (2014), on décide d'effectuer le même sondage, mais avec un échantillon stratifié et allocation optimale (déterminée à partir de l'échantillon actuel, en supposant que rien n'a changé.) Déterminer la répartition optimale d'un échantillon de taille 400.

f) On tire un échantillon stratifié avec l'allocation déterminée en e) avec les résultats suivants:

	Secteur				
	1	2	3	4	Somme
n_h	162	162	57	19	400
z_h	18	54	36	15	123

Estimer la proportion de ménages en 2014 qui sont équipés d'un détecteur de fumée et estimer l'écart-type de l'estimateur.

g) La différence entre les deux années est-elle assez importante pour conclure avec confiance que le pourcentage de logements équipés d'un détecteur de fumée a augmenté (une augmentation qu'on pourrait alors raisonnablement attribuer à la campagne menée par la municipalité)?

7.13 Une variable Y est distribuée uniformément sur l'intervalle [0 ; a]. On forme L strates en découpant l'intervalle en sous-intervalles de longueur a/L puis on tire un *éas* de taille n/L dans chaque strate. Soit V_{strat} la variance de l'estimateur stratifié $\bar{y}_{st}$ de μ_y, et V_a la variance de la moyenne $\bar{y}$ d'un *éas* de taille n. Montrer que $V_{strat} = V_a / L^2$. [Ignorer les facteurs de correction.]

7.14 D'une population constituée de deux strates, de tailles $N_1 = 250$ et $N_2 = 180$, respectivement, on tire un *éas* de taille $n = 20$. Soit n_1 le nombre d'unités échantillonnales qui appartiennent à la strate 1. L'estimateur par stratification a posteriori $\bar{y}_{st}$ est conditionnellement (étant donné n_1) sans biais, tout comme s'il s'agissait d'un tirage stratifié. De même, la variance conditionnelle est $V(\bar{y}_{st} \mid n_1) = W_1^2 \frac{S_1^2}{n_1} + W_2^2 \frac{S_2^2}{n_2} - W_1^2 \frac{S_1^2}{N_1} - W_2^2 \frac{S_1^2}{N_2}$, où $n_2 = n - n_1$. On vous demande de déterminer la variance inconditionnelle de $\bar{y}_{st}$ (mais néanmoins conditionnelle à ce que les n_h soient tous positifs).

a) Déterminer à l'aide d'Excel ou de tout autre logiciel la distribution conditionnelle de n_1 étant donné $n_1 \neq 0$ et $n_2 \neq 0$ et vérifier que $E(1/n_1 \mid n_1 \neq 0 \text{ et } n_2 \neq 0) = 0{,}0893772$ et $E(1/n_2 \mid n_1 \neq 0 \text{ et } n_2 \neq 0) = 0{,}1294379$.

b) Montrer que $0{,}02885925 s_1^2 + 0{,}02170789 s_2^2$ est un estimateur sans biais de $V(\bar{y}_{st} \mid n_1 \neq 0 \text{ et } n_2 \neq 0)$.

7.15 Soit $V_a = (1 - n/N)S^2$ la variance de $\bar{y}$ dans un *éas*.

a) Montrer que

$$\widehat{S^2} = \frac{1}{N-1}\left\{\sum N_h(\bar{y}_h - \bar{y}_{st})^2 + \sum N_h s_h^2\left[1 + \frac{N_h}{N n_h} - \frac{1}{N} - \frac{1}{n_h}\right]\right\}$$

est un estimateur sans biais de S^2 dans un échantillon aléatoire simple.

b) En déduire que $(1-f)\dfrac{\widehat{S^2}}{n}$ est un estimateur sans biais de V_a.

c) Utiliser les données de l'exercice 7.4 pour estimer la variance de $\bar{y}$ dans un *éas* de taille 30.

7.16 On ne peut estimer la variance de $\bar{y}_{st}$ lorsque $n_h = 1$ pour une strate ou plus. Cela est possible, cependant si les strates sont construites au hasard. Supposons qu'on répartit au hasard la population en L strates, de tailles N_h, $h = 1, \ldots, n$, et qu'ensuite on tire une seule observation dans chacune des strates. Soit y_h l'observation tirée dans la strate h.

a) Montrer que $\hat{\mu} = \sum_{h=1}^{n} W_h y_h$ est un estimateur sans biais de μ.

b) Montrer que $V(\hat{\mu}) = S^2[\sum_{h=1}^{n} W_h^2 - 1/N]$, où S^2 est la variance de la population entière.

c) Montrer que $\hat{V}(\hat{\mu}) = \dfrac{\sum_{h=1}^{n} W_h^2 - 1/N}{1-\sum_{h=1}^{n} W_h^2} \sum_{h=1}^{n} W_h (y_h - \hat{\mu})^2$ est un estimateur sans biais pour $V(\hat{\mu})$.

d) Montrer que si les strates sont de même taille, alors la procédure est équivalente à un tirage aléatoire simple (vous pouvez montrer soit que la variance est la même dans les deux cas, ou encore montrer que $p(\omega) = 1/\binom{N}{n}$ pour tout ensemble ω de n unités.)

e) Soit V_a la variance de $\bar{y}$ dans un *éas* de taille n. Montrer que la $V(\hat{\mu}) \geq V_a$ et que $V(\hat{\mu}) = V_a$ si et seulement si les N_h sont égaux.

7.17 Démontrer le Théorème 7.7.

7.18 [*Échantillonnage double*] On tire un *éas* de taille N d'une population stratifiée de taille M dont les strates sont de tailles $M_1, M_2, \ldots, M_L$. Soit $w_h = N_h/N$, où N_h est le nombre d'unités de l'échantillon appartenant à la strate h. Montrer que

a) La variance de w_h est donnée par $V(w_h) = \dfrac{g}{n} W_h (1 - W_h)$ et $E(w_h^2) = \dfrac{g}{N} W_h + \left(1 - \dfrac{g}{N}\right) W_h^2$, où $g = \dfrac{M-N}{M-1}$.

b) La covariance entre w_h et w_ℓ est donnée par $C(w_h; w_\ell) = -\dfrac{M-N}{N(M-1)} W_h W_\ell$.

c) $\hat{V}(w_h) = \dfrac{M-N}{M(N-1)} w_h (1 - w_h)$ est un estimateur sans biais de $V(w_h)$.

d) $\hat{C}(w_h; w_\ell) = -\dfrac{M-N}{M(N-1)} w_h w_\ell$ est un estimateur sans biais de $C(w_h; w_\ell)$.

8 Stratification autres estimateurs

L'estimateur de μ dans un échantillon stratifié est une combinaison linéaire des estimateurs des moyennes μ_h des strates. Au niveau des strates, le seul estimateur considéré au chapitre 7 était la moyenne $\bar{y}_h$. Mais les $\bar{y}_h$ peuvent parfaitement bien être remplacés par d'autres estimateurs, notamment ceux présentés au chapitre 6 : les estimateurs par la différence, par le quotient ou par la régression, et la théorie se développe de façon analogue à celle du dernier chapitre. Cette approche donne lieu à des estimateurs dits *séparés*.

Mais chacun de ces estimateurs (quotient, différence, régression) peut s'appliquer d'une deuxième façon et donne lieu à une nouvelle série d'estimateurs, appelés estimateurs *combinés*.

Estimateurs séparés et combinés

Par exemple, l'estimateur par le quotient séparé remplace $\bar{y}_h$ par $\hat{R}_h\mu_{xh}$, où $\hat{R}_h$ estime le quotient R_h dans la strate h. On a donc L quotients distincts à estimer, $R_h = \dfrac{\mu_{yh}}{\mu_{xh}}$.

Un estimateur combiné remplace les paramètres individuels des strates par un paramètre global. Ainsi, l'estimateur par le quotient combiné définit un seul quotient, $R = \dfrac{\mu_y}{\mu_x}$ et estime μ par $\hat{\mu}_{yq} = \hat{R}\mu_x$ où $\hat{R}$ est un estimateur de R propre à l'échantillonnage stratifié. L'estimateur par la régression combiné est $\hat{\mu}_y + \hat{B}(\mu_x - \hat{\mu}_x)$, où B, μ_y et μ_x sont estimés en tenant compte de la stratification.

8.1 ESTIMATEURS SÉPARÉS

Les estimateurs séparés s'inscrivent dans un modèle général qui les englobe tous. Supposons que pour tout h il existe un estimateur sans biais (ou *approximativement* sans biais) $\hat{\mu}_{yh}$ de μ_h, ainsi qu'un estimateur sans biais (ou approximativement sans biais) $\hat{V}(\hat{\mu}_{yh})$ de $V(\hat{\mu}_{yh})$. Les énoncés suivants, qui découlent des propriétés des espérances et des variances, ne dépendent pas de la nature de ces estimateurs. L'estimateur

(8.1) $$\hat{\mu}_y = \sum_{h=1}^{L} W_h \hat{\mu}_{yh}$$

est sans biais pour μ_{yh}. La variance de $\hat{\mu}_y$ est

(8.2) $$V(\hat{\mu}_y) = \sum_{h=1}^{L} W_h^2 V(\hat{\mu}_{yh}) ,$$

et un estimateur sans biais de $V(\hat{\mu}_y)$ est donné par

(8.3) $$\hat{V}(\hat{\mu}_y) = \sum_{h=1}^{L} W_h^2 \hat{V}(\hat{\mu}_{yh}) .$$

Remarque *À propos de la notation*

Dans ce qui précède, les notations $\hat{\mu}_{yh}$ et $\hat{\mu}_y$ désignent génériquement « l'estimateur de μ_{yh} » et « l'estimateur de μ_y »; de même, la notation « $\hat{V}(\hat{\mu}_y)$ » désigne « l'estimateur de la variance de $\hat{\mu}_y$ ». Elles ne représentent pas des formules spécifiques. Elles prennent, dans chaque application, une forme propre au contexte.

Nous supposons dans ce chapitre que l'échantillon tiré dans chaque strate est un *éas*, et nous considérons les propriétés de $\hat{\mu}_y$ lorsque $\hat{\mu}_{yh}$ est un estimateur pas la différence, par le quotient et par la régression.

Tableau 8.1
Notation

$V(\bar{y}_h)$ =	$V(\bar{y}_{st})$ =	$C(\bar{x}_h ; \bar{y}_h)$ =	$C(\bar{x}_{st} ; \bar{y}_{st})$ =
$(1-f_h)\dfrac{S_{yh}^2}{n_h}$	$\sum_{h=1}^{L} W_h^2 V(\bar{y}_h)$	$(1-f_h)\dfrac{S_{xyh}}{n_h}$	$\sum_{h=1}^{L} W_h^2 C(\bar{x}_h ; \bar{y}_h)$
$\hat{V}(\bar{y}_h)$ =	$\hat{V}(\bar{y}_{st})$ =	$\hat{C}(\bar{x}_h ; \bar{y}_h)$ =	
$(1-f_h)\dfrac{s_{yh}^2}{n_h}$	$\sum_{h=1}^{L} W_h^2 \hat{V}(\bar{y}_h)$	$(1-f_h)\dfrac{s_{xyh}}{n_h}$	$\hat{C}(\bar{x}_{st} ; \bar{y}_{st}) = \sum_{h=1}^{L} W_h^2 \hat{C}(\bar{x}_h ; \bar{y}_h)$

Considérons maintenant les trois cas particuliers annoncés : μ_{yh} estimé par la différence, par le quotient, et par la régression. Leurs propriétés, déjà établies au chapitre 6, sont présentées dans le Tableau 8.2 dans une notation ajustée au contexte: un indice *h* identifie la strate. Par exemple, μ_{xh} représente la moyenne de la variable auxiliaire dans la strate *h*. Une fois les $\hat{\mu}_{yh}$ et les $\hat{V}(\hat{\mu}_{yh})$ calculés, les équations (8.2) et (8.3) donnent la variance de $\hat{\mu}_{yh}$ et l'estimateur de $V(\hat{\mu}_{yh})$.

Tableau 8.2
Propriétés de quelques estimateurs séparés

Estimateur	$\hat{\mu}_{yh}$	$V(\hat{\mu}_{yh})$	$\hat{V}(\hat{\mu}_{yh})$
Par la différence	$\hat{\mu}_{ydh} = \overline{y}_h + (\mu_{xh} - \overline{x}_h)$	$V(\hat{\mu}_{ydh}) =$ $V(\overline{y}_h) + V(\overline{x}_h) - 2C(\overline{x}_h; \overline{y}_h)$	$\hat{V}(\hat{\mu}_{ydh}) =$ $\hat{V}(\overline{y}_h) + \hat{V}(\overline{x}_h) - 2\hat{C}(\overline{x}_h; \overline{y}_h)$
Par le quotient	$\hat{\mu}_{yqh} = \hat{R}\mu_{xh}$ où $\hat{R}_h = \frac{\overline{y}_h}{\overline{x}_h}$.	$V(\hat{\mu}_{yqh}) \approx$ $V(\overline{y}_h) + R_h^2 V(\overline{x}_h) - 2R_h C(\overline{x}_h; \overline{y}_h)$	$\hat{V}(\hat{\mu}_{yqh}) =$ $\hat{V}(\overline{y}_h) + \hat{R}_h^2 \hat{V}(\overline{x}_h) - 2\hat{R}_h \hat{C}(\overline{x}_h; \overline{y}_h)$
Par la régression	$\hat{\mu}_{yrh} = \overline{y}_h + \hat{B}_h(\mu_{xh} - \overline{x}_h)$ $\hat{B}_h = \frac{s_{xyh}}{s_{xh}^2}$	$V(\hat{\mu}_{yrh}) \approx V(\overline{y}_h)(1 - \rho_h^2)$. $\rho_h = \frac{S_{xyh}}{S_{xh}^2}$	$\hat{V}(\hat{\mu}_{yrh}) =$ $\frac{n_h - 1}{n_h - 2}\hat{V}(\overline{y}_h)(1 - r_h^2)$, $r_h = \frac{s_{xyh}}{s_{xh}^2}$.

Exemple 8.1 *Estimateurs séparés*

Une population de 1112 subdivisions de recensement, tirée du tableau A02, est classée selon le genre : Village, Paroisse, Municipalité, et Ville. (Le genre « Village » comprend également le genre « Canton »). On tire un échantillon de taille 50 stratifié selon le genre. Les variables observées sont :

Y : le nombre d'habitants en 2011

X : le nombre d'habitants en 2006

Les résultats sont présentés au Tableau 8.3.

Tableau 8.3

	Strate 1 Village	Strate 2 Paroisse	Strate 3 Municipalité	Strate4 Ville
x	507 1426 388 278 777 642 526 514 226 764	1213 431 1440 722 889 401 1312 1088 5401 832 2307 2644	1464 929 703 2934 1623 1560 1918 936 4002 299 1690 617 24	18088 1686 21599 7336 7322 5088 94703 17016 11584 3041 18586 3309 2587 5197 2395
y	542 1404 368 271 853 652 419 646 229 743	1155 399 1559 790 868 378 1373 1017 6177 826 2365 3035	1469 973 744 2851 1584 1616 1751 856 5122 270 1773 503 178	18208 1677 21555 7719 7462 5135 106322 19295 12889 3118 19447 3198 2457 5073 24980
n_h	10	12	13	15
N_h	92	179	619	222
μ_{xh}	1038,16	1118,63	1512,16	28227,40
$\overline{y}_h$	612,70	1661,83	1514,62	17235,67
$\overline{x}_h$	604,80	1556,67	1455,62	16073,27
s_{yh}	344,79	1619,91	1310,82	25886,11
s_{xh}	341,90	1388,45	1064,21	23022,15
s_{xyh}	115846,60	2246017	1371883	595638795
$\hat{R}_h$	1,01306	1,06756	1,040530	1,072320
r_h	0,98272	0,9986	0,98344	0,99947
$\hat{B}_h$	0,9910	1,16507	1,21133	1,123810

Estimation par la moyenne

Estimations …	Village	Paroisse	Municipalité	Ville
$\hat{\mu}_{yh} = \overline{y}_h$	612,7000	1661,8333	1514,6154	17235,667
$\sqrt{\hat{V}(\overline{y}_h)}$	102,9356	451,6821	359,7174	6454,013

$\hat{\mu}_y = \overline{y}_{st} = 4602{,}25$; $\sqrt{\hat{V}(\hat{\mu}_y)} = \sqrt{\sum_{h=1}^{L} W_h^2 \hat{V}(\overline{y}_h)} = 1306{,}00.$

Estimation par la différence

Estimations …	Village	Paroisse	Municipalité	Ville
$\hat{\mu}_{ydh} = \overline{y}_h + (\mu_{xh} - \overline{x}_h)$	1046,06304	1223,79236	1571,1551	29389,7964
$\sqrt{\hat{V}(\hat{\mu}_{ydh})}$ $= \sqrt{\hat{V}(\overline{y}_h) + \hat{V}(\overline{x}_h) - 2\hat{C}(\overline{x}_h; \overline{y}_h)}$	19,07302	68,22905	89,7746	740,9734

$\hat{\mu}_{yds} = 7025{,}52$; $\sqrt{\hat{V}(\hat{\mu}_{yds})} = \sqrt{\sum_{h=1}^{L} W_h^2 \hat{V}(\hat{\mu}_{ydh})} = 156{,}54.$

Estimation par le quotient

Estimations …	Village	Paroisse	Municipalité	Ville
$\hat{\mu}_{yqh} = \hat{R}\mu_{xh}$	1051,72370	1194,19880	1573,44679	30268,7690
$\sqrt{\hat{V}(\hat{\mu}_{yqh})} = \frac{\mu_{xh}}{\overline{x}_h}\sqrt{\hat{V}(\overline{y}_h) + \hat{R}_h^2\hat{V}(\overline{x}_h) - 2\hat{R}_h\hat{C}(\overline{x}_h;\overline{y}_h)}$	32,92893	32,10745	85,27676	636,52734

$\hat{\mu}_{yqs}$ = 7197,98 ; $\sqrt{\hat{V}(\hat{\mu}_{yqs})} = \sqrt{\sum_{h=1}^{L} W_h^2\hat{V}(\hat{\mu}_{yqh})}$ = 135,779.

Estimation par la régression

Estimations …	Village	Paroisse	Municipalité	Ville
$\hat{\mu}_{yrh} = \overline{y}_h + \hat{B}_h(\mu_{xh} - \overline{x}_h)$	1042,17445	1151,48375	1583,10350	30894,5488
$\sqrt{\hat{V}(\hat{\mu}_{yrh})} = \sqrt{(n-1)\hat{V}(\overline{y}_h)(1-r_h^2)/(n-2)}$	20,20665	25,06495	68,09495	217,75771

$\hat{\mu}_{yrs}$ = 7320,62 ; $\sqrt{\hat{V}(\hat{\mu}_{yqs})} = \sqrt{\sum_{h=1}^{L} W_h^2\hat{V}(\hat{\mu}_{yrh})}$ = 57,84299.

Remarque *Deux estimateurs de la variance*

Dans le dernier exemple, l'estimation de l'écart-type $\sqrt{\hat{V}(\hat{\mu}_{yqh})}$ correspond à la formule (6.8) et comprend le facteur $\frac{\mu_{xh}}{\overline{x}_h}$. Ce facteur aurait pu être omis, conformément à la formule (6.10), présentée comme alternative valable. Il se trouve que dans ce cas le résultat aurait été substantiellement diffèrent: on aurait estimé l'écart-type à 85,9 plutôt que 135,8. Ce résultat sous-estime sérieusement l'écart-type réel de 189 (tel que calculé approximativement à partir des données de la population). Ceci montre que le choix d'inclure ou pas ce facteur peut avoir d'importantes conséquences. En général, il semble que l'inclusion du facteur donne une meilleure estimation de l'écart-type.

8.2 ESTIMATEURS COMBINÉS

Théoriquement, les estimateurs *combinés* (par le quotient et par la régression) devraient être moins précis que les estimateurs séparés. L'estimateur par le quotient séparé est basé sur une estimation distincte de chaque quotient R_h, alors que l'estimateur combiné est basé sur l'estimation d'un quotient global R pour la population entière. Lorsque les R_h diffèrent substantiellement d'une strate à l'autre, on peut s'attendre à ce qu'un seul quotient R ne représente pas aussi bien le rapport des y sur les x. Même problème avec l'estimateur par la régression: l'estimateur combiné est basé sur un seul coefficient B plutôt que sur des coefficients distincts B_h. Mais en pratique, un obstacle se dresse devant les estimateurs séparés: les échantillons des strates peuvent être trop petits pour que, à l'intérieur des strates, les estimations par le quotient et par la régression, dont les propriétés sont approximatives, soient valables.

Les estimateurs par la moyenne et par la différence ne seront pas repris ici car les estimateurs combinés et séparés sont identiques.

La variance des estimateurs combinés fait intervenir la covariance entre deux moyennes stratifiées, définie par

(8.4) $$C(\bar{x}_{st};\bar{y}_{st}) = \sum_{h=1}^{L} W_h C(\bar{x}_h;\bar{y}_h) ,$$

et estimée par

(8.5) $$C(\bar{x}_{st};\bar{y}_{st}) = \sum_{h=1}^{L} W_h C(\bar{x}_h;\bar{y}_h) .$$

ESTIMATEUR PAR LE QUOTIENT COMBINÉ

L'estimateur par le quotient combiné est motivé, ici comme au Chapitre 5, par une décomposition de la moyenne globale de la population en un produit, soit $\mu_y = \frac{\mu_y}{\mu_x}\mu_x = R\mu_x$, où $R = \frac{\mu_y}{\mu_x}$ est le paramètre à estimer. On l'estime en remplaçant le numérateur et le dénominateur par leur estimateur, ce qui, dans un échantillon stratifié, devient

$$\hat{R}_c = \frac{\bar{y}_{st}}{\bar{x}_{xt}} .$$

L'estimateur de μ_y par le quotient combiné est alors

$$\hat{\mu}_{yqc} = \hat{R}_c \mu_x .$$

Comme tout estimateur par le quotient, $\hat{\mu}_{yqc}$ est biaisé mais le biais est négligeable lorsque l'échantillon est assez grand. Le théorème suivant donne une formule approximative de la variance de $\hat{\mu}_{yqc}$.

THÉORÈME 8.1 **VARIANCE DE** $\hat{R}_c$

Si n est assez grand, l'estimateur $\hat{R}_c$ est à peu près sans biais et sa variance est à peu près égale à

$$V(\hat{R}_c) \approx \frac{1}{\mu_x^2}[V(\bar{y}_{st}) + R^2 V(\bar{x}_{st}) - 2R\, C(\bar{x}_{st};\bar{y}_{st})] ,$$

et

$$V(\hat{\mu}_{yqc}) \approx \mu_x^2 V(\hat{R}_c) .$$

On remarque que la formule $V(\hat{\mu}_{yqc})$ est presque identique à celle de $V(\hat{\mu}_{yqs})$, la seule différence étant qu'ici le quotient R est constant : il est le même pour toutes les strates.

On estimera la variance de $\hat{R}_c$ par

$$\hat{V}(\hat{R}_c) \approx \frac{1}{\overline{x}_{st}^2}[\hat{V}(\overline{y}_{st}) + \hat{R}^2\hat{V}(\overline{x}_{st}) - 2\hat{R}_c\hat{C}(\overline{x}_{st};\overline{y}_{st})] \ ,$$

et celle de $\hat{\mu}_{yqc}$ par $\hat{V}(\hat{\mu}_{yqc}) \approx \mu_x^2\hat{V}(\hat{R}_c)$:

$$\hat{V}(\hat{\mu}_{yqc}) = \frac{\mu_x^2}{\overline{x}_{st}^2}[\hat{V}(\overline{y}_{st}) + \hat{R}_c^2\hat{V}(\overline{x}_{st}) - 2\hat{R}_c^2\hat{C}(\overline{x}_{st};\overline{y}_{st})] \, .$$

ESTIMATEUR PAR LA RÉGRESSION COMBINÉ

L'estimateur par la régression combiné est

$$\hat{\mu}_{yrc} = \overline{y}_{st} + \hat{B}_c(\mu_x - \overline{x}_{st}) \text{ , où } \hat{B}_c = \frac{\hat{C}(\overline{x}_{st};\overline{y}_{st})}{\hat{V}(\overline{x}_{st})} \, .$$

La variance de $\hat{\mu}_{yrc}$ est à peu près égale à

$$V(\hat{\mu}_{yrc}) \approx V(\overline{y}_{st})(1-\rho_{st}^2) \text{ , où } \rho_{st} = \frac{C(\overline{x}_{st};\overline{y}_{st})}{\sqrt{V(\overline{x}_{st})}\sqrt{V(\overline{y}_{st})}} \, .$$

La variance est estimée par

$$\hat{V}(\hat{\mu}_{yrc}) = \hat{V}(\overline{y}_{st})(1-r_{st}^2) \text{ , où } r_{st} = \frac{\hat{C}(\overline{x}_{st};\overline{y}_{st})}{\sqrt{\hat{V}(\overline{x}_{st})}\sqrt{\hat{V}(\overline{y}_{st})}} \ .$$

Tableau 8.4

Propriétés des estimateurs par le quotient et par la régression combinés

	Par le quotient	Par la régression
$\hat{\mu}_y$	$\hat{\mu}_{yqc} = \hat{R}_c\mu_x$, où $\hat{R}_c = \frac{\overline{y}_{st}}{\overline{x}_{xt}}$	$\hat{\mu}_{yrc} = \overline{y}_{st} + \hat{B}_c(\mu_x - \overline{x}_{st})$, où $\hat{B}_c = \frac{\hat{C}(\overline{x}_{st};\overline{y}_{st})}{\hat{V}(\overline{x}_{st})}$
$V(\hat{\mu}_y)$	$V(\hat{\mu}_{yqc}) \approx V(\overline{y}_{st}) + R^2V(\overline{x}_{st}) - 2R\,C(\overline{x}_{st};\overline{y}_{st})$	$V(\hat{\mu}_{yrc}) \approx V(\overline{y}_{st})(1-\rho_{st}^2)$, où $\rho_{st} = \frac{C(\overline{x}_{st};\overline{y}_{st})}{\sqrt{V(\overline{x}_{st})}\sqrt{V(\overline{y}_{st})}}$
$\hat{V}(\hat{\mu}_y)$	$\hat{V}(\hat{\mu}_{yqc}) = \frac{\mu_x^2}{\overline{x}_{st}^2}[\hat{V}(\overline{y}_{st}) + \hat{R}_c^2\hat{V}(\overline{x}_{st}) - 2\hat{R}_c\hat{C}(\overline{x}_{st};\overline{y}_{st})]$	$\hat{V}(\hat{\mu}_{yrc}) = \hat{V}(\overline{y}_{st})(1-r_{st}^2)$, où $r_{st} = \frac{\hat{C}(\overline{x}_{st};\overline{y}_{st})}{\sqrt{\hat{V}(\overline{x}_{st})}\sqrt{\hat{V}(\overline{y}_{st})}}$

Exemple 8.2 *Estimation par le quotient et par la régression combinés*

Déterminer les estimateurs de μ_y par le quotient combiné et par la régression combiné. On utilise les données calculées dans l'Exemple 8.1.

D'abord, les données de base :

$\bar{y}_{st}$ = 4602,25; $\bar{x}_{st}$ = 4319,763; $\sqrt{\hat{V}(\bar{y}_{st})}$ = 1306,001; $\sqrt{\hat{V}(\bar{x}_{st})}$ = 1159,109; $\hat{C}(\bar{x}_{st};\bar{y}_{st})$ = 1 512 334; μ_x = 6743,031.

Estimateur par le quotient

$$\hat{R}_c = \frac{4602,25}{4319,763} = 1,065394;$$

$$\hat{\mu}_{yqs} = \hat{R}_c \mu_x = 1,065394\ (6743,031) = 7183,986;$$

$$\sqrt{\hat{V}(\hat{\mu}_{yqc})} = \frac{\mu_x}{\bar{x}_{st}} \sqrt{\hat{V}(\bar{y}_{st}) + \hat{R}_c^2 \hat{V}(\bar{x}_{st}) - 2\hat{R}_c \hat{C}(\bar{x}_{st}, \bar{y}_{st})} = 141,1.$$

Estimateur par la régression

$$\hat{B} = \frac{\hat{C}(\bar{x}_{st};\bar{y}_{st})}{\hat{V}(\bar{x}_{st})} = \frac{1512334}{(1159,109)^2} = 1,12564;$$

$$r_{st} = \frac{\hat{C}(\bar{x}_{st};\bar{y}_{st})}{\sqrt{\hat{V}(\bar{x}_{st})}\sqrt{\hat{V}(\bar{y}_{st})}} = \frac{1512334}{(1159,109)(1306,001)} = 0,9990336;$$

$$\hat{\mu}_{yregc} = \bar{y}_{st} + \hat{B}(\mu_x - \bar{x}_{st}) = 4602,25 + (1,12564)(6743,031 - 4319,763) = 7329,977;$$

$$\sqrt{\hat{V}(\hat{\mu}_{yrc})} = \sqrt{\hat{V}(\bar{y}_{st})}\sqrt{1 - r_{st}^2} = 57,4.$$

8.3 DÉMONSTRATIONS

THÉORÈME 8.1 **ESTIMATEUR PAR LE QUOTIENT COMBINÉ**

Si n est assez grand, l'estimateur $\hat{R}_c$ est à peu près sans biais et sa variance est à peu près égale à $V(\hat{R}_c) \approx \frac{1}{\mu_x^2}[V(\bar{y}_{st}) + R_c^2 V(\bar{x}_{st}) - 2R\ C(\bar{x}_{st};\bar{y}_{st})]$, et $V(\hat{\mu}_{yqc}) \approx \mu_x^2 V(\hat{R}_c)$.

