Nº 12-001-XIF au catalogue

Techniques d'enquête

Juin 2006

Statistique Canada
Statistics Canada
Canada

Comment obtenir d'autres renseignements

Toute demande de renseignements au sujet du présent produit ou au sujet de statistiques ou de services connexes doit être adressée à : Division des méthodes d'enquêtes auprès des entreprises, Statistique Canada, Ottawa, Ontario, K1A 0T6 (téléphone : 1 800 263-1136).

Pour obtenir des renseignements sur l'ensemble des données de Statistique Canada qui sont disponibles, veuillez composer l'un des numéros sans frais suivants. Vous pouvez également communiquer avec nous par courriel ou visiter notre site Web.

Service national de renseignements	1 800 263-1136
Service national d'appareils de télécommunications pour les malentendants	1 800 363-7629
Renseignements concernant le Programme des services de dépôt	1 800 700-1033
Télécopieur pour le Programme des services de dépôt	1 800 889-9734
Renseignements par courriel	infostats@statcan.ca
Site Web	www.statcan.ca

Renseignements pour accéder au produit

Le produit nº 12-001-XIF au catalogue est disponible gratuitement. Pour obtenir un exemplaire, il suffit de visiter notre site Web à www.statcan.ca et de choisir la rubrique Nos produits et services.

Normes de service à la clientèle

Statistique Canada s'engage à fournir à ses clients des services rapides, fiables et courtois, et ce, dans la langue officielle de leur choix. À cet égard, notre organisme s'est doté de normes de service à la clientèle qui doivent être observées par les employés lorsqu'ils offrent des services à la clientèle. Pour obtenir une copie de ces normes de service, veuillez communiquer avec Statistique Canada au numéro sans frais 1 800 263-1136. Les normes de service sont aussi publiées dans le site www.statcan.ca sous À propos de Statistique Canada > Offrir des services aux Canadiens.

Statistique Canada

Division des méthodes d'enquêtes auprès des entreprises

Techniques d'enquête

Juin 2006

Publication autorisée par le ministre responsable de Statistique Canada

© Ministre de l'Industrie, 2006

Tous droits réservés. Le contenu de la présente publication électronique peut être reproduit en tout ou en partie, et par quelque moyen que ce soit, sans autre permission de Statistique Canada, sous réserve que la reproduction soit effectuée uniquement à des fins d'étude privée, de recherche, de critique, de compte rendu ou en vue d'en préparer un résumé destiné aux journaux et/ou à des fins non commerciales. Statistique Canada doit être cité comme suit : Source (ou « Adapté de », s'il y a lieu) : Statistique Canada, année de publication, nom du produit, numéro au catalogue, volume et numéro, période de référence et page(s). Autrement, il est interdit de reproduire le contenu de la présente publication, ou de l'emmagasiner dans un système d'extraction, ou de le transmettre sous quelque forme ou par quelque moyen que ce soit, reproduction électronique, mécanique, photographique, pour quelque fin que ce soit, sans l'autorisation écrite préalable des Services d'octroi de licences, Division des services à la clientèle, Statistique Canada, Ottawa, Ontario, Canada K1A 0T6.

Juillet 2006

Nº 12-001-XIF au catalogue
ISSN 1712-5685

Périodicité : semestriel

Ottawa

This publication is available in English upon request (catalogue no. 12-001-XIE)

Note de reconnaissance

Le succès du système statistique du Canada repose sur un partenariat bien établi entre Statistique Canada et la population, les entreprises, les administrations canadiennes et les autres organismes. Sans cette collaboration et cette bonne volonté, il serait impossible de produire des statistiques précises et actuelles.

Vol. 32, N° 1, pp. 11-26
Statistique Canada, N° 12-001 au catalogue

Plans de sondage à marche aléatoire ciblée

Steven K. Thompson [1]

Résumé

Les populations humaines cachées, Internet et d'autres structures en réseau conceptualisées mathématiquement sous forme de graphes sont intrinsèquement difficiles à échantillonner par les moyens conventionnels et les plans d'étude les plus efficaces comportent habituellement des procédures de sélection de l'échantillon par suivi adaptatif des liens reliant un nœud à un autre. Les données d'échantillon obtenues dans le cadre de telles études ne sont généralement pas représentatives au pied de la lettre de la population d'intérêt dans son ensemble. Cependant, un certain nombre de méthodes fondées sur le plan de sondage ou sur un modèle sont maintenant disponibles pour faire des inférences efficaces à partir d'échantillons de ce type. Les méthodes fondées sur le plan de sondage ont l'avantage de ne pas s'appuyer sur un modèle de population hypothétique, mais dépendent, en ce qui concerne leur validité, de la mise en œuvre du plan de sondage dans des conditions contrôlées et connues, ce qui est parfois difficile, voire impossible, en pratique. Les méthodes fondées sur un modèle offrent plus de souplesse quant au plan de sondage, mais requièrent que la population soit modélisée au moyen de modèles de graphes stochastiques et que le plan de sondage soit ignorable ou de forme connue, afin qu'il puisse être inclus dans les équations de vraisemblance ou d'inférence bayésienne. Aussi bien pour les méthodes basées sur le plan de sondage que celles fondées sur un modèle, le point faible est souvent le manque de contrôle concernant l'obtention de l'échantillon initial, à partir duquel débute le dépistage des liens. Les plans de sondage décrits dans le présent article offrent une troisième méthode, dans laquelle les probabilités de sélection de l'échantillon deviennent pas-à-pas moins dépendantes de la sélection de l'échantillon initial. Un modèle de « marche aléatoire » markovienne idéalise au moyen d'un graphe, les tendances d'un plan d'échantillonnage naturel d'une séquence de sélections par dépistage de liens à suivre. Le présent article présente des plans de sondage à marche uniforme ou ciblée dans lesquels la marche aléatoire est ajustée à chaque pas afin de produire un plan de sondage ayant les probabilités stationnaires souhaitées. On obtient ainsi un échantillon qui, à d'importants égards, est représentatif au pied de la lettre de la population d'intérêt dans son ensemble, ou qui ne nécessite que de simples facteurs de pondération pour qu'il en soit ainsi.

Mots clés : Échantillonnage adaptatif; échantillonnage déterminé selon les répondants (Respondent-driven sampling); échantillonnage d'une population cachée; échantillonnage en réseau; échantillonnage par graphes; marche aléatoire; méthode de Monte Carlo par chaîne de Markov; plans d'échantillonnage par dépistage de liens.

1. Introduction

Les populations comportant des liens ou une structure en réseau sont conceptualisées sous forme de graphes dans lesquels les nœuds (ou sommets) représentent les unités de la population et les arêtes ou les arcs, les relations ou liens entre ces unités. L'un des grands problèmes des études par établissement de graphes est qu'il est difficile, voire impossible, pour de nombreuses populations d'intérêt, d'obtenir des échantillons au moyen des plans de sondage conventionnels et que les échantillons sélectionnés peuvent être, tels qu'ils sont obtenus, fortement non représentatifs de la population d'intérêt dans son ensemble. En pratique, les seules méthodes d'échantillonnage applicables consistent souvent à suivre les liens à partir des nœuds sélectionnés, afin d'y ajouter des nœuds et des liens supplémentaires. Par exemple, lors de l'étude de populations humaines cachées, telles que les utilisateurs de drogues injectables, les travailleurs du sexe et d'autres populations courant le risque de contracter ou de transmettre le VIH/Sida ou l'hépatite C, les liens sociaux sont suivis en partant des répondants identifiés au départ, afin d'accroître l'échantillon de participants à l'étude. De même, dans les études des caractéristiques d'Internet, la procédure habituelle consiste à obtenir un échantillon de sites Web en suivant les liens allant des sites initiaux vers d'autres sites.

Klovdahl (1989) a utilisé l'expression « marche aléatoire » pour décrire une procédure conçue afin d'obtenir un échantillon à partir d'une population cachée en demandant à un répondant d'identifier plusieurs contacts, dont un est sélectionné au hasard pour être le répondant suivant, et en répétant le scénario pendant un certain nombre de pas. Heckathorn (1997) a décrit des méthodes d'échantillonnage déterminé selon les répondants « respondent-driven sampling » en appliquant des procédures de ce genre. En pratique, la raison qui motive l'utilisation de plans de sondage de ce type est de pénétrer plus en profondeur dans la population cachée afin d'obtenir des répondants plus « représentatifs » de la population que ne le sont peut-être les personnes plus visibles sélectionnées initialement. Dans les études d'Internet, l'idée parallèle est que l'« internaute aléatoire », qui choisit une page Web au hasard, clique ensuite au hasard sur l'un des liens figurant sur cette page, passant ainsi à une autre page, et ainsi de suite (Brin et Page 1998).

1. Steven K. Thompson, Département de statistique et de science actuarielle, Université Simon Fraser, Burnaby (Colombie-Britannique), Canada, V5A 1S6. Courriel : Thompson@stat.sfu.ca.

Le plan de sondage à marche aléatoire peut être conceptualisé comme une chaîne de Markov (Heckathorn 1997, 2002; Henzinger et coll. 2000; Salganik et Heckathorn 2004). Dans le présent article, nous décrivons certaines modifications apportées à ces plans de sondage à chaîne de Markov, dans le but d'obtenir des probabilités stationnaires de valeur égale ou spécifiée afin d'obtenir des estimations simples des caractéristiques du graphe de la population d'intérêt.

Les approches de l'inférence à partir d'échantillons provenant d'un graphe comprennent les méthodes fondées sur le plan de sondage, les méthodes fondées sur un modèle et les méthodes mixtes fondées sur une combinaison des deux. Dans l'approche fondée sur le plan de sondage, toutes les valeurs des variables de nœud et de lien du graphe sont considérées comme étant fixes ou données, et l'inférence est basée sur les probabilités induites par le plan de sondage intervenant dans la sélection de l'échantillon. Dans l'approche fondée sur un modèle, la population proprement dite est considérée comme une réalisation d'un modèle de graphe stochastique, qui fournit la loi de probabilité conjointe de toutes les variables de nœud et de lien. Les approches fondées sur le plan de sondage décrites antérieurement comprennent les méthodes d'échantillonnage en réseau ou basé sur la multiplicité (Birnbaum et Sirken 1965), l'échantillonnage en grappes adaptatif appliqué à un graphe (Thompson et Collins 2002), ainsi que quelques-unes des méthodes décrites dans la littérature sur l'échantillonnage en boule de neige (Frank 1977, 1978; Frank et Snijders 1994). Une méthode combinant les approches fondées sur le plan de sondage et sur un modèle est utilisée dans Felix-Medina et Thompson (2004) pour étudier une population cachée dans laquelle un échantillonnage par dépistage de liens est réalisé à partir d'un échantillon d'enquête probabiliste tiré d'une base de sondage couvrant uniquement une partie de la population.

