Estimation des poids dans l'approche contrefactuelle de la causalité en épidémiologie

François Séverac[1], Van Huyen Do[1], Frédéric Bertrand[2], Erik-A. Sauleau[1]

1 Laboratoire ICube, UMR 7357, Faculté de Médecine, 4, rue Kirschleger, 67085 Strasbourg Cédex.

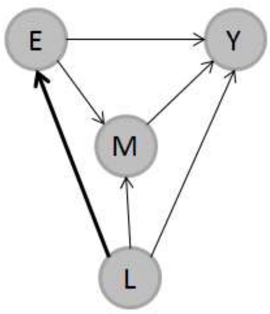
2 Institut de Recherche Mathématique Avancée, Université de Strasbourg, 7, rue René Descartes, 67000 Strasbourg.

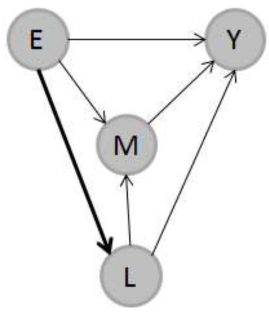
La recherche dans le domaine de l'épidémiologie est basée essentiellement sur les études observationnelles dans lesquelles la comparabilité des groupes n'est pas assurée. Pour pallier cette non-comparabilité, l'approche contrefactuelle permet, sous certaines hypothèses, d'estimer la distribution du résultat (Y), conditionnellement à l'exposition (E), de façon non biaisée. La détermination de poids permet la création d'une pseudo-population homogène en termes de facteurs de confusion.

Dans le cas simple où on veut estimer l'effet causal de E sur Y, en présence d'un facteur de confusion multinomial (L), sa modélisation dans la pseudopopulation permet son estimation sans biais, sous réserve que les poids utilisés respectent la relation : $\tilde{P}(Y = y|E = e) = \sum_m P(Y = y|L = l, E = e)\, P(L = l)$, $\tilde{P}(.)$ étant la probabilité dans la pseudopopulation et $P(.)$ la probabilité observée. Dans ce cas simple, les poids généralement utilisés sont $\frac{1}{P(E=e|L=l)}$ et $\frac{P(E=e)}{P(E=e|L=l)}$ (poids standardisés).

Théoriquement un résultat de Pearl [Pearl J. Causal inference in statistics: an overview. *Statistics surveys* 2009, 3:96-143.] permet l'extension de ce cas à toutes les situations. Dans un graph markovien, si X est l'ensemble des variables d'exposition d'intérêt, pour chaque autre variable (variable résultat, confusion, médiateur, variable d'ajustement, …) v_i, dont les parents sont l'ensemble des variables pa_i, on a $P\big(v_1, v_2, \dots, v_K|do(X = x)\big) = \prod_{v_i \notin X} P(v_i|pa_i)|_{X=x}$.Ici la notation $do(X = x)$ est un opérateur simulant une intervention sur X, en forçant leur valeur à x (même si ce n'est pas ce qui a été observé).

Nous avons appliqué cette formule au cas d'un médiateur et d'un facteur de confusion.





Les poids sont alors respectivement $\frac{1}{P(M=m|E=e,L=l)P(E=e|L=l)}$ et $\frac{1}{P(M=m|E=e,L=l)}$. Ces poids peuvent être estimés à partir des données en modélisant les probabilités ci-dessus.

Dans cette communication, nous montrerons sur des simulations que ces poids permettent d'estimer sans biais l'effet causal de E sur Y.

Pour des situations plus complexes, le résultat théorique de Pearl devrait pouvoir permettre de formuler une expression pour les poids. Ces travaux sont actuellement en cours.