UNIVERSITÉ DU QUÉBEC

MÉMOIRE PRÉSENTÉ À
L'UNIVERSITÉ DU QUÉBEC À CHICOUTIMI
COMME EXIGENCE PARTIELLE
DE LA MAÎTRISE EN INFORMATIQUE

par
Nancy Landry

Tests d'hypothèses et intervalles de confiance appliqués aux coefficients d'apparentement

Décembre 2003

Bibliothèque
Paul-Émile-Boulet

UQAC
Université du Québec
à Chicoutimi

Mise en garde/Advice

Afin de rendre accessible au plus grand nombre le résultat des travaux de recherche menés par ses étudiants gradués et dans l'esprit des règles qui régissent le dépôt et la diffusion des mémoires et thèses produits dans cette Institution, **l'Université du Québec à Chicoutimi (UQAC)** est fière de rendre accessible une version complète et gratuite de cette œuvre.

Motivated by a desire to make the results of its graduate students' research accessible to all, and in accordance with the rules governing the acceptation and diffusion of dissertations and theses in this Institution, the **Université du Québec à Chicoutimi (UQAC)** is proud to make a complete version of this work available at no cost to the reader.

L'auteur conserve néanmoins la propriété du droit d'auteur qui protège ce mémoire ou cette thèse. Ni le mémoire ou la thèse ni des extraits substantiels de ceux-ci ne peuvent être imprimés ou autrement reproduits sans son autorisation.

The author retains ownership of the copyright of this dissertation or thesis. Neither the dissertation or thesis, nor substantial extracts from it, may be printed or otherwise reproduced without the author's permission.

RÉSUMÉ

L'analyse des données généalogiques est un outil important pour comprendre la répartition du pool génétique d'une population sur un territoire ou pour établir si un trait (maladie ou autre) comporte une composante héréditaire majeure. La mesure des liens de parenté, utilisée dans notre étude, est le coefficient moyen d'apparentement.

Il existe peu d'études sur les propriétés statistiques du coefficient d'apparentement dû principalement à la dépendance inhérente et à la complexité de calcul de cette mesure ainsi qu'à l'indisponibilité de bases de données généalogiques fiables. Un test d'hypothèses faisant intervenir le coefficient moyen d'apparentement a été proposé par Hauck et Martin mais ce dernier est basé sur des résultats de normalité asymptotique et il n'est applicable que pour des tailles d'échantillons très grandes.

Le premier objectif de la recherche a été d'établir les limites des résultats de normalité asymptotique dont la statistique est le coefficient moyen d'apparentement dans un contexte plus réaliste : une population et des échantillons de tailles relativement petites. Le second objectif a été de construire des algorithmes permettant de calculer un intervalle de confiance et un test d'hypothèses sur cette statistique en utilisant des techniques de simulations spécifiques. Plusieurs méthodes de construction d'intervalles et de tests ont été comparées afin de dégager le meilleur estimateur par intervalle et le meilleur test.

Les meilleures méthodes de construction d'intervalles et de tests ont été obtenues à partir des techniques de rééchantillonnage. Elles ont permis d'obtenir d'excellents tests de comparaison. Cependant, les résultats concernant la détermination d'un intervalle de confiance sont mitigés. Il serait intéressant, lors de recherches ultérieures, de se pencher davantage sur ces techniques afin d'optimiser l'estimateur par intervalle.

Il n'existait pas de résultats de simulation portant sur les intervalles de confiance et les tests d'hypothèses faisant intervenir le coefficient moyen d'apparentement. Ces résultats sont un premier pas vers une meilleure compréhension des propriétés statistiques du coefficient moyen d'apparentement.

REMERCIEMENTS

Au terme de ce mémoire, je tiens à adresser mes remerciements et témoigner toute ma reconnaissance aux personnes qui ont participé de près ou de loin à la réalisation de cette recherche.

Tout d'abord, ma première pensée va à mon directeur de recherche, M. Louis Houde, professeur de statistiques à l'Université du Québec à Chicoutimi, sans qui, cette recherche n'aurait jamais abouti dans d'aussi bonnes conditions. Ses conseils, ses connaissances, son positivisme, ses encouragements, sa bonne humeur et sa grande disponibilité au quotidien m'ont permis de réaliser ce travail et surtout d'apprécier énormément la recherche. Il est selon moi, un directeur exceptionnel. Merci d'avoir accepté de me guider et d'être un « mentor » pour moi.

Mentionnons le support du projet BALSAC en collaboration avec le GRIG (Groupe de Recherche Interdisciplinaire en démographie et épidémiologie Génétique) pour l'utilisation de la banque de données contenant les reconstructions généalogiques. Je dois également souligner le support financier des différents organismes et autres soit le CRSNG, l'Industrielle Alliance et le syndicat des professeurs pour la bourse André Lebrun.

Je remercie également M. Richard Tremblay, professeur de mathématique à l'Université du Québec à Chicoutimi, qui m'a généreusement prêté son bureau et appuyé dans tous mes projets d'avenir. Merci d'être soucieux de la réussite des étudiants.

Mille mercis à mon copain pour ses encouragements et sa compréhension. Merci Francis d'être si patient avec moi. Un grand merci à M. Julien Bousquet, professeur de marketing à l'Université du Québec à Chicoutimi, un grand ami qui m'a soutenu tout au long de cette recherche. Merci Julien d'être toujours là.

Finalement, je ne peux clore cette partie sans remercier M. Richard Vézina, professeur de mathématique à l'Université du Québec à Chicoutimi, qui m'a aidé, conseillé et surtout transmis sa passion des mathématiques tout au long de mon bac et encore aujourd'hui. Mille mercis M. Vézina.

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

INTRODUCTION

Plusieurs recherches ont montré que la généalogie pouvait être au service de la recherche génétique. Les interactions entre structures génétiques et structures généalogiques sont un domaine en pleine expansion. Elles ont permis de faire avancer considérablement la recherche en génétique.

Les généalogies étant complexes à traiter directement, on s'intéresse au coefficient d'apparentement moyen d'une population. Cette mesure est une indication du patrimoine génétique que partagent les individus donc de la dépendance génétique existant dans la population (Jacquard, 1970). Cette mesure peut-être utilisée dans un contexte de recherche de maladie à prédominance génétique ou encore pour décrire la structure démo-génétique d'une population.

Scriver (2001) donne un recensement des études génétiques effectuées au Québec, en particulier au Saguenay Lac St-Jean. Plusieurs des études recensées ont utilisé le coefficient d'apparentement comme mesure du caractère héréditaire de la maladie. Dans un autre contexte, Tremblay et al. (2001) utilisent les mesures du coefficient d'apparentement pour caractériser certaines composantes de la population québécois.

Le traitement statistique de cette mesure se heurte à sa justification même à savoir la dépendance présente entre les mesures. Les techniques classiques ne pouvant être utilisées, on propose de faire l'étude des propriétés statistiques de l'apparentement et de

déduire des intervalles de confiance et des tests d'hypothèses qui seront valides, en tenant compte de la dépendance.

On s'intéresse dans un premier temps à la distribution du coefficient d'apparentement pour vérifier dans quelle mesure il est possible de se baser sur la normalité asymptotique pour traiter cette statistique. L'étude de la distribution du coefficient moyen d'apparentement se fait à l'aide de simulations.

Puisque les propriétés du coefficient d'apparentement moyen sont méconnues, on utilise divers recours tels que des transformations et des méthodes de rééchantillonnage, afin d'obtenir certaines estimations de paramètres nécessaires à l'élaboration des intervalles de confiance et des tests. Les méthodes de construction sont ensuite comparées en soulignant leurs forces et leurs faiblesses.

Les algorithmes de simulation de cet ouvrage sont compilés et exécutés avec le logiciel SPLUS pour Unix. Le développement des méthodes de simulations appliquées aux coefficients d'apparentement est possible grâce aux données du fichier généalogique BALSAC-RETRO. Ces données nous ont permis de construire une population d'individus suffisamment grande pour construire et valider les différentes méthodes de cet ouvrage. Le projet BALSAC a démarré en 1972 et vise essentiellement la reconstruction des histoires familiales et des généalogies ascendantes ou descendantes à partir des actes de l'état civil. Ces informations sont disponibles sous forme d'une banque de données (Bouchard, 2003).

CHAPITRE 1

COEFFICIENTS D'APPARENTEMENT

1.1 Introduction

Les généalogies permettent de retourner dans le passé des familles et apportent des renseignements utiles sur l'évolution du patrimoine génétique d'une population ainsi que sur son histoire démographique (Bouchard et al., 1990). Dans le contexte des généalogies ascendantes d'un ensemble d'individus, cette information est si dense qu'elle n'est pas interprétable directement dès que la profondeur de reconstitution est plus grande que cinq ou six générations. Des indices sont alors nécessaires pour appréhender ces structures ou du moins une partie de l'information qu'elles contiennent. Le coefficient d'apparentement introduit par Malécot en 1948 est un indice du lien de parenté qui existe entre deux individus. Ce coefficient d'apparentement est souvent utilisé afin d'obtenir une caractérisation démo-génétique d'une population quelconque ou encore pour étudier le caractère héréditaire d'un trait observé (maladie ou autre) dans une population.

Ce chapitre traite de la notion d'apparentement. Dans un premier temps, une définition formelle de l'apparentement est donnée. Le coefficient moyen d'apparentement est ensuite introduit. Finalement, les principales méthodes de calculs des coefficients d'apparentement sont abordées.

1.2 Liens de parenté et reconstruction généalogique

La mesure des liens d'apparentement entre deux individus quelconques nécessite la connaissance de l'arbre généalogique de ceux-ci. La généalogie ascendante d'un individu

est représentée par un arbre binaire dont les sommets représentent les individus et les arcs, les relations de parenté directes (parents-enfants). La profondeur k de cet arbre est le nombre de générations. Lorsque l'arbre généalogique est complet pour une profondeur k donnée, l'arbre comporte 2^h individus à la génération $h = 0,\ldots,k$ et $(2^{k+1}-1)$ individus au total.

Dans le domaine de la reconstruction généalogique, la méthode de numérotation des ancêtres Sosa-Stradonitz (Pence, 1994) est une méthode permettant d'identifier de façon unique les individus qui composent une généalogie. L'individu à la base de la généalogie porte le numéro 1, son père a le numéro 2, sa mère a le numéro 3 et ainsi de suite. Ainsi la généalogie X de l'individu Y_1 de profondeur k, peut alors être représentée par un vecteur de la forme

$$X = (Y_1, Y_2, Y_3, \ldots, Y_{2i}, Y_{2i+1}, \ldots, Y_{2m}, Y_{2m+1}),$$

où les Y_i, selon la numérotation Sosa-Stradonitz sont les individus i de la génération h, leurs pères et leurs mères formant la prochaine génération portent le numéro 2i et 2i + 1 respectivement. Les individus 2m et 2m + 1 sont donc les parents des individus m, marquant la fin de l'information généalogique (génération k).

Il est à noter que dans la pratique, l'information généalogique comporte souvent des valeurs manquantes. Les arbres d'ascendance sont donc incomplets, principalement dû à un manque d'information sur les ascendants lors de la reconstruction généalogique. Ce phénomène est évidemment accentué lorsque les générations s'éloignent.

1.3 Définition du coefficient d'apparentement

De façon intuitive, l'apparentement entre deux individus i et j est dû à la présence d'un ou de plusieurs ascendants communs aux généalogies de i et j et ils sont dits «apparentés» lorsque leurs arbres d'ascendance ont une partie commune.

Une mesure plus formelle du lien d'apparentement entre deux individus est basée sur la transmission du patrimoine génétique, à savoir l'existence et la quantification de liens génétiques. Ces liens génétiques entre deux individus résultent de la possibilité pour chacun d'eux d'avoir reçu des gènes qui sont la copie d'un même gène de leurs ancêtres communs. De tels gènes sont dits «identiques par ascendance». Malécot (1948) définit le coefficient d'apparentement comme suit : « le coefficient de parenté ϕ_{ij} de deux individus i et j est la probabilité pour qu'un gène pris au hasard chez i soit identique à un gène pris au hasard au même locus chez j », un locus étant l'emplacement exact d'un gène sur un chromosome.

La formule pour ϕ_{ij}, le coefficient d'apparentement entre l'individu i et l'individu j s'énonce de la façon suivante :

$$\phi_{ij} = \sum_{c} \left(\frac{1}{2}\right)^{k_c} (1 + f_{ci}) \qquad \text{(Malécot 1948),}$$

où k_c est la longueur de chaque liste d'individus reliant l'individu i à j dans les deux arbres généalogiques, c représente l'ensemble des listes possibles, ci est l'ancêtre commun permettant de lier les deux individus et f_{ci} est le coefficient d'apparentement des parents de l'ancêtre commun ci (l'indice i ne peut être utilisé qu'une seule fois).

Deux individus apparentés sont reliés l'un à l'autre par une liste d'individus comportant un ou plusieurs ancêtres communs. Sa longueur est le nombre d'individus la composant. Par exemple à la figure 1.1, la suite {1,4,6,4,2} reliant l'individu 1 à 2 par les ancêtres communs 4 et 6 n'est pas une liste d'individus valides puisque l'individu 4 est répété deux fois. La suite {1,4,6,3,2} est une liste valide de longueur 5.

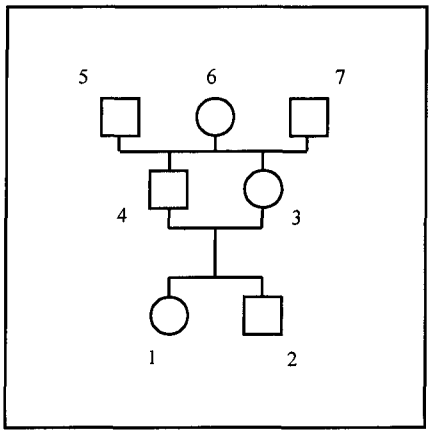

Figure 1.1 Généalogie des individus 1 et 2 sur deux générations

La figure 1.2 représente la généalogie ascendante de l'individu z s'étendant sur quatre générations. Pour évaluer le coefficient de parenté entre les individus i = 1 et j = 5, il est nécessaire de considérer toutes les possibilités pour ceux-ci de recevoir deux gènes identiques. Ainsi, les individus 1 et 5 ne peuvent avoir reçu la copie d'un même gène qu'en provenance de 9 ou de 10, leurs ancêtres communs. Pour que les individus 1 et 5 reçoivent le même gène provenant de 9, il faut que 6, 8 et 2 le reçoivent et le passent simultanément à 1 et 5. La probabilité est alors de $(1/2)^5$. De plus, les gènes doivent être identiques et ont une probabilité de 1/2 de l'être, la probabilité totale est donc de $(1/2)^6$. Le calcul étant le même pour 10, la probabilité devient $2 \cdot (1/2)^6$.

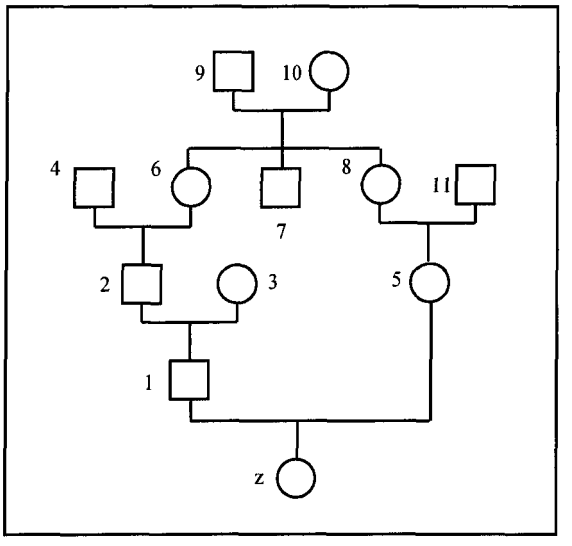

Figure 1.2 Généalogie ascendante de l'individu z sur quatre générations

Lorsque l'on considère la formule de Malécot (1948), il existe deux listes d'individus possibles: $\{1,2,6,9,8,5\}$ et $\{1,2,6,10,8,5\}$. Puisque 9 et 10 marquent la fin de l'information généalogique, $f_{ci} = 0$. Le coefficient d'apparentement entre i=1 et j=5 est alors,

$$\phi_{15} = \sum_{2}\left(\frac{1}{2}\right)^{6}(1+0) = 2*(1/2)^{6} = 0.03125.$$

Le coefficient de consanguinité de l'individu z est $f_z = \phi_{1,5} = 0.03125$.

Posons $X_1, X_2, \ldots, X_n$, un ensemble de n individus et leurs arbres d'ascendance. Le coefficient d'apparentement étant une mesure entre deux individus, il y a $n(n-1)/2$ coefficients d'apparentement différents. Pour un groupe de n individus, ces coefficients sont présentés dans une matrice d'apparentement de la forme

$$\Phi = \begin{pmatrix} \phi_{11} & \phi_{12} & \cdots & \phi_{1n} \\ \phi_{21} & \phi_{22} & \cdots & \phi_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \phi_{n1} & \phi_{n2} & \cdots & \phi_{nn} \end{pmatrix}.$$

Cette matrice est symétrique puisqu'elle possède des coefficients d'apparentement identiques entre les paires d'individus (i, j) et (j, i). De plus, les coefficients d'apparentement de sa diagonale principale sont ceux entre les paires d'individus $(i = j, j = i)$, c'est-à-dire un individu et lui-même. Leur valeurs sont $\phi_{i=j,j=i} = 1/2\,(1 + f_i)$ où f_i est le coefficient d'apparentement des parents de l'individu i.

