UNIVERSITÉ DU QUÉBEC À MONTRÉAL

L'ESTIMATION À L'AIDE D'UNE VARIABLE AUXILIAIRE

MÉMOIRE
PRÉSENTÉ
COMME EXIGENCE PARTIELLE
DE LA MAÎTRISE EN MATHÉMATIQUES

PAR

MALIKA ELBAZ

MARS 2010

UNIVERSITÉ DU QUÉBEC À MONTRÉAL

Service des bibliothèques

Avertissement

La diffusion de ce mémoire se fait dans le respect des droits de son auteur, qui a signé le formulaire *Autorisation de reproduire et de diffuser un travail de recherche de cycles supérieurs* (SDU-522 -Rév.01-2006). Cette autorisation stipule que «conformément à l'article **11** du Règlement no 8 des études de cycles supérieurs, [l'auteur] concède à l'Université du Québec à Montréal une licence non exclusive d'utilisation et de publication de la totalité ou d'une partie importante de [son] travail de recherche pour des fins pédagogiques et non commerciales. Plus précisément, [l'auteur] autorise l'Université du Québec à Montréal à reproduire, diffuser, prêter, distribuer ou vendre des copies de [son] travail de recherche à des fins non commerciales sur quelque support que ce soit, y compris l'Internet. Cette licence et cette autorisation n'entraînent pas une renonciation de [la] part [de l'auteur] à [ses] droits moraux ni à [ses] droits de propriété intellectuelle. Sauf entente contraire, [l'auteur] conserve la liberté de diffuser et de commercialiser ou non ce travail dont [il] possède un exemplaire.»

REMERCIEMENTS

Je dédie ce travail à ma mère, Mme Elbaz Zineb qui m'a toujours indiqué la bonne voie. Grâce à son soutien, j'ai appris que la volonté et la patience font des miracles. Aussi, à mes enfants, Oualid et Haytam, ma raison de vivre.

Mes sincères remerciements vont particulièrement à M Serge Alalouf, mon directeur de projet, sans qui, ce mémoire n'aurait pu voir le jour. Ses précieux conseils m'ont été d'un grand secours. Je lui suis infiniment reconnaissante de la bienveillance dont il a fait preuve à mon égard, de ses encouragements, de sa patience et de sa confiance. Merci monsieur.

TABLE DES MATIÈRES

LISTE DES TABLEAUX

RÉSUMÉ

Dans le domaine des sondages, il y a plusieurs méthodes d'estimer une moyenne ou un total d'une population. Une des techniques les plus pratiques est l'utilisation d'une variable auxiliaire ayant une corrélation avec la variable d'intérêt et dont les données sur toute la population sont connues.

Dans ce mémoire, on va s'intéresser particulièrement à l'estimation d'une moyenne par le quotient. Dans ce cadre, on va traiter largement cette approche, en détaillant premièrement tout ce qui concerne l'estimateur par le quotient afin de répondre à toute question sur le sujet.

Par la suite, on va étudier un cas particulier de l'estimation par le quotient lorsque la variable étudiée est dichotomique, en examinant les estimateurs possibles selon le mode de tirage, soit le tirage avec probabilités égales ou probabilités inégales avec ou sans remise.

Dans le dernier chapitre, on présentera le cas où la variable auxiliaire est aussi dichotomique. Quatre estimateurs seront proposés pour estimer la moyenne de la population. On va analyser leurs propriétés dans le but de pouvoir les comparer.

Afin que le traitement théorique soit valable, on va le concrétiser par des simulations avec des populations réelles et d'autres générées selon des paramètres donnés.

Mots clés : estimateur , quotient, variable auxiliaire, variable dichotomique.

.....

INTRODUCTION

Dans la pratique des sondages, lorsqu'on cherche à estimer dans une population finie un paramètre tel qu'un total, une moyenne ou un pourcentage, la théorie a montré qu'il est préférable, lorsqu'on connaît sur l'ensemble de la population les valeurs de la variable auxiliaire qui est bien corrélée avec la variable d'intérêt, d'utiliser cette information pour améliorer la qualité des résultats, mesurée par l'erreur quadratique moyenne de l'estimateur dans la population finie. On peut tenir compte de l'information, soit au niveau du tirage grâce à une technique de stratification, soit au niveau de l'estimation, le cas où on utilise essentiellement des estimateurs par la régression ou par le quotient. Ce dernier est le sujet de ce mémoire.

Définition du problème

L'estimation d'un quotient R (ou d'une moyenne ou d'un total par le quotient) a été étudiée en profondeur depuis le milieu du siècle dernier et continue à ce jour de faire l'objet de recherches dans plusieurs directions. Plusieurs raisons expliquent pourquoi ce sujet ne s'épuise pas. L'une d'elles est le fait que l'estimateur est biaisé et qu'il est généralement difficile de se faire une idée de l'ampleur du biais.

Étant donné la présence d'une variable aléatoire au dénominateur de l'estimateur, toutes les propriétés importantes ne sont établies que par des méthodes asymptotiques et donc approximatives, essentiellement par la technique de linéarisation de Taylor proposée par Goodman et Hartley (1958). À la question cruciale de savoir si une moyenne est mieux estimée par une moyenne échantillonnale ou par le quotient, nous n'avons que des réponses incertaines et surtout conditionnelles à certaines caractéristiques de la population, caractéristiques dont on ne sait pas trop si elles sont vérifiées ou non.

Notre but dans le mémoire est de voir si, en fixant notre attention sur certains cas particulier, nous pourrions établir quelques principes qui permettraient de décrire les populations pour lesquelles tel estimateur est supérieur à tel autre.

Le cas particulier qui nous concerne ici est celui où la variable d'intérêt est une variable dichotomique. Un précédent bien connu nous montre que les formules sont plus simples et de nouvelles propriétés surgissent lorsqu'on exploite le caractère dichotomique d'une variable. Le cas le plus évident est celui de l'estimation d'une proportion π. Une proportion est bien sûr une moyenne, et tout ce que l'on sait de l'estimation d'une moyenne s'applique tout aussi bien à l'estimation d'une proportion. Cependant de nouvelles possibilités s'ouvrent lorsqu'on tient compte du fait que, dans un échantillon aléatoire simple, la moyenne échantillonnale est, à un facteur constant près, de loi connue—binomiale ou hypergéométrique—et par conséquent que la variance peut s'exprimer en fonction du paramètre à estimer. Cette variance peut être bornée supérieurement et des intervalles de confiance peuvent être déterminés avec des échantillons petits, donc sans recours au théorème limite central, etc.

Nous nous proposons donc d'examiner ce que deviennent les méthodes actuelles dans le cas où la variable d'intérêt est dichotomique. On restreindra davantage notre attention, en examinant de près le cas encore plus particulier où la variable auxiliaire est, elle aussi, dichotomique.

Nous commencerons donc, au premier chapitre, par un survol des méthodes classiques d'estimation d'un quotient : les différents modes de tirage, les estimateurs, leurs propriétés. Au second chapitre, nous réduisons les formules aux formes plus simples qu'elles sont susceptibles de prendre dans les situations particulières étudiées.

Au troisième chapitre, nous traitons le cas où le numérateur et le dénominateur du quotient sont tous les deux des variables dichotomiques. Enfin, au dernier chapitre, nous procédons à certaines études de simulations afin de comparer l'efficacité de certaines des méthodes proposées.

Estimateur classique dans un échantillon aléatoire simple

Sukhatme (1953) est un des premiers à examiner de près la précision de la formule approximative classique V_C de la variance du quotient échantillonnal $\hat{R} = \overline{y}_s / \overline{x}_s$ dans un échantillon aléatoire simple, celle proposée par Cochran (1977) et utilisée sans changement depuis.

Il propose une approximation plus précise, V_S, qu'il compare à V_C, mais il conclut que V_C est adéquat dans de grands échantillons. L'applicabilité de ce résultat est probablement limitée, cependant, étant donné qu'elle est établie sous l'hypothèse que le couple $(X; Y)$ de variables est de loi normale bivariée. Cette hypothèse est certainement fausse dans la plupart des situations pratiques et encore moins vraie dans le cas qui nous préoccupe. Des Raj (1964) compare V_C à la variance réelle V telle qu'exprimée par Goodman (1958) en termes de certaines variances et non pas par une formule algébrique. Il détermine une borne inférieure pour V-V_C, ce qui lui permet d'énoncer une condition suffisante pour que V_C sous-estime V.

Sa conclusion est que V_C sous-estime V lorsque le coefficient de corrélation entre $1/\overline{x}_s$ et $(\overline{y}_s - R\overline{x}_s)^2$ est positif. L'opacité de cette condition témoigne de la difficulté du problème. En témoigne aussi le recours fréquent à des modèles. Des Raj propose le modèle $y = Rx + e$, où E(e) = 0 et où la moyenne échantillonnale $\overline{e}_s$ est de corrélation nulle avec certaines fonctions de $\overline{x}_s$. Dans ce modèle, Des Raj trouve des bornes inférieure et supérieure pour la différence V-V_C. Ce qu'on peut dire en l'absence d'un modèle n'est pas encore clair.

Certains auteurs montrent que les estimations peuvent se révéler très imprécises, comme le montre, au moyen d'un petit exemple, Koop (1968). D'autres, en revanche, comme Smith (1969), soutiennent que le scénario de Koop—un tout *petit* échantillon (quatre observations), dans une population dont on ne sait strictement rien—ne correspond à aucune situation réaliste.

Évaluation et réduction du biais

Le biais de l'estimateur d'un quotient a longuement préoccupé les chercheurs. David et Sukhatme (1974) ont estimé le biais dans un échantillon aléatoire simple par les moyens usuels de développement en série de Taylor. Ils expriment le biais en fonction de coefficients de variation et du quotient R lui-même, montrant ainsi que le biais est d'ordre $O_p(n^{-1})$ et présentent également une formule du biais relatif, c'est-à-dire, le rapport du biais sur la racine carrée de l'erreur quadratique moyenne. Ces approximations ont permis à Kish, Namboodiri et Pillai (1962) d'estimer le biais relatif (en remplaçant les paramètres par leur estimateur) de plusieurs populations réelles, à partir de divers sondages.

Les résultats, assez rassurants, donnent des rapports généralement assez petits, rarement au-dessus de 3 %. Une approche théorique de Hartley et Ross (1954) permet de définir une borne supérieure pour le rapport du biais sur l'écart-type de l'estimateur, une borne qui s'avère être d'ordre $(n^{-1/2})$.

Certains efforts ont été consacrés à la réduction du biais. La méthode du *Jackknife* de Quenouille (1956), proposée dans le cas d'un quotient par Durbin (1959), est la mieux connue. Dans sa plus simple expression, elle consiste à éliminer tour à tour chaque unité i de l'échantillon, $i = 1, \ldots, n$, calculer le quotient $\hat{R}_{(i)}$ des n-1 données restantes et estimer le quotient par $n\hat{R} - \frac{n-1}{n}\sum_{i=1}^{n}\hat{R}_{(i)}$. Durbin montre que le biais est alors d'ordre n^{-2}.

Estimateurs sans biais

Hartley et Ross (1954) ont réussi à proposer un estimateur qui est totalement sans biais dans un échantillon aléatoire simple. C'est un estimateur qui fait intervenir non seulement les moyennes $\bar{x}_s$ et $\bar{y}_s$ mais aussi la moyenne des quotients y_i/x_i (par opposition au quotient des moyennes). Un autre estimateur sans biais, dû à Mickey (1959), est apparenté à la méthode du *Jackknife*. Elle consiste en effet à retirer de l'échantillon une observation à la fois et

estimer le quotient par une certaine fonction de la moyenne des quotients $\hat{R}_{(i)}$, définis plus haut. La précision de cet estimateur a été étudiée par Rao (1967).

Plans d'échantillonnage permettant des estimateurs sans biais

Une approche entièrement différente permettant d'estimer R sans biais passe par une modification du mode de tirage. Il s'agirait de tirages avec probabilités inégales, c'est-à-dire, telle que $p(s)$, la probabilité que l'échantillon s soit tiré diffère d'un échantillon à l'autre. Parmi ceux-là, notons les modes de tirage proposés par Des Raj (1954), Nanjamma, Murthy et Sethi (1959) et Lahiri (1951).

Cette dernière a ceci de particulier qu'elle maintient le même estimateur, ce qui la distingue des approches qui modifient l'estimateur en fonction du mode de tirage. Elle est, au contraire, conçue de manière à ce que l'estimateur traditionnel $\hat{R} = \bar{y}_s / \bar{x}_s$ soit sans biais. On change donc le mode de tirage sans changer l'estimateur. La méthode consiste à tirer les échantillons de telle sorte que $p(s)$ soit proportionnelle à $\bar{x}_s$. On peut démontrer qu'alors $\hat{R}$ est sans biais.

Une façon simple de réaliser un tel échantillonnage a été proposée par Midzuno (1952). Les développements théoriques de Nanjamma, Murthy et Sethi (1959) montrent comment modifier un mode de tirage donné de telle sorte que l'estimation par le quotient de certains paramètres, dont la moyenne, soit sans biais. Ces travaux ont été suivis de ceux de Pathak (1964), qui a développé une technique (essentiellement une application du théorème de Rao-Blackwell) pour améliorer les estimateurs proposés par Nanjamma, Murthy et Sethi.

Études comparatives

La diversité des méthodes a évidemment provoqué des recherches comparatives. Celle de Rao (1968) compare l'approche de Lahiri à l'approche classique à l'aide d'un modèle de super population selon lequel $y_i = \alpha + \beta x_i + e_i$, avec les suppositions habituelles : non

corrélation des e_i et $E(e_i) = 0$ mais de variance $V(e_i)$ proportionnelle à une puissance g de x_i, $0 \leq g \leq 2$. Des conclusions assez générales sont possibles lorsque $\alpha = 0$ mais elles ne sont pas aisément décrites autrement. On rencontre les mêmes conclusions mitigées dans une autre étude de Rao (1971) qui compare cinq différentes approches, utilisant le même modèle. D'autres études, certaines empiriques, d'autres s'appuyant sur des modèles, ont été publiées par P. S. R. S. Rao (1968, 1969), J. N. K. Rao (1971), Tin (1965) et Hutchison (1971).

Tirages sans remise, avec probabilités inégales

Le tirage de Lahiri est un exemple particulier de tirage avec probabilités inégales. Nous traiterons ce cas particulier, mais en général, l'estimateur par le quotient sera défini comme $\frac{\bar{y}_{HT}}{\bar{x}_{HT}}$, où $\hat{\bar{y}}_{HT}$ et $\hat{\bar{x}}_{HT}$ sont les estimateurs de Horvitz-Thompson, $\hat{\bar{y}}_{HT} = \frac{1}{n}\sum_{i \in s} \frac{y_i}{\pi_i}$, π_i étant la probabilité que l'unité i se trouve dans l'échantillon.

Tirages avec remise

Les tirages sans remise avec probabilités inégales ne sont pas faciles à réaliser. Nous accorderons donc dans ce mémoire une certaine importance aux tirages *avec* remise. Ce qui soulève, d'emblée, la question suivante : si, dans un échantillon aléatoire simple avec remise, certaines unités sont tirées plus d'une fois, vaut-il mieux les compter autant de fois qu'elles se présentent ou est-il préférable de ne les compter qu'une seule fois? La réponse, donnée par Des Raj et Khamis (1958) est nette : il vaut mieux les considérer une seule fois (et donc accepter que l'échantillon soit de taille aléatoire). Des Raj et Khamis proposent des formules approximatives pour la variance et pour le biais de l'estimateur dans les deux cas.

Il est évident qu'en principe, un tirage avec remise est moins efficace qu'un tirage sans remise. Dans les applications les plus courantes, cependant, la population est de taille tellement supérieure à celle de l'échantillon que la différence d'efficacité est tout à fait négligeable. La facilité avec laquelle le tirage avec remise est exécuté et ses propriétés établies nous encouragent à le considérer en détail. À chaque tirage, l'unité k est tirée avec

probabilité p_k, $\sum_{k=1}^{N} p_k = 1$. L'estimateur d'une fonction des moyennes $f(\bar{x}_U; \bar{y}_U)$ est $f(\bar{x}_r; \bar{y}_r)$ où $\bar{x}_r = \frac{1}{n}\sum_{j=1}^{n} \frac{y_j}{p_j}$, où y_j est la valeur obtenue et p_j est la probabilité d'obtenir l'unité obtenue au j^e tirage.

Estimateurs considérés

Nous considérerons trois classes d'estimateur. La première est la classe des estimateurs d'une moyenne dans un échantillon aléatoire simple avec ou sans remise. Nous comparerons l'estimateur par le quotient aux estimateurs par la moyenne, par la différence et par la régression.

La deuxième classe sera celle des estimateurs issus d'un échantillon tiré avec probabilités inégales, avec remise. C'est la variable auxiliaire qui servira de probabilités de sélection. Nous traiterons là aussi des estimateurs par le quotient, par la différence et par la régression et montrerons qu'ils peuvent tous être unifiés à l'aide de la notion d'*estimation décalée* de Stuart (1986). Nous y ajouterons une modification de l'estimateur de Stuart.

Finalement nous examinerons les estimateurs issus du tirage spécialisé de Lahiri.

CHAPITRE I

ESTIMATION PAR LE QUOTIENT

Avant d'aborder le cas particulier que nous traiterons dans ce mémoire, à savoir, l'estimation d'un quotient, ou d'une moyenne ou d'un total par le quotient dans le cas de variables dichotomiques, nous présentons dans ce chapitre un survol des méthodes existantes générales. Elles comprennent des modes de tirage différents ainsi que, pour certains des modes de tirage, plusieurs estimateurs possibles. Dans le prochain chapitre, nous explorerons les simplifications qui pourraient découler de la nature particulière des variables.

L'estimateur d'un quotient n'étant pas, sauf rares exceptions, sans biais, notre attention portera entre autre sur l'estimation du biais, soit par des méthodes de développement en série, soit à l'aide de bornes supérieures. Dans le contexte d'échantillonnage avec probabilités inégales, on va examiner les propriétés de l'estimateur de Horvitz et Thomson (HT) et l'estimateur de la variance de Sen-Yates-Grundy (SYG).

Nous nous pencherons également sur une méthode particulière, proposée par Lahiri, qui consiste à éliminer le biais au moyen d'un mode de sélection particulier et nous évoquerons le schéma de Midzuno, qui est une façon de réaliser l'échantillonnage de Lahiri.

1.1 Les caractéristiques de l'estimateur par le quotient

Chaque estimateur soit il est biaisé ou non. Dans cette section, nous allons calculer l'espérance et la variance de l'estimateur par le quotient afin de déterminer ses propriétés.

1.1.1 Notations

Les notations suivantes seront utilisées tout au long de ce mémoire :

Pour toute la population on note par:

N : la taille de la population U : $\{1,2,.....,N\}$;

x_i : la valeur de la i^e unité correspondant à la variable auxiliaire;

$t_x = \sum_{i=1}^{N} x_i$: le total des valeurs x_i;

$\bar{x}_U = \frac{1}{N}\sum_{i=1}^{N} x_i = \frac{t_x}{N}$: la moyenne des valeurs x_i;

y_i : la valeur de la variable Y correspondant à la i^e unité;

$t_y = \sum_{i=1}^{N} y_i$: le total des valeurs y_i;

$\bar{y}_U = \frac{1}{N}\sum_{i=1}^{N} y_i = \frac{t_y}{N}$: la moyenne des valeurs y_i;

$R = \frac{t_y}{t_x} = \frac{\bar{y}_U}{\bar{x}_U}$: le quotient;

Pour l'échantillon on note par :

n : la taille de l'échantillon;

x_i : la valeur de la i^e unité correspondant à la variable auxiliaire;

$x_s = \sum_{i=1}^{n} x_i$: le total des valeurs x_i;

$\bar{x}_s = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_s}{n}$: la moyenne des valeurs x_i ;

y_i : la valeur de la variable Y correspondant à la i^e unité;

$y_s = \sum_{i=1}^{n} y_i$: le total des valeurs y_i ;

$\bar{y}_s = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{y_s}{n}$: la moyenne des valeurs y_i ;

$r = \frac{y_s}{x_s} = \frac{\bar{y}_s}{\bar{x}_s}$: le quotient;

$1 - \frac{n}{N} = 1 - f$: le facteur de correction dans un tirage aléatoire simple sans remise;

$\bar{q}_s = \frac{1}{n}\sum_{i=1}^{n} q_i = \frac{1}{n}\sum_{i=1}^{n} \frac{y_i}{x_i}$: la moyenne des quotients.

1.1.2 Principe de l'estimateur par le quotient

L'estimation par le quotient est une méthode de redressement de l'échantillon. Son principe est l'utilisation d'une information supplémentaire, une deuxième variable X, appelée *variable auxiliaire*, corrélée à Y, la variable étudiée, pour chaque unité de l'échantillon. Les valeurs de X peuvent et doivent être connues pour la population entière, mais pour certaines méthodes et certains paramètres, une connaissance du total t_x est suffisante.

Il s'agit d'une simple règle de trois : on porte la moyenne de X à sa valeur vraie et on modifie la moyenne $\bar{Y}$ dans la même proportion.

En général, notre objectif est d'estimer la moyenne générale $\bar{y}_U$ ou le total t_y, un pourcentage, tout en se basant sur le fait que l'échantillon nous fournit un estimateur usuel $\bar{y}_s$, la moyenne $\bar{x}_s$ de l'échantillon de la variable auxiliaire et aussi la vraie moyenne $\bar{x}_U$ pour l'ensemble de la population qui est supposée déjà connue.

Donc l'estimateur de la moyenne $\bar{y}_U$ par le quotient est : $\bar{Y}_Q = \frac{\bar{y}_s}{\bar{x}_s}\bar{x}_U$ et celui du total t_y est obtenu en multipliant par *N*, ce qui donne $Y_Q = \frac{y_s}{x_s}t_x = \frac{\bar{y}_s}{\bar{x}_s}t_x$.

Remarque : Bien que le recours à une variable auxiliaire permet souvent d'améliorer grandement la précision d'un estimateur, il n'est pas aisé de démontrer les propriétés des estimateurs qui exploitent l'information qu'elle contient. Ceci tient au fait que ces estimateurs font intervenir des statistiques comme $\bar{x}_s$ ou $\bar{y}_s$ de façon non linéaire. C'est le cas, entre autre, de l'estimateur par le quotient dont la présence de la moyenne $\bar{x}_s$ au dénominateur est cause de sérieuses difficultés.