Démonstration

On utilise la méthode de linéarisation (section 2.3) pour obtenir une approximation de la variance de $\hat{R}_c$. Soit $f(\bar{y}_{st};\bar{x}_{st}) = \frac{\hat{\mu}_{yst}}{\hat{\mu}_{xst}}$.

$$f(\bar{y}_{st};\bar{x}_{st}) \approx f(\mu_y;\mu_x) + \frac{\partial f(\mu_y;\mu_x)}{\partial \mu_y}(\bar{y}_{st} - \mu_y) + \frac{\partial f(\mu_y;\mu_x)}{\partial \mu_x}(\bar{x}_{st} - \mu_x) = \frac{\mu_y}{\mu_x} + \frac{1}{\mu_x}(\bar{y}_{st} - \mu_y) - \frac{\mu_y}{\mu_x^2}(\bar{x}_{st} - \mu_x) =$$

$$\frac{1}{\mu_x}\left[\mu_y + (\bar{y}_{st} - \mu_y) - R(\bar{x}_{st} - \mu_x)\right].$$

On approche donc $\hat{R}_c$ par

$$\hat{R}_c \approx \frac{1}{\mu_x}\left[\mu_y + (\bar{y}_{st} - \mu_y) - R(\bar{x}_{st} - \mu_x)\right] = R + \frac{1}{\mu_x}\left[(\bar{y}_{st} - \mu_y) - R(\bar{x}_{st} - \mu_x)\right].$$

Alors $E(\hat{R}_c) \approx R + \frac{1}{\mu_x}\left[E(\bar{y}_{st} - \mu_y) - RE(\bar{x}_{st} - \mu_x)\right] = R.$

On en conclut que $\hat{R}_c$ est à peu près sans biais et donc que $EQM(\hat{R}) \approx V(\hat{R}_c)$ et

$$V(\hat{R}_c) \approx \frac{1}{\mu_x^2}\left[V(\bar{y}_{st} - \mu_y) + R^2V(\bar{x}_{st} - \mu_x) - 2RE[(\bar{x}_{st} - \mu_x)(\bar{y}_{st} - \mu_y)]\right] =$$

$$\frac{1}{\mu_x^2}\left[V(\bar{y}_{st}) + R^2V(\bar{x}_{st}) - 2RC(\bar{x}_{st}; \bar{y}_{st})\right] \text{ et}$$

$$\hat{V}(\bar{y}_{st}) + \hat{R}_c^2\hat{V}(\bar{x}_{st}) - 2\hat{R}_c\hat{C}(\bar{x}_{st}; \bar{y}_{st})$$

RÉSUMÉ

- Le Tableau 8.1 résume la notation utilisée dans ce chapitre et sa suite.
- *Estimateurs séparés* Les estimateurs séparés sont de la forme $\hat{\mu}_y = \sum_{h=1}^{L} W_h\hat{\mu}_{yh}$; leur variance $V(\hat{\mu}_y) = \sum_{h=1}^{L} W_h^2V(\hat{\mu}_{yh})$ est estimée par $\hat{V}(\hat{\mu}_y) = \sum_{h=1}^{L} W_h^2\hat{V}(\hat{\mu}_{yh})$. Ici $\hat{\mu}_{yh}$ et $\hat{V}(\hat{\mu}_{yh})$ désignent génériquement des estimateurs sans biais (ou asymptotiquement sans biais) de la moyenne de la strate et de sa variance. Leur forme exacte, qui varie selon qu'il s'agit d'un estimateur par la moyenne, par la différence, par le quotient ou par la régression, est présentée au Tableau 8.2.
- *Estimateurs combinés* Les estimateurs combinés et leurs propriétés sont résumés au Tableau 8.4.

EXERCICES

8.1 Le Tableau 8.5 résume les données d'un échantillon tiré de la population présentée au tableau A04, stratifiée selon la région géographique. Les variables sont

X : Nombre d'habitants en 2006

Y : Nombre de logements en 2011.

Tableau 8.5
Statistiques descriptives d'un échantillon stratifié tiré de la population du tableau A04

		N_h	n_h	μ_{xh}	$\overline{x}_h$
Région	Est	851	43	2684,82	2565,84
	Centre	1729	88	11 391	7837,25
	Ouest	2254	114	4223,53	3847,55
	Territoires	98	10	1033,78	1025,90
		$\overline{y}_h$	s_{xh}	s_{yh}	s_{xyh}
Région	Est	2651,33	7661,25	8517,99	65 239 166
	Centre	8084,64	21 967	22 906	502 909 016
	Ouest	4129,85	17 483	18 858	329 638 821
	Territoires	1137,90	613,430	624,531	373 377

Soit μ_y le nombre moyen de logements en 2011.

a) Utiliser le nombre d'habitants en 2006 comme variable auxiliaire pour estimer μ_y en utilisant l'estimateur

   i) par la différence;

   ii) par le quotient séparé;

   iii) par la régression séparé.

   Comparer vos résultats avec l'estimateur par la moyenne.

b) Utiliser le nombre d'habitants en 2006 comme variable auxiliaire pour estimer μ_y en utilisant l'estimateur

   i) par le quotient combiné;

   ii) par la régression combiné.

8.2 Supposons, au numéro 8.1, qu'à la strate 4, étant donné la petitesse de l'échantillon, on préfère utiliser l'estimateur par la moyenne.

a) Estimer la moyenne μ_y en utilisant l'estimateur par la moyenne dans la strate 4 et l'estimation par le quotient dans les trois autres strates.

b) Estimer la moyenne μ_y en utilisant l'estimateur par la moyenne dans la strate 4 et l'estimateur par la régression dans les trois autres strates.

8.3 Les estimateurs décrits ici s'appliquent également lorsque les variables — X, Y ou les deux — sont dichotomiques. Considérer les données du tableau A03: une population de 8654 factures de restaurant stratifiée selon le jour de la semaine. On peut identifier d'avance les factures qui comprennent du vin. Utiliser cette information comme variable auxiliaire. Le Tablelau 8.6 présente les données pertinentes de la population ainsi que les résultats d'un échantillon stratifié.

Tableau 8.6
Données d'un échantillon stratifié tiré de la population du tableau A03

Effectifs de la population et de l'échantillon

	Population		Échantillon	
	Repas Sans vin	Repas avec vin	Repas Sans vin	Repas avec vin
Jour de semaine	1345	4075	36	116
Samedi	215	1320	2	33
Dimanche	261	1438	6	32

Sommes et sommes de carrés. Y : montant total de la facture

	Échantillon			
	Repas avec vin		Repas sans vin	
	$\sum_{j\in\omega_h} y_{hj}$	$\sum_{j\in\omega_h} y_{hj}^2$	$\sum_{j\in\omega_h} y_{hj}$	$\sum_{j\in\omega_h} y_{hj}^2$
Jour de semaine	11163,05	1267171	1771	104152
Samedi	3093,50	343448	100,05	5037,403
Dimanche	3278,65	392511	333,5	28143

Soit Y le montant total de la facture. Les montants observés dans l'échantillon sont indiqués dans le tableau. Utiliser l'information sur le nombre de factures comprenant du vin pour estimer μ_y.

a) par le quotient séparé;

b) par la régression séparé;

c) Par le quotient combiné;

d) Par la régression combiné.

8.4 Au numéro 8.3 soit U_d le domaine constitué de tous les repas comprenant du vin, et μ_{yd} le montant moyen des factures dans le domaine. On sait que 6 833 factures dans la population comprennent du vin, mais on n'a pas cette information pour les strates. Estimer μ_{yd} ainsi que l'écart-type de l'estimateur.

8.5 Le tableau A04 présente des données sur 4932 divisions de recensement du Canada.

Soit Y : le nombre d'habitants en 2011, et

X : le nombre d'habitants en 2006.

La population est stratifiée selon la région :

Chacune des strates 1, 2, et 3 sont à nouveau stratifiées en quatre strates selon la valeur de X (la première strate correspond aux plus grandes valeurs de X, la dernière aux plus petites).

Un échantillon stratifié est tiré de chacune des strates 1 à 3 et un *éas* est tiré de la strates 4. Le but de l'échantillonnage est d'estimer la moyenne μ_y. Les échantillons étant relativement petits dans les strates 1 et 2, on a préféré utiliser un estimateur par la moyenne. Dans la strate 3 on estime la moyenne par le quotient séparé. Finalement, dans la strate 4 on utilise l'estimation par le quotient. Voici un résumé des résultats.

Strate 1

n_h	N_h	$\overline{y}_h$	$\sqrt{\hat{V}(\overline{y}_h)}$
21	213	6556,43	1583,652
21	212	1211,380	76,244
21	212	478,90	30,246
21	214	171,476	18,7808

Strate 2

n_h	N_h	$\overline{y}_h$	$\sqrt{\hat{V}(\overline{y}_h)}$
21	432	62364,48	40591,64
21	432	2444,33	189,494
21	432	863,286	37,4598
21	433	373,571	62,0343

Strate 3

n_h	N_h	$\hat{R}_h$	μ_{xh}	$\sqrt{\hat{V}(\hat{\mu}_{yqh})}$
56	564	1,08849	15726	227,084
56	563	1,05614	794,906	15,594
56	562	1,00369	295,285	5,177
56	565	1,01806	65,294	3,230

Strate 4

n	N	$\overline{y}$	$\sqrt{\hat{V}(\overline{y})}$	$\hat{R}$	μ_x	$\sqrt{\hat{V}(\hat{\mu}_{yq})}$
20	98	676,95	101,4791	1,07138	1033,78	40,503

Estimer μ_y ainsi que l'écart-type de l'estimateur.

8.6 On tire un échantillon stratifié de telle sorte que la probabilité d'un échantillon donné est proportionnelle à la moyenne stratifiée $\bar{x}_{st}$ de l'échantillon (la variable X étant supposée connue pour la population entière). Montrer qu'alors l'estimateur de la moyenne par le quotient combiné est sans biais.

8.7 D'une population constituée de L strates de tailles N_h (h = 1,..., L) on compte tirer un échantillon de taille n_h dans la strate h (h = 1,..., L). On procède de la façon suivante: on tire une observation de la population entière avec probabilités proportionnelles aux valeurs x_k d'une variables auxiliaire X, dont les valeurs sont connues pour la population entière. Si cette première observation tombe dans la strate h, on procède au tirage d'un *éas* de taille n_h -1 parmi les N_h - 1 autres unités de la strate h. Ensuite un tire un *éas* de chacune des L-1 autres strates (de taille n_ℓ dans la strate $\ell \neq h$).

a) Soit $\omega = \omega_1 \cup \omega_2 \cup \ldots \cup \omega_L$ l'échantillon obtenu, ω_h étant l'échantillon de taille n_h tiré de la strate h. Montrer que $p(\omega) = \dfrac{\bar{x}_{st}}{\mu_x \prod_{h=1}^{L} \binom{N_h}{n_h}}$.

b) En déduire que $\hat{R} = \dfrac{\bar{y}_{st}}{\bar{x}_{st}}$ est un estimateur sans biais de $R = \dfrac{\mu_y}{\mu_x}$.

9 Échantillonnage par grappes à un degré et probabilités égales

Les unités d'une population appartiennent souvent à des groupes naturels : chaque logement urbain se trouve dans un quartier; chaque cégépien est inscrit à un cégep; chaque électeur vit dans une circonscription. Ces regroupements fournissent un moyen de sélectionner des unités lorsqu'un échantillonnage aléatoire simple se révèle impossible ou trop coûteux. Impossible, dans le cas des cégépiens, si on ne dispose pas d'une liste de tous les étudiants. Ou trop coûteux, comme dans le cas des logements, si l'information recherchée ne peut être prélevée que par une interview personnelle: les logements sélectionnés peuvent se trouver dispersés un peu partout dans la ville, ce qui entraîne d'importants coûts de déplacement. Il est normal, certes, que l'intervieweur rejoigne des logements un peu partout dans la ville, mais il peut mieux rentabiliser son déplacement s'il en profite pour inclure plusieurs logements d'un même voisinage.

On pourrait songer à un échantillonnage stratifié — dans lequel les quartiers, les cégeps, les circonscriptions servent de strates. Mais un échantillonnage stratifié exige qu'on tire un échantillon dans *chacune* des strates, ce qui risque d'être onéreux, le nombre de strates étant, dans les exemples mentionnés, beaucoup trop élevé (bien que peut-être pas dans le cas des cégeps). Pour ces raisons purement pratiques, on peut juger préférable de prélever un échantillon de *quartiers* plutôt que de logements; ou de *cégeps* plutôt que d'étudiants. Les quartiers et les cégeps sont alors appelés des *grappes* ou *unités primaires*.

9.1 INTRODUCTION, DÉFINITIONS ET NOTATION

Considérons l'ensemble des logements d'une ville. Afin d'échantillonner les logements, on procède d'abord par une sélection de quartiers — les unités primaires. L'échantillon est alors constitué de tous les logements — les unités secondaires — des quartiers sélectionnés.

Il existe plusieurs façons d'échantillonner par grappes et d'estimer les paramètres. Dans ce chapitre, on se concentre sur la procédure suivante :

- On tire un *éas* — donc avec *probabilités égales* — d'unités primaires ou grappes. D'autres modes d'échantillonnage, permettant des tirages avec probabilités *inégales* seront traités au chapitre 10.
- On saisit les données de *toutes* les sous-unités des grappes sélectionnées.
  Cela en fait un tirage *à un degré*, par opposition à un tirage à *deux degrés*, qui consiste à

échantillonner plutôt qu'à *recenser* les sous-unités des grappes sélectionnées. Au chapitre 11, nous développerons des méthodes d'échantillonnage à deux degrés (ou plus).

Dans le contexte particulier traité ici, aucune nouvelle technique n'est nécessaire. Il s'agit, après tout, d'un *éas*. Les techniques et les formules développées aux chapitres 4, 6 et 7 peuvent s'appliquer intégralement. On introduira quelques changements de notation afin de refléter le contexte. Ces changements ne sont pas indispensables mais ils faciliteront le passage progressif aux idées des prochains chapitres. Il sera question de trois estimateurs connus : les estimateurs par la moyenne, par le quotient et par la régression. Nous illustrons leur application dans un même cadre: on tire un échantillon de 15 quartiers dans une ville composée de 180 quartiers afin d'estimer le nombre d'habitants dans la ville. Dans chacun des quartiers sélectionnés, on observe le nombre de logements et le nombre total de personnes dans le quartier. Le Tableau 9.1 présente les résultats. Les observations et les paramètres sont les suivants :

N : Nombre de quartiers = 180.

M_k : Nombre de logements dans le quartier k.

t_k : Nombre total de personnes dans le quartier k.

τ : Le nombre total de personnes dans la population (le paramètre à estimer).

Les valeurs de la variable principale sont maintenant désignées par t_k, et celles d'une deuxième variable (qui jouera le rôle de variable auxiliaire) par M_k. On suppose les M_k connues pour la population entière. Soit

$$M_o = \sum_{k=1}^{N} M_k \text{ = Nombre de logements dans la population = 3 500.}$$

Outre le fait que nous écrivons t_k au lieu de y_k et M_k au lieu de x_k, la situation ici n'a rien de nouveau.

On se concentrera sur le paramètre $\tau = \sum_{k=1}^{N} t_k$, le nombre total de personnes dans la population. Deux autres paramètres sont intimement liés à τ, soit

$\mu_t = \dfrac{\tau}{N} = \dfrac{\text{Nombre de personnes}}{\text{Nombre de quartiers}}$: Le nombre moyen de personnes par quartier; et

$\mu_y = \dfrac{\tau}{M_o} = \dfrac{\text{Nombre de personnes}}{\text{Nombre de logements}}$: Le nombre moyen de personnes par logement.

Tableau 9.1
Échantillon de n =15 quartiers tirés d'une population de N = 180 quartiers ayant en tout M_0 = 3500 logements

Unité (i)	Nombre de logements (M_i)	Nombre de personnes (t_i)	Personnes par logement (t_i / M_i)
1	25	41	1,64
2	4	11	2,75
3	25	58	2,32
4	20	32	1,60
5	20	32	1,60
6	10	30	3,00
7	4	13	3,25
8	25	57	2,28
9	15	36	2,40
10	16	32	2,00
11	20	28	1,40
12	22	55	2,50
13	10	24	2,40
14	14	49	3,50
15	24	54	2,25
Somme	254	552	34,89
Moyenne	16,93333	36,8	2,3260
Écart-type	7,284687	15,21372	0,6212752

Tout estimateur $\hat{\tau}$ de τ mène naturellement à un estimateur de μ_t et à un estimateur de μ_y. soit

$$\hat{\mu}_t = \frac{\hat{\tau}}{N} \text{ et } \hat{\mu}_y = \frac{\hat{\tau}}{M_0}.$$

De même, un estimateur $\sqrt{\hat{V}(\hat{\tau})}$ de $\sqrt{V(\hat{\tau})}$ donne lieu aux estimateurs

$$\sqrt{\hat{V}(\hat{\mu}_t)} = \frac{\sqrt{\hat{V}(\hat{\tau})}}{N} \text{ et } \sqrt{\hat{V}(\hat{\mu}_y)} = \frac{\sqrt{\hat{V}(\hat{\tau})}}{M_0}.$$

Nous allons donc limiter les formules à celle concernant l'estimation de τ.

Remarque *Moyenne par grappe et moyenne pas sous-unité*

Les notations μ_t et μ_y pourraient laisser croire qu'il s'agit de deux variables distinctes, t et Y, mais ce n'est pas le cas: μ_t est une moyenne calculée sur les grappes, donc une somme de N nombres divisée par N; alors que μ_y est calculée sur les sous-unités, donc une somme de M_0 nombres divisée par M_0. Mais il s'agit du nombre de personnes dans les deux cas.

9.2 ESTIMATION PAR LA MOYENNE

Considérant chaque grappe comme une unité indivisible, nous sommes devant un problème identique à celui introduit au Chapitre 3: une population dont les unités sont des grappes, un *éas* de grappes, une variable dont les valeurs t_k sont associées aux grappes, et un total τ à estimer. La moyenne par grappe est estimée par $\bar{t} = \frac{1}{n}\sum_{i\in\omega} t_i$ (ω désigne l'ensemble des n grappes de l'échantillon); et le total τ est estimé par $\hat{\tau}_m = N\bar{t}$. L'écart-type de $\hat{\tau}_m$ est $\sqrt{V(\hat{\tau}_m)} = N\sqrt{1-f}\,\frac{S_t}{n}$, où $f = n/N$ et $S_t = \sqrt{\frac{1}{N-1}\sum_{k=1}^{N}(t_k - \mu_t)^2}$. L'écart-type $\sqrt{V(\hat{\tau}_m)}$ est bien sûr estimé par $\sqrt{\hat{V}(\hat{\tau}_m)} = N\sqrt{1-f}\,\frac{s_t}{n}$, où $s_t = \sqrt{\frac{1}{n-1}\sum_{i\in\omega}(t_i - \bar{t})^2}$.

Exemple 9.1 *Estimation par la moyenne*

Utiliser les données du Tableau 9.1 pour estimer par la moyenne le nombre moyen de personnes par quartier, le nombre total de personnes dans la ville, et le nombre moyen de personnes par logement. La population comprend $N = 180$ quartiers et un total de 3 500 personnes.

Solution

Quelques données: $n = 15$; $N = 180$; $\bar{M} = 16{,}9333$; $\bar{t} = 36{,}8$.

Estimations

La moyenne $\bar{t} = 36{,}8$ estime le nombre moyen de personnes par quartier.

Le nombre total de personnes dans la population est estimé par $\hat{\tau}_m = N\bar{t} = 180(36{,}8) = 6624$.

Le nombre moyen de personnes par logements, $\mu_y = \frac{\tau}{M_o}$, est estimé par

$$\hat{\mu}_{ym} = \frac{\hat{\tau}_m}{M_o} = \frac{6624}{3500} = 1{,}892571 .$$

Estimation des écarts-types

$$s_t = \sqrt{\frac{\sum_{i\in\omega}(t_i - \bar{t})^2}{n-1}} = 15{,}21372 .$$

Les écarts-type de ces estimateurs sont estimés par :

$$\sqrt{\hat{V}(\bar{t})} = \sqrt{1-f}\,\frac{s_t}{n} = \sqrt{1-15/180}\,\frac{15{,}21372}{\sqrt{15}} = 3{,}760931 \ ;$$

$$\sqrt{\hat{V}(\hat{\tau}_m)} = N\sqrt{\hat{V}(\bar{t})} = N\sqrt{1-f}\,\frac{s_t}{n} = 180\sqrt{1-15/180}\,\frac{15{,}21372}{\sqrt{15}} = 676{,}9676 ;$$

$$\sqrt{\hat{V}(\hat{\mu}_{ym})} = \frac{\sqrt{\hat{V}(\hat{\tau}_m)}}{M_o} = \frac{676{,}9676}{3500} = 0{,}1934193 .$$

9.3 ESTIMATION PAR LE QUOTIENT

L'estimateur par le quotient utilise la taille des unités primaires (les M_k) comme variable auxiliaire. Ainsi donc la moyenne par unité secondaire μ_y est estimée par $\hat{\mu}_{yq} = \bar{\bar{y}} = \frac{\sum_{i\in\omega} t_i}{\sum_{i\in\omega} M_i}$; τ est estimé par $\hat{\tau}_q = M_0\bar{\bar{y}}$; et la moyenne par unité primaire μ_t est estimée par $\hat{\mu}_{tq} = \frac{\hat{\tau}_q}{N} = \bar{\bar{y}}\mu_M$, où $\mu_M = \frac{\sum_{i=1}^{N} M_i}{N} = \frac{3500}{180} = 19{,}444$. L'écart-type de $\hat{\mu}_{yq}$ est $\sqrt{V(\hat{\mu}_{yq})}$ $= \sqrt{1-f}\,\frac{\sqrt{S_t^2 + \mu_y^2 S_m^2 - 2\mu_y S_{mt}}}{\sqrt{n}\,\mu_M}$. Il est estimé par $\sqrt{\hat{V}(\hat{\mu}_{yq})} = \sqrt{1-f}\,\frac{\sqrt{s_t^2 + \bar{\bar{y}}^2 s_m^2 - 2\bar{\bar{y}}s_{mt}}}{\sqrt{n}\,\bar{M}}$. Nous avons par conséquent les estimateurs suivants:

$$\sqrt{\hat{V}(\hat{\tau}_q)} = M_0\sqrt{1-f}\,\frac{\sqrt{s_t^2 + \bar{\bar{y}}^2 s_m^2 - 2\bar{\bar{y}}s_{mt}}}{\sqrt{n}\,\bar{M}} ; \ \sqrt{\hat{V}(\hat{\mu}_{tq})} = \mu_M\sqrt{1-f}\,\frac{\sqrt{s_t^2 + \bar{\bar{y}}^2 s_m^2 - 2\bar{\bar{y}}s_{mt}}}{\sqrt{n}\,\bar{M}} .$$

Exemple 9.2 *Estimation par le quotient*

Utiliser les données du Tableau 9.1 pour estimer par le quotient le nombre moyen de personnes par logement, le nombre total de personnes dans la ville et le nombre moyen de personnes par quartier.

Solution

Quelques calculs :

$$\bar{\bar{y}} = \frac{\sum_{i\in\omega} t_i}{\sum_{i\in\omega} M_i} = 2{,}173228 \; ; \; s_m = \sqrt{\frac{\sum_{i\in\omega}(M_i - \bar{M})^2}{n-1}} = 7{,}284687 \; ; \; s_{mt} = \frac{\sum_{i\in\omega}(M_i - \bar{M})(t_i - \bar{t})}{n-1} = 90{,}91429 \; .$$

Estimations

Le nombre moyen de personnes par logement est estimé par $\bar{\bar{y}} = 2{,}173228$; le nombre total de personnes est estimé par $\hat{\tau}_q = M_0 \bar{\bar{y}} = 7606{,}299$; et le nombre moyen par quartier est estimé par $\bar{\bar{y}}\mu_M = 42{,}25722$.

Estimation des écarts-types

$$\sqrt{\hat{V}(\bar{\bar{y}})} = \sqrt{1-f}\,\frac{\sqrt{s_t^2 + \bar{\bar{y}}^2 s_m^2 - 2\bar{\bar{y}} s_{mt}}}{\sqrt{n}\bar{M}} = 0{,}1361153 \; ; \sqrt{\hat{V}(\hat{\mu}_{yq})} = \mu_M \sqrt{\hat{V}(\bar{\bar{y}})} = \frac{3500}{180}(0{,}1361153) = 2{,}646687 \; ;$$

$$\sqrt{\hat{V}(\hat{\tau}_q)} = M_0 \sqrt{\hat{V}(\bar{\bar{y}})} = 476{,}4037 \; .$$

Remarque *$\bar{\bar{y}}$ est la moyenne pondérée des moyennes des sous-unités*

La notation $\bar{\bar{y}}$ se justifie par le fait que c'est une moyenne (pondérée) de moyennes : $\frac{\sum_{i\in\omega} t_i}{\sum_{i\in\omega} M_i}$ peut s'écrire comme $\sum_{i\in\omega} \alpha_i \bar{y}_i$, où $\bar{y}_i = \frac{t_i}{M_i}$ et $\alpha_i = \frac{M_i}{\sum_{j\in\omega} M_j}$.

9.4 ESTIMATION PAR LA RÉGRESSION

L'estimateur par la régression de la moyenne par unité primaire est $\hat{\mu}_t = \bar{t} + \hat{B}(\mu_M - \bar{M})$, où $\hat{B} = \frac{s_{mt}}{s_m^2}$. Le total est donc estimé par $\hat{\tau}_r = N\hat{\mu}_{tr}$ et la moyenne par unité secondaire par $\hat{\mu}_{yr} = \frac{\hat{\tau}_r}{M_0}$.

Les écarts-types approximatifs sont estimés par, respectivement,

$$\sqrt{\hat{V}(\hat{\mu}_t)} = \sqrt{1 - n/N}\,\frac{\sqrt{n-1}}{\sqrt{n-2}}\,\frac{s_t}{\sqrt{n}}\sqrt{1-r^2} \quad ,$$

où

$$r = \frac{s_{mt}}{s_m s_t} \; ; \; \sqrt{\hat{V}(\hat{\tau}_r)} = N\sqrt{\hat{V}(\hat{\mu}_{tr})} \; ; \; \sqrt{\hat{V}(\hat{\mu}_{yr})} = \frac{\sqrt{\hat{V}(\hat{\tau}_r)}}{M_0} \; .$$

Exemple 9.3 *Estimation par la régression*

Utiliser les données du Tableau 9.1 pour estimer par la régression le nombre moyen de personnes par quartier, le nombre total de personnes dans la ville, et le nombre moyen de personnes par logement.

Solution

s_m =7,284687; s_t =15,21372 ; s_{mt} = 90,91429.

$\hat{B} =$ 1,713209 ; $r = \dfrac{90,91429}{(7,284687)(15,21372)} = 0,820325.$

$\hat{\mu}_{tr} = \bar{t} + \hat{B}(\mu_M - \bar{M}) = 36,8 + 1,713209(19,44444 - 16,933333) = 41,10206$. $\hat{\tau}_r =$ $180(41,10206) = 7398,37$;
$\hat{\mu}_{yr} = \dfrac{7398,37}{3500} = 2,11382$.

$\sqrt{\hat{V}(\hat{\mu}_{tr})} = \sqrt{1-15/180}\,\dfrac{\sqrt{15-1}}{\sqrt{15-2}}\,\dfrac{15,21372}{\sqrt{15}}\sqrt{1-(0,820325)^2} = 2,232061$. $\sqrt{\hat{V}(\hat{\tau}_r)} = N\sqrt{\hat{V}(\hat{\mu}_{tr})} =$ 401,771 ;
$\sqrt{\hat{V}(\hat{\mu}_{yr})} = \dfrac{\sqrt{\hat{V}(\hat{\tau}_r)}}{M_o} = 0,1147917$.

COMPARAISON DES VARIANCES

On peut s'attendre à ce que les estimateurs par le quotient et par la régression soient plus efficaces que l'estimateur par la moyenne. Si le facteur de correction est négligeable, les variances approximatives des trois estimateurs de τ peuvent s'écrire comme ceci:

- *Estimateur par la moyenne* : $V_m \approx \dfrac{N^2 S_t^2}{n}$.
- *Estimateur par le quotient* : $V_q \approx \dfrac{N^2}{n(N-1)} \sum_{k=1}^{N} M_k^2 (\bar{y}_k - \mu_y)^2$ [$\bar{y}_k = t_k / M_k$].
- *Estimateur par la régression* : $V_r \approx \dfrac{N^2 S_t^2}{n}(1-\rho^2)$.

On sait que l'estimateur par le quotient est plus précis que l'estimateur par la moyenne lorsque la corrélation entre la variable d'intérêt et la variable auxiliaire est élevée. C'est sûrement le cas ici, puisque la variable auxiliaire M_k est le nombre d'unités secondaires dans la grappe k et la variable d'intérêt t_k est la somme des valeurs de Y des unités secondaires de la grappe. On peut s'attendre à ce que $V_q \leq V_m$. Dans la mesure où l'échantillon est assez grand pour valider les formules approximatives, il est clair que $V_r \leq V_m$ et que la différence dépend entièrement du coefficient de corrélation entre les totaux des strates et leur taille.

Remarque *Échantillon par grappes vs un éas d'unités secondaires*

En général, l'échantillonnage par grappes est moins efficace qu'un *éas* comprenant le même nombre d'unités secondaires. (Un échantillon de deux quartiers comprenant 20 maisons chacune ne vaut pas un *éas* de 40 maisons). On y recourt souvent quand même, et ce pour deux raisons :

1) Il est parfois impossible de faire autrement, à cause d'une base de sondage insuffisante. Dans le dernier exemple, il aurait fallu disposer d'une liste de logements pour tirer un *éas* de logements.

2) Un échantillon par grappes est généralement moins efficace qu'un *éas* dans le sens suivant: pour un même nombre d'unités secondaires, l'estimateur dans un *éas* a une plus petite variance. Mais il n'est pas nécessairement moins efficace lorsqu'on tient compte des coûts de l'échantillonnage. Le coût par unité secondaire d'un échantillon par grappes peut être moins élevé, ce qui permet de tirer de plus grands échantillons.

Remarque *La moyenne non pondérée des moyennes est biaisée*

La dernière colonne du Tableau 9.1 donne la moyenne par logement $\overline{y}_i$ dans chaque quartier. Leur moyenne $\hat{\mu}_y^* = \frac{1}{n}\sum_{i\in\omega}\overline{y}_i$ est-elle un estimateur sans biais de la moyenne par quartier $\mu_y = \frac{\sum_{i=1}^{N} t_i}{M_0}$? La réponse est, en général, non: l'espérance de $\hat{\mu}_y^*$ est $E\left(\frac{1}{n}\sum_{i\in\omega}\overline{y}_i\right) = \frac{1}{N}\sum_{i=1}^{N}\overline{y}_i = \frac{1}{N}\sum_{k=1}^{N}\frac{t_k}{M_k} \neq \frac{\sum_{k=1}^{N} t_k}{M_0}$. C'est la pondération (uniforme) des $\overline{y}_i$ dans l'estimateur $\hat{\mu}_y^*$ qui fait défaut. Car μ_y est une moyenne des μ_i *pondérée* par la taille des grappes : $\mu_y = \frac{\sum_{i=1}^{N} M_i\mu_i}{M_0} = \sum_{i=1}^{N}\left(\frac{M_i}{M_0}\right)\mu_i$, alors que $\hat{\mu}_y^*$ ne pondère pas les moyennes. Et si on pondérait les $\overline{y}_i$? Voyons : la moyenne pondérée est $\sum_{i\in\omega}\left(\frac{M_i}{\sum_{i\in\omega} M_i}\right)\overline{y}_i = \sum_{i\in\omega}\left(\frac{M_i\overline{y}_i}{\sum_{i\in\omega} M_i}\right) = \frac{\sum_{i\in\omega} M_i\overline{y}_i}{\sum_{i\in\omega} M_i} = \frac{\sum_{i\in\omega} t_i}{\sum_{i\in\omega} M_i}$: on tombe sur l'estimateur par le quotient.