Le coefficient moyen d'apparentement permet de résumer l'information contenue dans la matrice Φ. Il est donné par

$$\bar{\phi} = \frac{2}{n(n-1)} \sum_{i}^{n-1} \sum_{j>i}^{n} \phi_{ij}\,.$$

Ce coefficient est effectivement la moyenne des $n(n-1)/2$ coefficients d'apparentement différents, en excluant la diagonale principale soit éléments tels que $\phi_{i=j\ j=i} \geq 1/2$ de la matrice Φ.

1.4 Méthodes de calculs

Le calcul de la matrice des coefficients d'apparentement est fastidieux lorsque le nombre d'individus ainsi que la profondeur des généalogies sont grands. Dans un tel contexte, le recours à l'informatique s'impose. Diverses méthodes ont été développées afin de calculer les coefficients d'apparentement de façon exhaustive. Les principales sont : la

méthode des chemins, la méthode séquentielle et la méthode tabulaire classique (Thompson, 1974).

La méthode des chemins (Wright, 1922) évalue le coefficient d'apparentement ϕ_{ij} pour chaque paire d'individus, en ne considérant qu'une paire à la fois. Pour les individus i et j, ϕ_{ij} est calculé en recensant tous les ancêtres communs et toutes les longueurs de chemins. Un chemin est formé des individus reliant i et j et sa longueur est le nombre d'individus le composant. Puisque le coefficient de consanguinité des ancêtres communs est nécessaire au calcul, toutes les longueurs de chemins des parents des ancêtres communs à i et j sont aussi recensées. Cette méthode est particulièrement longue donc coûteuse en temps d'exécution lorsque le nombre d'individus n est grand.

La méthode séquentielle (Thompson, 1974) consiste à calculer l'ensemble des coefficients d'apparentement ϕ_{ij} entre tous les ancêtres, de génération en génération, jusqu'à ce que les descendants soient ceux recherchés. En d'autres termes, la matrice obtenue est celle de tous les individus et ancêtres calculés de façon séquentielle, d'où est extraite la matrice Φ des n individus recherchés. Cette méthode nécessite le calcul préalable du coefficient de consanguinité de tous les ancêtres. Cette méthode est coûteuse en terme d'espace mémoire lorsque le nombre d'individus est grand et que la profondeur des généalogies est grande. Par exemple, la taille de la matrice des coefficients d'apparentement, calculée à partir de 10 individus dont la profondeur de leur arbre d'ascendance respectif est de 12, est de l'ordre de 40950^2 soit 1.6769025×10^9 coefficients d'apparentement.

La méthode tabulaire (Lange, 2002) est basée sur la relation $\phi_{ij} = 1/2\,(\,\phi_{(\text{mère de i}),\,j} + \phi_{(\text{père de i}),\,j}\,)$. Cette fonction est appliquée sur chacune des paires d'individus i et j possibles, issue de l'ensemble, pour former la matrice Φ des coefficients d'apparentement. L'idée de base de la méthode tabulaire est de noter que les deux gènes présents chez un individu quelconque proviennent respectivement d'un tirage au sort parmi les deux gènes présents chez son père, d'une part, et d'un tirage au sort parmi les deux gènes présents chez sa mère, d'autre part. Cette méthode est celle qui a été retenue dans cette étude en raison de sa simplicité.

La matrice symétrique Φ de la généalogie ascendante présentée à la figure 1.2 se résume au coefficient d'apparentement entre les individus 1 et 5. Le tableau 1.4 montre en détail la procédure de calcul du coefficient d'apparentement entre ces individus, à l'aide de la méthode tabulaire. Il est à noter que cette méthode est valable seulement si la numérotation des individus est faite de façon à ce que les parents aient toujours un numéro plus grand que leurs enfants.

ϕ_{15} $= 1/2\,(\phi_{35} + \phi_{25})$
$= 1/2\,(\,1/2(\phi_{65} + \phi_{45}) + 0\,)$ Permutation de l'indice de ϕ_{65}
$= 1/4\,(\phi_{56} + \phi_{45})$
$= 1/4\,(1/2(\phi_{86} + \phi_{116}) + 0)$ Permutation des indices de ϕ_{86} et ϕ_{116}
$= 1/8\,(\phi_{68} + \phi_{611})$
$= 1/8\,(1/2(\phi_{98} + \phi_{108}) + 0\,)$ Permutation des indices de ϕ_{98} et ϕ_{108}
$= 1/16\,(\,\phi_{89} + \phi_{810}\,)$
$= 1/16\,(1/2(\phi_{99} + \phi_{109}) + 1/2(\phi_{910} + \phi_{1010})\,)$
$= 1/32\,((1/2 + 0) + (0 + 1/2\,))$
$= 1/32 = 0.03125$

Figure 1.3 Application de la méthode tabulaire sur les généalogies des individus 1 et 5 de la figure 1.2

CHAPITRE 2

DISTRIBUTION D'ÉCHANTILLONNAGE

2.1 Introduction

La distribution d'échantillonnage d'une statistique est la base pour construire un intervalle de confiance ou un test d'hypothèses. La normalité de la distribution des observations ou plus souvent la normalité approximative de la statistique sont souvent évoquées mais seulement pour établir les résultats. Dans le cas du coefficient d'apparentement il existe peu de référence sur la distribution de la statistique. Hauk et Martin (1984) proposent un test d'hypothèses en considérant que la distribution de la statistique $\bar{\phi}$ est asymptotiquement normale. Or, les tailles échantillonnales sont souvent petites dans le contexte de l'analyse des généalogies ce qui rend souvent l'utilisation des ces résultats hasardeuse.

Nous proposons donc de faire l'étude de la distribution échantillonnale du coefficient d'apparentement en regardant dans un premier temps le fondement théorique de la normalité asymptotique pour ensuite faire une étude de la distribution par simulation.

2.2 Statistique $\bar{\phi}$

Considérons une population composée d'individus et leur arbre d'ascendance respectif. La moyenne et la variance de la population sont respectivement μ_ϕ et σ^2_ϕ. Un échantillon aléatoire de taille n est tiré de la population et on s'intéresse à la statistique $\bar{\phi}$. Concrètement, la population est habituellement assez grande relativement à l'échantillon

pour considérer qu'un individu ne peut être tiré deux fois. La statistique $\bar{\phi}$ est un estimateur non biaisé de la moyenne de la population soit $E(\bar{\phi}) = \mu_\phi$ tandis que la variance de l'échantillon, S^2 est un estimateur biaisé soit $VAR(S^2) \neq \sigma_\phi^2$. Ce biais provient de la définition même du coefficient d'apparentement qui est une mesure de la dépendance entre les individus.

2.3 Normalité asymptotique

La statistique $\bar{\phi}$ étant une somme de variables aléatoires dépendantes, le théorème central limite ne s'applique pas afin de conclure à la normalité asymptotique. Hauk et Martin (1984) remarquent que la statistique est en fait une U-statistique et en se basant sur les résultats de Hoeffding (1948), ils déduisent que la statistique est asymptotiquement normale sous certaines conditions.

Considérons $\xi_n = X_1, X_2, \ldots, X_n$ n individus tirés aléatoirement d'une population de distribution inconnue et de moyenne μ_ϕ, où chaque individu v = 1..n est caractérisé par un vecteur de la forme $X_v = (Y_{v;1}, Y_{v;2}, Y_{v;3}, \ldots, Y_{v;2m}, Y_{v;2m+1})$, qui correspond à l'arbre d'ascendance de profondeur k selon la numérotation Sosa-Stradonitz. Le coefficient moyen d'apparentement $\bar{\phi}$ correspond effectivement à une U-statistique d'ordre deux donnée par

$$\bar{\phi} = U(X_1, X_2, \ldots, X_n) = \frac{2}{n(n-1)} \sum_{i<j} \phi_{ij}$$

où le coefficient d'apparentement ϕ_{ij} entre les individus X_i et X_j correspond au noyau $h(X_i, X_j)$ de la U-statistique (Serfling, 1980) comportant deux vecteurs.

Pour conclure à la normalité asymptotique d'une U-statistique (hoeffding, 1948), il faut que les vecteurs X_1, X_2, ... , X_n soient indépendants et que la taille de la population converge vers l'infini. Deux scénarios doivent être envisagés: une population dont uniquement les fondateurs sont fixés et une population dont les fondateurs et les descendants sont fixés.

Dans le cas où seulement les fondateurs sont fixés, la population est infinie puisqu'elle représente toutes les possibilités de naissances et d'associations entre les individus issus de ces fondateurs donc tous les enfants et descendants possibles. Lorsqu'un échantillon est prélevé dans la population contemporaine (population effectivement accessible pour l'échantillonnage), les ascendants se fixent (ses parents, ses grands parents, etc). Les arbres d'ascendance, à savoir les vecteurs observables, deviennent dépendants puisque chaque individu conditionne des liens de parenté. Dans un tel cas, les vecteurs sont dépendants et on ne peut utiliser les résultats de Hoeffding pour déduire la normalité asymptotique.

Dans l'autre scénario, on considère la population comme étant la population contemporaine. Les vecteurs sont effectivement indépendants mais les contraintes liées à la démographie humaine sont telles que la taille de la population est toujours restreinte. Il est donc difficile de justifier une taille d'échantillon convergeant vers l'infini. De plus, les

tailles des populations analysées dans la majorité des cas, lorsque l'étude s'intéresse aux traits héréditaires, sont de tailles relativement petites.

Dans un cas comme dans l'autre, l'approximation normale n'est pas justifiée par les résultats de distribution asymptotique. Il reste à déterminer si d'un point de vue pratique, la distribution normale est justifiable.

2.4 Distribution empirique

On propose une étude de la distribution de la statistique $\overline{\phi}$ par simulation puisqu'il est impossible d'obtenir des résultats théoriques sur la distribution. À priori, une population doit être disponible pour effectuer les simulations. La génération d'une population virtuelle « semblable » à une population contemporaine a été exclue car les paramètres de création d'une telle population sont trop complexes et mal connus (facteurs démographiques (migration, nuptialité, fécondité, mortalité), maladies, etc.). Il existe cependant des bases de données généalogiques assez grandes pour servir de population de référence.

Le Groupe de Recherche Interdisciplinaire en démographie et épidémiologie Génétique (GRIG) en collaboration avec le projet BALSAC de l'Université du Québec à Chicoutimi a créé une population de référence de 2600 individus et leur arbre d'ascendance respectif. Elle est constituée de 100 individus choisis dans 26 régions ou sous-régions différentes du Québec à partir d'actes de mariage datés entre 1935 et 1965. Elle comporte 150973 individus, une profondeur maximum de 14 générations et une profondeur moyenne de 10 générations. La matrice des coefficients d'apparentement Φ des 2600 individus a pu

être calculée pour constituer la population virtuelle de référence dont sa moyenne est $\mu_\phi = 6.638297 \times 10^{-4}$.

2.4.1 Méthodes

La distribution empirique de la statistique $\overline{\phi}$ a été simulée par tirage aléatoire de 10000 échantillons de mêmes tailles pour n = 10, 20, 30, 40, 50, 75 et 100 individus parmi la population de référence. À chaque tirage, l'apparentement des individus de l'échantillon a été calculé et la statistique $\overline{\phi}$ a été évaluée. Le calcul pour établir la matrice d'apparentement est long, c'est pourquoi la matrice d'apparentement de la population a été évaluée au préalable. Cette procédure nous a permis de minimiser le temps de calcul de la distribution empirique en évitant le calcul de la matrice des coefficients d'apparentement à chaque tirage.

Les échantillons aléatoires utilisés dans cette étude ont été prélevés avec le générateur de nombres aléatoires (pseudo aléatoire) du logiciel SPLUS (data analysis division mathsoft inc, 2000a). Ce générateur est un générateur hybride, souvent nommé «Super-Duper», développé par George Marsaglia et al. (1973). Il combine un générateur congruentiel linéaire et un générateur Tausworthe (shift-register generator). À des fins comparatives, le même germe de départ (seed) du générateur a été fixé afin de reproduire les mêmes échantillons.

Les simulations sont effectuées à partir de nombres aléatoires suivant une distribution uniforme discrète de taille 2600 puisque l'expérience consiste à choisir des

individus parmi la population de référence. Pour établir que le germe de départ ne génère pas une « mauvaise » séquence lors de notre expérimentation, l'évolution de la statistique du khi-deux comparant les nombres aléatoires à une distribution uniforme a été tracée en fonction du nombre de valeurs. La figure 2.1 permet de visualiser l'efficacité du générateur « Super-Duper » pour le germe fixé, en comparant les différentes valeurs χ^2 pour $1\times10^5, 2\times10^5, \ldots, 1\times10^6$ tirages. La statistique du khi-deux est calculée en comparant les fréquences relatives observées du générateur de nombre entier dans l'intervalle de 1 à 2600 et la probabilité théorique (1/2600).

On remarque, lorsque le nombre de tirages augmente, que les valeurs du khi-deux se stabilisent. Cela indique qu'il n'y a pas de nombres ou d'ensembles de nombres qui ont une probabilité plus forte que les autres.

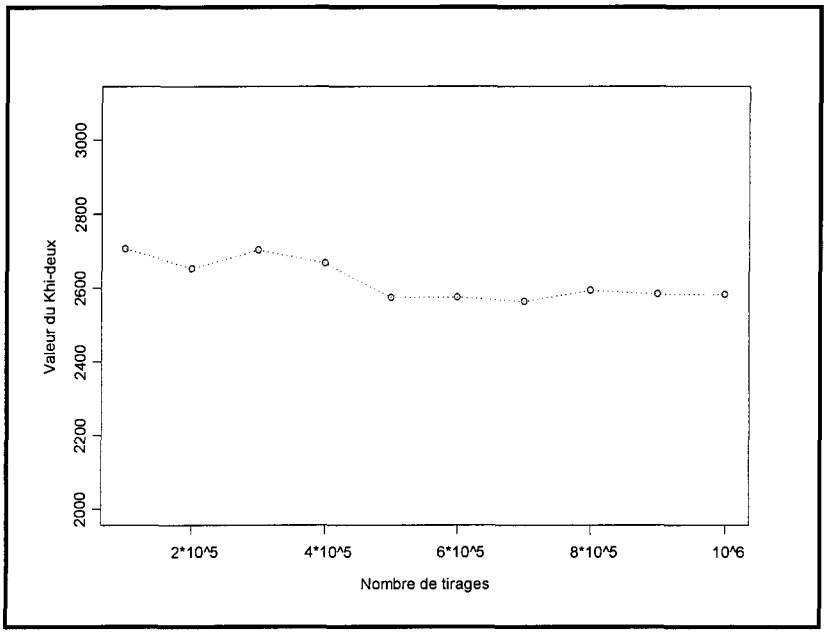

Figure 2.1 Efficacité du générateur « Super-Duper » (valeur χ^2 en fonction du nombre de tirages)

2.4.2 Validation de la population de référence

La population de référence est relativement petite principalement pour des raisons de complexité du calcul. Il est cependant possible de valider cette population pour quelques cas spéciaux. Le projet BALSAC dispose d'une base de données contenant 9899 individus pour lesquels les généalogies ont été reconstituées. Ces individus proviennent principalement de l'est du Québec et les années de mariages de ces individus pour lesquelles les ascendances sont disponibles s'étalent de 1900 a 1975. Cet ensemble de données a été utilisé pour évaluer si le fait de se restreindre à une population de 2600 individus apportait un biais important pour l'étude de la distribution de la statistique d'apparentement. Les comparaisons ont été faites pour une taille d'échantillon n =10, puisque c'était la seule taille pouvant se faire avec un temps de calcul raisonnable.

Considérons $F_R(\overline{\phi})$ et $F_C(\overline{\phi})$, les distributions empiriques respectives de la population de référence et de la population comparative pour une taille d'échantillon de n =10. Dans un premier temps, la comparaison a été réalisée en observant les fonctions de densité et de répartitions, ainsi que les principaux paramètres (quartiles, médiane, moyenne, etc) des distributions. Ensuite, un test de khi-deux et un test de Kolmogorov Smirnoff ont été utilisés. Puisque l'on s'intéresse à des probabilités caractéristiques lors de la construction d'un intervalle de confiance et d'un test, les distributions empiriques ont été comparées sur la base de ces probabilités.

2.4.2.1 Fonction de densité et fonction de répartition

La figure 2.2 présente les fonctions de densité des distributions $F_R(\bar{\phi})$ et $F_C(\bar{\phi})$. Elle révèle que les distributions tentent de se fondre uniquement au niveau des ailes des distributions. Le tableau 2.1 présente les principaux paramètres des distributions. Il témoigne du rapprochement des distributions. On peut observer un léger décalage du troisième et dernier quartile.