L'avantage des méthodes fondées sur le plan de sondage, dans le cas de populations humaines cachées qui ont leur propre réseau social et sont difficiles à modéliser de façon réaliste, est que certaines propriétés des inférence, comme l'absence de biais et la convergence des estimateurs, ne dépendent pas d'hypothèses de modélisation. Par contre, elles dépendent de la mise en œuvre du plan de sondage comme il a été prévu; or, l'application exacte d'un plan de sondage particulier peut constituer un très grand défi dans les études de populations humaines cachées. C'est ce qui a motivé l'élaboration d'une gamme de méthodes fondées sur un modèle pour l'inférence à partir d'échantillons de graphe, y compris les techniques du maximum de vraisemblance et les techniques bayésiennes (Thompson et Frank 2000; Chow et Thompson 2003). Fondées sur l'hypothèse que l'échantillon de départ est « ignorable » au sens de la vraisemblance (Rubin 1976) ou que le plan de sondage est de forme connue de sorte qu'il peut être inclus dans les équations de la vraisemblance et de l'inférence bayésienne, ces méthodes conviennent à une très grande gamme de procédures d'échantillonnage par dépistage de liens, y compris la plupart des variantes des méthodes d'échantillonnage en boule de neige et en réseau. Toutefois, en pratique, il se peut que l'échantillon initial soit sélectionné d'une façon loin d'être ignorable, avec probabilités de sélection dépendant de la valeur de nœud, du degré de nœud et d'autres facteurs. L'omniprésence du problème de la sélection de l'échantillon initial dans les études par dépistage de liens a été soulignée par Spreen (1992), entre autres.

L'approche poursuivie dans le présent article ne repose pas sur l'hypothèse d'un contrôle total sur toutes les possibilités de plan de sondage, mais vise plutôt à tirer parti de la façon dont les échantillons ont naturellement tendance à être sélectionnés dans les populations en réseau par les ethnographes ou d'autres spécialistes des sciences sociales, les membres de la population proprement dits ou les moteurs de recherche Web automatisés. Partant de ces processus naturels de sélection, nous introduisons des modifications itératives afin d'obtenir des procédures d'échantillonnage qui, pas à pas, s'approchent des probabilités de sélection souhaitées.

Quoique la structure sous-jacente des plans de sondage décrits dans l'article dépende de chaînes de Markov, les estimateurs et les paramètres présentant le plus d'intérêt pour les chercheurs pourraient en fait ne pas être markoviens. Par exemple, alors que la séquence de sélection d'unités d'échantillonnage peut ne dépendre, à chaque pas, que de l'unité sélectionnée le plus récemment, la séquence selon laquelle des unités distinctes sont ajoutées à l'échantillon dépend de toutes les unités sélectionnées jusqu'à ce moment-là. Par conséquent, nous étudions par simulation les propriétés de plusieurs estimateurs conjugués à divers plans de sondage, en sélectionnant répétitivement des échantillons à partir de réalisations d'un graphe stochastique et à partir d'une population empirique provenant d'une étude sur des personnes courant un grand risque de transmission du VIH/Sida.

À la section 2, nous décrivons les plans à marche aléatoire. Aux sections 3 et 4, nous présentons les plans à marche uniforme et ciblée, respectivement. À la section 5, nous donnons un exemple illustratif en prenant pour population une réalisation d'un modèle de graphe stochastique et un exemple empirique en utilisant des données provenant d'une étude portant sur une population présentant un risque élevé de transmission du VIH/Sida.

2. Marche aléatoire

La population d'intérêt est un graphe, donné par un ensemble de N nœuds portant les étiquettes $U = \{1, 2, \ldots, N\}$ et ayant les valeurs $\mathbf{y} = \{y_1, \ldots, y_N\}$, et une matrice **A** de dimensions $N \times N$ indiquant les relations ou les liens entre les nœuds. Un élément a_{ij} de **A** a la valeur 1 s'il existe un lien allant de i au nœud j et la valeur 0, autrement. Nous supposons que les éléments diagonaux a_{ii} sont nuls. Pour le nœud i, la somme de ligne $a_{\bullet i}$ est le « degré sortant » ou nombre de nœuds vers lesquels i possède un lien (successeurs) et la somme de colonne $a_{\bullet i}$ est le « degré entrant » ou nombre de nœuds qui ont un lien vers i (prédécesseurs). Dans le cas d'un graphe non orienté, la matrice **A** est symétrique et le degré entrant de tout nœud est égal à son degré sortant.

Soit W_k l'unité ou le nœud du graphique qui est sélectionné lors de la k^{e} vague. Si i est le nœud sélectionné à la k^{e} vague, alors à la vague $k+1$, l'un des nœuds reliés en partant de i est sélectionné au hasard. Donc, $\{W_0, W_1, W_2, \ldots\}$ est une chaîne de Markov avec

$$P(W_{k+1} = j \mid W_k = i) = a_{ij} / a_{i\bullet}. \qquad (1)$$

Soit **Q** la matrice de transition de la chaîne avec les éléments $q_{ij} = P(W_{k+1} = j \mid W_k = i)$. La chaîne est une marche aléatoire en ce sens qu'à chaque pas, l'un des états voisins de l'état courant est sélectionné au hasard.

Si le graphe est constitué d'une seule composante connectée, c'est-à-dire si chaque nœud du graphe peut être atteint à partir de chaque autre nœud selon un certain chemin, alors la chaîne est irréductible et ses probabilités stationnaires $(\pi_1, \ldots, \pi_N)$ satisfont $\pi_j = \sum \pi_i q_{ij}$ pour $j = 1, \ldots, N$. En fait, dans le cas du plan d'échantillonnage à marche aléatoire simple dans un graphe non orienté connecté, on peut montrer que les probabilités stationnaires (Salganik et Heckathorn 2004) sont

$$\pi_j \propto a_{\bullet j}.$$

Autrement dit, dans un graphe non orienté ne comportant qu'une seule composante connectée, la fréquence de sélection de long terme de tout nœud est proportionnelle à son degré entrant, qui est égal au degré sortant, puisque le graphe n'est pas orienté.

Supposons que l'on veuille estimer une caractéristique du graphe de population, telle que la moyenne de population des valeurs de nœud $\mu_y = \sum_{i=1}^{N} y_i / N$ en utilisant des données provenant d'un échantillon sélectionné par marche aléatoire. La moyenne d'échantillon $\bar{y} = \sum_{i \in s} y_i$ n'est généralement pas sans biais, parce que la valeur y_i d'un nœud peut être reliée au degré de celui-ci et, donc, à sa probabilité d'être sélectionné. Cependant, on peut obtenir une estimation approximativement sans biais en pondérant chaque valeur y de l'échantillon par l'inverse de son degré entrant, en supposant que cette information puisse être extraite des données (Salganik et Heckathorn 2004).

2.1 Marche aléatoire avec sauts aléatoires

Dans un graphe avec composantes distinctes ou avec nœuds non connectés, la marche aléatoire simple que nous venons de décrire n'a pas la propriété que chaque nœud peut, en dernière analyse, être atteint à partir de chaque autre nœud. Sans cette propriété, la loi limite de la marche aléatoire est sensible à la loi initiale, puisque la probabilité limite de sélection d'un nœud dépend de la probabilité initiale de démarrer dans la composante qui contient ce nœud. Une modification du plan d'échantillonnage qui permet de surmonter ce problème consiste à autoriser un saut avec faible probabilité vers un nœud choisi au hasard dans l'ensemble du graphe. À chaque pas, cette marche aléatoire suit un lien sélectionné au hasard avec la probabilité d et, avec la probabilité $1-d$, saute à un autre nœud du graphe au hasard ou avec une probabilité spécifiée. Dans la littérature traitant de la recherche au sujet d'Internet, d est appelé « facteur d'amortissement », puisqu'une valeur de d inférieure à 1 amortit l'effet du degré sortant d'un nœud donné (Brin et Page 1998).

Les probabilités de transition pour la marche aléatoire avec sauts sont

$$q_{ij} = \begin{cases} (1-d)/N + d a_{ij} / a_{i\bullet} & \text{si} \quad a_{i\bullet} > 0 \\ 1/N & \text{si} \quad a_{i\bullet} = 0. \end{cases} \qquad (2)$$

Dans le cas de la faible probabilité $1-d$ d'un saut aléatoire à n'importe quel pas, la marche aléatoire markovienne peut, en principe, atteindre tout nœud du graphe à partir de tout autre nœud, de sorte que la chaîne est irréductible. En outre, les sauts aléatoires, qui comprennent la possibilité d'aller du nœud i au nœud i, assurent que la chaîne soit apériodique de sorte que les probabilités stationnaires concordent avec les probabilités limites. Si $d < 1$, la probabilité stationnaire du nœud i n'est pas une fonction simple de son propre degré entrant et dépend aussi des probabilités stationnaires des nœuds qui s'y relient.

De façon plus générale, les sauts peuvent être faits avec n'importe quelle probabilité spécifiée $\mathbf{p} = (p_1, \ldots, p_N)$ et la probabilité d'un saut peut dépendre de l'état courant, de sorte que les probabilités de transition sont

$$q_{ij} = \begin{cases} (1-d_i) p_j + d_i a_{ij} / a_{i\bullet} & \text{si} \quad a_{i\bullet} > 0 \\ 1/N & \text{si} \quad a_{i\bullet} = 0. \end{cases}$$

Des estimations des caractéristiques du graphe de population approximativement sans biais par rapport au plan peuvent être obtenues en pondérant les valeurs d'échantillon par des facteurs inversement proportionnels aux probabilités limites de sélection de la chaîne de Markov, mais avec le problème supplémentaire que ces probabilités limites sont inconnues et doivent être estimées d'après les données

d'échantillon (voir Henzinger et coll. 2000, pour une approche de ce problème).

Dans la suite de l'article, les expressions « marche aléatoire » ou « marche aléatoire ordinaire » feront référence à la marche aléatoire avec sauts, sauf indication contraire explicite.

3. Marche uniforme

À la présente section, nous proposons une modification du plan d'échantillonnage à marche aléatoire qui mène à des probabilités stationnaires uniformes $\boldsymbol{\pi} = (\pi, \ldots, \pi)$.

