Non-réponse totale dans les enquêtes de surveillance épidémiologique

Gaëlle Santin

UNIVERSITÉ PARIS-SUD

ÉCOLE DOCTORALE 420 :
SANTÉ PUBLIQUE PARIS SUD 11, PARIS DESCARTES

Laboratoire : *DST-InVS*

THÈSE DE DOCTORAT

SANTÉ PUBLIQUE - EPIDÉMIOLOGIE

par

Gaëlle SANTIN

Non-réponse totale dans les enquêtes de surveillance épidémiologique

Date de soutenance : 09/02/2015

Composition du jury :

Directeur de thèse :	Jean BOUYER	Directeur de recherche (Inserm)
Co-directeur de thèse :	Alice GUEGUEN	Ingénieur de recherche (Inserm)
Rapporteurs :	Camelia GOGA	Maître de conférences (Université de Bourgogne)
	Fred PACCAUD	Directeur (IUMSP)
Examinateurs :	François BECK	Directeur (OFDT)
	David HAZIZA	Professeur agrégé (Université de Montréal
Présidente :	Laurence MEYER	PUPH (Inserm)

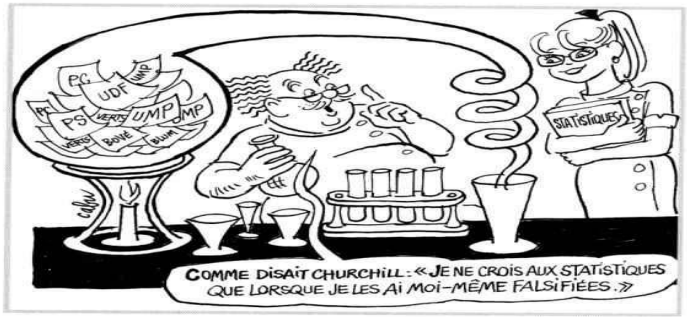
A Charlie Hebdo, Cabu, Charb, Honoré, Oncle Bernard et Wolinski

Pour leur longue expérience …

…des questionnaires…

…des résultats issus d'enquêtes…

… de la limite des statistiques…



« Jamais la litanie des chiffres n'a été aussi pesante, dévidée jour après jour. Sondage contre sondage, moyenne contre moyenne. Ubiquité du chiffre, qui ne signifie plus, qui efface tout et qui disparaît à son tour... »
Bernard Maris

… et puis pour tout le reste (surtout pour tout le reste)

REMERCIEMENTS

Cette thèse est loin d'être le résultat d'un travail solitaire ; c'est avant tout le fruit de collaborations et d'échanges qui m'ont énormément apportés, tant d'un point de vue professionnel que personnel. C'est pourquoi je tiens à remercier chaleureusement toutes les personnes qui, par leur soutien, leur amitié, leurs conseils et leur travail m'ont permis de mener à bien ce projet.

Je remercie très sincèrement Camelia Goga et Fred Paccaud d'avoir accepté d'être rapporteurs de ce travail, Laurence Meyer de me faire l'honneur de présider mon jury de thèse, François Beck et David Haziza d'avoir accepté d'en être les examinateurs. Je les remercie tous vivement d'avoir accepté d'évaluer ce travail.

Je remercie infiniment Alice Guéguen et Jean Bouyer d'avoir accepté de diriger cette thèse pas très bien définie initialement et de m'avoir autant apporté pendant ces quatre années et demie ! C'était un vrai bonheur pour moi et un réel privilège de les avoir comme directeurs de thèse. Je les remercie pour la liberté qui m'ont laissé, leurs conseils, la qualité des réunions dans le bureau de Jean (qui vont vraiment me manquer), leur disponibilité, leur bienveillance et leur complémentarité.

Je remercie vivement Ellen Imbernon qui a accepté et soutenu mon projet de thèse au sein du Département Santé Travail de l'Institut de veille sanitaire. Je la remercie pour tout ce qu'elle m'a appris et apporté durant ma décennie passée au DST.

Je remercie chaleureusement l'équipe Coset : Béatrice Geoffroy, Laetitia Bénézet, Juliette Chatelot et Pauline Delézire. Sans elles et leur implication dans le pilote Coset-MSA, cette thèse n'aurait pas pu exister. Je remercie particulièrement Béatrice Geoffroy de m'avoir fait confiance sur les questions méthodologiques relatives aux cohortes ayant des objectifs descriptifs (non je ne dirai pas le mot qui (me) fâche). Mettre en place les cohortes pilote Coset-MSA et Coset-RSI a été un vrai défi et nous y sommes arrivées ! C'est une vraie joie de voir que Coset-MSA va connaitre son extension nationale cette année, et j'espère de tout cœur que Coset-RSI aura ce même avenir prochainement. Je les remercie pour leur amitié, qui est le fruit de tout ce que nous avons partagé et échangé pendant toutes ces années.

Je remercie amicalement toutes les personnes que j'ai côtoyées au quotidien au Département santé travail au cours de ces dix dernières années ; elles m'ont permis d'apprendre énormément en épidémiologie des risques professionnels, de partager souvent de très bons moments mais également de rencontrer de vrai(e)s ami(e)s.

Je remercie chaleureusement mes nouveaux collègues de l'UMS011 auprès de qui une nouvelle aventure commence pour moi. Je remercie particulièrement Marie Zins pour m'avoir laissé terminer sereinement ce travail et pour ses encouragements, Marcel Goldberg pour ses relectures attentives des articles soumis, Diane Cyr pour ses relectures de mon anglais bien français et Eléonore Herquelot pour sa relecture des chapitres méthodos et sa solidarité de thésarde en fin de thèse comme moi !

Je remercie la MSA sans qui la mise en place de la cohorte Coset-MSA n'aurait pas été possible ; je remercie particulièrement Alain Pelc, Joël Roy et Nicolas Viarouge.

Je remercie Olivier Sautory de l'Insee d'avoir mis à ma disposition la macro Calker, ce qui m'a dispensé de développer un programme d'estimation de variance laborieux mais essentiel !

Je remercie également l'ED420 pour la qualité des enseignements dont j'ai pu bénéficier ; je remercie particulièrement Audrey Bourgeois pour son professionnalisme et sa gentillesse.

Je remercie toutes celles et ceux qui m'ont donné envie de me lancer tardivement dans une thèse, par ce qu'ils m'ont apporté en épidémiologie. Je remercie particulièrement Maria Martinez qui m'a donné le virus de l'épidémiologie et qui m'a réellement appris à travailler, Claire Julian-Reynier qui m'a permis de découvrir de nouveaux champs de l'épidémiologie et Christine Cohidon, pour tous nos échanges fructueux, pour la confiance qu'elle m'a accordée (sans laquelle je ne me serai probablement jamais sentie capable de mener à bien une thèse) et pour son soutien sans faille, jusqu'aux dernières lignes de ce manuscrit.

Je remercie mes parents pour m'avoir toujours encouragée et soutenue dans mes choix étudiants et professionnels.

Enfin, je remercie mes amies et mes amis qui m'ont supporté dans tous les sens du terme pendant quatre ans et demi. Sans eux, je ne serais probablement jamais venue au bout de cette aventure. J'espère néanmoins que les problèmes de non-réponse totale dans les enquêtes de surveillance épidémiologique ne leur manqueront pas trop !

RÉSUMÉ

La non-réponse, rencontrée dans la plupart des enquêtes épidémiologiques, est génératrice de biais de sélection (qui, dans ce cas est un biais de non-réponse) lorsqu'elle est liée aux variables d'intérêt. En surveillance épidémiologique, dont un des objectifs est d'estimer des prévalences, on a souvent recours à des enquêtes par sondage. On est alors confronté à la non-réponse totale et on peut utiliser des méthodes issues de la statistique d'enquête pour la corriger. Le biais de non-réponse peut être exprimé comme le produit de l'inverse du taux de réponse et de la covariance entre la probabilité de réponse et la variable d'intérêt. Ainsi, deux types de solution peuvent généralement être envisagés pour diminuer ce biais. La première consiste à chercher à augmenter le taux de réponse au moment de la planification de l'enquête. Cependant, la maximisation du taux de réponse peut entraîner d'autres types de biais, comme des biais de mesure. Dans la seconde, après avoir recueilli les données, on utilise des informations liées *a priori* aux variables d'intérêt et à la probabilité de réponse, et disponibles à la fois pour les répondants et les non-répondants pour calculer des facteurs correctifs. Cette solution nécessite donc de disposer d'informations sur l'ensemble de l'échantillon tiré au sort (que les personnes aient répondu ou non) ; or ces informations sont en général peu nombreuses. Les possibilités récentes d'accès aux bases médico-administratives (notamment celles de l'assurance maladie) ouvrent de nouvelles perspectives sur cet aspect.
Les objectifs de ce travail, qui sont centrés sur les biais de non-réponse, étaient d'étudier l'apport de données supplémentaires (enquête complémentaire auprès de non-répondants et bases médico-administratives) et de discuter l'influence du taux de réponse sur l'erreur de non-réponse et l'erreur de mesure.
L'analyse était centrée sur la surveillance épidémiologique des risques professionnels via l'exploitation des données de la phase pilote de la cohorte Coset-MSA à l'inclusion. Dans cette enquête, en plus des données recueillies par questionnaire (enquête initiale et enquête complémentaire auprès de non-répondants), des informations auxiliaires issues de bases médico-administratives (SNIIR-AM et MSA) étaient disponibles pour les répondants mais aussi pour les non-répondants à l'enquête par questionnaire.
Les résultats montrent que les données de l'enquête initiale, qui présentait un taux de réponse de 24%, corrigées pour la non-réponse avec des informations auxiliaires directement liées à la thématique de l'enquête (la santé et le travail) fournissent des estimations de prévalence en général proches de celles obtenues grâce à la combinaison des données de l'enquête initiale et de l'enquête complémentaire (dont le taux de réponse atteignait 63%) après correction de la non réponse par ces mêmes informations auxiliaires. La recherche d'un taux de réponse maximal à l'aide d'une enquête complémentaire n'apparait donc pas nécessaire pour diminuer le biais de non réponse. Cette étude a néanmoins mis en avant l'existence de potentiels biais de mesure plus importants pour l'enquête initiale que pour l'enquête complémentaire. L'étude spécifique du compromis entre erreur de non-réponse et erreur de mesure montre que, pour les variables qui ont pu être étudiées, après correction de la non-réponse, la somme de l'erreur de non-réponse de l'erreur de mesure est équivalente dans l'enquête initiale et dans les enquêtes combinées (enquête initiale et complémentaire).
Ce travail a montré l'intérêt des bases médico-administratives pour diminuer l'erreur de non-réponse et étudier les erreurs de mesure dans une enquête de surveillance épidémiologique.

Mots-clés : Surveillance, non-réponse totale, erreur de non-réponse, erreur de mesure, erreur totale, bases médico-administratives, repondération, biais

Thèse préparée au :
Département santé travail de l'Institut de veille sanitaire
12 rue du val d'Osne
94415 Saint-Maurice Cedex

ABSTRACT

Nonresponse occurs in most epidemiologic surveys and may generate selection bias (which is, in this case, a nonresponse bias) when it is linked to outcome variables. In epidemiologic surveillance, whose one of the purpose is to estimate prevalences, it is usual to use survey sampling. In this case, unit nonresponse occurs and it is possible to use methods coming from survey sampling to correct for nonresponse. Nonresponse bias can be expressed as the product of the inverse of the response rate and the covariance between the probability of response and the outcome variable. Thus, two options are available to reduce the effect of nonresponse. The first is to increase the response rate by developing appropriate strategies at the study design phase. However, the maximization of the response rate can prompt other kinds of bias, such as measurement bias. In the second option, after data collection, information associated with both nonresponse and the outcome variable, and available for both respondents and nonrespondents, can be used to calculate corrective factors. This solution requires having information on the complete random sample (respondents and nonrespondents); but this information is rarely sufficient. Recent possibilities to access administrative databases (particularly those pertaining to health insurance) offer new perspectives on this aspect.

The objectives of this work focused on the nonresponse bias were to study the contribution of supplementary data (administrative databases and complementary survey among nonrespondents) and to discuss the influence of the response rate on the nonresponse error and the measurement error.

The analyses focused on occupational health epidemiologic surveillance, using data (at inclusion) from the Coset-MSA cohort pilot study. In this study, in addition to the data collected by questionnaire (initial and complementary survey among nonrespondents), auxiliary information from health and occupational administrative databases was available for both respondents and nonrespondents.

Results show that the data from the initial survey (response rate : 24%), corrected for nonresponse with information directly linked to the study subject (health and work) produce estimations of prevalence close to those obtained by combining data from the initial survey and the complementary survey (response rate : 63%), after nonresponse adjustment on the same auxiliary information. Using a complementary survey to attain a maximal response rate does not seem to be necessary in order to decrease nonresponse bias. Nevertheless, this study highlights potential measurement bias which could be more consequential for the initial survey than for the complementary survey. The specific study of the trade-off between nonresponse error and measurement error shows that, for the studied variables and after correction for nonresponse, the sum of the nonresponse error and the measurement error is equivalent in the initial survey and in the combined surveys (initial plus complementary survey).

This work illustrated the potential of administrative databases for decreasing the nonresponse error and for evaluating measurement error in an epidemiologic surveillance survey.

Keywords: Surveillance, unit nonresponse, nonresponse error, measurement error, total error, administrative databases, reweighting, bias

PRODUCTION SCIENTIFIQUE

Article publié dans une revue internationale à comité de lecture

Santin G., Geoffroy B., Bénézet L., Delézire P., Chatelot J., Sitta R., Bouyer J, Gueguen A. SNIIR-AM Cohorts Group. In an occupational health surveillance study, auxiliary data from administrative health and occupational databases effectively corrected for nonresponse. JCE. 67 (2014) 722-730

Article soumis dans une revue internationale à comité de lecture

Santin G., Bénézet L., Geoffroy B., Bouyer J., Gueguen A., Coset team. A complementary survey and its paradata contributed to take into account nonresponse bias in an occupational health surveillance survey.

Article en préparation

Santin G., Delézire P., Geoffroy B., Bénézet L., Bouyer J, Gueguen A. Compromis entre erreur de non-réponse et erreur de mesure dans Coset-MSA.

Autre publication avec comité de lecture

Geoffroy-Perez B, Chatelot J, Santin G, Bénézet L, Delézire P, Imbernon E. 2012. Coset : un nouvel outil généraliste pour la surveillance épidémiologique des risques professionnels. Bull Epidémiol Hebd. 22-23:276-77

Autres publications

Rapport

Geoffroy-Perez B., Bénézet L., Santin G., Delabre L., Delézire P., Chatelot J. 2012. Programme Coset : Cohortes pour la surveillance épidémiologique en lien avec le travail. Premier bilan de la phase pilote pour la mise en place de la cohorte d'actifs relevant du régime agricole au moment de l'inclusion - cohorte Coset-MSA. Saint-Maurice : Institut de veille sanitaire.

Autres articles

Geoffroy-Perez B., Santin G., Chatelot J. Intérêt des cohortes pour la surveillance épidémiologique : exemples dans le domaine des risques professionnels. Actualité et dossier en santé publique n° 78, mars 2012.

Protocoles

Geoffroy-Perez B., Bénézet L., Santin G. 2009. Programme Coset : Cohortes pour la surveillance épidémiologique en lien avec le travail – Protocole d'inclusion d'actifs relevant de la Mutualité Sociale Agricole. Saint-Maurice : Institut de veille sanitaire.

Bénézet L., Geoffroy-Perez B., Santin G. 2010. Accès aux données des systèmes d'information existants dans le cadre de la phase pilote du projet Coset-MSA. Saint-Maurice : Institut de veille sanitaire.

Santin,G., Bénézet,L., Geoffroy-Perez B. 2010. Enquête complémentaire auprès d'un échantillon de non-participants lors de la phase pilote Coset-MSA. Saint-Maurice : Institut de veille sanitaire.

Communications orales

Santin,G., Bénézet,L., Gueguen,A., Sitta,R., Zins,M., Geoffroy-Perez,B., Goldberg,M. 2010. Stratégies pour prendre en compte la non-réponse à l'inclusion dans les cohortes Coset-MSA et Constances. 6ème colloque francophone sur les sondages, Tanger, Maroc.

Geoffroy-Perez B., Bénézet L., Roy J., Grillet J.P., Pelc A., Santin G., Chatelot J., Viarouge N., Chérie J.C., Brémaud F., Goldberg M., Imbernon E. 2010. Le programme Coset – Lancement de la cohorte Coset-MSA. Aderest, Pont-à-Mousson, France.

Gueguen,A., Sitta,R., Bénézet,L., Santin,G., Goldberg,M., Zins,M. 2010. L'apport des bases administratives et médico-administratives pour la prise en compte des effets de sélection dans la cohorte Constances. Adelf, Marseille, France.

Santin G., Bénézet L., Delézire P., Gueguen A., Groupe Cohortes-Sniiram, Equipe Coset. 2012. Biais de sélection à l'inclusion dans une cohorte de surveillance épidémiologique de risques professionnels en France. Adelf, Bruxelles, Belgique.

Santin G., Bénézet L., Geoffroy B., Delézire P., Chatelot J., Sitta R., Bouyer J., Gueguen A., Groupe Cohortes-Sniiram. 2012. Biais de sélection à l'inclusion dans une cohorte de surveillance épidémiologique de risques professionnels en France. 7ème colloque francophone sur les sondages, Rennes, France.

Chatelot J., Geoffroy B., Santin G., Bénézet L., Delézire P. 2013. Evaluer et suivre la santé au travail : les cohortes Coset-MSA et Coset-RSI. Colloque « Travail indépendant : santé et conditions de travail ». Paris, France.

Santin G., Geoffroy B., Bénézet L., Delézire P., Bouyer J, Gueguen A. 2013. Une enquête complémentaire auprès de non-répondants est-elle nécessaire si on dispose de données auxiliaires nombreuses et variées ? Résultats de l'étude Coset-MSA. . Symposium international de 2013 de Statistique Canada sur les questions de méthodologie, Ottawa, Canada.

Communication affichée

Santin G, Bouyer J, Sitta R, Gueguen A. 2012. Analyse de sensibilité de deux méthodes d'estimation de moyenne et de variance doublement robustes. Adelf, Bruxelles, Belgique

SOMMAIRE

TABLE DES FIGURES

LISTE DES TABLEAUX

LISTE DES ABRÉVIATIONS OU DES ACRONYMES

ATMP: Accident du Travail et Maladie Professionnelle
CAPI : Computed Assisted Personal Interview
CATI : Computed Assisted Telephone Interview
CAWI : Computed Assisted Web Interview
CHAID: Chi-Squared Automatic Interaction Detection
Cnav : Caisse nationale d'assurance vieillesse
Coset : Cohortes pour la surveillance épidémiologique en lien avec le travail
EC : Enquête Complémentaire
ECMS: Enquête Canadienne sur les Mesures de Santé
EGB : Echantillon Généraliste des Bénéficiaires
EI : Enquête Initiale
EQM : Erreur Quadratique Moyenne
ENSP : Enquête Nationale sur la Santé des Populations
IJ : Indemnités Journalières
Inpes : Institut national de la prévention et de l'éducation pour la santé
InVS : Institut de Veille Sanitaire
Insee : Institut national de la statistique et des études économiques
IPW: Inverse Probability Weighting
MAR : Missing At Random
MCAR : Missing Completely At Random
MNAR: Missing Not At Random
MSA : Mutualité Sociale Agricole
NHANES : National Health And Nutrition Survey
NHIS : National Health Interview Survey
NIR : Numéro d'Identification au Répertoire
PMSI : Programme de Médicalisation des Systèmes d'Information
PND : Pli Non Distribuable
RG : Régime Général de sécurité sociale
RSI : Régime Social des Indépendants
SNIIR-AM : Système d'Information Inter-régime de l'Assurance Maladie

CHAPITRE I. INTRODUCTION

I.1 LA SURVEILLANCE ÉPIDÉMIOLOGIQUE

La surveillance épidémiologique est définie comme « le suivi et l'analyse épidémiologique systématiques et permanents d'un problème de santé et de ses déterminants à l'échelle d'une population, afin de les contrôler par des interventions au niveau collectif, et d'identifier des phénomènes inconnus en termes d'effets ou de déterminants. » (44).

Les phénomènes pris en compte dans les programmes de surveillance sont en général des pathologies identifiées comme ayant un impact important en termes de santé publique de par leur gravité ou leur prévalence. Par ailleurs, les programmes de surveillance épidémiologique doivent également permettre d'identifier des événements de santé non surveillés.

Les systèmes de surveillance peuvent s'appuyer sur la remontée d'informations issues de systèmes d'informations existants, tels que par exemple le Programme de médicalisation des systèmes d'information (PMSI) (22, 102) ou les urgences (10) c'est-à-dire sur des systèmes passifs, ou bien sur des informations collectées avec des objectifs épidémiologiques spécifiques, tels que des registres du cancer, des réseaux de médecins (64, 77) ou des enquêtes, c'est-à-dire sur des systèmes actifs.

On s'intéresse ici plus spécifiquement à des systèmes de surveillance ayant recours à des enquêtes. La fiabilité de ce type d'enquêtes dans un objectif de surveillance s'apprécie par leur capacité d'inférer, à partir d'un échantillon, des résultats extrapolables à l'ensemble de la population d'intérêt et d'être en mesure d'assurer un suivi d'indicateurs pertinents, par des schémas d'enquêtes transversales répétées ou de cohorte. Dans ce qui suit sont présentées quelques enquêtes de santé publique en population générale.

L'enquête de santé publique en population générale la plus connue est l'enquête américaine NHIS (National Health Interview Survey) ; créée en 1957, elle a pour objectif de mesurer l'état de santé de la population américaine à partir des données d'un échantillon aléatoire de 75 000 à 100 000 personnes recueillies en face-à-face par questionnaire avec enquêteur (21). Depuis 1999, les Etats-Unis mènent par ailleurs l'enquête NHANES (National Health And Nutrition Survey) ; c'est une enquête en population générale plus spécifiquement dédiée à la nutrition. Son originalité vient du fait qu'en plus de recueillir des données par questionnaire, elle inclut le recueil de données de santé collectées à partir d'examens physiques et d'échantillons sanguins et d'urine (20). Compte tenu du schéma de l'enquête et de son coût, l'échantillon aléatoire est constitué de 5000 personnes.

Le Canada s'est inspiré de l'enquête NHANES pour mettre en place en 2007 l'enquête de santé publique en population générale ECMS (Enquête Canadienne sur les Mesures de Santé) (121). Cette enquête, dont le spectre est plus large que l'enquête NHANES, recueille, par questionnaire avec enquêteur, de nombreuses informations sur la santé et les comportements à risque pour la santé et, par des examens médicaux, certaines mesures physiques directes (telles que le poids ou la taille), mais inclut également le prélèvement d'échantillons de sang et d'urine. Compte tenu là aussi de son coût, l'enquête, transversale répétée, est conduite auprès d'un échantillon aléatoire de 5 000 personnes.

En France, les principales enquêtes de santé publique en population générale, exploitées en surveillance épidémiologique, sont les suivantes.

L'enquête décennale Santé, pilotée par l'Insee a débuté en 1960 ; la dernière édition date de 2003 (17). C'est une enquête en population générale, auprès d'un échantillon aléatoire d'environ 20 000 personnes pour l'édition 2003, dont les objectifs principaux sont de décrire

la santé de la population au travers de la morbidité déclarée. Elle a été exploitée entre autres pour la surveillance de la santé mentale selon l'activité professionnelle (26, 111).

Le Baromètre santé, piloté par l'Inpes, est une enquête quinquennale qui a débuté en 1992 ; c'est une enquête en population générale menée auprès d'un échantillon aléatoire d'environ 10 000 personnes (63). Le Baromètre a pour objectifs principaux l'étude des comportements, des attitudes et des perceptions liées à la prise de risque et l'étude de l'état de santé des personnes vivant en France. L'édition 2005 a permis de décrire les tentatives de suicide selon la profession dans un objectif de surveillance épidémiologique des risques professionnels (25).

I.2 LE DÉROULEMENT D'UNE ENQUÊTE

Une enquête, qu'elle ait un objectif de santé publique ou autre, comporte les étapes suivantes (2).

Elle nécessite de définir des objectifs clairs : quelle est la population d'intérêt, quelles sont les variables d'intérêt et les paramètres d'intérêt ? Ces critères doivent prendre en compte les contraintes suivantes : le coût, l'information déjà disponible, l'organisation et la logistique à mettre en place.

Il faut ensuite mettre en place un protocole d'enquête en adéquation avec les objectifs, qui respecte la confidentialité des données et qui prend en compte les éléments décrits ci-après.

Tout d'abord, il faut rechercher la base de sondage la plus adaptée, c'est-à-dire, le fichier de données qui permet d'accéder à une population qui soit la plus proche possible de la population d'intérêt.

Ensuite il faut réaliser le tirage au sort de l'échantillon à enquêter et concevoir le questionnaire. La collecte des données peut alors être lancée.

Une fois les données collectées, il faut traiter les données, c'est-à-dire les codifier et les saisir, puis contrôler la qualité des données saisies.

Enfin, on peut procéder à l'estimation des paramètres d'intérêt et de leur variance pour ensuite publier les résultats.

La qualité des estimateurs issus d'une enquête est formalisée par l'erreur totale.

I.3 L'ERREUR TOTALE DANS LES ENQUÊTES

I.3.1 DÉFINITION

Soit θ un paramètre (par exemple une moyenne ou une prévalence) et $\hat{\theta}$ un estimateur de θ.

L'erreur totale de $\hat{\theta}$ est définie par :

$$ETot(\hat{\theta}) = \hat{\theta} - \theta$$

Selon Biemer (9), l'erreur totale dans les enquêtes est un cadre conceptuel qui fait référence à l'accumulation de toutes les erreurs qui peuvent survenir dans le plan de sondage (base de sondage et tirage au sort), la collecte, le traitement et l'analyse des données d'une enquête. Une erreur totale importante peut être problématique car elle correspond à une diminution de la fiabilité des inférences issues des données d'une enquête. Une estimation sera considérée comme fiable si l'erreur totale est petite c'est-à-dire si la distribution de $\hat{\theta}$ est de variance faible et centrée sur θ.

I.3.2 MESURE DE L'ERREUR TOTALE

La mesure la plus répandue pour quantifier l'erreur totale est l'erreur quadratique moyenne (EQM) (9).

Elle est définie par l'espérance de la différence au carré entre l'estimateur $\hat{\theta}$ et le paramètre à estimer θ :

$$EQM(\hat{\theta}) = E(\hat{\theta} - \theta)^2$$

Elle peut par ailleurs s'écrire en fonction du biais et de la variance de l'estimateur (cf.Figure I-1).

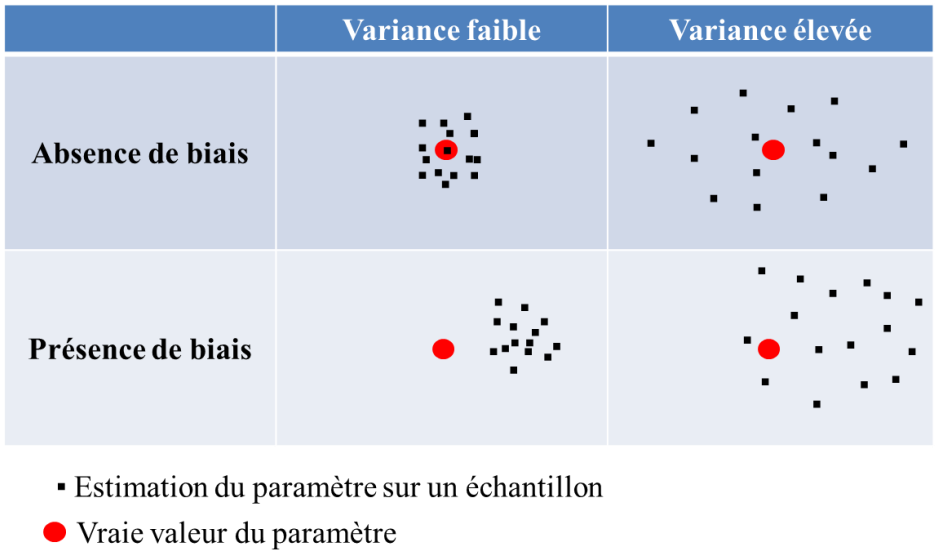
$$EQM(\hat{\theta}) = Biais^2(\hat{\theta}) + Var(\hat{\theta})$$

L'erreur quadratique moyenne résume les effets de toutes les sources d'erreur relatives à un estimateur ; elles seront décrites dans le paragraphe suivant (cf. I.3.3). Une EQM faible indique que l'erreur totale est petite. Une EQM élevée indique qu'une ou plusieurs sources d'erreurs qui constituent l'erreur totale affectent la qualité de l'estimateur.

Un biais est problématique car il fournit des estimations qui ne sont pas extrapolables à la population d'intérêt ; en termes de santé publique, un biais entraine un risque d'alerter à tort ou au contraire de ne pas identifier un problème de santé publique.

Une variance élevée est problématique car elle est le reflet de la volatilité de l'estimation d'intérêt.

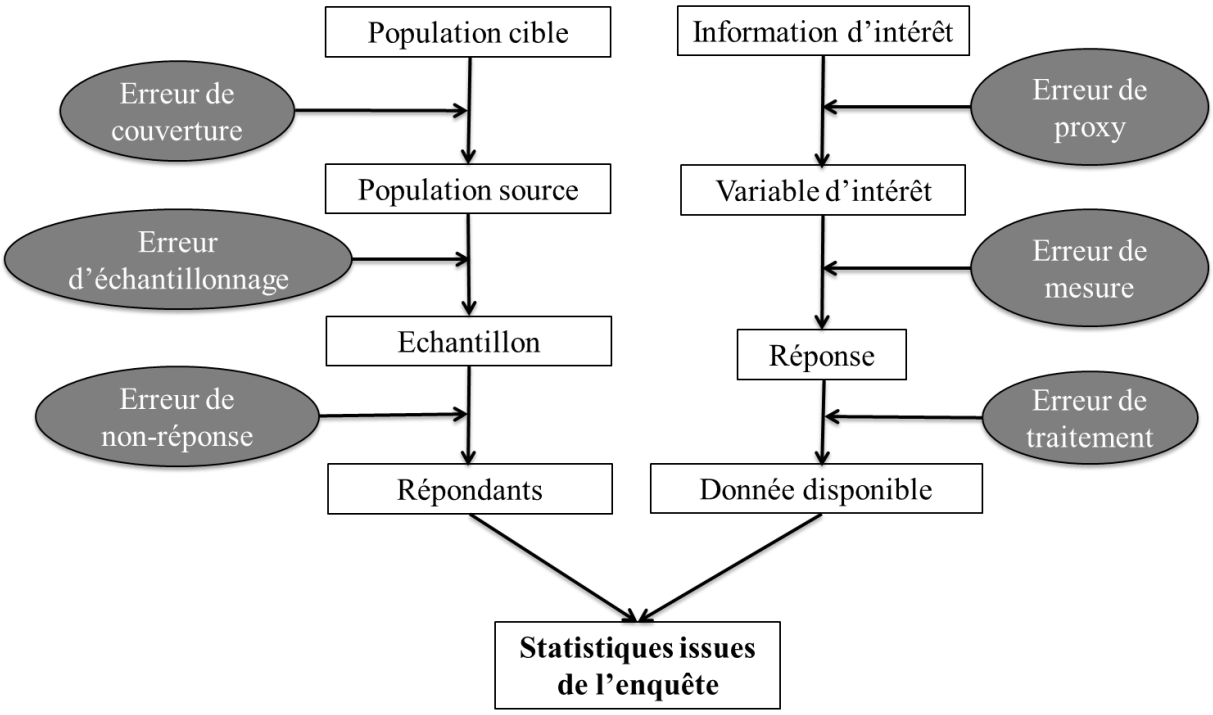
Figure I-1Représentation graphique du biais et de la variance (d'après (12))



I.3.3 PRINCIPALES COMPOSANTES DE L'ERREUR TOTALE

Le processus conduisant à l'erreur totale a été représenté graphiquement de manière synthétique (Figure I-2) par Groves (52). Les principales composantes de l'erreur totale sont l'erreur liée au fait que seule une partie de la population cible, ou population d'intérêt, est observée (partie gauche de la Figure I-2) et aux différences potentielles qui existent entre la vraie valeur de l'information d'intérêt et la valeur effectivement disponible (partie droite de la Figure I-2).

Figure I-2 (d'après (52)) : Eléments de l'erreur totale reliés aux étapes menant de la population cible (partie gauche) et l'information d'intérêt (partie droite) de à l'inférence statistique – pour une enquête donnée, avec un protocole d'enquête donné



I.3.3.1 <u>De la population cible à l'échantillon de répondants</u>

Le processus qui conduit de la population d'intérêt, appelée aussi population cible, à l'échantillon final de répondants sur lequel des statistiques seront calculées génère dans la plupart des cas des erreurs. Il est représenté dans la partie gauche de la Figure I-2.

Lorsque l'ensemble de la population cible constitue l'échantillon final, on est dans le cadre d'un recensement ; dans ce cas, les statistiques estimées ne sont pas entachées par les erreurs liées au fait que seule une partie de la population cible est enquêtée.

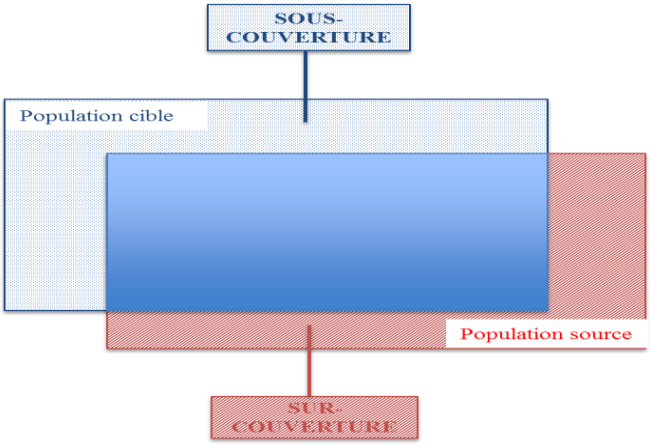
Dans le cas où on cherche à inférer des statistiques à une population cible à partir d'une partie de la population, ou échantillon, les estimations comprennent nécessairement des erreurs d'échantillonnage. En pratique, elles sont aussi souvent entachées par des erreurs de couverture et des erreurs de non-réponse.

I.3.3.1.1 Erreur de couverture

I.3.3.1.1.1 Définition

On observe une erreur de couverture lorsque la base de sondage, appelée aussi la population source, ne correspond pas exactement à la population cible (Figure I-3). On dit qu'il y a sur-couverture lorsque la base de sondage couvre plus de personnes que la population cible ; à l'inverse, on dit qu'il y a sous-couverture lorsque la base de sondage n'inclut pas l'ensemble de la population cible.

Figure I-3 : Erreurs de couverture



Les erreurs de couverture sont classiques, car il est souvent difficile de disposer d'une base de sondage qui coïncide parfaitement avec la population cible. Par exemple, réaliser une enquête auprès de la population en France n'est pas aisé car les données individuelles du recensement ne sont pas accessibles. C'est pour pallier ce problème que l'Inpes réalise ses enquêtes via des numéros de téléphone, en supposant qu'un numéro de téléphone correspond à un ménage ; par définition, les ménages ne disposant pas de numéro de téléphone, par exemple les sans domicile, ne peuvent donc pas être enquêtés. On observe dans ce cas une sous-couverture.

Des situations mêlant sur-couverture et sous-couverture, peuvent, en pratique, coexister. Imaginons que l'on souhaite réaliser une enquête auprès de femmes résidant en France par

une enquête utilisant les numéros de téléphone comme base de sondage, on aura sous-couverture car les personnes sans domicile ne pourront pas être enquêtées, et sur-couverture car les hommes pourront être contactés.

I.3.3.1.1.2 Conséquences sur les estimations

Si les unités non-incluses (sous-couverture) ou incluses à tort (sur-couverture) diffèrent de la population source, les erreurs de couverture peuvent entrainer un biais dans les estimations de prévalence.

I.3.3.1.1.3 Comment limiter les erreurs de couverture ?

Avant le recueil des données, la première façon de limiter les erreurs de couverture est de trouver une base de sondage la plus proche possible de la population d'intérêt.

Lorsque, comme dans l'exemple précédent, il est possible de savoir après la mise en œuvre du plan de sondage, qu'un individu ne fait pas partie du champ de l'enquête (qu'il s'agisse d'un homme et non d'une femme), on peut obtenir des estimations non biaisées en n'utilisant que les données sur les femmes ; on parle alors d'estimations sur un domaine (1). Quand il s'agit d'erreurs de sous-couverture ou d'erreurs de sur-couverture qu'on ne peut pas connaître, il est possible de corriger partiellement les erreurs de couverture une fois les données recueillies ; ce point sera abordé dans le chapitre II.1.2.3.

I.3.3.1.2 Erreur d'échantillonnage

Cette partie sera développée spécifiquement dans le chapitre II.

I.3.3.1.2.1 Définition de l'échantillonnage

L'échantillonnage est le processus qui conduit de la population source à un échantillon. Il y a deux types d'échantillonnage : l'échantillonnage probabiliste où la probabilité d'inclusion d'un individu dans l'échantillon est connue et l'échantillonnage empirique, qui inclut les enquêtes par quotas, où la probabilité d'inclusion d'un individu dans l'échantillon est inconnue. Nous nous plaçons ici exclusivement dans le cadre de l'échantillonnage probabiliste.

Dans un échantillonnage probabiliste, les unités constituant la population source sont sélectionnées selon un processus aléatoire connu. Autrement dit, les unités de la base de sondage sont sélectionnées par tirage au sort pour constituer l'échantillon. Une unité peut être par exemple un hôpital, une famille ou un individu. Dans tout le document, nous considérerons qu'une unité correspond exclusivement à un individu.

En statistique d'enquête, on considère que la population est de taille finie *N* et que la variable d'intérêt *Y* a une valeur fixée pour chaque individu de la population ; autrement dit, la variable *Y* n'est pas considérée comme aléatoire, mais comme un vecteur fixe. Dans un échantillon *(s)* de taille *n*, c'est le fait qu'un individu de la population a été sélectionné ou non qui est aléatoire ; cette sélection dépend du plan de sondage auquel on a eu recours.

Soit *I* la variable aléatoire « individu sélectionné par tirage au sort ».

On a : $I_i = \begin{cases} 1 \text{ si l'individu i est sélectionné dans l'échantillon} \\ 0 \text{ sinon} \end{cases}$

Habituellement, on note π_i la probabilité d'inclusion de l'individu *i* dans l'échantillon *(s)*. La loi de probabilité générant l'échantillon *(s)* est appelée plan de sondage.

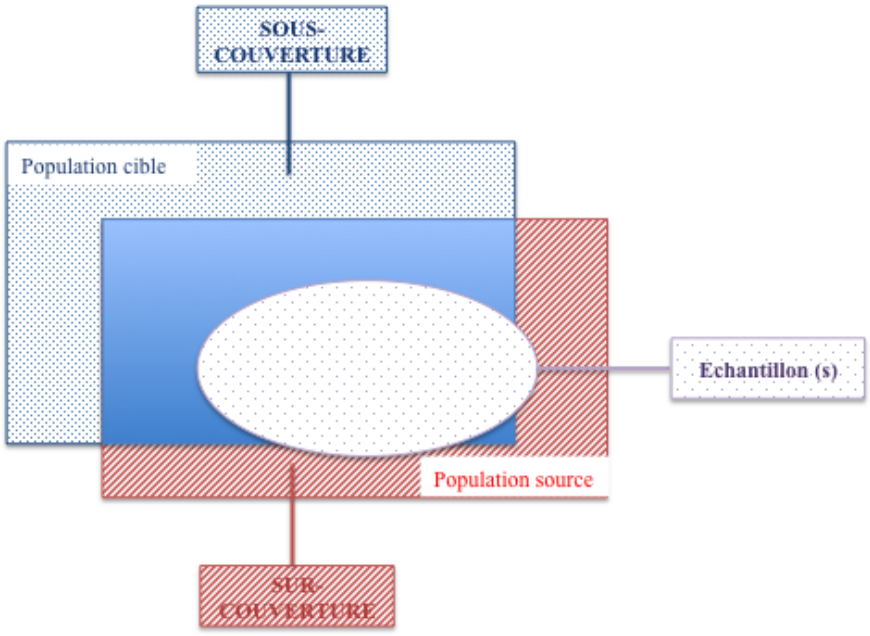
Remarque importante :

C'est une spécificité de la statistique d'enquête par rapport à la statistique classique. Plus précisément, en statistique classique, on suppose que la population est de taille infinie et que la variable d'intérêt *Y* est une variable aléatoire qui suit une loi de probabilité donnée (par exemple loi normale, loi de Poisson…). Un échantillon de taille *n* correspond à n réalisations de *Y* ; autrement dit, pour tout $i = \{1,...,n\}$, y_i prend une valeur aléatoire selon la loi de probabilité suivie par *Y*.

I.3.3.1.2.2 Définition de l'erreur d'échantillonnage

L'erreur d'échantillonnage correspond à l'écart entre la valeur du paramètre estimé sur l'échantillon et sa vraie valeur sur la population source (Figure I-4).

Figure I-4 : Erreur de d'échantillonnage



I.3.3.1.2.3 Conséquences sur les estimations

Le biais d'un estimateur correspond à la différence entre la moyenne de toutes les estimations obtenues sur tous les échantillons qui pourraient être possiblement sélectionnés selon un certain plan de sondage et la valeur réelle de ce paramètre dans la population.

Soit θ le paramètre qu'on cherche à estimer et $\hat{\theta}$ un estimateur de θ. Le biais de $\hat{\theta}$ lié à un plan de sondage *(ps)* s'exprime par :

$$Biais_{ps}(\hat{\theta}) = E_{ps}(\hat{\theta}) - \theta$$

L'échantillonnage probabiliste conduit à des estimations sans biais ou asymptotiquement sans biais quand le paramètre d'intérêt est une prévalence ou une moyenne (cf. II.1.2.2.1.1).

L'erreur d'échantillonnage ne comprend donc pas de biais (directement ou asymptotiquement) ; en revanche, elle est composée de variance, qu'on appelle fluctuation d'échantillonnage.

I.3.3.1.2.4 Comment limiter les erreurs d'échantillonnage ?

Avant le recueil des données, l'erreur d'échantillonnage peut être limitée par le plan de sondage : la taille de l'échantillon ou le type de sondage utilisé (par exemple un sondage stratifié conduit à des estimateurs ayant une variance plus faible que ceux issus d'un sondage aléatoire simple).

Il est possible de corriger partiellement les erreurs d'échantillonnage une fois les données recueillies ; ce point sera abordé dans le chapitre II.1.2.3.

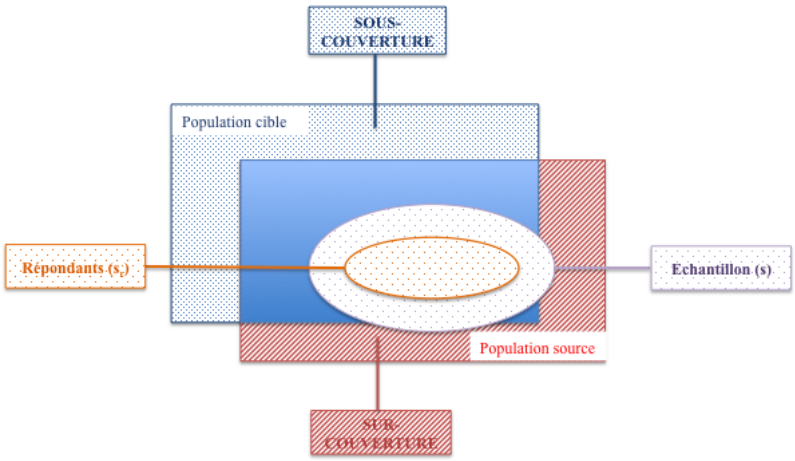
I.3.3.1.3 Erreur de non-réponse

I.3.3.1.3.1 Définition

L'erreur de non-réponse (Figure I-5) correspond à l'écart entre la valeur du paramètre estimé sur l'échantillon de répondants et la valeur qui serait estimée sur l'échantillon des individus tirés au sort ; on dit que l'erreur de non-réponse est conditionnelle à l'échantillon tiré au sort. L'écart entre la valeur du paramètre estimé sur l'échantillon de répondants et sa vraie valeur sur la population source (quand celle-ci est disponible) est également utilisé comme définition dans la littérature.

L'erreur de non-réponse est due au fait que les individus constituant l'échantillon sélectionné selon un certain plan de sondage ne répondent pas toutes à l'enquête.

Figure I-5 : Erreur de non-réponse



Remarque :

On dit que la non-réponse est :

- totale lorsque la personne enquêtée ne répond à aucune question de l'enquête ;
- partielle si elle répond à certaines questions de l'enquête, mais pas à toutes. On appelle plus communément ce type de non-réponse « données manquantes ».

Cette différence a été introduite par les statisticiens d'enquête, car le traitement de la non-réponse dépend du type de non-réponse (partielle ou totale), essentiellement pour des raisons pratiques. Ce point sera détaillé ultérieurement (cf. II.2.3).

I.3.3.1.3.2 Conséquences sur les estimations

I.3.3.1.3.2.1 Biais de non-réponse

Historiquement, la non-réponse a d'abord été appréhendée comme un processus fixe ; autrement dit, une personne non-répondante était considérée comme telle quelle que soit les conditions de l'enquête (sujet, protocole, période). Cette approche est *a priori* valable pour un petit groupe de personnes hermétiques aux enquêtes, mais probablement trop restrictive. Une personne peut en effet être répondante ou non-répondante selon les conditions de l'enquête et certains facteurs personnels. D'un point de vue plus formel, la non-réponse a été vue par Rubin (108), non plus comme un processus déterministe, mais comme un processus stochastique et décrite par une variable aléatoire binaire notée par la suite R (R vaut 1 en cas de réponse totale à l'enquête, 0 sinon).

Une fois qu'un échantillon est constitué selon un certain plan de sondage, le biais de non-réponse peut être défini comme la différence moyenne entre l'estimation calculée chez les répondants et l'estimation calculée sur l'échantillon complet si le processus de non-réponse était répété un grand nombre de fois. C'est une définition semblable à celle du biais d'échantillonnage, mis à part le fait qu'en échantillonnage, la valeur de référence de l'estimation est la vraie valeur obtenue à partir de la population, et que ce sont les échantillons complets qui sont répétés, et non pas les échantillons de répondants.

Une différence fondamentale existe entre échantillonnage et non-réponse. En échantillonnage, le processus qui conduit de la population à l'échantillon est défini *a priori* et contrôlé par le

plan de sondage, puisque les probabilités d'inclusion sont connues, et ceci permet d'obtenir des estimations sans biais ou approximativement sans biais de prévalences ; ce n'est pas le cas de la non-réponse, puisque le processus qui conduit de l'échantillon complet à des échantillons de répondants possibles est inconnu.

I.3.3.1.3.2.2 Variance

En présence de non-réponse, la taille de l'échantillon étant plus petite que prévue, la variance des estimateurs est en général plus élevée que celle qu'on aurait obtenue en l'absence de non-réponse.

I.3.3.1.3.3 Comment limiter les erreurs de non-réponse ?

Le traitement de la non-réponse sera spécifiquement traité dans le chapitre II.2.

I.3.3.2 De l'information d'intérêt à la donnée disponible

Une autre source d'erreurs peut survenir lors d'une enquête : l'écart entre la vraie valeur de l'information d'intérêt et la donnée disponible pour l'analyse. Elle se décompose en trois sources d'erreur possibles listées ci-dessous (cf. partie droite de la Figure I-2).

I.3.3.2.1 Erreur de proxy

I.3.3.2.1.1 Définition

L'erreur de proxy est définie comme l'écart entre l'information d'intérêt et la variable d'intérêt. Elle peut venir de la difficulté à mesurer ce qu'on l'on cherche à mesurer. Par exemple, l'âge est, dans la plupart des cas, directement mesurable ; en revanche il est beaucoup plus difficile de mesurer la symptomatologie dépressive (information d'intérêt) d'une personne. Pour mesurer ce type d'information d'intérêt, on peut utiliser des échelles,

comme le Ces-d (variable d'intérêt) (35) à partir desquelles on calcule un score ; au-delà d'un certain seuil, on peut considérer que la personne présente une symptomatologie dépressive. On peut se demander jusqu'à quel point ce qu'on mesure dans ce cas-là correspond à ce que l'on souhaite réellement mesurer ; cet exemple laisse par ailleurs entrevoir que la notion de vraie valeur n'est pas toujours facile à appréhender.

I.3.3.2.1.2 Conséquences sur les estimations

Les erreurs de proxy peuvent entrainer un biais dans les estimations de prévalence.

I.3.3.2.1.3 Comment limiter les erreurs de proxy ?

Les erreurs de proxy peuvent être limitées en définissant au mieux ce qu'on chercher à mesurer et en utilisant des outils de mesure validés (32).

I.3.3.2.2 Erreur de mesure

I.3.3.2.2.1 Définition

L'erreur de mesure correspond à l'écart entre la vraie valeur de la variable d'intérêt et la valeur mesurée (ou déclarée) pour cette variable lors de l'enquête. Les sources d'erreur de mesure sont très nombreuses. Sont listées ici les erreurs de mesure liées à la collecte des données par questionnaire.

Les erreurs de mesure peuvent venir du mode de collecte des données. Par exemple, lors d'une enquête avec enquêteur, des erreurs de mesure peuvent survenir si l'enquêteur est peu impliqué dans l'enquête ou si les questions posées sont sensibles ; dans ce dernier cas, un phénomène de désirabilité sociale peut survenir, les personnes enquêtées pouvant répondre préférentiellement de façon consensuelle à l'enquêteur. D'un autre côté, les questionnaires auto-administrés peuvent aussi générer des erreurs de mesure lorsque par exemple, le

questionnaire est long et complexe ; c'est moins le cas dans les enquêtes avec enquêteur car ce dernier peut aider à la compréhension des questions.

L'erreur de mesure peut également être induite par la personne enquêtée. Si cette dernière se sent peu concernée par l'enquête, si elle ne se souvient pas d'une exposition passée sur laquelle on l'interroge ou si elle ne souhaite pas donner la réponse exacte à la question posée, ses réponses peuvent contenir des erreurs de mesure.

L'ordre des questions peut également avoir une influence sur les réponses (29). Il peut entrainer un effet de halo : lorsque, par exemple, les questions sont orientées dans le même sens, les personnes ont tendance à répondre de manière semblable. Un autre effet est l'effet de contamination : les questions posées précédemment peuvent avoir une influence sur les réponses aux questions suivantes. Par exemple, dans un questionnaire de santé, poser une question sur l'état de santé général perçu ne produira probablement pas les mêmes estimations de prévalence si la question est posée en début ou en fin de questionnaire.

I.3.3.2.2.2 Conséquences sur les estimations

Les erreurs de mesure entrainent une augmentation de la variance de l'estimation d'une prévalence.

Elles peuvent par ailleurs entrainer un biais dans les estimations de prévalence.

I.3.3.2.2.3 Comment limiter les erreurs de mesure ?

Avant le recueil des données, les erreurs de mesure peuvent être limitées grâce à un questionnaire clair, de longueur acceptable et par un mode de recueil de données adapté au sujet de l'enquête. Par exemple, pour des thématiques sensibles comme les comportements à

risque pour la santé, il vaut mieux utiliser des autoquestionnaires ou avoir des recours à des enquêtes par téléphone pour limiter le phénomène de désirabilité sociale (97).

I.3.3.2.3 Erreur de traitement

I.3.3.2.3.1 Définition

Une fois que les données ont été recueillies, des erreurs de traitement sont également possibles, comme les erreurs de saisie ou de codage.

I.3.3.2.3.2 Conséquences sur les estimations

Les erreurs de traitement peuvent entrainer un biais dans les estimations des prévalences, mais également une augmentation de la variance de l'estimation d'une prévalence.

I.3.3.2.3.3 Comment limiter les erreurs de traitement ?

Les erreurs de traitement peuvent être limitées lors de la saisie de questionnaire papier, par exemple par de la double saisie. Le codage (ou recodage des données) peut également introduire des erreurs de traitement ; il est préférable de programmer les codages effectués et de conserver la base de données initiale afin de pouvoir y revenir après une potentielle erreur de codage.

I.3.3.3 Conclusion

La longue liste des erreurs possibles donne par elle-même une idée de leurs conséquences sur la qualité des estimations d'une prévalence : elles peuvent entrainer biais et augmentation de la variance. On peut cependant noter qu'alors que les variances s'additionnent, les biais peuvent éventuellement se compenser.

La diversité des efforts à consentir pour obtenir un schéma d'enquête optimal est assez conséquente. Il faut par ailleurs contrôler toutes ces erreurs pour un budget et des contraintes d'enquête donnés.

I.4 TAUX DE RÉPONSE ET ERREUR TOTALE

L'idée prédominante dans de nombreuses disciplines est qu'il faut autant que possible maximiser le taux de réponse afin de minimiser les conséquences de la non-réponse, c'est-à-dire l'existence d'un biais lié à la non-réponse et l'augmentation de la variance (14).

Ce point de vue traditionnel est cependant remis en question par certains auteurs, dont Groves (48), qui, dans une revue de la littérature, montre qu'en pratique, le taux de réponse n'est pas forcément lié au biais de non-réponse. Par ailleurs, augmenter à tout prix le taux de réponse à une enquête peut s'avérer contre-productif si les personnes qui ne souhaitent pas répondre spontanément, mais qui finissent par répondre après de nombreuses sollicitations, ont des caractéristiques liées directement à la variable d'intérêt (87).

En outre, des questions se posent sur le lien entre la propension à répondre et l'erreur de mesure. En effet, si on cherche à augmenter autant que possible le taux de réponse à une enquête, on va être amené à obtenir la réponse de personnes peu intéressées par l'enquête, donc pouvant potentiellement donner des réponses imprécises, voire erronées (9), les personnes les moins motivées répondant avec moins d'attention. Dans le pire des cas, les répondants « non spontanés » à des enquêtes sur un sujet sensible peuvent volontairement donner des réponses erronées (97).

Un échantillon de taille plus grande entraine une variance plus petite ; donc chercher à maximiser le taux de réponse à une enquête ne pose pas de problème de variance, au contraire.

On peut donc s'interroger sur les conséquences en termes d'erreur totale d'une recherche d'un taux de réponse maximal. En termes de biais, trois cas peuvent survenir :

- Une diminution ou une stagnation des biais de non-réponse et de mesure : il y a un réel intérêt à maximiser le taux de réponse ;
- Une diminution du biais de non-réponse et une augmentation biais de mesure (ou l'inverse) : se pose alors la question de la balance entre biais de non-réponse et biais de mesure ;
- Une augmentation ou une stagnation des biais de non-réponse et de mesure : il apparait dangereux de maximiser le taux de réponse. Ce cas, qui peut paraitre à première vue contre-intuitif, sera illustré par des exemples ultérieurement (cf. II.3).

Il est cependant difficile de situer en pratique dans quel cas de figure on se situe. Des propositions d'évaluation ont été émises par certains auteurs (74, 92) mais elles sont difficiles à mettre en œuvre car elles nécessitent de disposer de données de référence.

Cette question sera plus longuement développée dans le chapitre II.3.

I.5 COMMENT ESTIMER L'ERREUR TOTALE DANS UNE ENQUÊTE ET DISTINGUER SES COMPOSANTES ?

I.5.1 DISPOSER DE GOLD STANDARD

L'erreur totale est bien définie d'un point de vue théorique, mais il est plus problématique de l'estimer en pratique. La seule erreur pouvant être estimée est l'erreur d'échantillonnage car elle se déduit de la connaissance du plan de sondage ; c'est l'erreur qui est en général la mieux connue. Pour une étude donnée, si on ne s'intéresse qu'à l'erreur de non-réponse et à l'erreur de mesure, il faut disposer de valeurs de référence ou « gold standard » pour les sujets sélectionnés dans l'échantillon afin de pouvoir d'une part mesurer l'écart entre l'estimation

obtenue via les valeurs prises par la variable mesurée dans l'enquête et celle obtenue via les vraies valeurs de la variable dans l'échantillon de répondants (ce qui correspond à l'erreur de mesure), et d'autre part mesurer l'écart entre l'estimation obtenue via l'échantillon de répondants et l'estimation qui aurait été obtenue sur l'échantillon total (ce qui correspond à l'erreur de non-réponse). C'est en pratique, possible lorsqu'on dispose de systèmes d'information existants appariés sur l'échantillon sélectionné. Si on considère que la valeur du paramètre estimé sur ces données pour l'échantillon sélectionné est un gold standard, et que certaines variables issues de ces systèmes d'information ont également été recueillies par questionnaire, il sera possible d'estimer une erreur de non-réponse et une erreur de mesure, comme l'ont fait Olson et Kreuter (74, 92).

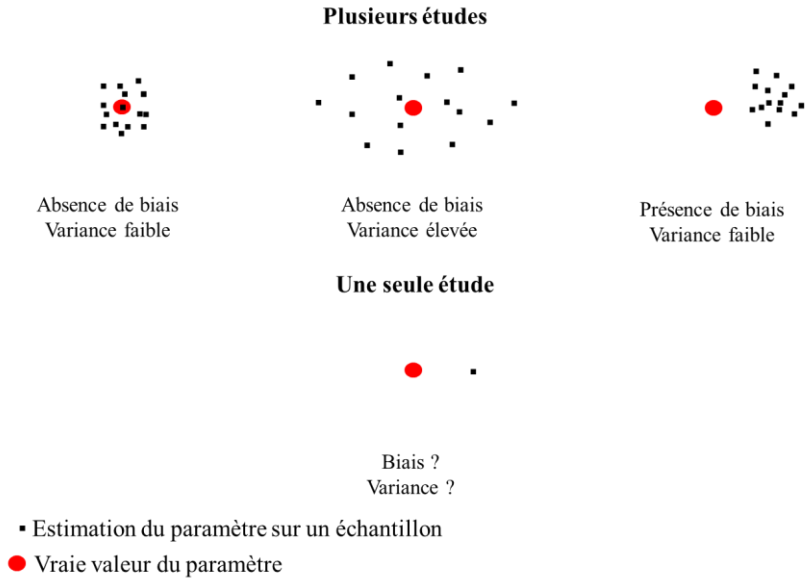
La disponibilité de données issues de systèmes d'information existants permet donc d'appréhender l'erreur totale sous un angle nouveau d'un point de vue pratique.

I.5.2 ERREUR, BIAIS ET VARIANCE SUR UNE SEULE ÉTUDE

En pratique, il est impossible, avec les données d'une seule étude, de distinguer le biais et la variance constitutifs de l'erreur totale.

Comme nous l'avons vu, pour pouvoir estimer un biais, il faudrait répéter une enquête un nombre de fois suffisant, estimer le paramètre d'intérêt pour chaque répétition, calculer la moyenne du paramètre estimé sur l'ensemble des répétitions et estimer la différence entre cette moyenne et la vraie valeur du paramètre d'intérêt. Cela suppose donc de disposer de la vraie valeur et d'un nombre suffisant d'enquêtes répétées. En général, on dispose au mieux de la vraie valeur du paramètre à estimer (Figure I-6).

Figure I-6 : De la difficulté de différencier le biais et la variance sur une seule étude



Sur une enquête, on mesure donc une erreur et il n'est pas possible de différencier le biais de la variance. Néanmoins, si on considère qu'une estimation biaisée est plus problématique qu'une estimation sujette à des fluctuations d'échantillonnage, on se place dans le cas le plus défavorable en considérant que l'erreur est avant tout un biais.

Dans le travail qui suit, on mesurera l'erreur en calculant la différence entre l'estimation du paramètre et sa valeur de référence, comme l'ont fait d'autres auteurs (74, 92, 97) ; il y a donc un certain abus de langage de parler de biais dans ce cas, mais nous le ferons nous aussi.

I.6 OBJECTIFS DE LA THÈSE

Comme nous l'avons vu en introduction, un des facteurs influant sur l'erreur totale dans une enquête est la non-réponse. Ce travail se focalise sur cette problématique dans le cadre de l'estimation d'une moyenne ou d'une prévalence et s'appuie sur les données de la cohorte Coset-MSA. Les objectifs sont les suivants :

- faire le point sur la problématique de la non-réponse totale en statistique d'enquête et sur les méthodes pour minimiser ses conséquences en terme de biais ;
- étudier l'apport de données supplémentaires pour corriger les biais de non-réponse totale : d'une part celle d'une enquête auprès de non-répondants et d'autre part celles des bases médico-administratives ;
- discuter le lien entre le taux de réponse et deux composantes de l'erreur totale : l'erreur de non-réponse et l'erreur de mesure.

CHAPITRE II. LA NON-RÉPONSE TOTALE EN STATISTIQUE D'ENQUÊTE

La non-réponse totale est systématiquement rencontrée dans les enquêtes. Elle peut être due à l'absence de la personne à enquêter au moment de la prise de contact ou à son refus de participer à l'enquête. La non-réponse peut entrainer un biais dans les estimations si les valeurs de la variable d'intérêt sont en moyenne différentes chez les répondants et les non répondants.

Pour comprendre le fondement des méthodes proposées pour corriger les biais de non-réponse dans les enquêtes, il faut tout d'abord se placer dans le cadre des enquêtes par sondage sans non-réponse, sujettes principalement aux erreurs d'échantillonnage. En effet, celles-ci proposent un cadre de travail permettant d'obtenir des estimateurs sans biais ou asymptotiquement sans biais de statistiques descriptives telles que des totaux, des moyennes ou des prévalences. Tout le travail de prise en compte de la non-réponse va consister à se ramener autant que possible à des schémas d'enquête bien connus en l'absence de non-réponse pour obtenir des estimateurs de statistiques descriptives les moins biaisés possibles.

C'est pourquoi ce chapitre aborde en premier lieu les enquêtes sans non-réponse, puis les enquêtes avec non-réponse.

Enfin, une discussion plus générale sera conduite pour discuter des liens entre la propension à répondre, les biais de non-réponse et les biais de mesure.

II.1 ENQUÊTE SANS NON-RÉPONSE : L'ERREUR D'ÉCHANTILLONNAGE

Cette partie s'inspire fortement du cours de David Haziza (58) et du livre de Pascal Ardilly (2). Pour simplifier, on se place dans le contexte idéal où il n'existe pas d'erreur de couverture, c'est-à-dire que la population d'intérêt (ou population cible) correspond exactement à la population source (ou base de sondage).

L'échantillonnage correspond à la manière dont on sélectionne l'échantillon à partir de la population d'intérêt.

Trois éléments essentiels sont à considérer:

- la méthode de sélection de l'échantillon, autrement dit le plan de sondage ;
- le paramètre à estimer (par exemple un total ou une moyenne) ;
- l'estimateur.

II.1.1 DÉFINITIONS ET NOTATIONS

Soit :

- U la population d'intérêt ; U est une population finie d'individus de taille connue N. On note $U = \{1,\dots,N\}$;
- $y = (y_1, y_2, \dots, y_i, \dots, y_n)$ les valeurs de la variable Y prise dans la population lorsqu'on s'intéresse à une seule variable d'intérêt.
- $y_a = (y_{a,1}, \dots, y_{a,i}, \dots, y_{a,n})$ et $y_b = (y_{b,1}, \dots, y_{b,i}, \dots, y_{b,n})$ si on s'intéresse à deux variables d'intérêt.

En statistique d'enquête, les paramètres d'intérêt sont des paramètres descriptifs, tels que des totaux, des moyennes, des prévalences ou des pourcentages, des ratios :

- total de *y* dans la population : $t_y = \sum_{i \in U} y_i$;
- moyenne ou prévalence de *y* dans la population : $\bar{y}_U = \frac{t_y}{N}$;
- ratio de deux totaux : $R = \frac{t_{ya}}{t_{y_b}} = \frac{\bar{y}_a}{\bar{y}_b}$.

Remarques importantes :

1. En statistiques d'enquête, le total est le principal paramètre d'intérêt. En effet, tous les autres paramètres présentés ici découlent de ce dernier. C'est pourquoi un accent particulier est mis sur le total, même si en épidémiologie descriptive, c'est un paramètre rarement utilisé.
2. Une moyenne ou une prévalence peuvent s'exprimer comme un ratio de deux totaux, le dénominateur correspondant dans ce cas-là au nombre total des individus composant la population.

II.1.2 MÉTHODES D'ÉCHANTILLONNAGE

Il existe deux grandes familles d'échantillonnage en statistique d'enquête : l'échantillonnage probabiliste, où la probabilité de sélection d'un individu est connue et l'échantillonnage non-probabiliste (appelé également échantillonnage empirique), où la probabilité de sélection d'un individu est inconnue. Les deux principaux types d'échantillonnage non-probabilistes sont :

- L'échantillonnage à participation volontaire où l'échantillon est recruté sur l'unique base du volontariat. Si les volontaires ont des caractéristiques différentes des non-volontaires, il y a un biais de sélection ;
- L'échantillonnage par quotas où le recrutement des volontaires se fait en fonction d'un effectif que l'on souhaite atteindre pour certains sous-groupes de personnes ; on peut par exemple, souhaiter obtenir un échantillon respectant la proportion

d'hommes et de femmes dans la population française (supposons que ce soit 48% pour les hommes et 52% pour les femmes). Si on souhaite un échantillon de 100 personnes avec un quota sur le sexe, on va donc chercher à recruter 48 hommes et 52 femmes. L'échantillonnage par quotas est préférable à l'échantillonnage à participation volontaire car il garantit l'inclusion de sous-groupes différents ; néanmoins si les variables utilisées pour les quotas ne sont pas associées aux variables d'intérêt, ce type d'échantillonnage est équivalent à l'échantillonnage à participation volontaire. Quoiqu'il en soit, il ne garantit pas d'obtenir des estimateurs sans biais.

Nous nous plaçons dans ce travail dans le cadre de sondage probabiliste (avec tirage au sort).

II.1.2.1 Cadre de travail des sondages probabilistes

Une base de sondage est définie comme une liste d'unités, ici des individus, qui couvre autant que possible la population cible et qui permet d'identifier chaque individu. Les qualités nécessaires à une base de sondage sont d'être une liste exhaustive et sans doublon. Il est par ailleurs toujours bienvenu de disposer, pour chaque individu de la liste, d'informations auxiliaires.

La base de sondage correspond idéalement à la population cible $U=\{1,\ldots,N\}$; en effet, dans le cas contraire, la différence conduit à des erreurs de couverture.

A partir de la base de sondage, on sélectionne un échantillon (s) à l'aide d'un plan de sondage $p(s)$ sur U. Un plan de sondage est défini comme une loi de probabilité telle que :

$\forall s \subset U, p(s) \geq 0$ et $\sum_{s \subset U} p(s) = 1$ avec $p(s)$ la probabilité de sélectionner l'échantillon (s).

A partir d'un plan de sondage, on peut calculer la probabilité d'inclusion de chaque individu de la population *U*. Soit π_i la probabilité d'inclusion d'un individu *i*. Elle est définie par :

$$\pi_i = P(i \in s) = P(I_i = 1) = \sum_{s \subset U, s \ni i} p(s)\,,$$

où la variable aléatoire $I_i = \begin{cases} 1 \text{ si l'individu i est sélectionné dans l'échantillon } (s) \\ 0 \text{ sinon} \end{cases}$

Le plan de sondage doit être tel que $\pi_i > 0$. Cette condition, également appelée condition de positivité est nécessaire pour obtenir un estimateur sans biais.

On note π_{ij} la probabilité que deux individus distincts *i* et *j* soient sélectionnés conjointement dans l'échantillon :

$$\pi_{ij} = P(i,j \in s) = P\big(I_i = 1; I_j = 1\big) = \sum_{s \subset U, s \ni i,j} p(s).$$

Le choix du plan de sondage dépend des objectifs de l'enquête et des informations disponibles dans la base de sondage.

Ne sont présentés ici que les plans de sondage utilisés dans la thèse.

II.1.2.2 Quelques plans de sondage probabiliste

II.1.2.2.1 Quelques sondages élémentaires

II.1.2.2.1.1 Sondage aléatoire simple sans remise

Le sondage aléatoire simple sans remise, noté SAS, consiste à tirer dans une population *U* de taille *N* un échantillon de taille *n* fixée, sans remise et de telle sorte que chaque individu ait la même probabilité d'inclusion. Ainsi, pour tout individu *i*, la probabilité d'inclusion est égale à $\pi_i = \frac{n}{N}$.

i) Estimateur d'un total

Un **estimateur sans biais du total t_y** est donné par : $\hat{t}_{y,SAS} = \sum_{i \in s} d_i y_i$ avec $d_i = \frac{1}{\pi_i} = \frac{N}{n}$.

Démonstration : en annexe

Remarque : Une autre façon d'exprimer $\hat{t}_{y,SAS}$ est $\hat{t}_{y,SAS} = \frac{N}{n} \sum_{i \in s} y_i$

Exemple illustratif :

Soit une population de taille $N = 4$ et y la variable nombre de jours d'arrêt de travail.

i	1	2	3	4
y_i	2	15	5	7

Le total de la variable y dans la population est $t_y = 2 + 15 + 5 + 7 = 29$.

On sélectionne par tirage aléatoire simple un échantillon de 2 individus dans cette population. Les probabilités d'inclusion sont $\pi_i = \frac{n}{N} = \frac{2}{4} = 0{,}5$. On veut estimer le nombre total de jours d'arrêt de travail y dans cette population.

Il existe 6 échantillons équiprobables.

Echantillon (s)	{1,2}	{1,3}	{1,4}	{2,3}	{2,4}	{3,4}
$\{y_i, y_j\}$	{2,15}	{2,5}	{2,7}	{15,5}	{15,7}	{5,7}
$\hat{t}_{y,SAS} = \frac{N}{n} \sum_{i \in s} y_i = 2 \sum_{i \in s} y_i$	34	14	18	40	44	24

$$E_{p(s)}(\hat{t}_{y,SAS}) = \frac{1}{6}(34 + 14 + 18 + 40 + 44 + 24) = \frac{174}{6} = 29 = t_y.$$

➔ $\hat{t}_{y,SAS}$ est un estimateur sans biais du total t_y.

Remarque :

Ne pas obtenir à partir d'un échantillon la vraie valeur de t_y ne signifie pas avoir un biais. L'écart à la vraie valeur de t_y reflète simplement les fluctuations d'échantillonnage, autrement dit la variance.

Un **estimateur sans biais de la variance de l'estimateur du total** est donné par :

$$\hat{V}(\hat{t}_{y,SAS}) = N^2\left(1-\frac{n}{N}\right)\frac{s^2}{n} \; avec \; s^2 = \frac{1}{n-1}\sum_{i\in s}(y_i-\hat{\bar{y}})^2 \text{ avec } \hat{\bar{y}} = \frac{1}{n}\sum_{i\in s} y_i$$

Remarque :

Plusieurs algorithmes de tirage au sort peuvent être utilisés pour sélectionner un échantillon aléatoire simple sans remise ; ils ne sont pas présentés ici. On peut se référer à des ouvrages existants (2, 113).

ii) Estimateur d'une moyenne ou d'une prévalence quand la taille de la population est connue

Un **estimateur sans biais d'une moyenne ou d'une prévalence $\bar{y}$** est donné par :

$$\hat{\bar{y}}_{SAS} = \frac{1}{n}\sum_{i\in s} y_i.$$

Un **estimateur sans biais de la variance de l'estimateur d'une moyenne ou d'une prévalence** est donné par:

$$\hat{V}(\hat{\bar{y}}_{SAS}) = \left(1-\frac{n}{N}\right)\frac{s^2}{n} \text{ avec } s^2 = \frac{1}{n-1}\sum_{i\in s}(y_i-\hat{\bar{y}})^2.$$

iii) Estimateur d'un ratio

Même si le ratio de deux totaux est rarement un paramètre d'intérêt en épidémiologie, son estimateur est présenté ici pour permettre la compréhension de l'estimateur d'une moyenne ou d'une prévalence quand la taille de la population est inconnue ou pour une estimation sur domaine. Dans ce cas, il est aussi nécessaire d'estimer la taille de la population, donc le numérateur et le dénominateur peuvent être considérés comme des totaux estimés et l'estimateur d'une moyenne ou d'une prévalence découle directement de l'estimateur du ratio.

L'espérance d'un ratio n'étant pas égale au ratio des espérances, l'estimateur $\hat{R}_{SAS} = \frac{\hat{t}_{y_{a,SAS}}}{\hat{t}_{y_{b,SAS}}}$ n'est pas un estimateur sans biais du ratio $R = \frac{t_{y_a}}{t_{y_b}}$. On peut néanmoins montrer que, pour une taille d'échantillon suffisamment grande, le biais de $\hat{R}_{SAS}$ est négligeable. On dit que $\hat{R}_{SAS}$ est un estimateur asymptotiquement sans biais de R.

Un estimateur approximativement sans biais de la variance de l'estimateur d'un ratio est donné par :

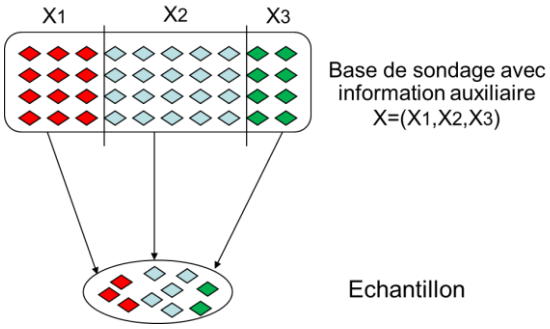
$$\hat{V}(\hat{R}_{SAS}) = \left(1 - \frac{n}{N}\right)\frac{s_e^2}{n\hat{\bar{y}}_{b,SAS}^2} \text{ avec } s_e^2 = \frac{1}{n-1}\sum_{i\in s}(y_{ai} - \hat{R}_{SAS}y_{bi})^2 \text{ et } \hat{\bar{y}}_{b,SAS} = \frac{1}{n}\sum_{i\in s} y_{bi}$$

Même si la taille de la population est connue, l'estimation d'un ratio est en fait courante en épidémiologie descriptive car il est fréquent qu'on veuille estimer des prévalences ou des moyennes dans un domaine particulier, dans une profession donnée par exemple, où le nombre de sujets dans la population est inconnu.

II.1.2.2.1.2 Sondage stratifié aléatoire simple

Le sondage stratifié est une procédure qui utilise une information disponible pour tous les individus de la base de sondage dans le but d'améliorer la qualité des estimateurs. Une telle information est dite information auxiliaire. Elle peut être par exemple l'âge, le sexe ou le département. A partir de cette information auxiliaire, on partitionne la population en sous-populations et on tire au sort dans chaque sous-population. Si, dans chaque strate, on réalise un sondage aléatoire simple dans chaque sous-population, on dit qu'on réalise un sondage stratifié aléatoire simple.

Figure II-1 : Représentation graphique d'un sondage stratifié aléatoire simple avec une probabilité d'inclusion de 0.25 dans chaque strate



Le sondage stratifié assure d'obtenir des individus issus de chaque strate. L'intérêt d'un sondage stratifié vient du fait que si la variable *Y* est faiblement dispersée dans chaque strate et fortement dispersée d'une strate à l'autre, le tirage aléatoire stratifié permet d'obtenir, un estimateur ayant une variance plus faible que celui issu d'un tirage aléatoire simple pour un même nombre d'individus *n*.

Soit $U = U_1 \cup U_2 \cup ... \cup U_H$; la population *U* est partitionnée en *H* strates.

Soit N_h la taille de la strate *h* pour $h = \{1,...,H\}$; on a $N = \sum_{h=1}^{H} N_h$.

On réalise un sondage aléatoire simple à allocation proportionnelle dans chaque strate U_h, en sélectionnant dans chacune d'elles un échantillon s_h de taille n_h (où $n_h = n\frac{N_h}{N}$) avec une probabilité d'inclusion $\pi_h = \frac{n_h}{N_h} = \frac{n}{N}$.

Un **estimateur sans biais du total t_y** est donné par :

$$\hat{t}_{y,strate} = \sum_{h=1}^{H} \hat{t}_h \text{ avec } \hat{t}_h = \sum_{i \in n_h} y_i$$

Un **estimateur sans biais de la variance de l'estimateur du total** est :

$$\hat{V}\left(\hat{t}_{y,strate}\right) = \sum_{h=1}^{H} {N_h}^2 \left(1 - \frac{n_h}{N_h}\right)\frac{s_h^2}{n_h} \text{ avec } s_h^2 = \frac{1}{n_h - 1}\sum_{i \in s_h}(y_{hi} - \hat{\bar{y}}_h)^2$$

Un **estimateur sans biais d'une moyenne ou d'une prévalence** $\hat{\bar{y}}$ (quand la taille *N* de la population est connue) est donné par :

$$\hat{\bar{y}}_{strate} = \sum_{h=1}^{H} \frac{N_h}{N}\hat{\bar{y}}_h$$

Un **estimateur sans biais de la variance de l'estimateur d'une moyenne ou d'une prévalence** est donné par:

$$\hat{V}(\hat{\bar{y}}_{strate}) = \sum_{h=1}^{H} \left(\frac{N_h}{N}\right)^2 \left(1 - \frac{n_h}{N_h}\right)\frac{s_h^2}{n_h}$$

II.1.2.2.1.3 Sondage à probabilités inégales

Dans un sondage aléatoire simple ou dans un sondage stratifié aléatoire simple à allocation proportionnelle, chaque individu a la même probabilité d'inclusion.

Il peut être néanmoins intéressant de tirer au sort les individus de l'échantillon avec une probabilité d'inclusion différente pour chaque individu.

Par exemple, dans un sondage stratifié, si on suppose qu'une strate s_a est moins homogène que les autres, on peut définir une probabilité d'inclusion plus élevée pour les individus appartenant à la strate s_a ; ainsi, le nombre d'individus tirés au sort dans la strate la moins homogène sera plus grand que celui des strates plus homogènes, et on peut espérer ainsi avoir une variance équivalente dans chaque strate.