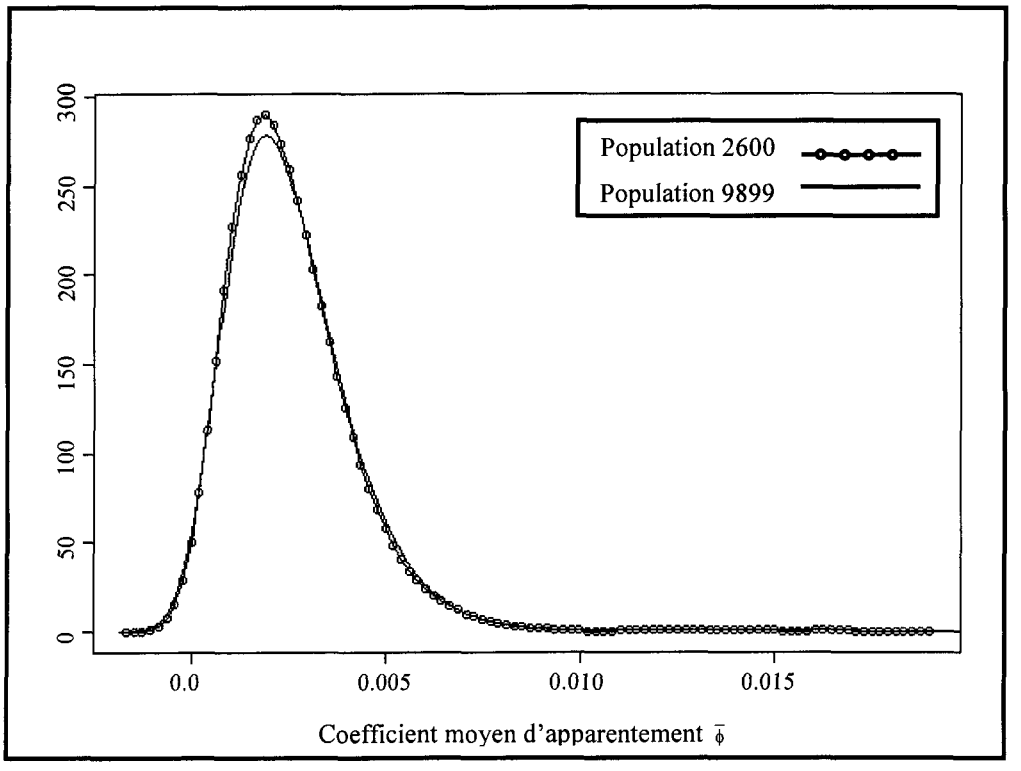

Figure 2.2 Fonction de densité des distribution empiriques de $F_R(\bar{\phi})$ et $F_C(\bar{\phi})$

	Quartile minimum	Premier quartile	Médiane	Troisième quartile	Dernier quartile	Moyenne $\mu_{\bar{\phi}}$
$F_R(\bar{\phi})$	0.000097683	0.0014270086	0.002216836	0.003256640	0.01716003	0.00251248
$F_C(\bar{\phi})$	0.000097683	0.0014475300	0.002257695	0.003321006	0.01883486	0.00257890

Tableau 2.1 : Paramètres des distributions empiriques $F_R(\bar{\phi})$ et $F_C(\bar{\phi})$

La figure 2.3 présente en détail les valeurs du coefficient moyen d'apparentement $\overline{\phi}$ aux différents quantiles des distributions. On peut observer une légère différence lorsque $\overline{\phi} \geq 0.0075$. Cette différence semble être causée par le fait que les 9899 individus de la population comparative ont été choisis dans des régions spécifiques où l'apparentement était plus important.

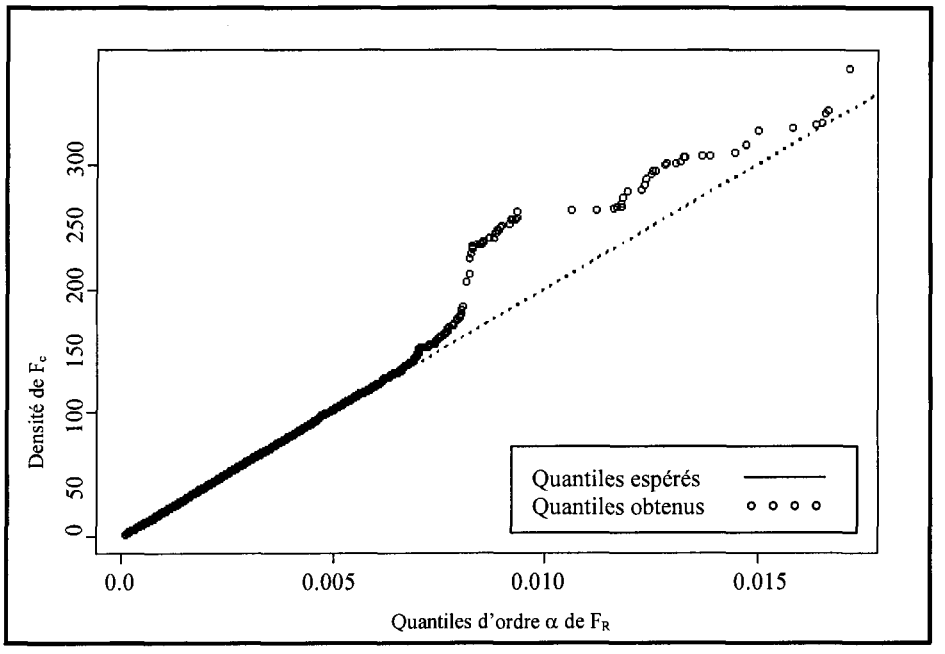

Figure 2.3 *Comparaison des valeurs de $\overline{\phi}$ aux différents quantiles d'ordre α des distributions $F_R(\overline{\phi})$ et $F_C(\overline{\phi})$*

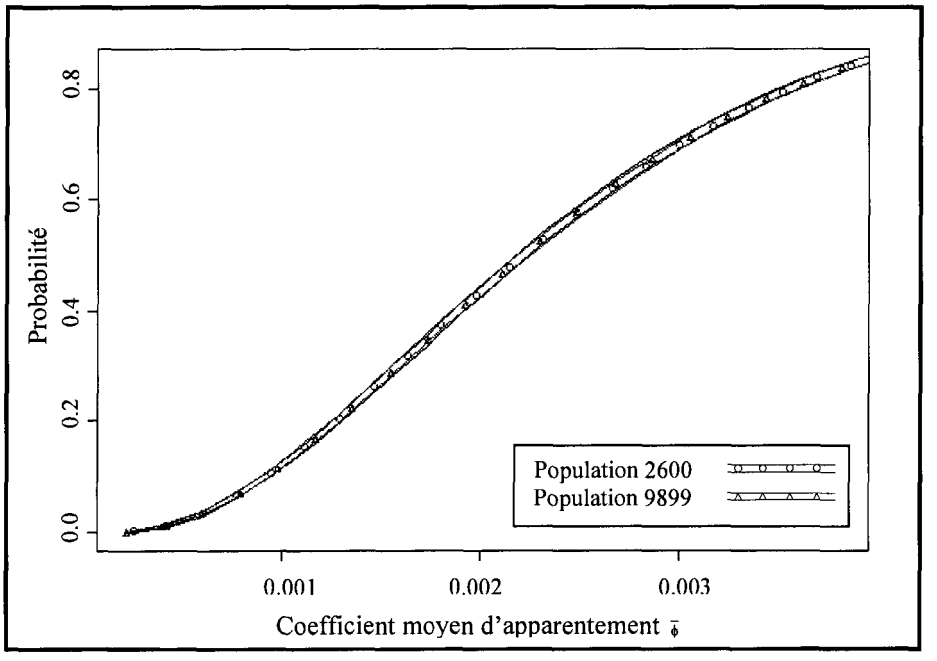

Figure 2.4 Fonctions de répartition des distributions empiriques $F_R(\bar{\phi})$ et $F_C(\bar{\phi})$ pour les probabilités de 0% à 85% et une précision de ± 0.0043

Les fonctions de répartition des distributions $F_R(\bar{\phi})$ et $F_C(\bar{\phi})$ ont été construites avec une précision de ± 0.0043 au pourtour de celles-ci pour contrer l'effet des fluctuations échantillonnales. La précision est donnée par la demie longueur d'un intervalle de confiance de niveau α pour la vraie valeur du paramètre soit $z_{\alpha/2}\sqrt{\alpha(1-\alpha)/m}$, où $z_{\alpha 2}$ est le point critique d'une loi normale centrée réduite et m, le nombre de simulations. Dans ce cas-ci, la précision a été calculée en considérant un demi intervalle au seuil α = 95 % et m = 10000. Il est à noter que toutes les précisions de cette étude sont calculées de cette façon.

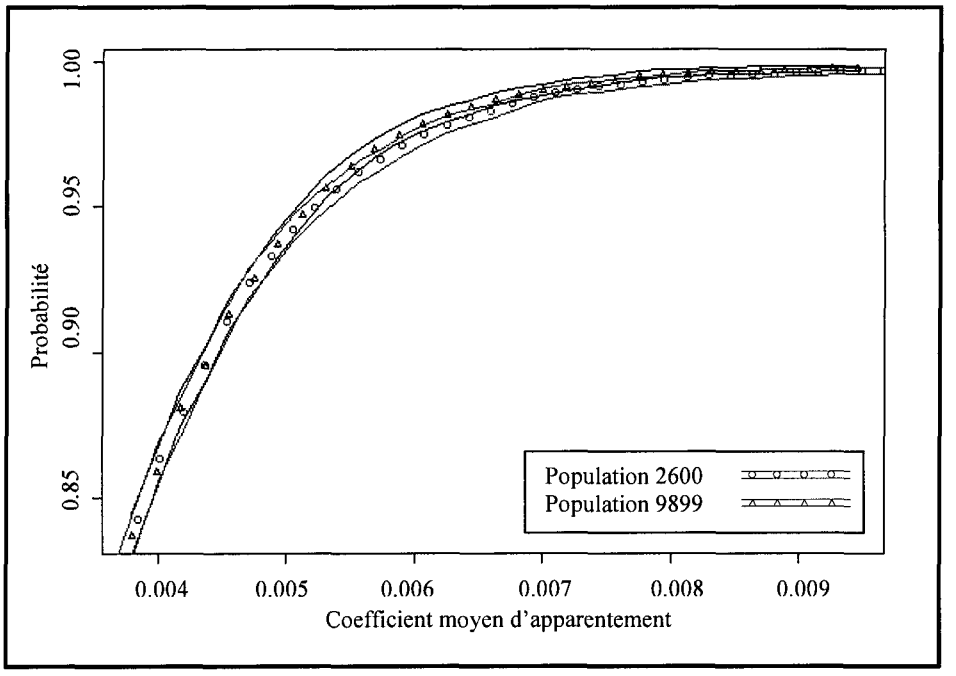

Figure 2.5 Fonctions de répartition des distributions empiriques $F_R(\bar{\phi})$ et $F_C(\bar{\phi})$ pour les probabilités de 85% à 100% et une précision de ± 0.0043

Les figures 2.4 et 2.5 présentent les fonctions de répartition des distributions $F_R(\bar{\phi})$ et $F_C(\bar{\phi})$ telles que décrites précédemment, pour les probabilités de 0 à 85 % et de 85 % à 100% respectivement. L'entrelacement des précisions inférieures et supérieures des fonctions de répartition confirme que les distributions sont similaires.

2.4.2.2 Test de Kolmogorov Smirnoff et du χ^2 (khi-deux)

La qualité de l'ajustement des deux distributions mesurée par le test de Kolmogorov Smirnoff est donnée par le niveau de signification (p-value) p = 0.1085. Le nombre de simulations étant grand, on peut donc conclure qu'il n'y a pas de différence entre les deux distributions.

Puisque l'on s'intéresse à des probabilités spécifiques, un test du khi-deux basé sur ces probabilités a été construit. Les polygones de fréquences associés aux distributions $F_R(\bar{\phi})$ et $F_C(\bar{\phi})$ ont été répartis suivant 14 classes, favorisant l'observation des fréquences absolues aux ailes des distributions (figure 2.6). Les fréquences théoriques sont celles de la distribution comparative $F_C(\bar{\phi})$ tandis que les fréquences observées sont celles de la distribution de référence $F_R(\bar{\phi})$. Le niveau de signification (p-value) est p = 0.2645103. Le test ne permet pas de rejeter l'hypothèse d'égalité des deux populations pour les niveaux α = 1%, 5%, 10% et 20%.

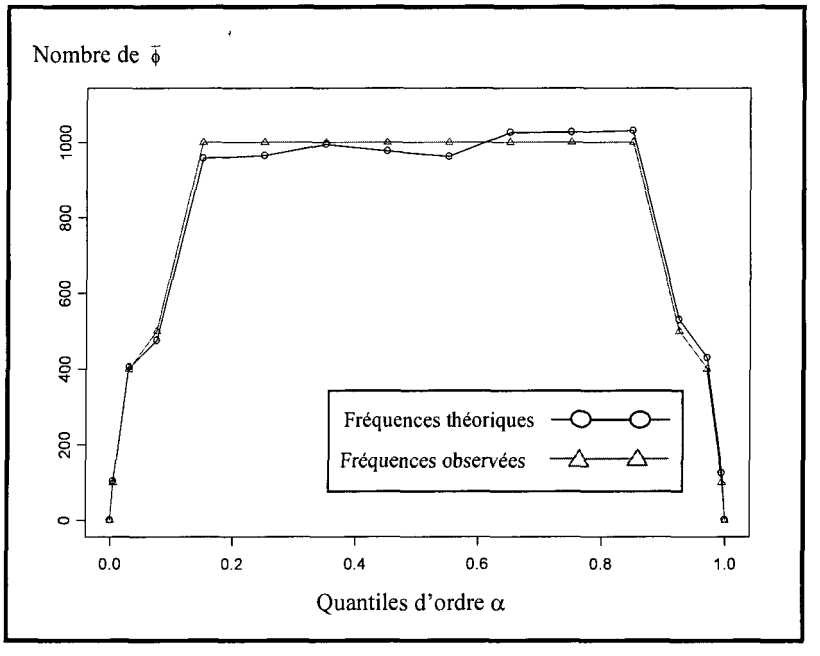

Figure 2.6 Polygones de fréquences des distributions empiriques $F_R(\bar{\phi})$ et $F_C(\bar{\phi})$

2.4.2.3 Quantiles empiriques

On veut déterminer si les probabilités $P(\bar{\phi} \leq Q_\alpha)$ et $P(\bar{\phi} \leq Q_{1-\alpha})$ pour les niveaux α = 1%, 2.5%, 5%, et 10% sont les mêmes sur les deux distributions. Pour ce faire, chaque quantile empirique d'ordre α de la distribution de référence $F_R(\bar{\phi})$ a été obtenu et chaque quantile inverse soit $\hat{\alpha}$ a été calculé à partir de la distribution comparative $F_C(\bar{\phi})$. Les niveaux α et $\hat{\alpha}$ ont ensuite été comparés.

Le tableau 2.2 présente les niveaux $\hat{\alpha}$ de la distribution comparative $F_C(\bar{\phi})$, en considérant des précisions de ± 0.0026, 0.0043, ± 0.0050 et ± 0.0051 pour les niveaux 1%, 2.5%, 5%, et 10% respectivement. On remarque que tous les niveaux $\hat{\alpha}$ sont acceptables.

Niveau α	Niveau $\hat{\alpha}$ de la distribution $F_C(\bar{\phi})$	$\lvert\alpha - \hat{\alpha}\rvert$
1 %	0.0103 ± 0.0026	0.0003
2.5 %	0.0248 ± 0.0035	0.0002
5 %	0.0507 ± 0.0049	0.0007
10 %	0.0982 ± 0.0051	0.0018
90 %	0.8915 ± 0.0051	0.0085
95 %	0.9445 ± 0.0049	0.0055
97.5 %	0.9727± 0.0043	0.0023
99 %	0.9874 ± 0.0026	0.0026

Tableau 2.2 Niveaux empiriques de la distribution empirique comparative $F_C(\bar{\phi})$ selon les quantiles empiriques d'ordre α = 1, 2.5, 5 et 10 % calculés à partir de la distribution empirique $F_R(\bar{\phi})$

2.4.2.4 Conclusion

Une distribution de la statistique $\bar{\phi}$ pour une taille d'échantillon de n=10 a été obtenue de la population de référence de 2600 individus et de la population de 9899 individus. Différentes méthodes de comparaison des distributions ont été utilisées afin de vérifier si la taille de la population de référence est assez grande pour être représentative d'une population réelle.

Les deux distributions sont semblables et la seule différence importante est un plus grand nombre de valeurs extrêmes pour la population comparative de 9899 individus. On peut conclure que la taille de la population de référence est suffisante pour effectuer les études par simulation.

2.4.3 Distribution de $\bar{\phi}$

La distribution normale étant une référence pour la construction d'intervalles de confiance et de tests d'hypothèses, on propose de comparer la distribution empirique de notre population de référence à une distribution normale. Comme l'ont suggéré Hauk et Martin (1984), une transformation logarithmique est appliquée sur la statistique et les distributions sont comparées de nouveau. Pour les comparer, les fonctions de densité et les quantiles empiriques tels que décrits antérieurement seront observés.

La distribution empirique est obtenue en générant m=10000 échantillons de taille n issue de la population de 2600 individus puis en calculant le coefficient d'apparentement moyen $\bar{\phi}$ sur chaque échantillon tiré.

2.4.3.1 Normalité

La distribution empirique de la population de référence $F_R(\bar{\phi})$ a été comparée à une distribution normale empirique $N(\hat{\mu}_{\bar{\phi}}, \hat{\sigma}^2_{\bar{\phi}})$ où $\hat{\mu}_{\bar{\phi}}$ et $\hat{\sigma}^2_{\bar{\phi}}$ sont respectivement la moyenne observée et la variance observée. Les figures 2.7 et 2.8 présentent les fonctions de densité des deux distributions pour les tailles d'échantillons n = 10, 20, 30, 50, 75 et 100. Elles démontrent une asymétrie visible donnant une proportion trop faible à gauche et trop forte à droite, avec un décalage notable de la moyenne observée.

Le tableau 2.3 présente les niveaux $\hat{\alpha}$ de la distribution de référence, calculés à partir des quantiles empiriques d'ordre α = 1%, 2.5%, 5%, et 10% de la distribution normale $N(\hat{\mu}_{\bar{\phi}}, \hat{\sigma}^2_{\bar{\phi}})$. Les niveaux $\hat{\alpha}$ confirment l'asymétrie observée à partir des fonctions de densité.