Commençons par considérer le cas du graphe de population constitué d'une seule composante connectée. Soit **Q** la matrice de transition pour la marche aléatoire simple avec les probabilités de transition q_{ij} données par (1). Supposons qu'au pas k, l'état du processus est i. Une sélection provisoire est faite en utilisant les probabilités de transition de la i^{e} ligne de **Q**. Supposons que la sélection provisoire soit le nœud j. Si le degré sortant $a_{j\bullet}$ du nœud j est inférieur au degré sortant $a_{i\bullet}$ du nœud i, alors la sélection pour la vague suivante est le nœud j, c'est-à-dire $W_{k+1} = j$. Si, par contre, le degré sortant du nœud j est supérieur au degré sortant du nœud i, alors un nombre aléatoire uniforme Z est sélectionné dans l'intervalle unitaire. Si $Z < a_{i\bullet} / a_{j\bullet}$, alors $W_{k+1} = j$. Sinon, $W_{k+1} = i$.

En utilisant la méthode de Hastings-Metropolis (Hastings 1970), nous construisons la matrice de transition pour la marche modifiée dans le graphe connecté au moyen des éléments

$$P_{ij} = q_{ij}\alpha_{ij} \qquad \text{pour} \quad i \neq j$$

et

$$P_{ii} = 1 - \sum_{j \neq i} P_{ij}$$

où

$$\alpha_{ij} = \min\left\{\frac{a_{i\bullet}}{a_{j\bullet}}, 1\right\}.$$

Dans le cas d'un graphe de population contenant des composantes distinctes ou des nœuds isolés, la marche aléatoire avec sauts, dont la matrice de transition **Q** est donnée par (2), peut être modifiée pour obtenir

$$P_{ij} = q_{ij}\alpha_{ij} \qquad \text{pour} \quad i \neq j$$

et

$$P_{ii} = 1 - \sum_{j \neq i} P_{ij}$$

où

$$\alpha_{ij} = \min\left\{\frac{q_{ji}}{q_{ij}}, 1\right\}.$$

Donc, pour deux nœuds mutuellement connectés i et j, la probabilité d'acceptation d'une transition de i à j est

$$\alpha_{ij} = \min\left\{\frac{(1-d)/N + d/a_{j\bullet}}{(1-d)/N + d/a_{i\bullet}}, 1\right\}.$$

Pour une transition d'une unité isolée à une unité dans une composante plus grande qu'un nœud, la probabilité d'acceptation est $\alpha_{ij} = 1 - d$. Pour les autres probabilités d'acceptation, $\alpha_{ij} = 1$. Notons aussi que dans un graphe orienté, la probabilité d'acceptation serait nulle pour le parcours d'un lien asymétrique.

La marche uniforme est appliquée, si l'état courant est i, en sélectionnant un prochain état candidat, disons j, d'après les probabilités de transition figurant sur la i^{e} ligne de **Q**. Un nombre aléatoire uniforme standard Z est sélectionné et, si $Z < \alpha_{ij}$, l'état suivant est j, tandis qu'autrement, la marche reste à l'état i pendant un pas supplémentaire.

La quantité α_{ij}, dans le cas des plans à marche uniforme, dépend des probabilités de transition connues de la marche aléatoire de base, si bien que sa mise en œuvre ne nécessite pas d'estimation.

4. Marche ciblée

La même approche peut être suivie pour construire une marche ayant n'importe quelle probabilité stationnaire spécifiée, comme la sélection des nœuds dont la valeur y est élevée avec de plus grandes probabilités ou la sélection de nœuds de telle façon que les probabilités soient strictement proportionnelles à leur degré, même si le graphe contient des composantes distinctes connectées. Soit $\pi_i(y)$ la probabilité de sélection stationnaire souhaitée pour le i^{e} nœud sous forme d'une fonction de sa valeur de y. Par exemple, lors d'une étude d'une population humaine cachée exposée au risque de transmission du VIH/Sida, supposons que l'on souhaite échantillonner les utilisateurs de drogues injectables $(y_i = 1)$ avec une probabilité double de celle appliquée pour les membres de cette population qui ne prenne pas ce genre de drogues $(y_i = 0)$. Les probabilités de transition pertinentes pour la marche à valeur ciblée, en utilisant de nouveau la méthode de Hastings-Metropolis, sont

$$P_{ij} = q_{ij}\alpha_{ij} \qquad \text{pour} \quad i \neq j$$

et

$$P_{ii} = 1 - \sum_{j \neq i} P_{ij}$$

où

$$\alpha_{ij} = \min\left\{\frac{\pi_j q_{ji}}{\pi_i q_{ij}}, 1\right\}.$$

Soulignons que la probabilité de transition de base est connue, puisqu'elle dépend uniquement du degré sortant des nœuds observés, de la probabilité choisie d et du ratio spécifié π_j / π_i.

Dans le cas d'une marche pour laquelle la probabilité de sélection relative dépend de la valeur de y, le ratio $\pi_j(y_j)/\pi(y_i)$ est spécifié et

$$\alpha_{ij} = \min\left\{\frac{\pi_j(y_j) q_{ji}}{\pi_i(y_i) q_{ij}}, 1\right\}.$$

Un autre exemple de marche ciblée pourrait être celui d'une distribution cible obtenue en sélectionnant les nœuds proportionnellement à leur degré sortant, c'est-à-dire au nombre de liens qui en partent. Puisque le degré d'un nœud isolé est nul, une possibilité que nous nommerons marche ciblée selon le « degré + 1 », consiste simplement à ajouter une unité à chaque degré, de sorte que $\pi_i \propto a_{i\bullet} + 1$ soit la probabilité de sélection cible.

Un choix légèrement différent, appelé simplement marche ciblée selon le degré, consiste à n'ajouter une unité qu'au degré des nœuds isolés, de sorte que $\pi_i \propto \max(a_{i\bullet}, 1)$. Pour une marche ciblée selon le degré de ce type, la probabilité d'acceptation d'une transition entre deux nœuds connectés mutuellement est

$$\alpha_{ij} = \min\left\{\frac{a_{j\bullet}(1-d)/N + 1}{a_{i\bullet}(1-d)/N + 1}, 1\right\}.$$

Pour une transition entre un nœud isolé et un nœud dont le degré est positif, la probabilité est

$$\alpha_{ij} = \min(a_{j\bullet}(1-d), 1).$$

La probabilité de transition entre deux nœuds ayant chacun un degré positif est

$$\alpha_{ij} = \min\left\{\frac{a_{j\bullet}}{a_{i\bullet}}, 1\right\}.$$

Dans ce cas,

$$\alpha_{ij} = \min\left\{\frac{a_{j\bullet} q_{ji}}{a_{i\bullet} q_{ij}}, 1\right\}.$$

Puisque les nœuds isolés, n'ayant aucun lien avec d'autres nœuds, sont de degré nul, afin de leur donner une probabilité de sélection positive, on peut attribuer arbitrairement à leur degré la valeur « 1 » dans le calcul de la marche ciblée selon le degré ou ajouter la valeur 1 au degré de chaque nœud.

5. Plan de sondage à marche sans remise

Les résultats relatifs aux lois limites des sections précédentes s'appliquent exactement au plan de sondage à marche aléatoire avec remise, de sorte que la sélection des nœuds peut se poursuivre indéfiniment au sein de la population finie. Certains estimateurs utilisés dans les exemples qui suivront sont toutefois fondés sur la séquence d'unités distinctes sélectionnées par ce processus. Dans le cas de la séquence d'unités distinctes, qui, en fait, fournit un échantillon à marche aléatoire sans remise, on ne peut ajouter des unités que jusqu'à ce que le nombre de nœuds distincts soit le même dans l'échantillon que dans la population finie, point auquel la moyenne d'échantillon et la moyenne de population coïncident.

Une autre procédure en vue de sélectionner un échantillon par marche aléatoire sans remise consiste à restreindre directement la sélection de l'unité suivante, à n'importe quel pas, à l'ensemble d'unités qui n'ont pas encore été sélectionnées, comme dans la « marche aléatoire auto-évitante » (Lovász 1993). Si l'on utilise une procédure sélection-rejet comme dans les marches ciblées, la sélection suivante est faite d'après l'ensemble d'unités qui n'ont fait l'objet d'aucune sélection provisoire, que l'unité ait été ou non acceptée.

6. Estimateurs fondés sur les valeurs des nœuds acceptés

Sous une marche aléatoire uniforme avec remise, la moyenne d'échantillon tirage par tirage de la série de valeurs acceptées est asymptotiquement sans biais par rapport à la moyenne de population, parce que les probabilités de sélection limites sont toutes égales. La moyenne de l'échantillon tirage par tirage est la moyenne nominale englobant les valeurs répétées, de sorte que la valeur d'un nœud est pondérée par le nombre de fois que le nœud est sélectionné. Si l'on utilise un plan de sondage sans remise, ce même estimateur n'est pas précisément asymptotiquement sans biais, parce que les probabilités limites ne sont pas exactement égales. L'estimateur de variance type fondé sur une variance d'échantillon à marche aléatoire n'est pas sans biais, à cause de l'interdépendance des marches aléatoires. Les estimateurs de la variance sont estimés empiriquement dans les exemples.

Dans le cas d'une marche ciblée dans laquelle la probabilité limite π_i du nœud i est proportionnelle à c_i, un estimateur asymptotiquement convergent, fondé sur les probabilités limites, est fourni par l'estimateur par le ratio généralisé

$$\hat{\mu} = \frac{\sum_{s_a} y_i / c_i}{\sum_{s_a} 1/q_i}.$$

Notons que l'estimateur d'Horvitz-Thompson ne peut être utilisé, parce que la constante de proportionnalité des probabilités d'inclusion est inconnue, tandis que dans l'estimateur par le ratio généralisé, elle s'annule. De nouveau, les probabilités limites sur lesquelles est fondé l'estimateur sont vérifiées exactement pour le plan de sondage avec remise. Pour la variante sans remise, l'estimateur est examiné empiriquement dans les exemples.

7. Exemples

7.1 Graphe stochastique réalisé

La figure 1 illustre, pour commencer, une petite population simulée de 60 nœuds. Les nœuds dont la valeur est $y = 1$ sont de couleur foncée et ceux dont la valeur est $y = 0$ sont clairs. Nous considérons la réalisation entière comme étant notre population d'intérêt. Le modèle qui produit la réalisation est un modèle stochastique en blocs dans lequel la probabilité d'un lien entre deux nœuds quels qu'ils soient dépend des valeurs de nœud. Des liens sont plus susceptibles d'exister entre des nœuds de même type et les nœuds foncés sont plus fortement connectés que les nœuds de couleur claire. Par exemple, on pourrait souhaiter estimer la proportion de nœuds positifs (c'est-à-dire de nœuds dont $y = 1$) dans le graphe. Dans le graphe de population, 24 des 60 nœuds sont positifs, si bien que la proportion réelle est de 0,4. Le même graphe est présenté à droite, mais avec les tailles de nœud proportionnelles aux probabilités de sélection limites de la marche aléatoire. Étant donné que les nœuds positifs ont davantage tendance à former des liens, nombre d'entre eux ont une probabilité de sélection supérieures à la moyenne.

