UNIVERSITÉ DU QUÉBEC À MONTRÉAL

Estimation de la borne supérieure par des approches statistiques et par la méthode de Stringer

MÉMOIRE

PRÉSENTÉ

COMME EXIGENCE PARTIELLE

DE LA MAÎTRISE EN MATHÉMATIQUES

PAR

SAMI JOUBIR

JANVIER 2010

REMERCIEMENTS

Je voudrais exprimer ma profonde gratitude au professeur Serge Alalouf qui a dirigé mes travaux de recherches. Ses conseils judicieux, ses explications ont été précieux.

Mes remerciements vont également à tous les professeurs avec qui j'ai suivi des cours, en particulier Mme Sorona Froda.

Merci aussi à toute ma famille à Tunis, mes cousins à Montréal et mes amis pour leur soutien.

TABLE DES MATIÈRES

LISTE DES TABLEAUX

RÉSUMÉ

Ce mémoire présente de nouvelles approches statistiques pour estimer la borne supérieure d'une population dans un contexte bien particulier, celui de la vérification comptable.
Étant donné que dans la plupart des cas on se retrouve avec des échantillons où le nombre d'erreurs est souvent faible ou nul, les méthodes classiques risquent fort d'être inadéquates.

Dans ce mémoire, nous allons revenir sur quelques méthodes classiques puis présenter différentes méthodes spécifiques proposées par des chercheurs et nous mettrons l'accent sur la méthode de Stringer qui est très utilisée dans la pratique de la profession. Notre objectif est de voir dans quels cas ces méthodes pourraient être plus efficaces que les méthodes classiques.

Les propriétés des méthodes classiques sont connues, contrairement à celles des approches spécifiques où plusieurs d'entre elles n'ont jamais été démontrées et, parmi elles, la méthode de Stringer qui nous intéresse particulièrement.

À cet effet, dans le chapitre 3, nous allons faire des simulations pour confirmer les comparaisons théoriques entre les méthodes dont on connait les propriétés et voir les résultats de celles qu'on ne connaît pas.

Mots clés : échantillonnage, estimation, borne supérieure, méthodes classiques, méthode de Stringer.

INTRODUCTION

Ce mémoire traite d'un problème de sondage classique--comment estimer, par intervalle de confiance, la moyenne (ou le total) d'une population—pour lequel les méthodes traditionnelles se sont souvent révélées inadéquates. Ces méthodes, basées sur des théorèmes asymptotiques, reposent sur le fait que les estimateurs habituels—l'estimateur par la moyenne, par la différence, par le quotient, ou par la régression—sont des moyennes ou des fonctions de moyennes. Par conséquent, il est possible d'invoquer un théorème limite permettant de conclure que l'estimateur est à-peu-près de loi normale. L'écart-type de l'estimateur étant généralement inconnu, on le remplace par une estimation calculée à partir des données de l'échantillon. Cela permet de définir un pivot, de loi (à peu près) normale centrée réduite et mène à un intervalle de confiance par les moyens habituels.

Bien que ces procédures dépendent fortement de la convergence des estimateurs vers une loi normale, il se trouve que dans la plupart des domaines, la convergence est assez rapide pour les besoins pratiques et la normalité approximative est atteinte avec des échantillons de taille modérée. Les intervalles de confiance déterminés de cette manière jouissent donc de cette propriété essentielle, à savoir, que la probabilité de recouvrement—la probabilité que l'intervalle contienne le paramètre—est à peu près égale au niveau de confiance nominal $1 - \alpha$. Évidemment, cette convergence vers la normalité peut être lente si la distribution des données de la population a une forme très différente d'une normale, mais dans la plupart des domaines d'application, la forme de la population, sans s'approcher d'une normale, se prête bien au traitement classique. C'est ce que révèlent de nombreuses études empiriques.

Il existe un domaine d'application, cependant, où la confiance que donnent ces études s'ébranle : la vérification comptable. Là la situation est la suivante. Un vérificateur tire un échantillon dans une population de comptes, les corrige s'il y a lieu et estime le montant total de l'erreur commise en vue d'une cotisation. Par exemple, afin de vérifier les demandes de remboursement de taxes (TVQ ou TPS) faites par des entreprises, Revenu Québec doit tirer un échantillon des factures soumises par l'entreprise et déterminer si les remboursements demandés sont éligibles et correctes. Le vérificateur doit estimer le montant total τ de l'erreur, c'est-à-dire la différence entre le montant réclamé et le montant éligible et ensuite—une obligation fondamentale—déterminer une borne supérieure pour la valeur de τ afin de justifier sa cotisation et de ne pas compromettre les droits du contribuable. Il s'agit donc d'établir un intervalle de confiance de la forme $[0; \tau]$. Un problème, donc, classique. Sauf que, dans ce cas précis, la population est généralement constituée majoritairement de valeurs nulles, la plupart des documents dans la population ne comportant aucune erreur. Une distribution, donc, très éloignée de la loi normale. C'est ce qui fait que les méthodes classiques risquent fort d'être inadéquates, à moins d'un échantillon démesuré.

Ce ci n'est pas la seule difficulté. Car il est également possible—fréquent, en fait—que l'échantillon ne contienne aucune erreur, c'est-à-dire que toutes les données de l'échantillon soient nulles. On peut bien, dans ce cas, déterminer une borne supérieure pour la proportion des comptes erronés, mais on ne dispose d'aucune information sur le montant des comptes erronés. Aucune des méthodes classiques d'échantillonnage ne fournit de solution à un tel problème, un problème qui semble, à première vue, insoluble. Certaines solutions ont néanmoins été proposées et, dans la plupart des cas, par des chercheurs oeuvrant dans le domaine de la vérification. Dans ce mémoire, nous faisons un exposé des approches proposées. Nous portons une attention particulière à celles conçues pour traiter le problème spécifique de vérification. L'une d'elles, qui fournit ce qu'on appelle en vérification les bornes de Stringer, mérite une place importante dans ce mémoire en vertu de la place qu'elle occupe dans le monde de la vérification : c'est la méthode

préconisée par certains ordres comptables et utilisée le plus couramment dans les cabinets de comptabilité.

Ces méthodes présentent plusieurs inconvénients. D'abord, il y en a certaines dont la difficulté d'application exclut leur utilisation routinière par des non spécialistes. Ensuite, certaines reposent sur des hypothèses concernant les valeurs de la population, des hypothèses qui sont souvent vérifiées, mais pas toujours. Finalement, leurs propriétés n'ont pas toujours été démontrées et c'est le cas, entre autres, de la méthode de Stringer. On a de bonnes raisons de croire qu'elles ont une probabilité de recouvrement assez élevée, mais que cette confiance se paie cher : les intervalles sont souvent extrêmement conservateurs, les bornes supérieures étant souvent très grandes, au point, parfois, d'être totalement inutiles. C'est pourquoi les méthodes classiques, malgré leurs défauts dans le contexte actuel, retiennent leur attrait et ne doivent pas être rejetées d'emblée. Nous avons donc tenté d'établir les conditions dans lesquelles telle méthode est préférable à telle autre et surtout de déterminer dans quelles conditions les méthodes classiques pouvaient malgré tout être retenues. Nous verrons que si l'échantillon est assez grand, sans être démesuré, les méthodes classiques ont leur place encore, même lorsque la population contient une forte proportion de valeurs nulles.

Bien que dans quelques cas des démonstrations théoriques ont été possibles, nous avons procédé par des moyens empiriques. Les propriétés théoriques sont parfois asymptotiques, ce qui ne donne pas assez d'information sur les propriétés d'un échantillon fini. Il y a aussi le fait que certaines des méthodes proposées sont restées jusqu'ici sans justification formelle. Les propriétés de bornes de Stringer, par exemple, n'ont jamais été démontrées. Nos simulations sont faites à partir de populations générées artificiellement, conçues pour couvrir un certain éventail de caractéristiques : moyennes variables, variances variables et surtout, une proportion variable de valeurs nulles, la source principale de difficulté. Nos simulations révéleront que les probabilités de recouvrement sont souvent adéquates, mais que le vrai problème, c'est l'extrême conservatisme de plusieurs

des méthodes. Nous avons donc tenté de répondre à deux questions. La première est : dans quelle mesure la probabilité de recouvrement approche-t-elle la probabilité nominale $1 - \alpha$? La deuxième est : dans quelle mesure la limite supérieure fournie est-elle raisonnablement petite ? Nous avons obtenu quelques réponses, avec, bien entendu, les réserves d'usage.

CHAPITRE I

HISTORIQUE

Le problème traité dans ce mémoire est un problème statistique dont le domaine principal d'application est celui de la vérification comptable. Aussi, la plupart des chercheurs qui ont contribué à l'étude de ce problème sont issus du monde de la comptabilité, un domaine où l'application de la statistique s'est fait attendre et où l'implantation des méthodes statistiques s'est faite avec une certaine réticence.

Néanmoins, l'application des méthodes statistiques en comptabilité prend de plus en plus d'importance de nos jours, en partie comme conséquence de certaines poursuites contre des cabinets de comptables, mettant en cause la méthode d'échantillonnage des comptes d'une entreprise. Un autre facteur contributif est le fait que les vérificateurs des agences gouvernementales ont elles-même employé des méthodes scientifiques d'échantillonnage, obligeant ainsi les entreprises privées à instaurer des procédures dont les propriétés pouvaient être établies de façon relativement objective.

Bien que ces préoccupations aient commencé à ce faire sentir dès les années quarante, le progrès a été plutôt lent au début. Ainsi, en 1947, un document du Americain Institute of Accountants « Tentative Statement of Auditing Standards »discute longuement des problèmes d'échantillonage. Il propose une approche basée sur le jugement du vérificateur, mais ne fait aucune allusion à l'application de méthodes statistiques. Pourtant, antérieurement, Hebert (1946) s'était déjà penché sur les questions de probabilité

et en particulier au problème qui nous concerne ici, soit les situations où les erreurs dans la population sont rares. Cependant, il s'est borné à simplement calculer la probabilité de déceler un erreur ou plus, en fonction du pourcentage d'erreur dans la population et de la taille de l'échantillon.

C'est au début des années cinquante que les chercheurs commencent à prendre conscience de l'utilité des méthodes statistiques en comptabilité. Warrimer (1951) insiste sur l'utilité possible des méthodes telles les nombres indices comme outil supplémentaire dont la profession aurait intérêt à se doter. Mais le problème d'échantillonnage n'est pas mentionné. Vance (1951) a été parmi les premiers à s'attarder spécifiquement sur les problèmes d'échantillonnage en vérification. Il recommande l'utilisation de tests d'hypothèse sur des taux d'erreurs, donc essentiellement des tests classiques sur une proportion p. Neter (1954) a également beaucoup fait pour adapter le langage des méthodes d'enquêtes au contexte du monde comptable. Entre autres il souligne la distinction qui doit être faite entre l'inférence concernant un paramètre et celle concernant une superpopulation infinie.

Il semblerait qu'à cette date les méthodes d'échantillonnages faisaient déjà partie des systèmes de vérification dans certaines grandes entreprises. Vance (1952) présente des exemples réels de sélection rigoureusement aléatoire appliquée par certaines entreprises, utilisant des tables de nombres aléatoires. Neter (1952) relate un certain nombre de tels cas. On y met l'accent sur différents modes de tirage, entre autres le tirage systématique et le tirage par grappes, bien que les marges d'erreurs ne sont pas calculées dans ces cas-là. Mais dans des tirages relativement simples, certaines entreprises évaluaient le risque d'erreur de première et seconde espèce. Trubelood et al. (1955) étudient quelques cas où l'échantillonnage a été effectué selon les règles, avec estimation et tests d'hypothèses sur des moyennes et des proportions. Le problème d'un grand nombre de zéros n'est cependant pas évoqué. Vance (1960) relate quelques applications réelles de méthodes d'estimation classiques (estimation de moyenne, avec intervalle de confiance basée sur

la loi normale).

Au courant des années soixante et jusqu'au début des années soixante-dix, bien que les méthodes classiques et connues aient pris de l'ampleur, aucune méthode particulière ne semble avoir été développée pour résoudre les problèmes propres à la vérification. Jusque là, la seule attention particulière qui semble avoir été consacrée aux cas qui nous concernent ici, ceux où le taux d'erreur est extrêmement faible, consistait à calculer la probabilité qu'au moins une erreur se retrouve dans l'échantillon. Mais les méthodes classiques se révélaient inadéquates pour certaines populations comptables.

Dans une étude comparative par simulation, Kaplan (1973) révèle des faiblesses dans la performance des estimateurs par le quotient, par la régression et par la différence. À partir de quelques hypothèses concernant la distribution des valeurs aux livres (uniforme et exponentielle) le taux d'erreur et la distribution des erreurs, il étudie la performance des méthodes d'estimation classiques et les trouve plutôt déficientes. Plus tard, Neter et Loebeckke(19750) ont montré (par simulation, également) que la méthode d'estimation par le quotient peut surestimer considérablement le niveau de confiance. Des résultats semblables sont confirmées plus tard par Forest et Tamura(1986), qui attribue le déficience à la non symétrie des populations. Reneau (1978) compare la performance de cinq méthodes et obtient des résultats mixtes. Ramage, Krieger et Spero (1979) tentent de classifier par des indices quantitatifs les populations selon les caractéristiques de la distribution d'erreurs. Ces analyses sont basées sur des données empiriques sur diverses industries. Ramage, Krieger et Spero Considèrent l'estimation par le quotient, avec l'approche usuelle, mais en remplaçant l'estimateur habituel de la variance par un estimateur Jackknife. Ils trouvent par simulation que dans les populations dont la corrélation entre la variable d'intérêt et la variable auxiliaire est forte, l'intervalle proposé est légèrement plus précis que l'intervalle habituel. Menzefricke et Smieliauska (1984) présentent une étude sur les diverses méthodes d'échantillonnage dites par *unité monétaire* conçues pour traiter des cas de-sous estimation ainsi que celles conçues pour

rendre les méthodes usuelles moins conservatrices.

Quelques auteurs, vers les années soixante-quinze, se sont intéressés à des questions plus fondamentales, touchant à l'objectif de la vérification et son lien avec les critères statistiques usuels. Teitlebaum et Robinson (1975) tentent de développer une théorie fondamentale de vérification statistique dans laquelle interviennent les notions de risque et d'utilité. Dans le même esprit, Loebbecke et Neter (1975) insistent sur l'importance de faire correspondre le plan de sondage à l'environnement et les procédures de vérification. Roberts (1975) s'intéresse au sens des probabilités calculées par d'autres auteurs et attire l'attention sur le fait que certaines des probabilités citées par Teitlebaum et Robinson étaient des probabilités conditionnelles.

1.1 Énoncé formel du problème et approches classiques

On doit procéder à la vérification d'une population de N *comptes*, et pour ce faire, on tire un échantillon de n *comptes*. Les valeurs aux livres des comptes $x_1, x_2, \cdots, x_N$, sont connues. Les valeurs *réelles* $y_1, y_2, \cdots, y_N$, sont inconnues. Le problème consiste à estimer le montant total des erreurs $t_d = \sum_{i=1}^{N}(x_i - y_i) = \sum_{i=1}^{N} d_i$.
Lorsque le tirage est aléatoire simple, les estimateurs habituels sont :
Estimateur par la moyenne : $\widehat{t}_{d_m} = N\overline{d}_s$, où $\overline{d}_s = \frac{1}{n}\sum_{i\in s} d_i$.
Estimateur par la différence : $\widehat{t}_{diff} = N(\overline{x}_U - \overline{y}_s)$, où $\overline{x}_U = \frac{1}{N}\sum_{i\in U} x_i$, $\overline{y}_s = \frac{1}{n}\sum_{i\in s} y_i$.
Estimateur par le ratio : $\widehat{t}_r = \frac{\widehat{t}_{d_m}}{\widehat{t}_x} t_x$, où $\widehat{t}_x = N\overline{x}_s$.
Quand le tirage se fait avec remise et probabilités proportionnelles aux valeurs aux livres x_i, l'estimateur habituel est $\widehat{t}_{d_r} = \frac{t_x}{n}\sum_{i\in s}\frac{d_i}{x_i}$.

La quantité cruciale dans ce contexte, c'est la borne supérieure d'un intervalle de confiance de la forme $[0; \widehat{T}_{dL}]$ pour le total t_d. Cette borne est normalement déterminée

en faisant la somme de l'estimateur et de son écart-type estimé multiplié par z_α. Cette façon de faire se justifie comme ceci :
L'estimateur étant une fonction de moyennes échantillonnales, il suit asymptotiquement une loi normale ; l'écart-type estimé de l'estimateur tend en probabilité vers l'écart-type réel ; et, finalement, la probabilité que t_d se situe dans l'intervalle $[0; \widehat{T}_{dL}]$ est à-peu-près de 95% si n est assez grand.

Tout cela est vrai, sauf que, dans certains cas extrêmes, la convergence vers la loi normale est très lente et l'approximation normale, même avec un grand échantillon, peut être très grossière. C'est ce à quoi fait face le vérificateur, du fait que les valeurs d_i de la population sont pour la plupart nulles, ce qui entraîne non seulement une distribution d'échantillonnage très éloignée d'une normale, mais en plus, le risque très réel d'un échantillon constitué entièrement de données nulles.