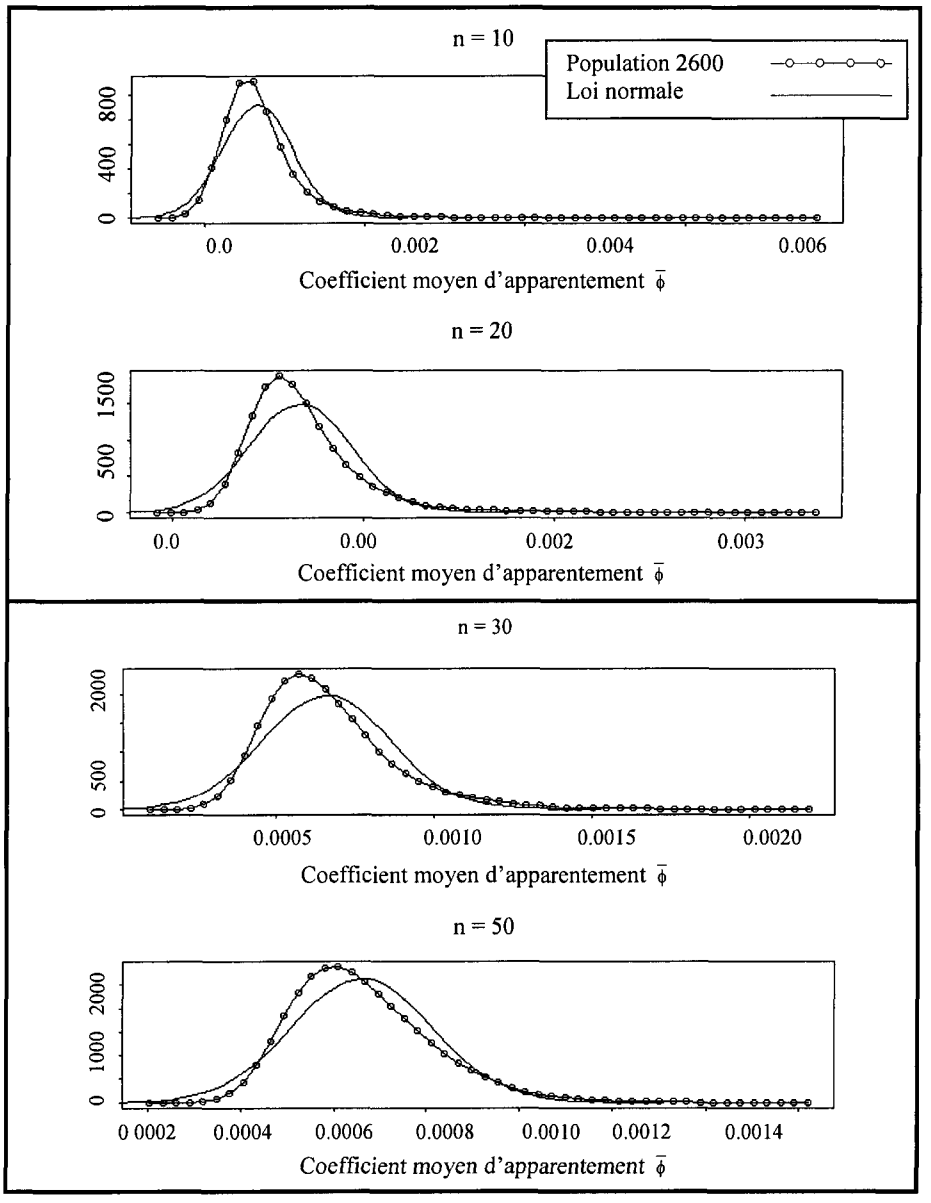

Figure 2.7 *Fonction de densité des distributions empiriques* $F_R(\bar{\phi})$ *et* $N(\hat{\mu}_{\bar{\phi}}, \hat{\sigma}^2_{\bar{\phi}})$ *pour des échantillons de taille n = 10, 20, 30 et 50*

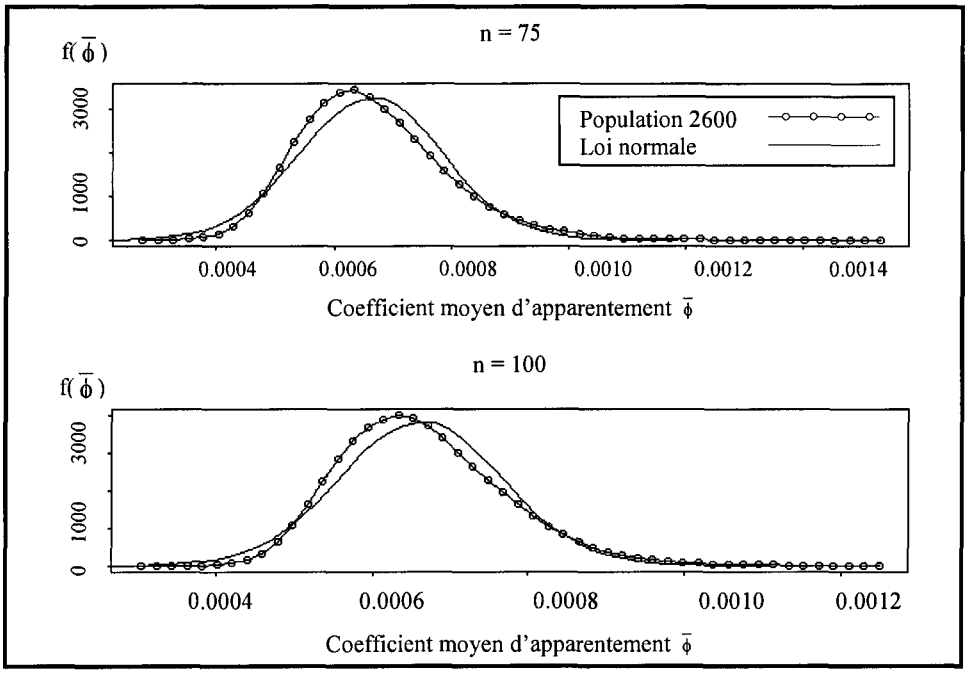

Figure 2.8 *Fonction de densité des distributions empiriques* $F_R(\bar{\phi})$ *et* $N(\hat{\mu}_{\bar{\phi}}, \hat{\sigma}^2_{\bar{\phi}})$ *pour des échantillons de taille n = 75 et 100*

Niveau $\hat{\alpha}$ de la distribution $F_R(\bar{\phi})$						
α \ n	10	20	30	50	75	100
1 %	0.0000	0.0000	0.0000	0.0000	0.0001	0.0007
2.5 %	0.0000	0.0000	0.0000	0.0019	0.0041	0.0053
5 %	0.0000	0.0001	0.0048	0.0125	0.0198	0.0230
10 %	0.0000	0.0149	0.0413	0.0614	0.0736	0.0784
90%	0.0822	0.1005	0.1086	0.1112	0.1095	0.1113
95 %	0.0555	0.0659	0.0694	0.0691	0.0646	0.0643
97.5 %	0.0415	0.0446	0.0500	0.0406	0.0405	0.0385
99 %	0.0296	0.0305	0.0319	0.0248	0.0249	0.0220

Tableau 2.3 *Niveaux empiriques de la distribution de référence* $F_R(\bar{\phi})$ *selon les quantiles empiriques d'ordre* α *= 1, 2.5, 5 et 10 % calculés à partir de la distribution empirique* $N(\hat{\mu}_{\bar{\phi}}, \hat{\sigma}^2_{\bar{\phi}})$

2.4.3.2 Transformation

Une transformation logarithmique de la statistique $\bar{\phi}$ a été considérée afin d'obtenir une distribution plus près d'une loi normale. Les figure 2.9 et 2.10 présentent les fonctions de densité dont la statistique a subit une transformation logarithmique pour les tailles d'échantillon. Elles démontrent une asymétrie moins importante tout en étant encore trop loin de la normalité aux queues des distributions. Le tableau 2.4 présente les niveaux $\hat{\alpha}$ de la distribution de référence, calculés à partir des quantiles empirique d'ordre α = 1%, 2.5%, 5%, et 10% de la distribution normale après transformation logarithmique.

Niveau $\hat{\alpha}$ de la distribution $F_R(\bar{\phi})$						
α \ n	10	20	30	50	75	100
1 %	0.0018	0.0017	0.0029	0.00240	0.0047	0.0007
2.5 %	0.0093	0.0088	0.0149	0.0158	0.0179	0.0173
5 %	0.0265	0.0309	0.0351	0.0378	0.0407	0.0417
10 %	0.0801	0.0871	0.0901	0.0946	0.0960	0.0975
90%	0.1131	0.1132	0.1143	0.1124	0.1094	0.1108
95 %	0.0658	0.0637	0.0635	0.0589	0.0593	0.0552
97.5 %	0.0397	0.0360	0.0368	0.0293	0.0309	0.0286
99 %	0.0216	0.0211	0.0181	0.0157	0.0152	0.0148

Tableau 2.4 Niveaux empiriques de la distribution de référence $F_R(\bar{\phi})$ selon les quantiles empiriques d'ordre α = 1, 2.5, 5 et 10 % calculés à partir de la distribution empirique $N(\hat{\mu}_{\bar{\phi}}, \hat{\sigma}^2_{\bar{\phi}})$ *après transformation logarithmique de la statistique*

On observe que les niveaux $\hat{\alpha}$ sont médiocres même en considérant une précision de ± 0.0026,± 0.0043, ± 0.0050 et ± 0.0051 pour les niveaux 1%, 2.5%, 5%, et 10% respectivement.

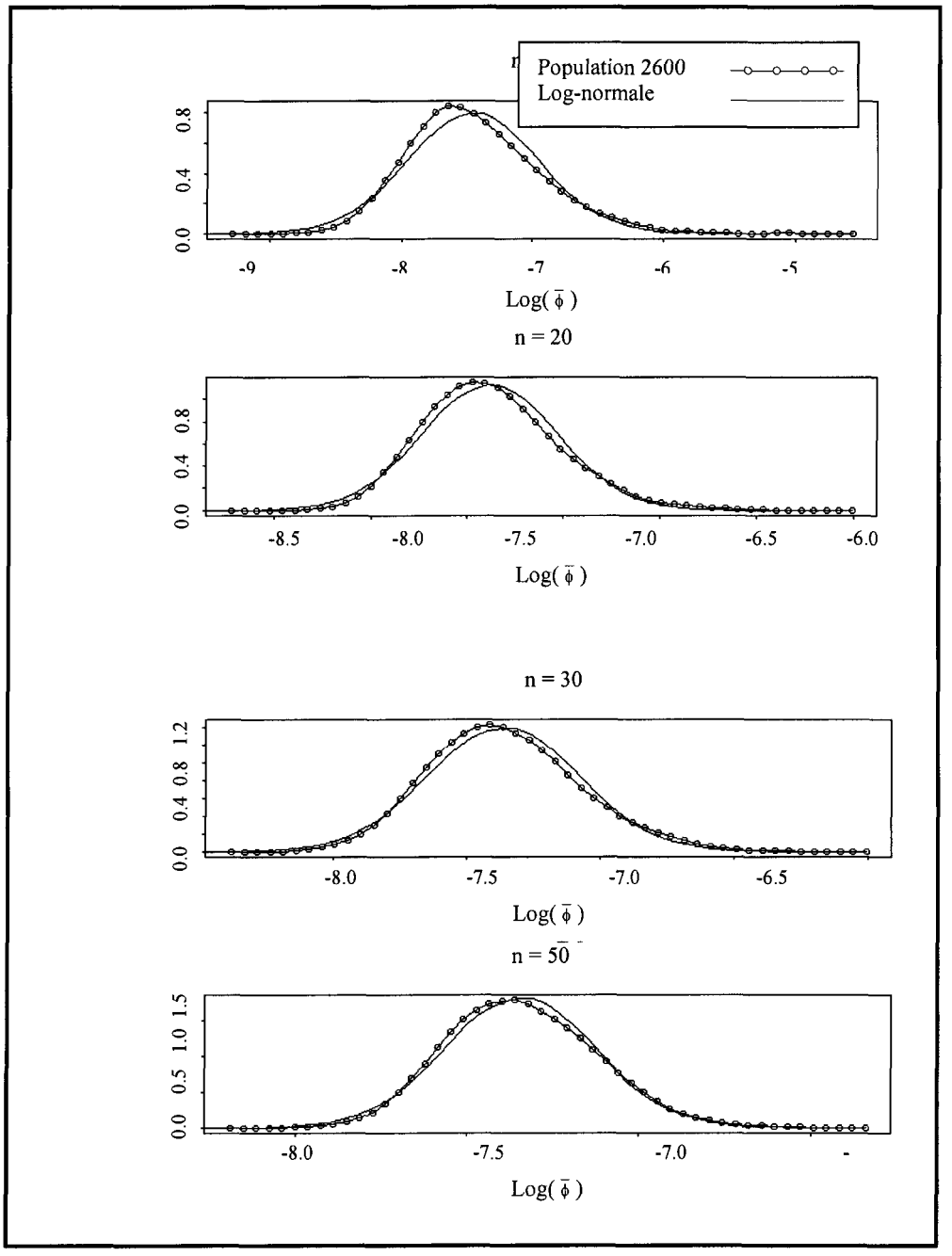

Figure 2.9 Fonction de densité des distributions empiriques $F_R(\bar{\phi})$ *et* $N(\hat{\mu}_{\bar{\phi}}, \hat{\sigma}^2_{\bar{\phi}})$ *après transformation logarithmique de la statistique pour des échantillons de taille n = 10, 20, 30 et 50*

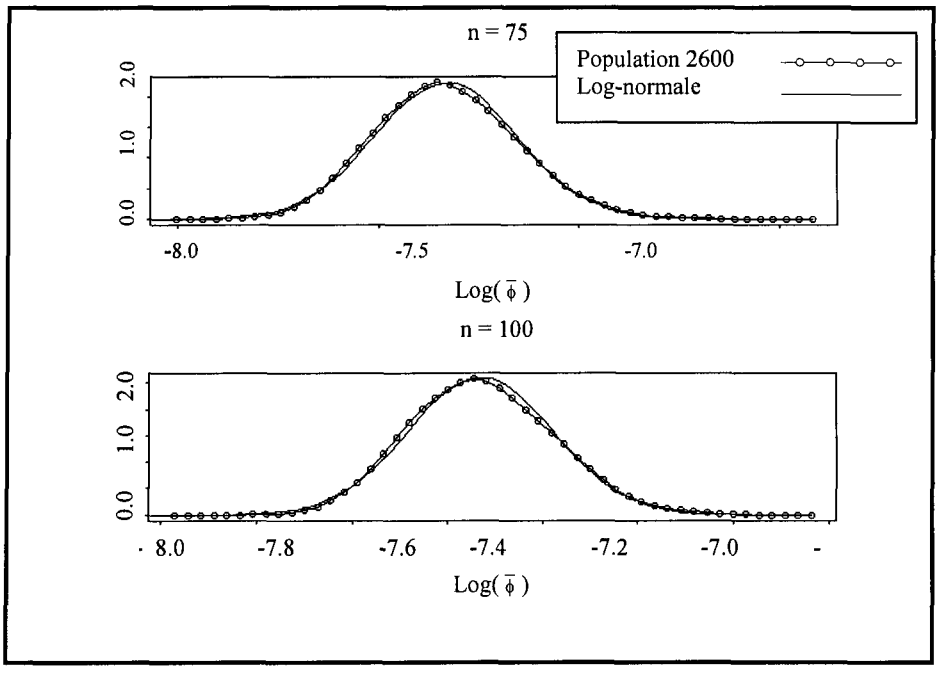

Figure 2.10 Fonction de densité des distributions empiriques $F_R(\bar{\phi})$ *et* $N(\hat{\mu}_{\bar{\phi}}, \hat{\sigma}^2_{\bar{\phi}})$ *après transformation logarithmique de la statistique pour des échantillons de taille n = 75 et 100*

2.5 Conclusion

Dans le cas d'échantillons de tailles restreintes, on ne peut utiliser l'approximation normale de la distribution de la statistique $\bar{\phi}$ d'autant plus que les différences observées se retrouvent aux ailes des distributions et que ces probabilités sont les plus importantes pour la construction d'un intervalle de confiance et d'un test.

Une transformation logarithmique a permis d'obtenir une distribution beaucoup plus proche d'une loi normale. Cependant, même en considérant une taille d'échantillon de n = 100, il existe encore une différence importante aux queues des distributions.

CHAPITRE 3

INTERVALLES DE CONFIANCE

3.1 Introduction

Le coefficient moyen d'apparentement est un indice de l'homogénéité du pool génétique d'une population. Les études qui utilisent cet indice ne présentent que cette valeur, sans tenir compte de la variation échantillonnale, étant donné que le calcul classique de cette variation ne s'applique pas dans ce contexte. On propose donc dans ce chapitre d'utiliser des méthodes de rééchantillonnage pour construire un intervalle de confiance pour μ_ϕ, l'apparentement moyen d'une population donnée.

Cinq méthodes ont été considérées pour le calcul d'un intervalle de confiance. Les trois premières méthodes sont basées sur une hypothèse de normalité du coefficient moyen tandis que les deux dernières sont entièrement non paramétriques. Les méthodes basées sur la normalité repose sur le mythe que pour des tailles d'échantillons n = 30, l'approximation normale est valide. Nous avons gardé cette règle puisque c'est celle qui est utilisée en pratique par les chercheurs.