9.5 TIRAGE SYSTÉMATIQUE

L'échantillonnage systématique — un mode d'échantillonnage très populaire — est un type particulier de tirage par grappes. Il suppose, d'abord, un certain positionnement des unités les unes par rapport aux autres, un certain ordre prédéfini. Par exemple, les livres d'une bibliothèque rangés sur des tablettes; les fiches médicales des patients d'une clinique

disposées en ordre alphabétique; les plants de tomates dans une rangée; les spectateurs d'un film défilant à la sortie. L'échantillon systématique consiste alors à tirer les unités de la population à des intervalles réguliers, ceci à partir d'un point de départ aléatoire.

Exemple 9.4 *Tirage systématique*

On doit prélever un échantillon de patients dans une clinique médicale où les dossiers des patients sont rangés dans des tiroirs en ordre alphabétique. La population est de 1 000 patients et on souhaite tirer un échantillon de taille $n = 20$. On découpe la population en 20 tranches de 50 dossiers consécutifs, puis on tire au hasard un dossier parmi les 50 premiers. Après quoi, on sélectionne chaque 50e dossier. Par exemple, si le premier dossier, tiré au hasard, est le 20e, l'échantillon sera constitué des dossiers numérotés 20, 70, 120, …, 920, et 970. L'échantillon est donc entièrement déterminé par le premier. Le seul aspect aléatoire du tirage est le choix de cette première unité.

Exemple 9.5 *Tirage systématique inévitable*

Pour sonder l'auditoire d'une pièce de théâtre, on s'installe à la porte du théâtre et on interpelle chaque dixième spectateur à la sortie de la salle. Remarquez qu'il n'est pas nécessaire de disposer d'une liste des membres de la population ou même d'en connaître la taille. C'est là un autre avantage de l'échantillonnage systématique.

Exemple 9.6 *Tirage systématique recommandable*

Afin de tirer 20 éprouvettes de terre dans une rangée de 800 mètres de long, on tire un échantillon à chaque 40e mètre de la rangée.

Un des attraits de l'échantillonnage systématique est purement pratique : il est facile à exécuter lorsque la position physique des unités s'y prête. Mais on peut aussi y voir un avantage théorique : un échantillon systématique donne l'assurance que chaque partie de la population sera représentée. Cet avantage n'est pas évident dans le dernier exemple, mais son intérêt est clair s'il s'agit, par exemple, d'échantillonner des plants de tomates alignées en rangées. Un échantillon systématique comprendra nécessairement des unités issues des différentes parties du terrain, alors que, théoriquement, un *éas* peut ne contenir que des plants dans les parties les plus fertiles (ou les moins fertiles) du terrain. C'est l'un des attraits de la méthode.

Mais si un tirage systématique a beaucoup d'attraits, il a par contre un désavantage de taille: *il ne permet pas d'estimer l'écart-type de l'estimateur*. Par conséquent, on ne peut pas estimer une marge d'erreur ou déterminer un intervalle de confiance.

Illustration de la méthode

Un fabricant de jus qui reçoit 200 lots de mangues en une semaine voudrait estimer la teneur en sucre Y du jus produit par l'ensemble des lots. Y se mesure en grammes par tasse et le paramètre à estimer est la moyenne par lot.

On compte estimer la moyenne μ_y à partir d'un échantillon systématique de 8 lots. On tirera donc un lot au hasard parmi les 25 premiers, après quoi on tirera chaque vingt-cinquième lot. Par exemple, si le lot tiré est le dix-septième, alors l'échantillon sera constitué des unités 17, 42, 67, 92, 117, 142, 167, et 192.

Les valeurs de Y sont présentées au Tableau 9.2 dans l'ordre d'arrivée (à lire colonne par colonne). La disposition est donc telle que chaque ligne est l'un des 25 échantillons possibles. L'espace échantillon Ω se trouve donc affiché dans le tableau. La probabilité de chaque échantillon ω est $p(\omega) = 1/25$. Les deux dernières colonnes donnent les moyennes et les variances des échantillons. Donc l'espérance de $\bar{y}$ est la moyenne de tous les $\bar{y}$, soit $\sum_{\omega\in\Omega} \bar{y}_\omega p(\omega) =$ 37,355. Cette moyenne étant évidemment égale à la moyenne μ de la population, on peut affirmer que $\bar{y}$ est sans biais pour μ.

Tableau 9.2
Échantillons systématiques d'une population de taille 200

Échantillon									$\bar{y}$	s^2
1	40	32	35	31	29	41	48	32	36,000	41,714
2	37	43	48	31	27	33	49	33	37,625	66,554
3	41	27	38	42	27	47	34	33	36,125	51,554
4	25	40	32	33	39	32	47	33	35,125	44,411
5	32	28	30	39	37	27	34	43	33,750	31,357
6	44	28	36	30	50	37	50	44	39,875	71,554
7	43	50	39	32	40	26	33	47	38,750	65,071
8	38	43	27	41	28	44	37	44	37,750	46,786
9	41	27	44	35	31	32	36	34	35,000	29,714
10	28	36	35	44	34	41	47	45	38,750	42,786
11	31	43	29	38	50	41	45	37	39,250	49,357
12	42	48	30	38	43	38	42	33	39,250	33,357
13	35	46	29	41	31	30	46	44	37,750	53,643
14	31	25	35	30	37	50	35	26	33,625	62,268
15	31	43	28	28	48	48	42	41	38,625	70,839
16	33	30	31	27	27	47	26	34	31,875	45,839
17	36	49	33	31	26	48	40	41	38,000	65,143
18	37	32	49	28	40	35	39	27	35,875	50,982
19	42	30	27	35	32	42	45	38	36,375	41,411
20	26	41	47	45	35	45	48	38	40,625	55,125
21	26	50	26	35	41	29	31	41	34,875	72,982
22	45	44	31	26	38	44	45	42	39,375	51,982
23	48	31	39	29	50	38	45	35	39,375	59,696
24	49	41	49	48	44	34	48	32	43,125	46,982
25	34	45	36	43	27	48	31	33	37,125	54,696
	$E(\bar{y})$ = Moyenne des moyennes :								37,335	52,23214

Remarques

- Entre l'échantillonnage aléatoire simple et l'échantillonnage systématique, la différence la plus frappante d'un point de vue théorique est dans la cardinalité de l'espace échantillon. Dans le cas présent, le nombre d'échantillons possibles est 25. Un *éas* aurait généré un espace échantillon de 55 098 996 177 225 échantillons.
- Un échantillon systématique est essentiellement un échantillon par grappes. Dans le Tableau 9.2, chaque ligne est une grappe, et la procédure consiste à tirer au hasard l'une des 25 grappes. C'est donc un échantillon de taille 1, et c'est pour cela que nous ne pouvons par estimer la variance.

Variance et estimation de la variance

La variance de $\bar{y}$ est $V(\bar{y}) = \sum_{\omega \in \Omega} (\bar{y}_\omega - \mu)^2 p(\omega)$. On peut la calculer ici car on dispose de toutes les données de la population. On obtient $V(\bar{y})$ = 6,08585. Un *éas* de taille n = 8 donnerait $V(\bar{y})$ = 6,245907, assez proche de la variance pour un échantillon systématique. Cet écart relativement peu important entre la variance d'un *éas* et d'un échantillon systématique est probablement la norme. Le problème est plutôt dans l'estimation de la variance. Dans le présent exemple, supposons qu'on estimait S^2 par s^2, comme on l'aurait fait avec un *éas*. Les valeurs de s^2 sont indiquées dans la dernière colonne du Tableau 9.2. L'espérance de s^2 (la moyenne des 25 valeurs de s^2) est 52,23214. La variance de la population est S^2 = 52,04922.

Donc dans cet exemple, les propriétés d'un échantillon systématique ont l'air comparables à celles d'un *éas*. Mais qu'en est-il de l'échantillonnage systématique *en général* ? Règle générale, on peut utiliser la moyenne échantillonnale $\bar{y}$ pour estimer la moyenne de la population, tout comme avec un *éas*. C'est un estimateur sans biais si N est un multiple de n; et « à peu près » sans biais sinon (voir l'exercice 9.10).

Discussion concernant l'estimation de la variance

En principe, s^2 n'est pas un estimateur sans biais de S^2 dans un échantillon systématique et donc l'estimateur $\sqrt{\hat{V}(\bar{y})} = \sqrt{1-\frac{n}{N}}\frac{s}{\sqrt{n}}$ de $\sqrt{V(\bar{y})}$ n'est pas valide dans un échantillon systématique. Il faut savoir que, pourtant, certains analystes et statisticiens l'utilisent quand même. Cette pratique est-elle défendable? Peut-on considérer que, malgré le mode de tirage particulier, l'échantillon est « tout comme » un *éas*? La réponse est oui si l'ordre dans lequel les unités sont disposées est aléatoire.

Commençons par une situation où l'ordre est incontestablement aléatoire. Une statisticienne est en mesure de placer les unités dans l'ordre qu'elle veut (ce qu'elle peut faire avec des fichiers en carton mais pas avec des plants de tomates). Elle dispose donc les données en utilisant un mécanisme de permutations aléatoires. Ensuite, elle tire un échantillon systématique. Dans ce cas, l'échantillon est réellement un *éas*. La procédure n'est qu'une

façon de réaliser un *éas*, une procédure en deux étapes: la première dispose les unités dans un ordre aléatoire, la deuxième tire un échantillon systématique.

En réalité, on ne manipule pas l'ordre dans lequel se trouvent les unités. Seule la deuxième étape est sous le contrôle de l'expérimentateur; la première a été effectuée par la nature. Reste à savoir si la nature a disposé les unités dans un ordre aléatoire. On peut probablement le considérer comme tel s'il s'agit de spectateurs défilant à la sortie d'un théâtre, parce que ce que dit le répondant n'a probablement pas grand-chose à voir avec le moment de sa sortie. Ce serait plus douteux si l'on parle de plants de tomates.

Nous n'irons pas plus loin dans cette discussion, sauf pour dire que certaines approches modernes en échantillonnage, basées sur des modèles, tentent de clarifier davantage cette problématique.

RÉSUMÉ

- D'une population constituée de N grappes on tire un *éas* de n unités primaires. Une variable Y est définie sur toutes les unités secondaires. Les estimateurs par la moyenne, par le quotient et par la régression sont basés sur les totaux t_i des unités secondaires et les tailles M_i comme variable auxiliaire.
- Dans les deux tableaux qui suivent, nous définissons la notation utilisée dans ce chapitre et particularisons entre parenthèses leur sens dans le contexte d'une population de maisons groupées en quartiers.
- Quantités définies sur la population :

Tableau 9.3

Notation, échantillonnage par grappes à un degré

Définition générale	*Cas particulier*
N : Nombre de grappes dans la population	Nombre de quartiers dans la population
M_k : Nombre d'unités secondaires dans la grappe k	Nombre de logements dans le quartier k
$M_o = \sum_{k=1}^{N} M_k$: Nombre d'unités secondaires dans la population	Nombre de logements dans la ville
$\mu_M = \frac{\sum_{k=1}^{N} M_k}{N}$: La taille moyenne des grappes	Nombre moyen de logements par quartier
y_{kj} : Valeur de y correspondant à l'unité secondaire j de l'unité primaire k	Nombre de personnes dans le logement j du quartier k
t_k : Somme des valeurs de y dans la grappe k	Nombre total de personnes dans le quartier k
$\tau = \sum_{k=1}^{N} t_k$: Total des y de la population	Nombre de personnes dans la ville
μ_t : Moyenne des totaux t_k des grappes : $\mu_t = \frac{\tau}{N}$	Nombre moyen de personnes par quartier dans la population
$\mu_y = \frac{\tau}{M_o}$: Moyenne par unité secondaire des valeurs de y	Nombre moyen de personnes par ménage dans la population
$S_t = \sqrt{\frac{\sum_{k=1}^{N} (t_k - \mu_k)^2}{N-1}}$: Écart-type des totaux des grappes	Dispersion du nombre total de personnes dans les quartiers (écarts entre les quartiers)
$S_m = \sqrt{\frac{\sum_{k=1}^{N} (M_k - \mu_M)^2}{N-1}}$: L'écart-type des tailles M_i des grappes	La dispersion des tailles des quartiers
$S_{tm} = \frac{\sum_{k=1}^{N} (t_k - \mu_t)(M_k - \mu_M)}{N-1}$: La covariance entre les t_i et les M_i	Covariance entre le nombre de personnes et le nombre de logements dans un quartier

- Quantités définies sur l'échantillon :

Tableau 9.3 (*suite*)

Définition générale	*Cas particulier*
n : Nombre de grappes dans un *éas* de grappes	Nombre de quartiers dans l'échantillon
Ω : Ensemble des échantillons de n grappes	Ensemble des échantillons de n quartiers
ω : Un élément de Ω (un échantillon de n grappes)	Un échantillon de n quartiers
$\bar{M} = \frac{\sum_{i\in\omega} M_i}{n}$: Taille moyenne des grappes de l'échantillon	Nombre moyen de logements par quartier dans l'échantillon
$\bar{t} = \frac{\sum_{i\in\omega} t_i}{n}$: Moyenne par grappe dans l'échantillon	Nombre moyen de personnes par quartier dans l'échantillon
$\bar{y}_i = \frac{t_i}{M_i}$: Moyenne par unité secondaire dans la grappe i	Nombre moyen de personnes par logement dans le quartier i
$\bar{\bar{y}} = \frac{\sum_{i\in\omega} t_i}{\sum_{i\in\omega} M_i}$: Moyenne par unité secondaire dans l'échantillon	Nombre moyen de personnes par logement dans l'échantillon
$s_t = \sqrt{\frac{\sum_{i\in\omega}(t_i - \bar{t})^2}{n-1}}$: Écart-type des totaux des grappes	Dispersion du nombre total de personnes dans les quartiers (écarts entre les quartiers)
$S_m = \sqrt{\frac{\sum_{i\in\omega}(M_i - \bar{M})^2}{n-1}}$: Écart-type des M_i	Dispersion des tailles des quartiers
$S_{tm} = \frac{\sum_{i\in\omega}(M_i - \bar{M})(t_i - \bar{t})}{n-1}$: Covariance entre les t_i et les M_i	Covariance entre le nombre de personnes et le nombre de logements dans un quartier

EXERCICES

9.1 Le tableau A07 porte sur un ensemble de 2 500 comptes de dépenses présentés par 257 représentants d'une compagnie. On tire un échantillon de 10 représentants puis on analyse tous les comptes soumis par ceux-là : on y apporte les correctifs nécessaires puis on prend note du montant corrigé Y. Le tableau 9.4 présente les montants corrigés des comptes soumis par les 10 représentants. Soit τ_y la valeur totale des 2 500 montants corrigés. Estimer par la moyenne chacun des paramètres suivants, ainsi que l'écart-type de l'estimateur.

a) Le montant total des valeurs corrigées;

b) Le montant moyen par représentant;

c) Le montant moyen par demande de remboursement.

Tableau 9.4
Montant des comptes présentés par un échantillon de 10 représentants d'une population constituée de 2 500 comptes présentés par 257 représentants

Repré-sentant	Nombre de comptes	Total	Montants des comptes
91	4	6429,37	695,21 59,47 5063,80 610,89
116	5	886,58	527,41 145,70 49,52 121,94 42,01
9	5	3911,09	754,51 460,52 574,38 608,44 1513,24
159	3	112,11	28,63 29,26 54,22
192	30	12013,12	23,59 19,63 2154,91 57,52 30,27 167,84 27,03 3151,17 32,35 39,98 75,41 728,84 36,51 184,90 34,53 104,47 75,72 334,41 110,60 90,38 78,57 2217,62 306,55 62,86 13,49 139,69 38,14 786,91 798,19 91,04
97	8	19315,43	784,30 1052,06 51,20 82,90 15835,60 581,08 698,67 229,62
218	22	3899,64	113,17 133,96 42,01 2,62 1069,76 37,77 116,65 134,80 268,68 57,14 39,16 1470,19 107,98 49,10 5,47 12,99 28,65 7,09 42,31 52,77 22,11 85,26
110	6	636,54	256,12 82,17 115,22 36,56 104,04 42,43
127	5	155,46	43,05 21,46 37,67 16,73 36,55
138	4	718,89	88,20 64,37 546,08 20,24

9.2 Dans la population décrite à l'exercice 9.1, estimer par le quotient (avec le nombre de réclamations comme variable auxiliaire) le montant corrigé total et estimer l'écart-type de l'estimateur.

9.3 Dans la population décrite à l'exercice 9.1, estimer par la régression (avec le nombre de réclamations comme variable auxiliaire) le montant corrigé total et estimer l'écart-type de l'estimateur.

9.4 Supposons, dans le contexte de l'exercice 9.1, que le montant réclamé est connu pour tous les comptes. Le montant total des réclamations est de 1 206 611 $. On peut alors utiliser l'échantillon pour estimer *l'erreur* totale par représentant et de là le montant total corrigé. Le Tableau 9.5 présente les montants des erreurs (les différences «montant réclamé» - «montant admissible», dans l'ordre du Tableau 9.4) dans les comptes de l'échantillon.

a) Utiliser cette information pour estimer par la moyenne le montant total (corrigé) des 2 500 comptes de la population et estimer l'écart-type de l'estimateur. Comparer la précision de cet estimateur à celle de l'exemple 9.1.

b) Pourquoi l'estimateur ici est-il nettement meilleur qu'au numéro 9.3?

Tableau 9.5

Corrections des comptes présentés par un échantillon de 10 représentants d'une population constituée de 2500 comptes présentés par 157 représentants

Nombre de comptes	Total	Montants des comptes
4	941,34	52,85 1,06 771,24 116,19
5	65,68	65,68 0 0 0 0
5	0	0 0 0 0 0
3	0	0 0 0
30	23,73	0 0,56 0 0 0 0 0 0 8,83 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 14,34 0 0 0 0
8	38,02	0,45 36,39 0 1,18 0 0 0 0
22	0	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
6	35,6	33,1 0,0 0,0 2,5 0,0 0,0
5	14,22	1,98 1,75 6,28 0 4,21
4	8,09	4,11 2,58 0 1,40

9.5 Dans le contexte de l'exercice 9.4, estimer par le quotient (avec le nombre de comptes comme variable auxiliaire) le montant total (corrigé) des 2 500 comptes de la population et estimer l'écart-type de l'estimateur. Comment se compare la précision ici à celle obtenue à l'exercice 9.4?

9.6 Dans le contexte de l'exercice 9.4, estimer par la régression (avec le nombre de comptes comme variable auxiliaire) le montant total (corrigé) des 2 500 comptes de la population et estimer l'écart-type de l'estimateur. Comment se compare la précision ici à celles obtenues aux exercices 9.3 et 9.4?

9.7 [*Suite du numéro* 9.1] L'échantillon au numéro 9.1 est constitué de 92 comptes.

a) Estimer τ_y par la moyenne ainsi que l'écart-type de l'estimateur sous la supposition que l'échantillon est un *éas* de 92 comptes tiré dans la population de 2 500 comptes.

b) Vous constaterez que l'écart-type estimé en a) est inférieur à celui obtenue au numéro 9.1. Vérifier qu'il ne s'agit pas d'un simple accident aléatoire en déterminant l'écart-type réel de l'estimateur dans les deux situations suivantes: i) $\hat{\tau}_1$ est l'estimateur du total par la moyenne basé sur un *éas* de 92 comptes tirés parmi les 2 500 comptes; et ii) $\hat{\tau}_2$ est l'estimateur du total par la moyenne basé sur un *éas* de 10 représentants. Les données de la population sont les suivantes: l'écart-type des 2 500 montants corrigés est S_y = 1 436,485; et l'écart-type des montants corrigés (totaux) des 257 représentants est S_t = 9 348,732.

9.8 Dans le contexte de l'exercice 9.1, estimer τ_y par le quotient en prenant pour variable auxiliaire non pas le nombre de comptes mais le montant réclamé par le représentant. Voici les valeurs de X (les montants réclamés) et de Y (les montants corrigés) dans l'échantillon:

x	7370,71	952,26	3911,09	112,11	12036,85
y	6429,37	886,58	3911,09	112,11	12013,12
x	19353,45	3899,64	672,14	169,68	726,98
y	19315,43	3899,64	636,54	155,46	718,89

Le montant total des réclamations est τ_x = 1 206 611 $.

9.9 Dans un tirage systématique, l'ordre dans lequel les unités sont disposées peut avoir d'importantes conséquences sur la qualité de l'estimation. Cet exercice illustre ce fait. On tire un échantillon systématique de taille 4 dans une population de taille 32. Dans chacune des situations suivantes, déterminer a) la variance $V(\bar{y})$, et b) le biais de l'« estimateur » $\sqrt{\hat{V}(\bar{y})} = \sqrt{1-n/N}\,\dfrac{s}{\sqrt{n}}$.

Les données sont présentées de telle sorte que chaque ligne constitue un échantillon.

Situation 1 : *Les données ont été alignées dans un ordre aléatoire.*

Échantillon				$\bar{y}$	s^2
475	473	430	356	433,50	3100,333
431	431	376	468	426,50	1437,667
426	377	509	276	397,00	9475,333
406	509	399	361	418,75	4010,917
406	463	364	507	435,00	3950,000
444	416	474	351	421,25	2754,250
446	303	278	479	376,50	10147,000
341	344	280	294	314,75	1060,917

Situation 2 : *Les données se trouvent alignées en ordre croissant.*

Échantillon				$\bar{y}$	s^2
276	356	416	468	379,00	6809,333
278	361	426	473	384,50	7149,667
280	364	430	474	387,00	7132,000
294	376	431	475	394,00	6084,667
303	377	431	479	397,50	5705,000
341	399	444	507	422,75	4932,250
344	406	446	509	426,25	4804,250
351	406	463	509	432,25	4708,917

Situation 3 : *Les données se trouvent alignées de telle sorte que les unités de chaque échantillon prennent des valeurs rapprochées:*

Échantillon				$\bar{y}$	s^2
276	278	280	294	282,00	66,66667
303	341	344	351	334,75	465,58333
356	361	364	376	364,25	72,250000
377	399	406	406	397,00	188,66667
416	426	430	431	425,75	46,916667
431	444	446	463	446,00	172,66667
468	473	474	475	472,50	9,666667
479	507	509	509	501,00	216,00000

a) Vérifier les calculs présentés dans le tableau suivant:

	$V(\bar{y})$	$E(s^2)$	Biais Relatif de s^2
Situation 1	1451,7334	4492	-9,73%
Situation 2	383,9521	5916	18,88 %
Situation 3	4704,6709	155	-96,89 %

b) Les observations suivantes étaient prévisibles. Essayer de les expliquer.

i) Dans un *éas* de taille 4, la variance de $\bar{y}$ est de 1 088. C'est dans la situation 1 que la variance s'approche le plus de 1 088.

ii) C'est aussi dans la situation 1 que le biais relatif de s^2 est à son plus bas.

iii) La variance $V(\bar{y})$ est très faible dans la situation 2.

iv) La variance $V(\bar{y})$ est très élevée dans la situation 3.

v) Dans la situation 3, l'estimation de S^2 présente un sérieux biais, et celui-ci est négatif.

9.10 Un tirage systématique qui consiste à tirer chaque k^e unité produit des échantillons de même taille seulement si N est un multiple de k. Considérer la population suivante de taille $N = 35$ (à lire ligne par ligne) :

344 403 283 451 420 434 449 426 426 434 415 422
416 438 440 337 435 260 510 439 511 317 267 467
331 539 287 253 507 437 472 510 352 432 459

On effectue un tirage systématique en tirant « chaque huitième unité » à partir d'une unité tirée au hasard parmi les 8 premières. Dresser une liste des échantillons possibles, qui seront alors de taille 4 ou 5.

a) Confirmer que la moyenne de l'échantillon n'est pas sans biais;

b) Déterminer l'erreur quadratique moyenne de $\bar{y}$

c) Confirmer que s^2 n'est pas un estimateur sans biais de S^2.

9.11 Il existe une procédure permettant de tirer un échantillon systématique de taille n fixe, même lorsque N n'est pas un multiple de n. Cet exercice illustre la procédure. Considérer le tirage d'un échantillon de taille $n = 4$ d'une population de taille $N = 13$ dont les unités sont numérotées de 1 à 13. On inscrit sur une liste de longueur 4(13) = 52 les nombres 1 à 13, chacun répété 4 fois, ainsi: 1,1,1,1,2,2,2,2,...,13,13,13,13. Ensuite on prélève chaque 13e nombre, à partir d'un premier sélectionné au hasard parmi les 13 premiers. Vérifier qu'aucune unité ne figure dans un échantillon plus d'une fois, et que chaque unité est contenue dans exactement 4 des 13 échantillons possibles.

9.12 Le tableau A03 présente les données d'une population de factures émises par un restaurant sur une période de 364 jours. Le Tableau 9.6 présente les données d'un *éa*s de 8 jours.

Tableau 9.6

Données sur les factures émises par un restaurant pendant 8 journées tirées au hasard parmi les 364 jours de la période soumise à une vérification

Nombre de factures	Nombre total de convives	Montant total des facture	Portion « Vin »
25	73	2040,10	319,70
24	83	2389,70	396,75
25	75	2099,90	392,15
21	84	2288,50	408,25
18	83	2363,25	402,50
26	74	2012,50	365,70
33	89	2239,05	377,20
18	77	2348,30	470,35

Soit $M_0 = 8\,654$ le nombre de factures dans la population, $N = 364$ le nombre de jours couverts par la population, y_{ij} le montant total de la j^e facture du jour i et t_i le montant total de toutes les factures du jour i. Soit $\tau_y = \sum_{k=1}^{364}\sum_{\ell} y_{k\ell} = \sum_{k=1}^{364} t_k$, S_y l'écart-type des $y_{k\ell}$ et S_t l'écart-type des t_k.

a) Estimer par la moyenne le montant moyen d'une facture et estimer l'écart-type de l'estimateur.

b) Les données suivantes portent sur la population elle-même: $\sum_{k=1}^{364} t_k = 763\,319{,}40$ et $S_t = 192{,}3355$. Déterminer le montant moyen d'une facture et l'écart-type de l'estimateur défini en a).

c) Le nombre de factures observées durant ces 8 jours est de 190. Supposons qu'on tire plutôt un échantillon de 190 factures dans la population de 8 654 factures, et qu'on estime la moyenne μ_y des montants des factures par la moyenne $\bar{y}$ des 190 factures de l'échantillon. Sachant que l'écart-type des montants des factures de la population est $S_y = 39{,}6646369$, déterminer l'écart-type de l'estimateur du montant moyen par facture.

d) Estimer par le quotient le montant moyen d'une facture et estimer l'écart-type de l'estimateur avec, pour variable auxiliaire

   i) Le nombre de factures;

   ii) Le nombre de convives. Le nombre moyen de convives par jour est 76,17582.

e) Estimer par la régression le montant moyen d'une facture et estimer l'écart-type de l'estimateur avec, pour variable auxiliaire.

   i) Le nombre de factures;

   ii) Le nombre de convives.

9.13 Dans la population décrite au numéro 9.12, on compare la précision de deux modes d'échantillonnage, soit un tirage de 50 jours ou un tirage de 1 189 factures (le nombre moyen de factures dans un échantillon de 50 jours); et trois estimateurs, soit par la moyenne, par le quotient ou par la régression. Les variables auxiliaires sont le nombre de convives et le nombre de factures dans le cas d'un échantillon de 50 jours; et le nombre de convives dans le cas d'un échantillon de 1189 factures. Le tableau suivant présente l'écart-type de l'estimateur dans chacune de ces hypothèses.

Estimateur	Variable auxiliaire	*Éas* de 50 jours	*Éas* de 1189 factures
Par le quotient	Convives	0,4425	0,4317
	Factures	3,4114	
Par la régression	Convives	0,4461	0,4318
	Factures	1,0729	
Par la moyenne	-	1,0626	1,0684

Pourquoi les observations suivantes étaient-elles prévisibles?

a) Dans presque tous les cas, les estimateurs par le quotient et par la régression sont plus précis que l'estimateur par la moyenne.

b) Dans le cas d'un échantillon de 50 jours, l'utilisation du nombre de convives comme variable auxiliaire est nettement plus avantageux que l'utilisation du nombre de factures.

c) L'avantage d'un échantillon de 1 189 factures par rapport à un échantillon de 50 jours est nettement moins prononcé dans le cas de l'estimation par la moyenne.

9.14 Dans le contexte du numéro 9.12, supposons qu'on décide de procéder par stratification selon le jour de la semaine. On a donc 7 strates, chacune comprenant 52 jours. On tire un *éas* de deux jours dans chaque strate. Voici un résumé des résultats (M_i est le nombre de factures, et t_i le montant total des factures de la journée)

Jour de la semaine	Nb de factures (pop.)	M_i	t_i	s_m	s_t	s_{mt}
Lundi	1059	15	1204,05	11,3137	1110,79	12567
		31	2774,95			
Mardi	1158	28	2455,25	3,5355	488,72	1728
		23	1764,10			
Mercredi	1049	25	2204,55	3,5355	411,47	1455
		20	1622,65			
Jeudi	1097	23	2287,35	3,5355	99,21	-350.8
		28	2147,05			
Vendredi	1057	20	1623,80	0,7071	187,84	132.8
		21	1889,45			
Samedi	1535	28	2543,80	3,5355	354,54	1254
		23	2042,40			
Dimanche	1699	33	3114,20	4,2426	556,21	2360
		27	2327,60			

a) Estimer le montant total des factures en utilisant l'estimateur par la moyenne dans chaque strate, et estimer l'écart-type de l'estimateur.

b) Estimer le montant moyen des factures en utilisant l'estimateur par le quotient (avec le nombre de factures comme variable auxiliaire) dans chaque strate, et estimer l'écart-type de l'estimateur.

