

INFÉRENCE POUR L'ÉCHANTILLONNAGE LIEUX-MOMENTS

Lucie Léon¹, Marie Jauffret-Roustide² & Yann Le Strat³

¹*Institut de Veille Sanitaire - 12 rue du Val d'Osne - 94415 Saint-Maurice Cedex France*
l.leon@invs.sante.fr, ²*Cermes3 (Inserm U988/CNRS UMR 8236/EHESS/Université*
Paris Descartes) et Institut de Veille Sanitaire - 12 rue du Val d'Osne - 94415
Saint-Maurice Cedex France et m.jauffret@invs.sante.fr, ³*Institut de Veille Sanitaire -*
12 rue du Val d'Osne - 94415 Saint-Maurice Cedex France et y.lestrat@invs.sante.fr

Résumé. L'échantillonnage lieux-moments est une technique d'échantillonnage largement utilisée en épidémiologie auprès de populations particulièrement à risque d'être infectées notamment par des maladies infectieuses. Son principe consiste à échantillonner des lieux particuliers à des moments précis puis à échantillonner des personnes fréquentant ces lieux-moments. Cette technique d'échantillonnage soulève un point important dans le calcul des estimations : la fréquentation multiple et hétérogène des personnes dans les lieux. La méthode généralisée du partage des poids, peu utilisée dans le domaine de la santé, apporte une réponse adaptée. L'objectif de cette présentation est de formaliser l'échantillonnage lieux-moments dans le cadre d'un sondage indirect et de proposer un estimateur tenant compte du comportement hétérogène des personnes fréquentant les lieux d'enquêtes. Les propriétés de cet estimateur et celles de l'estimateur d'Horvitz-Thompson ont été évaluées à travers une étude de simulation et appliquées aux données de l'enquête ANRS-Coquelicot réalisée en 2011 dont l'objectif principal était d'estimer la prévalence du virus de l'hépatite C chez les usagers de drogues fréquentant des structures en France. Les résultats montrent que l'estimateur d'une prévalence ou d'un total, qui prend en compte les fréquentations, est sans biais et peut, même si des erreurs interviennent parfois dans la déclaration des fréquentations, demeurer sans biais alors qu'il est fortement biaisé lorsque cette fréquentation multiple est occultée. Ce biais est d'autant plus important quand cette fréquentation dépend du statut sérologique des personnes.

Mots-clés. Echantillonnage lieux-moments, Sondage indirect, Méthode généralisée du partage des poids, Populations difficiles d'accès.

1 Introduction

L'échantillonnage lieux-moments, aussi connu sous le nom anglophone de time-location sampling (TLS), est une technique d'échantillonnage largement utilisée en épidémiologie auprès de populations particulièrement à risque d'être infectées notamment par des maladies infectieuses (Stueve *et al.* (2000), Paquette et De Wit (2010)). Les usagers de

drogues, les hommes ayant des relations sexuelles avec des hommes, les travailleurs du sexe, les personnes sans domicile ou certains migrants sont des exemples de populations particulièrement exposées et vulnérables vis-à-vis de maladies transmissibles de personnes à personnes (VIH, IST, etc.). Or, réaliser des études séro-épidémiologiques dans ces populations s'avère difficile notamment en raison du caractère illicite de certaines pratiques (recours à la prostitution, utilisation de drogues), de la stigmatisation de certaines populations et de la difficulté à les contacter. Ces populations sont généralement accessibles dans les centres spécialisés qu'elles fréquentent ce qui rend difficile l'estimation sans biais d'indicateurs épidémiologiques tels que la prévalence notamment en raison de la grande mobilité des personnes dans les centres. Par exemple, les hommes ayant des relations sexuelles avec des hommes se rencontrent dans des lieux de convivialité gays (Wejnert *et al.* (2013)). Les usagers de drogues se rendent dans des lieux dédiés (consultations médicales, échanges de seringues, etc.) à certains moments de la journée (Jauffret-Roustide *et al.* (2009)) et Sutton *et al.* (2012)).