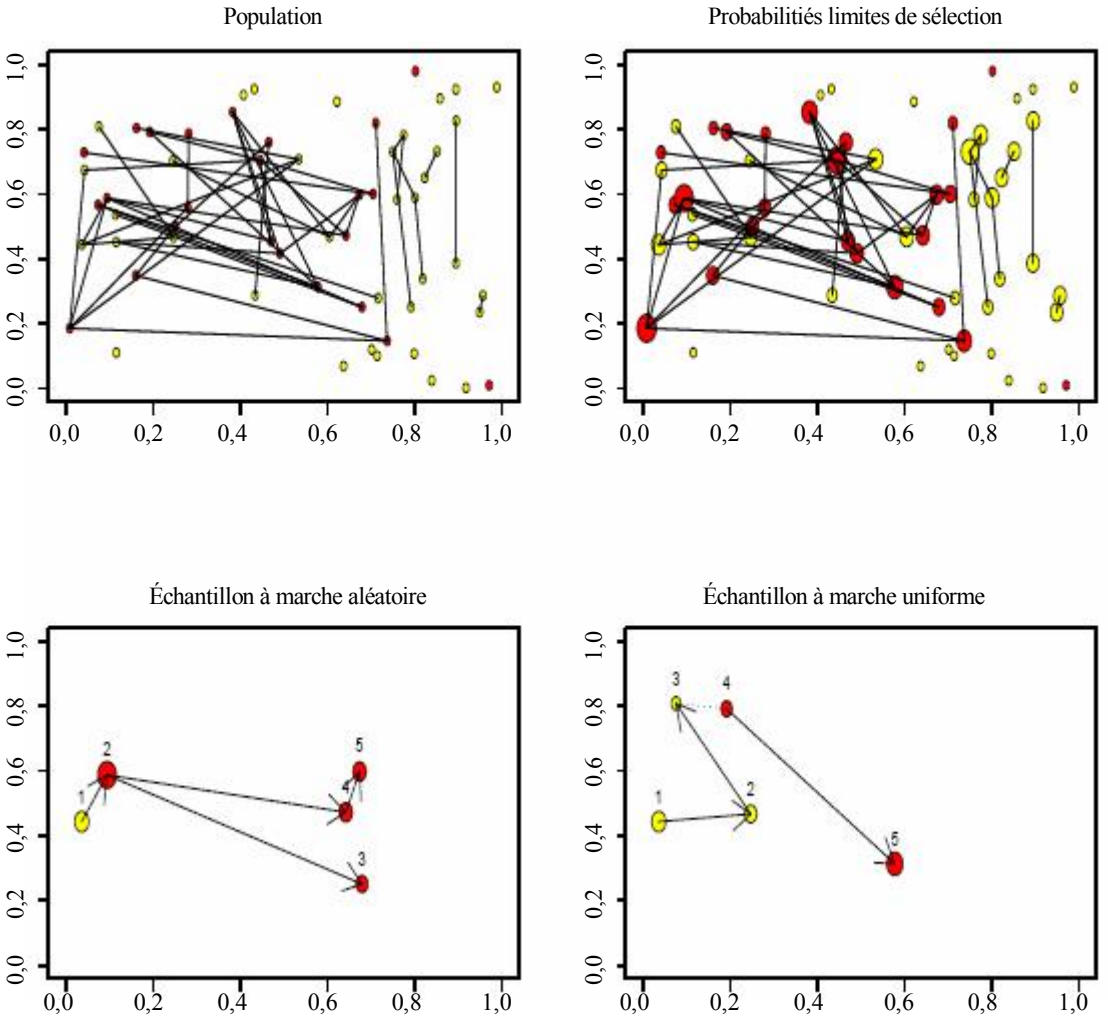

Figure 1. En haut à gauche : La population est la réalisation du modèle de graphe stochastique en blocs. En haut à droite : Probabilités limites des nœuds de la marche aléatoire. En bas à gauche : Marche aléatoire de cinq pas. En bas à droite : Marche uniforme de cinq pas. Des échelles d'axe arbitraires sont fournies comme aides visuelles pour distinguer les nœuds d'échantillon des nœuds de population.

À la rangée inférieure de la figure 1 sont présentées une marche aléatoire et une marche uniforme sélectionnées à partir de la population, comme il est illustré. Chacune a pour point de départ le même nœud sélectionné au hasard, dénoté « 1 », et se poursuit jusqu'à ce que cinq nœuds distincts soient sélectionnés. Les flèches indiquent la direction dans laquelle sont suivis les liens et un saut vers un nouveau nœud sélectionné au hasard dans le graphe est indiqué parune ligne en pointillé. Notons que la marche aléatoire revient en arrière du troisième nœud sélectionné vers le deuxième avant de suivre un nouveau lien vers le quatrième nœud échantillonné. À partir du premier nœud échantillonné, la marche uniforme passe le nœud à probabilité plus élevée sélectionné par la marche aléatoire et accepte à sa place un des nœuds qui y sont liés. Ces marches peuvent l'une et l'autre, à tout moment, faire un saut aléatoire, quoique dans les exemples illustrés, seule la marche uniforme en fasse un, lors de la transition du troisième au quatrième nœud échantillonné.

7.2 Population empirique

Des données provenant d'une étude sur la transmission hétérosexuelle du VIH/Sida dans une population à risque élevé à Colorado Springs (Potterat et coll. 1993; Rothenberg et coll. 1995) sont présentées aux figures 2 et 3. Les 595 membres de la population étudiée qui ont été interviewés sont représentés par les nœuds du graphe, et les relations sexuelles déclarées entre ces personnes sont représentées par des liens (arcs) entre les nœuds. (Les liens d'ordre sexuel supplémentaires de n'importe laquelle de ces 595 personnes avec des personnes qui n'ont pas été interviewées subséquemment ne sont pas présentés.) La population étudiée comprend les personnes à risque, c'est-à-dire les utilisateurs de drogues injectables, les travailleurs du sexe, leurs partenaires sexuels et d'usage de drogues, ainsi que d'autres personnes avec lesquelles ils ont des contacts sociaux étroits. La variable de nœud illustrée est celle de la prostitution, avec une couleur foncée pour une valeur positive ($y = 1$). Seuls les liens de nature sexuelle sont représentés, quoique bon nombre d'entre eux coïncident avec les liens relatifs à la consommation de drogues. La composante sexuellement connectée la plus importante du graphique contient 219 personnes. La composante connectée suivante, par ordre décroissant de taille, contient 12 personnes, et est suivie par plusieurs composantes de 4, 3 et 2 personnes. Les nœuds restants représentent des personnes n'ayant aucun contact sexuel déclaré au sein de la population interviewée.

Le profil observé de cette population, dont une composante connectée est beaucoup plus grande que les autres, a été décrit par divers chercheurs comme n'étant pas atypique des études portant sur des populations cachées à risque. Nous utilisons la population susmentionnée uniquement à titre de population empirique à partir de laquelle nous sélectionnons des échantillons afin de comparer des plans d'échantillonnage et des estimateurs.

La figure 3 représente la même population avec la taille de nœud tirée proportionnellement à la probabilité de sélection limite de la marche aléatoire.

Chaque tracé de la figure 4 montre la moyenne d'échantillon cumulative d'une marche unique poursuivie jusqu'à ce que 120 nœuds distincts soient sélectionnés. La proportion réelle de nœuds positifs (valeur 1) dans la population empirique (0,2235) est représentée par la droite horizontale dans chaque tracé.

Les tracés de la rangée supérieure de la figure 4 représentent une marche aléatoire ordinaire dont le nœud de départ est sélectionné aléatoirement. Le tracé de gauche montre la moyenne d'échantillon cumulative des unités distinctes. Le tracé de droite montre les mêmes données, mais avec la moyenne d'échantillon tirage par tirage, qui comprend les sélections répétées d'un même nœud, de sorte que chaque valeur de nœud soit pondérée par le nombre de fois que le nœud a été sélectionné durant la marche aléatoire.

Dans la rangée inférieure de la figure 4, nous montrons les deux mêmes types de moyenne d'échantillon pour une marche uniforme poursuivie jusqu'à ce que 120 nœuds distincts soient sélectionnés. Notons que, pour la marche aléatoire ordinaire, les fluctuations de la moyenne d'échantillon ont lieu principalement au-dessus de la moyenne réelle, ce qui représente le biais positif résultant de la sélection préférentielle des personnes plus fortement connectées, à plus haut risque, dans la population. Dans le cas de la marche uniforme, la moyenne d'échantillon fluctue plus près de la valeur réelle, sa valeur étant parfois supérieure et parfois inférieure. Chacun de ces tracés donne aussi une idée de l'autocorrélation présente dans une chaîne de Markov unique.

Les tracés de la figure 5 représentent la valeur attendue des nœuds à mesure qu'une marche progresse vague par vague, pour divers types de marches et diverses lois initiales à partir desquelles est sélectionné le premier nœud, pour la population empirique de 595 nœuds. Donc, pour la k^{e} vague, les tracés représentent $\mathrm{E}(Y_k)$, où Y_k est la valeur du nœud sélectionné à la k^{e} vague. La ligne en trait interrompu montre la moyenne réelle pour la population de Colorado Springs (0,2235). Les trois autres courbes représentent trois lois initiales différentes. Dans tous les cas, la courbe qui part du nœud le plus bas est la loi initiale uniforme, puisque la moyenne pour le nœud initial sélectionné au hasard est égale à la moyenne de la population. La loi initiale fondée sur la valeur de nœud, selon laquelle les nœuds positifs ($y = 1$) ont une probabilité initiale de sélection double des nœuds

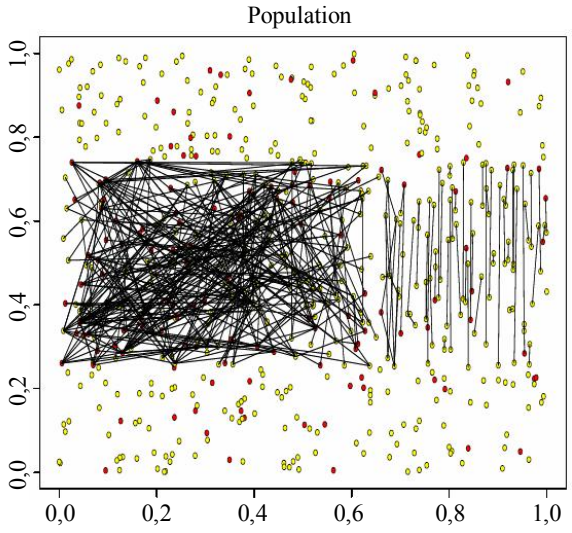

Figure 2. Population à haut risque de l'étude de Colorado Springs sur la transmission hétérosexuelle du VIH/Sida (Potterat et coll. 1993; Rothenberg et coll. 1995, et communications personnelles). Les cercles foncés représentent les individus présentant le risque le plus élevé, ici ceux qui se sont prostitués. Les liens entre les individus sont ceux de relations sexuelles et d'injection de drogues.