9.15 Dans le contexte du numéro 9.12, supposons qu'on décide de stratifier la population selon qu'il s'agit d'un jour de semaine, d'un samedi ou d'un dimanche. On tire un *éas* de 15 jours parmi les 260 jours de semaine; un échantillon de 5 samedis et un échantillon de 5 dimanches. Voici les résultats (M_i est le nombre de factures, x_i le nombre total de convives, t_i est le montant total des factures de la journée.):

Données de la strate 1: Jour de semaine

[5420 factures et 17 117 convives dans la strate 1 de la population]

x_i	t_i	x_i	t_i	x_i	t_i
66	1823,90	46	1271,90	57	1624,95
52	1382,30	70	1892,90	58	1631,85
71	1934,30	77	2126,35	80	2140,15
57	1541,00	71	1903,25	72	1897,50
80	2086,10	54	1294,90	56	1497,30

$\sum x_i = 967; \sum t_i = 26\,049$

Résumé - strate 1:

	Estimateur (Lundi-Vendredi)		
	Par la moyenne	Par le quotient	Par la régression
$\hat{\tau}_{1y}$	451 510	461 091	460 974
$\sqrt{\hat{V}(\hat{\tau}_{1y})}$	19 215	4060	4119

Données - strate 2: Samedis

[1535 factures et 5022 convives dans la strate 2 de la population]

	x_i	t_i
	112	3321,20
	103	2832,45
	84	2385,10
	76	2210,30
	113	3096,95

$\sum x_i = 488; \sum t_i = 13\,846$

	Estimateur (Samedis)		
	Par la moyenne	Par le quotient	Par la régression
$\hat{\tau}_{1y}$	143 998	142 489	142 550
$\sqrt{\hat{V}(\hat{\tau}_{1y})}$	10347	2281	2617

Données de la strate 3: Dimanches

[1699 factures et 5 589 convives dans la strate 3 de la population]

	x_i	t_i
	117	3279,80
	85	2438,00
	94	2669,15
	107	3164,80
	99	2889,95

$\sum x_i = 502; \sum t_i = 14\,442$

	Estimateur (Dimanches)		
	Par la moyenne	Par le quotient	Par la régression
$\hat{\tau}_{1y}$	150 194	160 786	160 430
$\sqrt{\hat{V}(\hat{\tau}_{1y})}$	7663	1585	1683

a) Vérifier les calculs effectués au niveau des strates.

b) Utiliser les estimations ci-dessus pour estimer la valeur totale τ_y des factures, ainsi que l'écart-type de l'estimateur. Dans le cas de la strate 1, utiliser l'estimateur i) par la moyenne, et puis ii) par le quotient.

9.16 Une population est constituée de 8 654 factures émises par un restaurant au courant d'une période de 364 jours. On tire un *éas* de 10 jours afin d'estimer le montant total des factures émises durant cette période. Voici pour chaque journée sélectionnée les valeurs observées des trois variables suivantes:

- t_i: le montant total de la facture
- x_i : le nombre total de convives
- M_i : le nombre de factures dans la journée

M_{is}	13	30	28	40	18	29	15	18	21	23
t_i	12623	2717	2523	3781	1604	2884	1351	1639	1727	2153
x_i	45	96	92	127	56	97	50	59	68	81

Par ailleurs, $M_0 = 8654$; $N = 364$; $n = 10$; $\mu_x = 76{,}17582$.

a) Estimer par la moyenne le montant total τ des factures de la population.

b) Estimer l'écart-type de l'estimateur en a).

c) Estimer τ par le quotient, avec le nombre de factures comme variable auxiliaire.

d) Estimer l'écart-type de l'estimateur en c).

e) Estimer τ par le quotient, avec le nombre de convives comme variable auxiliaire.

f) Estimer l'écart-type de l'estimateur en e).

g) Le nombre de factures dans l'échantillon est de 235. Déterminer l'écart-type de l'estimateur de τ (par la moyenne) dans un *éas* de 235 factures. L'écart-type de montants des factures dans la population est de 39,66.

h) Pourquoi les estimateurs en c) et e) sont-ils moins bons que celui en a)?

i) Pourquoi l'estimateur en e) est-il meilleur que celui en c)?

j) Pourquoi l'estimateur en g) est-il meilleur que celui en a)?

10 Échantillonnage par grappes à un degré et probabilités inégales

Dans un échantillonnage par grappes, les unités primaires sont parfois de grandeur très variable, et on devine qu'il faudrait accorder plus d'importance aux grosses grappes — dont la contribution au total est plus importante — qu'aux plus petites. Les grosses grappes sont souvent peu nombreuses, et par conséquent il y a de fortes chances qu'un *éas* n'en contienne aucune. Une façon d'éviter cela est d'effectuer les tirages de telle sorte que les probabilités d'inclusion des unités primaires reflètent leur taille — qu'elles soient à peu près proportionnelles aux tailles des grappes.

L'idée de probabilités d'inclusion variables ne se limite pas nécessairement aux populations constituées de grappes : elle est parfois imposée par la nature même de la population. Considérons, par exemple, l'ensemble des lacs dans un parc national. Si ces lacs ne sont pas répertoriés et si on ne connaît pas d'avance où ils se situent, comment en tirer un échantillon? Une façon de faire est la suivante : sur un plan du parc, on choisit une paire de coordonnées — un point sur le parc. Si le point tombe sur un lac, on le sélectionne; sinon, on recommence. On applique cette procédure autant de fois que nécessaire pour créer un échantillon de n lacs. Dans ce cas, à chaque tirage, la probabilité qu'un lac soit sélectionné est proportionnelle à sa surface.

Ou encore, pour échantillonner les usagers d'une piscine publique, on interpelle une personne (au hasard, dans la mesure du possible) parmi ceux qui s'y trouvent à un moment choisi aléatoirement. Il est évident que l'échantillon contiendra un nombre disproportionné de ceux qui passent beaucoup de temps à la piscine.

Autre exemple : si pour tirer un échantillon de ménages on ne dispose que d'un bottin téléphonique, on joindra les ménages par des numéros de téléphone tirés au hasard. Chaque numéro de téléphone a la même probabilité de sélection, mais la probabilité de sélection d'un ménage dépend du nombre de lignes qu'il possède (et évidemment un ménage sans ligne téléphonique n'a aucune chance de se trouver dans l'échantillon: il n'appartient pas de fait à la population).

Une application particulière

Une telle approche, répandue en vérification comptable et portant dans ce domaine le nom d'*échantillonnage par unité monétaire* (*Dollar Unit Sampling*), est attribuée à Anderson et al (Anderson, 1973) . En fait elle est mentionnée par Arkin (Arkin, 1963) et a été discutée

longuement dans les années soixante-dix et quatre-vingts. Il a été entre autres traité par Neter et al (Neter, 1978), Menzefricke (Menzefricke, 1984), McCray (McCray, 1984), Reneau (Reneau, 1978), Ramage et al (Ramage, 1979). Lorsqu'on échantillonne une population de comptes, la mesure d'un compte est une certaine valeur monétaire qui lui est associée et qui est connue pour la population entière. Cela serait, par exemple, la valeur d'un remboursement de la TPS *réclamé* au ministère du Revenu. La probabilité de sélection d'un compte (en un tirage) serait proportionnelle au montant réclamé. Le terme utilisé en vérification pour ce type de sélection vient de ce qu'on imagine la population comme étant constituée de *dollars*, leur nombre M_0 étant la valeur totale de tous les comptes. Chacun de ces dollars appartient à un compte. On procède alors comme ceci :

- On tire au hasard, avec une probabilité de $1/M_0$, l'un de ces *dollars*;
- On retient dans l'échantillon le compte auquel il appartient;
- On recommence n fois, toujours à partir de la population entière.

Le sujet est très vaste, car il existe un grand nombre de façons possibles de sélectionner les unités, et peu d'indices généraux pour guider le choix. Nous allons donc décrire un petit nombre de procédures de sélection, mais l'estimateur que nous proposerons est assez général. Il peut s'adapter à presque tout mode de sélection qui permet le calcul des probabilités d'inclusion. Il s'agit de l'*estimateur de Horvitz-Thompson* (Horvitz, 1952).

Nous commençons, cependant, par une méthode qui se distingue des autres par sa très grande simplicité : le tirage avec remise et probabilités proportionnelles à la taille.

10.1 TIRAGE AVEC REMISE ET PROBABILITÉS PROPORTIONNELLES À LA TAILLE

Lorsque N est grand par rapport à n, un tirage avec remise n'est pas tellement moins efficace qu'un tirage sans remise. Dans ce cas, l'idée de tirer l'échantillon avec remise est, comme on le verra tout de suite, attrayante par sa simplicité. Nous considérons donc la procédure suivante :

1) On effectue n tirages successifs, *avec* remise;

2) À chaque tirage, les unités primaires sont sélectionnées avec probabilités proportionnelles à leur taille ψ_k où les ψ_k sont normalisées de sorte que $\sum_{k=1}^{N} \psi_k = 1$.

Normalement, la taille d'une grappe est le nombre d'unités secondaires qu'elle comprend et alors

$$\psi_k = \frac{M_k}{M_o}, \text{ où } M_o = \sum_{k=1}^{N} M_k .$$

Mais ce n'est pas toujours le cas, et donc nous exprimerons les formules en fonction d'une mesure quelconque ψ_k.

On observe une variable Y définie sur chaque unité secondaire. Soit M_k le nombre d'unités secondaires et t_k la somme des valeurs de Y dans l'unité primaire k.

THÉORÈME 10.1 **TIRAGES AVEC REMISE ET PROBABILITÉS PROPORTIONNELLES À LA TAILLE**

Dans un tirage avec remise et probabilités ψ_k, soit $R_i = \frac{t_i}{\psi_i}$, où t_i est le total de la i^e grappe sélectionnée et ψ_i est sa probabilité de sélection. L'estimateur

$$\hat{\tau} = \frac{1}{n}\sum_{i=1}^{n} R_i \tag{10.1}$$

est sans biais pour τ. Sa variance est

$$V(\hat{\tau}) = \frac{\sigma_r^2}{n}, \text{ où } \sigma_r^2 = \sum_{k=1}^{N}(r_k - \tau)^2\psi_k = \sum_{k=1}^{N}\frac{t_k^2}{\psi_k} - \tau^2 \text{ et } r_k = \frac{t_k}{\psi_k}.$$

Un estimateur sans biais de $V(\hat{\tau})$ est donné par

$$\hat{V}(\hat{\tau}) = \frac{s_r^2}{n}, \text{ où } s_r^2 = \frac{\sum_{i\in\omega}(R_i - \bar{R})^2}{n-1}. \tag{10.2}$$

COROLLAIRE : **PROBABILITÉS PROPORTIONNELLES AU NOMBRE D'UNITÉS**

Les estimateurs de μ_t et μ_y et leurs propriétés découlent directement de ceux de τ : $\hat{\mu}_t = \frac{\hat{\tau}}{N}$,

$$\hat{\mu}_y = \frac{\hat{\tau}}{M_0}, \ \sqrt{\hat{V}(\hat{\mu}_t)} = \frac{\sqrt{\hat{V}(\hat{\tau})}}{N}, \ \sqrt{\hat{V}(\hat{\mu}_y)} = \frac{\sqrt{\hat{V}(\hat{\tau})}}{M_o}.$$

Remarque *La moyenne non-pondérée des moyennes est sans biais*

Lorsque le tirage se fait avec probabilités proportionnelles au nombre d'unités secondaires, c'est-à-dire, lorsque $\psi_k = M_k / M_o$, l'estimateur de μ_y devient simplement la moyenne des $\bar{y}_i = \frac{t_i}{M_i}$, $\hat{\mu}_y = \frac{1}{n}\sum_{i=1}^{n}\bar{y}_i$. Cette même moyenne non pondérée aurait été biaisée dans un échantillon tiré avec probabilités égales.

Exemple 10.1 *Tirages avec remise et probabiilités ψ_i*

Considérons la population présentée au tableau A07 constituée de M_o = 2 500 comptes de dépenses soumis par N = 257 représentants d'une compagnie. Les représentants sont les unités primaires et les comptes sont les unités secondaires. On tire un échantillon de n = 15 représentants, avec remise et probabilités proportionnelles aux tailles, c'est-à-dire, aux nombres de comptes M_i. Les résultats sont présentés au Tableau 10.1. Estimer le montant moyen par compte.

Solution

Dans le cas présent, les probabilités de sélection sont les valeurs $\psi_k = M_k / M_o$. L'estimateur (10.1) est donc $\hat{\tau} = M_0 \frac{1}{n}\sum_{i=1}^{n} \frac{t_i}{M_i} = M_0 \frac{1}{n}\sum_{i=1}^{n} \bar{y}_i$. La moyenne par compte est estimée par $\hat{\mu}_y = \frac{1}{M_o}\hat{\tau} = \frac{1}{n}\sum_{i=1}^{n} \bar{y}_i =$ 505,6128. La variance de l'estimateur est estimée par $\hat{V}(\hat{\mu}_y) = \frac{s_r^2}{nM_o^2}$. Puisque $R_i = \frac{t_i}{M_i / M_0} = M_0 \bar{y}_i$. $s_r = M_0 s_{\bar{y}}$ et donc $\hat{V}(\hat{\mu}_y) = \frac{M_o^2 s_{\bar{y}}^2}{nM_o^2} = \frac{1}{n} s_{\bar{y}}^2$, où $s_{\bar{y}}^2 = \frac{1}{n-1}\sum_{i=1}^{n} (\bar{y}_i - \bar{\bar{y}})^2 =$ 364 344,7. L'écart-type est $\sqrt{V(\hat{\mu}_y)} = \frac{\sqrt{364344,7}}{\sqrt{15}} = 155,8514$.

Remarques *Concernant le dernier exemple*

1) Lorsque $\psi_i = M_i / M_0$, la procédure se résume très simplement sans formules : on calcule $\bar{\bar{y}}$, la moyenne des moyennes $\bar{y}_i$, et l'écart-type $s_{\bar{y}}$ de ces moyennes. On estime la moyenne μ_y par $\hat{\mu}_y = \bar{\bar{y}}$ et son écart-type par $\frac{s_{\bar{y}}}{\sqrt{n}}$.

2) En nous référant aux données de la population, on trouve que l'échantillon a donné d'assez bons résultats. La moyenne réelle de la population est 472,5 (estimée par l'échantillon à 505,6), et l'écart-type de l'estimateur basé sur un échantillon de 15 représentants est de 145 (estimé par l'échantillon à 155,9).

3) On constate qu'un échantillon par grappes de taille 15 avec probabilités égales aurait été moins efficace. L'estimation par la moyenne aurait donné un écart-type de 114 alors que les estimateurs par le quotient et par la régression (avec la taille de la grappe comme variable auxiliaire) auraient donné un écart-type de l'ordre de 203.

4) Qui plus est, le nombre d'unités secondaires dans un échantillon de 15 grappes est beaucoup plus grand quand on tire avec probabilités proportionnelles aux tailles (de 329 en moyenne contre 146 lorsqu'on tire avec probabilités égales).

5) Un *éas* d'unités secondaires de même taille aurait donné de bien meilleurs résultats. Si on avait plutôt tiré un *éas* de 329 comptes, l'écart-type aurait été de 74, nettement inférieur à celui d'un échantillon de 15 grappes. Ceci confirme qu'un échantillon par grappes n'est généralement pas aussi efficace qu'un *éas* comprenant le même nombre d'unités secondaires.

Tableau 10.1
Échantillon de 15 représentants tiré d'une population de 257 représentants ayant soumis en tout 2 500 comptes de dépenses

k	Nombre de comptes (M_i)	Montant total des comptes du représentant i (t_i)	Probabilité de sélection ($\psi_i = M_i/M_o$)	Moyenne par représentant ($\bar{y}_i = t_i/M_i$)	$r_i = \frac{t_i}{\psi_i} = M_o\bar{y}_i$
17	8	4708	0,0032	588,56	1471391
33	22	11978	0,0088	544,47	1361175
55	17	7050	0,0068	414,70	1036738
67	40	3314	0,0160	82,84	207096
74	16	17116	0,0064	1069,77	2674419
76	28	17180	0,0112	613,58	1533942
78	2	696	0,0008	348,03	870075
87	5	512	0,0020	102,34	255840
123	4	214	0,0016	53,52	133788
134	46	9960	0,0184	216,51	541278
161	40	97957	0,0160	2448,93	6122336
170	8	2013	0,0032	251,58	628947
181	47	25575	0,0188	544,15	1360371
214	51	4490	0,0204	88,05	220120
221	73	15855	0,0292	217,19	542964
Total	407	218618	0,1628	7584,1917	
Moyenne	27,133	14575		505,6128	
Écart-type				603,6097	

Comparaison avec un tirage avec probabilités égales

Comment se compare l'estimateur (10.1) à celui d'un estimateur par le quotient dans un *éas* de même taille? Voici donc la variance de l'estimateur de τ dans les deux cas :

Tirage avec remise et probabilités proportionnelles aux $\psi_i = M_i / M_0$:

$$V_1 = \frac{N\mu_M}{n}\sum_{k=1}^{N} M_k(\bar{y}_k - \mu_y)^2 \ .$$

Tirages sans remise, probabilités égales et estimateur par le quotient (grands échantillon et population):

$$V_2 \approx \frac{N}{n}\sum_{k=1}^{N} M_k^2(\bar{y}_k - \mu_y)^2 \ .$$

Si les M_k ne varient pas beaucoup, la différence entre V_1 et V_2 est faible. On peut montrer que $V_2 = V_1 + \frac{N(N-1)}{n}\rho S_a S_b$, où S_a et S_b sont les écarts-types des ψ_k et $\psi_k(\bar{y}_k - \mu_y)^2$,

respectivement et ρ est leur coefficient de corrélation. Donc $V_2 \geq V_1$ si et seulement si $\rho \geq 0$. On ne peut pas savoir dans une situation particulière si cette condition est vérifiée, mais elle est probablement plus commune que l'inverse.

COMMENT GÉNÉRER L'ÉCHANTILLON

Supposons que nous disposons des moyens pour générer des nombres de loi uniforme sur l'intervalle [0 ; 1) — des tables de nombres au hasard, des boules numérotées, ou d'un programme d'ordinateur. Comment utiliser ces facilités pour choisir les grappes avec probabilités proportionnelles à leurs tailles ? Soit $\psi_k = M_k / M_0$, pour $k = 1, 2, \ldots, N$.

Une procédure relativement simple est la suivante. On découpe l'intervalle [0 ; 1) en n sous-intervalles. Le premier est de longueur ψ_1, le deuxième est de longueur ψ_2, ainsi de suite, jusqu'au dernier, qui est de longueur ψ_N. On génère un nombre U de loi uniforme sur l'intervalle [0;1).

Ce nombre appartiendra à l'un des sous-intervalles définis ci-dessus. On choisira la k^e unité primaire si U est dans le k^e intervalle.

Exemple 10.2 *Choix d'un échantillon à probabilités inégales*

Afin de tirer un échantillon de taille 4 d'une population de 10 grappes dont les tailles sont

1, 4, 7, 9, 10, 11, 12, 13, 15, 18,

on génère 4 nombres au hasard. On obtient les nombres suivants :

0,0274 0,5122 0,6329 0,8675.

Quelles unités primaires doit-on alors tirer de la population ?

Solution.

Les ψ_k et leurs valeurs cumulées sont données dans le tableau suivant

ψ_k	0,01	0,04	0,07	0,09	0,10	0,11	0,12	0,13	0,15	0,18
ψ_k cumulés	0,01	0,05	0,12	0,21	0,31	0,42	0,54	0,67	0,82	1,00

Les 10 intervalles sont donc

[0 ; 0,01), [0,01 ; 0,05), [0,05 ; 0,12), [0,12 ; 0,21), [0,21 ; 0,31), [0,31 ; 0,42), [0,42 ; 0,54), [0,54 ; 0,67), [0,67 ; 0,82), [0,82 ; 1,00).

Puisque

0,0274 ∈ [0,01 ; 0,05), *intervalle* 2
0,5122 ∈ [0,42 ; 0,54), *intervalle* 7
0,6329 ∈ [0,54 ; 0,67), *intervalle* 8
0,8675 ∈ [0,82 ; 1,00), *intervalle* 10

les grappes à retirer sont les unités 2, 7, 8 et 10.

Remarque

Noter que puisqu'on tire avec remise, si deux des nombres au hasard tombent dans le même intervalle, l'unité correspondante est tirée deux fois. Par exemple, si les 4 nombres au hasard sont 0,0274 ; 0,0682 ; 0,8329 ; et 0,8675,

0,0274 ∈ [0,01 ; 0,05), intervalle 2;

0,0682 ∈ [0,05 ; 0,12), intervalle 3;

0,8329 ∈ [0,82 ; 1,00), intervalle 10;

0,8675 ∈ [0,82 ; 1,00), intervalle 10.

Les grappes à retirer sont les unités 2, 3, et 10 (2 fois).

10.2 TIRAGE SANS REMISE ET PROBABILITÉS INÉGALES — L'ESTIMATEUR DE HORVITZ-THOMPSON

Considérons maintenant un mode de tirage sans remise, et avec des probabilités d'inclusion qui reflètent les tailles des grappes.

Rappelons la motivation. Soit $\psi_1, \psi_2, \ldots, \psi_N$ les tailles normalisées des unités primaires, des nombres positifs tels que $\sum_{k=1}^{N}\psi_k = 1$. Considérons un mode de tirage bien défini qui génère *N probabilités d'inclusion* $\pi_1, \pi_2, \ldots . \pi_N$, où π_k est la probabilité que l'unité k soit incluse dans l'échantillon:

$$\pi_k = P(k \in \omega).$$

Il génère aussi $N(N\text{-}1)/2$ probabilités d'inclusion doubles, les probabilités $\pi_{k\ell}$ que deux unités données soient incluses:

$$\pi_{k\ell} = P(k \in \omega;\ \ell \in \omega).$$

On souhaite développer un mode de sélection tel que les π_k soient proportionnels aux ψ_k. On ne s'en doute peut-être pas, mais ceci est moins évident qu'il ne paraît. On pourrait croire que ce but serait atteint en effectuant n tirages sans remise, à chaque fois avec probabilités proportionnelles aux taille des unités (autres que celles déjà sélectionnées). Par exemple, lorsque $n = 2$, la première unité k est tirée avec probabilités proportionnelles aux ψ_k; si l'unité k est tirée, l'unité ℓ ($\ell \neq k$) est tirée avec probabilité $\frac{\psi_\ell}{1-\psi_k}$. Comme le montre le prochain exemple, cette façon de procéder ne mène pas à des π_k proportionnelles aux ψ_k.

Exemple 10.3 *Tirages sans remise et probabilités inégales*

Considérons une population de $N = 4$ unités dont les tailles sont $\psi_1 = 0{,}1$; $\psi_2 = 0{,}2$; $\psi_3 = 0{,}3$; et $\psi_4 = 0{,}4$. On tire une première unité avec probabilités ψ_k; puis une deuxième avec probabilités proportionnelles aux probabilités des unités restantes.

a) Déterminer π_{12}, la probabilité que l'échantillon soit constitué des unités 1 et 2.

b) Déterminer π_k, la probabilité que l'unité k se trouve dans l'échantillon, pour $k = 1, 2, 3, 4$.

Solution

a) $\pi_{12} = \psi_1 \frac{\psi_2}{1-\psi_1} + \psi_2 \frac{\psi_1}{1-\psi_2} = 0{,}02222 + 0{,}025 = 0{,}047222.$

b) $\pi_1 = \psi_1 + \psi_2 \frac{\psi_1}{(1-\psi_2)} + \psi_3 \frac{\psi_1}{(1-\psi_3)} + \psi_4 \frac{\psi_1}{(1-\psi_4)} = 0{,}2345238.$

De la même manière, on obtient $\pi_2 = 0{,}4412698$; $\pi_3 = 0{,}6083333$; $\pi_4 = 0{,}7158730$. Les π_k ne sont pas proportionnels aux ψ_k: $\frac{\pi_k}{\sum_{j=1}^{4} \pi_j} \neq \psi_k$.

L'ESTIMATEUR DE HORVITZ-THOMPSON

Le nombre de façons de tirer un échantillon avec probabilités inégales est quasiment illimité, et la littérature en propose un grand nombre. Un contexte précis peut faire préférer une façon de procéder plutôt qu'une autre; ou encore, on pourrait créer un mode de tirage sur mesure qui n'est répertorié nulle part. Il serait donc utile de pouvoir compter sur un estimateur qui puisse s'adapter à différents mode de tirage. Il en existe un, en effet, qu'on peut déterminer pour peu qu'on puisse calculer les probabilités d'inclusion simple π_k pour $k = 1, 2, \ldots, N$. La variance de l'estimateur peut être estimée à condition de pouvoir calculer les probabilités d'inclusion double $\pi_{k\ell} = P(k \in \omega \,\&\, \ell \in \omega)$, $k \neq \ell$, et que ces probabilités soient non nulles.

C'est l'estimateur de Horvitz-Thompson.

Estimateur Horvitz-Thompson

L'estimateur de Horvitz-Thomson (HT) est défini par

(10.3) $$\hat{\tau} = \sum_{i \in \omega} \frac{t_i}{\pi_i}.$$

THEOREME 10.2 **L'ESTIMATEUR DE HT EST SANS BIAIS**

L'estimateur (10.3) est sans biais pour le total τ: $E(\hat{\tau}) = \tau$.

Exemple 10.4 *L'estimateur $\hat{\tau} = N\overline{y}$ dans un éas est un estimateur de HT*

Dans un tirage aléatoire simple, $\pi_k = n/N$ pour tout k. Donc l'estimateur de HT est $\hat{\tau} = \sum_{i\in\omega} \frac{t_i}{n/N} = N\overline{t}$.

Variance de $\hat{\tau}$

THÉORÈME 10.3 **DEUX EXPRESSION DE LA VARIANCE D'UN ESTIMATEUR HT**

La variance de l'estimateur de HT peut s'écrire de deux façons :

Première expression :

$$V_1 = V(\hat{t}) = \sum_{k=1}^{N} t_k^2 \frac{(1-\pi_k)}{\pi_k} + 2\sum_{k}^{N-1}\sum_{k>\ell} t_k t_\ell \frac{(\pi_{k\ell} - \pi_k\pi_\ell)}{\pi_k\pi_\ell} . \tag{10.4}$$

Deuxième expression

$$V_2 = V(\hat{t}) = \sum_{k=1}^{N-1}\sum_{k>\ell} (\pi_k\pi_\ell - \pi_{ij})\left(\frac{t_k}{\pi_k} - \frac{t_\ell}{\pi_\ell}\right)^2 . \tag{10.5}$$

Estimateurs de la variance :

Chacune des deux expressions de la variance donne lieu à une estimation distincte.

THEOREME 10.4 **ESTIMATEURS DE LA VARIANCE DE L'ESTIMATEUR DE HT**

Les deux estimateurs suivants sont sans biais pour la variance de $\hat{\tau} = \sum_{i\in\omega} \frac{t_i}{\pi_i}$

$$\hat{V}_1(\hat{\tau}) = \sum_{i\in\omega} t_i^2 \frac{(1-\pi_i)}{\pi_i^2} + 2\sum_{i\in\omega}\sum_{\substack{j\in\omega \\ j>i}} t_i t_j \frac{\pi_{ij} - \pi_i\pi_j}{\pi_i\pi_j\pi_{ij}} . \tag{10.6}$$

La deuxième expression donne lieu à celui-ci :

$$\hat{V}_2(\hat{\tau}) = \sum_{i\in\omega}\sum_{\substack{j\in\omega \\ j>i}} \frac{\pi_i\pi_j - \pi_{ij}}{\pi_{ij}}\left(\frac{t_i}{\pi_i} - \frac{t_j}{\pi_j}\right)^2 . \tag{10.7}$$

Ce deuxième estimateur est appelé *estimateur de Yates-Grundy*.

Remarque *Les estimateurs peuvent prendre des valeurs négatives*

Un inconvénient de ces deux estimateurs de variance est qu'ils peuvent prendre des valeurs négatives. On s'attend à ce que $\hat{V}_2(\hat{\tau})$ le soit moins souvent, puisque, en général $\pi_{k\ell} \leq \pi_k\pi_\ell$: si k est dans l'échantillon la probabilité (conditionnelle) que ℓ y soit aussi est

généralement réduite. Par exemple, dans un *éas*, $\pi_k = n/N$ et $\pi_{k\ell} = \frac{n(n-1)}{N(N-1)} \leq \frac{n}{N}\frac{n}{N} = \pi_k\pi_\ell$.

Exemple 10.5 *Estimations de la variance de l'estimateur HT*

Considérons encore l'Exemple 10.3. Supposons que les valeurs totales d'une variable (les t_k) sont 12, 13, 25, 22.

a) Déterminer les π_k et les $\pi_{k\ell}$.

b) Déterminer la variance de l'estimateur du total basé sur un échantillon de taille 2;

c) Un échantillon de taille 2 est constitué des unités 2 et 4. Estimer τ.

d) Estimer la variance de l'estimateur utilisant les deux formules présentées.

Solution

a) Utilisant l'approche illustrée à l'Exemple 10.3, les valeurs des π_k sont

$\pi_1 = 0{,}2345238$; $\pi_2 = 0{,}4412698$; $\pi_3 = 0{,}6083333$; $\pi_4 = 0{,}7158730$

et celles des $\pi_{k\ell}$ sont :

$\pi_{12} = 0{,}04722222$; $\pi_{13} = 0{,}07619048$; $\pi_{14} = 0{,}111111$; $\pi_{23} = 0{,}16071429$; $\pi_{24} = 0{,}2333333$; $\pi_{34} = 0{,}3714286$.

b) $\sum_{k=1}^{N} t_k^2 \frac{(1-\pi_k)}{\pi_k} = 1278{,}491$; $2\sum_{k=1}^{N-1}\sum_{k>\ell} t_k t_\ell \frac{(\pi_{j\ell} - \pi_k\pi_\ell)}{\pi_k\pi_\ell} = -1199{,}924$;

$$V(\hat{\tau}) = \sum_{k=1}^{N-1} t_k^2 \frac{(1-\pi_k)}{\pi_k} + 2\sum_{k=1}^{N-1}\sum_{\ell>k} t_k t_\ell \frac{(\pi_{k\ell} - \pi_k\pi_\ell)}{\pi_k\pi_\ell} = 78{,}56644.$$

On peut vérifier que la formule $V(\hat{t}) = \sum_{k=1}^{N-1}\sum_{k>\ell}(\pi_k\pi_\ell - \pi_{k\ell})\left(\frac{t_k}{\pi_k} - \frac{t_\ell}{\pi_\ell}\right)^2$ donne le même résultat.

c) $t_2 = 13$; $t_4 = 22$; $\pi_2 = 0{,}4412698$; $\pi_4 = 0{,}7158730$; $\pi_{24} = 0{,}233333$.

$$\hat{\tau} = \frac{t_2}{\pi_2} + \frac{t_4}{\pi_4} = 60{,}19214.$$

d) Premier estimateur : $\hat{V}_1(\hat{\tau}) = \sum_{i\in\omega} t_i^2 \frac{(1-\pi_i)}{\pi_i^2} + 2\sum_{i\in\omega}\sum_{\substack{j\in\omega \\ j>i}} t_i t_j \frac{\pi_{ij} - \pi_i\pi_j}{\pi_i\pi_j\pi_{ij}}$.