On considère μ_ϕ, le coefficient moyen d'apparentement d'une population, ξ_n un échantillon de n individus et le coefficient moyen de l'échantillon $\bar{\phi}$. On cherche à comparer les différentes méthodes par simulation. On s'intéresse alors aux niveaux de confiance réels des différentes méthodes pour des niveaux généralement utilisés soit 80%, 90%, 95% et 99% et pour des tailles d'échantillons relativement petites n = 10, 20, 30, 40, 50, 75 et 100.

3.2 Intervalles de confiance

3.2.1 Classique

Soit un échantillon de n valeurs telles que les mesures sont relativement symétriques. L'intervalle de confiance classique, suggéré dans les cours élémentaires de statistique (Hogg et al., 1970) est donné par

$$\left(\overline{X} - z_{\alpha/2}\, S/\sqrt{n}\,, \overline{X} + z_{\alpha/2}\, S/\sqrt{n}\right),$$

où $\overline{X}$ est la moyenne des n valeurs, S est l'écart-type et $z_{\alpha/2}$ est le point critique de niveau $\alpha/2$ pour une distribution normale centrée réduite. C'est l'intervalle pour une taille d'échantillons de n > 30. Lorsque la taille de l'échantillon est plus petite et que les mesures suivent une loi approximativement normale, l'intervalle est donné en substituant le point critique $z_{\alpha/2}$ par celui d'une loi de Student à n – 1 degrés de liberté.

En considérant $\xi_n = X_1, X_2, \ldots, X_n$ un échantillon de n individus, l'intervalle de confiance classique équivalent de niveau $1-\alpha$ pour μ_ϕ est donné par $\left(\overline{\phi} \pm z_{\alpha/2}\, S/\sqrt{n}\right)$ où

$$S^2 = \frac{2}{(n(n-1)-2)} \sum_{i}^{n-1} \sum_{j>i}^{n} \left(\phi_{ij} - \overline{\phi}\right)^2 .$$

L'estimation de la variance de ϕ est faite en utilisant les $n(n-1)/2$ valeurs différentes de la matrice Φ triangulaire supérieure, la diagonale principale étant exclue. Cette méthode ne devrait pas fournir de bons résultats puisque la variance est biaisée, étant donné la dépendance entre les éléments ϕ_{ij}.

3.2.2 Méthode du Jackknife

Lorsque les données n'offrent pas la possibilité d'obtenir un estimateur non biaisé de la variance, la méthode du Jackknife (Tukey, 1958) permet de contourner cette difficulté en fournissant une estimation de la variance.

Considérons un échantillon de n valeurs, θ la statistique calculée sur l'échantillon, $\theta_{(i)}$ la statistique calculée sur l'échantillon en omettant la i-ième valeur et θ_J la moyenne des n-1 valeurs de $\theta_{(i)}$. L'écart-type estimé par la méthode du Jackknife est donné par

$$\hat{S}_J = \sqrt{\frac{n-1}{n}\sum_{i=1}^{n}\left(\theta_{(i)} - \overline{\theta}_J\right)^2}\ .$$

En supposant une distribution proche d'une distribution normale pour l'estimateur, l'intervalle de confiance, est donné par $\left(\overline{X} \pm z_{\alpha/2}\hat{S}_J\right)$.

Dans le contexte du coefficient d'apparentement, l'estimateur de la variance par la méthode du Jackknife est donné par

$$\hat{S}_J = \sqrt{\frac{n-1}{n}\sum_{i=1}^{n}\left(\overline{\phi}_{(i)} - \overline{\phi}_J\right)^2}\ ,$$

où $\overline{\phi}_{(i)}$ est le i-ième coefficient d'apparentement moyen des $n-1$ individus obtenu en omettant l'individu i et $\overline{\phi}_J$ est la moyenne des n valeurs de $\overline{\phi}_{(i)}$. L'intervalle de confiance de niveau α est alors donné par $\left(\overline{\phi} \pm z_{\alpha/2}\hat{S}_J\right)$. Pour des échantillons de tailles $n \leq 30$, une loi

de Student a été utilisée pour déterminer le point critique permettant l'amélioration du niveau de l'intervalle.

Cette méthode sera adéquate dans la mesure où la taille de l'échantillon est assez grande pour obtenir une bonne estimation de la variance et que la distribution de la statistique $\bar{\phi}$ soit près d'une loi normale.

3.2.3 Transformation log-normale

La méthode classique et la méthode Jackknife sont basées sur l'hypothèse de normalité du coefficient $\bar{\phi}$. L'étude de la distribution de $\bar{\phi}$ indique une distribution légèrement asymétrique. Il a été démontré au chapitre deux qu'une transformation logarithmique permet de rendre la distribution de la moyenne plus près d'une distribution normale.

Suite à cette transformation, l'intervalle de confiance considéré de niveau α pour μ_ϕ est $\left(e^{\log(\bar{\phi})-z_{\alpha/2}\hat{S}_L}, e^{\log(\bar{\phi})+z_{\alpha/2}\hat{S}_L}\right)$ où $\hat{S}_L$ est obtenu par la méthode du Jackknife :

$$\hat{S}_L = \sqrt{\frac{n-1}{n}\sum_{i=1}^{n}\left(\log(\bar{\phi}_{(i)}) - \log(\bar{\phi}_J)\right)^2},$$

où $\log(\bar{\phi}_{(i)})$ est le logarithme du coefficient moyen d'apparentement en omettant l'individu i et $\log(\bar{\phi}_J)$ est le logarithme de la moyenne des n valeurs de $\bar{\phi}_{(i)}$.

3.2.4 Méthode du Bootstrap

Une autre forme de rééchantillonnage permettant l'estimation des limites de confiance est la méthode du Bootstrap (Palm, 2002). Ses limites sont données par les quantiles de la distribution d'échantillonnage empirique construite par la méthode du Bootstrap (Efron, 1979).

Soit un échantillon de n individus ξ_n. Pour obtenir un intervalle de confiance de niveau α pour μ_ϕ par la technique du Boostrap, il faut rééchantillonner avec remise les n individus afin de construire la distribution empirique de $\overline{\phi}$. Le nombre de répétitions suggéré pour évaluer les quantiles de la distribution est de $B \geq 1000$ (Efron et Tibshirani, 1993). L'algorithme est donné à la figure 3.1.

Pour i de 1 à B
Tirer ξ_i^*, un échantillon aléatoire avec remise de n individus.
Calculer Φ_i^*, l'apparentement des n individus.
Évaluer l'apparentement moyen $\overline{\phi}_i^*$ de la matrice Φ_i^*.
Calculer les quantiles d'ordre $\alpha/2$ et $(1-\alpha/2)$ des valeurs de $(\overline{\phi}_1^*, \overline{\phi}_2^*, \ldots, \overline{\phi}_B^*)$.

Figure 3.1 : Algorithme de construction d'un intervalle de confiance par la méthode du bootstrap

3.2.5 Méthode du Bootstrap Bca

La méthode Bca «**B**iais **c**orrected **a**cceleration» est dérivée directement du Bootstrap (Efron et Tibshirani, 1993). Elle consiste à transformer la distribution empirique de manière à obtenir une distribution qui suit une loi normale centrée réduite. Cette

transformation est basée sur l'ajustement de deux paramètres $\hat{z}_0$ et $\hat{a}$ qui sont respectivement le biais corrigé, assurant l'ajustement de la moyenne et l'accélération, assurant la correction de la variance. Les bornes de l'intervalle de confiance sont données par l'application de la transformation inverse aux points critiques $z_{\alpha/2}$ et $z_{1-\alpha/2}$ de la distribution normale centrée réduite.

Soit un échantillon de n individus ξ_n et soit $\bar{\phi}$ la statistique observée. Pour obtenir un intervalle de confiance de niveau $1-\alpha$ pour μ_ϕ par la technique du Boostrap Bca, il faut rééchantillonner avec remise les n individus afin de construire la distribution empirique de la statistique. Le biais corrigé est donné par la proportion de la distribution empirique qui est inférieure à la statistique observée soit :

$$\hat{z}_0 = \frac{\sum_{i=1}^{B} I_{\bar{\phi}_i^* < \bar{\phi}}}{B}$$

où I est la fonction indicatrice, $\bar{\phi}_i^*$ est la statistique estimée par la méthode du bootstrap et B le nombre de répétitions. Efron et Tibshirani (1993) proposent d'utiliser la méthode du Jackknife pour estimer l'accélération. Elle est donnée par :

$$\hat{a} = \frac{\sum_{i=1}^{n} \left(\bar{\phi}_{(i)} - \bar{\phi}_J \right)^3}{6 \cdot \left[\sum_{i=1}^{n} \left(\bar{\phi}_{(i)} - \bar{\phi}_J \right)^2 \right]^{3/2}}$$

où $\bar{\phi}_{(i)}$ est l'estimation du paramètre observé à partir de l'échantillon ξ_n en omettant la i-ième observation et $\bar{\phi}_J$ est la moyenne des n valeurs de $\bar{\phi}_{(i)}$. Les bornes de l'intervalle de confiance de niveau $1-\alpha$ sont alors données par

$$x_1 = \hat{z}_0 + \frac{\hat{z}_0 + z_{\alpha/2}}{1-\hat{a}(\hat{z}_0 + z_{\alpha/2})} \qquad \text{et} \qquad x_2 = \hat{z}_0 + \frac{\hat{z}_0 + z_{1-\alpha/2}}{1-\hat{a}(\hat{z}_0 + z_{1-\alpha/2})},$$

où $z_{\alpha/2}$ et $z_{1-\alpha/2}$ sont les points critiques d'une loi normale centrée réduite. Le nombre de répétitions suggéré pour cette méthode est de $B \geq 5000$ (Efron et Tibshirani, 1993) pour l'évaluation des quantiles de la distribution. L'algorithme est donné à la figure 3.2.

Pour i de 1 à B (rééchantillonnage Bootstrap)
- Tirer ξ_i^*, un échantillon aléatoire avec remise de n individus.
- Calculer Φ_i^*, l'apparentement des n individus.
- Évaluer l'apparentement moyen $\bar{\phi}_i^*$ de la matrice Φ_i^*.
- Évaluer la fonction indicatrice $I_{\bar{\phi}_i^* < \bar{\phi}}$.

Évaluer $\hat{z}_0$

Pour i de 1 à B (rééchantillonnage Jaccknife)
- Tirer ξ_i^*, un échantillon aléatoire de n-1 individus en omettant l'individu i
- Calculer Φ_i^*, l'apparentement des individus.
- Évaluer l'apparentement moyen $\bar{\phi}_i^*$ de la matrice Φ_i^*.

Évaluer $\bar{\phi}_J$, la moyenne des $\bar{\phi}_i^*$.

Calculer l'accélération $\hat{a}$.

Calculer les bornes de l'intervalle x_1 et x_2.

Calculer $\hat{\alpha}/2$ et $(1-\hat{\alpha}/2)$ donnés par les quantiles inverses $Q_{x_1}^{-1}$ et $Q_{x_2}^{-1}$ d'une distribution normale centrée réduite aux points critiques x_1 et x_2.

Évaluer $Q_{\hat{\alpha}/2}$ et $Q_{1-\hat{\alpha}/2}$ de la distribution des $\bar{\phi}_i^*$.

Figure 3.2 : Algorithme de construction d'un intervalle de confiance par la méthode du Bootstrap Bca

3.3 Résultats

Les cinq méthodes proposées ont été comparées par simulation afin de valider les niveaux réels de chaque intervalle de confiance. Des échantillons de tailles $n = 10$, 20, 30, 40, 50, 75 et 100 ont été tirés de la population de référence et les intervalles de confiance construits à des niveaux de confiance $1-\alpha = 99\%$, 95%, 90% et 80%. Pour chaque taille d'échantillons et chaque niveau fixé, 10000 répétitions ont été effectuées afin d'obtenir une grande précision.

3.3.1 Niveaux

Pour chaque taille et chaque niveau de confiance, le niveau réel de l'intervalle de confiance est estimé par la proportion d'intervalles simulés qui contiennent la vraie valeur du paramètre μ_ϕ soit 6.638297×10^{-4}, la moyenne de la population de référence. En utilisant 10000 répétitions, les niveaux réels sont estimés avec des précisions de ± 0.0026,± 0.0043, ± 0.0050 et ± 0.0051 pour les niveaux respectifs 99%, 95%, 90% et 80%.

Niveau $\alpha = 0.01$							
Intervalles	n = 10	n = 20	n = 30	n = 40	n = 50	n = 75	n = 100
Classique	0.0822	**0.0329**	**0.0153**	**0.0095**	**0.0061**	0.0022	0.0009
Jackknife	0.1154	0.0851	0.0727	0.0683	0.0565	0.0386	0.0351
Log-normal	**0.0419**	0.0422	0.0347	0.0376	0.0336	0.0234	0.0191
Bootstrap	0.0984	0.0579	0.0414	0.0344	0.0286	0.0200	0.0199
Bootstrap Bca	0.0813	0.0384	0.0254	0.0213	0.0192	**0.0142**	**0.0120**
Niveau $\alpha = 0.05$							
Intervalles	n = 10	n = 20	n = 30	n = 40	n = 50	n = 75	n = 100
Classique	0.1628	**0.0703**	**0.0418**	0.0315	0.0232	0.0114	0.0084
Jackknife	0.1978	0.1465	0.1244	0.1181	0.1017	0.0854	0.0749
Log-normal	**0.1105**	0.0966	0.0918	0.0885	0.0774	0.0671	0.0620
Bootstrap	0.1878	0.1237	0.1041	0.0908	0.0790	0.0663	0.0623
Bootstrap Bca	0.1574	0.0943	0.0764	**0.0680**	**0.0669**	**0.0534**	**0.0526**
Niveau $\alpha = 0.10$							
Intervalles	n = 10	n = 20	n = 30	n = 40	n = 50	n = 75	n = 100
Classique	0.2157	**0.1057**	**0.0711**	0.0569	0.0442	0.0246	0.0190
Jackknife	0.2469	0.1913	0.1673	0.1587	0.1444	0.1261	0.1165
Log-normal	**0.1792**	0.1505	0.1389	0.1373	0.1254	0.1137	0.1084
Bootstrap	0.2509	0.1869	0.1642	0.1489	0.1339	0.1164	0.1100
Bootstrap Bca	0.2184	0.1467	0.1343	**0.1219**	**0.1162**	**0.1030**	**0.0999**
Niveau $\alpha = 0.20$							
Intervalles	n = 10	n = 20	n = 30	n = 40	n = 50	n = 75	n = 100
Classique	0.2891	**0.1717**	0.1305	0.1088	0.0933	0.0660	0.0553
Jackknife	0.3151	0.2643	0.2430	0.2372	0.2268	0.2102	0.205
Log-normal	**0.2746**	0.2421	**0.2276**	0.2292	0.2179	0.2051	0.2028
Bootstrap	0.3478	0.2869	0.2667	0.2399	0.2327	0.2132	0.2035
Bootstrap Bca	0.3163	0.2523	0.2335	**0.2183**	**0.2136**	**0.2020**	**0.2024**

Tableau 3.1: Niveaux empiriques des cinq méthodes de construction d'intervalles de confiance pour 10000 simulations. Le gras souligne les meilleurs niveaux empiriques.

Le tableau 3.1 présente l'estimation des niveaux réels pour les cinq méthodes. On remarque que les niveaux réels de toutes les méthodes, à l'exception de la méthode classique, convergent vers les niveaux théoriques lorsque la taille des échantillons augmente. L'intervalle classique semble donner des résultats intéressants pour quelques tailles d'échantillons mais une étude approfondie montre que, conformément à la théorie, les résultats sont médiocres. Bien qu'étant très près des niveaux espérés pour certaines tailles d'échantillons, les niveaux réels de l'intervalle classique deviennent moins précis lorsque la taille des échantillons augmente. Cette caractéristique est en fait un estimateur par intervalle à rejeter en toutes circonstances.

La transformation log-normale est la méthode qui procure les meilleurs niveaux pour de petites tailles d'échantillons (10 et 20) tandis que dans tous les autres cas, c'est la méthode du Bootstrap Bca qui donne les résultats les plus intéressants. Il faut cependant noter que pour des tailles $n \geq 20$, dans presque tous les cas la précision de l'estimation des niveaux réels suffit à expliquer la différence observée. En considérant les niveaux les plus utilisés soit 90% et 95%, les meilleurs estimateurs par intervalles obtenus sont libéraux, c'est-à-dire de longueur plus petite que les intervalles espérés.

3.3.2 Symétrie

L'étude de la distribution de $\overline{\phi}$ indique une asymétrie importante pour de petites tailles d'échantillons. Cette asymétrie devrait se refléter dans les intervalles de confiance. C'est pourquoi la partie droite de l'intervalle a été observée séparément de celle de gauche lors de l'évaluation du niveau réel des intervalles.

Posons α_d, la probabilité que μ_ϕ soit à droite de la borne aléatoire supérieure de l'intervalle de confiance et α_g, la probabilité que le paramètre soit à gauche de la borne aléatoire inférieure de l'intervalle de confiance. La somme de α_g et α_d correspond au niveau réel des intervalles, soit $\alpha = \alpha_g + \alpha_d$. Bien que l'asymétrie ne soit pas une contrainte nécessaire pour la construction d'un intervalle de confiance, il est souhaitable que la probabilité que le paramètre estimé se situe en deçà de la borne inférieure ou supérieure soit $\alpha_d = \alpha_g = \alpha/2$.

Le tableau 3.2 présente une estimation des niveaux réels α_g et α_d. On observe effectivement une asymétrie prononcée pour toutes les tailles d'échantillons. En considérant l'intervalle classique proscrit, l'intervalle Bootstrap Bca s'avère le moins asymétrique d'entre tous pour tous les niveaux. L'intervalle log-normal suit l'intervalle Bootstrap Bca pour de petites tailles d'échantillons.