L'utilité de $\bar{Y}_Q$ vient de la dépendance entre la variable auxiliaire et la variable d'intérêt. Ainsi, si le quotient $\frac{y_i}{x_i}$ est à peu près le même pour toutes les unités de l'échantillon, alors les valeurs $\frac{\bar{y}_s}{\bar{x}_s}$ ne varient pas d'un échantillon à l'autre et par conséquent, on a une grande précision de l'estimation du quotient.

1.1.3 Le biais de l'estimateur par le quotient

L'estimateur par le quotient est biaisé et généralement il n'est pas possible de calculer ce biais avec exactitude, ce qui a incité plusieurs statisticiens à l'approcher à l'aide du développement en série de Taylor.

Néanmoins, ce biais est négligeable dans les échantillons de grande taille.

Par définition, le biais c'est l'espérance de:

$$\bar{Y}_Q - \bar{y}_U = \bar{x}_U \frac{\bar{y}_s}{\bar{x}_s} - \bar{y}_U = \bar{x}_U \frac{\bar{y}_s - R\bar{x}_s}{\bar{x}_s} \quad \text{avec} \quad R = \frac{\bar{y}_U}{\bar{x}_U}$$

Pour simplifier, on désignera pas $\bar{x}$ et $\bar{y}$ (sans l'indice s) les moyennes dans l'échantillon.

$\bar{x}$ est un estimateur sans biais de $\bar{x}_U$ et généralement la statistique $\frac{\bar{x} - \bar{x}_U}{\bar{x}_U}$ est petite, d'ordre $O_p\left(n^{-1/2}\right)$, donc une quantité petite pour n suffisamment grande.

D'après Cochran (1977), on effectue un développement limité:

$$\bar{Y}_Q - \bar{y}_U = \frac{\bar{y} - R\bar{x}}{1 + \frac{\bar{x} - \bar{x}_U}{\bar{x}_U}} = \frac{\bar{y} - R\bar{x}}{1 + \varepsilon} = (\bar{y} - R\bar{x})(1 - \varepsilon + \varepsilon^2 - \varepsilon^3 + \ldots)$$

$$\approx (\bar{y} - R\bar{x})(1 - \varepsilon) \approx (\bar{y} - R\bar{x})\left(1 - \frac{\bar{x} - \bar{x}_U}{\bar{x}_U}\right)$$

L'approximation du biais de cet estimateur :

$$E(\bar{Y}_Q - \bar{y}_U) \approx E\left[(\bar{y} - R\bar{x})\left(1 - \frac{\bar{x} - \bar{x}_U}{\bar{x}_U}\right)\right] = E(\bar{y} - R\bar{x}) - \frac{E\left[(\bar{y} - R\bar{x})(\bar{x} - \bar{x}_U)\right]}{\bar{x}_U}$$

$$= -\frac{E\left[\bar{y}\left(\bar{x}-\bar{x}_U\right)\right]-RE\left[\bar{x}\left(\bar{x}-\bar{x}_U\right)\right]}{\bar{x}_U} \quad \text{car} \quad E(\bar{y}-R\bar{x})=0\,.$$

$$= \frac{RE\left[\bar{x}\left(\bar{x}-\bar{x}_U\right)\right]-E\left[\bar{y}\left(\bar{x}-\bar{x}_U\right)\right]}{\bar{x}_U} = \frac{RV(\bar{x})-Cov(\bar{x},\bar{y})}{\bar{x}_U}$$

$$= \frac{1-f}{n}\frac{RS^2{}_x - S_{xy}}{\bar{x}_U} = \frac{1-f}{n\bar{x}_U}S_x^2\left(R-\frac{S_{xy}}{S_x^2}\right) \text{ où } B=\frac{S_{xy}}{S_x^2} \text{ est la pente de la}$$

droite des moindres carrés entre y et x dans la population.

Ce biais est nul si et seulement si la droite des moindres carrés ajustée sur la population (la droite de régression) au travers les points$(x_i;y_i)$ passe par l'origine. Aussi il est négligeable quand n est grande.

Finalement, l'espérance de l'estimateur par le quotient est approchée par:

$$E\left(\bar{Y}_Q\right)=\bar{y}_U+\frac{1-f}{n\bar{x}_U}S_x^2\left(R-\frac{S_{xy}}{S_x^2}\right).$$

1.1.4 La variance de l'estimateur par le quotient

Puisque l'estimateur par le quotient est biaisé, on va chercher une approximation de l'erreur quadratique moyenne qu'on peut assimiler à la variance si la taille n de l'échantillon est grande.

L'erreur quadratique moyenne est définie par $EQM(\bar{Y}_Q)=E\left(\bar{Y}_Q-\bar{y}_U\right)^2$ et selon le développement limité d'ordre 2 de $(\bar{Y}_Q-\bar{y}_U)$ déjà effectué, une première approximation s'écrit : $EQM(\bar{Y}_Q)\approx E\left(\bar{y}-R\bar{x}\right)^2 \approx E\left[\left(\bar{y}-\bar{y}_U\right)-RE\left(\bar{x}-\bar{x}_U\right)\right]^2$

$$\approx V(\bar{y})+R^2V(\bar{x})-2RCov(\bar{x},\bar{y})$$

Dans le cas où les tirages ont été effectués à probabilités égales sans remise PESR, on a : $V(\bar{Y}_Q)=\frac{1-f}{n}(S_y^2+R^2S_x^2-2RS_{xy})=\bar{y}_U^2\frac{1-f}{n}\left(C_y^2+C_x^2-2\rho C_xC_y\right)$

avec $C_x=\frac{S_x}{\bar{x}_U}$; $C_y=\frac{S_y}{\bar{y}_U}$ et $\rho=\frac{S_{xy}}{S_xS_y}$

Dans ce qui suit, en se basant sur l'espérance et la variance de l'estimateur par le quotient, nous allons trouver les caractéristiques de l'estimateur r du quotient R

1.2 L'estimateur $r=\frac{\bar{y}}{\bar{x}}$ du quotient $R=\frac{\bar{y}_U}{\bar{x}_U}$

1.2.1 Calcul du biais de r

On a : $\bar{Y}_Q=\frac{\bar{y}}{\bar{x}}\bar{x}_U=r\bar{x}_U\Rightarrow E\left(\bar{Y}_Q\right)=\bar{x}_U E(r)$ alors,

$Biais\,(r)=E(r)-R=\frac{E(\bar{Y}_Q)}{\bar{x}_U}-R$.

On a déjà montré que $E\left(\bar{Y}_Q\right)=\bar{y}_U+\frac{1-f}{n}\frac{RS_x^2-S_{xy}}{\bar{x}_U}$

Donc le biais est : $Biais(r)=R\frac{1-f}{n}(\frac{S_x^2}{\bar{x}_U^2}-\rho\frac{S_xS_y}{\bar{x}_U\bar{y}_U})=R\frac{1-f}{n}(C_x^2-\rho C_xC_y)$.

1.2.2 Calcul de la variance de r

De même, pour la variance, on a: $V(\bar{Y}_Q)=\bar{x}_U^2V(r)\Rightarrow V(r)=\frac{V(\bar{Y}_Q)}{\bar{x}_U^2}=\frac{1}{\bar{x}_U^2}E\left[\left(\bar{Y}_Q-E(\bar{y}_U)\right)^2\right]$

D'où $V(r)\approx\frac{1}{\bar{x}_U^2}E\left[\left(\bar{Y}_Q-\bar{y}_U\right)^2\right]\approx\frac{EQM(\bar{Y}_Q)}{\bar{x}_U^2}$.

1.3 Les différents estimateurs d'une moyenne

L'estimation d'une caractéristique de la population à partir d'un échantillon aléatoire pose un problème au niveau du choix de l'estimateur qui soit le plus efficace à mesurer le paramètre visé.

Nous avons déjà étudié l'estimateur par le quotient et dans ce qui suit, nous allons nous intéresser à trois autres types d'estimateurs de la moyenne (et par le fait même du total) de la population les plus utilisés soit par la moyenne, par la régression, par la différence et qui sont faites selon le type d'échantillonnage, soit un tirage à probabilité égales avec ou sans remise (PEAR ou PESR) et avec probabilités inégales avec ou sans remise (PIAR ou PISR).

Finalement, nous évaluerons la qualité de ces estimateurs en examinant leurs variances afin de déduire le meilleur.

Avant de procéder à un échantillonnage, il y a certaines démarches à suivre selon les objectifs et les caractéristiques de l'enquête en question. D'abord, il faut préciser au départ le mode de tirage des éléments qui vont composer l'échantillon (avec et/ou sans remise), ensuite déterminer la nature de la probabilité de tirage de chaque élément (à probabilités égales et/ou inégales).

Dans la section suivante, nous allons étudier les propriétés des différents estimateurs selon le mode d'échantillonnage et le type de tirage.

1.3.1 Echantillonnage avec probabilités égales

1.3.1.1 Probabilités égales avec remise PEAR:

Lorsque les tirages se font avec probabilités égales et sans remise, on retombe sur le modèle d'échantillonnage simple décrit en statistique classique. Un échantillon est alors défini comme une suite de n variables aléatoires indépendantes de même loi. Les propriétés qui en découlent sont faciles à démontrer.

Dans ce cas, on note que: $V(\bar{x}) = \frac{\sigma_x^2}{n}$ et $Cov(\bar{x};\bar{y}) = \frac{\sigma_{xy}}{n}$ où $\sigma_x^2 = \frac{1}{N}\sum_{i=1}^{N}\left(x_i - \bar{x}_U\right)^2$;

$\sigma_{xy} = \frac{1}{N}\sum_{i=1}^{N}\left(x_i - \bar{x}_U\right)\left(y_i - \bar{y}_U\right)$. Alors les différents estimateurs de la moyenne sont :

Estimateur par la moyenne : c'est l'estimateur habituel, il est défini: $\bar{y}_{SAS} = \frac{1}{n}\sum_{i=1}^{n} y_i$ et sa variance est égale à : $V(\bar{y}_{SAS}) = \frac{\sigma_U^2}{n}$ estimée par $\hat{V}(\bar{y}_{SAS}) = \frac{S_s^2}{n}$ où $S_s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(y_i - \bar{y}s\right)^2$

Estimateur par le quotient : on a déjà déterminé l'estimateur par le quotient et son biais s'écrit : $E(\bar{Y}_Q) - \bar{y}_U \approx \frac{1}{n}\left(\frac{R\sigma_x^2}{\bar{x}_U} - \frac{\sigma_{xy}}{\bar{x}_U}\right)$ et sa variance est $V(\bar{Y}_Q) \approx \frac{1}{n}(\sigma_y^2 + R^2\sigma_x^2 - 2R\sigma_{xy})$.

L'estimateur par le quotient peut souvent se justifier intuitivement : si la corrélation entre X et Y est positive et forte, de grandes valeurs de $\bar{x}_s$ seront accompagnées de grandes valeurs de $\bar{y}_s$, de sorte que le quotient $r = \frac{\bar{y}_s}{\bar{x}_s}$ sera stable.

Cette stabilité est particulièrement marquée lorsque la relation entre X et Y peut s'exprimer par une droite passant par l'origine.

On peut aussi justifier l'estimateur en faisant appel à un *modèle*, un modèle qui tente d'exprimer plus rigoureusement les suppositions qu'on vient d'énoncer informellement. On suppose que les x_i sont fixes et que les y_i sont des réalisations de variable Y dont on suppose que $y_i = ax_i + \varepsilon_i$ où a est un paramètre à estimer et $E(\varepsilon_i) = 0$ et $V(\varepsilon_i) = x_i\sigma^2$.

Alors le principe des moindres carrés, qui minimise $\sum\left(\frac{y_i - a\,x_i}{\sigma\sqrt{x_i}}\right)^2$ par rapport à a donne

$\hat{a} = \frac{\bar{y}}{\bar{x}}$. Donc le modèle estimé est $\hat{y}_i = \frac{\bar{y}}{\bar{x}} x_i$ d'où l'estimateur par le quotient est obtenu en calculant l'espérance, $\bar{Y}_Q = \frac{\bar{y}}{\bar{x}} \bar{x}_U$.

Estimateur par la différence: L'estimateur par la différence est défini par : $\bar{Y}_D = \bar{x}_U + (\bar{y} - \bar{x})$ et c'est un cas particulier de l'estimateur par la régression avec coefficient de régression égal à 1.

Intuitivement, on peut justifier cet estimateur de deux façons :

La première consiste à décomposer $\bar{y}_U$ en deux parties, l'une connue, l'autre pas : $\bar{y}_U = \bar{x}_U + (\bar{y}_U - \bar{x}_U)$. La première partie $\bar{x}_U$ est connue et n'a pas à être estimée. La deuxième partie $\bar{y}_U - \bar{x}_U$, la différence entre la moyenne des *X* et celle des *Y*, n'est pas connue et doit être estimée. Naturellement, on l'estime par la différence entre les deux moyennes échantillonnées, $\bar{y} - \bar{x}$.

La seconde se base sur le fait que $\bar{y}$ est l'estimateur naturel et privilégié de $\bar{y}_U$. Dans l'estimateur par la différence, écrit comme $\bar{Y}_D = \bar{y} + (\bar{x}_U - \bar{x})$, l'ajout du terme $\bar{x}_U - \bar{x}$ peut s'interpréter comme un ajustement à l'estimateur $\bar{y}$. Grâce à notre information sur la variable *X*, on peut deviner si, en l'occurrence, l'estimateur $\bar{y}$ a surestimé ou sous-estimé la moyenne $\bar{y}_U$.

L'estimateur par la différence est sans biais et sa variance est: $V(\bar{Y}_D) = E\left[(\bar{y} - \bar{y}_U) - (\bar{x} - \bar{x}_U)\right]^2$

$$V(\bar{Y}_D) = V(\bar{y}) - 2Cov(\bar{x}, \bar{y}) + V(\bar{x}) = \frac{1}{n}(\sigma_y^2 - 2\sigma_{xy} + \sigma_x^2).$$

L'estimateur se justifie par l'hypothèse que les écarts $y_i - x_i$ sont des variables de moyenne nulle et de même variance.

Estimateur par la régression : Cette méthode suppose une relation de type affine linéaire entre x et y. Supposons que cette relation suit le modèle, $y_i = a + bx_i + \varepsilon_i$, $V(\varepsilon_i) = \sigma^2$.

On va estimer les paramètres a et b, ensuite utiliser la grandeur $\overline{x}_U$ pour redresser et fournir l'estimateur par la régression de la moyenne $\overline{Y}$. Par la méthode des moindres carrés ordinaires appliquée à l'échantillon, on a : $\hat{a} = \overline{y} - \hat{b}\,\overline{x}$ et $\hat{b} = \sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) \Big/ \sum_{i=1}^{n}(x_i - \overline{x})^2$.

Dans ce cas $\hat{b} = \dfrac{S_{xy}}{S_x^2}$, avec $S_{xy} = \dfrac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$ et $S_x^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$.

D'où le modèle estimé est $\hat{y}_i = \overline{y} + \dfrac{S_{xy}}{S_x^2}(x_i - \overline{x})$ alors, l'estimateur par la régression est $\overline{Y}_{Rg} = \overline{y} + \dfrac{S_{xy}}{S_x^2}(\overline{x}_U - \overline{x})$.

La variance de cet estimateur est exprimée par : $V(\overline{Y}_{Rg}) = E\left[(\overline{y} - \overline{y}_U) - \hat{b}(\overline{x} - \overline{x}_U)\right]^2$

$$V(\overline{Y}_{Rg}) = V(\overline{y}) - 2\hat{b}Cov(\overline{x}, \overline{y}) + \hat{b}^2 V(\overline{x}) = \frac{1}{n}\left(\sigma_y^2 - 2\hat{b}\sigma_{xy} + \hat{b}^2\sigma_x^2\right) = \frac{1}{n}\sigma_y^2(1 - \rho^2).$$

Remarque : Selon l'expression de l'estimateur par la régression, on remarque, par un choix adéquat de b que cet estimateur comprend notamment les deux cas, l'estimateur par la moyenne et celui par le quotient. Évidemment si b est égal à zéro, $\overline{Y}_{Rg}$ se réduit à $\overline{y}$ et si $b = \dfrac{\overline{y}}{\overline{x}}$ alors $\overline{Y}_{Rg} = \overline{y} + \dfrac{\overline{y}}{\overline{x}}(\overline{x}_U - \overline{x}) = \dfrac{\overline{y}}{\overline{x}}\overline{x}_U = \overline{Y}_Q$ et aussi, c'est évident que $\overline{Y}_{Rg} = \overline{Y}_D$ si $b = 1$.

1.3.1.2 Probabilités égales sans remise PESR:

Lorsque les tirages se font sans remise (avec probabilités égales), les estimateurs ne changent pas. Seule la variance change et ce, par le simple ajout d'un facteur de correction.

On a : $V(\bar{x}) = \frac{1-f}{n} s_x^2$ et $Cov(\bar{x}, \bar{y}) = \frac{1-f}{n} S_{xy}$ où $S_x^2 = \frac{1}{N-1} \sum_{i=1}^{N} \left(x_i - \bar{x}_U\right)^2$;

$S_{xy} = \frac{1}{N-1} \sum_{i=1}^{N} \left(x_i - \bar{x}_U\right)\left(y_i - \bar{y}_U\right)$. Donc, les formules des estimateurs s'écrivent comme suit :

Estimateur par la moyenne : il reste le même $\bar{y}_{SAS} = \frac{1}{n} \sum_{i=1}^{n} y_i$ et sa variance est égale à: $V(\bar{y}_{SAS}) = \frac{N-n}{N-1} \frac{\sigma_U^2}{n}$ estimée par $\hat{V}(\bar{y}_{SAS}) = \frac{N-n}{N-1} \frac{S_s^2}{n}$ où $S_s^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(y_i - \bar{y}s\right)^2$.

Estimateur par le quotient : son biais s'écrit : $E(\bar{Y}_Q) - \bar{y}_U = \frac{1-f}{n} \left(\frac{RS^2{}_x}{\bar{x}_U} - \frac{S_{xy}}{\bar{x}_U} \right)$ et sa variance est $V(\bar{Y}_Q) = \frac{1-f}{n} (S_y^2 + R^2 S_x^2 - 2RS_{xy})$.

Estimateur par la différence: $V\left(\bar{Y}_D\right) = V(\bar{y}) - 2Cov(\bar{x}, \bar{y}) + V(\bar{x}) = \frac{1-f}{n} \left(S_y^2 - 2S_{xy} + S_x^2\right)$

Estimateur par la régression : dans ce cas $\hat{b} = \frac{S_{xy}}{S_x^2}$, alors :

$$V\left(\bar{Y}_{Rg}\right) = V(\bar{y}) - 2\hat{b}Cov(\bar{x}, \bar{y}) + \hat{b}^2 V(\bar{x}) = \frac{1-f}{n} \left(S_y^2 - 2\hat{b}S_{xy} + \hat{b}^2 S_x^2\right) = \frac{1-f}{n} S_y^2 \left(1 - \rho^2\right)$$

1.3.2 Echantillonnage avec probabilités inégales

Les sondages à probabilités inégales se justifient par le fait que dans certains cas et pour certains domaines d'étude, il est intéressant de donner à certaines unités à échantillonner une probabilité plus forte d'être tirée.

1.3.2.1 Plans à probabilités inégales

Les plans à probabilités inégales ont fait l'objet de très nombreuses études consistant à introduire un « effet de taille » sur les probabilités d'inclusion. Lorsque les unités sont de tailles très variables, il est utile de les sélectionner avec des probabilités de variables.

En d'autres termes, si une variable auxiliaire X permet de mesurer approximativement cet effet, il est particulièrement intéressant de sélectionner les unités d'observations avec des probabilités d'inclusion proportionnelles à cette variable auxiliaire. Le gain de précision sera alors très important.

On utilise cette méthode d'échantillonnage lorsque les unités de la population étudiée contribuent inégalement au total d'intérêt.

Donc, dans un sondage à PIAR, chaque unité i de la population U a la probabilité p_i d'être tirée à chacun des tirages.

De plus, on a: $\sum_{i\in U} p_i = 1$ et p_i est souvent proportionnelle à une mesure de la taille de l'unité i. Si y_i est sa taille, alors on choisit $p_i = \frac{y_i}{\sum_{j\in U} y_j} = \frac{y_i}{t_y}$ et $\sum_{i\in U} p_i = 1$.

1.3.2.2 Probabilités inégales avec remise PIAR:

L'estimateur de la moyenne $\bar{y}_U$ se définit par: $\bar{y} = \frac{1}{nN}\sum_{i=1}^{n}\frac{y_i}{p_i}$ où y_i est la variable aléatoire pour l'unité qui sera sélectionnée au i^e tirage et p_i sa probabilité d'être sélectionnée à chaque tirage. Cet estimateur est sans biais $E(\bar{y}) = \frac{1}{N}\sum_{i=1}^{N} y_i = \bar{y}_U$.

La variance de $\bar{y}$ est égale à $V(\bar{y}) = \frac{1}{nN^2}\sum_{i\in U} p_i\left[\frac{y_i}{p_i} - \left(\sum_{i\in U} y_i\right)\right]^2 = \frac{1}{nN^2}\left(\sum_{i\in U}\frac{y_i^2}{p_i} - t_y^2\right)$ et peut être estimée par : $\hat{V}(\bar{y}) = \frac{1}{n(n-1)N^2}\left(\sum_{i\in s}\frac{y_i}{p_i} - \hat{t}_y\right)^2$

où $\hat{t}_y$ est l'estimateur du total t_y défini par : $\hat{t}_y = \frac{1}{n}\sum_{i=1}^{n}\frac{y_i}{p_i}$ et il est sans biais.

La variance de $\hat{t}_y$ est égale à : $V(\hat{t}_y) = \frac{1}{n}\sum_{i\in U} p_i\left[\frac{y_i}{p_i} - \left(\sum_{i\in U} y_i\right)\right]^2 = \frac{1}{n}\left(\sum_{i\in U}\frac{y_i^2}{p_i} - t_y^2\right)$ et on l'estime sans biais par: $\hat{V}(\hat{t}_y) = \frac{1}{n(n-1)}\left(\sum_{i\in s}\frac{y_i}{p_i} - \hat{t}_y\right)^2$.