La plupart des méthodes que nous allons discuter sont basées sur l'hypothèse que $d_i \geq 0$ pour toute facture de la population. C'est une hypothèse que l'on défend de deux façons. D'abord, il est possible qu'elle soit strictement vraie lorsque les erreurs ne sont pas vraiment involontaires et constituent une tentative de fraude, ou bien, surévaluer plutôt que sous-évaluer le montant réclamé, dans le cas de doute. C'est le cas des demandes de remboursement en particulier. À cela s'ajoute le fait qu'il n'appartient pas au vérificateur de se préoccuper des sous-évaluations, puisque son but est de s'assurer qu'il n'y ait pas de surévaluation—l'estimation des sous-évaluations étant considérée comme la responsabilité du demandeur.

Dans la partie qui suit, nous décrirons, dans les grandes lignes, les principales nouvelles méthodes proposées.

1.2 Les approches DUS et CAV

La méthode d'échantillonnage dite par *unité monétaire* (« Dollar-unit sampling », ou DUS) est généralement attribuée à Anderson et Teitlebaum(1973) alors qu'elle était déja mentionnée par Arkin (1963). Anderson et Teitlebaum ont également introduit l'expression « Combined attributes-variables procedures »(CAV) pour désigner une approche par laquelle une valeur monétaire est associée à chaque unité monétaire érronée. Bien qu'ils ne l'aient pas présenté ainsi, l'échantillonnage DUS est essentiellement un tirage avec probabilités proportionnelles aux montants des comptes. Leur article a donné lieu à une activité fébrile dans une direction nouvelle. Une approche dite d'« attributs »est à l'origine de l'échantillonnage par *unité monétaire.* Par opposition à un *échantillonnage de variables*, l'échantillonnage d'attributs se limite à estimer uniquement la *proportion* et par suite le *nombre* de documents erronés. Une borne supérieure pour ce nombre, $N\widehat{\pi}$, déterminée par des moyens statistiques classiques, est multipliée par une certaine estimation $\overline{d}_L$ du *montant* moyen des erreurs non nulles. $\overline{d}_L$ n'est pas une estimation statistique. Il s'agit plutôt d'une limite supérieure découlant de certaines hypothèses supplémentaires que le vérificateur doit être en mesure de valider. L'hypothèse la plus courante—et reprise dans plusieurs recherches subséquentes—est que $0 \leq d_i \leq x_i$ pour toute unité de la population : toute erreur est une erreur de surestimation et le montant d'une erreur ne peut excéder la valeur aux livres. Dans ce cas, l'erreur moyenne est inférieure à la plus grande valeur aux livres, x_L, disons. La borne supérieure pour l'erreur totale est alors $N\widehat{\pi}_L x_L$. Évidemment, cette borne est extrêmement conservatrice, en particulier lorsque la population est très grande et x_L est excessivement grand par rapport aux autres valeurs.

Une façon de pallier ces difficultés, proposée (sans justification) par Meikle (1972), consiste à stratifier la population. La stratification permet de réduire la taille des populations ainsi que de modérer la surestimation que constitue x_L . L'allocation proposée par Meikle est une allocation proportionnelle à la valeurs aux livres.

Si les valeurs aux livres de comptes d'une même strate sont presque égales, la stratification revient à choisir les comptes avec probabilités proportionnelles aux valeurs aux livres. C'est une technique connue, dont l'application à la vérification a d'abord été proposée par Haskins & Sells (1970). Mais elle a été popularisée par Anderson et Teitlebaum, qui l'ont présentée sous le nom d'*échantillonnage d'unités monétaires* ou DUS, en anglais. Ils considèrent chacun des t_x dollars—et non chaque *compte*— comme unité et un échantillon est un tirage aléatoire simple de n dollars. Autrement, la procédure décrite ci-dessus reste intacte. L'intérêt de la méthode est que le montant maximal d'erreur, borné par x_L ci-dessus, est maintenant borné par 1. C'est ce qui permet de prévoir que la borne supérieure sera moins conservatrice.

Les méthodes mentionnées ci-dessus sont toutes basées sur l'échantillonnage d'attributs, ces méthodes sont considérées en réaction aux bénéfices de l'échantillonnage des variables dans la situation problématique traitée dans ce mémoire. Mais les bornes obtenues par l'échantillonnage d'attributs se révélant souvent trop conservatrices, plusieurs chercheurs ont tenté d'utiliser un mélange d'échantillonnage d'attributs et de variables. C'est ce qu'on nomme aujourd'hui l'approche CAV. L'un des avantages qu'on lui attribue est qu'elle permet de déterminer des bornes monétaires même lorsque aucune erreur ne se présente dans l'échantillon.

1.3 La méthode de Stringer

La plus connue et la plus répandue des méthodes CAV est la méthode de Stringer (1963). Anderson et Leslie (1975) affirment que les bornes de Stringer, appliquées dans le cadre DUS devraient être appliquées dans presque tous les cas étant donné qu'elles sont moins conservatrices que bien d'autres.

Pourtant, elles aussi semblent être très conservatrices et leurs propriétés n'ont jamais été démontrées. Nous avons choisi de nous concentrer sur les bornes de Stringer justement parce qu'elles sont bien implantées dans la pratique de la profession alors que ses propriétés demeurent relativement inconnues. Clayton (1994) compare deux façons

de calculer les limites supérieures, l'une basée sur les bornes de Hoeffding, l'autre sur la distribution *bootstrap*. Nous décrirons cette approche en détail dans le prochain chapitre. Il y a eu plusieurs autres tentatives faites dans le monde académique mais qui ne semblent pas avoir eu d'écho dans la profession. Nous en rapportons quelques unes brièvement.

1.4 Approches classiques modifiées

Certains chercheurs ont tenté d'améliorer, à l'aide de quelques ajustements, les méthodes classiques. Gartska et Olson (1979) ont proposé de remplacer z_α par une constante C à fin de construire l'intervalle de confiance pour t_d. La justification est une heuristique et s'appuie sur quelques simulations. Leur approche est critiquée par Tamura (1985), qui affirme que les bornes proposées par Gartska et Olson souffrent de ne pas tenir compte de l'asymétrie de la distribution de l'estimateur. Frost et Tamura (1982) examinent l'effet d'une estimation de l'erreur type par le jackknife. Ils concluent que cette dernière estimation donne de meilleurs résultats quand le taux d'erreur n'est pas trop faible. Bickel (1992) offre une justification intuitive de la méthode de Stringer et propose deux autres méthodes, l'une basée sur les bornes de Hoeffding (1963), l'autre calquée sur les tests de Kolmogorov-Smirnov.

1.5 Modèles paramétriques et approches bayésienne

Les difficultés rencontrées avec les approches ci-dessus, qui sont toutes essentiellement non paramétriques, ont encouragé certains chercheurs à recourir à des modèles. Gratska (1977) modélise par la loi de Poisson le taux d'erreurs (ce qui est raisonnable comme approximation de la loi binomiale) et traite ensuite le montant d'erreur comme une variable discrète —nombre d'*unités* d'*erreur*—de la loi géométrique. Il propose une estimation bayésienne empirique pour estimer le paramètre de la loi géométrique. La loi géométrique ne peut se justifier que par des considérations contextuelles et il n'en pro-

pose aucune, outre les résultats de certaines simulations comparant quelques variantes de cette approche à certaines approches classiques. Lillestol (1981), suivant la piste de Gartska, examine l'importance du modèle dans le calcul des bornes supérieures. À cette fin, il compare les limites calculées à partir de modèles géométriques à celles calculées à partir de modèles de série logarithmique $P(X = k) = \dfrac{\theta}{-k\ln(1-\theta)}$, $k = 1, 2, \cdots$. Il trouve des différences assez importantes et conclut que l'idée même d'utiliser des modèles est douteuse. Il étaye sa conclusion en montrant qu'un modèle géométrique n'est pas aisément rejeté, même s'il est assez faux pour donner des résultats nettement erronées. Cox et Snell (1979) présentent une analyse bayésienne d'un modèle dans lequel le nombre d'erreurs est de loi de Poisson et la densité des montants d'erreurs est exponentielle. Ils prennent la loi gamma pour loi à-*priori* et discutent des valeurs possibles de ses paramètres. La sensibilité de ce modèle au choix de l'à-*priori* est étudié par Godfrey et Neter (1984) qui considèrent, par exemple, la loi bêta comme alternative à la loi gamma. Ils constatent que l'effet de cette modification est modéré. Dworin et Grimland (1984) utilisent la loi khi-deux pour modéliser la distribution des erreurs. Neter et Godfrey (1985) proposent d'autres loi à-*priori* qui semblent donner de meilleurs résultats. Tamura et Frost (1986) proposent une loi de densité proportionnelle à une puissance de x et utilisent le bootstrap pour estimer la distribution de l'estimateur. Kvanly, Shen et Deng (1998) modélisent la distribution des erreurs non nulles par la loi normale et par la loi exponentielle et définissent la vraisemblance en fonction des paramètres (moyenne des données non nulles) et l'espérance d'une observation, ainsi que de l'écart-type des données non nulles dans le cas du modèle normal. L'intervalle de confiance pour τ est défini par l'ensemble de toutes les valeurs τ_0 pour lesquelles l'hypothèse $H_0 : \tau = \tau_0$ n'est pas rejetée par un test du rapport de vraisemblances. On utilise le fait que, asymptotiquement, $-2\ln\frac{L_0}{L}$ est de loi χ^2 sous H_0, L étant la fonction de vraisemblance maximisée sur l'espace paramétrique entier et L_0 est le maximum sous la contrainte $\tau = \tau_0$. À partir de quelques simulations, ils concluent que leur approche est beaucoup plus exacte que la méthode classique basée sur le théorème limite central, ceci dans les deux modèles étudiés, le modèle normal et le modèle exponentiel. Ces conclusions, cependant, n'ont rien d'étonnant, puisque les simulations sont faites à

partir de populations qui correspondent aux hypothèses. Cette critique a été évitée par Chen, Chen et Rao (2002) qui adoptent la même approche sans toutefois imposer un modèle : le test du rapport de vraisemblances est effectué à partir de la fonction de vraisemblance *empirique*. Ils comparent, par simulations, leurs taux de recouvrement à ceux des méthodes classiques et à ceux de Kvanly, Shen et Deng. Ils trouvent que la performance de la méthode par la vraisemblance empirique est proche de celle des méthodes de Kvanly et al. lorsque le modèle adopté est bon, mais que sinon, leur approche est généralement meilleure. Menzefricke (1983) suit optimise par des tests de contrôle basés sur un tirage DUS : cette approche consiste à déterminer une taille d'échantillon n et un point critique de façon à minimiser le coût total sous une contrainte de risque de type II. Il propose deux méthodes mais Finley (1985) montre que l'une d'elles n'est pas optimale.

1.6 Bornes basées sur la loi multinomiale

Une méthode très particulière a été proposée par Fienberg, Neter et Leitch (1977). Elle est basée sur un tirage d'unités monétaires et l'erreur relative est considérée comme une variable discrète dont les valeurs sont $0, 1, \cdots, 100$. Un dollar échantillonné prend la valeur i avec probabilité p_i et le nombre d'observations prenant chacune des valeurs est un vecteur de loi multinomiale. Le paramètre à estimer est $t_d = \dfrac{t_x}{100} \displaystyle\sum_{i=1}^{100} i p_i$. L'approche consiste à déterminer d'abord une région de confiance pour le vecteur de paramètres, $\boldsymbol{p} = [p_0; p_1; \cdots; p_{100}]$ et par la suite obtenir la valeur maximale de t_d lorsque p est dans la région de confiance. C'est une approche extrêmement difficile à appliquer, à moins que le nombre d'erreurs soit très faible. Et comme avec plusieurs des méthodes considérées, celle-ci ne s'applique qu'au cas où les erreurs sont des erreurs de surévaluation. Neter, Leitch et Fienberg (1978) et Plante, Neter et Leitch (1984) ont proposé une modification permettant de l'utiliser pour des erreurs de sous-évaluation également ; Leitch, Neter, Plante et Sinha (1981, 1982) ont adapté la méthode au cas où le nombre d'erreurs est plus important, constituant les erreurs en grappes. Tsui, Matsumura, et Tsui (1985) définissent les mêmes paramètres, qu'ils estiment par des méthodes bayésiennes

avec pour distribution à-*priori* la loi de Dirichlet. Matsumura, Plante, Tsui et Kannan (1991) comparent l'approche bayésienne à l'approche originale au moyen de simulations. Grimlund et Felix (1987) comparent par simulation un grand nombre de méthodes.

CHAPITRE II

ESTIMATION

Plusieurs modes d'échantillonnage et estimateurs classiques existent et nous allons les considérer au même titre que les méthodes développées spécifiquement pour le contexte qui nous préoccupe dans ce mémoire. Les méthodes classiques dépendent dans une large mesure de l'hypothèse que l'estimateur, étant toujours une moyenne ou une fonction de moyennes, est asymptotiquement de loi normale. Or la convergence vers la normalité, quand la population est fortement asymétrique, est très lente et c'est ce qui a motivé la recherche d'autres méthodes mieux adaptées. Ce qui ne veut pas dire que les méthodes classiques doivent être abandonnées. C'est pourquoi nous en présentons un bref aperçu d'abord, pour ensuite décrire les méthodes spécifiques.

2.1 Théorie de l'échantillonnage

Considérons une population U composée de N comptes, de laquelle on tire un échantillon s, de taille n. Soit :

$x_i, i \in \{1; \cdots; N\}$ la valeur comptable pour l'unité i

$y_i, i \in \{1; \cdots; N\}$ la valeur vérifiée pour l'unité i

$t_x = \sum_{i \in U} x_i$

$t_y = \sum_{i \in U} y_i$

Le montant d'erreur dans un compte donné i que l'on désigne par d_i est défini selon $d_i = x_i - y_i$

Le montant d'erreur total dans la population que l'on désigne par t_d est défini selon

$$\begin{aligned} t_d &= t_x - t_y \\ &= \sum_{i \in U} (x_i - y_i) \\ &= \sum_{i \in U} d_i \end{aligned}$$

Dans le cas d'un échantillonnage aléatoire simple sans remise, l'échantillon est tiré à partir d'une population de taille finie N.

C'est-à-dire que contrairement à la statistique classique, le vecteur

$$\mathbf{D} = \begin{pmatrix} d_1 \\ . \\ . \\ . \\ . \\ d_N \end{pmatrix}$$

est non aléatoire mais ce qui est aléatoire c'est la sélection d'une unité dans l'échantillon : ce qui est généralement le cas sauf en agriculture et autres domaines d'applications analogues.

Une technique classique de démonstration, que nous utiliserons souvent dans ce mémoire, passe par des variables indicatrices.

Notons par Z_i la variable indicatrice de sélection définie par :

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

Tous nos estimateurs pourront s'exprimer en fonction de ces Z_i et donc les propriétés de ces estimateurs dépendent de celles des Z_i. Les paramètres du vecteur $[Z_1; \cdots; Z_N]$ qui nous concernent sont les probabilités d'inclusion dans l'échantillon.
Soit Ω l'ensemble de tous les échantillons s possibles et soit $p(s)$ la probabilité de sélection de l'échantillon s, avec :

$$\text{i) } p(s) \geq 0 \qquad \forall s \in \Omega$$
$$\text{ii) } \sum_{s \in \Omega} p(s) = 1$$

π_i est définie selon

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

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

$$\pi_{ij} = P(i \in s \,\&\, j \in s | i \neq j) = \sum_{\substack{s \in \Omega \\ (i,j) \in s^{i \neq j}}} p(s)$$

Notre objectif dans ce mémoire est double, car il s'agit non seulement d'évaluer certaines des méthodes proposées mais également de déterminer dans quelles conditions il est nécessaire d'y avoir recours. Car ce qui n'a pas encore été clairement établi, c'est dans quelle mesure les méthodes classiques font défaut dans le contexte examiné ici.

2.2 Méthodes classiques

Les méthodes classiques sont toutes fondées sur une même prémisse : tous les estimateurs utilisés sont des moyennes ou des fonctions de moyennes. En invoquant donc un théorème central limite, on peut conclure que l'estimateur est asymptotiquement de loi normale. Cela étant dit, on sait qu'en pratique les échantillons sont souvent de taille

modeste et on devine que la convergence vers la normalité peut être lente, surtout lorsqu'on a affaire à une population qui est loin d'être normale, ce qui est précisément la situation qui nous concerne dans ce mémoire.

Nous allons néanmoins passer en revue les techniques classiques car ce sont celles auxquelles nous comparerons les techniques spécialisées.