Un autre exemple intéressant est le sondage à probabilités inégales et appelé proportionnel à une « taille » dans la littérature sur les sondages : supposons que l'on cherche à estimer le total d'une variable quantitative *y* et qu'on dispose par ailleurs d'une information auxiliaire quantitative *X* dans la base de sondage. Il est alors possible de réaliser un sondage à probabilités inégales tel que :

$$\pi_i = n\frac{x_i}{\sum_{i=1}^{N} x_i}.$$

Un estimateur sans biais du total est alors : $\hat{t}_{prop} = \sum_{i\in s}\frac{y_i}{\pi_i}$.

Si *y* et *x* sont fortement corrélées, le rapport $\frac{y_i}{\pi_i}$ est proche d'une même valeur quel que soit *i*. L'estimateur $\hat{t}_{prop}$ a donc une variance faible. Néanmoins, dans le cas où *y* et *x* sont indépendants, la variance de $\hat{t}_{prop}$ peut être au contraire très grande dans certains cas. Ce point sera développé en II.2.3.4.1.

Supposons que l'on souhaite étudier la gravité des pathologies de patients hospitalisés. Si on fait l'hypothèse que plus la taille d'un hôpital est grand, plus il admet des patients avec des pathologies graves, il est dans ce cas intéressant de réaliser un sondage proportionnel à la taille de l'hôpital.

II.1.2.2.1.4 Expression unifiée de l'estimateur d'un total issu d'un sondage aléatoire simple ou d'un sondage aléatoire stratifié

Les estimateurs de totaux issus d'un sondage aléatoire simple sans remise, d'un sondage stratifié ou d'un sondage à probabilités inégales sont des cas particuliers de l'estimateur de Horvitz-Thompson (62), appelé aussi estimateur par expansion ou par dilatation qui s'écrit :

$$\hat{t}_{y,\pi} = \sum_{i \in s} \frac{y_i}{\pi_i} = \sum_{i \in s} d_i \, y_i \text{ avec } d_i = \frac{1}{\pi_i}.$$

Cet estimateur est sans biais.

Démonstration : en annexe

On dit que d_i est le poids de sondage de l'individu *i*.

La variance de l'estimateur de Horvitz-Thompson est égale à :

$$V(\hat{t}_{y,\pi}) = \sum_{i \in U} \sum_{j \in U} (\pi_{ij} - \pi_i \pi_j) \frac{y_i}{\pi_i} \frac{y_j}{\pi_j}.$$

II.1.2.2.2 Un exemple de sondage non élémentaire : le sondage en deux phases

Un sondage en deux phases est défini par le fait de tirer au sort un premier échantillon, puis de tirer au sort un deuxième échantillon dans le premier échantillon. En pratique, ce type d'enquêtes n'a d'intérêt que dans le cas des sondages en deux phases stratifiés.

Les sondages en deux phases stratifiés sont utilisés lorsqu'on cherche à réaliser un sondage stratifié, mais que les variables sur lesquelles on souhaite réaliser cette stratification ne sont pas disponibles dans la base de sondage. Dans ce qui suit, nous parlerons de sondage en deux phases pour évoquer les sondages en deux phases stratifiés pour simplifier la lecture.

En santé publique, il est en général utilisé pour des enquêtes qui combinent des objectifs descriptifs et analytiques dans lesquelles les variables d'intérêt sont des « caractéristiques » rares. Il a été appliqué par exemple dans l'enquête Cocon dont un des objectifs était de décrire les pratiques contraceptives et les échecs de contraception, et où des femmes ayant eu une interruption volontaire de grossesse dans les 5 ans devaient être sur-échantillonnées (76). Il a par ailleurs était appliqué dans l'enquête handicap santé 2008 où des personnes présentant un handicap ou une incapacité étaient sur-échantillonnées en deuxième phase selon la gravité de leur handicap (11).

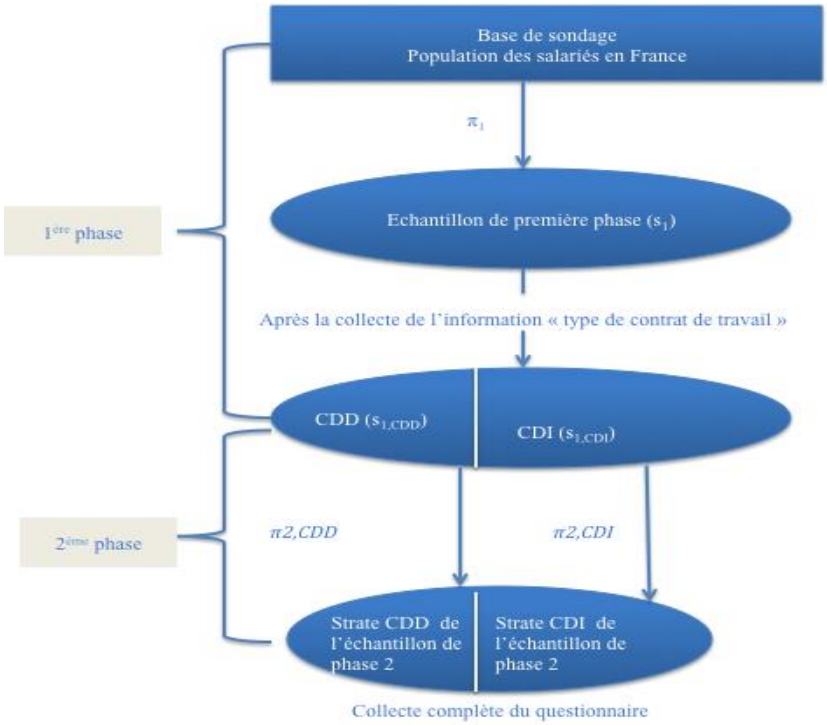
Le sondage en deux phases consiste dans un premier temps à tirer au sort un premier échantillon aléatoire avec une probabilité d'inclusion $\pi_{1,i}$ et de recueillir auprès de cet échantillon la variable X sur laquelle on veut stratifier. Cet échantillon est appelé échantillon de première phase. Dans un deuxième temps, une fois que le recueil de la variable X a été réalisé dans ce premier échantillon, un tirage au sort stratifié sur celle-ci est effectué sur cet échantillon de première phase avec des probabilités d'inclusion différentes dans chaque modalité de la variable X (par exemple, $\pi_{2,X_1,i}$ et $\pi_{2,X_2,i}$ si X est constituée de deux modalités). Ce deuxième échantillon est appelé échantillon de deuxième phase. Il est important de noter qu'à chaque phase de l'enquête, les probabilités d'inclusion sont connues *a priori*, car définies par la personne ayant réalisé le plan de sondage.

Pour l'illustrer, prenons l'exemple fictif suivant. On souhaite réaliser une enquête auprès de travailleurs salariés en France sur leur insécurité d'emploi ; on souhaite stratifier sur le type de contrat (CDD ou CDI) afin de pouvoir sur-échantillonner les travailleurs en CDD. En effet, il est préférable que l'échantillon sur lequel les analyses explicatives seront effectuées contienne suffisamment de personnes en insécurité d'emploi pour disposer d'une bonne puissance statistique. Comme on sait que l'insécurité d'emploi est liée au fait d'être en CDD,

il est pertinent de surreprésenter les personnes en CDD. Supposons que nous disposions d'une base de sondage des salariés en France mais que celle-ci ne contienne pas d'information sur le type de contrat. Le sondage en deux phases est dans ce cas un plan de sondage adapté.

Dans une première phase, on tire au sort un échantillon (s_1) avec une probabilité d'inclusion $\pi_{1,i}$. On recueille auprès de cet échantillon l'information sur le type de contrat. Une fois que les données ont été recueillies, dans une deuxième phase de l'enquête, on tire au sort un sous-échantillon $(s_{2,CDD})$ de personnes ayant déclaré être en CDD dans l'enquête de première phase avec une probabilité d'inclusion $\pi_{2,CDD,i}$ et un sous-échantillon $(s_{2,CDI})$ de personnes ayant déclaré être en CDI dans l'enquête de première phase avec une probabilité d'inclusion $\pi_{2,CDI,i}$.

Figure II-2 : Plan de sondage de l'exemple fictif



Soient $(X_1,\ldots,X_g,\ldots,X_G)$ les modalités de la variable X recueillie dans l'échantillon de première phase et qui seront utilisées pour stratifier le tirage au sort pour l'échantillon de deuxième phase.

Un estimateur sans biais du total, appelé estimateur par double dilatation ou par double expansion est donné par :

$$\hat{t}_{y,2phases} = \sum_{i\in s_2} \frac{y_i}{\pi_{1,i}\pi_{2,i}} \text{ où } \hat{t}_{y,2phases} = \sum_{g=1}^{G} \sum_{i\in s_{2g}} \frac{y_i}{\pi_{1,i}\pi_{2,g,i}}.$$

Un estimateur asymptotiquement sans biais de la variance de l'estimateur du total est :

$$\hat{V}\left(\hat{t}_{y,2phases}\right) = \sum_{i\epsilon s_1}\sum_{j\epsilon s_1} \frac{\Delta_{1,ij}}{\pi_{1,ij}} y_i y_j + \sum_{i\epsilon s_1}\sum_{j\epsilon s_2} \frac{\Delta_{1,ij}}{\pi_{1,ij}} \frac{1}{\pi_{2,j}} y_i y_j + \sum_{i\epsilon s_2}\sum_{j\epsilon s_1} \frac{\Delta_{1,ij}}{\pi_{1,ij}} \frac{1}{\pi_{2,i}} y_i y_j$$

$$+ \sum_{i\epsilon s_2}\sum_{j\epsilon s_2} \frac{\Delta_{1,ij}}{\pi_{1,ij}} \frac{1}{\pi_{2,ij}} y_i y_j + \sum_{i\epsilon s_2}\sum_{j\epsilon s_2} \frac{\Delta_{2,ij}}{\pi_{1,i}\pi_{1,j}\pi_{2,ij}} y_i y_j$$

avec $\Delta_{1,ij} = \frac{\pi_{1,ij} - \pi_{1,i}\pi_{1,j}}{\pi_{1,i}\pi_{1,j}}$ et $\Delta_{2,ij} = \frac{\pi_{2,ij} - \pi_{2,i}\pi_{2,j}}{\pi_{2,i}\pi_{2,j}}$

II.1.2.3 Amélioration des estimateurs par calage

Dans cette partie, pour simplifier la présentation, on se place dans le cadre de l'estimation d'un total.

Dans les conditions idéales où la seule erreur rencontrée dans une enquête est l'erreur d'échantillonnage, le système de pondération issu du plan de sondage garantit que, quelle que soit la variable d'intérêt étudiée, on peut obtenir un estimateur sans biais d'un total.

D'après G. Chauvet et D. Haziza (23), « Le calage consiste à ajuster les poids de sondage des unités (ici des individus) de manière à ce que les estimations de totaux (ou de moyenne) puissent coïncider avec les totaux (ou les moyennes) connus dans la population et disponibles grâce à une source externe (recensement, bases administratives, enquête de référence). » Dans le cas d'un sondage sans erreurs autres que celle due à l'échantillonnage, l'intérêt du calage

est double : il permet « d'assurer la cohérence entre les estimations issues d'une enquête et les totaux connus au niveau de la population et d'améliorer la précision des estimateurs. »

Soi tX_{tj} le total de la variable X_j dans la population :

$$X_{tj} = \sum_{i \in U} x_{ji}.$$

En général, à cause des fluctuations d'échantillonnage, $\hat{X}_{tj} = \sum_{i \epsilon s} d_i x_{ji} \neq X_{tj}$ avec d_i poids de sondage de l'individu i tel que $d_i = \frac{1}{\pi_i}$. On obtient donc des incohérences entre les totaux connus dans la population et leurs estimations. Le calage permet de faire coïncider les estimations des totaux issus de l'enquête avec les totaux connus.

Soit $X = (X_1, \dots, X_j, \dots, X_Q)$ une matrice de Q variables appelées informations auxiliaires. Soit $x_i = (x_{1i}, \dots, x_{ji}, \dots, x_{Qi})^T$ le vecteur de dimension Q associé à l'individu i.

On suppose que le vecteur x_i est connu pour tout $i \in s$ et que le vecteur des totaux sur la population U, $X_t = (X_{t1}, \dots, X_{tj}, \dots, X_{tQ})^T$ est également connu.

Le calage consiste à modifier les poids de sondage d_i. Soit w_i le poids final associé à l'individu i après calage.

L'estimateur de calage d'un total est un estimateur linéaire donné par (28) :

$$\hat{t}_{y,C} = \sum_{i \epsilon s} w_i y_i$$

Où les poids de calage w_i :

- sont aussi proches que possible des poids avant calage d_i (pour que le biais introduit soit le plus petit possible) ;
- satisfont les équations de calage $\sum_{i \in s} w_i x_i = X_t$ (pour assurer la cohérence).

La contrainte de proximité nécessite l'introduction d'une fonction de distance à minimiser. On montre que quelle que soit la fonction de distance choisie, l'estimateur par calage est asymptotiquement sans biais.

La variance de l'estimateur par calage peut être approchée par :

$$V(\hat{t}_{y,C}) \approx \sum_{i \in U} \sum_{j \in U} (\pi_{ij} - \pi_i \pi_j) \frac{E_i}{\pi_i} \frac{E_j}{\pi_j}$$

où $E_i = y_i - x_i' B$ et $B = \left(\sum_{i \in U} c_i^{-1} x_i x_i' \right)^{-1} \sum_{i \in U} c_i^{-1} x_i y_i$.

L'expression de la variance d'un estimateur calé est semblable à celle de l'estimateur non calé donné dans le chapitre II.1.2.2.1.4 excepté que les y_i ont été remplacés par les ε_i, dont les valeurs sont beaucoup plus faibles puisqu'il s'agit de résidus. On voit donc que plus les variables auxiliaires *X* seront liées à la variable d'intérêt *Y*, plus le calage permettra de diminuer la variance de l'estimateur. Il existe plusieurs fonctions de distance qui peuvent être utilisées pour le calage mais elles ne sont pas présentées ici car, quelle que soit la fonction de distance choisie, on obtient asymptotiquement la même formule de variance.

Remarque :

Le calage peut également être utilisé pour réduire les erreurs non dues à l'échantillonnage telles que les erreurs de couverture ou les erreurs de non-réponse (cf. I.3.3.1.1 et I.3.3.1.3).

II.1.2.4 Retour sur les informations auxiliaires

Nous avons vu dans cette partie qu'on utilisait la terminologie « information auxiliaire » pour deux types d'informations disponibles différentes. Ce sont des variables disponibles :

- soit pour tous les individus de la population (cf. II.1.2.2.1.1 et II.1.2.2.1.3) ;
- soit uniquement pour les individus tirés au sort, mais dont les totaux dans la population sont connus (cf. II.1.2.3).

Une information auxiliaire est aussi nommée donnée auxiliaire, ou variable auxiliaire. Elle est définie classiquement en statistique d'enquête comme une information supplémentaire permettant soit d'être plus efficace au moment de la mise en place du plan de sondage, soit d'être plus efficace (d'obtenir une variance plus petite) une fois les données recueillies et d'assurer la cohérence entre les estimations issues de l'enquête et les valeurs connues sur une population.

Une information auxiliaire peut aussi servir à corriger certaines erreurs au moment de l'estimation (défaut de couverture ou non-réponse).

II.2 MÉTHODES POUR MINIMISER LES BIAIS DUS À LA NON-RÉPONSE TOTALE DANS LES ENQUÊTES

II.2.1 LA NON-RÉPONSE DANS LES ENQUÊTES DE SANTÉ PUBLIQUE

En épidémiologie comme dans d'autres domaines, les taux de réponse aux enquêtes diminuent de plus en plus. Dans sa revue de la littérature sur cette question, Galea avance les raisons suivantes (36) : l'augmentation du nombre d'enquêtes, la baisse de la participation à des activités sociales en général et l'augmentation de la complexité des enquêtes qui, en plus d'un questionnaire, peuvent aussi collecter des examens médicaux ou des échantillons sanguins pour effectuer des tests biologiques.

L'étude des biais de non-réponse, qui sont des biais de sélection, n'est pas nouvelle en épidémiologie. Il a été 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 (42, 45, 71, 85).

Dans les enquêtes de santé publique ayant pour objectif de produire des estimations de prévalence extrapolables à une population d'intérêt, la non-réponse est corrigée en général en faisant un calage. Dans la plupart des cas, ce sont des variables sociodémographiques qui sont utilisées (sexe, âge, département ou région) ; l'hypothèse qui est faite implicitement est que la non-réponse est complètement expliquée par ces variables. On peut citer par exemple en France une enquête de prévalence des marqueurs sériques des infections dues aux virus des hépatites B et C (86), le Baromètre santé (63), l'enquête ESPS (65). Pour l'enquête National Health Interview Survey (21), l'utilisation de paradonnées (89), qui constituent un nouveau type d'informations auxiliaires disponibles pour l'échantillon tiré au sort, est régulièrement testée. Des enquêtes présentant plusieurs vagues peuvent par ailleurs utiliser les informations

recueillies lors de la vague précédente pour corriger la non-réponse ; en France, on peut citer l'Enquête décennale santé 2003 de l'Insee (17), et au Canada l'Enquête Nationale sur la Santé de la Population qui est une enquête longitudinale (122).

Dans ce contexte, il apparait nécessaire de proposer des protocoles alternatifs pour prendre en compte du mieux possible la non-réponse dans les enquêtes transversales ou dans la première vague d'une enquête à plusieurs vagues (36, 38, 119).

Dans cette partie, nous présenterons d'un point de vue formel ce qu'est un biais de non-réponse (pour une moyenne ou une prévalence), ce qui permettra de comprendre en quoi deux approches sont possibles pour essayer de les limiter.

II.2.2 BIAIS DE NON-RÉPONSE

II.2.2.1 Définition du biais de non-réponse

Dans le cadre d'enquête par sondage, le biais lié à la non-réponse pour l'estimation d'une moyenne ou d'une prévalence peut être approché par (7) :

$$Biais(\hat{\bar{y}}) \equiv \frac{1}{\bar{\delta}}\frac{1}{N}\sum_{i\epsilon U}(\delta_i - \bar{\delta})(y_i - \bar{y}).$$

avec :

- $y = (y_1, \dots, y_N)$ le vecteur des valeurs de la variable d'intérêt dans la population *U*
- $\bar{y}$ la moyenne (ou la prévalence) de *y* dans la population *U*
- δ_i les valeurs de la probabilité de réponse dans la population *U*
- $\bar{\delta}$ la probabilité de réponse moyenne dans la population *U*
- $\hat{\bar{y}}$ la moyenne (ou la prévalence) estimée dans l'échantillon des répondants

Le biais de non-réponse dépend de la variable d'intérêt étudiée ; il s'exprime comme le produit de l'inverse de la probabilité de réponse moyenne et de la covariance entre la probabilité de réponse et la variable d'intérêt. Il y aura absence de biais de non-réponse 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 dans la population.

II.2.2.2 Typologie des non-réponses (d'après la classification de Rubin)

La classification de Rubin (108) est définie dans le cadre de la statistique classique (dans lequel on considère que la variable d'intérêt *Y* et la réponse *R* sont des variables aléatoires) et quand on s'intéresse à n'importe quel paramètre concernant la distribution entière de *Y*.

Soit *X* l'information auxiliaire mesurée.

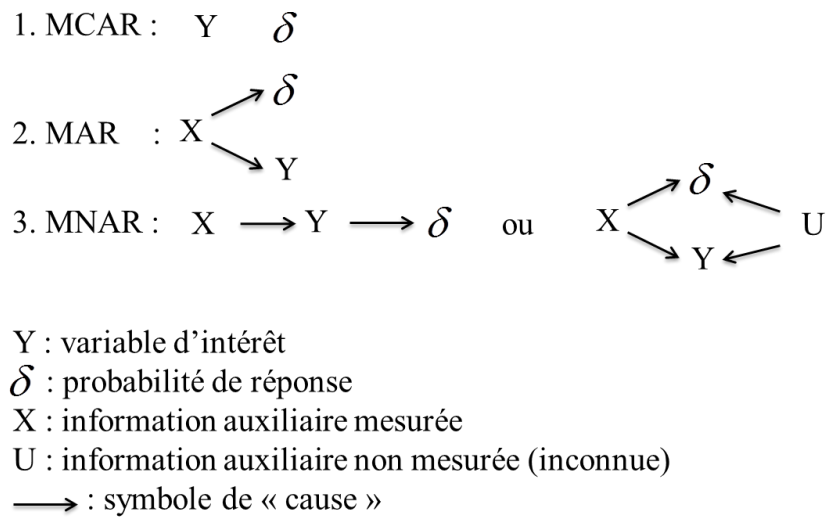
Les non-réponses peuvent être classées entre trois types :

- la non-réponse complètement aléatoire (Missing Completely At Random ou MCAR) : il y a indépendance entre la réponse *R* et la variable d'intérêt *Y* ;
- la non-réponse aléatoire (Missing At Random ou MAR) : il y a indépendance entre la réponse *R* et la variable d'intérêt *Y* conditionnellement à *X* ;
- la non-réponse non aléatoire (Missing Not At Random ou MNAR) : il n'y a pas indépendance entre *R* et *Y* conditionnellement à *X*.

Nous nous intéressons à des moyennes ou des prévalences dans le cadre de la statistique d'enquête (dans lequel les y_i et les δ_i sont des valeurs fixes). Cela conduit à réécrire une classification de Rubin adaptée à ce cadre (Figure II-3) :

- MCAR : la covariance entre la probabilité de réponse δ_i et la variable d'intérêt y_i est nulle dans la population. L'estimation de la moyenne des y_i est sans biais (Figure II-3.1) ;
- MAR : la covariance entre la probabilité de réponse δ_i et la variable d'intérêt y_i est non nulle dans la population ; néanmoins, après stratification sur *X*, cette covariance est nulle. Dans ce cas, il est possible de proposer un nouvel estimateur pour lequel à *X* fixé, la covariance entre la variable d'intérêt et la probabilité de réponse sera nulle ; autrement dit, cette prise en compte permet donc de rendre nul le biais de non-réponse initialement causé par le lien entre la probabilité de réponse et la variable d'intérêt (Figure II-3.2) ;
- MNAR : la covariance entre la probabilité de réponse δ_i et la variable d'intérêt y_i est non nulle dans la population. Cette covariance peut résulter soit d'un lien direct entre la variable d'intérêt et la probabilité de réponse, soit d'une situation MAR conditionnellement à un ensemble de variables *U* non mesurées. Dans ce cas, le biais de non-réponse ne peut être nul (Figure II-3.3).

Figure II-3 : Diagrammes de causalité des typologies de non-réponse



II.2.2.3 Diminution du biais de non-réponse : apport de données supplémentaires

Plusieurs solutions peuvent être envisagées pour prendre en compte la non-réponse.

Une première approche consiste à réaliser les analyses sur les cas complets, donc sans données supplémentaires. Cette approche est justifiée sous l'hypothèse que les données sont manquantes complètement aléatoirement (MCAR). Cette hypothèse est très forte et en général non vérifiée. Dans ce cas, l'estimateur est sans biais, mais la variance est plus grande que celle attendue initialement.

Une deuxième approche est de garder le protocole d'enquête initial et d'augmenter la taille de l'échantillon tiré au sort pour, au final avoir un échantillon de répondants de taille plus importante dans le but de diminuer la variance, mais sans pour autant augmenter le taux de réponse ; ceci peut s'avérer encore plus dangereux que de ne rien faire, car non seulement cette stratégie est inefficace en terme de correction de biais de non-réponse, mais en plus elle entraine une diminution de la variance estimée. Autrement dit, la vraie valeur du paramètre estimé a encore moins de chance d'être contenue dans l'intervalle de l'estimation.

En fait, pour orienter son choix de données supplémentaires à utiliser pour prendre en compte la non-réponse, il faut se référer à la formule du biais de non-réponse. En effet, d'après cette dernière, on peut procéder de deux façons : soit en maximisant le taux de réponse, donc, pour un échantillon de taille donné, obtenir des répondants supplémentaires, soit en corrigeant la non-réponse en utilisant des variables supplémentaires, appelées informations auxiliaires, causes communes à la fois des variables d'intérêt et de la probabilité de réponse. En pratique, ces deux approches sont souvent combinées. Cependant, elles ne sont pas toujours faciles à mettre en œuvre en raison d'une part de la difficulté d'accès à des variables supplémentaires pertinentes pour corriger la non-réponse et d'autre part des coûts supplémentaires que peuvent engendrer les efforts pour obtenir une réponse de personnes tirées au sort.

Quoi qu'il en soit, il est en général impossible d'obtenir un taux de réponse de 100% et il est très difficile de faire l'hypothèse que les variables expliquant le lien entre la probabilité de réponse et les variables d'intérêt ont toutes été prises en compte. Il est donc fort possible que quelle que soit la stratégie mise en œuvre, un biais résiduel subsiste et que l'on se retrouve dans un processus de non-réponse MNAR.

II.2.3 TRAITEMENT DE LA NON-RÉPONSE PAR REPONDÉRATION

Comme l'indique sa formulation mathématique, une manière de réduire le biais de non-réponse une fois que les données d'enquête ont été recueillies est d'ajuster sur les variables X causes communes de la propension à répondre et de la variable d'intérêt et ainsi de diminuer la covariance entre la propension à répondre et la variable d'intérêt. Pour cela, deux approches sont possibles : la première approche consiste à modéliser la propension à répondre, appelée aussi probabilité de réponse et recourir à un estimateur par repondération. La deuxième approche consiste à modéliser la variable d'intérêt et recourir à un estimateur par imputation. Quelle que soit l'approche choisie, elle nécessite la présence d'informations auxiliaires qui correspondent aux causes communes notées X dans la Figure II-3 (50, 114). La repondération consiste à augmenter le poids des répondants de façon à ce qu'ils représentent les non-répondants alors que l'imputation consiste à imputer les données manquantes de la variable d'intérêt, donc à créer un jeu de données complet partiellement « fictif ».

En pratique, on préfère avoir recours à la repondération pour traiter la non-réponse totale et à l'imputation pour traiter la non-réponse partielle (50, 114). Ces recommandations sont avant tout d'ordre pratique. En effet, contrairement à l'imputation qui nécessite autant de modélisations que de variables d'intérêt présentes dans l'enquête, il est possible de n'effectuer qu'un seul traitement lorsqu'on a recours à la repondération en modélisant la probabilité de réponse totale à l'enquête ; elle peut donc être utilisée quelle que soit la variable d'intérêt Y.

Néanmoins la repondération peut poser certains problèmes qui seront évoqués par la suite (cf. II.2.3.4.1 et II.2.3.4.3).

Comme nous nous intéressons à la non-réponse totale, ce qui suit concerne la repondération, donc le calcul des facteurs corrigeant les poids de sondage pour prendre en compte la non-réponse.

II.2.3.1 Notations

U : population d'intérêt

N : taille de la population U

Y variable d'intérêt

$y = (y_1, \ldots y_N)$ les valeurs prises par Y dans la population U

$\bar{y}$: la moyenne ou la prévalence de Y dans la population U : $\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$

ps : processus aléatoire correspondant au plan de sondage

(s) : échantillon aléatoire tiré au sort de taille fixée n

I : variable aléatoire « Sélection par tirage au sort » $I_i = \begin{cases} 1 \text{ si i sélectionnée} \\ 0 \text{ sinon} \end{cases}$

π_i : probabilité d'inclusion de l'individu i

π_{ij} : probabilité d'inclusion double des individus i et j

X_i : vecteur d'informations auxiliaires disponibles pour l'individu i appartenant à (s)

q : processus aléatoire correspondant à la non-réponse

s_r : échantillon des répondants de taille aléatoire n_r

R : variable aléatoire « Réponse à l'enquête de l'individu i » $R_i = \begin{cases} 1 \text{ si i répond} \\ 0 \text{ sinon} \end{cases}$

δ_i : probabilité de réponse (inconnue) de l'individu i

$\hat{\delta}_i$: estimation de δ_i

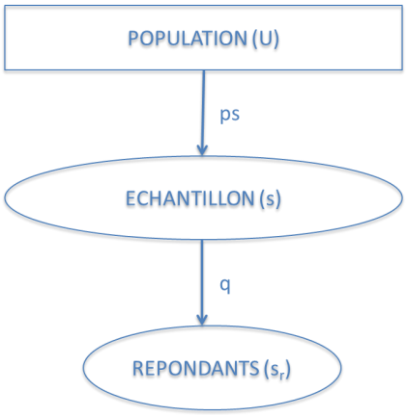
$\hat{\bar{y}}_{hypo}$: estimation de $\bar{y}$ sous l'hypothèse « *hypo* »

II.2.3.2 La non-réponse vue comme un sondage en deux phases

Le processus qui conduit de la population à l'échantillon de répondants peut être assimilé à un sondage en deux phases (Figure II-4). En effet, à l'issue du recueil de données, on peut considérer que deux aléas sont intervenus :

- au niveau de l'échantillon *(s)*, qui a été sélectionné aléatoirement selon un certain plan de sondage *(ps)*. Ici, l'aléa vient du fait d'avoir été sélectionné ou non ;
- au niveau des répondants *(s_r)*, si on fait l'hypothèse que le processus *(q)* qui conduit à une réponse est aléatoire.

Figure II-4 : Le processus de non-réponse vu comme un sondage en deux-phases



Si on se réfère à une enquête en deux phases, l'échantillon tiré au sort *(s)* correspond alors à une enquête de première phase et l'échantillon de répondants observé parmi les personnes sélectionnées par tirage au sort correspond à l'échantillon de deuxième phase *(s_r)*. La différence fondamentale avec le sondage en deux phases provient du fait qu'on ne connaît pas le processus qui a conduit de l'échantillon de première phase *(s)* à l'échantillon de deuxième phase *(s_r)*. La repondération consiste dans un premier temps à faire des hypothèses pour modéliser le processus de non-réponse et dans un deuxième temps à modifier les poids de sondage initiaux des répondants en fonction de cette modélisation. Si cette modélisation était parfaite, le processus de non-réponse pourrait être assimilé à une deuxième phase d'un sondage en deux phases, et donc permettrait d'obtenir des estimations asymptotiquement sans biais.

Par conséquent, pour un individu *i*, la probabilité de réponse δ_i doit avoir la même propriété que la probabilité d'inclusion de deuxième phase π_{2i} ; comme π_{2i}, δ_i doit être strictement positive. De plus, de la même façon que π_{2i} est l'espérance conditionnelle de la variable aléatoire I_2 « sélectionné pour l'enquête de deuxième phase », δ_i est l'espérance conditionnelle de la variable aléatoire R_i « répondant à l'enquête » :

$$R_i = \begin{cases} 1 \text{ si i répond} \\ 0 \text{ sinon} \end{cases}.$$

On fait l'hypothèse que R_i suit une loi de Bernoulli de paramètre δ_i, et que les réponses de deux individus *i* et *j* sont indépendantes. Cette hypothèse d'indépendance des réponses de deux individus est vraisemblable dans le type d'enquête étudié ici, ce qui ne serait pas le cas dans des enquêtes de type « ménage » où tous les individus d'un même ménage tiré au sort sont sollicités pour participer à l'enquête.

Si on se réfère aux différentes typologies de non-réponse, on a :

- sous l'hypothèse MCAR, δ_i est constante quel que soit l'individu *i* et l'enquête de deuxième phase correspond à un sondage aléatoire simple avec tirage de Bernoulli :

$$P(R_i = 1/y_i, X_i) = P(R_i = 1) = \delta_i.$$

- sous l'hypothèse MAR(X), δ_i est fonction des informations auxiliaires X_i qui sont les causes communes de la probabilité de réponse et de la variable d'intérêt *Y* et l'enquête de deuxième phase correspond à un sondage aléatoire stratifié sur les informations auxiliaires *X* avec tirage de Bernoulli dans chacune des strates :

$$P(R_i = 1/y_i, X_i) = P(R_i = 1/X_i) = \delta(X_i) = \delta_i.$$

- sous l'hypothèse MNAR, le processus de non-réponse n'est pas assimilable à un sondage en deux phases car δ_i dépend soit de y_i, soit d'informations auxiliaires non disponibles.

Pour rappel, dans une enquête en deux phases, un estimateur sans biais du total est donné par :

$$\hat{t}_{y,2phases} = \sum_{i \in s_2} \frac{y_i}{\pi_{1,i}\pi_{2,i}}.$$

En présence de non-réponse, un estimateur sans biais du total est donné par (plus de précisions sont données en II.2.3.3.3) :

$$\hat{t}_{y,2phases} = \sum_{i \in s_r} \frac{y_i}{\pi_i \delta_i}.$$

En pratique, la probabilité de réponse δ_i d'un individu i n'est jamais connue ; elle doit donc être estimée et son estimation sera notée par la suite $\hat{\delta}_i$.

Exemple illustratif :

On cherche à estimer le nombre total de fumeurs dans une population de taille N = 1000 (Figure II-5). Pour cela, on tire au sort dans une base de sondage et on réalise un sondage aléatoire simple avec une probabilité d'inclusion $\pi_1 = 0{,}20$ soit n = 200 personnes. La variable d'intérêt y est « Etes-vous fumeur ? ». Sur les 200 personnes sollicitées, seulement 50% participent, soit $n_r = 100$ personnes. Parmi ces 100 personnes, 40 ont répondu être fumeuses.

Sous l'hypothèse que les données sont manquantes complètement aléatoirement (MCAR), tous les individus de l'échantillon ont la même probabilité de réponse $\hat{\delta}_i$ estimée à 0,50.

Le nombre total de fumeurs dans la population est égal à :

$$\hat{t}_{y,MCAR} = \sum_{i=1}^{n_r} \frac{1}{\pi_{1,i}} \times \frac{1}{\hat{\delta}_i} \times y_i = \frac{1}{\pi_1} \times \frac{1}{\hat{\delta}} \times n_{r,fumeur} = \frac{1}{0{,}20} \times \frac{1}{0{,}50} \times 40 = 400$$

Sous l'hypothèse MCAR, on a :

$$Biais(\hat{t}_{y,MCAR}) \equiv \frac{1}{\bar{\delta}} \sum_{i \epsilon U} (\delta_i - \bar{\delta})(y_i - t_y) = 0$$

D'après la littérature, la consommation de tabac est liée au sexe. Supposons que le sexe explique complètement le lien entre la probabilité de réponse à l'enquête et le statut tabagique, donc que les données sont manquantes aléatoirement conditionnellement au sexe noté X={h pour homme, f pour femme}.

Supposons que nous disposons de l'information auxiliaire sur le sexe des personnes pour toutes les 200 personnes tirées au sort : 120 hommes et 80 femmes. Parmi les répondants, 40 sont des hommes et $n_{r,h,fumeur} = 30$ ont déclaré être fumeurs et 60 sont des femmes et

$n_{r,f,fumeur} = 10$ ont déclaré être fumeuses. La probabilité de réponse pour les hommes est donc estimée à $\hat{\delta}_h = \frac{40}{120} = 0{,}333$ et la probabilité de réponse pour les femmes est estimée à $\hat{\delta}_f = \frac{60}{80} = 0{,}75$.

Sous l'hypothèse MAR conditionnellement au sexe, le nombre total de fumeurs dans la population est estimée à :

$$\hat{t}_{MAR/sexe} = \sum_{i=1}^{n_{r,h}} \frac{1}{\pi_{1,i}} \times \frac{1}{\hat{\delta}_{h,i}} \times y_{h,i} + \sum_{i=1}^{n_{r,f}} \frac{1}{\pi_{1,i}} \times \frac{1}{\hat{\delta}_{f,i}} \times y_{f,i}$$

$$= \frac{1}{\pi_1} \times \frac{1}{\hat{\delta}_h} \times n_{r,h,fumeur} + \frac{1}{\pi_1} \times \frac{1}{\hat{\delta}_f} \times n_{r,f,fumeur}$$

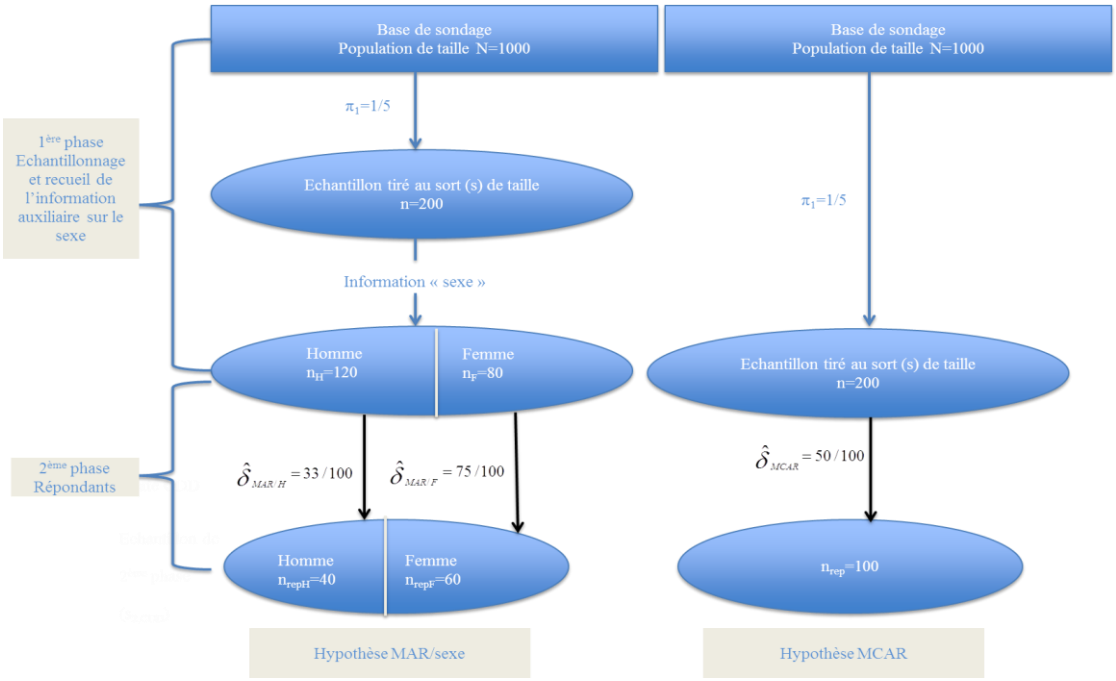
$$= \frac{1}{0{,}20} \times \frac{1}{0{,}333} \times 30 + \frac{1}{0{,}20} \times \frac{1}{0{,}75} \times 10 = 517$$

Sous l'hypothèse MAR(X), on a :

$$Biais(\hat{t}_{y,MAR/sexe}) \equiv \frac{1}{\bar{\delta}_f} \sum_{i \epsilon U_f} (\delta_{f,i} - \bar{\delta}_f)(y_{f,i} - t_{f,y}) + \frac{1}{\bar{\delta}_h} \sum_{i \epsilon U_h} (\delta_{h,i} - \bar{\delta}_h)(y_{h,i} - t_{h,y}) = 0$$

On remarque que les estimations sous les hypothèses MCAR ou MAR conditionnellement au sexe sont sensiblement différentes.

Figure II-5 : Représentation graphique du plan de sondage et des différentes hypothèses sur le processus de non-réponse



II.2.3.3 Méthodes pour estimer la probabilité de réponse

II.2.3.3.1 Introduction

Les probabilités de réponse des répondants δ_i étant inconnues, il faut les estimer correctement par $(\hat{\delta}_i)$ Quelle que soit la méthode utilisée, des informations auxiliaires X sont nécessaires pour traiter le cas MAR. Pour qu'elles soient pertinentes pour corriger la non-réponse, ces informations auxiliaires doivent être des causes communes X de la probabilité de réponse et de la variable d'intérêt. Si elles permettent de prendre en compte complètement la covariance entre la probabilité de réponse et la variable d'intérêt, le processus de données manquantes est aléatoire conditionnellement aux X (MAR/X).

Il existe deux familles de méthodes de repondération pour corriger la non-réponse : la repondération basée sur l'inverse de la probabilité de réponse (IPW) et le calage.

II.2.3.3.2 Estimation de la probabilité de réponse

II.2.3.3.2.1 Information auxiliaire nécessaire

Les informations auxiliaires doivent être disponibles pour l'ensemble des individus tirés au sort.

Elles peuvent être issues de la base de sondage, comme par exemple les variables de stratification. Elles peuvent également être relatives au processus de collecte de données, telles que par exemple le nombre d'appels téléphoniques ; ces informations appelées paradonnées (72) seront discutées dans le chapitre IV.3.

Depuis quelques années, des protocoles d'enquête épidémiologique originaux permettent l'accès à des données issues de systèmes d'informations, appelées également bases médico-administratives, pour l'ensemble de l'échantillon tiré au sort, que les personnes aient répondu ou non à l'enquête (39, 131) : des données de santé de l'assurance maladie et des données de l'assurance vieillesse qui informent sur l'historique professionnel d'une personne (informations pertinentes dans les enquêtes de surveillance épidémiologique de risques professionnels).

II.2.3.3.2.2 Méthodes

Il existe de nombreuses méthodes pour estimer la probabilité de réponse ; seules quelques-unes sont présentées ici.

II.2.3.3.2.2.1 La modélisation

Si on suppose que la probabilité de réponse suit (par exemple) un modèle de régression logistique, on a :

$$\delta(X;\alpha) = \frac{\exp(\sum_{j=1}^{p} \alpha_j' X_j)}{1 + \exp(\sum_{j=1}^{p} \alpha_j' X_j)}$$

Comme $\delta(X,\alpha)$ est inconnue, on l'estime par $\hat{\delta} = \delta(X,\hat{\alpha})$, ou $\hat{\alpha}$ est l'estimateur du maximum de vraisemblance de α. Autrement dit, on prend la variable « réponse à l'enquête » comme variable dépendante que l'on explique par les variables auxiliaires *X*. A partir de ce modèle, on estime des probabilités de réponse prédites par le modèle et ce sont ces dernières qui seront utilisées comme facteurs d'ajustement de la non-réponse.

La limite de cette méthode est qu'elle repose sur un modèle ; il peut être difficile de spécifier correctement celui-ci si la taille de l'échantillon est faible par rapport au nombre d'informations auxiliaires (effet non linéaire d'une variable continue, présence d'interaction, etc.). Une mauvaise spécification du modèle pose particulièrement un problème pour les individus ayant des prédictions de probabilités de réponse faibles ; en effet, ces derniers pourront avoir une influence importante sur l'estimation de la moyenne de *y* car ils présenteront une pondération importante leur permettant de représenter tous les non-répondants ayant les mêmes caractéristiques *X* qu'eux-mêmes. De plus, que le modèle spécifié soit ou non correct, si des estimations de probabilités de réponse sont très faibles, cela peut entrainer une inflation de la variance telle que décrite en II.2.3.4.1.

II.2.3.3.2.2.2 La méthode des scores par quantile égaux

La méthode des scores par quantile égaux (30, 59, 80) se base sur les probabilités de réponse prédites par un modèle (de régression logistique par exemple) pour constituer des groupes homogènes de réponse. Elle consiste à trier, pour tous les individus échantillonnés, les valeurs prédites par le modèle de non-réponse et de les grouper en k groupes de taille égale puis de calculer, dans chacun de ces groupes, des taux de réponse observés. Il est recommandé de constituer entre 5 et 25 groupes de taille égale. $\hat{\delta}(X)$ est ensuite estimé par le taux de réponse observé dans chaque groupe.

II.2.3.3.2.2.3 La segmentation

La segmentation consiste à partitionner le fichier de données selon un processus itératif afin de former un arbre de décision (78). Le premier nœud de l'arbre contient l'ensemble des répondants et des non-répondants. Un premier embranchement est créé à partir de la variable discriminant le mieux la non-réponse selon un critère statistique (par exemple une fonction de distance). Pour chaque nœud formé, le même mécanisme est répété. Les itérations sont stoppées lorsque plus aucune variable ne s'avère significative pour expliquer la non-réponse. A la fin de ce processus, on calcule un taux de réponse observé pour chaque groupe de réponse homogène.

Plusieurs méthodes peuvent être utilisées pour déterminer la variable la plus influente pour expliquer la non-réponse. Statistique Canada utilise préférentiellement l'algorithme CHAID (Chi-Squared Automatic Interaction Detection) qui utilise comme critère statistique pour sélectionner la variable la plus influente la statistique du Chi-Deux de Pearson ; la variable dont la valeur de cette statistique est la plus élevée est utilisée pour créer un nouvel embranchement (78).

II.2.3.3.2.2.4 Avantages et inconvénients

Comme nous l'avons décrit plus haut, la modélisation explicite de la probabilité de réponse, en utilisant les données sur l'échantillon tiré au sort, pose problème si certaines estimations des probabilités de réponse sont très faibles ; en effet, on ne peut pas écarter l'hypothèse que le modèle soit mal spécifié et que ces estimations soient incorrectes. Cependant, la probabilité de réponse estimée, même si elle est issue d'un modèle paramétrique (régression logistique dans la plupart des cas), peut servir de base pour construire des groupes de réponse homogènes et se ramener à un modèle semi-paramétrique ; il a été montré (60) que la repondération par constitution de groupes de réponse homogène était robuste à une mauvaise spécification du modèle paramétrique initial, ce qui est un vrai point fort de cette technique. Les méthodes semi-paramétriques, telles que la méthode des scores et certaines méthodes non paramétriques, telles que l'algorithme CHAID, permettent par ailleurs de contrôler la volatilité des facteurs d'ajustement pour la non-réponse, donc à être moins sujettes à des inflations de variance (cf. II.2.3.4.1). Les premières sont préférées lorsque la taille de l'échantillon de répondants est faible, alors que les méthodes non-paramétriques sont plutôt préférées dans le cas contraire.

II.2.3.3.3 Estimateur d'un total, d'une moyenne ou d'une prévalence

Quelle que soit la méthode choisie pour estimer la probabilité de réponse $\hat{\delta}_i$, on peut écrire que celle-ci est une fonction des variables auxiliaires X_i et d'un vecteur de paramètres qu'on note $\hat{\alpha}$ où $\hat{\alpha}$ est une fonction des (R_i, X_i) pour $i=\{1,\ldots,n\}$.

i) Estimateur d'un total

Nous avons vu que si les probabilités de réponse δ_i sont connues, on peut se ramener à un plan de sondage en deux phases. Un estimateur sans biais du total est :

$$\hat{t}_{y,MAR(X),pond} = \sum_{i \in s_r} \frac{y_i}{\pi_i \delta_i}.$$

En réalité, les probabilités de réponse δ_i ne sont pas connues et sont estimées (cf. partie précédente) par $\hat{\delta}_i$. Sous l'hypothèse que le modèle pour estimer δ_i est correctement spécifié, un estimateur asymptotiquement sans biais du total est :

$$\hat{\hat{t}}_{y,MAR(X),pond} = \sum_{i \in s_r} \frac{y_i}{\pi_i \hat{\delta}_i}.$$

Démonstration en annexe

ii) Estimateur d'une moyenne ou d'une prévalence

Sous l'hypothèse que δ_i est correctement estimé par $\hat{\delta}_i$,

Si la taille de la population *N* est connue, un estimateur asymptotiquement sans biais de la moyenne ou de la prévalence de *y* est :

$$\hat{\bar{y}}_{MAR(X),pond} = \frac{1}{N} \sum_{i \in s_r} \frac{y_i}{\pi_i \hat{\delta}_i}.$$

Si la taille de la population *N* est inconnue, un estimateur approximativement sans biais de la moyenne ou de la prévalence de *Y* est :

$$\hat{\bar{y}}_{MAR(X),pond} = \frac{\sum_{i \in s_r} \frac{y_i}{\pi_i \hat{\delta}_i}}{\sum_{i \in s_r} \frac{1}{\pi_i \hat{\delta}_i}}.$$

II.2.3.3.4 Estimateur de la variance de l'estimateur d'un total, d'une moyenne ou d'une prévalence

Dans le cas où δ_i est connu, un estimateur asymptotiquement sans biais de la variance de l'estimateur du total est :

$$\hat{V}\left(\hat{t}_{y,pond}\right) = \sum_{i \in s_r} \sum_{j \in s_r} \frac{(\pi_{ij} - \pi_i \pi_j)}{\pi_i \pi_j} \frac{1}{\pi_{ij} \delta_i \delta_j} y_i y_j + \sum_{i \in s_r} \frac{1}{\pi_i^2} \frac{(1 - \delta_i)}{\delta_i^2} y_i^2$$

Démonstration : en annexe

Un estimateur asymptotiquement sans biais de la variance de l'estimateur de la moyenne ou de la prévalence quand δ_i est connu et la taille de la population *N* est connue est :

$$\hat{V}\left(\hat{\bar{y}}_{MAR(X),pond}\right) = \frac{1}{N^2} \left\{ \sum_{i \in s_r} \sum_{j \in s_r} \frac{(\pi_{ij} - \pi_i \pi_j)}{\pi_i \pi_j} \frac{1}{\pi_{ij} \delta_i \delta_j} y_i y_j + \sum_{i \in s_r} \frac{1}{\pi_i^2} \frac{(1 - \delta_i)}{\delta_i^2} y_i^2 \right\}.$$

II.2.3.4 Discussion sur la repondération

II.2.3.4.1 IPW et calage

En l'absence d'informations auxiliaires individuelles disponibles à la fois chez les répondants et les non-répondants, il n'est plus possible de recourir à des méthodes de repondération par IPW. Cette configuration est très fréquente en pratique ; dans ce cas, il est possible de recourir au calage pour corriger la non-réponse si, pour des variables collectées dans l'enquête, on dispose des totaux de ces variables sur la population cible. Dans les enquêtes en population générale, on utilise classiquement les variables sociodémographiques recueillies lors de recensement comme variables de calage (63, 128) (par exemple pour le sexe le nombre total d'hommes et le nombre total de femmes dans la population source).

Cependant, certaines propriétés des estimateurs par calage ne sont plus valides en présence de non-réponse. En particulier, contrairement au cas complet, les fonctions de distance ne sont plus équivalentes. Ainsi, en présence de non-réponse, le calage correspond à une modélisation implicite de la probabilité de réponse qui dépend de la fonction de distance utilisée. Si la fonction de distance utilisée ne conduit pas à une modélisation correcte de la non-réponse, il peut subsister un biais résiduel de non-réponse.

Par ailleurs, pour les repondérations par IPW, on peut sélectionner les informations auxiliaires nécessaires ; la segmentation effectue naturellement cette sélection, et la modélisation permet de construire le modèle en testant statistiquement le lien entre la probabilité de réponse et les informations auxiliaires. Par contre, on ne peut pas procéder de la même façon avec le calage, car la modélisation est implicite et les variables à intégrer doivent donc être choisies *a priori*.

Enfin, l'utilisation de certaines distances peut poser des problèmes de convergence ; les méthodes linéaires ne posent pas de problème de convergence, mais elles peuvent conduire à des pondérations négatives, ce qui est le signe d'une mauvaise estimation implicite de la probabilité de réponse. Il faut alors utiliser d'autres méthodes qui, pour converger, demandent d'utiliser un nombre d'informations auxiliaires restreint.

Pour toutes ces raisons, pour minimiser les biais de non-réponse, il est conseillé d'utiliser la repondération par IPW plutôt que par calage lorsque c'est possible.

En pratique, lorsqu'on dispose de données individuelles sur les répondants et les non-répondants et de totaux sur la population, il est conseillé de corriger la non-réponse par repondération IPW, puis d'effectuer un calage. Le calage permet en effet, de fournir des estimations cohérentes avec des statistiques déjà connues et de potentiellement réduire la variance des estimations produites. D'autre part, le calage pourra être utile pour corriger les erreurs de couverture si ce type d'erreur existe dans l'enquête (cf. II.1.2.3).

II.2.3.4.1 Inflation de la variance

La principale limite de la repondération est qu'elle peut conduire inutilement à des estimations de variance très grandes lorsque les poids corrigés pour la non-réponse sont très dispersés et non corrélés aux variables d'intérêt (104). C'est pour cette raison qu'il est vivement conseillé de modéliser la probabilité de réponse par des variables *X* associées à la fois à la non-réponse et aux variables d'intérêt, donc de ne pas inclure dans le modèle des variables qui, bien que prédictrices de la non réponse, ne soient associées à aucune variable d'intérêt. D'autre part, il faut noter que l'inflation de la variance est particulièrement importante lorsque la variable d'intérêt est quantitative ; ce phénomène posera donc moins de problème dans le cas d'estimations de prévalence.

Le phénomène d'inflation de la variance est bien connu en statistique d'enquête car il a été étudié dans le cadre des enquêtes par sondage avec probabilités inégales proportionnelles à une taille (cf. II.1.2.2.1.3). Dans ce contexte, si la variable d'intérêt *Y* est corrélée à la variable auxiliaire *Z* grâce à laquelle on réalise un tirage proportionnel à sa taille, la variabilité de l'estimateur de la moyenne de *y* sera petite par construction. Dans le cas contraire, sa variabilité peut être très grande si les *Z* sont très dispersés avec des valeurs extrêmes petites. On peut se référer à l'exemple de Basu pour une explication appliquée (4).

II.2.3.4.2 Choix des informations auxiliaires

Le choix des informations auxiliaires est primordial lorsqu'on cherche à minimiser les biais de non-réponse. Pour Rizzo, Kalton et Brick (103), « le choix des variables auxiliaires est important, probablement plus important que celui de la méthode de pondération ». Ce point est clairement fondamental.

Quelle que soit la technique utilisée aussi sophistiquée soit elle, si les informations auxiliaires utilisées ne sont pas à la fois causes de la propension à répondre et de la variable d'intérêt les biais de non-réponse ne seront pas réduits (dans certains cas une amplification du biais pourrait même se produire cf. II.2.3.4.3) et la variance des estimateurs pourrait être dégradée (cf. II.2.3.4.1).

En santé publique, le choix *a priori* des informations auxiliaires à inclure doit être fait en collaboration directe avec l'épidémiologiste qui a une expertise sur les effets de sélection dans les enquêtes de sa spécialité.

Le choix des informations auxiliaires à inclure dans le modèle de non-réponse totale peut être assez simple si le nombre de variables d'intérêt de l'enquête est restreint et si elles sont corrélées entre elles. En général, le nombre d'informations auxiliaires disponibles n'est pas très élevé. Néanmoins, avec le développement de protocoles originaux incluant le recueil de paradonnées ou l'appariement avec des bases de données issues de systèmes d'information existants, la quantité d'informations auxiliaires peut être gigantesque. Il se pose alors deux nouveaux problèmes : le premier est de rendre exploitable ces informations qui nécessitent en général un énorme travail de nettoyage et de construction d'indicateurs pertinents (129), et le deuxième de réduire de manière pertinente toutes ces informations.

Dans le cas où les variables d'intérêt sont nombreuses et représentent des dimensions différentes, il faut inclure des informations auxiliaires avec parcimonie, en excluant celles qui ne sont liées à aucune des variables d'intérêt afin que l'inclusion d'un trop grand nombre d'informations auxiliaires inutiles, ne produise des effets indésirables de la repondération cités plus haut (cf. II.2.3.4.1).

Une façon classique de réduire l'information de manière pertinente est de constituer des groupes de variables (par exemple, sociodémographiques, relatives à la santé, relatives au

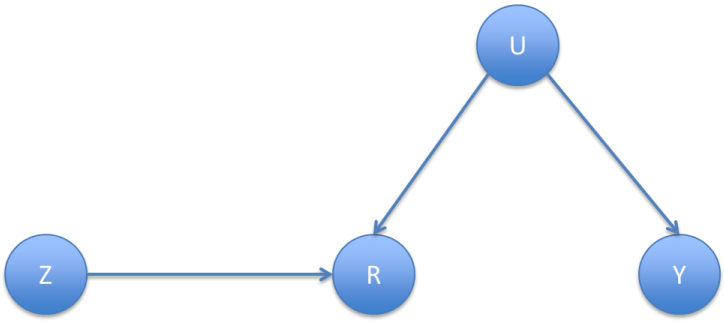
travail) et de construire des modèles de non-réponse pour chaque groupe de variables puis à l'aide de ces modèles de construire un modèle final. Cette procédure permet de mesurer l'apport des différents groupes d'informations auxiliaires dans la réduction de biais de non-réponse.

Il est aussi possible d'avoir recours à des approches de statistiques descriptives adaptées à la réduction d'informations telles que les analyses en composantes principales ou les analyses en composantes multiples (112) ; certains auteurs proposent d'utiliser ces méthodes dans le cas où le nombre d'informations auxiliaires pour réaliser un calage est important (40).

II.2.3.4.3 Amplification du biais

La repondération est une méthode très utile pour corriger la non-réponse une fois que les données d'enquête ont été recueillies, mais elle peut entrainer une amplification du biais quand d'une part tous les facteurs explicatifs communs de la non-réponse et de la variable d'intérêt n'ont pas été pris en compte et d'autre part que les variables utilisées pour modéliser la non-réponse sont très prédictrices de la non-réponse et peu associées à la variable d'intérêt (Figure II-6) (90, 96). Cependant, l'amplification du biais a en général une valeur assez faible, ce qui en fait un phénomène peu connu en épidémiologie.

Figure II-6 (90) : Diagramme de causalité montrant des facteurs explicatifs non observés (U) de la réponse (R) et de la variable d'intérêt (Y) et une variable observée (Z) explicative de la réponse mais pas à la variable d'intérêt.



Cette configuration peut se produire lorsqu'on ne dispose que de paradonnées pour corriger la non-réponse, puisqu'elles sont en général fortement associées à la non-réponse et peuvent être très peu liées aux variables d'intérêt. Mais elle peut éventuellement se produire lorsqu'on utilise des informations auxiliaires de qualité pour corriger la non-réponse totale ; en effet, certaines variables peuvent être fortement liées à la probabilité de réponse mais pas nécessairement liées à toutes les variables d'intérêt recueillies dans l'enquête.

II.2.3.4.4 Repondération et imputation

Comme nous l'avons vu en introduction de la partie II.2.3, l'imputation permet, comme la repondération, de diminuer les biais liés à la non-réponse et pourrait, en pratique, être utilisé pour corriger la non-réponse totale. A la différence de la repondération, qui se base sur la modélisation de la probabilité de réponse, l'imputation consiste à remplacer directement les valeurs manquantes de la variable d'intérêt par des valeurs artificielles (prédites par un modèle ou déterminées par jugement d'experts). Une fois réalisée, elle permet donc de travailler sur un jeu de données complet, contenant des valeurs observées et des valeurs imputées. Il existe de nombreuses méthodes d'imputation (déterministes ou aléatoires, prédites par un modèle ou non, simples ou multiples) (57) qui ne sont pas détaillées ici ; à chaque méthode d'imputation correspond une estimation de la variance.

Comme pour la repondération, le remplacement des valeurs manquantes par des valeurs prédites par un modèle nécessite que le processus de données manquantes soit (MAR/X) et que ces informations auxiliaires X soient disponibles. De même que pour la pondération, il faut en outre que le modèle de pondération soit correctement spécifié, il faut pour l'imputation que le modèle d'imputation soit lui aussi correctement spécifié.

On note *m(X)* l'espérance conditionnelle de *Y* sachant *X*. Sous l'hypothèse MAR/X, on a :

$$m(X) \equiv E(Y/X) = E(Y/X, R = 1)$$

Supposons que l'espérance de la variable d'intérêt *Y* soit correctement modélisée par un modèle paramétrique $m(X, \beta)$, alors un estimateur sans biais de la prévalence $\bar{y}$ de *Y* est :

$$\hat{\bar{y}}_{MAR(X),imp} = \frac{1}{N}\sum_{i=1}^{n}\frac{1}{\pi_i}(r_i y_i + (1 - r_i)m(x_i; \hat{\beta}))$$

où $\hat{\beta}$ est l'estimateur du maximum de vraisemblance de β.

Si la variable *Y* est binaire, on peut par exemple choisir le modèle de régression logistique :

$$E(Y/X, R = 1) = m(X; \beta) = \frac{\exp(\sum_{j=1}^{p}\beta_j \ x_j)}{1 + \exp(\sum_{j=1}^{p}\beta_j \ x_j)}$$

On peut noter que si on s'intéresse à l'estimation d'une moyenne ou d'un total, des imputations simples sont suffisantes alors que des imputations multiples seraient nécessaires si le paramètre d'intérêt représentait le lien entre deux variables (par exemple un odds-ratio) ce qui est le cas en épidémiologie analytique (61, 109) , domaine dans lequel les imputations multiples sont très largement utilisées. Ces dernières sont cependant peu recommandées dans les enquêtes avec plan de sondage car elle peuvent conduire à des estimations de variance biaisées (69).

II.2.4 MAXIMISATION DU TAUX DE RÉPONSE DANS LES ENQUÊTES

En théorie, le taux de réponse est un facteur qui influe sur le biais de non-réponse. On comprend d'ailleurs bien qu'en l'absence de non-réponse, il n'y a pas de biais de non-réponse ! Si on se réfère à la formule du biais de non-réponse, à covariance constante entre la propension à répondre et la variable d'intérêt, le biais de non-réponse est d'autant plus grand que le taux de réponse est petit. C'est pour cette raison que la façon la plus répandue de diminuer les biais de non-réponse consiste à minimiser le taux de non-réponse d'une part par manque d'informations auxiliaires, et d'autre part parce que maximiser le taux de réponse ne nécessite pas de modélisations, donc d'hypothèses sur le lien entre la variable d'intérêt et la propension à répondre. Comme la théorie des sondages permet de ne pas faire d'hypothèses en l'absence de non-réponse pour obtenir des estimations sans biais ou approximativement sans biais, on fait généralement le choix de maximiser le taux de réponse pour approcher autant que possible un taux de réponse de 100%. Dans le contexte où on cherche à atteindre un taux de réponse de 100%, on peut citer les conseils de Singleton & Straits (118) qui considèrent que dans les enquêtes avec enquêteur, des taux de réponse inférieurs à 70% sont problématiques.

II.2.4.1 Facteurs influençant le taux de réponse

La littérature référençant les facteurs influant sur le taux de réponse est riche ; elle converge sur les points suivants (82).

La thématique de l'enquête est un facteur important lié à la réponse. Par exemple, on peut s'attendre à un taux de réponse inférieur dans une enquête relative à des comportements à risque pour la santé que dans une enquête relative aux enfants, où ce sont les parents qui sont interrogés.

Pour les enquêtes avec enquêteur, l'enquêteur peut également avoir une influence sur le taux de réponse. Il faut d'une part lui assurer une bonne formation et d'autre part choisir des profils d'enquêteur qui sont en accord avec le thème de l'enquête.

Le mode de collecte de données influe sur le taux de réponse ; les enquêtes en face-à-face sont en général plus propices à un taux de réponse élevé, mais cela peut dépendre de la thématique de l'enquête. Sur par exemple des sujets sensibles, les personnes enquêtées peuvent répondre plus facilement par questionnaire auto-administré que par questionnaire en face-à-face (97).

La présentation du contexte et des objectifs de l'enquête est également à considérer pour que les personnes enquêtées trouvent un intérêt personnel ou plus altruiste à participer.

La proposition de contreparties (cadeaux,…) aux répondants peut également influer sur le taux de réponse mais peut entrainer des problèmes qui seront évoqués dans la suite du document (cf. II.3.1).

Le questionnaire en tant que tel a aussi beaucoup d'importance ; il faut qu'il soit suffisamment court pour que les personnes enquêtées veuillent y répondre et suffisamment long pour que toute l'information nécessaire puisse être collectée. Les questions doivent par ailleurs être bien formulées et la mise en page agréable pour les questionnaires auto-administrés.

Enfin, le protocole d'enquêtes, en particulier le nombre de relance, influence le taux de réponse.

Pour un protocole donné, la propension à répondre dépend également de caractéristiques individuelles, notamment de variables sociodémographiques ou socioprofessionnelles (par

exemple, les femmes, respectivement les cadres, participent aux enquêtes en moyenne plus que les hommes, respectivement les ouvriers) (45).

II.2.4.2 Enquête auprès de non-répondants

Quels que soient les efforts mis en œuvre pour maximiser le taux de réponse à une enquête, on observe en général de la non-réponse. On peut alors essayer de contacter tous les non-répondants, et tenter d'obtenir une réponse de chacun d'entre eux, mais parce que cela représente un travail long, fastidieux et onéreux, il est rarement entrepris. Plutôt que de chercher la réponse de tous les non-répondants, Hansen et Hurwitz (54) ont montré qu'à partir d'un sous-échantillon aléatoire de non-répondants, il était possible d'obtenir un estimateur sans biais d'un total ou d'une moyenne en combinant les données des répondants de l'enquête initiale et du sous-échantillon de non-répondants, à la condition que tous les non-répondants sous-échantillonnés répondent à l'enquête. Ils ont ainsi montré que sans obtenir la réponse de tous les non-répondants initiaux, il était possible d'obtenir un estimateur sans biais d'un total ou d'une moyenne lorsque la taille de la population est connue (ou asymptotiquement sans biais si la taille de la population est inconnue). Par ailleurs, la variance de l'estimateur dépend de la proportion de non-répondants sous-échantillonnés.

II.2.4.2.1 Notations

EI : enquête initiale

EC : enquête complémentaire auprès de non-répondants

s_{EI} : échantillon aléatoire de l'enquête initiale de taille n_{EI}

I_{EI} variable aléatoire « Sélection par tirage au sort à l'enquête initiale » telle que $I_{EI,i} = \begin{cases} 1 \text{ si } i \text{ sélectionnée} \\ 0 \text{ sinon} \end{cases}$

$\pi_{EI,i}$ la probabilité d'inclusion de l'individu *i* dans l'enquête initiale

$\pi_{EI,ij}$ probabilité d'inclusion double des individus *i* et *j* à l'enquête initiale

R_{EI} variable aléatoire « Réponse à l'enquête initiale » telle que $R_{EI,i} = \begin{cases} 1 \text{ si } i \text{ a répondu} \\ 0 \text{ sinon} \end{cases}$

$\delta_{EI,i}$ probabilité de réponse (inconnue) de l'individu i à l'enquête initiale et $\hat{\delta}_{EI,i}$ son estimation

$s_{EI,r}$ échantillon de répondants à l'enquête initiale et $s_{EI,nr}$ échantillon de non-répondants à l'enquête initiale. On a $s_{EI} = s_{EI,r} \cup s_{EI,nr}$.

$n_{EI,r}$ nombre de répondants à l'enquête initiale et $n_{EI,nr}$ nombre de non-répondants à l'enquête initiale. On a $n_{EI} = n_{EI,r} + n_{EI,nr}$.

s_{EC} : échantillon aléatoire de l'enquête complémentaire

n_{EC} : taille de l'échantillon de l'enquête complémentaire

I_{EC} variable aléatoire « Sélection par tirage au sort à l'enquête complémentaire » telle que $I_{EC,i} = \begin{cases} 1 \text{ si } i \text{ sélectionnée} \\ 0 \text{ sinon} \end{cases}$

$\pi_{EC/s_{EI,nr},i}$ la probabilité d'inclusion de l'individu *i* dans l'enquête *EC* ; elle sera notée par la suite pour simplifier les notations $\pi_{EC,i}$

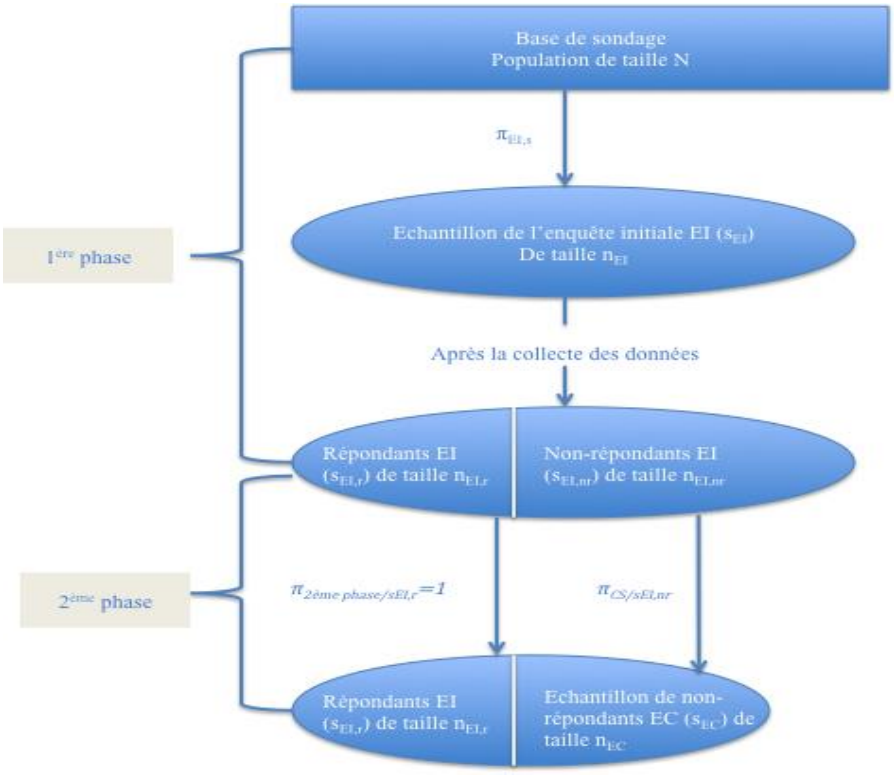
$\pi_{EC,ij}$ la probabilité d'inclusion double des individus *i* et *j* à l'enquête complémentaire

II.2.4.2.2 Principe

L'enquête en deux phases pour non-réponse (Figure II-7) est similaire à l'enquête en deux phases présentée en II.1.2.2.2.

Elle consiste dans un premier temps à réaliser une enquête par sondage classique (par exemple un sondage aléatoire simple ou un sondage stratifié) appelée enquête initiale. Après la collecte des données, l'échantillon initial est divisé en une strate de répondants et une strate de non-répondants. Dans un deuxième temps, un sous-échantillon de non-répondants est sélectionné aléatoirement puis enquêté selon un protocole construit pour obtenir un taux de réponse maximal ; cette deuxième enquête est appelée en statistique d'enquête « enquête de suivi » ou « enquête follow-up » mais sera appelée par la suite enquête complémentaire pour éviter une confusion avec les enquêtes de cohortes prospectives. Un taux de réponse de 100% à l'enquête complémentaire permettrait d'obtenir une estimation sans biais de totaux, de moyennes et de prévalences en combinant les réponses obtenues lors de l'enquête initiale et de l'enquête complémentaire. La différence entre le sondage en deux phases pour non-réponse et le sondage en deux phases « classique » vient du fait que la variable strate recueillie à l'issue de la première phase est la réalisation de la variable aléatoire R dans le premier cas, alors qu'elle est une valeur fixe dans le deuxième.

Figure II-7 : Représentation graphique d'un échantillonnage en deux phases pour non-réponse



II.2.4.2.3 Estimateur d'un total, d'une moyenne ou d'une prévalence

iii) Estimateur d'un total

Un estimateur sans biais du total est (54) :

$$\hat{t}_{y,2phases,nr} = \sum_{i \epsilon s_{EI,r}} \frac{y_i}{\pi_{EI,i}} + \sum_{i \epsilon s_{EC}} \frac{y_i}{\pi_{EI,i}\pi_{EC/s_{EI,nr},i}}$$

Démonstration : en annexe

iv) Estimateur d'une moyenne ou d'une prévalence

Si la taille de la population *N* est connue, un estimateur asymptotiquement sans biais de la moyenne ou de la prévalence de y est :

$$\hat{\bar{y}}_{2phases,nr} = \frac{1}{N}\left(\sum_{i\epsilon s_{EI,r}} \frac{y_i}{\pi_{EI,i}} + \sum_{i\epsilon s_{EC}} \frac{y_i}{\pi_{EI,i}\pi_{EC/s_{EI,nr},i}}\right)$$

Si la taille de la population *N* est inconnue, un estimateur asymptotiquement sans biais de la moyenne ou de la prévalence de y est :

$$\hat{\bar{y}}_{2phases,nr} = \frac{\sum_{i\epsilon s_{EI,r}} \frac{y_i}{\pi_{EI,i}} + \sum_{i\epsilon s_{EC}} \frac{y_i}{\pi_{EI,i}\pi_{EC/s_{EI,nr},i}}}{\sum_{i\epsilon s_{EI,r}} \frac{1}{\pi_{EI,i}} + \sum_{i\epsilon s_{EC}} \frac{1}{\pi_{EI,i}\pi_{EC/s_{EI,nr},i}}}$$

II.2.4.2.4 Estimateur de la variance de l'estimateur d'un total, d'une moyenne ou d'une prévalence

$\hat{V}\left(\hat{t}_{y,2phases,nr}\right)$

$$= \sum_{i\epsilon s_{EI,r}} \sum_{j\epsilon s_{EI,r}} \frac{\Delta_{EI,ij}}{\pi_{EI,ij}} y_i y_j + \sum_{i\epsilon s_{EI}} \sum_{j\epsilon s_{EC}} \frac{\Delta_{EI,ij}}{\pi_{EI,ij}} \frac{1}{\pi_{EC,j}} y_i y_j$$

$$+ \sum_{i\epsilon s_{EC}} \sum_{j\epsilon s_{EI}} \frac{\Delta_{EI,ij}}{\pi_{EI,ij}} \frac{1}{\pi_{EC,i}} y_i y_j + \sum_{i\epsilon s_{EC}} \sum_{j\epsilon s_{EC}} \frac{\Delta_{EI,ij}}{\pi_{EI,ij}} \frac{1}{\pi_{EC,ij}} y_i y_j$$

$$+ \sum_{i\epsilon s_{EC}} \sum_{j\epsilon s_{EC}} \frac{\Delta_{EC,ij}}{\pi_{EI,i}\pi_{EI,j}\pi_{EC,ij}} y_i y_j$$

avec $\Delta_{EI,ij} = \frac{\pi_{EI,ij} - \pi_{EI,i}\pi_{EI,j}}{\pi_{EI,i}\pi_{EI,j}}$ et $\Delta_{EC,ij} = \frac{\pi_{EC,ij} - \pi_{EC,i}\pi_{EC,j}}{\pi_{EC,i}\pi_{EC,j}}$

Démonstration : en annexe

II.2.4.2.5 Mises en garde

En pratique, il est très souvent impossible d'obtenir un taux de réponse de 100% à l'enquête complémentaire. Néanmoins, les enquêtes complémentaires auprès de non-répondants fournissent des données supplémentaires qui permettent de diminuer les biais de non-réponse. Pour cela, il faut être vigilant sur plusieurs points.

Il faut choisir un protocole pour l'enquête complémentaire qui permet de maximiser autant que possible le taux de réponse. Cependant, ce n'est pas toujours évident d'obtenir des taux de réponse élevés puisqu'on enquête des non-répondants : cela dépend surtout du schéma d'enquête initial et c'est pourquoi les enquêtes complémentaires sont particulièrement recommandées lorsque le mode de recueil des données pour l'enquête initiale se fait par autoquestionnaire postal (54, 79) ; les enquêtes complémentaires permettent de pallier le faible taux de réponse généralement rencontré, à la condition d'utiliser un mode de recueil de données plus performant lors de l'enquête complémentaire d'un point de vue du taux de réponse. Ainsi, pour que ce type d'enquête fonctionne, il est conseillé d'une part de choisir un questionnaire restreint sur les questions essentielles lorsque le questionnaire de l'enquête initiale est long et d'autre part un mode de collecte des données plus propice à la réponse que celui utilisé pour l'enquête initiale. Cela se traduit généralement par un coût important pour les enquêtes complémentaires, et dans ce cas-là par des tailles d'échantillon à l'enquête complémentaire qui sont relativement modestes.

Cependant, différents modes de collecte de données à l'enquête initiale et à l'enquête complémentaire peuvent introduire des biais de mesure (13). Le fait d'insister auprès de non-répondants initiaux peut également entrainer des biais de mesure, les non-répondants initiaux étant potentiellement moins intéressés par l'enquête ou plus réticents donc pouvant potentiellement répondre avec moins d'application que les répondants initiaux (9).

II.3 LA PROPENSION À RÉPONDRE ET DEUX COMPOSANTES DE L'ERREUR TOTALE : LE BIAIS DE NON-RÉPONSE ET LE BIAIS DE MESURE

Il est maintenant intéressant de s'interroger sur les liens entre propension à répondre et erreur totale. La propension à répondre δ_i est synonyme de probabilité de réponse à une enquête donnée et est définie pour chaque individu. Le taux de réponse correspond à la propension à répondre moyenne $\bar{\delta}$ dans la population.

L'idée qu'il faut maximiser autant possible le taux de réponse aux enquêtes est encore largement dominante malgré certains articles récents remettant en cause cette façon de procéder. Comme nous l'avons rapidement évoqué plus haut, maximiser le taux de réponse peut être, en pratique, inefficace pour minimiser les biais de non-réponse (49) et peut potentiellement augmenter les biais de mesure. Cette partie se propose d'étayer ce propos.

II.3.1 PROPENSION À RÉPONDRE ET BIAIS DE NON-RÉPONSE

De plus en plus d'articles récents suggèrent que la baisse des taux de réponse n'altère pas nécessairement les estimations issues d'une enquête (49) ; après avoir résumé les résultats de 30 articles ayant produit plus de 200 estimations, Groves conclut que la plupart des variations des biais de non-réponse surviennent à l'intérieur d'une même enquête mesurant des variables d'intérêt différentes plus qu'entre différentes études ayant des taux de réponse variant de 15% à 70%. Autrement dit, maximiser le taux de réponse global sans autre critère semble peu efficace en termes de biais de non-réponse. Par ailleurs, chercher à maximiser autant que possible le taux de réponse à une enquête peut s'avérer contre-productif. L'exemple tiré de l'article de Merkle et coll. (87) est significatif en ce sens. Il est issu d'une enquête pré-électorale qui avait pour objectif d'estimer les intentions de vote aux élections présidentielles américaines de 2008. Sans contrepartie, la participation des sympathisants démocrates et des

sympathisants républicains à l'enquête était équivalente. En revanche, en offrant un stylo en cas de participation à l'enquête, les sympathisants démocrates participaient plus à l'enquête que les sympathisants républicains, ce qui conduisait finalement à une surestimation des intentions de vote envers le candidat démocrate. Le recours à un cadeau reposait sur l'hypothèse qu'à covariance constante, le taux de réponse pouvait augmenter, avec en corollaire, une diminution du biais. L'exemple ici montre que le recours au stylo a augmenté le taux de réponse, ce qui était attendu, mais a également augmenté la covariance entre la probabilité de réponse et la variable d'intérêt, et a ainsi augmenté le biais de non-réponse, malgré l'augmentation du taux de réponse.