$t_2^2 \frac{(1-\pi_2)}{\pi_2^2} = 484{,}9314$; $t_4^2 \frac{(1-\pi_4)}{\pi_4^2} = 268{,}3403$; $\sum_{i\in\omega} t_i^2 \frac{(1-\pi_i)}{\pi_i^2} = 753{,}2717$;

$$2\sum_{i\in\omega}\sum_{\substack{j\in\omega\\ j>i}} t_i t_j \frac{\pi_{ij}-\pi_i\pi_j}{\pi_i\pi_j\pi_{ij}} = \frac{2t_2t_4\dfrac{\pi_{24}-\pi_2\pi_4}{\pi_2\pi_4\pi_{24}}}{} = -640{,}6898$$

$$\hat{V}_1(\hat{\tau}) = \sum_{i\in\omega} t_i^2 \frac{(1-\pi_i)}{\pi_i^2} + 2\sum_{i\in\omega}\sum_{\substack{j\in\omega\\ j>i}} t_i t_j \frac{\pi_{ij}-\pi_i\pi_j}{\pi_i\pi_j\pi_{ij}} = 112{,}5819.$$

Deuxième estimateur de la variance (Estimateur de Yates-Grundy) :

$$\hat{V}_2(\hat{\tau}) = \frac{\pi_2\pi_4-\pi_{24}}{\pi_{24}}\left(\frac{t_2}{\pi_2}-\frac{t_4}{\pi_4}\right)^2 = 0{,}571836.$$

Il se trouve que pour cet échantillon particulier, l'estimateur de Yates-Grundy est étonnamment optimiste. Mais c'est un cas extrême, et bien sûr, on ne devrait pas trop attendre d'un échantillon de taille 2. Le tableau suivant présente la distribution des deux estimateurs.

ω	$p(\omega)$	$\hat{V}_1$	$\hat{V}_2$
(1,2)	0,04722222	-1103,1892	561,439488
(1,3)	0,07619048	-1003,8719	88,506926
(1,4)	0,11111111	665,3745	213,406335
(2,3)	0,16071429	-476,6327	90,746213
(2,4)	0,23333333	112,5819	0,571836
(3,4)	0,37142857	494,1705	18,526250

On remarque que $\hat{V}_1$ est quelquefois négatif (avec probabilité 0,284), alors que, dans cet exemple, $\hat{V}_2$ ne l'est jamais.

LA MÉTHODE DE LAHIRI

Une méthode particulière proposée par Lahiri (Lahiri, 1951) est conçue non pas pour que les probabilités π_k soient proportionnelles aux ψ_k, mais plutôt pour que la probabilité d'un échantillon ω soit proportionnelle à sa taille totale. Le mode de tirage est le suivant: on tire une première unité avec probabilités ψ_k; puis on tire un *éas* de taille n-1 parmi les N-1 unités restantes.

THÉORÈME 10.5 **PROBABILITÉS D'INCLUSION DANS UN TIRAGE LAHIRI**

Les probabilités d'inclusion simple et double dans un tirage effectué par la méthode de Lahiri sont :

(10.8) $$\pi_k = \frac{n-1}{N-1} + \frac{N-n}{N-1}\psi_k \; ; \; \pi_{k\ell} = \frac{(n-1)(n-2)}{(N-1)(N-2)} + \frac{(n-1)(N-n)}{(N-1)(N-2)}(\psi_k+\psi_\ell)$$

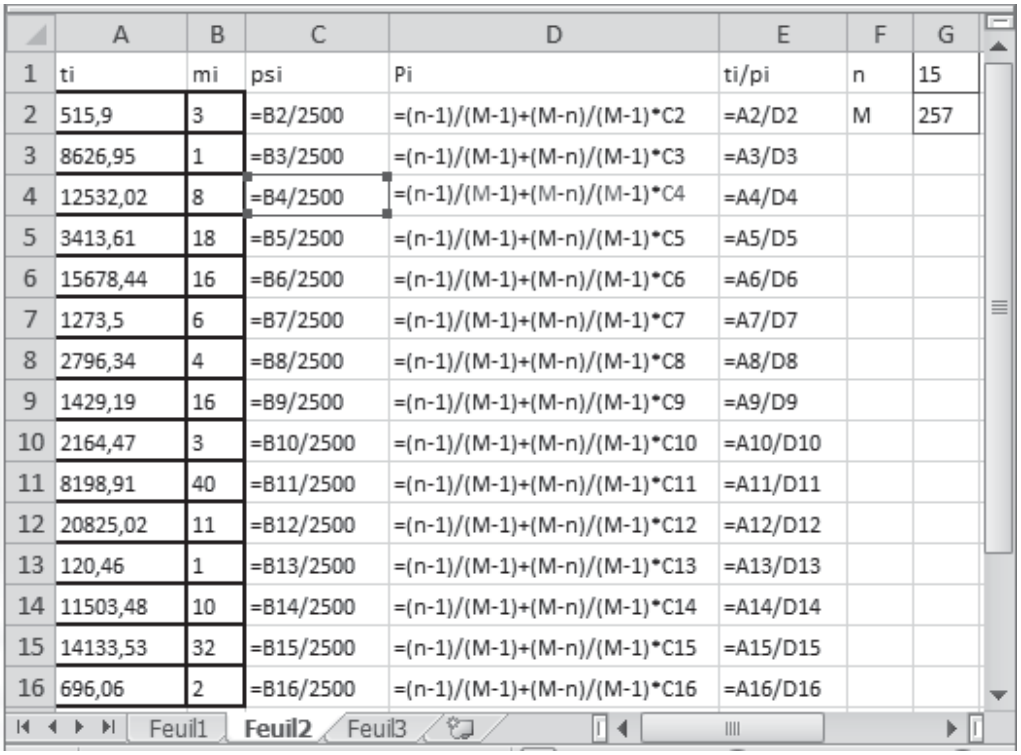
COROLLAIRE **L'ESTIMATEUR DE YATES-GRUNDY EST SANS BIAIS DANS UN TIRAGE LAHIRI**

$$\pi_k\pi_\ell - \pi_{k\ell} = \frac{N-n}{(N-1)^2}\left\{\frac{n-1}{N-2}[1-(\psi_k+\psi_\ell)]+(N-n)\psi_k\psi_\ell\right\}.$$

Cette dernière expression montre que l'estimateur de Yates-Grundy est toujours positif.

Exemple 10.6 *Lahiri: calculs avec Excel*

Données A07 N = 257 représentants, M_0 = 2 500 comptes de dépenses, sélections des représentants par la méthode de Lahiri



	A	B	C	D	E	F	G
1	ti	mi	psi	Pi	ti/pi	n	15
2	515,9	3	=B2/2500	=(n-1)/(M-1)+(M-n)/(M-1)*C2	=A2/D2	M	257
3	8626,95	1	=B3/2500	=(n-1)/(M-1)+(M-n)/(M-1)*C3	=A3/D3		
4	12532,02	8	=B4/2500	=(n-1)/(M-1)+(M-n)/(M-1)*C4	=A4/D4		
5	3413,61	18	=B5/2500	=(n-1)/(M-1)+(M-n)/(M-1)*C5	=A5/D5		
6	15678,44	16	=B6/2500	=(n-1)/(M-1)+(M-n)/(M-1)*C6	=A6/D6		
7	1273,5	6	=B7/2500	=(n-1)/(M-1)+(M-n)/(M-1)*C7	=A7/D7		
8	2796,34	4	=B8/2500	=(n-1)/(M-1)+(M-n)/(M-1)*C8	=A8/D8		
9	1429,19	16	=B9/2500	=(n-1)/(M-1)+(M-n)/(M-1)*C9	=A9/D9		
10	2164,47	3	=B10/2500	=(n-1)/(M-1)+(M-n)/(M-1)*C10	=A10/D10		
11	8198,91	40	=B11/2500	=(n-1)/(M-1)+(M-n)/(M-1)*C11	=A11/D11		
12	20825,02	11	=B12/2500	=(n-1)/(M-1)+(M-n)/(M-1)*C12	=A12/D12		
13	120,46	1	=B13/2500	=(n-1)/(M-1)+(M-n)/(M-1)*C13	=A13/D13		
14	11503,48	10	=B14/2500	=(n-1)/(M-1)+(M-n)/(M-1)*C14	=A14/D14		
15	14133,53	32	=B15/2500	=(n-1)/(M-1)+(M-n)/(M-1)*C15	=A15/D15		
16	696,06	2	=B16/2500	=(n-1)/(M-1)+(M-n)/(M-1)*C16	=A16/D16		

Feuil1 Feuil2 Feuil3

	A	B	C	D	E	F	G
1	ti	mi	psi	Pi	ti/pi	n	15
2	515,9	3	0,0012	0,055821875	9241,8967	M	257
3	8626,95	1	0,0004	0,055065625	156666,7		
4	12532	8	0,0032	0,0577125	217145,68		
5	3413,61	18	0,0072	0,06149375	55511,495		
6	15678,4	16	0,0064	0,0607375	258134,43		
7	1273,5	6	0,0024	0,05695625	22359,267		
8	2796,34	4	0,0016	0,0562	49756,94		
9	1429,19	16	0,0064	0,0607375	23530,603		
10	2164,47	3	0,0012	0,055821875	38774,584		
11	8198,91	40	0,016	0,0698125	117441,86		
12	20825	11	0,0044	0,058846875	353884,89		
13	120,46	1	0,0004	0,055065625	2187,5716		
14	11503,5	10	0,004	0,05846875	196745,78		
15	14133,5	32	0,0128	0,0667875	211619,39		
16	696,06	2	0,0008	0,05544375	12554,346		

Feuil1 Feuil2 Feuil3

Classeur1 - Microsoft Excel

Fichier Accueil Insertion Mise en page Formules Données Révision Affichage

C8

	A	B	C	D
1		Mi	1	8
2	Mi	Psi	=C1/2500	=D1/2500
3	3	=A3/2500	=(n-1)*(n-2)/(M-1)/(M-2)+(n-1)*(M-n)/(M-1)/(M-2)*($B3+C$2)	=(n-1)*(n-2)/(M-1)/(M-2)+(n-1)*(M-n)/(M-1)/(M-2)*($B3+D$2)
4	1	=A4/2500		=(n-1)*(n-2)/(M-1)/(M-2)+(n-1)*(M-n)/(M-1)/(M-2)*($B4+D$2)
5	8	=A5/2500		
6	18	=A6/2500		
7	16	=A7/2500		
8	6	=A8/2500		

Feuil1 Feuil2 Feuil3

Prêt 100 %

Grappe (i)	Total de la grappe (t_i)	Nombre d'unités secondaires (M_i)	Taille relative ($\psi_i = M_i/M_o$)	Probabilité d'inclusion de l'unité i (π_i)	Totaux pondérés (t_i/π_i)
114	515,90	3	0,0012	0,05582187	9241,897
80	8626,95	1	0,0004	0,05506562	156666,705
53	12532,02	8	0,0032	0,05771250	217145,679
160	3413,61	18	0,0072	0,06149375	55511,495
176	15678,44	16	0,0064	0,06073750	258134,431
94	1273,50	6	0,0024	0,05695625	22359,267
226	2796,34	4	0,0016	0,05620000	49756,940
82	1429,19	16	0,0064	0,06073750	23530,603
143	2164,47	3	0,0012	0,05582187	38774,584
88	8198,91	40	0,0160	0,06981250	117441,862
166	20825,02	11	0,0044	0,05884687	353884,892
86	120,46	1	0,0004	0,05506562	2187,572
47	11503,48	10	0,0040	0,05846875	196745,783
113	14133,53	32	0,0128	0,06678750	211619,390
78	696,06	2	0,0008	0,05544375	12554,346
					1 725 555

Estimation de τ: $\hat{\tau} = \sum_{i\in\omega} t_i / \pi_i = 9241{,}897 + 156\,666{,}705 + \ldots + 12\,554{,}346 = 1\,725\,555$;

Estimation de la moyenne par représentant: $\hat{\mu}_y = \dfrac{\hat{\tau}}{M_o} = \dfrac{1725555}{2500} = 690{,}2222$;

Estimation de $V(\hat{\tau})$ selon (10.6) :

$$\hat{V}(\hat{\tau}) = 171\,274\,119\,872\ ;\ \sqrt{\hat{V}(\hat{\tau})} = 413\,852{,}8\ ;\ \sqrt{\hat{V}(\hat{\mu}_y)} = \frac{\sqrt{\hat{V}(\hat{\tau})}}{M_o} = 165{,}5411\ .$$

Estimation de $V(\hat{\tau})$ selon (10.7) :

$$\hat{V}(\hat{\tau}) = 170\,856\,071\,727\ ;\ \sqrt{\hat{V}(\hat{\tau})} = 413\,347{,}4\ ;\ \sqrt{\hat{V}(\hat{\mu}_y)} = 165{,}339\ .$$

LA MÉTHODE D'ÉCHANTILLONNAGE SYSTÉMATIQUE

L'échantillonnage systématique, appliquée aux unités *secondaires* fournit une méthode de sélection des unités *primaires* avec probabilités inégales proportionnelles aux tailles. On aligne les M_o unités secondaires, puis on tire un échantillon systématique de n unités secondaires. Chaque unité secondaire sélectionnée entraîne l'inclusion de l'unité primaire qui la contient.

Exemple 10.7 *Méthode systématique*

Considérons une population de $N = 8$ unités primaires de tailles M_i égales à 6; 8; 17; 13; 3; 11; 12; 10. Le nombre d'unités secondaires est $M_0 = \sum_{k=1}^{N} M_k = 100$. On forme une chaîne, en maintenant l'identité de la grappe (les 6 « 1 » représentent les unités de la grappe 1, etc.):

1 1 1 1 1 1 2 2 2 2 2 2 2 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 4 4 4 4 4 4 4 4 4 4 4 4 4 5 5 5 6 6 6 6 6 6 6 6 6 6 6 7 7 7 7 7 7 7 7 7 7 7 8 8 8 8 8 8 8 8 8.

Il y a 80 unités secondaires. Afin de sélectionner un échantillon de $n = 4$ grappes, on tire au hasard un entier entre 1 et 20. Après quoi, on sélectionne chaque 20e unité. Par exemple, si le premier chiffre choisi est 13, on sélectionne la 13e, la 33e, la 53e et 73e.

Ces unités portent les numéros 2, 4, 6 et 8. L'échantillon sera donc constitué des unités primaires 2, 4, 6, et 8. Le Tableau 10.2-a présente une disposition des 80 unités secondaires sous la forme d'une matrice d'ordre 20 × 4 (colonne par colonne). Chaque ligne présente un des échantillons possibles.

Lorsque n est un diviseur de M_0, cette façon de procéder fait que les probabilités d'inclusion sont proportionnelles aux tailles. Mais elle n'exclut pas que certaines unités paraissent plus d'une fois dans l'échantillon. C'est le cas, par exemple, si les tailles étaient les suivantes : 2, 4, 25, 13, 3, 23, 4 et 6 (Tableau 10.2-b). Entre autres répétitions, les échantillons 7 à 11 contiennent l'unité 3 deux fois.

Lorsque n n'est pas un diviseur de M_0, les échantillons possibles ne seront pas tous de même taille. Dans le Tableau 10.2-c) le nombre d'unités secondaires est 87, ce qui fait que certains échantillons sont de taille 5, d'autres de taille 4.

L'approche systématique souffre de l'inconvénient commun à tout tirage systématique: elle ne peut fournir d'estimateur de la variance.

Tableau 10.2
Échantillons de grappes par la méthode systématique, selon la nature de la population et la taille de l'échantillon

Échantillon	a	b	c
1	1 3 4 7	1 3 4 6	1 2 4 6 8
2	1 3 4 7	1 3 4 6	1 2 4 6 8
3	1 3 4 7	2 3 4 6	1 2 4 6 8
4	1 3 4 7	2 3 4 6	1 2 4 7 8
5	1 3 5 7	2 3 5 6	1 2 4 7 8
6	1 3 5 7	2 3 5 6	1 2 4 7 8
7	2 3 5 7	3 3 5 6	2 2 4 7 8
8	2 3 6 7	3 3 6 6	2 2 4 7
9	2 3 6 7	3 3 6 6	2 2 4 7
10	2 3 6 7	3 3 6 6	2 2 4 7
11	2 3 6 8	3 3 6 7	2 2 5 7
12	2 4 6 8	3 4 6 7	2 3 5 8
13	2 4 6 8	3 4 6 7	2 3 5 8
14	2 4 6 8	3 4 6 7	2 3 5 8
15	3 4 6 8	3 4 6 8	2 3 5 8
16	3 4 6 8	3 4 6 8	2 3 5 8
17	3 4 6 8	3 4 6 8	2 3 5 8
18	3 4 6 8	3 4 6 8	2 4 5 8
19	3 4 7 8	3 4 6 8	2 4 6 8
20	3 4 7 8	3 4 6 8	2 4 6 8

10.3 LES BORNES DE STRINGER

Cette section traite d'un problème particulier qui surgit dans le domaine de la vérification comptable, et d'une solution particulière qui semble répandue dans le domaine, bien que ses propriétés n'aient pas été établies rigoureusement. Le problème est le suivant. Un vérificateur tire un échantillon de comptes, les corrige s'il y a lieu, et estime ensuite le montant total de l'erreur, τ. Par exemple, afin de vérifier les demandes de remboursement de taxes faites par des entreprises, les agences du revenu gouvernementales tirent un échantillon des factures soumises et déterminent si les remboursements demandés sont éligibles. À partir de l'échantillon on estime le montant total de l'erreur, soit la différence totale entre le montant réclamé et le montant dû. La raison pour laquelle les méthodes classiques sont inadéquates, c'est que la distribution des erreurs est souvent très loin d'être

normale. Typiquement, un échantillon comprendra un grand nombre de valeurs nulles, laissant un trop petit nombre de valeurs positives avec lesquelles estimer l'erreur totale. À moins que l'échantillon soit de très grande taille, le montant total de l'erreur est estimé à partir d'un très petit nombre de données quantitatives.

À chaque compte sont associées deux variables : X la valeur aux livres, et Y, la valeur corrigée. L'erreur est donc $X - Y$ et l'erreur relative est $D = \frac{X - Y}{X}$. On effectue un tirage par unité monétaire : la population est constituée des τ_x *dollars*. On tire avec remise un échantillon de n *dollars*. Chaque observation est l'erreur relative $d = \frac{x - y}{x}$ du compte auquel appartient le dollar sélectionné. D prend les valeurs $d_k = \frac{x_k - y_k}{x_k}$ avec probabilités $\frac{x_k}{\tau_x}$, k = 1,…, N, N étant le nombre de comptes dans la population. Si d est une observation échantillonnale, τ_{xd} estime le total. Cette procédure revient à sélectionner les comptes avec probabilités proportionnelles à la valeur aux livres. Nous supposerons que $d_i = \frac{x_i - y_i}{x_i} \geq 0$ pour toutes les unités de la population.

On tire donc n comptes avec remise. Soit $0 < d_m \leq d_{m-1} \leq \ldots \leq d_1 \leq 1$ les $m \leq n$ observations non-nulles. En général, m est nettement inférieur à n. L'objectif est de déterminer une borne supérieure à l'erreur relative totale $\tau = \sum_{k=1}^{\tau_x} d_k$. La procédure que nous décrivons ici est due à Teitlebaum (Teitlebaum, 1973); elle est décrite entre autres par Anderson et Teitlebaum (Anderson, 1973). Elle produit une borne supérieure au total des erreurs appelée *borne de Stringer*.

Nous décrivons la procédure et sa justification dans le cadre d'un petit exemple.

Exemple 10.8 *Les bornes de Stringer*

Dans un échantillon de n = 100 comptes, 3 comptes ont été trouvés erronés. La valeur aux livres est connue pour tous les comptes et leur total est τ_x = 1 000 000 $. Le tirage est un tirage par unité monétaire. Les n valeurs constituent un *éas* de taille 100 d'une population de taille τ_x. Les 3 valeurs relatives des erreurs sont d_3 = 0,3; d_2 = 0,6; d_1 = 0,9. Posons d_4 = 0.

Soit X_i le nombre de données dans l'échantillon supérieures à d_{i+1} (i = 0, 1, 2, 3). X_i est une variable de loi binomiale de paramètres n et p_i.

	Nombre de valeurs supérieures à la limite dans l'échantillon	
Limites	Observé	Loi
$d_4 = 0$	$X_3 = 3$	$\mathcal{B}(n; p_3)$
$d_3 = 0{,}3$	$X_2 = 2$	$\mathcal{B}(n; p_2)$
$d_2 = 0{,}6$	$X_1 = 1$	$\mathcal{B}(n; p_1)$
$d_1 = 0{,}9$	$X_0 = 0$	$\mathcal{B}(n; p_0)$

La proportion des valeurs de D comprise dans un intervalle $]d_{i+1}\ ;\ d_i]$ est $p_i - p_{i-1}$. Une borne supérieure aux valeurs de D dans cet intervalle est déterminée en supposant que toutes les valeurs dans un intervalle $]d_{i+1}\ ;\ d_i]$ sont égales à d_i :

Intervalle	Effectif	Valeur maximale
$]\ d_4\ ;\ d_3]$	$\tau_x(p_3 - p_2)$	$\tau_x(p_3 - p_2)d_3$
$]\ d_3\ ;\ d_2]$	$\tau_x(p_2 - p_1)$	$\tau_x(p_2 - p_1)d_2$
$]\ d_2\ ;\ d_1]$	$\tau_x(p_1 - p_0)$	$\tau_x(p_1 - p_0)d_1$
$]\ d_1\ ;\ 1]$	$\tau_x p_0$	$\tau_x p_0$

Donc la valeur totale est bornée par LS = $\tau_x[\ (p_3 - p_2)d_3 + (p_2 - p_1)d_2 + (p_1 - p_0)d_1 + p_0]$.

Les probabilités p_0, p_1, p_2 et p_3 n'étant pas connues, on remplace chacune par la borne supérieure d'un intervalle de confiance unilatéral (borné à droite seulement), utilisant les techniques de la section 4.5. Le paramètre p_0 est le paramètre de la distribution de X_0, dont la valeur observée est $X_0 = 0$. Nous allons donc déterminer un intervalle de confiance $[0\ ;\ p_0^*]$ et remplacer p_0 par p_0^* dans la formule ci-dessus. De même, nous remplacerons p_1 par p_1^*, la limite d'un intervalle de confiance $[0\ ;\ p_1^*]$ sachant que $X_1 = 1$. On fait de même avec p_2 et p_3. La borne de Stringer est alors

$$\text{Total} \le \hat{\tau}^*, \text{ où } \hat{\tau}^* = \tau_x[(p_3^* - p_2^*)d_3 + (p_2^* - p_1^*)d_2 + (p_1^* - p_0^*)d_1 + p_0^*].$$

Si on fixe $\alpha = 0{,}05$, les limites supérieures sont :

$$p_0^* = 0{,}02951305;\ p_1^* = 0{,}04655981;\ p_2^* = 0{,}06161920;\ p_3^* = 0.$$

$(p_3^* - p_2^*)d_3 + (p_2^* - p_1^*)d_2 + (p_1^* - p_0^*)d_1 + p_0^* = 0{,}05811825$ et la limite supérieure est $\hat{\tau}^* = 100\,000(0{,}05811825) = 5811{,}83$ \$.

Remarquez qu'il n'est pas évident, à partir de l'expression LS = $\tau_x[(p_3 - p_2)d_3 + (p_2 - p_1)d_2 + (p_1 - p_0)d_1 + p_0]$, que cette quantité est bornée supérieurement par $\hat{\tau}^*$. Mais LS peut s'écrire, après un réarrangement des termes, comme LS = $\tau_x[p_0(1-d_1) + p_1(d_1-d_2) + p_2(d_2-d_3) + p_3 d_3]$. Donc si on affirme que chaque p_i est borné supérieurement par p_i^*, on affirme du coup que LS est borné supérieurement par $\hat{\tau}^*$.

Voici une description formelle de la procédure

Bornes de Stringer

Soit $0 < d_m \leq d_{m-1} \leq \ldots \leq d_1 < 1$ les $m \leq n$ les m erreurs relatives non-nulles dans un échantillon de taille n tiré avec remise d'une population de τ_x dollars. La borne de Stringer de niveau 100(1-a) % est donnée par

$$\tau \leq \tau_x p_0^* + \tau_x \sum_{i=1}^{m} [p_i^* - p_{i-1}^*] d_i$$

où p_i^* est la valeur de p_i pour laquelle $P(X_i \leq i \mid p_i) = a$.

La technique des bornes de Stringer demeure controversée, car sa validité n'a pas été établie formellement : il n'a pas été démontré que l'intervalle $[0\,;\hat{\tau}^*]$ contient τ avec probabilité au moins égale à 1-a. Néanmoins, la justification intuitive présentée dans l'exemple tend à montrer que cette propriété est souvent vérifiée : la procédure prévoit des estimations pessimistes qui tendent à élargir l'intervalle de confiance, ce qui accroît ses chances de contenir le paramètre. L'autre côté de la médaille, c'est que la borne de Stringer s'est souvent révélée trop conservatrice, c'est-à-dire, inutilement large. C'est la conclusion qui découle des études de simulation, par Plante et al. (Plante & Neter, 1985) et par Joubir (Joubir, 2010). Le lecteur peut vérifier que, dans le dernier exemple, si on considérait les 100 observations comme un échantillon aléatoire avec remise, et déterminé un intervalle de confiance à 95 %, on aurait obtenu une borne supérieure de 4 141 $. Si la borne de Stringer est trop conservatrice par rapport à celle-ci, celle-ci ne l'est probablement pas assez.

10.4 DÉMONSTRATIONS

THÉORÈME 10.1

$\hat{\tau} = \frac{1}{n}\sum_{i=1}^{n} R_i$ est sans biais pour τ. Sa variance est $V(\hat{\tau}) = \frac{\sigma_r^2}{n}$, où $r_k = \frac{t_k}{\psi_k}$ et $\sigma_r^2 = \sum_{k=1}^{N} \frac{t_k^2}{\psi_k} - \tau^2$.

Un estimateur sans biais de $V(\hat{\tau})$ est donné par $\hat{V}(\hat{\tau}) = \frac{s_r^2}{n}$, où $s_r^2 = \frac{\sum_{i \in \omega} (r_i - \bar{r})^2}{n-1}$.

Démonstration

Soit R une variable dont les valeurs sont $r_k = \frac{t_k}{\psi_k}$, $k = 1, \ldots, N$ et R_i ($i = 1, \ldots, n$) la valeur de R observée au i^e tirage. Ces variables constituent un *éas* dans le sens classique : elles sont indépendantes et identiquement distribuées.

Chaque R_i prend les valeurs $r_1 = \frac{t_1}{\psi_1}, r_2 = \frac{t_2}{\psi_2}, \ldots, r_N = \frac{t_N}{\psi_N}$ avec probabilités ψ_1, ψ_2, …, ψ_N. Donc $E(R_i) = \sum_{k=1}^{N} r_k \psi_k = \sum_{k=1}^{N} \frac{t_k}{\psi_k} \psi_k = \tau$. La variance de R_i est $\sigma_r^2 = V(Z_i) = \sum_{k=1}^{N} [r_k - \tau]^2 \psi_k =$ $\sum_{k=1}^{N} r_k^2 \psi_k - \tau^2 = \sum_{k=1}^{N} \frac{t_k^2}{\psi_k} - \tau^2$.

La variance de $\hat{\tau}$ est $\frac{\sigma_r^2}{n}$: $V(\hat{\tau}) = V\left(\frac{1}{n}\sum_{i=1}^{n} R_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} V(R_i) = \frac{1}{n^2}\sum_{i=1}^{n} \sigma_r^2 = \frac{\sigma_r^2}{n}$. Par ailleurs, $s_r^2 = \frac{\sum_{i=1}^{n} (R_i - \bar{R})^2}{n-1}$ est un estimateur sans biais de σ_r^2 et donc $\hat{V}(\hat{\tau}) = \frac{s_r^2}{n}$ est un estimateur sans biais de $V(\hat{\tau})$.

THEOREME 10.2

$\hat{t} = \sum_{i \in s} \frac{y_i}{\pi_i}$ est un estimateur sans biais de τ.

Démonstration

Soit $Z_k = \begin{cases} 1 & \text{si } k \in \omega \\ 0 & \text{sinon} \end{cases}$. Alors $P(Z_k=1) = \pi_k$. et $E(Z_k) = \pi_k$.

$\hat{\tau} = \sum_{i \in \omega} \frac{t_i}{\pi_i} = \sum_{k=1}^{N} \frac{t_k}{\pi_k} Z_k$ et $E(\hat{\tau}) = \sum_{k=1}^{N} \frac{t_k}{\pi_k} E(Z_k) = \sum_{k=1}^{N} \frac{t_k}{\pi_k} \pi_k = \tau$.

La variance de Z_k est $V(Z_k) = \pi_k(1-\pi_k)$. La covariance entre Z_k et Z_ℓ est donnée par $C(Z_k ; Z_\ell) = E(Z_k Z_\ell) - E(Z_k)E(Z_\ell) = P(Z_k Z_\ell = 1) - \pi_k \pi_\ell = \pi_{k\ell} - \pi_k \pi_\ell$.

Propriétés des π_k et des $\pi_{k\ell}$:

a) $\sum_{i=1}^{N} \pi_i = n$; b) $\sum_{\ell \neq k}^{N} \pi_{k\ell} = (n-1)\pi_k$; c) $\sum_{\ell \neq k}^{N} (\pi_{k\ell} - \pi_k \pi_\ell) = -\pi_k(1-\pi_k)$.

Démonstration

a) $\sum_{k=1}^{N} \pi_k = \sum_{k=1}^{N} E(Z_k) = E[\sum_{k=1}^{N} Z_k] = E(n) = n$.

b) $\sum_{\ell \neq k}^{N} \pi_{k\ell} = \sum_{\ell \neq k}^{N} E(Z_k Z_\ell) = E[\sum_{\ell \neq k}^{N} Z_k Z_\ell] = E[Z_k \sum_{\ell \neq k}^{N} Z_\ell] = E[Z_k (\sum_{\ell=1}^{N} Z_\ell - Z_k)] = E[Z_k (n - Z_k] =$
$E[nZ_k - Z_k^2] = n\pi_k - \pi_k = (n-1)\pi_k$.

c) $\sum_{\ell \neq k}^{N} (\pi_{k\ell} - \pi_k \pi_\ell) = \sum_{\ell \neq k}^{N} \pi_{k\ell} - \sum_{\ell \neq k}^{N} \pi_k \pi_\ell = (n-1)\pi_k - \pi_k \sum_{\ell \neq k}^{N} \pi_\ell =$
$(n-1)\pi_k - \pi_k [\sum_{\ell=1}^{N} \pi_\ell - \pi_k] = (n-1)\pi_k - \pi_k [n - \pi_k] = -\pi_k(1-\pi_k)$.