2.2.1 Tirage avec probabilités égales

Supposons que les unités sont tirées à partir d'un échantillonnage aléatoire simple sans remise alors :

La probabilité d'inclusion de l'unité i dans l'échantillon est donnée par :

$$\pi_i = \sum_{\substack{s\in\Omega \\ i\in s}} P(s) = \frac{1}{C_n^N} \sum_{\substack{s\in\Omega \\ i\in s}} 1 = \frac{C_{n-1}^{N-1}}{C_n^N} = \frac{n}{N}$$

La probabilité d'inclusion jointe des unités i et j $(i \neq j)$ est donnée par :

$$\pi_{ij} = \sum_{\substack{s\in\Omega \\ i,j\in s}} P(s) = \frac{1}{C_n^N} \sum_{\substack{s\in\Omega \\ i,j\in s}} 1 = \frac{C_{n-2}^{N-2}}{C_n^N} = \frac{n(n-1)}{N(N-1)}$$

où Ω est l'ensemble des C_n^N échantillons équiprobables possibles. En plus on a :

$$\begin{aligned}
E(Z_i) &= \pi_i = \frac{n}{N} \\
V(Z_i) &= E(Z_i^2) - E(Z_i)^2 \\
&= \pi_i - \pi_i^2 \\
&= \frac{n}{N}(1 - \frac{n}{N}) \\
Cov(Z_i, Zj) &= E(Z_i Z_j) - E(Z_i)E(Z_j) \\
&= \pi_{ij} - \pi_i^2 \\
&= -\frac{1}{N-1}\frac{n}{N}(1 - \frac{n}{N})
\end{aligned}$$

Parmi les estimateurs classiques, nous considérons les suivants : l'estimation par la moyenne, l'estimation par la différence, l'estimation par le ratio et l'estimation par grappe.

2.2.1.1 Estimation par la moyenne

Soit : .

$$\overline{x}_U = \frac{1}{N}\sum_{i\in U} x_i$$
$$\overline{y}_U = \frac{1}{N}\sum_{i\in U} y_i$$
$$\overline{d}_U = \overline{x}_U - \overline{y}_U = \frac{1}{N}\sum_{i\in U} d_i$$
$$\overline{x}_s = \frac{1}{n}\sum_{i\in s} x_i$$
$$\overline{y}_s = \frac{1}{n}\sum_{i\in s} y_i$$
$$\overline{d}_s = \overline{x}_s - \overline{y}_s = \frac{1}{n}\sum_{i\in s} d_i$$

Proposition 2.2.1. *Un estimateur par la moyenne pour t_d est donné par $\widehat{t}_{dm} = N\overline{d}_s$, $\widehat{t}_{dm}$ est sans biais pour t_d*

Démonstration.

$$\widehat{t}_{dm} = \frac{N}{n}\sum_{i\in s} d_i = \frac{N}{n}\sum_{i\in U} d_i Z_i$$
$$E(\widehat{t}_{dm}) = \frac{N}{n}\sum_{i\in s} d_i E(Z_i) = \frac{N}{n}\sum_{i\in U} d_i \pi_i$$
$$= \sum_{i\in U} d_i = t_d$$

□

Proposition 2.2.2. *La variance de $\widehat{t}_{dm}$ est donnée par*

$$V(\widehat{t}_{dm}) = N^2(1 - \frac{n}{N})\frac{S_d^2}{n}$$

où, $S_d^2 = \frac{1}{N-1}\sum_{i\in U}(d_i - \overline{d_U})^2$, désigne la dispersion de la variable d_i dans l'ensemble de la population

Démonstration.

$$\begin{aligned} V(\widehat{t}_{dm}) &= V\left(\frac{N}{n}\sum_{i\in s} d_i\right) = \frac{N^2}{n^2}V\left(\sum_{i\in U} d_i Z_i\right) \\ &= \frac{N^2}{n^2}\left(\sum_{i\in U} V(Z_i)d_i^2 + \sum_{i\in U}\sum_{\substack{j\in U \\ i\neq j}} Cov(Z_i, Z_j)d_i d_j\right) \\ &= \frac{N^2}{n}(1 - \frac{n}{N})\frac{1}{N-1}\left(\sum_{i\in U} d_i^2 - \frac{\left[\sum_{i\in U} d_i\right]^2}{N}\right) \\ &= N^2(1 - \frac{n}{N})\frac{S_d^2}{n} \end{aligned}$$

□

La dispersion des d_i dans la population est inconnue mais on peut l'estimer par $s_d^2 = \frac{1}{n-1}\sum_{i\in s}(d_i - \overline{d}_s)^2$ qui est un estimateur sans biais de S_d^2.

Sous la supposition que la variable $\frac{\overline{d}_s - \overline{d}_U}{\sqrt{(1-\frac{n}{N})\frac{s_d^2}{n}}}$ est de loi normale centrée réduite, une borne supérieure est donnée par $N\overline{d}_s + z_\alpha N\sqrt{1 - \frac{n}{N}}\, s_d/\sqrt{n}$.

Dans presque tout ce qui suit, nous supposons que N est assez grand pour nous permettre de négliger le facteur de correction $1 - \frac{n}{N}$.

2.2.1.2 Estimation par la différence

Si on remplace $\overline{x}_s$ dans l'expression de $\widehat{t}_{dm}$ par $\overline{x}_U$ qu'on connaît, on obtient alors l'expression de notre estimateur par la différence qui est donnée par

$$\widehat{t}_{diff} = N(\overline{x}_U - \overline{y}_s)$$

$\widehat{t}_{diff}$ est un estimateur sans biais pour t_d et sa variance estimée est donnée par

$$\widehat{V}(\widehat{t}_{diff}) = N^2(1 - \frac{n}{N})\frac{s_y^2}{n}$$

où $s_y^2 = \dfrac{1}{n-1}\sum_{i\in s}(y_i - \overline{y}_s)^2$
La borne supérieure est donnée par

$$N\widehat{t}_{diff} + z_\alpha N\sqrt{1 - \frac{n}{N}}\, s_y/\sqrt{n}$$

2.2.1.3 Estimation par le ratio

Soit le ratio de deux totaux donné par $B = \dfrac{t_d}{t_x}$.
Un estimateur naturel de B serait :

$$\widehat{B} = \frac{\widehat{t}_{dm}}{\widehat{t}_x}$$

où :

$$\widehat{t}_x = N\frac{\sum_{i=1}^{n} x_i}{n} = N\overline{x}_s$$

Malheureusement cet estimateur n'est pas sans biais pour B. En effet :

$$E(\widehat{B}) = E(\frac{\widehat{t}_{dm}}{\widehat{t}_x}) \neq \frac{E(\widehat{t}_{dm})}{E(\widehat{t}_x)} = \frac{t_{dm}}{t_x} = B$$

L'estimateur $\widehat{B}$ est un ratio de deux variables aléatoires. Trouver une expression exacte de son biais et de sa variance s'avère donc pratiquement impossible.
La linéarisation par séries de Taylor de la fonction $f(\widehat{t}_{dm}, \widehat{t}_x) = \frac{\widehat{t}_{dm}}{\widehat{t}_x} = \widehat{B}$ permettra de

donner les expressions approximatives du biais et de la variance de $\widehat{B}$. Nous résumons ici cette approche.

Linéarisation par séries de Taylor :

En général le développement par les séries de Taylor du premier ordre de $f(\widehat{t}_1, \widehat{t}_2, \cdots, \widehat{t}_q)$ est donné par

$$f(\widehat{t}_1, \widehat{t}_2, \cdots, \widehat{t}_q) = f(t_1, t_2, \cdots, t_q) + \sum_{i=1}^{q} \frac{\partial f(\widehat{t}_1, \widehat{t}_2, \cdots, \widehat{t}_q)}{\partial \widehat{t}_i}|_{\widehat{t}_1=t_1, \cdots, \widehat{t}_q=t_q}(\widehat{t}_i - t_i) + Reste$$

Le développement par séries de Taylor du deuxième ordre de $f(\widehat{t}_1, \widehat{t}_2, \cdots, \widehat{t}_q)$ est donné par

$$\begin{aligned} f(\widehat{t}_1, \widehat{t}_2, \cdots, \widehat{t}_q) = f(t_1, t_2, \cdots, t_q) &+ \sum_{i=1}^{q} \frac{\partial f(\widehat{t}_1, \widehat{t}_2, \cdots, \widehat{t}_q)}{\partial \widehat{t}_i}|_{\widehat{t}_1=t_1, \cdots, \widehat{t}_q=t_q}(\widehat{t}_i - t_i) \\ &+ \frac{1}{2!} \sum_{i=1}^{q} \sum_{j=1}^{q} \frac{\partial^2 f(\widehat{t}_1, \widehat{t}_2, \cdots, \widehat{t}_q)}{\partial \widehat{t}_i \partial \widehat{t}_j}|_{\widehat{t}_1=t_1, \cdots, \widehat{t}_q=t_q}(\widehat{t}_i - t_i)(\widehat{t}_j - t_j) + \text{Reste} \end{aligned}$$

Proposition 2.2.3. *Dans le cas d'un échantillonnage aléatoire simple sans remise, le biais approximatif de $\widehat{B}$ est donné par*

$$E(\widehat{B} - B) \approx (1 - \frac{n}{N}) \frac{1}{n\overline{x}_u^2}(BS_x^2 - RS_dS_x)$$

où $R = \dfrac{\sum_{i \in U}(d_i - \overline{d}_U)(x_i - \overline{x}_U)}{(N-1)S_dS_x}$ *désigne le coefficient de corrélation entre x_i et d_i*

Démonstration. Soit $f(\widehat{t}_{dm}, \widehat{t}_x) = \frac{\widehat{t}_{dm}}{\widehat{t}_x} = \widehat{B}$ alors

$$\begin{aligned}
\frac{\partial f(\widehat{t}_{dm}, \widehat{t}_x)}{\partial \widehat{t}_{dm}}|_{\widehat{t}_{dm}=t_{dm}, \widehat{t}_x=t_x} &= \frac{1}{t_x} \\
\frac{\partial f(\widehat{t}_{dm}, \widehat{t}_x)}{\partial \widehat{t}_x}|_{\widehat{t}_{dm}=t_{dm}, \widehat{t}_x=t_x} &= -\frac{t_{dm}}{t_x^2} = -\frac{B}{t_x} \\
\frac{\partial^2 f(\widehat{t}_{dm}, \widehat{t}_x)}{\partial^2 \widehat{t}_{dm}}|_{\widehat{t}_{dm}=t_{dm}, \widehat{t}_x=t_x} &= 0 \\
\frac{\partial^2 f(\widehat{t}_{dm}, \widehat{t}_x)}{\partial^2 \widehat{t}_x}|_{\widehat{t}_{dm}=t_{dm}, \widehat{t}_x=t_x} &= \frac{2t_{dm}}{t_x^3} \\
\frac{\partial^2 f(\widehat{t}_{dm}, \widehat{t}_x)}{\partial \widehat{t}_{dm} \partial \widehat{t}_x}|_{\widehat{t}_{dm}=t_{dm}, \widehat{t}_x=t_x} &= -\frac{1}{t_x^2}
\end{aligned}$$

Le développement par séries de Taylor du deuxième ordre de $f(\widehat{t}_{dm}, \widehat{t}_x)$ nous donne

$$\begin{aligned}
\widehat{B} &= B + \frac{1}{t_x}(\widehat{t}_{dm} - t_{dm}) - \frac{t_{dm}}{t_x^2}(\widehat{t}_x - t_x) + \frac{2t_{dm}}{2!t_x^3}(\widehat{t}_x - t_x)^2 - \frac{2}{2!t_x^2}(\widehat{t}_{dm} - t_{dm})(\widehat{t}_x - t_x) \\
\widehat{B} - B &= \frac{1}{t_x}(\widehat{t}_{dm} - t_{dm}) - \frac{B}{t_x}(\widehat{t}_x - t_x) + \frac{B}{t_x^2}(\widehat{t}_x - t_x)^2 - \frac{1}{t_x^2}(\widehat{t}_{dm} - t_{dm})(\widehat{t}_x - t_x)
\end{aligned}$$

Il s'ensuit que

$$\begin{aligned}
E(\widehat{B} - B) &\approx \frac{B}{t_x^2} E((\widehat{t}_x - t_x)^2) - \frac{1}{t_x^2} E\left[(\widehat{t}_{dm} - t_{dm})(\widehat{t}_x - t_x)\right] \qquad (\text{car } E(\widehat{t}_{dm} - t_{dm}) = E(\widehat{t}_x - t_x) = 0) \\
&= \frac{1}{t_x^2}\left[BV(\widehat{t}_x) - Cov(\widehat{t}_{dm}, \widehat{t}_x)\right]
\end{aligned}$$

On sait déjà que

$$V(\widehat{t}_x) = N^2\left(1 - \frac{n}{N}\right)\frac{S_x^2}{n}$$

Et on a

$$\begin{aligned}
Cov(\widehat{t}_{dm}, \widehat{t}_x) &= Cov(N\frac{\sum_{i\in s} d_i}{n}, N\frac{\sum_{i\in s} x_i}{n}) \\
&= \frac{N^2}{n^2} Cov(\sum_{i\in U} d_i Z_i, \sum_{j\in U} x_j Z_j) \\
&= \frac{N^2}{n^2}\left[\sum_{i\in U} V(Z_i) d_i x_i + \sum_{i\in U}\sum_{\substack{j\in U \\ i\neq j}} Cov(Z_i, Z_j) d_i x_j\right] \\
&= \frac{N}{n}(1-\frac{n}{N})\frac{1}{N-1}\left[N\sum_{i\in U} d_i x_i - \sum_{i\in U} x_i \sum_{j\in U} d_j\right] \\
&= \frac{N^2}{n}(1-\frac{n}{N})\frac{1}{N-1}\left[\sum_{i\in U} d_i x_i - N\overline{d}_U\,\overline{x}_U\right] \\
&= \frac{N^2}{n}(1-\frac{n}{N})\frac{1}{N-1}\left[\sum_{i\in U}(d_i-\overline{d}_U)(x_i-\overline{x}_U)\right] \\
&= N^2(1-\frac{n}{N})\frac{S_{dx}}{n}
\end{aligned}$$

où $S_{dx} = \dfrac{1}{N-1}\sum_{i\in U}(d_i-\overline{d}_U)(x_i-\overline{x}_U)$

et donc

$$\begin{aligned}
E(\widehat{B}-B) &\approx N^2(1-\frac{n}{N})\frac{1}{nt_x^2}(BS_d^2 - S_{dx}) \\
&= (1-\frac{n}{N})\frac{1}{n\overline{x}_U^2}(BS_x^2 - RS_dS_x) \qquad (\text{car } \overline{x}_U^2 = \frac{t_x^2}{N^2})
\end{aligned}$$

□

Proposition 2.2.4. *Dans le cas d'un échantillonnage aléatoire simple sans remise, la variance approximative estimée de $\widehat{B}$ est donnée par*

$$\widehat{V}(\widehat{B}) = (1-\frac{n}{N})\frac{1}{n\overline{x}_U{}^2}\frac{s_e^2}{n}$$

où $s_e^2 = \dfrac{1}{n-1}\sum_{i\in s}(d_i - \widehat{B}x_i)^2$

Démonstration. Soit $e_i = d_i - Bx_i$ et $\widehat{t_e} = N\dfrac{\sum_{i\in s} e_i}{n} = N\overline{e_s}$.
Le développement par séries de Taylor du premier ordre de $f(\widehat{t}_{dm}, \widehat{t}_x)$ nous donne

$$\begin{aligned}
\widehat{B} &= f(t_{dm}, t_x) + \frac{1}{t_x}(\widehat{t}_{dm} - t_{dm}) - \frac{t_y}{t_x^2}(\widehat{t}_x - t_x) \\
\widehat{B} - B &= \frac{1}{t_x}(\widehat{t}_{dm} - t_{dm}) - \frac{B}{t_x}(\widehat{t}_x - t_x) \\
&= \frac{1}{t_x}(\widehat{t}_{dm} - B\widehat{t}_x) \\
&= \frac{1}{t_x}\frac{N}{n}\sum_{i\in s}(t_{dm} - Bx_i) \\
&= \frac{1}{t_x}\frac{N}{n}\sum_{i\in s} e_i \\
&= \frac{\widehat{t}_e}{t_x}
\end{aligned}$$

et donc

$$\begin{aligned}
V(\widehat{B}) = V(\widehat{B} - B) &= \frac{1}{t_x^2}V(\widehat{t}_e) \\
&= N^2(1 - \frac{n}{N})\frac{S_e^2}{nt_x^2} \\
&= (1 - \frac{n}{N})\frac{S_e^2}{n\overline{x}_U^2} \qquad (\text{car}\overline{x}_U^2 = \frac{t_x^2}{N^2})
\end{aligned}$$

Où $S_e^2 = \dfrac{1}{N-1}\sum_{i\in U}(e_i - \overline{e_U})^2 = \dfrac{1}{N-1}\sum_{i\in U} e_i^2 = \dfrac{1}{N-1}\sum_{i\in U}(d_i - Bx_i)^2$, car $\overline{e}_U = \dfrac{1}{N}\sum_{i\in U}(d_i - Bx_i) = \dfrac{1}{N}(d_m - \dfrac{d_m}{t_x}t_x) = 0$.

