Santin G, Soullier N, Guéguen A

Traitement des données manquantes pour l'estimation de prévalences

Atelier de formation - 17ème colloque de l'ADEREST, 24-25 novembre 2016, Brest

RESUME

CONTEXTE - En santé comme dans d'autres domaines, les taux de réponse aux enquêtes diminuent de plus en plus. La non-réponse est souvent problématique car elle peut entrainer des biais ; elle a également des conséquences délétères sur la variance des estimateurs. L'étude des biais de non-réponse, n'est pas nouvelle en épidémiologie : plusieurs études ont déjà montré que la participation à une enquête épidémiologique est liée à l'âge, à la catégorie sociale, à l'état de santé de la personne et aux comportements à risque pour la santé tels que la consommation d'alcool et de tabac. Nous nous plaçons ici dans le cadre d'enquêtes de santé publique ou de surveillance épidémiologique ayant pour objectif de produire des estimations de prévalence extrapolables à une population d'intérêt. On distingue la non-réponse totale qui survient lorsque la personne enquêtée ne répond à aucune question de l'enquête à de la non-réponse partielle, appelée plus communément donnée manquante, qui est rencontrée si la personne enquêtée répond à certaines questions de l'enquête, mais pas à toutes. Le biais de non-réponse dépend de la variable d'intérêt, puisqu'il est fonction du produit de l'inverse de la probabilité de réponse moyenne et de la covariance entre la probabilité de réponse et la variable d'intérêt. Autrement dit, il y aura absence de biais si la probabilité de réponse est égale à 1 (tout le monde répond), ou si la covariance entre la probabilité de réponse et la variable d'intérêt est nulle. Les non-réponses peuvent être classées selon trois types : non-réponse complètement aléatoire (Missing Completely At Random ou MCAR), non-réponse aléatoire (Missing At Random ou MAR), non-réponse non aléatoire (Missing Not At Random ou MNAR). Dans le cas MCAR, il y a indépendance entre la probabilité de réponse et la variable d'intérêt. La prévalence estimée de la variable d'intérêt est sans biais. Dans le cas MAR, il y a indépendance entre la probabilité de réponse et la variable d'intérêt conditionnellement à d'autres variables notées X. Autrement dit, la probabilité de réponse et la variable d'intérêt partagent des causes communes X. Après prise en compte de ces causes communes par des méthodes appropriées, la prévalence estimée de la variable d'intérêt est sans biais. Quelle que soit la technique utilisée, la correction est possible à condition que l'ensemble des causes communes X soient disponibles pour l'ensemble des répondants et des non-répondants. Dans le cas MNAR, il n'y a pas indépendance entre la probabilité de réponse et la variable d'intérêt. Elle peut résulter soit d'un lien direct entre la variable d'intérêt et la probabilité de réponse, soit lorsqu'il n'a pas été possible de prendre en compte l'ensemble des causes communes X. Dans ce cas, la prévalence estimée de la variable d'intérêt ne peut être sans biais. Les méthodes qui permettent de prendre en compte les causes communes X dans le cas MAR seront présentées : la repondération (préférentiellement utilisée pour traiter la non-réponse totale) et l'imputation (préférentiellement utilisée pour traiter les données manquantes). CORRECTION DE LA NON REPONSE TOTALE PAR REPONDERATION : L'EXEMPLE DE LA PHASE PILOTE DE LA COHORTE COSET-RSI - La non réponse totale traitée par repondération est présentée ici et sera illustrée à partir des données de la phase pilote de la cohorte Coset-RSI, cohorte pour la surveillance épidémiologique en lien avec le travail auprès d'actifs relevant du Régime Social des Indépendants au moment de l'inclusion. La repondération pour correction de la non-réponse consiste à augmenter les poids de sondage des répondants afin de compenser l'absence de réponse des non-répondants. La première étape est d'expliquer le mécanisme de non-réponse en le modélisant à partir des informations qui sont disponibles à la fois pour les répondants et pour les non-répondants. Cela suppose d'avoir de telles informations, dites variables auxiliaires, qui, par définition, ne sont pas issues du questionnaire. Parmi les variables auxiliaires disponibles, on sélectionne les variables qui expliquent la non-réponse et qui sont liées aux variables d'intérêt évaluées par le questionnaire. La cohorte Coset-RSI a la particularité de disposer de nombreuses variables auxiliaires. En effet, outre les données de la base de sondage qui sont classiquement utilisées, les données de l'assurance maladie (SNIIRAM) sont

renseignées pour tous les individus sélectionnés pour être enquêtés. Pour être utilisées, les données du SNIIRAM sont synthétisées sous la forme d'indicateurs disponibles à la fois pour les répondants et les non-répondants. La non-réponse est ensuite modélisée selon ces indicateurs via une régression logistique, qui permet d'obtenir pour chaque individu une probabilité de réponse prédite. La deuxième étape est la constitution de groupes de réponse homogène. Pour les constituer, la méthode des scores est utilisée : les groupes correspondent aux quantiles de la distribution des probabilités de réponse prédites. Dans chaque groupe ainsi constitué, on calcule ensuite le taux de réponse observé. L'inverse de ce taux de réponse est appliqué à tous les répondants du groupe, en multipliant leur poids de sondage initial par ce facteur correctif. Chaque étape de la méthode de repondération sera illustrée à partir des données de la phase pilote de la cohorte Coset-RSI. L'impact de la correction de la non-réponse sera observé sur des prévalences mesurées dans l'enquête telles que le tabagisme, l'obésité ou la symptomatologie dépressive. TRAITEMENT DE LA NON REPONSE PARTIELLE PAR IMPUTATION : L'exemple DE LA COHORTE CONSTANCES - La non-réponse partielle traitée par imputation est présentée ici et sera illustrée à partir des données d'inclusion de la cohorte Constances, cohorte épidémiologique « généraliste » constituée à terme d'un échantillon de 200 000 adultes âgés de 18 à 69 ans à l'inclusion et affiliés au Régime général de la sécurité sociale. L'imputation consiste à remplacer les valeurs manquantes par des valeurs plausibles. Les méthodes d'imputation reposent sur les mêmes hypothèses que les méthodes de pondération, à savoir que les données sont manquantes au hasard, c'est-à-dire manquantes aléatoirement conditionnellement aux données observées. Les méthodes d'imputation sont nombreuses : on distingue les méthodes déterministes des méthodes probabilistes et les imputations simples des imputations multiples. Nous présenterons deux méthodes d'imputation qui peuvent être utilisées si l'objectif est d'estimer la prévalence d'une variable d'intérêt et que toutes les variables auxiliaires sont non manquantes. La méthode « hot deck par donneur » consiste à remplacer les valeurs manquantes d'un sujet « receveur » par celle d'un sujet « donneur » ne présentant pas de donnée manquante et ayant les mêmes caractéristiques (i.e les mêmes variables auxiliaires) que celles du sujet « receveur ». L'imputation par modèle de régression consiste à remplacer la valeur manquante d'un sujet par sa prédiction issue d'un modèle de régression expliquant la variable d'intérêt par les variables auxiliaires. Les prévalences sur données complètes et sur données imputées seront comparées pour la symptomatologie dépressive (mesurée par le CESD) et pour le questionnaire de Siegrist.

MOTS CLES : -

INFORMATIONS COMPLEMENTAIRES, ICI