D'après la formule de $V(\hat{t}_y)$, il est évident que la variance est nulle, c'est-à-dire minimale, si: $p_i = \frac{y_i}{\sum_{j\in U} y_j}$.

Rappel de la variance du total dans le cas d'un tirage à PEAR :

$$V(\hat{t}_{y\,PEAR}) = N^2\frac{\sigma_U^2}{n} = \frac{1}{n}\left(\sum_{i=1}^{N} N y_i^2 - t_y^2\right).$$

Donc $V(\hat{t}_{y\,PEAR}) > V(\hat{t}_{y\,PIAR})$ implique $\sum_{i=1}^{N} N y_i^2 > \sum_{i=1}^{N}\frac{y_i^2}{p_i} \Leftrightarrow \sum_{i=1}^{N} y_i^2\left(\frac{1}{p_i} - \frac{1}{1/N}\right) < 0$

La différence est d'autant plus grande si : $p_i > 1/N$ et y_i^2 est grand.

$p_i < 1/N$ et y_i^2 est petit.

1.3.2.3 Probabilités inégales sans remise PISR:

Dans ce cas, le problème se complique du fait que chaque tirage modifie les conditions du tirage suivant. Ainsi, en plus des probabilités de sortie au premier tirage, il faut connaître les probabilités de sortie de l'unité i au deuxième tirage, sachant que l'unité j est sortie au premier tirage, et ainsi de suite... Les procédures proposées pour le tirage d'un échantillon PISR sont très nombreuses et souvent complexes. L'estimateur de Horvitz-Thompson (1952) permet de contourner (provisoirement) le problème, car son point de départ est l'ensemble des probabilités d'inclusion, quelles qu'elles soient. Cette approche est une approche générale, pas seulement limitée aux sondages à probabilités inégales. Néanmoins, elle est la seule utilisable dans un sondage PISR.

Soit :

n : la taille de l'échantillon.

π_i : probabilité que l'unité i appartient à l'échantillon ou probabilité d'inclusion d'ordre 1.

$d_i = 1/\pi_i$: le poids de sondage. Intuitivement, c'est le nombre d'individus de la population représentés par l'individu i dans l'échantillon.

π_{ij} : probabilité que les unités i et j appartiennent simultanément à l'échantillon ou probabilité d'inclusion d'ordre 2.

s : un échantillon ou un sous-ensemble quelconque de la population $U=\{1\ ;\ 2\ ;\ ;\ N\}$.

S : l'ensemble des échantillons possibles.

$p(s)$: la loi de probabilité sur tous les échantillons possibles, $\sum_{s \in S} p(s) = 1$.

y_i : la valeur prise par la variable Y pour la i^e unité.

Les relations suivantes sont vérifiées si l'échantillon est de taille fixe ;

$$\sum_{1}^{N}\pi_i = n \;\; ; \;\; \sum_{j(\neq i)=1}^{N}\pi_{ij} = (n-1)\pi_i \;\; ; \;\; \sum_{j=1}^{N}\pi_{ij} = n\pi_i \;\; \text{avec} \;\; \pi_{ii} = \pi_i \;\; ; \;\; \sum_{i=1}^{N}\sum_{j>i}\pi_{ij} = \frac{1}{2}n(n-1)$$

$$\sum_{i=1}^{N}\sum_{j\neq i}\pi_{ij} = n(n-1) \;\; ; \;\; \sum_{i=1}^{N}\sum_{j\neq i}\pi_i\pi_j = n^2 - \sum_{i=1}^{N}\pi_i^2 \;\; ; \;\; \sum_{(i\neq j)=1}^{N}\pi_i\pi_j = (n-\pi_i)\pi_i .$$

Etant donné que la probabilité π_i est connue, si la i^e unité appartient à l'échantillon s, donc, on peut calculer pour tout $i \in s, \tilde{y}_i = \frac{y_i}{\pi_i}$, avec y_i, $i=1,\ldots,n$ sont les valeurs de la variable Y correspondant à la sélection de n individus selon le plan de sondage.

Cette quantité est appelée *la valeur dilatée de la variable Y* pour la i^e unité.

Donc, l'estimateur de Horvitz et Thompson du total t_y sur la population $\left(t_y = \sum_{1}^{N} y_i\right)$ est défini par le total sur l'échantillon des valeurs dilatées, c'est-à-dire, par $\tilde{t}_y = \sum_{s}\tilde{y}_i = \sum_{i=1}^{n}\frac{y_i}{\pi_i}$ et elle est appelée *somme dilatée en* y_i.

Une technique, introduite par Cornfield (1944) permet de démontrer facilement certaines propriétés des estimateurs :

Soit I_i une variable indicatrice de la sélection de l'individu i

où $I_i = 1$ si $i \in s$ et $I_i = 0$ sinon.

Donc, pour $i=1,\ldots,N$, on a : $\tilde{t}_y = \sum_{U} I_i\tilde{y}_i$.

Cette somme apparaît comme une statistique linéaire aléatoire en I_i .

Elle peut servir comme estimateur sans biais de $t_y = \sum_{1}^{N} y_i$ car les $\tilde{y}_i$ étant considérés comme des nombres fixes : $E(t_y) = E\left(\sum I_i \tilde{y}_i\right) = \sum \pi_i \tilde{y}_i = \sum y_i = t_y$; $i=I,.....,N$.

Avec : $\pi_i = \Pr(i \in s) = E(I_i)$; $V(I_i) = \pi_i(1-\pi_i)$; $\pi_{ij} = P(i \in s et\ j \in s) = E(I_{ij}) = I_{ij}$;

$$\text{où } I_{ij} = \begin{cases} I_i * I_j \ si\ i \in s \ \ et \ \ j \in s \\ 0 \quad \sin on \end{cases}$$

$\Delta_{ij} = Cov(I_i, I_j) = E(I_{ij}) - E(I_i)E(I_j) = \pi_{ij} - \pi_i\pi_j$. (si $i = j$; $\Delta_{ii} = \pi_i(1-\pi_i)$)

1.3.2.4 Propriétés de l'estimateur de Horvitz et Thompson

Son espérance

L'estimateur d'Horvitz et Thompson du total est sans biais car :

$$E(\tilde{t}_y) = E(\sum_s \tilde{y}_i) = E(\sum_{i=1}^{n} \frac{y_i}{\pi_i}) = E(\sum_{i \in U} \frac{y_i}{\pi_i} I_i) = \sum_{i \in U} \frac{y_i}{\pi_i} E(I_i) = \sum_{i \in U} y_i = t_y$$

Si certaines probabilités d'inclusion sont nulles, alors dans ce cas, l'estimateur est biaisé. Mais ce biais ne dépend que des unités qui n'ont aucune chance d'être dans l'échantillon : c'est un problème de couverture.

Sa variance

Dans le cas général, si $\pi_i > 0$ pour tout i de la population U, alors la variance de l'estimateur d'Horvitz et Thompson du total est :

$$V(\tilde{t}_y) = \sum_{i \in U} \sum_{j \in U} \frac{y_i}{\pi_i} \frac{y_j}{\pi_j} \Delta_{ij} = \sum_{i \in U} \sum_{j \in U} \frac{y_i}{\pi_i} \frac{y_j}{\pi_j} (\pi_{ij} - \pi_i\pi_j) \text{ car}$$

$$V(\tilde{t}_y) = V(\sum_s \tilde{y}_i) = V(\sum_{i=1}^{n} \frac{y_i}{\pi_i}) = V(\sum_{i\in U} \frac{y_i}{\pi_i} I_i) = \sum_{i\in U} V(\frac{y_i}{\pi_i} I_i) + \sum_{i\in U} \sum_{j\in U, i\neq j} Cov(\frac{y_i}{\pi_i} I_i, \frac{y_j}{\pi_j} I_j)$$

$$= \sum_{i\in U} (\frac{y_i}{\pi_i})^2 V(I_i) + \sum_{i\in U} \sum_{j\in U, i\neq j} \frac{y_i}{\pi_i} \frac{y_j}{\pi_j} Cov(I_i, I_j) = \sum_{i\in U} (\frac{y_i}{\pi_i})^2 \pi_i(1-\pi_i) + \sum_{i\in U} \sum_{j\in U, i\neq j} \frac{y_i}{\pi_i} \frac{y_j}{\pi_j} Cov(I_i, I_j)$$

$$= \sum_{i\in U} \sum_{j\in U} \frac{y_i}{\pi_i} \frac{y_j}{\pi_j} \Delta_{ij} \quad \text{car} \quad \Delta_{ii} = \pi_i(1-\pi_i)$$

Cette variance est estimée par $\hat{V}(\tilde{t}_y) = \sum_{i\in s} \sum_{j\in s} \frac{y_i}{\pi_i} \frac{y_j}{\pi_j} \frac{\Delta_{ij}}{\pi_{ij}}$ qui présente l'inconvénient de pouvoir prendre des valeurs négatives. Alors Sen, Yates et Grundy (SYG 1953) ont proposé un autre estimateur basé sur le fait que si $\pi_i > 0$ pour tout i de la population U et le plan est de taille fixe, la variance du total peut s'écrire comme suit : $V(\tilde{t}_y) = -\frac{1}{2} \sum_{i\in U} \sum_{j\in U} (\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j})^2 \Delta_{ij} = -\frac{1}{2} \sum_{i\in U} \sum_{j\in U} (\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j})^2 (\pi_{ij} - \pi_i \pi_j)$

Preuve :

$$-\frac{1}{2} \sum_{i\in U} \sum_{j\in U} (\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j})^2 \Delta_{ij} = -\frac{1}{2} \sum_{i\in U} \sum_{j\in U} \left[\left(\frac{y_i}{\pi_i} \right)^2 + \left(\frac{y_j}{\pi_j} \right)^2 - 2 \frac{y_i}{\pi_i} \frac{y_j}{\pi_j} \right] \Delta_{ij}$$

$$= -\sum_{i\in U} \sum_{j\in U} (\frac{y_i}{\pi_i})^2 \Delta_{ij} + \sum_{i\in U} \sum_{j\in U} \frac{y_i}{\pi_i} \frac{y_j}{\pi_j} \Delta_{ij}$$

$$= -\sum_{i\in U} \left(\frac{y_i}{\pi_i} \right)^2 \sum_{j\in U} \Delta_{ij} + \sum_{i\in U} \sum_{j\in U} \frac{y_i}{\pi_i} \frac{y_j}{\pi_j} \Delta_{ij} = 0 + \sum_{i\in U} \sum_{j\in U} \frac{y_i}{\pi_i} \frac{y_j}{\pi_j} \Delta_{ij} = V(\tilde{t}_y)$$

car pour i fixe, $\sum_{j\in U} \Delta_{ij} = 0$ dans un plan de taille fixe.

Son estimateur sans biais est $\hat{V}(\bar{Y}_{HT}) = -\frac{1}{2}\sum_{i\in s}\sum_{j\in s}(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j})^2 \frac{\Delta_{ij}}{\pi_{ij}}$ qui est positif sous la condition (dite de Sen-Yates-Grundy) $\pi_{ij} \leq \pi_i \pi_j$ pour tout i et j.

1.3.2.5 Estimation d'une moyenne par l'estimateur d'Horvitz et Thompson

Lorsque la taille de la population est connue, on peut estimer la moyenne de Y avec l'estimateur de Horvitz et Thompson : $\bar{Y}_{HT} = \frac{1}{N}\tilde{t}_y = \frac{1}{N}\sum_s \tilde{y}_i = \frac{1}{N}\sum_{i=1}^{n}\frac{y_i}{\pi_i}$

Toutes les propriétés vues pour l'estimateur de Horvitz et Thompson d'un total s'appliquent à une moyenne. Il suffit d'adapter les formules pour tenir compte du coefficient N. Donc, l'estimateur de la moyenne est sans biais $E(\bar{Y}_{HT}) = \bar{Y}$ et sa variance $V(\bar{Y}_{HT}) = \frac{1}{N^2}\sum_{i\in U}\sum_{j\in U}\frac{y_i}{\pi_i}\frac{y_j}{\pi_j}Cov(I_i, I_j)$; $V(\bar{Y}_{HT}) = -\frac{1}{2N^2}\sum_{i\in U}\sum_{j\in U}(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j})^2 Cov(I_i, I_j)$ est estimée par $\hat{V}(\bar{Y}_{HT}) = \frac{1}{N^2}\sum_{i\in s}\sum_{j\in s}\frac{y_i}{\pi_i}\frac{y_j}{\pi_j}\frac{Cov(I_i, I_j)}{\pi_{ij}}$; $\hat{V}(\bar{Y}_{HT}) = -\frac{1}{2N^2}\sum_{i\in s}\sum_{j\in s}(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j})^2 \frac{Cov(I_i, I_j)}{\pi_{ij}}$

1.3.2.6 Estimation d'une moyenne par l'estimateur de Hajek

Lorsque la taille de la population est inconnue, on utilise l'estimateur de Hajek (1971) pour estimer la moyenne de Y. Il est défini par :

$$\bar{Y}_H = \frac{1}{\hat{N}}\tilde{t}_y = \frac{1}{\hat{N}}\sum_s \tilde{y}_i = \frac{1}{\hat{N}}\sum_{i=1}^{n}\frac{y_i}{\pi_i} = \sum_{i=1}^{n}\frac{y_i}{\pi_i}\Big/\sum_{i=1}^{n}\frac{1}{\pi_i}$$

Cet estimateur est un quotient de deux estimateurs et comme tel il peut s'avérer plus précis, même lorsque N est connu. Comme tout estimateur par le quotient, il est biaisé, mais son biais est asymptotiquement nul et négligeable lorsque n est grand.

1.3.3 Echantillonnage avec probabilités variables

1.3.3.1 Méthode de Lahiri-Midzuno

Lahiri (1951) a proposé un mode de tirage conçu pour que le quotient échantillonnal soit un estimateur sans biais du quotient de la population. Midzuno (1952) a proposé une façon de procéder pour réaliser un tirage de Lahiri. Ce tirage consiste à s'assurer que la probabilité $p(s)$ qu'un échantillon s soit tiré soit proportionnelle à la moyenne échantillonnale $\bar{x}_s$ de la variable auxiliaire. La constante de proportionnalité α doit satisfaire la condition $\sum_{s\in S} \Pr(\textit{échantillon } s \textit{ soit tiré}) = 1$, donc $\sum_{s\in S} p(s) = \sum_{s\in S} \alpha \bar{x}_s = 1$.

On sait que dans un échantillonnage aléatoire simple SAS, on a $E(\bar{x}_s) = \sum_{s\in S} \frac{\bar{x}_s}{\binom{N}{n}} = \bar{x}_U$

et par conséquent $\alpha \bar{x}_U \binom{N}{n} = 1$ d'où $\alpha = \frac{1}{\bar{x}_U \binom{N}{n}}$, avec $\binom{N}{n} = \frac{N!}{n!(N-n)!}$.

Théorème : Selon la méthode de Lahiri, le quotient échantillonnal r est un estimateur sans biais du quotient R.

Preuve : On a $E(r) = \sum_{s\in S} r(s)P(s) = \sum_{s\in S} \frac{\bar{y}_s}{\bar{x}_s}(\alpha \bar{x}_s) = \alpha \binom{N}{n} \sum_{s\in S} \frac{\bar{y}_s}{\binom{N}{n}} = \frac{\bar{y}_U}{\bar{x}_U} = R$

Ce qui signifie que $\bar{Y}_Q$ l'estimateur par quotient de la moyenne de la population est sans biais: $E(\bar{Y}_Q) = \sum_{s\in S} \bar{Y}_Q P(s) = \sum_{s\in S} \frac{\bar{y}_s}{\bar{x}_s} \bar{x}_U (\alpha \bar{x}_s) = \alpha \binom{N}{n} \bar{x}_U \sum_{s\in S} \frac{\bar{y}_s}{\binom{N}{n}} = \bar{y}_U$

Le principe proposé par Lahiri stipule seulement que $p(s)$ soit proportionnelle à $\bar{x}_s$. C'est Midzuno qui a proposé une façon de réaliser ce type d'échantillonnage. La procédure proposée est la suivante : tirer une première unité avec probabilité proportionnelle aux valeurs de la variable auxiliaire (les x_i) et ensuite tirer un échantillon aléatoire simple de taille $(n-1)$ unités parmi les $(N-1)$ unités restantes de la population.

On démontre qu'en procédant ainsi, on a bien $p(s) = \dfrac{\bar{x}_s}{\binom{N}{n}\bar{x}_U}$.

Soit s un échantillon aléatoire tiré selon cette méthode. On a :

$$P(s) = \sum_{i=1}^{n} \Pr\ (l'unité\ i\ soit\ tirée) * \Pr(les(n-1) autres\ unités\ soient\ tirées\ par\ la\ suite)$$

$$= \sum_{i=1}^{n} \frac{x_i}{t_x} \frac{1}{\binom{N-1}{n-1}} = \frac{n\bar{x}_s}{t_x \binom{N-1}{n-1}} = \frac{1}{\bar{x}_U \binom{N}{n}} \bar{x}_s$$

1.4 Estimateurs sans biais : cas d'échantillonnage aléatoire simple

Dans ce qui suit, nous présentons quelques estimateurs sans biais dans le cas d'échantillonnage aléatoire simple.

1.4.1 Estimateur de Mickey

Si le biais d'un estimateur est identifié, il est logique d'envisager de corriger l'estimateur en ajoutant à celui-ci ou en soustrayant de celui-ci le biais. Ce qui conduit à une classe d'estimateurs sans biais du quotient dans un échantillon aléatoire simple.

Dans ce contexte, Mickey (1959) propose une approche qui ressemble à celle de *Jackknife* dont l'idée consiste à éliminer chacune des observations tour à tour de l'échantillon d'origine et à recalculer la statistique d'intérêt sur les observations restantes.

Mickey suggère d'utiliser une partie de l'échantillon pour estimer le paramètre et une autre partie pour corriger le biais.

Considérons un estimateur de la moyenne de la forme $\bar{y} - a(\bar{x} - \bar{x}_U)$. Si a est une constante, l'estimateur est sans biais, mais a pourrait bien être fonction des observations, comme dans l'estimateur par la régression. Dans ce cas, l'estimateur est généralement biaisé.

Considérons donc un espace échantillon formé de suites de n éléments *ordonnés*, c'est-à-dire dans lequel on tient compte de l'ordre des tirages. Soit les n_1 ($< n$) premiers tirages et Z_1 une fonction des n_1 premières observations. L'estimateur $\bar{y} - a(Z_1)(\bar{x} - \bar{x}_U)$ est biaisé, mais le biais peut être estimé par un multiple de la différence entre $\bar{y}_1 - a(Z_1)(\bar{x}_1 - \bar{x}_U) - \bar{y}_1 - \bar{y} - a(Z_1)(\bar{x}_1 - \bar{x}) = \bar{y}_1 - \bar{y} - a(Z_1)(\bar{x}_1 - \bar{x})$, où $\bar{y}_1$ et $\bar{x}_1$ sont les moyennes de n_1 premières observations.

En effet, l'estimateur qu'on pourrait proposer est :

$$T^* = \bar{y} - a(Z_1)(\bar{x} - \bar{x}_U) - \frac{n_1(N-n)}{(n-n_1)N}\left[\bar{y}_1 - \bar{y} - a(Z_1)(\bar{x}_1 - \bar{x})\right], \quad n_2 = n - n_1$$

Montrons que T^* est sans biais.

Considérons l'estimateur T^* basé sur les n_2 dernières observations tirées, les n_1 premières n'étant utilisées que pour estimer a (par $a(Z_1)$), soit $u_1 = \bar{y}_2 - a(Z_1)(\bar{x}_2 - \bar{x}_{U_2})$, où $\bar{x}_{U_2}$ est la moyenne de la population de taille (N-n_1) de laquelle les n_1 premières observations sont retirées, où $\bar{y}_2 = \dfrac{n\bar{y} - n_1\bar{y}_1}{n_2}$, $\bar{x}_2 = \dfrac{n\bar{x} - n_1\bar{x}_1}{n_2}$, et $\bar{x}_{U_2} = \dfrac{N\bar{x}_U - n_1\bar{x}_1}{N - n_1}$.

Alors $u_1 = \dfrac{n\bar{y} - n_1\bar{y}_1}{n_2} - a(Z_1)\left[\dfrac{n\bar{x} - n_1\bar{x}_1}{n_2} - \dfrac{N\bar{x}_U - n_1\bar{x}_1}{N - n_1}\right]$.

Conditionnellement à Z_1, on a $E(\bar{y}_2 \mid Z_1) = \dfrac{N\bar{y}_U - n_1\bar{y}_1}{N-n_1}$, $E(\bar{x}_2 \mid Z_1) = \dfrac{N\bar{x}_U - n_1\bar{x}_1}{N-n_1} = \bar{x}_{U_2}$.

Donc $E(u_1|Z_1) = E\left[\dfrac{N\bar{y}_U - n_1\bar{y}_1}{N-n_1}\right] - a(Z_1)\,E(\bar{x}_2 - \bar{x}_{U_2} \mid Z_1) = \dfrac{N\bar{y}_U - n_1\bar{y}_1}{N-n_1}$ et par conséquent $E(u_1) = \bar{y}_U$.

Or $T^* = \dfrac{(N-n_1)u_1 + n_1\bar{y}_1}{N}$, donc $E(T^*) = E\left(\dfrac{(N-n_1)u_1 + n_1\bar{y}_1}{N}\right) = \dfrac{(N-n_1)\bar{y}_U + n_1\bar{y}_U}{N} = \bar{y}_U$.

L'estimateur $T^* = T^*(s)$ dépend de l'échantillon ordonné s^*, une permutation des éléments de l'échantillon non ordonné s^*. On peut améliorer cette estimation en utilisant une technique semblable à celle qui mène au théorème de Rao-Blackwell.

Alors l'estimateur amélioré est défini par:

$\hat{\bar{y}}_M$=Moyenne des $n!$ valeurs de $T^*(s^*)$ correspondant aux $n!$ permutations s^* des éléments de s.