□

La dispersion des e_i dans la population est inconnue mais on peut la remplacer par $s_e^2 = \dfrac{1}{n-1}\sum_{i\in s}(e_i - \overline{e}_s)^2$ qui est un estimateur sans biais de S_e^2 et on obtient alors la variance approximative estimée de $\widehat{B}$.

L'estimateur par le ratio de t_d est donné par $\widehat{t}_r = \widehat{B}t_x = \frac{\widehat{t}_{d_m}}{\widehat{t}_x}t_x$. Puisque t_x est une constante, le biais approximatif de $\widehat{t}_r$ est donné par

$$E(\widehat{t}_r - t_d) = N(1 - \frac{n}{N})\frac{1}{n\overline{x}_U^2}(BS_x^2 - RS_dS_x)$$

La variance approximative de $\widehat{t}_r$ est donnée par

$$\widehat{V}(\widehat{t}_r) = N^2(1 - \frac{n}{N})\frac{s_e^2}{n}$$

La borne supérieure est donnée par

$$\widehat{t}_r + z_\alpha N\sqrt{1 - \frac{n}{N}}\, s_e/\sqrt{n}$$

Remarque 2.1. L'estimateur par le ratio est fréquemment utilisé afin d'améliorer la précision des estimateurs. En effet, lorsque la variable d est approximativement proportionnelle à la variable x, l'estimateur par le ratio peut être considérablement plus précis que les estimateurs par la moyenne et par la différence. En fait, si $y_i = cx_i$, pour une constante donnée c, l'estimateur par le ratio est égale à $\widehat{t}_r = ct_x$, qui est une constante. Dans ce cas, la variance de l'estimateur par le ratio est égale à zéro. Bien sûr, il n'est pas réaliste de supposer que la variable d est parfaitement proportionnelle à la variable x. Il faut cependant s'attendre à ce que l'estimateur par le ratio soit très précis lorsque la variable d est approximativement proportionnelle à x.

2.2.2 Tirage avec probabilités inégales

Parfois il est plus judicieux de tirer les éléments avec probabilités proportionnelles à une mesure de leur taille. En fait, on s'attend à ce que l'erreur dans les comptes dont les valeurs sont grandes soit plus grande que celle des comptes dont les valeurs sont petites. Si on utilise l'échantillonnage aléatoire simple et si par exemple, notre population contient en majorité des comptes dont les valeurs sont petites, dans ce cas on s'attend à obtenir, la plupart du temps, un échantillon composé d'un grand nombre de comptes dont les valeurs sont petites, et par la suite, l'estimation de l'erreur totale dans la

population risque d'être en deçà de la vraie valeur. Par contre, si l'échantillon contient plusieurs comptes dont les valeurs sont grandes, alors l'estimation de l'erreur totale risque d'être bien au dessus de la vraie valeur. L'estimateur sera donc vraisemblablement instable. Par contre, si l'on tire un échantillon de telle manière que les comptes dont les valeurs sont grandes ont une grande probabilité d'être tirés dans l'échantillon, on s'attend à obtenir un estimateur plus stable.

2.2.2.1 Échantillonnage par grappes

Dans cette méthode, on considère le dollar comme unité d'échantillonnage secondaire dans un tirage par grappes dont l'unité primaire est un compte et donc la taille de la population est $t_x = \sum_{i=1}^{N} x_i$ et à chaque $j^{\text{ème}}$ élément dans la population des dollars ($j \in \{1; \cdots; t_x\}$) correspond un i ème élément dans la population des comptes.
Une façon d'accorder aux comptes l'importance qu'ils méritent, c'est de les tirer avec probabilités proportionnelles à leurs taille, où la « taille » est la valeur nominale du compte. Il est cependant extrêmement compliqué de le faire sans remise. Car pour qu'une telle procédure atteigne l'objectif visé, il faudrait que les π_i soient proportionnels aux x_i. Les différentes méthodes proposées pour le faire sont généralement compliquées et le calcul des π_{ij}, indispensables pour estimer la variance est également compliqué. Puisque nous supposons que N est grand comparé à n, nous allons nous limiter aux tirages avec remise.
À chaque tirage les résultats possibles sont $y_1, \cdots, y_n$ avec probabilités correspondantes $\Psi_1 = \frac{x_1}{t_x}, \cdots, \Psi_N = \frac{x_N}{t_x}$.

Soit z_i la variable aléatoire qui prend les valeurs $\frac{d_1}{x_1}, \cdots, \frac{d_N}{x_N}$, avec probabilitées respectives $\Psi_1 = \frac{x_1}{t_x}, \cdots, \Psi_N = \frac{x_N}{t_x}$, $\overline{z} = \frac{1}{n}\sum_{i=1}^{n} z_i$ $V(z_i) = S_z^2$. $\widehat{t}_{dr} = \frac{t_x}{n}\sum_{i=1}^{n} z_i$ est un estimateur sans biais de l'erreur totale .

Les z_i sont des variables indépendantes, alors

$$\begin{aligned}
E(\widehat{t}_{dr}) &= t_x E(z_i) = t_x \sum_{i=1}^{N} \frac{d_i}{x_i} \Psi_i \\
&= t_x \sum_{i=1}^{N} \frac{d_i}{x_i} \frac{x_i}{t_x} = \sum_{i=1}^{N} d_i \\
&= t_d \\
V(\widehat{t}_{dr}) &= \frac{t_x^2}{n} [\sum_{i=1}^{N} \frac{d_i^2}{x_i^2} \frac{x_i}{t_x} - (\sum_{i=1}^{N} \frac{d_i}{x_i} \frac{x_i}{t_x})^2] \\
&= \frac{t_x^2}{n} [\sum_{i=1}^{N} \frac{d_i^2}{x_i t_x} - (\sum_{i=1}^{N} \frac{d_i}{t_x})^2] \\
&= \frac{1}{n} [\sum_{i=1}^{N} \frac{d_i^2 t_x}{x_i} - (\sum_{i=1}^{N} d_i)^2] \\
&= \frac{1}{n} \sum_{i=1}^{N} [(\frac{d_i^2}{\Psi_i}) - t_d^2] \\
&= \frac{1}{n} \sum_{i=1}^{N} [\Psi_i (\frac{d_i}{\Psi_i} - t_d)^2]
\end{aligned}$$

Et donc la borne supérieure est donnée par $\widehat{t}_{dr} + z_\alpha \sqrt{V(\widehat{t}_{dr})}$.

2.3 Méthodes spécifiques

Parmi les méthodes spécifiques, nous considérons les suivantes : la méthode par attributs, la stratification, la méthode combinée, la méthode suggérée par Bickel, la méthode de Stringer, la méthode de Hoeffding et la méthode basée sur la loi de Kolmogrov-Smirnov.

2.3.1 Tirage avec probabilités égales

La méthode par attributs, la stratification, la méthode combinée et la méthode suggérée par Bickel sont employées lors d'un tirage avec probabilités égales.

2.3.1.1 Méthode par attributs

Ce qu'en vérification on appelle « échantillonnage d'attributs » est en fait un échantillonnage destiné à estimer une *proportion*. Appliquée au problème qui nous concerne, cette méthode consiste à concentrer les efforts sur l'estimation de la proportion π des d_i non nuls. Puisque le paramètre à estimer est $N\pi\mu_1$, où μ_1 est la moyenne des d_i non nuls dans la population, on utilise des moyens statistiques pour estimer π et on substitue à μ_1 une borne supérieure découlant de diverses considérations contextuelles.

Rappelons les hypothèses à la base de plusieurs des méthodes proposées.
On suppose d'abord que l'erreur est toujours une erreur de *surestimation*. C'est à dire, que $x_i - y_i \geq 0$ pour tout i. Les montants en question ne sont jamais négatifs, ce qui entraîne que $x_i - y_i \leq x_i$. Soit x_l le plus grand montant parmi les comptes comptables, alors

$$d_i = x_i - y_i \leq x_l \qquad \forall i = 1, \cdots, N$$

Soit π la proportion des comptes dans la population qui contient des erreurs de *surestimation*, alors

$$t_d = \sum_{i=1}^{N} (x_i - y_i) \leq N\pi x_l$$

On note par $B = N\pi x_l$ la borne supérieure de l'erreur totale de *surestimation*. N et x_l sont connues, π est inconnue, donc construire un intervalle de confiance pour t_d revient à construire un intervalle de confiance pour π.
Comme $\pi \geq 0$, alors l'intervalle à $(1 - \alpha)\%$ de confiance pour π serait de la forme $[0, P_U(k, 1 - \alpha)]$, où k est le nombre d'erreurs de *surestimation* trouvé dans l'échantillon et $P_U(k, 1 - \alpha)$ est la valeur de π pour laquelle $P(X \leq k|\pi) = \alpha$.

Cette méthode se justifie par l'argument suivant :
Soit X une variable aléatoire dont la distribution dépend d'un paramètre θ, de telle sorte que pour toute valeur x de X, $P(x \leq x|\theta)$ est une fonction décroissante de θ. Alors l'intervalle $[0; \theta_s(x)]$ est un intervalle de confiance à $100(1-\alpha)\%$ lorsque $\theta_s(x)$ est la solution de $P(X \leq x|\theta) = \alpha$.

Démonstration. Soit θ_0 la vraie valeur de θ. Alors $\theta_0 \in [0; \theta_s(x)]$ si et seulement si $P(X \leq x|\theta_0) > \alpha$. La probabilité de l'ensemble $\{x|P(X \leq x|\theta_0) > \alpha\}$ est alors supérieure ou égale à $1-\alpha$, ce qui complète la démonstration. □

On note par $\widehat{B} = NP_U(k, 1-\alpha)x_l$ la borne supérieure de l'erreur totale de *surestimation* à $(1-\alpha)\%$ de confiance. La valeur de $P_U(k, 1-\alpha)$ change selon le plan de sondage utilisé.

On voit immédiatement ce qui s'avère être le plus grand défaut de cette approche : la borne supérieure sera extrêmement conservatrice à moins que les valeurs de x soient très peu dispersées.

Échantillonnage aléatoire simple avec remise :

Soit N le nombre total de comptes dans la population et à partir de cette population, on tire un échantillon aléatoire simple sans remise de taille n. Dans cet échantillon, on observe k erreurs de *surestimation*. Soit X la variable aléatoire qui prend la valeur 1 s'il y a une erreur de *surestimation*, 0 sinon. Donc X suit une loi binomiale de paramètres $(n; p)$, où p est inconnue.

Alors pour trouver $\widehat{B}$, dans le cas d'un échantillonnage aléatoire simple avec remise, on aurait besoin de méthodes numériques pour trouver $P_U(k, 1-\alpha)$, telle que

$$P(X \leq k|P_U(k, 1-\alpha)) = \alpha$$

Échantillonnage aléatoire simple sans remise :

Il'est relativement aisé d'adopter cette approche au tirage aléatoire simple sans remise.

1) Le cas ou N est petit :

Dans ce cas X suit une loi hypergéométrique $H(n, N_1, N_2)$ où

n : la taille de l'échantillon tiré

N_1 : le nombre de comptes dans la population qui ont des erreurs de *surestimation*

N_2 :le nombre de comptes dans la population qui n'ont pas d'erreurs de *surestimation*

La *proportion* des comptes qui ont des erreurs de surestimation dans la population est de la forme $\pi = \frac{N_1}{N}$ et puisque N_1 ne prend que des valeurs comprises entre 0 et N, il est donc possible de les essayer toutes pour savoir celles qui vont faire partie de l'intervalle de confiance.

On sait déjà que $P(X \leq k|p)$ est une fonction décroissante de p et, soit N_1' la plus grande valeur de N_1 telle que $P_U(k, 1-\alpha) = P(X \leq k|\frac{N_1}{N}) \geq \alpha$, alors toutes les valeurs de p telle que $p \leq P_U(k, 1-\alpha)$ sont contenues dans l'intervalle de confiance car dans ce cas, $P(X \leq k|p) \geq \alpha$.

Les valeurs de p telle que $p > P_U(k, 1-\alpha)$ ne feront pas partie de l'intervalle de confiance , car dans ce cas, $P(X \leq k|p) < \alpha$ et par la suite, on rejette l'hypothèse que $p \leq P_U(k, 1-\alpha)$.

2) Le cas ou N est grand :

En général, si $f = \frac{n}{N} < 0,1$, on considère alors dans ce cas que N est grande par rapport à n et par la suite, la loi hypergéométrique peut être approchée par la loi binomiale.

Souvent on préfère les calculs par la loi binomiale à ceux de la loi hypergéométrique car cette dernière exige la connaissance de N, ce qui n'est pas toujours le cas, et aussi parce que les calculs utilisés pour construire des intervalles de confiance sont plus faciles avec la loi binomiale.

Cependant, la loi binomiale présente des difficultés particulières, dans la mesure où, le paramètre p étant un paramètre continu, l'approche qui consisterait à essayer toutes les valeurs possibles est exclue.
Nous devons donc recourir à des méthodes numériques, telle que la méthode de Newton-Raphson, pour résoudre l'équation $P(X \leq k|p) = \alpha$.

2.3.1.2 Échantillonnage stratifié

Comme nous l'avons noté la méthode d'attributs est extrêmement conservatrice, en particulier lorsque la population est très grande et x_l est excessivement grand par rapport aux autres valeurs. Une solution consiste à stratifier la population, de façon à ce que la borne supérieure soit moins élevée dans chaque strate, car elle permet de réduire les tailles des populations ainsi que de modérer la surestimation que constitue x_l.

L'échantillonnage stratifié consiste à répartir la population en strates et dans chacune des strates, on tire un échantillon selon un plan de sondage donné qui sera, dans la

plupart des cas, un échantillonnage aléatoire simple sans remise ou avec remise.

Soit N_h le nombre total des comptes dans la strate h tel que $h = \{1; \cdots; H\}$ et $\sum_{h=1}^{H} N_h = N$ et soit x_{hl} le plus grand montant parmi les comptes comptables dans la strate h et p_h la *proportion* des comptes dans la strate h qui contient des erreurs de *surestimation*. Alors, la borne supérieure de l'erreur de *surestimation* devient

$$B = \sum_{h=1}^{H} N_h p_h x_{hL}$$

Soit n_h la taille de l'échantillon tiré dans la strate h, k_h le nombre d'erreurs de *surestimation* trouvées dans cet échantillon. Si on suppose que les tailles des échantillons tirés dans les strates sont grandes et que les *proportions* p_h sont petites, alors k_h peut être considérée comme une variable de Poisson de paramètre $\lambda_h = n_h p_h$.

Puisque les échantillons tirés dans chacune des strate sont indépendants, alors $k = \sum_{h=1}^{H} k_h$ suit une loi de Poisson de paramètre $\lambda = \sum_{h=1}^{H} n_h p_h$.
On peut déterminer la stratification de telle sorte que B soit un multiple connu de k. Il suffit de choisir les N_h strates de telle sorte que $n_h = \frac{N_h x_{hl}}{\sum_{h=1}^{H} N_h x_{hl}} n$. Dans ce cas, $B = \frac{\sum_{h=1}^{H} N_h x_{hl}}{n} \lambda$ et il suffira alors de déterminer une borne supérieure pour λ. En effet, si les strates sont construites de façon étroite, c'est-à-dire que dans chacune des strates il n'y a pas une grande différence entre les montants des comptes comptables, alors la probabilité qu'un compte de la strate h soit tiré dans l'échantillon s_h est

$$\frac{n_h}{N_h} = \frac{n x_{hL}}{\sum_{h=1}^{H} N_h x_{hl}}$$

On dit alors que les comptes sont tirés approximativement avec une probabilité proportionnelle aux montants des comptes et donc la taille de l'échantillon n_h tiré dans la

strate h devrait satisfaire l'égalité

$$n_h = \frac{N_h x_{hl}}{\sum_{h=1}^{H} N_h x_{hl}} n$$

Et par la suite on aura

$$\begin{aligned}
\lambda &= \sum_{h=1}^{H} n_h p_h \\
&= \frac{\sum_{h=1}^{H} N_h x_{hl} p_h}{\sum_{j=1}^{H} N_j x_{jl}} n \\
&= \frac{n}{\sum_{j=1}^{H} N_j x_{jl}} \sum_{h=1}^{H} N_h x_{hl} p_h \\
&= \frac{n}{\sum_{h=1}^{H} N_h x_{hl}} B \\
\Rightarrow B &= \frac{\sum_{h=1}^{H} N_h x_{hl}}{n} \lambda
\end{aligned}$$

N_h, n et x_{hl} sont connues, donc trouver une borne supérieure de confiance pour t_d revient à trouver une bande supérieure de confiance pour λ .

Si $[0, \lambda_U(k, 1-\alpha)]$ est l'intervalle avec $(1-\alpha)\%$ de confiance pour λ, alors par analogie avec les résultats précédents, $\lambda_U(k, 1-\alpha)$ est la valeur telle que $P(X \leq k | \lambda_U(k, 1-\alpha)) = \alpha$.

2.3.1.3 Estimation par la méthode combinée

La méthode combinée est un mélange d'échantillonnage d'attributs et de variables, le but de cette méthode étant de pallier le défaut majeur des méthodes d'attributs, à savoir

l'extrême conservatisme .