L'objectif de cette présentation est de formaliser l'échantillonnage lieux-moments dans le cadre d'un sondage indirect et de proposer un estimateur tenant compte du comportement hétérogène des personnes fréquentant les lieux d'enquêtes. Les propriétés de cet estimateur et celles de l'estimateur d'Horvitz-Thompson ont été évaluées à travers une étude de simulation et appliquées aux données de l'enquête ANRS-Coquelicot réalisée en 2011 dont l'objectif principal était d'estimer la prévalence du virus de l'hépatite C chez les usagers de drogues fréquentant des structures en France (Jauffret-Roustide *et al.* (2013)).

2 Méthode

2.1 Enquête ANRS-Coquelicot 2011

En 2011, l'enquête ANRS-Coquelicot a été lancée pour suivre la dynamique de l'épidémie du virus de l'hépatite C chez les usagers de drogues fréquentant des structures en France. C'est la seconde édition, une première enquête ayant été réalisée en 2004 (Jauffret-Roustide *et al.* (2009)), avec des prélèvements biologiques sur buvard en plus d'un questionnaire long en face-à-face. Elle a été menée auprès de 1568 usagers de drogues ayant sniffé ou injecté au moins une fois dans leur vie et fréquentant les 121 centres recensés dans 5 agglomérations françaises (Lille, Strasbourg, Paris, Marseille et Bordeaux) et deux départements (Seine-et-Marne et Seine-Saint-Denis) durant 11 semaines. Lors des interviews, deux questions sur les fréquentations ont été posées: 1/ Hier et les 3 jours précédents, avez-vous fréquenté un ou plusieurs centres spécialisés pour les usagers de drogues? et si oui, lesquels et combien de fois? 2/ aujourd'hui, en comptant l'endroit où nous sommes, quel(s) centre(s) avez-vous fréquenté ou allez-vous fréquenté? En effet, un usager de drogues pouvait bénéficier de différentes prestations (une consultation médicale, un hébergement post-cure, un échange de seringue) proposées dans les centres au cours

de l'enquête.

2.2 Estimation

Le principe du TLS pour atteindre ces populations consiste à échantillonner des lieux particuliers à des moments précis puis à échantillonner des personnes fréquentant ces lieux-moments. Ces lieux et ces moments dépendent de la population d'intérêt. Cette technique d'échantillonnage soulève un point important dans le calcul des estimations : la fréquentation multiple et hétérogène des personnes dans les lieux (Gustafson *et al.* (2013), Karon (2012)).

On s'intéresse donc à une population d'individus fréquentant des centres spécialisés durant les heures d'ouverture. Ces centres proposent différentes prestations en fonction de la population ciblée: une consultation médicale, un hébergement, un repas, un échange de matériel ou toute autre prestation. Pour simplifier mais sans perte de généralité, on suppose que l'unité de temps d'ouverture d'un centre est la demi-journée. Quelle que soit la population d'intérêt, les lieux d'enquêtes, leur nombre et l'unité de temps, un TLS peut être vu comme un sondage aléatoire à 3 degrés où, les unités primaires sont les centres spécialisés, les unités secondaires sont les demi-journées d'ouverture et les unités tertiaires sont les individus. En pratique, la construction d'un calendrier de visites est souvent nécessaire en amont de l'enquête en listant l'ensemble des centres recevant la population d'intérêt, leurs files actives et l'ensemble des demi-journées d'ouverture des centres pour la période d'enquête.

Un même individu peut fréquenter plusieurs fois un ou plusieurs centres au cours de l'enquête et donc prétendre à plusieurs prestations offertes dans ces centres. Ainsi, un individu bénéficiant de plusieurs prestations dans différents centres a plus de chance d'être tiré au sort qu'un individu bénéficiant d'une seule prestation dans un centre particulier. La méthode généralisée du partage des poids (MGPP), peu utilisée dans le domaine de la santé, apporte une réponse adaptée à la prise en compte de la fréquentation multiple des individus enquêtés (Lavallée (1995), Lavallée (2007)). La fréquentation d'un individu est généralement collectée lors de l'entretien en face-à-face avec l'individu enquêté.