Dans ce contexte, au lieu de viser un taux de réponse maximal global, de nouveaux protocoles d'enquête développés actuellement (51, 83) cherchent, pour un budget donné, à maximiser le taux de réponse de groupes de personnes particulières, potentiellement génératrices de biais de non-réponse. Ces protocoles d'enquête sont appelés « plan de collecte adaptative ». Ils sont définis comme une « approche adaptative où l'information disponible est utilisée pour modifier la collecte des données pour les personnes restantes » (83). Ils ont été formalisés par Groves et Heeringa (51). Deux notions sont associées à ce schéma d'enquête : la phase et la capacité d'une phase. Une phase correspond à une période de la collecte de données pendant laquelle le même protocole est utilisé (par exemple, un questionnaire postal avec plusieurs relances). La capacité d'une phase est la condition de stabilité d'une estimation dans une phase spécifique ; autrement dit, une phase a atteint sa capacité lorsqu'il n'est plus utile de continuer la collecte des données selon le protocole choisi pour cette phase (par exemple lorsqu'on considère que le nombre maximal de répondants a été atteint sous un protocole d'enquête donné pour un budget donné).

Ainsi, même si ce formalisme est assez récent (2008), la pratique du responsive design est ancienne ; il est en effet courant depuis longtemps de changer de mode de recueil de données au cours de la collecte pour enquêter le plus de personnes possibles. De la même manière, les enquêtes en deux phases pour non-réponse entrent également dans la définition du responsive design.

Depuis quelques années, des responsive design plus élaborés sont en pleine expansion. Ils nécessitent en général d'avoir précédemment réalisé une enquête similaire permettant d'accéder à des informations auxiliaires également disponibles dans l'enquête à réaliser.

Peytchev (98) propose d'utiliser l'enquête précédente pour modéliser la probabilité de réponse en fonction des informations auxiliaires, d'affecter ce modèle de prédiction à l'échantillon auprès duquel l'enquête doit être réalisée afin de disposer d'une estimation de la propension à répondre avant la collecte et de faire plus d'efforts de collecte auprès des personnes avec une propension à répondre *a priori* faible. C'est un point important car il considère qu'une personne avec une propension de réponse initialement faible peut avoir une propension à répondre plus élevée avec un processus de collecte de données différent. Il conclut cependant que cette stratégie n'est pas forcément efficace en terme de biais et qu'il vaudrait mieux identifier les groupes de personnes induisant potentiellement le plus de biais dans les estimations d'un paramètre pour une variable d'intérêt donnée plutôt que les groupes de personnes avec une probabilité de réponse basse.

L'approche adaptative proposée par Lundquist et Sarndal (83) se base sur une distance à minimiser entre la moyenne estimée chez les répondants et la moyenne estimée pour l'échantillon tiré au sort pour des variables auxiliaires expliquant la variable d'intérêt.

Schouten (116) propose l'indicateur R construit à partir d'une probabilité de réponse prédite par un modèle ; l'indicateur R est dit de bonne qualité si les probabilités de réponse prédites

ont une petite variabilité. La collecte des données peut être conduite en fonction de cet indicateur afin de réduire au maximum la variabilité des réponses prédites par ce modèle en utilisant, comme dans le cas de Peytchev, les probabilités de réponse prédites dans une enquête similaire. Cette approche est critiquée par Beaumont et col. (5) pour deux raisons : si le modèle de non-réponse inclut uniquement des variables peu associées à la non-réponse alors qu'il existe des variables associées à la probabilité de réponse et aux variables d'intérêt, les probabilités de réponse prédites par le modèle vont être peu dispersées et l'indicateur de représentativité va conduire à un indicateur R de bonne qualité alors que les « vraies » probabilités de réponse sont en fait dispersées et n'ont pas été correctement estimées. Une autre limite vient du fait que des probabilités de réponse dispersées ne signifient pas nécessairement qu'il y a un biais de non-réponse ; en effet, si les variables associées à la non-réponse ne sont pas associées à la variable d'intérêt, il n'y aura pas de biais de non-réponse.

Quelle que soit l'approche choisie, elle permet de diminuer le biais lié à la non-réponse si les informations auxiliaires utilisées expliquent le lien entre la propension à répondre et la variable d'intérêt. Mais, comme le soulignent Beaumont et coll. (5), cette approche, qui nécessite des informations auxiliaires disponibles chez les répondants et les non-répondants, n'apporte pas plus en terme de correction du biais de non-réponse qu'un ajustement pour la non-réponse par repondération qui utiliserait ces mêmes variables. Beaumont et col (5), plutôt que d'utiliser des responsive design pour diminuer les biais de non-réponse, proposent de les utiliser plutôt pour diminuer la variance liée à la non-réponse.

La stratégie optimale pour recueillir les données afin de limiter les effets du lien entre la propension à répondre et le biais de non-réponse est donc assez complexe. Pour Beaumont et

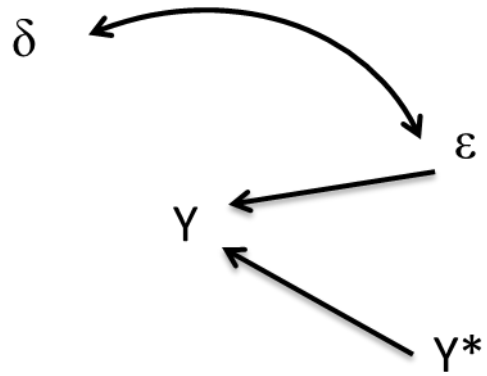
col (5), il vaut mieux avoir recours à un plan de sondage adaptatif pour diminuer le biais de non-réponse tel que l'enquête en deux phases pour non-réponse ; dans ce cas, il suffit de tirer au sort des personnes parmi les non-répondants et que celles-ci répondent pour obtenir une estimation sans biais. Groves suggère pour sa part en conclusion de son article relatif à cette question (48) que, dans un contexte où les taux de réponse aux enquêtes diminuent de plus en plus, il faut avant tout rechercher autant que possible des informations auxiliaires qui peuvent être exploitées pour réduire les effets de la covariance entre la probabilité de réponse et les variables d'intérêt d'une enquête.

II.3.2 PROPENSION À RÉPONDRE ET BIAIS DE MESURE

Le lien entre la propension à répondre et le biais de mesure a été évoqué dès 1963 (16) et peut s'expliquer ainsi : les répondants les moins motivés pour participer à une enquête, que l'on suppose avoir une propension à répondre petite, ont tendance à répondre de manière plus imprécise voire erronée que les répondants spontanés (120). Si on suppose que ce manque motivation est corrélé avec la difficulté à joindre les personnes, alors les personnes qui n'auraient pas répondu spontanément mais qui répondent après plusieurs relances seraient susceptibles de générer des biais de mesure plus grands que les personnes ayant répondu spontanément (27, 34).

Ce lien potentiel a été formalisé par Groves (49) et représenté par la Figure II-8.

Figure II-8 : Représentation graphique du lien entre probabilité de réponse et biais de mesure (d'après (49))



Soit Y^* la vraie valeur de la variable d'intérêt et Y sa mesure. Supposons qu'il n'existe pas de lien entre Y^* et δ et que l'erreur de mesure notée ε soit liée à la propension à répondre δ. Ce lien va engendrer un lien entre la probabilité de réponse et la variable d'intérêt mesurée Y, même si en réalité il n'existe pas de lien entre la probabilité de réponse et la variable d'intérêt Y^*.

Il est difficile de quantifier ce biais de mesure car il faudrait pour cela disposer, pour chaque répondant, à la fois de la vraie valeur de la variable d'intérêt et de sa valeur mesurée. Comme c'est rarement le cas, des études par simulation ont été réalisées pour évaluer quantitativement ce problème (120).

En pratique, on peut mesurer la proportion de données manquantes partielles selon la propension à répondre pour avoir une idée des biais de mesure potentiels ; Fricker et Tourangeau (34) qui étudient des variables socioéconomiques et Dalhammer (27) qui étudie des variables relatives au recours au soin ou à l'emploi, ont trouvé que la proportion de données manquantes partielles tendait à augmenter chez les personnes avec une propension à répondre faible en comparaison aux personnes avec une propension à répondre élevée.

Peu d'études ont mesuré le lien entre propension à répondre et biais de mesure. Les rares qui ont pu le faire disposaient pour les répondants à leurs enquêtes de variables identiques disponibles à la fois dans le questionnaire posé et dans des systèmes d'information existants pouvant servir de gold standard (74, 92). Deux études ont montré que généralement le biais de mesure augmentait légèrement ou restait stable lorsqu'on le comparait chez les personnes avec une propension à être contactée élevée ou dans l'échantillon de répondants complet ; les variables étudiées étaient relatives au statut matrimonial (durée du mariage, nombre de mois depuis le divorce, nombre de mariages dans l'étude de Olson) et sociodémographiques ou de santé (bénéficiaire de l'aide sociale, statut vis-à-vis de l'emploi, âge, nationalité étrangère). Olson a trouvé des résultats similaires en étudiant le biais de mesure selon la propension à coopérer (probabilité de réponse des personnes ayant réussi à être contactées par un enquêteur). Ces deux études ne montrent pas d'augmentation significative des biais de mesure selon la propension à répondre.

II.3.3 VARIANCE

La non-réponse génère une variance supplémentaire assimilable à celle observée dans une enquête en deux phases. Ainsi, plus le taux de réponse est élevé, plus la variance liée à la non-réponse est petite. Néanmoins, si, pour obtenir un taux de réponse maximal, on a recours à une enquête en deux phases pour non-réponse, la variance a plutôt tendance à augmenter.

L'erreur de mesure génère une variance supplémentaire ; si on suppose que le biais de mesure augmente lorsque la propension à répondre diminue, on peut également supposer qu'il engendre une plus grande variabilité dans les estimations donc une augmentation de la variance.

II.3.4 RELATION PROPENSION À RÉPONDRE, BIAIS DE MESURE, BIAIS DE NON-RÉPONSE

Les points précédents montrent que les relations entre propension à répondre et biais de non-réponse, biais de mesure ou variance ne sont pas simples à comprendre même lorsqu'ils sont appréhendés indépendamment les uns des autres. Quelques auteurs ont tenté de les étudier conjointement en supposant disposer de données gold standard.

On peut effectivement se demander quel est l'intérêt de maximiser des taux de réponse si ces derniers influent peu sur le biais de non-réponse et sont par ailleurs associés à une augmentation du biais de mesure. Est-ce que cela entraine une augmentation du biais total, une augmentation de l'erreur totale mesurée par l'erreur quadratique moyenne ?

Peytchev (97), dans une étude cherchant à quantifier la prévalence d'avortements à partir des données d'une enquête nationale américaine (taux de réponse 68%), a trouvé que les personnes avec une propension à répondre faible étaient plus susceptibles de sous-déclarer une expérience d'avortement que les personnes avec une propension à répondre élevée. Il suppose que biais de mesure et biais de non-réponse proviennent d'une cause commune : la désirabilité sociale qui incite les personnes ayant connu un avortement de ne pas répondre spontanément et qui les pousse à répondre de manière erronée lorsqu'on fait des efforts particuliers pour qu'elles répondent. Même si la donnée « gold standard » pour cette étude provient d'une information déclarée au cours de l'enquête, l'étude permet des discussions intéressantes. Après approximation de l'erreur quadratique moyenne (EQM), Peytchev conclut que l'EQM est supérieure lorsqu'il inclut le groupe de personnes avec une propension à répondre faible pour estimer la prévalence d'avortement que lorsqu'il ne l'inclut pas. En conclusion, l'auteur met en garde contre des maximisations naïves des taux de réponse qui

peuvent ne pas améliorer la qualité générale des prévalences estimées d'avortement du fait de liens entre biais de non-réponse et biais de mesure.

Olson (92) est, à notre connaissance, la première à étudier et à mesurer la balance entre biais de non-réponse et biais de mesure selon la propension à répondre à partir de variables gold standard disponibles via des systèmes d'information existants pour l'ensemble des personnes tirées au sort et de variables identiques mesurées par questionnaire, donc uniquement pour les répondants à l'enquête. Elle se base sur une étude relative au statut matrimonial de la personne (taux de contact 80% et taux de coopération, égal au nombre de répondants divisé par le nombre de personnes contactées, 88%) et étudie trois variables : la durée de mariage, la durée écoulée depuis le divorce, le nombre total de mariages. Elle considère par ailleurs deux composantes de la propension à répondre : la propension à être contacté et la propension à coopérer après avoir été contacté. Le biais de mesure diminue avec la propension à être contacté pour l'estimation de la moyenne de deux variables ; il augmente avec la propension à coopérer pour l'estimation de la moyenne d'une seule variable. Quelle que soit la variable et la composante utilisée pour la propension à répondre, le biais de non-réponse tend à diminuer après inclusion des personnes avec la propension à répondre la plus faible. Par ailleurs, quelle que soit la variable d'intérêt, le biais total (estimé ici par la somme du biais de mesure et du biais de non-réponse) diminue lorsque sont rajoutées dans l'analyse les personnes avec la propension d'être contacté la plus faible ; en revanche, le biais total ne diminue que pour une seule variable lorsque les personnes avec la propension à coopérer la plus faible sont incluses dans l'analyse. Olson conclut que le lien entre probabilité de réponse, biais de non-réponse et biais de mesure dépend de la variable étudiée, de la statistique estimée et du type de non-réponse étudié.

Kreuter (74) étudie elle aussi les liens entre biais de mesure et biais de non-réponse grâce à l'utilisation de variables gold-standard issues de systèmes d'information existants. A partir d'une enquête relative à l'emploi (taux de réponse 27%), elle étudie plus exactement l'évolution du biais de non-réponse, du biais de mesure, de la variance et de l'erreur quadratique moyenne selon la propension à être contactées. Pour les trois variables relatives à l'emploi, elle trouve d'une part que les biais de non-réponse diminuent avec l'inclusion de personnes avec une faible propension à être contacté et d'autre part que les biais de mesure ont tendance à augmenter avec l'inclusion de personnes avec une faible propension de contact ; au final, le biais total est inchangé. Pour la variable relative au fait d'être bénéficiaire de l'aide sociale, elle observe un résultat à première vue surprenant : alors que le biais de non-réponse diminue avec la propension à être contactées et que le biais de mesure et la variance restent stables, l'erreur quadratique moyenne augmente. Cela se produit parce que les biais de non-réponse et les biais de mesure qui affectent cette variable sont contraires : si les personnes avec une faible propension à être contacté sont des personnes avec une probabilité de bénéficier de l'aide sociale élevée et si tous les répondants, quelle que soit leur propension à être contactés, surestiment leur probabilité de bénéficier de l'aide sociale, alors le biais total va être supérieur si on inclut les personnes avec une propension à être contactées faible.

Ces trois études montrent donc que pris conjointement, les liens entre propension à répondre, biais de non-réponse et biais de mesure sont complexes et que leur impact sur le biais total varie. Aucune d'entre elles cependant n'étudie le lien entre propension à répondre et biais de mesure après correction de la non-réponse ; on peut se demander s'il est intéressant de maximiser autant que possible le taux de réponse alors qu'il est possible de corriger le biais du à la non-réponse grâce à des informations auxiliaires pertinentes et que la maximisation du taux de réponse entraine des biais de mesure.

II.4 SYNTHÈSE

Nous avons traité dans ce chapitre différentes manières d'aborder la non-réponse totale dans les enquêtes avec des objectifs descriptifs.

En résumé, il existe deux familles de méthodes pour minimiser les biais de non-réponse dans les enquêtes : maximiser le taux de réponse ou utiliser de l'information auxiliaire permettant de diminuer la covariance entre la propension à répondre et la variable d'intérêt.

Utiliser de l'information auxiliaire une fois que les données ont été recueillies suppose que l'on peut « faire confiance » au modèle de non-réponse sous-jacent pour diminuer les biais de non-réponse. C'est pourquoi l'approche la plus utilisée en pratique est la maximisation du taux de réponse. Néanmoins, cette dernière est de plus en plus discutée car d'une part elle peut être inutile, voire contre-productive, pour minimiser les biais de non-réponse et d'autre part qu'elle peut augmenter les biais de mesure. La question qui se pose alors est de savoir si la maximisation du taux de réponse offre, au final, un gain en termes d'erreur totale.

Au-delà de ces considérations, il est nécessaire de donner des éclairages sur la stratégie optimale à adopter en pratique.

Nous nous proposons, dans la suite, de prendre en compte au mieux toutes ces dimensions de la non-réponse dans la phase pilote de la cohorte Coset-MSA à l'inclusion ; ce travail permettra d'émettre des recommandations, au moins pour cette étude lors de sa phase de généralisation, et autant que possible, plus généralement pour les enquêtes de surveillance en population.

CHAPITRE III. LA COHORTE PILOTE COSET-MSA

Cette thèse repose sur l'exploitation des données de la cohorte pilote Coset-MSA, qui s'inscrit dans le programme Coset.

III.1 LE PROGRAMME COSET

III.1.1 CONTEXTE

La surveillance épidémiologique des risques professionnels consiste à décrire de façon systématique et permanente la fréquence et la survenue de problèmes de santé en lien avec des facteurs de risque professionnels (44). Ses principaux objectifs sont d'établir des indicateurs permettant de quantifier le poids de l'activité professionnelle sur l'état de santé de la population générale, de repérer des secteurs et des professions à risque élevé, d'alerter sur d'éventuels problèmes en relation avec le travail, connus ou émergents et enfin d'évaluer les dispositifs de prévention et de réparation.

La surveillance épidémiologique des risques professionnels est complexe de par la difficulté à établir qu'une pathologie est d'origine professionnelle. En effet, les expositions à risque pour la santé ne sont en général pas spécifiques de l'activité professionnelle. Si on prend l'exemple des troubles musculo-squelettiques, ils peuvent être causés par des gestes répétitifs, qui peuvent avoir été exécutés dans le cadre d'activités domestiques comme le bricolage ou le jardinage ou bien dans le cadre d'activité professionnelle comme le passage d'articles en caisse pour les caissières de supermarché (105). Par ailleurs, un agent pathogène, qu'il soit d'origine professionnelle ou non, n'induit pas nécessairement à lui seul la survenue d'une maladie ; celle-ci est plus souvent liée à la combinaison de plusieurs facteurs d'exposition professionnels ou non professionnels voire d'une susceptibilité génétique. Pour toutes ces

raisons, il est souvent difficile d'isoler la contribution spécifique des facteurs professionnels et les facteurs d'interaction possibles pour une pathologie donnée. Ceci se traduit d'un point de vue méthodologique par la nécessité de recueillir dans les systèmes de surveillance épidémiologique des risques professionnels à la fois les emplois, les expositions professionnelles et extra-professionnelles et les pathologies. Une autre difficulté de ce type de surveillance vient du temps de latence parfois important entre une exposition à risque et la survenue de la maladie. Si on prend l'exemple du cancer de la plèvre, il est presque sûrement causé par une exposition à l'amiante, qui est, chez les hommes, d'origine professionnelle dans 90% des cas et qui survient environ 20 ans après une exposition à l'amiante (24). Ce temps de latence important nécessite de surveiller les personnes en emploi mais aussi celles qui ne le sont plus (chômeurs, inactifs, retraités) et de recueillir les expositions tout au long de l'historique professionnel. C'est pourquoi les cohortes sont un outil essentiel de la surveillance épidémiologique des risques professionnels.

Dans ce contexte, le Département santé travail de l'Institut de veille sanitaire a décidé de mettre en place le programme Coset « COhortes pour la Surveillance Epidémiologique en lien avec le Travail » (39).

III.1.2 OBJECTIFS

L'objectif principal du programme Coset est la connaissance et la surveillance de la morbidité et de la mortalité de la population active en France. En d'autres termes, il s'agit de pouvoir décrire et suivre l'état de santé des actifs selon leur activité professionnelle et son évolution dans le temps, l'activité professionnelle étant définie par la combinaison de deux informations : la profession et le secteur d'activité. Ce programme a vocation de permettre une surveillance épidémiologique multi-professions multi-secteurs. Son objectif principal peut se décliner en objectifs spécifiques :

- décrire à un instant « *t* » la morbidité/mortalité des actifs selon l'activité professionnelle ;
- décrire l'évolution dans le temps de la morbidité/mortalité des actifs selon l'activité professionnelle ;
- décrire et surveiller les liens entre la morbidité/mortalité des actifs et les expositions professionnelles qu'elles soient d'origine physique, chimique, psychosociale ou organisationnelle ;
- calculer des fractions de morbidité/mortalité attribuables aux facteurs d'exposition professionnelle ;
- aider au repérage de problèmes émergents et faciliter la mise en place d'études *ad hoc* en cas de repérage de problèmes émergents ou mal documentés sur le plan scientifique.

III.1.3 POPULATION ET MÉTHODES

III.1.3.1 Population cible

La population cible du programme Coset à l'inclusion correspond à l'ensemble de la population active en France âgée de 18 à 65 ans, qu'elle soit effectivement en activité ou bien dans une période d'inactivité (au chômage par exemple), salariés et non salariés, quels que soient la catégorie socioprofessionnelle, le secteur d'activité et le type de contrat de travail.

III.1.3.2 Populations sources

Les données du programme Coset seront issues de différentes études de cohorte incluant des actifs affiliés aux trois principaux régimes de protection sociale : le Régime général de sécurité sociale (RG), la Mutualité Sociale Agricole (MSA) et le Régime Social des Indépendants (RSI), qui, à eux trois, couvrent 95% de la population cible.

III.1.3.3 Schéma d'enquête

Une cohorte par régime sera mise en place avec un suivi minimal de 20 ans (Tableau III-1):

- pour la population active salariée, affiliée au Régime général de sécurité sociale, soit environ 80 % des actifs en France, les données seront issues de la cohorte Constances (131), cohorte généraliste des affiliés à ce régime, en cours de mise en place par l'Institut national de la santé et de la recherche médicale (Inserm). Les objectifs et les domaines de recherche de Constances sont larges ; ce sont les informations les informations nécessaires à la surveillance des risques professionnels qui seront exploitées dans le cadre du programme Coset ;
- une cohorte de la population active des travailleurs agricoles affiliés à la Mutualité sociale agricole (MSA), Coset-MSA et une cohorte de la population active des travailleurs indépendants affiliés au Régime social des indépendants (RSI), Coset-RSI, seront mises en place par l'Institut de veille sanitaire en partenariat avec les régimes de protection sociale concernés. Elles sont actuellement en phase pilote, l'extension à l'échelle nationale étant prévue en 2015 pour Coset-MSA et, au mieux en 2016 pour Coset-RSI.

La méthodologie des trois cohortes est, autant que possible, similaire. La constitution de l'échantillon des personnes invitées à participer est réalisée par tirage au sort. On peut néanmoins noter que la cohorte Constances comprend, en plus des actifs, des ayants droits, des retraités et des étudiants ; par ailleurs, les participants à la cohorte Constances bénéficient d'un examen de santé via les Centres d'Examens de Santé de la Sécurité Sociale, ce qui n'est pas le cas des participants aux cohortes Coset-MSA et Coset-RSI. Des réinclusions régulières sont envisagées.

III.1.3.4 Mode de recueil de données et données recueillies

Le dispositif général prévoit, lors de l'inclusion, le recueil sur la santé et l'activité professionnelle par auto-questionnaire, ainsi que le recueil des informations nécessaires pour recontacter la personne dans le cadre du suivi de cohorte. Pour les sujets recrutés au sein de la cohorte Constances, ces informations seront complétées par des données cliniques et paracliniques. Chaque année, un nouvel auto-questionnaire sera proposé aux participants afin de suivre l'évolution de leur état de santé et de leur parcours professionnel. En parallèle, il est à l'inclusion et au fil du suivi, un recueil « passif » pour les participants et un sous-échantillon aléatoire de non-participants, par la consultation des données issues des systèmes d'information existants contenant des informations exploitables sur les remboursements de soins, les événements de santé et les événements professionnels.

III.1.3.5 Défaut de couverture

Par construction, les régimes spéciaux (RATP, EDF,…) ne seront pas inclus dans la cohorte.

Par ailleurs, seront exclus les actifs ayant eu seulement des emplois de moins de six mois pour les cohortes du Régime général et du RSI, et de moins de trois mois pour la cohorte de la MSA.

III.1.3.6 Remarque sur les taux de réponse attendus

Les taux de réponse attendus pour les trois cohortes correspondent aux taux de réponse observés lors des phases pilote (8% pour Constances, 23% pour Coset-MSA, 15% pour Coset-RSI). Le taux de 8% pour Constances s'explique notamment par le fait que les personnes sollicitées pour participer à la cohorte doivent se déplacer dans un Centre d'Examen de Santé. Le taux de 15% pour Coset-RSI était attendu du fait du peu d'engouement en général des professions concernées pour les enquêtes réalisées dans cette

population. Le taux de 23% pour Coset-MSA était lui aussi attendu du fait, au contraire, du lien fort qui existe entre la MSA et ses affiliés.

Tableau III-1 : Récapitulatif des trois cohortes constituant le programme Coset

	Régime général	**M.S.A.**	**R.S.I.**
Population au travail couverte	80%	9%	6%
	Salariés	Travailleurs du monde agricoles (salariés et non salariés)	Commerçants, artisans, professions libérales (non salariés)
Nom de la cohorte	Constances	Coset-MSA	Coset-RSI
Mise en œuvre	Inserm	InVS-DST	InVS-DST
Modalités d'inclusion	Tirage au sort		
Effectif initial tiré au sort	2 000 000	150 000	235 000
Mode de recueil de données	Actif : Questionnaire (auto et administré), bilan de santé CES	Actif : auto-questionnaire	
	Passif : Données issues des systèmes d'information existants		
Effectif attendu de répondants au questionnaire	200 000	35 000	35 000
Effectif suivi passivement	400 000	70 000	70 000
Phase pilote (année)	2009	2010	2012
Extension nationale (année)	2012-2017	2015	2016

III.2 LA PHASE PILOTE DE LA COHORTE COSET-MSA

Avant de mettre en place la cohorte Coset-MSA à l'échelle nationale, les modalités de recrutement d'actifs du régime agricole ont été testées en 2010 sur cinq départements lors d'une phase pilote. Compte tenu du faible taux de réponse attendu et de la durée du suivi envisagée (au moins 20 ans), l'objectif de la phase pilote était par ailleurs d'étudier le mieux possible les biais potentiels de non-réponse à l'inclusion. C'est pour cela qu'en plus du recueil par questionnaire postal et de la consultation des bases de données issues de systèmes d'information existants, une enquête complémentaire auprès d'un échantillon de non-répondants à l'enquête postale a été réalisée spécifiquement pour l'enquête pilote. Les accords de la Cnil nécessaires ont été obtenus préalablement à la mise en œuvre des différentes étapes (dossiers Cnil n°909091, Cnil n°910191 et Cnil n°910176).

III.2.1 POPULATION SOURCE

La Mutualité Sociale Agricole assure en maladie et en vieillesse des non salariés (agriculteurs ou chefs d'entreprise agricole) et des salariés (ouvriers agricoles, personnels de certaines banques et assurances, employés de coopératives agricoles, enseignants en lycée agricole etc.).

La population source pour la phase pilote correspond aux non-salariés agricoles et aux salariés affiliés à une des cinq caisses départementales pilotes (Bouches-du-Rhône, Finistère, Pas-de-Calais, Pyrénées Atlantiques, Saône-et-Loire), âgés de 18 à 65 ans et ayant travaillé au moins 90 jours en 2008 en tant qu'affiliés au Régime agricole, quel que soit leur type d'activité. Elle comprend environ 100 000 personnes.

III.2.2 MÉTHODES

III.2.2.1 Plan de sondage

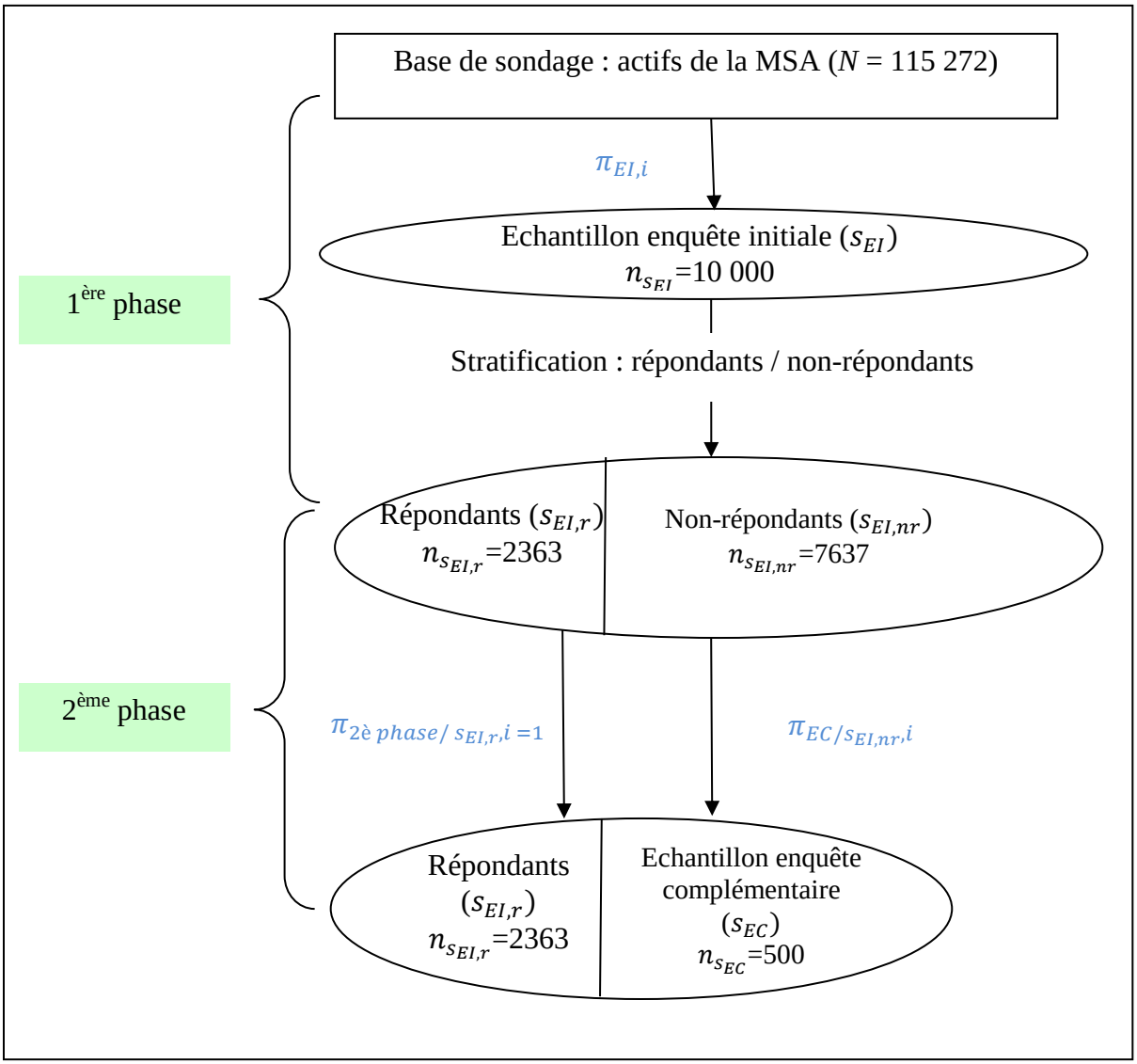
Pour réaliser le tirage au sort, la base de sondage utilisée était la base d'assurance retraite, qui contient toutes les personnes ayant travaillé au moins une fois en tant qu'affiliées à la MSA. Le plan de sondage mis en place pour la phase pilote était un plan de sondage en deux-phases pour non-réponse (Figure III-1).

Pour le pilote de Coset-MSA, la première phase a consisté à tirer au sort 10 000 personnes, soit 2 000 personnes dans chacun des cinq départements participant à la phase pilote. Dans chaque département, le tirage au sort, par sondage aléatoire simple sans remise, était stratifié selon le sexe, l'âge et le statut d'emploi (salarié ou non-salarié) et était proportionnel à la taille des strates. Cette première phase correspondait au plan de sondage envisagé pour l'extension nationale, France entière. L'enquête de deuxième phase a été réalisée auprès d'un échantillon de 500 non-répondants ; 100 non-répondants par département ont été tirés au sort par sondage aléatoire simple.

L'enquête de première phase sera appelée par la suite enquête initiale, et l'enquête additionnelle menée lors de la deuxième phase, l'enquête complémentaire.

Pour simplifier la lecture du texte, l'enquête en deux-phases pour non-réponse pourra être appelée par la suite « enquête en deux phases » ou « enquêtes combinées ».

Figure III-1 : Sondage en deux phases pour non-réponse pour la cohorte pilote Coset-MSA



N : taille de la population
EI : enquête initiale
EC : enquête complémentaire
s_{EI} *: échantillon issu de EI et* $n_{s_{EI}}$ *sa taille*
$s_{EI,r}$ *: sous-échantillon de répondants à EI et* $n_{s_{EI,r}}$ *sa taille*
$s_{EI,nr}$ *: sous-échantillon de non-répondants à EI et* $n_{s_{EI,nr}}$ *sa taille*
s_{EC} *: échantillon issu de EC et* $n_{s_{EC}}$ *sa taille*
$\pi_{EI,i}$ *: probabilité d'inclusion d'un individu i à l'enquête initiale*
$\pi_{2è\ phase/s_{EI,r},i}$ *: probabilité d'inclusion d'un individu i dans l'échantillon de deuxième phase sachant qu'il appartient à* $s_{EI,r}$
$\pi_{EC/s_{EI,nr},i}$ *: probabilité d'inclusion d'un individu i dans l'échantillon de deuxième phase sachant qu'il appartient à* $s_{EI,nr}$

III.2.2.2 Données recueillies et mode de recueil des données

Deux types de données ont été recueillis : des données nécessitant l'interrogation directe des personnes, appelées par la suite données de questionnaire ainsi que des données ne nécessitant pas l'interrogation des personnes, appelées par la suite données passives ou informations auxiliaires. Les premières ne sont disponibles que chez les répondants alors que les deuxièmes le sont pour la totalité des répondants et des non-répondants à l'exception des personnes pour lesquelles il n'y a pas eu d'appariement avec les données passives (cf. III.2.2.2.2.1).

III.2.2.2.1 Données nécessitant l'interrogation directe des personnes : données de questionnaire

- Enquête initiale

Pour l'enquête initiale, le recueil des données a été réalisé par questionnaire postal auto-administré de 40 pages avec une relance postale un mois après. En complément du questionnaire, chaque envoi comprenait une plaquette de présentation de l'étude, une lettre annonce, une enveloppe T pour le retour prépayé du questionnaire et un formulaire de refus d'accès aux systèmes d'information existants.

Les informations recueillies par questionnaire concernaient l'état de santé (problèmes musculaires et articulaires, symptômes dépressifs, problèmes cardio-vasculaires et respiratoires, cancer…) , les comportements de santé (consommations d'alcool et de tabac), l'activité professionnelle actuelle et passée (statut, type de contrat, temps de travail) et les expositions à certaines nuisances passées ou actuelles (contraintes organisationnelles et psychosociales, pénibilité, bruit, nuisances d'origine chimique, physique ou biologique) subies sur le lieu de travail.

- Enquête complémentaire

Le schéma d'étude (Figure III-2) de l'enquête complémentaire a été construit de façon à maximiser autant que possible le taux de réponse. C'est pourquoi il a été choisi de réaliser une enquête par téléphone plutôt qu'une enquête postale. La MSA ne disposant pas des numéros de téléphone de ses bénéficiaires, il était probable qu'un pourcentage non nul de numéros de téléphone ne serait pas retrouvé. La MSA possédant les adresses postales de ses affiliés, il a donc été décidé d'enquêter en face-à-face les personnes non-jointes par téléphone. Cependant, afin de tenir compte de l'influence des différents modes de collecte des données (par téléphone et en face-à-face), il a été décidé qu'un groupe de personnes serait enquêté en première intention en face-à-face, que son numéro de téléphone ait été retrouvé ou non.

Sur les 500 personnes échantillonnées, 350 ont été affectées au groupe 1 (enquête par téléphone en première intention) et 150 ont été affectées au groupe 2 (enquête en face-à-face en première intention). Les personnes échantillonnées ayant expressément exprimé un refus de participer lors de l'invitation à l'enquête initiale n'ont pas été enquêtées (n = 15).

Le terrain de l'enquête a été réalisé par un prestataire, sous la supervision de l'équipe Coset.

Un courrier d'annonce a été envoyé à toutes les personnes échantillonnées. Pour les 500 personnes tirées au sort, une recherche des numéros de téléphone (fixe ou portable) a été réalisée par le prestataire à partir des noms, prénoms et adresses des personnes transmises par la Mutualité Sociale Agricole et mises-à-jour par le prestataire. Ce dernier, à partir des noms, prénoms et adresses des personnes, a recherché leurs numéros de téléphone.

Avant le terrain de l'enquête, les enquêteurs ont reçu une formation assurée par l'équipe Coset sur les objectifs et le déroulement de l'enquête.

- Enquête auprès des personnes du groupe 1 (téléphone en première intention)

Cas 1 : Lorsque les coordonnées téléphoniques étaient retrouvées, l'enquêteur tentait de contacter la personne avec deux issues possibles. Si l'enquêteur réussissait à contacter la personne, il réalisait l'enquête par téléphone ou obtenait un refus. Si l'enquêteur ne réussissait pas à contacter la personne ou si celle-ci n'avait pas exprimé clairement un refus, une enquête en face-à-face était alors mise en place (cf. enquête auprès des personnes du groupe 2).

Cas 2 : Lorsque les coordonnées téléphoniques n'étaient pas retrouvées, la personne était alors enquêtée directement en face-à-face (cf. enquête auprès des personnes du groupe 2).

- Enquête auprès des personnes du groupe 2 (en face-à-face en première intention)

L'objectif initial était d'enquêter toutes les personnes y compris celles ne résidant plus dans les départements pilotes, mais pour des raisons de coût, une zone géographique limitée (jusqu'à 20 kilomètres autour de chaque département pilote) a été définie pour les personnes ne résidant pas ou plus dans un des départements pilotes.

Cas 1 : Si la personne résidait dans cette zone et si ses coordonnées téléphoniques étaient retrouvées, elle était contactée par téléphone pour une prise de rendez-vous. En cas de refus, l'enquête était terminée. Si l'enquêteur ne réussissait pas à joindre la personne par téléphone, il se rendait directement à son domicile pour l'interviewer.

Cas 2 : Si la personne résidait dans cette zone et si ses coordonnées téléphoniques n'étaient pas retrouvées, l'enquêteur se rendait directement au domicile de la personne pour l'interviewer.

Cas 3 : Si la personne ne résidait pas dans cette zone et si ses coordonnées téléphoniques étaient retrouvées, la personne était alors enquêtée par téléphone (cf. enquête auprès des personnes du groupe 1).

Cas 4 : Si la personne ne résidait pas dans cette zone et si ses coordonnées téléphoniques n'étaient pas retrouvées, elle n'était pas interviewée. L'enquête complémentaire présente donc un défaut de sous-couverture pour ces personnes.

Remarque : Pour les personnes du groupe 1 ayant basculé dans le groupe 2, les cas 1 et 3 sont sans objet.

- Recueil des données par téléphone

Pour la collecte des données de questionnaire par téléphone, la méthode de CATI (Computer Assisted Telephone Interview) a été utilisée, où l'enquêteur remplissait directement les réponses au questionnaire via un ordinateur. Etant donné qu'il s'agissait d'une population d'actifs, les créneaux horaires privilégiés pour les appels étaient de 12h à 14h et de 17h à 21h en semaine et de 10h à 18h le samedi. En cas de non-réponse ou d'absence (pas de réponse, occupé, répondeur) jusqu'à 15 tentatives étaient réalisées à des jours et heures différents (dont au moins un samedi). Si la personne était jointe mais indisponible au moment de l'appel, un rendez-vous téléphonique était proposé.

- Recueil des données en face-à-face

Pour la collecte des données de questionnaire en face-à-face, l'enquêteur saisissait directement les réponses via la méthode CAPI (Computer Assisted Personal Interview). Les personnes dont le numéro de téléphone a pu être retrouvé étaient préalablement contactées par téléphone pour fixer un rendez-vous. Pour les personnes dont le numéro de téléphone n'avait pas pu être retrouvé, les visites étaient effectuées de préférence à des créneaux horaires où les

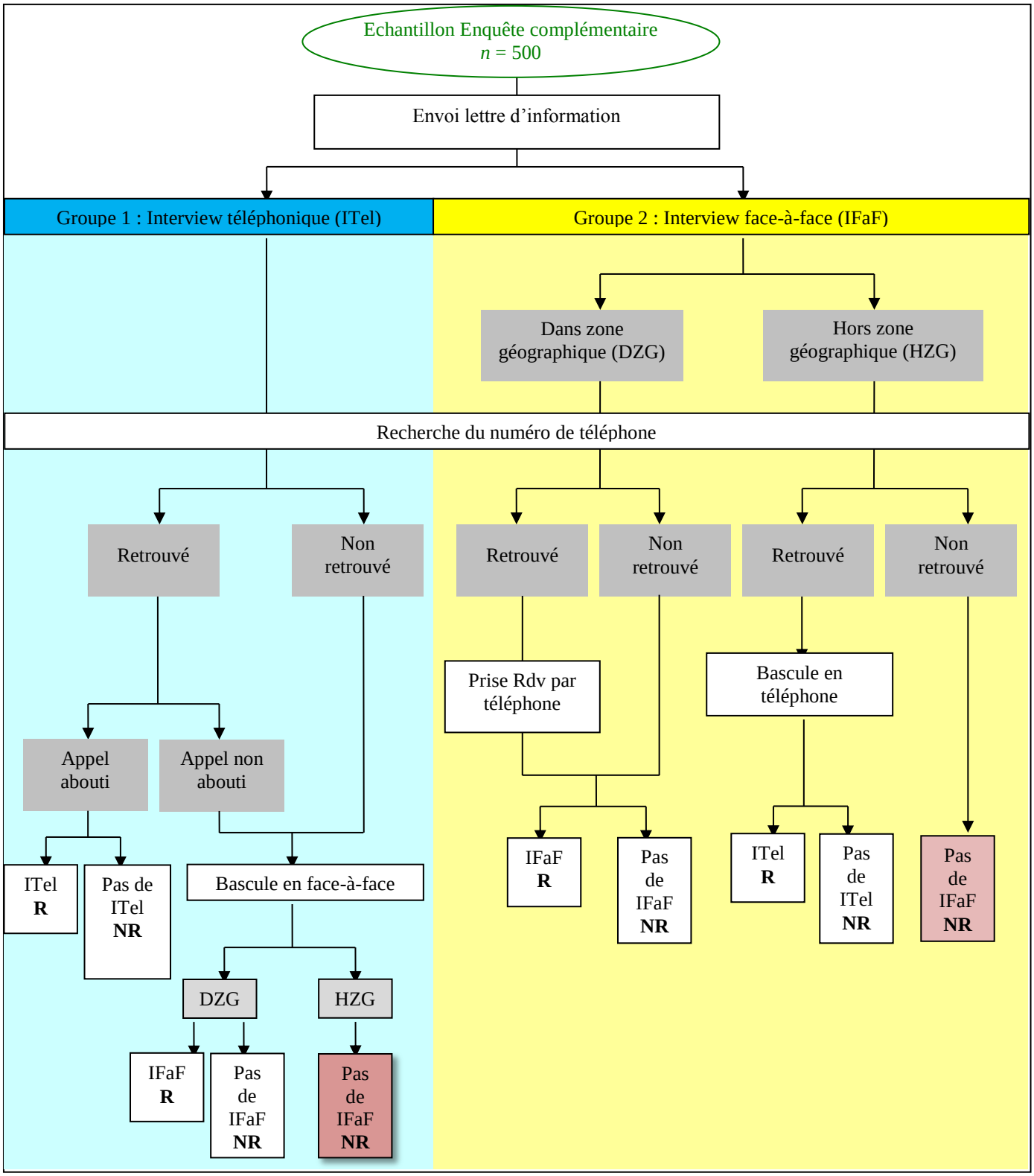
personnes avaient plus de chance de se trouver à leur domicile, c'est-à-dire les soirées en semaine ou la journée du samedi. En cas d'absence, l'enquêteur laissait un avis de passage et un numéro vert auquel le joindre. Jusqu'à trois passages étaient effectués.

Pour l'enquête complémentaire, un questionnaire plus court issu du questionnaire de l'enquête initiale a été utilisé (cf. annexes). En effet, la longueur du questionnaire initial étant probablement à l'origine de certaines non-réponses et afin de limiter les abandons en cours d'interview, la taille du questionnaire a été conçue pour que sa durée de collecte n'excède pas un quart d'heure. Les questions sélectionnées ont été retenues selon les critères suivants :

- **Variables *a priori* associées à la non-participation (45)** : sexe, âge (mois et année de naissance), catégorie sociale, statut marital, niveau d'étude, comportements (alcool, tabac), état de santé général ;
- **Principales variables d'intérêts** dans Coset-MSA : état de santé général, statut vis-à-vis de l'emploi, principales expositions professionnelles, emploi actuel ou dernier emploi en date (profession et secteur d'activité), nombre d'épisodes professionnels.

Les variables étudiées dans ce travail sont les variables présentes à la fois dans le questionnaire de l'enquête initiale et de l'enquête complémentaire.

Figure III-2 : Modalités de réalisation de l'enquête complémentaire



R : répondants
NR : non-répondants
: défaut de couverture

III.2.2.2.2 Données ne nécessitant pas l'interrogation directe des personnes

III.2.2.2.2.1 Données issues des systèmes d'information existants

Les données issues des systèmes d'information existants provenaient soit de la MSA, soit du SNIIR-AM. Les données issues de la MSA ont été extraites de deux systèmes d'information différents : le système des bases « retraite », qui a servi de base de sondage, a permis d'accéder à des données sociodémographiques et le système des bases « contrats » pour les salariés et « cotisations » pour les non-salariés a permis d'extraire des données relatives à la profession de la personne. Les données issues du SNIIR-AM (Système d'Information Inter-régime de l'Assurance Maladie) sont des données relatives à la de santé de la personne.

Les données sont présentées selon leur nature (en gras), leur provenance (en gras italique) et leur année (en italique) ; leur notation, qui sera largement reprise par la suite, est pour sa part soulignée. Elles sont par ailleurs synthétisées dans le Tableau III-2.

Tableau III-2 : Données issues des systèmes d'information existants

Nature	Contenu	Année ou date	Provenance
Variables sociodémographiques	Sexe Age Département Statut	31/12/2008	MSA : base retraite
Variables relatives à la santé	Remboursement de soins Hospitalisations Indemnités journalières	2008 à 2010	SNIIR-AM (dont PMSI)
Variables relatives à l'emploi	Dernier emploi principal en tant qu'affilié à la MSA ATMP	2008, 2009 ou 2010 2003 à 2008	MSA : bases cotisations et contrats MSA : bases ATMP

- **Variables sociodémographiques :** ***Variables de stratification issues de la base de sondage*** *(2008)* notées *X*

La base de sondage a été construite via les bases d'assurance retraite de la MSA. Par construction, les données issues de la base de sondage, qui ont été utilisées comme variables de stratification, étaient disponibles pour l'ensemble des personnes tirées au sort après l'extraction de la base de sondage.

Les variables sociodémographiques étaient le sexe, l'âge (18-34 ans, 35-49 ans, 50-65 ans), le département (Bouches-du-Rhône, Saône et Loire, Pyrénées Atlantique, Finistère, Nord-Pas de Calais), et le statut (salarié exclusif SA, non salarié NSA).

Pour chacune de ces variables, on disposait par ailleurs de leurs totaux croisés dans l'ensemble de la population d'étude décrite en III.2.1 (par exemple, on connaissait le nombre exact d'hommes salariés âgés de 18 à 29 ans affiliés à la MSA et travaillant dans le département des Bouches-du-Rhône).

- ***Données issues des bases de données du SNIIR-AM et de cotisations de la MSA*** notées *V*

L'accès aux bases de données du SNIIR-AM et de cotisations de la MSA repose sur un schéma complexe, ayant reçu un accord de la Cnil, détaillé par ailleurs (6). Il fait entrer en jeu un tiers de confiance, des numéros de confidentialité avec des correspondances et des transferts de fichiers cryptés. Ainsi, l'accès à ces données respecte leur confidentialité ; il a été conçu de façon à ce que les organismes possesseurs des données d'intérêt n'entrent jamais en possession d'informations identifiantes qu'ils ne détiennent déjà.

Ces données étaient disponibles pour l'ensemble des personnes tirées au sort, à l'exclusion de 693 personnes pour une des deux raisons suivantes :

- leur adresse postale a été considérée comme invalide (donc ces personnes n'ont pas été informées de l'étude et n'ont pas pu exprimer d'éventuel refus d'accès à ces bases) (75%) ;
- elles ont explicitement exprimé un refus d'accès aux bases SNIIR-AM ou MSA (17%) ;
- leurs données SNIIR-AM ou MSA n'ont pas pu être appariées car leur état civil était incomplet à la date du tirage au sort (jour et mois de naissance manquants) (8%).

Il n'a pas été fait d'appariement pour ces personnes.

✓ **Données de santé** : ***SNIIR-AM*** *(de 2008 à 2010)*

Le SNIIR-AM (Système National d'Information Inter-régime d'Assurance Maladie) est une base de données contenant les remboursements de soins et les indemnités journalières des principaux régimes d'assurance maladie, ainsi que les données d'hospitalisation issues du PMSI (Programme de Médicalisation des systèmes d'Information). Cette base de données, très riche, a l'avantage d'être multi-régimes, donc d'être alimentée même si une personne a changé de régime d'assurance maladie. Le type de données utilisées pour l'analyse est brièvement décrit ci-après.

- Remboursements de soins :

Nombre de recours à un professionnel de santé, par type de professionnel de santé ;

Nombre de boites de médicaments, par type (classification « Anatomique, Thérapeutique et Chimique » (ATC)) ;

Montant total des prestations.

- Absentéisme pour raison de santé – Indemnités :

Montant total des versements, durée totale, type d'arrêts (pour maladie, accident du travail ou maternité).

- Hospitalisations :

Nombre d'hospitalisations, type d'établissement (public ou privé), motif (chirurgical ou non chirurgical), pathologie (grands chapitres de la CIM10).

✓ **Données professionnelles : *Contrats et cotisations professionnelles, MSA*** :

La MSA, en tant qu'assureur en retraite et en Accident du Travail et Maladie Professionnelle (branche ATMP), recueille, via plusieurs systèmes d'information centralisés internes, de nombreuses informations relatives à l'emploi des personnes lorsque ce dernier relève de leur régime. Ainsi, contrairement aux données du SNIIR-AM, les informations relatives aux emplois des personnes en tant qu'affiliées à un autre régime d'assurance maladie ne sont pas disponibles dans les bases de la MSA. Les données de la MSA utilisées pour l'étude sont brièvement décrites ci-après.

- Données relatives au dernier emploi principal en date en tant qu'affilié à la MSA *(2008, 2009 ou 2010*)

Statut (salarié SA, non-salarié NSA), secteur d'activité (selon la classification « code risque » interne à la MSA), durée d'emploi, nombre d'emplois salariés dans l'année.

- Accidents du Travail et Maladies Professionnelles (ATMP) reconnus *(2003 à 2008*)

Nombre d'ATMP, par gravité, par année.

III.2.2.2.2.2 Données enregistrées pour la gestion de l'enquête complémentaire notées *Z*

Pour l'enquête complémentaire, en amont du recueil des données de questionnaire, un outil de gestion et de suivi des inclusions a été mis en place, accessible par les enquêteurs et par les superviseurs de l'enquête. Son objectif était d'enregistrer toutes les étapes et les informations relatives au déroulement de terrain de l'enquête (72).

De cet outil a pu être extrait un fichier de données, appelées paradonnées qui étaient par définition disponibles pour les répondants et les non-répondants à l'enquête complémentaire.

Deux types de paradonnées ont été collectés : des données recueillies avant le processus de collecte proprement dit, et des données recueillies pendant le processus de collecte.

Les paradonnées disponibles avant la collecte des données de questionnaire étaient *a priori* relatives à la qualité des informations disponibles pour contacter les personnes. La première était un indicateur de la fiabilité de l'adresse postale transmise par la MSA. Suite à la recherche par le prestataire, si l'adresse était considérée comme correcte, sa fiabilité était considérée comme bonne, et mauvaise sinon. La deuxième concernait les numéros de téléphone. A partir des noms, prénoms et adresse postale, le numéro de téléphone de la

personne à enquêter était recherché. Un indicateur de la fiabilité du numéro de téléphone retrouvé a été construit ; si un numéro de téléphone était associé à un même nom, prénom que la personne tirée au sort et une même adresse postale, sa fiabilité était considérée comme bonne. Si un numéro de téléphone était retrouvé avec un prénom ou un détail d'adresse non correspondants, sa fiabilité était considérée comme mauvaise. Une troisième modalité « téléphone non retrouvé » était par ailleurs présente pour l'indicateur.

Les paradonnées enregistrées pendant la collecte des données étaient : le nombre d'appels téléphoniques (pour un rendez-vous ou une interview), le nombre d'appels après 17 heures, le nombre d'appels le samedi, le nombre de visites d'un enquêteur, la visite d'un enquêteur le samedi.

Une autre information, à l'interface entre le schéma de l'enquête et la paradonnée a été construite : l'évolution du mode de collecte des données (assignation au groupe téléphone sans changement de mode de recueil, assignation au groupe téléphone et changement de mode de recueil, assignation au groupe face-à-face sans changement de mode de recueil, et assignation au groupe face-à-face et changement de mode de recueil).

Tableau III-3 : Résumé enquêtes initiale et complémentaire

	Enquête initiale (EI)	**Enquête complémentaire (EC)**
Période d'enquête	Février-mars 2010	Novembre 2010-février 2011
Mode de recueil des données	Autoquestionnaire postal	Téléphone ou face-à-face avec enquêteur
Taille de l'échantillon :		
Tiré au sort	10 000	500
Avec données du SNIIR-AM et de la MSA	9 307	454
Répondants au questionnaire	2 363	313
Données recueillies pour :		
Population d'étude (base de sondage)	Totaux croisés des variables sociodémographiques (sexe, âge, caisse, statut d'emploi)	
Personnes tirées au sort	Variables sociodémographiques	
		Paradonnées
Personnes tirées au sort n'ayant pas exprimé de refus d'accès aux bases médico-administratives	Données du SNIIR-AM (de 2008 à 2010) : remboursements de soins, indemnités journalières, hospitalisations via le PMSI Données de la MSA : données relatives au dernier emploi principal en date (2008, 2009 ou 2010), accidents du travail et maladies professionnelles reconnus (de 2003 à 2008),	
Répondants au questionnaire	Santé perçue, échelles de santé (TMS, symptômes dépressifs, asthme), consommation de tabac et d'alcool, historique d'emplois et d'expositions professionnelles	Santé perçue, consommation de tabac et d'alcool, emploi actuel, quelques expositions professionnelles actuelles

III.3 DONNÉES ÉTUDIÉES SELON LES CHAPITRES DE LA THÈSE

Les données exploitées dans le cadre de ce travail étaient de deux sortes : des données de questionnaire et des données auxiliaires (variables sociodémographiques, variables issues des données du SNIIR-AM ou de la MSA, paradonnées). Selon l'objectif propre à chaque chapitre ou sous-chapitre, les données utilisées étaient différentes. Pour faciliter la lecture du manuscrit, le Tableau III-4 pourra être utilisé pour se référer aux données exploitées selon le chapitre étudié.

Tableau III-4 : Données utilisées selon le chapitre étudié

	Variables d'intérêt			**Informations auxiliaires**		
	Données de questionnaire		Données issues de systèmes d'information existants	Données issues de systèmes d'information existants		Paradonnées
	EI	EC	SNIIR-AM et MSA	Variables de stratification	SNIIR-AM et MSA	
	Répondants	*Répondants*	*Répondants et non-répondants*	*Répondants et non-répondants*		*Répondants et non-répondants à l'EC*
Chapitre IV : Apport de données supplémentaires dans la prise en compte des biais de non-réponse						
Chapitre IV.2 : Présence d'informations auxiliaires de qualité : l'apport des bases médico-administratives	**X**		X	**X**	X	
Chapitre IV.3 : Absence d'informations auxiliaires de qualité (provenant de bases médico-administratives) : l'apport d'une enquête auprès de non-répondants	**X**	X	X	**X**		X
Chapitre IV.4 : Apport d'une enquête auprès de non-répondants en présence d'informations auxiliaires de qualité	**X**	X	X	**X**	X	
Chapitre V : Niveau d'efforts consentis pour obtenir une réponse, erreur de non-réponse, erreur de mesure	**X**	X	X	**X**	X	

EI : enquête initiale, EC : enquête complémentaire

*X : données utilisées ;**X** : données utilisées quel que soit le chapitre*

CHAPITRE IV. APPORT DE DONNÉES SUPPLÉMENTAIRES DANS LA PRISE EN COMPTE DES BIAIS DE NON-RÉPONSE DANS LA COHORTE PILOTE COSET-MSA

Cette partie est consacrée à l'étude de l'apport de données supplémentaires dans la prise en compte des biais de non-réponse dans la phase pilote Coset-MSA pour l'estimation de prévalences.

Le biais de non-réponse étant fonction du taux de réponse et de la covariance entre la probabilité de réponse et la variable d'intérêt, deux options sont classiquement possibles pour réduire les biais de non-réponse : utiliser de l'information auxiliaire directement liée à la thématique de l'enquête pour corriger la non-réponse une fois les données recueillies, ou bien maximiser autant que possible le taux de réponse à l'enquête.

On considère le scénario de référence comme étant celui où les seules données disponibles sont les données de questionnaire des répondants de l'enquête initiale ainsi que les variables sociodémographiques pour l'ensemble de l'échantillon (cf. Tableau III-4 ; données représentées par **X**).

On considère que les données supplémentaires (cf. Tableau III-4 ; données représentées par X) sont soit des répondants supplémentaires (les répondants à l'enquête complémentaire) soit des informations auxiliaires supplémentaires (les données du SNIIR-AM et de la MSA ou les paradonnées pour l'enquête complémentaire).

Trois scénarios ont été envisagés pour minimiser les biais de non-réponse :

- scénario 1 : étude de l'apport d'informations auxiliaires (les données du SNIIR-AM et de la MSA) dont le lien est direct avec la thématique de l'enquête ;

- scénario 2 : étude de l'apport d'une enquête réalisée auprès d'un échantillon de non-répondants (enquête complémentaire), construite de manière à obtenir un taux de réponse maximal, ainsi que d'informations auxiliaires collectées sur l'ensemble de cet échantillon (les paradonnées) qui n'ont pas (ou pas toujours) de lien direct avec la thématique de l'enquête ;
- scénario 3 : étude de l'apport d'une enquête réalisée auprès d'un échantillon de non-répondants lorsqu'on dispose par ailleurs d'informations auxiliaires (les données du SNIIR-AM et de la MSA) dont le lien est direct avec la thématique de l'enquête.

IV.1 MÉTHODOLOGIE COMMUNE

Quel que soit le scénario considéré, la méthodologie commune suivante a été suivie :

IV.1.1 ETUDE DE LA NON-RÉPONSE SELON LES INFORMATIONS AUXILIAIRES CONSIDÉRÉES

La réponse à l'enquête (initiale ou complémentaire) a été modélisée par régression logistique en fonction des informations auxiliaires. Compte tenu de leur nombre et de leur corrélation, les informations auxiliaires ont dans un premier temps été étudiées par groupe ; pour un groupe donné (par exemple le groupe des variables relatives à la santé) les variables qui étaient associées à la réponse avec un degré de signification inférieur à 0,20 ont été sélectionnées.

Dans un deuxième temps, les variables sélectionnées provenant des différents groupes ont été incluses dans un même modèle. Pour éviter l'inclusion de trop de variables dans le modèle final, seules celles associées à la réponse avec un degré de signification inférieur à 0,05 pour l'enquête initiale et inférieure à 0,20 pour l'enquête complémentaire ont été retenues.

IV.1.2 CONSTRUCTION DES GROUPES HOMOGÈNES DE RÉPONSE

La méthode des scores (30, 59, 81) a été choisie pour estimer les facteurs correctifs de non-réponse ; elle consiste à trier les probabilités de réponse prédites par le modèle, à découper l'échantillon en k groupes de taille égale et à calculer un taux de réponse dans chacun des k groupes. L'inverse du taux de réponse observé est ensuite utilisé comme facteur correctif de la non-réponse.

IV.1.3 EVALUATION DE LA CONTRIBUTION DES DONNÉES SUPPLÉMENTAIRES POUR RÉDUIRE LE BIAIS DE NON-RÉPONSE

Deux types de variables ont été utilisés pour étudier la contribution de données supplémentaires : des variables issues du questionnaire et des variables issues des systèmes d'information existants. L'utilisation de variables issues des systèmes d'information existant présente des atouts et des limites. Le principal intérêt est qu'elles permettent de disposer d'une prévalence gold standard. De plus, elles permettent d'étudier spécifiquement le biais de non-réponse, indépendamment du biais de mesure ; ce point est important pour les scénarios 2 et 3 dans lesquels les données de questionnaire sont recueillies par plusieurs modes. Leur principale limite est qu'il est délicat de les utiliser à la fois comme informations auxiliaires et comme gold standard ; ce point concerne les scénarios 1 et 3. Aussi l'accent sera mis sur les données de questionnaire pour le scénario 1 et sur les données issues des systèmes d'information existant pour le scénario 2. Pour le scénario 3, on étudiera principalement les données de questionnaire.

IV.1.3.1 Variables d'intérêt issues du questionnaire

Les variables suivantes, issues du questionnaire, ont été étudiées :

- trois variables relatives à la santé ou à des comportements à risque pour la santé :
    - état de santé général perçu comme très bon
    - fumeur
    - consommation d'alcool supérieure ou égale à deux fois par semaine
- deux variables « sociodémographiques » :
    - marié
    - niveau d'études atteint inférieur au baccalauréat
- cinq variables relatives à l'emploi :
    - catégorie sociale : agriculteurs exploitants, artisans ou commerçants, cadres ou professions libérales, professions intermédiaires, employés, ouvriers
    - secteur d'activité primaire
    - faible ressenti dans l'intensité des efforts physiques
    - « Je reçois le respect que je mérite au travail » issu du questionnaire de Siegrist (117)
    - exposition à des bruits intenses.

IV.1.3.2 Variables d'intérêt issues des systèmes d'information existants

Deux types de variables existantes ont été utilisées : des variables relatives à la santé, issues du SNIIR-AM et des variables relatives à l'emploi, extraites des bases de données de la MSA. Ces variables sont toutes binaires.

- au moins 100 remboursements d'actes pour soins de ville entre 2008 et 2010 ;
- hospitalisation entre 2008 et 2010 ;
- travail dans le secteur primaire ;
- durée d'emploi inférieure à 10 ans ;
- absentéisme au travail (indemnités journalières) ;
- accident du travail ou maladie professionnelle entre 2002 et 2008.

IV.1.3.3 Estimations de prévalence

Plusieurs prévalences ont été estimées grâce aux sujets répondants à l'enquête (initiale ou initiale et complémentaire selon le scénario) sous plusieurs hypothèses spécifiques sur le processus de non-réponse :

- sous l'hypothèse MCAR, on supposait qu'il n'y avait pas d'association entre la probabilité de réponse et la variable d'intérêt du questionnaire. Aucune correction pour la non-réponse n'était donc nécessaire ;
- sous l'hypothèse MAR(info_aux), on supposait que les causes communes expliquant le lien entre la probabilité de réponse et la variable d'intérêt correspondaient aux informations auxiliaires (info_aux) incluses.

Ces prévalences sont notées $\hat{p}_{rep\_enq,nr\_process}$ où rep_enq indique à quelle enquête appartiennent les sujets répondants et nr_procees indique le processus de non réponse. L'expression de $\hat{p}_{rep\_enq,nr\_process}$ sera explicitée dans chaque scénario. L'estimation des intervalles de confiance prenait en compte la variabilité liée au plan de sondage et à la non-réponse.

De plus, pour les variables issues des systèmes d'information existant, la prévalence gold standard, notée $\hat{p}_{GS}$, était également calculée :

$$\hat{p}_{GS} = \frac{\sum_{i \in s_{EI}} \frac{y_i}{\pi_{EI,i}}}{\sum_{\in s_{EI}} \frac{1}{\pi_{EI,i}}}$$

Cette prévalence $\hat{p}_{GS}$ prend en compte les probabilités d'inclusion à l'enquête initiale et l'estimation des intervalles de confiance qui lui sont associés prend en compte la variabilité liée au plan de sondage.

IV.1.3.4 Erreurs relatives

Pour les variables issues des systèmes d'information existant, une erreur relative de non-réponse était ensuite calculée pour chaque prévalence estimée :

$$ER = \frac{\hat{p}_{rep\_enq,nr\_process} - \hat{p}_{GS}}{\hat{p}_{GS}} * 100$$

Une erreur relative inférieure à 10% était considérée comme acceptable (127). Une erreur relative était supposée correspondre à un biais de non-réponse, comme l'avait considéré Olson auparavant (92).

IV.2 PRÉSENCE D'INFORMATIONS AUXILIAIRES DE QUALITÉ : L'APPORT DES BASES MÉDICO-ADMINISTRATIVES

IV.2.1 CONTEXTE ET OBJECTIFS

Cette partie correspond au scénario 1, donc à l'étude de l'apport d'informations auxiliaires (les données du SNIIR-AM et de la MSA) dont le lien est direct avec la thématique de l'enquête.

Comme nous l'avons vu plus haut, quand un biais de non-réponse est dû au fait que le lien entre la probabilité de réponse et la variable d'intérêt est complètement expliqué par des causes communes, il peut être totalement corrigé par repondération par l'inverse de la probabilité de réponse (IPW) ; cette méthode nécessite que les causes communes soient connues à la fois pour les répondants et les non-répondants. Le problème est alors de modéliser correctement la probabilité de réponse en fonction de des causes communes connues. Ces dernières ne sont pas toujours très bien documentées et l'accès à des informations pertinentes disponibles à la fois chez les répondants et les non-répondants est en général difficile. L'accès à des bases de données médico-administratives telles que les données du SNIIR-AM et de la MSA offrent de nouvelles perspectives.

Plusieurs études épidémiologiques ont montré que la non-réponse était associée à l'âge, au sexe, au statut marital, à des comportements à risque pour la santé, à des remboursements de soins médicaux ou au statut d'emploi (42, 71, 85, 91, 127). Cependant, peu d'études ont utilisé ces résultats pour corriger pour la non-réponse des prévalences estimées (3, 110) et les méthodes de repondération sont de fait rarement utilisées et peu connues dans la communauté épidémiologique.

L'objectif principal est de montrer comment la repondération peut corriger la non-réponse totale dans une enquête épidémiologique de surveillance des risques professionnels, en utilisant des données issues des bases de données administratives relatives à la santé et au travail, en supplément des variables sociodémographiques classiquement utilisées.

IV.2.2 POPULATION ET MÉTHODES

La plupart des informations sur la méthodologie suivie a été développée en IV.1. Les spécificités pour cette partie sont développées ci-après.

IV.2.2.1 Données étudiées

On peut se référer au Tableau III-4 pour une information résumée.

- *Enquête*

L'échantillon est l'ensemble des répondants à l'enquête initiale.

- *Informations auxiliaires pour corriger la non-réponse*

Les variables sociodémographiques, notées X, ainsi que les variables du SNIIR-AM et de la MSA issues des systèmes d'information existants, notées V, ont été utilisées pour corriger la non-réponse (cf. III.2.2.2.2.1).

- *Variables d'intérêt*

L'analyse principale concerne les prévalences des variables issues du questionnaire. Les prévalences des variables issues des bases de données administratives sont utilisées ici pour étayer les résultats de l'analyse principale.

IV.2.2.2 Analyses statistiques

- *Taux de réponse à l'enquête initiale*

Les taux de réponse ont été estimés en s'inspirant des recommandations de l'American Association for Public Opinion Research (AAPOR) (124).

Le taux de contact a été estimé en divisant le nombre de plis distribuables à l'enquête par la taille de l'échantillon (n = 10 000).

Deux taux de réponse ont été estimés : un taux de réponse brut, qui a été estimé en divisant le nombre de répondants à l'enquête par la taille de l'échantillon (n = 10 000) et un taux de réponse parmi les personnes ayant été contactées, qui a été estimé en divisant le nombre de répondants à l'enquête par le nombre de plis distribuables.

Les probabilités d'inclusion des personnes variant peu d'une personne à l'autre, tous ces taux ont été estimés sans prendre en compte les pondérations issues du plan de sondage.

- *Probabilités de réponse à l'enquête initiale*

La probabilité de réponse a été étudiée selon les variables sociodémographiques, puis selon les variables sociodémographiques et les variables du SNIIR-AM et de la MSA.

- *Constitution des groupes homogènes de réponse*

Afin d'étudier différentes hypothèses sur le processus de non-réponse, deux types de groupes homogènes de réponse ont été construits par la méthode des scores.

A partir des modèles finaux construits dans l'étape précédente, les probabilités de réponse suivantes ont été estimées :

- $\hat{\delta}_{MAR_{EI}(X)}$, sous l'hypothèse notée $MAR_{EI}(X)$ que pour l'enquête initiale, la non-réponse est aléatoire conditionnellement aux variables sociodémographiques (*ie* les causes communes expliquant le lien entre la probabilité de réponse à l'enquête initiale et la variable d'intérêt considérée sont les variables sociodémographiques) ;
- $\hat{\delta}_{MAR_{EI}(X,V)}$, sous l'hypothèse notée $MAR_{EI}(X,V)$.que pour l'enquête initiale, la non-réponse est aléatoire conditionnellement aux variables sociodémographiques et aux variables du SNIIR-AM et de la MSA.

- *Estimation des prévalences*

Les prévalences ont été estimées selon trois hypothèses sur le processus de non-réponse : $MCAR_{EI}$, $MAR_{EI}(X)$ et $MAR_{EI}(X,V)$.

- Sous l'hypothèse $MCAR_{EI}$:

$$\hat{p}_{s_{EI,r};MCAR_{EI}} = \frac{\sum_{i \in s_{EI,r}} w_{MCAR_{EI},i} y_i}{\sum_{i \in s_{EI,r}} w_{MCAR_{EI},i}} \text{ où } w_{MCAR_{EI},i} = \frac{1}{\pi_{EI,i}}$$

- Sous l'hypothèse $MAR_{EI}(X)$:

$$\hat{p}_{s_{EI,r};MAR_{EI}(X)} = \frac{\sum_{i \in s_{EI,r}} w_{MAR_{EI}(X),i} y_i}{\sum_{i \in s_{EI,r}} w_{MAR_{EI}(X),i}} \text{ où } w_{MAR_{EI}(X),i} = \frac{1}{\pi_{EI,i}} * \frac{1}{\hat{\delta}_{MAR_{EI}(X),i}}$$

- Sous l'hypothèse $MAR_{EI}(X,V)$:

$$\hat{p}_{s_{EI,r};MAR_{EI}(X,V)} = \frac{\sum_{i \in s_{EI,r}} w_{MAR_{EI}(X,V),i} y_i}{\sum_{i \in s_{EI,r}} w_{MAR_{EI}(X,V),i}} \text{ où } w_{MAR_{EI}(X,V),i} = \frac{1}{\pi_{EI,i}} * \frac{1}{\hat{\delta}_{MAR_{EI}(X,V),i}}$$

L'estimation de la variance de ces estimations de prévalence a pris en compte le plan de sondage et la variabilité liée à la non-réponse (cf. II.2.3.3.4) ; elle a été calculée grâce au programme Calker de l'Insee (15).

IV.2.3 RÉSULTATS

IV.2.3.1 Description de l'échantillon tiré au sort

Parmi les 10 000 personnes tirées au sort, 56,3% étaient salariées, 67,8% étaient des hommes et leur âge médian était de 43 ans.

Parmi les 9 307 personnes ayant des informations auxiliaires complètes, entre 2008 et 2010, 90% ont eu un remboursement pour un acte médical, 29,2% ont eu une hospitalisation et 89,2% avaient travaillé en 2010 en tant qu'actif affilié à la MSA.

IV.2.3.2 Taux de réponse

Le taux de réponse à l'enquête initiale était de 23,6% (95% des questionnaires ont été distribués et 24,8% des personnes ayant reçu un questionnaire ont répondu à l'enquête).

IV.2.3.3 Propension à répondre

IV.2.3.3.1 Selon les variables sociodémographiques

L'étude de la réponse selon les variables sociodémographiques montre qu'elle est significativement plus élevée dans les groupes suivants : les femmes (OR=1,3 ; IC 95% [1,2 ; 1,5]), les personnes les plus âgées (OR=1,6 ; IC 95% [1,4 ; 1,8] pour les personnes âgées entre 50 et 65 ans vs les personnes âgées entre 18 et 34 ans), les salariés (OR=0,8; IC 95% [0,7 ; 0,9] pour les non-salariés vs les salariés) et les personnes habitant en Saône-et-Loire

(OR=1,7 ; IC 95% [1,4 ; 1,9] pour les personnes vivant en Saône-et-Loire vs personnes vivant dans les Bouches-du-Rhône) (Tableau IV-1).

Tableau IV-1 : Variables associées à la réponse au questionnaire postal dans le modèle final (hypothèse MAR(X))

		n	OR	IC 95%
Sexe	Homme	6775	1	
	Femme	3225	1,3	(1,2 ; 1,5)
Age	18-34 ans	2677	1	
	35-49 ans	4323	1,4	(1,2 ; 1,6)
	50-65 ans	3000	1,6	(1,4 ; 1,8)
Statut d'emploi	Salarié	5845	1	
	Non salarié	4155	0,8	(0,7 ; 0,8)
Département	Bouches-du-Rhône	2000	1	
	Finistère	2000	1,4	(1,2 ; 1,7)
	Pas-de-Calais	2000	1,2	(1,0 ; 1,4)
	Pyrénées-Atlantiques	2000	1,3	(1,1 ; 1,5)
	Saône-et-Loire	2000	1,7	(1,4 ; 1,9)

IV.2.3.3.2 Selon les variables sociodémographiques, du SNIIR-AM et de la MSA

Les résultats de l'étude de la propension à répondre selon les variables sociodémographiques, du SNIIR-AM et de la MSA sont présentés dans le Tableau IV-2 et le Tableau IV-3.

Parmi les 9 307 personnes avec informations auxiliaires complètes, 2 320 ont répondu à l'enquête postale.

Un nombre important de variables a été étudié, la plupart d'entre elles étaient associées significativement à la probabilité de réponse au questionnaire dans les analyses bivariées. Les variables sélectionnées dans le modèle final sont décrites ci-après.

La propension à répondre à l'enquête initiale est, dans le modèle multivarié, associée à plusieurs variables sociodémographiques, à plusieurs variables relatives aux recours aux soins, aux hospitalisations, aux arrêts de travail et au dernier emploi exercé. Elle n'est en revanche pas associée à la notion de déclaration en accidents du travail ou maladie professionnelle.

Concernant les variables sociodémographiques, la propension à répondre est associée au département et est plus élevée chez les 35 ans et plus en comparaison aux 18-34 ans (OR=1,3 IC 95% [1,1 ; 1,5] pour les 35-49 ans ; OR=1,4 IC 95% [1,2 ; 1,6] pour les 50-65 ans).

Pour les variables relatives au recours aux soins entre 2008 et 2010, la propension à répondre est plus élevée chez les personnes ayant recours à un médecin généraliste (OR=1,6 IC 95% [1,2 ; 2,1] entre 5 et 9 fois versus aucun) ou spécialiste (OR=1,3 IC 95% [1,1 ; 1,6] entre 5 et 9 fois versus aucun), à un dentiste (OR=1,3 IC 95% [1,2 ; 1,5] entre 1 et 9 fois versus aucun) ou un auxiliaire médical (OR=1,2 IC 95% [1,1 ; 1,4] au moins 10 fois versus aucun), à un laboratoire d'analyses (OR=1,3 IC 95% [1,1 ; 1,5] entre 1 et 9 fois versus aucun) ; elle est en revanche moins élevée chez les personnes ayant eu des médicaments concernant la sphère "voies digestives et métabolisme" (OR=0,8 IC 95% [0,8 ; 1,0] pour au moins 100 boîtes versus aucune) ou la sphère "système nerveux" (OR=0,8 IC 95% [0,7 ; 1,0] pour au moins 50 boîtes versus aucune). Pour ce qui concerne les hospitalisations entre 2008 et 2010, la propension à répondre est plus élevée chez les personnes hospitalisées pour maladie de l'appareil circulatoire (OR=1,3 IC 95% [1,0 ; 1,8] versus aucune) et moins élevée chez les personnes ayant eu au moins une hospitalisation (OR=0,8 IC 95% [0,7 ; 0,9] versus aucune).

Pour les variables relatives aux arrêts de travail, la propension à répondre est plus élevée chez les personnes arrêtées pour maternité (OR=1,5 IC 95% [1,2 ; 1,8] versus aucun arrêt) et moins

élevée chez les personnes arrêtées moins de 60 jours pour maladie (OR=0,7 IC 95% [0,6 ; 0,9] pour des arrêts cumulés de deux mois versus aucun arrêt).

Enfin, concernant le dernier emploi exercé, la propension à répondre est plus élevée chez les personnes travaillant dans des organismes de service en comparaison de personnes travaillant dans les cultures spécialisées (OR=2,1 IC 95% [1,3 ; 3,5]), ayant des durées d'emploi supérieures à 6 mois (OR=1,5 IC 95% [1,1 ; 1,9] pour des durées entre 1 et 6 ans versus moins de 6 mois), salariées (OR=1,5 IC 95% [1,3 ; 1,8]), et dont le dernier emploi en date datait de 2010 (OR=1,3 IC 95% [1,1 ; 1,7] versus 2008).