Nous pouvons montrer que $\hat{\bar{y}}_M$ est également sans biais et que sa variance est inférieure ou égale à celle de T^*. Notons que $\hat{\bar{y}}_M = E[T^*|s]$, donc $E[\hat{\bar{y}}_M] = E[E(T^* \mid s)] = E(T^*) = \bar{y}_U$.

Montrons que la variance de $\hat{\bar{y}}_M$ est inférieure à celle de T^* :

$V(T^*) = E[(T^* - \bar{y}_U)^2] = E[(T^* - \hat{\bar{y}}_M + \hat{\bar{y}}_M - \bar{y}_U)^2] = E[(\hat{\bar{y}}_M - \bar{y}_U)^2] + E[(T^* - \hat{\bar{y}}_M)^2] +$

$2E[(\hat{\bar{y}}_M - \bar{y}_U)(T^* - \hat{\bar{y}}_M)]$.

Le troisième terme disparaît, car :

$E[(\hat{\bar{y}}_M - \bar{y}_U)(T^* - \hat{\bar{y}}_M)] = E\{E[(\hat{\bar{y}}_M - \bar{y}_U)(T^* - \hat{\bar{y}}_M)] \mid s\} =$

$E\{(\hat{\bar{y}}_M - \bar{y}_U)E[(T^* - \hat{\bar{y}}_M)] \mid s\} = 0$ puisque, étant donné s, $E[(T^* - \hat{\bar{y}}_M)] = E(T^* \mid s) - \hat{\bar{y}}_M$

$= \hat{\bar{y}}_M - \hat{\bar{y}}_M = 0$.

1.4.2 Estimateur de Hartley-Ross

Par la méthode de Mickey, on a ainsi défini une grande famille d'estimateurs dont on peut tirer un grand nombre d'estimateurs particuliers. L'estimateur de Hartley-Ross en est un. Il est défini par :

$\hat{\bar{y}}_{HR} = \bar{r}\,\bar{x}_U + \frac{n(N-1)}{(n-1)N}(\bar{y} - \bar{r}\,\bar{x})$, où $\bar{r} = \frac{1}{n}\sum_{i\in s}\frac{y_i}{x_i}$ désigne la moyenne des quotients.

Si on prend n_1=1 et $a(Z_1)$=$\frac{y_1}{x_1}$, alors l'estimateur T^* devient

T^*=$\frac{y_1}{x_1}\bar{x}_U + \frac{n(N-1)}{(n-1)N}\left(\bar{y} - \frac{y_1}{x_1}\bar{x}\right)$,

et $\hat{\bar{y}}_M = \frac{1}{n}\sum_{i\in s}\left\{\frac{y_i}{x_i}\bar{x}_U + \frac{n(N-1)}{(n-1)N}\left(\bar{y} - \frac{y_i}{x_i}\bar{x}\right)\right\} = \frac{1}{n}\left\{\bar{x}_U\sum_{i\in s}\frac{y_i}{x_i} + \frac{n(N-1)}{(n-1)N}\sum_{i\in s}\left(\bar{y} - \frac{y_i}{x_i}\bar{x}\right)\right\}$

$= \bar{x}_U\frac{1}{n}\sum_{i\in s}\frac{y_i}{x_i} + \frac{n(N-1)}{(n-1)N}\frac{1}{n}\sum_{i\in s}\left(\bar{y} - \frac{y_i}{x_i}\bar{x}\right) = \bar{x}_U\bar{r} + \frac{n(N-1)}{(n-1)N}\left[\bar{y} - \bar{x}\frac{1}{n}\sum_{i\in s}\left(\frac{y_i}{x_i}\right)\right]$

$= \bar{x}_U\bar{r} + \frac{n(N-1)}{(n-1)N}[\bar{y} - \bar{x}\bar{r}]$.

Est-ce qu'on aurait intérêt à prendre n_1 = 2, 3, … , ou n-1? La réponse est non, car l'estimateur ne changera pas si on prend d'autres valeurs que n_1 = 1.

Preuve :

T^*= $\bar{y} - a(Z_1)(\bar{x} - \bar{x}_U) - \frac{n_1(N-n)}{(n-n_1)N}\left[\bar{y}_1 - \bar{y} - a(Z_1)(\bar{x}_1 - \bar{x})\right]$ et avec $a(Z_1)$=$\frac{y_1}{x_1}$ et n_1 quelconque,

on a $T^* = \bar{y} - \frac{y_1}{x_1}\bar{x} + \frac{y_1}{x_1}\bar{x}_U - \frac{n_1(N-n)}{N(n-n_1)}\left(\frac{y_1}{x_1}\bar{x} - \bar{y}\right) = \frac{y_1}{x_1}\bar{x}_U + \frac{n(N-n_1)}{N(n-n_1)}\left(\bar{y} - \frac{y_1}{x_1}\bar{x}\right)$.

Une autre application est la suivante : une estimation par la régression où on utilise les deux premières observations pour estimer a. Alors l'estimateur qu'on obtient ainsi est

$$\bar{y} - \frac{2}{n(n-1)}\left(\bar{x} - \bar{x}_U\right)\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\frac{y_i - y_j}{x_i - x_j} + \frac{4(N-n)}{n(n-1)(n-2)N}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\frac{y_i - y_j}{x_i - x_j}\left(\frac{x_i + x_j}{2} - \bar{x}\right)$$

Lorsque $x_i = x_j$, on laisse tomber les deux derniers termes.

CHAPITRE II

DIFFERENTS ESTIMATEURS D'UNE PROPORTION

Dans ce chapitre, on va traiter un cas particulier des estimateurs par le quotient, celui où la variable d'intérêt Y est dichotomique. Les estimateurs présentés peuvent alors prendre une forme plus simple. Deux cas se présentent : celui où la variable auxiliaire X est elle aussi dichotomique et celui où elle est quantitative. Dans le premier cas, la moyenne de la population est une *proportion* qu'on désignera par π. On va déterminer les formules particulières des estimateurs: par la moyenne, par la différence, par le quotient et par la régression. Et nous montrerons finalement, que ces estimateurs peuvent être unifiés par la notion de *décalage* de Stuart.

Par la suite et dans ce contexte, nous ferons des simulations et tenterons dans la mesure du possible d'arriver à des recommandations particulières.

2.1 Formules des estimateurs lorsque la variable auxiliaire est dichotomique

Y est une variable dichotomique et la moyenne de la population $\bar{y}_U$ est une proportion notée par π qu'on veut estimer. Par exemple, considérons une population de N ménages dont on voudrait estimer la proportion π qui sont mono-parentaux. Il est possible qu'une information auxiliaire X, comme par exemple, le fait qu'un ménage soit ou non sur l'assistance sociale puisse servir à mieux estimer π.

Évidemment, il faut connaître la valeur de X pour chaque ménage de la population, ou du moins connaître sa moyenne. L'estimation pourrait s'avérer meilleure s'il se trouve qu'une forte proportion des ménages sur l'assistance sociale sont mono-parentaux.

Nous considérerons tout particulièrement le cas où la variable auxiliaire est un effectif, comme dans l'exemple suivant, où il s'agit en fait d'un tirage par grappes.

Soit une population de N unités

$y_i = 1 \quad ou \quad 0 \qquad i = 1,.....,N$

z_i = une mesure quantitative de l'importance de l'unité i : $\sum_{i=1}^{N} z_i = 1$

Lorsque les N unités sont des grappes (ou unités primaires), $z_i = \frac{M_i}{M}$ où M_i est le nombre d'unités secondaires dans l'unité primaire i et $M = \sum_{i=1}^{N} M_i$.

Considérons le cas où les tirages se font avec remise et la probabilité de sélection de l'unité i est z_i. Le paramètre à estimer est $\pi = \frac{1}{N}\sum_{i=1}^{N} y_i$ et l'estimateur de Hajek est : $\hat{\pi}_m = \frac{1}{n}\sum_{i=1}^{n} \frac{y_i}{N z_i}$.

Mais on pourrait également commencer par estimer plutôt le paramètre $(1-\pi)$, pour ensuite le soustraire de 1, ce qui donne l'estimateur $\hat{\pi}_d = 1 - \frac{1}{n}\sum_{i=1}^{n} \frac{1-y_i}{N z_i}$.

Or, contrairement à ce qui se produit dans l'échantillonnage avec probabilités égales, ces deux estimateurs ne coïncident pas. Il se trouve qu'on peut considérer $\hat{\pi}_d$ comme un

estimateur par la différence, puisque $\hat{\pi}_d = \hat{\pi}_m + \left(1 - \frac{1}{n}\sum_{i=1}^{n} \frac{1}{Nz_i}\right)$. Notons $\hat{\pi}_x = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{Nz_i}$ donc $\hat{\pi}_d = \hat{\pi}_m + (1 - \hat{\pi}_x)$.

Le premier terme à droite est l'estimateur par la moyenne, le deuxième est la moyenne (connue) d'une variable auxiliaire X qui prendrait la valeur 1 pour toutes les unités et le troisième terme est une estimation de la moyenne de la variable auxiliaire.

En fait, les deux estimateurs $\hat{\pi}_m$ et $\hat{\pi}_d$ ne sont que deux cas particuliers d'une classe d'estimateurs résultant d'une translation et définie par: $\tilde{\pi}(\beta) = \frac{1}{n}\sum_{i=1}^{n} \frac{y_i - \beta}{N z_i} + \beta$.

Ces estimateurs sont sans biais pour chaque β fixé et $\hat{\pi}_m$ et $\hat{\pi}_d$ sont des cas spéciaux, avec β égal à 0 et 1, respectivement.

On peut également définir un estimateur par le quotient, avec X encore comme variable auxiliaire $\hat{\pi}_q = \frac{\hat{\pi}_m}{\hat{\pi}_x} = \frac{\frac{1}{n}\sum_{i=1}^{n} \frac{y_i}{Nz_i}}{\frac{1}{n}\sum_{i=1}^{n} \frac{1}{Nz_i}} = \frac{\sum_{i=1}^{n} y_i / z_i}{\sum_{i=1}^{n} 1/z_i}$.

Pour ce qui est de l'estimateur par la régression, on va simplifier la notation en posant $u_i = \frac{y_i}{N z_i}$; $v_i = \frac{1}{N z_i}$ avec $i = 1,\ldots, n$. Alors $\bar{u} = \frac{1}{n}\sum_{i=1}^{n} u_i$; $\bar{v} = \frac{1}{n}\sum_{i=1}^{n} v_i$.

Ainsi, $\hat{\pi}_m = \bar{u}$; $\hat{\pi}_d = \bar{u} + (1 - \bar{v})$ et $\hat{\pi}_q = \frac{\bar{u}}{\bar{v}}$

Ce qui montre que $\hat{\pi}_m, \hat{\pi}_d$ et $\hat{\pi}_q$ sont, respectivement, les estimateurs standards par la moyenne, par la différence et par le quotient.

On a $E(\bar{u})=\pi$, $E(\bar{v})=1$

D'où $Var(\hat{\pi}_m)=Var(\bar{u})=\frac{1}{n}Var(u_i)=\frac{\sigma_u^2}{n}$ avec :

$$\sigma_u^2=\sum_{i=1}^{N}(u_i-E(u_i))^2 z_i=\sum_{i=1}^{N}\left(\frac{y_i}{N z_i}-\pi\right)^2 z_i=\sum_{i=1}^{N}z_i u_i^2-(E(u_i))^2=\frac{1}{N}\sum_{i=1}^{N}\frac{y_i}{N z_i}-\pi^2$$

De même ;

$$\sigma_v^2=\sum_{i=1}^{N}(v_i-E(v_i))^2 z_i=\sum_{i=1}^{N}\left(\frac{1}{N z_i}-1\right)^2 z_i=\sum_{i=1}^{N}z_i v_i^2-(E(v_i))^2=\frac{1}{N}\sum_{i=1}^{N}\frac{1}{N z_i}-1$$

$$\sigma_{uv}=\sum_{i=1}^{N}(u_i-E(u_i))(v_i-E(v_i))z_i=\sum_{i=1}^{N}(\frac{y_i}{N z_i}-\pi)(\frac{1}{N z_i}-1)z_i=\frac{1}{N}\sum_{i=1}^{N}\frac{y_i}{N z_i}-\pi$$

Et $V(\hat{\pi}_d)=V(\bar{u}+(1-\bar{v}))=V(\bar{u})+V(\bar{v})-2Cov(\bar{u},\bar{v})=\frac{1}{n}(\sigma_u^2+\sigma_v^2-2\sigma_{uv})$

Pour la variance de $\hat{\pi}_q$, Sarndal, Swensson et Wretman (1992) l'ont approchée par $V(\hat{\pi}_q)\approx\frac{1}{n}(\sigma_u^2+\pi^2\sigma_v^2-2\pi\sigma_{uv})$.

L'estimateur par la régression avec un β fixe est $\tilde{\pi}(\beta)=\bar{u}+\beta(1-\bar{v})$. Sa variance est minimisée lorsque $\beta=\beta_0=\frac{\sigma_{uv}}{\sigma_v^2}$ et $V[\tilde{\pi}(\beta_0)]=\frac{\sigma_u^2}{n}(1-\rho^2)$ où $\rho=\frac{\sigma_{uv}}{\sigma_u\sigma_v}$.

Un estimateur possible de β_0 est $\hat{\beta}_r = \frac{s_{uv}}{s_v^2}$ où : $s_{uv} = \frac{\sum_{i=1}^{n}(u_i - \bar{u})(v_i - \bar{v})}{n-1}$ et $s_v^2 = \frac{\sum_{i=1}^{n}(v_i - \bar{v})^2}{n-1}$ donc $\hat{\beta}_r = \frac{\sum_{i=1}^{n}(y_i / N^2 z_i^2) - (1/n)(\sum_{i=1}^{n} y_i / N z_i)(\sum_{i=1}^{n} 1/N z_i)}{(\sum_{i=1}^{n} \frac{1}{N^2 z_i^2}) - (1/n)(\sum_{i=1}^{n} 1/N z_i)^2}$

L'estimateur $\hat{\pi}_r$ est alors défini par $\hat{\pi}_r = \bar{u} + \hat{\beta}_r(1 - \bar{v})$.

Estimateur avec décalage : Stuart (1986) a proposé un estimateur général qui regroupe tous ceux que nous avons présentés : $\tilde{\pi}(\beta) = \frac{1}{n}\sum_{i=1}^{n}\frac{(y_i - \beta)}{N z_i} + \beta$.

On a alors que $\beta = 0$ donne $\hat{\pi}_m$; $\beta = 1$ donne $\hat{\pi}_d$. Mais, en général, $\tilde{\pi}(\beta)$ est notre estimateur par la régression. Ce que propose Stuart c'est d'estimer la valeur optimale de β par $\hat{\beta}_s = \frac{\sum_{i=1}^{n}\frac{y_i}{N z_i}\left(\frac{1}{N z_i} - 1\right)}{\frac{1}{N}\sum_{i=1}^{n}\frac{1}{N z_i} - 1}$

L'estimateur de Stuart alors devient : $\hat{\pi}_s = \frac{1}{n}\sum_{i=1}^{n}\frac{(y_i - \hat{\beta}_s)}{N z_i} + \hat{\beta}_s = \pi_m + \beta_s(1 - \bar{v})$

2.2 Détermination des estimateurs

Dans cette section, on a mis les estimateurs sous une forme qui montre que $\hat{\pi}_m$ est l'estimateur standard de la moyenne, $\hat{\pi}_d$ est l'estimateur par la différence, $\hat{\pi}_q$ est l'estimateur par quotient et $\hat{\pi}_r, \hat{\pi}_s$ sont deux versions de l'estimateur par la régression.

Stuart (1986) a montré que la valeur du paramètre de translation β qui minimise la variance, est aussi le coefficient de régression entre u et v se définit comme suit :

$$\beta_s = \frac{\sigma_{uv}}{\sigma_v^2} = \frac{\sum_{i=1}^{N} (\frac{y_i}{N z_i} - \pi)(\frac{1}{N z_i} - 1) z_i}{\sum_{i=1}^{N} (\frac{1}{N z_i} - 1)^2 z_i} = \frac{\frac{1}{N}\sum_{i=1}^{N} \frac{y_i}{N z_i} - \pi}{\frac{1}{N}\sum_{i=1}^{N} \frac{1}{N z_i} - 1} = \frac{\sum_{i=1}^{N} \frac{y_i}{z_i} - N^2\pi}{\sum_{i=1}^{N} \frac{1}{z_i} - N^2}$$

On sait que l'estimateur par la régression $\hat{\pi}_r$ doit être asymptotiquement le meilleur des estimateurs.

En effet, $V(\hat{\pi}_r) \cong V(\tilde{\pi}(\beta_0)) = \frac{\sigma_u^2}{n}(1-\rho^2) = V(\hat{\pi}_m)(1-\rho^2) \le V(\hat{\pi}_m)$, donc pour n grand, $\hat{\pi}_r$ est meilleur que $\hat{\pi}_m$.

Utilisant les mêmes arguments avec $y_i' = 1 - y_i$,on a:

$V(\hat{\pi}_r) \cong \frac{\sigma_u^2}{n}(1-\rho^2) = V(\hat{\pi}_d)(1-\rho^2) \le V(\hat{\pi}_d)$ où $v_i' = \frac{y_i'}{N z_i}$.

En utilisant les expressions $V(\hat{\pi}_q) = \frac{1}{n}\left(\sigma_u^2 + \pi^2\sigma_v^2 - 2\pi\sigma_{uv}\right)$ et $V(\tilde{\pi}(\beta)) = \frac{1}{n}\left(\sigma_u^2 + \beta^2\sigma_v^2 - 2\beta\sigma_{uv}\right)$

Pour l'approximation des variances, on a également : $V(\hat{\pi}_r) \le V(\hat{\pi}_q)$.

En outre, il est facile de voir que : $V(\hat{\pi}_m) \le V(\hat{\pi}_q) \le V(\hat{\pi}_d)$.

L'estimateur de Stuart $\hat{\pi}_s$ est analytiquement plus difficile à évaluer et l'expression de sa variance est aussi compliquée.

Ce qui distingue $\hat{\beta}_s$ de l'estimateur du coefficient de régression habituel $\hat{\beta}_r$,c'est le fait que dans ce dernier, on remplace tous les paramètres par leur estimation, alors que dans

l'estimateur de Stuart, on estime seulement les paramètres qui sont inconnus dans l'expression du paramètre de translation optimal β_s .

Donc, Stuart estime $\beta_s = \frac{\sigma_{uv}}{\sigma_v^2} = \frac{E(uv) - E(u)E(v)}{\sigma_v^2}$ par $\frac{\left(\frac{1}{n}\sum_{i=1}^{n} u_i v_i\right) - \bar{u}}{\sigma_v}$.

Il s'avère, comme c'est souvent le cas, que les estimateurs sont plus efficaces quand on remplace les paramètres par leurs estimés.

Dans ce qui suit, nous développons les propriétés de la méthode de décalage de Stuart.

2.3 Méthode de décalage de Stuart (1986)

On va expliquer amplement la notion de décalage, tout en justifiant les formules déjà présentées. Cette méthode est utilisée dans le cas d'échantillonnage à probabilités inégales avec remise où se pose le problème de choix de probabilités. La moyenne $\bar{y}_U$ peut être estimée plus efficacement en décalant (faire une translation) toutes les unités de la population, $y' = y - \lambda$ où λ est une constante appelée paramètre de translation.

D'une population de taille N , on tire avec remise un échantillon de taille n . Chaque unité i est tirée avec probabilité z_i où $\sum_{i=1}^{N} z_i = 1$.

D'après Cochran (1977), il existe un estimateur sans biais de la moyenne $\bar{y}_U$ de la population et il se définit comme suit: $\hat{\bar{Y}} = \frac{1}{n}\sum_{i=1}^{n} \frac{y_i}{Nz_i}$ avec une variance égale à $\hat{V}(\hat{\bar{Y}}) = \frac{1}{2N^2 n}\sum_{i=1}^{N}\sum_{j=1}^{N} z_i z_j \left(\frac{y_i}{z_i} - \frac{y_j}{z_j}\right)^2$ et estimée par: $\hat{V}(\hat{\bar{Y}}) = \frac{1}{N^2 n(n-1)}\sum_{i=1}^{n}\left(\frac{y_i}{z_i} - \frac{1}{n}\sum_{j=1}^{n}\frac{y_j}{z_j}\right)^2$.

Dans le cas du tirage aléatoire simple où toutes les probabilités z_i sont égales à $\frac{1}{N}$, on a $\overline{Y'} = \hat{\overline{Y}} - \beta$ et $V\left(\overline{Y'}\right) = V\left(\hat{\overline{Y}}\right)$ mais ces résultats ne tiennent plus quand les probabilités sont inégales car si les z_i sont proportionnelles aux y_i, elles ne le sont pas forcément pour les $(y_i - \beta)$ pour tout $\beta \neq 0$, qui peut faire augmenter la variance.

Une conséquence de ceci est qu'on peut chercher la valeur de β qui réduit au minimum la variance $V\left(\overline{Y'}\right)$ où $y' = y - \beta$. On peut écrire $V\left(\overline{Y'}\right)$ comme $V_\beta\left(\hat{\overline{Y}}\right)$.

Une fois que $\overline{y'}$ est estimé par $\overline{Y'}$, il suffit d'ajouter la constante β pour obtenir : $\hat{\overline{Y}} = \overline{Y'} + \beta$ et évidemment, l'ajout de la constante n'affectera pas la variance, ainsi : $V_\lambda\left(\hat{\overline{Y}}\right) = V\left(\overline{Y'}\right)$. De cette façon, on va estimer $\overline{Y}$ par $\hat{\overline{Y}}'(\beta) = \frac{1}{n}\sum_{i=1}^{n}\frac{(y_i - \beta)}{Nz_i} + \beta$.

Cependant, selon cette définition, pour chaque valeur de β on a un estimateur sans biais, par conséquent, une infinité d'estimateurs possibles mais pas tous de même importance et efficacité. Donc, on doit trouver la valeur optimale de β qui minimise la variance de l'estimateur.