Un TLS peut alors être défini comme un sondage indirect à 3 degrés où, au troisième degré, les individus sont associés aux prestations dont ils bénéficient dans les centres qu'ils fréquentent. Par définition, l'approche par sondage indirect consiste à utiliser la base de sondage d'une population A disponible et reliée indirectement aux unités d'intérêt, afin de produire une estimation dans la population d'intérêt B en s'appuyant sur les liens qui existent entre les unités de la population A et les unités de la population B . Ici, la population A correspond à l'ensemble des prestations proposées dans les centres spécialisés recevant la population d'intérêt B et les fréquentations pour bénéficier de ces prestations définissent les liens. La MGPP fournit alors pour chaque individu enquêté

un poids de sondage pondéré par sa fréquentation au cours de l'enquête et permet la construction d'un estimateur tenant compte du comportement hétérogène des personnes fréquentant les lieux d'enquêtes. Ainsi, l'estimation d'une prévalence selon le nouvel estimateur s'écrit: $\hat{P} = \frac{\sum_i \tilde{w}_i y_i}{\sum_i \tilde{w}_i}$ où y_i est la variable d'intérêt correspondant au statut

sérologique de l'individu i ($y_i = 1$ si i est infecté et 0 sinon), $\tilde{w}_i = \frac{1}{L_i} \sum_i w_i$, w_i est son poids de sondage et L_i est son nombre total de fréquentations au cours de l'enquête. Sa variance est donnée par :

$$\widehat{Var}(\hat{P}) = \frac{1}{\hat{N}^2} \left[\widehat{Var}(\hat{T}) - 2\hat{P}\widehat{Cov}(\hat{T}, \hat{N}) + \hat{P}^2\widehat{Var}(\hat{N}) \right] \text{ où } \hat{T} = \sum_i \tilde{w}_i y_i \text{ est le nombre estimé d'individus infectés, et } \hat{N} = \sum_i \tilde{w}_i \text{ est la taille estimée de la population d'intérêt.}$$

2.3 Etude de simulation

Afin d'évaluer les propriétés de cet estimateur et de les comparer à celles de l'estimateur d'Horvitz-Thompson, une étude de simulation a été réalisée en générant plusieurs scénarios. Pour cela, des populations d'individus infectés ou non et se rendant dans des lieux d'enquêtes pour bénéficier de prestations ont été générées en faisant varier la prévalence de 1% à 90% et le nombre de fréquentations suivant une distribution binomiale négative de moyenne μ et de variance $\theta\mu$ où $\theta \geq 1$. Pour chaque population générée, 10000 échantillons ont été tirés au sort selon un TLS. Puis, pour chacun des échantillons, la prévalence et la taille de la population ont été calculées selon les deux estimateurs. Pour aller plus loin, des erreurs potentielles sur la déclaration des fréquentations ont été générées puis les estimations de la prévalence et de la taille de la population ont été à nouveau calculées à partir des 10000 échantillons. En effet, on sait qu'en pratique, il est souvent difficile de se rappeler de toutes les visites surtout si elles sont nombreuses et variées et la consommation de produits (utilisation de drogues) peut rendre encore plus difficile cette mémorisation.

3 Résultats

3.1 Enquête ANRS-Coquelicot 2011

La taille de la population des usagers de drogues fréquentant des centres dédiés a été estimée à 43710 individus, $IC95\% = [39667; 47753]$ selon l'estimateur nouvellement construit et à 48147 individus, $IC95\% = [43741; 52553]$ selon l'estimateur d'Horvitz-Thompson. Dans cette population, la prévalence du virus de l'hépatite C a été estimée à 43.7%, $IC95\% = [39.5\%, 47.9\%]$ selon l'estimateur nouvellement construit et à 43.4%, $IC95\% =$

[39.3%, 47.6%] selon l'estimateur d'Horvitz-Thompson. Ces deux prévalences estimées ont des valeurs proches, certainement dues à une variance faible du nombre de fréquentations déclarées par les usagers de drogues en 2011.