Posons $\pi = P(d > 0)$, $\mu = E(d)$, $\sigma_d^2 = V(d)$, $\mu_1 = E(d|d > 0)$, $S_{d_1}^2 = V(d|d > 0)$.

Et on a :

$$E(d) = \sum_{i \in U} \frac{d_i}{N} = \frac{t_d}{N}$$

$$\Longrightarrow t_d = N\mu = N\pi\mu_1$$

Construire une borne supérieure pour t_d revient à trouver une borne supérieure pour π et μ_1.

Soit k le nombre de comptes qui contiennent des erreurs de *surestimation* dans l'échantillon et $\overline{d}_1 = \sum_{i=1}^{n} \frac{d_i}{k}$ est la moyenne des différences non nulles dans l'échantillon. Nous supposons ici que la probabilité que k soit nul est négligeable.

$\overline{d}_1$ est alors conditionnellement sans biais pour μ_1.

$$E(\overline{d}_1|k > 0) = \mu_1$$

$$V(\overline{d}_1|k > 0) = \frac{S_{d_1}^2}{k}$$

Une borne supérieure pour μ_1 est donnée par $\overline{d}_1 + z_\alpha \frac{s_{d_1}}{\sqrt{k}}$ où $s_{d_1}^2$ est la variance échantillonnale des données non nulles qui est un estimateur sans biais de $S_{d_1}^2$. Notons cependant que $s_{d_1}^2$ n'existe que lorsque $k \geq 2$

$P_u(k, 1-\alpha)$ est la borne supérieure pour π, obtenue en utilisant la méthode d'attributs.

La borne supérieure de confiance pour t_d est alors donnée par :

$$NP_u(k, 1-\alpha)(\overline{d}_1 + z_\alpha \frac{s_{d_1}}{\sqrt{k}})$$

Une borne supérieure pour π peut aussi être obtenue en utilisant l'estimateur d'une

proportion $\widehat{\pi} = \frac{k}{n}$. Cet estimateur est sans biais pour π et sa variance est donnée par $\sigma^2_{\widehat{\pi}} = \frac{\widehat{\pi}(1-\widehat{\pi})}{n}$.

Dans ce cas, la borne supérieure de confiance est alors donnée par

$$N(\widehat{\pi} + z_1\sigma_{\widehat{\pi}})(\overline{d}_1 + z_2\frac{s_{d_1}}{\sqrt{k}})$$

où z_1, z_2 sont définis par

$z_1 = \Phi^{-1}(\sqrt{1-\alpha_1})$, $z_2 = \Phi^{-1}(\sqrt{1-\alpha_2})$

Φ est la fonction de répartition d'une normale $N(0,1)$ et $\sqrt{1-\alpha_1}\sqrt{1-\alpha_2} = 1-\alpha$

Cette approche se justifie par l'argument suivant

On a $P(\widehat{\pi} + z_1\sigma_{\widehat{\pi}} > \pi) \approx \sqrt{1-\alpha_1}$

$P(\overline{d_1} + z_2\frac{s_{d_1}}{\sqrt{k}} > \mu_1) \approx \sqrt{1-\alpha_2}$

et comme $\widehat{\pi}$ et μ_1 sont indépendants alors

$$P(\widehat{\pi} + z_1\sigma_{\widehat{\pi}} > \pi \& \overline{d_1} + z_2\frac{s_{d_1}}{\sqrt{k}} > \mu_1) = P(\widehat{\pi} + z_1\sigma_{\widehat{\pi}} > \pi)P(\overline{d_1} + z_2\frac{s_{d_1}}{\sqrt{k}} > \mu_1)$$

$$\Rightarrow P(\widehat{\pi} + z_1\sigma_{\widehat{\pi}}\overline{d_1} + z_2\frac{s_{d_1}}{\sqrt{k}} > \pi\mu_1) \approx \sqrt{1-\alpha_1}\sqrt{1-\alpha_2} = 1-\alpha$$

2.3.1.4 La méthode suggérée par Bickel

L'estimateur par la moyenne $\widehat{t}_{d_m} = N\overline{d}_s$ qui est un estimateur sans biais de t_d est égal aussi à $N\widehat{\pi}\overline{d}_1$.

La variance de $N\widehat{\pi}\overline{d}_1$ est donnée par

$$\begin{aligned} V(\widehat{\pi}\overline{d}_1) &= E[V(\widehat{\pi}\overline{d}_1|\widehat{\pi})] + V[E(\widehat{\pi}\overline{d}_1|\widehat{\pi})] \\ &= E[\widehat{\pi}^2 V(\overline{d}_1|\widehat{\pi})] + V[\widehat{\pi}E(\overline{d}_1|\widehat{\pi})] \\ &= E[(\frac{k}{n}^2)\frac{S_{d_1}^2}{k}] + V[\widehat{\pi}\mu_1] \\ &= \frac{S_{d_1}^2}{k}E[\widehat{\pi}] + V[\widehat{\pi}\mu_1] \\ &= \frac{\pi}{n}S_{d_1}^2 + {\mu_1}^2\frac{\pi(1-\pi)}{n} \\ &= \frac{\pi}{n}[S_{d_1}^2 + \mu_1^2(1-\pi)] \end{aligned}$$

On remplace $S_{d_1}^2$ par $s_{d_1}^2$ et π par la borne supérieure $P_u(k, 1-\alpha)$ pour obtenir la quantité suivante comme borne supérieure de la variance de l'estimateur $\widehat{V}(\widehat{\pi}\overline{d}_1) = \frac{P_u(k,1-\alpha)}{n}[s_{d_1}^2 + {\mu_1}^2(1 - P_u(k, 1-\alpha))]$.

La borne supérieure pour le total est alors $N[\widehat{\pi}\overline{d}_1 + z_\alpha\sqrt{\widehat{V}(\widehat{\pi}\overline{d}_1)}]$

2.3.2 Tirage avec probabilités inégales

La méthode de Stringer, la méthode de Hoeffding,la méthode basée sur la loi multinomiale et la méthode basée sur la loi de Kolmogrov-Smirnov sont utilisées lors d'un tirage avec probabilités inégales.

2.3.2.1 La méthode de Stringer

La méthode par attributs tend à être très conservatrice du fait que toute information concernant la valeur monétaire des erreurs dans l'échantillon est inutilisée : ces valeurs sont remplacées par une valeur maximale qui peut être très éloignée de ce qui devrait être la moyenne.

La méthode de Stringer tente d'utiliser des valeurs observées dans l'échantillon.

Soit $z_1 \geq z_2 \geq \cdots \geq z_k$ les k valeurs non nulles observées dans l'échantillon.

La borne supérieure obtenue par la méthode de Stringer est donnée par

$$t_x P_U(0, 1-\alpha) + t_x \sum_{j=1}^{k} [P_U(j, 1-\alpha) - P_U(j-1, 1-\alpha)] z_j$$

Une façon d'expliquer la vraie motivation de Stringer, qui est inconnue, est la suivante. Soit G est la fonction de répartition de z étant donné $z > 0$, $\pi = p(z > 0)$ et $\mu = E(z) = P(z > 0)E(z|z > 0) = \pi \int_0^1 z g(z) dz$.
On suppose que $U = z$ et $V = G(z)$, alors en faisant une intégration par parties, on obtient que

$$\begin{aligned}
\mu &= \pi \Big[[zG(z)]_0^1 - \int_0^1 G(z) dz \Big] \\
&= \pi \Big[G(1) - \int_0^1 G(z) dz \Big] \\
&= \pi \Big[1 - \int_0^1 G(z) dz \Big] \\
&= \pi \int_0^1 [1 - G(z)] dz \\
&= \pi \sum_{j=0}^{k} \int_{z_{j+1}}^{z_j} [1 - G(z)] dz \\
&\leq \pi \sum_{j=0}^{k} \int_{z_{j+1}}^{z_j} [1 - G(z_{j+1})] dz \qquad \text{car } G \text{ est croissante} \\
&\leq \pi \sum_{j=0}^{k} (z_j - z_{j+1})(1 - G(z_{j+1}))
\end{aligned}$$

Maintenant on suppose que n_j est le nombre d'observations supérieures à z_j, donc $x_j \sim B(n; \pi_j)$, où $\pi_j = P(z \geq z_{j+1}) = P(z > 0)P(z \geq z_{j+1}|z > 0) = \pi[1 - G(z_{j+1})]$. On a observé $n_j = j$ la borne supérieure de π_j étant donnée par $P_U(j; 1-\alpha)$ et elle vérifie $P(n_j \leq j | P_U(j; 1-\alpha)) = \alpha$. Si on remplace chaque π_j par $P_U(j; 1-\alpha)$ on obtient la borne supérieure pour μ qu'on multiplie par t_x pour obtenir la borne supérieure.

2.3.2.2 La méthode basée sur la loi multinomiale

La borne supérieure basée sur la méthode d'attributs est conservatrice car elle ne tient pas compte de l'ampleur des erreurs. C'est pour cela que Neter et Leich et Fienberg dans l'article « Estimating the Total Overstatement Error in Accounting Populations » paru dans le journal « Journal of the American Statistical Association, Vol .72, No. 358 (Jun, 1977), pp. 295-302 » ont proposé de calculer la borne supérieure en se basant sur la loi multinomiale qui, comme la méthode de Stringer, tient compte de la grandeur des erreurs dans la population. En plus, l'approche par la loi multinomiale est essentiellement non paramétrique : sa distribution et ses propriétés ne dépendent pas de la distribution des erreurs dans la population , chose qui n'est pas acquise lorsqu'on utilise la méthode de Stringer.

Pour la simplicité des calculs, on va considérer le dollar comme unité d'échantillonnage et on suppose que le nombre de dollars dans la population est très grand par rapport au nombre de dollars dans l'échantillon.

Chaque dollar tiré dans l'échantillon correspond à une erreur relative qui est considérée comme une variable discrète dont les valeurs sont $0, 1, \cdots, 100$ et qui peut être placé dans l'une des 101 catégories . Soit $\boldsymbol{p} = [p_0; p_1; \cdots; p_{100}]$ le vecteur des proportions correspondant aux erreurs relatives, avec $p_i \geq 0$ et $\sum_{i=1}^{100} p_i = 1$. Supposons qu'on tire un échantillon aléatoire simple de taille n, soit w_i le nombre d'erreurs trouvées dans l'échantillon correspondant à la catégorie i. La distribution du vecteur $\boldsymbol{w} = [w_0; w_1; \cdots; w_{100}]$ est multinomiale de paramètres n et p_i. L'erreur totale de surestimation peut s'écrire de la façon suivante

$$t_d = \frac{t_x}{100} \sum_{i=1}^{100} i p_i \tag{2.3.1}$$

La borne supérieure pour t_d peut être calculée en se basant sur la loi multinomiale car t_d est une combinaison linéaire du paramètre multinomiale p_i. L'approche consiste tout d'abord à trouver une ensemble de confiance pour les paramètres p_i basé sur les effectifs observés w_i. Cet ensemble est solution de l'équation

$$\sum_{S} \frac{n!}{\delta_0!\delta_1!\cdots\delta_{100}!} \prod_{i=0}^{100} p_i^{\delta_i} = \alpha \tag{2.3.2}$$

où S est l'ensemble des valeurs $\delta_0, \cdots, \delta_{100}$ telle que $\frac{\sum_{i=1}^{100} i\delta_i}{n} \leq \frac{\sum_{i=1}^{100} iw_i}{n}$.

La deuxième étape est de trouver l'ensemble des paramètres p_i qui appartiennent à la région de confiance et qui maximisent t_d.
Formellement cette approche consiste à maximiser

$$t_d = \frac{t_x}{100} \sum_{i=0}^{100} ip_i \tag{2.3.3}$$

sous les contraintes suivantes

$$\sum_{S} \frac{n!}{\delta_0!\delta_1!\cdots\delta_{100}!} \prod_{i=0}^{100} p_i^{\delta_i} = \alpha$$

$$p_i \geq 0 \ , \ i = 0, 1, \cdots, 100$$

$$\sum_{i=0}^{100} p_i = 1$$

Dans la plupart des cas, les échantillons tirés contiennent des comptes où le nombre d'erreurs est très proche ou égal à zéro, ce qui veut dire que w_0 est proche de n et la plupart des w_i, $i = 1, 2, \cdots, 100$, seront alors égales à zéro, nous facilitant ainsi la tâche de trouver l'ensemble S et maximiser t_d. Dans ce qui suit, on va traiter les cas où on se retrouve avec zéro erreur, une erreur, deux erreurs ou plus.

Premier cas : zéro erreur

Si on ne trouve pas d'erreur dans l'échantillon, alors $w_0 = n$ et $w_i = 0 \quad \forall i = 1, 2, \cdots, 100$.
Dans ce cas, on doit maximiser t_d sous l'hypothèse que $p_0^n = \alpha$.

Il est clair que le maximum est atteint lorsqu'on prend

$$\widehat{p}_0 = \alpha^{1/n}$$
$$\widehat{p}_i = 0 \quad \forall i = 1, \cdots, 99$$
$$\widehat{p}_{100} = 1 - \widehat{p}_0 = P_U(0; 1-\alpha),$$

où $P_U(0; 1-\alpha)$ est une borne supérieure pour la proportion d'une binomiale lorsqu'on n'observe aucun succès parmi n tirages .
Donc la borne supérieure lorsqu'il ny'a aucune erreur dans l'échantillon est

$$D \leq (t_x/100)[100P_U(0; 1-\alpha)] = t_x P_u(0, 1-\alpha) \tag{2.3.4}$$

Deuxième cas : une erreur

Supposons maintenant qu'on a trouvé une seule erreur de grandeur c dans l'échantillon, alors $w_0 = n-1$, $w_c = 1$ et les w_i restant sont égales à zéro. Dans ce cas S est l'ensemble des éléments tel que

$$\sum_{i=1}^{100} i\delta_i \leq c \tag{2.3.5}$$
$$\sum_{i=1}^{100} \delta_i \leq 1$$

La première inégalité exige que la somme des valeurs des comptes qui contiennent des erreurs et qui sont inclus dans S ne doivent pas dépasser la valeur de l'erreur observée. L'utilisation de la première inégalité toute seule est la façon la plus rigoureuse pour déterminer S mais c'est très compliqué de déterminer cet ensemble dans ce cas. Par exemple, si on suppose que l'erreur dans l'échantillon qu'on a tiré est de 60, alors l'ensemble S ne va pas contenir seulement les échantillons avec une erreur de $60, 59, \ldots\ldots 1$, mais aussi les échantillons avec deux erreurs de 59 et 1, 58 et 2, etc., les échantillons

avec trois erreurs comme 55, 3 et 2 et ainsi de suite. La seconde inégalité exige que les échantillons qui sont inclus dans S contiennent au plus une erreur.

Les contraintes dans (1.6) impliquent qu'on doit maximiser t_d sous la contrainte

$$p_0^n + np_0^{n-1}(p_1 + \cdots + p_c) = \alpha \tag{2.3.6}$$

ou tout simplement

$$p_0^{n-1}(p_0 + n\sum_{i=1}^{c} p_i) = \alpha \tag{2.3.7}$$

avec $p_i \geq 0$ et $\sum_{i=0}^{100} p_i = 1$

Si $c = 100$, la contrainte qu'on doit maximiser devient alors

$$p_0^n + np_0^{n-1}(1 - p_0) = \alpha \tag{2.3.8}$$

La solution de 1.9 est donnée par $\widehat{p}_{100} = P_U(1, 1-\alpha)$ qui est la borne supérieure pour la proportion d'une binomiale lorsqu'on observe 1 succès parmi n tirages. Connaissant $\widehat{p}_{100}$, la maximisation de 1.2 est obtenue facilement en prenant $\widehat{p}_0 = \widehat{p}_1 = \cdots = \widehat{p}_{99} = 0$ et $\widehat{p}_0 = 1 - \widehat{p}_{100}$. Donc, si $c = 100$, la borne supérieure de confiance est alors donnée par

$$t_d \leq \frac{t_x}{100}[100P_U(1; 1-\alpha)] = t_x P_U(1; 1-\alpha)$$

Maintenant en supposant que $c < 100$, la solution serait donnée en prenant $\widehat{p}_0 = \widehat{p}_1 = \cdots = \widehat{p}_{c-1} = \widehat{p}_{c+1} = \widehat{p}_{99} = 0$ et $\widehat{p}_{100} = 1 - \widehat{p}_0 - \widehat{p}_c$, alors notre problème serait réduit à maximiser

$$dp_c + 100p_{100} = 100(1 - p_0) - (100 - d)p_c \tag{2.3.9}$$

sous les hypothèses que

$$p_0^{n-1}(p_0 + np_c) = \alpha \tag{2.3.10}$$

$$p_0 + p_c + p_{100} = 1$$

$$p_0,\ p_c,\ p_{100} \geq 0$$

Les équations (1.10) et (1.11) ont réduit le problème multinomiale à une trinomiale dans ce cas. La solution de la première contrainte dans (1.11) est donnée par $\widehat{p}_c = 0$ ou bien

$\widehat{p}_{100} = 0$. Dans le premier cas, on aura $\widehat{p}_0^n = \alpha$, $\widehat{p}_d = 0$ et $\widehat{p}_{100} = 1 - \widehat{p}_0$ et par la suite, la borne supérieure sera la même que celle dans (1.5). Dans le deuxième cas, $\widehat{p}_0 = 1 - \widehat{p}_c$ est la solution de l'équation 1.9.