Tableau IV-2 : Variables associées à la réponse au questionnaire postal dans le modèle final

		n	OR	IC 95%
Variables relatives à des remboursements de soins de ville entre 2008 et 2010				
Recours à un médecin généraliste	Aucun	813	1	
	Entre 1 et 4 fois	2439	1,4	(1,1 ; 1,8)
	Entre 5 et 9 fois	2375	1,6	(1,2 ; 2,1)
	Au moins 10 fois	3680	1,4	(1,1 ; 1,9)
Recours à un médecin spécialiste	Aucun	1955	1	
	Entre 1 et 4 fois	3318	1,1	(0,9 ; 1,3)
	Entre 5 et 9 fois	1866	1,3	(1,1 ; 1,6)
	Au moins 10 fois	2168	1,7	(1,4 ; 2,1)
Recours à un dentiste	Aucun	3069	1	
	Entre 1 et 9 fois	4895	1,3	(1,2 ; 1,5)
	Au moins 10 fois	1343	1,2	(1,0 ; 1,4)
Recours à un laboratoire d'analyses	Aucun	3245	1	
	Entre 1 et 9 fois	4732	1,3	(1,1 ; 1,5)
	Au moins 10 fois	1330	1,3	(1,1 ; 1,6)
Recours à un auxiliaire médical	Aucun	4628	1	
	Entre 1 et 9 fois	2404	1,0	(0,9 ; 1,2)
	Au moins 10 fois	2275	1,2	(1,1 ; 1,4)
Nb total de boites facturées : VOIES DIGESTIVES ET METABOLISME	Aucun	2930	1	
	Entre 1 et 99	5223	0,9	(0,8 ; 1,0)
	Au moins 100	1154	0,8	(0,7 ; 1,0)
Nb total de boites facturées : SYSTEME NERVEUX	Aucun	1731	1	
	Entre 1 et 49	3701	0,9	(0,8 ; 1,1)
	Au moins 50	3875	0,8	(0,7 ; 1,0)
Variables relatives aux hospitalisations entre 2008 et 2010				
Nombre d'hospitalisations	Aucune	6587	1	
	Une	1747	0,8	(0,7 ; 1,0)
	Au moins 2	973	1,0	(0,8 ; 1,2)
Hospitalisation pour maladie de l'appareil circulatoire	Non	9082	1	
	Oui	225	1,3	(1,0 ; 1,8)
Variables relatives aux indemnités journalières (IJ) entre 2008 et 2010				
Durée IJ pour maladie (en jours)	0	7007	1	
	1 à 29	1313	0,9	(0,7 ; 1,0)
	30 à 59	388	0,7	(0,6 ; 1,0)
	Au moins 60	599	1,0	(0,8 ; 1,3)
IJ pour maternité	Non	8750	1	
	Oui	557	1,5	(1,2 ; 1,8)

Tableau IV-3 : Variables associées à la réponse au questionnaire postal dans le modèle final

		n	OR	IC 95%
Variables sociodémographiques au 31/12/2008				
Caisse d'affiliation MSA	Bouches du Rhône	1743	1	
	Pyrénées Atlantiques	1865	1,1	(0,9 ; 1,3)
	Finistère	1881	1,1	(1,0 ; 1,4)
	Pas de Calais	1911	1,4	(1,2 ; 1,6)
	Saône et Loire	1907	1,5	(1,3 ; 1,8)
Age	18 à 34 ans	2417	1	
	35 à 49 ans	4063	1,3	(1,1 ; 1,5)
	50 à 65 ans	2827	1,4	(1,2 ; 1,6)
Variables relatives au dernier emploi en date en 2008, 2009 ou 2010				
Secteur d'activité du dernier emploi principal	Cultures spécialisées	1832	1	
	Viticulture	500	1,6	(1,0 ; 2,7)
	Elevage spécialisé gros animaux	1976	1,5	(0,9 ; 2,6)
	Elevage spécialisé petits animaux	311	1,6	(0,9 ; 2,6)
	Polyculture, polyélevage	1101	1,7	(1,0 ; 3,0)
	Haras	122	1,3	(0,8 ; 2,2)
	Travaux forestiers	191	1,5	(0,8 ; 2,7)
	Entreprises de travaux agricoles	994	1,2	(0,7 ; 2,0)
	Coopération	973	1,5	(0,9 ; 2,5)
	Organismes de service	1092	2,1	(1,3 ; 3,5)
	Activités diverses	215	1,2	(0,7 ; 2,2)
Durée du dernier emploi principal	<6 mois	561	1	
	[6 mois-1 an[	544	1,1	(0,8 ; 1,5)
	[1-6[ans	2522	1,5	(1,1 ; 1,9)
	[6-10[ans	1491	1,3	(1,0 ; 1,7)
	[10-20[ans	2082	1,7	(1,3 ; 2,2)
	>=20 ans	2107	1,6	(1,2 ; 2,1)
Statut du dernier emploi principal	Non salarié	4050	1	
	Salarié	5257	1,5	(1,3 ; 1,8)
Année du dernier emploi principal	2008	478	1	
	2009	531	1,0	(0,7 ; 1,3)
	2010	8298	1,3	(1,1 ; 1,7)

IV.2.3.4 Construction des groupes homogènes de réponse

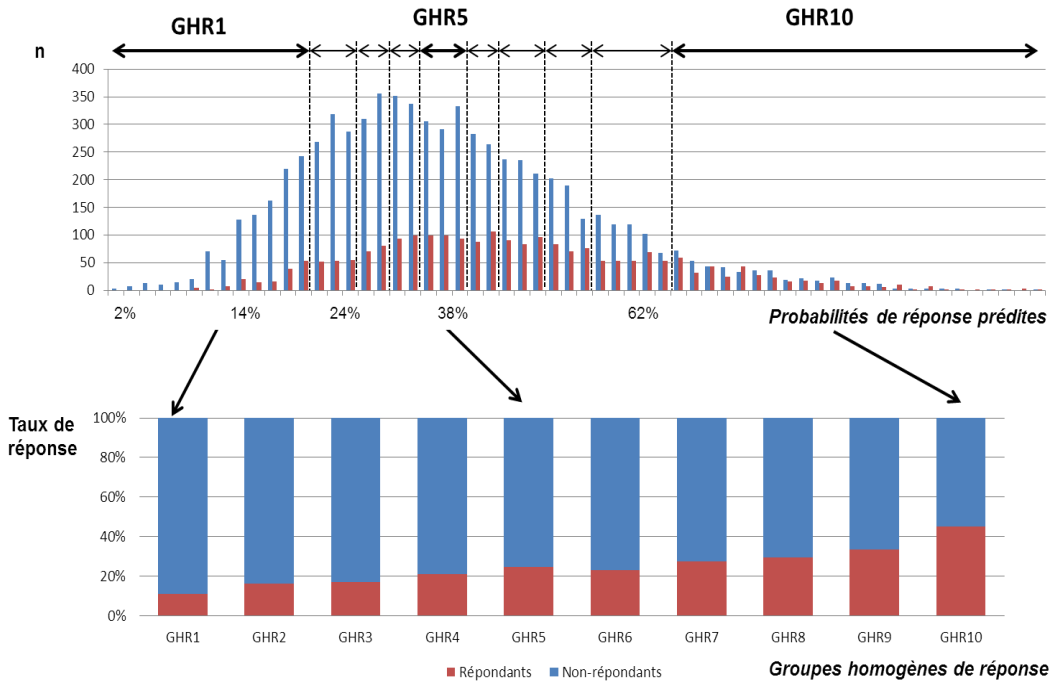
IV.2.3.4.1 Sous l'hypothèse MAR(X)

Les probabilités de réponse prédites par le modèle, qui étudie le lien entre la réponse et la variables sociodémographiques varient d'environ 12% à 39%. Apres constitution des groupes homogènes de réponse, le taux de réponse des groupes varie d'environ 17% à 32%.

IV.2.3.4.2 Sous l'hypothèse MAR(X,V)

Les probabilités de réponse prédites par le modèle, qui étudie le lien entre la réponse et la variables sociodémographiques, du SNIIR-AM et de la MSA, varient d'environ 2% à 62%. La Figure IV-1 montre que les distributions des répondants et des non-répondants sont différentes mais se chevauchent suffisamment. Ce point est important pour construire des groupes homogènes de réponse : en effet, il est nécessaire que chaque groupe contienne un pourcentage non nul de répondants, de manière à ce que ces répondants, une fois pondérés puissent représenter les non répondants du même groupe. Apres constitution des groupes homogènes de réponse, le taux de réponse des groupes varie d'environ 10% à 45%.

Figure IV-1 : Construction de 10 groupes homogènes de réponse (GHR1-GHR10) par la méthode des scores par quantiles



IV.2.3.5 Estimation des prévalences des variables du questionnaire

Les prévalences des variables du questionnaire estimées sous différentes hypothèses sur le processus de non-réponse sont présentées dans le Tableau IV-4 et le Tableau IV-5.

Pour les variables « fumeur », « niveau d'études inférieur au bac », et « ouvriers agricoles », alors que les prévalences estimées sous les hypothèses MCAR et MAR(X) varient peu, elles sont assez différentes sous l'hypothèse MAR(X,V). Par exemple, la prévalence pour fumeur augmente de 23,7% sous l'hypothèse MCAR à 26,1% sous l'hypothèse MAR(X,V). Cette différence est à souligner puisque la prévalence sous l'hypothèse MAR(X,V) n'est pas incluse dans l'intervalle de confiance estimé sous l'hypothèse MCAR. Il en est de même pour la prévalence de niveau d'études et d'ouvriers agricoles.

Pour les variables « état de santé », « agriculteur » et « secteur activité primaire », les prévalences augmentent quand on passe de l'hypothèse MCAR à l'hypothèse MAR(X) et de

l'hypothèse MAR(X) à l'hypothèse MAR(X,V). Pour les variables « marié », « artisan », « cadre », « profession intermédiaire », « employé » et « faible ressenti », on observe le phénomène inverse, c'est-à-dire une diminution. Pour toutes ces variables d'intérêt, les deux types d'informations auxiliaires, sociodémographiques et issues des bases médico-administratives contribuent à la diminution du biais de non-réponse. De plus, les estimations obtenues sous MAR(X,V) ne sont pas dans l'intervalle de confiance des prévalences estimées sous l'hypothèse MCAR.

Pour les autres variables (« consommation d'alcool supérieure à deux fois par semaine » « respect que je mérite au travail » ou « exposition à des bruits intenses »), les prévalences estimées sont peu modifiées quelle que soit l'hypothèse sur le processus de non-réponse.

IV.2.3.6 Estimation des prévalences des variables issues des bases de données existantes

Les erreurs relatives comparant les prévalences corrigées pour des variables issues de bases de données existantes sous les hypothèses MCAR, MAR(X) et MAR(X,V) avec un gold-standard sont présentées dans le Tableau IV-6. Elles sont clairement les plus faibles sous l'hypothèse MAR(X,V).

Après correction de la non-réponse sur les variables sociodémographiques, les erreurs relatives diminuent, sauf pour l'absentéisme. Pour la variable « au moins 100 remboursements d'actes pour soins de ville », l'erreur relative reste néanmoins élevée (ER=29,6% sous l'hypothèse MCAR et ER=22,1% sous l'hypothèse MAR(X)).

Après la correction de la non-réponse sur les variables sociodémographiques, relatives à la santé et à l'emploi, toutes les erreurs relatives sont pratiquement inférieures à 5%, ce qui signifie que les prévalences estimées sous l'hypothèse MAR(X,V) sont très proches de la prévalence gold-standard. La seule variable pour laquelle l'erreur relative est supérieure à 5%

est la variable « au moins 100 remboursements d'actes pour soins de ville » ; néanmoins l'erreur relative est nettement diminuée, puisqu'elle passe de 22,1% sous l'hypothèse MAR(X) à 7,5% sous l'hypothèse MAR(X,V).

Tableau IV-4 : Prévalences des variables du questionnaire (variables de santé et sociodémographiques) estimées à partir des répondants de l'enquête initiale sous les hypothèses MCAR et MAR(X) et MAR(X,V)

	n	**MCAR**	**MAR(X)**	**MAR(X,V)**
		% **IC 95%**	**%** **IC 95%**	**%** **IC 95%**
Etat de santé général perçu comme très bon	1307	55,8 (53,7 ; 57,9)	57,0 (54,9 ; 59,1)	58,2 (55,9 ; 60,5)
Fumeur	518	23,7 (21,9 ; 25,5)	24,5 (22,6 ; 26,5)	26,1 (24,0 ; 28,3)
Consommation d'alcool ≥ deux fois/semaine	721	30,1 (28,2 ; 32,0)	30,4 (28,4 ; 32,4)	29,8 (27,7 ; 31,9)
Marié	1557	66,1 (64,1 ; 68,1)	64,0 (61,9 ; 66,1)	63,2 (60,9 ; 65,5)
Niveau d'études < bac	1455	64,0 (61,9 ; 66,0)	63,7 (61,5 ; 65,8)	67,5 (65,3 ; 69,6)

X : variables sociodémographiques, V : variables du SNIIR-AM et de la MSA

Tableau IV-5 : Prévalences des variables du questionnaire (variables relatives à l'emploi) estimées à partir des répondants de l'enquête initiale sous les hypothèses MCAR et MAR(X) et MAR(X,V)

	n	MCAR	MAR(X)	MAR(X,V)
		% IC 95%	% IC 95%	% IC 95%
Catégorie sociale				
Agriculteur exploitant	650	27,7 (25,8 ; 29,7)	29,7 (27,6 ; 31,7)	30,5 (28,3 ; 32,7)
Artisan, commerçant	43	1,8 (1,2 ; 2,3)	2,4 (1,6 ; 3,1)	2,7 (1,8 ; 3,7)
Cadre, profession libérale	155	7,7 (6,4 ; 8,8)	7,1 (6,0 ; 8,3)	6,1 (5,1 ; 7,2)
Profession intermédiaire	364	17,0 (15,4 ; 18,6)	16,6 (15,0 ; 18,3)	14,4 (12,9 ; 16,0)
Employé	310	14,0 (12,5 ; 15,5)	12,6 (11,2 ; 14,0)	11,6 (10,2 ; 13,1)
Ouvrier	675	31,8 (29,8 ; 33,8)	31,6 (29,5 ; 33,7)	34,6 (32,2 ; 36,9)
Secteur d'activité primaire	1362	63,1 (61,0 ; 65,2)	65,2 (63,0 ; 67,3)	68,7 (66,6 ; 70,9)
Faible ressenti dans l'intensité des efforts physiques	669	34,2 (32,0 ; 36,4)	32,9 (30,7 ; 35,1)	30,2 (27,9 ; 32,4)
Respect que je mérite au travail	1144	59,4 (57,2 ; 61,7)	59,4 (57,0 ; 61,8)	60,4 (57,9 ; 62,9)
Exposition à des bruits intenses	364	19,1 (17,3 ; 20,9)	19,8 (17,9 ; 21,7)	20,7 (18,6 ; 22,8)

X : variables sociodémographiques, V : variables du SNIIR-AM et de la MSA

Tableau IV-6 : Prévalences des variables de systèmes d'information existants estimées sur l'échantillon complet (gold standard), à partir des répondants de l'enquête initiale sous les hypothèses MCAR, MAR(X) et MAR(X,V)

	Répondants à l'enquête initiale				**Echantillon complet (gold standard)**		**Erreur relative**		
	n	**MCAR (a)**	**MAR(X) (b)**	**MAR(X,V) (c)**	**n**	**(f)**	**(a-f) /f**	**(b-f) /f**	**(c-f) /f**
		% IC 95%	**% IC 95%**	**% IC 95%**		**% IC 95%**			
Au moins 100 remboursements d'actes pour soins de ville	585	25,8 (23,9 ; 27,6)	24,3 (22,5 ; 26,2)	21,4 (19,6 ; 23,1)	1856	19,9 (19,1 ; 20,7)	29,6	22,1	7,5
Hospitalisation	534	23,4 (21,6 ; 25,2)	22,9 (21,1 ; 24,7)	20,9 (19,1 ; 22,7)	1950	20,9 (20,1 ; 21,7)	12,0	9,6	0,0
Travail dans le secteur primaire	1349	57,2 (55,1 ; 59,3)	58,8 (56,7 ; 61,0)	61,8 (59,6 ; 64,0)	5869	61,1 (60,1 ; 62,1)	-6,4	-3,8	1,1
Durée d'emploi inférieure à 10 ans	1193	53,4 (51,3 ; 55,4)	53,7 (51,5 ; 55,9)	56,6 (54,4 ; 58,9)	5137	56,4 (55,4 ; 57,4)	-5,3	-4,8	0,4
Absentéisme au travail	803	36,0 (34,0 ; 38,0)	35,8 (33,6 ; 37,9)	33,8 (31,6 ; 35,9)	3089	33,4 (33,0 ; 34,8)	7,8	7,2	1,2
Accident du travail ou maladie professionnelle	600	25,6 (23,8 ; 27,4)	26,4 (24,5 ; 28,4)	26,0 (23,9 ; 28,0)	2514	26,9 (25,9 ; 27,9)	-4,8	-1,9	-3,3

X : variables sociodémographiques, V : variables du SNIIR-AM et de la MSA

IV.2.4 DISCUSSION

Dans l'étude Coset-MSA, la probabilité de réponse était non seulement associée aux variables sociodémographiques, mais aussi à des variables relatives à la santé et au travail. La comparaison des prévalences estimées pour les variables d'intérêt collectées par questionnaire ou via les systèmes d'information existants, sous différentes hypothèses de processus de non-réponse a montré des différences modérées mais notables dont la magnitude variait selon la variable d'intérêt étudiée ; ces différences reflètent l'association entre les probabilités de réponse estimées et les variables d'intérêt.

Les facteurs associés à la probabilité de réponse sont globalement consistants avec ceux reportés dans la littérature. Beaucoup d'études épidémiologiques ont trouvé que les personnes les plus âgées participaient plus que les personnes les plus jeunes (45, 71, 75, 101, 127). Nous avons trouvé par ailleurs que les hommes participaient moins que les femmes, ce qui est en accord avec d'autres publications (45, 71, 85, 127). Cependant, dans notre étude, cette association ne persistait pas après ajustement sur les données relatives à la santé. Globalement, les résultats de l'analyse peuvent s'interprétés par l'hypothèse suivante : les personnes prenant soin de leur santé (celles qui ont eu un remboursement pour un soin médical ou paramédical) avaient une probabilité plus élevée de répondre, alors qu'à nombre de recours aux soins équivalent, les personnes avec des problèmes de santé plus sérieux (celles qui étaient remboursées pour des prescriptions médicamenteuses ou des hospitalisations) avaient une probabilité de réponse moins élevée. La plupart de ces associations ont déjà été rencontrées dans des études épidémiologiques (31, 71, 75, 101, 127), mis-à-part pour la prescription de médicaments (75, 91, 101, 127) où une association positive entre cette variable et la participation a été trouvée. Cependant, dans ces études, le remboursement de médicaments prescrits était souvent étudié séparément des autres types de

recours aux soins ; dans notre étude, on observait dans un premier temps une association brute positive entre la probabilité de réponse et la prescription d'un médicament remboursé, mais le sens de l'association changeait de sens après ajustement dans le modèle sur le nombre de visites chez le médecin généraliste. De manière semblable, la probabilité de réponse significativement plus élevée chez les femmes en univarié et non différente de celle des hommes dans le modèle final peut s'expliquer entièrement par leur attitude plus positive vis-à-vis de la prévention. Lorsqu'on s'intéresse aux variables relatives à la profession, il est plus difficile de comparer nos résultats avec ceux reportés dans la littérature. La catégorie sociale n'était pas disponible parmi les données auxiliaires et seuls des proxys de la catégorie sociale ont été étudiés auparavant (42, 43, 75, 85, 101). Compte tenu de la taille du questionnaire (40 pages) et du mode de collecte des données (questionnaire auto-administré), on peut supposer que les salariés d'organismes de service avaient plus de disponibilité, donc étaient plus enclins à répondre à l'enquête. Par ailleurs, les personnes ayant un emploi en tant qu'affilié à la MSA en 2010 avaient de meilleurs taux de réponse, que celles dont la notification d'un emploi datait d'avant, possiblement parce qu'elles se sont senties plus impliquées par l'étude que celles qui avaient quitté la MSA entre temps. Par ailleurs, la participation à l'enquête était plus faible parmi les personnes ayant un emploi peu stable (d'une durée inférieure à 6 mois).

Même si toutes ces informations auxiliaires n'ont pas été collectées dans un objectif épidémiologique et qu'elles peuvent également être affectées d'erreurs de mesure, ces données présentent l'avantage essentiel d'avoir été collectées indépendamment du fait que les personnes ou non aient répondu à l'enquête. Par ailleurs, pour expliquer la non-réponse par des variables associées à la fois à la probabilité de réponse et aux variables d'intérêt, une autre force de notre étude était que des variables nombreuses et variées ont pu être utilisées : des variables sociodémographiques, relatives à la santé et à l'emploi. Ceci est donc un atout majeur dans cette étude portant sur la santé et le travail.

Notre analyse sur les prévalences de variables issues des bases de données administratives confirme que le modèle de non-réponse a été correctement spécifié, puisque nous avons pu modifier correctement les valeurs des prévalences dont nous connaissions la vraie valeur. Il est cependant important de noter que les variables issues des systèmes d'information existants ont été utilisées dans deux optiques : estimer des prévalences de variables d'intérêt et disposer d'informations auxiliaires de qualité pour corriger la non-réponse. Il était donc attendu que les estimations obtenues sous l'hypothèse MAR(X,V) pour les variables d'intérêt issues des systèmes d'information existants soient proches des prévalences gold-standard ; néanmoins, étant donné que même pour les variables d'intérêt non incluses dans le modèle final de probabilité de réponse (par exemple « au moins 100 remboursements d'actes pour soins de ville » ou « travail dans le secteur primaire »), les erreurs sous l'hypothèse MAR(X,V) sont faibles, cela renforce l'utilité des informations auxiliaires pour corriger les biais de non réponse des variables de questionnaire tout en jugeant raisonnable qu'il subsiste un biais résiduel.

L'accès à des informations auxiliaires pertinentes n'est pas aisé. Celles-ci doivent être fortement associées au sujet de l'étude, recueillies de manière standardisée, centralisées et il doit être possible de les apparier au niveau individuel avec les données de l'étude ; par ailleurs, la mise à disposition par les détenteurs de ces données et les autorisations légales pour les exploiter peuvent être difficile à obtenir.

Le taux de réponse à notre enquête est plutôt faible (24%). Cependant, même si le taux de réponse est théoriquement un élément prédicteur du biais de non-réponse (7), une revue de la littérature (49) a montré que le taux de réponse n'était pas principal facteur dans le biais de non-réponse : le principal facteur est la non disponibilité d'informations auxiliaires pertinentes pour corriger la non-réponse. Dans notre étude, nous nous sommes intéressés à

l'épidémiologie des risques professionnels et comme les données utilisées pour corriger la non-réponse étaient relatives à la santé et au travail on peut penser raisonnablement que le biais de non réponse a été fortement réduit, mais sans écarter l'existence d'un biais résiduel.

IV.3 ABSENCE D'INFORMATIONS AUXILIAIRES DE QUALITÉ : L'APPORT D'UNE ENQUÊTE AUPRÈS DE NON-RÉPONDANTS

IV.3.1 CONTEXTE ET OBJECTIF

Dans le scénario 1, nous supposions que nous disposions d'informations auxiliaires de qualité, directement liées à la thématique de l'enquête. Cependant, cette configuration nécessite à la fois l'existence d'un système d'information centralisé et un accès à ces bases de données, ce qui n'est pas évident en pratique.

C'est pour cette raison que nous avons proposé le scénario 2 qui se place dans les conditions où les informations auxiliaires issues des bases administratives ne sont pas disponibles. Ce deuxième scénario suppose que les données supplémentaires sont les données de questionnaire d'une enquête auprès de non-répondants, dont le protocole d'enquête tentait de maximiser autant que possible le taux de réponse, ainsi que les paradonnées relatives à cette enquête. Rappelons que les paradonnées n'ont pas toujours de liens avec la thématique de l'enquête.

L'objectif de cette étude est double : évaluer la contribution de l'enquête complémentaire dans la phase pilote de Coset-MSA pour diminuer les biais de non-réponse, et évaluer la contribution additionnelle des paradonnées prises comme informations auxiliaires.

IV.3.2 POPULATION ET MÉTHODES

La plupart des informations sur la méthodologie suivie a été développée en IV.1. Les spécificités pour cette partie sont développées ci-après.

IV.3.2.1 Données étudiées

- *Enquêtes*

L'échantillon est constitué soit de l'ensemble des répondants à l'enquête initiale pour le scénario de référence soit de l'ensemble des répondants à l'enquête initiale et à l'enquête complémentaire pour le scénario 2.

- *Variables d'intérêt*

Les variables d'intérêt étaient d'une part des variables de systèmes d'information existants et d'autre part des variables de questionnaire. L'intérêt d'exploiter des variables issues de systèmes d'information existants est ici très grand car celles-ci permettent d'étudier les biais de non-réponse indépendamment des biais de mesure induits potentiellement par les différents modes de collecte de données (face-à-face, téléphone, enquête par autoquestionnaire postal) (13).

- *Informations auxiliaires pour corriger la non-réponse*

Que ce soit pour l'enquête initiale ou pour l'enquête complémentaire, les variables sociodémographiques (cf. III.2.2.2.2.1), notées X, ont été utilisées pour corriger la non-réponse.

Pour l'enquête complémentaire, les paradonnées (cf. III.2.2.2.2.2), notées Z, ont également été exploitées pour corriger la non-réponse.

IV.3.2.2 Analyses statistiques

- *Taux de réponse à l'enquête complémentaire et aux enquêtes combinées*

Les taux de réponse ont été estimés en s'inspirant des recommandations de l'American Association for Public Opinion Research (AAPOR) (124). Comme pour l'enquête initiale, des taux de contact et des taux de réponse non pondérés ont été estimés (cf IV.2.2.2).

Pour l'enquête en deux phases pour non-réponse (EDPNR), qui combine les deux enquêtes, un taux de réponse $\widehat{TR}_{EDPNR}$ pondéré par les poids de sondage a été estimé :

$$\widehat{TR}_{EDPNR} = \frac{\sum_{i \in s_{EI}} \frac{1}{\pi_{EI,i}} R_{EI,i} + \sum_{i \in s_{EC}} \frac{1}{\pi_{EI,i}\pi_{EC,i}} R_{EC,i}}{\sum_{i \in s_{EI}} \frac{1}{\pi_{EI,i}}}$$

avec $R_{EI,i} = \begin{cases} 1 \; si \; i \; répondant \text{ à } EI \\ 0 \; sinon \end{cases}$ $et \; R_{EC,i} = \begin{cases} 1 \; si \; i \; répondant \text{ à } EC \\ 0 \; sinon \end{cases}$

- *Probabilités de réponse à l'enquête initiale et à l'enquête complémentaire*

Pour l'enquête initiale, la probabilité de réponse a été étudiée selon les variables sociodémographiques.

Pour l'enquête complémentaire, la probabilité de réponse a été étudiée dans un premier temps selon les variables sociodémographiques seules, et dans un deuxième temps selon les variables sociodémographiques et les paradonnées. Lors de la construction de ce deuxième modèle, il aurait été préférable de construire deux sous-modèles (un pour chaque groupe selon le mode de recueil planifié) mais cela n'a pas été possible à cause du faible nombre de sujets par rapport au nombre de variables. De plus, toutes les catégories des paradonnées ont été construites de façon à ce qu'il puisse y avoir dans chacune d'elles des répondants et des non-répondants. Par exemple, les répondants de la modalité « pas d'appel » pour la paradonnée

« nombre d'appels téléphoniques » correspondent à des personnes interviewées en face-à-face et qui n'avaient pas pu être contactées par téléphone.

- *Constitution des groupes homogènes de réponse*

Afin d'étudier différentes hypothèses sur le processus de non-réponse, plusieurs types de groupes homogènes de réponse ont été construits en utilisant la méthode des scores.

A partir des modèles finaux construits dans l'étape précédente, les probabilités de réponse suivantes ont été estimées :

- $\hat{\delta}_{MAR_{EI}(X)}$, sous l'hypothèse notée $MAR_{EI}(X)$ que pour l'enquête initiale, la non-réponse est aléatoire conditionnellement aux variables sociodémographiques (scénario de référence) ;
- $\hat{\delta}_{MAR_{EC}(X)}$, sous l'hypothèse notée $MAR_{EC}(X)$ que pour l'enquête complémentaire, la non-réponse est aléatoire conditionnellement aux variables sociodémographiques ;
- $\hat{\delta}_{MAR_{EC}(X,Z)}$, sous l'hypothèse, notée $MAR_{EC}(X,Z)$.que pour l'enquête complémentaire, la non-réponse est aléatoire conditionnellement aux variables sociodémographiques et aux paradonnées.

- *Estimation des prévalences*

Les prévalences suivantes ont été estimées.

- Données issues de l'enquête initiale

Ces prévalences ont déjà été calculées au chapitre IV.2 Elles ont été estimées selon deux hypothèses sur le processus de non-réponse : $MCAR_{EI}$ et $MAR_{EI}(X)$.

- Sous l'hypothèse $MCAR_{EI}$:

$$\hat{p}_{s_{EI,r};MCAR_{EI}} = \frac{\sum_{i \in s_{EI,r}} w_{MCAR_{EI},i} y_i}{\sum_{i \in s_{EI,r}} w_{MCAR_{EI},i}} \text{ où } w_{MCAR_{EI},i} = \frac{1}{\pi_{EI,i}}$$

- Sous l'hypothèse $MAR_{EI}(X)$:

$$\hat{p}_{s_{EI,r};MAR_{EI}(X)} = \frac{\sum_{i \in s_{EI,r}} w_{MAR_{EI}(X),i} y_i}{\sum_{i \in s_{EI,r}} w_{MAR_{EI}(X),i}} \text{ où } w_{MAR_{EI}(X),i} = \frac{1}{\pi_{EI,i}} * \frac{1}{\hat{\delta}_{MAR_{EI}(X),i}}$$

Quelle que soit l'hypothèse considérée, l'estimation de la variance de l'estimation de ces prévalences a pris en compte le plan de sondage et la variabilité liée à la non-réponse (cf. II.2.3.3.4) ; elle a été calculée grâce au programme Calker de l'Insee (15).

- Données issues de l'enquête en deux phases pour non-réponse, ou enquêtes combinées

Pour les enquêtes combinées, les prévalences ont été estimées selon trois hypothèses sur le processus de non-réponse: $MCAR_{EC}$, $MAR_{EC}(X)$ et $MAR_{EC}(X, Z)$.

- Sous l'hypothèse $MCAR_{EC}$:

$$\hat{p}_{s_{EI,r} \cup s_{EC,r};MCAR_{EC}} = \frac{\sum_{i \in s_{EI,r} \cup s_{EC,r}} w_{MCAR_{EC},i} y_i}{\sum_{i \in s_{EI,r} \cup s_{EC,r}} w_{MCAR_{EC},i}} \quad \text{où}$$

$$w_{MCAR_{EC},i} = \begin{cases} \frac{1}{\pi_{EI,i}} \; si \; i \in s_{EI,r} \\ \frac{1}{\pi_{EI,i} * \pi_{EC/s_{EI,nr},i}} \; si \; i \in s_{EC,r} \end{cases}$$

- Sous l'hypothèse $MAR_{EC}(X)$:

$$\hat{p}_{s_{EI,r} \cup s_{EC,r}; MAR_{EC}(X)} = \frac{\sum_{i \in s_{EI,r} \cup s_{EC,r}} w_{MAR_{EC}(X),i} y_i}{\sum_{i \in s_{EI,r} \cup s_{EC,r}} w_{MAR_{EC}(X),i}} \quad \text{où}$$

$$w_{MAR_{EC}(X),i} = \begin{cases} \dfrac{1}{\pi_{EI,i}} \; si \; i \in s_{EI,r} \\ \dfrac{1}{\pi_{EI,i} * \pi_{EC/s_{EI,nr},i} * \hat{\delta}_{MAR_{EC}(X),i}} if \; i \in s_{EC,r} \end{cases}$$

- Sous l'hypothèse $MAR_{EC}(X, Z)$:

$$\hat{p}_{s_{EI,r} \cup s_{EC,r}; MAR_{EC}(X,Z)} = \frac{\sum_{i \in s_{EI,r} \cup s_{EC,r}} w_{MAR_{EC}(X,Z),i} y_i}{\sum_{i \in s_{EI,r} \cup s_{EC,r}} w_{MAR_{EC}(X,Z),i}} \quad \text{où}$$

$$w_{MAR_{EC}(X,Z),i} = \begin{cases} \dfrac{1}{\pi_{EI,i}} \; si \; i \in s_{EI,r} \\ \dfrac{1}{\pi_{EI,i} * \pi_{EC/s_{EI,nr},i} * \hat{\delta}_{MAR_{EC}(X,Z),i}} if \; i \in s_{EC,r} \end{cases}$$

Quelle que soit l'hypothèse considérée, l'estimation de la variance de l'estimation de ces prévalences a pris en compte le plan de sondage et la variabilité liée à la non-réponse (Ici, nous avons établi la démonstration spécifiquement pour ce point, car il n'en n'existait pas à notre connaissance dans la littérature sur les sondages. La démonstration se trouve en annexe). En pratique, la prévalence et la variance ont été estimées à l'aide d'un programme *ad hoc*.

IV.3.3 RÉSULTATS

IV.3.3.1 Taux de réponse à l'enquête complémentaire et aux enquêtes combinées

Le taux de réponse à l'enquête complémentaire était de 62,6% (77,0% pour le taux de contact et 81,3% pour le taux de réponse parmi les personnes contactées). Parmi les répondants, 57% ont été interrogés par téléphone et 43% en face-à-face.

Le taux de réponse aux enquêtes combinées (enquête initiale et enquête complémentaire) a été estimé à 70,4%.

IV.3.3.2 Propension à répondre

Les variables sociodémographiques expliquant la probabilité de réponse à l'enquête initiale ont été présentées en IV.2.3.3.1.

Dans le modèle multivarié incluant les variables sociodémographiques, la probabilité de réponse à l'enquête complémentaire est significativement plus élevée dans les groupes suivants : les non-salariés (à l'inverse de l'enquête initiale) (OR=2,5 ; IC 95% : [1,7 ; 3,7] pour les non-salariés vs les salariés), et pour les personnes habitant en Saône-et-Loire (OR=2,5 ; IC 95% : [1,3 ; 4,5] pour les personnes vivant en Saône-et-Loire vs les personnes vivant dans les Bouches-du-Rhône) (Tableau IV-7).

Dans les régressions logistiques univariées, la plupart des paradonnées étaient significativement associées à la non-réponse.

Tableau IV-7 : Variables sociodémographiques associées à la réponse à l'enquête complémentaire (hypothèse MAR(X))

		n	OR	IC 95%
Statut d'emploi	Salarié	281	1	
	Non salarié	219	2,5	(1,7 ; 3,7)
Département	Bouches-du-Rhône	100	1	
	Finistère	100	1,6	(0,9 ; 2,9)
	Pas-de-Calais	100	1,4	(0,8 ; 2,5)
	Pyrénées-Atlantiques	100	2,3	(1,3 ; 4,3)
	Saône-et-Loire	100	2,5	(1,3 ; 4,5)

Après avoir pris en compte les variables sociodémographiques et les paradonnées dans l'analyse multivariée (Tableau IV-8), la probabilité de réponse à l'enquête complémentaire était plus élevée dans les groupes suivants : les non-salariés (OR=2,4 ; IC 95% : [1,6 ; 3,6] pour les non-salariés vs les salariés), les personnes n'ayant pas eu de visite d'un enquêteur le

samedi (OR=2,2 ; IC 95% : [1,0 ; 5,2]) ou les personnes affectées au groupe « mode de collecte par téléphone » sans changement de mode de recueil par rapport aux autres personnes.

Tableau IV-8 : Variables sociodémographiques et paradonnées associées à la réponse à l'enquête complémentaire dans le modèle final (hypothèse MAR(X,Z))

		n	OR	IC 95%
Statut d'emploi	Salarié	281	1	
	Non salarié	219	2,4	(1,6 ; 3,6)
Visite d'enquêteur le samedi (enquête en face-à-face)	Oui	31	1	
	Non	469	2,2	(1,0 ; 5,2)
Evolution du mode de collecte	Assignation au groupe téléphone ; pas de changement de mode de recueil	199	1	
	Assignation au groupe téléphone ; changement de mode de recueil	137	0,2	(0,1 ; 0,3)
	Assignation au groupe face à face ; pas de changement de mode de recueil	139	0,2	(0,1 ; 0,4)
	Assignation au groupe face à face ; changement de mode de recueil	25	0,1	(0,0 ; 0,2)

IV.3.3.3 Contribution de l'enquête complémentaire et des paradonnées pour réduire le biais de non-réponse

La contribution est évaluée grâce aux variables issues des bases administratives et considérées comme variables d'intérêt et pour lesquelles on dispose d'un gold standard.

IV.3.3.3.1 Contribution de l'enquête complémentaire

Comme cela a déjà été vu au chapitre IV.2.3.6, pour ce qui concerne l'enquête initiale, les erreurs relatives diminuent après avoir corrigé la non-réponse en utilisant les variables sociodémographiques (Tableau IV-9 ; colonnes (a-f)/f et (b-f)/f). Les mêmes résultats sont observés pour les enquêtes combinées, excepté pour la variable « hospitalisation » (Tableau IV-9 ; colonnes (c-f)/f et (d-f)/f).

Quelle que soit la variable étudiée, la prévalence estimée via les enquêtes combinées est toujours plus proche de la prévalence gold standard que celle estimée via l'enquête initiale (Tableau IV-9 ; colonnes (b-f)/f et (d-f)/f). Les erreurs relatives restent élevées pour « au moins 100 remboursements d'actes pour soins de ville » (ER=22,3 pour l'enquête initiale; ER=14,5 pour les enquêtes combinées), modérées pour « hospitalisation » (ER=9,6 pour l'enquête initiale ; ER=-6,5 pour les enquêtes combinées), et l'absentéisme (ER=5,3 pour l'enquête initiale ; ER=-3,4 pour les enquêtes combinées) et petites pour les autres variables.

On peut noter par ailleurs que les intervalles de confiance pour les enquêtes combinées sont deux à trois fois plus larges que ceux estimés pour l'enquête initiale (Tableau IV-9).

IV.3.3.3.2 Contribution des paradonnées

Quelle que soit la variable d'intérêt considérée, les prévalences estimées à partir des enquêtes combinées sont très peu différentes, que l'on prenne en compte ou non les paradonnées pour corriger la non-réponse (Tableau IV-9 ; colonnes (d-f)/f et (e-f)/f).

Tableau IV-9 : Prévalences des variables de systèmes d'information existants estimées sur l'échantillon complet (gold standard), à partir des répondants de l'enquête initiale sous les hypothèses MCAR et MAR(X), et à partir des répondants à l'enquête initiale et à l'enquête complémentaire sous les hypothèses MCAR, MAR(X) et MAR(X,Z)

	Répondants à l'enquête initiale		**Répondants à l'enquête initiale et à l'enquête complémentaire**			**Echantillon complet (gold standard)**	**Erreur relative**				
	MCAR (a)	**MAR(X) (b)**	**MCAR (c)**	**MAR(X) (d)**	**MAR(X,Z) (e)**	**(f)**	**(a-f) /f**	**(b-f) /f**	**(c-f) /f**	**(d-f) /f**	**(e-f) /f**
	% IC 95%	**% IC 95%**	**% IC 95%**	**% IC 95%**	**% IC 95%**	**% IC 95%**					
Au moins 100 remboursements d'actes pour soins de ville	25,8 (23,9 ; 27,6)	24,3 (22,5 ; 26,2)	23,1 (16,5 ; 29,6)	22,8 (17,7 ; 27,9)	22,0 (17,1 ; 26,9)	19,9 (19,1 ; 20,7)	29,4	22,3	15,7	14,5	11,5
Hospitalisation	23,4 (21,6 ; 25,2)	22,9 (21,1 ; 24,7)	20,0 (13,9 ; 26,0)	19,5 (14,9 ; 24,2)	19,1 (14,5 ; 23,7)	20,9 (20,1 ; 21,7)	12,0	9,6	-4,3	-6,5	-8,6
Travail dans le secteur primaire	57,2 (55,1 ; 59,3)	58,8 (56,7 ; 61,0)	64,9 (54,9 ; 74,9)	61,8 (54,7 ; 68,9)	64,2 (56,2 ; 72,2)	61,1 (60,1 ; 62,1)	-6,3	-3,8	6,2	1,1	5,1
Durée d'emploi inférieure à 10 ans	53,4 (51,3 ; 55,4)	53,7 (51,5 ; 55,9)	50,8 (41,6 ; 60,0)	54,8 (46,7 ; 62,9)	55,7 (46,9 ; 64,5)	56,4 (55,4 ; 57,4)	-5,5	-4,8	-10,1	-2,9	-1,2
Absentéisme au travail	36,0 (34,0 ; 38,0)	35,8 (33,6 ; 37,9)	30,3 (23,1 ; 37,5)	32,8 (26,2 ; 39,4)	33,2 (26,2 ; 40,2)	33,4 (33,0 ; 34,8)	6,0	5,3	-10,7	-3,4	-0,6
Accident du travail ou maladie professionnelle	25,6 (23,8 ; 27,4)	26,4 (24,5 ; 28,4)	26,0 (18,9 ; 33,0)	26,4 (20,8 ; 32,0)	26,5 (20,9 ; 32,1)	26,9 (25,9 ; 27,9)	-4,7	-1,6	-3,2	-1,7	-1,5

X : variables sociodémographiques, Z : paradonnées

IV.3.3.4 Estimation des prévalences pour les variables du questionnaire

Nous nous intéressons maintenant à l'estimation de prévalences pour des variables recueillies questionnaire.

Après correction de la non-réponse avec les variables sociodémographiques, la prévalence de l'état de santé perçu comme très bon varie de 57,0% (IC 95% [54,9 ; 59,1]) pour l'enquête initiale à 60,7% (IC 95% [54,6 ; 66,8]) pour les enquêtes combinées (Tableau IV-10 ; colonnes MAR(X)). La proportion d'agriculteurs exploitants varie de 29,7% (IC 95% [27,6 ; 31,7]) pour l'enquête initiale à 34,1% (IC 95% [29,1 ; 39,1]) pour les enquêtes combinées (Tableau IV-11 ; colonnes MAR(X)). Pour les cadres et les professions libérales, la proportion varie de 7,1% (IC 95% [6,0 ; 8,3]) pour l'enquête initiale à 3,3% (IC 95% [1,6 ; 5,0]) pour les enquêtes combinées. Des différences importantes sont également observées pour les variables « niveau d'étude atteint inférieur au bac » et « exposition à des bruits intenses ». Ces différences sont réellement importantes car, après correction de la non-réponse par les variables sociodémographiques, les prévalences estimées pour les enquêtes combinées ne sont pas incluses dans les intervalles de confiance estimés pour l'enquête initiale.

Après correction de la non-réponse par les variables sociodémographiques et les paradonnées, la plupart des prévalences estimées sont inchangées (Tableau IV-10 et Tableau IV-11 ; colonnes MAR(X,Z)).

On peut noter que pour les deux variables « faible ressenti dans l'intensité des efforts physiques » et « consommation d'alcool supérieure ou égale à deux fois par semaine », les prévalences sont pratiquement les mêmes, quelles que soient l'enquête (enquête initiale ou enquêtes combinées) et l'hypothèse considérée sur le processus de non-réponse (i.e. MCAR, MAR conditionnellement aux les variables sociodémographiques, MAR conditionnellement aux variables sociodémographiques et aux paradonnées).

Tableau IV-10 : Prévalences des variables du questionnaire (variables de santé et sociodémographiques) estimées à partir des répondants de l'enquête initiale sous les hypothèses MCAR et MAR(X), et à partir des répondants à l'enquête initiale et à l'enquête complémentaire sous les hypothèses MCAR, MAR(X) et MAR(X,Z)

	Enquête initiale			**Enquête initiale et enquête complémentaire**			
	n	**MCAR**	**MAR(X)**	**n**	**MCAR**	**MAR(X)**	**MAR(X,Z)**
		% IC 95%	**% IC 95%**		**% IC 95%**	**% IC 95%**	**% IC 95%**
Etat de santé général perçu comme très bon	1307	55,8 (53,7; 57,9)	57,0 (54,9 ; 59,1)	1501	59,6 (51,6 ; 67,6)	60,7 (54,6 ; 66,8)	59,4 (52,5 ; 66,3)
Fumeur	518	23,7 (21,9 ; 25,5)	24,5 (22,6 ; 26,5)	606	27,6 (20,1 ; 35,2)	30,1 (23,7 ; 36,5)	29,1 (22,7 ; 35,5)
Consommation d'alcool ≥ deux fois/semaine	721	30,1 (28,2 ; 32,0)	30,4 (28,4 ; 32,4)	818	29,5 (22,4 ; 36,6)	27,8 (22,7 ; 32,8)	27,4 (22,6 ; 32,2)
Marié	1557	66,1 (64,1 ; 68,1)	64,0 (61,9 ; 66,1)	1766	66,7 (56,9 ; 76,6)	66,2 (58,6 ; 73,8)	66,9 (58,6 ; 75,2)
Niveau d'études < bac	1455	64,0 (61,9 ; 66,0)	63,7 (61,5 ; 65,8)	1693	71,7 (61,4 ; 82,0)	71,3 (63,4 ; 79,3)	73,3 (64,5 ; 82,1)

X : variables sociodémographiques, Z : paradonnées

Tableau IV-11 : Prévalences des variables du questionnaire (variables relatives à l'emploi) estimées à partir des répondants de l'enquête initiale sous les hypothèses MCAR et MAR(X), et à partir des répondants à l'enquête initiale et à l'enquête complémentaire sous les hypothèses MCAR, MAR(X) et MAR(X,Z)

	Enquête initiale			**Enquête initiale et enquête complémentaire**			
	n	**MCAR**	**MAR(X)**	**n**	**MCAR**	**MAR(X)**	**MAR(X,Z)**
		% IC 95%	**% IC 95%**		**% IC 95%**	**% IC 95%**	**% IC 95%**
Catégorie sociale							
Agriculteur exploitant	650	27,7 (25,8 ; 29,7)	29,7 (27,6 ; 31,7)	788	38,5 (30,0 ; 46,9)	34,1 (29,1 ; 39,1)	34,1 (29,1 ; 39,1)
Artisan, commerçant	43	1,8 (1,2 ; 2,3)	2,4 (1,6 ; 3,1)	47	2,4 (0,1 ; 4,8)	2,4 (0,1 ; 4,3)	2,5 (0,1 ; 4,4)
Cadre, profession libérale	155	7,7 (6,4 ; 8,8)	7,1 (6,0 ; 8,3)	163	3,7 (1,7 ; 5,7)	3,3 (1,6 ; 5,0)	2,8 (1,8 ; 3,8)
Profession intermédiaire	364	17,0 (15,4 ; 18,6)	16,6 (15,0 ; 18,3)	403	14,7 (9,2 ; 20,2)	16,5 (11,4 ; 21,6)	15,4 (10,6 ; 20,2)
Employé	310	14,0 (12,5 ; 15,5)	12,6 (11,2 ; 14,0)	341	11,2 (6,6 ; 15,8)	11,6 (7,7 ; 15,5)	10,5 (6,9 ; 13,9)
Ouvrier	675	31,8 (29,8 ; 33,8)	31,6 (29,5 ; 33,7)	762	29,6 (22,1 ; 37,0)	32,1 (25,6 ; 38,6)	34,8 (27,2 ; 42,4)
Secteur d'activité primaire	1362	63,1 (61,0 ; 65,2)	65,2 (63,0 ; 67,3)	1584	69,1 (58,6 ; 79,5)	66,8 (59,2 ; 74,4)	69,2 (60,5 ; 77,8)
Faible ressenti dans l'intensité des efforts physiques	669	34,2 (32,0 ; 36,4)	32,9 (30,7 ; 35,1)	744	29,6 (21,9 ; 37,4)	31,1 (24,4 ; 37,9)	29,2 (22,9 ; 35,6)
Respect que je mérite au travail	1144	59,4 (57,2 ; 61,7)	59,4 (57,0 ; 61,8)	1329	64,8 (53,7 ; 75,9)	66,4 (57,5 ; 75,3)	66,9 (57,2 ; 76,5)
Exposition à des bruits intenses	364	19,1 (17,3 ; 20,9)	19,8 (17,9 ; 21,7)	452	26,6 (18,7 ; 34,6)	31,1 (21,0 ; 32,8)	27,7 (21,3 ; 34,1)

X : variables sociodémographiques, Z : paradonnées

IV.3.4 DISCUSSION

Même si les variables sociodémographiques sont associées à la probabilité de réponse, et malgré le fait que de nombreuses études rapportent que de telles variables sont aussi associées à la santé et au travail (36, 42, 45), nos résultats suggèrent que ces données ne sont pas suffisantes pour corriger la non-réponse à l'enquête initiale. Cela est particulièrement vrai pour la variable « au moins 100 remboursements d'actes pour soins de ville ». Nos résultats montrent par ailleurs que l'enquête complémentaire auprès d'un échantillon de non-répondants est utile pour corriger le biais de non-réponse. Par ailleurs, même si les paradonnées sont fortement associées à la probabilité de réponse, elles sont moins utiles pour corriger la non-réponse.

Le taux de réponse à l'enquête complémentaire (62,6%) est trois fois plus élevé que celui observé à l'enquête initiale. Ce résultat était attendu, compte tenu du fait que le taux de réponse à l'enquête initiale était particulièrement faible (79, 123) ; il démontre l'efficacité du protocole de l'enquête complémentaire, qui a été construit de manière à maximiser le taux de réponse (mode de collecte plus incitatif, questionnaire réduit). Ainsi, le taux de réponse pour les enquêtes combinées est estimé à 70,4%, ce qui permet *a priori* d'augmenter la fiabilité des résultats si on les compare à ceux issus de l'enquête initiale seule.

Les erreurs de mesure peuvent être accentuées par l'utilisation de plusieurs modes de collecte de données (face-à-face, téléphone, postal), et par les nombreux efforts consentis pour contacter les personnes à l'enquête complémentaire. Cependant, comme nous avons étudié les biais de non-réponse pour des variables non collectées par questionnaire, mais issues de bases de données médico-administratives, notre étude est robuste car elle analyse exclusivement les erreurs de non-réponse. D'autres études ont déjà utilisé des données administratives en ce sens (74, 92).

Si on s'intéresse aux variables issues du questionnaire, les erreurs de mesure peuvent expliquer certaines différences observées entre les valeurs des prévalences estimées à l'enquête initiale ou aux enquêtes combinées (51, 93). La proportion d'exploitants agricoles estimée via les enquêtes combinées est proche de la proportion exacte de travailleurs non-salariés, qui est connue à partir des bases de données de la MSA (37%), et qui peut être considéré comme un proxy de la catégorie sociale « agriculteurs exploitants ». Ce résultat suggère que l'enquête complémentaire est utile pour corriger les biais de non-réponse. La proportion de « cadres et professions libérales » est probablement mieux estimée avec les enquêtes combinées. En effet, les agriculteurs exploitants étaient peut-être moins enclins à répondre à l'enquête initiale que les cadres et les professions libérales, probablement à cause de la longueur et de la complexité du questionnaire initial. Donc, il semble que la proportion de cadres et de professions libérales peut être surestimée à l'enquête initiale. De plus, les différences observées peuvent également provenir d'erreurs de mesure à l'enquête initiale ; en effet, la catégorie sociale a été codée en grande partie par le logiciel Sicore (18) qui nécessite en entrée un libellé de profession informatif et plusieurs variables annexes renseignées. Il est certain que la qualité des informations fournies en entrée était meilleure pour l'enquête complémentaire où les données étaient recueillies par enquêteur, que pour l'enquête initiale où les données étaient recueillies par auto-questionnaire papier.

L'utilité de l'enquête complémentaire doit également être discutée en fonction de son coût. Le coût de l'enquête complémentaire, réalisée auprès de 500 personnes, était aussi élevé que le coût de l'enquête initiale, réalisée auprès de 10 000 personnes, alors que le nombre de variables collectées à l'enquête complémentaire était nettement plus faible.

Même si le taux de réponse à l'enquête complémentaire est beaucoup plus élevé que celui de l'enquête initiale, cela ne signifie pas en soi que l'enquête complémentaire était utile pour

corriger la non-réponse ; en effet, les répondants à l'enquête complémentaire auraient pu avoir le même profil que les répondants à l'enquête initiale. Les résultats montrent que ces groupes de répondants partagent un certain nombre de caractéristiques sociodémographiques. Cependant, ils diffèrent probablement pour d'autres caractéristiques qui n'ont pas été mesurées. C'est la raison pour laquelle certaines différences persistent même après correction de la non-réponse sur les variables sociodémographiques.

La contribution de l'enquête complémentaire doit aussi être discutée en fonction de la largeur des intervalles de confiance estimés à partir des enquêtes combinées. Ils sont au moins deux fois plus larges que ceux estimés via l'enquête initiale seule. La plus grande part de perte de précision s'explique par le fait que les poids de sondage sont vraiment différents pour les répondants de l'enquête initiale et ceux de l'enquête complémentaire (ils varient de 7 à 220). Pour réduire cette différence entre poids, il aurait fallu tirer au sort un échantillon plus grand pour l'enquête complémentaire, mais les coûts de l'enquête auraient augmenté d'autant. Cependant, cette perte de précision, qui a été également trouvée dans une étude épidémiologique similaire (66), peut être considérée comme acceptable au regard du gain dans la fiabilité des estimations, étant donné que la préoccupation principale dans les études descriptives est le biais de non-réponse.

Comme nous l'avons mentionné ci-dessus, les paradonnées sont fortement associées à la non-réponse dans cette étude. Ce résultat peu surprenant a déjà été trouvé dans de nombreuses études (73, 94). Il est important de rappeler que la variable « visite d'un enquêteur le samedi pour une enquête en face-à-face » était incluse dans le modèle final expliquant la propension à répondre. Ce résultat est en accord avec la recommandation d'inclure des indicateurs de jour et d'heure d'enquête comme paradonnées (73). Il en est de même pour la variable « disponibilité du numéro de téléphone » qui est prédictive de la propension à répondre dans

les analyses univariées (8, 123). Cependant, pour être vraiment pertinentes, les paradonnées doivent aussi être des facteurs fortement corrélés aux variables d'intérêt. Si ce n'est pas le cas, elles ne corrigeront pas le biais de non-réponse et pourraient peut-être générer une inflation de la variance des prévalences estimées (104). Généralement, les paradonnées sont faiblement associées aux variables d'intérêt (94). Dans notre étude, la contribution des paradonnées pour corriger la prévalence de variables issues de bases de données de santé ou d'emploi s'est révélée faible, mais il faut noter que cette contribution a été étudiée après l'utilisation de variables sociodémographiques pertinentes pour corriger la non-réponse. Si les prévalences estimées à partir des enquêtes combinées avaient été corrigées par les paradonnées seules, leur contribution aurait été plus élevée pour certaines variables, comme par exemple pour les variables « au moins 100 remboursements d'actes pour soins de ville » ou « durée d'emploi inférieure à 10 ans ».

Plus généralement, l'utilisation des paradonnées relatives au processus de collecte de données dans le but de modéliser la participation à une enquête n'est pas facile, pour plusieurs raisons. On peut distinguer les problèmes liés à la construction du modèle de ceux liés à un possible effet enquêteur. De plus, les paradonnées telles que « nombre de tentatives d'appels téléphoniques » peuvent être considérées comme un processus continu pour lequel on pourrait supposer une relation linéaire entre le nombre de tentatives d'appels et la propension à répondre et donc, n'utiliser que les répondants avec un nombre maximal de tentatives d'appels pour représenter les non-répondants (95). Par ailleurs, dans notre cas, comme plusieurs modes de collecte de données ont été utilisés pour l'enquête complémentaire, il aurait été judicieux de construire un modèle prédictif de non-réponse qui différencie les enquêtes par téléphone et les enquêtes en face-à-face. Cependant, nous avons été contraints de construire un modèle plus simple pour des raisons de puissance statistique. De plus, une certaine réserve doit être prise sur la signification potentielle des paradonnées associées à la

collecte des données. Elles peuvent être enquêteur-dépendantes (par exemple, un enquêteur motivé peut avoir moins de refus qu'un enquêteur moins motivé) (19). Dans l'étude Coset-MSA, ce type de problème a été probablement limité parce que les enquêteurs ont été bien formés, supervisés et motivés.

L'utilisation des paradonnées pour corriger la non-réponse est relativement récente et de nombreuses questions sur leur exploitation subsistent (72). Même si elles ne semblent pas pertinentes dans cette étude, qui disposait de variables sociodémographiques pour corriger la non-réponse, les paradonnées sont cependant potentiellement des informations auxiliaires utiles pour corriger les biais de non-réponse (84, 94). Collecter des paradonnées est utile d'autant plus que le recueil de celles qui sont relatives au processus de collecte des données sont peu onéreuses.

Même si notre analyse montre l'intérêt de l'enquête complémentaire, l'hypothèse MAR conditionnellement aux variables sociodémographiques et aux paradonnées n'est pas vérifiée pour toutes les variables d'intérêt issues des bases médico-administratives. Par conséquent, un biais résiduel subsiste pour quelques variables et probablement pour certaines variables du questionnaire.

IV.4 APPORT D'UNE ENQUÊTE AUPRÈS DE NON-RÉPONDANTS EN PRÉSENCE D'INFORMATIONS AUXILIAIRES DE QUALITÉ

IV.4.1 CONTEXTE ET OBJECTIFS

Les deux précédents scénarios proposaient deux stratégies différentes pour minimiser les biais de non-réponse : l'utilisation d'informations auxiliaires de qualité ou le recours à une enquête complémentaire auprès de non-répondants. On peut maintenant s'interroger sur l'intérêt (c'est-à-dire la "valeur ajoutée") de mener une enquête complémentaire quand on dispose d'informations auxiliaires de qualité.

Cette partie correspond au scénario 3, donc à l'étude de l'apport d'une enquête réalisée auprès d'un échantillon de non-répondants lorsqu'on dispose par ailleurs d'informations auxiliaires pour minimiser les biais de non-réponse (les données du SNIIR-AM et de la MSA) dont le lien est direct avec la thématique de l'enquête.

Autrement dit, l'objectif de ce travail est de comparer les estimations obtenues après correction de la non-réponse grâce à des informations auxiliaires directement liée à la thématique de l'enquête dans une enquête en deux phases pour non réponse et les estimations obtenues dans la première phase de cette même enquête.

IV.4.2 POPULATION ET MÉTHODES

La méthodologie générale (cf. IV.1) a été suivie ; sont reprises par ailleurs certaines estimations calculées dans les deux précédentes applications.

La probabilité de réponse à l'enquête complémentaire a été étudiée selon les variables sociodémographiques et les variables du SNIIR-AM et de la MSA.

Grâce au modèle expliquant la non réponse établi à l'étape précédente, les probabilités de réponse $\hat{\delta}_{MAR_{EC}(X,V)}$ ont été estimées par la méthode des scores. A partir de ces groupes homogènes de réponse, des prévalences ont été estimées sous l'hypothèse, notée $MAR_{EC}(X,V)$ que pour l'enquête complémentaire, la non-réponse était aléatoire conditionnellement aux variables sociodémographiques, aux variables du SNIIR-AM et de la MSA.

Via les données de l'enquête initiale, les prévalences $\hat{p}_{s_{EI,r};MAR_{EI}(X,V)}$ ont été déjà été estimées et présentées (cf. IV.2.2.2).

Via les données des enquêtes combinées, les prévalences $\hat{p}_{s_{EI,r}\cup s_{EC,r};MAR_{EC}(X,V)}$ ont été estimées par :

$$\hat{p}_{s_{EI,r}\cup s_{EC,r};MAR_{EC}(X,V)} = \frac{\sum_{i\in s_{EI,r}\cup s_{EC,r}} w_{MAR_{EC}(X,V),i} y_i}{\sum_{i\in s_{EI,r}\cup s_{EC,r}} w_{MAR_{EC}(X,V),i}} \quad \text{où}$$

$$w_{MAR_{EC}(X,V),i} = \begin{cases} \dfrac{1}{\pi_{EI,i}} \text{ si } i \in s_{EI,r} \\ \dfrac{1}{\pi_{EI,i} * \pi_{EC/s_{EI,nr},i} * \hat{\delta}_{MAR_{EC}(X,V),i}} \text{ si } i \in s_{EC,r} \end{cases}$$

IV.4.3 RÉSULTATS

IV.4.3.1 Propension à répondre à l'enquête complémentaire

La probabilité de réponse à l'enquête complémentaire est, dans le modèle multivarié, associée à plusieurs variables sociodémographiques, à plusieurs variables relatives au recours aux soins, aux hospitalisations, aux accidents du travail et maladies professionnelles et au dernier emploi exercé. Elle n'est en revanche pas associée aux arrêts de travail (Tableau IV-12).

Concernant les variables sociodémographiques, la probabilité de réponse est associée au département et est plus élevée chez les 35 ans et plus en comparaison aux 18-34 ans (OR=1,7 IC 95% [1,0 ; 2,8] pour les 35-49 ans ; OR=1,6 IC 95% [0,9 ; 2,9] pour les 50-65 ans).

Pour les variables relatives au recours aux soins entre 2008 et 2010, la probabilité de réponse est plus élevée chez les personnes ayant eu des médicaments pour des problèmes du "système génito-urinaire" (OR=2,2 IC 95% [1,2 ; 3,9] versus aucun) et moins élevée chez les personnes ayant eu moins de 30 boites de médicaments pour des problèmes de "muscles et squelettes" (OR=0,6 IC 95% [0,4 ; 1,1] versus aucun). Pour ce qui concerne les hospitalisations entre 2008 et 2010, la probabilité de réponse est plus élevée chez les personnes ayant eu au moins une hospitalisation (OR=2,0 IC 95% [1,0 ; 3,9] versus aucune) et moins élevée chez les personnes hospitalisées pour une "maladie du système ostéo-articulaire" (OR=0,2 IC 95% [0,1 ; 0,6] versus aucune hospitalisation).

Pour les variables relatives aux accidents du travail et maladie professionnelle (ATMP), la probabilité de réponse est moins élevée chez les personnes ayant eu au moins un ATMP (OR=0,3 IC 95% [0,1 ; 0,8] versus aucun ATMP).

Enfin, concernant le dernier emploi exercé, la probabilité de réponse est moins élevée chez les personnes salariées (OR=0,4 IC 95% [0,3 ; 0,6] vs non-salariées).

Tableau IV-12 : Variables associées à la réponse à l'enquête complémentaire dans le modèle final

		n	OR	IC 95%
Variables sociodémographiques				
Caisse d'affiliation MSA	Bouches du Rhône	82	1	
	Pyrénées Atlantiques	90	1,1	(0,9 ; 1,3)
	Finistère	94	1,1	(1,0 ; 1,4)
	Pas de Calais	94	1,4	(1,2 ; 1,6)
	Saône et Loire	94	1,5	(1,3 ; 1,8)
Age	18 à 34 ans	110	1	
	35 à 49 ans	196	1,7	(1,0 ; 2,8)
	50 à 65 ans	148	1,6	(0,9 ; 2,9)
Variables relatives à des remboursements de soins de ville				
Nb total de boites facturées : système génito-urinaire	Aucun	363	1	
	Au moins 1	91	2,2	(1,2 ; 3,9)
Nb total de boites facturées : muscle et squelette	Aucun	138		
	Entre 1 et 29	157	0,6	(0,4 ; 1,1)
	Au moins 30	159	1,5	(0,8 ; 2,7)
Variables relatives aux hospitalisations				
Nombre d'hospitalisations	Aucune	329	1	
	Une	78	2,0	(1,0 ; 3,9)
	Au moins 2	47	1,0	(0,4 ; 2,4)
Hospitalisation pour maladie du système ostéo-articulaire	Non	431	1	
	Oui	23	0,2	(0,1 ; 0,6)
Variables relatives aux accidents du travail et maladies professionnelles (ATMP)				
Au moins un ATMP entre 2003 et 2008	Non	422	1	
	Oui	32	0,3	(0,1 ; 0,8)
Variables relatives au dernier emploi en date				
Statut du dernier emploi principal	Non salarié	211	1	
	Salarié	243	0,4	(0,3 ; 0,6)

IV.4.3.2 Estimation des prévalences des variables du questionnaire

Les résultats relatifs à la comparaison des estimations obtenues via l'enquête initiale et l'enquête en deux phases pour non-réponse sont présentés dans le Tableau IV-13, la Figure IV-2 et la Figure IV-3. Les différences entre les prévalences estimées via l'enquête initiale et

via l'enquête en deux phases pour non-réponse sont petites pour l'état de santé perçu comme très bon, la situation maritale « marié(e) ou pacsé(e) », la consommation d'alcool supérieure à deux fois par semaine, le secteur d'activité primaire et un faible ressenti des efforts physiques. Elle est modérée (écart compris entre 4 à 5%), pour le fait d'être fumeur, le niveau d'études, les catégories sociales et le respect mérité au travail. En revanche, les différences sont élevées pour l'exposition à des bruits intenses ; la prévalence estimée via l'enquête initiale est de 20,7% alors qu'elle est de 27,4% via l'enquête en deux phases pour non-réponse, soit un écart d'environ 7%.

IV.4.3.3 Estimation des prévalences des variables issues des bases médico-administratives existantes

Les résultats relatifs à la comparaison des estimations obtenues via l'enquête initiale et l'enquête en deux phases pour non-réponse sont présentés dans le Tableau IV-14 et dans la Figure IV-4. Quelle que soit la prévalence estimée, les erreurs relatives pour les deux estimations sont du même ordre de grandeur : la plupart sont inférieures à 5%, et les plus élevées sont autour de 10%.

Figure IV-2 : Prévalences des variables issues du questionnaire estimées à partir des répondants de l'enquête initiale et des répondants aux enquêtes combinées sous différentes hypothèses sur le processus de non-réponse (1)

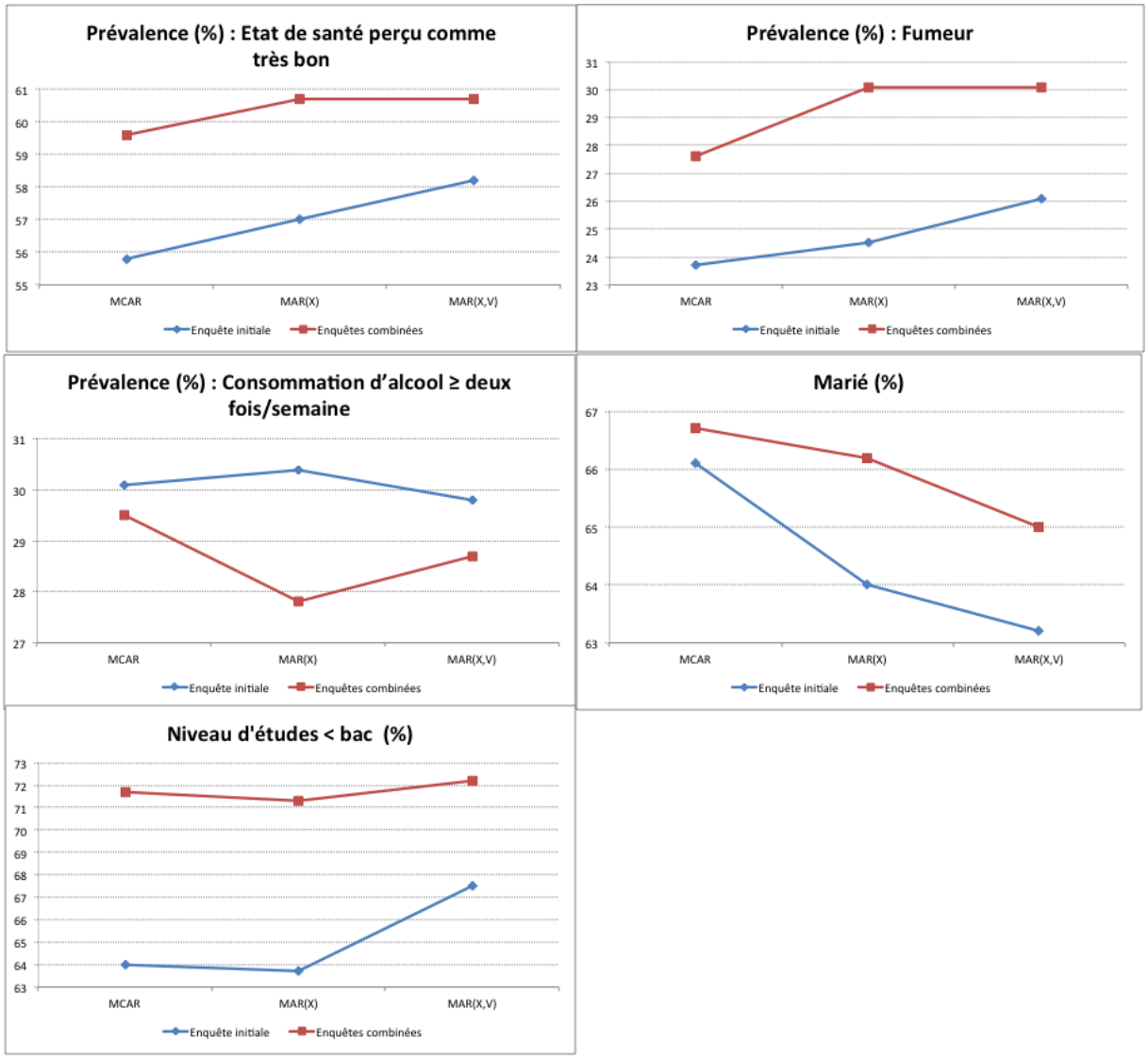


Figure IV-3 : Prévalences des variables issues du questionnaire estimées à partir des répondants de l'enquête initiale et des répondants aux enquêtes combinées sous différentes hypothèses sur le processus de non-réponse (2)

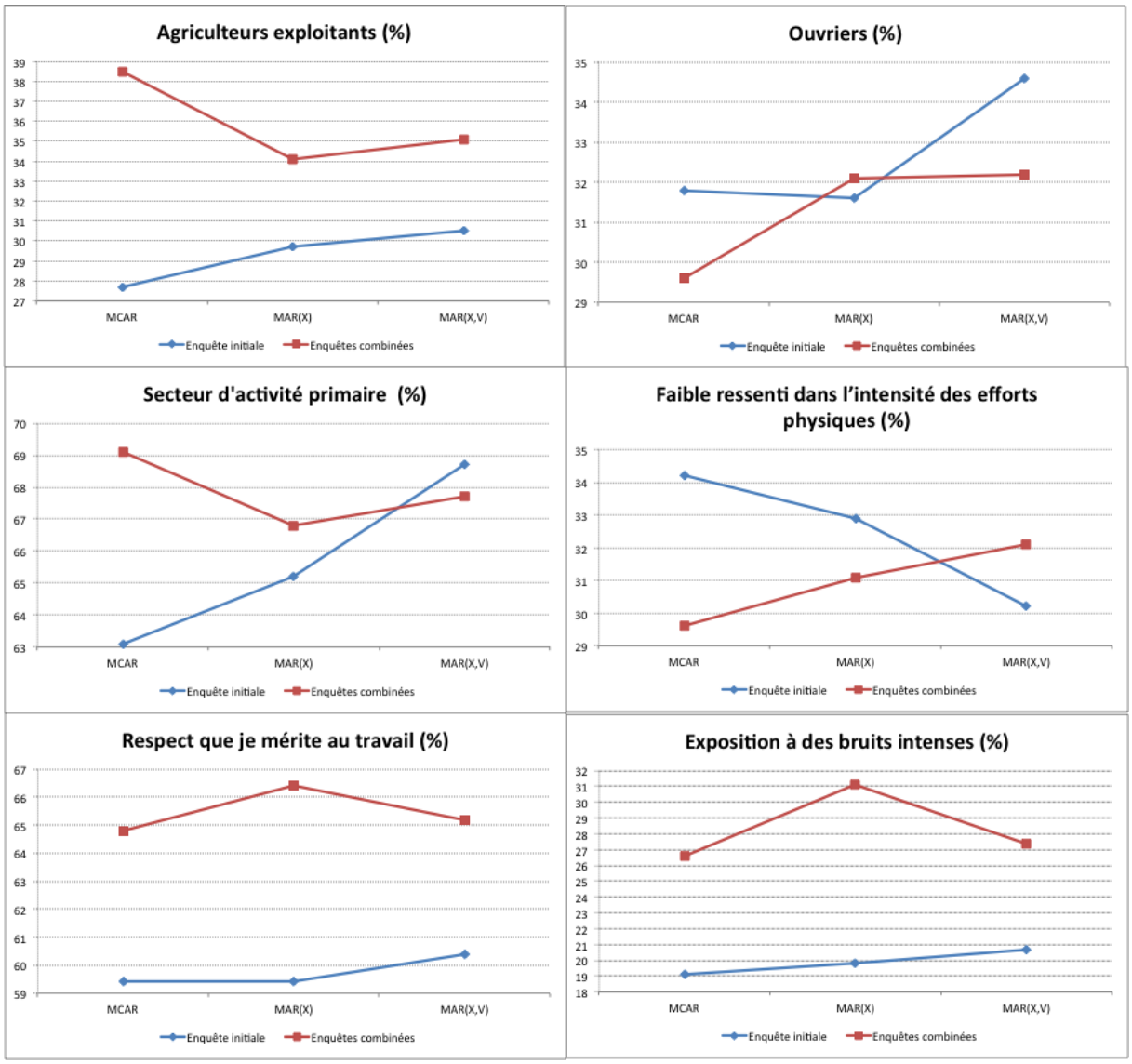


Figure IV-4 : Prévalences des variables issues des bases médico-administratives estimées sur l'échantillon complet et à partir des répondants de l'enquête initiale et des répondants aux enquêtes combinées pour non-réponse sous différentes hypothèses sur le processus de non-réponse

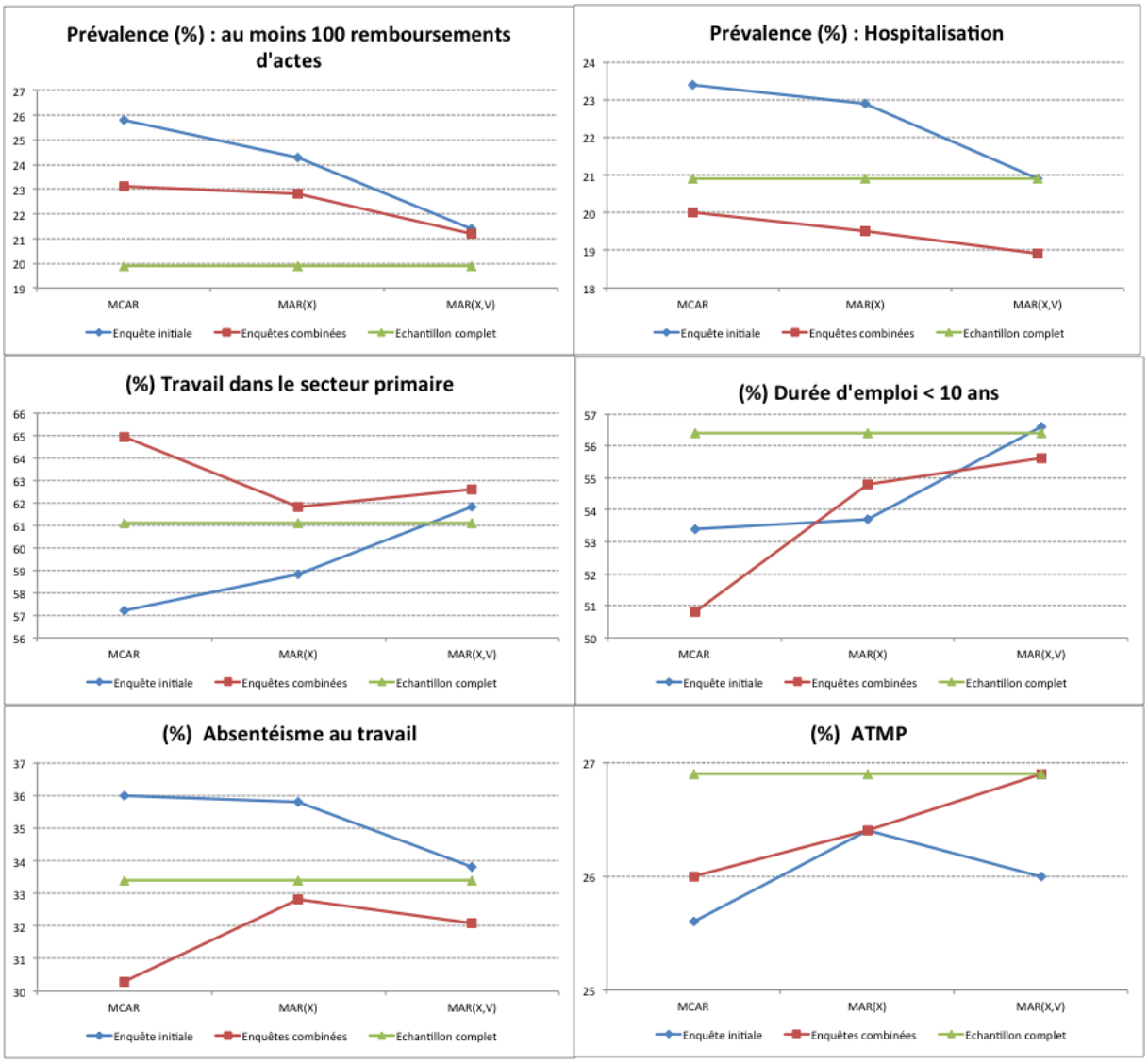


Tableau IV-13 : Prévalences des variables du questionnaire estimées à partir des répondants de l'enquête initiale sous l'hypothèse MAR(X,V), et à partir des répondants à l'enquête en deux phases pour non-réponse (enquêtes initiale et complémentaire combinées) sous l'hypothèse MAR(X,V)

	Enquête initiale		**Enquêtes combinées**	
	n	**% IC 95%**	**n**	**% IC 95%**
Etat de santé général perçu comme très bon	1307	58,2 (55,9 ; 60,5)	1501	60,7 (54,0 ; 67,4)
Fumeur	518	26,1 (24,0 ; 28,3)	606	30,1 (23,2 ; 37,1)
Consommation d'alcool ≥ deux fois/semaine	721	29,8 (27,7 ; 31,9)	818	28,7 (23,2 ; 34,2)
Marié	1557	63,2 (60,9 ; 65,5)	1766	65,0 (57,1 ; 73,0)
Niveau d'études < bac	1455	67,5 (65,3 ; 69,6)	1693	72,2 (66,2 ; 78,2)
Catégorie sociale				
Agriculteur exploitant	650	30,5 (28,3 ; 32,7)	788	35,1 (30,0 ; 40,3)
Artisan, commerçant	43	2,7 (1,8 ; 3,7)	47	2,7 (0,6 ; 4,8)
Cadre, profession libérale	155	6,1 (5,1 ; 7,2)	163	3,8 (1,4 ; 6,2)
Profession intermédiaire	364	14,4 (12,9 ; 16,0)	403	15,3 (10,5 ; 20,1)
Employé	310	11,6 (10,2 ; 13,1)	341	10,9 (7,1 ; 14,7)
Ouvrier	675	34,6 (32,2 ; 36,9)	762	32,2 (24,9 ; 39,5)
Secteur d'activité primaire	1362	68,7 (66,6 ; 70,9)	1584	67,7 (59,5 ; 76,0)
Faible ressenti dans l'intensité des efforts physiques	669	30,2 (27,9 ; 32,4)	744	32,1 (24,6 ; 39,6)
Respect que je mérite au travail	1144	60,4 (57,9 ; 62,9)	1329	65,2 (55,8 ; 74,6)
Exposition à des bruits intenses	364	20,7 (18,6 ; 22,8)	452	27,4 (20,8 ; 33,9)

X : variables sociodémographiques, V : variables du SNIIR-AM et de la MSA

Tableau IV-14 : Prévalences des variables de systèmes d'information existants estimées sur l'échantillon complet (gold standard), à partir des répondants de l'enquête initiale sous l'hypothèse MAR(X,V), et à partir des répondants à l'enquête en deux phases pour non-réponse (enquêtes initiale et complémentaire combinées) sous l'hypothèse MAR(X,V)

	Enquête initiale (a)		**Enquêtes combinées (b)**		**Echantillon complet (f)**		**Erreur relative**	
	n	**% IC 95%**	**n**	**% IC 95%**	**n**	**% IC 95%**	**(a-f) /f**	**(b-f) /f**
Au moins 100 remboursements d'actes pour soins de ville	585	21,4 (19,6 ; 23,1)	650	21,2 (16,6 ; 25,9)	1856	19,9 (19,1 ; 20,7)	7,5	6,5
Hospitalisation	534	20,9 (19,1 ; 22,7)	589	18,9 (14,4 ; 23,3)	1950	20,9 (20,1 ; 21,7)	0,0	-9,6
Travail dans le secteur primaire	1349	61,8 (59,6 ; 64,0)	1563	62,6 (55,0 ; 70,2)	5869	61,1 (60,1 ; 62,1)	1,1	2,5
Durée d'emploi inférieure à 10 ans	1193	56,6 (54,4 ; 58,9)	1344	55,6 (46,7 ; 64,5)	5137	56,4 (55,4 ; 57,4)	0,4	-1,4
Absentéisme au travail	803	33,8 (31,6 ; 35,9)	890	32,1 (25,3 ; 38,9)	3089	33,4 (33,0 ; 34,8)	1,2	-3,9
Accident du travail ou maladie professionnelle	600	26,0 (23,9 ; 28,0)	682	26,9 (20,9 ; 32,9)	2514	26,9 (25,9 ; 27,9)	-3,3	0,0

X : variables sociodémographiques, V : variables du SNIIR-AM et de la MSA

IV.4.4 DISCUSSION

Que ce soit pour l'enquête initiale ou l'enquête complémentaire, les variables utilisées pour corriger la non-réponse sont pertinentes puisqu'elles sont associées à la non-réponse et à la thématique de l'enquête (la santé et le travail). Après correction de la non-réponse à l'enquête initiale et à l'enquête en deux phases pour non-réponse, on observe des différences plus ou moins importantes pour les estimations des prévalences des variables issues du questionnaire. Elles sont particulièrement élevées pour l'exposition à des bruits intenses.