α_g et α_d pour un niveau $\alpha = 0.01$							
Intervalles	n = 10	n = 20	n = 30	n = 40	n = 50	n = 75	n = 100
Classique	0.0000 0.0822	0.0000 0.0329	**0.0000** **0.0153**	**0.0000** **0.0095**	**0.0000** **0.0061**	**0.0000** **0.0022**	**0.0000** **0.0009**
Jackknife	0.0000 0.1154	0.0000 0.0851	0.0000 0.0727	0.0000 0.0683	0.0000 0.0565	0.0000 0.0386	0.0000 0.0351
Log-normal	**0.0004** **0.0415**	0.0003 0.0419	0.0005 0.0342	0.0012 0.0364	0.0014 0.0322	0.0013 0.0221	0.0014 0.0177
Bootstrap	0.0009 0.0975	0.0003 0.0570	0.0011 0.0403	0.0017 0.0327	0.0015 0.0271	0.0017 0.0183	0.002 0.0179
Bootstrap Bca	0.0037 0.0776	**0.0049** **0.0335**	0.0037 0.0217	0.0039 0.0174	0.0032 0.0160	0.0035 0.0107	0.0031 0.0089
α_g et α_d pour un niveau $\alpha = 0.05$							
Intervalles	n = 10	n = 20	n = 30	n = 40	n = 50	n = 75	n = 100
Classique	0.0001 0.1627	0.0002 0.0711	0.0000 0.0418	0.0002 0.0313	0.0001 0.0231	0.0001 0.0113	0.0002 0.0082
Jackknife	0.0001 0.1977	0.0002 0.1463	0.0002 0.1242	0.0003 0.1178	0.0011 0.1006	0.0018 0.0836	0.0024 0.0725
Log-normal	0.0037 0.1068	0.0049 0.0917	0.0079 0.0839	0.0075 0.0810	0.0093 0.0681	0.0097 0.0574	0.0001 0.0520
Bootstrap	0.0053 0.1825	0.0072 0.1165	0.0068 0.0973	0.0089 0.0819	0.0080 0.0710	0.0087 0.0576	0.0107 0.0516
Bootstrap Bca	**0.0397** **0.1177**	**0.0223** **0.0720**	**0.0216** **0.0548**	**0.0228** **0.0452**	**0.0232** **0.0437**	**0.0214** **0.0320**	**0.0225** **0.0301**

Tableau 3.2 Niveaux empiriques de la borne inférieure α_g et de la borne supérieure α_d des cinq méthodes de construction d'intervalles de confiance pour des niveaux α =0.01 et 0.05 et 10000 répétitions. Le gras souligne les meilleurs niveaux empiriques

3.4 Conclusion

L'intervalle classique s'avère effectivement une méthode de construction à proscrire puisque celle-ci se caractérise par l'obtention de niveaux estimés médiocres à mesure que la taille de l'échantillon augmente. En se basant sur l'estimation des niveaux réels des intervalles de confiance construits à l'aide des autres méthodes, l'intervalle de confiance obtenu après transformation logarithmique a obtenu les meilleurs niveaux pour n = 10 et l'asymétrie de cet intervalle est acceptable.

La méthode du Bootstrap Bca s'avère la plus intéressante. Elle donne des niveaux réels plus près des vraies valeurs bien qu'elle comporte une asymétrie notable mais la moins prononcée de toutes les méthodes. Même si l'algorithme de construction de l'intervalle Bootstrap Bca est le plus compliqué, cette méthode vaut la peine d'être utilisée. L'intervalle obtenu sera plus petit que l'intervalle espéré donc il aura moins de chance de recouvrir la vraie valeur du paramètre estimé μ_ϕ.

α_g et α_d pour un niveau $\alpha = 0.10$							
Intervalles	n = 10	n = 20	n = 30	n = 40	n = 50	n = 75	n = 100
Classique	0. 0009 0.2148	0.0008 0.1049	0.0013 0.0698	0.0017 0.0552	0.0020 0.0422	0.0017 0.0229	0.0014 0.0176
Jackknife	0.0002 0.2467	0.0012 0.1901	0.0025 0.1648	0.0038 0.1549	0.0064 0.1380	0.0093 0.1168	0.0109 0.1056
Log-normal	0.0118 0.1674	0.0150 0.1355	0.0185 0.1204	0.0212 0.1161	0.0225 0.1089	0.0221 0.0916	0.0248 0.0836
Bootstrap	0.0120 0.2389	0.0156 0.1713	0.0171 0.1471	0.0194 0.1295	0.0193 0.1146	0.0218 0.0946	0.0247 0.0853
Bootstrap Bca	**0.0399** **0.1785**	**0.0426** **0.1041**	**0.0478** **0.0865**	**0.0471** **0.0748**	**0.0446** **0.0716**	**0.0445** **0.0585**	**0.0451** **0.0548**
α_d et α_g pour un niveau $\alpha = 0.20$							
Intervalles	n = 10	n = 20	n = 30	n = 40	n = 50	n = 75	n = 100
Classique	0.0082 0.2809	0.0093 0.1624	0.0104 0.1201	0.0108 0.0980	0.0108 0.0825	0.0097 0.0563	0.0098 0.0455
Jackknife	0.0055 0.3096	0.0122 0.2521	0.0195 0.2235	0.0271 0.2101	0.0326 0.1942	0.0378 0.1724	0.0440 0.161
Log-normal	0.0309 0.2437	0.0409 0.2012	0.0441 0.1835	0.0531 0.1761	0.0560 0.1619	0.0581 0.1470	0.0630 0.1398
Bootstrap	0.0279 0.3199	0.0370 0.2499	0.0395 0.2272	0.0419 0.1980	0.0474 0.1853	0.0536 0.1596	0.0579 0.1456
Bootstrap Bca	**0.0816** **0.2347**	**0.0923** **0.1600**	**0.0961** **0.1374**	**0.0921** **0.1262**	**0.0923** **0.1213**	**0.0931** **0.1089**	**0.0985** **0.1079**

Tableau 3.3 Niveaux empiriques de la borne inférieure α_g et de la borne supérieure α_d des cinq méthodes de construction d'intervalles de confiance pou des niveaux rα =0.10 et 0.20 et 10000 répétions. Le gras souligne les meilleurs niveaux empiriques

CHAPITRE 4

COMPARAISON DE DEUX MOYENNES

4.1 Introduction

La comparaison de deux coefficients moyens d'apparentement peut se faire en comparant les intervalles de confiance mais ces derniers ne sont pas très fiables. De plus, la technique est moins puissante qu'avec un test d'hypothèse. On s'intéresse donc aux hypothèses

$$\begin{aligned} H_0 &: \mu_{\phi_1} = \mu_{\phi_2} \\ H_1 &: \mu_{\phi_1} \neq \mu_{\phi_2} \end{aligned} \quad (1) \qquad \text{et} \qquad \begin{aligned} H_0 &: \mu_{\phi_1} = \mu_{\phi_2} \\ H_1 &: \mu_{\phi_1} > \mu_{\phi_2} \end{aligned} \quad (2),$$

où μ_{ϕ_1} et μ_{ϕ_2} sont les coefficients moyens d'apparentement de deux populations respectives, pour un test bilatéral dans le premier cas et un test unilatéral dans le second cas.

Puisque le problème de dépendance des mesures est toujours présent, les tests conventionnels ne sont pas appropriés. On considère deux tests non paramétriques qui permettent de contrôler la dépendance: le test de permutation et le test du Bootstrap. Ces tests seront comparés au test classique de Student.

Dans ce chapitre, on cherche à établir les niveaux réels et la puissance relative des différents tests d'hypothèses pour les niveaux 99%, 95%, 90% et 80% et des tailles d'échantillons relativement petites n = 10, 20, 30, 40, 50, 75 et 100.

4.2 Tests d'hypothèses

4.2.1 Student

Soit deux échantillons de n et m valeurs provenant de deux populations distinctes. Lorsque les mesures des échantillons sont assez près d'une loi normale et que leurs variances sont considérées égales, le test de Student, tel que décrit dans la littérature, consiste à calculer la statistique

$$T = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{(\frac{1}{n} + \frac{1}{m})(\frac{(n-1)S_1^{\,2} + (m-1)S_2^{\,2}}{n+m-2})}},$$

où $S_1^{\,2}$ et $S_2^{\,2}$ sont les variances échantillonnales. La statistique T, sous l'hypothèse nulle, suit une distribution de Student à (n – m – 2) degrés de liberté et la règle de décision du test bilatéral est de rejeter H_0 si $|T| > t_{\alpha/2;(n-m-2)}$.

Soit ξ_1 et ξ_2 deux échantillons de n et m individus prélevés respectivement des populations 1 et 2. L'équivalent du test de Student pour les hypothèses (1) et (2) est basé sur la statistique

$$T = \frac{\overline{\phi}_1 - \overline{\phi}_2}{\sqrt{(\frac{1}{n} + \frac{1}{m})(\frac{(n-1)S_1^{\,2} + (m-1)S_2^{\,2}}{n+m-2})}},$$

où $\overline{\phi}_1$ et $\overline{\phi}_2$ sont les coefficients d'apparentement moyen des échantillons ξ_1 et ξ_2 respectivement et $S_1^{\,2}$ et $S_2^{\,2}$ sont les variances des échantillons.

Les variances ne peuvent être estimées sans biais par S_1^2 et S_2^2, c'est pourquoi il est préférable de remplacer ses estimateurs par $\hat{S}_{J1}^2$ et $\hat{S}_{J2}^2$, où ces variances sont respectivement celles de ξ_1 et ξ_2 estimées par la méthode du Jackknife. Dans la mesure où la statistique $\bar{\phi}_1 - \bar{\phi}_2$ est proche d'une loi normale, ce test devrait donner des résultats acceptables surtout si les tailles d'échantillons sont grandes.

4.2.2 Test de permutation

L'étude de la distribution de $\bar{\phi}$ laisse présager que la statistique du test est loin d'une loi normale. Il serait alors avantageux d'utiliser un test non paramétrique. Le test de permutation est un test non paramétrique qui consiste à simuler la distribution de la statistique du test sous l'hypothèse H_0, en permutant les deux échantillons aléatoirement. La région de rejet, pour un niveau α fixé, est estimée à l'aide de cette distribution (Efron et al., 1993).

Dans le cas de deux échantillons indépendants ξ_1 et ξ_2 de taille n et m respectivement, la statistique observée est $\bar{\phi}_d = \bar{\phi}_1 - \bar{\phi}_2$. Le test de permutation consiste à permuter les n + m individus de l'échantillon et à recalculer la statistique un certain nombre de fois. Posons $\bar{\phi}_d^*$, la statistique obtenue lors de la i-ième permutation. La règle de décision pour le test bilatéral (1) est d'accepter H_0 si la différence $\bar{\phi}_d$ se situe dans l'intervalle $Q_{\alpha/2} \leq \bar{\phi}_d \leq Q_{(1-\alpha/2)}$ où Q_α est le quantile d'ordre α de la distribution de $\bar{\phi}_d^*$.

Pour le test unilatéral (2), la région d'acceptation est donnée par $\bar{\phi}_d \leq Q_{(1-\alpha)}$. L'algorithme est présenté à la figure 4.1.

Évaluer $\bar{\phi}_d = \bar{\phi}_1 - \bar{\phi}_2$.
Pour i de 1 à B fois
Permuter les (n + m) individus de $\xi_1 \cup \xi_2$.
Prendre $\xi^*_{1(i)}$, les n premier individus.
Prendre $\xi^*_{2(i)}$, les m individus restant.
$\Phi^*_{1(i)}$: apparentement des n individus.
$\Phi^*_{2(i)}$: apparentement des m individus.
Évaluer l'apparentement moyen de $\xi^*_{1(i)}$ et $\xi^*_{2(i)}$.
Évaluer $\bar{\phi}^*_{d(i)} = \bar{\phi}^*_{1(i)} - \bar{\phi}^*_{2(i)}$.
Accepter H_0 si $Q_{\bar{\phi}^*_d;\alpha/2} \leq \bar{\phi}_d \leq Q_{\bar{\phi}^*_d;(1-\alpha/2)}$.

Figure 4.1 Algorithme du test de permutation bilatéral pour deux échantillons indépendants.

4.2.3 Méthode du Bootstrap

La méthode du Bootstap est aussi une méthode non paramétrique semblable au test de permutation qui se base sur l'intervalle de confiance pour la différence des deux paramètres. Soit deux échantillons de n et m individus provenant de deux populations. La statistique du test est la même que pour le test de permutation soit $\bar{\phi}_d = \bar{\phi}_1 - \bar{\phi}_2$.

Cette méthode est cependant approximative car la règle de décision est basée sur les quantiles empiriques et la précision dépend du nombre de répétitions (paramètre B de l'algorithme). On suggère B = 1000 pour obtenir une précision suffisante (Efron et al., 1993). L'algorithme est présenté à la figure 4.2.

Répéter B fois

Tirer $\xi^*_{1(i)}$, un échantillon aléatoire avec remise de n individus de ξ_1.

Tirer $\xi^*_{2(i)}$, un échantillon aléatoire avec remise de m individus de ξ_2.

Évaluer l'apparentement moyen de $\xi^*_{1(i)}$ et $\xi^*_{2(i)}$.

Évaluer $\bar{\phi}^*_{d(i)} = \bar{\phi}^*_{1(i)} - \bar{\phi}^*_{2(i)}$.

Accepter H_0 si $Q_{\bar{\phi}^*_d;\alpha/2} \leq 0 \leq Q_{\bar{\phi}^*_d;(1-\alpha/2)}$.

Figure 4.2 Algorithme du test de Bootstrap bilatéral pour deux échantillons indépendants

4.3 Résultats

Les trois tests proposés ont été comparés par simulation afin de valider les niveaux réels, à savoir la probabilité de rejeter H_0 étant donné l'égalité des paramètres $\bar{\phi}_1$ et $\bar{\phi}_2$. Afin de simuler l'hypothèse nulle, deux échantillons de même taille (n = m) ont été prélevés sans remise parmi la population de référence de 2600 individus. Pour chaque paire d'échantillons ainsi prélevée, les tests statistiques ont été réalisés. Cela nous a permis d'estimer les niveaux réels sous l'hypothèse H_0. Cette procédure a été répétée 10000 fois afin d'obtenir une précision suffisante des niveaux réels.

La puissance des trois tests a aussi été comparée dans le but de vérifier s'il y avait une différence entre les tests. Pour ce faire, une seule contre hypothèse a été étudiée en divisant la population de référence en deux sous-populations de 1000 individus chacune. Les moyennes des deux populations sont respectivement $\mu_{\phi 1}$ = 0.0012437 et $\mu_{\phi 2}$ = 0.0005034. Pour un niveau donné et une taille d'échantillon fixé, le rejet ou

l'acceptation de l'hypothèse H_0 a été vérifié. Le nombre de répétitions de cette procédure a été fixé à 1000.

Il est important de rappeler que l'objectif visé, lors du calcul de la puissance, est d'obtenir un aperçu global de la relation existante entre les tests étudiés, et non l'avantage d'un test sur un autre. C'est la raison pour laquelle on ne considère qu'une seule contre hypothèse et que l'on effectue seulement 1000 répétitions. Une analyse exhaustive de la puissance aurait été trop longue à simuler dans le cadre de cette recherche.

Les niveaux réels et la puissance de chaque test ont été obtenus en utilisant des échantillons de tailles $n = 10, 20, 30, 40, 50, 75$ et 100 et des niveaux les plus couramment utilisés soit $\alpha = 1\%, 5\%, 10\%$ et 20%.

4.3.1 Niveaux

Les tableaux 4.1 et 4.2 présentent les niveaux empiriques (bilatéraux et unilatéraux) des trois tests de comparaison des moyennes. Les niveaux des tests bilatéraux et unilatéraux sont semblables pour les différents tests. On observe qu'effectivement le test de Student donne des résultats médiocres confirmant la présence de la dépendance entre les observations. Cependant les niveaux des tailles d'échantillons 10 et 20 obtenus par la méthode du Jackknife sont relativement précis.

Le test du Bootstrap et le test de permutation obtiennent les meilleurs niveaux, avec une légère avance pour le test de permutation. Dans la littérature, le test du Bootstrap

est réputé pour être peu fiable pour de petites tailles d'échantillons, dû à la variation échantillonnale. Dans notre cas, le fait qu'un échantillon de 10 individus génère une matrice d'apparentement de 45 valeurs avantage certainement l'obtention de niveaux précis.

4.3.2 Puissance

Les tableaux 4.3 et 4.4 présentent la puissance des différents tests de comparaison des moyennes (bilatéraux et unilatéraux). La puissance semble être du même ordre de grandeur pour tous les tests à l'exception du test de Student (avec la variance de l'échantillon) qui obtient une puissance médiocre traduisant encore une fois la dépendance des observations.

4.4 Conclusion

Deux tests non paramétriques ont été comparés au test classique de Student. Le test de Student, couplé à la méthode du Jackknife, offre des résultats tout aussi performants tant pour les niveaux que pour la puissance lorsque les échantillons sont de petites tailles. Par contre, le test de Student construit avec la variance de l'échantillon est à exclure en tout temps. Le test du Bootstrap et le test de permutation sont cependant nettement meilleurs. Puisque le test de permutation obtient les niveaux les plus précis et une puissance légèrement supérieure, ce test semble la meilleure alternative dans tous les cas.