Donc

$$P_L(n-1; 1-\alpha) \leq \widehat{p}_0 \leq P_L(n; 1-\alpha) \tag{2.3.11}$$

On peut maximiser l'équation 1.10 sous la première contrainte de (1.11) en utilisant la méthode de multiplicateur de Lagrange.

Soit

$$L = 100(1-p_0) - (100-d)p_c + \lambda(p_0^n + np_0^{n-1}p_c - \alpha)$$

et si on suppose que

$$\begin{aligned}
\frac{\partial L}{\partial p_0} &= -100 + \lambda n \widehat{p_0}^{n-1}\Big(1 + \frac{(n-1)\widehat{p}_c}{\widehat{p}_0}\Big) = 0 \\
\frac{\partial L}{\partial p_c} &= -(100-c) + \lambda n \widehat{p_0}^{n-1} = 0 \\
\frac{\partial L}{\partial p_c} &= \widehat{p_0}^{\,n} + n\widehat{p_0}^{n-1}\widehat{p}_c - \alpha = 0
\end{aligned}$$

alors

$$\begin{aligned}
\widehat{p}_c &= \frac{1}{n}\Big(\frac{\alpha}{\widehat{p}_0^{n-1}} - \widehat{p}^0\Big) \\
\widehat{p}_0 &= \left(\frac{\alpha}{1 + \frac{dn}{(100-c)(n-1)}}\right)
\end{aligned}$$

et sachant que

$$\widehat{p}_0 \geq P_L(n-1; 1-\alpha)$$

on aura alors

$$\widehat{p}_0 = \max\left[\left(\frac{\alpha}{1 + \frac{cn}{(100-d)(n-1)}}\right)^{1/n}, P_L(n-1; 1-\alpha)\right]$$

si $\widehat{p}_0 = P_L(n-1; 1-\alpha)$ alors la borne supérieure de confiance sera donnée par

$$t_d \leq t_x \frac{d}{100}[1 - P_L(n-1; 1-\alpha)]$$

si $\widehat{p}_0 > P_L(n-1; 1-\alpha)$ alors la borne supérieure de confiance sera donnée par

$$t_d \leq \frac{t_x}{100}\Big[100 - \alpha^{1/n}\Big(100 + \frac{c}{n-1}\Big)^{(n-1)/n}(100-c)^{1/n}\Big]$$

Troisième cas : deux erreurs ou plus

La détermination de la borne supérieure de confiance devient plus difficile lorsque le nombre d'erreurs trouvé dans l'échantillon est supérieur ou égal à deux. Supposons qu'on observe deux erreurs de grandeurs respectives c_1 et c_2 , avec $c_1 \geq c_2$. Si $c_1 \neq c_2$, alors $w_{c_1} = w_{c_2} = 1$ et tout les autres w_i sont égales à zéro. Si $c_1 = c_2$, alors $w_0 = n-2$, $w_{c_1} = 2$ et tout les autres w_i sont égales à zéro. Dans ce cas, S est l'ensemble des éléments tel que

$$\begin{aligned} &\sum_{i=1}^{100} i\delta_i \leq c_1 + c_2 \\ &\sum_{i=1}^{100} \delta_i \leq 2 \end{aligned} \tag{2.3.12}$$

Les contraintes dans 1.13 impliquent qu'on doit maximiser t_d sous la contrainte

$$p_0^n + np_0^{n-1}(p_1 + \cdots + p_{c_1} + p_{c_2}) + \binom{n}{2} p_0^{n-2} \prod\nolimits^{*} p_i p_j = \alpha$$

où $\prod^{*} p_i p_j$ est le produit des proportions qui vérifient $2 \leq i+j \leq \min(100, c_1+c_2)$. Il est très difficile de maximiser t_d sous cette contrainte alors, une autre façon pour résoudre le problème est de considérer seulement les erreurs qui ne dépassent pas celles observées dans l'échantillon, et dans ce cas, le problème sera de maximiser t_d sous la contrainte

$$\begin{aligned} p_0^n + np_0^{n-1}(p_1 + \cdots + p_{c_1}) + \binom{n}{2} p_0^{n-2}(p_1 + \cdots + p_{c_2}) \\ [(p_1 + \cdots + p_{c_2}) + 2(p_{c_2+1} + \cdots + p_{c_1})] = \alpha \end{aligned} \tag{2.3.13}$$

Il serait alors plus facile de maximiser t_d en prenant, $\widehat{p}_1 = \widehat{p}_2 = \cdots = \widehat{p}_{c_2-1} = \widehat{p}_{c_2+1} = \cdots = \widehat{p}_{c_1-1} = 0$, et donc (1.14) devient

$$p_0^n + np_0^{n-1}(p_{c_2} + p_{c_1}) + \binom{n}{2} p_0^{n-2}(p_{c_2} + 2p_{c_1})p_{c_2} = \alpha \tag{2.3.14}$$

Dans le cas de deux erreurs il n'est pas facile de trouver une solution explicite pour maximiser t_d, cependant on peut utiliser des méthodes numériques pour trouver la borne supérieure en tenant compte des contraintes suivantes

$$\begin{aligned} P_L(n-2; 1-\alpha) \leq \widehat{p}_0 &\leq P_L(n; 1-\alpha) \\ 0 \leq \widehat{p}_{c_1} &\leq P_U(1; 1-\alpha) \\ 0 \leq \widehat{p}_{c_2} &\leq P_U(2; 1-\alpha) \\ 0 \leq \widehat{p}_{c_{100}} &\leq P_U(0; 1-\alpha) \end{aligned}$$

2.3.2.3 La méthode de Hoeffding

Dans cette méthode, l'échantillon est tiré avec probabilités proportionnelles à la taille. Soit $\mu = E(z_i) = E(\overline{z}) = \dfrac{t_d}{t_x}$. Nous déterminerons la borne supérieure pour μ, que nous multiplierons par t_x pour obtenir la borne supérieure pour t_d.

Hoeffding a montré le résultat suivant : Si Z une variable aléatoire $0 \leq Z \leq 1$, $\mu = E(Z)$, alors $p(\overline{Z} > a) \leq V(a; \mu), \mu \leq a < 1$ où $V(a; \mu) = [\frac{1-\mu}{1-a}]^{n(1-a)}[\frac{\mu}{a}]^{na}$.

Considérons la fonction $a(\mu)$ définie implicitement par

$$V(a(\mu); \mu) = \begin{cases} \alpha & \text{si } 0 < \mu \leq \alpha^{1/n} \\ 1 & \text{si } \mu > \alpha^{1/n} \end{cases}$$

La borne supérieure proposée est

$$\overline{\mu}_H = 1 - a^{-1}(1 - \overline{z})$$

Ceci équivaut à $a(1 - \overline{\mu}_H) = 1 - \overline{z}$. En d'autres termes, il faut que $\overline{\mu}_H$ satisfasse $V(1 - \overline{z}; 1 - \overline{\mu}_H) = \alpha$. Puisque $V(1 - a; 1 - \mu) = V(a; \mu)$, on peut écrire plus simplement que

$\overline{\mu}_H$ est solution de $V(\overline{z};\overline{\mu}_H) = \alpha$.

Donc la procédure consiste à résoudre $[\frac{1-\overline{\mu}_H}{1-\overline{z}}]^{n(1-\overline{z})}[\frac{\overline{\mu}_H}{\overline{z}}]^{n\overline{z}} = \alpha$.

Nous montrons que $P(\overline{\mu}_H > \mu) \leq 1 - \alpha$

En effet

$$\begin{aligned} p(\mu > \overline{\mu}_H) &= p[1 - \mu < a^{-1}(1 - \overline{z})] \\ &= p[1 - \overline{z} > a(1 - \mu)] \\ &\leq V[a(1 - \mu); (1 - \mu)] = \alpha \qquad \forall \mu. \end{aligned}$$

Il est nécessaire pour valider cet argument de vérifier que la fonction a est croissante.

Soit

$$\begin{aligned} W(a;\mu) &= lnV(a;\mu) - ln\alpha \\ W_1(x;y) &= \frac{\partial W(x;y)}{\partial x} \\ W_2(x;y) &= \frac{\partial W(x;y)}{\partial y} \end{aligned}$$

Puisque

$$\begin{aligned} & W(a(\mu);\mu) = 0 \\ & \text{et } \frac{dW(a(\mu);\mu)}{d\mu} = W_1(a(\mu);\mu)\frac{da}{d\mu} + W_2(a(\mu);\mu) = 0 \\ & \Rightarrow \frac{da}{d\mu} = -\frac{W_2(a(\mu);\mu)}{W_1(a(\mu);\mu)}. \end{aligned}$$

On a

$$\begin{aligned} & W_1(a(\mu);\mu) = nln[\frac{(1-a)\mu}{(1-\mu)a}] \\ & \text{et } W_2(a(\mu);\mu) = \frac{n(a-\mu)}{\mu(1-\mu)} \\ & \Rightarrow \frac{da}{d\mu} = -\frac{W_2(a(\mu);\mu)}{W_1(a(\mu);\mu)} = \frac{n(\mu-a)}{\mu(1-\mu)ln[\frac{\mu(1-a)}{(1-\mu)a}]}. \end{aligned}$$

On voit bien que le numérateur et le dénominateur ont toujours le même signe, ce qui entraîne que la dérivée de a est positive.

Application

Pour voir la fiabilité de la méthode de Hoeffding par rapport à une des méthodes classiques, on considère une population de comptes comptables générée selon une loi gamma. Les valeurs des comptes vérifiés sont identiques aux valeurs des comptes comptables lorsqu'il n'y a pas d'erreurs. Les restes sont obtenus en multipliant les valeurs comptables des comptes erronés par une variable de loi bêta. Après cela, on tire des échantillons où on fait varier leur taille n et le pourcentage d'erreur p.

On note par b_{sup} la borne supérieure , pour chaque échantillon tiré le pourcentage de recouvrement et le rapport $t'_d = \frac{b_{sup}}{t_d}$ sont présentés dans le tableau 2.1.

Les résultats montrent que la méthode de Hoeffding à un bon pourcentage de recouvrement comparé à la méthode classique, mais elle est conservatrice.

Tableau 2.1 Comparaison entre la méthode de Hoeffding et une méthode classique

		Méthode de Hoeffding		Méthode classique	
p	n	% de recouvrement	t'_d	% de recouvrement	t'_d
0.3	40	99.7	2.84	85.2	1.75
	60	99.6	2.42	86.3	1.64
	100	99.3	2.02	89	1.48
0.2	40	100	3.33	80.9	1.9
	60	99.5	2.84	85.7	1.81
	100	99.1	2.3	87.5	1.6
0.1	40	98	4.99	62.8	2.1
	60	99.9	4.08	73.1	2.01
	100	100	3.72	79.5	1.78

2.3.2.4 La méthode basée sur la loi de Kolmogrov-Smirnov

La borne proposée est $t_x\overline{\mu}_{KS}$, où $\overline{\mu}_{KS} = \overline{z} + q_\alpha$ et q_α est le point tel que $P\{sup_x[F_n(x) - F(x)] \leq q_\alpha\} \geq 1 - \alpha$, où $F_n(x)$ est la fonction de répartition empirique d'un échantillon de taille n et F_x est la fonction de répartition de la population.

Justification

$P(\mu \leq \overline{\mu}_{KS}) = P(\mu \leq \overline{z} + q_\alpha) = P(\overline{z} - \mu \geq -q_\alpha)$

Or $\overline{z} = \int_0^1 [1 - F_n(x)]dx$ et $\mu = \int_0^1 [1 - F(x)]dx$

Donc

$$\begin{aligned} P(\mu \leq \overline{\mu}_{KS}) &= P(\overline{Z} - \mu \geq -q_\alpha) \\ &= P(\int_0^1 [1 - F_n(x)]dx - \int_0^1 [1 - F(x)]dx \geq -q_\alpha) \\ &= P(\int_0^1 [F_n(x) - F_x]dx \leq q_\alpha) \\ &\geq P(sup_x[F_n(x) - F_x] \leq q_\alpha) \end{aligned}$$

Kolmogrov a montré que lorsque n est grand :

$P\{sup_x[F_n(x) - F(x)] \leq \frac{z}{\sqrt{n}}\} \rightarrow 1 - e^{-2z^2}$

Donc

$$\begin{aligned} & P\{sup_x[F_n(x) - F_x] \leq q_\alpha\} = 1 - \alpha \\ & \Rightarrow P\{sup_x[F_n(x) - F_x] \leq \frac{\sqrt{n}q_\alpha}{\sqrt{n}} = 1 - \alpha \\ & \Rightarrow 1 - e^{-2nq_\alpha^2} = 1 - \alpha \\ & \Rightarrow q_\alpha = \sqrt{\frac{-log\alpha}{2n}} \end{aligned}$$

La borne supérieure pour le total est donc $t_x(\overline{z} + \sqrt{\frac{-log\alpha}{2n}})$

Application

Dans cette application, on considère la même population qu'on a utilisée pour obtenir les résultats dans le tableau précédent et les résultas sont résumés dans le tableau 2.2. Nous remarquons que la méthode de Kolmogrov-Smirnov est encore plus conservatrice que celle de Hoeffding.

Tableau 2.2 Comparaison entre la méthode de Kolmogrov-Smirnov et une méthode classique

		Méthode de Kolmogrov-Smirnov		Méthde classique	
p	n	% de recouvrement	t'_d	% de recouvrement	t'_d
0.3	40	100	3.79	78.4	1.58
	60	100	3.31	81.8	1.51
	100	100	2.75	82	1.36
0.2	40	100	4.98	76.7	1.74
	60	100	4.23	80.6	1.62
	100	100	3.52	83.8	1.5
0.1	40	100	9.55	60.7	1.87
	60	100	7.95	68.1	1.73
	100	100	6.4	78.1	1.65

CHAPITRE III

SIMULATION

Dans ce chapitre, nous allons faire des simulations pour comparer sept méthodes présentées dans le chapitre précédent, soit : la méthode d'attributs, la méthode de la moyenne, la méthode par le ratio, la méthode de la différence et la méthode combinée lors d'un tirage avec probabilités égales, la méthode par grappes et la méthode de Stringer lors d'un tirage avec probabilités inégales.

Nous commençons par générer une population constituée de N valeurs $x_1, \cdots, x_N$, les valeurs aux livres, connues pour la population entière. Une hypothèse réaliste pour des valeurs comptables est que la distribution de la population est asymétrique. C'est pour cela que nous avons généré les valeurs des x selon une loi gamma de paramètres k et *theta*. Ces deux paramètres, permettent d'avoir une assez bonne diversité de populations étant donné que, leur moyennes et varriances sont des fonctions de k et *theta*. On a $\mu = \frac{k}{\theta}$ et $\sigma^2 = k\theta^2$. Nous considérons ensuite quelques valeurs de p, la proportion des comptes erronés : les $N(1-p)$ premiers ne sont pas erronés et donc y_i sera identique à x_i pour ceux-là. Les Np autres comptes contiendront des valeurs y_i distinctes, que l'on génère de la façon suivante : $y_i = \beta_i x_i$, où β_i est une variable de loi bêta de paramètres α et β et de moyenne $\mu_1 = \frac{\alpha}{\alpha+\beta}$ et de variance $\sigma_1^2 = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$, ce qui nous permettra d'obtenir des $0 \leq y_i \leq x_i \quad \forall i = Np+1, \cdots, N$.

Ayant généré cette population, nous générons M échantillons tirés de cette population. Pour chacune des sept méthodes, deux indices de qualité seront retenus : l'un est bien sûr le pourcentage de recouvrement. Les points critiques à partir d'hypothèses de normalité manifestement fausses, il est important de savoir dans quelle mesure la probabilité de recouvrement s'approche de la probabilité nominale, que nous avons fixée à 95%.
Le second critère de qualité est le rapport $t'_d = \frac{b_{sup}}{t_d}$ où b_{sup} est la borne supérieure. Cette information est importante car le défaut majeur de certaines des méthodes considérées c'est d'être beaucoup trop conservatrices ; la probabilité de recouvrement est parfaitement satisfaisante, mais à un prix parfois inacceptable : une borne supérieure très élevée. On veut bien avoir un t'_d proche de 1 car dans ce cas la valeur de la borne supérieure serait proche de la vraie valeur t_d mais si t'_d est trop grand, la méthode serait alors trop conservatrice.
La taille de la population et le nombre d'échantillons tirées sont fixées respectivement à 10000 et 1000 préalablement. Nous comparerons les différentes méthodes avec différents valeurs de n, p, μ, σ, μ_1, σ_1.

La variation du rapport $\tau = \frac{\sigma}{\mu}$ peut avoir une influence sur les résultats.