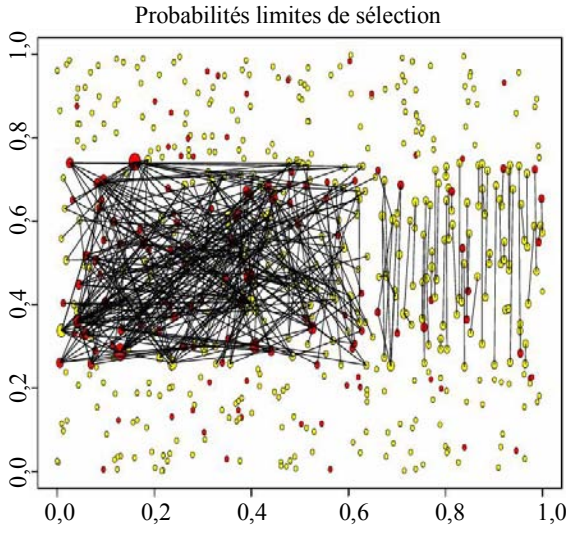

Figure 3. Probabilités limites de sélection par marche aléatoire pour la population de Colorado Springs. Soulignons que, dans la population réelle, un grand nombre d'individus présentant le comportement à risque le plus élevé ont aussi une forte probabilité d'être sélectionnés dans le cas de la marche aléatoire ordinaire et auront donc tendance à être surreprésentés dans l'échantillon.

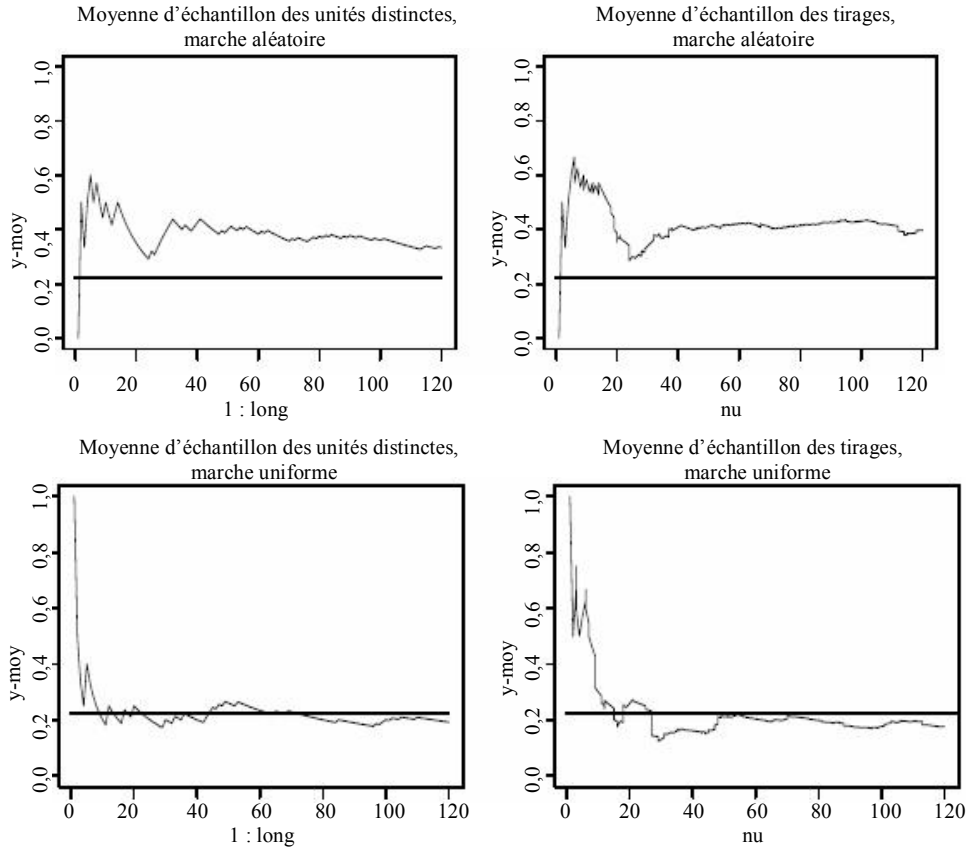

Figure 4. Chemins d'échantillon des moyennes d'échantillon pour une marche aléatoire unique de 120 nœuds de long. Les deux tracés supérieurs correspondent à une marche aléatoire ordinaire, et les deux tracés inférieurs, à une marche uniforme. La moyenne d'échantillon des unités distinctes, jusqu'à la vague donnée par l'axe des x, est représentée à gauche. La moyenne d'échantillon des tirages nominaux est représentée à droite, de sorte que la valeur de nœud soit pondérée par le nombre de fois que le nœud est sélectionné.

nuls ($y = 0$), donne la courbe de l'espérance qui est, dans tous les cas, principalement au milieu au départ et manifeste la tendance la plus forte vers une périodicité initiale. La loi initiale fondée sur le degré, selon laquelle la probabilité initiale de sélection d'un nœud est proportionnelle à son degré (plus une unité, puisque le degré des nœuds isolés est nul), forme la courbe supérieure dans chacun des tracés.

Les six tracés de la figure 5 montrent les espérances des valeurs de nœud pour six différents types de marches. Pour une marche aléatoire qui suit seulement les liens, sans possibilité de sauts aléatoires, la loi à long terme est fonction du point de départ, lequel dépend de la loi initiale. Les trois lignes séparées du premier tracé reflètent la sensibilité à la loi initiale. Par ailleurs, la marche aléatoire avec sauts permet à n'importe quel nœud d'être atteint par n'importe quel autre de sorte que la loi limite est atteinte assez rapidement peut importe la loi initiale. Avec la marche uniforme, celle qui débute avec la loi uniforme demeure dans la loi uniforme, vague après vague tandis que les marches qui débutent avec les autres lois inégales décrites, tendent assez rapidement vers la loi uniforme. Chacune de marches qui dépendent de la valeur ou du degré atteint sa loi limite assez rapidement, avec une espérance de la valeur du nœud passablement plus élevée que la valeur moyenne dans la population. La marche « degré +1 » atteint une loi dont les probabilités de sélection sont proportionnelles à un plus la valeur de chaque nœud tandis que la marche selon le degré tend vers une loi dont les probabilités limites sont proportionnelles au degré de chaque nœud sauf que les nœuds isolés se voient attribuer une valeur de degré égale à un.

Les tableaux 1 et 2 montrent les valeurs calculées de d'espérance de y pour la population de l'étude de Colorado Springs pour chaque type de marche, vague par vague, et avec diverses lois de départ pour la sélection des nœuds. Les résultats pour les marches aléatoires ordinaires sont présentés au tableau 1 et pour les marches uniformes, au tableau 2. Les espérances sont présentées pour les sélections initiales, les vagues 1, 2, 3, 4, 5, 6, 8, 16 et 32, et pour la limite à mesure que le nombre de vagues tend vers l'infini. Les trois lois initiales considérées pour la sélection du premier nœud d'une marche sont la sélection aléatoire, la sélection avec probabilité deux fois plus élevée pour les nœuds positifs que pour les nœuds à valeur nulle, et la sélection proportionnelle au degré entrant de chaque nœud plus 1. Notons que, dans le cas de k marches indépendantes

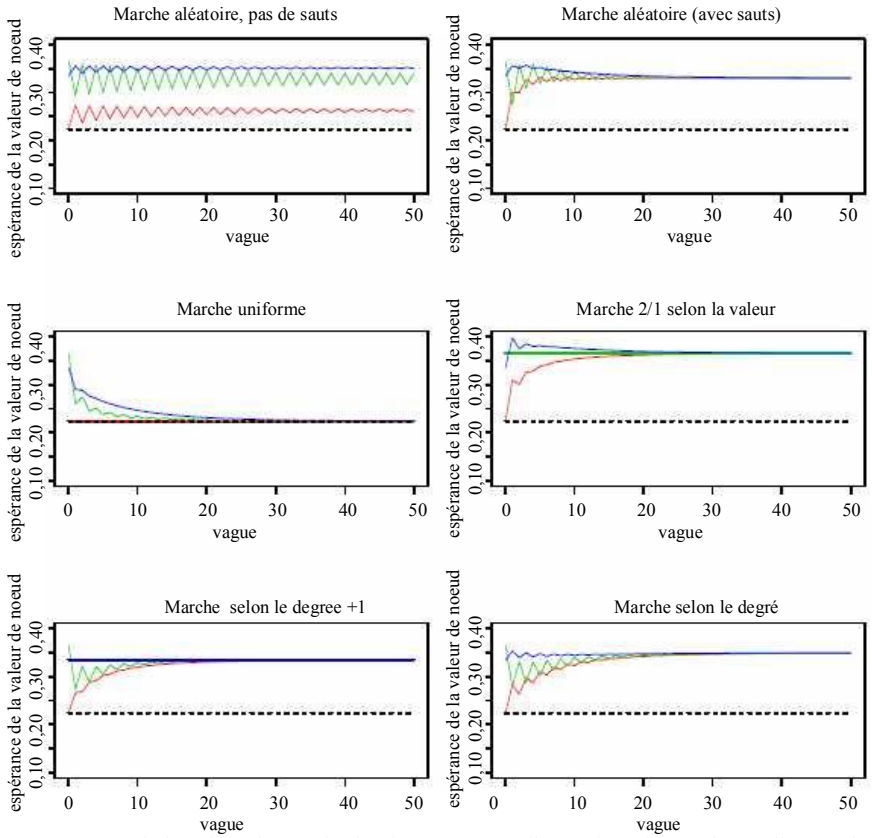

Figure 5. Espérance de la valeur de nœud selon la vague pour divers plans de marche appliqués à la population empirique de Colorado Springs. Chaque tracé illustre un plan de marche. La courbe en trait interrompu représente la moyenne réelle. Les trois autres courbes représentent l'espérance pour les trois distributions initiales examinées. Dans chaque cas, la courbe inférieure débute par la loi uniforme, celle du milieu, par la loi de probabilité 2/1 selon la valeur et la courbe supérieure, par la loi de probabilité selon le degré.

selon un plan donné, les espérances à la vague j s'appliqueraient à la moyenne d'échantillon des k valeurs de y à la vague j provenant de chacune des marches.