Niveau $\alpha = 0.01$ bilatéral							
Tests	n = 10	n = 20	n = 30	n = 40	n = 50	n = 75	n = 100
Student	0.0927	0.1913	0.2131	0.2020	0.1530	0.0723	0.0495
Student -jackknife	0.0066	**0.0106**	0.0161	0.0224	0.0281	0.0403	0.0480
Permutation	**0.0126**	0.0126	**0.0120**	**0.0117**	**0.0115**	**0.0126**	**0.0105**
Bootstrap	0.0165	0.0151	0.0141	0.0140	0.0128	0.0127	0.0107
Niveau $\alpha = 0.05$ bilatéral							
Tests	n = 10	n = 20	n = 30	n = 40	n = 50	n = 75	n = 100
Student	0.2129	0.3139	0.3059	0.2736	0.2207	0.1533	0.1329
Student -jackknife	**0.0458**	0.0780	0.1008	0.1158	0.1206	0.1386	0.1471
Permutation	0.0543	**0.0508**	**0.0565**	**0.0520**	0.0462	**0.0515**	**0.0487**
Bootstrap	0.0624	0.0577	0.0611	0.0568	**0.0522**	0.0538	0.0479
Niveau $\alpha = 0.10$ bilatéral							
Tests	n = 10	n = 20	n = 30	n = 40	n = 50	n = 75	n = 100
Student	0.3088	0.3874	0.3665	0.3345	0.2868	0.2282	0.2028
Student -jackknife	0.1096	0.1675	0.1890	0.2061	0.2058	0.2224	0.2310
Permutation	**0.1043**	**0.1034**	**0.1028**	**0.0995**	**0.0973**	**0.1003**	**0.0962**
Bootstrap	0.1108	0.1094	0.1077	0.1041	0.1030	0.1025	**0.0962**
Niveau $\alpha = 0.20$ bilatéral							
Tests	n = 10	n = 20	n = 30	n = 40	n = 50	n = 75	n = 100
Student	0.4390	0.4896	0.4751	0.4611	0.4286	0.3497	0.3210
Student -jackknife	0.2616	0.3152	0.3241	0.3453	0.3376	0.3500	0.3592
Permutation	**0.2114**	**0.2083**	0.2068	0.2034	**0.2046**	**0.1968**	**0.1941**
Bootstrap	0.2115	0.2106	**0.2061**	**0.2029**	0.2063	0.1960	0.1921

Tableau 4.1 Niveaux empiriques des tests bilatéraux de comparaison des moyennes pour deux échantillons indépendant et 10000 simulations. Le gras souligne les meilleurs niveaux réels

Niveau $\alpha = 0.01$ unilatéral							
Tests	n = 10	n = 20	n = 30	n = 40	n = 50	n = 75	n = 100
Student	0.0639	0.1118	0.1203	0.1159	0.0851	0.0488	0.0337
Student -jackknife	0.0069	0.0129	0.0191	0.0241	0.0274	0.0382	0.0370
Permutation	**0.0121**	**0.0106**	**0.0121**	**0.0118**	**0.0105**	**0.0106**	**0.0092**
Bootstrap	0.0147	0.0128	0.0130	0.0131	0.0122	0.0117	0.0106
Niveau $\alpha = 0.05$ unilatéral							
Tests	n = 10	n = 20	n = 30	n = 40	n = 50	n = 75	n = 100
Student	0.1529	0.1879	0.1831	0.1686	0.1439	0.1152	0.0987
Student -jackknife	**0.0521**	0.0800	0.0925	0.1048	0.1044	0.1137	0.1150
Permutation	0.0533	**0.0517**	**0.0516**	**0.0503**	**0.0472**	**0.0508**	**0.0493**
Bootstrap	0.0549	0.0543	0.0531	0.0530	0.0530	0.0550	0.0471
Niveau $\alpha = 0.10$ unilatéral							
Tests	n = 10	n = 20	n = 30	n = 40	n = 50	n = 75	n = 100
Student	0.2168	0.2403	0.2399	0.2311	0.2148	0.1753	0.1584
Student -jackknife	0.1304	0.1522	0.1593	0.1761	0.1675	0.1762	0.1776
Permutation	**0.1052**	0.1058	0.1044	**0.1014**	0.1044	0.0994	0.0987
Bootstrap	0.1053	**0.1042**	**0.1028**	0.1010	**0.1023**	**0.0971**	**0.0961**
Niveau $\alpha = 0.20$ unilatéral							
Tests	n = 10	n = 20	n = 30	n = 40	n = 50	n = 75	n = 100
Student	0.3263	0.3531	0.3698	0.3448	0.3190	0.2683	0.2564
Student -jackknife	0.2630	0.2624	0.2727	0.2775	0.2740	0.2730	0.2752
Permutation	**0.2021**	0.2067	0.2036	**0.2006**	**0.1994**	0.1976	0.1997
Bootstrap	0.2053	**0.2019**	**0.2035**	**0.2006**	0.2028	**0.1940**	**0.1970**

Tableau 4.2 Niveaux empiriques des tests unilatéraux de comparaison des moyennes pour deux échantillons indépendants et 10000 simulations. Le gras souligne les meilleurs niveaux réels

Puissance pour un niveau $\alpha = 0.01$ bilatéral							
Tests	n = 10	n = 20	n = 30	n = 40	n = 50	n = 75	n = 100
Student	0.3630	0.5310	0.3850	0.2190	0.1950	0.2770	0.4420
Student -jackknife	0.1960	0.3810	0.6490	0.8250	0.8870	0.9800	0.9970
Permutation	0.2180	0.4570	0.6980	0.7980	0.8890	0.9840	0.9980
Bootstrap	0.1880	0.4070	0.5390	0.6500	0.7410	0.8670	0.9440
Puissance pour un niveau $\alpha = 0.05$ bilatéral							
Tests	n = 10	n = 20	n = 30	n = 40	n = 50	n = 75	n = 100
Student	0.5430	0.5940	0.4400	0.3610	0.3050	0.4560	0.6220
Student -jackknife	0.2980	0.6790	0.8570	0.9130	0.9490	0.9910	0.9980
Permutation	0.3920	0.6780	0.8480	0.9080	0.9600	0.9960	0.9990
Bootstrap	0.3560	0.6170	0.7420	0.8005	0.8700	0.9540	0.9890
Puissance pour un niveau $\alpha = 0.10$ bilatéral							
Tests	n = 10	n = 20	n = 30	n = 40	n = 50	n = 75	n = 100
Student	0.6110	0.6200	0.5200	0.4390	0.3780	0.5530	0.7250
Student -jackknife	0.4580	0.7870	0.9110	0.9470	0.9680	0.9960	0.9980
Permutation	0.5090	0.7750	0.8900	0.9400	0.9770	0.9970	0.9990
Bootstrap	0.4680	0.7240	0.8260	0.8820	0.9200	0.9760	0.9970
Puissance pour un niveau $\alpha = 0.20$ bilatéral							
Tests	n = 10	n = 20	n = 30	n = 40	n = 50	n = 75	n = 100
Student	0.6860	0.6980	0.6250	0.4940	0.5030	0.6640	0.8280
Student -jackknife	0.6420	0.8690	0.9490	0.9690	0.9870	0.9990	0.9980
Permutation	0.6500	0.8510	0.9340	0.9690	0.9880	0.9990	0.9990
Bootstrap	0.6100	0.8180	0.9010	0.9300	0.9670	0.9940	0.9990

Tableau 4.3 Puissance des tests bilatéraux de comparaison des moyennes pour deux échantillons indépendants et 1000 simulations.

Puissance pour un niveau $\alpha = 0.01$ unilatéral							
Tests	n = 10	n = 20	n = 30	n = 40	n = 50	n = 75	n = 100
Student	0.4380	0.5630	0.4020	0.2720	0.2520	0.3510	0.5180
Student -jackknife	0.1580	0.5160	0.7600	0.8570	0.9230	0.9880	0.9970
Permutation	0.2820	0.5450	0.7550	0.8370	0.9210	0.9890	0.9990
Bootstrap	0.2440	0.4590	0.6390	0.7310	0.8250	0.9030	0.9690
Puissance pour un niveau $\alpha = 0.05$ unilatéral							
Tests	n = 10	n = 20	n = 30	n = 40	n = 50	n = 75	n = 100
Student	0.6070	0.6180	0.5190	0.4380	0.3770	0.5520	0.7240
Student -jackknife	0.4560	0.7870	0.9110	0.9470	0.9680	0.9960	0.9980
Permutation	0.5060	0.7750	0.8900	0.9400	0.9770	0.9970	0.9990
Bootstrap	0.4740	0.7000	0.8300	0.8840	0.9270	0.9840	0.9990
Puissance pour un niveau $\alpha = 0.10$ unilatéral							
Tests	n = 10	n = 20	n = 30	n = 40	n = 50	n = 75	n = 100
Student	0.6780	0.6910	0.6130	0.4880	0.4960	0.6620	0.8260
Student -jackknife	0.6380	0.8670	0.9490	0.9690	0.9870	0.9990	0.9980
Permutation	0.6440	0.8510	0.9340	0.9690	0.9870	0.9990	0.9990
Bootstrap	0.6080	0.8120	0.9000	0.9360	0.9730	0.9960	0.9990
Puissance pour un niveau $\alpha = 0.20$ unilatéral							
Tests	n = 10	n = 20	n = 30	n = 40	n = 50	n = 75	n = 100
Student	0.8000	0.8060	0.6530	0.5800	0.6290	0.7880	0.9001
Student -jackknife	0.7730	0.9280	0.9660	0.9810	0.9940	1.0000	0.9990
Permutation	0.7890	0.9200	0.9640	0.9780	0.9950	0.9990	0.9990
Bootstrap	0.7690	0.9000	0.9540	0.9730	0.9920	0.9990	0.9990

Tableau 4.4 Puissance des tests unilatéraux de comparaison des moyennes pour deux échantillons indépendants et 1000 simulations.

CHAPITRE 5

ÉCHANTILLONS APPARIÉS

5.1 Introduction

Les différences de coefficient d'apparentement entre deux groupes sont habituellement d'origines multiples. Il existe un partage de gènes mais aussi plusieurs facteurs à considérer tels que l'âge des proposants, le sexe, la localisation géographique, etc. Ces facteurs peuvent être contrôlés par l'utilisation de témoins appariés aux cas selon une variable quelconque. On propose dans ce chapitre de valider quelques tests de comparaison des moyennes sur des échantillons appariés.

On considère deux échantillons tirés de deux populations telles que chaque observation du premier échantillon est appariée à une ou plusieurs observations du deuxième échantillon selon une variable de contrôle. On s'intéresse au coefficient d'apparentement moyen du premier groupe en fonction du deuxième groupe. Le premier groupe est appelé le « groupe des cas » tandis que le deuxième groupe est appelé le « groupe témoin ». On suppose que le nombre de témoins par cas est toujours identique, soit k.

On considère μ_C et μ_T, les coefficients moyens d'apparentement dans la population des cas et des témoins respectivement et ξ_C et ξ_T les échantillons, en supposant l'existence d'une variable intermédiaire permettant d'apparier les témoins aux cas. On veut confronter les hypothèses $H_0 : \mu_C = \mu_T$, les moyennes sont égales contre $H_1 : \mu_C \neq \mu_T$. Dans cette étude les moyennes sont les coefficients moyens d'apparentement. Le test est souvent utilisé pour déterminer si la population des cas a

un apparentement moyen plus élevé que la population des contrôles. Cela explique pourquoi les hypothèses retenues pour un test unilatéral sont $H_0 : \mu_C = \mu_T$ contre $H_1 : \mu_C > \mu_T$.

Un test de comparaison des coefficients d'apparentement dans un contexte cas / témoins a été proposé par Hauk et Martin (1984). Ce dernier est basé sur la normalité de la différence de coefficients d'apparentement et l'estimation de la variance par la technique du Jackknife. Dans ce chapitre, ce test sera comparé au test de permutation en ce qui a trait au niveau.

5.2 Statistiques du test

Considérons deux échantillons appariés tirés de deux populations. Le premier échantillon $\xi_C = X_1, X_2, \ldots, X_n$ tiré de la première population de moyenne μ_C, est composé de n cas. Le deuxième échantillon, tiré de la seconde population de moyenne μ_T et composé de kn témoins est donné par

$$\xi_T = Y_{11}, Y_{12}, \ldots, Y_{1j}, \ldots, Y_{1k}, \ldots, Y_{i1}, Y_{i2}, \ldots, Y_{ij}, \ldots, Y_{ik}, \ldots, Y_{n1}, Y_{n2}, \ldots, Y_{nj}, \ldots, Y_{nk},$$

où Y_{ij} est le j-ième témoin apparié au cas i. Pour comparer ces deux séries appariées, Hauck et Martin (1984) proposent la statistique suivante :

$$\overline{D}_{inter} = \frac{2}{n(n-1)} \sum_{i<j} \sum \left[\phi(X_i, X_j) - \frac{1}{2k} \sum_{r=1}^{k} \left(\phi(X_i, Y_{jr}) + \phi(X_j, Y_{ir}) \right) \right]$$

où $\phi(x, y)$ est le coefficient d'apparentement entre les individus x et y. La statistique $\overline{D}$ dite «inter groupe» mesure l'apparentement résiduel entre les cas et les contrôles en soustrayant l'apparentement spécifique à la variable de contrôle. Pour une

paire de cas, l'apparentement dû à la variable de contrôle est donné par la moyenne des 2k coefficients d'apparentement entre le premier cas et les contrôles du deuxième cas.

Une autre technique, pour éliminer l'effet de la variable de contrôle pour une paire de cas, consiste à évaluer l'apparentement moyen entre chaque témoin du premier cas et chaque témoin du deuxième cas. On obtient alors la statistique « intra groupe » et sa formule est donnée par

$$\overline{D}_{intra} = \frac{2}{n(n-1)} \sum_{i<j} \sum \left[\phi(X_i, X_j) - \frac{1}{k^2} \sum_{b=1}^{k} \sum_{l=1}^{k} \phi(Y_{ib}, Y_{jl}) \right].$$

D'un point de vue pratique, l'évaluation de la statistique $\overline{D}_{inter}$ nécessite les n(n-1) coefficients de la matrice Φ des cas et les nk coefficients de la matrice inter groupe. De plus, si on souhaite visualiser la différence d'apparentement entre les cas et les témoins, il est nécessaire de calculer les coefficients de la matrice intra témoin. Cela veut dire que la matrice d'apparentement pour l'ensemble des individus de l'étude doit être évaluée. Lorsque le nombre de cas et de témoins est grand, le calcul devient laborieux. La statistique $\overline{D}_{intra}$ quant à elle ne requiert que la matrice d'apparentement intra cas et intra témoin.

D'un point de vue théorique, la différence majeure entre les deux statistiques est la suivante : la statistique « intra groupe » permet de mesurer la différence d'apparentement entre deux populations, même si elles n'ont aucune souche commune, tandis que la statistique « inter groupe » ne mesure cette différence que lorsque les populations ont les mêmes ancêtres. Ainsi, deux populations n'ayant aucun ancêtre en

commun auront toujours une statistique « inter groupe » égale au coefficient moyen d'apparentement du groupe de cas même si les structures généalogiques sont semblables.

5.3 Tests de comparaison de deux échantillons appariés

5.3.1 Méthode du Jackknife

La méthode du Jackknife est basée sur la normalité de la statistique $\overline{D}$ (intra ou inter) et elle permet d'évaluer la variance de cette statistique. Le test d'hypothèses est alors basé sur :

$$D = \frac{\overline{D}}{\hat{S}_j}$$

où $\hat{S}_j^{\,2}$ est la variance de la distribution de la statistique $\overline{D}$ obtenue par la méthode du jackknife. Cette variance est donnée par :

$$\hat{S}_J = \sqrt{\frac{n-1}{n}\sum_{i=1}^{n}\left(\overline{D}_{(i)} - \overline{D}_J\right)^2},$$

où $\overline{D}_{(i)}$ est la moyenne de la différence calculée sur l'échantillon ξ_C et ξ_T en omettant le i-ième cas et ses contrôles. Le paramètre $\overline{D}_J$ est la moyenne des n valeurs de $\overline{D}_{(i)}$.

La règle de décision du test bilatéral basé sur la méthode du Jackknife est de rejeter l'hypothèse nulle H_0 si $|Z| > z_{\alpha/2}$ où $z_{\alpha/2}$ est le point critique de niveau $\alpha/2$

d'une loi normale centrée réduite. Dans le cas du test unilatéral, la région de rejet est donnée par $Z > z_\alpha$.

5.3.2 Test de permutation

Soit ξ_C un échantillon de n cas et ξ_T un échantillon apparié de k témoins pour chaque cas. Le test de permutation permet d'obtenir une approximation de la distribution de la statistique $\overline{D}$ d'un test sous l'hypothèse H_0 en permutant les échantillons un certain nombre de fois. Dans le cas d'un échantillon apparié, chaque bloc, composé d'un cas et de ses contrôles, est permuté aléatoirement. La statistique $\overline{D}^*$ est ensuite évaluée. Posons $\overline{D}_1^*, \overline{D}_1^*, \ldots, \overline{D}_B^*$, les valeurs obtenues après B répétitions de cette procédure. La règle de décision du test de permutation bilatéral est d'accepter H_0 si la valeur de la statistique $\overline{D}$ est telle que $Q_{\alpha/2} \leq \overline{D} \leq Q_{(1-\alpha/2)}$, où $Q_{\alpha/2}$ est le quantile d'ordre $\alpha/2$ de la distribution de la statistique $\overline{D}^*$. Dans le cas d'un test unilatéral, la région d'acceptation est donnée par $\overline{D} \leq Q_{(1-\alpha)}$.