On déduit de c) et du fait que $\pi_{k\ell} - \mu_k\pi_\ell = C(Z_k ; Z_\ell)$, que

$$V(Z_k) = -\sum_{\ell\neq k}^{N} C(Z_k;Z_\ell)\,. \tag{0.1}$$

Variance de $\hat{\tau}$

THÉORÈME 10.3

Première expression :

$$V(\hat{t}) = \sum_{k=1}^{N}\left(\frac{t_k}{\pi_k}\right)^2 V(Z_k) + 2\sum_{k=1}^{N-1}\sum_{\ell>k}\frac{t_k}{\pi_k}\frac{t_\ell}{\pi_\ell}C(Z_k\,;\,Z_\ell)$$

$$= \sum_{k=1}^{N}\left(\frac{t_k}{\pi_k}\right)^2 \pi_k(1-\pi_k) + 2\sum_{k=1}^{N-1}\sum_{\ell>k}\frac{t_k}{\pi_k}\frac{t_\ell}{\pi_\ell}(\pi_{k\ell}-\pi_k\pi_\ell)$$

Deuxième expression:

$$V_2 = V(\hat{t}) = \sum_{k=1}^{N-1}\sum_{\ell\neq k}(\pi_k\pi_\ell - \pi_{ij})\left(\frac{t_k}{\pi_k} - \frac{t_\ell}{\pi_\ell}\right)^2 .$$

Démonstration

La première expression est immédiate. Nous développons la deuxième.

$$V(\hat{t}) = \sum_{k=1}^{N}\left(\frac{t_k}{\pi_k}\right)^2 V(Z_k) + \sum_{k=1}^{N}\sum_{\ell\neq k}\frac{t_k}{\pi_k}\frac{t_\ell}{\pi_\ell}C(Z_k;Z_\ell)$$

Substituant l'expression (0.1) à $V(Z_k)$, on a

$$V(\hat{t}) = -\sum_{k=1}^{N}\left(\frac{t_k}{\pi_k}\right)^2\sum_{\ell\neq k}C(Z_k;Z_\ell) + \sum_{k=1}^{N}\sum_{\ell\neq k}\frac{t_k}{\pi_k}\frac{t_\ell}{\pi_\ell}C(Z_k;Z_\ell)$$

Par symétrie, le premier terme est

$$-\sum_{k=1}^{N}\left(\frac{t_k}{\pi_k}\right)^2\sum_{\ell\neq k}C(Z_k;Z_\ell) = -\sum_{k=1}^{N-1}\left(\frac{t_k}{\pi_k}\right)^2\sum_{\ell>k}C(Z_k;Z_\ell) - \sum_{k=1}^{N-1}\left(\frac{t_k}{\pi_k}\right)^2\sum_{\ell<k}C(Z_k;Z_\ell)\,.$$

Le deuxième terme est

$$\sum_{k=1}^{N}\sum_{\ell\neq k}\frac{t_k}{\pi_k}\frac{t_\ell}{\pi_\ell}C(Z_k;Z_\ell) = 2\sum_{k=1}^{N}\sum_{\ell>k}\frac{t_k}{\pi_k}\frac{t_\ell}{\pi_\ell}C(Z_k;Z_\ell)$$

On a donc

$$V(\hat{t}) = -\sum_{k=1}^{N-1}\left(\frac{t_k}{\pi_k}\right)^2\sum_{\ell>k}C(Z_k;Z_\ell) - \sum_{\ell=1}^{N-1}\left(\frac{t_\ell}{\pi_\ell}\right)^2\sum_{k<\ell}C(Z_k;Z_\ell) + 2\sum_{k=1}^{N}\sum_{\ell>k}\frac{t_k}{\pi_k}\frac{t_\ell}{\pi_\ell}C(Z_k;Z_\ell)$$

$$= -\sum_{k=1}^{N-1}\sum_{\ell>k}\left[\left(\frac{t_k}{\pi_k}\right)^2 + \left(\frac{t_\ell}{\pi_\ell}\right)^2 - 2\frac{t_k}{\pi_k}\frac{t_\ell}{\pi_\ell}\right]C(Z_k;Z_\ell) = -\sum_{k=1}^{N-1}\sum_{\ell>k}\left(\frac{t_k}{\pi_k} - \frac{t_\ell}{\pi_\ell}\right)^2 C(Z_k;Z_\ell)$$

$$= \sum_{k=1}^{N} \sum_{\ell>k} (\pi_k \pi_\ell - \pi_{k\ell}) \left(\frac{t_k}{\pi_k} - \frac{t_\ell}{\pi_\ell} \right)^2 .$$

Nous avons donc deux expressions pour $V(\hat{t})$:

- $V_1 = V(\hat{t}) = \sum_{k=1}^{N} t_k^2 \frac{(1-\pi_k)}{\pi_k} + \sum_{k=1}^{N-1} \sum_{\ell \neq k} \frac{t_k}{\pi_k} \frac{t_\ell}{\pi_\ell} (\pi_{k\ell} - \pi_k \pi_\ell)$; et
- $V_2 = V(\hat{t}) = \sum_{k=1}^{N-1} \sum_{\ell>k} (\pi_k \pi_j - \pi_{k\ell}) \left(\frac{t_k}{\pi_k} - \frac{t_\ell}{\pi_\ell} \right)^2 .$

Estimateurs de la variance :

THEOREME 10.4

L'estimateur $\hat{V}_1(\hat{\tau})$ s'écrit comme

$$\hat{V}_1(\hat{t}) = \sum_{k=1}^{N} t_k^2 \frac{(1-\pi_k)}{\pi_k^2} Z_k + 2 \sum_{k=1}^{N-1} \sum_{\ell>k} t_k t_\ell \frac{\pi_{k\ell} - \pi_k \pi_\ell}{\pi_k \pi_\ell \pi_{k\ell}} Z_k Z_\ell .$$

Son espérance est donc

$$E[\hat{V}_1(\hat{t})] = \sum_{k=1}^{N} t_k^2 \frac{(1-\pi_k)}{\pi_k^2} E(Z_k) + 2 \sum_{k=1}^{N-1} \sum_{\ell>k} t_k t_\ell \frac{\pi_{k\ell} - \pi_k \pi_\ell}{\pi_k \pi_\ell \pi_{k\ell}} E(Z_k Z_\ell) =$$

$$\sum_{k=1}^{N} t_k^2 \frac{\pi_k (1-\pi_k)}{\pi_k^2} \pi_k + 2 \sum_{k=1}^{N-1} \sum_{\ell>k} t_k t_\ell \frac{\pi_{k\ell} - \pi_k \pi_\ell}{\pi_k \pi_\ell \pi_{k\ell}} \pi_{k\ell} = \sum_{k=1}^{N} t_k^2 \frac{1-\pi_k}{\pi_k} + 2 \sum_{k=1}^{N-1} \sum_{\ell>k} t_k t_\ell \frac{\pi_{k\ell} - \pi_k \pi_\ell}{\pi_k \pi_\ell} = V_1.$$

De même,

$$\hat{V}_2(\hat{t}) = \sum_{k=1}^{N-1} \sum_{\ell>k} \frac{\pi_k \pi_\ell - \pi_{k\ell}}{\pi_{k\ell}} \left(\frac{t_k}{\pi_k} - \frac{t_\ell}{\pi_\ell} \right)^2 Z_k Z_\ell$$

$$\Longrightarrow E[\hat{V}_2(\hat{t})] = \sum_{k=1}^{N-1} \sum_{\ell>k} \frac{\pi_k \pi_\ell - \pi_{k\ell}}{\pi_{k\ell}} \left(\frac{t_k}{\pi_k} - \frac{t_\ell}{\pi_\ell} \right)^2 E(Z_k Z_\ell)$$

$$= \sum_{k=1}^{N-1} \sum_{\ell>k} (\pi_k \pi_\ell - \pi_{k\ell}) \left(\frac{t_k}{\pi_k} - \frac{t_\ell}{\pi_\ell} \right)^2 = V_2.$$

THÉORÈME 10.5

Si ω est un échantillon qui contient k, alors $p(\omega) = \sum_{i \in \omega} \psi_i \frac{1}{\binom{N-1}{n-1}} = \sum_{i \in \omega, i \neq k} \psi_i \frac{1}{\binom{N-1}{n-1}} + \psi_k \frac{1}{\binom{N-1}{n-1}} =$

$\frac{(n-1)\bar{\psi}^*}{\binom{N-1}{n-1}} + \frac{\psi_k}{\binom{N-1}{n-1}}$, où $\bar{\psi}^*$ est la moyenne des $(n\text{-}1)$ ψ_i de l'échantillon différentes de k. Alors

$\pi_k = \sum_{\omega \ni k} p(\omega) = \sum_{\omega \ni k}\left[\frac{(n-1)\bar{\psi}*}{\binom{N-1}{n-1}} + \frac{\psi_k}{\binom{N-1}{n-1}}\right] = \sum_{\omega \ni k}\left[\frac{(n-1)\bar{\psi}*}{\binom{N-1}{n-1}} + \frac{\psi_k}{\binom{N-1}{n-1}}\right]$. La somme $\sum_{\omega \ni k}\bar{\psi}*\frac{1}{\binom{N-1}{n-1}}$ peut être interprétée comme l'espérance, dans un *éas*, de la moyenne d'un échantillon de taille *n*-1 tiré de la population constituée de tous les ψ_ℓ autres que *k*. Elle est égale donc à la moyenne de la population, soit $\frac{\sum_{\ell \neq k}\psi_\ell}{N-1} = \frac{1-\psi_k}{N-1}$. Le nombre d'échantillons qui contiennent *k* est $\binom{N-1}{n-1}$, et donc $\sum_{\omega \ni k}\frac{\psi_k}{\binom{N-1}{n-1}} = \psi_k$. Rassemblant les résultats, on obtient $\pi_k = \psi_k + \frac{n-1}{N-1}(1-\psi_k)$

$= \frac{N-n}{N-1}\psi_k + \frac{n-1}{N-1}$.

Soit ω un échantillon comprenant les unités *k* et ℓ. Sa probabilité est

$p(\omega) = \sum_{i \in \omega, i \neq k, i \neq \ell}\psi_i \frac{1}{\binom{N-1}{n-1}} + \psi_k \frac{1}{\binom{N-1}{n-1}} + \psi_\ell \frac{1}{\binom{N-1}{n-1}} = (n-2)\bar{\psi}^* \frac{1}{\binom{N-1}{n-1}} + \psi_k \frac{1}{\binom{N-1}{n-1}} + \psi_\ell \frac{1}{\binom{N-1}{n-1}}$, où $\bar{\psi}^*$ est la moyenne des *n*-2 données de ω excluant *k* et ℓ. Alors $\pi_{k\ell} =$

$\frac{1}{\binom{N-1}{n-1}}\sum_{\omega \ni k, \omega \ni \ell}[(n-2)\bar{\psi}^* + \psi_k + \psi_\ell] = \frac{(n-2)\binom{N-2}{n-2}}{\binom{N-1}{n-1}}\sum_{\omega \ni k, \omega \ni \ell}\bar{\psi}^* \frac{1}{\binom{N-2}{n-2}} + \frac{\binom{N-2}{n-2}\psi_k}{\binom{N-1}{n-1}} + \frac{\binom{N-2}{n-2}\psi_\ell}{\binom{N-1}{n-1}}$

$= \frac{(n-2)\binom{N-2}{n-2}}{\binom{N-1}{n-1}}\frac{[1-(\psi_k+\psi_\ell)]}{N-2} + \frac{\binom{N-2}{n-2}\psi_k}{\binom{N-1}{n-1}} + \frac{\binom{N-2}{n-2}\psi_\ell}{\binom{N-1}{n-1}}$

$= \frac{(n-1)(n-2)}{(N-1)(N-2)}[1-(\psi_k+\psi_\ell)] + \frac{n-1}{N-1}(\psi_k+\psi_\ell)$

$= \frac{(n-1)(n-2)}{(N-1)(N-2)}[1-(\psi_k+\psi_\ell)] + \frac{n-1}{N-1}(\psi_k+\psi_\ell)$

$= \frac{(n-1)(n-2)}{(N-1)(N-2)} + \frac{(n-1)(N-n)}{(N-1)(N-2)}(\psi_k+\psi_\ell)$.

EXERCICES

10.1 Une petite ville est constituée de 20 quartiers. On doit tirer un échantillon de quartiers afin d'estimer le nombre total τ d'enfants dans la ville. Voici les données de la population. M_i est le nombre de logements dans le quartier $i.$, t_i le nombre total d'enfants.

#	M_i	t_i	#	M_i	t_i	#	M_i	t_i	#	M_i	t_i
1	37	58	6	38	58	11	33	49	16	34	36
2	34	34	7	36	37	12	40	53	17	37	51
3	33	36	8	35	43	13	41	59	18	36	43
4	33	30	9	40	38	14	41	64	19	31	33
5	33	37	10	39	52	15	38	41	20	36	46

On tire un échantillon de deux grappes de la façon suivante: on tire une grappe avec probabilités $\psi_k = M_k/M_0$, où $M_0 = \sum_{k=1}^{20} M_k = 725$. Si i est l'unité tirée en premier, on tire une seconde avec probabilités $\dfrac{\psi_\ell}{1-\psi_k}$.

a) Montrer que $\pi_{k\ell} = \psi_k\psi_\ell[1/(1-\psi_k)+1/(1-\psi_\ell)]$.

b) Montrer que $\pi_k = \dfrac{\psi_k(1-2\psi_k)}{(1-\psi_k)} + \psi_k\sum_{\ell=1}^{N}\dfrac{\psi_\ell}{(1-\psi_\ell)}$.

c) Vérifier que $\pi_1 = 0{,}102029635$, $\pi_2 = 0{,}093971515$, et que $\pi_{12} = 0{,}005033156$

$[\sum_{\ell=1}^{N}\psi_\ell/(1-\psi_\ell) = 1{,}053008]$.

d) Supposons que les deux grappes sélectionnées sont les grappes 1 et 2. Estimer τ par l'estimateur de Horvitz-Thompson.

e) Estimer la variance de l'estimateur à l'aide de l'estimateur (10.6) [la variance réelle est 10 837,6].

f) Estimer la variance de l'estimateur à l'aide de l'estimateur (10.7).

10.2 Au numéro 10.1. supposons qu'on tire un échantillon de trois quartiers par la même méthode. On tombe sur les grappes 14, 12, et 16. Voici quelques données sur l'échantillon:

Quartier #	14	12	16
M_i	41	40	34
π_i	0,1684589	0,1646098	0,1412346
t_i	64	53	36

Voici les probabilités de sélection doubles: π_{12} = 0,01969343; π_{13} = 0,01675199; π_{23} = 0,01634538, où les indices 1, 2, et 3 désignent les quartiers 14, 12 et 16. respectivement.

a) À partir des données de l'échantillon, évaluer l'estimateur de Horvitz-Thompson.

b) Estimer la variance de l'estimateur à l'aide de la formule (10.6) [La variance réelle est 6908].

c) Estimer la variance de l'estimateur à l'aide de la formule(10.7) .

10.3 On vérifie une population de 8 654 factures émises par un restaurant sur une période de 364 jours. À cette fin, on tire un échantillon de 5 journées, utilisant la méthode de Lahiri: un premier tirage avec probabilités proportionnelles au nombre de factures M_i. puis un échantillon aléatoire simple de taille n-1 parmi les journées restantes. Le tableau suivant présente les données de l'échantillon, les M_i, et les montants totaux t_i des factures de la i^e journée :

M_i	15	25	23	31	18
t_i	1298,35	2273,55	2202,25	2718,6000	1491,55

a) Estimer le montant total des factures.

b) Estimer l'écart-type de l'estimateur à l'aide de l'estimateur (10.6).

c) Estimer l'écart-type de l'estimateur à l'aide de l'estimateur (10.7).

10.4 Même contexte qu'au numéro précédent. On procède, cependant, par une stratification selon le jour de la semaine. On a donc 52 journées dans chaque strate, et on tire un échantillon de taille 2 dans chaque strate.

	Première unité		Deuxième unité			
	Montant total	Nombre de factures	Montant total	Nombre de factures	Nombre total de factures	$\sum_{i=1}^{N_h} \frac{\psi_{hi}}{1-\psi_{hi}}$
Lundi	1373,10	16	1144,25	13	1059	1,020922
Mardi	2420,75	28	2188,45	24	1158	1,020261
Mercredi	1158,05	15	1246,60	15	1049	1,020357
Jeudi	1713,50	21	1353,55	18	1097	1,020457
Vendredi	1343,20	15	1580,10	20	1057	1,020246
Samedi	2949,75	35	2539,20	31	1535	1,020386
Dimanche	3105,00	36	3796,15	35	1699	1,020017

La procédure dans chaque strate consiste à tirer une unité avec probabilités proportionnelles au nombre de factures, puis une deuxième parmi les 51 journées restantes. avec probabilités proportionnelles au nombre de factures. L'ensemble des résultats est présenté au tableau précédent.

a) Vérifiez les probabilités de sélection :

π_i	0,030301	0,048250	0,028682	0,038304	0,028465	0,045535	0,042343
π_j	0,024656	0,041432	0,028682	0,032879	0,037861	0,040386	0,041180
π_{ij}	0,000376	0,001025	0,000415	0,000640	0,000546	0,000941	0,000891

b) Vérifier les estimateurs des totaux des strates.

c) Déterminer les valeurs de l'estimateur de l'écart-type selon (10.6).

d) Déterminer les valeurs de l'estimateur de l'écart-type selon (10.7).

e) Déterminer l'estimateur du total de la population.

f) Estimer l'écart-type de l'estimateur selon (10.7).

10.5 Une procédure naturelle pour tirer un échantillon avec probabilités inégales procède comme ceci: on effectue un premier tirage avec probabilités ψ_k; si l'unité k est sélectionnée en premier, la deuxième est ensuite sélectionnée parmi les N-1 unités restantes avec probabilités $\frac{\psi_\ell}{1-\psi_k}$, $\ell \neq k$; une troisième est ensuite sélectionnée avec probabilités $\frac{\psi_h}{1-\psi_k-\psi_\ell}$; ainsi de suite.

a) Soit $\hat{\tau}^* = \frac{t_i + t_j}{2(N-1)\psi_i\psi_j/(1-\psi_i)}$ un estimateur de τ dans un échantillon de taille n = 2, où i est la première unité sélectionnée et j la deuxième. Montrer que τ^* est sans biais.

b) Montrer que l'estimateur $\hat{\tau} = \frac{t_i + t_j}{(N-1)\psi_i\psi_j[1/(1-\psi_i)+1/(1-\psi_j)]}$ est aussi sans biais.

c) Montrer que $V(\hat{\tau}) \leq V(\hat{\tau}^*)$.

10.6 Au numéro 10.5, l'estimateur $\hat{\tau}^*$, qui dépend de l'ordre dans lequel les données ont été sélectionnées, est moins bon que $\hat{\tau}$, qui ne dépend pas de l'ordre. Ce résultat est un cas particulier d'un théorème plus général selon lequel pour tout estimateur sans biais $\hat{\tau}^*$ qui tient compte de l'ordre, il existe un estimateur sans biais $\hat{\tau}$ qui ne tient pas compte de l'ordre et tel que $V(\hat{\tau}) \leq V(\hat{\tau}^*)$. Désignons par ω^* un élément de l'ensemble Ω^* des $\frac{N!}{(N-n)!}$ échantillons *ordonnés* sans biais et par ω un élément de l'ensemble Ω des $\binom{N}{n}$ échantillons non ordonnés. Soit $\hat{\tau}^*(\omega^*)$ un estimateur sans biais de τ défini sur Ω^*. Montrer que l'estimateur $\hat{\tau}(\omega) = \hat{\tau}^*(\omega^*)p(\omega^*|\omega)$ est sans biais. Précisons la notation. À chaque ω correspondent $n!$ échantillons ω^*, et l'indice de sommation $\omega^*|\omega$ désigne l'ensemble des ω^* qui correspondent à un même ω. On écrit $\hat{\tau}^*(\omega^*)$ pour souligner le fait que $\hat{\tau}^*$ dépend de l'ordre, contrairement à $\hat{\tau}(\omega)$ qui en est indépendant. Montrer que $V(\hat{\tau}) \leq V(\hat{\tau}^*)$.

11 Tirage par grappes à deux degrés

Le tirage par grappes à *un degré* décrit dans les deux derniers chapitres prévoit qu'une grappe sélectionnée est recensée entièrement. Mais il arrive qu'on doive se contenter d'un *échantillon*, soit parce que la grappe est trop importante, soit parce qu'un tirage avec moins d'unités secondaires et plus d'unités primaires constitue une utilisation plus efficace des ressources. On effectue alors un *tirage à deux degrés*: en un premier temps, on sélectionne les grappes; et dans chacune d'elles on tire un échantillon d'unités secondaires. Une nouvelle source d'erreur est donc introduite et les propriétés des estimateurs s'en trouvent modifiées.

On conçoit facilement à quel point un plan d'échantillonnage peut se compliquer par les combinaisons de modes de tirage: les grappes peuvent être tirées de plusieurs façons et les unités secondaires aussi. Les unités secondaires peuvent elles-mêmes être des grappes et l'on pourrait effectuer un tirage à deux degrés à l'intérieur des grappes, ainsi de suite. Il est impossible d'établir des formules pour toutes les combinaisons possibles. Heureusement, cela n'est pas nécessaire. Nous décrirons ici une procédure qui permet de passer d'une méthode d'échantillonnage à un degré à une méthode d'échantillonnage correspondante dans laquelle les paramètres des grappes sélectionnés sont eux-mêmes estimés.

Nous commençons par le cas le plus simple, n'exigeant que très peu d'outillage nouveau: le cas des tirages avec remise.

11.1 PREMIER DEGRÉ: TIRAGE AVEC REMISE ET PROBABILITÉS PROPORTIONNELLES À LA TAILLE

On effectue n tirages consécutifs, avec remise et probabilités ψ_k, $\sum_{k=1}^{N}\psi_k = 1$. Dans un tirage à un degré, l'estimateur du total est $\hat{\tau} = \frac{1}{n}\sum_{i=1}^{n}\frac{t_i}{\psi_i}$. Dans un tirage à deux degrés, on remplace t_i par un estimateur $\hat{t}_i$ déterminé à partir de l'échantillon tiré dans la grappe i. Outre ce changement, la procédure est identique à celle présentée à la section 10.1.

THÉORÈME 11.1 **TIRAGE AVEC REMISE ET PROBABILITÉS ψ_i**

Supposons qu'il existe un estimateur sans biais $\hat{t}_i$ dans la grappe *i*. Alors

(11.1) $$\hat{\hat{\tau}} = \frac{1}{n}\sum_{i=1}^{n} \hat{R}_i \text{ , où } \hat{R}_i = \frac{\hat{t}_i}{\psi_i} \quad .$$

Un estimateur sans biais de la variance $V(\hat{\hat{\tau}})$ est donné par

(11.2) $$\hat{V}(\hat{\hat{\tau}}) = \frac{s_{\hat{r}}^2}{n} \text{ , où } s_{\hat{r}}^2 = \frac{\sum_{i=1}^{n}(\hat{R}_i - \bar{\hat{R}})^2}{n-1} \quad .$$

Remarquez que nous n'avons pas spécifié la nature de l'estimateur $\hat{t}_i$. La conclusion du théorème demeure vraie, quel que soit l'estimateur $\hat{t}_i$, pourvu qu'il soit sans biais (ou à peu près.)

Remarque *Une même grappe peut être échantillonnée plus d'une fois*

Étant donné que le tirage au premier degré se fait avec remise, une même grappe peut se retrouver plus d'une fois dans l'échantillon. Il est entendu dans le Théorème 11.2 que le tirage au deuxième degré est repris autant de fois que la grappe se présente dans l'échantillon. C'est ce qui permet d'affirmer que $\hat{t}_i$ et $\hat{t}_j$ sont indépendants, même si *i* et *j* sont la même grappe.

Exemple 11.1 *Tirage avec remise et probabilités proportionnelles au premier degré*

Afin d'estimer le nombre total de femmes dans une population de 6 623 logements répartis en 180 quartiers, on tire avec remise un échantillon de 10 quartiers, avec probabilités proportionnelles au nombre de logements. Dans chaque quartier, on tire un *éas* de logements; on estime le nombre total de femmes dans le quartier par la moyenne. Voici les résultats:

Quartier #	$M_i = 6623\psi_i$	m_i	Échantillon	$\hat{t}_i$	$\hat{R}_i = \hat{t}_i / \psi_i$
64	47	9	1 1 1 0 2 2 1 2 1	57,44444	8094,778
65	35	7	1 0 1 1 1 1 1	30,00000	5676,857
98	38	8	3 1 1 0 1 2 1 2	52,25000	9106,625
105	35	7	3 1 1 1 0 1 1	40,00000	7569,143
105	35	7	2 1 1 1 1 3 1	50,00000	9461,429
139	36	7	1 0 0 1 0 0 0	10,28571	1892,286
149	40	8	1 1 2 3 1 2 0 0	50,00000	8278,750
166	33	7	1 0 1 1 1 1 3	37,71429	7569,143
172	36	7	2 1 1 2 2 2 2	61,71429	11353,714
173	35	7	1 0 1 1 1 2 1	35,00000	6623,000

Vous remarquez que l'unité 105 est présente deux fois. C'est l'inconvénient d'un échantillonnage avec remise. Le nombre total de femmes dans la population est estimé à $\hat{\tau} = \frac{1}{10}\sum_{i\in\omega} \hat{R}_i$ = 7563. L'écart-type des $\hat{R}_i$ est $s_{\hat{r}}$ = 2533,106. Donc $\hat{V}(\hat{\tau}) = \frac{s_{\hat{r}}}{\sqrt{n}} = 801$.

11.2 UN THÉORÈME FONDAMENTAL

Considérons maintenant le cas plus complexe où les tirages se font sans remise, avec probabilités égales ou pas. Supposons que pour un tirage à un degré il existe un estimateur $H[\boldsymbol{t}(\omega)]$ sans biais de $\tau = \sum_{k=1}^{N} t_k$, où $\boldsymbol{t}(\omega)$ est le vecteur des totaux t_i dans l'échantillon de grappes ω. De même. on suppose qu'il existe un estimateur $Q[\boldsymbol{t}(\omega)]$ de la variance de $H[\boldsymbol{t}(\omega)]$. Donc H et Q satisfont les conditions suivantes:

$$E_1\{H[\boldsymbol{t}(\omega)]\} = \tau \ , \quad \text{et} \quad E_1\{Q[\boldsymbol{t}(\omega)]\} = V_1\{H[\boldsymbol{t}(\omega)]\}$$

L'indice « 1 » dans E_1 et V_1 indique que l'espérance et la variance sont prises par rapport aux tirages au premier degré — le tirage des grappes.

Supposons maintenant que t_i n'est pas observé, mais qu'il est plutôt estimé à partir d'un échantillon tiré dans la grappe. On suppose ceci:

- Il existe un estimateur sans biais $\hat{t}_i$ du total t_i de la grappe sélectionnée i. C'est-à-dire, $E_2(\hat{t}_i) = t_i$.
- Il existe un estimateurs sans biais $\hat{\sigma}_i^2$ de la variance $\sigma_i^2 = V_2(\hat{t}_i)$.

L'indice « 2 » signifie qu'il s'agit de l'espérance et la variance par rapport au tirage des unités secondaires. On suppose, comme c'est généralement le cas, que les $\hat{t}_i$ ainsi que les $\hat{\sigma}_i^2$ sont mutuellement indépendants. Le théorème suivant propose un estimateur de τ ainsi qu'un estimateur de la variance de l'estimateur.

THÉORÈME 11.2 **VARIANCE DE L'ESTIMATEUR DU TOTAL DANS UN ÉCHANTILLON À DEUX DEGRÉS**

Soit $\boldsymbol{t}(\omega)$ le vecteur des t_i dans ω. Supposons que $\hat{\tau} = H[\boldsymbol{t}(\omega)]$ est un estimateur sans biais de τ; et $Q[\boldsymbol{t}(\omega)]$ un estimateur sans biais de $V(\hat{\tau})$ dans un tirage à un degré. Dans un échantillon à deux degrés, soit $\hat{t}_i$ un estimateur sans biais de t_i et $\hat{\sigma}_i^2$ un estimateur sans biais de la variance $\sigma_i^2 = V_2(\hat{t}_i)$. Alors $\hat{\hat{\tau}} = H[\hat{\boldsymbol{t}}(\omega)]$ est un estimateur sans biais de τ, et

$$\hat{V}(\hat{\hat{\tau}}) = Q[\hat{\boldsymbol{t}}(\omega)] + H[\hat{\boldsymbol{\sigma}}^2(\omega)] \tag{11.3}$$

est un estimateur sans biais de la variance de $\hat{\hat{\tau}}$, où $\hat{\boldsymbol{t}}(\omega)$ est le vecteur des $\hat{t}_i$ dans ω et $\hat{\sigma}^2(\omega)$ le vecteur des $\hat{\sigma}_i^2$ dans ω.

Voici, concrètement, ce que dit le théorème:

$H[\mathtt{t}(\omega)]$ est la formule de l'estimateur de τ à un degré.

$Q\left[\boldsymbol{t}(\omega)\right]$ est la formule de l'estimateur de $\hat{V}(\hat{\tau})$ à un degré.

$H[\hat{\mathtt{t}}(\omega)]$ est l'estimateur de τ dans un échantillon à deux degrés: il consiste à substituer les estimateurs $\hat{t}_i$ aux t_i inconnus.

$Q\left[\hat{\boldsymbol{t}}(\omega)\right]$ est la formule $Q\left[\boldsymbol{t}(\omega)\right]$ dans laquelle on substitue les $\hat{t}_i$ aux t_i

$H[\hat{\boldsymbol{\sigma}}^2(\omega)]$ est la formule $H[\mathtt{t}(\omega)]$ dans laquelle on substitue les $\hat{\sigma}_i^2$ aux t_i

Nous illustrons la procédure dans quelques cas particuliers.

11.3 PREMIER DEGRÉ: ÉCHANTILLONNAGE ALÉATOIRE SIMPLE

On considère ici un tirage aléatoire simple aux deux degrés: un échantillon aléatoire simple de n grappes, puis un échantillon aléatoire simple de m_i unités secondaires dans la grappe i, $i = 1, \ldots, n$. Le Tableau 11.1 donne le sens des formules définies ci-dessus dans ce cas particulier .

Tableau 11.1
Échantillon aléatoire simple à deux degrés

Estimation par la moyenne

Formule	Signification
M_i	La taille de la grappe *i*
m_i	La taille de l'échantillon tiré dans la grappe *i*
$\overline{y}_i$	La moyenne de l'échantillon tiré dans la grappe *i*
$\hat{t}_i = M_i\overline{y}_i$	L'estimateur du total de la grappe *i*
s_i	Écart-type des données de la grappe *i*
$\hat{\sigma}_i^2 = M_i^2(1-m_i/M_i)s_i^2/m_i$	L'estimateur de variance de $\hat{t}_i$
$\hat{\tau} = H[\boldsymbol{t}(\omega)] = \frac{N}{n}\sum_{i\in\omega} t_i$	Estimateur du total dans un tirage à un degré. Ne peut être calculé dans un tirage à deux degrés
$\hat{\hat{\tau}} = H[\hat{\boldsymbol{t}}(\omega)] = \frac{N}{n}\sum_{i\in\omega}\hat{t}_i$	Estimateur du total dans un tirage à deux degrés.
$s_t^2 = \frac{\sum_{i\in\omega} t_i^2 - (\sum_{i\in\omega} t_i)^2/n}{(n-1)}$	Variance des totaux des grappes sélectionnées. Ne peut être calculé dans un tirage à deux degrés
$s_{\hat{t}}^2 = \frac{\sum_{i\in\omega}\hat{t}_i^2 - (\sum_{i\in\omega}\hat{t}_i)^2/n}{(n-1)}$	Variance des totaux estimés des grappes sélectionnées.
$Q[\boldsymbol{t}(\omega)] = V(\hat{\tau}) = N^2(1-f)\frac{s_t^2}{n}$	Variance estimée de $\hat{\tau}$ dans un échantillon à un degré
$Q[\hat{\boldsymbol{t}}(\omega)] = N^2(1-f)\frac{s_{\hat{t}}^2}{n}$	La formule de $Q[\boldsymbol{t}(\omega)]$ dans laquelle on remplace s_t par $s_{\hat{t}}$
$\hat{V}(\hat{\hat{\tau}}) = Q\left[\hat{\boldsymbol{t}}(\omega)\right] + H[\hat{\boldsymbol{\sigma}}^2(\omega)]$	Estimateur de la variance de $\hat{\hat{\tau}}$.