Tableau 1
Marches aléatoires : espérance de y pour les vagues 0, 1, 2, 3, 4, 5, 6, 8, 16, 32 et à l'infini. La vague 0 correspond à la sélection initiale. Trois hypothèses de sélection initiale différentes sont appliquées : sélection initiale aléatoire ($\pi_0 = 1/N$ pour tous les nœuds), probabilité de sélection des nœuds de valeur $y = 1$ double de celle des nœuds de valeur $y = 0 (\pi_0 \propto y+1)$, et probabilité de sélection initiale proportionnelle au degré entrant du nœud plus 1 ($\pi_0 \propto a_{\bullet j}+1$). La moyenne réelle des valeurs de nœud pour cette population est égale à 0,2235294

valeur	$\pi_0 = 1/N$	$\pi_0 \propto y+1$	$\pi_0 \propto a_{\bullet j}+1$
0	0,2235294	0,3653846	0,3349894
1	0,2998771	0,2752690	0,3560839
2	0,3005446	0,3587093	0,3507451
3	0,3273606	0,3082865	0,3570490
4	0,3177081	0,3594697	0,3500041
5	0,3320705	0,3179675	0,3528395
6	0,3231213	0,3542086	0,3469835
8	0,3256034	0,3490933	0,3440449
16	0,3291087	0,3372548	0,3363884
32	0,3302606	0,3313908	0,3315119
∞	0,3303787	0,3303787	0,3303787

Tableau 2
Marches uniformes : espérance de y pour les vagues 0, 1, 2, 3, 4, 5, 6, 8, 16, 32 et à l'infini, pour trois hypothèses de sélection initiale différentes

valeur	$\pi_0 = 1/N$	$\pi_0 \propto y+1$	$\pi_0 \propto a_{\bullet j}+1$
0	0,2235294	0,3653846	0,3349894
1	0,2235294	0,2590239	0,2903147
2	0,2235294	0,2741356	0,2877974
3	0,2235294	0,2447258	0,2761270
4	0,2235294	0,2511473	0,2707929
5	0,2235294	0,2372440	0,2646280
6	0,2235294	0,2420866	0,2600923
8	0,2235294	0,2371714	0,2522952
16	0,2235294	0,2285370	0,2352150
32	0,2235294	0,2243635	0,2256228
∞	0,2235294	0,2235294	0,2235294

Dans le cas des marches aléatoires ordinaires, qui ont pour point de départ l'échantillon initial, la valeur observée n'est sans biais par rapport à la valeur de population que pour la sélection initiale, puis le biais augmente rapidement pour atteindre sa valeur limite de 0,3303787−0,223594. Comme les échantillons initiaux présentent un biais en faveur des nœuds positifs, le biais change moins à mesure que la marche progresse.

Dans le cas de la marche uniforme, la sélection aléatoire initiale coïncide avec la distribution stationnaire, de sorte que la marche demeure sans biais vague après vague. Dans le cas de la sélection initiale des nœuds positifs avec probabilité double de celle des nœuds de valeur nulle, le biais est réduit considérablement après chacune des quelques premières vagues et les valeurs des nœuds sélectionnés s'approchent de leur état limite sans biais. Dans le cas de la sélection initiale proportionnelle au degré entrant plus 1, quelques vagues de plus sont nécessaires pour que le biais devienne petit. Le rapprochement initial rapide de l'espérance vers la valeur limite donne à penser qu'il pourrait être souhaitable de considérer une période initiale « de rodage » qui ne sera pas utilisée dans l'estimation. Même un rodage très court d'une à trois vagues pourrait réduire sensiblement le biais des estimateurs fondés sur de courtes marches.

Les figures 6 à 9 illustrent les distributions d'échantillonnage des moyennes d'échantillon et des estimateurs pondérés pour divers plans à marche aléatoire pour l'ensemble de données de Colorado Springs. Chaque histogramme est basé sur 1 000 simulations du plan d'échantillonnage appliqué à la population empirique. Pour les plans illustrés aux figures 6 et 7, chaque échantillon comprend 24 marches, chacune de 5 pas, c'est-à-dire continuant jusqu'à ce que 5 nœuds distincts soient sélectionnés. La figure 5 représente les distributions des moyennes d'échantillon pour les marches aléatoires (rangée supérieure) et pour les marches uniformes (rangée inférieure). La distribution de la moyenne des 24 moyennes d'échantillon des 5 unités distinctes est donnée à gauche. À droite est donnée la moyenne des 24 moyennes tirage par tirage, qui intègre les sélections répétées.

La proportion réelle (0,2235) de nœuds ayant la valeur y dans la population empirique est indiquée par le triangle plein, tandis que la moyenne de la distribution d'échantillonnage est indiquée par le triangle vide. Dans le cas des marches aléatoires, les moyennes d'échantillon présentent un biais par excès, tandis que dans le cas de la marche uniforme, elles sont presque sans biais. La moyenne n'est précisément sans biais ni dans l'un ni dans l'autre cas parce que la marche se poursuit jusqu'à ce qu'un nombre fixe de nœuds distincts soit sélectionné au lieu de se poursuivre pendant un nombre fixe de vagues.

La figure 7 illustre la distribution de l'estimateur par le ratio généralisé pour les marches ciblées dont les probabilités stationnaires sont reliées à la valeur des nœuds et à leur degré (degré du nœud plus 1). Aux fins de comparaison, chacune de ces marches a débuté dans sa propre loi stationnaire, ce qui donne en fait les distributions des estimateurs après le « rodage ». Ces estimateurs ne sont pas dépourvus de biais, parce que la taille effective de l'échantillon est fixe, ce qui affecte les probabilités réelles avec lesquelles les nœuds distincts sont sélectionnés en série et que le dénominateur de l'estimateur est aléatoire, puisqu'il est égal à la somme des poids de sondage.

Les figures 8 et 9 donnent les distributions des mêmes estimateurs et plans de sondage qu'aux figures 6 et 7, mais dans le cas où chaque échantillon consiste en une longue marche de 120 nœuds distincts.

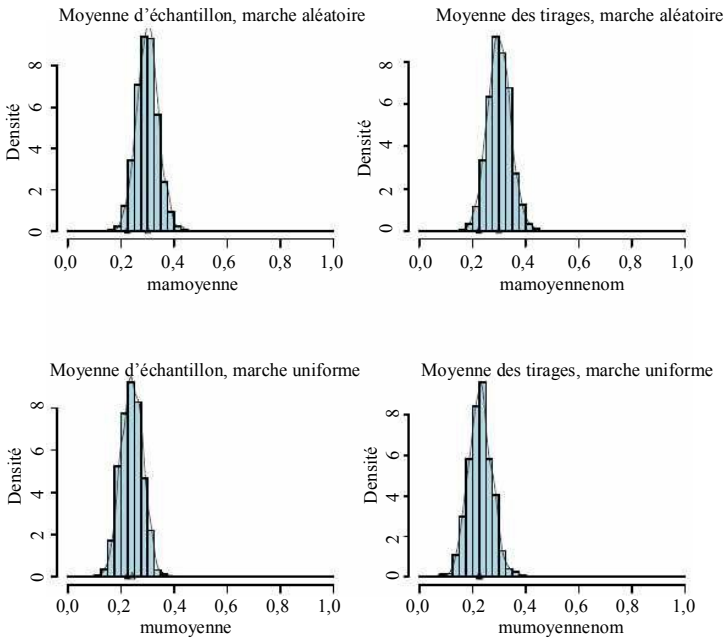

Figure 6. Distributions des moyennes d'échantillon en tant qu'estimateurs de la proportion de personnes qui se sont prostituées dans la population empirique de l'étude de Colorado Springs, sous les marches aléatoires et uniformes. Le triangle plein représente la proportion réelle dans la population. Le triangle vide représente la moyenne de la distribution de l'estimateur. À noter la surestimation dans le cas des moyennes d'échantillon pour les marches aléatoires ordinaires. Les marches aléatoires sont représentées à la partie supérieure de la figure et les marches uniformes, à la partie inférieure. Le plan comporte 24 marches, chacune de 5 pas, et l'ensemble des 120 observations sont utilisées dans l'estimateur. Le nombre de réalisations de la simulation est égal à 1 000.

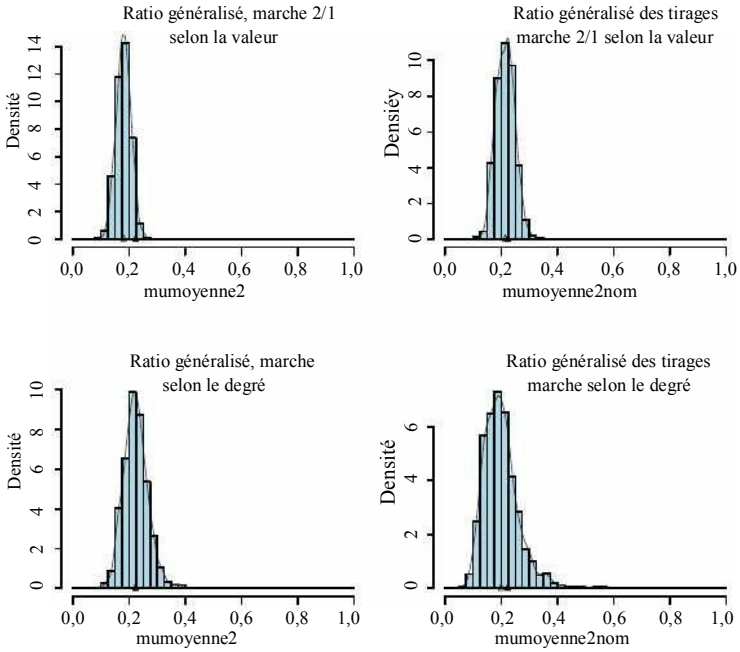

Figure 7. Distributions des estimateurs par le ratio généralisé de la proportion de personnes qui se sont prostituées dans la population empirique de l'étude de Colorado Springs, sous marches ciblées. Le triangle plein représente la proportion réelle dans la population. Le triangle vide représente la moyenne de la distribution de l'estimateur. À noter la surestimation dans le cas des moyennes d'échantillon pour les marches aléatoires ordinaires. Les marches aléatoires sont représentées à la partie supérieure de la figure et les marches uniformes, à la partie inférieure. Le plan comprend 24 marches, de 5 pas chacune, et l'ensemble des 120 observations sont utilisées dans l'estimateur. Le nombre de réalisations de la simulation est égal à 1 000.