Calcul de la matrice Φ pour $\xi_c \cup \xi_t$.
Évaluer $\overline{D}$
Pour i de 1 à B
    Permuter les k+1 individus (le cas et ses contrôles)
    Évaluer $\overline{D}_i^*$
Accepter H_0 si $|\overline{D}| \leq Q_{\overline{D}^*}$.

Figure 5.1 Algorithme du test de permutation d'un échantillon apparié.

5.4 Simulations

Les simulations pour évaluer le niveau réel des tests sont basées sur la population de référence comportant 2600 individus issus de 26 régions ou sous-régions distinctes du Québec (100 individus par région). Les échantillons de n cas ont été constitués en prélevant n individus sans remise parmi les 2600 individus. Pour chaque cas, les k témoins ont été appariés selon la région d'origine du cas. Seules les tailles d'échantillons n = 10, 20, 50 et 75 cas ont été considérées puisque les simulations n'affichaient aucune différence entre les résultats pour les tailles n = 20 et 50. Le nombre du contrôle a été fixé à k = 1, 2 et 4 et les niveaux utilisés ont été α = 1%, 5%, 10% et 20% soit les plus courants.

Pour chaque taille d'échantillon, de contrôle et chaque niveau fixé, 5000 échantillons de cas et de contrôles ont été tirés de la population de référence et les quatre tests ont été calculés afin de valider les niveaux réels. Le nombre d'échantillons simulés a été fixé à 5000 dû à la complexité de calcul associée à la méthode du test de permutation.

5.5 Résultats

La méthode du Jackknife et la méthode du test de permutation ont été comparées en considérant la statistique $\overline{D}$ (intra ou inter). Les tableaux 5.1 et 5.2 présentent les niveaux empiriques (bilatéraux et unilatéraux) des méthodes pour la statistique inter groupe tandis que les tableaux 5.3 et 5.4 présentent les niveaux empiriques pour la statistique intra groupe.

Les niveaux (bilatéraux et unilatéraux) de la méthode du test de permutation pour la statistique $\overline{D}$ sont très près des niveaux espérés. La précision augmente en fonction de la taille des échantillons. Quant au nombre de témoins, il ne semble pas affecter de manière significative la précision des niveaux réels en considérant la variation échantillonnale. Dans le cas de la méthode du Jackknife, les niveaux sont médiocres. La précision des niveaux se détériore lorsque la taille de l'échantillon augmente et que le nombre de témoins augmente.

Test bilatéral	**Jackknife cas vs contrôles (inter)**				**Permutation cas vs contrôles (inter)**			
Niveau 1%	n = 10	n = 20	n = 50	n = 75	n = 10	n = 20	n = 50	n = 75
k=1	0.0030	0.0016	0.0018	0.0060	**0.0112**	**0.0108**	**0.0078**	**0.0092**
k=2	0.0041	0.0030	0.0064	0.0086	**0.0110**	**0.0100**	**0.0104**	**0.0090**
k=4	0.0056	0.0054	0.0074	0.0126	**0.0100**	**0.0114**	**0.0090**	**0.0088**
Niveau 5%	n = 10	n = 20	n = 50	n = 75	n = 10	n = 20	n = 50	n = 75
k=1	0.0132	0.0098	0.0244	0.0326	**0.0548**	**0.0544**	**0.0488**	**0.0482**
k=2	0.0178	0.0232	0.0350	0.0400	**0.0546**	**0.0472**	**0.0492**	**0.0466**
k=4	0.0248	0.0322	0.0444	0.0470	**0.0450**	**0.0496**	**0.0508**	**0.0496**
Niveau 10%	n = 10	n = 20	n = 50	n = 75	n = 10	n = 20	n = 50	n = 75
k=1	0.0338	0.0334	0.0620	0.0716	**0.1032**	**0.1044**	**0.1026**	**0.0998**
k=2	0.0423	0.0570	0.0708	0.0828	**0.1082**	**0.0900**	**0.0952**	**0.0878**
k=4	0.0533	0.0732	0.0884	0.0872	**0.0986**	**0.0974**	**0.0994**	**0.1034**
Niveau 20%	n = 10	n = 20	n = 50	n = 75	n = 10	n = 20	n = 50	n = 75
k=1	0.0966	0.1100	0.1440	0.1638	**0.2042**	**0.2028**	**0.1996**	**0.1990**
k=2	0.1110	0.1464	0.1606	0.1732	**0.2048**	**0.1916**	**0.2028**	**0.1958**
k=4	0.1255	0.1578	0.1784	0.1677	**0.2032**	**0.1954**	**0.1942**	**0.2044**

Tableau 5.1 Niveaux empiriques bilatéraux de la statistique inter groupe pour 5000 échantillons de taille n = 10, 20, 50 et 75 et k = 1, 2 et 4 témoins

Test unilatéral	**Jackknife cas vs contrôles (inter)**				**Permutation cas vs contrôles (inter)**			
Niveau 1%	n = 10	n = 20	n = 50	n = 75	n = 10	n = 20	n = 50	n = 75
k=1	0.0004	0.0000	0.0004	0.0006	**0.0096**	**0.0092**	**0.0080**	**0.0120**
k=2	0.0008	0.0002	0.0002	0.0000	**0.0108**	**0.0098**	**0.0104**	**0.0104**
k=4	0.0004	0.0000	0.0000	0.0000	**0.0078**	**0.0100**	**0.0078**	**0.0096**
Niveau 5%	n = 10	n = 20	n = 50	n = 75	n = 10	n = 20	n = 50	n = 75
k=1	0.0054	0.0048	0.0122	0.0168	**0.0526**	**0.0450**	**0.0462**	**0.0500**
k=2	0.0046	0.0052	0.0054	0.0100	**0.0488**	**0.0498**	**0.0490**	**0.0498**
k=4	0.0034	0.0016	0.0062	0.0060	**0.0480**	**0.0504**	**0.0512**	**0.0482**
Niveau 10%	n = 10	n = 20	n = 50	n = 75	n = 10	n = 20	n = 50	n = 75
k=1	0.0266	0.0304	0.0430	0.0514	**0.1002**	**0.0924**	**0.0908**	**0.1016**
k=2	0.0134	0.0236	0.0270	0.0382	**0.1030**	**0.1002**	**0.1018**	**0.1028**
k=4	0.0136	0.0150	0.0240	0.0290	**0.0978**	**0.1004**	**0.1044**	**0.0978**
Niveau 20%	n = 10	n = 20	n = 50	n = 75	n = 10	n = 20	n = 50	n = 75
k=1	0.1130	0.1214	0.1350	0.1454	**0.1976**	**0.1874**	**0.1920**	**0.1942**
k=2	0.0794	0.1068	0.1170	0.1286	**0.2026**	**0.2066**	**0.1984**	**0.2058**
k=4	0.0662	0.0824	0.1122	0.1172	**0.1934**	**0.1964**	**0.1980**	**0.1924**

Tableau 5.2 Niveaux empiriques unilatéraux de la statistique inter groupe pour 5000 échantillons de taille n = 10, 20, 50 et 75 et k = 1, 2 et 4 témoins

Test bilatéral	**Jackknife contrôles vs contrôles (intra)**				**Permutation contrôles vs contrôles (intra)**			
Niveau 1%	n = 10	n = 20	n = 50	n = 75	n = 10	n = 20	n = 50	n = 75
k=1	0.0020	0.0000	0.0010	0.0012	**0.0090**	**0.0146**	**0.0088**	**0.0090**
k=2	0.0056	0.0010	0.0006	0.0034	**0.0102**	**0.0072**	**0.0092**	**0.0080**
k=4	0.0068	0.0016	0.0064	0.0078	**0.0108**	**0.0098**	**0.0114**	**0.0086**
Niveau 5%	n = 10	n = 20	n = 50	n = 75	n = 10	n = 20	n = 50	n = 75
k=1	0.0120	0.0078	0.0108	0.0078	**0.0464**	**0.0558**	**0.0458**	**0.0458**
k=2	0.0128	0.0096	0.0190	0.0232	**0.0536**	**0.0448**	**0.0436**	**0.0498**
k=4	0.0224	0.0160	0.0328	0.0388	**0.0540**	**0.0492**	**0.0476**	**0.0468**
Niveau 10%	n = 10	n = 20	n = 50	n = 75	n = 10	n = 20	n = 50	n = 75
k=1	0.0288	0.0302	0.0366	0.0320	**0.0978**	**0.1084**	**0.1016**	**0.0950**
k=2	0.0330	0.0312	0.0514	0.0832	**0.1064**	**0.0966**	**0.0952**	**0.0960**
k=4	0.0458	0.0560	0.0732	0.0802	**0.1082**	**0.1008**	**0.0922**	**0.0980**
Niveau 20%	n = 10	n = 20	n = 50	n = 75	n = 10	n = 20	n = 50	n = 75
k=1	0.0918	0.1112	0.1110	0.1376	**0.1962**	**0.2092**	**0.2020**	**0.1878**
k=2	0.0928	0.1032	0.1448	0.1680	**0.2156**	**0.1920**	**0.1990**	**0.1944**
k=4	0.1138	0.1716	0.1620	0.1622	**0.2066**	**0.1980**	**0.1940**	**0.2034**

Tableau 5.3 Niveaux empiriques bilatéraux de la statistique intra groupe pour 5000 échantillons de taille n = 10, 20, 50 et 75 et k = 1, 2 et 4 témoins

Test unilatéral	**Jackknife contrôles vs contrôles (intra)**				**Permutation contrôles vs contrôles (intra)**			
Niveau 1%	n = 10	n = 20	n = 50	n = 75	n = 10	n = 20	n = 50	n = 75
k=1	0.0020	0.0002	0.0006	0.0008	**0.0144**	**0.0056**	**0.0110**	**0.0108**
k=2	0.0000	0.0002	0.0004	0.0004	**0.0116**	**0.0088**	**0.0108**	**0.0114**
k=4	0.0002	0.0002	0.0002	0.0002	**0.0112**	**0.0078**	**0.0106**	**0.0108**
Niveau 5%	n = 10	n = 20	n = 50	n = 75	n = 10	n = 20	n = 50	n = 75
k=1	0.0136	0.0148	0.0196	0.0226	**0.0534**	**0.0434**	**0.0464**	**0.0526**
k=2	0.0058	0.0068	0.0074	0.0090	**0.0522**	**0.0488**	**0.0512**	**0.0516**
k=4	0.0044	0.0014	0.0072	0.0058	**0.0486**	**0.0520**	**0.0494**	**0.0490**
Niveau 10%	n = 10	n = 20	n = 50	n = 75	n = 10	n = 20	n = 50	n = 75
k=1	0.0446	0.0570	0.0518	0.0518	**0.1020**	**0.0882**	**0.0964**	**0.0998**
k=2	0.0218	0.0290	0.0178	0.0312	**0.1036**	**0.1008**	**0.1034**	**0.1098**
k=4	0.0166	0.0074	0.0240	0.0280	**0.0966**	**0.1004**	**0.095**	**0.0972**
Niveau 20%	n = 10	n = 20	n = 50	n = 75	n = 10	n = 20	n = 50	n = 75
k=1	0.1538	0.1510	0.1640	0.1712	**0.2002**	**0.1868**	**0.1966**	**0.2020**
k=2	0.1042	0.1064	0.1286	0.1454	**0.2004**	**0.1976**	**0.2008**	**0.2080**
k=4	0.0806	0.0452	0.0800	0.1144	**0.1948**	**0.1944**	**0.2048**	**0.1870**

Tableau 5.4 Niveaux empiriques unilatéraux de la statistique intra groupe pour 5000 échantillons de taille n = 10, 20, 50 et 75 et k = 1, 2 et 4 témoins.

5.6 Conclusion

Deux méthodes ont été proposées afin de comparer les coefficients moyens d'apparentement de deux populations à partir d'échantillons appariés. Pour chacune de ces méthodes, deux statistiques ont été considérées : la statistique intra groupe $\overline{D}_{intra}$ et la statistique inter groupe $\overline{D}_{inter}$. En considérant les résultats obtenus lors des simulations, les niveaux empiriques de la méthode du test de permutation ont démontré qu'il est nettement plus intéressant d'utiliser cette méthode pour les deux statistiques même si le coût en temps est plus grand que la méthode du Jackknife.

Il aurait été intéressant de comparer la puissance des tests en considérant les deux statistiques, afin d'observer si une différence évidente existe entre ceux-ci. Cependant, cette démarche était trop complexe dans le cadre de cette étude.

CONCLUSION

L'étude de la distribution du coefficient moyen d'apparentement a démontré qu'on ne peut utiliser la loi normale pour cette statistique lorsque la population est du même type que celle visée dans cette étude et que les tailles d'échantillons sont relativement petites. Une transformation logarithmique a permis d'obtenir une distribution plus proche d'une loi normale mais encore trop différente pour considérer la normalité asymptotique. Bien qu'une étude en profondeur pourrait être réalisée afin de déterminer la distribution théorique de cette statistique, la construction d'un estimateur par intervalle et d'un test d'hypothèse demeure problématique. La solution réside dans l'utilisation des techniques de rééchantillonnage.

La construction d'un intervalle de confiance pour le coefficient d'apparentement moyen selon la méthode classique en utilisant une approximation normale n'est pas fiable autant du point de vue théorique que pratique. La meilleure méthode s'avère celle du Bootstrap Bca. Il faut cependant noter que même avec cette méthode, les intervalles sont fortement asymétriques et pour de petits échantillons, les intervalles ont un niveau de confiance beaucoup trop petit par rapport à celui recherché. L'interprétation de ces intervalles doit tenir compte de ces propriétés.

Pour ce qui est des tests d'hypothèses, les résultats sont excellents. En considérant des échantillons indépendants et des échantillons appariés, la méthode conseillée est celle du test de permutation qui est fiable et robuste.

Les techniques de rééchantillonnage sont à l'origine de ce succès. Dans le cas des tests d'hypothèses, elles ont été efficaces tandis que dans le cas des estimateurs par intervalle, l'exploration de ces techniques sera certainement une voie intéressante lors de recherches ultérieures.

BIBLIOGRAPHIE

Bouchard, G. (2003). *Rapport annuel du Projet BALSAC 2002-2003*. Septembre, (p.52).

Bouchard, G. et De Braekeleer, M. (1990) *Histoire d'un génome*, Presse de l'Université, (p.634).

Data Analysis Division MathSoft Inc. (2000a), *S-PLUS 6.0 for Unix Programmer's Guide*, Seattle: Data Analysis Division MathSoft Inc, (p.534).

Data Analysis Division MathSoft Inc. (2000b), *S-PLUS 6.0 for Unix Guide to Statistics,* Vol.1, Seattle: Data Analysis Division MathSoft Inc, (p.704).

Data Analysis Division MathSoft Inc. (2000c), *S-PLUS 6.0 for Unix Guide to Statistics,* Vol.2, Seattle: Data Analysis Division MathSoft Inc, (p.618).

Efron, B. (1979), "Bootstrap methods: another look at the jackknife", Ann. Statist., vol. 7, p.1-26.

Efron, B., Tibshirani, R.J. (1993), *An Introduction to the Bootstrap*. New York: Chapman and Hall, p. 436

Falconer, D.S. (1972), *Introduction à la Génétique Quantitative*, Paris: Masson & Cie, (p.284), chap.3, p.36.

Hauk, W.W., Martin, A.O. (1984), " A statistical test for detection of ancestral genetic contributions to disease occurrence in finite populations", Genetic Epidemiology, vol. 1, p.383-400.

Hoeffding, W. (1948), "A class of statistics with asymptotically normal distribution", Ann. Math. Statistics, vol. 19, p.293-325.

Hogg, R.V., Craig, A.T. (1970), *Introduction To Mathematical Statistics*, 3ème Édition, New York : Mc Millan, (p.415).

Jacquard, J. (1970), *Structures Génétiques des Populations*, Paris: Masson & Cie, (p.399), chap.6, p.107.

Lange, K., (2002), *Mathematical and Statistical Methods for Genetic Analysis*, New York: Springer, (p.265), chap.5, p.70.

Malécot, G. (1948). *Les mathématiques de l'hérédité*, Paris: Masson & Cie.

Marsaglia, G. et al. (1973), "Random Number Package: « Super-Duper »". School of Computer Science, McGill University.

Palm, R. (2002), "Utilisation du bootstrap pour les problèmes statistiques liés à l'estimation des paramètres", Biotechnol. Agron. Soc. Environ., vol. 6, n.3, p.143-153.

Pence, R.A. (1994), "Still more heresy by the numbers", NGC Newsletter, vol. 20, n. 1-2.

Scriver, C.R. (2001), "Human genetics: lessons from Quebec populations", Annu. Rev. Genomics Hum. Genet, (p.2:69-101)

Serfling, R.J. (1980), *Approximation Thorems of Mathematical Statistics*, États-Unis: John Wiley & Sons, (p.371), chap.5, p.171.

Thompson, E.A. (1974), *Annals of Human Genetics*, chap. 2, p.299-305.

Tukey, J. 1958, "Bias and confidence in not quite large samples". Ann. Math. Stat. vol 29, p.614.

Tremblay, M., Jomphe, M. et Vézina H. (2001), "Comparaison des structures patronymiques et génétiques dans la population québécoise". Brunet G., Darlu P. et Zei G (dir.), Paris, (p.367-389).

Weir, B.S. (1990). *Genetic Data Analysis*, Sunderland, MA: Sinauer Associates, (p.377), chap.5, p.135.

Wright, S. (1922), "Coefficients of inbreeding and relationship". Am. Naturalist vol 56, p.330-338.