Dans un échantillon aléatoire simple de *n* grappes, l'estimateur à un degré du total τ est $\hat{\tau} = H[\boldsymbol{t}(\omega)] = \frac{N}{n}\sum_{i\in\omega} t_i$ et l'estimateur de la variance de $\hat{\tau}$ est $Q[\boldsymbol{t}(\omega)] = N^2(1-f)\frac{s_t^2}{n}$, où $f = n/N$. Dans un tirage à deux degrés, l'estimateur devient $\hat{\hat{\tau}} = H[\hat{\boldsymbol{t}}(\omega)] = \frac{N}{n}\sum_{i\in\omega}\hat{t}_i$, où $\hat{t}_i$ est une estimation du total de la grappe *i*. L'estimateur de sa variance est $Q[\hat{\boldsymbol{t}}(\omega)] + H[\widehat{\boldsymbol{\sigma}}^2(\omega)]$, où $Q[\hat{\boldsymbol{t}}(\omega)] = N^2(1-f)\frac{s_{\hat{t}}^2}{n}$, $s_{\hat{t}}^2 = \frac{\sum_{i\in\omega}\hat{t}_i^2 - (\sum_{i\in\omega}\hat{t}_i)^2/n}{(n-1)}$ et $H[\widehat{\boldsymbol{\sigma}}^2(\omega)] = \frac{N}{n}\sum_{i\in\omega}\hat{\sigma}_i^2$. Si dans la strate *i* l'échantillon est un *éas* et $\hat{t}_i$ est l'estimateur de t_i par la moyenne, $\hat{t}_i = M_i\overline{y}_i$, où $\overline{y}_i$ est la moyenne de l'échantillon tiré dans la strate *i*, alors $\hat{\sigma}_i^2 = M_i^2(1-m_i/M_i)s_i^2/m_i$. On a alors

$$\hat{V}(\hat{\hat{\tau}}) = \frac{N^2}{n}\frac{\sum_{i\in\omega}\hat{t}_i^2 - (\sum_{i\in\omega}\hat{t}_i)^2/n}{n-1} + \frac{N}{n}\sum_{i\in\omega} M_i^2(1-m_i/M_i)s_i^2/m_i \quad .$$

Exemple 11.2 *Échantillonnage à deux degrés; éas à deux degrés*

Le tableau A10 présente des données sur une population de 6 623 logements répartis en 180 quartiers. Afin d'estimer le nombre total de femmes dans le quartier, on tire un *éas* de 10 quartiers. Dans chaque quartier, on tire un *éas* d'environ 10 %, et on estime le total par la moyenne. Voici les résultats:

M_i	m_i	Échantillon	$\bar{y}_i$	$\hat{t}_i = M_i \bar{y}_i$	s_i^2	$\hat{\sigma}_i^2 = \hat{V}(\hat{t}_i)$
36	3	2 0 1	1,000	36	1,0000	396
43	4	1 1 2 1	1,250	53,75	0,2500	104,8
33	3	2 0 1	1,000	32	1,0000	330
37	4	1 2 1 2	1,500	55,5	0,3333	101,75
34	3	2 1 1	1,333	45,3333	0,3333	117,11
34	3	2 1 1	1,333	45,3333	0,3333	117,11
42	4	2 1 1 3	1,750	73,5	0,9167	365,75
34	4	1 1 1 1	1,000	34	0,0000	0
44	5	1 2 1 1 2	1,400	61,6	0,3000	102,96
42	4	1 1 0 1	0,750	31,5	0,2500	99,75

Si les grappes avaient été recensées au complet, l'estimateur de τ aurait été $\hat{\tau}$ = $H[t(\omega)] = N\dfrac{\sum_{i=1}^{n} t_i}{n}$. Donc l'estimateur ici sera $\hat{\hat{\tau}}$ = $H[\hat{t}(\omega)] = N\dfrac{\sum_{i=1}^{n} \hat{t}_i}{n}$ =180(46,95167) = 8 451. Dans un tirage à un degré, la variance de $\hat{\tau}$ aurait été estimée par $Q[\boldsymbol{t}(\omega)]$ = $N^2(1-n/N)\dfrac{\sum_{i\in\omega}(t_i-\bar{t})^2}{n(n-1)} = N^2(1-n/N)\dfrac{\sum_{i\in\omega} t_i^2 - (\sum_{i\in\omega} t_i)^2/n}{n(n-1)}$. Ici elle sera estimée par $Q\left[\hat{\boldsymbol{t}}(\omega)\right] + H[\hat{\boldsymbol{v}}(\omega)]$. On a $Q\left[\hat{\boldsymbol{t}}(\omega)\right]$ = $N^2(1-n/N)\dfrac{\sum_{i\in\omega}\hat{t}_i^2 - (\sum_{i\in\omega}\hat{t}_i)^2/n}{n(n-1)}$ = 600 101,6. $H[\hat{\boldsymbol{\sigma}}^2(\omega)] = N\dfrac{\sum_{i\in\omega}\hat{\sigma}_i^2}{n}$ = 31 234,4. Donc $\hat{V}(\hat{\tau})$ = 600 101,6 + 31 234,4 = 631 336 et $\sqrt{\hat{V}(\hat{\tau})}$ = 794.

Exemple 11.3 *Estimation par le quotient au premier degré*

Dans l'Exemple 11.2 utilisons au premier degré l'estimateur par le quotient (avec les tailles des grappes comme variables auxiliaires). Au premier degré le total est estimé par $\hat{\tau}$ = $H[\hat{\boldsymbol{t}}(\omega)] = M_0\bar{\bar{y}}$, où $\bar{\bar{y}} = \dfrac{\sum_{i\in\omega} t_i}{\sum_{i\in\omega} M_i}$, et la variance de l'estimateur par $Q[t(\omega)]$ = $M_0^2(1-f)\dfrac{s_t^2 + \bar{\bar{y}}^2 s_m^2 - 2\bar{\bar{y}}s_{mt}}{n\bar{M}^2}$. Au deuxième degré, $\hat{\hat{\tau}}$ = $H[\hat{t}(\omega)] = M_0\dfrac{\sum_{i\in\omega}\hat{t}_i}{\sum_{i\in\omega} M_i} = 6623(1,23883) =$ 8205. Pour évaluer $Q[\hat{\boldsymbol{t}}(\omega)]$, nous remplaçons t_i par $\hat{t}_i$ dans s_t et dans s_{mt}. Le lecteur peut vérifier

que nous obtenons s_t = 14,00399, s_m = 4,357624, et s_{mt} = 34,90389 et $\bar{M}$ = 37,9. Finalement, $Q[\hat{\boldsymbol{t}}(\omega)]$ = 400 235. $H[\boldsymbol{\sigma}^2(\omega)]$ = $M_0 \frac{\sum_{i\in\omega}\hat{\sigma}_i^2}{\sum_{i\in\omega}M_i} = 6623\frac{1735{,}245}{379}$ = 30 323,29. Donc $\hat{V}\left(\hat{\hat{\tau}}\right)$ = 400 235 + 30 323,29 = 430 558 et $\sqrt{\hat{V}\left(\hat{\hat{\tau}}\right)}$ = 656.

Noter que l'estimateur utilisé dans le dernier exemple est risqué, étant donné la petite taille (n = 10) de l'échantillon. Le seul but de l'exemple est d'illustrer les calculs.

11.3 ESTIMATEUR DE HORVITZ-THOMPSON AU PREMIER DEGRÉ

Considérons un tirage général de grappes dans lequel les probabilités de sélection des unités π_k sont connues. Les propriétés générales ont été établies dans la section 10.2. L'estimateur de τ dans un tirage à un degré, (10.3), mène à l'estimateur à deux degrés suivant:

(11.4) $$\hat{\hat{\tau}} = \sum_{i\in\omega}\frac{\hat{t}_i}{\pi_i}.$$

Les équations mènent aux estimateurs suivants de la variance:

(11.5) $$\hat{V}_1(\hat{\hat{\tau}}) = \sum_{i\in\omega}\hat{t}_i^2\frac{(1-\pi_i)}{\pi_i^2} + 2\sum_{i\in\omega}\sum_{\substack{j\in\omega\\ j>i}}\hat{t}_i\hat{t}_j\frac{\pi_{ij}-\pi_i\pi_j}{\pi_i\pi_j\pi_{ij}} + \sum_{i\in\omega}\frac{\hat{\sigma}_i^2}{\pi_i},$$

et l'estimateur de Yates-Grundy

(11.6) $$\hat{V}_2(\hat{\hat{\tau}}) = \sum_{i\in\omega}\sum_{\substack{j\in\omega\\ j>i}}\frac{\pi_i\pi_j-\pi_{ij}}{\pi_{ij}}\left(\frac{\hat{t}_i}{\pi_i}-\frac{\hat{t}_j}{\pi_j}\right)^2 + \sum_{i\in\omega}\frac{\hat{\sigma}_i^2}{\pi_i}.$$

Exemple 11.4 *Estimateur HT au premier degré*

D'une population de 725 logements répartie en 20 quartiers, on tire un échantillon de 3 quartiers. Le premier quartier est tiré avec probabilités proportionnelles aux nombres M_k de logements; les deux quartiers suivants avec probabilités proportionnelles aux tailles des quartiers restants. Les tailles M_k sont:

37 ; 34 ; 33 ; 33 ; 33 ; 38 ; 36 ; 35 ; 40 ; 39 ; 33 ; 40 ; 41 ; 41 ; 38 ; 34 ; 37 ; 36 ; 31 ; 36

Le tirage donne les quartiers 19, 1, et 12, dans cet ordre. Le calcul des probabilités de sélection est ardu:

$$\pi_i = \psi_i + \psi_i\sum_{j\neq i}\frac{\psi_j}{(1-\psi_j)} + \sum_{j\neq i}\sum_{\substack{k\neq j\\ k\neq i}}\frac{\psi_j\psi_k\psi_i}{(1-\psi_j)(1-\psi_j-\psi_k)} \;;\; \pi_{ij} = \pi_{ij}^* + \pi_{ji}^*, \text{ où}$$

$$\pi_{rs}^* = \frac{\psi_r\psi_s}{1-\psi_r} + \frac{\psi_r\psi_s}{1-\psi_r}\sum_{\substack{k\neq r\\ k\neq s}}\frac{\psi_k}{(1-\psi_r-\psi_k)} + \psi_r\psi_s\sum_{\substack{k\neq r\\ k\neq s}}\frac{\psi_k}{(1-\psi_r)(1-\psi_r-\psi_k)}.$$

Nous obtenons les probabilités suivantes (les quartiers sélectionnés, 19, 1, et 12, sont maintenant désignés par 1, 2, 3):

π_1 = 0,1293681; π_2 = 0,1529821; π_3 = 0,1646098; π_{12} = 0,01379411; π_{13} = 0,01490786; π_{23} = 0,01778139.

La variable Y est le nombre de femmes dans le logement; la variable auxiliaire est le nombre de ménages dans le quartier, supposé connu pour la population.

Les données sont présentées dans le tableau ci-dessous. Elles portent sur les échantillons tirés dans les trois grappes sélectionnées.

	Unité 19 (i = 1)	Unité1 (i = 2)	Unité 12 (i = 3)
M_i	31	37	40
m_i	13	21	21
Nombre de femmes	1,181818	1,615385	1,5
s_y	0,4045199	0,9607689	0,6504436

Estimations au deuxième degré

Les totaux et les écarts-types estimés des quartiers sont

	Unité 19 (i = 1)	Unité1 (i = 2)	Unité 12 (i = 3)
$\hat{t}_i = M_i \bar{y}_i$	36,63636	59,76923	60,00000
$\hat{\sigma}_i^2 = M_i^2(1 - m_i / M_i)\frac{s_{yi}^2}{m_i}$	9,22314	63,05325	31,42857

On estime donc le total de la population par $\hat{\hat{\tau}} = \sum_{i\in\omega} \frac{\hat{t}_i}{\pi_i}$ = $\frac{36,63636}{0,1293681} + \frac{59,76923}{0,1529821} + \frac{60}{0,1646098}$

= 1038,387.

Estimation de la variance:

Selon la formule (11.5) ,

$\frac{\hat{t}_1^2(1-\pi_1)}{\pi_i^2} + \frac{\hat{t}_2^2(1-\pi_2)}{\pi_2^2} + \frac{\hat{t}_2^2(1-\pi_2)}{\pi_2^2}$ = 310 103,8; $\hat{t}_1\hat{t}_2\frac{\pi_{12}-\pi_1\pi_2}{\pi_1\pi_2\pi_{12}} + \hat{t}_1\hat{t}_3\frac{\pi_{13}-\pi_1\pi_3}{\pi_1\pi_3\pi_{13}} + \hat{t}_2\hat{t}_3\frac{\pi_{23}-\pi_2\pi_3}{\pi_2\pi_3\pi_{23}}$ =

- 151 600,9; $\frac{\hat{\sigma}_1^2}{\pi_1} + \frac{\hat{\sigma}_2^2}{\pi_2} + \frac{\hat{\sigma}_2^2}{\pi_2}$ = 674,328. Donc $\hat{V}_1(\hat{\hat{\tau}})$ =310 103,8-2*(151600,9) + 674,3825 =

7 576,317

Selon la formule (11.6) :

$\frac{\pi_1\pi_2-\pi_{12}}{\pi_{12}}\left(\frac{\hat{t}_1}{\pi_1} - \frac{\hat{t}_2}{\pi_2}\right)^2 + \frac{\pi_1\pi_3-\pi_{13}}{\pi_{13}}\left(\frac{\hat{t}_1}{\pi_1} - \frac{\hat{t}_3}{\pi_3}\right)^2 + \frac{\pi_2\pi_3-\pi_{23}}{\pi_{23}}\left(\frac{\hat{t}_2}{\pi_2} - \frac{\hat{t}_3}{\pi_3}\right)^2$ = 8 142. Donc $\hat{V}_2(\hat{\hat{\tau}})$ = 8 141,79

+ 674,3825 = 8 816.

11.4 DEMONSTRATIONS

THÉORÈME 11.2 **THEOREME FONDAMENTAL**

$\hat{\hat{\tau}} = H[\hat{\boldsymbol{t}}(\omega)]$ est un estimateur sans biais de τ, et $\hat{V}(\hat{\hat{\tau}}) = H[\hat{\boldsymbol{\sigma}}^2(\omega)] + Q[\hat{\boldsymbol{t}}(\omega)]$ est un estimateur sans biais de $V(\hat{\hat{\tau}})$.

Démonstration

Un tirage à deux degrés est constitué de deux étapes: on tire *n* unité primaires, puis un échantillon d'unités secondaires dans chacune des unités primaires. Il n'est pas interdit, cependant, d'envisager la procédure inverse : On tire un échantillon dans chacune des grappes de la *population*, puis on tire *n* grappes. Cette interprétation est légitime dans la mesure où le mode de tirage au second degré est indépendant des résultats du premier. On a donc une population à laquelle sont associées les valeurs t_i, $\hat{t}_i$, σ_i^2 , et $\hat{\sigma}_i^2$. Si au premier degré, une fonction $H[\boldsymbol{t}(\omega)]$ satisfait $E_1\{H[\boldsymbol{t}(\omega)]\} = \sum_{i=1}^N t_i$, alors il est également vrai que

(11.7) $$E_{1|2}\{H[\hat{\boldsymbol{t}}(\omega)]\} = \sum_{i=1}^N \hat{t}_i \ ,$$

et

(11.8) $$E_{1|2}\{H[\hat{\boldsymbol{\sigma}}^2(\omega)]\} = \sum_{k=1}^N \hat{\sigma}_k^2 \ .$$

Par ailleurs, si au premier degré $E\{Q[\boldsymbol{t}(\omega)]\} = V\{H[\boldsymbol{t}(\omega)]\}$, alors

(11.9) $$E_{1|2}\{Q[\hat{\boldsymbol{t}}(\omega)]\} = V_{1|2}\{H[\hat{\boldsymbol{t}}(\omega)]\} \ .$$

L'indice « 1|2 » dans $E_{1|2}$ et $V_{1|2}$ signifie qu'il s'agit de l'espérance et la variance conditionnelles étant donné les valeurs $\hat{\boldsymbol{t}}(\Omega)$ et $\hat{\sigma}^2(\Omega)$.

L'équation (11.7) entraîne que $H[\hat{\boldsymbol{t}}(\omega)]$ est sans biais pour τ, car $E\{H[\hat{\boldsymbol{t}}(\omega)]\} = E_2 E_{1|2}\{H[\hat{\boldsymbol{t}}(\omega)]\} = E_2\left[\sum_{i=1}^N \hat{t}_i\right]$, et par l'indépendance des tirages intra-grappes, $E_2\left[\sum_{k=1}^N \hat{t}_k\right] = \sum_{k=1}^N E_2(\hat{t}_k) = \sum_{k=1}^N t_k = \tau$. Par le même argument, (11.8) entraîne que

(11.10) $$E\{H[\hat{\boldsymbol{\sigma}}^2(\omega)]\} = \sum_{k=1}^N \sigma_k^2 \ ,$$

et (11.9) entraîne que

(11.11) $$E\{Q[\hat{\boldsymbol{t}}(\omega)]\} = E_2 V_{1|2}\{H[\hat{\boldsymbol{t}}(\omega)]\} \ .$$

En prenant la variance au deuxième degré de (11.7) on obtient

(11.12) $$V_2\left[E_{1|2}\left\{H[\hat{\boldsymbol{t}}(\omega)]\right\}\right] = V_2\left[\sum_{k=1}^{N}\hat{t}_k\right] = \sum_{k=1}^{N}\sigma_k^2 ,$$

et par (11.10) et (11.12), il s'ensuit que

(11.13) $$E\{H[\hat{\boldsymbol{\sigma}}^2(\omega)]\} = V_2[E_{1|2}(H[\hat{\boldsymbol{t}}(\omega)])] .$$

Maintenant, la variance de $\hat{\hat{\tau}}$ est

(11.14) $$V(\hat{\hat{\tau}}) = V\left\{H\left[\hat{\boldsymbol{t}}(\omega)\right]\right\} = E_2\left[V_{1|2}\left\{H[\hat{\boldsymbol{t}}(\omega)]\right\}\right] + V_2\left[E_{1|2}\left\{H[\hat{\boldsymbol{t}}(\omega)]\right\}\right] .$$

L'équation (11.10) entraîne que $Q[\hat{\boldsymbol{t}}(\omega)]$ est un estimateur sans biais du premier terme et que (11.13) est un estimateur sans biais du deuxième terme. Donc l'estimateur

(11.15) $$\hat{V}(\hat{\hat{\tau}}) = Q[\hat{\boldsymbol{t}}(\omega)] + H[\boldsymbol{\sigma}^2(\omega)]$$

est sans biais pour $V(\hat{\hat{\tau}})$.

Généralisation

Le théorème peut être généralisé de façon à être applicable à des estimateurs basés sur des variables auxiliaires. Sans entrer dans les détails, nous indiquons les ajustements nécessaires. Au premier degré, on exprime $\hat{\tau}$ comme fonction $u(\boldsymbol{t}\,;\,\boldsymbol{x})$, où $\boldsymbol{x}_i$ est le vecteur des valeurs d'une variable auxiliaire. De même, on exprime σ_i^2 comme fonction $\sigma_i^2 = v(\boldsymbol{t}\,;\,\boldsymbol{x})$. À partir d'ici, les énoncés des hypothèses et des conclusions demeurent les mêmes: $u(\hat{\boldsymbol{t}}\,,\boldsymbol{x})$ est un estimateur sans biais de τ; $\hat{\sigma}_i^2 = v(\hat{\boldsymbol{t}}\,,\boldsymbol{x})$ est sans biais pour $V_2(\hat{t}_i)$; etc.

EXERCICES

11.1 Afin d'échantillonner une population de 5 544 logements répartis en 200 quartiers on tire un *éas* de 10 quartiers. La variable d'intérêt est *Y*, la consommation d'énergie (en giga joules). On dispose, par ailleurs, d'une information complète sur une variable auxiliaire *X*, la superficie du logement (en m^2). Dans chaque quartier, on tire un *éas* de logements, et on estime son total t_i par la moyenne. Voici, au tableau qui suit les données recueillies.

Toutes les statistiques dans le tableau, sauf celles des première et dernière colonnes, portent sur les échantillons de ménages tirés dans les quartiers sélectionnés: On désigne par ω_i l'échantillon tiré dans la grappe *i*:

M_i : la taille de la grappe (le quartier) *i*;

m_i : la taille de l'échantillon tiré dans la grappe *i*;

Σx, Σy : les sommes des *X* et des *Y* dans les ω_i;

s_x et s_y : les écarts-types des *X* et des *Y* dans les ω_i;

s_{xy} : la covariance entre X et Y dans les ω_i;

μ_x : la moyenne des X dans le quartier sélectionné.

M_i	m_i	Σx	Σy	s_x	s_y	s_{xy}	μ_x
24	10	1561,69	1095,13	9,221774	5,945516	37,80096	152,73333
33	13	811,28	799,99	5,864493	6,614463	28,41832	61,38515
41	16	1028,78	1009,04	7,457107	9,207889	59,22779	63,04976
23	9	1076,30	884,21	6,285279	6,744865	17,36679	117,90913
21	8	1726,09	1032,26	23,745186	16,496970	376,41974	206,05190
16	6	542,19	492,04	7,231962	9,392394	56,90468	89,45687
31	12	1408,21	1133,21	6,255597	7,248030	38,95916	118,26032
25	10	393,59	453,49	6,584373	7,811916	38,18491	40,69520
31	12	3555,37	1777,47	25,308825	13,482344	327,91456	294,00677
35	14	2790,35	1678,37	20,639611	11,749135	228,69186	206,50086

a) Estimer τ lorsqu'on utilise l'estimation par la moyenne au premier degré ainsi qu'au second degré.

b) Estimer l'écart-type de l'estimateur en a).

c) Estimer τ lorsqu'on utilise l'estimation par le quotient au premier degré (avec la superficie des logements comme variables auxiliaire) et l'estimation par la moyenne au second degré (superficie moyennes des logements: 137,4236 m^2.

d) Estimer l'écart-type de l'estimateur en c).

11.2 Une population comprend 3 343 fermes regroupées en 200 régions géographiques. Afin d'estimer la production totale τ d'orge dans la saison, on prélève un *éas* de 15 régions, puis un *éas* de fermes dans chacune des régions sélectionnées. Les totaux des régions sont estimés par la moyenne.

Voici les données (la notation est définie au numéro 11.1):

M_i	m_i	Σy	s_y	$\hat{t}_i$	$\sqrt{\hat{V}(\hat{t}_i)}$
16	6	773	74,34895	2061,3333	383,9363
16	6	885	120,17279	2360,0000	620,5696
18	7	597	110,38073	1535,1429	587,0523
20	8	223	43,36370	557,5000	237,5128
18	7	324	68,47558	833,1429	364,1827
27	11	295	53,15227	724,0909	333,0944
17	7	142	34,66369	344,8571	170,8245
11	4	333	96,51036	915,7500	423,4375
13	5	379	85,65454	985,4000	390,6448
17	7	461	93,34065	1119,5714	459,9877
15	6	617	109,80786	1542,5000	520,8644
18	7	165	41,36769	424,2857	220,0112
15	6	738	111,51681	1845,0000	528,9707
20	8	651	95,35338	1627,5000	522,2720
20	8	188	66,46804	470,0000	364,0604

a) Confirmer les valeurs des $\hat{t}_i$ présentées dans le tableau.

b) Confirmer les valeurs des $\sqrt{\hat{V}(\hat{t}_i)}$ présentées dans le tableau.

c) Estimer la production totale de la population.

d) Estimer l'écart-type de l'estimateur.

L'écart-type réel est 25 629 si on fixe $m_i \approx 0{,}4M_i$.

11.3 [*Suite*] Estimer les totaux des régions par la moyenne, puis estimer τ par le quotient utilisant pour variable auxiliaire la surface X d'orge cultivée, soit $\hat{\hat{\tau}} = \dfrac{\sum_{i\in\omega}\hat{t}_i}{\sum_{i\in\omega}t_{xi}}\sum_{k=1}^{N}t_{xk}$,

où t_{xi} est le total réel de la strate i et non le total estimé. Les valeurs t_{xi} dans les régions échantillonnées sont les suivantes (dans l'ordre de présentation du numéro précédent):

330,0	492,6	478,2	429,5	439,3	496,1	230,8	327,4
248,0	491,1	461,4	458,0	348,7	475,0	698,9	

Vous supposerez que cette information est connue pour toutes les grappes et que la superficie moyenne par ferme (dans la population) est 24,76069.

a) Estimer la production totale de la population.

b) Estimer l'écart-type de l'estimateur.

11.4 Une population comprend 3 343 fermes regroupées en 200 régions géographiques. Afin d'estimer la production totale τ d'orge dans la saison, on tire un échantillon de 15 régions avec remise et probabilités proportionnelles au nombre de fermes. On estime les totaux des régions par la moyenne.

Voici les données:

M_i	m_i	$\sum y_i$	$\hat{t}_i$
27	10	722	1949,4
18	7	442	1136,571
17	6	236	573,143
18	7	408	1049,143
16	6	738	1968,000
13	5	237	616,200
20	8	672	1680,000
15	6	249	622,500
20	8	370	925,000
20	8	590	1475,000
16	6	521	1389,333
11	4	410	1127,500
18	7	888	2283,429
15	6	468	1170,000
17	6	287	813,1667

a) Confirmer les valeurs des $\hat{t}_i$ présentées dans le tableau.

c) Estimer la production totale de la population.

d) Estimer l'écart-type de l'estimateur.

11.5 Reprendre l'exercice précédent en utilisant dans chaque région une estimation par le quotient avec pour variable auxiliaire la superficie d'orge cultivée.

a) Confirmer les valeurs des $\hat{t}_i$ présentées dans le tableau.

b) Confirmer les valeurs des $\sqrt{\hat{V}(\hat{t}_i)}$ présentées dans le tableau.

c) Estimer la production totale de la population.

d) Estimer l'écart-type de l'estimateur.

Σx	s_x	s_y	s_{xy}	μ_x	$\hat{t}_i$	$\sqrt{\hat{V}(\hat{t}_i)}$
237,5	23,75384	72,32458	1713,1778	18,37407	1508,1440	28,843243
150,4	33,92027	102,45069	3471,9857	25,44444	1345,9840	32,282836
79,4	14,89653	43,91659	653,5267	13,57647	686,0050	12,032500
130,4	32,74150	103,48384	3385,7238	26,56667	1496,2086	30,975774
243,6	36,68602	112,53977	4110,6000	30,78750	1492,3596	41,300512
80,2	18,47574	55,33353	1021,1550	19,07692	732,8678	14,829098
233,6	50,66701	144,96995	7334,0143	34,94500	2010,5342	52,820159
89,1	24,15663	67,89330	1639,8300	23,24667	974,4815	9,123781
123,1	23,19916	70,45718	1634,1036	23,75000	1427,7011	15,122152
187,7	32,92966	103,68737	3413,1036	21,47500	1350,0533	14,286268
160,8	23,24900	74,19546	1714,4600	20,62500	1069,2164	33,104946
138,4	41,27639	121,75796	5025,0000	29,76364	969,8988	8,082583
285,3	26,85882	83,31752	2228,4929	24,40556	1367,3270	24,270726
142,6	20,47854	65,18589	1325,5000	30,76000	1514,2721	49,825589
87,3	25,80107	81,99614	2111,0100	28,88824	1614,4983	70,442318

11.6 Dans le contexte de l'exercice 10.1, on procède à un tirage au deuxième degré (plutôt que de recenser la grappe entière). Dans chaque strate, on tire un *éas* et on estime le total t_i par la moyenne. Voici les données (nombre de femmes) des deux échantillons:

Quartier	M_i	Données
1	37	1 2 1 1 3 2 1 3 1 1 2
2	34	0 1 3 0 2 2 2 3 2 2

a) Estimer le total τ de la population.

b) Estimer l'écart-type de l'estimateur de τ à l'aide de la formule (11.5).

c) Estimer l'écart-type de l'estimateur de τ à l'aide de la formule (11.6).

11.7 Dans le contexte de l'exercice 10.2, on procède à un tirage au deuxième degré (plutôt que de recenser la grappe entière). Dans chaque strate, on tire un *éas* et on estime le total par la moyenne. Voici les données des trois échantillons:

M_i	41	40	34
m_i	18	18	15
$\sum y_i$	23	24	11
s_y^2	0,6830065	0,5882353	0,3523810

a) Estimer t_i par la moyenne pour chacune des trois grappes.

b) Estimer $\hat{\sigma}_i = \sqrt{\hat{V}(\hat{t}_i)}$ pour chacune des deux strates.

c) Estimer le total τ de la population.

d) Estimer l'écart-type de l'estimateur de τ.

11.8 Démontrer les énoncés suivants sous les hypothèses du Théorème 11.1 (ω est l'échantillon d'unités primaires).

a) $E(\hat{\tau} \mid \omega) = \frac{1}{n}\sum_{i=1}^{n} t_i / \psi_i$.

b) $E(\hat{\tau}) = \tau$.

c) $V[E(\hat{\tau} \mid \omega)] = \frac{1}{n}[\sum_{k=1}^{N}(t_k^2 / \psi_k) - \tau^2]$.

d) $V(\hat{\tau} \mid \omega) = \frac{1}{n^2}\sum_{i=1}^{n}(\sigma_i^2 / \psi_i^2)$.

e) $E[V(\hat{\tau} \mid \omega)] = \frac{1}{n}\sum_{k=1}^{N}(\sigma_k^2 / \psi_k)$.

f) $V(\hat{\tau}) = \frac{1}{n}\sum_{k=1}^{N}(\sigma_k^2 / \psi_k) + \frac{1}{n}\sum_{k=1}^{N}(t_k / \psi_k - \tau)^2 \psi_k$.

g) $E(s_{\hat{r}}^2 \mid \omega) = s_r^2 + \frac{1}{n}\sum_{i=1}^{n}(\sigma_i^2 / \psi_i^2)$.

h) $E(s_{\hat{r}}^2) = \sum_{k=1}^{N}(t_k / \psi_k - \tau)^2 \psi_k + \sum_{k=1}^{N}(\sigma_k^2 / \psi_k)$.

i) L'estimateur (11.2) est sans biais pour $V(\hat{\tau})$.