La variation de μ_1 et σ_1 nous permet de comparer les méthodes selon l'ampleur des erreurs et de la variation du rapport $\beta_i = \frac{y_i}{x_i}$.

Nous attachons une importance particulière à la méthode de Stringer, car ses propriétés n'ont jamais été démontrées.

Toutes les simulations sont faites avec Logiciel R, et elles sont présentés sous forme de tableaux.

Dans les 4 premiers tableaux (3.1- 3.2- 3.3- 3.4), nous visons essentiellement à déterminer une taille d'échantillon limite à partir de laquelle le problème soulevé dans ce mémoire ne se pose plus. C'est-à-dire, nous voulons savoir à partir de quel moment l'échantillon peut être considéré comme assez grand pour qu'on puisse faire confiance aux méthodes classiques.

Dans le tableau 3.1, on considère une population où la moyenne des valeurs des comptes comptables est grande, τ est petit, le pourcentage d'erreur dans les comptes est élevé, les différences entre les vraies valeurs et les valeurs vérifiées sont petites et le rapport $\beta_i = \frac{y_i}{x_i}$ est stable.

Dans le tableau 3.2, on considère une population où la moyenne des valeurs des comptes comptables est grande, τ est grand, le pourcentage d'erreur dans les comptes est petit, les différences entre les vraies valeurs et les valeurs vérifiées sont petites et le rapport $\beta_i = \frac{y_i}{x_i}$ est stable.

Dans le tableau 3.3, on considère une population où la moyenne des valeurs des comptes comptables est petite, τ est grand, le pourcentage d'erreur dans les comptes est assez élevé, les différences entre les vraies valeurs et les valeurs vérifiées sont assez grandes et le rapport $\beta_i = \frac{y_i}{x_i}$ est peu variable.

Dans le tableau 3.4, on considère une population où la moyenne des comptes comptables est petite, τ est petit, le pourcentage d'erreur dans les comptes est très petit, les différences entre les vraies valeurs et les valeurs vérifiées sont grandes et le rapport $\beta_i = \frac{y_i}{x_i}$ est stable.

Tableau 3.1 Résultats de simulation lorsque $p = 0.5$, $\mu = 3000$, $\sigma = 300$, $\mu_1 = 0.8$, $\sigma_1^2 = 0.1$

	Attributs		Moyenne		Ratio		Différence		Combinée		Grappes		Stringer	
n	%	t'_d	%	t'_d	%	t'_d	%	t'_d	%	t'_d	%	t'_d	%	t'_d
30	100	9.14	87.8	1.7	87.7	1.7	89.3	1.75	94.3	2.17	87.8	1.73	99.8	2.33
70	100	8.27	90.9	1.47	90.7	1.47	91.6	1.49	96.8	1.72	90.3	1.46	99.1	1.74
200	100	7.67	93.4	1.28	93.5	1.28	93.4	1.3	97.9	1.41	93.2	1.28	98.1	1.39
500	100	7.37	92.9	1.17	92.9	1.17	94.2	1.18	97.5	1.25	94.3	1.18	97.5	1.23

Tableau 3.2 Résultats de simulation lorsque $p = 0.1$, $\mu = 3000$, $\sigma = 3000$, $\mu_1 = 0.8$, $\sigma_1^2 = 0.1$

	Attributs		Moyenne		Ratio		Différence		Combinée		Grappes		Stringer	
n	%	t'_d	%	t'_d	%	t'_d	%	t'_d	%	t'_d	%	t'_d	%	t'_d
30	100	99.6	50.2	4.24	50.6	4.2	97.5	16.9	70	15.3	58.7	2.33	100	6.52
70	100	75.1	62.8	2.22	62.8	2.22	97.4	1.11	75.2	3.88	77.5	2.03	100	3.75
200	100	60.1	76.4	1.82	76.9	1.81	96.9	6.88	86.7	2.5	86.9	1.64	99.5	2.23
500	100	53.2	86.6	1.58	86.7	1.58	95.3	4.77	94.1	1.92	90.8	1.41	98.1	1.66

Tableau 3.3 Résultats de simulation lorsque $p = 0.3$, $\mu = 200$, $\sigma = 200$, $\mu_1 = 0.5$, $\sigma_1^2 = 0.2$

	Attributs		Moyenne		Ratio		Différence		Combinée		Grappes		Stringer	
n	%	t'_d	%	t'_d	%	t'_d	%	t'_d	%	t'_d	%	t'_d	%	t'_d
30	100	27.5	77.4	1.86	80.1	1.82	97.1	3.03	88.7	2.72	90.6	1.69	98.6	2.03
70	100	23.9	86.6	1.63	87.1	1.59	96	2.28	94.8	2.09	90.2	1.44	97	1.59
200	100	21.3	89.6	1.36	90.6	1.35	96	1.79	96.6	1.57	92.8	1.25	96.7	1.31
500	100	20	91.7	1.24	92.3	1.23	95.6	1.49	97	1.35	92.4	1.16	96.5	1.19

Tableau 3.4 Résultats de simulation lorsque $p = 0.05$, $\mu = 200$, $\sigma = 20$, $\mu_1 = 0.5$, $\sigma_1^2 = 0.1$

	Attributs		Moyenne		Ratio		Différence		Combinée		Grappes		Stringer	
n	%	t'_d	%	t'_d	%	t'_d	%	t'_d	%	t'_d	%	t'_d	%	t'_d
30	100	5.84	95.1	2.9	95.1	2.9	97	3.06	99	4.98	72.6	2.2	100	3.85
70	100	3.96	88.4	1.9	88.3	1.9	92.8	2.02	99.8	2.99	83.1	1.82	100	2.47
200	99.9	2.84	90.1	1.53	90.2	1.53	92.4	1.61	99	1.98	90.7	1.52	96.7	1.74
500	100	2.35	92.3	1.33	92.3	1.33	94	1.37	99.4	1.54	92.1	1.33	96.1	1.42

Dans tous les tableaux qu'on a présentés ci-dessus, on remarque que toutes les méthodes deviennent moins conservatrices lorsqu'on augmente la taille de l'échantillon.Ainsi la méthode d'attributs a un pourcentage de recouvrement très proche ou égal à 100% quelle que soit la taille de l'échantillon et la méthode de Stringer a un pourcentage toujours supérieur à 95% mais le problème avec la méthode d'attributs c'est qu'elle est souvent très conservatrice.

Dans le tableau 3.1, on remarque que les méthodes classiques donnent un pourcentage de recouvrement proche de 95% et une petite borne supérieure quand la taille de l'échantillon est supérieure à 70. La méthode combinée et la méthode de Stringer donnent de bons résultats lorsque la taille d'échantillon est assez grande.

Dans le tableau 3.2, on remarque que quelle que soit la taille de l'échantillon, la méthode de la différence à un pourcentage de recouvrement toujours supérieur à 95% mais elle est toujours très conservatrice. Les autres méthodes classiques ont un pourcentage de recouvrement qui est loin de la probabilité nominale même lorsque la taille de l'échantillon est égale à 200. Les méthodes d'attribut et de Stringer deviennent moins conservatrices lorsque la taille de l'échantillon est très grande.

Dans le tableau 3.3, lorsque la taille de l'échantillon est égale à 500, toutes les méthodes donnent une petite borne supérieure et un pourcentage de recouvrement supérieur ou proche de 95% .

Dans le tableau 3.4, les méthode classiques sont inadéquates lorsque la taille de l'échantillon est trop petite mais deviennent plus efficaces lorsqu'on augmente la taille de l'échantillon.

.

Donc, étant donné que l'augmentation de la taille de l'échantillon et du pourcentage d'erreurs contribue à l'amélioration de la plupart des méthodes pour différentes types de populations, on va considérer par la suite des populations avec des tailles d'échantillons relativement petites et de faibles pourcentages d'erreurs.

Dans le tableau 3.5, on considère une population où la moyenne des valeurs des comptes comptables est grande, τ est petit, les différences entre les vraies valeurs et les valeurs vérifiées sont petites et le rapport $\beta_i = \frac{y_i}{x_i}$ est stable.

Dans le tableau 3.6, on considère une population où la moyenne des valeurs des comptes comptables est grande, τ est petit, les différences entre les vraies valeurs et les valeurs vérifiées sont assez grandes et le rapport $\beta_i = \frac{y_i}{x_i}$ est peu variable.

La population comptable et les différences entre les vraies valeurs et les valeurs vérifiées dans le tableau 3.7 sont identiques à celles dans tableau 3.6 et le rapport $\beta_i = \frac{y_i}{x_i}$ est très stable.

Le tableau 3.8 considère une population où la moyenne des valeurs des comptes comptables est grande, τ est petit, les différences entre les vraies valeurs et les valeurs vérifiées sont grandes et le rapport $\beta_i = \frac{y_i}{x_i}$ est stable.

Le tableau 3.9 considère une population où la moyenne des valeurs des comptes comptables est grande, τ est grand, les différences entre les vraies valeurs et les valeurs vérifiées sont petites et le rapport $\beta_i = \frac{y_i}{x_i}$ est stable.

Le tableau 3.10 considère une population où la moyenne des valeurs des comptes comptables est grande, τ est grand, les différences entre les vraies valeurs et les valeurs vérifiées sont grandes et le rapport $\beta_i = \frac{y_i}{x_i}$ est stable.

Le tableau 3.11 considère une population où la moyenne des valeurs des comptes comptables est petite, τ est petit, les différences entre les vraies valeurs et les valeurs vérifiées sont petites et le rapport $\beta_i = \frac{y_i}{x_i}$ est stable.

Le tableau 3.12 considère une population où la moyenne des valeurs des comptes comptables est petite, τ est grand, les différences entre les vraies valeurs et les valeurs vérifiées sont assez grandes et le rapport $\beta_i = \frac{y_i}{x_i}$ est peu variable.

Tableau 3.5 Résultats de simulation lorsque $\mu = 2000$, $\sigma = 200$, $\mu_1 = 0.8$, $\sigma_1^2 = 0.1$

		Attributs	Moyenne	Ratio	Différence	Combinée	Grappes	Stringer
p	n	% t'_d	% t'_d	% t'_d	% t'_d	% t'_d	% t'_d	% t'_d
0.2	20	100 14	68.7 2.29	68.9 2.26	85.6 2.69	83.5 4.58	66.5 2.23	100 4.93
	50	100 10	82.6 1.86	82.7 1.86	88.3 2.02	91.3 2.78	83.1 1.84	100 2.87
	100	100 9.69	85.2 1.63	85 1.62	90.4 1.73	95.1 2.15	87.1 1.63	99.4 2.16
0.05	20	100 30.9	68.4 10.1	68.4 10.3	95.7 11.2	93.5 22.6	28.7 2.26	100 15.4
	50	100 19.1	58.6 3	58.7 3	93.5 4.54	80.9 8.23	48.1 2.25	100 7.65
	100	100 14.4	72.1 2.19	72.3 2.19	90.9 3.09	84.1 4.26	74.1 2.21	100 4.58

Tableau 3.6 Résultats de simulation lorsque $\mu = 2000$, $\sigma = 200$, $\mu_1 = 0.5$, $\sigma_1^2 = 0.2$

		Attributs	Moyenne	Ratio	Différence	Combinée	Grappes	Stringer
p	n	% t'_d	% t'_d	% t'_d	% t'_d	% t'_d	% t'_d	% t'_d
0.2	20	100 5.64	87.8 2	87.9 2	88.4 2.08	94.8 3.43	85.7 2	100 2.82
	50	100 4.48	87.7 1.61	87.6 1.61	89.1 1.65	98.2 2.27	89.8 1.64	98.5 1.97
	100	100 4.04	90.7 1.46	90.5 1.46	91.7 1.49	98.6 1.84	92.7 1.47	97.4 1.66
0.05	20	100 12.6	87.6 5.53	81.8 5.53	94.3 5.95	97.5 10.6	45.5 2.48	100 7.14
	50	100 7.05	84.4 2.43	84.4 2.43	91.6 2.74	96.4 5	74.9 2.11	100 3.71
	100	100 6.25	85.1 1.88	85.1 1.88	89.5 2.15	93.9 3.35	86.2 1.92	100 2.8

Tableau 3.7 Résultats de simulation lorsque $\mu = 2000$, $\sigma = 200$, $\mu_1 = 0.5$, $\sigma_1^2 = 0.1$

		Attributs		Moyenne		Ratio		Différence		Combinée		Grappes		Stringer	
p	n	%	t'_d	%	t'_d	%	t'_d	%	t'_d	%	t'_d	%	t'_d	%	t'_d
0.2	20	100	6.03	88	1.91	87.8	1.91	89.4	1.97	98.3	3.1	85.9	1.85	100	2.77
	50	100	4.33	88.4	1.53	88.4	1.53	90	1.57	98.6	2.06	92.5	1.57	99	1.94
	100	100	3.84	90.4	1.4	90.6	1.4	91.6	1.43	99	1.71	93	1.41	98.7	1.61
0.05	20	100	11.5	90.9	5.29	90.9	5.29	97.4	5.62	100	9.72	56.8	2.41	100	6.96
	50	100	8.5	85	2.41	84.9	2.41	93.9	2.78	98.3	4.86	77.3	2.12	100	4
	100	100	5.44	85.2	1.78	84.9	1.78	91.1	2.01	98.4	2.91	85.3	1.79	100	2.61

Tableau 3.8 Résultats de simulation lorsque $\mu = 2000$, $\sigma = 200$, $\mu_1 = 0.2$, $\sigma_1^2 = 0.1$

		Attributs		Moyenne		Ratio		Différence		Combinée		Grappes		Stringer	
p	n	%	t'_d	%	t'_d	%	t'_d	%	t'_d	%	t'_d	%	t'_d	%	t'_d
0.2	20	100	3.57	87.9	1.81	87.8	1.81	88.8	1.84	98.9	2.59	86.2	1.76	98.1	2.18
	50	99.9	2.8	90.8	1.5	90.9	1.49	91.1	1.5	98.8	1.88	90.4	1.49	97.7	1.67
	100	100	2.4	93	1.36	93	1.36	94	1.37	99.4	1.58	92	1.35	97.7	1.44
0.05	20	100	7.67	92.3	3.99	92.3	3.99	96	4.14	99.8	6.99	59.1	2.31	100	4.97
	50	100	4.63	91.9	2.15	91.9	2.15	93.4	2.3	99.4	3.68	85.3	2.05	100	3.03
	100	100	3.69	87.6	1.73	87.7	1.73	91	1.83	99.3	2.57	90	1.77	98	2.24

Tableau 3.9 Résultats de simulation lorsque $\mu = 2000$, $\sigma = 2000$, $\mu_1 = 0.8$, $\sigma_1^2 = 0.1$

		Attributs		Moyenne		Ratio		Différence		Combinée		Grappes		Stringer	
p	n	%	t'_d	%	t'_d	%	t'_d	%	t'_d	%	t'_d	%	t'_d	%	t'_d
0.2	20	100	80.3	53	2.89	53.5	2.76	97.4	8.96	68.4	7.29	67.4	2.07	100	4.42
	50	100	82.1	67.2	2.61	68.5	2	97.6	7.1	79.6	3.1	82	1.85	100	2.97
	100	100	82.5	78.1	1.8	78.2	1.79	95.5	5.04	88.2	2.43	87.8	1.63	99.3	2.17
0.05	20	100	242	62.2	58.7	62.1	58.6	98.5	85.57	92.1	58.8	31	2.74	100	18.8
	50	100	115	51.5	6.73	51.4	6.72	96.6	26.7	76.8	26.9	53.2	2.53	100	7.85
	100	100	82.4	60.3	2.42	60.1	2.4	97.1	1.51	74.2	5.85	74.7	2.08	100	4.2

Tableau 3.10 Résultats de simulation lorsque $\mu = 2000$, $\sigma = 2000$, $\mu_1 = 0.2$, $\sigma_1^2 = 0.1$

		Attributs		Moyenne		Ratio		Différence		Combinée		Grappes		Stringer	
p	n	%	t'_d	%	t'_d	%	t'_d	%	t'_d	%	t'_d	%	t'_d	%	t'_d
0.2	20	100	20.2	74.3	2	76.5	1.95	96.2	3.14	92.7	4.07	86.2	1.79	97.8	2.2
	50	100	20.7	80.6	1.66	82	1.62	97	2.34	93.2	2.38	90.7	1.48	96.4	1.66
	100	100	16.1	86	1.52	87.5	1.49	96.3	1.98	97.1	1.95	94.2	1.36	98.2	1.45
0.05	20	100	49.6	85.1	14.1	85.5	14.1	98.7	18.66	98.8	32.88	58.8	2.4	100	5.22
	50	100	37.6	76.2	3.6	76.4	3.56	96.9	7.2	96	10.1	84.2	1.95	100	2.89
	100	100	22.8	76.2	1.94	76.4	1.9	96.9	5.04	96	3.61	84.2	1.7	100	2.18

Tableau 3.11 Résultats de simulation lorsque $\mu = 300$, $\sigma = 30$, $\mu_1 = 0.8$, $\sigma_1^2 = 0.1$

		Attributs		Moyenne		Ratio		Différence		Combinée		Grappes		Stringer	
p	n	%	t'_d	%	t'_d	%	t'_d	%	t'_d	%	t'_d	%	t'_d	%	t'_d
0.2	20	100	14.1	67.2	2.26	67.3	2.26	84.2	2.26	79.1	4.51	64.4	2.14	100	4.99
	50	100	10.4	82.3	1.83	82.3	1.83	86.9	2	90.6	2.74	85	1.9	100	2.92
	100	100	9.33	86.5	1.61	86.8	1.61	89.8	1.71	95.9	2.12	88.4	1.62	98.8	2.14
0.05	20	100	30.6	76.3	9.68	76.3	9.68	94.9	11.31	92.3	22.16	32.4	2.6	100	15.7
	50	100	16.5	63.8	2.81	63.8	2.81	91.6	4.03	84.5	7.46	56.3	2.39	100	6.89
	100	100	13.2	72.6	2.12	72.6	2.12	91.6	2.97	84.3	4.19	70.9	2.11	100	4.31

Tableau 3.12 Résultats de simulation lorsque $\mu = 300$, $\sigma = 300$, $\mu_1 = 0.5$, $\sigma_1^2 = 0.2$

		Attributs		Moyenne		Ratio		Différence		Combinée		Grappes		Stringer	
p	n	%	t'_d	%	t'_d	%	t'_d	%	t'_d	%	t'_d	%	t'_d	%	t'_d
0.2	20	100	34.3	67.7	2.27	69.8	2.21	97.2	4.55	86.4	4.9	86.7	1.99	100	2.78
	50	100	38.7	79.2	1.87	81.1	1.84	96.4	3.25	90.8	2.79	86.4	1.64	97.4	1.97
	100	100	21.8	85	1.58	85.6	1.57	96.4	2.57	94.2	2.08	91.1	1.45	97.4	1.62
0.05	20	100	94.7	74	25.3	74	25.2	98.5	35.5	94.7	62.9	40.4	2.32	100	7.52
	50	100	56	60.7	4.14	61.2	4.13	97.6	11.39	86.9	14.6	72.5	2.15	100	3.93
	100	100	36.3	69.4	2.17	70.4	2.15	96.2	7.38	86.1	4.45	87.7	1.89	100	2.7

On remarque que lorsque le pourcentage de comptes erronés est trop petit, toutes les méthodes sont inadéquates. Lorsque la moyenne des valeurs des comptes comptables et τ diminuent et les erreurs dans les comptes deviennent plus grandes, la méthode d'attributs devient moins conservatrice.