3.2 Etude de simulation

Les résultats ont montré que l'estimateur d'une prévalence ou d'un total, qui prenait en compte les fréquentations, était sans biais quel que soit le scénario. Au contraire, selon l'estimateur d'Horvitz-Thompson, les prévalences estimées pouvaient être fortement biaisées surtout lorsque la fréquentation des lieux d'enquêtes dépendait du statut sérologique des personnes, le biais relatif associé à ces estimations variant de 0% à 222% par rapport à la vraie prévalence. De plus, même si des erreurs intervenaient parfois dans la déclaration des fréquentations, un biais faible voire une absence de biais a été observé pour les estimations obtenues à partir de l'estimateur nouvellement construit contrairement à celles obtenues à partir de l'estimateur d'Horvitz-Thompson, et cela quel que soit le scénario.

4 Conclusion

Un échantillonnage lieux-moments ou TLS peut être défini comme un sondage indirect, et dans ce cadre, l'utilisation d'un estimateur tenant compte de la fréquentation multiple des lieux d'enquêtes doit être privilégiée pour l'estimation d'indicateurs épidémiologiques. En effet, cet estimateur est sans biais et peut le demeurer malgré des erreurs potentielles sur le nombre de fréquentations déclarées. L'étude de simulation a montré que le biais observé de l'estimateur occultant cette fréquentation dépendait de la prévalence et de la distribution du nombre de fréquentations en termes de valeur, de variance et de qualité. La fréquentation des individus est généralement obtenue à partir des informations collectées lors des entretiens en face-à-face au cours de l'enquête.

Bibliographie

- [1] Gustafson, P. (2013), Impact of statistical adjustment for frequency of venue attendance in a venue-based survey of men who have sex with men, *American Journal of Epidemiology*, 177(10), 1157-1164.
- [2] Jauffret-Roustide, M. and Le Strat, Y. and Couturier, E. and Thierry, D. and Rondy, M. and Quaglia, M. and Razafandratsima, N. and Emmanuelli, J. and Guibert, G. and Barin, F. and Desenclos, J. C. (2009), A national cross-sectional study among drug-users in France: epidemiology of HCV and highlight on practical and statistical aspects of the design, *BMC Infectious Diseases*, 9, Pages 113.

- [3] Jauffret-Roustide, M. and Pillonel, J. and Weill-Barillet, L. and Léon, L. and Le Strat, Y. and Brunet, S. and Benoit, T. and Chauvin, C. and Lebreton, M. and Barin, F. and Semaille, C. (2013), Estimation de la séroprévalence du VIH et de l'hépatite C chez les usagers de drogues en France - premiers résultats de l'enquête ANRS-COQUELICOT 2011, *Bulletin Epidémiologique Hebdomadaire*, 39-40, 504-509.
- [4] Karon, J. M. and Wejnert, C. (2012), Statistical methods for the analysis of time-location sampling data, *Journal of Urban Health*, 89,565-586.
- [5] Lavallée, P. (2007), *Indirect sampling*, Springer, New York.
- [6] Lavallée, P. (1995), Cross-sectional weighting of longitudinal surveys of individuals and households using the weight share method, *Survey Methodology*, 21, 25–32.
- [7] Paquette, D. et De Wit, J. (2010), Sampling methods used in developed countries for behavioural surveillance among men who have sex with men, *AIDS and Behavior*, 14, 1252-1264.
- [8] Stueve, A. et O'Donnell, L. N. et Duran, R. et Doval, A. S. et Blome, J.(2001), Time-space sampling in minority communities: results with young latino men who have sex with men, *American Journal of Public Health*, 91, 922-926.
- [9] Sutton, A. J. and McDonald, S. A. and Palmateer, N. and Taylor, A. and Hutchinson, S. J. (2012), Estimating the variability in the risk of infection for hepatitis C in the Glasgow injecting drug user population, *Epidemiology and Infection*, 140, 2190-2198.
- [10] Wejnert, C. and Le, B. and Rose, C. E. and Oster, A. M. and Smith, A. J. and Zhu, J. and Paz-Bailey, G. for the NHBS study group (2013), HIV Infection and Awareness among men who have sex with men-20 cities, United States, 2008 and 2011, 8 (Issue 10), *Plosone* 8.