11.9 Rappelons une particularité de la procédure décrite dans la section 11.1: si une même grappe est tirée k fois, on tire indépendamment k échantillons d'unités secondaires dans cette même grappe. Considérons une alternative à cette procédure: on tire un échantillon d'unités secondaires dans chaque grappe *distincte* une seule fois: si i et j sont la même grappe, $\hat{t}_i$ et $\hat{t}_j$ sont égaux. Dans ce cas, certaines propriétés changent. En particulier, la composante $E_1[V_{2|}(\hat{\tau})]$ n'est plus celle établie au numéro 11.8. Pour établir les propriétés de $\hat{\hat{\tau}}$ dans ce cas, il est utile de l'écrire autrement, soit $\hat{\hat{\tau}} = \frac{1}{n}\sum_{k=1}^{N} n_k(\hat{t}_k / \psi_k)$, où n_k est le nombre de fois où l'unité primaire k figure dans l'échantillon. Les n_k sont des variables aléatoires et chaque échantillon possible ω correspond à une et une seule suite $[n_1, n_2, \ldots, n_N]$. Montrer que

a) $E(\hat{\hat{\tau}}) = \tau$.

b) $V(\hat{\hat{\tau}} \mid \omega) = \frac{1}{n^2}\sum_{k=1}^{N} n_k^2(\sigma_k^2 / \psi_k^2)$.

c) $E[V(\hat{\tau} \mid \omega)] = \frac{1}{n}\sum_{k=1}^{N}(\sigma_k^2 / \psi_k) + \frac{n-1}{n}\sum_{k=1}^{N}\sigma_k^2$.

d) $V[E(\hat{\tau} \mid \omega)] = \frac{W(\boldsymbol{t})}{n}$, où $W(\boldsymbol{t}) = \sum_{k=1}^{N}[t_k / \psi_k - \tau]^2 \psi_k$ et $\boldsymbol{t}$ = [t_1; t_2; …; t_N].

e) $V(\hat{\tau}) = \frac{1}{n}\sum_{k=1}^{N}(\sigma_k^2 / \psi_k) + \frac{n-1}{n}\sum_{k=1}^{N}\sigma_k^2 + \frac{1}{n}W(\boldsymbol{t})$.

[Le vecteur [n_1; n_2;…; n_N] est de loi multinomiale (voir le Théorème 1.16).]

11.10 La procédure alternative proposée au numéro 11.9 peut être modélisée d'une autre manière: on suppose que le tirage des unités secondaires se fait *avant* le tirage des unités primaires: on tire un échantillon dans toutes les unités primaires de la population, ce qui permet de calculer les estimateurs $\hat{t}_1, \hat{t}_2, ..., \hat{t}_N$. On tire ensuite un échantillon de n grappes avec remise. On s'assure ainsi qu'un seul échantillon soit tiré dans une grappe donnée. Soit $W(\boldsymbol{t}) = \sum_{i=1}^{N}(t_i / \psi_i - \tau)^2 \psi_i$, où $\boldsymbol{t}$ = [t_1; t_2; … ; t_N]. Montrer que

a) $E_{1|2}(\hat{\tau}) = \sum_{k=1}^{N}\hat{t}_k$;

b) $E(\hat{\tau}) = \tau$;

c) $E_2[\sum_{k=1}^{N}(\hat{t}_k / \psi_k)^2] = \sum_{k=1}^{N}(t_k^2 + \sigma_k^2) / \psi_k^2$;

d) $E_{1|2}(s_{\hat{z}}^2) = W(\hat{\boldsymbol{t}})$;

e) $E[W(\hat{\boldsymbol{t}})] = E(s_{\hat{r}}^2) = W(\boldsymbol{t}) + \sum_{k=1}^{N}\frac{\sigma_k^2(1-\psi_k)}{\psi_k}$;

f) $V_{1|2}(\hat{\tau}) = \frac{W(\hat{\boldsymbol{t}})}{n}$, où $\hat{\boldsymbol{t}} = [\hat{t}_1; \hat{t}_2; ...; \hat{t}_N]$;

g) $E_2[V_{1|2}(\hat{\tau})] = \frac{W(\boldsymbol{t})}{n} + \frac{1}{n}\sum_{k=1}^{N}\frac{\sigma_k^2(1-\psi_k)}{\psi_k}$;

h) $V_2[E_{1|2}(\hat{\tau})] = \sum_{k=1}^{N}\sigma_k^2$;

i) $V(\hat{\tau}) = \frac{1}{n}W(\boldsymbol{t}) + \frac{1}{n}\sum_{k=1}^{N}\frac{\sigma_k^2}{\psi_k}(1-\psi_k) + \sum_{k=1}^{N}\sigma_k^2$;

j) $\frac{s_{\hat{r}}^2}{n}$ est un estimateur sans biais de $E_2[V_{1|2}(\hat{\tau})]$;

k) $\hat{V}(\hat{\tau}) = \frac{s_{\hat{r}}^2}{n} + \frac{1}{n}\sum_{i=1}^{n}(\hat{\sigma}_i^2 / \psi_i)$ est un estimateur sans biais de $V(\hat{\tau})$.

11.11 Montrer que l'approche décrite à la section 11.1 (chaque unité primaire est échantillonnée autant de fois qu'elle est sélectionnée) est plus efficace que celle décrite à l'exercice 11.9.

Annexe A

Jeux de données

Nous décrivons dans cette annexe les données utilisées dans ce manuel: nous définissons les variables et présentons une ébauche du tableau de données afin d'en illustrer le format. Les données elles-mêmes sont disponibles en formats texte et Excel à l'adresse Internet suivante:

www.lozedion.com/introduction-aux-sondages-donnees

TABLEAU A01

Une population de 3 343 fermes dans un territoire réparti en 200 régions géographiques (identifiées par un numéro de 1 à 200.)

Variables

x : Superficie consacrée à la culture de l'orge, en hectares

w : Superficie consacrée à la culture, d'orge ou autre, en hectares

y : Production d'orge, en tonnes (donc nulles lorsque $X = 0$)

Format du fichier

x	*w*	*y*	*Région*
0	280,5	0	1
30	86,5	100	1
57	136	182	1
0	127,9	0	1
0	161,3	0	1
62	62	172	1
28,2	134,6	82	1
26,6	176,6	79	1
0	114	0	1
8,9	19,7	17	2
49,6	227,2	162	2
0	269,4	0	2

Sommes des variables

x	*w*	*y*
82775	619605	248843

Variances-covariances

	x	*w*	*y*
x	1006,690	96,496	3023,020

w	96,496	13241,200	291,465
y	3023,020	291,465	9132,070

TABLEAU A01.1

Un sous-ensemble des fermes du tableau A01, comprenant uniquement les 1579 fermes qui cultivent de l'orge.

Sommes des variables

x	*w*	*y*
82775	295275	248843

Variances-covariances

	x	w	y
x	681,025	117,448	2040,25
w	117,448	13241,300	355,99
y	2040,250	355,990	6226,89

TABLEAU A01.2

Un échantillon de taille 50 tiré de la population A01

Sommes des variables

x	w	y
1366,100	8721,112	4112

Variances-covariances

	x	w	y
x	1062,7842	-698,5616	3216,070
w	-698,5616	12313,6301	-2011,410
y	3216,0701	-2011,4098	9809,247

TABLEAU A01.3

Un échantillon de taille 50 tiré de la population A01.1

Sommes des variables

x	w	y
2483,9	8127,7	7534,0

Variances-covariances

	x	w	y
x	472,849	148,449	1426,180
w	148,449	12291,800	203,303
y	1426,180	203,303	4370,590

TABLEAU A02

Population: Un ensemble de 1183 subdivisions de recensement au Canada

Source: Statistique Canada, Information disponible dans

http://www12,statcan,gc,ca/census-recensement/2011/dp-pd/hlt-fst/pd-pl/Table-Tableau,cfm?LANG=Fra&T=307&S=11&O=A&RPP=100

Variables :

Pop11	Nombre d'habitants en 2011
Pop06	Nombre d'habitants en 2006
Loge11	Nombre de logements privés en 2011
Superficie11	Superficie des terres en kilomètres carrés, 2011
Densité11	Densité de la population au kilomètre carré, 2011

Format du fichier

Pop11	Pop06	Loge11	Superficie11	Densité11
12291	12560	5924	168,1115	73,1122
490	531	229	37,2906	13,14
1055	1109	470	33,4965	31,4958
3456	3409	1566	87,4859	39,5035
7703	7914	3506	419,5285	18,3611
2453	2586	1173	302,6934	8,1039
15163	14834	6897	1121,074	13,5254
743	764	412	160,2229	4,6373
178	248	97	39,2801	4,5316
1137	1230	572	144,371	7,8755
764	812	508	61,1771	12,4883

Sommes des variables

Pop11	Pop06	Loge11	Superficie11	Densité11
7902244	7546013	3684790	674330	168033

Variances-covariances

	Pop11	Pop06	Loge11	Superficie11	Densité11
Pop11	2884196772	2809255293	1399670307	-1204341	9046414
Pop06	2809255293	2737661775	1364548939	-1062146	8779416
Loge11	1399670307	1364548939	681210322	-449056	4262216
Superficie11	-1204341	-1062146	-449056	78338770	-74348
Densité11	9046414	8779416	4262216	-74348	296856

TABLEAU A02.1

Échantillon de taille 45 tiré de la population A02.

Sommes des variables

pop11	pop06	loge11	superficie11	densité11
191662	174213	85449	7508	7906

Variances-covariances

	pop11	pop06	loge11	superficie11	densité11
pop11	58179418	49149799	22200704	-138728	1144816
pop06	49149799	41997158	18785924	-116847	1031502
loge11	22200704	18785924	8575290	-47036	428668
superficie11	-138728	-116847	-47036	173679	-25623
densité11	1144816	1031502	428668	-25623	201702

TABLEAU A03

Population : L'ensemble des 8 654 factures émises par un restaurant sur une période de 364 jours

Variables :

Semaine	Jour de la semaine (1 = lundi; …; 7 = dimanche)
Convives	Le nombre de convives compris dans la facture
Vin	La partie du montant de la facture attribuable au vin (ou autre boisson alcoolique)
Total	Le montant total de la facture
Journée	La journée (désignée chronologiquement de 1 à 364)

Format du fichier

Semaine	Convives	Vin	Total	Journée
1	2	9,2	75,9	1
1	2	9,2	54,05	1
1	3	9,2	65,55	1
1	4	9,2	80,5	1
1	4	37,95	139,15	1
1	2	13,8	46	1
1	2	9,2	46	1
1	2	0	55,2	1
1	2	9,2	69	1
1	4	9,2	88,55	1
1	3	9,2	103,5	1
1	4	9,2	85,1	1
1	3	9,2	67,85	1
1	3	9,2	77,05	1
1	5	9,2	135,7	1

Sommes des variables

Convives	Vin	Total
27728	128887	763319

Variances-covariances

	Convives	Vin	Total
Convives	1,75519	8,06288	48,068
Vin	8,06288	136,87800	319,921
Total	48,06800	319,92100	1573,280

TABLEAU A03.1

Données du tableau A03, montants cumulés par jour.

Sommes des variables

Convives	Vin	Total
27728	128887	763319

Variances-covariances

	Convives	Vin	Total
Convives	42,8836	181,166	1145,63
Vin	181,1660	2945,370	7065,58
Total	1145,6300	7065,580	36993,00

TABLEAU A03.2

Échantillon de 50 factures tirées de la population A03

Sommes des variables

convives	vin	total
151,00	624,45	4046,85

Variances-covariances

	convives	vin	total
convives	1,57102	6,69206	40,5462
vin	6,69206	114,92300	253,8570
total	40,54620	253,85700	1234,4500

TABLEAU A04

Population : 4 932 subdivisions de recensement, Canada.

Source: Statistique Canada, 2012, Chiffres de population et des logements, Canada et subdivisions de recensement (municipalités), recensements de 2011 et 2006 (tableau), Chiffres de population et des logements - Faits saillants en tableaux, Recensement de 2011,

Produit no 98-310-XWF2011002 au catalogue de Statistique Canada, Ottawa, Diffusé le 8 février 2012

Variables :

Pop11	Nombre d'habitants en 2011
Pop06	Nombre d'habitants en 2006
Loge	Nombre de logements privés en 2011
Résid	Nombre de en 2011 (habités par leur propriétaires?)
Superf	Superficie du territoire, en km2, 2011
Densité	Densité de la population par km2, 2011
Prov	Province
Strate	Stratification en quatre grandes régions géographiques (Est, Centre, Ouest, Nord)

Format du fichier

Pop11	Pop06	Loge	Résid	Superf	Densité	Prov	Strate
62	65	83	28	894,14	0,0693	TN	1
905	931	497	364	82,06	11,029	IPÉ	1
518	408	161	155	3,7062	139,77	NÉ	1
738	758	430	365	24,417	30,225	NB	1
12291	12560	5924	5173	168,11	73,112	Qc	2
725	606	251	183	2,5418	285,23	Qc	2
260	252	100	82	20,578	12,635	Ont	2
1535	1629	799	618	1161,6	1,3214	Man	3
40	57	14	7	0,4222	94,742	Sask	3
7214	6709	3040	2518	13160	0,5482	Alb	3

Sommes des variables

Pop11	Pop06	Loge	Résid	Superf	Densité
33460145	31601665	14562850	13314956	7940738	714975

Variances-covariances

	Pop11	Pop06	Loge	Résid	Superf	Densité
Pop11	3287914835	3110920088	1411277324	1333036007	-4210115	7787659
Pop06	3110920088	2947696669	1338970322	1264492674	-3987151	7354516
Loge	1411277324	1338970322	612024206	577519222	-1232589	3348624
Résid	1333036007	1264492674	577519222	545144222	-1723754	3170691
Superf	-4210115	-3987151	-1232589	-1723754	447310675	-226664
Densité	7787659	7354516	3348624	3170691	-226664	157321

TABLEAU A05

Population : 1160 municipalités québécoises en 2011

Source : Statistique Canada, Recensement 2011, adapté par l'Institut de la statistique du Québec,; Statistique Canada, Recensement de la population de 2011

Variables

E0 Nombre de ménages sans enfants

E1 à E3 : Nombre de ménages avec 1, 2, et 3 enfants ou plus,

F Nombre d'unilingues francophones

A Nombre d'unilingues anglophones

D Nombre d'unilingues de langue autre que français ou anglais

FA Nombre de bilingues français et anglais

FD Nombre de bilingues français et autre

AD Nombre de bilingues anglais et autre

FAD Nombre de trilingues français, anglais et autre

Fam Nombre total de familles

Hab Nombre total de personnes

Format du fichier

E0	E1	E2	E3	F	A	D	FA	FD	AD	FAD	Fam	Hab
150	70	45	25	920	5	0	0	0	0	0	290	925
115	50	35	15	675	0	0	0	0	0	0	215	675
975	540	500	170	6840	170	55	35	5	0	0	2185	7105
65	35	40	5	270	185	0	15	0	0	0	145	470
35	30	15	5	260	5	5	0	0	0	0	85	270
15	30	45	50	10	20	585	0	0	0	0	140	615
45	30	25	10	140	185	0	10	0	0	0	110	335
415	160	160	70	2570	30	15	5	5	0	0	805	2625
510	330	215	55	3935	30	45	10	0	0	0	1110	4020
215	75	65	20	1145	75	20	10	0	5	5	375	1260
45	20	15	10	85	170	10	5	0	0	0	90	270
20	15	10	5	180	0	0	0	0	0	0	50	180

Sommes des variables

E0	E3	F	A	fam	hab
930100	189715	6101455	596360	2202210	7810400

Variances-covariances

	E0	E3	F	A	fam	hab
E0	29007111	6585624	167980111	30291664	74125729	286339098
E3	6585624	1561875	37279767	7467352	17143714	66507491
F	167980111	37279767	994580464	163341731	425221915	1632769884
A	30291664	7467352	163341731	39622218	80265446	315498400
fam	74125729	17143714	425221915	80265446	190950223	739032485
hab	286339098	66507491	1632769884	315498400	739032485	2866737065

TABLEAU A06

Un ensemble de 222 maisons vendues dans la région de Montréal

Le tableau suivant présente des données sur un ensemble de 222 maisons vendues par l'entremise d'une certaine agence.

Variables:

Demandé	Le prix demandé dans l'annonce de mise en marché
Cac	Le nombre de chambres à coucher
Bains	Le nombre de salles de bains
Prix	Le prix auquel la maison a été vendue
Strate	Les valeurs 1 à 4 correspondent à des quartiers, Une 5e strate réunit les maisons n'appartenant pas à la région habituellement couverte par l'agence
Espéré	Le prix espéré par le vendeur, une information obtenue suite à une entrevue personnelle avec celui-ci

Format du fichier

Quartier	Demandé	Cac	Bains	Prix	Espéré	Strate
Ahuntsic-Cartierville	626	3	2	501	506	1
Côte-des-Neiges/Notre-Dame-de-Grâce	1354	4	2	1115	1294	2
Mont-Royal	4655	5	6	4421	4629	3
Outremont	1116	3	1	1023	1055	5
Rosemont	455	2	1	385	413	5
St-Laurent	598	4	2	535	552	5
Villeray/Saint-Michel/Park-Extension	1215	4	3	1052	1191	4

Sommes des variables

Demandé	Cac	Bains	Prix	Espéré
159662	650	373	143599	150931

Variances-covariances

	Demandé	Cac	Bains	Prix	Espéré
Demandé	377822	476,127	400,421	348374	366234
Cac	476,127	1,596590	0,714402	434,120	457,868
Bains	400,421	0,714402	0,725307	368,696	388,085
Prix	348374	434,120	368,696	322845	338696
Espéré	366234	457,868	388,085	338696	356182

TABLEAU A07

Population: 2500 comptes de dépenses soumises par 257 représentants (identifiés ici par un numéro de 1 à 257) d'une compagnie,

Variables:

Prov	La province d'appartenant du représentant
Rep	L'identité du représentant
x	Montant total du remboursement demandé
y	Montant total du remboursement autorisé, après vérification.

Format du fichier

Prov	Rep	x	y
Ontario	167	7,8	7,8
Ontario	168	61,99	61,99
Saskatchewan	130	134,4	134,4
Saskatchewan	202	618,2	499,71
Alberta	32	1640,22	1640,22
Alberta	184	51,58	51,58
Colombie-Britannique	214	7,12	7,12
Colombie-Britannique	152	433,89	433,89
Nouveau-Brunswick	173	145,7	136,27
Terre-Neuve	88	87,67	87,67
Québec	118	46,68	46,68
Nouvelle-Écosse	163	47,28	47,28
Nouvelle-Écosse	224	27,1	22,36
Manitoba	23	37,75	37,75
Manitoba	117	2,91	2,91

Sommes des variables

xpop	ypop
1210366	1181289

Variances-covariances

	xpop	ypop
xpop	2216759	2127357
ypop	2127357	2063489

TABLEAU A07.1

Données du tableau A07, cumulées par représentant,

Sommes des variables

xpop	ypop
1210366	1181289

Variances-covariances

	xpop	ypop
xpop	21007586	20177373
ypop	20177373	19599728

TABLEAU A08

Population: 182 pays

Source: United Nations, Department of Economic and Social Affairs, Population Division (2011), *World Population Prospects: The 2010 Revision, Disponible sur* http://www,un,org/esa/population/unpop,htm

Demographic Yearbook, Disponible sur le site

http://unstats,un,org/unsd/demographic/products/dyb/default,htm

Secretariat of the Pacific Community (SPC) Statistics and Demography Programme, Disponible sur le site http://www,spc,int/sdp,

Variables

Pays	Nom du pays
EV0F	Espérance de vie des femmes à la naissance
EV0H	Espérance de vie des hommes à la naissance
EV60F	Espérance de vie des femmes à 60 ans
EV60H	Espérance de vie des hommes à 60 ans
MortMat	Taux de mortalité maternelle
MortInf	Taux de mortalité infantile
Mort5	Taux de mortalité à 5 ans

Format du fichier

Pays	EV0F	EV0H	EV0F	EV0H	MortMat	MortInf	Mort5
Afghanistan	49	49	15	14	460	125	184
Afrique du Sud	54	53	18	14	300	46	64
Albanie	80	74	24	19	27	17	19
Algérie	75	72	20	18	97	21	27
Allemagne	83	78	25	22	7	3	4
Angola	53	50	16	15	450	96	156
Arabie Saoudite	76	73	20	19	24	16	19
Argentine	80	72	24	19	77	12	14
Arménie	77	71	22	18	30	24	27
Australie	84	80	26	23	7	4	5
Autriche	84	78	26	22	4	4	5
Azerbaïdjan	74	68	20	16	43	38	43
Bahamas	79	73	23	18	47	14	18

Sommes des variables

EV0F	EV0H	EV60F	EV60H	MortMat	MortInf	Mort5
13021	12208	3733	3230	30620	5720	8157

Variances-covariances

	EV0F	EV0H	EV60F	EV60H	MortMat	MortInf	Mort5
EV0F	105,90	90,56	30,480	20,960	-1974	-287,0	-456,0
EV0H	90,56	81,30	26,340	19,730	-1697	-249,0	-396,0
EV60F	30,48	26,34	11,100	7,869	-538	-82,2	-128,0
EV60H	20,96	19,73	7,869	7,009	-349	-55,0	-84,8
MortMat	-1974,00	-1697,00	-538	-349	50198	5908,0	9537,0
MortInf	-287,00	-249,00	-82,200	-55	5908	885,3	1401,0
Mort5	-456,00	-396,00	-128	-84,800	9537	1401,0	2263,0

TABLEAU A08.1

Échantillon de taille 30 tiré de la population A08,

Sommes des variables

EV0F	EV0H	EV60F	EV60H	MortMat	MortInf	Mort5
2234	2084	650	558	3689	770	1045

Variances-covariances

	EV0F	EV0H	EV60F	EV60H	MortMat	MortInf	Mort5
EV0F	90,12	78,46	26,400	20,740	-1597	-228,0	-349,0
EV0H	78,46	72,95	24,260	20,570	-1363	-196,0	-300,0
EV60F	26,40	24,26	9,747	8,103	-424	-68,0	-99,8
EV60H	20,74	20,57	8,103	8,248	-288	-49,8	-72,3
MortMat	-1597,00	-1363,00	-424	-288	37973	4470,0	7194,0
MortInf	-228,00	-196,00	-68	-49,800	4470	644,5	982,0
Mort5	-349,00	-300,00	-99,800	-72,300	7194	982,0	1525,0

TABLEAU A09

Population: 708 subdivisions de recensement (municipalités) avec une population de 5000 ou plus.

Source: Statistique Canada, 2012; Faits saillants en tableaux Familles et ménages, Recensement de 2011.

Statistique Canada catalogue no 98-312-XWF2011002,Ottawa, Diffusé le 19 septembre, 2012.

Variables

Tous	Nombre de couples
Mar0	Nombre de couples mariées sans enfants
Mar1	Nombre de couples mariées avec enfants
NMar0	Nombre de couples en union libre sans enfants
NMar1	Nombre de couples en union libre avec enfants

Format du fichier

Tous	Mar0	Mar1	Nmar0	Nmar1
31810	16095	12780	1255	1675
1830	355	665	455	355
2830	1365	1055	200	210
7685	1820	2855	1665	1345
1500	545	620	165	170
1885	760	855	105	165
2150	725	985	185	260

Sommes des variables

Tous	Mar0	Mar1	Nmar0	Nmar1
6828840	3070975	2446500	572175	739135

Variances-covariances

	Tous	Mar0	Mar1	Nmar0	Nmar1
Tous	978475378	493542312	321936575	57460041	105526199
Mar0	493542312	255956443	162085701	25909423	49585500
Mar1	321936575	162085701	106848551	18493978	34504812
Nmar0	57460041	25909423	18493978	5173560	7882493
Nmar1	105526199	49585500	34504812	7882493	13552528

TABLEAU A09.1

UN ÉCHANTILLON DE TAILLE 80 TIRÉ DE A09

Sommes des variables

tous	mar0	mar1	nmar0	nmar1
631700	223975	232790	81740	93200

Variances-covariances

	tous	mar0	mar1	nmar0	nmar1
tous	295769388	91372291	101773193	45264712	57345556
mar0	91372291	32217982	32486736	11283428	15382018
mar1	101773193	32486736	35599378	14593839	19088843
nmar0	45264712	11283428	14593839	9050150	10333956
nmar1	57345556	15382018	19088843	10333956	12536985

TABLEAU A10

Population : Les 6623 logements d'une ville, regroupés en 180 quartiers identifiés par un numéro allant de 1 à 180,

Variables

Femmes Le nombre de femmes dans le logement

Hommes Le nombre d'hommes dans le logement

Quartier Le quartier

Format du fichier

Femmes	Hommes	Quartier
1	1	1
1	0	2
0	1	6
3	2	7
0	1	111
1	1	154
1	0	156
1	0	161

0	1	164

TABLEAU A10.1

Données du tableau A10 cumulées par quartier; n = 120 quartiers.

Sommes des variables

Femmes	Hommes	Personnes
8262	8314	16576

Variances-covariances

	Femmes	Hommes	Personnes
Femmes	21,52067	10,86816	32,38883
Hommes	10,86816	18,00881	28,87697
Personnes	32,38883	28,87697	61,26580

Travaux cités

Anderson, R. e. (1973). Dollar-unit Sampling. *Canadian Chartered Accountant*(4), 30-39.

Arkin, H. (1963). *Handbook of Sampling for Auditing.* New York: McGraw-Hill.

Cochran, W. G. (1977). *Samplign Techniques, 3rd Ed.* John Wiley and Sons.

Hajek, J. (1960). Limiting distributions in simple random sampling from a finite population. *Publ. Math. Inst. Hung. Acad. Sci.*, pp. 361-374.

Horvitz, D. e. (1952). A generalization of sampling without replacement from a fiite universe. *Journal of the American Statistical Association, 47*, 663-685.

Joubir, S. (2010). *Estimation de la borne supérieure par des approches statistiques et par la méthode de Stringer.* Montréal, Québec, Canada: Mémoire de maîtrise, Département de mathématiques, Université du Québec à Montréal.

Lahiri, D. (1951). A method for sample selection providing unbiased unbiased ratio estimates. *Bull. Inst. Stat.Inst., 33*(2), 133-140.

McCray, J. H. (1984). quasi-Bayesian Audit Risk Model for Dollar Unit Sampling. *American Accounting Association, 59*(1), 35-51.

Menzefrickem, U. e. (1984). A Simulation Studyof the Performance of Parametric Dollar Unit Sampling Statistical Procedures. *Journal of Accounting Research, 22*(2), 588-604.

Midzuno, H. (1951). On the sampling system with probability proportionate to sum of sizes. *Ann. Inst. Stat. Math., 2*, 99-108.

Neter, J. L. (1978). Dollar Unit Sampling: Multinomial bounds for Total Overstatement Errors. *The Accounting Review*, 77-93.

Plante, R. J., & Neter, J. e. (1985). Comparative performance of multinomial, cell, and Stringer bounds. *Auditing, 5*, 40-56.

Raj, D. e. (1958). Some Remarks on Sampling with Replacement. *The Annals of Mathematical Statistics 29 (Number 2)*, 550--557.

Ramage, J. G. (1979). An Empirical Study of Error Characteristics in Audit Populations. *Journal of Accounting Research, 17*, 72-102.

Rao, J. N. (1973). On double sampling for stratification and analytical surveys. *Biometrika, 60*, 125-133.

Reneau, J. H. (1978). CAV Bounds in Dollar Unit Sampling: Some Simulation Results. *American Accounting Association*, 669-680.

Särndal, C. E. (1972). Sample survey theory vs general statistical theory: Estimation of the population mean. *Rev. Int. Stat. Inst,, 40*, 1-12.

Teitlebaum, A. D. (1973). Dollar-unit Sampling in Auditing. *Annual Meeting of American Statistical Association.* New York.

Serge Alalouf détient un doctorat en mathématiques (statistique) de l'Université McGill. Il est professeur à l'UQAM depuis 1969 et consultant en statistique auprès d'organismes gouvernementaux, d'entreprises privées et de cabinets de comptables.

Une lente évolution amorcée dans la quatrième décennie du siècle dernier — la tendance à remplacer l'énumération complète d'une population par un échantillon — s'est accélérée au cours des vingt dernières années. Une nouvelle exigence acquiert aujourd'hui une importance croissante: l'échantillonnage dit «de raison» cède la place à l'échantillonnage dit «scientifique». Dans plusieurs domaines — en particulier, dans un contexte de litige — la nécessité de justifier une conclusion de façon objective commande un échantillonnage rigoureusement probabiliste. Les gestionnaires, les vérificateurs comptables, les actuaires et les analystes de divers domaines sont souvent confrontés à cette nécessité.

C'est à eux, ainsi qu'aux futurs spécialistes, que ce livre s'adresse. Car si certains sondages complexes exigent des connaissances très poussées, le type d'échantillonnage que ces professionnels sont appelés à faire exige essentiellement les connaissances de base de la théorie classique. Celle-ci repose sur un cadre clair et simple: une population bien définie; et un mode de tirage qui permet de calculer certaines probabilités essentielles. Or ces conditions se trouvent réunies dans le contexte d'un échantillonnage interne: une population accessible (généralement sous la forme d'un fichier électronique) et bien identifiée; une base de sondage souvent riche et féconde; un mécanisme de sélection aisément applicable. Les difficultés pour lesquelles les techniques les plus avancées ont été créées — qui nécessitent un spécialiste — ne se présentent pas souvent. Le paradigme de base est donc applicable. Et c'est à cela que ce livre s'en tient.

Si cette limitation semble radicale, elle se révèle peu contraignante pour le public visé — les professionnels cités plus haut (gestionnaires, comptables, actuaires, et analystes divers ainsi que les spécialistes débutants). En revanche, chaque sujet est traité en grand détail, suivi de nombreux exemples. Suivi également d'un développement théorique que le lecteur peut décider de lire assidûment, en diagonale ou pas du tout.

Illustration de la couverture : l'amas globulaire M22, vieux de plus de 12 milliards d'années, se situe dans la constellation du Sagitaire. Autour de cet amas on dénombre 100 000 fois plus d'étoiles que dans l'environnement du Soleil.
Gracieuseté de la NASA, Kailash Sahu, Stefano Casertano, Mario Livio, Ron Gilliland (*Institut des sciences du télescope spatial*), Nino Panagia (*Agence spatiale européenne/Institut des sciences du télescope spatial*), Michael Albrow and Mike Potter (*Institut des sciences du télescope spatial*) et de Nigel A.Sharp, REU program/AURA/NOAO/NSF.

Loze-Dion éditeur inc.