Cas particuliers : Dans le cas d'échantillonnage à probabilités égales, où $z_i = \frac{1}{N}$,

$$\hat{\overline{Y}}'(\beta) = \frac{1}{n}\sum_{i=1}^{n}\frac{(y_i - \beta)}{N(1/N)} + \beta = \frac{1}{n}\sum_{i=1}^{n} y_i = \overline{y}_n \text{ indépendante de } \beta .$$

Aussi, si la variable Y est dichotomique, la moyenne $\overline{y}_U$ est une proportion, le choix de la valeur de β est égal à 1.

Si la moyenne $\bar{y}_U = \frac{Y}{N}$ est considérée comme un quotient avec N définit le total d'une variable auxiliaire X qui prend la valeur 1 pour toutes les unités de la population. Alors l'estimateur $\hat{\bar{Y}}$ est donné par la division de l'estimateur $\frac{1}{n}\sum_{i=1}^{n}\frac{y_i}{z_i}$ de Y par N.

Mais N peut être remplacé par son estimateur $\frac{1}{n}\sum_{i=1}^{n}\frac{1}{z_i}$, dans ce cas, on aura un nouvel estimateur (un quotient): $\hat{R} = \frac{\frac{1}{n}\sum_{i=1}^{n}\frac{y_i}{z_i}}{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{z_i}} = \frac{\sum_{i=1}^{n}\frac{y_i}{z_i}}{\sum_{i=1}^{n}\frac{1}{z_i}}$ qui vaut $\hat{R} = \frac{1}{n}\sum_{i=1}^{n} y_i = \bar{y}_n$ pour $z_i = \frac{1}{N}$.

Maintenant, on va déterminer la valeur du paramètre de translation β qui minimise la variance.

2.3.1 La valeur optimale de λ

On considère la famille des estimateurs $\hat{\bar{Y}}(\beta) = \frac{1}{n}\sum_{i=1}^{n}\frac{(y_i - \beta)}{Nz_i} + \beta$.

On va déterminer la valeur de β qui minimise la variance de l'estimateur.

Si on dérive : $V = 2N^2 n V_\beta(\hat{\bar{Y}}) = \sum_{i=1}^{N}\sum_{j=1}^{N} z_i z_j (\frac{y_i - \beta}{z_i} - \frac{y_j - \beta}{z_j})^2$ selon β, on trouve:

La première dérivée : $\frac{\partial V}{\partial \beta} = 2\sum_{i=1}^{N}\sum_{j=1}^{N} z_i z_j (\frac{y_i - \beta}{z_i} - \frac{y_j - \beta}{z_j})(\frac{1}{z_j} - \frac{1}{z_i})$

La deuxième dérivée : $\frac{\partial^2 V}{\partial^2 \beta} = 2\sum_{i=1}^{N}\sum_{j=1}^{N} z_i z_j (\frac{1}{z_j} - \frac{1}{z_i})^2$ est toujours positive.

Donc le minimum de V est la solution de $\frac{\partial V}{\partial \beta} = 0$, ce qui donne :

$$\beta_{op} = \frac{\sum_{i=1}^{N}\sum_{j=1}^{N} z_i z_j (\frac{y_i}{z_i} - \frac{y_j}{z_j})(\frac{1}{z_i} - \frac{1}{z_j})}{\sum_{i=1}^{N}\sum_{j=1}^{N} z_i z_j (\frac{1}{z_i} - \frac{1}{z_j})^2} = \frac{A}{B}$$

Pour simplifier, on sait que $\sum_{i=1}^{N} z_i = 1$ et $\sum_{i=1}^{N} y_i = N\overline{Y}$ donc après calcul, on a :

$A = 2\left(\sum_{i=1}^{N}\frac{y_i}{z_i} - N^2\overline{Y}\right)$ et $B = 2\left(\sum_{i=1}^{N}\frac{1}{z_i} - N^2\right)$ ainsi $\beta_{op} = \frac{\sum_{i=1}^{N}\frac{y_i}{z_i} - N^2\overline{Y}}{\sum_{i=1}^{N}\frac{1}{z_i} - N^2}$.

On remarque que seul le numérateur de β_{op} dépend de la variable y tandis que le numérateur ne dépend que des z_i et N, alors, il est facile de prouver que β_{op} est le coefficient de régression de $\frac{y}{z}$ sur $\frac{1}{z}$, ce qui est équivalent aussi à la régression de

$P = \frac{1}{n}\sum_{i=1}^{n}\frac{y_i}{z_i}$ sur $Q = \frac{1}{n}\sum_{i=1}^{n}(\frac{1}{z_i} - N)$

Preuve :

Le coefficient de régression est égal à $\frac{Cov(P,Q)}{V(Q)}$.

D'abord, on calcule la variance de Q et la covariance entre P et Q :

On a : $E\left(\frac{1}{z_i}\right) = N$ par suite $E(Q) = 0$ donc :

$$Cov(P,Q) = E(PQ) = E\left(\frac{1}{n^2}\sum_{i=1}^{n}\frac{y_i}{z_i}\sum_{j=1}^{n}\frac{1}{z_j} - \frac{N}{n}\sum_{i=1}^{n}\frac{y_i}{z_i}\right)$$

$$= \frac{n-1}{n} E\left(\frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i=1}^{n} \frac{y_i}{z_i z_j} \right) + \frac{1}{n^2} E\left(\sum_{i=1}^{n} \frac{y_i}{z_i^2} \right) - \frac{N}{n} E\left(\sum_{i=1}^{n} \frac{y_i}{z_i} \right)$$

$$= \frac{n-1}{n} \sum_{i=1}^{N} \sum_{j \neq i=1}^{N} y_j + \frac{1}{n} \sum_{i=1}^{N} \frac{y_i}{z_i} - N^2 \overline{Y} = \frac{n-1}{n} N^2 \overline{Y} + \frac{1}{n} \sum_{i=1}^{N} \frac{y_i}{z_i} - N^2 \overline{Y}$$

Alors, $Cov(P,Q) = \frac{1}{n}\left(\sum_{i=1}^{N} \frac{y_i}{z_i} - N^2 \overline{Y} \right)$

Et, $V(Q) = V\left(\frac{1}{n} \sum_{i=1}^{n} \frac{1}{z_i} \right) = \frac{1}{n} V\left(\frac{1}{z_i} \right) = \frac{1}{n}\left(\sum_{i=1}^{N} \frac{1}{z_i} - N^2 \right)$

Finalement, $\beta_{op} = \frac{Cov(P,Q)}{Var(Q)}$ (*) qui est le coefficient de régression entre P et Q .

La proportionnalité des y_i au z_i fait de P une constante, ainsi (*) est égale à 0.

2.3.2 La réduction de la variance

Suite à la translation de toutes les unités de la population, la réduction de la variance peut être exprimée par : $V_{réd} = 2N^2 n\left[V(\hat{\overline{Y}}) - V_{\beta_{op}}(\hat{\overline{Y}}) \right]$.

Après développement des composantes de cette variance, on aura :

$$V_{réd} = 2\sum_{i=1}^{N} \sum_{j=1}^{N} z_i z_j \left(\frac{y_i}{z_i} - \frac{y_j}{z_j}\right)\left(\frac{1}{z_i} - \frac{1}{z_j}\right)\beta_{op} - \sum_{i=1}^{N} \sum_{j=1}^{N} z_i z_j \left(\frac{1}{z_i} - \frac{1}{z_j}\right)^2 \beta_{op}^2$$

D'après la formule de β_{op}, on prouve que $V_{réd} = 2A\beta_{op} - B\beta_{op}^2 = A\beta_{op}$.

Donc, la réduction de la variance est :

$$Réd = V(\hat{\bar{Y}}) - V_{\beta_{op}}(\hat{\bar{Y}}) = \frac{A\beta_{op}}{2N^2 n} = \frac{\left(\sum_{i=1}^{N} \frac{y_i}{z_i} - N^2 \bar{Y}\right)^2}{N^2 n\left(\sum_{i=1}^{N} \frac{1}{z_i} - N^2\right)}$$

Là aussi, seul le numérateur dépend de la variable y .

Bien que l'expression de la valeur optimale du paramètre de translation et celle de la réduction de la variance sont toutes les deux simples avec un dénominateur constant, on ne peut les utiliser que si la population est connue. Par conséquence, on doit chercher leurs estimateurs non biaisés.

On remarque qu'on peut estimer facilement le numérateur de β_{op} par :

$$w = \frac{1}{n}\sum_{i=1}^{n} \frac{y_i}{z_i^2} - \frac{N}{n}\sum_{i=1}^{n} \frac{y_i}{z_i} = \frac{1}{n}\sum_{i=1}^{n} w_i \quad \text{avec} \quad w_i = \frac{y_i}{z_i}\left(\frac{1}{z_i} - N\right) \; ; \; i = 1,\ldots..,\, n$$

Ainsi l'estimateur sans biais de β_{op} est le suivant: $\hat{\beta}_{op} = w \Big/ \sum_{i=1}^{N} \frac{1}{z_i} - N^2$ dont la variance estimée par: $\hat{V}\left(\hat{\beta}_{op}\right) = \hat{V}(w) \Big/ \left(\sum_{i=1}^{N} \frac{1}{z_i} - N^2\right)^2 = \sum_{i=1}^{n} (w_i - w)^2 \Big/ n(n-1)\left(\sum_{i=1}^{N} \frac{1}{z_i} - N^2\right)^2$.

Pour ce qui est de la réduction de la variance, son dénominateur est w^2 où w est la moyenne de n variables w_i indépendantes et identiquement distribuées (*iid*). Ainsi, l'estimateur non biaisé de la variance de w est donné par : $\hat{V}(w) = \frac{\sum_{i=1}^{n} (w_i - w)^2}{n(n-1)}$.

Sachant que $V(w) = E(w^2) - [E(w)]^2$ d'où $w^2 = E(w^2) - V(w)$ estime son biais par $\hat{w}^2 = w^2 - \hat{V}(w)$ et selon la formule de w et $\hat{V}(w)$, on a le résultat :

$$\hat{w}^2 = \left(\frac{1}{n}\sum_{i=1}^{n} w_i\right)^2 - \frac{1}{n(n-1)}\left\{\sum_{i=1}^{n} w_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} w_i\right)^2\right\} = \frac{1}{n(n-1)}\left\{\left(\sum_{i=1}^{n} w_i\right)^2 - \sum_{i=1}^{n} w_i^2\right\}.$$

Donc, l'estimateur de la réduction de la variance est défini par :

$$\hat{R}\acute{e}d = \left\{ \left(\sum_{i=1}^{n} w_i \right)^2 - \sum_{i=1}^{n} w_i^2 \right\} \Big/ N^2 n^2 (n-1) \left(\sum_{i=1}^{N} \frac{1}{z_i} - N^2 \right)$$

Par suite, on peut calculer facilement la proportion de la réduction de la variance.

On a : $2nN^2 V(\hat{\bar{Y}}) = \sum_{i=1}^{N} \sum_{j=1}^{N} z_i z_j \left(\frac{y_i}{z_i} - \frac{y_j}{z_j} \right)^2 = 2 \left(\sum_{i=1}^{N} \frac{y_i^2}{z_i} - N^2 \bar{Y}^2 \right)$

Selon la définition de la réduction, sa proportion est:

$$P = \frac{R\acute{e}d}{V(\hat{\bar{Y}})} = \frac{\left(\sum_{i=1}^{N} \frac{y_i}{z_i} - N^2 \bar{Y} \right)^2}{\left(\sum_{i=1}^{N} \frac{1}{z_i} - N^2 \right) \left(\sum_{i=1}^{N} \frac{y_i^2}{z_i} - N^2 \bar{Y}^2 \right)}.$$

Mais, on sait que $\beta_{op} = \frac{Cov(P,Q)}{V(Q)} = \frac{A}{B}$ et $\rho^2(P,Q) = \frac{Cov^2(P,Q)}{V(P)V(Q)} = \beta_{op} \frac{Cov(P,Q)}{V(P)}$.

Ainsi $\rho^2(P,Q) = \frac{A \beta_{op}}{nV(P)}$ et $P = \frac{R\acute{e}d}{V(\hat{\bar{Y}})} = \frac{A^2 / B N^2 n}{V(\hat{\bar{Y}})} = \frac{A \beta_{op}}{N^2 n V(\hat{\bar{Y}})} = \frac{A \beta_{op}}{nV(P)}$

Finalement : $P = \rho^2(P,Q)$.

Remarque : La proportion doit satisfaire $0 \leq P \leq 1$

Il est évident que la proportion est supérieure ou égale à zéro puisque c'est une division des sommes de carrés.

Soit : $p_i = \sqrt{z_i} \left(\frac{y_i}{z_i} - N \bar{Y} \right)$ et $q_i = \sqrt{z_i} \left(\frac{1}{z_i} - N \right)$

On a : $\sum_{i=1}^{N} p_i^2 = \sum_{i=1}^{N} \frac{y_i^2}{z_i} - N^2 \overline{Y}^2$; $\sum_{i=1}^{N} q_i^2 = \sum_{i=1}^{N} \frac{1}{z_i} - N^2$ et $\sum_{i=1}^{N} p_i q_i = \sum_{i=1}^{N} \frac{y_i}{z_i} - N^2 \overline{Y}$

Selon l'inégalité de Cauchy-Schwartz, on a : $\left(\sum_{i=1}^{N} p_i q_i \right)^2 \leq \left(\sum_{i=1}^{N} p_i^2 \right) \left(\sum_{i=1}^{N} q_i^2 \right)$

D'où l'inégalité : $\left(\sum_{i=1}^{N} \frac{y_i}{z_i} - N^2 \overline{Y} \right)^2 \leq \left(\sum_{i=1}^{N} \frac{y_i^2}{z_i} - N^2 \overline{Y}^2 \right) \left(\sum_{i=1}^{N} \frac{1}{z_i} - N^2 \right)$

L'égalité est atteinte $(p=1)$ si et seulement si: $y_i - \chi_{op} = a\, z_i$, proportionnalité des $(y_i - \beta_{op})$ par rapport aux z_i dont la constante de proportionnalité est obtenue par sommation $a=N(\overline{Y}-\beta_{op})$, d'où $P=1$ si et seulement si $y_i - \beta_{op} = N(\overline{Y}-\beta_{op}) z_i$ .Cela est équivalent à écrire $\frac{y_i}{z_i} = \beta_{op} \frac{1}{z_i} + N(\overline{Y}-\beta_{op})$ ce qui explique notre raison de dire que β_{op} est le coefficient de régression de $\frac{y}{z}$ sur $\frac{1}{z}$.

Exemple numérique : on va prendre des exemples simples pour mieux illustrer le rôle du paramètre de translation optimal.

Soit une population de taille $N = 8$, $\{y_1, y_2, ..., y_8\}$, et $\{z_1, z_2, ..., z_8\}$ les probabilités correspondantes aux valeurs $\{y_i\}_{i=1,....,8}$. On va étudier trois cas selon le type de corrélation entre les $\{z_i\}$ et les $\{y_i\}$. Pour plus d'exactitude, β_{op} sera calculée sur les données de toute la population.

1er cas : les probabilités sont corrélées positivement aux unités de la population (tableau 2.1)

On a : $V(\hat{\overline{Y}}) = \frac{1}{N^2} \left(\sum_{i=1}^{N} \frac{y_i^2}{z_i} - N^2 \overline{Y}^2 \right) = \frac{1}{64} \left(1\,081\,347 - 64 * 123^2 \right) = 1798$

$$\beta_{op} = \frac{\sum_{i=1}^{N} \frac{y_i}{z_i} - N^2 \overline{Y}}{\sum_{i=1}^{N} \frac{1}{z_i} - N^2} = \frac{6034 - 64 * 1253}{95 - 64} = -59$$

$$P = \frac{\left(\sum_{i=1}^{N} \frac{y_i}{z_i} - N^2 \overline{Y} \right)^2}{\left(\sum_{i=1}^{N} \frac{1}{z_i} - N^2 \right) \left(\sum_{i=1}^{N} \frac{y_i^2}{z_i} - N^2 \overline{Y}^2 \right)} = \frac{(6034 - 64 * 123)^2}{(95 - 64)(1\,081\,347 - 64 * 123^2)} = 0.9461 = 95\%$$

La proportion de la réduction de la variance suite à la translation des unités de la population est 95%, ce qui signifie que la translation a permis d'enlever la majorité de la variance.

On calcule la variance de la variable étudiée après translation en utilisant les deux dernières colonnes, donc $V(\hat{\overline{Y}}') = \frac{1}{N^2} \left(\sum_{i=1}^{N} \frac{y_i'^2}{z_i} - N^2 \overline{Y}'^2 \right) = \frac{1}{64} (2134590 - 64 * 182^2) = 97$, ainsi le quotient des deux variances est $\frac{V(\hat{\overline{Y}}')}{V(\hat{\overline{Y}})} = \frac{97}{1798} = 0.05$ qui est le complément de la proportion 0.95.

2^e^ cas : les probabilités sont corrélées négativement aux unités de la population (tableau 2.2)

Comme dans le premier cas, on va calculer le paramètre de translation, la variance et sa proportion de réduction.

$$V(\hat{\overline{Y}}) = 64\,905 \; ; \; \beta_{op} = 329 \; ; \; P = 85\% \; ; \; V(\hat{\overline{Y}}') = 9756 \; ; \; \frac{V(\hat{\overline{Y}}')}{V(\hat{\overline{Y}})} = 0.15 \cdot$$

On remarque que la variance est plus grande, ce qui est dû au fait que les plus grandes probabilités sont associées aux plus petites valeurs des $\{y_i\}$. Aussi, la valeur du paramètre de translation est grande, par conséquence, les valeurs des $\{y_i'\}$ sont négatives avec une variance plus améliorée et la proportion de la réduction de la variance reste assez importante 85%.

3ème cas : une faible corrélation entre les probabilités et les unités de la population (tableau 2.3)

C'est le cas d'une faible corrélation entre les $\{z_i\}$ et les $\{y_i\}$.

De même, on a :

$$V(\hat{\bar{Y}}) = 41387 \ ; \ \beta_{op} = 150 \ ; \ P = 26\% \ ; \ V(\hat{\bar{Y}}') = 30429 ; \ \frac{V(\hat{\bar{Y}}')}{V(\hat{\bar{Y}})} = 0.74 .$$

La valeur du paramètre de translation a diminué tout en rendant la majorité des valeurs de $\{y_i'\}$ négatives mais on a toujours une réduction de la variance de 26%.

Pour récapituler, l'importance de décalage dans ces trois cas peut être comprise du fait que si la corrélation entre les $\{z_i\}$ et les $\{y_i\}$ est élevée, cela signifie qu'il y a presque une relation linéaire entre eux.

Dans le tableau 2.1, la corrélation est élevée, positive et les $\{z_i\}$ sont dans le même ordre que les $\{y_i\}$ ainsi, seulement un petit décalage est nécessaire pour obtenir aussi étroitement la proportionnalité possible. Si le décalage était sensiblement plus grand, il y aurait des valeurs négatives des $\{y_i'\}$ rendant la proportionnalité plus importante.

Dans le tableau 2.2, les $\{z_i\}$ sont dans l'ordre inverse aux $\{y_i\}$ ainsi, la seule manière d'approcher la proportionnalité c'est de faire un décalage assez grand pour que les valeurs des $\{y_i'\}$ soient négatives de telle sorte qu'on puisse avoir une constante négative de la proportionnalité.

Dans le tableau 2.3, avec une corrélation entre les $\{z_i\}$ et les $\{y_i\}$ près de zéro, le paramètre de décalage ou de translation a une valeur près de la moyenne $\bar{Y}$. Ceci résulte du fait que si les $\{y_i\}$ et les $\left\{\frac{1}{z_i}\right\}$ sont non corrélatifs, la formule du paramètre de translation optimal se réduit à $\beta_{op} = \bar{Y}$ et si les $\{y_i\}$, $\{z_i\}$ sont indépendamment déterminés, alors $\{y\}$ ne

sera corrélé ni avec $\{z\}$ ni avec $\left\{\frac{1}{z}\right\}$. Même dans ce cas de faible corrélation, on obtient toujours une réduction de 26% de la variance parce que $\left\{\frac{y}{z}\right\}$ et $\left\{\frac{1}{z}\right\}$ ont une corrélation positive modérée, principalement parce que la plus grande valeur de $\left\{\frac{y}{z}\right\}$ est si extrême.

CHAPITRE III

ESTIMATEURS PAR LE QUOTIENT LORSQUE Y ET X SONT DICHOTOMIQUES

Dans ce chapitre, on va traiter le cas particulier des estimateurs usuels quand les variables sont dichotomiques. On peut le schématiser comme suit :

	Paramètres	
	X=1	X=0
Y=1	P_1	P_2
Y=0	P_3	P_4

On veut estimer le nombre d'observations pour lesquelles $Y=1$, c'est-à-dire P_1+P_2. Pour cette raison, il existe plusieurs estimateurs de ce paramètre, mais il reste à analyser les propriétés de chaque estimateur afin de déduire lequel est le plus fiable et dans quelles circonstances l'un est meilleur que l'autre.

3.1 Les estimateurs

Il y a quatre choix possibles d'estimateurs par le quotient selon qu'on estime π ou $1-\pi$ et selon que la variable auxiliaire est x_i ou $1-x_i$:

$$\hat{R}_1 = \frac{X_1+X_2}{X_1+X_3}(P_1+P_3) = \frac{\bar{y}}{\bar{x}} x_U \qquad ; \quad \hat{R}_2 = 1-\frac{X_3+X_4}{X_1+X_3}(P_1+P_3) = 1-\frac{1-\bar{y}}{\bar{x}} x_U$$

$$\hat{R}_3 = \frac{X_1 + X_2}{X_2 + X_4}(P_2 + P_4) = \frac{\bar{y}}{1-\bar{x}}\left(1-\bar{x}_U\right);\ \hat{R}_4 = 1 - \frac{X_3 + X_4}{X_2 + X_4}(P_2 + P_4) = 1 - \frac{1-\bar{y}}{1-\bar{x}}\left(1-\bar{x}_U\right)$$

Les variables X_1 ,X_2 ,X_3 et X_4 sont les effectifs correspondant aux quatre valeurs possibles du couple (X, Y);

On va développer les propriétés de chacun de ces estimateurs, et pour cette raison, il faut calculer leur biais et leurs variances.