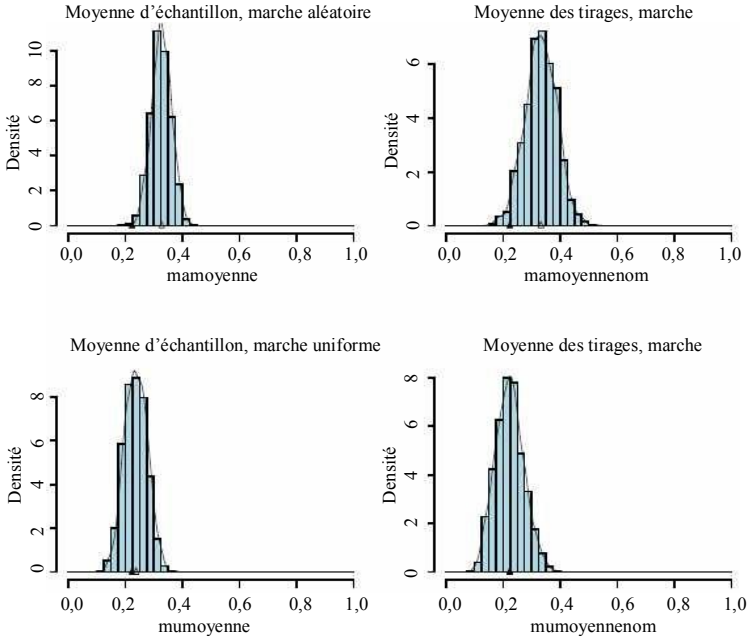

Figure 8. Distributions des moyennes d'échantillon en tant qu'estimateurs de la proportion de personnes qui se sont prostituées dans la population empirique de l'étude de Colorado Springs, sous marches aléatoires et uniformes. Le triangle plein représente la proportion réelle dans la population. Le triangle vide représente la moyenne de la distribution de l'estimateur. À noter la surestimation dans le cas des moyennes d'échantillon pour les marches aléatoires ordinaires. Les marches aléatoires sont représentées à la partie supérieure de la figure et les marches uniformes, à la partie inférieure. Le plan comprend une marche unique de 120 pas. Le nombre de réalisations de la simulation est égal à 1 000.

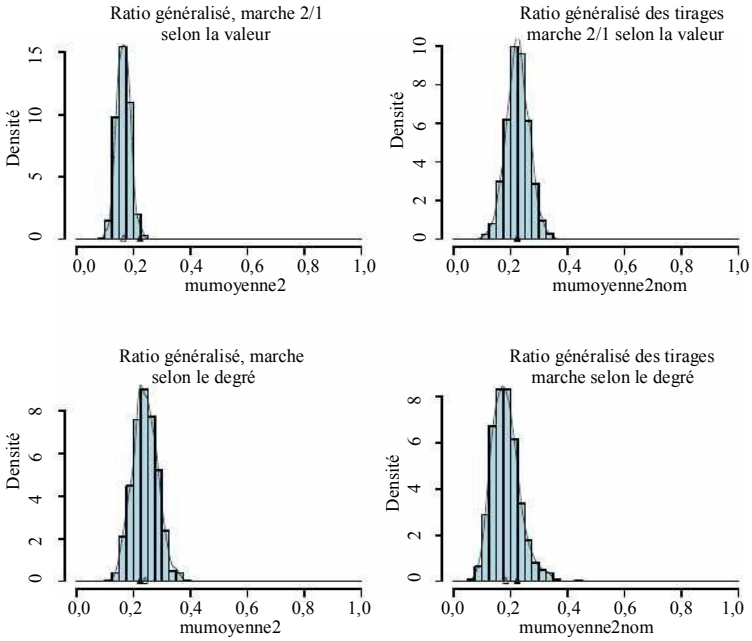

Figure 9. Distributions des estimateurs par le ratio généralisé en tant qu'estimateurs de la proportion de personnes qui se sont prostituées dans la population empirique de l'étude de Colorado Springs, sous marches ciblées. Le triangle plein représente la proportion réelle dans la population. Le triangle vide représente la moyenne de la distribution de l'estimateur. À noter la surestimation dans le cas des moyennes d'échantillon pour les marches aléatoires ordinaires. Les marches aléatoires sont représentées à la partie supérieure de la figure et les marches uniformes, à la partie inférieure. Le plan comprend une marche unique de 120 pas. Le nombre de réalisations de la simulation est égal à 1 000.

Les tableaux 3 à 6 résument les espérances et les erreurs quadratiques moyennes des estimateurs calculées pour les diverses stratégies d'après les 1 000 simulations exécutées en prenant l'ensemble de données de Colorado Springs comme population.

Tableau 3
Moyennes et erreurs quadratiques moyennes des moyennes d'échantillon des unités distinctes et moyennes tirage par tirage pour les marches aléatoires et les marches uniformes. Le plan de sondage comporte 24 marches se poursuivant chacune jusqu'à ce que 5 nœuds distincts soient inclus

Plan :	Marche aléatoire	Marche aléatoire	Marche uniforme	Marche uniforme
Estimateur :	Moyenne d'échantillon	Moyenne du tirage	Moyenne d'échantillon	Moyenne du tirage
moyenne	0,3008000	0,2994872	0,2423000	0,2289125
e.q.m.	0,007617465	0,007608868	0,002016378	0,001974826

Tableau 4
Moyennes et erreurs quadratiques moyennes pour les moyennes pondérées (estimateur par le ratio généralisé), en utilisant les unités distinctes dans chaque marche ou les sélections tirage par tirage pour les marches en fonction de la valeur et les marches en fonction du degré. Le plan comprend 24 marches se poursuivant chacune jusqu'à ce que 5 nœuds distincts soient inclus

Plan :	Marche selon la valeur	Marche selon la valeur	Marche selon le degré	Marche selon le degré
Estimateur :	Unités distinctes	Tirage par tirage	Unités distinctes	Tirage par tirage
moyenne	0,1805114	0,2144555	0,2235257	0,1994530
e.q.m.	0,002546968	0,001195507	0,001807981	0,004382568

Tableau 5
Moyennes et erreurs quadratiques moyennes pour les moyennes d'échantillon des unités distinctes et les moyennes tirage par tirage pour les marches aléatoires et les marches uniformes. Le plan comprend une marche se poursuivant jusqu'à ce que 120 nœuds distincts soient inclus

Plan :	Marche aléatoire	Marche aléatoire	Marche uniforme	Marche uniforme
Estimateur	Moyenne d'échantillon	Moyenne du tirage	Moyenne d'échantillon	Moyenne du tirage
moyenne	0,3274083	0,3325171	0,2379333	0,2232534
e.q.m.	0,012004961	0,014902382	0,001777285	0,002442825

Tableau 6
Moyennes et erreurs quadratiques moyennes pour les moyennes pondérées (estimateur par le ratio généralisé), en utilisant les unités distinctes dans chaque marche ou les sélections tirage par tirage pour les marches selon la valeur et selon le degré. Le plan comprend une marche se poursuivant jusqu'à ce que 120 nœuds distincts soient inclus

Plan :	Marche selon la valeur	Marche selon la valeur	Marche selon le degré	Marche selon le degré
Estimateur :	Unités distinctes	Tirage par tirage	Unités distinctes	Tirage par tirage
moyenne	0,1652275	0,2254267	0,2404622	0,835336
e.q.m.	0,003952703	0,001578039	0,002115518	0,03951540

Les tableau 7 et 8 donnent la variance et l'espérance des variances d'échantillon inter-marches, lorsqu'elles existent, et des variances d'échantillon intramarche pour les plans à marche uniforme.

Tableau 7
Variance des estimateurs et espérance des variances d'échantillon inter-marches et intramarche pour la marche aléatoire uniforme, pour le plan comportant 24 marches de 5 nœuds distincts chacune

Estimateur :	Moyenne d'échantillon	Moyenne tirage par tirage
Variance de l'estimateur :	0,001665709	0,001947796
E (variance inter-marches)	0,001584203	0,001919005
E (variance intramarche moyenne)	0,001515521	0,001231983

Tableau 8
Variance des estimateurs et espérance de la variance d'échantillon intramarche pour la marche aléatoire uniforme, pour le plan comportant une seule marche de 120 nœuds distincts (aucune variance d'échantillon inter-marches n'est disponible pour ce plan)

Estimateur	Moyenne d'échantillon	Moyenne tirage par tirage
Variance de l'estimateur	0,001571384	0,002445194
E (variance intramarche moyenne)	0,001510515	0,001429126

Tableau 9
Taux d'acceptation pour les marches uniforme et ciblée dans la population empirique

Plan :	Marche uniforme	Marche selon la valeur	Marche selon le degré+1	Marche selon le degré
Taux d'acceptation	0,62	0,60	0,85	0,88

8. Taux d'acceptation

Les principaux avantages des plans d'échantillonnage à chaîne de Markov contrôlée, comme les marches uniformes et ciblées, sont les suivants : 1) ils permettent de connaître les probabilités limites de sélection d'après les données, de sorte que celles-ci peuvent être utilisées dans l'estimation, 2) les probabilités limites sont choisies de sorte que certains types de nœuds ou de caractéristiques des graphes puissent être sélectionnés de manière préférentielle, 3) les estimations sont fondées sur le plan d'échantillonnage, de sorte que certaines de leurs propriétés essentielles ne dépendent pas des hypothèses, qui pourraient s'avérer incorrectes, au sujet du graphe de population proprement dit et 4) à mesure que la longueur de la chaîne augmente, l'espérance des estimations a tendance à évoluer vers la quantité correspondante des graphes, même lorsque la loi de sélection initiale diffère de la loi limite. En outre, les plans à marche uniforme produisent un échantillon qui, sans pondération ni analyse, est au pied de la lettre « représentatif » à certains égards de l'ensemble de la population.

Dans le cas des marches uniformes et ciblées, l'une des questions pratiques importantes est celle du taux d'acceptation, c'est-à-dire la probabilité moyenne qu'un nœud sélectionné provisoirement soit accepté. Les nœuds sélectionnés provisoirement qui sont rejetés ne contribuent pas aux estimateurs simples. Dans le cas d'une population telle qu'Internet, pour laquelle les sélections provisoires et les décisions d'acceptation/rejet peuvent être automatisées et exécutées rapidement, le taux d'acceptation n'est pas nécessairement critique. L'échantillonnage se poursuit simplement jusqu'à ce qu'un nombre approprié de nœuds soit accepté. Par contre, lors des études de populations humaines cachées, les tailles d'échantillon sont généralement faibles. Les membres de la population sont difficiles à atteindre et les interviews peuvent prendre beaucoup de temps. Toutefois, dans certaines études, la décision d'accepter ou de rejeter une unité d'après le degré sortant d'une personne sélectionnée provisoirement peut être prise assez rapidement au moyen d'une brève interview de filtrage. Il est malgré tout souhaitable de disposer d'une méthode d'échantillonnage dont le taux d'acceptation est aussi élevé que possible.