Comme nous ne disposons pas de gold standard pour les variables issues du questionnaire, les différences sont difficiles à interpréter. Elles peuvent provenir de deux sources différentes : d'un biais de non-réponse résiduel plus important pour l'enquête initiale ou d'un biais de mesure en particulier pour l'enquête en deux phases pour non-réponse : elles peuvent aussi s'expliquer par la volatilité des estimations, les intervalles de confiance via l'enquête en deux phases étant en moyenne trois fois plus larges que ceux estimés via l'enquête initiale.

Il est possible qu'un biais résiduel existe pour l'estimation de la prévalence de fumeur via les données de l'enquête initiale. On peut néanmoins noter que, que ce soit pour l'enquête initiale ou pour l'enquête en deux phases, les informations auxiliaires ont corrigé l'estimation de la proportion de fumeurs. En effet, pour l'enquête initiale, la proportion de fumeurs est estimée à 23,7% sans correction de la non-réponse et à 26,1% avec correction de la non-réponse ; pour l'enquête en deux phases, elle est estimée à 27,6% sans correction de la non-réponse et à 30,1% avec correction de la non-réponse. Il serait intéressant de comparer dans plusieurs années les hospitalisations ou la mortalité pour des pathologies liées à la consommation de tabac, comme par exemple les cancers des voies aérodigestives supérieures (56), chez les répondants et les non-répondants pour documenter les potentiels biais de non-réponse résiduels.

Concernant la catégorie sociale, les écarts observés pour les catégories « agriculteur exploitant » et « cadre et profession libérale » proviennent probablement d'un mélange d'erreur de non-réponse et de biais de mesure ; ce point a déjà été discuté en IV.3.4. Le biais de non-réponse pourrait venir du fait que les agriculteurs exploitants ont répondu préférentiellement à l'enquête complémentaire, où le recueil des données était par téléphone ou en face-à-face, alors que le biais de mesure pourrait provenir de la meilleure qualité des informations collectées à l'enquête complémentaire qu'à l'enquête initiale pour réaliser codage de la catégorie sociale par le logiciel Sicore (18).

Concernant les différences observées pour le niveau d'études, elles peuvent potentiellement provenir d'un biais résiduel de non-réponse ; en effet, il est possible que les personnes avec un niveau d'études inférieur au baccalauréat aient moins répondu à l'enquête initiale qu'à l'enquête complémentaire compte tenu de la longueur de l'autoquestionnaire de l'enquête initiale. Cette hypothèse est probablement valide pour expliquer les différences des prévalences estimées à l'enquête initiale et aux enquêtes combinées sous les hypothèses MAR(X) ; néanmoins, le fait que la proportion de personnes avec un niveau d'études inférieur au bac estimé via les données de l'enquête initiale sous l'hypothèse MAR(X,V) se rapproche fortement de celle estimée via les données des enquêtes combinées sous l'hypothèse MAR(X,V) laisse supposer que les données du SNIIR-AM et de la MSA corrigent efficacement la non-réponse. Il est donc fort possible que d'autres phénomènes (biais de non réponse résiduel, biais de mesure, variance) interviennent pour expliquer l'écart résiduel entre les deux proportions estimées.

Les écarts concernant la variable « respect que je mérite au travail » sont d'environ 5% quelle que soit l'hypothèse sur le processus de non-réponse, la proportion étant toujours supérieure pour les enquêtes combinées (65%) que pour l'enquête initiale seule (60%). Cette question,

issue d'une échelle mesurant les risques psychosociaux, est une question sensible. Il est possible que les personnes y aient répondu avec plus de sincérité via un autoquestionnaire postal que via un enquêteur. Le fait que les informations auxiliaires, qu'elles soient sociodémographiques, de santé ou relatives à l'emploi, ne modifient pas les proportions estimées peut signifier que ces dernières ne sont pas vraiment adaptées pour corriger les biais de non-réponse pour ce type de variables.

Pour l'estimation de la prévalence d'exposition à des bruits intenses, l'hypothèse de l'existence d'un biais de mesure pour expliquer la différence de 7% nous semble prédominante. Cependant, cette question semble peu sujette à une désirabilité sociale liée à un mode de recueil de données par enquêteur (29) ; si un biais de mesure existe, il ne provient donc vraisemblablement pas de l'enquête complémentaire. En revanche, il peut s'expliquer par la place de cette question dans le questionnaire de l'enquête initiale. En effet, la variable relative à l'exposition à des bruits intenses était recueillie à la 33ème page du questionnaire (sur 40) et était incluse dans une liste de questions relative à des historiques d'exposition professionnelle formatées de la même façon ; pour chaque exposition, il fallait dans un premier temps dire si on avait été exposé ou non, la première modalité proposée étant « non », et, en cas de réponse positive, donner dans un deuxième temps les années où cette exposition était survenue.

Des études antérieures ont montré qu'il existait un lien entre la difficulté à joindre les personnes et le biais de mesure ; les personnes les moins coopératives, ont tendance à répondre « ne sait pas » ou à présenter un taux de non-réponse partielle plus élevé (27) ou bien donnent des réponses moins fiables (16, 120) que les personnes plus coopératives. Dans Coset-MSA, le pourcentage de participants ayant répondu de façon incohérente ou n'ayant pas répondu du tout à la partie sur les historiques professionnels est plus élevé chez les

personnes ayant une probabilité de réponse estimée inférieure à 30% que chez les personnes avec une probabilité de réponse estimée supérieure à 30% ; il est donc possible que les personnes ayant répondu à la partie sur les expositions professionnelles avec un faible taux de réponse aient répondu plus facilement « non » aux questions sur les historiques d'exposition, d'une part parce que c'était la première modalité de réponse proposée, et d'autre part pour ne pas avoir à remplir la partie sur les années d'exposition. Il est également possible qu'un effet de halo soit survenu ; en effet, dans les échelles orientées dans le même sens, les répondants ont tendance à répondre de manière semblable (29). Compte tenu du nombre important d'expositions recueillies dans le questionnaire, la réponse à ces questions est en général négative, donc les répondants ont pu répondre négativement par erreur à cause de cet effet. Un autre effet relatif à l'ordre des questions et à la taille du questionnaire est l'effet de contamination (29) ; les questions précédentes peuvent avoir une influence sur les réponses aux questions suivantes. Il est possible que les répondants aient considéré que l'exposition à des bruits était une exposition négligeable par rapport aux autres expositions recensées telles que les horaires atypiques ou les expositions chimiques. En résumé, il est donc possible que les répondants à l'enquête initiale aient répondu à la fin du questionnaire avec moins d'attention. Ceci expliquerait que la proportion de personnes exposées à des bruits intenses via l'enquête initiale est inférieure à celle estimée via l'enquête en deux phases.

Concernant les variables issues des systèmes d'information, après correction de la non-réponse sur les variables de santé et relatives à l'emploi, les prévalences estimées via les données de l'enquête initiale se rapprochent de celles estimées via les données des enquêtes combinées et des prévalences gold-standard. Ce résultat était attendu puisque les variables d'intérêt faisaient partie des variables candidates pour calculer les facteurs correcteurs de la non réponse.

Pour la plupart des autres variables incluses dans les questionnaires de l'enquête initiale et de l'enquête complémentaire mais non présentées ici, la contribution de l'enquête complémentaire pour l'estimation d'une prévalence était soit faible soit modérée. La seule variable présentant une contribution élevée de l'enquête complémentaire était relative au contact avec le public ; compte tenu de la place de cette question dans le questionnaire de l'enquête initiale, les hypothèses avancées sur les potentiels biais de mesure pour la variable relative aux bruits intenses sont également plausibles pour la question relative au contact avec le public.

Pour ce qui est des variables incluses dans l'enquête initiale mais pas dans l'enquête complémentaire, on peut supposer que pour la plupart des variables relatives à la santé, les résultats seraient aussi fiables que ceux obtenus par une enquête complémentaire puisque l'état de santé général perçu, dont les estimations de prévalence sont semblables à l'enquête initiale et aux enquêtes complémentaires, est bien corrélé à l'état de santé réel de la personne (47, 130). Il en est de même pour les autres variables, mises à part peut-être les questions des historiques d'expositions compte tenu des hypothèses évoquées pour l'exposition à des bruits intenses. Néanmoins, dans une optique de cohorte prospective, il suffirait peut-être de ne pas changer l'ordre des questions dans les questionnaires de suivi pour éviter que des biais de mesure potentiels surviennent lorsque serait discutée l'évolution de prévalences.

IV.5 DISCUSSION GÉNÉRALE

Les prévalences estimées sous plusieurs hypothèses sur les processus de non-réponse sont reportées dans la Figure IV-2, la Figure IV-3 et la Figure IV-4.

La première étude a non seulement montré l'intérêt d'apparier des données de santé et professionnelles issues de systèmes d'information existants pour étudier les biais de non-réponse, mais a aussi montré comment elles pouvaient être utilisées pour prendre en compte ces biais de non-réponse pour estimer des prévalences. Ces résultats sont encourageants, même avec un taux de réponse de 25%. Ils renforcent la démarche qui consiste à rechercher des informations auxiliaires pertinentes pour corriger les biais de non-réponse en présence de taux de réponse faible. Dans notre étude, les informations auxiliaires relatives à la santé et au travail issues de systèmes d'informations existants corrigent efficacement la non-réponse, même si on ne peut pas exclure qu'un biais résiduel persiste.

La deuxième étude a, pour sa part, montré qu'avec un faible taux de réponse à une enquête initiale avec des objectifs descriptifs, il est utile de mener une enquête complémentaire parmi des non-répondants lorsque seulement quelques informations auxiliaires sont disponibles pour corriger la non-réponse. Lorsque les données sont collectées par enquêteur, il peut être avantageux de collecter des paradonnées car elles ne sont pas onéreuses et elles peuvent être utiles pour corriger la non-réponse quand il existe une corrélation entre les paradonnées et les variables d'intérêt. Néanmoins, des travaux sur l'utilisation des paradonnées doivent être poursuivis et développés.

La comparaison des résultats obtenus pour les deux premiers scénarios montre qu'à première vue il est plus intéressant de corriger la non-réponse avec des informations auxiliaires de qualité à l'enquête initiale plutôt que de combiner les données d'une enquête initiale et d'une enquête complémentaire ; en effet, dans la première étude, les erreurs relatives sont toutes

inférieures à 10%, ce qui n'est pas le cas dans la deuxième étude. Cependant, les bons résultats de la première étude étaient plus attendus puisque les données auxiliaires utilisées pour corriger la non-réponse et certaines variables d'intérêt utilisées provenaient des mêmes sources de données.

Quoiqu'il en soit, il n'est pas suffisant d'utiliser seulement les données de l'enquête initiale et les variables sociodémographiques pour estimer correctement des prévalences.

Enfin, la dernière étude sur l'intérêt de réaliser une enquête complémentaire auprès de non-répondants quand on dispose d'informations auxiliaires directement liées à la thématique de l'enquête a montré que les écarts de prévalences estimées à l'enquête initiale et l'enquête en deux phases étaient dans la plupart des cas faibles à modérés. Par ailleurs, même si les différences observées sur les variables issues de questionnaire peuvent être expliquées en partie par la potentielle existence de biais de non-réponse et de biais de mesure, elles sont aussi probablement liées à la volatilité des estimations de prévalence issues de l'enquête en deux phases, les intervalles de confiance via l'enquête en deux phases étaient en moyenne trois fois plus larges que ceux estimés via l'enquête initiale. On constate ainsi que l'enquête complémentaire n'a plus la même utilité quand on dispose d'informations auxiliaires aussi pertinentes que celles du SNIIR-AM et de la MSA. De plus, il faut noter que le coût marginal généré par l'inclusion d'un sujet supplémentaire dans l'enquête complémentaire est fixe, alors que le coût en termes d'investissement lié à la recherche des informations auxiliaires du SNIIR-AM et de la MSA ne dépend pas du nombre de sujets inclus.

Cette étude a néanmoins relevé la probable existence de potentiels biais de mesure différentiels à l'enquête initiale et à l'enquête complémentaire. Ce point sera étudié plus en détail dans la troisième partie de ce chapitre, qui sera consacré à l'analyse de la balance entre erreur de mesure et erreur de non-réponse.

CHAPITRE V. DIFFICULTÉ À JOINDRE LES PERSONNES, ERREUR DE NON-RÉPONSE ET ERREUR DE MESURE DANS LE PILOTE COSET-MSA

V.1 CONTEXTE ET OBJECTIFS

Cette étude reprend la problématique exposée dans le chapitre II.3. Nous considérons ici qu'une personne difficile à joindre est une personne pour laquelle un niveau important d'efforts dans la collecte doit être mis en place pour obtenir sa réponse. En résumé, l'hypothèse que l'inclusion de personnes difficiles à joindre diminue l'erreur de non-réponse est de plus en plus remise en question (49, 53). De plus, chercher à joindre autant que possible des personnes difficiles à joindre peut augmenter l'erreur de mesure car ces dernières peuvent répondre avec moins de précision, voire de manière sciemment erronée aux questions posées. On peut donc se demander si en termes d'erreur totale, il est vraiment utile de chercher à augmenter autant que possible le taux de réponse aux enquêtes si le gain en termes d'erreurs de non-réponse est discutable et s'il peut entrainer une augmentation des erreurs de mesure.

Les quelques enquêtes ayant discuté cette balance entre erreur de non-réponse et erreur de mesure selon la difficulté à joindre les personnes montrent des résultats différents. Par ailleurs, à notre connaissance, aucune de ces études n'a étudié cette balance après correction de la non-réponse.

L'objectif de ce travail est d'étudier la balance entre erreur de non-réponse et erreur de mesure selon la difficulté à joindre une personne dans Coset-MSA avant et après correction de la non-réponse. Nous avons considéré que la difficulté à joindre les personnes est faible pour les répondants à l'enquête initiale et élevée pour les répondants à l'enquête complémentaire.

V.2 POPULATION ET MÉTHODES

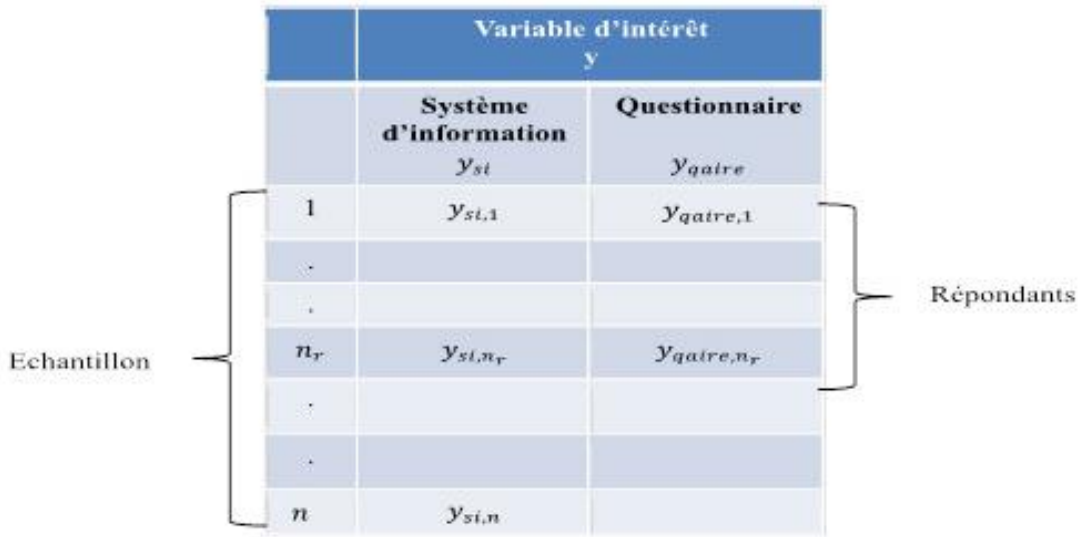
V.2.1 PRINCIPE GÉNÉRAL DE LA MÉTHODE

Cette méthode a été proposée par Olson (92). L'objectif est de quantifier les erreurs de mesure et les erreurs de non-réponse selon la difficulté à joindre les personnes. Dans la littérature, la difficulté à joindre les personnes a été définie de deux manières selon les auteurs : par l'estimation de la probabilité de réponse d'une personne (une personne avec une probabilité de réponse petite étant considérée comme difficile à joindre), ou par des informations relevées pendant la collecte des données (par exemple, un répondant après une relance ou après un changement de protocole d'enquête est considéré comme plus difficile à joindre).

Pour quantifier les erreurs de mesure et les erreurs de non-réponse selon la difficulté à joindre les personnes, deux variables mesurant la variable d'intérêt, notée y, sont nécessaires (Figure V-1) :

- une variable dont la mesure correspond à la vraie valeur de la variable d'intérêt, disponible pour l'ensemble de l'échantillon. Elle ne présente donc pas d'erreur de mesure pour y ; on l'appellera variable gold-standard, ou permettant d'obtenir des estimations gold-standard. En général elle provient de systèmes d'information existants. On la note y_{si}. On a donc $y = y_{si}$;
- une variable qui mesure la variable d'intérêt par questionnaire, et qui n'est donc disponible que pour les répondants. On la note y_{qaire}. C'est cette variable qui présente potentiellement des erreurs de mesure pour y.

Figure V-1 : Représentation d'un fichier de données nécessaire pour l'étude des erreurs de non-réponse et de mesure



Soit :

- $\hat{\bar{y}}_{s,si}$ la moyenne de y_{si} estimée à partir de l'échantillon tiré au sort *s* ; $\hat{\bar{y}}_{s,si}$ est une estimation sans biais de la moyenne de *Y* (cf. II.1.2.2.1.4) ;
- $\hat{\bar{y}}_{s_r,qaire}$ la moyenne de y_{qaire} estimée à partir de l'échantillon de répondants (s_r) ;
- $\hat{\bar{y}}_{s_r,si}$ la moyenne de y_{si} estimée à partir de l'échantillon de répondants (s_r).

L'erreur de non-réponse notée $\hat{E}_{NR}$, est estimée par $\hat{E}_{NR} = \hat{\bar{y}}_{s,si} - \hat{\bar{y}}_{s_r,si}$.

L'erreur de mesure, notée $\hat{E}_M$, est estimée par $\hat{E}_M = \hat{\bar{y}}_{s_r,si} - \hat{\bar{y}}_{s_r,qaire}$.

L'erreur totale, notée $\hat{E}_{tot}$, est estimée par $\hat{E}_{tot} = \hat{E}_{NR} + \hat{E}_M$.

Supposons que la difficulté à joindre les personnes soit constituée de k catégories, la première catégorie correspondant aux personnes les plus faciles à joindre, la dernière catégorie aux plus difficiles à joindre (l'ensemble des catégories correspond à l'échantillon complet de répondants).

Selon Olson, quantifier les erreurs de mesure et les erreurs de non-réponse selon la difficulté à joindre les personnes revient à estimer $\hat{E}_{NR}$ et $\hat{E}_{M}$ une première fois chez les personnes faciles à joindre (de la catégorie 1), puis de recalculer ces estimations en incorporant à la catégorie précédente les données de la catégorie suivante (de la catégorie 2) et de recommencer les calculs *k-1* fois jusqu'à ce que l'échantillon total des répondants soit intégré (jusqu'à l'intégration de la catégorie *k*).

V.2.2 APPLICATION À COSET-MSA

V.2.2.1 Données étudiées

Afin d'appliquer le principe général de la méthode aux données de Coset-MSA, il faut, d'une part disposer de variables gold standard et d'autre part de variables de questionnaire mesurant la même variable d'intérêt que les variables gold standard.

Les variables issues des systèmes d'information existants, qui sont disponibles chez les répondants et les non-répondants, et qui sont mesurées de façon indépendante des personnes tirées au sort sont de bonnes candidates pour être des variables gold standard.

V.2.2.1.1 Disposer de variables identiques dans les variables gold standard et dans les variables de questionnaire

Il n'existe aucune variable mesurée à la fois par questionnaire et par le SNIIR-AM.

On dispose en revanche de 4 variables mesurées à la fois par questionnaire et par la MSA :

- le statut d'emploi salarié ;
- le secteur d'activité primaire ;
- la surface agricole utile en ares pour les non-salariés (variable quantitative) ;
- le contrat de travail en CDI pour les salariés.

V.2.2.1.2 Disposer de variables gold standard

Même si les variables issues des systèmes d'information de la MSA sont disponibles pour 93,7% des répondants et des non-répondants, elles ne sont pas directement acceptables comme variables gold standard. En effet, d'une part, les systèmes d'information de la MSA n'incluent pas les emplois affiliés aux autres régimes d'assurance vieillesse (Cnav, RSI) et d'autre part il est difficile de faire le lien entre un emploi décrit dans les systèmes d'information de la MSA et un emploi décrit dans le questionnaire quand la personne a eu plusieurs emplois.

Pour ces raisons, l'étude a été restreinte aux personnes ayant eu un seul emploi affilié à la MSA en 2010 dans les bases de la MSA (n = 7 484), soit environ 80% de l'échantillon initial ; parmi elles, 1 896 ont participé à l'enquête initiale et 237 à l'enquête complémentaire

V.2.2.1.3 Indicateur de difficulté à joindre les personnes

L'indicateur considéré pour la difficulté à joindre les personnes comprend deux modalités : réponse à l'enquête initiale (facile à joindre) vs réponse à l'enquête complémentaire (difficile à joindre).

Cet indicateur correspond bien à un niveau d'efforts consentis pour obtenir la réponse d'une personne.

V.2.2.2 Analyses statistiques

L'analyse consiste à estimer les erreurs de non-réponse et de mesure ainsi que l'erreur totale via les données de l'enquête initiale et des enquêtes combinées (ou enquête en deux phases) avant et après correction de la non-réponse (sous l'hypothèse MAR(X,V)), puis à comparer ces estimations selon les différentes configurations.

Sans correction de la non-réponse, $\hat{\bar{y}}_{s_r,qaire}$ et $\hat{\bar{y}}_{s_r,si}$ ont été estimées par :

- Pour l'enquête initiale :

$$\hat{\bar{y}}_{s_{EI,r},MCAR_{EI}} = \frac{\sum_{i \in s_{EI,r}} w_{MCAR_{EI},i} y_i}{\sum_{i \in s_{EI,r}} w_{MCAR_{EI},i}} \text{ où } w_{MCAR_{EI},i} = \frac{1}{\pi_{EI,i}}$$

- Pour l'enquête en deux phases :

$$\hat{\bar{y}}_{s_{EI,r} \cup s_{EC,r};MCAR_{EC}} = \frac{\sum_{i \in s_{EI,r} \cup s_{EC,r}} w_{MCAR_{EC},i} y_i}{\sum_{i \in s_{EI,r} \cup s_{EC,r}} w_{MCAR_{EC},i}} \text{ où}$$

$$w_{MCAR_{EC},i} = \begin{cases} \frac{1}{\pi_{EI,i}} \text{ si } i \in s_{EI,r} \\ \frac{1}{\pi_{EI,i} * \pi_{EC/s_{EI,nr},i}} \text{ si } i \in s_{EC,r} \end{cases}$$

Avec correction de la non-réponse, $\hat{\bar{y}}_{s_r,qaire}$ et $\hat{\bar{y}}_{s_r,si}$ ont été estimées par :

- Pour l'enquête initiale :

$$\hat{\bar{y}}_{s_{EI,r};MAR_{EI}(X,V)} = \frac{\sum_{i \in s_{EI,r}} w_{MAR_{EI}(X,V),i} y_i}{\sum_{i \in s_{EI,r}} w_{MAR_{EI}(X,V),i}} \text{ où } w_{MAR_{EI}(X,V),i} = \frac{1}{\pi_{EI,i}} * \frac{1}{\hat{\delta}_{MAR_{EI}(X,V),i}}$$

- Pour l'enquête en deux phases :

$$\hat{\bar{y}}_{s_{EI,r} \cup s_{EC,r};MAR_{EC}(X,V)} = \frac{\sum_{i \in s_{EI,r} \cup s_{EC,r}} w_{MAR_{EC}(X,V),i} y_i}{\sum_{i \in s_{EI,r} \cup s_{EC,r}} w_{MAR_{EC}(X,V),i}} \text{ où}$$

$$w_{MAR_{EC}(X,V),i} = \begin{cases} \frac{1}{\pi_{EI,i}} si\ i \in s_{EI,r} \\ \frac{1}{\pi_{EI,i} * \pi_{EC/s_{EI,nr},i} * \hat{\delta}_{MAR_{EC}(X,V),i}} \text{ si } i \in s_{EC,r} \end{cases}$$

V.3 RÉSULTATS

Sans correction de la non-réponse (partie gauche de la Figure V-2), l'erreur de non-réponse est représentée par l'écart entre les courbes bleu clair (correspondant à la prévalence « gold standard » dans l'échantillon complet) et bleu foncé (correspondant à la prévalence « gold standard » chez les répondants) alors que l'erreur de mesure est représentée par l'écart entre les courbes bleu foncé et rouge (correspondant à la prévalence mesurée chez les répondants).

Avec correction de la non-réponse (partie droite de la Figure V-2), l'erreur de non-réponse est représentée par l'écart entre la courbe bleu clair et la courbe verte (correspondant à la prévalence « gold standard » chez les répondants) alors que l'erreur de mesure est représentée par l'écart entre la courbe verte et la courbe violette (correspondant à la prévalence mesurée chez les répondants).

- Secteur d'activité primaire

Avant correction de la non-réponse, l'erreur de non-réponse est de 7,6% pour l'enquête initiale et de 0,1% pour les enquêtes combinées alors que l'erreur de mesure est de 1% pour l'enquête initiale et de 0,4% pour l'enquête combinée. L'erreur de non-réponse est donc prépondérante par rapport à l'erreur de mesure pour l'enquête initiale.

Après correction de la non-réponse, l'erreur de non-réponse est nettement réduite pour l'enquête initiale puisqu'elle est estimée à 1,2% et elle reste stable pour les enquêtes combinées. L'erreur de mesure reste également stable pour les deux enquêtes. L'erreur totale est estimée à environ 1,8% pour l'enquête initiale versus 0,6% pour les enquêtes combinées.

- Statut salarié

Sans correction de la non-réponse, l'erreur de non-réponse est de 5,7% à l'enquête initiale et de 4,6% aux enquêtes combinées alors que l'erreur de mesure est de 7,2% à l'enquête initiale

et de 4,4% aux enquêtes combinées. Quelle que soit l'enquête, la part de l'erreur de non-réponse est équivalente à celle de l'erreur de mesure.

Avec correction de la non-réponse, l'erreur de non-réponse est nettement réduite à l'enquête initiale et aux enquêtes combinées puisqu'elle est estimée respectivement à 1,6% et 0,6%. L'erreur de mesure reste également stable dans les deux enquêtes. L'erreur totale est estimée à environ 8% à l'enquête initiale versus 4,5% aux enquêtes combinées.

- Surface agricole utile

Sans correction de la non-réponse, l'erreur de non-réponse est estimée à 514 ares à l'enquête initiale et à 435 ares aux enquêtes combinées alors que l'erreur de mesure est estimée à 1033 ares à l'enquête initiale et à 1143 ares aux enquêtes combinées. Quelle que soit l'enquête, la part de l'erreur de mesure est prépondérante par rapport à l'erreur de non-réponse.

Avec correction de la non-réponse, l'erreur de non-réponse est nettement réduite à l'enquête initiale et aux enquêtes combinées puisqu'elle est estimée respectivement à 285 et 346 ares. L'erreur de mesure reste également stable dans les deux enquêtes. L'erreur totale est estimée à environ 1500 ares quelle que soit l'enquête considérée.

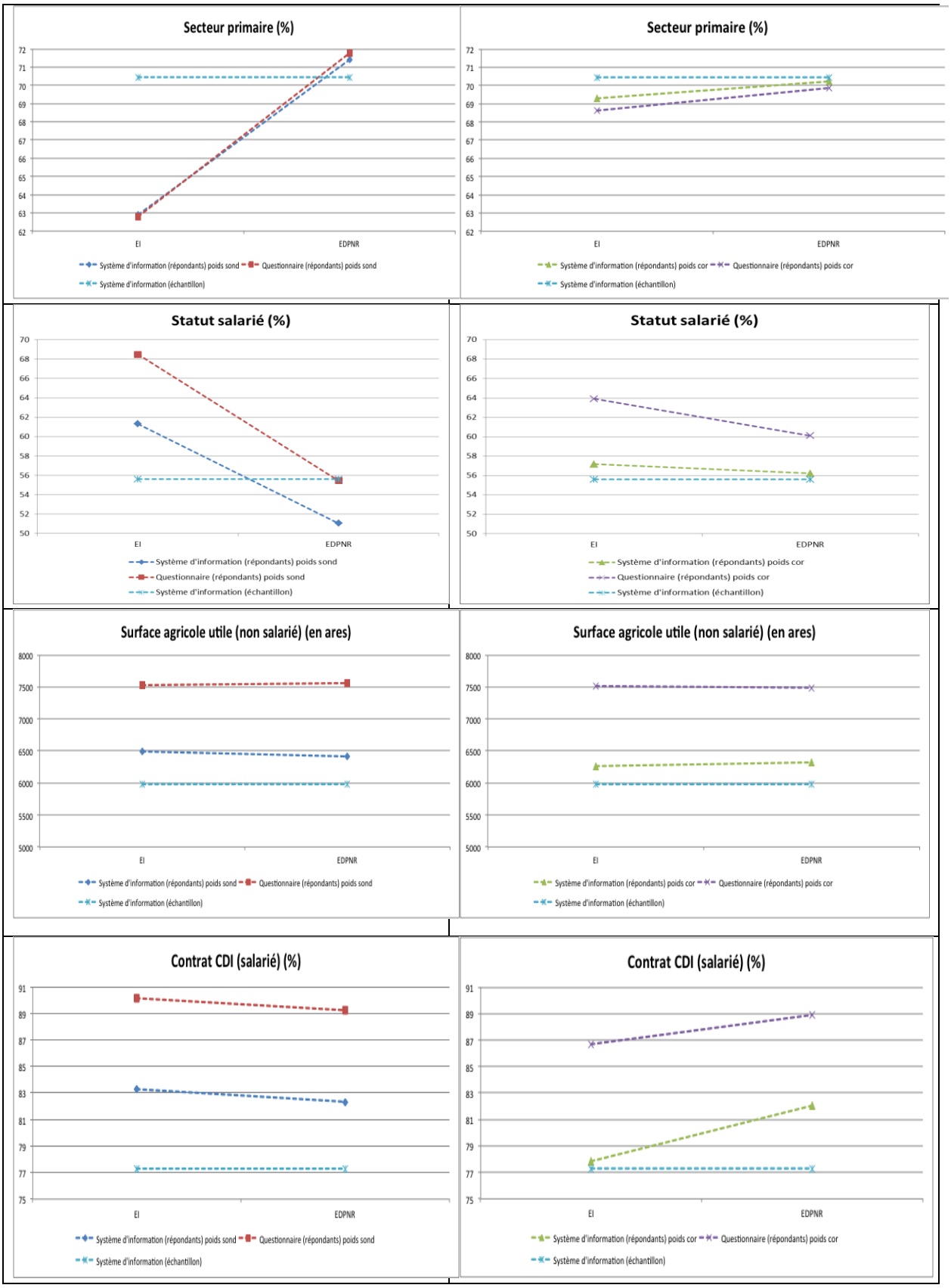
- Contrat de travail en CDI

Sans correction de la non-réponse, l'erreur de non-réponse est estimée à 6% à l'enquête initiale et à 5% aux enquêtes combinées alors que l'erreur de mesure est estimée à 6,9% à l'enquête initiale et aux enquêtes combinées. Quelle que soit l'enquête, la part de l'erreur de mesure est supérieure à celle de l'erreur de non-réponse.

Avec correction de la non-réponse, l'erreur de non-réponse est nettement réduite à l'enquête initiale puisqu'elle est estimée 0,5% et reste stable pour les enquêtes combinées. L'erreur de

mesure reste également stable pour les enquêtes combinées et augmente de 2% pour l'enquête initiale. L'erreur totale est estimée à environ 10% quelle que soit l'enquête considérée.

Figure V-2 : Moyenne ou prévalence à l'enquête initiale (EI) ou à l'enquête en deux phases (EDPNR) sans correction de la non-réponse (partie gauche) et avec correction de la non-réponse (partie droite)



V.4 DISCUSSION

V.4.1 DISCUSSION GÉNÉRALE

Avant correction de la non réponse, quelle que soit la variable considérée, l'erreur de non-réponse est soit équivalente, soit plus élevée pour l'enquête initiale que pour l'enquête en deux phases ; hormis pour le statut salarié, l'erreur de mesure est soit équivalente, soit légèrement supérieure pour l'enquête en deux phases comparativement à l'enquête initiale. L'erreur totale est plus élevée pour l'enquête initiale que pour l'enquête en deux phases, sauf pour la surface agricole utile où elle est équivalente dans les deux enquêtes.

Après correction de la non-réponse, l'erreur de non-réponse est nettement réduite, que ce soit à l'enquête initiale ou à l'enquête en deux phases. Ce résultat était attendu puisque les variables d'intérêt étaient candidates dans les modèles de non réponse.

Après correction de la non-réponse, les erreurs de mesure sont du même ordre de grandeur à l'enquête initiale et à l'enquête en deux phases. Elles sont à peu près équivalentes pour deux variables (secteur primaire et surface agricole utile), légèrement plus importante à l'enquête initiale pour le contrat CDI et légèrement moins importante à l'enquête initiale pour le statut salarié.

A priori, on aurait pu s'attendre à peu d'erreur de mesure pour les variables « secteur d'activité primaire », « statut salarié » et « emploi en CDI » pour les salariés, qui sont des variables factuelles, binaires et *a priori* faciles à renseigner. Cependant, il n'y a que pour le « secteur d'activité primaire » qu'on observe une faible erreur de mesure. Ce résultat à première vue étonnant pour les variables « statut salarié » et « emploi en CDI » montre que des questions qui nous semblent simples ne le sont pas nécessairement pour les personnes enquêtées. Néanmoins, avant de lancer l'étude pilote, des tests de questionnaire ont été réalisés en face-à-face par l'équipe Coset auprès de personnes travaillant en tant qu'affiliées à

la MSA et nous avons pu constater dans ce contexte que ce n'était pas toujours évident pour les personnes interrogées de situer leur position vis-à-vis de leur emploi (en tant que salarié, ou bien de leur temps de travail ou de leur type de contrat pour les personnes salariées).

Pour la variable « surface agricole utile », on observe une surestimation de la taille de l'exploitation par les répondants. On peut supposer que la question a mal été comprise et que les exploitants ont renseigné, non pas la surface agricole utile qui correspond à la surface agricole réellement exploitée, mais la surface agricole totale, qui correspond à la surface agricole existante, sans qu'elle soit forcément complètement exploitée.

Cette étude comporte certaines limites. En effet, elle est restreinte à l'étude de variables relatives à l'emploi. Il aurait bien entendu été intéressant d'étudier les variables relatives à la santé, mais ce travail a été envisagé une fois les données recueillies ; la correspondance entre les variables recueillies par questionnaire et issues des systèmes d'information a été réalisée *a posteriori* et aucune variable disponible dans le SNIIR-AM n'avait d'équivalent recueilli par questionnaire. C'est aussi pour cette raison que l'étude a porté sur peu de variables. Une autre limite vient du fait que l'enquête complémentaire a été construite pour étudier l'erreur de non-réponse sans que les données soient collectées de la même manière qu'à l'enquête initiale. Il est donc difficile de différencier les erreurs de mesure liées à la difficulté à joindre les personnes ou à un changement de mode de collecte des données ; pour ce faire, il aurait fallu intégrer un troisième groupe « enquête par questionnaire postal » à l'enquête complémentaire. Il nous semble néanmoins plus probable que les différences entre les erreurs de mesure nettement plus faibles à l'enquête complémentaire qu'à l'enquête initiale sont dues aux protocoles d'enquête différents, les enquêteurs pouvant expliciter une question mal comprise, ce qui n'était pas le cas pour les variables recueillies par questionnaire postal (malgré le numéro vert via lequel les personnes tirées au sort pouvaient contacter l'équipe Coset).

Cette étude présente tout de même certains atouts. Elle montre, comme dans les études d'Olson (92) et de Kreuter (74), un nouvel intérêt d'exploiter des variables issues des systèmes d'information, qui n'avait pas été anticipé initialement. Elles peuvent permettre d'étudier la qualité des informations collectées, en termes d'erreur de mesure et d'erreur de non-réponse. Si cette possibilité avait été anticipée initialement, le questionnaire aurait pu être construit en prenant en compte cet objectif ; il aurait en effet été possible d'ajouter, tout au long du questionnaire, des questions disponibles également dans les systèmes d'information pour suivre l'évolution de l'erreur de mesure et de l'erreur de non-réponse en fonction de la taille et de la complexité du questionnaire.

L'apport de notre étude par rapport aux études d'Olson et Kreuter, est que l'évolution des erreurs de mesure et de non-réponse en fonction de la difficulté à joindre les personnes a été estimée sans et avec correction de la non-réponse. Autant pour comprendre le processus, il est intéressant de comparer l'erreur totale de l'enquête initiale et des enquêtes combinées sans correction de la non-réponse autant pour les aspects appliqués (choisir de faire une enquête complémentaire ou non), il est plus judicieux de comparer l'erreur totale de l'enquête initiale et des enquêtes combinées avec correction de la non-réponse, car ce sont ces estimations qui sont présentées en pratique.

Par ailleurs, cette étude nous a amené à nous interroger sur la notion de difficulté à joindre les personnes et de propension à répondre. Certains auteurs considèrent que la propension à répondre et que les efforts consentis pour obtenir une réponse sont des notions proches (34, 126). Notre avons cherché à documenter ce point à partir des données de l'enquête initiale en prenant deux indicateurs de difficulté à joindre les personnes. Ce point, qui est marginal par rapport à l'ensemble de ce travail, est traité spécifiquement dans le paragraphe suivant.

V.4.2 FOCUS SUR LA DIFFICULTÉ À JOINDRE LES PERSONNES ET LA PROPENSION À RÉPONDRE

V.4.2.1 Méthodes

L'étude principale ayant montré que l'enquête initiale seule, après correction de la non-réponse, permettait d'estimer des moyennes ou des prévalences entachées d'erreur totale équivalentes à celles estimées par les enquêtes combinées, cette étude secondaire est restreinte à l'enquête initiale. Par ailleurs comme l'idée ici est de mieux comprendre ce que recouvre la notion de difficulté à joindre les personnes, l'étude est restreinte aux estimations sans correction de la non-réponse.

Deux indicateurs de difficulté à joindre les personnes ont été utilisés :

- réponse à l'enquête avant relance (facile à joindre) ou après relance (difficile à joindre) ;
- groupes homogènes de réponse qui correspondent à la propension à répondre construite en IV.2.3.4.2 (1 pour facile à joindre à 10 pour difficile à joindre).

V.4.2.2 Résultats

L'association entre les deux indicateurs de difficulté à joindre les personnes n'est pas significative (p = 0,69).

Pour l'enquête initiale, si on considère l'indicateur de la difficulté à joindre une personne par le fait d'avoir répondu avant ou après relance, on observe peu de différences en termes d'erreur de non-réponse ou d'erreur de mesure. La relance n'améliore donc pas la qualité des estimations (Figure V-3).

Si on examine les erreurs de non-réponse et les erreurs de mesure selon le deuxième indicateur (la propension à répondre à l'enquête initiale), on observe un résultat attendu en ce

qui concerne l'erreur de non réponse : quelle que soit la variable, l'erreur de non-réponse diminue dès qu'on inclut des personnes avec une propension à répondre plus petite (Figure V-4). L'erreur de mesure est petite et reste stable pour le secteur d'activité primaire après inclusion des personnes avec une propension à répondre plus faible. En revanche, pour les trois autres variables, l'erreur de mesure augmente dès qu'on inclut des personnes avec une propension à répondre plus faible.

V.4.2.3 Discussion

On peut considérer qu'une relance est efficace si elle contribue à diminuer le biais de non réponse sans augmenter d'autant le biais de mesure. Pour l'enquête initiale, sans correction de la non-réponse, la relance n'apporte ni une diminution de l'erreur de non-réponse ni une augmentation de l'erreur de mesure. La relance ne semble donc utile que pour diminuer la variance, puisqu'elle permet d'obtenir des répondants supplémentaires. L'étude des erreurs de mesure selon la propension à répondre montre que l'erreur de mesure augmente lorsqu'on intègre successivement les personnes avec une propension à répondre plus petite ; cela met en évidence le lien entre propension à répondre et erreur de mesure, qui correspond à la situation représentée par Groves dans la Figure II-8.

Ce travail complémentaire qui étudie deux définitions de la difficulté à joindre les personnes montre qu'on ne peut pas considérer qu'un indicateur relatif aux efforts consentis pour obtenir la réponse d'une personne et la propension à répondre recouvrent la même notion (d'ailleurs, il n'y a pas de lien statistique entre ce deux indicateurs). Il est donc important de différencier les deux.

Figure V-3 : Moyenne ou prévalence à l'enquête initiale (EI) sans correction de la non-réponse avant et après relance

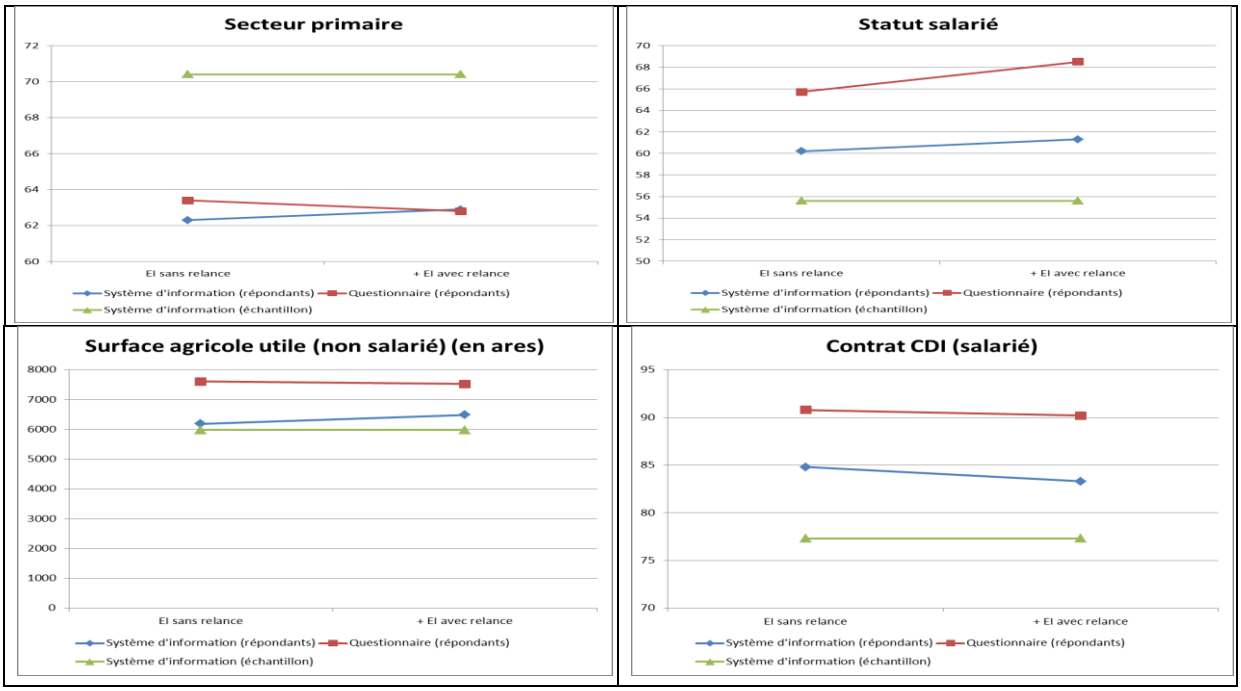
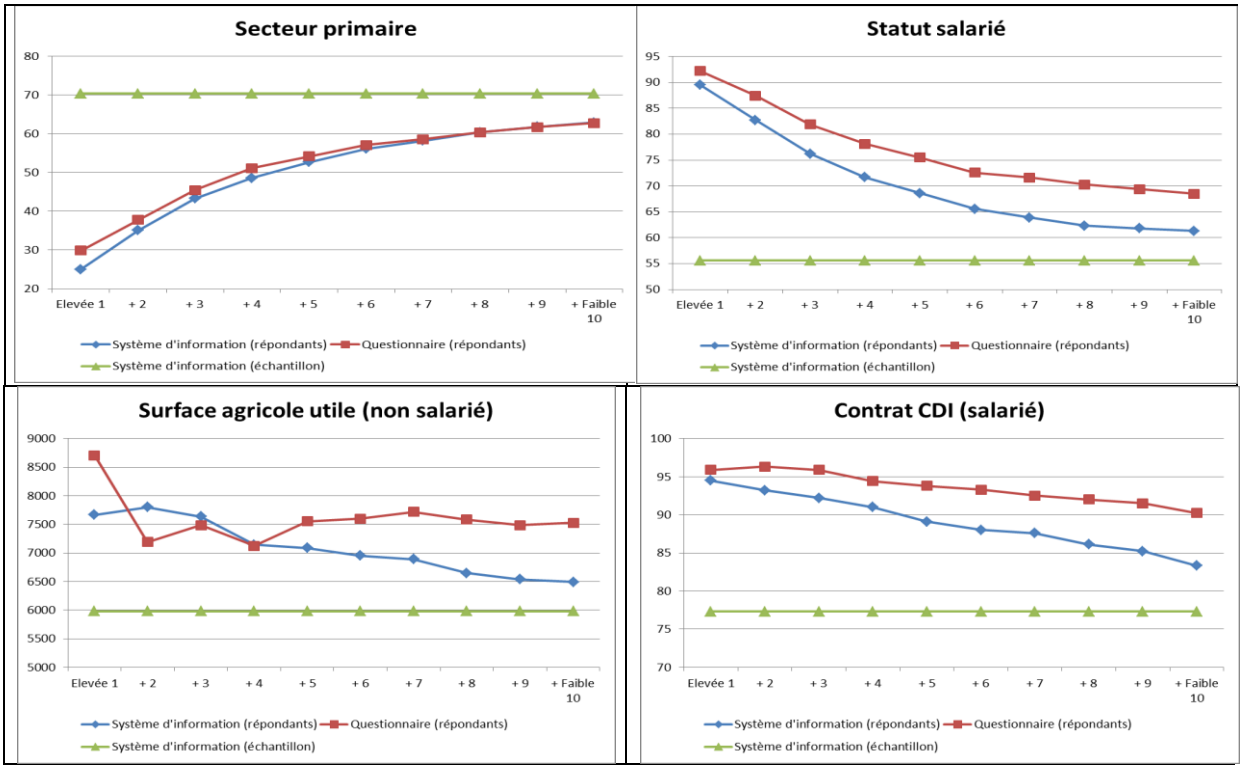


Figure V-4 : Moyenne ou prévalence à l'enquête initiale (EI) sans correction de la non-réponse selon les groupes homogènes de réponse



CHAPITRE VI. DISCUSSION GÉNÉRALE

VI.1 LEÇONS À TIRER POUR COSET-MSA

VI.1.1 SYNTHÈSE DES RÉSULTATS

Nos résultats suggèrent que, même si les variables sociodémographiques sont associées à la probabilité de réponse, elles ne sont pas suffisantes pour corriger la non-réponse à l'enquête initiale.

Deux scénarios ont montré leur intérêt pour minimiser les biais de non-réponse et sont donc des alternatives intéressantes de protocole d'enquête en population pour la surveillance épidémiologique. Dans le premier scénario, on utilise l'apport d'informations auxiliaires (données du SNIIR-AM et de la MSA) dont le lien était direct avec la thématique de l'enquête. Dans le deuxième scénario, on utilise l'apport d'une enquête réalisée auprès d'un échantillon de non-répondants, construite de manière à obtenir un taux de réponse maximal, ainsi que d'informations auxiliaires collectées sur l'ensemble de cet échantillon (les paradonnées) qui n'ont pas toujours un lien direct avec lien avec la thématique de l'enquête scénario.

Dans le troisième scénario, on montre que la correction des données de l'enquête initiale grâce à des informations auxiliaires directement liées à la thématique de l'enquête (la santé et le travail) est suffisante pour obtenir des estimations de prévalence aussi satisfaisantes que celles obtenues grâce à une enquête complémentaire. Dans ce cas, il n'est donc pas nécessaire de réaliser une enquête complémentaire pour tenter d'obtenir un taux de réponse maximal. Ce troisième scénario a néanmoins mis en avant l'existence de potentiels biais de mesure plus importants à l'enquête initiale qu'à l'enquête complémentaire, liés aux différents modes de collecte mais aussi à la longueur du questionnaire de l'enquête initiale.

L'étude spécifique de la balance entre erreur de non-réponse et erreur de mesure indique que, après correction de la non-réponse, l'erreur totale est équivalente dans l'enquête initiale et avec les enquêtes combinées ; il faut néanmoins préciser que les quatre variables étudiées dans ce dernier travail sont seulement des variables relatives à l'emploi. Par ailleurs, cette dernière étude a permis de montrer que la difficulté à joindre les personnes pendant la collecte et la probabilité de réponse sont deux notions différentes.

Trois types d'informations auxiliaires ont été utilisés dans ces analyses : des variables sociodémographiques, des variables relatives à la santé et au travail et des paradonnées. Toutes étaient associées à la probabilité de réponse aux enquêtes (initiale ou complémentaire). Les variables sociodémographiques permettaient de corriger partiellement les biais de non-réponse à l'enquête initiale ou à l'enquête en deux phases. Les variables relatives à la santé et au travail, issues des systèmes d'information du SNIIR-AM et de la MSA ont été très utiles pour corriger la non-réponse à l'enquête initiale. Leur apport a été moindre pour corriger la non-réponse à l'enquête en deux phases dès lors qu'on disposait des variables sociodémographiques pour l'ensemble des personnes. Les paradonnées, uniquement disponibles pour l'enquête complémentaire, paraissent peu intéressantes pour corriger la non-réponse lorsqu'on dispose déjà de variables sociodémographiques ; leur apport semble un peu plus intéressant dans le cas contraire.

VI.1.2 DISCUSSION SUR LA STRATÉGIE ADOPTÉE

Dans nos analyses, nous avons considéré que nous ne disposions comme informations auxiliaires de base que des variables sociodémographiques qui correspondaient aux variables de stratification issues du plan de sondage de l'enquête initiale. Dans la pratique, sauf pour un plan de sondage de même type, il est assez rare que ces informations soient disponibles individuellement pour tous les individus tirés au sort. Disposer de variables au niveau

individuel permet de modéliser la non-réponse de façon fine, mais on peut considérer qu'il est équivalent de réaliser un calage sur les totaux de ces variables, qui sont souvent disponibles au niveau de la population source. Or c'est cette dernière stratégie qui est en général adoptée pour les enquêtes en population générale (63, 65) ; nous considérons donc que ce scénario de référence (données d'enquête avec un protocole classique et variables sociodémographiques comme informations auxiliaires pour corriger la non-réponse) correspond à ce qui est rencontré dans la pratique.

Même si nous disposions des totaux croisés des variables de stratification sur la population source (cf. Tableau III-3), nous avons fait le choix ici de ne pas présenter dans le corps du manuscrit les résultats corrigés pour la non-réponse après calage bien que cela soit fréquemment réalisé en pratique. (2). Après avoir corrigé la non-réponse dans Coset-MSA, un calage a néanmoins été effectué dans l'optique de présenter les résultats définitifs de l'étude pilote (certains d'entre eux sont présentés en annexe). Dans le premier scénario (enquête initiale et données du SNIIR-AM et de la MSA pour corriger la non-réponse), le calage a permis de prendre en compte les 693 personnes sans informations auxiliaires issues du SNIIR-AM et de la MSA, parmi lesquelles 55 avaient répondu à l'enquête initiale ; on peut noter que les résultats sans ou avec calage différent peu. Ces résultats n'ont pas été intégrés dans le cœur de la thèse puisque le calage n'a pas été utilisé ici pour corriger la non-réponse.

L'intérêt de l'enquête complémentaire peut également être critiqué ; en effet, même si elle permet de réduire les biais de non-réponse, elle augmente la variance des estimateurs (9). Comme nous l'avons déjà discuté dans le deuxième scénario (enquête initiale et enquête complémentaire), les intervalles de confiance aux enquêtes combinées sont deux à trois fois plus larges que ceux issus de l'enquête initiale. Il aurait fallu tirer au sort et enquêter plus de personnes pour les réduire, mais les coûts de l'enquête auraient été encore plus élevés. Nous

pensons néanmoins que l'enquête complémentaire a été utile puisqu'elle a permis d'atteindre un taux de réponse très élevé et que notre stratégie initiale qui correspondait simplement à la mise en place de l'enquête initiale avec appariement d'informations auxiliaires issues du SNIIR-AM et des bases de données de la MSA, n'était pas suffisante du fait du faible taux de réponse envisagé avant la mise en place de l'étude (entre 20% et 30%). Une autre faiblesse de l'enquête complémentaire venait des différents modes de collecte des données qui pouvaient générer des biais de mesure (13) ; néanmoins, le fait de disposer de variables issues de systèmes d'information existants pouvant être considérées comme des variables d'intérêt a permis de pallier ce problème. Quoi qu'il en soit, cette enquête complémentaire a eu un intérêt tout particulier dans le cadre de la phase pilote : en effet, comme c'était la première fois que ce type de données passives étaient utilisées pour corriger la non-réponse, il fallait étudier du mieux que l'on pouvait leur capacité à compenser le faible taux de réponse attendu à l'enquête initiale.

L'utilisation des variables issues du SNIIR-AM et des systèmes d'information de la MSA à la fois comme informations auxiliaires pour corriger la non-réponse et comme variables d'intérêt servant à évaluer cette correction engendre probablement un certain optimisme. Mais, dans les scénarios où cela aurait pu poser problème (scénario 1 cf. chapitre IV.2 et scénario 3 cf. chapitre IV.4), l'accent a surtout été mis sur les prévalences des variables issues du questionnaire.

La dernière étude relative à la balance entre erreur de non-réponse et erreur de mesure a comme limite principale de n'avoir été réalisée que pour des variables relatives à l'emploi et aucune relative à la santé. Comme nous l'avons déjà évoqué plus haut, cette limite vient du fait que cette étude n'avait pas été anticipée initialement.

Néanmoins, malgré toutes ces limites, il est très rare qu'un protocole d'enquête soit construit dans le but d'étudier les problèmes liés à la non-réponse de façon aussi poussée et sous autant d'angles, ce qui est un réel atout. Cet effort a été fait car la cohorte Coset-MSA doit être généralisée sur toute la France et devra suivre les personnes pendant plusieurs années ; il était donc important d'étudier minutieusement les effets de sélection à l'inclusion dans la cohorte pilote afin d'adopter un protocole d'enquête efficient pour l'extension.

VI.1.3 RECOMMANDATIONS POUR L'EXTENSION NATIONALE

L'étude approfondie de la non-réponse lors de la phase pilote Coset-MSA permet d'émettre quelques recommandations pour le protocole mis en œuvre pour l'extension de Coset-MSA à la France entière.

La recommandation principale est qu'en termes d'erreur totale pour l'estimation de prévalences, il semble suffisant de ne réaliser que l'enquête initiale puisqu'on dispose des données du SNIIR-AM et de la MSA pour l'échantillon tiré au sort. Il ne parait pas très utile de reconduire une enquête complémentaire auprès d'un échantillon de non-répondants au regard du coût que cela engendre.

Les données du SNIIR-AM et de la MSA étant d'un intérêt majeur, un effort particulier doit être fait pour qu'elles puissent être appariées au maximum de personnes sélectionnées pour participer à l'enquête.

Un effort devra être réalisé pour que la proportion de plis non distribuables soit la plus faible possible ; cela réduirait de manière conséquente la proportion de personnes sans informations auxiliaires disponibles. En effet, pour la phase pilote, 6,9% de l'échantillon tiré au sort n'a pu être apparié les systèmes d'information du SNIIR-AM et de la MSA ; la raison prédominante pour laquelle cet appariement n'a pu être réalisé était que les plis étaient non distribuables

(75% des non-appariés) et il ne nous était pas possible de de recueillir ce type de données chez des personnes non informées. Ce fort pourcentage pourra certainement être diminué pour la généralisation : il était dû à un problème lié au routeur, qui n'a pas su faire le lien entre les plis non distribuables qu'il recevait et le fichier adresse que la MSA lui avait transmis.

Les données du SNIIR-AM et de la MSA paraissent très pertinentes pour minimiser les erreurs de non-réponse comme le montrent les résultats des études IV.2 et IV.4 ; c'est un résultat essentiel car ces données avaient été collectées initialement uniquement dans cette optique et que leur intérêt pour corriger la non-réponse dans une étude épidémiologique n'avait jamais été évalué jusqu'alors.

Des efforts devront être mis en œuvre pour minimiser les erreurs de mesure, qui semblent liées à la longueur et à la complexité du questionnaire. Une première recommandation est de pouvoir les estimer tout au long du questionnaire : pour cela, il sera pertinent de rajouter des questions relatives à la santé et au travail également disponibles dans les systèmes d'information existants. Par ailleurs, pour des contraintes budgétaires, le questionnaire de l'extension sera administré par internet (questionnaire CAWI pour « Computer Assisted Web Interview ») ; on peut espérer que ce mode de recueil de données permettra de réduire les erreurs de mesure, puisque, comme pour les questionnaires CATI ou CAPI, des filtres automatiques et des aides au remplissage du questionnaire (en particulier pour les emplois et les expositions professionnelles) pourront être mis en œuvre. Enfin, dans une optique de suivi annuel de la cohorte, l'ordre des questions ne devra pas être modifié pour éviter des erreurs de mesure liées à l'emplacement des questions dans le questionnaire.

VI.2 INTÉRÊT DES BASES MÉDICO ADMINISTRATIVES

Comme en témoigne la littérature récente (37, 46, 106, 125), l'utilisation des bases médico-administratives est en pleine expansion en épidémiologie. Nous illustrons ici des remarques générales retrouvées dans cette littérature par ce que nous avons pu observer dans Coset-MSA.

Dans Coset-MSA, comme dans les deux autres cohortes du programme Coset (Coset-RSI, Constances) dont les objectifs sont avant tout descriptifs, le fait que les bases du SNIIR-AM et de la MSA puissent être appariées à l'échantillon tiré au sort est un atout majeur pour ce programme. Comme nous l'avons vu tout au long de ce travail, elles permettent de corriger la non-réponse de manière satisfaisante. En pratique, nos résultats suggèrent qu'une enquête postale classique avec appariement sur des bases médico-administratives de qualité permet d'obtenir des estimations de prévalence suffisamment correctes, sans qu'il soit nécessaire de recourir à des protocoles d'enquêtes onéreux pour obtenir un meilleur taux de réponse. Par ailleurs, les bases du SNIIR-AM et de la MSA permettent d'estimer l'erreur totale comme la somme de l'erreur de non-réponse et l'erreur de mesure pour les variables mesurées par questionnaire lorsqu'une valeur gold-standard est disponible dans les systèmes d'information. Ce dernier point n'avait pas été anticipé au début de ce travail et c'est un réel point fort de l'utilisation des bases de données médico-administratives.

Un autre atout des trois cohortes composantes de Coset résulte de l'utilisation d'un numéro d'identification unique, le Numéro d'Identification au Répertoire (NIR) :il permet d'apparier plusieurs bases de données différentes (SNIIR-AM, MSA pour la cohorte Coset-MSA) avec un taux d'appariement pratiquement égal à 100% même si cela nécessite des circuits de données complexes afin de préserver l'anonymat des personnes mais cet effort porte vraiment ses fruits. Sans numéro unique sur lequel on peut apparier des données issues de sources

différentes, il faut passer par des techniques d'appariement probabiliste qui nécessitent de disposer d'un certain nombre de variables identiques dans les différentes sources, et qui fournissent rarement un taux d'appariement de 100% (106). Le fait de pouvoir apparier plusieurs sources de données de natures différentes est un réel avantage (37).

Dans une perspective de suivi longitudinal, le fait que le SNIIR-AM soit un système centralisé qui recueille les données de remboursement de soins de la plupart des régimes d'assurance maladie est un véritable atout (33), car les consommations de soins des personnes incluses dans la cohorte Coset-MSA pourront être suivies via le SNIIR-AM même en cas de changement de régime de protection sociale. Ce n'est en revanche pas le cas des données professionnelles issues des systèmes d'information de la MSA ; en effet, en cas de changement de régime, aucune information professionnelle ne remontera dans les systèmes d'information de la MSA. C'est pour cela que l'équipe Coset s'est rapprochée de la Caisse Nationale d'Assurance Vieillesse (Cnav) qui dispose des informations professionnelles de tous les salariés affiliés au Régime général, ce dernier couvrant plus de 80% de la population active en France. Ceci permettra de limiter les perdus de vue.

Cependant, il faut avoir conscience que ces données, recueillies à des fins non épidémiologiques, présentent certaines contraintes et limites. Tout d'abord, elles nécessitent un travail de consolidation et de nettoyage gigantesque avant de pouvoir être exploitées. Par ailleurs, ces bases ne disposent pas de toutes les variables nécessaires pour mener des études épidémiologiques de qualité sur certaines variables d'intérêt ; c'est particulièrement le cas pour les variables mesurant les comportements à risque pour la santé comme la consommation de tabac ou d'alcool (45).

On peut par ailleurs s'interroger sur le devenir à long terme des bases médico-administratives. Dans un contexte économique et politique où les lois évoluent, il est possible que les

modalités de remboursement ou de cotisations sociales soient également modifiées et que les bases s'appauvrissent en informations auxiliaires pertinentes.

CHAPITRE VII. PERSPECTIVES

VII.1 QUESTIONS MÉTHODOLOGIQUES POUR COSET-MSA

VII.1.1 ERREUR DE NON-RÉPONSE

La revue des méthodes pour prendre en compte la non-réponse totale à l'inclusion et la réflexion autour du sens d'un taux de réponse réalisées dans cette thèse, ont permis de bien avancer dans notre compréhension de l'erreur de non-réponse. Néanmoins, il reste encore des points à approfondir et à aborder sur ce sujet.

VII.1.1.1 Modélisation

Afin de prendre en compte la non-réponse, cette dernière a été modélisée de manière très simple par régression logistique. De nouvelles façons de modéliser la non-réponse pourraient être testées. En ce sens, l'algorithme du « superlearner » (100) pourrait être utilisé car il est particulièrement intéressant quand l'objectif est prédictif. A partir d'un jeu de données et d'un certain nombre de modèles candidats, cet algorithme détermine quel est, parmi les modèles candidats, le modèle le plus adapté pour modéliser la variable d'intérêt, en utilisant la validation croisée et une fonction de perte appropriée. Les modèles sont classés en fonction de leur performance et un nouvel algorithme hybride et original est construit sous la forme d'une combinaison pondérée des meilleurs modèles candidats (99).

Par ailleurs, les variables quantitatives ont été catégorisées, ce qui n'est pas très recommandé en pratique (55, 107). Pour la prise en compte des variables explicatives quantitatives, des outils tels que les polynômes fractionnaires ou les splines pourraient être utilisés.

VII.1.1.2 Non-réponse partielle

Notre travail a traité de la non-réponse totale à l'enquête. Néanmoins, de la non-réponse partielle a également été rencontrée puisqu'un répondant à l'enquête n'a pas nécessairement rempli complètement le questionnaire.

La non-réponse partielle devra donc être traitée. Il est en général conseillé de recourir à des méthodes d'imputation pour corriger la non-réponse partielle (50, 114). Des méthodes récentes de type « double robustesse », qui combine imputation et repondération (67), pourront être testées.

VII.1.1.3 Attrition

Comme dans toutes les cohortes, toutes les personnes répondantes à l'inclusion pourront être répondantes ou non au cours du suivi. Ce type de non-réponse, appelée attrition, sera également à prendre en compte. En effet, il est possible que la probabilité de réponse aux questionnaires de suivi soit liée à des événements (de santé ou d'exposition) antérieurs. Pour corriger cette attrition, les informations des bases administratives pourront été exploitées, mais également les données de questionnaire des vagues précédentes, ce qui représente une source d'information auxiliaire supplémentaire d'un grand intérêt.

VII.1.2 ERREUR DE MESURE

Ce travail, initialement axé sur l'erreur de non-réponse, a mis l'accent sur le fait qu'il était très difficile de l'étudier sans aborder l'erreur de mesure. Cette analyse n'avait pas été envisagée lors de l'élaboration des protocoles et des outils de recueil. Ainsi la perspective principale est probablement d'étudier de manière approfondie ce type d'erreur. Nous avons vu qu'en introduisant dans le questionnaire des variables dont on connait la valeur « gold standard » par les données des bases administratives, il était possible d'estimer l'erreur de

mesure pour ces variables. Ces estimations d'erreurs de mesure pour des variables « gold standard » pourraient être répercutées pour les estimations de prévalence de variables de questionnaire à condition de déterminer à quelles variables gold standard ces dernières sont corrélées. La littérature sur le sujet reste donc à explorer. Dans les enquêtes par enquêteur, où on fait l'hypothèse qu'il existe une erreur de mesure liée à l'enquêteur, on peut avoir recours à des modèles à effets aléatoires où un aléa supplémentaire est ajouté au niveau « enquêteur » (19). Un article récent de Keogh propose une boite à outils pour corriger les erreurs de mesure, en utilisant, par exemple, des approches de régression par calage ou par imputation multiple (68). Il semble donc qu'il soit possible de recourir au calage pour corriger les erreurs de mesure, sous l'hypothèse que les variables de calage sont corrélées aux variables présentant des erreurs de mesure, ou bien par imputation multiple, en supposant que les variables qui permettent de modéliser la variable d'intérêt présentant des erreurs de mesure ne présentent pas elles-mêmes d'erreurs de mesure.

VII.1.3 ESTIMATION DE LA VARIANCE

Les problèmes relatifs à l'estimation de la variance des estimations des prévalences ont été peu évoqués dans ce manuscrit. Les logiciels classiques permettent d'estimer correctement la variance pour des plans de sondage classiques mais ils ne permettent pas d'estimer la variance d'estimateurs issus d'un plan de sondage en deux phases, ni la variance générée par la non-réponse. C'est pour cela qu'un certain soin a été apporté dans ce travail afin d'obtenir des estimations de variance les plus correctes possibles.

Pour l'enquête initiale avec non-réponse, nous avons utilisé la macro Calker développée par l'Insee (15).

Pour l'enquête en deux phases pour non-réponse avec de la non-réponse, nous avons calculé la variance spécifique à ce plan de sondage puis nous avons écrit le programme

correspondant. Nous avons considéré que la taille de la population était connue. C'est effectivement le cas lorsqu'on infère des prévalences à toute la population d'intérêt. En revanche, ce n'est plus le cas pour les estimations par domaine. Il faudrait dans ce cas considérer que la taille du domaine est estimée, et estimer la variance par une linéarisation de Taylor (115). Ces améliorations seront programmées ultérieurement. De plus, nous avons en effet considéré que les probabilités de réponse étaient connues, alors qu'elles ont été estimées. Même si la littérature cependant que dans ce cas les variances sont surestimées (70) ; nous avons considéré que cette erreur est acceptable.

VII.2 L'AVENIR DES ENQUÊTES EN POPULATION POUR LA SURVEILLANCE ÉPIDÉMIOLOGIQUE

Dans un contexte économique difficile, l'accès aux bases médico-administratives présente de belles perspectives pour la surveillance épidémiologique. Par exemple, à partir du SNIIR-AM, il est possible de fournir des proxy de certaines pathologies ; il est par exemple possible de repérer des personnes atteintes de la maladie de Parkinson en combinant certaines consommations médicamenteuses sous un modèle statistique donné (88). En ce sens, le réseau REDSIAM (41) a pour objectif de recenser tous les algorithmes créés à partir des données du SNIIR-AM pour repérer certaines pathologies. Ainsi, les algorithmes consolidés pourront être utilisés pour estimer des prévalences sans avoir à interroger les personnes, donc sans rencontrer de non-réponse. Néanmoins il paraît optimiste de penser que ces algorithmes puissent recouvrir toutes les pathologies d'intérêt de la surveillance ; c'est pour cela que le recours aux enquêtes directement auprès des personnes auront toujours leur place dans le futur et, dans cette thèse, nous mettons en évidence tout l'intérêt d'apparier les échantillons tirés au sort à des bases médico-administratives pertinentes afin de pouvoir d'une part minimiser les biais de non-réponse et d'autre part évaluer les erreurs de mesure.

En ce sens, pour les enquêtes en santé publique en population générale, il faudrait par exemple faciliter l'accès aux bases de sondage qui permettent de récupérer les données du SNIIR-AM pour l'ensemble de l'échantillon tiré au sort. Apparier des données recueillies par questionnaire avec des données issues de systèmes d'information existants permettrait ainsi d'appréhender l'erreur totale en termes à la fois d'erreur de non-réponse et d'erreur de mesure, ce qui serait une réelle avancée dans l'analyse des données d'enquête de surveillance épidémiologique.

RÉFÉRENCES

1. Ardilly P. Estimation sur des domaines. In: Technip, ed. Les techniques de sondage. Paris: 1994:289-97.

2. Ardilly P. Les techniques de sondage. Paris: Technip, 1994.

3. Assogba GF, Couchoud C, Roudier C, et al. Prevalence, screening and treatment of chronic kidney disease in people with type 2 diabetes in France: The ENTRED surveys (2001 and 2007). Diabetes Metab 2012;38:558-66.

4. Basu D. An essay on the logical foundation of survey sampling. In: Godambe VP, Sprott DA, eds. Foundations of statistical inference. Toronto: Holt, Rinehart, Winston, 1971:203-33.

5. Beaumont JF, Haziza D, Bocci C. An adaptive data collection procedure for call prioritization. Journal of Official Statistics 2014, in press

6. Bénézet, L., Geoffroy-Perez, B., and Santin, G.Accès aux données des systèmes d'information existants dans le cadre de la phase pilote du projet Coset-MSA. Saint-Maurice: InVS,2010.

7. Bethlehem J.G. Weighting nonresponse adjustments based on auxiliary information. 2013.

8. Bethlehem JG, Cobben F, Schouten B. Handbook of nonresponse in household surveys. Hoboken: Wiley, 2011.

9. Biemer P. Total survey error - Design, implementation, and evaluation. Public Opin Q 2010;74:817-48.

10. Bousquet, V. and Caserio-Schönemann, C.La surveillance des urgences par le réseau OSCOUR® (Organisation de la surveillance coordonnée des urgences) . Saint-Maurice: InVS,2012.

11. Bouvier, G.L'enquête handicap santé - Présentation générale. Paris: Insee,2011.

12. Bouyer J. Méthodes statistiques. Médecine - Biologie. Paris: ESTEM, 2000.

13. Bowling A. Mode of questionnaire administration can have serious effects on data quality. J Public Health (Oxf) 2005;27:281-91.

14. Brick M. Unit nonresponse and weighting adjustments: a critical review. Journal of Official Statistics 2013;29:329-53.

15. Brion, Ph. and Gros, E.Macro Calker. Paris: Insee,2013.

16. Cannell, C. and Fowler, F.A Study of the Reporting Visits to Doctors in the National Health Survey. Ann Arbor: MI: Survey research centre,1963.

17. Caron, N. and Rousseau, S.Correction de la non-réponse et calage de l'enquête santé 2002. Paris: Insee,2005.

18. Carton, M., Santin, G., Geoffroy-Perez, B., and Chanut, A.Contribution des variables annexes au codage des libellés de profession par le logiciel Sicore. Paris: InVS; Inserm; Insee,2007.

19. Casas-Cordero C, Kreuter F, Wang Y, et al. Assessing the measurement error properties of interviewer observations of neighbourhood characteristics. J R Statist Soc A 2013;176:227-49.

20. Centers for Disease Control and Prevention. National Health And Nutrition Survey. http://www.cdc.gov/nchs/nhanes.htm. 2014.

21. Centers for Disease Control and Prevention. National Health Interview Survey. http://www.cdc.gov/nchs/nhis.htm. 2014.

22. Chan Chee, C. and Jezewski-Serra, D.Hospitalisations et recours aux urgences pour tentative de suicide en France métropolitaine à partir du PMSI-MCO 2004-2011 et d'Oscour® 2007-2011. Saint-Maurice: InVS,2014.

23. Chauvet G, Haziza D. Échantillonnage et repondération par calage dans les enquêtes. Presented at the Ateliers statistiques de la Société Française de Statistique, 2011.

24. Chevalier, A., Ducamp, S., Gilg Soit Ilg, A., Goldberg, M., Goldberg, S., Houot, M., Imbernon, E., Marchand, J. L., Rolland, P., and Santin, G.Des indicateurs en santé travail - Risques professionnels dus à l'amiante. Saint-Maurice: InVS,2010.

25. Cohidon C, Santin G, Geoffroy-Perez B, et al. Suicide and occupation in France. Rev Epidemiol Sante Publ 2010;58:139-50.

26. Cohidon C, Santin G, Imbernon E, et al. Working conditions and depressive symptoms in the 2003 decennial health survey: the role of the occupational category. Soc Psychiatry Psychiatr Epidemiol 2009.

27. Dahlhamer JM. The Intersection of Response Propensity and Data Quality in the National Health Interview Survey (NHIS). Presented at the JSM - Section on survey research methods, 2012.

28. Deville JC, Sarndal CE. Calibration estimators in survey sampling. JASA 1992;87:376-82.

29. Dussaix AM. La qualité dans les enquêtes. MODULAD 2009;39:137-71.

30. Eltinge JL, Yansaneh IS. Diagnostics for formation of nonresponse adjustment cells, with an application to income nonresponse in the U.S. consumer expenditure survey. Survey Methodol 1997;23:33-40.