On commence par l'estimateur $\hat{R}_1$. Selon sa structure, on doit le modifier dans le but d'éviter d'avoir un dénominateur nul, ce qui mène à l'estimateur suivant :

$\hat{R}_1 = \frac{X_1 + X_2}{X_1 + X_3 + 1}(P_1 + P_3)$ qui peut s'écrire comme suit: $\hat{R}_1 = \left[\frac{X_1}{X_1 + X_3 + 1} + \frac{X_2}{X_1 + X_3 + 1}\right](P_1 + P_3)$

On va utiliser la méthode des espérances conditionnelles pour calculer l'espérance et la variance de cet estimateur.

Pour simplifier l'écriture des formules, soit $\pi = P_1 + P_3$, π est un paramètre connu, soit $\theta_1 = \frac{P_1}{\pi}$ et $\theta_2 = \frac{P_2}{1-\pi}$. Alors le paramètre à estimer est $P_1 + P_2 = \pi\theta_1 + (1-\pi)\theta_2$.

On pose aussi $h = h(m,\pi) = \frac{1-(1-\pi)^m}{m\pi}$ et $\varphi = \varphi(m,\pi) = \frac{1}{m\pi}\sum_{y=1}^{m}\frac{1}{y}\binom{m}{y}\pi^y(1-\pi)^{m-y}$ avec $Y \sim Bin(m,\pi)$ et $m = n+1$.

$$\varphi = \varphi(m,\pi) = \frac{1}{m\pi}\sum_{y=1}^{m}\frac{1}{y}\binom{m}{y}\pi^y(1-\pi)^{m-y} = E\left(Y^{-1} \middle| Y \geq 1,\ Y \sim B(m,\pi)\right)$$

Par développement de Taylor jusqu'au 3ème terme, on a :

$$\varphi = E\left(Y^{-1} \middle| Y \geq 1,\ Y \sim B(m,\pi)\right) = E\left\{\frac{1}{\mu}\left[1 - \frac{Y-\mu}{\mu} + \frac{(Y-\mu)^2}{\mu^2}\right]\right\} = \frac{1}{\mu} + \frac{\sigma^2}{\mu^3}$$

Avec $\mu=\frac{m\pi}{1-(1-\pi)^m}=\frac{1}{h}$ et $\sigma^2=\frac{m\pi(1-\pi)}{1-(1-\pi)^m}-\frac{(m\pi)^2(1-\pi)^m}{\left[1-(1-\pi)^m\right]^2}=(1-\pi)\mu-(1-\pi)^m\mu^2=\frac{1-\pi}{h}-\frac{(1-\pi)^m}{h^2}$

car :

$$\mu = E(Y|Y\geq 1)=\frac{\sum_{y=1}^{m} y\binom{m}{y}\pi^y(1-\pi)^{m-y}}{P(Y\geq 1)}=\frac{m\pi\sum_{y=1}^{m}\frac{(m-1)!}{(y-1)!(m-y)!}\pi^{y-1}(1-\pi)^{m-y}}{1-(1-\pi)^m}$$

$$=\frac{m\pi}{1-(1-\pi)^m}\sum_{z=0}^{n}\binom{n}{z}\pi^z(1-\pi)^{n-z}=\frac{m\pi}{1-(1-\pi)^m}.$$

Et $\sigma^2 = E(Y^2|Y\geq 1)-\left[E(Y|Y\geq 1)\right]^2$

$$\sigma^2 = E(Y^2|Y\geq 1)-\mu^2=\frac{\sum_{y=1}^{m} y^2\binom{m}{y}\pi^y(1-\pi)^{m-y}}{1-(1-\pi)^m}-\mu^2=\frac{\sum_{y=1}^{m} y\frac{m!}{(y-1)!(m-y)!}\pi^y(1-\pi)^{m-y}}{1-(1-\pi)^m}-\mu^2$$

$$\sigma^2 =\frac{1}{1-(1-\pi)^m}\left[\sum_{y=1}^{m}\frac{m!}{(y-2)!(m-y)!}\pi^y(1-\pi)^{m-y}+\sum_{y=1}^{m}\frac{m!}{(y-1)!(m-y)!}\pi^y(1-\pi)^{m-y}\right]-\mu^2$$

$$=\frac{1}{1-(1-\pi)^m}\left[m(m-1)\pi^2\sum_{z=0}^{m-2}\frac{(m-2)!}{z!(m-2-y)!}\pi^z(1-\pi)^{m-2-z}+m\pi\sum_{z=0}^{m-1}\frac{(m-1)!}{z!(m-1-y)!}\pi^z(1-\pi)^{m-1-y}\right]-\mu^2$$

$$=\frac{1}{1-(1-\pi)^m}\left[m(m-1)\pi^2+m\pi\right]-\mu^2=\frac{m^2\pi^2}{1-(1-\pi)^m}-\frac{m\pi^2}{1-(1-\pi)^m}+\mu-\mu^2=\left(1-(1-\pi)^m\right)\mu^2-\pi\mu+\mu-\mu^2$$

$$=(1-\pi)\mu-(1-\pi)^m\mu^2$$

Ce qui donne: $\varphi(m,\pi)\approx h+h^3\left(\frac{1-\pi}{h}-\frac{(1-\pi)^m}{h^2}\right)=h+h^2(1-\pi)-h(1-\pi)^m=h\left(1-(1-\pi)^m\right)+h^2(1-\pi)$

3.1.1 Calcul de l'espérance

On sait que le vecteur $(X_1; X_2; X_3; X_4)$ suit une loi multinomiale de paramètres n et (P_1, P_2, P_3, P_4) ainsi, sous la condition $X_1 + X_3 = t$, la variable aléatoire X_1 suit une loi binomiale de paramètres t et θ_1, $X_1 \sim Bin(t, \theta_1)$ et la variable aléatoire X_2 suit une loi binomiale de paramètres $(n-t)$ et θ_2, $X_2 \sim Bin(n-t, \theta_2)$.

Donc, $E(\hat{R}_1 | X_1 + X_3 = t) = \pi\left[E\left(\frac{X_1}{X_1 + X_3 + 1} | X_1 + X_3 = t \right) + E\left(\frac{X_2}{X_1 + X_3 + 1} | X_1 + X_3 = t \right) \right]$

$$= \pi\left[\frac{t\theta_1}{t+1} + \frac{(n-t)\theta_2}{t+1} \right] = \pi\left(\frac{t}{t+1}(\theta_1 - \theta_2) + \frac{n}{t+1}\theta_2 \right)$$

$$= \pi\left((\theta_1 - \theta_2) + \frac{(n+1)\theta_2 - \theta_1}{t+1} \right)$$

Pour compléter le calcul, on va calculer l'espérance de $\frac{1}{t+1}$ et $\frac{t}{t+1}$ lorsque t suit une binomiale de paramètres $(m-1)$ et π, $t \sim Bin((m-1), \pi)$.

Propriétés : Si X est de loi binomiale de paramètres $(m-1)$ et π , alors :

$$E\left(\frac{1}{X+1} \right) = \sum_{i=0}^{m-1} \frac{1}{i+1} \binom{m-1}{i} \pi^i (1-\pi)^{m-1-i}$$

$$= \sum_{i=0}^{m-1} \frac{(m-1)!}{(i+1)i!(m-1-i)!} \pi^i (1-\pi)^{m-1-i} = \sum_{i=0}^{m-1} \frac{(m-1)!}{(i+1)!(m-1-i)!} \pi^i (1-\pi)^{m-1-i}$$

On fait un changement de variable $j = i + 1$ et on aura :

$$= \frac{1}{m\pi} \sum_{j=1}^{m} \frac{m!}{j!(m-j)!} \pi^j (1-\pi)^{m-j} = \frac{1}{m\pi} \sum_{j=0}^{m} \frac{m!}{j!(m-j)!} \pi^j (1-\pi)^{m-j} - \frac{(1-\pi)^m}{m\pi}$$

$$= \frac{1}{m\pi} - \frac{(1-\pi)^m}{m\pi} = \frac{1-(1-\pi)^m}{m\pi} = h(m; \pi) .$$

Et $E\left(\frac{X}{X+1}\right)=E\left(1-\frac{1}{X+1}\right)=1-E\left(\frac{1}{X+1}\right)=1-\frac{1-(1-\pi)^m}{m\pi}=1-h(m;\pi)$.

Donc, $E\left(\frac{1}{t+1}\right)=h(m;\pi)$ et $E\left(\frac{t}{t+1}\right)=1-h(m;\pi)$.

Finalement $E\left(E\left(\hat{R}_1|X_1+X_3=t\right)\right)=\pi\left[(\theta_1-\theta_2)(1-h(m;\pi))+nh(m;\pi)\theta_2\right]$.

Pour $m=n+1$ assez grand, on remarque que :

$(1-\pi)^m$ et $\frac{1}{m}$ tendent vers 0 et $\frac{n}{m}$ tend vers 1.

Par conséquent $h(m;\pi)$ tend vers 0 et $nh(m;\pi)$ tend vers $\frac{1}{\pi}$.

Donc, $E\left(E\left(\hat{R}_1|X_1+X_3=t\right)\right)=\pi\theta_1+(1-\pi)\theta_2$, ce qui signifie que l'estimateur $\hat{R}_1$ est presque sans biais.

3.1.2 Calcul de la variance

La variance de l'estimateur modifié est définie par :

$$Var\left(\hat{R}_1\right)=E\left[Var\left(\hat{R}_1|X_1+X_3=t\right)\right]+Var\left[E\left(\hat{R}_1|X_1+X_3=t\right)\right]$$

$$Var\left(\hat{R}_1|X_1+X_3=t\right)=\pi^2 Var\left[\left(\frac{X_1}{X_1+X_3+1}|X_1+X_3=t\right)+\left(\frac{X_2}{X_1+X_3+1}|X_1+X_3=t\right)\right]$$

$$=\pi^2\left[Var\left(\frac{X_1}{X_1+X_3+1}|X_1+X_3=t\right)+Var\left(\frac{X_2}{X_1+X_3+1}|X_1+X_3=t\right)+2Cov\left(\frac{X_1}{X_1+X_3+1},\frac{X_2}{X_1+X_3+1}\right)|X_1+X_3=t\right]$$

On va calculer chacun des termes de cette expression :

$$Var\left(\frac{X_1}{X_1+X_3+1}\middle|X_1+X_3=t\right)=\frac{t}{(t+1)^2}\theta_1(1-\theta_1)$$

$$Var\left(\frac{X_2}{X_1+X_3+1}\middle|X_1+X_3=t\right)=\frac{(n-t)}{(t+1)^2}\theta_2(1-\theta_2)=\frac{1}{(t+1)^2}n\theta_2(1-\theta_2)-\frac{t}{(t+1)^2}\theta_2(1-\theta_2)$$

Et la covariance conditionnelle entre les deux variables est nulle car X_1 et X_2 sont conditionnellement indépendants.

Donc :

$$Var\left(\hat{R}_1|X_1+X_3=t\right)=\pi^2\left[\frac{t}{(t+1)^2}[\theta_1(1-\theta_1)-\theta_2(1-\theta_2)]+\frac{n\theta_2(1-\theta_2)}{(t+1)^2}\right]$$

$$=\pi^2\left[\frac{1}{t+1}[\theta_1(1-\theta_1)-\theta_2(1-\theta_2)]+\frac{(n+1)\theta_2(1-\theta_2)-\theta_1(1-\theta_1)}{(t+1)^2}\right]$$

Et

$$E\left[Var\left(\hat{R}_1|X_1+X_3=t\right)\right]=\pi^2\left\{E\left(\frac{1}{t+1}\right)[\theta_1(1-\theta_1)-\theta_2(1-\theta_2)]+E\left(\frac{1}{(t+1)^2}\right)[(n+1)\theta_2(1-\theta_2)-\theta_1(1-\theta_1)]\right\}$$

Il reste à calculer, $E\left(\frac{1}{(t+1)^2}\right)$ et $E\left(\frac{t}{(t+1)^2}\right)$.

Propriétés : Si X est de loi binomiale de paramètres $(m-1)$ et π, alors :

$$\frac{X}{(X+1)^2}=\frac{1}{X+1}\frac{X}{X+1}=\frac{1}{X+1}\left(1-\frac{1}{X+1}\right)=\frac{1}{X+1}-\frac{1}{(X+1)^2}$$

$$E\left(\frac{1}{(X+1)^2}\right)=\sum_{i=0}^{m-1}\frac{1}{(i+1)^2}\binom{m-1}{i}\pi^i(1-\pi)^{m-1-i}$$

$$= \sum_{i=0}^{m-1} \frac{(m-1)!}{(i+1)(i+1)i!(m-1-i)!} \pi^{i} (1-\pi)^{m-1-i} = \sum_{i=0}^{m-1} \frac{1}{i+1} \frac{(m-1)!}{(i+1)!(m-1-i)!} \pi^{i} (1-\pi)^{m-1-i}$$

On fait un changement de variable $j = i + 1$ et on aura :

$$= \frac{1}{m\pi} \sum_{j=1}^{m} \frac{1}{j} \frac{m!}{j!(m-j)!} \pi^{j} (1-\pi)^{m-j} = \frac{1}{m\pi} \sum_{j=1}^{m} \frac{1}{j} \binom{m}{j} \pi^{i} (1-\pi)^{m-j} = h(m,\pi)\varphi(m,\pi)$$

Donc,

$$E\left[Var\left(\hat{R}_1 | X_1 + X_3 = t\right)\right] = \pi^2 \{h(m;\pi)[\theta_1(1-\theta_1) - \theta_2(1-\theta_2)] + h(m;\pi)\varphi(m;\pi)[(n+1)\theta_2(1-\theta_2) - \theta_1(1-\theta_1)]\}$$

$$= \pi^2 \{[\theta_1(1-\theta_1) - \theta_2(1-\theta_2)]h(m;\pi)(1-\varphi(m;\pi)) + n\theta_2(1-\theta_2)h(m;\pi)\varphi(m;\pi)\}$$

Et $Var\left(E\left(\hat{R}_1 | X_1 + X_3 = t\right)\right) = Var\left[\pi\left((\theta_1 - \theta_2) + \frac{(n+1)\theta_2 - \theta_1}{t+1}\right)\right] = Var\left[\frac{\pi((n+1)\theta_2 - \theta_1)}{t+1}\right]$

$$= [\pi((n+1)\theta_2 - \theta_1)]^2 Var\left[\frac{1}{t+1}\right] = [\pi((n+1)\theta_2 - \theta_1)]^2 \left[E\left(\frac{1}{(t+1)^2}\right) - \left\{E\left(\frac{1}{t+1}\right)\right\}^2\right]$$

$$= [\pi((n+1)\theta_2 - \theta_1)]^2 h(m;\pi)[\varphi(m;\pi) - h(m;\pi)]$$

Finalement, l'expression de la variance est :

$$Var\left(\hat{R}_1\right) = \pi^2 \{[\theta_1(1-\theta_1) - \theta_2(1-\theta_2)]h(m;\pi)(1-\varphi(m;\pi)) + n\theta_2(1-\theta_2)h(m;\pi)\varphi(m;\pi)\}$$

$$+ [\pi((n+1)\theta_2 - \theta_1)]^2 h(m;\pi)[\varphi(m;\pi) - h(m;\pi)]$$

Maintenant, on va simplifier cette expression en éliminant les termes trop petits.

Pour $m = n+1$ assez grand, on remarque que :

h et φ tendent vers 0; nh et $n\varphi$ tendent vers $\frac{1}{\pi}$; $n^2 h\varphi$ et $n^2 h^2$ tendent vers $\frac{1}{\pi^2}$ et $nh\varphi$ tend vers 0.

Alors, la variance s'annule pour n assez grand.

Par symétrie, les estimateurs $\hat{R}_2$; $\hat{R}_3$ et $\hat{R}_4$ ont les mêmes propriétés que $\hat{R}_1$ lorsque les valeurs des probabilités P_i sont permutées :

Sans refaire le calcul, on déduira que :

$$E\left(\hat{R}_2 \middle| X_1 + X_3 = t\right) = 1 - \pi\left([(1-\theta_1) - (1-\theta_2)] + \frac{(n+1)(1-\theta_2) - (1-\theta_1)}{t+1} \right)$$

$$E\left(E\left(\hat{R}_2 \middle| X_1 + X_3 = t\right)\right) = 1 - \pi\left[[(1-\theta_1) - (1-\theta_2)](1 - h(m;\pi)) + nh(m;\pi)(1-\theta_2) \right]$$

$$Var\left(\hat{R}_2\right) = \pi^2 \{ [\theta_1(1-\theta_1) - \theta_2(1-\theta_2)]h(m;\pi)(1-\varphi(m;\pi)) + n\theta_2(1-\theta_2)h(m;\pi)\varphi(m;\pi) \}$$

$$+ \left[\pi\left((n+1)(1-\theta_2) - (1-\theta_1)\right)\right]^2 h(m;\pi)\left[\varphi(m;\pi) - h(m;\pi)\right]$$

Pour les estimateurs $\hat{R}_3$ et $\hat{R}_4$, on note que $t \sim Bin((m-1), 1-\pi)$ ce qui donne :

$$h = H(m, 1-\pi) = \frac{1-\pi^m}{m(1-\pi)} \quad \text{et} \quad \varphi = E\left(Y^{-1} \middle| Y \geq 1, Y \sim B(m, 1-\pi)\right) = E\left\{ \frac{1}{\mu}\left[1 - \frac{Y-\mu}{\mu} + \frac{(Y-\mu)^2}{\mu^2} \right] \right\} = \frac{1}{\mu} + \frac{\sigma^2}{\mu^3}$$

Avec $\mu = \frac{m(1-\pi)}{1-\pi^m} = \frac{1}{h}$ et $\sigma^2 = \frac{m\pi(1-\pi)}{1-\pi^m} - \frac{(m(1-\pi))^2 \pi^m}{[1-\pi^m]^2} = \frac{\pi}{h} - \frac{\pi^m}{h^2}$

D'où: $\varphi(m,\pi) = h + h^3\left(\frac{\pi}{h} - \frac{\pi^m}{h^2}\right) = h + h^2\pi - h\pi^m = h\left(1-\pi^m\right) + h^2\pi$

Sous la condition $X_2 + X_4 = t$, $X_2 \sim Bin(t, \theta_2)$ et $X_1 \sim Bin(n-t, \theta_1)$.

Alors :|

$$E\left(\hat{R}_3 | X_2 + X_4 = t\right) = (1-\pi)\left((\theta_2 - \theta_1) + \frac{(n+1)\theta_1 - \theta_2}{t+1}\right)$$

$$E\left(E\left(\hat{R}_3 | X_2 + X_4 = t\right)\right) = (1-\pi)\left[(\theta_2 - \theta_1)(1 - h(m;1-\pi)) + nh(m;1-\pi)\theta_1\right]$$

$$Var\left(\hat{R}_3\right) = (1-\pi)^2\left\{\left[\theta_2(1-\theta_2) - \theta_1(1-\theta_1)\right]h(m;1-\pi)\left(1-\varphi(m;1-\pi)\right) + n\theta_1(1-\theta_1)h(m;1-\pi)\,\varphi(m;1-\pi)\right\}$$
$$+ \left[(1-\pi)\left((n+1)\theta_1 - \theta_2\right)\right]^2 h(m;1-\pi)\left[\varphi(m;1-\pi) - h(m;1-\pi)\right]$$

Et

$$E\left(\hat{R}_4 | X_2 + X_4 = t\right) = 1 - (1-\pi)\left(\left[(1-\theta_2) - (1-\theta_1)\right] + \frac{(n+1)(1-\theta_1) - (1-\theta_2)}{t+1}\right)$$

$$E\left(E\left(\hat{R}_4 | X_2 + X_4 = t\right)\right) = 1 - (1-\pi)\left[\left[(1-\theta_2) - (1-\theta_1)\right](1 - h(m;1-\pi)) + nh(m;1-\pi)(1-\theta_1)\right]$$

$$Var\left(\hat{R}_4\right) = (1-\pi)^2\left\{\left[\theta_2(1-\theta_2) - \theta_1(1-\theta_1)\right]h(m;1-\pi)\left(1-\varphi(m;1-\pi)\right) + n\theta_1(1-\theta_1)h(m;1-\pi)\,\varphi(m;1-\pi)\right\}$$
$$+ \left[(1-\pi)\left((n+1)(1-\theta_1) - (1-\theta_2)\right)\right]^2 h(m;1-\pi)\left[\varphi(m;1-\pi) - h(m;1-\pi)\right]$$

Généralement, en comparant les biais de ces estimateurs, on remarque qu'ils tendent vers zéro au fur et à mesure que la taille de l'échantillon augmente, de même pour le biais de leurs variances. Cependant, étant donné que les expressions des variances sont très complexes, on ne peut pas comparer les quatre estimateurs.

Pour cette raison, on va procéder autrement afin de trouver des formules plus faciles permettant de conclure les circonstances dans lesquelles un estimateur est plus fiable qu'un autre.

On désigne les estimateurs par $\hat{R}_{ij}$ dont les observations sont X_{ij} et les paramètres sont P_{ij}:

		j=1		j=2			
		Y=1		Y=0		Total	
i=1	X=1	X_{11}	P_{11}	X_{12}	P_{12}	$X_{1.}$	$P_{1.}$
i=2	X=0	X_{21}	P_{21}	X_{22}	P_{22}	$X_{2.}$	$P_{2.}$
	Total	$X_{.1}$	$P_{.2}$	$X_{.2}$	$P_{.2}$	n	1

Donc $\hat{R}_{ij} = \frac{X_{.j}}{X_{i.}} P_{i.}$

D'après les formules approximatives utilisées par Cochran pour l'estimation de la variance d'un quotient, on peut écrire : $V\left(\hat{R}_{ij}\right) = P_{i.}^2 V\left(\frac{X_{.j}}{X_{i.}}\right) = P_{i.}^2 V\left(\frac{\overline{Y}}{\overline{X}}\right)$ ainsi :

$$V\left(\hat{R}_{ij}\right) = P_{i.}^2 \left(\sum Y^2 - 2\frac{\sum Y}{\sum X} \sum XY + \frac{(\sum)}{(\sum)} \sum X^2 \right) \Big/ (\sum X)^2$$

$$V\left(\hat{R}_{ij}\right) = P_{i.}^2 \left(P_{.j} - 2\frac{P_{.j}}{P_{i.}} P_{ij} + \frac{P_{.j}^2}{P_{i.}^2} P_{i.} \right) \Big/ P_{i.}^2 = P_{.j} - \frac{P_{.j}}{P_{i.}} \left(2P_{ij} - P_{.j}\right)$$

Donc, la variance de l'estimateur $\hat{R}_{ij}$ est $V\left(\hat{R}_{ij}\right) = P_{.j} - \frac{P_{.j}}{P_i} \left(2P_{ij} - P_{.j}\right)$.