Pour une même population de comptes comptables, lorsque les différences entres les valeurs vérifiées et les vraies valeurs deviennent plus grandes, la méthode de Stringer devient légèrement meilleure et toutes les autres méthodes deviennent nettement meilleures.

Lorsqu'on augmente τ sans varier les autres paramètres, on remarque que toutes les méthodes sauf la méthode par grappes et la méthode de Stringer deviennent inadéquates.

Finalement on remarque que la méthode de Stringer est à son meilleur et peut être très efficace lorsque le pourcentage d'erreur n'est pas trop petit.

CONCLUSION

Dans ce mémoire, on a présenté des approches spécifiques pour l'estimation de l'erreur totale dans une population de comptes comptables. Et on a essayé de faire une comparaison entre ces méthodes et les méthodes classiques, dont on a fait un bref rappel.

Parmi ces méthodes spécifiques, la méthode de Stringer est celle qui nous intéresse particulièrement, vu la place qu'elle occupe dans le monde de la vérification. Cependant, ses propriétés n'ont jamais été démontrées.

Les études par simulation que nous avons faites ont montré que la méthode de Stringer est, dans la plupart des cas, meilleure que les autres méthodes spécifiques proposée, mais que dans certains cas, les méthodes classiques ont donné de bons résultats. En effet, comme on l'a déja vu, quand la taille de l'échantillon est assez grande, par exemple 500, les méthodes classiques restent valablesceci est vrai d'ailleurs pour toutes les autres méthodes présentées sauf pour la méthode d'attributs mais à condition que l'on tombe pas sur une population où la moyenne et la variance des x sont trop grandes comme dans le tableau 3.2. Cependant même avec une taille d'échantillon comprise entre 50 et 100, les méthodes classiques donnent des résultats satisfaisants lorsque les différences entre les vraies valeurs et les valeurs vérifiées dans les comptes erronés sont grandeset τ est petit. Dans ce même cas, on remarque que la méthode de Stringer est légèrement plus conservatrice que les méthodes classiques mais son pourcentage de recouvrement est supérieur à 95%. La méthode de Stringer devient plus conservatrice lorsque les différences entres les vraies valeurs et les valeurs vérifiées dans les comptes erronés sont petites, mais même dans ce cas, elle reste meilleure par rapport à toutes les autres méthodes proposées.

La méthode combinée donne aussi de très bons résultats mais elle est toujours légèrement plus conservatrice que la méthode de Stringer et devient inadéquate lorsque τ est grand et la différence entre les vraies valeurs et les valeurs vérifiées est petite.

Comme on l'a remarqué, même avant de faire lès simulations, la méthode d'attributs est trop conservatrice surtout lorsqu'on se trouve dans une population de comptes comptables avec une grande moyenne et leurs valeurs sont dispersées.

La méthode multinomiale est trop difficile à appliquer lorsque le nombre d'erreurs est supérieur ou égal à deux.

Les méthodes de Hoeffding et de Kolmogrov-Smirnov ont un bon pourcentage de recouvrement mais elles sont conservatrices.

RÉFÉRENCES

1989« Statistical Models and Analysis in Auditing : Panel on Nonstandard Mixtures of Distributions »*Statistical Science*, Vol. 4, No. 1 , pp. 2-33

Anderson, Rod et Teitlebaum, A. D., 1973 « Dollar-unit Sampling »*Canadian Chartered Accountant* pp.30-39

Anderson, R. J. et Leslie, D. A., 1975 « Discussion of Considerations in Choosing Statistical Sampling Procedures in Auditing »*Journal of Accounting Research*, vol.13, (suppl.) pp.53-64

Arkin, H., 1963 « Handbook of Sampling for Auditing »*The Accounting Review*, Vol. 39, No.1, pp. 233

Barnett, Anndrew H. ; Read, William J., 1986« How are auditors implementing SAS no.39 ? »*Journal Of Accountancy* pp.78-88

Barnett, Anndrew H. ; Read, William J., 1986« Attributes Sampling : A Local Firm's Experience »*Journal Of Accountancy* pp.130-135

Bickel, Peter J., 1992 « Inference and auditing : The Stringer Bound »*International Statistical Review* , Vol60, No.2, pp. 197-209.

Cox, D. R. ; Snell, E. J., 1979 « On Sampling and the Estimation of Rate Errors »*New Biometrika Trust* Vol.66, No.1, pp.125-132

Chen, Jiahua, Shun-Yi Chen et J. N. K. Rao, 2002 « Empirical likelihood condifence intervals for the mean of a population containing many zero values »*The Canadian Journal of Statistics* , Vol.31, pp.1-15

Clayton, Howard R., 1994 « A Combined Bound for Errors in Auditing Based on Hoeffding's Inequality and the Bootstrap »*Journal of Business and Economic Statistics*, Vol.12, No. 4, pp.437-448

Diciccio, Thomas J., Romano, Joseph P., 1988 « A Review of Bootstrap Confidence Intervals »*Journal of the Royal Statistical Society. Series B (Methodological*, Vol. 50, No.3, pp. 338-354

Dworin, L. et Grimlund, R. A., 1984 « Dollar Unit Sampling for Accounts Receivable and Inventory »*Accounting Review*, Vol. 59, No.2, pp. 218-241

Dworin, L. et Grimlund, R. A., 1986 « Dollar Unit Sampling : A Comparison of the Quasi-Bayesian and Moment Bounds »*Accounting Review* , Vol.61, No.1, pp.36-57

Fienberg, Stephen E., Holland, Paul W., 1973 « Simultaneous Estimation of Multinomial Cell Probabilities »*Journal of The American Statistical Association* , Vol.68, No.343, pp.683-691

Finley, David R., 1985 « Counterexamples to Proposed Dollar-Unit Sampling Algorithm »*Journal of Accounting Research* , Vol.23, No.1, pp.402-404

Frost, P. A. et Tamura, H, 1982 « Jackknifed Ratio Estimation in Statistical Auditing »*Journal of Accounting Research* , Vol. 20, No.1, pp.103-120

Frost, P. A. et Tamura, H, 1986 « Accuracy of Auxiliary Information Interval Estimation in Statistical Auditing »*Journal of Accounting Research*, Vol.24, No.1, pp.57-75

Garstka, Stanley J., 1977 « Models for Computing Upper Error Limits in Dollar-Unit Sampling »*Journal of Accounting Research*, Vol.15, No.2, pp.179-192

Gartska, S. J. et Ohlson, P. A., 1979 « Ratio Estimation in Accounting Populations with Probabilities of sample selection proportional to size of book values », Vol. 17, No.1, pp.23-59

Grimlund, R. A. et Felix, W. L., 1987 « Simulation Evidence and Analysis of Alternative Methods of Evaluating Dollar-Unit Samples »*Accounting Review*, Vol. 62, No.3, pp.455-479

Hartley, H.O. ; Rao, J.N.K. 1968 « A New Estimation Theory for Sample Surveys »*Biometrika*, Vol.55, No.3, pp.547-557

Hoeffding, W., 1963 « Probability inequalities for sums of random variables »*Journal of the American Statistical Association* , Vol.58, No.301, pp.13-30

Haskins & Sells, 1970 « Audit Sampling, A Programmed Instruction course, New York : Haskins & Sells. ».

Herbert, Leo, 1946« Practical Sampling For Auditors »*The accounting review* Vol.21, No 4, pp.386-390

Hurst, D.C. ; Quesenberry, C.P. 1964« Large Sample Simultaneous Confidence Intervals for Multinomial Proportions »*The accounting review* Vol.21, No 4, pp.386-390

Kaplan, Robert S., 1973 « Statistical Sampling in Auditing with Auxiliary Information Estimators »*Journal of Accounting Research*, Vol.11, No.2, pp.238-258

Kvanly, Alain H. ; Shen, Yaung Kaung ; Deng, Lih Yuan, 1998 « Construction of Confidence Intervals for the Mean of a Population Containing Many Zero Values »*Journal of Business & Economic Statistics* Vol.16, No.3, pp.362-368

Leitch, Robert A. ; Neter, John ; Plante, Robert ; Sinha, Prabhakant, 1981 « Implementation of Upper Multinomial Bound Using Clustering »*Journal of the American Statistical Association* Vol.16, No.375, pp.530-533

Leitch, Robert A. ; Neter, John ; Plante, Robert ; Sinha, Prabhakant, 1982 « Modified Multinomial Bounds for Larger Number of Errors in Audits »*The Accounting Review* pp. 384-400.

Lillestol, Jostein, 1981 « A Note on Computing Upper Error Limits in Dollar-Unit Sampling »*Journal of Accounting Research*, Vol.19, No.1, pp.263-267

Loebeckke, James K. et Neter, John, 1975 « Considerations in Choosing Statistical Sampling Procedures in Auditing »*Journal of Accounting Research*, Vol.13, Studies on Statistical Methodology in Auditing, pp.38-52

Matsumura, Ella M. ; Plante, Robert ; Tsui, Kam-wah ; Knan, P. 1991 « Comparative Performance of Two Multinomial-Based Methods for Obtaining Lower Bounds on the Total Overstatement Error in Accounting Populations »*Journal of Business & Economic Statistics*, Vol.9, No.4, pp.423-429

McCray, John H. A, 1984 « Quasi-Bayesian Audit Risk Model for Dollar Unit Sampling »*The Accounting Review*, Vol.59, No.1, pp.35-51

Meikle, Giles R., 1972 « Statistical Sampling in an Audit Context, Toronto : The Canadian Institute of Chartered Accountants ».

Menzefricke, Ulrich., 1983 « On Sampling Plan Selection with Dollar-Unit Sampling »*Journal of Accounting Research* , Vol. 21, No.1, pp. 96-105

Menzefricke, Ulrich ; Smieliauskas, Wally., 1984 « A Simulation Study of the Performance of Parametric Dollar Unit Sampling Statistical Procedures »*Journal of Accounting Research*, Vol.22, No.2, pp.588-604

Neter, John, 1952 « Some Applications of Statistics for Auditing »*Journal of The American Statistical Association*, Vol.47, No.257, pp.6-24

Neter, John, 1954 « Problems in Experimenting with the Application of Statistical Techniques in Auditing »*The Accounting Review*, Vol.29, No.4, pp.591-600

Neter, John, 1984 « Bayesian Bounds for Monetary Unit Sampling in Accounting and Auditing »*Journal of Accounting Research*, Vol.22, No.2, pp.497-525

Neter, John ; Godfrey, James ; Wrust Jhon 1991 « Effectivness of Rectification in Audit Sampling »*The Accounting Review*, Vol.66, No.2, pp.333-346

Neter, John ; Godfrey, James ; Wrust Jhon 1989 « Comparaison of Sieve Sampling with Random and Cell Sampling of Monetary Units »*The Statistician*, Vol.38, No.4, pp.235-249

Neter, John ; Godfrey, James ; Wrust Jhon 1989 « Efficiency of Sieve Sampling in Auditing »*Journal of Business & Economic Statistics*, Vol.7, No.2, pp.199-205

Neter, John, Loebbecke, James K., 1975 « Behavior of major Statistical estimators in Sampling Accounting Populations : An Empirical Study (manuscrit, 1975) Rapporté dans Loebbecke et Neter (1975) ».

Neter, John ; Leitch, Robert A. ; Fienberg, Stephen E., 1978 « Dollar Unit Sampling : Multinomial Bounds for Total Overstatement and Understatement Errors »*The Accounting Review* pp.77-93

Owen, Art B., 1988 « Empirical Likelihood Ratio Confidence Intervals for a Single Functional »*Biometrika* , Vol.75, No.2, pp.237-249

Pyke, R., 1965 « Spacings »*Journal of the royal Statistical Society. Series B (methodological*, Vol.27, No.3, pp.395-449

Plante, R., J. Neter, et R. A. Leitch, 1984 « A Lower Multinomial Bound for the Total Overstatement Error in Accounting Populations »*Management Science*, Vol.30, No.1, pp.37-50

Plante, R., J. Neter, et R. A. Leitch, 1985 « Comparative performance of multinomial, cell, and Stringer bounds »Auditing **5** (Fall) , Studies in Statistical Methodology in Auditing, pp.40-56

Ramage, John G. ; Krieger, Abba M. ; Spero, Leslie L., 1979 «An Empirical Study of Error Characteristics in Audit Populations »*Journal of Accounting Research*, Vol.17, Studies on Auditing-Selections from the"Research in Auditing "Program, pp.103-107

Rao, J.N.K, 1985 « Conditional Inference in Survey Sampling »*Statistics Canada : Survey Methodology* Vol.11, No.1, pp.15-31

Reneau, J. Hal, 1978 « CAV Bounds in Dollar Unit Sampling : Some Simulation Results »*The Accounting Review* Vol.53, No.3, pp.669-680

Roberts, Donald M., 1975 « Discussion of The Real Risks in Audit Sampling »*Journal of Accounting Research*, Vol.13, Studies On Statistical Méthodologie in Auditing, pp.92-94

Tamura, Hirokuni et Frost, Peter A., 1986 « Tightenening CAV (DUS) Bounds by Using a Parametric Model »*Journal of Accounting Research* **24** pp. 364-371

Teitlebaum, A. D. et Robinson, C.F., 75 « The Real Risks in Audit Sampling »*Journal of Accounting Research*, Vol.13, Studies On Statistical Méthodologie in Auditing, pp.70-91

Trueblood, Robert M. , Cooper, W. W., 1955 « Research and Practice in Statistical Applications to Accounting, Auditing, and Management Control »*The Accounting Review* , Vol.30, No.2, pp.221-229

Tsui, Kam-Wah ; Matsumura, Ella Mae ; Tsui, Kwok-Leung, 1985 « Multinomial-Dirichlet Bounds for Dollar-Unit Sampling in Auditing »*The Accounting Review*, Vol.60,No.1, pp.76-96

Stephen E. fienberg ; John Neter ; R.A.Leitch, 1977« Estimating the Total Overstatement Error in Accounting Populations »*Journal of the American Statistical Association*, Vol.72, No.358, pp.295-302

Vance, Lawrence L., 1951 « How Much Test Checking is enough ? »*The accounting Review*, Vol.26, No.1, pp.22-30

Vance, Lawrence L., 1952 « An Experience with Small Random Samples in Auditing »*The accounting Review*, Vol.27, No.4, pp.472-474

Vance, Lawrence L., 1960 « A review of Developments in Statistical Sampling for Accountants »*The accounting Review*, Vol.35, No.1, pp.19-28

Warrimer, Philip, 1951 « How Statistical Analysis Can Serve Accountants »*The accounting Review*, Vol.35, No.1, pp.362-370