31. Etter JF, Perneger TV. Analysis of non-response bias in a mailed health survey. J Clin Epidemiol 1997;50:1123-8.

32. Falissard B. Mesurer la subjectivité en santé - Perspective méthodologique et statistique. Paris: MASSON, 2008.

33. Fender P, Weill A. Epidemiology, public health and medical rates databases. Rev Epidemiol Sante Publique 2004;52:113-7.

34. Fricker S, Tourangeau R. Examining the relationship between nonresponse propensity and data quality in two national household surveys. Public Opin Q 2010;74:934-55.

35. Fuhrer R, Rouillon F. La version française de l'échelle CES-D (Center for Epidemiologic Studies-Depression Scale). Description et traduction de l'échelle d'autoévaluation. Psychiatrie et psychobiologie 1989;4:163-6.

36. Galea S, Tracy M. Participation Rates in Epidemiologic Studies. Ann Epidemiol 2007;17:643-53.

37. Gavrielov-Yusim N, Friger M. Use of administrative medical databases in population-based research. J Epidemiol Community Health 2014;68:283-7.

38. Geneletti S, Mason A, Best N. Adjusting for selection effects in epidemiologic studies: Why sensitivity analysis is the only solution. Epidemiology 2011;22:36-9.

39. Geoffroy-Perez B, Chatelot J, SG, et al. Coset : un nouvel outil généraliste pour la surveillance épidémiologique des risques professionnels. Bull Epidemiol Hebd 2012;22-23:276-7.

40. Goga C, Cardot H, Shehzad MA. Régression sur composantes principales à l'aide des données d'enquête. Presented at the Symposium international de Statistique Canada sur les questions de méthodologie, 2013.

41. Goldberg M. Le réseau REDSIAM. Presented at the Adelf EMOIS, 2014.

42. Goldberg M, Chastang JF, Leclerc A, et al. Socioeconomic, demographic, occupational, and health factors associated with participation in a long-term epidemiologic survey: a prospective study of the French GAZEL cohort and its target population. Am J Epidemiol 2001;154:373-84.

43. Goldberg M, Chastang JF, Zins M, et al. Health problems were the strongest predictors of attrition during follow-up of the GAZEL cohort. J Clin Epidemiol 2006;59:1213-21.

44. Goldberg M, Imbernon E. Quels dispositifs épidémiologiques d'oservation de la santé en relation avec le travail ? Le rôle de l'Institut de veille sanitaire. RFAS 2008;2-3:21-44.

45. Goldberg M, Luce D. Selection effects in epidemiological cohorts: Nature, causes and consequences. Rev Epidemiol Sante Publ 2001;49:477-92.

46. Goldberg M, Quantin C, Gueguen A, et al. Bases de données médico-administratives et épidémiologie : intérêts et limites. Courrier des statistiques 2008;124:59-70.

47. Goldberg P, Gueguen A, Schmaus A, et al. Longitudinal study of associations between perceived health status and self reported diseases in the French Gazel cohort. J Epidemiol Community Health 2001;55:233-8.

48. Groves RM. Nonresponse rates and nonresponse bias in household surveys. Public Opin Q 2006;70:646-75.

49. Groves RM. Nonresponse rates and nonresponse bias in household surveys. Public Opin Q 2006;70:646-75.

50. Groves RM, Dillman DA, Eltinge JL, et al. Survey nonresponse. New York: Wiley, 2002.

51. Groves RM, Heeringa SG. Responsive design for household surveys: tools for actively controlling survey errors ans costs. J R Statist Soc A 2006;169:457.

52. Groves RM, Lyberg L. Total survey error: past, present, future. Public Opin Q 2010;74:849-79.

53. Groves RM, Peytcheva E. The impact of nonresponse rates on nonresponse bias: A meta-analysis. Public Opin Q 2008;72:167-89.

54. Hansen MH, Hurwitz WN. The problem of nonresponse in sample surveys. JASA 1946;41:517-29.

55. Harrell FE. Regression modeling strategies with applications to linear models, logistic regression, and survival analysis. Springer, 2002.

56. Hashibe M, Brennan P, Chuang SC, et al. Interaction between tobacco and alcohol use and the risk of head and neck cancer: pooled analysis in the International Head and Neck Cancer Epidemiology Consortium. Cancer Epidemiol Biomarkers Prev 2009;18:541-50.

57. Haziza D. Inférence en présence d'imputation : un survol. Presented at the Journées de Méthodologie Statistique, 2002.

58. Haziza, D.Echantillonnage - Note de cours STT2000. Université de Montréal,2010.

59. Haziza D, Beaumont JF. On the construction of imputation classes in surveys. Int Stat Rev 2007;75:25-43.

60. Haziza, D. and Beliveau, A.Estimation non-paramétrique des probabilités de réponse. Université de Montréal,2010.

61. Herquelot E, Gueguen A, Roquelaure Y, et al. Work-related risk factors for incidence of lateral epicondylitis in a large working population. Scand J Work Environ Health 2013;39:578-88.

62. Horvitz DG, Thompson DJ. A generalization of sampling without replacement from a finite universe. JASA 1952;47:663-85.

63. Inpes. Les Baromètres santé, un observatoire des comportements des Français pour orienter les politiques de santé publique. http://www.inpes.sante.fr/Barometres/index.asp. 2014.

64. Inserm, UPMC, and InVS. Le réseau Sentinelles. https://websenti.u707.jussieu.fr/sentiweb/?site=fr. 2014.

65. Irdes. Enquête sur la santé et la protection sociale (ESPS). http://www.irdes.fr/recherche/enquetes/esps-enquete-sur-la-sante-et-la-protection-sociale/actualites.html. 2014.

66. Jenkins P, Scheim C, Wang JT, et al. Assessment of coverage rates and bias using double sampling methodology. J Clin Epidemiol 2004;57:123-30.

67. Kang JDY, Schafer JL. Demystifying Double Robustness: A Comparison of Alternative Strategies for Estimating a Population Mean from Incomplete Data. Statistical Science 2007;22:523-39.

68. Keogh RH, White IR. A toolkit for measurement error correction, with a focus on nutritional epidemiology. Stat Med 2014;33:2137-55.

69. Kim JK, Brick JM, Fuller WA, et al. On the bias of multiple-imputation variance estimator in survey sampling. J R Statist Soc B 2006;68:509-21.

70. Kim JK, Kim JJ. Nonresponse weighting adjustment using response probability. The Canadian Journal of Statistics 2007;35:501-14.

71. Knudsen AK, Hotopf M, Skogen JC, et al. The health status of nonparticipants in a population-based health study. Am J Epidemiol 2010;172:1306-14.

72. Kreuter F. Improving surveys with paradata: analytic uses of process information. Hoboken, New Jersey: Wiley, 2013.

73. Kreuter F, Kohler U. Analysing contact sequences in call record data. Potential and limitations of sequence indicators for nonresponse adjustments in the European Social Survey. Journal of Official Statistics 2009;25:203-26.

74. Kreuter F, Muller G, Trappmann M. Nonresponse and Measurement Error in Employment Research. Making use of Administrative Data. Public Opin Q 2010;74:880-906.

75. Lamers LM. Medical consumption of respondents and non-respondents to a mailed health survey. EUR J PUBLIC HEALTH 1997;7:267-71.

76. Lelong N, Moreau C, Kaminski M, et al. Induced abortion in France: results of the COCON study. J Gynecol Obstet Biol Reprod (Paris) 2005;34:53-61.

77. Lemaître, A. and Valenty, M.Programme de surveillance des maladies à caractère professionnel (MCP) en France. Résultats des Quinzaines MCP 2008 à 2011. Saint-Maurice: InVS,2014.

78. Lévesque, I. and Franklin, S.Pondérations longitudinale et transversale de l'Enquête sur la Dynamique du travail et du revenu –Année de référence 1997. Statistique Canada,2000.

79. Levy PS, Lemeshow S. Sampling of populations, Methods and Applications. New York: Wiley & Sons, 1991.

80. Little RJA. Survey nonresponse adjustments for estimates of means. Int Stat Rev 1986;54:139-57.

81. Little RJA, Rubin DB. Statistical analysis with missing data. New York: Wiley, 1987.

82. Lohr SL. Sampling: Design and analysis. Duxbury press, 1999.

83. Lundquist P, Särndal CE. Aspects of Responsive Design with Applications to the Swedish Living Conditions Survey. Journal of Official Statistics 2013;29:557-82.

84. Maitland A, Casas Cordero C, Kreuter F. An exploration into the use of paradata for nonresponse adjustment in a health survey. Presented at the Joint statistical meetings-Section on survey research methods, 2008.

85. Martikainen P, Laaksonen M, Piha K, et al. Does survey non-response bias the association between occupational social class and health? Scand J Public Health 2007;35:212-5.

86. Meffre, C.Prévalence des hépatites B et C en 2004. Saint-Maurice: InVS,2007.

87. Merkle D, Edelman M, Dykeman K, et al. An Experimental Study of Ways to Increase Exit Poll Response Rates and Reduce Survey Error. Presented at the Annual Conference of the American Association for Public Opinion Research, Saint-Louis (MO), 1998.

88. Moisan F, Gourlet V, Mazurie JL, et al. Prediction model of Parkinson's disease based on antiparkinsonian drug claims. Am J Epidemiol 2011;174:354-63.

89. Moriarty C, Dahlhamer JM. Adjustment for unit nonresponse in the National Health Interview Survey. Presented at the JSM - Section survey research methods, 2012.

90. Myers JA, Rassen JA, Gagne JJ, et al. Effects of adjusting for instrumental variables on bias and precision of effect estimates. Am J Epidemiol 2011;174:1213-22.

91. Nummela O, Sulander T, Helakorpi S, et al. Register-based data indicated nonparticipation bias in a health study among aging people. J Clin Epidemiol 2011;64:1418-25.

92. Olson K. Survey partipication, nonresponse bias, measurement error bias, and total bias. Public Opin Q 2006;70:737-58.

93. Olson K. Do non-response follow-ups improve or reduce data quality? A review of existing literature. J R Statist Soc A 2013;176:129-45.

94. Olson K. Paradata for nonresponse adjustment. The Annals of the American Academy of Political and Social Science 2013;645:142-70.

95. Olson K, Groves RM. An examination of within-person variation in response propensity over the data collection field period. Journal of Official Statistics 2012;28:29-51.

96. Pearl J. Invited commentary: understanding bias amplification. Am J Epidemiol 2011;174:1223-7.

97. Peytchev A, Peytcheva E, Groves RM. Measurement error, unit nonresponse, and self-reports of abortion experiences. Public Opin Q 2010;74:319-27.

98. Peytchev A, Rosen J, Riley S, et al. Reduction of nonresponse bias in surveys through case prioritization. Survey research methodology paper 2010;4:21-9.

99. Pirracchio R. Nouveautés en modélisation non paramétrique - Apports du Super Learner. Presented at the Adelf, 2014.

100. Polley, E. C. and van der Lann, M. J.Super learner in prediction. U.C. Berkeley Division of Biostatistics Working Paper Series,2010.

101. Reijneveld SA, Stronks K. The impact of response bias on estimates of health care utilization in a metropolitan area: The use of administrative data. Int J Epidemiol 1999;28:1134-40.

102. Rigou, A. and Thélot, B.Hospitalisations pour brûlures à partir des données du Programme de médicalisation des systèmes d'information, France métropolitaine, 2009 – Synthèse . Saint-Maurice: InVS,2011.

103. Rizzo L, Kalton G, Brick JM. Comparaison de quelques méthodes de correction de la non-réponse d'un panel. Techniques d'enquête 1996;22:43-53.

104. Robins JM, Sued M, Lei-Gomez Q, et al. Comment: performance of double-robust estimators when "inverse probability" weights are highly variable. Statistical Science 2013;22:544-59.

105. Roquelaure Y, Ha C, Pélier-Cady MC, et al. Surveillance en population générale du syndrome du canal carpien dans le Maine-et-Loire en 2002 et 2003. Bull Epidemiol Hebd 2005;44-45.

106. Rosendaal FR. National registers and their use for medical research. Eur J Epidemiol 2014;29:539-40.

107. Royston P, Sauerbrei W. Stability of multivariable fractional polynomial models with selection of variables and transformations: a bootstrap investigation. Stat Med 2003;22:639-59.

108. Rubin DB. Inference and missing data. Biometrika 1976;63:581-90.

109. Rubin DB. Multiple imputations for nonresponse in surveys. New-York: Wiley, 1987.

110. Saez M, Barcelo MA, de Tuero GC. A selection-bias free method to estimate the prevalence of hypertension from an administrative primary health care database in the Girona Health Region, Spain. Comput Methods Programs Biomed 2009;93:228-40.

111. Santin G, Cohidon C, Goldberg M, et al. Depressive symptoms and atypical jobs in France, from the 2003 Decennial health survey. Am J Ind Med 2009;52:799-810.

112. Saporta G. La statistique exploratoire. Probabilités, analyse des données et statistique. Paris: Technip, 1990:115-261.

113. Särndal CE, Swensson B, Wretman J. Model assisted survey sampling. New York: Springer, 1992.

114. Särndal CE, Swensson B, Wretman J. Nonresponse. Model assisted survey sampling. New-York: Springer, 1992:556-600.

115. Särndal CE, Swensson B, Wretman J. Variance estimation. Model assisted survey sampling. New-York: Springer, 1992:418-46.

116. Schouten B, Cobben F, Bethlehem J. Indicators for the representativeness of survey response. Survey Methodol 2009;35:101-13.

117. Siegrist J, Wege N, Pohlhofer F, et al. A short generic measure of work stress in the era of globalization: Effort-reward imbalance. Int Arch Occup Environ Health 2009;82:1005-13.

118. Singleton R, Straits B. Approaches to social research. New-York: Oxford university press, 2005.

119. Stang A. Nonresponse research--an underdeveloped field in epidemiology. Eur J Epidemiol 2003;18:929-31.

120. Stang A, Jockel KH. Studies with low response proportions may be less biased than studies with high response proportions. Am J Epidemiol 2004;159:204-10.

121. Statistique Canada. Enquête Canadienne sur les Mesures de Santé. http://www23.statcan.gc.ca/imdb/p2SV_f.pl?Function=getSurvey&SDDS=5071. 2014.

122. Statistique Canada. Enquête nationale sur la santé de la population - volet ménages - longitudinal (ENSP). http://www23.statcan.gc.ca/imdb/p2SV_f.pl?Function=getSurvey&SDDS=3225. 2014.

123. Stoop, I. A. L.The hunt for the last respondent. The Hague: Social and cultural planning office,2014.

124. The American Association for Public Opinion Research.Standard definitions: final dispositions of case codes and outcome rates for surveys. 7th edition. AAPOR,2011.

125. Thygesen LC, Ersboll AK. When the entire population is the sample: strengths and limitations in register-based epidemiology. Eur J Epidemiol 2014;29:551-8.

126. Tourangeau R, Groves RM, Redline CD. Sensitive topics and reluctant respondents : demonstrating a link between nonresponse bias ans measurement error. Public Opin Q 2010;74:413-32.

127. Vercambre MN, Gilbert F. Respondents in an epidemiologic survey had fewer psychotropic prescriptions than nonrespondents: An insight into health-related selection bias using routine health insurance data. J Clin Epidemiol 2012;65:1181-9.

128. Warszawski J, Messiah A, Lellouch J, et al. Estimating means and percentages in a complex sampling survey: application to a French national survey on sexual behaviour (ACSF). Analyse des Comportements Sexuels en France. Stat Med 1997;16:397-423.

129. Winkler, WE. Cleaning and using administrative lists: Enhanced practices and computational algorithms for record linkage and modeling/editing/imputation. Presented at the JSM - Section on survey research methodology, 2011.

130. Wu S, Wang R, Zhao Y, et al. The relationship between self-rated health and objective health status: a population-based study. BMC Public Health 2013;13:320.

131. Zins M, Bonenfant S, Carton M, et al. The CONSTANCES cohort: An open epidemiological laboratory. BMC Public Health 2010;10.

ANNEXES

ANNEXE I. DÉMONSTRATIONS

I.1 NOTATIONS

U : la population

$U = U_r \cup U_m$ avec U_r la population de répondants et U_m la population de non-répondants

N : taille de la population U

$y=(y_1,\ldots,y_i,\ldots,y_N)$ N réalisations de la variable d'intérêt Y dans la population (U) de taille N

On cherche à estimer le total de Y $t_y = \sum_{i \in U} y_i$ dans la population par $\hat{t}_y$

Pour les parties I.2. à I.4., on utilise les notations suivantes :

(s) : échantillon tiré au sort dans (U)

n : taille de l'échantillon (s)

p : variabilité liées au plan de sondage

I_i : individu i tiré au sort si $I_i = 1$, 0 sinon

π_i : probabilité d'inclusion de l'individu i

q : variabilité liée à la non-réponse

(s_r) : répondants de l'échantillon (s)

(s_m) : non-répondants de l'échantillon (s)

$s = s_r \cup s_m$

R_i : individu i répondant si $R_i = 1$, 0 sinon

$\delta = (\delta_1, \dots, \delta_n)$ probabilité de réponse connue

Pour les parties I.5. à I.9., on utilise les notations suivantes :

EI :enquête initiale

(s_{EI}) : échantillon tiré au sort dans *(U)*

$$I_{EI,i} = \begin{cases} 1 \text{ si } i \in (s_{EI}) \\ 0 \text{ sinon} \end{cases}$$

$\pi_{EI,i}$: probabilité d'inclusion de l'unité *i* dans l'échantillon (s_{EI})

$$d_{EI,i} = \frac{1}{\pi_{EI,i}}$$

$(s_{EI,r})$: répondants de l'échantillon (s_{EI})

$(s_{EI,m})$: non-répondants de l'échantillon (s_{EI})

$$R_{EI,i} = \begin{cases} 1 \text{ si } i \in (s_{EI,r}) \\ 0 \text{ sinon} \end{cases}$$

$\delta_{EI,i}$: probabilité de réponse de l'unité *i* dans (s_{EI})

EC :enquête complémentaire auprès d'un échantillon de non-répondants à *EI*

(s_{EC}) : échantillon de non-répondants tiré au sort dans $(s_{EI,m})$

$$I_{EC,i} = \begin{cases} 1 \text{ si } i \in (s_{EC}) \\ 0 \text{ sinon} \end{cases}$$

$\pi_{EC,i}$: probabilité d'inclusion de l'unité *i* dans l'échantillon (s_{EC})

($s_{EC,r}$) : répondants de l'échantillon de non-répondants *(s_{EC})*

($s_{EC,m}$) : non-répondants de l'échantillon de non-répondants *(s_{EC})*

$$R_{EC,i} = \begin{cases} 1 \text{ si } i \in (s_{EC,r}) \\ 0 \text{ sinon} \end{cases}$$

$\delta_{EC,i}$: probabilité de réponse de l'unité *i* dans (s_{EC})

I.2 ESTIMATEUR SANS BIAIS D'UN TOTAL SAS SANS REMISE

Dans un sondage aléatoire simple sans remise, un **estimateur sans biais du total t_y** est donné par : $\hat{t}_{y,SAS} = \sum_{i \in s} d_i y_i$ avec $d_i = \frac{1}{\pi_i} = \frac{N}{n}$

Démonstration :

$$E(\hat{t}_{y,SAS}) = \frac{N}{n} E\left(\sum_{i \in s} y_i\right) = \frac{N}{n} E\left(\sum_{i=1}^{N} I_i y_i\right) = \frac{N}{n} \sum_{i=1}^{N} E(I_i) y_i = \frac{N}{n} \sum_{i=1}^{N} \pi_i y_i = \frac{N}{n} \sum_{i=1}^{N} \frac{n}{N} y_i$$

$$= \frac{N}{n} \frac{n}{N} \sum_{i=1}^{N} y_i = \sum_{i=1}^{N} y_i = t_y$$

→ $\hat{t}_{y,SAS}$ est un estimateur sans biais de t_y

I.3 ESTIMATEUR DE HORVITZ-THOMPSON

Sous l'hypothèse que $\pi_i > 0$ pour tout *i* (condition de positivité), l'estimateur de Horvitz-Thompson est donné par : $\hat{t}_{y,\pi} = \sum_{i \in s} \frac{y_i}{\pi_i}$

$\hat{t}_{y,\pi}$ est-il un estimateur sans biais de t_y ?

$$E(\hat{t}_{y,\pi}) = E\left(\sum_{i \in U} \frac{y_i}{\pi_i} I_i\right) = \sum_{i \in U} \frac{y_i}{\pi_i} E(I_i) = \sum_{i \in U} \frac{y_i}{\pi_i} \pi_i = \sum_{i \in U} y_i = t_y$$

→ $\hat{t}_{y,\pi}$ est un estimateur sans biais de t_y

Remarque sur la condition de positivité

S'il existe des unités pour lesquelles $\pi_i = 0$, alors :

$$\hat{t}_{y,\pi} = \sum_{i \in s} \frac{y_i}{\pi_i} = \sum_{i \in U/\pi_i > 0} \frac{y_i}{\pi_i} I_i$$

$$E(\hat{t}_{y,\pi}) = E\left(\sum_{i \in U/\pi_i > 0} \frac{y_i}{\pi_i} I_i\right) = \sum_{i \in U/\pi_i > 0} \frac{y_i}{\pi_i} E(I_i) = \sum_{i \in U/\pi_i > 0} \frac{y_i}{\pi_i} \pi_i = \sum_{i \in U/\pi_i > 0} y_i$$

$$E(\hat{t}_{y,\pi}) \neq \sum_{i \in U/\pi_i > 0} y_i + \sum_{i \in U/\pi_i = 0} y_i \neq t_y$$

I.4 ESTIMATEUR ASYMPTOTIQUEMENT SANS BIAIS D'UN TOTAL AVEC NON-RÉPONSE PAR REPONDÉRATION

1. Supposons que la probabilité de réponse δ à l'enquête est connue

Un estimateur sans biais du total t_y est donné par :

$$\hat{t}_{y,MAR(X),pond} = \sum_{i \in s_r} \frac{y_i}{\pi_i \delta_i}$$

Démonstration :

$$E\left(\hat{t}_{y,MAR(X),pond}\right) = E\left(\sum_{i \in s_r} \frac{y_i}{\pi_i \delta_i}\right) = E_p\left(\sum_{i \in U} \frac{y_i}{\pi_i \delta_i} I_i R_i\right)$$

$$= \sum_{i \in U} \frac{y_i}{\pi_i \delta_i} E_p(I_i R_i)$$

$$= \sum_{i \in U} \frac{y_i}{\pi_i \delta_i} E_p E_q(I_i R_i / I_i)$$

$$= \sum_{i \in U} \frac{y_i}{\pi_i \delta_i} E_p(I_i \delta_i) = \sum_{i \in U} \frac{y_i}{\pi_i \delta_i} E_p(I_i)\delta_i = \sum_{i \in U} \frac{y_i}{\pi_i \delta_i} \pi_i \delta_i = \sum_{i \in U} y_i = t_y$$

$\hat{t}_{y,MAR(X),pond}$ est un estimateur sans biais de t_y

2. Supposons que la probabilité de réponse δ à l'enquête est estimée

Sous l'hypothèse que le modèle pour estimer δ_i est correctement spécifié, un estimateur asymptotiquement sans biais du total est :

$$\hat{\hat{t}}_{y,MAR(X),pond} = \sum_{i \in s_r} \frac{y_i}{\pi_i \hat{\delta}_i(\hat{\alpha}_i, X_i)}$$

Démonstration :

$$E\left(\hat{\hat{t}}_{y,MAR(X),pond}\right) = E_p E_q \left(\sum_{i \in s_r} \frac{y_i}{\pi_i \hat{\delta}_i(\hat{\alpha}_i, X_i)} \right) = E_p E_q \left(\sum_{i \in U} \frac{y_i}{\pi_i \hat{\delta}_i(\hat{\alpha}_i, X_i)} I_i R_i \right)$$

$$= E_p \left(\sum_{i \in U} \frac{y_i}{\pi_i \hat{\delta}_i(\hat{\alpha}_i, X_i)} I_i E_q(R_i) \right)$$

sous l'hypothèse MAR(X), on a $E_q(R_i) = \delta(X_i)$

$$E\left(\hat{\hat{t}}_{y,MAR(X),pond}\right) = E_p \left(\sum_{i \in U} \frac{y_i}{\pi_i \hat{\delta}_i(\hat{\alpha}_i, X_i)} I_i \delta(X_i) \right)$$

sous l'hypothèse que le modèle est correctement spécifié, on a $\delta(X_i) = \delta(\alpha, X_i)$

$$E\left(\hat{\hat{t}}_{y,MAR(X),pond}\right) = E_p \left(\sum_{i \in U} \frac{y_i}{\pi_i \hat{\delta}_i(\hat{\alpha}_i, X_i)} I_i \delta(\alpha, X_i) \right)$$

$$E\left(\hat{\hat{t}}_{y,MAR(X),pond}\right) = \sum_{i \in U} \frac{y_i}{\pi_i} \delta(\alpha, X_i) E_p \left(\frac{I_i}{\hat{\delta}_i(\hat{\alpha}_i, X_i)} \right)$$

asymptotiquement, le rapport de l'espérance est l'espérance des rapports

$$E\left(\hat{\hat{t}}_{y,MAR(X),pond}\right) \equiv \sum_{i \in U} \frac{y_i}{\pi_i} \delta(\alpha, X_i) \left(\frac{E_p(I_i)}{E_p(\hat{\delta}_i(\hat{\alpha}_i, X_i))} \right)$$

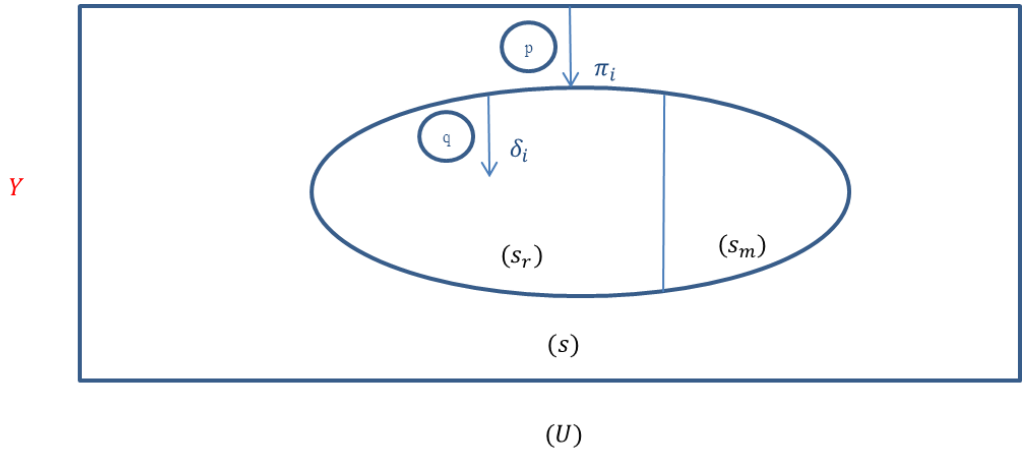
$$E\left(\hat{\hat{t}}_{y,MAR(X),pond}\right) \equiv \sum_{i \in U} \frac{y_i}{\pi_i} \delta(\alpha, X_i) \left(\frac{\pi_i}{\delta(\alpha, X_i)} \right)$$

$$E\left(\hat{\hat{t}}_{y,MAR(X),pond}\right) \equiv \sum_{i \epsilon U} y_i \equiv t_y$$

→ $\hat{\hat{t}}_{y,MAR(X),pond}$ est un estimateur asymptotiquement sans biais de t_y

I.5 ESTIMATEUR DE LA VARIANCE D'UN TOTAL AVEC NON-RÉPONSE

Figure I-1 : Représentation graphique des sources de variabilité dans un sondage à probabilités d'inclusion inégales avec non-réponse



$$\hat{t}_{y,pond} = \sum_{i\in U} \frac{1}{\pi_i \delta_i} y_i I_i R_i$$

$$\hat{V}(\hat{t}_{y,pond}) = \sum_{i\in s_r} \sum_{j\in s_r} \frac{(\pi_{ij}-\pi_i\pi_j)}{\pi_i\pi_j} \frac{1}{\pi_{ij}\delta_i\delta_j} y_i y_j + \sum_{i\in s_r} \frac{1}{\pi_i^2} \frac{(1-\delta_i)}{\delta_i^2} y_i^2$$

Démonstration

D'où vient la variance ? De la différence entre la vraie valeur du total et de son estimateur qui est due au plan de sondage et à la non-réponse (représentés en bleu et numérotés de 1 à 2 dans la représentation graphique).

$$\hat{t}_{y,pond} - t_y = (\hat{t}_{y,\pi} - t_y) + (\hat{t}_{y,pond} - \hat{t}_{y,\pi})$$

Pour calculer la variance de $\hat{t}_{y,pond}$*, on va utiliser la propriété suivante :*

$$V(Y) = V(E(Y/X)) + E(V(Y/X))$$

1. Variance de $\hat{t}_{y,pond}$

$E_q(R) = E(R/I) = \delta$

$E_p(I) = \pi$

$V(\hat{t}_{y,pond}) = V_p E_q(\hat{t}_{y,pond}) + E_p V_q(\hat{t}_{y,pond})$

- $V_p E_q(\hat{t}_{y,pond}) = V_p E_q\left(\sum_{i\epsilon U} \frac{1}{\pi_i \delta_i} y_i I_i R_i\right)$

$$E_q\left(\sum_{i\in U} \frac{1}{\pi_i \delta_i} y_i I_i R_i\right) = \sum_{i\in U} \frac{1}{\pi_i \delta_i} y_i I_i E_q(R_i)$$

$$E_q\left(\sum_{i\in U} \frac{1}{\pi_i \delta_i} y_i I_i R_i\right) = \sum_{i\in U} \frac{1}{\pi_i \delta_i} y_i I_i \delta_i$$

$$E_q\left(\sum_{i\in U} \frac{1}{\pi_i \delta_i} y_i I_i R_i\right) = \sum_{i\in U} \frac{1}{\pi_i} y_i I_i$$

$V_p E_q(\hat{t}_{y,pond}) = V_p\left(\sum_{i\epsilon U} \frac{1}{\pi_i} y_i I_i\right)$ *(variance de l'estimateur de Horvitz-Thompson)*

$$V_p E_q(\hat{t}_{y,pond}) = \sum_{i\epsilon U} \sum_{i\epsilon U} (\pi_{ij} - \pi_i \pi_j) \frac{y_i}{\pi_i} \frac{y_j}{\pi_j}$$

- $E_p V_q(\hat{t}_{y,pond}) = E_p V_q\left(\sum_{i\epsilon U} \frac{1}{\pi_i \delta_i} y_i I_i R_i\right)$

Au niveau de l'échantillon de répondants, $\frac{1}{\pi_i \delta_i} y_i I_i$ est fixe

On suppose que R_i suit une loi de Bernoulli de moyenne δ_i et de variance $\delta_i(1-\delta_i)$

($Var(\sum_{i=1}^{N} aX_i) = \sum_{i=1}^{N} a^2 Var(X_i)$; a constante si les X_i sont indépendantes)

$$E_p V_q\left(\hat{t}_{y,pond}\right) = E_p\left(\sum_{i\epsilon U}\left(\frac{1}{\pi_i\delta_i}y_i I_i\right)^2 V_q(R_i)\right)$$

$$E_p V_q\left(\hat{t}_{y,pond}\right) = E_p\left(\sum_{i\epsilon U}\left(\frac{1}{\pi_i\delta_i}y_i I_i\right)^2 \delta_i(1-\delta_i)\right)$$

I binaire en 0/1 → $I^2=I$

$$E_p V_q\left(\hat{t}_{y,pond}\right) = E_p\left(\sum_{i\epsilon U}\left(\frac{1}{\pi_i}y_i\right)^2 \frac{I_i}{\delta_i}(1-\delta_i)\right)$$

$$E_p V_q\left(\hat{t}_{y,pond}\right) = \sum_{i\epsilon U}\left(\frac{1}{\pi_i}y_i\right)^2 \frac{(1-\delta_i)}{\delta_i}E_p(I_i)$$

$$E_p V_q\left(\hat{t}_{y,pond}\right) = \sum_{i\epsilon U}\left(\frac{1}{\pi_i}y_i\right)^2 \frac{(1-\delta_i)}{\delta_i}\pi_i$$

$$E_p V_q\left(\hat{t}_{y,pond}\right) = \sum_{i\epsilon U}\frac{(1-\delta_i)}{\pi_i\delta_i}y_i^2$$

Donc,

$$V\left(\hat{t}_{y,pond}\right) = \sum_{i\epsilon U}\sum_{i\epsilon U}(\pi_{ij}-\pi_i\pi_j)\frac{y_i}{\pi_i}\frac{y_j}{\pi_j} + \sum_{i\epsilon U}\frac{(1-\delta_i)}{\pi_i\delta_i}y_i^2$$

2. Estimateur sans biais de la variance de $\hat{t}_{y,pond}$

$$\hat{V}\left(\hat{t}_{y,pond}\right) = \sum_{i\in s_r}\sum_{j\in s_r}\frac{(\pi_{ij}-\pi_i\pi_j)}{\pi_i\pi_j}\frac{1}{\pi_{ij}\delta_i\delta_j}y_i y_j + \sum_{i\in s_r}\frac{1}{\pi_i^2}\frac{(1-\delta_i)}{\delta_i^2}y_i^2$$

Démonstration :

On va utiliser la propriété suivante : $E(Y) = E(E(Y/X))$

$$E\left(\hat{V}\left(\hat{t}_{y,pond}\right)\right) = E_p E_q\left(\sum_{i\in s_r}\sum_{j\in s_r}\frac{(\pi_{ij}-\pi_i\pi_j)}{\pi_i\pi_j}\frac{1}{\pi_{ij}\delta_i\delta_j}y_i y_j + \sum_{i\in s_r}\frac{1}{\pi_i^2}\frac{(1-\delta_i)}{\delta_i^2}y_i^2\right)$$

$$E\left(\hat{V}\left(\hat{t}_{y,pond}\right)\right) = E_p E_q\left(\sum_{i\in s_r}\sum_{j\in s_r}\frac{(\pi_{ij}-\pi_i\pi_j)}{\pi_i\pi_j}\frac{1}{\pi_{ij}\delta_i\delta_j}y_i y_j\right) + E_p E_q\left(\sum_{i\in s_r}\frac{1}{\pi_i^2}\frac{(1-\delta_i)}{\delta_i^2}y_i^2\right)$$

$$E\left(\hat{V}\left(\hat{t}_{y,pond}\right)\right)$$

$$= E_p E_q\left(\sum_{i\in U}\sum_{j\in U}\frac{(\pi_{ij}-\pi_i\pi_j)}{\pi_i\pi_j}\frac{1}{\pi_{ij}\delta_i\delta_j}y_i y_j I_i I_j R_i R_j\right)$$

$$+ E_p E_q\left(\sum_{i\in U}\frac{1}{\pi_i^2}\frac{(1-\delta_i)}{\delta_i^2}y_i^2 I_i R_i\right)$$

$$E\left(\hat{V}\left(\hat{t}_{y,pond}\right)\right)$$

$$= E_p\left(\sum_{i\in U}\sum_{j\in U}\frac{\left(\pi_{ij}-\pi_i\pi_j\right)}{\pi_i\pi_j}\frac{1}{\pi_{ij}\delta_i\delta_j}y_i y_j I_i I_j E_q(R_i R_j)\right)$$

$$+ E_p\left(\sum_{i\in U}\frac{1}{\pi_i^2}\frac{(1-\delta_i)}{\delta_i^2}y_i^2 I_i E_q(R_i)\right)$$

$$E\left(\hat{V}\left(\hat{t}_{y,pond}\right)\right)$$

$$= E_p\left(\sum_{i\in U}\sum_{j\in U}\frac{\left(\pi_{ij}-\pi_i\pi_j\right)}{\pi_i\pi_j}\frac{1}{\pi_{ij}\delta_i\delta_j}y_i y_j I_i I_j \delta_i\delta_j\right)$$

$$+ E_p\left(\sum_{i\in U}\frac{1}{\pi_i^2}\frac{(1-\delta_i)}{\delta_i^2}y_i^2 I_i\delta_i\right)$$

$$E\left(\hat{V}\left(\hat{t}_{y,pond}\right)\right) = E_p\left(\sum_{i\in U}\sum_{j\in U}\frac{\left(\pi_{ij}-\pi_i\pi_j\right)}{\pi_i\pi_j}\frac{1}{\pi_{ij}}y_i y_j I_i I_j\right) + E_p\left(\sum_{i\in U}\frac{1}{\pi_i^2}\frac{(1-\delta_i)}{\delta_i}y_i^2 I_i\right)$$

$$E\left(\hat{V}\left(\hat{t}_{y,pond}\right)\right) = \sum_{i\in U}\sum_{j\in U}\frac{\left(\pi_{ij}-\pi_i\pi_j\right)}{\pi_i\pi_j}\frac{1}{\pi_{ij}}y_i y_j E_p(I_i I_j) + \sum_{i\in U}\frac{1}{\pi_i^2}\frac{(1-\delta_i)}{\delta_i}y_i^2 E_p(I_i)$$

$$E\left(\hat{V}\left(\hat{t}_{y,pond}\right)\right) = \sum_{i\in U}\sum_{j\in U}\frac{\left(\pi_{ij}-\pi_i\pi_j\right)}{\pi_i\pi_j}\frac{1}{\pi_{ij}}y_i y_j \pi_{ij} + \sum_{i\in U}\frac{1}{\pi_i^2}\frac{(1-\delta_i)}{\delta_i}y_i^2 \pi_i$$

$$E\left(\hat{V}\left(\hat{t}_{y,pond}\right)\right) = \sum_{i\in U}\sum_{j\in U}\frac{\left(\pi_{ij}-\pi_i\pi_j\right)}{\pi_i\pi_j}y_i y_j + \sum_{i\in U}\frac{1}{\pi_i}\frac{(1-\delta_i)}{\delta_i}y_i^2$$

$$E\left(\hat{V}\left(\hat{t}_{y,pond}\right)\right) = V\left(\hat{t}_{y,pond}\right)$$

→ $\hat{V}\left(\hat{t}_{y,pond}\right)$ estimateur sans biais de $V\left(\hat{t}_{y,pond}\right)$

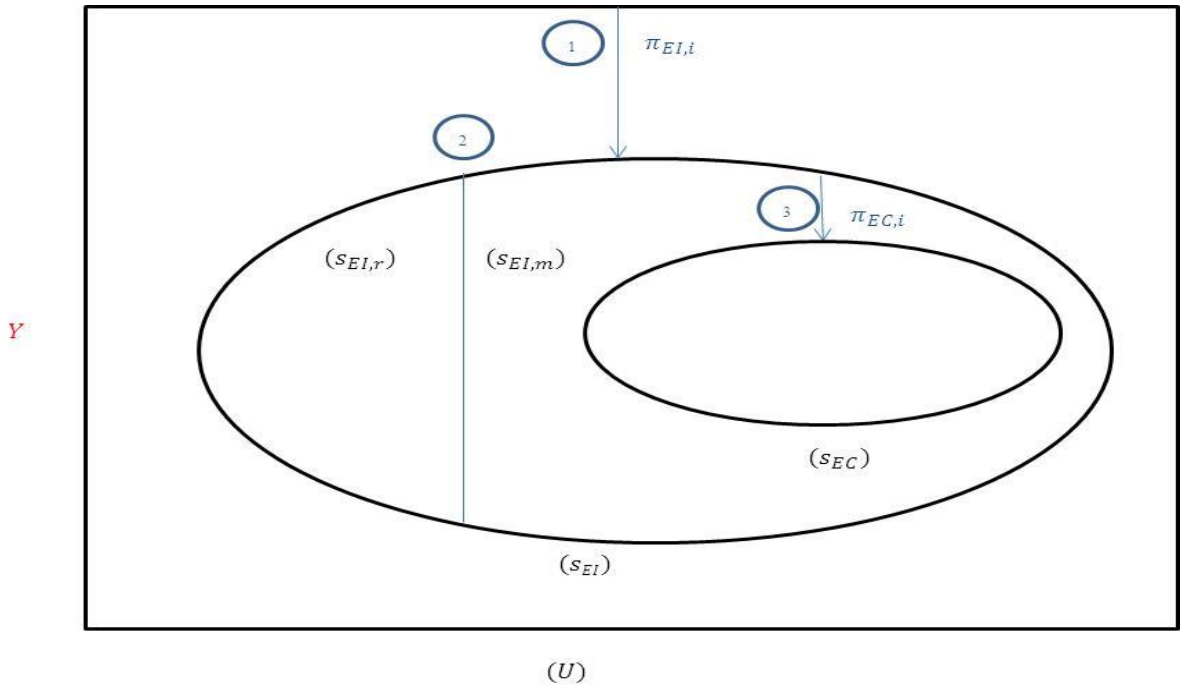
I.6 ESTIMATEUR D'UN TOTAL POUR UN PLAN DE SONDAGE EN DEUX PHASES POUR NON-RÉPONSE

$$\hat{t}_{y,2phases,nr} = \sum_{i \epsilon s_{EI,r}} \frac{y_i}{\pi_{EI,i}} + \sum_{i \epsilon s_{EC}} \frac{y_i}{\pi_{EI,i}\pi_{EC/s_{EI,nr},i}}$$

Démonstration :

$\hat{t}_{y,2phases,nr}$ estimateur sans biais de t_y ?

Figure I-2 : Représentation graphique des sources de variabilité dans un sondage en deux phases pour non-réponse



Soit :

$E_3 = E(./I_{EI}, R_{EI})$

$E_2 = E(./I_{EI})$

$E_1 = E_p(.)$, espérance liée au plan de première phase

On a :

$$E_3(I_{EC}) = E(I_{EC}/I_{EI}, R_{EI}\ ) = \pi_{EC/s_{EI,nr}} = \pi_{EC}$$

$$E_2(R_{EI}) = E(R_{EI}/I_{EI}) = \delta_{EI}$$

$$E_1(I_{EI}) = \pi_{EI}$$

On va montrer que $\hat{\bar{p}}_{s_{EI,r} \cup s_{EC}}$ est un estimateur sans biais de $\bar{p}_{s_{EI,r} \cup s_{EC}}$ en utilisant la propriété suivante :

$$E(Y) = E(E(Y/X))$$

Ainsi, on a $E(\hat{t}_{y,2phases,nr}) = E_1E_2E_3(\hat{t}_{y,2phases,nr})$

$$E_3\big(\hat{t}_{y,2phases,nr}\big) = E_3\Big(\sum_{i\epsilon U} \frac{1}{\pi_{EI,i}} y_i I_{EI,i} R_{EI,i} + \sum_{i\epsilon U} \frac{1}{\pi_{EI,i}} \frac{1}{\pi_{EC,i}} y_i I_{EI,i}\big(1 - R_{EI,i}\big) I_{EC,i}\Big)$$

$$E_3\big(\hat{t}_{y,2phases,nr}\big) = \sum_{i\epsilon U} \frac{1}{\pi_{EI,i}} y_i I_{EI,i} R_{EI,i} + \sum_{i\epsilon U} \frac{1}{\pi_{EI,i}} \frac{1}{\pi_{EC,i}} y_i I_{EI,i}\big(1 - R_{EI,i}\big) E_3(I_{EC,i})$$

$$E_3\big(\hat{t}_{y,2phases,nr}\big) = \sum_{i\epsilon U} \frac{1}{\pi_{EI,i}} y_i I_{EI,i} R_{EI,i} + \sum_{i\epsilon U} \frac{1}{\pi_{EI,i}} \frac{1}{\pi_{EC,i}} y_i I_{EI,i}\big(1 - R_{EI,i}\big) \pi_{EC,i}$$

$$E_3\big(\hat{t}_{y,2phases,nr}\big) = \sum_{i\epsilon U} \frac{1}{\pi_{EI,i}} y_i I_{EI,i} R_{EI,i} + \sum_{i\epsilon U} \frac{1}{\pi_{EI,i}} y_i I_{EI,i}\big(1 - R_{EI,i}\big)$$

$$E_3\big(\hat{t}_{y,2phases,nr}\big) = \sum_{i\epsilon U} \frac{1}{\pi_{EI,i}} y_i I_{EI,i}$$

$$E_2E_3\big(\hat{t}_{y,2phases,nr}\big) = E_2\Big(\sum_{i\epsilon U} \frac{1}{\pi_{EI,i}} y_i I_{EI,i}\Big)$$

$$E_2E_3\left(\hat{t}_{y,2phases,nr}\right) = \sum_{i\epsilon U} \frac{1}{\pi_{EI,i}} y_i I_{EI,i}$$

$$E_1E_2E_3\left(\hat{t}_{y,2phases,nr}\right) = E_1\left(\sum_{i\epsilon U} \frac{1}{\pi_{EI,i}} y_i I_{EI,i}\right)$$

$$E_1E_2E_3\left(\hat{t}_{y,2phases,nr}\right) = \sum_{i\epsilon U} \frac{1}{\pi_{EI,i}} y_i E_1(I_{EI,i})$$

$$E_1E_2E_3\left(\hat{t}_{y,2phases,nr}\right) = \sum_{i\epsilon U} \frac{1}{\pi_{EI,i}} y_i \pi_{EI,i}$$

$$E_1E_2E_3\left(\hat{t}_{y,2phases,nr}\right) = \sum_{i\epsilon U} y_i$$

$$E_1E_2E_3\left(\hat{t}_{y,2phases,nr}\right) = t_y$$

→ $\hat{t}_{y,2phases,nr}$ estimateur sans biais de t_y

I.7 ESTIMATEUR DE LA VARIANCE D'UN TOTAL POUR UN PLAN DE SONDAGE EN DEUX PHASES POUR NON-RÉPONSE

$\hat{V}\left(\hat{t}_{y,2phases,nr}\right)$

$$= \sum_{i\epsilon s_{EI,r}} \sum_{j\epsilon s_{EI,r}} \frac{\Delta_{EI,ij}}{\pi_{EI,ij}} y_i y_j + \sum_{i\epsilon s_{EI}} \sum_{j\epsilon s_{EC}} \frac{\Delta_{EI,ij}}{\pi_{EI,ij}} \frac{1}{\pi_{EC,j}} y_i y_j$$

$$+ \sum_{i\epsilon s_{EC}} \sum_{j\epsilon s_{EI}} \frac{\Delta_{EI,ij}}{\pi_{EI,ij}} \frac{1}{\pi_{EC,i}} y_i y_j + \sum_{i\epsilon s_{EC}} \sum_{j\epsilon s_{EC}} \frac{\Delta_{EI,ij}}{\pi_{EI,ij}} \frac{1}{\pi_{EC,ij}} y_i y_j$$

$$+ \sum_{i\epsilon s_{EC}} \sum_{j\epsilon s_{EC}} \frac{\Delta_{EC,ij}}{\pi_{EI,i}\pi_{EI,j}\pi_{EC,ij}} y_i y_j$$

avec $\Delta_{EI,ij} = \frac{\pi_{EI,ij} - \pi_{EI,i}\pi_{EI,j}}{\pi_{EI,i}\pi_{EI,j}}$ et $\Delta_{EC,ij} = \frac{\pi_{EC,ij} - \pi_{EC,i}\pi_{EC,j}}{\pi_{EC,i}\pi_{EC,j}}$

Démonstration

1. Variance de $\hat{t}_{y,2phases,nr}$

Pour calculer la variance de $\hat{t}_{y,2phases,nr}$, on va utiliser la propriété suivante :

$$V(Y) = V(E(Y/X)) + E(V(Y/X))$$

$$\hat{t}_{y,2phases,nr} = \sum_{i\epsilon s_{EI,r}} \frac{1}{\pi_{EI,i}} y_i + \sum_{i\epsilon s_{EC}} \frac{1}{\pi_{EI,i}\pi_{EC,i}} y_i$$

$E_3 = E(I_{EC}/I_{EI}, R_{EI}) = \pi_{EC}$

$E_2 = E(R_{EI}/I_{EI}) = \delta_{EI}$

$E_1 = E(I_{EI}) = \pi_{EI}$

$V(\hat{t}_{y,2phases,nr}) = V_1E_2E_3(\hat{t}_{y,2phases,nr}) + E_1V_2E_3(\hat{t}_{y,2phases,nr}) + E_1E_2V_3(\hat{t}_{y,2phases,nr})$

- $E_1E_2V_3(\hat{t}_{y,2phases,nr}) = E_1E_2V_3\left(\sum_{i\epsilon s_{EI,r}} \frac{1}{\pi_{EI,i}} y_i + \sum_{i\epsilon s_{EC}} \frac{1}{\pi_{EI,i}\pi_{EC,i}} y_i\right)$

$$E_1E_2V_3(\hat{t}_{y,2phases,nr}) = E_1E_2V_3\left(\sum_{i\epsilon s_{EI,r}} \frac{1}{\pi_{EI,i}} y_i + \sum_{i\epsilon s_{EC}} \frac{1}{\pi_{EI,i}\pi_{EC,i}} y_i\right)$$

$$E_1E_2V_3(\hat{t}_{y,2phases,nr}) = E_1E_2V_3\left(\sum_{i\epsilon s_{EI,r}} \frac{1}{\pi_{EI,i}} y_i\right) + E_1E_2V_3\left(\sum_{i\epsilon s_{EC}} \frac{1}{\pi_{EI,i}\pi_{EC,i}} y_i\right)$$

$$E_1E_2V_3(\hat{t}_{y,2phases,nr}) = E_1E_2V_3\left(\sum_{i\epsilon s_{EC}} \frac{1}{\pi_{EI,i}\pi_{EC,i}} y_i\right)$$

$$E_1E_2V_3(\hat{t}_{y,2phases,nr}) = E_1E_2V_3\left(\sum_{i\epsilon U} \frac{1}{\pi_{EI,i}\pi_{EC,i}} y_i I_{EI,i}(1 - R_{EI,i}) I_{EC,i}\right)$$

$$E_1E_2V_3(\hat{t}_{y,2phases,nr}) = E_1E_2V_3\left(\sum_{i\epsilon U} \frac{1}{\pi_{EI,i}\pi_{EC,i}} y_i I_{EI,i}(1 - R_{EI,i}) I_{EC,i}\right)$$

$Soit\ \phi_i = \frac{1}{\pi_{EI,i}} y_i I_{EI,i}(1 - R_{EI,i})$

$E_1E_2V_3(\hat{t}_{y,2phases,nr}) = E_1E_2V_3\left(\sum_{i\epsilon U} \frac{1}{\pi_{EC,i}} \phi_i I_{EC}\right)$ *(variance de l'estimateur de Horvitz-Thompson)*

$$E_1E_2V_3(\hat{t}_{y,2phases,nr}) = E_1E_2\left(\sum_{i\epsilon U}\sum_{i\epsilon U} (\pi_{EC,ij} - \pi_{EC,i}\pi_{EC,j}) \frac{\phi_i}{\pi_{EC,i}} \frac{\phi_j}{\pi_{EC,j}}\right)$$

- $E_1V_2E_3\left(\hat{t}_{y,2phases,nr}\right) = E_1V_2E_3\left(\sum_{i\epsilon s_{EI,r}} \frac{1}{\pi_{EI,i}} y_i + \sum_{i\epsilon s_{EC}} \frac{1}{\pi_{EI,i}\pi_{EC,i}} y_i\right)$

$$E_1V_2E_3\left(\hat{t}_{y,2phases,nr}\right) = E_1V_2E_3\left(\sum_{i\epsilon U} \frac{1}{\pi_{EI,i}} y_i I_{EI,i} r_{EI,i} + \sum_{i\epsilon U} \frac{1}{\pi_{EI,i}\pi_{EC,i}} y_i I_{EI,i}(1 - R_{EI,i}) I_{EC,i}\right)$$

$$E_1V_2E_3\left(\hat{t}_{y,2phases,nr}\right) = E_1V_2E_3\left(\sum_{i\epsilon U} \frac{1}{\pi_{EI,i}} y_i I_{EI,i} R_{EI,i}\right) + E_1V_2E_3\left(\sum_{i\epsilon U} \frac{1}{\pi_{EI,i}\pi_{EC,i}} y_i I_{EI}(1 - R_{EI,i}) I_{EC,i}\right)$$

$$E_1V_2E_3\left(\hat{t}_{y,2phases,nr}\right) = E_1V_2\left(\sum_{i\epsilon U} \frac{1}{\pi_{EI,i}} y_i I_{EI,i} R_{EI,i}\right) + E_1V_2\left(\sum_{i\epsilon U} \frac{1}{\pi_{EI,i}\pi_{EC,i}} y_i I_{EI,i}(1 - R_{EI,i}) E_3(I_{EC,i})\right)$$

$$E_1V_2E_3\left(\hat{t}_{y,2phases,nr}\right) = E_1V_2\left(\sum_{i\epsilon U} \frac{1}{\pi_{EI,i}} y_i I_{EI,i} R_{EI,i}\right) + E_1V_2\left(\sum_{i\epsilon U} \frac{1}{\pi_{EI,i}\pi_{EC,i}} y_i I_{EI,i}(1 - R_{EI,i}) \pi_{EC,i}\right)$$

$$E_1V_2E_3\left(\hat{t}_{y,2phases,nr}\right) = E_1V_2\left(\sum_{i\epsilon U} \frac{1}{\pi_{EI,i}} y_i I_{EI,i} R_{EI,i}\right) + E_1V_2\left(\sum_{i\epsilon U} \frac{1}{\pi_{EI,i}} y_i I_{EI,i}(1 - R_{EI,i})\right)$$

$$E_1V_2E_3\left(\hat{t}_{y,2phases,nr}\right) = E_1V_2\left(\sum_{i\epsilon U} \frac{1}{\pi_{EI,i}} y_i I_{EI,i}(R_{EI,i} + (1 - R_{EI,i}))\right)$$

$$E_1V_2E_3\left(\hat{t}_{y,2phases,nr}\right) = E_1V_2\left(\sum_{i\epsilon U} \frac{1}{\pi_{EI,i}} y_i I_{EI,i}\right)$$

Plus de $R_{EI,i}$ → Pas de variable aléatoire au niveau 2 → Variance nulle

$E_1 V_2 E_3(\hat{t}_{y,2phases,nr}) = 0$

- $V_1 E_2 E_3(\hat{t}_{y,2phases,nr}) = V_1 E_2 E_3\left(\sum_{i\epsilon s_{EI,r}} \frac{1}{\pi_{EI,i}} y_i + \sum_{i\epsilon s_{EC}} \frac{1}{\pi_{EI,i}\pi_{EC,i}} y_i\right)$

$$V_1 E_2 E_3(\hat{t}_{y,2phases,nr}) = V_1 E_2 E_3\left(\sum_{i\epsilon U} \frac{1}{\pi_{EI,i}} y_i I_{EI,i} R_{EI,i} + \sum_{i\epsilon U} \frac{1}{\pi_{EI,i}\pi_{EC,i}} y_i I_{EI,i}(1 - r_{EI,i}) I_{EC,i}\right)$$

$$V_1 E_2 E_3(\hat{t}_{y,2phases,nr}) = V_1 E_2\left(\sum_{i\epsilon U} \frac{1}{\pi_{EI,i}} y_i I_{EI,i} R_{EI,i} + \sum_{i\epsilon U} \frac{1}{\pi_{EI,i}\pi_{EC,i}} y_i I_{EI,i}(1 - R_{EI,i}) E_3(I_{EC,i})\right)$$

$$V_1 E_2 E_3(\hat{t}_{y,2phases,nr}) = V_1 E_2\left(\sum_{i\epsilon U} \frac{1}{\pi_{EI,i}} y_i I_{EI,i} R_{EI,i} + \sum_{i\epsilon U} \frac{1}{\pi_{EI,i}\pi_{EC,i}} y_i I_{EI,i}(1 - R_{EI,i}) \pi_{EC,i}\right)$$

$$V_1 E_2 E_3(\hat{t}_{y,2phases,nr}) = V_1 E_2\left(\sum_{i\epsilon U} \frac{1}{\pi_{EI,i}} y_i I_{EI,i} R_{EI,i} + \sum_{i\epsilon U} \frac{1}{\pi_{EI,i}} y_i I_{EI,i}(1 - R_{EI,i})\right)$$

$$V_1 E_2 E_3(\hat{t}_{y,2phases,nr}) = V_1 E_2\left(\sum_{i\epsilon U} \frac{1}{\pi_{EI,i}} y_i I_{EI,i}(R_{EI,i} + (1 - R_{EI,i}))\right)$$

$$V_1 E_2 E_3(\hat{t}_{y,2phases,nr}) = V_1 E_2\left(\sum_{i\epsilon U} \frac{1}{\pi_{EI,i}} y_i I_{EI,i}\right)$$

$$V_1 E_2 E_3(\hat{t}_{y,2phases,nr}) = V_1\left(\sum_{i\epsilon U} \frac{1}{\pi_{EI,i}} y_i I_{EI,i}\right)$$

(variance de l'estimateur de Horvitz-Thompson)

$$V_1 E_2 E_3(\hat{t}_{y,2phases,nr}) = \sum_{i\epsilon U}\sum_{i\epsilon U} (\pi_{EI,ij} - \pi_{EI,i}\pi_{EI,j}) \frac{y_i}{\pi_{EI,i}} \frac{y_j}{\pi_{EI,j}}$$

$$\rightarrow V(\hat{t}_{y,2phases,nr}) = V_1 E_2 E_3(\hat{t}_{y,2phases,nr}) + E_1 V_2 E_3(\hat{t}_{y,2phases,nr}) + E_1 E_2 V_3(\hat{t}_{y,2phases,nr})$$

$$V\left(\hat{t}_{y,2phases,nr}\right)$$

$$= \sum_{i\epsilon U}\sum_{i\epsilon U}(\pi_{EI,ij} - \pi_{EI,i}\pi_{EI,j})\frac{y_i}{\pi_{EI,i}}\frac{y_j}{\pi_{EI,j}} + 0$$

$$+ E_1E_2\left(\sum_{i\epsilon U}\sum_{i\epsilon U}(\pi_{EC,ij} - \pi_{EC,i}\pi_j)\frac{\phi_i}{\pi_{EC,i}}\frac{\phi_j}{\pi_{EC,j}}\right)$$

$$V\left(\hat{t}_{y,2phases,nr}\right)$$

$$= \sum_{i\epsilon U}\sum_{i\epsilon U}(\pi_{EI,ij} - \pi_{EI,i}\pi_{EI,j})\frac{y_i}{\pi_{EI,i}}\frac{y_j}{\pi_{EI,j}}$$

$$+ E_1E_2\left(\sum_{i\epsilon U}\sum_{i\epsilon U}(\pi_{EC,ij} - \pi_{EC,i}\pi_{EC,j})\frac{\phi_i}{\pi_{EC,i}}\frac{\phi_j}{\pi_{EC,j}}\right)$$

$$avec\ \phi_i = \frac{1}{\pi_{EI,i}}y_iI_{EI,i}(1 - R_{EI,i})$$

Notons pour la suite :

$$V_1\left(\hat{t}_{y,2phases,nr}\right) = \sum_{i\epsilon U}\sum_{i\epsilon U}(\pi_{EI,ij} - \pi_{EI,i}\pi_{EI,j})\frac{y_i}{\pi_{EI,i}}\frac{y_j}{\pi_{EI,j}}$$

$$V_2\left(\hat{t}_{y,2phases,nr}\right) = E_1E_2\left(\sum_{i\epsilon U}\sum_{i\epsilon U}(\pi_{EC,ij} - \pi_{EC,i}\pi_{EC,j})\frac{\phi_i}{\pi_{EC,i}}\frac{\phi_j}{\pi_{EC,j}}\right)$$

2. Estimateur sans biais de la variance de $\hat{t}_{y,2phases,nr}$

$\hat{V}\left(\hat{t}_{y,2phases,nr}\right)$

$$= \sum_{i\epsilon s_{EI,r}} \sum_{j\epsilon s_{EI,r}} \frac{\Delta_{EI,ij}}{\pi_{EI,ij}} y_i y_j + \sum_{i\epsilon s_{EI}} \sum_{j\epsilon s_{EC}} \frac{\Delta_{EI,ij}}{\pi_{EI,ij}} \frac{1}{\pi_{EC,j}} y_i y_j$$

$$+ \sum_{i\epsilon s_{EC}} \sum_{j\epsilon s_{EI}} \frac{\Delta_{EI,ij}}{\pi_{EI,ij}} \frac{1}{\pi_{EC,i}} y_i y_j + \sum_{i\epsilon s_{EC}} \sum_{j\epsilon s_{EC}} \frac{\Delta_{EI,ij}}{\pi_{EI,ij}} \frac{1}{\pi_{EC,ij}} y_i y_j$$

$$+ \sum_{i\epsilon s_{EC}} \sum_{j\epsilon s_{EC}} \frac{\Delta_{EC,ij}}{\pi_{EI,i}\pi_{EI,j}\pi_{EC,ij}} y_i y_j$$

avec $\Delta_{EI,ij} = \frac{\pi_{EI,ij} - \pi_{EI,i}\pi_{EI,j}}{\pi_{EI,i}\pi_{EI,j}}$ et $\Delta_{EC,ij} = \frac{\pi_{EC,ij} - \pi_{EC,i}\pi_{EC,j}}{\pi_{EC,i}\pi_{EC,j}}$

Démonstration :

On va utiliser la propriété suivante : $E(Y) = E(E(Y/X))$

$E(\hat{V}\left(\hat{t}_{y,2phases,nr}\right)) = E_1E_2E_3(\hat{V}\left(\hat{t}_{y,2phases,nr}\right))$

$E\left(\hat{V}\left(\hat{t}_{y,2phases,nr}\right)\right) = E_1E_2E_3\left(\hat{V}_1\left(\hat{t}_{y,2phases,nr}\right)\right) + E_1E_2E_3(\hat{V}_2\left(\hat{t}_{y,2phases,nr}\right))$

$\boldsymbol{E_1E_2E_3\left(\hat{V}_1\left(\hat{t}_{y,2phases,nr}\right)\right)}$

$$= E_1E_2E_3\left(\sum_{i\epsilon s_{EI,r}} \sum_{j\epsilon s_{EI,r}} \frac{\Delta_{EI,ij}}{\pi_{EI,ij}} y_i y_j + \sum_{i\epsilon s_{EI}} \sum_{j\epsilon s_{EC}} \frac{\Delta_{EI,ij}}{\pi_{EI,ij}} \frac{1}{\pi_{EC,j}} y_i y_j \right.$$

$$\left. + \sum_{i\epsilon s_{EC}} \sum_{j\epsilon s_{EI}} \frac{\Delta_{EI,ij}}{\pi_{EI,ij}} \frac{1}{\pi_{EC,i}} y_i y_j + \sum_{i\epsilon s_{EC}} \sum_{j\epsilon s_{EC}} \frac{\Delta_{EI,ij}}{\pi_{EI,ij}} \frac{1}{\pi_{EC,ij}} y_i y_j \right)$$

$$E_1E_2E_3\left(\hat{V}_1\left(\hat{t}_{y,2phases,nr}\right)\right)$$

$$= E_1E_2E_3\left(\sum_{i\epsilon U}\sum_{j\epsilon U}\frac{\Delta_{EI,ij}}{\pi_{EI,ij}}y_iy_jI_{EI,i}I_{EI,j}R_{EI,i}R_{EI,j}\right.$$

$$+\sum_{i\epsilon U}\sum_{j\epsilon U}\frac{\Delta_{EI,ij}}{\pi_{EI,ij}}\frac{1}{\pi_{EC,j}}y_iy_jI_{EI,i}I_{EI,j}R_{EI,i}(1-R_{EI,j})I_{EC,j}$$

$$+\sum_{i\epsilon U}\sum_{j\epsilon U}\frac{\Delta_{EI,ij}}{\pi_{EI,ij}}\frac{1}{\pi_{EC,i}}y_iy_jI_{EI,i}I_{EI,j}R_{EI,j}(1-R_{EI,i})I_{EC,i}$$

$$\left.+\sum_{i\epsilon U}\sum_{j\epsilon U}\frac{\Delta_{EI,ij}}{\pi_{EI,ij}}\frac{1}{\pi_{EC,ij}}y_iy_jI_{EI,i}I_{EI,j}(1-R_{EI,i})(1-R_{EI,j})I_{EC,i}I_{EC,j}\right)$$

$$E_1E_2E_3\left(\hat{V}_1\left(\hat{t}_{y,2phases,nr}\right)\right)$$

$$= E_1E_2\left(\sum_{i\epsilon U}\sum_{j\epsilon U}\frac{\Delta_{EI,ij}}{\pi_{EI,ij}}y_iy_jI_{EI,i}I_{EI,j}R_{EI,i}R_{EI,j}\right.$$

$$+\sum_{i\epsilon U}\sum_{j\epsilon U}\frac{\Delta_{EI,ij}}{\pi_{EI,ij}}\frac{1}{\pi_{EC,j}}y_iy_jI_{EI,i}I_{EI,j}R_{EI,i}(1-R_{EI,j})E_3(I_{EC,j})$$

$$+\sum_{i\epsilon U}\sum_{j\epsilon U}\frac{\Delta_{EI,ij}}{\pi_{EI,ij}}\frac{1}{\pi_{EC,i}}y_iy_jI_{EI,i}I_{EI,j}R_{EI,j}(1-R_{EI,i})E_3(I_{EC,i})$$

$$\left.+\sum_{i\epsilon U}\sum_{j\epsilon U}\frac{\Delta_{EI,ij}}{\pi_{EI,ij}}\frac{1}{\pi_{EC,ij}}y_iy_jI_{EI,i}I_{EI,j}(1-R_{EI,i})(1-R_{EI,j})E_3(I_{EC,i}I_{EC,j})\right)$$

$$E_1E_2E_3\left(\hat{V}_1\left(\hat{t}_{y,2phases,nr}\right)\right)$$

$$= E_1E_2\left(\sum_{i\epsilon U}\sum_{j\epsilon U}\frac{\Delta_{EI,ij}}{\pi_{EI,ij}}y_iy_jI_{EI,i}I_{EI,j}R_{EI,i}R_{EI,j}\right.$$

$$+\sum_{i\epsilon U}\sum_{j\epsilon U}\frac{\Delta_{EI,ij}}{\pi_{EI,ij}}\frac{1}{\pi_{EC,j}}y_iy_jI_{EI,i}I_{EI,j}R_{EI,i}(1-R_{EI,j})\pi_{EC,j}$$

$$+\sum_{i\epsilon U}\sum_{j\epsilon U}\frac{\Delta_{EI,ij}}{\pi_{EI,ij}}\frac{1}{\pi_{EC,i}}y_iy_jI_{EI,i}I_{EI,j}R_{EI,j}(1-R_{EI,i})\pi_{EC,i}$$

$$\left.+\sum_{i\epsilon U}\sum_{j\epsilon U}\frac{\Delta_{EI,ij}}{\pi_{EI,ij}}\frac{1}{\pi_{EC,ij}}y_iy_jI_{EI,i}I_{EI,j}(1-R_{EI,i})(1-R_{EI,j})\pi_{EC,ij}\right)$$

$$E_1E_2E_3\left(\hat{V}_1\left(\hat{t}_{y,2phases,nr}\right)\right)$$

$$= E_1E_2\left(\sum_{i\epsilon U}\sum_{j\epsilon U}\frac{\Delta_{EI,ij}}{\pi_{EI,ij}}y_iy_jI_{EI,i}I_{EI,j}R_{EI,i}R_{EI,j}\right.$$

$$+\sum_{i\epsilon U}\sum_{j\epsilon U}\frac{\Delta_{EI,ij}}{\pi_{EI,ij}}y_iy_jI_{EI,i}I_{EI,j}R_{EI,i}(1-R_{EI,j})+\sum_{i\epsilon U}\sum_{j\epsilon U}\frac{\Delta_{EI,ij}}{\pi_{EI,ij}}y_iy_jI_{EI,i}I_{EI,j}R_{EI,j}(1$$

$$\left.-R_{EI,i})+\sum_{i\epsilon U}\sum_{j\epsilon U}\frac{\Delta_{EI,ij}}{\pi_{EI,ij}}y_iy_jI_{EI,i}I_{EI,j}(1-R_{EI,i})(1-R_{EI,j})\right)$$

$$E_1E_2E_3\left(\hat{V}_1\left(\hat{t}_{y,2phases,nr}\right)\right)$$

$$= E_1E_2\left(\sum_{i\epsilon U}\sum_{j\epsilon U}\frac{\Delta_{EI,ij}}{\pi_{EI,ij}}y_iy_jI_{EI,i}I_{EI,j}(R_{EI,i}R_{EI,j}+R_{EI,i}\left(1-R_{EI,j}\right)\right.$$

$$\left.+R_{EI,j}\left(1-R_{EI,i}\right)+\left(1-R_{EI,i}\right)\left(1-R_{EI,j}\right)\right)\Bigg)$$

$$E_1E_2E_3\left(\hat{V}_1\left(\hat{t}_{y,2phases,nr}\right)\right)$$

$$= E_1E_2\left(\sum_{i\epsilon U}\sum_{j\epsilon U}\frac{\Delta_{EI,ij}}{\pi_{EI,ij}}y_iy_jI_{EI,i}I_{EI,j}(R_{EI,i}R_{EI,j} + R_{EI,i} - R_{EI,i}R_{EI,j} + R_{EI,j}\right.$$

$$\left. - R_{EI,j}R_{EI,i} + 1 - R_{EI,i} - R_{EI,j} + R_{EI,i}R_{EI,j})\right)$$

$$E_1E_2E_3\left(\hat{V}_1\left(\hat{t}_{y,2phases,nr}\right)\right) = E_1E_2\left(\sum_{i\epsilon U}\sum_{j\epsilon U}\frac{\Delta_{EI,ij}}{\pi_{EI,ij}}y_iy_jI_{EI,i}I_{EI,j}\right)$$

$$E_1E_2E_3\left(\hat{V}_1\left(\hat{t}_{y,2phases,nr}\right)\right) = E_1\left(\sum_{i\epsilon U}\sum_{j\epsilon U}\frac{\Delta_{EI,ij}}{\pi_{EI,ij}}y_iy_jI_{EI,i}I_{EI,j}\right)$$

$$E_1E_2E_3\left(\hat{V}_1\left(\hat{t}_{y,2phases,nr}\right)\right) = \sum_{i\epsilon U}\sum_{j\epsilon U}\frac{\Delta_{EI,ij}}{\pi_{EI,ij}}y_iy_jE_1\left(I_{EI,i}I_{EI,j}\right)$$

$$E_1E_2E_3\left(\hat{V}_1\left(\hat{t}_{y,2phases,nr}\right)\right) = \sum_{i\epsilon U}\sum_{j\epsilon U}\frac{\Delta_{EI,ij}}{\pi_{EI,ij}}y_iy_j\pi_{EI,ij}$$

$$E_1E_2E_3\left(\hat{V}_1\left(\hat{t}_{y,2phases,nr}\right)\right) = \sum_{i\epsilon U}\sum_{j\epsilon U}\Delta_{EI,ij}y_iy_j$$

$$E_1E_2E_3\left(\hat{V}_1\left(\hat{t}_{y,2phases,nr}\right)\right) = \sum_{i\epsilon U}\sum_{j\epsilon U}(\pi_{EI,ij} - \pi_{EI,i}\pi_{EI,j})\frac{y_i}{\pi_{EI,i}}\frac{y_j}{\pi_{EI,j}}$$

$$E_1E_2E_3\left(\hat{V}_1\left(\hat{t}_{y,2phases,nr}\right)\right) = V_1\left(\hat{t}_{y,2phases,nr}\right)$$

$$\boldsymbol{E_1E_2E_3}\left(\widehat{\boldsymbol{V}}_{\boldsymbol{2}}\left(\hat{t}_{y,2phases,nr}\right)\right) = E_1E_2E_3\left(\sum_{i\epsilon s_{EC}}\sum_{j\epsilon s_{EC}}\frac{\pi_{EC,ij}-\pi_{EC,i}\pi_{EC,j}}{\pi_{EC,i}\pi_{EC,j}}\frac{1}{\pi_{EI,i}\pi_{EI,j}\pi_{EC,ij}}y_iy_j\right)$$

$$E_1E_2E_3\left(\widehat{V}_2\left(\hat{t}_{y,2phases,nr}\right)\right)$$

$$= E_1E_2E_3\left(\sum_{i\epsilon U}\sum_{j\epsilon U}\frac{\pi_{EC,ij}-\pi_{EC,i}\pi_{EC,j}}{\pi_{EC,i}\pi_{EC,j}}\frac{1}{\pi_{EI,i}\pi_{EI,j}\pi_{EC,ij}}y_iy_jI_{EI,i}I_{EI,j}(1\right.$$

$$\left.- R_{EI,i})(1-R_{EI,j})I_{EC,i}I_{EC,j}\right)$$

$$E_1E_2E_3\left(\widehat{V}_2\left(\hat{t}_{y,2phases,nr}\right)\right)$$

$$= E_1E_2\left(\sum_{i\epsilon U}\sum_{j\epsilon U}\frac{\pi_{EC,ij}-\pi_{EC,i}\pi_{EC,j}}{\pi_{EC,i}\pi_{EC,j}}\frac{1}{\pi_{EI,i}\pi_{EI,j}\pi_{EC,ij}}y_iy_jI_{EI,i}I_{EI,j}(1-R_{EI,i})(1\right.$$

$$\left.- R_{EI,j})E_3(I_{EC,i}I_{EC,j})\right)$$

$$E_1E_2E_3\left(\widehat{V}_2\left(\hat{t}_{y,2phases,nr}\right)\right)$$

$$= E_1E_2\left(\sum_{i\epsilon U}\sum_{j\epsilon U}\frac{\pi_{EC,ij}-\pi_{EC,i}\pi_{EC,j}}{\pi_{EC,i}\pi_{EC,j}}\frac{1}{\pi_{EI,i}\pi_{EI,j}\pi_{EC,ij}}y_iy_jI_{EI,i}I_{EI,j}(1-R_{EI,i})(1\right.$$

$$\left.- R_{EI,j})\pi_{EC,ij}\right)$$

$$E_1E_2E_3\left(\hat{V}_2\left(\hat{t}_{y,2phases,nr}\right)\right)$$

$$= E_1E_2\left(\sum_{i\epsilon U}\sum_{j\epsilon U}\frac{\pi_{EC,ij}-\pi_{EC,i}\pi_{EC,j}}{\pi_{EC,i}\pi_{EC,j}}\frac{1}{\pi_{EI,i}\pi_{EI,j}}y_iy_jI_{EI,i}I_{EI,j}(1-R_{EI,i})(1\right.$$

$$\left.-R_{EI,j})\right)$$

$$E_1E_2E_3\left(\hat{V}_2\left(\hat{t}_{y,2phases,nr}\right)\right)$$

$$= E_1E_2\left(\sum_{i\epsilon U}\sum_{j\epsilon U}(\pi_{EC,ij}-\pi_{EC,i}\pi_{EC,j})\frac{y_iI_{EI,i}(1-R_{EI,i})}{\pi_{EI,i}\pi_{EC,i}}\frac{y_jI_{EI,j}(1-R_{EI,j})}{\pi_{EI,j}\pi_{EC,j}}\right)$$

$$E_1E_2E_3\left(\hat{V}_2\left(\hat{t}_{y,2phases,nr}\right)\right) = E_1E_2\left(\sum_{i\epsilon U}\sum_{i\epsilon U}(\pi_{EC,ij}-\pi_{EC,i}\pi_{EC,j})\frac{\phi_i}{\pi_{EC,i}}\frac{\phi_j}{\pi_{EC,j}}\right)$$

$$avec\ \phi_i = \frac{1}{\pi_{EI,i}}\mathrm{y_i}I_{EI,i}(1-R_{EI,i})$$

$$E_1E_2E_3\left(\hat{V}_2\left(\hat{t}_{y,2phases,nr}\right)\right) = V_2\left(\hat{t}_{y,2phases,nr}\right)$$

$$\rightarrow E\left(\hat{V}\left(\hat{t}_{y,2phases,nr}\right)\right) = E_1E_2E_3\left(\hat{V}_1\left(\hat{t}_{y,2phases,nr}\right)\right) + E_1E_2E_3\left(\hat{V}_2\left(\hat{t}_{y,2phases,nr}\right)\right) =$$

$$V_1\left(\hat{t}_{y,2phases,nr}\right) + V_2\left(\hat{t}_{y,2phases,nr}\right) = V\left(\hat{t}_{y,2phases,nr}\right)$$

→ $\hat{V}\left(\hat{t}_{y,2phases,nr}\right)$ estimateur sans biais de $V\left(\hat{t}_{y,2phases,nr}\right)$

I.8 ESTIMATEUR D'UN TOTAL POUR UN PLAN DE SONDAGE EN DEUX PHASES POUR NON-RÉPONSE AVEC NON-RÉPONSE

On estime t_y par $\hat{t}_{y,2phases,nr,pond}$ tel que :

$$\hat{t}_{y,2phases,nr,pond} = \sum_{i \in s_{EI,r}} d_{EI,i} y_i + \sum_{i \in s_{EC,r}} \frac{d_{EI,i}}{\pi_{EI,i}} \times \frac{1}{\delta_{EC,i}} y_i$$

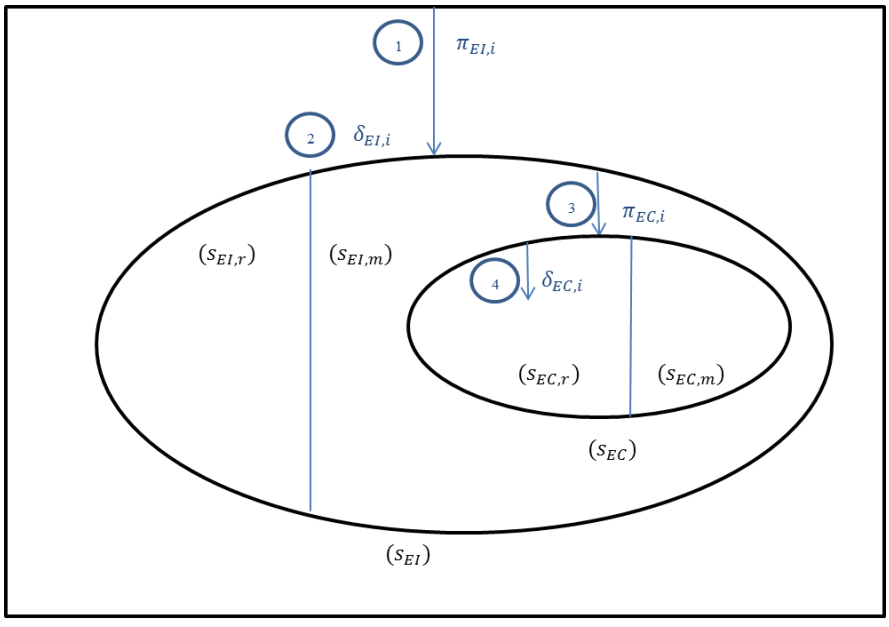
Autrement dit,

$$\hat{t}_{y,2phases,nr,pond} = \sum_{i \in U} d_{EI,i} y_i I_{EI,i} R_{EI,i} + \sum_{i \in U} \frac{d_{EI,i}}{\pi_{EC,i}} \times \frac{1}{\delta_{EC,i}} y_i I_{EI,i} (1 - R_{EI,i}) I_{EC,i} R_{EC,i}$$

Démonstration :

$\hat{t}_{y,2phases,nr,pond}$ est-il un estimateur sans biais de t_y ?