On remarque que si $i' \neq i$ alors $P_{i'j} = P_{.j} - P_{ij}$; $P_{i'.} = 1 - P_{i.}$ et $P_{ij'} = P_{i.} - P_{ij}$; $P_{.j'} = 1 - P_{.j}$.

Alors, on peut calculer les variances des trois autres estimateurs :

$$V\left(\hat{R}_{i'j}\right) = P_{.j} - \frac{P_{.j}}{P_{i'.}}\left(2P_{i'j} - P_{.j}\right) = P_{.j} - \frac{P_{.j}}{1 - P_{i.}}\left(2P_{.j} - 2P_{ij} - P_{.j}\right) = P_{.j} + \frac{P_{.j}}{1 - P_{i.}}\left(2P_{ij} - P_{.j}\right).$$

On compare cette variance à celle de $\hat{R}_{ij}$:

$$V\left(\hat{R}_{i'j}\right) - V\left(\hat{R}_{ij}\right) = P_{.j} + \frac{P_{.j}}{1 - P_{i.}}\left(2P_{ij} - P_{.j}\right) - P_{.j} + \frac{P_{.j}}{P_{i.}}\left(2P_{ij} - P_{.j}\right) = \frac{P_{.j}\left(2P_{ij} - P_{.j}\right)}{P_{i.}\left(1 - P_{i.}\right)}$$

Donc $V\left(\hat{R}_{i'j}\right) \geq V\left(\hat{R}_{ij}\right)$ si et seulement si $2P_{ij} - P_{.j} \geq 0 \Leftrightarrow \frac{P_{ij}}{P_{.j}} \geq \frac{1}{2}$

Aussi, $V\left(\hat{R}_{ij'}\right) = P_{j'} - \frac{P_{j'}}{P_{i.}}\left(2P_{ij'} - P_{.j'}\right) = 1 - P_{.j} - \frac{1 - P_{.j}}{P_{i.}}\left(2P_{i.} - 2P_{ij} - 1 + P_{.j}\right)$

$$= \frac{1 - P_{.j}}{P_{i.}}\left(2P_{ij} - P_{.j} + 1 - P_{i.}\right) = \frac{1}{P_{i.}}\left(2P_{ij} - P_{.j} + 1 - P_{i.}\right) - \frac{P_{.j}}{P_{i.}}\left(2P_{ij} - P_{.j}\right) - \frac{P_{.j}}{P_{i.}}\left(1 - P_{i.}\right)$$

$$= \frac{2\left(P_{ij} - P_{.j}\right) + 1 - P_{i.}}{P_{i.}} + V\left(\hat{R}_{ij}\right)$$

Donc $V\left(\hat{R}_{ij'}\right) \geq V\left(\hat{R}_{ij}\right)$ si et seulement si $2\left(P_{ij} - P_{.j}\right) + 1 - P_{i.} \geq 0 \Leftrightarrow \frac{P_{.j} - P_{ij}}{1 - P_{i.}} \leq \frac{1}{2} \Leftrightarrow \frac{P_{i'j}}{P_{i'.}} \leq \frac{1}{2}$

Et pour finir, on calcule :

$$V\left(\hat{R}_{i'j'}\right)= P_{.j'} - \frac{P_{.j'}}{P_{i'.}}\left(2P_{i'j'} - P_{.j'}\right)= 1 - P_{.j} - \frac{1-P_{.j}}{1-P_{i.}}\left(2P_{.j'} - 2P_{ij'} - 1 + P_{.j}\right)$$

$$= 1 - P_{.j} - \frac{1-P_{.j}}{1-P_{i.}}\left(2 - 2P_{.j} - 2P_{i.} + 2P_{ij} - 1 + P_{.j}\right)$$

$$= 1 - P_{.j} - \frac{1-P_{.j}}{1-P_{i.}}\left(1 - 2P_{i.} + 2P_{ij} - P_{.j}\right)= \frac{1-P_{.j}}{1-P_{i.}}\left(P_{i.} - 2P_{ij} + P_{.j}\right)$$

$$= \frac{1}{1-P_{i.}}\left(P_{i.} - 2P_{ij} + P_{.j}\right)+ \frac{P_{.j}}{1-P_{.j}}\left(2P_{ij} - P_{.j} - P_{i.}\right)= \frac{P_{i.} - 2P_{ij} + P_{.j}}{1-P_{i.}} - \frac{P_{.j}}{1-P_{i.}} + V\left(\hat{R}_{i'j}\right)$$

$$= \frac{P_{i.} - 2P_{ij}}{1-P_{i.}} + V\left(\hat{R}_{i'j}\right)$$

Alors, $V\left(\hat{R}_{i'j'}\right)- V\left(\hat{R}_{ij}\right)= \frac{P_{i.} - 2P_{ij}}{1-P_{i.}} + V\left(\hat{R}_{i'j}\right)- V\left(\hat{R}_{ij}\right)= \frac{P_{i.} - 2P_{ij}}{1-P_{i.}} + \frac{P_{.j}\left(2P_{ij} - P_{.j}\right)}{P_{i.}\left(1-P_{i.}\right)}$

$$= \frac{\left(P_{i.} - P_{.j}\right)\left(P_{i.} + P_{.j}\right)- 2P_{ij}\left(P_{i.} - P_{.j}\right)}{P_{i.}\left(1-P_{i.}\right)} = \frac{\left(P_{i.} - P_{.j}\right)\left(P_{i.} + P_{.j} - 2P_{ij}\right)}{P_{i.}\left(1-P_{i.}\right)} = \frac{\left(P_{i.} - p_{.j}\right)\left(P_{ij'} + P_{i'j}\right)}{P_{i.}\left(1-p_{i.}\right)}$$

Donc $V\left(\hat{R}_{i'j'}\right)\geq V\left(\hat{R}_{ij}\right)$ si et seulement si $P_{i.} - P_{.j} \geq 0 \Leftrightarrow P_{i.} \geq P_{.j}$

En termes des événements, soit A l'événement X=1 et B l'événement Y=1, ainsi, $P(A)=P_{1.}$; $P(B)=P_{.1}$; $P(A\cap B)=P_{11}$; $P(A \text{ sans } B) = P_{12}$; $\sum X^2 = P_{1.}$, $\sum Y^2 = P_{.1}$, $\sum XY = P_{11}$.

Calculons les variances de $\hat{R}_{11}$; $\hat{R}_{12}$; $\hat{R}_{21}$ et $\hat{R}_{22}$:

$$\hat{R}_{11} = \frac{X_{.1}}{X_{1.}} P_1 \quad ; \quad \hat{R}_{12} = \frac{X_{.2}}{X_1} P_{1.} \quad ; \quad \hat{R}_{21} = \frac{X_1}{X_2} P_{2.} \quad ; \quad \hat{R}_{22} = \frac{X_{.2}}{X_2} P_2$$

$$V\left(\hat{R}_{11}\right) = \frac{P(B)}{P(A)}\left[P(A \cup B) - P(A \cap B)\right] = \frac{P(B)}{P(A)}\left[P(A) + P(B) - 2P(A \cap B)\right] = \frac{P_{.1}}{P_1}\left(P_{1.} + P_{.1} - 2P_{11}\right)$$

$$= P_{.1} - \frac{P_{.1}}{P_{1.}}\left(2P_{11} - P_{.1}\right)$$

De même $V\left(\hat{R}_{12}\right) = \frac{P\left(\overline{B}\right)}{P(A)}\left[P(A) + P\left(\overline{B}\right) - 2P\left(A \cap \overline{B}\right)\right]$

$$= \frac{1 - P_{.1}}{P_1}\left(P_1 + 1 - P_1 - 2P_1 + 2P_{11}\right) = \frac{2\left(P_{11} - P_1\right) + 1 - P_1}{P_1} + V\left(\hat{R}_{11}\right)$$

$$V\left(\hat{R}_{21}\right) = \frac{P(B)}{P\left(\overline{A}\right)}\left[P\left(\overline{A}\right) + P(B) - 2P\left(\overline{A} \cap B\right)\right]$$

$$= \frac{P_{.1}}{1 - P_{1.}}\left(1 - P_{1.} + P_1 - 2P_1 + 2P_{11}\right) = P_{.1} + \frac{P_1}{1 - P_{1.}}\left(2P_{11} - P_{.1}\right)$$

$$V\left(\hat{R}_{22}\right) = \frac{P\left(\overline{B}\right)}{P\left(\overline{A}\right)}\left[P\left(\overline{A}\right) + P\left(\overline{B}\right) - 2P\left(\overline{A} \cap \overline{B}\right)\right]$$

$$= \frac{1 - P_{.1}}{1 - P_{1.}}\left(1 - P_{1.} + 1 - P_{.1} - 2 + 2P_{.1} + 2P_{1.} - 2P_{11}\right) = \frac{P_{1.} - 2P_{11}}{1 - P_{1.}} + V\left(\hat{R}_{21}\right)$$

Ces résultats nous permettent de savoir dans quelles circonstances un estimateur est meilleur qu'un autre selon le vecteur des paramètres $P_{ij.}$ comme exemple :

$V(\hat{R}_{21}) \geq V(\hat{R}_{11})$ si et seulement si $\frac{P_{11}}{P_1} \geq \frac{1}{2}$

$V(\hat{R}_{12}) \geq V(\hat{R}_{11})$ si et seulement si $\frac{P_{.1} - P_{11}}{1 - P_1} \leq \frac{1}{2}$

$V(\hat{R}_{22}) \geq V(\hat{R}_{11})$ si et seulement si $P_1 \geq P_1$

CHAPITRE IV

LES SIMULATIONS

L'un des facteurs les plus importants dans une estimation d'un paramètre, soit une moyenne, un total d'une population, c'est la taille de l'échantillon. Plus elle est grande, plus l'échantillon est représentatif. En pratique, les échantillons de petites tailles entraînent toujours de gros biais ainsi que d'importantes erreurs quadratiques moyennes.

Dans ce chapitre, nous présentons les différents résultats de certaines études de simulations afin de prouver l'efficacité des méthodes élaborées précédemment. Pour cette raison, nous avons pris des tailles assez petites par rapport à la taille de la population dans le but de comparer certains estimateurs de la moyenne sous quelques conditions exprimées en terme de paramètres.

Afin d'étudier les estimateurs possibles, nous avons pris huit populations réelles dont les caractéristiques sont différentes et la variable étudiée Y est dichotomique. D'après la structure de ces populations (Tableau 1.1), le nombre total d'observations est identique pour chaque deux populations mais la répartition des observations de la variable d'intérêt Y reste différente. Ce qui influence effectivement celles de la variable auxiliaire et par conséquence, les résultats des estimateurs utilisés pour estimer la moyenne de la population.

Aussi, le coefficient de variation des observations de la variable auxiliaire est petit pour les populations 1 et 2, par contre il est très élevé pour les autres populations. Ceci indique que ces observations sont très dispersées vis-à-vis de leur moyenne, ce qui a un effet direct sur certains estimateurs. Nous avons exclu la population 6 du traitement car selon sa composition, toutes les valeurs de Y=1 correspondent à X=0, ce qui fait que certains

estimateurs ne peuvent pas être utilisés. C'est un cas particulier néanmoins intéressant concernant certaines populations où peu d'estimateurs peuvent être utilisés.

Pour commencer, deux modes de tirage se manifestent, soit avec probabilités égales ou probabilités inégales. Dans les deux cas, plusieurs estimateurs sont possibles. Pour le premier cas, nous avons opté pour cinq estimateurs : l'estimateur usuel par la moyenne, l'estimateur logistique, l'estimateur par le quotient, l'estimateur par la régression et l'estimateur de Hartley-Ross.

En examinant bien le tableau 1.2 qui représente le biais relatif et l'erreur quadratique moyenne des estimateurs pour les populations 1 et 2, les cinq estimateurs donnent des résultats proches de la réalité, leurs biais sont faibles et leurs variances sont à peu près sans biais car les EQM sont proches de zéro. L'estimateur logistique donne toujours de bons résultats, tout comme celui par la régression. Cependant, l'estimateur par le quotient et de Hartley- Ross sont très mauvais pour les populations ayant de grandes valeurs de cœfficient de variation de la variable auxiliaire. Cependant, il ne faut pas négliger le fait que même avec le même coefficient de variation, si on compare simultanément les populations 3 avec 4 et 7 avec 8, le biais de l'estimateur de Hartley- Ross et son EQM sont très divergents.

Aussi, on remarque que d'une part la présence du ratio dans la formulation des deux estimateurs et d'autre part la moyenne de la variable auxiliaire correspondant à Y=1 est très élevée dans les populations 3, 4, 5, 7 et 8 et aussi la faible différence entre les moyennes générales $\overline{X}$ dans toutes les populations. Tout ça, a engendré le fait que le ratio $\frac{y_n}{x_n}$ est très petit, d'où tous ces facteurs influencent les estimations issues du quotient et de Hartley-Ross qui sont très grandes.

Mais la raison la plus importante est due au coefficient de variation des observations de la variable auxiliaire, celui-ci est grand pour toutes les populations sauf 1 et 2.

Dans ce cas, l'estimateur par le quotient et celui de Hartley-Ross sont de bons estimateurs pour un cœfficient de variation très petit et de mauvais estimateurs dans le cas contraire. Donc, dans certaines circonstances, il est préférable de ne pas les utiliser et se

contenter soit de l'estimateur par la moyenne, l'estimateur logistique ou l'estimateur par régression afin d'estimer la moyenne de la population.

Donc, pour récapituler, l'estimation d'une moyenne peut être faite à l'aide de plusieurs estimateurs qui sont fiables et pratiques mais certains d'entre eux exigent une faible dispersion des observations dans le but d'obtenir de bons résultats.

Pour ce qui est du tirage avec probabilités inégales, les estimateurs étudiés sont : l'estimateur par la moyenne, par la différence, par la régression, l'estimateur de Stuart, l'estimateur de Lahiri et l'estimateur de Hartley-Ross. D'après le tableau 1.3, la première remarque qui saute aux yeux est que l'estimateur de Stuart et celui de Hartley-Ross sont loin d'être de bons estimateurs pour les populations 3, 4, 5, 7 et 8.

On note aussi que les estimateurs par la moyenne et par la différence donnent à peu prés les mêmes résultats pour toutes les populations. Ceci est logique, du fait que les deux estimateurs sont comparés parallèlement à cause de leur relation. Cependant pour certaines populations, le biais des estimateurs est assez remarquable.

Ces différences peuvent être dues aux compositions des populations en terme de moyenne de la variable d'intérêt, la grandeur des observations de la variable auxiliaire qui influence les valeurs des probabilités d'inclusion et surtout l'effet de la dispersion de ces observations autour de la moyenne, ce qui veut dire, le coefficient de variation. Ceci se voit clairement dans le tableau, une fois le CV est supérieur à 1, les estimations sont aberrantes.

La question qui se pose est : est ce que la valeur du coefficient de variation influence directement les estimations des différents estimateurs?

Pour répondre à cette question et afin de prouver que le cœfficient de variation a un effet direct sur les estimateurs, nous avons généré des populations avec une variété de valeurs de coefficient de variation et la moyenne de la variable auxiliaire et avec deux tailles d'échantillon 10 et 100. Les résultats sont illustrés dans les tableaux 1.4 et 1.5. En analysant ces résultats, on peut déduire qu'effectivement, le coefficient de variation a un impact direct sur la plupart des estimateurs mais surtout sur les estimateurs de Hartley-Ross et celui de

Stuart qui sont très sensibles vis-à-vis à un petit changement de paramètres. Donc, il faut les éviter dans ces cas, même si l'échantillon est assez grand. Pour l'effet de la moyenne de la population, il est plus au moins négligeable par rapport à l'impact de la dispersion des observations autour de cette moyenne.

Pour résumer, quand la variable d'intérêt est dichotomique, afin d'estimer une moyenne, un total, il y a plusieurs estimateurs à utiliser. Cependant, il faut étudier les propriétés de la population étudiée car chaque estimateur a sa particularité et ses circonstances pour être un bon estimateur.

Finalement, si le l'objectif est d'estimer par le quotient, un paramètre dans le cas où la variable d'intérêt et la variable auxiliaire sont toutes les deux dichotomiques, il existe plusieurs estimateurs. Mais, il reste à savoir lequel est le plus efficace et dans quelles circonstances l'un est meilleur que l'autre. Pour cette raison, nous avons analysé les quatre estimateurs par le quotient, estimant le nombre d'observations pour lesquelles $Y=1$.

Les propriétés de ces estimateurs sont illustrées dans le tableau 3.1, exprimées par le biais relatif et le biais relatif de la variance de chaque estimateur selon la taille de l'échantillon et un vecteur fixe des probabilités P = (0.1,0.3,0.4,0.2). La remarque générale pour les estimateurs est que le biais relatif de l'estimateur et celui de sa variance sont de moins en moins faibles au fur et à mesure que la taille de l'échantillon augmente. Cependant, l'estimateur $\hat{R}_1$ est le plus approprié en terme de biais dégagé et de variance calculée. Pour mieux comprendre, il faut aller plus loin dans le but de vérifier si le choix des probabilités n'a pas un effet sur ces estimateurs.

Dans ce sens, nous avons élaboré des simulations avec des vecteurs de probabilités différentes et de taille d'échantillon variable. D'après les résultats du tableau 3.2, effectivement, le vecteur des probabilités influence les estimateurs. Pour des valeurs très faibles, quelques estimations deviennent très grandes. Ce qui et du logiquement aux expressions de ces quatres estimateurs, c'est-à-dire le calcul de chacun. Nous avons trouvé des expressions pour les variances qui permettent de comparer les estimateurs selon les valeurs des probabilités. Ainsi, pour qu'un estimateur soit meilleur qu'un autre, il suffit que

les probabilités respectent des conditions reliées aux inégalités entre sa variance et celles des autres estimateurs.

Généralement, le sondage offre plusieurs méthodes pour estimer un paramètre donné. Mais, il est très important d'étudier minutieusement les propriétés de chaque estimateur afin de déterminer le meilleur à utiliser car le but de n'importe quel sondage, c'est approximer la valeur réelle du paramètre et par conséquence plusieurs décisions seront basées là-dessus.

Ce qui veut dire que la comparaison des erreurs quadratiques moyennes est importante dans le choix de l'estimateur à utiliser mais l'importance des calculs exigés par les estimateurs eux-mêmes est aussi un facteur du choix. Donc, le choix de l'estimateur dépend en fait de divers arguments concurrents liés à la connaissance du domaine d'étude, des facilités d'enquêtes et de calculs, élément qu'il n'est pas possible de placer dans un cadre théorique général.

CONCLUSION

Dans la théorie des sondages, l'estimation d'un quotient ou d'une moyenne par le quotient occupe une grande place du fait qu'elle constitue une des façons les plus naturelles d'exploiter une information auxiliaire. Elle est par conséquent très efficace dans les situations qui s'y prêtent. Mais en même temps, elle présente un sérieux défi à qui cherche à démontrer ses propriétés. Malgré les très nombreuses recherches faites sur la question, les réponses demeurent partielles et incertaines.

Dans ce présent mémoire, nous avons essayé de limiter l'étude à certains cas particuliers dans l'espoir d'obtenir des résultats plus clairs. En premier lieu, nous avons parcouru tout ce qui concerne l'estimation d'un paramètre à l'aide d'une variable auxiliaire, en examinant les méthodes existantes pour estimer une moyenne à partir d'un échantillon aléatoire. Nous nous sommes intéressés à quelques types d'estimateurs selon le type d'échantillonnage. Ensuite, nous avons traité le cas particulier où la variable d'intérêt est dichotomique et nous avons déterminé les formules des estimateurs dans le cas d'estimation d'une proportion notée par π où la variable auxiliaire est aussi dichotomique. Ces cas qui, quoique particuliers, ont des applications importantes.

D'une part, lorsque les deux variables sont dichotomiques, nous avons développé des formules appropriées correspondant aux divers estimateurs et dans des plans d'expérience variés: tirages avec probabilités égales et tirages avec probabilités proportionnelles. Les résultats sont intéressants dans la mesure où une population peut dans ce cas être caractérisée par un petit nombre de paramètres dont les valeurs peuvent être connues dans un contexte particulier, du moins assez pour permettre un choix judicieux d'estimateur.

D'autre part, lorsque l'une des variables est quantitative, la gamme possible de populations est beaucoup plus large. C'est d'ailleurs ce qui rend les conclusions des diverses

recherches difficiles à utiliser: un même mode de tirage, un même estimateur, peuvent être adéquat dans une population est pas du tout dans une autre. C'est pour cela que nus avons choisi de tester plusieurs méthodes sur des populations réelles.