Les marches aléatoires sont caractérisées par une probabilité d'acceptation égale à un, mais n'ont généralement pas de probabilités limites connues ou contrôlées. Si l'on se représente la marche aléatoire sous-jacente comme le cheminement naturel, non contrôlé, au sein d'une population, alors on pourrait s'attendre à ce qu'une marche contrôlée ayant une loi limite proche de la marche aléatoire naturelle de la population produise un taux d'acceptation plus élevé qu'une marche contrôlée dont la loi limite diffère considérablement de cette marche aléatoire naturelle. Autrement dit, une marche contrôlée dont la loi stationnaire s'écarte peu de la loi de la marche aléatoire sous-jacente devrait nécessiter moins de modifications par rejet de nœuds sélectionnés provisoirement qu'une marche dont la loi stationnaire s'écarte considérablement de la marche aléatoire représentant la tendance naturelle.

Comme il est mentionné plus haut, les probabilités stationnaires d'une marche aléatoire ordinaire dans un graphe non orienté à une seule composante sont proportionnelles au degré des nœuds. Lorsqu'il existe plus d'une composante connectée, l'ajout du saut aléatoire est nécessaire pour assurer que chaque nœud puisse être atteint, pour produire une loi stationnaire unique ne dépendant pas de la loi initiale et pour faire en sorte que les probabilités limites soient influencées par le degré des nœuds, mais qu'elles n'y soient pas strictement proportionnelles. Même avec l'introduction du saut aléatoire et des probabilités d'acceptation induites, les marches ciblées produisant des probabilités stationnaires proportionnelles au degré de nœud pourraient s'approcher davantage de la loi naturelle de la marche aléatoire que les autres marches contrôlées considérées dans le présent article. En effet, si l'on examine la figure 5, il est évident que, pour la population empirique, la loi d'équilibre de la valeur de nœud espérée obtenue pour la marche selon le degré + 1 est plus proche de la loi d'équilibre de la marche aléatoire avec saut que de celle de tout autre plan contrôlé étudié.

Dans le cas de la population empirique tirée de l'étude sur la transmission hétérosexuelle du VIH/Sida, les taux d'acceptation obtenus pour les divers plans d'échantillonnage sont donnés au tableau 9. Pour le plan à marche uniforme, le taux d'acceptation est de 62 %. Pour la marche selon la valeur de nœud, où la probabilité limite est deux fois plus élevée pour les personnes à haut risque que pour celles à faible risque, le taux d'acceptation est de 60 %. Pour la marche selon le degré de nœud, dans laquelle la probabilité limite est proportionnelle au degré+1, il est de 85 %. Enfin, pour la marche selon le degré, avec une unité ajoutée uniquement au degré des nœuds isolés, il est de 88 %.

9. Discussion

Les plans d'échantillonnage à marche uniforme et à marche ciblée ont pour but de permettre de déterminer les probabilités limites de sélection d'après les données, afin de pouvoir les utiliser dans l'estimation. En outre, les probabilités limites sont choisies de sorte que certains types de nœuds ou de caractéristiques de graphe puissent être sélectionnés de manière préférentielle. La dépendance à l'égard de la sélection initiale, qui peut ne pas être contrôlée, diminue pas à pas.

Les estimateurs utilisés dans le présent article avec les plans d'échantillonnage à marche uniforme et à marche ciblée peuvent être considérés comme des estimateurs fondés sur le plan de sondage. Le plan exact fondé sur les probabilités de sélection pourrait ne pas être connu, si les probabilités de sélection initiales sont inconnues, mais on utilise les probabilités de sélection stationnaires dans les estimateurs. À mesure qu'augmente la longueur de la chaîne, ces probabilités deviennent plus exactes et l'espérance des estimations se rapproche de la quantité de graphe correspondante. L'avantage des méthodes d'estimation fondées sur le plan de sondage est que certaines de leurs propriétés, comme l'absence de biais ou la convergence par rapport au plan, ne dépendent pas d'hypothèses fondées sur un modèle qui pourraient être incorrectes. Les estimations fondées sur le plan de sondage ont la qualité intéressante supplémentaire d'être très simples et faciles à comprendre et à expliquer, et elles peuvent même produire des données qu'il est possible de présenter sans analyse ou interprétation comme étant représentatives de caractéristiques importantes de la population d'intérêt dans son ensemble.

L'utilisation d'algorithmes de Monte Carlo par chaîne de Markov pour l'analyse des données associées à des modèles compliqués est fréquente en statistique. Les approches décrites ici sont inhabituelles en ce sens que les méthodes par chaîne de Markov sont appliquées à des populations réelles pour obtenir effectivement des données, qui peuvent être facilement analysées manuellement. En fait, on pourrait aller une étape plus loin et construire un modèle de graphe stochastique bayésien complexe de la population en utilisant des méthodes de Monte Carlo par chaîne de Markov de la façon classique pour l'analyse des données, ainsi que pour leur collecte.

Les plans d'échantillonnage à marche uniforme ou ciblée sont utiles pour obtenir des échantillons de nœuds acceptés présentant certaines propriétés désirables en ce qui concerne la population, qui fournissent des estimateurs très simples des quantités de population ou qui pourraient fournir un échantillon initial pour un autre plan d'échantillonnage. Il convient de souligner que les nœuds qui ont été observés, puis « rejetés » sous les conditions du plan continuent de faire effectivement partie des données. Leur valeur peut encore être intégrée dans les estimations, au besoin, en appliquant la méthode de Rao-Blackwell, une fois que la chaîne a atteint approximativement l'équilibre, mais dans ces conditions, le calcul des estimations est complexe.

Une autre option consiste à utiliser des méthodes fondées sur un modèle, comme les méthodes d'estimation bayésiennes. En plus de la modélisation appropriée de la population par graphe stochastique, ces méthodes requièrent une procédure de sélection initiale ignorable, condition qui n'est généralement pas satisfaite sous sélections initiales biaisées par les valeurs ou les degrés de nœud, ou bien la modélisation adéquate de la procédure de sélection non ignorable dans les équations de vraisemblance. Les plans d'échantillonnage à marche ciblée produisant une loi asymptotique non corrélée à la procédure de sélection non ignorable et, donc, approximativement non corrélée aux valeurs ou aux degrés de nœud en dehors de l'échantillon pourraient fournir les sélections initiales pour un échantillon auquel les méthodes d'inférence basées sur un modèle pourraient ensuite être appliquées.

Remerciements

La présente étude a été financée par le National Center for Health Statistics, la National Science Foundation (DMS-9626102 et DMS-0406229) et les National Institutes of Health (R01-DA09872). Je tiens à remercier John Potterat et Steve Muth de m'avoir prodigué des conseils et permis d'utiliser les données provenant de l'étude de Colorado Springs.

Bibliographie

Birnbaum, Z.W., et Sirken, M.G. (1965). Design of Sample Surveys to Estimate the Prevalence of Rare Diseases: Three Unbiased Estimates. *Vital and Health Statistics*, Série 2, No.11. Washington: Government Printing Office.

Brin, S., et Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Proceedings of the 7th International World Wide Web Conference*, Elsevier, 107-117.

Chow, M., et Thompson, S.K. (2003). Estimation avec plans d'échantillonnage par dépistage de liens – Une approche bayésienne. *Techniques d'enquête*, 29, 221-230.

Felix-Medina, M.H., et Thompson, S.K. (2004). Combining link-tracing sampling and cluster sampling to estimate the size of hidden populations. *Journal of Official Statistics*, 20, 19-38.

Frank, O. (1977). Survey sampling in graphs. *Journal of Statistical Planning and Inference*, 1, 235-264.

Frank, O. (1978). Sampling and estimation in large social networks. *Social Networks*, 1, 91-101.

Frank, O., et Snijders, T. (1994). Estimating the size of hidden populations using snowball sampling. *Journal of Official Statistics*, 10, 53-67.

Hastings, W.K. (1970). Monte-Carlo sampling methods using Markov chains and their application. *Biometrika*, 57, 97-109.

Heckathorn, D.D. (1997). Respondent driven sampling: A new approach to the study of hidden populations. *Social Problems*, 44, 174-199.

Heckathorn, D.D. (2002). Respondent driven sampling II: Deriving valid population estimates from chain-referral samples of hidden populations. *Social Problems*, 49, 11-34.

Henzinger, M.R., Heydon, A., Mitzenmacher, M. et Najork, M. (2000). On near-uniform URL sampling. *Proceedings of the Ninth International World Wide Web Conference*, Elsevier, 295-308.

Klovdahl, A.S. (1989). Urban social networks: Some methodological problems and possibilities. Dans *The Small World*, (Éd. M. Kochen) Norwood, NJ: Ablex Publishing, 176-210.

Lovász, L. (1993). Random walks on graphs: A survey. Dans *Combinatorics*, *Paul Erdös is Eighty*, (Éds. D. Miklós, D. Sós et T. Szöni), János Bolyai Mathematical Society, Keszthely, Hungary, 2, 1-46.

Potterat, J.J., Woodhouse, D.E., Rothenberg, R.B., Muth, S.Q., Darrow, W.W., Muth, J.B. et Reynolds, J.U. (1993). AIDS in Colorado Springs: Is there an epidemic? *AIDS*, 7, 1517-1521.

Rothenberg, R.B., Woodhouse, D.E., Potterat, J.J., Muth, S.Q., Darrow, W.W. et Klovdahl, A.S. (1995). Social networks in disease transmission: The Colorado Springs study. Dans *Social Networks*, (Éds. R.H. Needle, S.G. Genser et R.T. Trotter) Drug Abuse, and HIV Transmission, NIDA Research Monograph 151. Rockville, MD: National Institute of Drug Abuse, 3-19.

Rubin, D.B. (1976). Inference and missing data. *Biometrika*, 63, 581-592.

Salganik, M.J., et Heckathorn, D.D. (2004). Sampling and Estimation in Hidden Populations Using Respondent-Driven Sampling. *Sociological Methodology*, 34, 193-239.

Spreen, M. (1992). Rare populations, hidden populations, and link-tracing designs: what and why? *Bulletin de Méthodologie Sociologique*, 36, 34-58.

Thompson, S.K., et Collins, L.M. (2002). Adaptive sampling in research on risk-related behaviors. *Drug and Alcohol Dependence*, 68, S57-S67.

Thompson, S.K., et Frank, O. (2000). Estimation fondée sur un modèle et comportant des plans d'échantillonnage à dépistage de liens. *Techniques d'enquête*, 26, 99-112.