Figure I-3 : Représentation graphique des sources de variabilité dans un sondage en deux phases pour non-réponse avec non-réponse



Soit :

$E_4 = E(./I_{EI}, R_{EI}, I_{EC})$

$E_3 = E(./I_{EI}, R_{EI})$

$E_2 = E(./I_{EI})$

$E_1 = E_p(.)$, espérance liée au plan de première phase

On a :

$E_4(R_{EC}) = E(R_{EC}/I_{EI}, R_{EI}, I_{EC})$

R_{EC} variable aléatoire qui suit une loi de Bernoulli

➔ $E_4(R_{EC}) = E(R_{EC}/I_{EI}, R_{EI}, I_{EC}) = \delta_{EC}$

$E_3(I_{EC}) = E(I_{EC}/I_{EI}, R_{EI}) = \pi_{EC}$

$E_2(R_{EI}) = E(R_{EI}/I_{EI}) = \delta_{EI}$

$E_1(I_{EI}) = \pi_{EI}$

On va montrer que $\hat{t}_{y,2phases,nr,pond}$ est un estimateur sans biais de t_y *en utilisant la propriété suivante :*

$$E(Y) = E(E(Y/X))$$

Ainsi, on a $E(\hat{t}_{y,2phases,nr,pond}) = E_1E_2E_3E_4(\hat{t}_{y,2phases,nr,pond})$

$\boldsymbol{E_4}(\hat{t}_{y,2phases,nr,pond})$

$$= \sum_{i\in U} d_{EI,i} y_i I_{EI,i} R_{EI,i} + \sum_{i\in U} \frac{d_{EI,i}}{\pi_{EC,i}} \times \frac{1}{\delta_{EC,i}} y_i I_{EI,i}(1 - R_{EI,i}) I_{EC,i} E_4(R_{EC,i})$$

$$E_4\left(\hat{t}_{y,2phases,nr,pond}\right) = \sum_{i\in U} d_{EI,i} y_i I_{EI,i} R_{EI,i} + \sum_{i\in U} \frac{d_{EI,i}}{\pi_{EC,i}} \times \frac{1}{\delta_{EC,i}} y_i I_{EI,i} (1 - R_{EI,i}) I_{EC,i} \delta_{EC,i}$$

$$E_4\left(\hat{t}_{y,2phases,nr,pond}\right) = \sum_{i\in U} d_{EI,i} y_i I_{EI,i} R_{EI,i} + \sum_{i\in U} \frac{d_{EI,i}}{\pi_{EC,i}} y_i I_{EI,i} (1 - R_{EI,i}) I_{EC,i}$$

$$\boldsymbol{E_3 E_4}\left(\hat{t}_{y,2phases,nr,pond}\right) = \sum_{i\in U} d_{EI,i} y_i I_{EI,i} R_{EI,i} + \sum_{i\in U} \frac{d_{EI,i}}{\pi_{EC,i}} y_i I_{EI,i} (1 - R_{EI,i}) E_3(I_{EC,i})$$

$$E_3 E_4\left(\hat{t}_{y,2phases,nr,pond}\right) = \sum_{i\in U} d_{EI,i} y_i I_{EI,i} R_{EI,i} + \sum_{i\in U} \frac{d_{EI,i}}{\pi_{EC,i}} y_i I_{EI,i} (1 - R_{EI,i}) \pi_{EC,i}$$

$$E_3 E_4\left(\hat{t}_{y,2phases,nr,pond}\right) = \sum_{i\in U} d_{EI,i} y_i I_{EI,i} R_{EI,i} + \sum_{i\in U} d_{EI,i} y_i I_{EI,i} (1 - R_{EI,i})$$

$$E_3 E_4\left(\hat{t}_{y,2phases,nr,pond}\right) = \sum_{i\in U} d_{EI,i} y_i I_{EI,i} R_{EI,i} - \sum_{i\in U} d_{EI,i} y_i I_{EI,i} R_{EI,i} + \sum_{i\in U} d_{EI,i} y_i I_{EI,i}$$

$$E_3 E_4\left(\hat{t}_{y,2phases,nr,pond}\right) = \sum_{i\in U} d_{EI,i} y_i I_{EI,i}$$

$$\boldsymbol{E_2 E_3 E_4}\left(\hat{t}_{y,2phases,nr,pond}\right) = \sum_{i\in U} d_{EI,i} y_i I_{EI,i}$$

(les $R_{EI,i}$ ont disparu à l'étape précédente → plus de variable aléatoire)

$$\boldsymbol{E_1 E_2 E_3 E_4}\left(\hat{t}_{y,2phases,nr,pond}\right) = \sum_{i\in U} d_{EI,i} y_i E_1(I_{EI,i})$$

$$E_1 E_2 E_3 E_4\left(\hat{t}_{y,2phases,nr,pond}\right) = \sum_{i\in U} d_{EI,i} y_i \pi_{EI,i}$$

$$E_1 E_2 E_3 E_4\left(\hat{t}_{y,2phases,nr,pond}\right) = \sum_{i\in U} y_i$$

$E_1 E_2 E_3 E_4 \left(\hat{t}_{y,2phases,nr,pond}\right) = t_y$

→ $\hat{t}_{y,2phases,nr,pond}$ est un estimateur sans biais de t_y

I.9 ESTIMATEUR DE LA VARIANCE D'UN TOTAL POUR UN PLAN DE SONDAGE EN DEUX PHASES POUR NON-RÉPONSE AVEC NON-RÉPONSE

1. Quelle est la variance de $\hat{t}_{y,2phases,nr,pond}$?

D'où vient la variance ? De la différence entre la vraie valeur du total et de son estimateur qui est due au plan de sondage et à la non-réponse (représentés en bleu et numérotés de 1 à 4 dans la représentation graphique).

$$\hat{t}_{y,2phases,nr,pond} - Y_t$$
$$= \left(\hat{t}_{y,\pi} - t_y\right) + \left(\hat{t}_{y,pond} - \hat{t}_{y,\pi}\right) + \left(\hat{t}_{y,2phases,nr} - \hat{t}_{y,pond}\right)$$
$$+ \left(\hat{t}_{y,2phases,nr,pond} - \hat{t}_{y,2phases,nr}\right)$$

Pour calculer la variance de $\hat{t}_{y,2phases,nr,pond}$, on va utiliser les propriétés suivantes :

$$E(Y) = E(E(Y/X))$$

$$V(Y) = V(E(Y/X)) + E(V(Y/X))$$

Ainsi, on a :

$$V\left(\hat{t}_{y,2phases,nr,pond}\right)$$
$$= V_1E_2E_3E_4\left(\hat{t}_{y,2phases,nr,pond}\right) + E_1V_2E_3E_4\left(\hat{t}_{y,2phases,nr,pond}\right)$$
$$+ E_1E_2V_3E_4\left(\hat{t}_{y,2phases,nr,pond}\right) + E_1E_2E_3V_4\left(\hat{t}_{y,2phases,nr,pond}\right)$$

Démonstration :

$$V\left(\hat{t}_{y,2phases,nr,pond}\right) = V\left(E\left(\hat{t}_{y,2phases,nr,pond}/I_{EI}\right)\right) + E\left(V\left(\hat{t}_{y,2phases,nr,pond}/I_{EI}\right)\right)$$

$$V\left(\hat{t}_{y,2phases,nr,pond}\right)$$

$$= V\left(EEE\left(\hat{t}_{y,2phases,nr,pond}/I_{EI}, R_{EI}, I_{EC}\right)\right) + EVE\left(\hat{t}_{y,2phases,nr,pond}/I_{EI}, R_{EI}\right)$$

$$+ EEV\left(\hat{t}_{y,2phases,nr,pond}/I_{EI}, R_{EI}\right)$$

$$V\left(\hat{t}_{y,2phases,nr,pond}\right)$$

$$= V\left(EEE\left(\hat{t}_{y,2phases,nr,pond}/I_{EI}, R_{EI}, I_{EC}\right)\right)$$

$$+ EVEE\left(\hat{t}_{y,2phases,nr,pond}/I_{EI}, R_{EI}, I_{EC}\right)$$

$$+ EEVE\left(\hat{t}_{y,2phases,nr,pond}/I_{EI}, R_{EI}, I_{EC}\right)$$

$$+ EEEV\left(\hat{t}_{y,2phases,nr,pond}/I_{EI}, R_{EI}, I_{EC}\right)$$

$$V\left(\hat{t}_{y,2phases,nr,pond}\right)$$

$$= V_1E_2E_3E_4\left(\hat{t}_{y,2phases,nr,pond}\right) + E_1V_2E_3E_4\left(\hat{t}_{y,2phases,nr,pond}\right)$$

$$+ E_1E_2V_3E_4\left(\hat{t}_{y,2phases,nr,pond}\right) + E_1E_2E_3V_4\left(\hat{t}_{y,2phases,nr,pond}\right)$$

- $\boldsymbol{V_1E_2E_3E_4}\left(\hat{t}_{y,2phases,nr,pond}\right)$ ***(V_1 : variance liée à l'échantillonnage de première phase)***

$\boldsymbol{V_1E_2E_3E_4}\left(\hat{t}_{y,2phases,nr,pond}\right) = V_1\left(\sum_{i\in U} d_{EI,i}\, y_i I_{EI,i}\right)$

$$V_1E_2E_3E_4\left(\hat{t}_{y,2phases,nr,pond}\right) = \sum_{i\in U}\sum_{j\in U}\left(\pi_{EI,ij} - \pi_{EI,i}\pi_{EI,j}\right)\frac{y_i}{\pi_{EI,i}}\frac{y_j}{\pi_{EI,j}}$$

Rq : On retombe sur la variance de l'estimateur de Horvitz-Thompson

- $\boldsymbol{E_1V_2E_3E_4}\left(\hat{t}_{y,2phases,nr,pond}\right)$ ***(V_2 : variance liée à la non-réponse de première phase)***

$\boldsymbol{E_1 V_2 E_3 E_4}\left(\hat{t}_{y,2phases,nr,pond}\right) = E_1 V_2 (\sum_{i\in U} d_{EI,i}\, y_i I_{EI,i})$

Au niveau 2, I_{EI} est fixé et R_{EI} est absent

$E_1 V_2 E_3 E_4\left(\hat{t}_{y,2phases,nr,pond}\right) = E_1(0)$

$E_1 V_2 E_3 E_4\left(\hat{t}_{y,2phases,nr,pond}\right) = 0$

- $\boldsymbol{E_1 E_2 V_3 E_4}\left(\hat{t}_{y,2phases,nr,pond}\right)$ ***(V_3 : variance liée à l'échantillonnage de deuxième phase)***

$\boldsymbol{E_1 E_2 V_3 E_4}\left(\hat{t}_{y,2phases,nr,pond}\right) = E_1 E_2 V_3 (\sum_{i\in U} d_{EI,i}\, y_i I_{EI,i} R_{EI,i} + \sum_{i\in U} \frac{d_{EI,i}}{\pi_{EC,i}} y_i I_{EI,i}(1 - R_{EI,i}) I_{EC,i})$

Au niveau 3, I_{EI} et R_{EI} sont fixés donc $V_3\left(\sum_{i\in U} d_{EI,i}\, y_i I_{EI,i} R_{EI,i}\right) = 0$

$$E_1 E_2 V_3 E_4\left(\hat{t}_{y,2phases,nr,pond}\right) = E_1 E_2 V_3 \left(\sum_{i\in U} \frac{d_{EI,i}}{\pi_{EC,i}} y_i I_{EI,i}(1 - R_{EI,i}) I_{EC,i}\right)$$

On cherche une transformation qui nous ramène à l'estimateur de la variance de Horvitz-Thompson.

Soit $\phi_i = d_{EI,i} y_i I_{EI,i}(1 - R_{EI,i})$ *Rq : tout est fixe*

$$E_1 E_2 V_3 E_4\left(\hat{t}_{y,2phases,nr,pond}\right) = E_1 E_2 V_3 \left(\sum_{i\in U} \frac{1}{\pi_{EC,i}} \phi_i I_{EC,i}\right)$$

$$E_1 E_2 V_3 E_4\left(\hat{t}_{y,2phases,nr,pond}\right) = E_1 E_2 \sum_{i\in U}\sum_{j\in U} (\pi_{EC,ij} - \pi_{EC,i}\pi_{EC,j}) \frac{\phi_i}{\pi_{EC,i}} \frac{\phi_j}{\pi_{EC,j}}$$

- $\boldsymbol{E_1E_2E_3V_4}\left(\hat{t}_{y,2phases,nr,pond}\right)$ ***(V_4 : variance liée à la non-réponse de deuxième phase)***

$\boldsymbol{E_1E_2E_3V_4}\left(\hat{t}_{y,2phases,nr,pond}\right)$

$$= E_1E_2E_3V_4\left(\sum_{i\in U} d_{EI,i}\, y_i I_{EI,i} R_{EI,i} + \sum_{i\in U} \frac{d_{EI,i}}{\pi_{EC,i}} \frac{1}{\delta_{EC,i}} y_i I_{EI,i}(1-R_{EI,i}) I_{EC,i} R_{EC,i}\right)$$

Au niveau 4, I_{EI} et R_{EI} et sont fixés donc $V_4(\sum_{i\in U} d_{EI,i}\, y_i I_{EI,i} R_{EI,i})$=0

$$E_1E_2E_3V_4\left(\hat{t}_{y,2phases,nr,pond}\right) = E_1E_2E_3V_4\left(\sum_{i\in U} \frac{d_{EI,i}}{\pi_{EC,i}} \frac{1}{\delta_{EC,i}} y_i I_{EI,i}(1-R_{EI,i}) I_{EC,i} R_{EC,i}\right)$$

Au niveau 4, I_{EI}, R_{EI} et I_{EC} sont fixés donc $\frac{d_{EI,i}}{\pi_{EC,i}} \frac{1}{\delta_{EC,i}} y_i I_{EI,i}(1-R_{EI,i}) I_{EC,i}$ est une constante

Or, $Var(\sum_{i=1}^{N} aX_i) = \sum_{i=1}^{N} a^2 Var(X_i)$ *avec a constante.*

Donc :

$$E_1E_2E_3V_4\left(\hat{t}_{y,2phases,nr,pond}\right) = E_1E_2E_3 \left(\sum_{i\in U} \frac{d_{EIi}^2}{\pi_{EC,i}^2} \frac{1}{\delta_{EC,i}^2} y_i^2 I_{EI,i}^2 \left(1-R_{EI,i}\right)^2 I_{EC,i}^2 V_4\left(R_{EC,i}\right)\right)$$

$$E_1E_2E_3V_4\left(\hat{t}_{y,2phases,nr,pond}\right) = E_1E_2E_3 \left(\sum_{i\in U} \frac{d_{EIi}^2}{\pi_{EC,i}^2} \frac{1}{\delta_{EC,i}^2} y_i^2 I_{EI,i}\left(1-R_{EI,i}\right) \quad I_{EC,i} V_4\left(R_{EC,i}\right)\right)$$

Or, on suppose que le processus de non-réponse R_{EC} suit une loi de Bernoulli. Donc :

$$E_1E_2E_3V_4\left(\hat{t}_{y,2phases,nr,pond}\right) = E_1E_2E_3 \left(\sum_{i\in U} \frac{d_{EIi}^2}{\pi_{EC,i}^2} \frac{1}{\delta_{EC,i}^2} y_i^2 I_{EI,i}\left(1-R_{EI,i}\right) \quad I_{EC,i}\delta_{EC,i}(1-\delta_{EC,i})\right)$$

$$E_1E_2E_3V_4\left(\hat{t}_{y,2phases,nr,pond}\right) = E_1E_2E_3 \left(\sum_{i\in U} \frac{d_{EIi}^2}{\pi_{EC,i}^2} \frac{(1-\delta_{EC,i})}{\delta_{EC,i}} y_i^2 I_{EI,i}\left(1-R_{EI,i}\right) \quad I_{EC,i}\right)$$

- **Retour à $\boldsymbol{V}\left(\hat{t}_{y,2phases,nr,pond}\right)$**

$$V\left(\hat{t}_{y,2phases,nr,pond}\right) = V_1E_2E_3E_4\left(\hat{t}_{y,2phases,nr,pond}\right) + E_1V_2E_3E_4\left(\hat{t}_{y,2phases,nr,pond}\right) + E_1E_2V_3E_4\left(\hat{t}_{y,2phases,nr,pond}\right) + E_1E_2E_3V_4\left(\hat{t}_{y,2phases,nr,pond}\right)$$

$$\boldsymbol{V}\left(\hat{t}_{y,2phases,nr,pond}\right) = \sum_{i\in U}\sum_{j\in U}(\boldsymbol{\pi_{EI,ij}} - \boldsymbol{\pi_{EI,i}\pi_{EI,j}})\frac{\boldsymbol{y_i}}{\boldsymbol{\pi_{EI,i}}}\frac{\boldsymbol{y_j}}{\boldsymbol{\pi_{EI,j}}} + \boldsymbol{E_1E_2}\ \sum_{i\in U}\sum_{j\in U}(\boldsymbol{\pi_{EC,ij}} - \boldsymbol{\pi_{EC,i}\pi_{EC,j}})\frac{\boldsymbol{\phi_i}}{\boldsymbol{\pi_{EC,i}}}\frac{\boldsymbol{\phi_j}}{\boldsymbol{\pi_{EC,j}}} + \boldsymbol{E_1E_2E_3}\ \left(\sum_{i\in U}\frac{d_{EIi}^2}{\pi_{EC,i}^2}\frac{(1-\delta_{EC,i})}{\delta_{EC,i}}y_i^2 I_{EI,i}(1-R_{EI,i})\ \ I_{EC,i}\right)$$

2. Quel est l'estimateur de la variance de $\hat{t}_{y,2phases,nr,pond}$?

$$\boldsymbol{V}\left(\hat{t}_{y,2phases,nr,pond}\right) = \sum_{i\in U}\sum_{j\in U}(\boldsymbol{\pi_{EI,ij}} - \boldsymbol{\pi_{EI,i}\pi_{EI,j}})\frac{\boldsymbol{y_i}}{\boldsymbol{\pi_{EI,i}}}\frac{\boldsymbol{y_j}}{\boldsymbol{\pi_{EI,j}}} \qquad \boldsymbol{V_1}(\hat{t}_{y,2phases,nr,pond})$$

$$+\boldsymbol{E_1E_2}\ \sum_{i\in U}\sum_{j\in U}(\boldsymbol{\pi_{EC,ij}} - \boldsymbol{\pi_{EC,i}\pi_{EC,j}})\frac{\boldsymbol{\phi_i}}{\boldsymbol{\pi_{EC,i}}}\frac{\boldsymbol{\phi_j}}{\boldsymbol{\pi_{EC,j}}} \qquad \boldsymbol{V_2}(\hat{t}_{y,2phases,nr,pond})$$

$$+\boldsymbol{E_1E_2E_3}\ \left(\sum_{i\in U}\frac{\boldsymbol{d_{EI,i}^2}}{\boldsymbol{\pi_{EC,i}^2}}\frac{\boldsymbol{(1-\delta_{EC,i})}}{\boldsymbol{\delta_{EC,i}}}\boldsymbol{y_i^2\, I_{EI,i}(1-R_{EI,i})I_{EC,i}}\right) \qquad \boldsymbol{V_3}(\hat{t}_{y,2phases,nr,pond})$$

avec $\boldsymbol{\phi_i = d_{EI,i}y_iI_{EI,i}(1-R_{EI,i})}$

- $\boldsymbol{V_3}\left(\hat{t}_{y,2phases,nr,pond}\right)$

$$\boldsymbol{V_3}\left(\hat{t}_{y,2phases,nr,pond}\right) = E_1E_2E_3\ \left(\sum_{i\in U}\frac{d_{EI,i}^2}{\pi_{EC,i}^2}\frac{(1-\delta_{EC,i})}{\delta_{EC,i}}y_i^2\, I_{EI,i}(1-R_{EI,i})I_{EC,i}\right)$$

Soit $\Delta_{3,i}=\frac{d_{EI,i}^2}{\pi_{EC,i}^2}\frac{(1-\delta_{EC,i})}{\delta_{EC,i}}$

$E_1E_2E_3$ → On cherche à représenter s_{EC}

Pour cela, on a les données de $s_{EC,r}$ et on doit repondérer cet échantillon observé par l'inverse de $E_4(R_{EC,i})$ soit $\frac{1}{\delta_{EC,i}}$

→ $\boldsymbol{\hat{V}_3}\left(\hat{t}_{y,2phases,nr,pond}\right)=\sum_{\boldsymbol{i\in s_{EC,r}}}\frac{\Delta_{3i}}{\delta_{EC,i}}\boldsymbol{y_i^2}$

- $\boldsymbol{V_2}\left(\hat{t}_{y,2phases,nr,pond}\right)$

$$\boldsymbol{V_2}\left(\hat{t}_{y,2phases,nr,pond}\right)=E_1E_2\sum_{i\in U}\sum_{j\in U}(\pi_{EC,ij}-\pi_{EC,i}\pi_{EC,j})\frac{\phi_i}{\pi_{EC,i}}\frac{\phi_j}{\pi_{EC,j}}$$

$$V_2\left(\hat{t}_{y,2phases,nr,pond}\right)=E_1E_2\sum_{i\in U}\sum_{j\in U}(\pi_{EC,ij}-\pi_{EC,i}\pi_{EC,j})\frac{d_{EI,i}y_iI_{EI,i}(1-R_{EI,i})}{\pi_{EC,i}}\frac{d_{EI,j}y_jI_{EI,j}(1-R_{EI,j})}{\pi_{EC,j}}$$

Soit $\Delta_{2,ij}=\frac{(\pi_{EC,ij}-\pi_{EC,i}\pi_{EC,j})}{\pi_{EC,i}\pi_{EC,j}}d_{EI,i}d_{EI,j}$

$$V_2\left(\hat{t}_{y,2phases,nr,pond}\right)=E_1E_2\sum_{i\in U}\sum_{j\in U}\Delta_{2ij}y_iy_jI_{EI,i}(1-R_{EI,i})I_{EI,j}(1-R_{EI,j})$$

E_1E_2 → On cherche à représenter $s_{EI,m}$

Pour cela, on a les données de $s_{EC,r}$ et on doit repondérer cet échantillon observé par l'inverse de :

- $E_4(R_{EC,i}R_{EC,j})$ soit $\frac{1}{\delta_{EC,i}}\frac{1}{\delta_{EC,j}}$
- $E_3\left(I_{EC,i}I_{EC,j}\right)$ soit $\frac{1}{\pi_{EC,ij}}$

→ $\hat{V}_2\left(\hat{t}_{y,2phases,nr,pond}\right)=\sum_{i\in s_{EC,r}}\sum_{j\in s_{EC,r}}\frac{\Delta_{2ij}}{\pi_{EC,ij}}\frac{1}{\delta_{EC,i}\delta_{EC,j}}y_iy_j$

- $\boldsymbol{V_1}(\hat{t}_{y,2phases,nr,pond})$

$$\boldsymbol{V_1}(\hat{t}_{y,2phases,nr,pond}) = \sum_{i\in U}\sum_{j\in U}(\pi_{EI,ij} - \pi_{EI,i}\pi_{EI,j})\frac{y_i}{\pi_{EI,i}}\frac{y_j}{\pi_{EI,j}}$$

Soit $\Delta_{1ij} = \frac{(\pi_{EI,ij} - \pi_{EI,i}\pi_{EI,j})}{\pi_{EI,i}\pi_{EI,j}}$

$$V_1(\hat{t}_{y,2phases,nr,pond}) = \sum_{i\in U}\sum_{j\in U}\Delta_{1ij}y_i y_j$$

On cherche à représenter la population U=$U_r \cup U_m$.

Pour cela on a les données de $s_{EI,r}$ pour représenter U_r, les données de $s_{EC,r}$ pour représenter U_m et 4 combinaisons possibles pour la double somme :

- $i \in U_r \; et \;\; j \in U_r$ → pour se ramener à cette partie de la population, il faudra repondérer par l'inverse de $E\;\; (I_{EI,i}I_{EI,j}) = \frac{1}{\pi_{EI,ij}}$
- $i \in U_r \; et \;\; j \in U_m$→ pour se ramener à cette partie de la population, il faudra repondérer par les inverses de $E\;\; (I_{EI,i}I_{EI,j}) = \frac{1}{\pi_{EI,ij}}\;\; E_4(R_{EC,j}) = \frac{1}{\delta_{EC,j}}$ et $E_3(I_{EC,j}) = \frac{1}{\pi_{EC,j}}$
- $i \in U_m \; et \; j \in U_r$ → pour se ramener à cette partie de la population, il faudra repondérer par les inverses de $E\;\; (I_{EI,i}I_{EI,j}) = \frac{1}{\pi_{EI,ij}}\;\; E_4(R_{EC,j}) = \frac{1}{\delta_{EC,j}}$ et $E_3(I_{EC,j}) = \frac{1}{\pi_{EC,j}}$
- $i \in U_m \; et \; j \in U_m$ → pour se ramener à cette partie de la population, il faudra repondérer par les inverses de $E\;\; (I_{EI,i}I_{EI,j}) = \frac{1}{\pi_{EI,ij}}\;\; E_4(R_{EC,i}R_{EC,j}) = \frac{1}{\delta_{EC,i}\delta_{EC,j}}$ et $E_3(I_{EC,i}I_{EC,j}) = \frac{1}{\pi_{EC,ij}}$

→

$\hat{V}_1(\hat{Y}_{NR}) =$

$\sum_{i\in s_{EI,r}}\sum_{j\in s_{EI,r}}\frac{\Delta_{1,ij}}{\pi_{EI,ij}}y_i y_j + \sum_{i\in s_{EI,r}}\sum_{j\in s_{EC,r}}\frac{\Delta_{1,ij}}{\pi_{EI,ij}}\frac{1}{\pi_{EC,j}\delta_{EC,j}}y_i y_j + \sum_{i\in s_{EC,r}}\sum_{j\in s_{EI,r}}\frac{\Delta_{1,ij}}{\pi_{EI,ij}}\frac{1}{\pi_{EC,i}\delta_{EC,i}}y_i y_j +$

$\sum_{i\in s_{EC,r}}\sum_{j\in s_{EC,r}}\frac{\Delta_{1,ij}}{\pi_{EI,ij}}\frac{1}{\pi_{EC,ij}\delta_{EC,i}\delta_{EC,j}}y_i y_j$

- **Retour à $\widehat{V}\ \left(\hat{t}_{y,2phases,nr,pond}\right)$**

$$\hat{V}\ \left(\hat{t}_{y,2phases,nr,pond}\right) = \hat{V}_1\left(\hat{t}_{y,2phases,nr,pond}\right) + \hat{V}_2\left(\hat{t}_{y,2phases,nr,pond}\right) + \hat{V}_3\left(\hat{t}_{y,2phases,nr,pond}\right)$$

$$\widehat{\boldsymbol{V}}\ \left(\hat{t}_{y,2phases,nr,pond}\right)$$

$$= \sum_{i\in s_{EI,r}} \sum_{j\in s_{EI,r}} \frac{\Delta_{1,ij}}{\boldsymbol{\pi_{EI,ij}}} \boldsymbol{y_i y_j} + \sum_{i\in s_{EI,r}} \sum_{j\in s_{EC,r}} \frac{\Delta_{1,ij}}{\boldsymbol{\pi_{EI,ij}}} \frac{\boldsymbol{1}}{\boldsymbol{\pi_{EC,j}\delta_{EC,j}}} \boldsymbol{y_i y_j}$$

$$+ \sum_{i\in s_{EC,r}} \sum_{j\in s_{EI,r}} \frac{\Delta_{1,ij}}{\boldsymbol{\pi_{EI,ij}}} \frac{\boldsymbol{1}}{\boldsymbol{\pi_{EC,i}\delta_{EC,i}}} \boldsymbol{y_i y_j} + \sum_{i\in s_{EC,r}} \sum_{j\in s_{EC,r}} \frac{\Delta_{1,ij}}{\boldsymbol{\pi_{EI,ij}}} \frac{\boldsymbol{1}}{\boldsymbol{\pi_{EC,ij}\delta_{EC,i}\delta_{EC,j}}} \boldsymbol{y_i y_j}$$

$$+ \sum_{i\in s_{EC,r}} \sum_{j\in s_{EC,r}} \frac{\Delta_{2,ij}}{\boldsymbol{\pi_{EC,ij}}} \frac{\boldsymbol{1}}{\boldsymbol{\delta_{EC,i}\delta_{EC,j}}} \boldsymbol{y_i y_j} + \sum_{i\in s_{EC,r}} \frac{\Delta_{3,i}}{\boldsymbol{\delta_{EC,i}}} \boldsymbol{y_i^2}$$

avec $\Delta_{1,ij} = \frac{(\pi_{EI,ij} - \pi_{EI,i}\pi_{EI,j})}{\pi_{EI,i}\pi_{EI,j}}$

$\Delta_{2,ij} = \frac{(\pi_{EC,ij} - \pi_{EC,i}\pi_{EC,j})}{\pi_{EC,i}\pi_{EC,j}} d_{EI,i} d_{EI,j}$

$\Delta_{3,i} = \frac{d_{EI,i}^2}{\pi_{EC,i}^2} \frac{(1-\delta_{EC,i})}{\delta_{EC,i}}$

Elle peut se décomposer ainsi :

- $i \neq j$

$$\hat{V}\ \left(\hat{t}_{y,2phases,nr,pond}\right) = \hat{V}_1\left(\hat{t}_{y,2phases,nr,pond}\right) + \hat{V}_2\left(\hat{t}_{y,2phases,nr,pond}\right) + \hat{V}_3\left(\hat{t}_{y,2phases,nr,pond}\right)$$

$$\widehat{\boldsymbol{V}}\ \left(\hat{t}_{y,2phases,nr,pond}\right)$$

$$= \sum_{i\in s_{EI,r}} \sum_{j\in s_{EI,r}(j\neq i)} \frac{\Delta_{1,ij}}{\boldsymbol{\pi_{EI,ij}}} \boldsymbol{y_i y_j} + \sum_{i\in s_{EI,r}} \sum_{j\in s_{EC,r}} \frac{\Delta_{1,ij}}{\boldsymbol{\pi_{EI,ij}}} \frac{\boldsymbol{1}}{\boldsymbol{\pi_{EC,j}\delta_{EC,j}}} \boldsymbol{y_i y_j}$$

$$+ \sum_{i\in s_{EC,r}} \sum_{j\in s_{EI,r}} \frac{\Delta_{1,ij}}{\boldsymbol{\pi_{EI,ij}}} \frac{\boldsymbol{1}}{\boldsymbol{\pi_{EC,i}\delta_{EC,i}}} \boldsymbol{y_i y_j} + \sum_{i\in s_{EC,r}} \sum_{j\in s_{EC,r}(j\neq i)} \frac{\Delta_{1,ij}}{\boldsymbol{\pi_{EI,ij}}} \frac{\boldsymbol{1}}{\boldsymbol{\pi_{EC,ij}\delta_{EC,i}\delta_{EC,j}}} \boldsymbol{y_i y_j}$$

$$+ \sum_{i\in s_{EC,r}} \sum_{j\in s_{EC,r}(j\neq i)} \frac{\Delta_{2,ij}}{\boldsymbol{\pi_{EC,ij}}} \frac{\boldsymbol{1}}{\boldsymbol{\delta_{EC,i}\delta_{EC,j}}} \boldsymbol{y_i y_j} + \sum_{i\in s_{EC,r}} \frac{\Delta_{3,i}}{\boldsymbol{\delta_{EC,i}}} \boldsymbol{y_i^2}$$

avec $\Delta_{1,ij} = \frac{(\pi_{EI,ij} - \pi_{EI,i}\pi_{EI,j})}{\pi_{EI,i}\pi_{EI,j}}$

$$\Delta_{2,ij}=\frac{(\pi_{EC,ij}-\pi_{EC,i}\pi_{EC,j})}{\pi_{EC,i}\pi_{EC,j}}d_{EI,i}d_{EI,j}$$

$$\Delta_{3,i}=\frac{d_{EI,i}^2}{\pi_{EC,i}^2}\frac{(1-\delta_{EC,i})}{\delta_{EC,i}}$$

- i = j

$$\hat{V}\ \left(\hat{t}_{y,2phases,nr,pond}\right)=\hat{V}_1\left(\hat{t}_{y,2phases,nr,pond}\right)+\hat{V}_2\left(\hat{t}_{y,2phases,nr,pond}\right)+\hat{V}_3\left(\hat{t}_{y,2phases,nr,pond}\right)$$

$$\widehat{\boldsymbol{V}}\ \left(\hat{t}_{y,2phases,nr,pond}\right)$$

$$=\sum_{i\in s_{EI,r}}\sum_{j\in s_{EI,r}(j=i)}\frac{\boldsymbol{1-\pi_{EI,i}}}{\boldsymbol{\pi_{EI,i}^2}}\boldsymbol{y_i^2}\ +\sum_{i\in s_{EC,r}}\sum_{j\in s_{EC,r}(j=i)}\frac{\boldsymbol{1-\pi_{EI,i}}}{\boldsymbol{\pi_{EI,i}^2}}\frac{\boldsymbol{1}}{\boldsymbol{\pi_{EC,i}\delta_{EC,i}^2}}\boldsymbol{y_i^2}$$

$$+\sum_{i\in s_{EC,r}}\sum_{j\in s_{EC,r}(j=i)}\frac{\boldsymbol{1-\pi_{EC,i}}}{\boldsymbol{\pi_{EC,i}^2}}\left(\frac{\boldsymbol{d_{EI,i}}}{\boldsymbol{\delta_{EC,i}}}\right)^{\boldsymbol{2}}\boldsymbol{y_i^2}+\sum_{i\in s_{EC,r}}\frac{\boldsymbol{d_{EI,i}^2}}{\boldsymbol{\pi_{EC,i}^2}}\frac{\boldsymbol{(1-\delta_{EC,i})}}{\boldsymbol{\delta_{EC,i}^2}}\boldsymbol{y_i^2}$$

Propriétés utilisées :

- $\pi_{ii}=\pi_i$

- $Var(I_iI_i)=\pi_i(1-\pi_i)$ *(Loi de Bernoulli)*

3. $\hat{V}(\hat{t}_{y,2phases,nr,pond})$ est-il un estimateur sans biais de $V(\hat{t}_{y,2phases,nr,pond})$?

- $\hat{V}_3$

$$E_1E_2E_3E_4\left(\hat{V}_3\right)=E_1E_2E_3E_4(\sum_{i\in s_{EC,r}}\frac{\Delta_{3,i}}{\delta_{EC,i}}y_i^2)$$

$$E_1E_2E_3E_4\left(\hat{V}_3\right)=E_1E_2E_3(\sum_{i\in s_{EC}}\frac{\Delta_{3,i}}{\delta_{EC,i}}y_i^2E_4\left(R_{EC,i}\right))$$

$$E_1E_2E_3E_4\left(\hat{V}_3\right)=E_1E_2E_3(\sum_{i\in s_{EC}}\frac{\Delta_{3,i}}{\delta_{EC,i}}y_i^2\delta_{EC,i})$$

$$E_1E_2E_3E_4(\hat{V}_3) = E_1E_2E_3(\sum_{i\in s_{EC}} \Delta_{3,i}y_i^2)$$

$$E_1E_2E_3E_4(\hat{V}_3) = E_1E_2E_3\ \ (\sum_{i\in U} \Delta_{3,i}y_i^2\, I_{EI,i}(1-R_{EI,i})I_{EC,i})$$

Or $\Delta_{3,i}=\frac{d_{EI,i}^2}{\pi_{EC,i}^2}\frac{(1-\delta_{EC,i})}{\delta_{EC,i}}$

$$E_1E_2E_3E_4(\hat{V}_3) = E_1E_2E_3\ \ (\sum_{i\in U} \frac{d_{EI,i}^2}{\pi_{EC,i}^2}\frac{\left(1-\delta_{EC,i}\right)}{\delta_{EC,i}}y_i^2\, I_{EI,i}(1-R_{EI,i})I_{EC,i})$$

$$E_1E_2E_3E_4(\hat{V}_3) = V_3$$

- $\hat{V}_2$

$$\boldsymbol{E_1E_2E_3E_4(\hat{V}_2)} = E_1E_2E_3E_4(\sum_{i\in s_{EC,r}}\sum_{j\in s_{EC,r}} \frac{\Delta_{2,ij}}{\pi_{EC,ij}}\frac{1}{\delta_{EC,i}\delta_{EC,j}}y_i\ \ y_j)$$

$$E_1E_2E_3E_4(\hat{V}_2) = E_1E_2E_3E_4(\sum_{i\in s_{EC}}\sum_{j\in s_{EC}} \frac{\Delta_{2,ij}}{\pi_{EC,ij}}\frac{1}{\delta_{EC,i}\delta_{EC,j}}y_i\ \ y_j\, R_{EC,i}R_{EC,j})$$

$$E_1E_2E_3E_4(\hat{V}_2) = E_1E_2E_3(\sum_{i\in s_{EC}}\sum_{j\in s_{EC}} \frac{\Delta_{2,ij}}{\pi_{EC,ij}}\frac{1}{\delta_{EC,i}\delta_{EC,j}}y_i\ \ y_j\, E_4(R_{EC,i}R_{EC,j}))$$

$$E_1E_2E_3E_4(\hat{V}_2) = E_1E_2E_3(\sum_{i\in s_{EC}}\sum_{j\in s_{EC}} \frac{\Delta_{2,ij}}{\pi_{EC,ij}}\frac{1}{\delta_{EC,i}\delta_{EC,j}}y_i\ \ y_j\, \delta_{EC,i}\delta_{EC,j})$$

$$E_1E_2E_3E_4(\hat{V}_2) = E_1E_2E_3(\sum_{i\in s_{EC}}\sum_{j\in s_{EC}} \frac{\Delta_{2,ij}}{\pi_{EC,ij}}y_i\ \ y_j)$$

$$E_1E_2E_3E_4(\hat{V}_2) = E_1E_2E_3(\sum_{i\in s_{EI,m}}\sum_{j\in s_{EI,m}} \frac{\Delta_{2,ij}}{\pi_{EC,ij}}y_i\ \ y_j\, I_{EC,i}I_{EC,j})$$

$$E_1E_2E_3E_4(\hat{V}_2) = E_1E_2(\sum_{i\in s_{EI,m}}\sum_{j\in s_{EI,m}}\frac{\Delta_{2,ij}}{\pi_{EC,ij}}y_i \quad y_j\, E_3(I_{EC,i}I_{EC,j}))$$

$$E_1E_2E_3E_4(\hat{V}_2) = E_1E_2(\sum_{i\in s_{EI,m}}\sum_{j\in s_{EI,m}}\frac{\Delta_{2,ij}}{\pi_{EC,ij}}y_i \quad y_j\, \pi_{EC,i})$$

$$E_1E_2E_3E_4(\hat{V}_2) = E_1E_2(\sum_{i\in s_{EI,m}}\sum_{j\in s_{EI,m}}\Delta_{2,ij}y_i \quad y_j)$$

$$E_1E_2E_3E_4(\hat{V}_2) = E_1E_2(\sum_{i\in U}\sum_{j\in U}\Delta_{2,ij}y_i \quad y_j\, I_{EI,i}I_{EI,j}(1-R_{EI,i})(1-R_{EI,j}))$$

Or, $\phi_i = d_{EI,i}y_iI_{EI,i}(1-R_{EI,i})$ et $\Delta_{2,ij}=\frac{(\pi_{EC,ij}-\pi_{EC,i}\pi_{EC,j})}{\pi_{EC,i}\pi_{EC,j}}d_{EI,i}d_{EI,j}$

Donc

$$E_1E_2E_3E_4(\hat{V}_2) = E_1E_2 \quad \sum_{i\in U}\sum_{j\in U}(\pi_{EC,ij}-\pi_{EC,i}\pi_{EC,j})\frac{\phi_i}{\pi_{EC,i}}\frac{\phi_j}{\pi_{EC,j}}$$

$$E_1E_2E_3E_4(\hat{V}_2) = V_2$$

- $\hat{V}_1$

$$\boldsymbol{E_1E_2E_3E_4(\hat{V}_1)}$$

$$= E_1E_2E_3E_4(\sum_{i\in s_{EI,r}}\sum_{j\in s_{EI,r}}\frac{\Delta_{1,ij}}{\pi_{EI,ij}}y_iy_j + + \sum_{i\in s_{EI,r}}\sum_{j\in s_{EC,r}}\frac{\Delta_{1,ij}}{\pi_{EI,ij}}\frac{1}{\pi_{EC,j}\delta_{EC,j}}y_iy_j$$

$$+ \sum_{i\in s_{EC,r}}\sum_{j\in s_{EI,r}}\frac{\Delta_{1,ij}}{\pi_{EI,ij}}\frac{1}{\pi_{EC,i}\delta_{EC,i}}y_iy_j + \sum_{i\in s_{EC,r}}\sum_{j\in s_{EC,r}}\frac{\Delta_{1,ij}}{\pi_{EI,ij}}\frac{1}{\pi_{EC,ij}\delta_{EC,i}\delta_{EC,j}} \quad y_iy_j)$$

$$E_1E_2E_3E_4\left(\hat{V}_1\right) = E_1E_2E_3E_4(\hat{V}_1(a)) + E_1E_2E_3E_4(\hat{V}_1(b)) + E_1E_2E_3E_4(\hat{V}_1(c)) + E_1E_2E_3E_4(\hat{V}_1(d))$$

$$\boldsymbol{E_1E_2E_3E_4}\left(\boldsymbol{\hat{V}_1(d)}\right) = E_1E_2E_3E_4\Big(\sum_{i\in s_{EC,r}}\sum_{j\in s_{EC,r}}\frac{\Delta_{1,ij}}{\pi_{EI,ij}}\frac{1}{\pi_{EC,ij}\delta_{EC,i}\delta_{EC,j}}\quad y_iy_j\Big)$$

$$E_1E_2E_3E_4\left(\hat{V}_1(d)\right) = E_1E_2E_3E_4\Big(\sum_{i\in s_{EC}}\sum_{j\in s_{EC}}\frac{\Delta_{1,ij}}{\pi_{EI,ij}}\frac{1}{\pi_{EC,ij}\delta_{EC,i}\delta_{EC,j}}\quad y_iy_j\,R_{EC,i}R_{EC,j}\Big)$$

$$E_1E_2E_3E_4\left(\hat{V}_1(d)\right) = E_1E_2E_3\Big(\sum_{i\in s_{EC}}\sum_{j\in s_{EC}}\frac{\Delta_{1,ij}}{\pi_{EI,ij}}\frac{1}{\pi_{EC,ij}\delta_{EC,i}\delta_{EC,j}}\quad y_iy_j\,E_4\big(R_{EC,i}R_{EC,j}\big)\Big)$$

$$E_1E_2E_3E_4\left(\hat{V}_1(d)\right) = E_1E_2E_3\Big(\sum_{i\in s_{EC}}\sum_{j\in s_{EC}}\frac{\Delta_{1,ij}}{\pi_{EI,ij}}\frac{1}{\pi_{EC,ij}\delta_{EC,i}\delta_{EC,j}}\quad y_iy_j\,\delta_{EC,i}\delta_{EC,j}\Big)$$

$$E_1E_2E_3E_4\left(\hat{V}_1(d)\right) = E_1E_2E_3\Big(\sum_{i\in s_{EC}}\sum_{j\in s_{EC}}\frac{\Delta_{1,ij}}{\pi_{EI,ij}}\frac{1}{\pi_{EC,ij}}\quad y_iy_j\Big)$$

$$E_1E_2E_3E_4\left(\hat{V}_1(d)\right) = E_1E_2E_3\Big(\sum_{i\in s_{EI,m}}\sum_{j\in s_{EI,m}}\frac{\Delta_{1,ij}}{\pi_{EI,ij}}\frac{1}{\pi_{EC,ij}}\quad y_iy_j\,I_{EC,i}I_{EC,j}\Big)$$

$$E_1E_2E_3E_4\left(\hat{V}_1(d)\right) = E_1E_2\Big(\sum_{i\in s_{EI,m}}\sum_{j\in s_{EI,m}}\frac{\Delta_{1,ij}}{\pi_{EI,ij}}\frac{1}{\pi_{EC,ij}}\quad y_iy_j\,E_3(I_{EC,i}I_{EC,j})\Big)$$

$$E_1E_2E_3E_4\left(\hat{V}_1(d)\right) = E_1E_2\Big(\sum_{i\in s_{EI,m}}\sum_{j\in s_{EI,m}}\frac{\Delta_{1,ij}}{\pi_{EI,ij}}\frac{1}{\pi_{EC,ij}}\quad y_iy_j\,\pi_{EC,ij}\Big)$$

$$E_1E_2E_3E_4\left(\hat{V}_1(d)\right) = E_1E_2\Big(\sum_{i\in s_{EI,m}}\sum_{j\in s_{EI,m}}\frac{\Delta_{1,ij}}{\pi_{EI,ij}}\quad y_iy_j\Big)$$

$$\boldsymbol{E_1E_2E_3E_4}\left(\boldsymbol{\hat{V}_1(c)}\right) = E_1E_2E_3E_4\Big(\sum_{i\in s_{EC,r}}\sum_{j\in s_{EI,r}}\frac{\Delta_{1,ij}}{\pi_{EI,ij}}\frac{1}{\pi_{EC,i}\delta_{EC,i}}y_iy_j\Big)$$

$$E_1E_2E_3E_4\left(\widehat{V}_1(c)\right) = E_1E_2E_3(\sum_{i\in s_{EC}}\sum_{j\in s_{EI,r}}\frac{\Delta_{1,ij}}{\pi_{EI,ij}}\frac{1}{\pi_{EC,i}\delta_{EC,i}}y_iy_j\,E_4(R_{EC,i}))$$

$$E_1E_2E_3E_4\left(\widehat{V}_1(c)\right) = E_1E_2E_3(\sum_{i\in s_{EC}}\sum_{j\in s_{EI,r}}\frac{\Delta_{1,ij}}{\pi_{EI,ij}}\frac{1}{\pi_{EC,i}\delta_{EC,i}}y_iy_j\,\delta_{EC,i})$$

$$E_1E_2E_3E_4\left(\widehat{V}_1(c)\right) = E_1E_2E_3(\sum_{i\in s_{EC}}\sum_{j\in s_{EI,r}}\frac{\Delta_{1,ij}}{\pi_{EI,ij}}\frac{1}{\pi_{EC,i}}y_iy_j)$$

$$E_1E_2E_3E_4\left(\widehat{V}_1(c)\right) = E_1E_2(\sum_{i\in s_{EI,m}}\sum_{j\in s_{EI,r}}\frac{\Delta_{1,ij}}{\pi_{EI,ij}}\frac{1}{\pi_{EC,i}}y_iy_j\,E_3(I_{EC,i}))$$

$$E_1E_2E_3E_4\left(\widehat{V}_1(c)\right) = E_1E_2(\sum_{i\in s_{EI,m}}\sum_{j\in s_{EI,r}}\frac{\Delta_{1,ij}}{\pi_{EI,ij}}\frac{1}{\pi_{EC,i}}y_iy_j\,\pi_{EC,i})$$

$$E_1E_2E_3E_4\left(\widehat{V}_1(c)\right) = E_1E_2(\sum_{i\in s_{EI,m}}\sum_{j\in s_{EI,r}}\frac{\Delta_{1,ij}}{\pi_{EI,ij}}y_iy_j)$$

$$\boldsymbol{E_1E_2E_3E_4}\left(\widehat{\boldsymbol{V}}_{\boldsymbol{1}}(\boldsymbol{b})\right) = E_1E_2E_3E_4(\sum_{i\in s_{EI,r}}\sum_{j\in s_{EC,r}}\frac{\Delta_{1,ij}}{\pi_{EI,ij}}\frac{1}{\pi_{EC,j}\delta_{EC,j}}y_iy_j)$$

Même raisonnement que pour $\widehat{V}_1(c)$ en intervertissant i et j

$$E_1E_2E_3E_4\left(\widehat{V}_1(b)\right) = E_1E_2(\sum_{i\in s_r}\sum_{j\in s_m}\frac{\Delta_{1ij}}{\pi_{ij}}y_iy_j)$$

$$\boldsymbol{E_1E_2E_3E_4}\left(\widehat{\boldsymbol{V}}_{\boldsymbol{1}}(\boldsymbol{a})\right) = E_1E_2E_3E_4(\sum_{i\in s_{EI,r}}\sum_{j\in s_{EI,r}}\frac{\Delta_{1,ij}}{\pi_{EI,ij}}y_iy_j)$$

$$E_1E_2E_3E_4\left(\widehat{V}_1(a)\right) = E_1E_2(\sum_{i\in s_{EI,r}}\sum_{j\in s_{EI,r}}\frac{\Delta_{1,ij}}{\pi_{EI,ij}}y_iy_j)$$

$$\boldsymbol{E_1E_2E_3E_4(\hat{V}_1)}$$

$$= E_1E_2E_3E_4(\hat{V}_1(a)) + E_1E_2E_3E_4(\hat{V}_1(b)) + E_1E_2E_3E_4(\hat{V}_1(c)) + E_1E_2E_3E_4(\hat{V}_1(d))$$

$$E_1E_2E_3E_4(\hat{V}_1) = E_1E_2\left(\sum_{i\in s_{EI,r}}\sum_{j\in s_{EI,r}}\frac{\Delta_{1,ij}}{\pi_{EI,ij}}y_iy_j\right) + E_1E_2\left(\sum_{i\in s_{EI,r}}\sum_{j\in s_{EI,m}}\frac{\Delta_{1,ij}}{\pi_{EI,ij}}y_iy_j\right) + E_1E_2(\sum_{i\in s_{EI,m}}\sum_{j\in s_{EI,r}}\frac{\Delta_{1,ij}}{\pi_{EI,ij}}y_iy_j) + E_1E_2(\sum_{i\in s_{EI,m}}\sum_{j\in s_{EI,m}}\frac{\Delta_{1,ij}}{\pi_{EI,ij}}\ y_iy_j)$$

$$E_1E_2E_3E_4(\hat{V}_1) = E_1E_2\left(\sum_{i\in U}\sum_{j\in U}\frac{\Delta_{1,ij}}{\pi_{EI,ij}}y_iy_j\, I_{EI,i}I_{EI,j}R_{EI,i}R_{EI,j}\right) + E_1E_2\left(\sum_{i\in U}\sum_{j\in U}\frac{\Delta_{1,ij}}{\pi_{EI,ij}}y_iy_j\, I_{EI,i}I_{EI,j}R_{EI,i}(1-R_{EI,j})\right) + E_1E_2\left(\sum_{i\in U}\sum_{j\in U}\frac{\Delta_{1,ij}}{\pi_{EI,ij}}y_iy_j\, I_{EI,i}I_{EI,j}(1-R_{EI,i})R_{EI,j}\right) + E_1E_2(\sum_{i\in U}\sum_{j\in U}\frac{\Delta_{1,ij}}{\pi_{EI,ij}}\ y_iy_j\, I_{EI,i}I_{EI,j}(1-R_{EI,i})(1-R_{EI,j}))$$

$$E_1E_2E_3E_4(\hat{V}_1) = E_1E_2\left(\sum_{i\in U}\sum_{j\in U}\frac{\Delta_{1,ij}}{\pi_{EI,ij}}y_iy_j\, I_{EI,i}I_{EI,j}(R_{EI,i}R_{EI,j} + R_{EI,i}(1-R_{EI,j}) + (1-R_{EI,i})R_{EI,j} + (1-R_{EI,i})(1-R_{EI,j})\right)$$

$$E_1E_2E_3E_4(\hat{V}_1) = E_1E_2\left(\sum_{i\in U}\sum_{j\in U}\frac{\Delta_{1,ij}}{\pi_{EI,ij}}y_iy_j\,I_{EI,i}I_{EI,j}(2R_{EI,i}R_{EI,j} - 2R_{EI,i}R_{EI,j} + R_{EI,i} - R_{EI,i} + R_{EI,j} - R_{EI,j} + 1)\right)$$

$$E_1E_2E_3E_4(\hat{V}_1) = E_1E_2\left(\sum_{i\in U}\sum_{j\in U}\frac{\Delta_{1,ij}}{\pi_{EI,ij}}y_iy_j\,I_{EI,i}I_{EI,j}\right)$$

$$E_1E_2E_3E_4(\hat{V}_1) = E_1\left(\sum_{i\in U}\sum_{j\in U}\frac{\Delta_{1,ij}}{\pi_{EI,ij}}y_iy_j\,I_{EI,i}I_{EI,j}\right)$$

$$E_1E_2E_3E_4(\hat{V}_1) = \sum_{i\in U}\sum_{j\in U}\frac{\Delta_{1,ij}}{\pi_{EI,ij}}y_iy_j\,E_1(I_{EI,i}I_{EIj})$$

$$E_1E_2E_3E_4(\hat{V}_1) = \sum_{i\in U}\sum_{j\in U}\frac{\Delta_{1,ij}}{\pi_{EI,ij}}y_iy_j\,\pi_{EI,ij}$$

$$E_1E_2E_3E_4(\hat{V}_1) = \sum_{i\in U}\sum_{j\in U}\Delta_{1,ij}y_iy_j$$

Or $\Delta_{1,ij} = \frac{(\pi_{EI,ij} - \pi_{EI,i}\pi_{EI,j})}{\pi_{EI,i}\pi_{EI,j}}$

$$E_1E_2E_3E_4(\hat{V}_1) = \sum_{i\in U}\sum_{j\in U}\frac{(\pi_{EI,ij} - \pi_{EI,i}\pi_{EI,j})}{\pi_{EI,i}\pi_{EI,j}}y_iy_j$$

$$E_1E_2E_3E_4(\hat{V}_1) = V_1$$

- **Retour à** $\hat{V}(\hat{t}_{y,2phases,nr,pond})$

$$\rightarrow E_1E_2E_3E_4\left(\hat{V}(\hat{t}_{y,2phases,nr,pond})\right) = V(\hat{t}_{y,2phases,nr,pond})$$

→ $\hat{V}(\hat{t}_{y,2phases,nr,pond})$ est un estimateur sans biais de $V(\hat{t}_{y,2phases,nr,pond})$.

ANNEXE II. QUESTIONNAIRE DE L'ENQUÊTE COMPLÉMENTAIRE

Le questionnaire de l'enquête complémentaire était extrait du questionnaire de l'enquête initiale. Ce dernier n'est pas intégré dans les annexes compte tenu de la longueur (40 pages). Il est consultable à l'adresse suivante : http://www.coset.fr/documentation_msa.htm

QUESTIONNAIRE ENQUETE COMPLEMENTAIRE

INFORMATIONS SOCIO-DEMOGRAPHIQUES

Q3 **Quels sont vos mois et année de naissance ?** |_||_| |1||9||_||_|

mois année

FOYER

F1 **Quelle est votre situation de famille au sens de l'état civil ?**

□1 Célibataire (jamais marié)

□2 Marié(e) ou pacsé(e)

□3 Veuf (ve)

□4 Divorcé(e) ou séparé

ETAT DE SANTE GENERAL

G3 **Comment jugez-vous votre état de santé général ?**

Mettre une croix sur l'échelle de A à H. A signifiant un très bon état de santé général et H un très mauvais état de santé général

Très bon	□	□	□	□	□	□	□	□	**Très mauvais**
	A	B	C	D	E	F	G	H	

G4 **Comment jugez-vous votre état de santé général par rapport à une personne de votre entourage du même âge ?**

Mettre une croix sur l'échelle de A à H. A signifiant un très bon état de santé général et H un très mauvais état de santé général

Très bon	□	□	□	□	□	□	□	□	**Très mauvais**
	A	B	C	D	E	F	G	H	

COMPORTEMENT ET SANTE

CONSOMMATION DE TABAC

T1 **Etes-vous ?**

□1 Fumeur (au moins une cigarette par jour)

□2 Non fumeur (ou moins d'une cigarette par jour)

□3 Ex-fumeur (arrêt du tabagisme depuis au moins 1 an)

CONSOMMATION DE BOISSONS ALCOOLISÉES : VIN, BIÈRE, APÉRITIF, DIGESTIF, CIDRE, …

Pour chacune des 2 questions suivantes, cochez la case qui répond le mieux à votre cas, **pour les 12derniers mois**.

A1 Quelle est la fréquence de votre consommation d'alcool ?

□1 Jamais

□2 1 fois par mois ou moins

□3 2 à 4 fois par mois

□4 2 à 3 fois par semaine

□5 Au moins 4 fois par semaine

A2 Combien de verres contenant de l'alcool consommez-vous un jour typique où vous buvez ?

□1 1 ou 2

□2 3 ou 4

□3 5 ou 6

□4 7 à 9

□5 10 ou plus

NIVEAU D'ETUDES

H1 Quel est le niveau d'études le plus élevé que vous ayez atteint (avec ou sans obtention de diplôme)

□0 Enseignement primaire ou 1er cycle de l'enseignement secondaire (niveau collège)

□3 Enseignement technique court : CAP, BEP

□4 Niveau lycée (enseignement général, technologique ou professionnel)

□5 Premier cycle de l'enseignement supérieur

□6 Deuxième et troisième cycle de l'enseignement supérieur, grande école, école d'ingénieur, de commerce …

□7 Autre :

Précisez : |_||_||_||_||_||_||_||_||_||_||_||_||_||_||_||_||_||_||_||_||_||_||_||_||_||_||_||_||_||_||_||_||_||_||_||_|

ACTIVITE PROFESSIONNELLE

P0 Au cours de votre carrière professionnelle combien d'emplois de plus de 4 mois consécutifs avez-vous occupé ?

□1 1 □2 2 à 4 □3 5 à 8 □4 9 ou plus

P1 Etes-vous actuellement en activité (y compris à temps partiel, intérimaire, intermittent du spectacle ou saisonnier) ?

□0 Non □1 Oui

Si non,

Quelle est votre situation actuelle ?

□1 Demandeur d'emploi ou à la recherche d'un emploi

□2 Retraité(e), préretraité(e) ou retiré(e) des affaires

□3 En arrêt volontaire temporaire d'activité (congés parental d'éducation, convenance personnelle…)

□4 En formation (apprenti, étudiant, stage…)

□5 Au foyer, sans profession

□6 Ne travaille pas pour raison de santé (longue maladie, invalidité,…)

□7 Autre, *Précisez :* |_||_||_||_||_||_||_||_||_||_||_||_||_||_||_||_||_||_||_||_||_||_||_||_||_||_||_||_||_||_|

Quelle était votre dernière profession ?

|_||_||_||_||_||_||_||_||_||_||_||_||_||_||_||_||_||_||_||_||_||_||_||_||_||_||_||_||_||_||_||_||_||_||_||_||_||_||_||_||_||_|

Quelle était la production ou le secteur d'activité de votre entreprise?

|_||_||_||_||_||_||_||_||_||_||_||_||_||_||_||_||_||_||_||_||_||_||_||_||_||_||_||_||_||_||_||_||_||_||_||_||_||_||_||_||_||_|

P3 Etiez-vous ?

□1 Salarié

□2 A votre compte Précisez □1 exploitant agricole □2 travailleur indépendant

□3 Aide familial

□4 Conjoint-collaborateur

Travailliez-vous à :

□1 Temps complet

□2 Temps partiel

Si vous étiez à votre compte, merci de préciser :

- le nombre de salariés employés

□1 0 □2 1 ou 2 □3 3 à 9 □4 10 salariés ou plus

Si vous étiez exploitant agricole :

- la surface agricole utile |_||_||_| ha

Si vous étiez salarié(e) :

- votre type de contrat

□1 CDI

□2 CDD

□3 Saisonnier/intérimaire

- le statut de votre employeur

□1 Etat

□2 Collectivités territoriales, HLM, hôpitaux

□3 Sécurité sociale

□4 Entreprises publiques nationalisées

□5 Privé

□6 Ne sait pas

- votre niveau de qualification

□1 Manœuvre ou ouvrier spécialisé

□2 Ouvrier qualifié ou ouvrier hautement qualifié ou technicien d'atelier

□3 Agent de maîtrise

□4 Directeur général ou adjoint direct au directeur

□5 Technicien, dessinateur, VRP

□6 Instituteur, assistante sociale, infirmière et autres personnels de catégorie B de la fonction publique

□7 Ingénieur ou cadre

□8 Professeur et personnel de catégorie A de la fonction publique

□9 Employés de bureau, de commerce, agents de service, aides soignantes, gardiennes d'enfants, personnels de catégorie C ou D de la fonction publique

□10 Autres

□11 Ne sait pas

Puis passer à la question N1

Si oui,

Quelle est votre profession ?

|_||_||_||_||_||_||_||_||_||_||_||_||_||_||_||_||_||_||_||_||_||_||_||_||_||_||_||_||_||_||_||_||_||_||_||_||_||_||_|

Quelle est la production ou le secteur d'activité de votre entreprise (de votre employeur)?

|_||_||_||_||_||_||_||_||_||_||_||_||_||_||_||_||_||_||_||_||_||_||_||_||_||_||_||_||_||_||_||_||_||_||_||_||_||_||_|

P3 **Etes-vous ?**

□1 Salarié

□2 A votre compte Précisez □1 exploitant agricole □2 travailleur indépendant

□3 Aide familial

□4 Conjoint-collaborateur

Travaillez-vous à :

□1 Temps complet

□2 Temps partiel

Si vous êtes à votre compte, merci de préciser :

- le nombre de salariés employés

□1 0 □2 1 ou 2 □3 3 à 9 □4 10 salariés ou plus

Si vous êtes exploitant agricole :

- la surface agricole utile |_||_||_| ha

Si vous êtes salarié(e) :

- votre type de contrat

□1 CDI

□2 CDD

□3 Saisonnier/intérimaire

- le statut de votre employeur

□1 Etat

□2 Collectivités territoriales, HLM, hôpitaux

□3 Sécurité sociale

□4 Entreprises publiques nationalisées

□5 Privé

□6 Ne sait pas

- votre niveau de qualification

□1 Manœuvre ou ouvrier spécialisé

□2 Ouvrier qualifié ou ouvrier hautement qualifié ou technicien d'atelier

□3 Agent de maîtrise

□4 Directeur général ou adjoint direct au directeur

□5 Technicien, dessinateur, VRP

□6 Instituteur, assistante sociale, infirmière et autres personnels de catégorie B de la fonction publique

□7 Ingénieur ou cadre

□8 Professeur et personnel de catégorie A de la fonction publique

□9 Employés de bureau, de commerce, agents de service, aides soignantes, gardiennes d'enfants, personnels de catégorie C ou D de la fonction publique

□10 Autres

□11 Ne sait pas

EXPOSITIONS PROFESSIONNELLES

P15 Êtes-vous en contact physique ou téléphonique avec le public (usagers, patients, clients, voyageurs...) ?

□0 Non □1 Oui

Si OUI vivez-vous des situations de tension dans vos rapports avec le public ?

□1 Non ou presque jamais

□2 Rarement

□3 Souvent

□4 Toujours ou presque

Pour les propositions suivantes, cochez la case correspondant le mieux à votre situation professionnelle actuelle (une case et une seule par question) :

S1 Je suis constamment pressé(e) par le temps à cause d'une forte charge de travail

□1 Pas d'accord

□2 D'accord mais ça ne me perturbe pas du tout

□3 D'accord et ça me perturbe un peu

□4 D'accord et ça me perturbe

□5 D'accord et ça me perturbe beaucoup

S10 Ma sécurité d'emploi est menacée

□1 Pas d'accord

□2 D'accord mais ça ne me perturbe pas du tout

□3 D'accord et ça me perturbe un peu

□4 D'accord et ça me perturbe

□5 D'accord et ça me perturbe beaucoup

S11 Vu tous mes efforts, je reçois le respect et l'estime que je mérite à mon travail

□1 D'accord

□2 Pas d'accord mais ça ne me perturbe pas du tout

□3 Pas d'accord et ça me perturbe un peu

□4 Pas d'accord et ça me perturbe

□5 Pas d'accord et ça me perturbe beaucoup

*Les questions suivantes se rapportent à une **journée de travail typique au cours des 12 derniers mois.***

J1 Comment évaluez-vous l'intensité des efforts physiques de votre travail au cours d'une journée de travail typique, sur une échelle allant de 6 à 20 ? ..... |_|

Cochez la modalité correspondant à votre choix sur l'échelle de 6 à 20 ci-dessous, qui va de « pas d'effort du tout » à « épuisant ».

6	☐	Pas d'effort du tout
7	☐	Extrêmement léger
8	☐	
9	☐	Très léger
10	☐	
11	☐	Léger
12	☐	
13	☐	Un peu dur
14	☐	
15	☐	Dur
16	☐	
17	☐	Très dur
18	☐	
19	☐	Extrêmement dur
20	☐	Épuisant

J2 Votre travail nécessite-t-il habituellement de répéter les mêmes actions plus de 2 à 4 fois environ par minute ?

☐$_1$ Non ou presque jamais

☐$_2$ Rarement (moins 2 heures par jour)

☐$_3$ Souvent (2 à 4 heures par jour)

☐$_4$ Toujours ou presque

J8 Au cours d'une journée typique de travail, devez-vous conduire un engin de chantier, un tracteur, un chariot automoteur ou autre machine mobile sur votre lieu de travail ?

☐$_1$ Non ou presque jamais

☐$_2$ Rarement (moins 2 heures par jour)

□3 Souvent (2 à 4 heures par jour)

□4 Toujours ou presque

J11 Au cours d'une journée typique de travail, utilisez-vous des outils vibrants ou devez-vous poser la(les) main(s) sur des machines vibrantes ?

□1 Non ou presque jamais

□2 Rarement (moins 2 heures par jour)

□3 Souvent (2 à 4 heures par jour)

□4 Toujours ou presque

E2 EXPOSITIONS AUX BRUITS

12) Etes-vous exposé(e) à des bruits intenses (tels qu'ils vous empêchent d'entendre la voix d'une personne placée à deux ou trois mètres, même si elle crie) ?

□0 Non □1 Oui

ANNEXE III. DISTRIBUTION DES POIDS

		Min	1%	5%	10%	25%	50%	75%	90%	95%	99%	Max
Enquête initiale	d	7,6	7,6	7,6	7,6	8,7	13,4	13,6	14,4	14,5	14,5	14,5
	w(X)	23,9	23,9	23,9	23,9	32,8	45,4	64,2	79,8	81,0	86,0	86,2
	w(X,V)	17,2	17,2	17,2	19,9	29,2	43,7	54,2	73,6	94,5	126,6	134,8
Enquêtes combinées	d	7,6	7,6	7,6	7,6	8,7	13,4	14,4	109,4	207,5	220,3	220,6
	w(X)	7,6	7,6	7,6	7,6	8,7	13,4	14,4	162,3	289,4	506,2	533,6
	w(X,V)	7,6	7,6	7,6	7,6	8,7	13,4	14,4	131,7	263,9	629,2	669,2
	w(X,Z)	7,6	7,6	7,6	7,6	8,7	13,4	14,4	148,1	266,4	536,3	667,5

Légende : d : poids de sondage, w : poids corrigé pour la non-réponse, X : variables sociodémographiques, V : variables du SNIIR-AM et de la MSA, Z : paradonnées

ANNEXE IV. RÉSULTATS SUPPLÉMENTAIRES : QUELQUES PRÉVALENCES ESTIMÉES APRÈS CORRECTION DE LA NON-RÉPONSE ET CALAGE

Une correction de la non-réponse a été effectuée pour les personnes avec informations auxiliaires issues du SNIIR-AM et de la MSA, soit pour XX répondants, où 10 groupes homogènes de réponse ont été constitués (cf. chapitre V.2.).

Afin de mettre à disposition aux utilisateurs les données du pilote Coset-MSA, deux autres traitements ont été effectués. Dans un premier temps, un onzième groupe de réponse homogène a été constitué pour prendre en compte les répondants personnes sans informations auxiliaires issues du SNIIR-AM et de la MSA. Dans un deuxième temps, un calage a été réalisé sur le sexe, l'âge, le statut d'emploi et le département.

- Prise en compte des personnes sans informations auxiliaires issues du SNIIR-AM et de la MSA

Il y avait 693 personnes tirées au sort sans informations auxiliaires issues du SNIIR-AM et de la MSA. Parmi elles, 55 étaient répondantes à l'enquête postale ; leur taux de réponse observé dans ce groupe était donc 7,9%. Un onzième groupe de réponse homogène a été constitué à partir de ces 693 personnes avec, comme facteur correctif, l'inverse du taux de réponse observé dans ce groupe.

- Calage sur le sexe, l'âge, le statut d'emploi et le département

Pour rappel, le calage consiste à considérer une population de référence, ici les affiliés à la MSA des cinq départements pilotes, âgés de 18 à 65 ans et ayant travaillé au moins 90 jours en 2008, qui fournit des distributions de référence pour certaines variables clés, ici les variables de stratification – caisse d'affiliation, sexe, âge, statut d'emploi. Les poids corrigés pour la non-réponse ont été modifiés de

telle sorte que la distribution de ces variables clés dans Coset-MSA coïncide exactement avec la distribution de référence.

Comme nous disposions dans la population de référence des distributions croisées pour les variables de stratification, la méthode de calage utilisée était la post-stratification. Le calage a été réalisé à l'aide de la macro Calmar de l'Insee.

La prévalence estimée après ces deux corrections est donnée par :

$$\hat{p}_{s_r;cal} = \frac{\sum_{i=1}^{n_r} w_{cal,i} y_i}{\sum_{i=1}^{n_r} w_{cal,i}}$$

Avec $w_{cal,i} = d_i * corr_{MAR(X,V,+GHR11),i} * corr_{cal,i}$ et $corr_{cal,i}$ facteur de correction obtenu après calage.

Tableau : Quelques prévalences variables issues du questionnaire avant et après calage

		MAR(X,V)			MAR(X,V,+GHR11) puis calage		
	n	**%**	**IC 95%**		**n**	**%**	**IC 95%**	
Etat de santé général perçu comme moyennement / pas bon	1010	41,8	39,5	44,1	1036	41,4	39,2	43,7
Catégorie sociale								
Agriculteur exploitant	635	30,5	28,3	32,7	650	29,9	27,8	32,1
Artisans commerçants chef d'entreprise	41	3,1	2,1	4,1	43	3,2	2,1	4,2
Cadre, profession intellectuelle supérieure	150	6,0	5,0	7,1	155	6,2	5,1	7,3
Profession intermédiaire	359	14,4	12,8	15,9	364	14,3	12,7	15,8
Employé	303	11,6	10,2	13,0	310	11,5	10,0	13,0
Ouvrier	654	34,4	32,1	36,8	675	35,0	32,7	37,4
J'ai le respect que je mérite au travail	1112	60,4	57,9	62,9	1144	61,1	58,4	63,8
Faible ressenti dans l'intensité des efforts physiques	654	30,2	27,9	32,4	669	30,4	28,0	32,8

MAR(X,V) : processus de non-réponse aléatoire conditionnellement aux variables sociodémographiques, relatives à la santé et au travail

ANNEXE V. PUBLICATIONS

Santin G., Bénézet L., Geoffroy B., Bouyer J., Gueguen A., Coset team. A complementary survey and its paradata contributed to take into account nonresponse bias in an occupational health surveillance survey. Soumis à l'American Journal of Epidemiology (under review)

Santin G., Geoffroy B., Bénézet L., Delézire P., Chatelot J., Sitta R., Bouyer J, Gueguen A. SNIIR-AM Cohorts Group. In an occupational health surveillance study, auxiliary data from administrative health and occupational databases effectively corrected for nonresponse. Journal of Clinical Epidemiology. 67 (2014) 722-730

A two-phase sampling survey for nonresponse, including paradata, corrected for nonresponse bias in an occupational health surveillance survey

ABSTRACT

Study objective

To study the contribution of a complementary survey among a subsample of nonrespondents, and the additional contribution of paradata in correcting for nonresponse bias in an occupational health surveillance survey.

Methods

In 2010, 10,000 workers were randomly selected and sent a postal questionnaire. Sociodemographic data were available for the whole sample. After data collection, a complementary survey among a random subsample of 500 nonrespondents was performed using a questionnaire administered by an interviewer. Paradata were collected for the complete subsample. Variables of interest were variables from administrative databases, not subject to differential measurement errors, and questionnaire variables. Corrected prevalences were estimated by first using the initial survey alone and then the initial and complementary surveys combined, under several assumptions regarding the missing data process. Results were compared by computing relative errors.

Results

The response rates of the initial and complementary surveys were 23.6% and 62.6%, respectively. For the combined surveys without paradata, relative errors decreased compared with the initial survey. The contribution of the paradata was weak.

Conclusions

When a complex descriptive survey has a low response rate, a short complementary survey among nonrespondents is useful. However, the contribution of paradata in correcting for nonresponse bias is questionable.

KEYWORDS

Unit nonresponse, Selection bias, Two-phase sampling for nonresponse, Paradata, Surveillance

In recent years, the decline in participation rates in surveys, including epidemiological surveillance surveys, has become a real concern since it may increase nonresponse bias (1). Since nonresponse bias is a function of both the response rate and the covariance between the probability of response and outcome variables (2), two options are available to reduce the effect of nonresponse.

The first is to increase the response rate by developing designs that make it possible to contact hard-to-reach persons, sometimes called "responsive design" protocols (3, 4). As Groves & Heeringa (3) pointed out, despite being developed several decades before the use of the term "responsive design", the oldest responsive design protocol in survey methodology is the use of two-phase sampling for nonresponse (5). Briefly, this design consists in first carrying out a classic random survey that leads to the division of the original sample into two strata: respondents and nonrespondents. Second, a subsample of nonrespondents is randomly selected and surveyed with a protocol designed to obtain a maximal response rate. This second survey, called a follow-up survey in the survey methodology literature, will be called a complementary survey in the present paper to avoid confusion with cohort studies aimed at following subjects over time. A 100% response rate to the complementary survey makes it possible to estimate unbiased prevalences by combining the responses of the first and the complementary surveys. Nevertheless, such a response rate is unrealistic, and therefore two-phase sampling cannot guarantee to reach an unbiased estimate (and even a reduction of nonresponse bias) (6). Furthermore, this kind of design increases the variance (7).

The second option to reduce nonresponse bias is to use reweighting techniques (7). The major difficulty in effectively reweighting correction is identifying available relevant auxiliary data that are correlated both with probability of response and with outcome variables. One example of auxiliary data is known population totals; in this case, bias can be reduced using calibration techniques (8). If auxiliary data are individually available for respondents and nonrespondents, they can be used to estimate response probability, and the inverse probability weighting (IPW) technique can be used. Often, the sampling frame provides such auxiliary data (for instance sociodemographic data such as age or gender). However, auxiliary data may be unavailable or insufficient (9). In this context, one source increasingly used in survey statistics is paradata (10-13). These data are mainly collected in surveys using interviewers, and were originally defined as data related to the interview process, such as the frequency or timing of phone calls. In recent years, their definition has been extended to include information directly recorded by the interviewer during face-to-face interviews, for instance observations that characterize the interviewee's neighborhood, such as the density of stores (13).

Paradata are easily collected for respondents and nonrespondents and are generally strongly correlated with participation (13).

The objective of this study was first to evaluate the contribution of a complementary survey in decreasing nonresponse bias in order to estimate prevalences, and second, to assess the additional contribution of complementary survey's paradata in reducing this bias. We used the data of the Coset-MSA pilot study, a French occupational health surveillance survey.

The Coset-MSA study (14) aims to study health characteristics and morbidity trends in relation to occupational factors among workers in agriculture and related occupations, covered by a specific insurance fund, entitled Mutualité Sociale Agricole (MSA). It includes non-salaried workers (such as farmers and stud farm managers) and salaried workers (farm workers and employees of certain banks, insurance companies and food cooperatives).

A pilot study was planned with the objective of studying several methods to take nonresponse bias into account, in order to apply the most efficient protocol for the Coset-MSA generalization. The first method, whose results have been already published (14), illustrates the use of administrative health and occupational databases to correct for nonresponse. The present study focused on the contribution of a complementary survey and of paradata.

METHODS

The Coset-MSA study

The pilot study included workers aged between 18 and 65 years on December 31, 2008, who had worked at least 90 days in a workplace affiliated to the MSA insurance fund in 2008, in one of five French administrative areas (Bouches-du-Rhône, Pas-de-Calais, Pyrénées-Atlantiques, Saône-et-Loire, Finistère). The survey design featured two-phase sampling in a responsive design context (hereafter "two-phase sampling survey" or "combined surveys") (Figure 1) (3, 5). In the first phase, 10,000 people were randomly selected with stratification on gender, age, employment status (salaried vs. non-salaried worker) and geographical area. They received a 40-page self-administered postal questionnaire about working conditions and health. A postal reminder was mailed one month later. This first phase of the survey (hereafter the "initial survey") was conducted in February 2010, the response rate being 23.6%. In the second phase (hereafter "complementary survey"), a subsample of 100 nonrespondents in each area was then randomly selected and questioned by an interviewer in November 2010. Data collection was designed so as to achieve a maximum response rate (3, 7). All

the persons in this subsample received an information letter. A shorter questionnaire was administered by interviewer, which consisted of a selected number of questions from the questionnaire of the initial survey. To maximize the response rate to the complementary survey, our initial intention was to collect data by phone or by face-to-face interview if telephone contact was impossible. However, this kind of strategy can lead to problems interpreting results as differences in responses could be linked to the difference in data collection methods. We therefore decided to create two random groups, as follows: in each geographical area, 70 and 30 nonrespondents were randomly selected and designated to have phone and face-to-face interviews, respectively. If the designated type of interview was impossible (for instance, no phone number was available, or no one answered the telephone call), the interviewer switched to the alternative type. Up to 20 attempts for phone interviews and up to 3 visits for face-to-face interviews were made, at different times and days of the week, including Saturdays.