Tableau 2.1

Paramètre de translation optimal, cas de corrélation positive

	y	z	$\frac{y}{z}$	$\frac{1}{z}$	$\frac{y^2}{z}$	y'	$\frac{y'^2}{z}$
	3	0,04	75	25	225	62	97618
	11	0,05	220	20	2420	70	99370
	35	0,07	500	14	17500	94	127542
	72	0,1	720	10	51840	131	172890
	100	0,1	1000	10	100000	159	254363
	152	0,14	1086	7	165029	211	319479
	240	0,2	1200	5	288000	299	448464
	370	0,3	1233	3	456333	429	614865
Total	**983**	**1**	**6034**	**95**	**1081347**	**1459**	**2134590**
Moyenne	**123**	**-**	**754**	**12**	**135168**	**182**	**266824**

Tableau 2.2

Paramètre de translation optimal, cas de corrélation négative

	y	z	$\frac{y}{z}$	$\frac{1}{z}$	$\frac{y^2}{z}$	y'	$\frac{y'^2}{z}$
	3	0,3	10	3	30	-326	353610
	11	0,2	55	5	605	-318	504679
	35	0,15	233	7	8167	-294	575081
	72	0,1	720	10	51840	-257	658970
	100	0,1	1000	10	100000	-229	523056
	152	0,06	2533	17	385067	-177	520406
	240	0,05	4800	20	1152000	-89	157368
	370	0,04	9250	25	3422500	41	42634
Total	**983**	**1**	**18602**	**97**	**5120208**	**-1647**	**3335804**
Moyenne	**123**	**-**	**2325**	**12**	**640026**	**-206**	**416975**

Tableau 2.3

Paramètre de translation optimal, cas de faible corrélation

	y	z	$\frac{y}{z}$	$\frac{1}{z}$	$\frac{y^2}{z}$	y'	$\frac{y'^2}{z}$
	3	0,05	60	20	180	-147	434586
	11	0,05	220	20	2420	-139	388695
	35	0,15	233	7	8167	-115	88794
	72	0,2	360	5	25920	-78	30740
	100	0,3	333	3	33333	-50	8470
	152	0,1	1520	10	231040	2	25
	240	0,1	2400	10	576000	90	80266
	370	0,05	7400	20	2738000	220	964408
Total	**983**	**1**	**12527**	**95**	**3615060**	**-220**	**1995984**
Moyenne	**123**	**-**	**1566**	**12**	**451883**	**-28**	**249498**

Tableau 3.1

Évolution du biais et de la variance des estimateurs

selon la taille de l'échantillon avec P = (0.1,0.3,0.4,0.2)

	$P_1 = 0.1$		$P_2 = 0.3$		$P_3 = 0.4$		$P_4 = 0.2$	
n	$\hat{R}_1$		$\hat{R}_2$		$\hat{R}_3$		$\hat{R}_4$	
	% Biais	**% B.Var**	**% Biais**	**% B.Var**	**% Biais**	**% B.Var**	**% Biais**	**% B.Var**
10	-4.62	-19.48	18.22	-21.43	-13.65	-24.51	9.18	-27.64
50	-0.98	-3.32	3.92	-4.75	-2.94	-5.77	1.96	-5.55
100	-0.5	-1.61	1.98	-2.41	-1.49	-2.94	0.99	-2.77
150	-0.33	-1.07	1.32	-1.61	-0.99	-1.97	0.66	-1.84
200	-0.25	-0.8	1	-1.21	-0.75	-1.49	0.50	-1.38
300	-0.17	-0.53	0.66	-0.81	-0.50	-0.99	0.33	-0.92

Tableau 3.2

Le biais et EQM des quatre estimateurs selon la taille de l'échantillon et le vecteur des probabilités

(n;p)	$\hat{R}_1$		$\hat{R}_2$		$\hat{R}_3$		$\hat{R}_4$	
	%Biais	EQM	%Biais	EQM	%Biais	EQM	%Biais	EQM
(20;0.4;0.4;0.1;0.1)	-4.76	0.0418	1.19	0.0098	-4.76	0.0418	1.19	0.0098
(100;0.4;0.4;0.1;0.1)	-3.31	0.0080	0.25	0.0019	-0.99	0.0080	0.25	0.0019
(200;0.4;0.4;0.1;0.1)	-0.50	0.0040	0.12	0.0009	-0.50	0.0040	0.12	0.0009
(20;0.9;0.07;0.01;0.02)	-4.86	0.0061	0.05	0.0014	-17.36	0.2360	1.10	0.0008
(100;0.9;0.07;0.01;0.02)	-3.31	0.0080	0.01	0.0003	-0.80	0.1279	0.23	0.0002
(200;0.9;0.07;0.01;0.02)	-0.51	0.0004	0.01	0.0001	-0.40	0.0550	0.11	0.0001
(20;0.3;0.3;0.1;0.3)	-5.95	0.0323	1.99	0.0332	-3.97	0.0300	3.97	0.0129
(100;0.3;0.3;0.1;0.3)	-1.24	0.0061	0.41	0.0061	-0.83	0.0060	0.83	0.0027
(200;0.3;0.3;0.1;0.3)	-0.62	0.0030	0.21	0.0030	-0.41	0.0030	0.41	0.0013
(20;0.2;0.3;0.1;0.4)	-6.39	0.0387	3.24	0.0619	-4.08	0.0208	5.44	0.0139
(100;0.2;0.3;0.1;0.4)	-1.32	0.0068	0.66	0.0103	-0.85	0.0043	1.13	0.0028
(200;0.2;0.3;0.1;0.4)	-0.66	0.0034	0.33	0.0051	-0.43	0.0021	0.57	0.0014
(20;0.3;0.1;0.4;0.2)	-5.1	0.0136	6.8	0.0209	-4.03	0.0399	8.01	0.0599
(100; 0.3;0.1;0.4;0.2)	-1.06	0.0028	1.41	0.0043	-0.83	0.0068	1.65	0.0103
(200;0.3;0.1;0.4;0.2)	-0.53	0.0014	0.71	0.0021	-0.41	0.0034	0.83	0.0051
(20;0.25;0.25;0.25;0.25)	-4.76	0.0253	4.76	0.0253	-4.76	0.0253	4.76	0.0253
(100;0.25;0.25;0.25;0.25)	-0.99	0.0050	0.99	0.0050	-0.99	0.0050	0.99	0.0050
(200;0.25;0.25;0.25;0.25)	-0.50	0.0025	0.50	0.0025	-0.50	0.0025	0.50	0.0025
(20;0.5;0.01;0.4;0.09)	-5.19	0.0112	4.15	0.0152	-12.75	0.0926	17.02	0.0598
(100;0.5;0.01;0.4;0.09)	-0.99	0.0050	0.99	0.0050	-0.20	0.0375	1.75	0.0242
(200;0.5;0.01;0.4;0.09)	-0.54	0.0011	0.43	0.0015	-0.10	0.0165	0.88	0.0108

Tableau 4.1

Les caractéristiques des populations réelles traitées dans les simulations

Pop	Nombre d'observations			Moy (x/y=1)	Moy (x/y=0)	Moy(y)	Moy(x)	Ecart type(x)	Cor(x,y)	CV(x)
	y=0	y=1	Total							
1	3177	1915	5092	786.5504	1143.062	0.3760801	1008.985	1325.5	-0.1303	1,314
2	3173	1919	5092	733.2751	1199.398	0.3768657	1023.733	1358.7	-0.1663	1,327
3	892	8493	9385	13747.38	3952.496	0.9049547	12816.42	219396.2	0.0131	17,118
4	7097	2288	9385	23599.97	9339.916	0.2437933	12816.42	219396.2	0.0279	17,118
5	10321	34792	45113	12473.25	6140.004	0.771219	11024.32	155274.6	0.0171	14,085
6	43586	1527	45113	0	771.4588	0.03384834	745.3462	11042.5	-0.0126	14,815
7	3201	39215	42416	14823.34	5389.686	0.9245332	14111.41	234334.1	0.0106	16,606
8	29269	13147	42416	21753.49	10678.76	0.3099538	14111.41	234334.1	0.0219	16,606

Tableau 4.2

Les estimateurs de la moyenne sous un tirage avec probabilités égales et une taille de l'échantillon n=100

Pop	n=100	Estimateur				
		Moyenne	Logistique	Quotient	Hartley-Ross	Régression
1	%Biais	0.249	0.265	2.352	0.664	-0.036
	EQM	0.002242297	0.002269265	0.005661029	0.019183963	0.002263268
2	%Biais	-0.211	-0.186	1.858	-0.019	-0.583
	EQM	0.002282990	0.002352536	0.005940497	0.022438471	0.002297086
3	%Biais	-0.673	-0.433	197.380	-191.042	-0.437
	EQM	0.0009500974	0.0009038922	8.400586	190494.7	0.002059643
4	%Biais	-0.347	0.141	209.932	-86.809	-5.816
	EQM	0.001809043	0.001838022	0.7465323	33396.01	0.006364403
5	%Biais	-0.548	-0.430	107.146	-195.207	0.530
	EQM	0.001834314	0.001826440	1.819857	28503.20	0.002679729
7	%Biais	-0.540	-0.320	182.657	-609.940	-0.046
	EQM	0.0007059684	0.0006693901	7.323299	215621.4	0.001485169
8	%Biais	-0.090	0.329	191.867	160.966	-6.872
	EQM	0.002120452	0.002143029	0.9221180	47237.39	0.007187363

Tableau 4.3

Les estimateurs de la moyenne sous un tirage avec probabilités inégales

et une taille de l'échantillon n=100

Pop	n=100	Estimateur					
		Moyenne	Différence	Régression	Stuart	Lahiri	Hartley-Ross
1	%Biais	0.152	0.152	0.174	-3.473	0.550	0.172
	EQM	0.007945562	0.007943388	0.005686346	0.005779988	0.005487718	0.019377641
	CV	0.2366581	0.2366255	0.2001615	0.2094287	0.1959005	0.3695092
2	%Biais	-0.089	-0.089	0.084	-3.938	-0.785	0.438
	EQM	0.008509256	0.008506895	0.005969962	0.006200828	0.005340005	0.021698087
	CV	0.2449888	0.2449544	0.2048495	0.2175144	0.1954362	0.3891588
3	%Biais	6.661	6.660	-2.928	-716.837	-4.121	734.103
	EQM	2.128553	2.128092	0.04289178	11308.76	1.516438	170781.5
	CV	1.5115023	1.5113498	0.2357577	-19.0506544	1.4192653	54.7487410
4	%Biais	-0.286	-0.285	-37.736	-1063.735	4.691	1500.165
	EQM	0.8027401	0.8025710	0.06118140	3453.070	0.1372805	26547.47
	CV	3.685613	3.685195	1.629486	-25.010548	1.451687	41.766193
5	%Biais	-6.347	-6.347	-0.619	-304.298	-3.237	-721.260
	EQM	0.6123109	0.6122843	0.04888842	2390.670	0.6212913	31916.27
	CV	1.0833972	1.0833726	0.2884854	-31.0326491	1.0562327	-37.2867742
7	%Biais	0.443	0.443	-0.899	-1251.103	1.883	751.257
	EQM	2.255961	2.255855	0.02750890	60451.51	1.649424	198849.6
	CV	1.6174236	1.6173856	0.1810238	-23.1029228	1.3634534	56.6603089
8	%Biais	-14.526	-14.525	-38.076	-222.265	-0.296	-590.926
	EQM	0.4696847	0.4696624	0.07668760	50.80157	0.2023659	51476.57
	CV	2.586845	2.586773	1.442799	-18.807826	1.455654	-149.104906

Tableau 4.4

Comparaison des estimateurs dans le cas des probabilités inégales

Populations générées selon le coefficient de variation et la moyenne de la variable auxiliaire n=10

CV	Estimateur	0.35		5		20.5	
		%Biais	EQM	%Biais	EQM	%Biais	EQM
0.5	Moyenne	1.170	0.02059161	2.734	0.02152182	-1.096	0.02077021
	Différence	1.170	0.02059095	2.735	0.02152103	-1.096	0.02076963
	Régression	0.462	0.02472578	1.694	0.02389024	-2.299	0.02467649
	Stuart	-9.954	0.02925298	-19.179	0.39171075	-13.199	0.04945924
	Lahiri	0.704	0.01798599	-0.271	0.01714655	-0.587	0.01891429
	Hartley-Ross	0.125	0.01869503	0.882	0.01841183	1.773	0.01937157
1	Moyenne	4.272	0.06317623	-2.697	0.03730049	-2.662	0.06779300
	Différence	4.271	0.06315689	-2.696	0.03729127	-2.662	0.06777084
	Régression	-2.139	0.02753799	-1.869	0.02849180	-7.491	0.02510217
	Stuart	-89.886	10.99631464	-50.685	5.27793301	-324.614	219.56160184
	Lahiri	0.103	0.01440060	-1.480	0.01339110	2.404	0.01453111
	Hartley-Ross	45.506	5.00269877	59.604	64.03055209	-12.043	3.15293939
1.5	Moyenne	-8.056	0.05121468	-9.346	0.04942690	-6.516	0.08962877
	Différence	-8.054	0.05119960	-9.349	0.04941646	-6.501	0.08959668
	Régression	-2.357	0.03630206	3.658	0.04499450	-2.355	0.03866087
	Stuart	-8.069	0.05112448	-9.347	0.04941745	-6.518	0.08960269
	Lahiri	3.994	0.01795416	1.790	0.01636550	1.014	0.01593050
	Hartley-Ross	252192.468	$1.046747\ 10^{9}$	-90688.529	$8.170452\ 10^{7}$	-40542.473	$4.701023\ 10^{6}$
2	Moyenne	-7.414	0.1830199	19.047	-0.1768276	-7.781	0.3258549
	Différence	-7.398	0.1829506	-19.011	0.1767572	-7.765	0.3257212
	Régression	3.628	0.04508684	-3.221	0.048655	-2.777	0.05342796
	Stuart	-7.414	0.1830199	-19.047	0.1768276	-7.781	0.3258549
	Lahiri	2.041	-0.01586852	-2.835	0.01965917	0.705	0.02394777
	Hartley-Ross	$-1.06043\ 10^{8}$	$2.125838\ 10^{15}$	$-1.61614\ 10^{17}$	$7.907088\ 10^{31}$	1.4260710^{7}	$2.072792\ 10^{12}$

Tableau 4.5

Comparaison des estimateurs dans le cas des probabilités inégales

Populations générées selon le coefficient de variation et la moyenne de la variable auxiliaire n=100

CV	Estimateur	0.35		5		20.5	
		%Biais	EQM	%Biais	EQM	%Biais	EQM
0.5	Moyenne	0.107	0.002127428	-0.245	0.001842979	0.386	0.002122613
	Différence	0.108	0.002127381	-0.245	0.001842950	0.385	0.002122619
	Régression	-0.063	0.002171177	-0.461	0.001878840	0.281	0.002149912
	Stuart	-1.410	0.002118446	-1.719	0.001901503	-0.790	0.002090632
	Lahiri	-0.268	0.001824242	-0.234	0.001656642	-0.574	0.001802958
	Hartley-Ross	0.016	0.001821401	-0.431	0.001660530	-0.022	0.001799015
1	Moyenne	0.333	0.005364118	-0.688	0.003690619	-0.076	0.007373202
	Différence	0.333	0.005362541	-0.686	0.003689511	-0.076	0.007370732
	Régression	-0.335	0.003120272	-0.333	0.003126026	-2.040	0.003101363
	Stuart	-9.564	0.069527277	-4.481	0.007871269	-68.772	14.594434674
	Lahiri	-0.202	0.001260165	-0.718	0.001349023	-0.682	0.001287340
	Hartley-Ross	-3.019	0.012540846	-2.842	0.285809214	-3.281	0.027723503
1.5	Moyenne	-3.263	0.02204768	-0.113	0.03874521	-3.892	0.01434280
	Différence	-3.258	0.02203990	-0.115	0.03873007	-3.887	0.01433724
	Régression	-0.005	0.007424159	2.380	0.007729598	-0.446	0.006705386
	Stuart	-3.478	0.02069471	-0.133	0.03839892	-3.893	0.01433932
	Lahiri	-1.040	0.001316596	-1.262	0.001226974	-0.030	0.001337614
	Hartley-Ross	-1904.750	110879.1	-1171.507	172130.4	597.296	35887.90
2	Moyenne	5.011	0.4361149	-14.466	0.07344055	-15.200	0.04126652
	Différence	5.004	0.4359470	-14.447	0.07341040	-15.194	0.04124749
	Régression	-3.303	0.01067713	-8.941	0.01183794	-3.476	0.01182109
	Stuart	5.011	0.4361148	-14.466	0.07344026	-15.200	0.04126651
	Lahiri	-0.018	0.001307034	0.577	0.001494448	0.459	0.001434053
	Hartley-Ross	$-1.754053\ 10^{8}$	$2.884548\ 10^{13}$	$3.785450\ 10^{8}$	$4.847668\ 10^{16}$	$1.8015\ 10^{6}$	$1.262022\ 10^{10}$

BIBLIOGRAPHIE

Ardilly P, *Les techniques de sondage*, Technip, 1994.

Basu, D.1958. «On sampling with and without replacement». *Sankhya* 20, 287-294.

Clairin R, et P Brion. *Manuel de sondages, Applications aux pays en développement*, deuxième édition. Paris 1997.

Cochran, W. G. 1977. *Sampling Techniques*, Third Edition, New York : John Wiley and Sons, Inc.

Cornfield, J. 1944. «On samples from finite populations». *Journal of the American Statistical Association*. 37, 236-239.

David, I. P., et B. V Sukhatme. 1974. «On the bias and mean square error of the ratio estimator». *Journal of the American Statistical Association* 69, 464-466.

Durbin, J. 1959. «A note on the application of Quenouille's method of bias reduction to the estimation of ratios». *Biometrika* 46, 477-480.

Goodman, L. A. 1960. «On the exact variance of products». *Journal of the American Statistical Association*, 55 708-13.

Goodman, L. A., et H. O. Hartley. 1958. «The precision of unbiased ratio-type estimators». *Journal of the American Statistical Association*, 53 491-508.

Hàjek, J. 1971 «Comment on "An Essay on the Logical Foundations of Survey Sampling, Part One," ». *The Foundations of Survey Sampling*. Godambe, V.P., et Sprott, D.A. eds., 236, Holt, Rinehart et Winston.

Hartley, H. O., et A.Ross. 1954. «Unbiased ratio estimates». *Nature* 174, 270-271.

Hutchison, M. C. 1971. «A Monte Carlo comparison of some ratio estimators». *Biometrika* 58, 313-0321.

Isaki, C.T. 1983 «Variance estimation using auxiliary information». *Journal of the American Statistical Association*, 78(381):117-123.

Kish, L., N. K. Namboodiri, et R.K. Pillai, 1962. «The ratio bias in surveys». *Journal of the American Statistical Association* 57, 863-876.

Koop, J.C. 1951. «A note on the bias of the ratio estimate». *Bulletin of the International Statistical Institute*, 33: 141-146.

_____. 1968. «An exercise in ratio estimation». *The American Statistician* 22, 21-30.

Lahiri, D. B. 1951. «A method of sample selection providing unbiased ratio estimators». *Bull. Inst. Internat. Statist.* 33 2, 133-140.

Mickey, M. R. 1959. «Some finite population unbiased ratio and regression estimators». *Journal of the American Statistical Association* 54,694-612.

Midzuno, H. 1952. «On the sampling system with probability proportionate to sum of sizes». *Ann. Inst. Statist. Math.* 3, 99-107.

Murakami, M. 1950 «Some consideration on the ratio and regression estimates». *Bulletin of Mathematical Statistics*, 4(1-2): 39-42.

Nanjamma, N. S., M. N Murthy, et V. K. Sethi, 1959. «Some sampling aystems providing unbiased ratio estimators». *Sankhya* 21, 299-314.

Pascual, J.N. 1961. «Unbiased ratio estimators in stratified sampling». *Journal of the American Statistical Association*, 56(293): 70-87.

Pathak, P.K. 1964. «On inverse sampling with unequal probabilities». *Biometrika* 51,(1-2), 185-193.

Quenouille, M. H. 1956 «Notes on bias estimation». *Biometrika* 43, 353-360.

Raj Des. 1954. «Ratio Estimation in sampling with equal and unequal probabilities». *J. Indian. Soc. Agric. Statist.*, 6 127-138.

_____, 1964. «A Note on the Variance of the Ratio Estimate». *Journal of the American Statistical Association* 59, 895-898.

Raj Des., et S.H Khamis,. (1958). «Some Remarks on Sampling with Replacement». *The Annals of Mathematical Statistics* 29, 550-557.

Rao, J. N. K. 1965. «A note on estimation of ratios by Quenouille's method». *Biometrika*, 52, 647 – 649.

_____. 1967. «The precision of Mickey's unbiased ratio estimator». *Biometrika* 54, 321-324.

_____.1969. «Ratio and regression estimators». In *New Developments in Survey Sampling*, eds. N. L. Johnson et H. Smith, pp. 213-234. New York: Wiley.

Rao, P. S. R. S. 1968. «On three procedures of sampling from finite populations». *Biometrika* 55 2, 438-441.

_____. 1969. «Comparison of four ratio-type estimates under a model». *Journal of the American Statistical Association* 64, 574-480.

Rao, P S. R. S., et J. N. K. Rao,1971. «Small Sample Results for Ratio Estimators». *Biometrika* 3, 625-630.

Rao, J.N.K., et V. Ramachandran, 1974. «Comparison of the separate and combined ratio estimators». *Sankhyā*, Volume 36, Series C, Pt. 3: 151-156.

Royall, R.M., et K.R. Eberhardt, 1975. «Variance estimates for the ratio estimator». *Sankhyā*, 37(C,1): 43-52.

Sàrndal, C., B.Swensson., et J. Wretman, 1992 «Model Assisted Survey Sampling». *Springer Verlag*,New York.

Sen, P. K. 1953. «On the estimate of the variance in sampling with varying probabilities». *Journal of the Indian Society of Agricultural Statistics* 5, 119-127.

Smith, H, Fairchild 1969. «Approximate Formulae for Bias and Variance Ratios», *The American Statistician* 23, 29-31.

Stuart, A. 1986. «Location-shifts in sampling with unequal probabilities». *Journal Royal of Statistical Society*. A, vol 149, Part 4, p 349-365.

Sukhatme, P. V. 1953. «Sampling Theory of Surveys with Applications», New Delhi, India: *Indian Society of Agricultural Statistics*.

Sukhatme, B.V. 1962. «Some ratio-type estimators in two-phase sampling». *Journal of the American Statistical Association*, 57(299): 628-632.

Sukhatme, B. V., et I. P. David, 1974. «On the bias and mean square error of the ratio estimator». *Journal of the American Statistical Association* 69, 464 – 466.

Tillé, Y, *Théorie des sondages*, Dunod, 2001.

Tin, M 1965. «Comparison of some ratio estimators». *Journal of the American Statistical Association* 60, 294-307.

Wu, C.F. 1982. «Estimation of variance of the ratio estimator». *Biometrika*, **69**(1): 183-189.

Yates, F., et P. M. Grundy, 1953. «Selection without replacement from within strata with probability proportional to size». *Journal of the Royal Statistical Society,* Série B1, 253-261.