The study protocol was approved by the French institutional review committee (CNIL number 909091 and DR-2010-148).

Data

Outcome variables

Data from health and occupational databases

We studied variables from administrative databases for two reasons:

- they were available for respondents and nonrespondents to the postal questionnaire and so they enabled us to define a prevalence gold standard;
- they allowed us to specifically study nonresponse bias, independently of measurement bias. This is an important point given the possibility of differential measurement bias being induced because of the three modes of data collection used in the survey (postal mail, phone and face-to-face interviews) (15).

Two types of existing data were used: health-related data (hospitalization and reimbursement claims for medical services), extracted from the French Health Insurance Information System database (SNIIR-AM) and work-related data, extracted from the MSA databases.

In the present study, the variables used were:

- more than 100 reimbursement claims for medical services (including consultations, prescribed medicines, etc.) between 2008 and 2010;

- hospitalization between 2008 and 2010;
- working in a primary economic activity;
- job duration less than 10 years;
- sickness absence (at work);
- worker compensation for accident at work or occupational disease between 2002 and 2008.

Data from the questionnaire (available only for respondents to the initial or complementary surveys)

Questionnaire data included information on health status, current and past jobs and occupational exposure. In the present work, only three variables were analyzed for studying correction for nonresponse bias: self-rated health status as good; most recent occupational category (farmers, tradespeople and shopkeepers, managers and professionals, intermediate white-collar occupations, office and sales personnel, and manual workers) and perceiving intensity of physical effort at work as low. .

Data used to correct for nonresponse

Data from the MSA administrative database

Sociodemographic data (available for all individuals in the MSA database, so for both respondents and nonrespondents to the initial survey) were collected in 2008 and corresponded to stratification variables: gender, age, employment status (salaried vs. non-salaried worker) and geographical area.

Paradata from the interview process

Paradata (available for both respondents and nonrespondents to the complementary survey) were recorded just before data collection and at the same time as the questionnaire data collection.

Paradata available before the questionnaire data collection reflected the quality of the information available to contact the persons: accuracy of the postal mailing address and of phone number. A specific postal service was used to verify and update the postal mailing addresses provided by the MSA one year before the beginning of the study. If verification concluded that no change of address had occurred, the accuracy of the address provided by the MSA was considered good, otherwise it was considered poor. Phone numbers were checked in the phone directory using each individual's name (first name and last name) and the updated mailing address. If a phone number corresponding to the same name and mailing address was found, its accuracy was considered good. If a phone number

corresponded to the same name but not the same address, its accuracy was considered bad. The third modality of this variable was when no phone number was found.

At the moment of data collection, the following paradata were recorded: frequency of phone calls, frequency of phone calls after 5 p.m., frequency of phone calls on Saturdays, frequency of interviewer visits, and visits of an interviewer on Saturdays. A final variable corresponded the two planned modes of data collection (face-to-face and phone) and the possibility of switching mode.

Statistical analysis

Response rates

Response rates were estimated following the AAPOR's recommendations (16).

For the initial and complementary surveys, unweighted contact rates and response rates were computed.

For the two-phase sampling survey, weighted response rate was estimated.

Determination of the response probability models

Three multivariate logistic regressions were computed: one for the probability of response to the initial survey, and two for the probability of response to the complementary survey according whether paradata were used or not.

In each case, variables with a P-value less than 5% were kept in the final model.

Contribution of complementary survey and of paradata in reducing nonresponse bias

The study of the nonresponse bias was performed using variables derived from health and occupational databases. First, six estimations of the prevalence were computed (see supplementary material):

- with the whole sample to provide a prevalence gold standard: $\hat{p}_{GS}$;
- with data from the initial survey
    - without correction for nonresponse: $\hat{p}_1$
    - with correction for nonresponse using the sociodemographic variables: $\hat{p}_2$;
- with data of the combined surveys
    - without correction for nonresponse: $\hat{p}_3$;
    - with correction for nonresponse using the sociodemographic variables: $\hat{p}_4$

- with correction for nonresponse using the sociodemographic variables and paradata: $\hat{p}_5$

The validity of the correction for nonresponse to estimate $\hat{p}_2$ or $\hat{p}_4$ implicitly assumed that the nonresponse processes were missing at random (MAR) conditionally on the sociodemographic variables (or, for $\hat{p}_5$, sociodemographic variables and paradata. To correct for nonresponse, we used a reweighting technique called the equal-quantile score method (14, 17-19). The confidence intervals took into account the sampling design and the variance due to the nonresponse.

Then, the relative nonresponse errors were then computed:

$$\widehat{RE} = \frac{\hat{p}_i - \hat{p}_{GS}}{\hat{p}_{GS}} * 100 \; for \; i = 1, \ldots, 5$$

A relative error of less than 10% was considered acceptable (20). In line with Olson (21), these relative errors were assumed to correspond to nonresponse bias.

Prevalence estimates of questionnaire variables

Prevalence estimates of variables from questionnaires were computed both for the respondents of the initial survey with correction for nonresponse using the sociodemographic variables, and for the respondents of the initial and complementary surveys with correction for nonresponse using the sociodemographic variables or sociodemographic variables and paradata, respectively (see formulas in the supplementary materials section).

RESULTS

Response rate to the initial and complementary surveys

The response rate to the initial survey was 23.6% (95.0% of questionnaires were successfully delivered to the address provided and 24.8% of persons answered the questionnaire).

The response rate to the complementary survey was 62.6% (77.0% for the contact rate and 81.3% for the response rate among those contacted). Among respondents, 57% were interviewed by phone *vs* 43% by face-to-face interview.

The weighted response rate to combined surveys was estimated at 70.4%.

Response probability models

Initial survey

The response to the initial survey was significantly higher in the following groups: women (odds ratio (OR), 1.3; 95% confidence interval (CI): [1.2, 1.5]), older persons (OR, 1.6; 95% CI: [1.4, 1.8] for people aged between 50 and 65 years vs people aged between 18 and 34 years), salaried workers (OR, 0.8; 95% CI: [0.7, 0.8] for non-salaried vs salaried workers) and persons living in the Saône-et-Loire area (OR, 1.7; 95% CI: [1.4, 1.9] for people living in Saône-et-Loire vs people living in Bouches-du-Rhône) (Table 1).

Complementary survey

In the final model including only sociodemographic variables, response to the complementary survey was significantly greater in the following groups: non-salaried workers (unlike the initial survey) (OR, 2.5; 95% CI: [1.7, 3.7] for non-salaried vs salaried workers), and persons living in the Saône-et-Loire area (OR, 2.5; 95% CI: [1.3, 4.5] for people living in Saône-et-Loire vs people living in Bouches-du-Rhône) (Table 1).

Most of the paradata were significantly associated with the response to the complementary survey (results not shown). After taking into account paradata and sociodemographic data in a multivariate analysis (Table 2), the response to the complementary survey was higher in the following groups: non-salaried workers (OR, 2.4; 95% CI: [1.6, 3.6] for non-salaried vs salaried workers), for persons who were not visited by an interviewer visit on a Saturday (OR, 2.2; 95% CI: [1.0, 5.2]) or for whom planned data collection mode was the phone mode and there were no switch compared with other persons.

Contribution of complementary survey and of paradata in reducing nonresponse bias

Contribution of the complementary survey

For the six variables from administrative databases the relative errors (Table 3) using the data from the respondents to the initial survey decreased after correcting for nonresponse using sociodemographic variables. The same results were observed for the combination of the initial and complementary surveys, except for the variable “hospitalizations”.

After nonresponse correction by sociodemographic variables, the differences between the prevalence values estimated with the initial survey and with the combination of the initial and complementary surveys differed according to the dependent variables considered (Table 3). Relative errors for “over 100 reimbursement claims for medical services” were high (RE=22.3 for the initial survey; RE= 14.5

for the combined surveys), moderate for “hospitalizations” (RE=9.6 for the initial survey; RE= -6.5 for the combined surveys), and “sickness absence” (RE=5.3 for the initial survey; RE= -3.4 for the combined surveys) and low for the other variables. Nevertheless, irrespective of the variable studied, the prevalence estimated for the combined surveys was always closer to the gold standard prevalence than that estimated for the initial survey. It is worth noting that the confidence intervals for the combined surveys were at least twice as wide as those estimated for the initial survey.

Contribution of the paradata

Irrespective of the outcome variable considered, the prevalences estimated from the combined survey when taking and not taking paradata into account (Table 3) were slightly different.

Prevalence estimates of questionnaire variables

After correction for nonresponse using sociodemographic variables, the prevalence of self- rated health status as “good” ranged from 57.0% (95% CI [54.9, 59.1]) for the initial survey to 60.7% (95% CI [54.6, 66.8]) for the combined surveys. The proportion of farmers ranged from 29.7% (95% CI [27.6, 31.7]) for the initial survey to 34.1% (95% CI [29.1, 39.1]) for the combined surveys. For managers and professionals, the proportion ranged from 7.1% (95% CI [6.0, 8.3]) for the initial survey to 3.3% (95% CI [1.6, 5.0]) for the combined surveys. These differences were meaningful because the prevalences estimated after correction for nonresponse with sociodemographic variables in the combined surveys were not included in the CI estimated for the initial survey.

After correction for nonresponse using sociodemographic variables and paradata, the prevalences of those who self-rated health status as “good” and of farmers were unchanged.

Small differences were observed for the variable "perceiving intensity of physical effort at work as low” irrespective of the surveys (initial or combined surveys) and the nonresponse process assumption used (Table 4).

DISCUSSION

Our results suggest that although sociodemographic variables were associated with survey response rate, and despite the fact that numerous studies report that such variables are also related to health and occupation (1, 22, 23), they were not sufficient to correct for nonresponse bias in our initial survey. This is particularly true for the variable “over 100 reimbursement claims for medical services”. Our results also showed that a complementary survey is useful to correct for nonresponse bias. Moreover,

although paradata were strongly associated with response to the complementary survey, they were less useful in correcting for nonresponse after taking sociodemographic variables into account.

The response rate to the complementary survey (62.6%) was three times higher than that to the initial survey. This was to be expected since the response rate to the initial survey was particularly low (24, 25) and demonstrates the efficacy of the complementary study's protocol, which was designed to maximize the response rate.

Measurement errors may exist and may have been increased by the use of several types of data collection and the strong focus placed on contacting people. However, as we studied nonresponse bias for outcome variables not collected by questionnaires, but from administrative databases, our study is reliable for studying nonresponse error. Administrative databases were already used in previous studies for this purpose (21, 26). With regard to prevalence values of variables from the questionnaires, measurement errors could explain some of the differences observed between the values estimated for the initial and the combined surveys (3, 27).

The proportion of farmers estimated with the combined surveys was close to the exact proportion of non-salaried workers known from the MSA database (37%), which could be considered as a proxy of the "farmers" occupational category. This result suggests that the complementary survey was useful in correcting this proportion for nonresponse bias. The proportion of "managers and professionals" was probably better estimated with the combined survey. Indeed, farmers were perhaps less inclined to participate in this kind of survey than managers and professionals, perhaps because of the length and the complexity of the initial postal questionnaire. Thus, it seems that the proportion of the managers and professionals could be overestimated in the initial survey. Indeed, the question on the social category was probably most reliable with an interviewer, who can ask some precisions if the response is too vagueness, which is not the case for self-administered questionnaire.The usefulness of the complementary survey should be balanced with its cost. In the Coset-MSA study, the cost of collecting data from the 500 persons in the complementary survey and in the initial postal survey of 10 000 persons was the same, even though fewer variables were collected in the former.

Even though the response rate to the complementary survey was much higher than that to the initial survey, this does not imply that the complementary survey may have been useful in correcting for nonresponse bias if the respondents of the complementary survey shared the same profile as their initial survey counterparts. The results showed that these groups shared certain sociodemographic characteristics. However, they were probably different for others characteristics which we did not

measure. This is the reason why, some differences between the estimated prevalences remained, even after correction for nonresponse bias using sociodemographic variables.

The contribution of the complementary survey must also be discussed in terms of the range of the confidence intervals estimated from the combined surveys. They were at least as twice as wide as those estimated in the initial survey. The major part of the loss in precision came from the sampling frame, which led to very different weights for the respondents to the initial and complementary surveys, respectively (range 7 to 220). To reduce this range of weights, it would have been necessary to select a larger sample for the complementary survey, and the cost of the survey would have increased. Nevertheless, the loss of precision, which was already found in a similar epidemiological study (9), may be considered acceptable.

As mentioned above, paradata were strongly associated with nonresponse in the present study. This unsurprising result has already been found in numerous studies (11, 13). It is important to remember that the variable "interviewer visits for face-to-face interviews on a Saturday" is included in our final model. Our results support the recommendation of including indicators of time and day of interview in paradata (11). The result of the univariate analysis also supports the recommendation of including the availability of a phone number (28, 29). However, to be really relevant, paradata also have to be strongly correlated with outcome variables but it is generally not the case (13). In our study, the contribution of the paradata in correcting for the prevalence of the variables coming from health and occupational databases was weak, but this contribution was studied after a relevant correction for nonresponse bias on sociodemographic variables. If the prevalences estimated from the combined surveys had been corrected for solely using paradata, their contribution would have been greater for certain variables, for example "over 100 reimbursement claims for medical services" and "job duration less than 10 years" (results not shown). The use of paradata to correct for nonresponse is relatively recent and numerous questions about their exploitation remain (12). Even though they were not relevant in the present study, paradata are nevertheless potentially useful auxiliary data to consider when correcting for nonresponse bias (13, 30). Collecting paradata is worthwhile especially as those paradata related to the data collection process have the advantage of usually being inexpensive once the system is set up.

Even though our analysis showed the usefulness of the complementary survey, the MAR assumption, where the link between the variables of interest and nonresponse is completely explained by both sociodemographic variables and paradata, was not verified for all outcome variables coming from

administrative databases. Consequently, a residual bias remained for some outcomes and probably also for questionnaire variables.

REFERENCES

1. Galea S, Tracy M. Participation Rates in Epidemiologic Studies. Ann Epidemiol 2007; 17(9): 643-53
2. Bethlehem J. Weighting nonresponse adjustments based on auxiliary information. In: Groves RM, Dillman DA, Eltinge JL, et al., eds. Survey nonresponse. New-York: Wiley, 2002:275-88
3. Groves RM, Heeringa SG. Responsive design for household surveys: tools for actively controlling survey errors ans costs. J R Statist Soc A 2006; 169(439): 457
4. Wagner J, West BT, Kirgis N, et al. Use of Paradata in a Responsive Design Framework to Manage a Field Data Collection. Journal of Official Statistics 2012; 28(4): 477-99
5. Hansen M, Hurwitz W. The problem of nonresponse in sample surveys. JASA 1946; 41: 517-29
6. Groves RM, Peytcheva E. The impact of nonresponse rates on nonresponse bias: A meta-analysis. Public Opin Q 2008; 72(2): 167-89
7. Sarndal CE, Swensson B, Wretman J. Nonresponse. Model assisted survey sampling. New-York: Srpinger, 1992:556-600
8. Särndal KE, Lundström S. Estimation in Surveys with nonresponse. Chichester: Wiley-Blackwell; 2005
9. Jenkins P, Scheim C, Wang JT, et al. Assessment of coverage rates and bias using double sampling methodology. J Clin Epidemiol 2004; 57(2): 123-30
10. Beaumont JF. On the use of data collection process information for the treatment of unit nonresponse through weight adjustment. Survey Methodol 2005; 31(2): 227-31
11. Kreuter F, Kohler U. Analysing contact sequences in call record data. Potential and limitations of sequence indicators for nonresponse adjustments in the European Social Survey. Journal of Official Statistics 2009; 25(2): 203-26
12. Kreuter F. Improving surveys with paradata: analytic uses of process information. Wiley; 2013
13. Olson K. Paradata for nonresponse adjustment. The Annals of the American Academy of Political and Social Science 2013; 645: 142-70
14. Santin G, Geoffroy B, Benezet L, et al. In an occupational health surveillance study, auxiliary data from administrative health and occupational databases effectively corrected for nonresponse. J Clin Epidemiol 2014; 67(6): 722-30
15. Bowling A. Mode of questionnaire administration can have serious effects on data quality. J Public Health (Oxf) 2005; 27(3): 281-91
16. The American Association for Public Opinion Research. Standard definitions: final dispositions of case codes and outcome rates for surveys. 7th edition. AAPOR. 2011. Report
17. Eltinge JL, Yansaneh IS. Diagnostics for formation of nonresponse adjustment cells, with an application to income nonresponse in the U.S. consumer expenditure survey. Survey Methodol 1997; 23: 33-40
18. Haziza D, Beaumont JF. On the construction of imputation classes in surveys. Int Stat Rev 2007; 75(1): 25-43
19. Little RJA. Survey nonresponse adjustments for estimates of means. Int Stat Rev 1986; 54: 139-57
20. Vercambre MN, Gilbert F. Respondents in an epidemiologic survey had fewer psychotropic prescriptions than nonrespondents: An insight into health-related selection bias using routine health insurance data. J Clin Epidemiol 2012; 65(11): 1181-9
21. Olson K. Survey Participation, Nonresponse Bias, Measurement Error Bias, and Total Bias. Public Opin Q 2006; 70(5): 737-58

22. Goldberg M, Chastang JF, Leclerc A, et al. Socioeconomic, demographic, occupational, and health factors associated with participation in a long-term epidemiologic survey: a prospective study of the French GAZEL cohort and its target population. Am J Epidemiol 2001; 154(4): 373-84
23. Goldberg M, Luce D. Selection effects in epidemiological cohorts: Nature, causes and consequences. Rev Epidemiol Sante Publ 2001; 49(5): 477-92
24. Levy P, Lemeshow S. Sampling of populations, Methods and Applications. 2nd ed. New York: Wiley & Sons.; 1991
25. Stoop IAL, Billiet J, Koch A, et al. Improving survey response. Lessons learned from the European Social Survey. Chichester: Wiley; 2010
26. Kreuter F, Muller G, Trappmann M. Nonresponse and Measurement Error in Employment Research. Making use of Administrative Data. Public Opin Q 2010; 74(5): 880-906
27. Olson K. Do non-response follow-ups improve or reduce data quality? A review of existing literature. J R Statist Soc A 2013; 176: 129-45
28. Bethlehem JG, Cobben F, Schouten B. Handbook of nonresponse in household surveys. Hoboken: Wiley; 2011
29. Stoop IAL. The hunt for the last respondent. The Hague: Social and cultural planning office. 2014. Report
30. Maitland A, Casas Cordero C, Kreuter F. An exploration into the use of paradata for nonresponse adjustment in a health survey. Presented at Joint statistical meetings-Section on survey research methods,

FIGURE LEGENDS

Figure1. Two-phase sampling for nonrespondents

N: population size

IS: initial survey; CS: complementary survey

s_{IS}: random sample of IS (size: n_{IS})

s_{CS}: random sample of CS (size n_{CS})

$\pi_{IS,i}$: known probability sampling weight of the individual i for the initial survey

$\pi_{2ndphase/s_{IS,nr},i}$: known probability sampling weight of the individual i for the second phase conditionally on the respondents sample coming from the initial survey

$\pi_{CS/s_{IS,nr},i}$: known probability sampling weight of the individual i for CS conditionally on the nonrespondents sample coming from the initial survey

$s_{survey,r}$: the respondent sample from s_{survey} and $n_{survey,r}$ its size where survey is equal to IS or CS

$\delta_{CS,i}$: the unknown response probability for the survey CS that generates $s_{CS,r}$

Figure 1. Two-phase sampling for nonrespondents

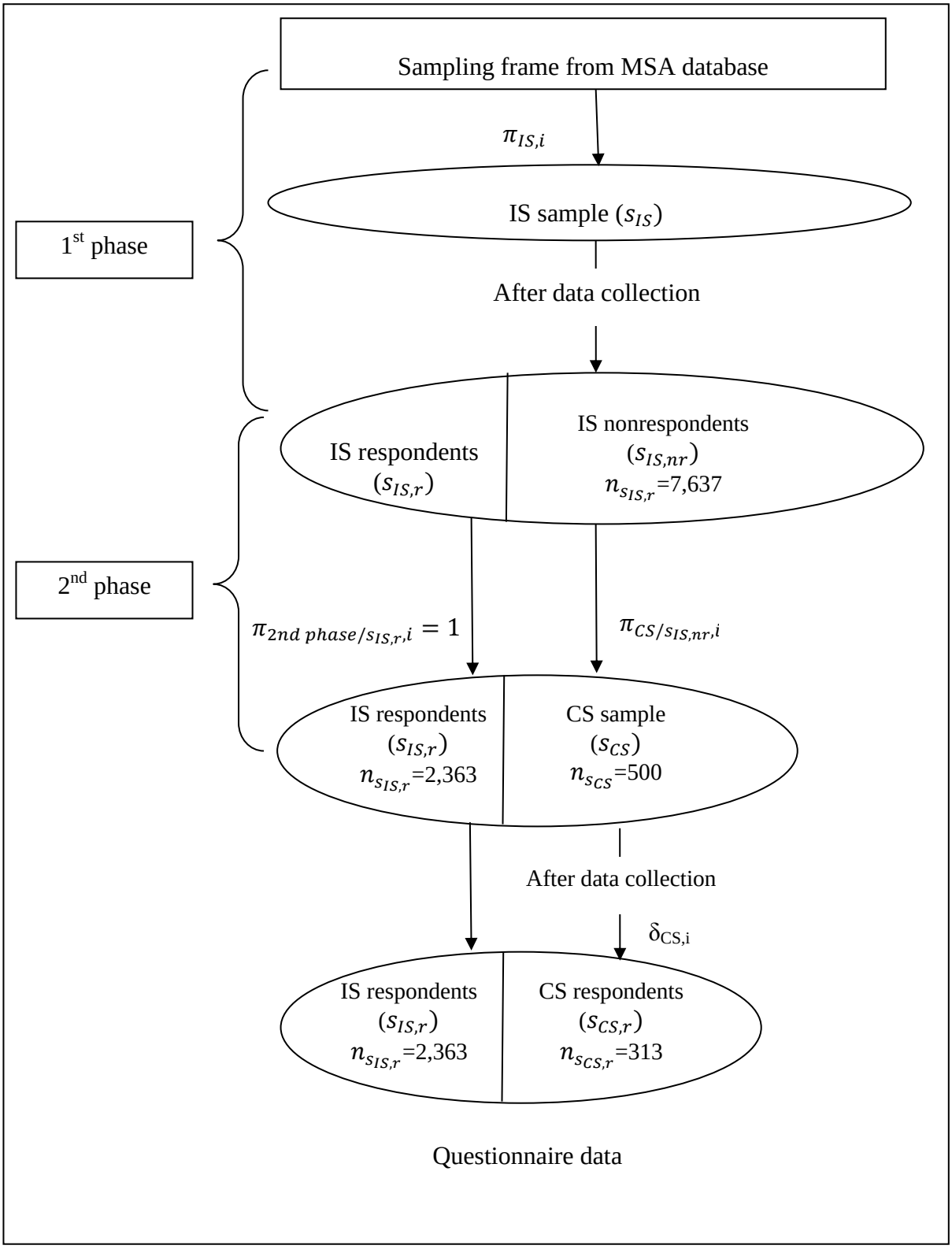


Table 1. Response to the surveys according to sociodemographic variables (final models)

	Initial survey			Complementary survey		
	n	**OR**	**95% CI**	**n**	**OR**	**95% CI**
Gender						
Male	6,775	1				
Female	3,225	1.3	1.2, 1.5			
Age (yr)						
18-24	2,677	1				
25-49	4,323	1.4	1.2, 1.6			
50-65	3,000	1.6	1.4, 1.8			
Employment status						
Salaried	5,845	1		281	1	
Nonsalaried	4,155	0.8	0.7, 0.8	219	2.5	1.7, 3.7
Geographic area						
Bouches-du-Rhône	2,000	1		100	1	
Finistère	2,000	1.4	1.2, 1.7	100	1.6	0.9, 2.9
Pas-de-Calais	2,000	1.2	1.0, 1.4	100	1.4	0.8, 2.5
Pyrénées-Atlantiques	2,000	1.3	1.1, 1.5	100	2.3	1.3, 4.3
Saône-et-Loire	2,000	1.7	1.4, 1.9	100	2.5	1.3, 4.5

Table 2. Response to the complementary survey according to sociodemographic variables and paradata (final models)

	n	**OR**	**95% CI**
Employment status			
Salaried	281	1	
Nonsalaried	219	2.4	1.6, 3.6
Interviewer visit for face-to-face interview on Saturdays			
Yes	31	1	
No	469	2.2	1.0, 5.2
Data collection evolution			
Phone at the beginning, phone at the end	199	1	
Phone at the beginning, face-to-face at the end	137	0.2	0.1, 0.3
Face-to-face at the beginning, face-to-face at the end	139	0.2	0.1, 0.4
Face-to-face at the beginning, phone at the end	25	0.1	0.0, 0.2

Table 3. Gold-standard prevalence and estimated prevalences of administrative database variables under several assumptions on nonresponse processes

	Respondents to the initial survey		Respondents to the initial survey and the complementary survey			Complete sample (gold standard)	Relative error				
	MCAR (a)	MAR(X) (b)	MCAR (c)	MAR(X) (d)	MAR(X,Z) (e)	(f)	(a-f) /f	(b-f) /f	(c-f) /f	(d-f) /f	(e-f) /f
	% 95% CI	% 95% CI	% 95% CI	% 95% CI	% 95% CI	% 95% CI					
More than 100 reimbursement claims for medical services	25.8 23.9, 27.6	24.3 22.5, 26.2	23.1 16.5, 29.6	22.8 17.7, 27.9	22.0 17.1, 26.9	19.9 19.1, 20.7	29.4	22.3	15.7	14.5	11.5
Hospitalization between 2008 and 2010	23.4 21.6, 25.2	22.9 21.1, 24.7	20.0 13.9, 26.0	19.5 14.9, 24.2	19.1 14.5, 23.7	20.9 20.1, 21.7	12.0	9.6	-4.3	-6.5	-8.6
Primary economic activity	57.2 55.1, 59.3	58.8 56.7,61.0	64.9 54.9, 74.9	61.8 54.7, 68.9	64.2 56.2, 72.2	61.1 60.1,62.1	-6.3	-3.8	6.2	1.1	5.1
Job duration less than 10 years	53.4 51.3, 55.4	53.7 51.5, 55.9	50.8 41.6, 60.0	54.8 46.7, 62.9	55.7 46.9, 64.5	56.4 55.4, 57.4	-5.5	-4.8	-10.1	-2.9	-1.2
Sickness absence at work	36.0 34.0, 38.0	35.8 33.6, 37.9	30.3 23.1, 37.5	32.8 26.2, 39.4	33.2 26.2, 40.2	33.4 33.0, 34.8	6.0	5.3	-10.7	-3.4	-0.6
Worker compensation for accident at work or occupational disease between 2002 and 2008	25.6 23.8, 27.4	26.4 24.5, 28.4	26.0 18.9, 33.0	26.4 20.8, 32.0	26.5 20.9, 32.1	26.9 25.9, 27.9	-4.7	-1.6	-3.2	-1.7	-1.5

X: set of sociodemographic variables
Z: set of paradata

Table 4. Estimated prevalences of questionnaire variables under several assumptions on nonresponse processes

	Initial survey			Combined surveys			
	n	MCAR % 95% CI	MAR(X) % 95% CI	n	MCAR % 95% CI	MAR(X) % 95% CI	MAR(X,Z) % 95% CI
Self-rated health status as good	1,307	55.8 53.7, 57.9	57.0 54.9, 59.1	1,501	59.6 51.6, 67.6	60.7 54.6, 66.8	59.4 52.5, 66.3
Occupational category							
Farmers	650	27.7 25.8, 29.7	29.7 27.6, 31.7	788	38.5 30.0, 46.9	34.1 29.1, 39.1	34.1 29.1, 39.1
Tradespeople ans shopkeepers	43	1.8 1.2, 2.3	2.4 1.6, 3.1	47	2.4 0.1, 4.8	2.4 0.1, 4.3	2.5 0.1, 4.4
Managers and professionals	155	7.7 6.4, 8.8	7.1 6.0, 8.3	163	3.7 1.7, 5.7	3.3 1.6, 5.0	2.8 1.8, 3.8
Intermediate white collar occupations	364	17.0 15.4, 18.6	16.6 15.0, 18.3	403	14.7 9.2, 20.2	16.5 11.4, 21.6	15.4 10.6, 20.2
Office and sales personnel	310	14.0 12.5, 15.5	12.6 11.2, 14.0	341	11.2 6.6, 15.8	11.6 7.7, 15.5	10.5 6.9, 13.9
Manual workers	675	31.8 29.8, 33.8	31.6 29.5, 33.7	762	29.6 22.1, 37.0	32.1 25.6, 38.6	34.8 27.2, 42.4
Low perceived intensity of physical effort	669	34.2 32.0, 36.4	32.9 30.7, 35.1	744	29.6 21.9, 37.4	31.1 24.4, 37.9	29.2 22.9, 35.6

X: set of sociodemographic variables
Z: set of paradata

Supplementary material: prevalences estimated from the initial survey $\widehat{p}_1$ and $\widehat{p}_2$, the combined surveys $\widehat{p}_3$ and $\widehat{p}_4$, and with the paradata $\widehat{p}_5$

Notations:

- Y: dichotomous variable and y_i value (0 or 1) for the individual i;

- p the population prevalence of Y we wish to estimate;

- X: set of the sociodemographic variables;

- Z: set of the paradata;

- *survey* = IS for initial suvey, CS for complementary survey;

- s_{survey} the random sample of IS or CS and n_{survey} its size;

- $\pi_{IS,i}$ the known probability sampling weight of the individual i for the initial survey. It takes into account the initial sampling framework (e.g. stratification);

- $\pi_{CS/s_{IS,nr},i}$ the known probability sampling weight of the individual i for CS conditionally on the nonrespondents sample coming from the initial survey;

- $s_{survey,r}$ the respondent sample from s_{survey} and $n_{survey,r}$ its size where *survey* is equal to IS or CS;

- MCAR[*survey*] for missing completely at random process for *survey* IS or CS;

- MAR(X)[*survey*] for missing at random process conditionally on X for *survey* IS or CS;

- $\delta_{[survey],i}$ the unknown response probability for the survey IS or CS that generates $s_{survey,r}$ and $\hat{\delta}_{MAR(X)[survey],i}$ its estimation under the MAR(X)[*survey*] assumption;

- $w_{IS,MCAR[IS],i}$ the weight used to estimate prevalence from the initial respondent sample $s_{IS,r}$ survey data under the MCAR assumption on IS;

- $w_{IS,MAR(X)[IS],i}$ the weight used to estimate prevalence from the initial respondent sample $s_{IS,r}$ survey data under the MAR(X) assumption on IS;

- $w_{(IS \cup CS),MCAR[CS],i}$ the weight used to estimate prevalence from the combined sample $s_{IS,r} \cup s_{CS,r}$ when nonresponse occurs under the MCAR assumption on CS;

- $w_{(IS \cup CS),MAR(X)[CS],i}$ the weight used to estimate prevalence from the combined sample $s_{IS,r} \cup s_{CS,r}$ when nonresponse occurs under the MAR(X) assumption on CS;

- MAR(X,Z)[CS] for missing at random process conditionally on X and Z for *survey* CS;

- $\hat{\delta}_{MAR(X,Z)[CS],i}$ the estimation of the response probability for the survey CS under the MAR(X,Z)[*CS*] assumption;

- $w_{(IS \cup CS),MAR(X,Z)[CS],i}$ the weight used to estimate prevalence from the combined sample $s_{IS,r} \cup s_{CS,r}$ when nonresponse occurs under the MAR(X,Z) assumption on CS;

- Prevalence estimated from the initial survey under MCAR assumption $\hat{p}_1$

This was estimated by:

$$\hat{p}_1 = \frac{\sum_{i \in s_{IS,r}} w_{IS,MCAR[IS],i} y_i}{\sum_{i \in s_{IS,r}} w_{IS,MCAR[IS],i}}$$

where $w_{IS,MCAR[IS],i} = \frac{1}{\pi_{IS,i}}$

- Prevalence estimated from the initial survey under the MAR(X) assumption $\hat{p}_2$

This was estimated by a reweighting technique that corrects for nonresponse on the sociodemographic variables:

$$\hat{p}_2 = \frac{\sum_{i \in s_{IS,r}} w_{IS,MAR(X)[IS],i} y_i}{\sum_{i \in s_{IS,r}} w_{IS,MAR(X)[IS],i}}$$

where $w_{IS,MAR(X)[IS],i} = \frac{1}{\pi_{IS,i}} * \frac{1}{\hat{\delta}_{MAR(X)[IS],i}}$

The use of the sociodemographic variables to correct for nonresponse corresponded to the implicit assumption that the nonresponse process is missing at random conditionally on the

sociodemographic variables X, labeled MAR(X)[IS]. In other words, associations between response probability and outcome variables were assumed to be completely explained by the sociodemographic variables. Here, correction for nonresponse was performed using a reweighting technique called the equal-quantile score method (1-3). The application of this method was explained in a previous article (4).

- Prevalence estimated from the combined surveys under the MCAR assumption $\hat{p}_3$

The design allowed us to use the two-phase sampling formula to estimate prevalence by combining the responses to the initial and complementary surveys (5).

The data from both surveys were combined and prevalence values estimated by:

$$\hat{p}_3 = \frac{\sum_{i \in s_{IS,r} \cup s_{CS,r}} w_{(IS \cup CS),MCAR[CS],i} y_i}{\sum_{i \in s_{IS,r} \cup s_{CS,r}} w_{(IS \cup CS),MCAR[CS],i}}$$

where $w_{(ISUCS),MCAR[CS],i} = \begin{cases} \frac{1}{\pi_{IS,i}} \; if \; i \in s_{IS,r} \\ \frac{1}{\pi_{IS,i} * \pi_{CS/s_{IS,nr},i}} if \; i \in s_{CS,r} \end{cases}$

- Prevalence estimated from the combined surveys under the MAR(X) assumption $\hat{p}_4$

To correct for nonresponse bias, the MAR(X) assumption on the nonresponse process $\delta_{[CS],i}$ was integrated to estimate the prevalence value of the combined samples $s_{IS,r}$ and $s_{CS,r}$. The response probability $\delta_{[CS],i}$ was estimated by $\hat{\delta}_{MAR(X)[CS],i}$ by the equal-quantile score method.

The data from the initial and the complementary surveys were combined and prevalence values were estimated by:

$$\hat{p}_4 = \frac{\sum_{i \in s_{IS,r} \cup s_{CS,r}} w_{(IS \cup CS),MAR(X)[CS],i} y_i}{\sum_{i \in s_{IS,r} \cup s_{CS,r}} w_{(IS \cup CS),MAR(X)[CS],i}}$$

where $w_{(IS \cup CS),MAR(X)[CS],i} = \begin{cases} \frac{1}{\pi_{IS,i}} \; if \; i \in s_{IS,r} \\ \frac{1}{\pi_{IS,i} * \pi_{CS/s_{IS,nr},i} * \hat{\delta}_{MAR(X)[CS],i}} if \; i \in s_{CS,r} \end{cases}$

- Prevalence estimated under the MAR(X,Z) assumption $\hat{p}_5$

In this case, the missing data process forthe complementary survey was assumed to be missing at random conditionally on sociodemographic variables (X) and paradata (Z), i.e. MAR(X,Z)[CS]. $\delta_{[CS],i}$ was then estimated by $\hat{\delta}_{MAR(X,Z)[CS],i}$ by the equal-quantile score method.

Prevalence values under the MAR(X,Z) process were then estimated by:

$$\hat{p}_5 = \frac{\sum_{i \in s_{IS,r} \cup s_{CS,r}} w_{(IS \cup CS),MAR(X,Z)[CS],i} y_i}{\sum_{i \in s_{IS,r} \cup s_{CS,r}} w_{(IS \cup CS),MAR(X,Z)[CS],i}}$$

where $w_{(IS \cup CS),MAR(X,Z)[CS],i} = \begin{cases} \frac{1}{\pi_{IS,i}} \; if \; i \in s_{IS,r} \\ \frac{1}{\pi_{IS,i} * \pi_{CS/s_{IS,nr},i} * \hat{\delta}_{MAR(X,Z)[CS],i}} if \; i \in s_{CS,r} \end{cases}$

REFERENCES

1. Eltinge JL, Yansaneh IS. Diagnostics for formation of nonresponse adjustment cells, with an application to income nonresponse in the U.S. consumer expenditure survey. Survey Methodol 1997; 23: 33-40

2. Haziza D, Beaumont JF. On the construction of imputation classes in surveys. Int Stat Rev 2007; 75(1): 25-43

3. Little RJA. Survey nonresponse adjustments for estimates of means. Int Stat Rev 1986; 54: 139-57

4. Santin G, Geoffroy B, Benezet L, et al. In an occupational health surveillance study, auxiliary data from administrative health and occupational databases effectively corrected for nonresponse. J Clin Epidemiol 2014; 67(6): 722-30

5. Hansen M, Hurwitz W. The problem of nonresponse in sample surveys. JASA 1946; 41: 517-29

ELSEVIER

Journal of Clinical Epidemiology 67 (2014) 722−730

Journal of Clinical Epidemiology

In an occupational health surveillance study, auxiliary data from administrative health and occupational databases effectively corrected for nonresponse

Gaëlle Santin[a,*], Béatrice Geoffroy[a], Laetitia Bénézet[a], Pauline Delézire[a], Juliette Chatelot[a], Rémi Sitta[b], Jean Bouyer[c], Alice Gueguen[b], SNIIR-AM Cohorts Group[a,b]

[a]*Department of Occupational Health, InVS French Institute for Public Health Surveillance, 12 rue du Val d'Osne, F-94415 Saint-Maurice, France*
[b]*CESP Center for Research in Epidemiology and Population Health U1018—Population-Based Cohorts' Research Platform, INSERM, Hôpital Paul Brousse, 16 Avenue Paul Vaillant-Couturier, F-94807 Villejuif, France*
[c]*CESP Center for Research in Epidemiology and Population Health U1018—Reproduction, Child Development Research Platform, INSERM, Hôpital Kremlin Bicêtre, 82 rue du Général Leclerc F-94276 Le Kremlin-Bicêtre, France*

Accepted 25 October 2013; Published online 31 January 2014

Abstract

Objectives: To show how reweighting can correct for unit nonresponse bias in an occupational health surveillance survey by using data from administrative databases in addition to classic sociodemographic data.

Study Design and Setting: In 2010, about 10,000 workers covered by a French health insurance fund were randomly selected and were sent a postal questionnaire. Simultaneously, auxiliary data from routine health insurance and occupational databases were collected for all these workers. To model the probability of response to the questionnaire, logistic regressions were performed with these auxiliary data to compute weights for correcting unit nonresponse. Corrected prevalences of questionnaire variables were estimated under several assumptions regarding the missing data process. The impact of reweighting was evaluated by a sensitivity analysis.

Results: Respondents had more reimbursement claims for medical services than nonrespondents but fewer reimbursements for medical prescriptions or hospitalizations. Salaried workers, workers in service companies, or who had held their job longer than 6 months were more likely to respond. Corrected prevalences after reweighting were slightly different from crude prevalences for some variables but meaningfully different for others.

Conclusion: Linking health insurance and occupational data effectively corrects for nonresponse bias using reweighting techniques. Sociodemographic variables may be not sufficient to correct for nonresponse.

Keywords: Unit nonresponse; Selection bias; Reweighting; Health insurance data; Occupational data; Surveillance

1. Introduction

A decline in participation rates in epidemiological studies has been observed in recent decades [1]. It is a particular concern in epidemiological surveillance surveys that aim to provide descriptive statistics that may be extrapolated to a target population. A nonresponse bias occurs when the response probability (also called response propensity) and the outcome variable are correlated [2]. It can be corrected when this correlation may be completely explained by a known set of variables. Two main techniques can be used for dealing with nonresponse [3]. The first is imputation, which consists of modeling the outcome variable and replacing each missing item of data by its predicted value. The second is reweighting, which broadly consists of modeling the response probability and then reweighting data by the inverse of the estimated response probability for each subject so-called inverse probability weighting (IPW). The use of imputation is generally recommended for partial nonresponse (subjects answered a questionnaire but did not fill in all the questions), and reweighting is recommended for unit nonresponse (subjects did not answer a questionnaire at all) [3,4]. As we focus on participation in epidemiological studies, we are specifically interested in unit nonresponse and thus in reweighting. Still, it should be noted that imputation, as well as reweighting, require that some data should be known on both respondents and nonrespondents. This may be particularly challenging for unit nonresponse, but it can be done

Conflict of interest: None.
Funding: None.

* Corresponding author. Tel.: +33-(0)155-12-54-28; fax: +33-(0)141-79-67-88.
E-mail address: g.santin@invs.sante.fr (G. Santin).

0895-4356/$ - see front matter
http://dx.doi.org/10.1016/j.jclinepi.2013.10.017

What is new?

What this adds to what was known?

- This study shows not only the interest of linking routine health insurance and occupational data to study nonresponse bias but also how these data can be taken into account to use response probability to estimate prevalences by reweighting techniques.

What is the implication and what should change now?

- In an epidemiological surveillance survey, it is not sufficient to correct nonresponse bias solely with sociodemographic variables. The health and occupation-related data available for both respondents and nonrespondents should also be used.

when survey data can be linked to existing databases such as medical administrative databases containing health-care, occupational, or sociodemographic information [5]. The aim is then to model accurately the probability of response using the variables available. Several epidemiological studies have already addressed this issue and have shown that nonresponse is associated with gender, age, marital status, unhealthy lifestyle, healthcare reimbursement, or occupational status [6−10]. Few studies, however, have used these results to correct the prevalence estimates for nonresponse bias [11,12]. Reweighting methods are in fact rarely used and are poorly known in the epidemiological community.

The principal objective of the present study was to show how reweighting can correct for unit nonresponse bias in an occupational surveillance survey by using data from administrative databases related to health and occupation, in addition to the sociodemographic data traditionally used. We then evaluated the impact on prevalence estimates of reweighting corrections with these auxiliary data.

2. Population and methods

2.1. The Coset-MSA cohort

The Coset-MSA study is part of the overall Coset program (Cohort for Epidemiological Surveillance in Connection with Occupation), which aims to study health characteristics and morbidity trends in relation to occupational factors in the French working population [13]. This program relies on data from three cohorts of individuals insured through the three main social welfare funds in France, which cover 95% of the population: the Constances cohort [14], conducted by the French National Institute for Health and Medical Research (INSERM), and the Coset-MSA and the Coset-RSI cohorts conducted by the French National Institute for Health Surveillance (InVS).

The Coset-MSA cohort focuses on workers in agriculture and related occupations covered by the corresponding insurance fund, the Mutualité Sociale Agricole (MSA). It includes nonsalaried workers (such as farmers and stud farm managers) and salaried workers (farm workers and some bank, insurance, or agricultural cooperative employees). Before setting up the Coset-MSA cohort in the whole of France, a pilot study was conducted in 2010 and was used as a basis for the present report.

The pilot study included workers aged between 18 and 65 years on December 31, 2008, who had worked at least 90 days in a workplace affiliated to the MSA insurance fund in 2008, in one of five French administrative areas (Bouches-du-Rhône, Pas-de-Calais, Pyrénées-Atlantiques, Saône-et-Loire, and Finistère). In each area, 2,000 individuals were randomly selected from the MSA database after stratification for gender, age, and employment status (salaried vs. nonsalaried worker).

The study protocol was approved by the French Institutional Review Committee (CNIL number 909091 and DR-2010-321).

Finally, after excluding persons who could not be contacted by post ($n = 406$) and refusals ($n = 236$) of data extraction from the SNIIR-AM and MSA databases, a total of 9,358 persons were included in the present study.

2.2. Data

Two types of data were collected and matched (only 0.2% of linkages failed because of the change in personal data in the interval between the sampling procedure and database extraction):

1. First, a self-completed postal questionnaire (40 pages) with a postal reminder 1 month later.

This concerned information on health status, current and past jobs, and current and past occupational exposures. In the present study, only three questions were analyzed: self-rated health status (very good/good vs. moderate/poor), last occupational category for persons who had worked more than 4 months in their life (farmers, tradespeople and shopkeepers, managers and professionals, intermediate white-collar occupations, office and sales personnel, and manual workers), and a question taken from the effort/reward imbalance at work questionnaire for persons at work in date of the study [15] (“Considering all my efforts and achievements, I receive the respect and prestige I deserve at work”). The first two questions were chosen because they are global indicators of health and occupation, respectively. The last one was chosen because it is a subjective constraint at work, which could be difficult to correct for nonresponse. Because of the small partial nonresponse to these questions (<5%), item nonresponse was not treated here.

2. Second, data extracted from existing administrative databases; these data will be referred to as auxiliary data.

Two types of existing databases were used: health-related data (hospitalization and reimbursement claims for medical services) extracted from the French Health Insurance Information System database (SNIIR-AM), which covers the whole French population, and work-related and sociodemographic data extracted from the MSA databases.

Health-related data covered information on hospitalizations in 2008 or 2009, reimbursement claims for medical services (general practitioner, specialist, laboratory analyses, medical auxiliary, or medical supplier), and the number of boxes of drugs reimbursed between 2008 and 2010. The cutoffs of the variables were chosen according to the distribution variables and to distinguish small, moderate, and big medical consumptions. Work-related data were recorded in 2008 or 2009 according to the date of the last entry, corresponding to an individual's last contribution to the MSA health insurance fund, and included indicators of most recent employment (employment status, last entry, economic activity, and job duration). Sociodemographic data were recorded in 2008 and corresponded to stratification variables (gender, age, geographical area, and employment status).

Because of the linkage failure ($n = 19$) and some missing data related to the last principal job at date ($n = 32$), the study was in fact based on 9,307 persons.

2.3. Statistical analysis

Auxiliary data were known for the whole sample, and data from the questionnaire were only known for respondents.

Statistical analysis was conducted in three steps, as detailed in the following.

1. Determination of the variables to be used to model the probability of response to the postal questionnaire

We used logistic regression with ''respondent'' as the dependent variable. In view of the number of variables and their correlations, we first studied variables by groups (sociodemographic characteristics, health-care utilization indicators, and work-related indicators) and selected those associated with response at $P < 0.2$.

In the second stage, the selected variables were considered together. To avoid including too many variables in the final model, only those associated with response at $P < 0.05$ were finally retained.

2. Estimations of the prevalence of questionnaire variables after correction for the nonresponse bias.

This was the principal analysis. We used the following notations:

- s: the sample selected by random drawing and n its size;
- s_r the respondent sample and n_r its size;
- y the variable whose prevalence we wish to estimate and y_i its value for the individual i;
- d_i the sampling weight of the individual i. This is the weight of the initial sampling framework (eg, stratification), independent of the nonresponse process.

Different estimations of prevalence can be derived according to the assumption of the nonresponse process: missing completely at random (MCAR) and missing at random (MAR).

2.1. In the case of an MCAR process, there is no association between the response probability and the outcome variable. Thus, no correction for nonresponse is needed and the estimation is

$$\overline{y}_{s_r;\mathrm{MCAR}} = \frac{\sum_{i=1}^{n_r} d_i y_i}{\sum_{i=1}^{n_r} d_i},$$

where $\overline{y}_{s_r;\mathrm{MCAR}}$ is the mean of y estimated on the sample s_r with the assumption MCAR.

2.2. In the case of an MAR process, associations between the response probability and the outcome variables of the postal questionnaire are completely explained by auxiliary data, for example, the variables studied in step 1. Outcome prevalence can be estimated by a reweighting technique:

$$\overline{y}_{s_r;\mathrm{MAR}} = \frac{\sum_{i=1}^{n_r} w_i y_i}{\sum_{i=1}^{n_r} w_i},$$

where $w_i = d_i \times \mathrm{corr}_i$ and corr_i is the weight correction for nonresponse.

The principle of weight correction consists in modifying original survey weights to correct for nonresponses, assuming that for a given response probability, nonrespondents would have, on average, given the same answer as respondents [16]. Robustness considerations lead to a recommendation of categorizing predicted response probabilities for reweighting using the equal-quantile score method [16−18]. This consists in sorting predicted values of the nonresponse model and grouping them in k equal size groups. It is recommended to constitute between 5 and 25 equal size groups [16] as estimations were stable between 10 and 25 groups, here k equals 10. The inverse observed response rate in each group is then computed and used as corr_i. It should be noted that in the survey statistics literature, weight correction is traditionally called weight adjustment, but as adjustment is often used in epidemiology in regression analysis with several covariates, we decided to use ''weight correction'' to avoid confusion.

3. Sensitivity analyses

Two sensitivity analyses were conducted to assess the reliability and usefulness of step 2 in correcting for nonresponse.

3.1. First, we first evaluated the contribution of health and work-related data. For this purpose, we corrected for nonresponse only with sociodemographic data. This corresponds to the MAR2 assumption that the associations between response probability and outcome variables were completely explained by sociodemographic data (ie, stratification data in our study), which are the classic variables used to correct for nonresponse when it is not possible to collect other auxiliary data.

The estimator was thus

$$\overline{y}_{s_r;\mathrm{MAR2}} = \frac{\sum_{i=1}^{n_r} w_{2,i} y_i}{\sum_{i=1}^{n_r} w_{2,i}},$$

where $w_{2,i} = d_i \times \mathrm{corr}_{2,i}$ and $\mathrm{corr}_{2,i}$ is the correction factor for nonresponse under a MAR2 assumption.

3.2. Second, we took advantage of the fact that some variables were known in the administrative databases that were recorded for the whole sample and thus provided a gold standard with which corrected estimation on the respondent sample could be compared.

Two variables were considered: frequency of reimbursement claims for medical services and economic activity. Letting $\overline{y}_s$ be the gold standard (unbiased) prevalence, we computed the relative error (RE) for each nonresponse process assumption:

$$\mathrm{RE} = \frac{\overline{y}_{s_r;\mathrm{nr\_process}} - \overline{y}_s}{\overline{y}_s} \times 100,$$

where nr_process = MCAR, MAR2, or MAR.

An RE less than 10% was considered as acceptable [10].

3. Results

Of the 9,307 persons included in the study, 57.2% were salaried, 67.6% were men, and their median age was 43 years. Around 90% of the sample had claimed for reimbursement of medical services and 10% had been hospitalized. The date of last entry was in 2008 for 4% of the sample and 2009 for the remaining 96%. The participation rate in the postal questionnaire was 24.8% (2,320 respondents).

3.1. Determination of the variables used to model the probability of response

A large number of variables were studied, most of which were significantly associated with response to the questionnaire in bivariate analysis. Variables selected in the final logistic model are displayed in Table 1.

Response to the questionnaire was associated with age and geographical area; people living in Finistère [odds ratio (OR), 1.1; 95% confidence interval (CI): [1.0, 1.4]], Pas-de-Calais (OR, 1.4; 95% CI: [1.2, 1.6]), and Saône-et-Loire (OR, 1.5; 95% CI: [1.3, 1.8]) and older people were more likely to respond. In the final model, which included health-related data, gender was no longer associated with response.

People who had received reimbursement for medical or paramedical services (general practitioner, specialist, dentist, laboratory analyses, medical auxiliary, or medical suppliers) responded more frequently than those who had not, whereas reimbursement for a large number of boxes of drugs during the year (OR, 0.6; 95% CI: [0.4, 0.8]) or hospitalization (OR, 0.9; 95% CI: [0.8, 1.0]) were associated with lower participation.

Salaried employees were more likely to respond than nonsalaried workers (OR, 1.4; 95% CI: [1.2, 1.7]). People with a recent last entry (2009) in the MSA occupational database responded more frequently than people with less recent entries. The length of employment in the last main job was also associated with nonresponse; people with job duration greater than 6 months responded more frequently than others (OR, 1.3; 95% CI: [1.0, 1.8] for job duration between 6 months and 1 year). Furthermore, people who worked in a service company (such as a bank or insurance company; OR, 1.3; 95% CI: [1.1, 1.6]) were more likely to respond than specialist crop producers.

3.2. Estimations of the prevalence of questionnaire variables under MCAR and MAR assumptions

The predicted values of the response probabilities yielded by the regression model varied between 2% and 62%. Fig. 1 shows that the distributions of the respondents and the nonrespondents overlapped each other sufficiently to allow the use of the equal-quantile score method to build robust homogeneous response groups with response rates ranging from 10% to 45%.

The prevalence of self-rated health status as good/very good ranged from 55.8% (95% CI: [53.8, 57.9]) under MCAR assumption (corresponding to no correction for nonresponse) to 58.5% (95% CI: [56.2, 60.7]) under MAR assumption. These differences were meaningful because the prevalence under MAR assumption was not included in the CI estimated under MCAR assumption (Table 2).

Relative differences between MCAR and MAR assumptions on the estimation of the prevalence of occupational category were greater. For instance, under MCAR and MAR assumptions, the proportions of farmers were 27.7% and 30.8%, respectively and the proportions of manual workers were 31.7% and 34.1%, respectively.

Whatever the assumptions used, the proportion of persons who considered that they had the respect they deserved at work remained almost unchanged.

3.3. Sensitivity analysis

Differences between prevalence estimates from questionnaire variables under MAR and MAR2 assumptions

Table 1. Variables associated with response to postal questionnaire in the final model

Variable	N	OR (95% CI)	P
Gender			
Male	6,284	1.0	
Female	3,023	1.0 (0.9, 1.1)	
Age (yr)			
18–34	2,416	1.0	**
35–49	4,063	1.2 (1.1, 1.4)	
50–65	2,828	1.3 (1.1, 1.5)	
Geographical area			
Bouches-du-Rhône	1,742	1.0	***
Pyrénées-Atlantiques	1,865	1.1 (0.9, 1.3)	
Finistère	1,882	1.1 (1.0, 1.4)	
Pas-de-Calais	1,911	1.4 (1.2, 1.6)	
Saône-et-Loire	1,907	1.5 (1.3, 1.8)	
Reimbursement claims for medical services between 2008 and 2010			
General practitioner			
None	813	1.0	**
1–4	2,437	1.5 (1.1, 2.0)	
5–9	2,378	1.6 (1.2, 2.2)	
≥10	3,679	1.4 (1.0, 2.0)	
Specialist			
None	1,954	1.0	***
1–4	3,319	1.1 (0.9, 1.2)	
5–9	1,865	1.2 (1.0, 1.4)	
≥10	2,169	1.6 (1.3, 1.9)	
Dentist			
None	3,068	1.0	***
1–9	4,895	1.3 (1.2, 1.5)	
At least 10	1,344	1.2 (1.0, 1.4)	
Laboratory analyses			
None	3,243	1.0	***
1–9	4,733	1.3 (1.1, 1.5)	
≥10	1,331	1.4 (1.2, 1.7)	
Medical auxiliary			
None	4,626	1.0	*
1–9	2,405	1.0 (0.9, 1.1)	
≥10	2,276	1.2 (1.0, 1.4)	
Medical supplier			
None	5,437	1.0	*
≥1	3,870	1.1 (1.0, 1.3)	
Medical prescription boxes of drugs, *N*			
0	611	1.0	**
1–199	4,076	0.8 (0.6, 1.1)	
200–799	3,253	0.7 (0.5, 1.0)	
≥800	1,367	0.6 (0.4, 0.8)	
Hospitalization between 2008 and 2009			
None	7,361	1.0	*
≥1	1,946	0.9 (0.8, 1.0)	
Last principal job to date (2008 or 2009)			
Employment status			
Nonsalaried	4,031	1.0	***
Salaried	5,276	1.4 (1.2, 1.7)	
Last entry			
2008	482	1.0	*
2009	8,825	1.4 (1.1, 1.7)	
Economic activity			
Specialist crop production	1,833	1.0	***
Wine growing	490	0.9 (0.7, 1.2)	
Large livestock raising	1,969	1.0 (0.9, 1.2)	
Small livestock raising	316	1.1 (0.8, 1.5)	
Mixed crops and livestock farming	1,118	0.9 (0.7, 1.0)	
Stud farming	161	0.7 (0.4, 1.1)	
Forestry and logging	189	0.9 (0.6, 1.3)	
Agricultural works company	997	0.7 (0.6, 0.9)	
Rural handicraft production	50	0.2 (0.1, 0.6)	
Agricultural cooperatives	971	0.9 (0.8, 1.1)	
Service company	1,070	1.3 (1.1, 1.6)	
Miscellaneous occupations	143	1.1 (0.7, 1.6)	
Job duration			
<6 mo	531	1.0	**
6 mo–1 yr	639	1.3 (1.0, 1.8)	
1–6 yr	2,660	1.4 (1.1, 1.9)	
6–10 yr	1,429	1.2 (0.9, 1.6)	
10–20 yr	2,092	1.6 (1.2, 2.1)	
>20 yr	1,956	1.5 (1.1, 2.0)	

Abbreviations: OR, odds ratio; CI, confidence interval.
*$P < 0.05$; **$P < 0.01$; ***$P < 0.001$.

were generally moderate (Table 2). The greatest difference was for the occupational category, in particular salaried occupations (intermediate white-collar occupations, office and sales personnel, and manual workers).

Relative errors when comparing corrected prevalence values of variables from existing databases under MCAR, MAR2, and MAR assumptions with a gold standard are presented in Table 3. They are clearly smaller under MAR assumption, in which not only sociodemographic variables but also health-care utilization indicators and work-related indicators are taken into account to estimate the corrected prevalences.

4. Discussion

In the Coset-MSA study, response probability was related to not only sociodemographic variables but also health and occupational variables. Comparison of the estimated prevalence on outcome variables yielded by questionnaires, under different assumptions on the missing data process (MCAR or MAR), showed moderate but noticeable differences whose magnitude varied according to the variables studied; these differences reflect the association between estimated response probabilities and outcome variables.

The factors associated with the probability of response are globally consistent with those reported in the literature. Many epidemiological studies have found that older people participate more than younger people [7,10,19–21]. In agreement with other publications [7,8,10,19], we found that men participated less than women. But in our study, this association did not persist after adjustment on health-related data. People who take care of their health (those who had been reimbursed for medical or paramedical services) were more likely to respond, and people with serious health problems (those who had been reimbursed for medical prescriptions or hospitalization) were less likely to respond. Most of these associations have already been encountered in

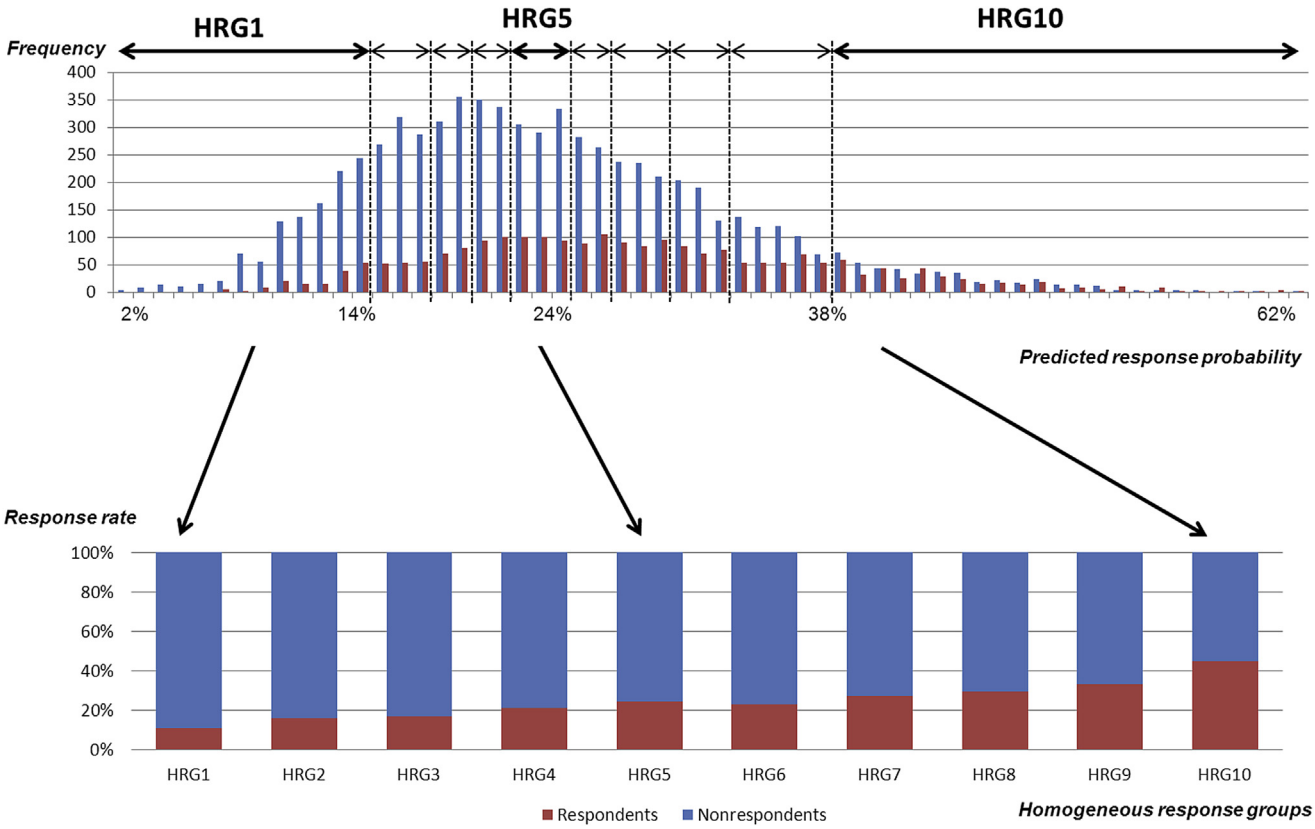


Fig. 1. Construction of 10 homogeneous response groups (HRG1 to HRG10) by the equal-quantile score method.

epidemiological studies [7,10,20–22], except for medical prescription reimbursement [9,10,20,21] in which a positive association between this variable and participation had been found. Nevertheless, in these studies, medical prescription reimbursement was often studied separately from the other types of medical care; in our study, we first found a crude positive association between response and medical prescription reimbursement, but the direction of this association changed after correction for the number of general practitioner visits in the model. With regard to occupational variables, it is more difficult to compare our results with those reported in the literature. The occupational category was not available in our data, and only proxies of occupational category have previously been studied [6,8,20,23]. Given the size of the questionnaire (40 pages) and the data collection procedure (self-administered questionnaire), we hypothesize that salaried people and service company workers have more availability so were more prone to respond. Moreover, people whose last job was affiliated to the MSA health insurance fund in 2009 had a better response rate probably because they felt more involved in the study. Furthermore, participation was lower among people with a less stable job (duration less than 6 months).

Even if all the auxiliary data were not collected for an epidemiological purpose and could also be affected by measurement errors, these data have the crucial advantage of being collected from the random sample independently of response or nonresponse to the survey. Furthermore, to model the response probability with variables also associated with the outcomes of interest, another strength of our study was that numerous and various data can be used: sociodemographic, health-related, and occupation-related data. This is all the more important because our study is related to health and work and because it is necessary to correct for variables associated with both outcome variables and response probability. The second sensitivity analysis on the prevalences of variables derived from administrative databases supported that our response model is well specified as we were in fact able to modify correctly the prevalence values for variables for which we knew the true value. The data we used appear strongly relevant to correct for nonresponse. In particular, the value of using work- and health-related variables in addition to classic sociodemographic ones is clearly shown by the differences between MAR and MAR2 estimates and by sensitivity analyses. The collection of efficient auxiliary data is not easy. They must be strongly related to the topic of the study, homogenous, and centralized with a linkage possibility; furthermore, an agreement with the data keepers and a legal authorization can be difficult. Nevertheless, efforts are recommended [24] and fructuous because they allow in one hand to take into account the unit nonresponse and in other hand to conduct efficient sensitivity analysis on the complete sample, with observable data strongly correlated with the

Table 2. Corrected prevalence of variables from postal questionnaires under MCAR, MAR,[a] and MAR2[b] processes of missing data among respondents

Variable	*N*	MCAR % (95% CI)	MAR % (95% CI)	MAR2 % (95% CI)
Perceived general health status				
Very good/good	1,283	55.8 (53.7, 57.9)	58.4 (56.2, 60.7)	56.9 (54.8, 59.1)
Moderate/poor	1,036	44.1 (42.0, 46.2)	41.5 (39.2, 43.7)	43.0 (40.8, 45.1)
Occupational category				
Farmers	650	27.7 (25.8, 29.6)	30.8 (28.6, 33.0)	29.6 (27.6, 31.7)
Tradespeople and shopkeepers	43	1.9 (1.3, 2.5)	3.1 (2.0, 4.1)	2.3 (1.6, 3.0)
Managers and professionals	155	7.5 (6.3, 8.7)	5.9 (4.9, 6.9)	7.1 (5.9, 8.2)
Intermediate white-collar occupations	364	16.9 (15.3, 18.6)	14.4 (12.9, 16.0)	16.6 (14.9, 18.2)
Office and sales personnel	310	13.9 (12.4, 15.4)	11.4 (10.0, 12.8)	12.5 (11.1, 14.0)
Manual workers	675	31.7 (29.7, 33.7)	34.1 (31.8, 36.4)	31.6 (29.5, 33.7)
I receive the respect and prestige I deserve at work				
Agree	1,144	59.4 (57.1, 61.7)	60.2 (57.7, 62.7)	59.4 (57.0, 61.7)
Disagree and not distressed	256	13.3 (11.7, 14.9)	13.5 (11.7, 15.2)	13.4 (11.8, 15.1)
Disagree and somewhat distressed	315	16.5 (14.8, 18.2)	16.3 (14.4, 18.1)	16.5 (14.7, 18.3)
Disagree and distressed	131	6.8 (5.6, 7.9)	6.2 (5.1, 7.3)	6.6 (5.4, 7.8)
Disagree and very distressed	75	3.8 (2.9, 4.7)	3.6 (2.7, 4.6)	3.9 (2.9, 4.8)

Abbreviations: MCAR, missing completely at random; MAR, missing at random; CI, confidence interval.

[a] MAR assumption: the relation between outcome variables and response probability is completely explained by sociodemographic, health-, and occupation-related variables.

[b] MAR2 assumption: the relation between outcome variables and response probability is completely explained by sociodemographic variables.

questionnaire data. Furthermore, auxiliary data also can be used to estimate disease frequency [25].

The response rate in our study was rather low (25%). Nevertheless, even if response rate is theoretically a factor in the nonresponse bias formula [2], a literature review [26] has shown that nonresponse rate is not the major concern in nonresponse bias. The main concern is the relevance of the data used to correct for nonresponse bias. In our study, the focus was on occupational health, and the data used to correct nonresponse concerned health and occupation, but it is possible that a small residual bias remains.

This study focused on unit nonresponse in a descriptive perspective. Unit nonresponse occurs when a subject does not respond to a survey (refusal or inability to contact). In this situation, reweighting, a technique originating from survey statistics, is recommended to correct for nonresponse bias. This consists in correcting survey weights of respondents to compensate for the absence of the nonrespondents [16]. There are two main methods for computing corrected weights. First, calibration [27], which consists in considering a reference population (census or reference survey) that provides reference distributions for some key variables. The initial survey weights are then modified so that the distribution of these key variables in the present survey coincides with the reference distribution. Calibration does not require individual information on nonrespondents, but convergence problems can arise when numerous auxiliary data are used. Therefore, calibration is usually used when no or little individual auxiliary data are available, which was not the case in our study. The second method is reweighting by IPW, which also consists in correcting initial survey weights. The corrected weights result from the

Table 3. Prevalence of existing database variables estimated from the complete sample (gold standard) and respondents to the questionnaire under MCAR, MAR,[a] and MAR2[b] processes of missing data

Variable	MCAR % (95% CI)	MAR2 % (95% CI)	MAR % (95% CI)	Gold standard (GS) % (95% CI)	Relative error MCAR/GS	MAR2/GS	MAR/GS
Reimbursement claims for medical services, *N*							
None	2.8 (2.1, 3.5)	2.7 (2.0, 3.4)	4.3 (3.1, 5.5)	4.4 (3.9, 4.8)	−35.1	−37.3	−2.0
1–49	43.4 (41.4, 45.5)	45.6 (43.4, 47.8)	49.0 (46.7, 51.3)	49.5 (48.5, 50.6)	−12.2	−7.9	−1.0
50–499	27.8 (26.0, 29.7)	27.2 (25.3, 29.1)	25.1 (23.2, 27.0)	26.0 (25.1, 27.0)	6.8	4.3	−3.6
≥500	25.7 (23.9, 27.6)	24.3 (22.5, 26.2)	21.4 (19.7, 23.2)	19.9 (19.0, 20.7)	29.3	22.3	7.8
Economic activity							
Primary	57.2 (55.1, 59.3)	58.8 (56.6, 60.9)	61.8 (59.6, 64.0)	61.1 (60.1, 62.1)	−6.4	−3.8	1.0
Secondary	6.2 (5.2, 7.2)	5.6 (4.6, 6.6)	5.8 (4.8, 6.9)	6.5 (6.0, 7.0)	−4.7	−13.6	−9.8
Tertiary	36.5 (34.5, 38.6)	35.5 (33.4, 37.6)	32.2 (30.1, 34.3)	32.3 (31.3, 33.2)	13.1	9.9	−0.1

Abbreviations: MCAR, missing completely at random; MAR, missing at random; CI, confidence interval.

[a] MAR assumption: the relation between outcome variables and response probability is completely explained by sociodemographic, health-, and occupation-related variables.

[b] MAR2 assumption: the relation between outcome variables and response probability is completely explained by sociodemographic variables.

comparison between respondents and nonrespondents. Thus, the choice of the variables that potentially generate nonresponse bias relies on modeling the response probability. Because we had numerous auxiliary variables available from administrative databases, we chose this method. IPW may lead to very large estimations of variance for the outcome variable prevalence when the response probabilities include extreme values and are not associated with the outcome variable [28]. This is why approaches such as reweighting with weight truncation have been developed [29]. This yields acceptable variance estimates but may lead to biased estimations (balance between bias and variance). Research in survey statistics for the best method for reweighting from response probabilities has led to recommend the construction of homogeneous response groups from response probabilities estimated by a model and then the use of observed response rates in these groups as correction factors. Several methods may be used to construct homogeneous response groups. It has been shown that the score method we chose in this study is the most robust for a poorly specified function [16,30]. Reweighting techniques are efficient in an MAR assumption on the response process. Some statistical tools are also available for not MAR assumption, but they were not used here because of the quality and the frequency of the auxiliary data available and the good results of our second sensitivity analysis, which is most important than the method used [24]. If it was not the case, more sophisticated methods could be used such as Heckman-type selection models but would require strongest assumptions on the distribution of the variables [31].

5. Conclusion

This study not only demonstrates the interest of linking routine health insurance and occupational data to study nonresponse bias but also shows how these data can be used to take into account the nonresponse bias for estimating prevalence. The results are quite promising even with a response rate as low as 25%. They indicate that in addition to the response rate, the major concern is the relevance of the data used to correct for nonresponse bias [24,26]. In our study, work- and health-related auxiliary data from administrative databases appeared to correct effectively for nonresponse. The magnitude of the corrections shows that it is not sufficient to correct nonresponse on sociodemographic variables because they only partially capture health and other determinants of participation in an epidemiological survey. Further research including the analysis of a follow-up survey among nonrespondents will give us better understanding of selection bias in our study.

Acknowledgments

The authors thank the Mutualité Sociale Agricole personnel (Alain Pelc, Nicolas Viarouge, Florian Brémaud, and Yves Cosset) for their fruitful collaboration on the Coset-MSA project, the CnamTS for the access to their databases, David Haziza and Jean-Luc Marchand for their precious advice during the analysis step, Marie Zins and her Constances team for the exchanges during the study, and Ellen Imbernon and Marcel Goldberg for their fruitful comments on the manuscript.

References

[1] Morton LM, Cahill J, Hartge P. Reporting participation in epidemiologic studies: a survey of practice. Am J Epidemiol 2006;163:197−203.

[2] Bethlehem J. Weighting nonresponse adjustments based on auxiliary information. In: Groves RM, Dillman DA, Eltinge JL, Little RJA, editors. Survey nonresponse. New York, NY: Wiley; 2002:275−88.

[3] Dillman DA, Eltinge JL, Groves RM, Little RJA. Survey nonresponse in design, data collection and analysis. In: Groves RM, Dillman DA, Eltinge JL, Little RJA, editors. Survey nonresponse. New York, NY: Wiley; 2002:3−26.

[4] Sarndal CE, Swensson B, Wretman J. Nonresponse. In: Model assisted survey sampling. New York, NY: Springer; 1992:556−600.

[5] Zanutto E, Zaslavsky A. Using administrative records to impute for nonresponse. In: Groves RM, Dillman DA, Eltinge JL, Little RJA, editors. Survey nonresponse. New York, NY: Wiley; 2002:403−16.

[6] Goldberg M, Chastang JF, Leclerc A, Zins M, Bonenfant S, Bugel I, et al. Socioeconomic, demographic, occupational, and health factors associated with participation in a long-term epidemiologic survey: a prospective study of the French GAZEL cohort and its target population. Am J Epidemiol 2001;154:373−84.

[7] Knudsen AK, Hotopf M, Skogen JC, Overland S, Mykletun A. The health status of nonparticipants in a population-based health study. Am J Epidemiol 2010;172:1306−14.

[8] Martikainen P, Laaksonen M, Piha K, Lallukka T. Does survey nonresponse bias the association between occupational social class and health? Scand J Public Health 2007;35:212−5.

[9] Nummela O, Sulander T, Helakorpi S, Haapola I, Uutela A, Heinonen H, et al. Register-based data indicated nonparticipation bias in a health study among aging people. J Clin Epidemiol 2011;64:1418−25.

[10] Vercambre MN, Gilbert F. Respondents in an epidemiologic survey had fewer psychotropic prescriptions than nonrespondents: an insight into health-related selection bias using routine health insurance data. J Clin Epidemiol 2012;65:1181−9.

[11] Assogba GF, Couchoud C, Roudier C, Pornet C, Fosse S, Romon I, et al. Prevalence, screening and treatment of chronic kidney disease in people with type 2 diabetes in France: the ENTRED surveys (2001 and 2007). Diabetes Metab 2012;38:558−66.

[12] Saez M, Barcelo MA, de Tuero GC. A selection-bias free method to estimate the prevalence of hypertension from an administrative primary health care database in the Girona Health Region, Spain. Comput Methods Programs Biomed 2009;93:228−40.

[13] Geoffroy-Perez B, Chatelot J, Santin G, Bénézet L, Delézire P, Imbernon E. Coset: un nouvel outil généraliste pour la surveillance épidémiologique des risques professionnels. Bull Epidemiol Hebd (Paris) 2012;22-23:276−7. [in French].

[14] Zins M, Bonenfant S, Carton M, Coeuret-Pellicer M, Guéguen A, et al. The CONSTANCES cohort: an open epidemiological laboratory. BMC Public Health 2010;10:479.

[15] Siegrist J, Wege N, Pühlhofer F, Wahrendorf M. A short generic measure of work stress in the era of globalization: effort-reward imbalance. Int Arch Occup Environ Health 2009;82:1005−13.

[16] Haziza D, Beaumont JF. On the construction of imputation classes in surveys. Int Stat Rev 2007;75:25−43.

[17] Eltinge JL, Yansaneh IS. Diagnostics for formation of nonresponse adjustment cells, with an application to income nonresponse in the U.S. consumer expenditure survey. Survey Methodol 1997;23: 33−40.
[18] Little RJA. Survey nonresponse adjustments for estimates of means. Int Stat Rev 1986;54:139−57.
[19] Goldberg M, Luce D. Selection effects in epidemiological cohorts: nature, causes and consequences. Rev Epidemiol Sante Publique 2001;49:477−92.
[20] Lamers LM. Medical consumption of respondents and non-respondents to a mailed health survey. Eur J Public Health 1997;7: 267−71.
[21] Reijneveld SA, Stronks K. The impact of response bias on estimates of health care utilization in a metropolitan area: the use of administrative data. Int J Epidemiol 1999;28:1134−40.
[22] Etter JF, Perneger TV. Analysis of non-response bias in a mailed health survey. J Clin Epidemiol 1997;50:1123−8.
[23] Goldberg M, Chastang JF, Zins M, Niedhammer I, Leclerc A. Health problems were the strongest predictors of attrition during follow-up of the GAZEL cohort. J Clin Epidemiol 2006;59:1213−21.
[24] Geneletti S, Mason A, Best N. Adjusting for selection effects in epidemiologic studies: why sensitivity analysis is the only solution. Epidemiology 2011;22:36−9.
[25] Moisan F, Gourlet V, Mazurie JL, Dupupet JL, Houssinot J, Goldberg M, et al. Prediction model of Parkinson's disease based on antiparkinsonian drug claims. Am J Epidemiol 2011;174:354−63.
[26] Groves RM. Nonresponse rates and nonresponse bias in household surveys. Public Opin Q 2006;70:646−75.
[27] Deville JC, Sarndal CE. Calibration estimators in survey sampling. J Am Stat Assoc 1992;87:376−82.
[28] Robins JM, Sued M, Lei-Gomez Q, Rotnitzky A. Comment: performance of double-robust estimators when ''inverse probability'' weights are highly variable. Stat Sci 2013;22:544−59.
[29] Seaman SR, White IR. Review of inverse probability weighting for dealing with missing data. Stat Methods Med Res 2011;22:278−95.
[30] Haziza D, Beliveau A. Estimation non-paramétrique des probabilités de réponse. 2010. [Report].
[31] Barnighausen T, Bor J, Wandira-Kazibwe S, Canning D. Correcting HIV prevalence estimates for survey nonparticipation using Heckman-type selection models. Epidemiology 2011;22:27−35.