L'enchaînement de non-réponses dans une enquête couplée : le cas de l'enquête Famille-Employeurs

Nicolas Razafindratsima (razafind@ined.fr)
INED, Service enquêtes et sondages

Journées de Méthodologie Statistique de l'INSEE (JMS-2009)
Paris, mars 23-25 mars 2009

Plan de la présentation

- Introduction
- L'enquête Famille-employeurs
- La non-réponse au volet ménages
- La non-réponse au volet employeurs
- L'enchaînement des non-réponses
- Conclusion

Introduction

- La non-réponse : problème récurrent dans les enquêtes par sondage en population générale
- Problèmes engendrés :
  - Un biais potentiel, si les non-répondants ont des caractéristiques différentes des répondants
  - Une baisse de la précision (diminution de la taille d'échantillon)
- Objectif : décrire les biais engendrés par la succession de non-réponses (totales) différentielles dans une enquête couplée

L'enquête Famille-employeurs

- INED et INSEE, 2004-2005
- Objectif : Etudier la conciliation entre la vie professionnelle et la vie familiale, du point de vue des salariés et de leurs employeurs
- Enquête en 2 volets :
  - Une enquête auprès des ménages : le « volet ménage ». 9 547 individus de 20 à 49 ans interrogés en face à face sur un questionnaire portant notamment sur les conditions de travail
  - Une enquête auprès des employeurs des personnes interrogées dans le volet ménage : « le volet employeurs ». 2 673 établissements de 20 salariés ou plus enquêtés par voie postale ou par internet (questionnaire auto-administré)

Le volet ménage

- Plan de sondage :
  - Tirage de logements dans l'échantillon-maître de l'Insee en 2 vagues. Les résidences principales de 1999 ayant 2 individus de la génération ciblée (nés entre 1958 et 1986) sont sur-représentées, celles sans individu de cette tranche d'âge exclues
  - L'enquêteur contacte le ménage et pose 1 questionnaire sur sa composition (tronc commun des ménages ou TCM). Ménages éligibles=ceux comprenant au moins 1 individu âgé 20 à 49 ans
  - On enquête tous les 20-49 ans du ménage s'il y en a 2 ou moins. Si plus de 2 individus éligibles : tirage au sort de 2 d'entre eux (par le système CAPI)
  - Interrogation séparée des individus sélectionnés

La non-réponse des ménages

- Tirage de 11 759 logements dans l'échantillon-maître : 6 775 en vague 1, 2 213 en vague 2
- 9 765 ménages identifiés (83%), parmi lesquels 7 552 (77,3%) acceptent de répondre au TCM
- La non-réponse (non contact ou refus) concerne surtout (caractéristiques en 1999) :
  - Les logements collectifs (immeubles) (25,7% vs 18,6%)
  - Les grandes agglomérations,
  - La région Ile de France
  - Les ménages de 1 ou de 2 personnes

Les non-réponses des individus

- Au sein des ménages éligibles, 10 189 personnes ont été sélectionnées pour l'enquête
- 9 547 d'entre eux ont effectivement répondu. Taux de non-réponse de 6,3%
- La non-réponse concerne surtout :
  - Les hommes (taux de non-réponse : 8,7% vs 4,1% pour les femmes)
  - Les personnes nées à l'étranger (taux de non-réponse de 10% vs 5,8% pour les nés en France)
  - Les Franciliens (10,2%)
  - Les personnes vivant dans un ménage avec 3 éligibles ou plus (non-réponse : 11,3%)

Comparaison des données du volet ménage avec l'enquête Emploi

- Comparaison entre individus EFE pondérés (pondération de base) et EE- 2004/2005
- On constate une sous-représentation :
  - des hommes : 46,7% vs 49,5% pour l'EE
  - des peu diplômés (sans ou non déclaré) : 15,3% vs 19,2% pour l'EE
- A contrario, on a une sur-représentation des femmes, et des très diplômés (> bac+2 : 19,6% vs 14,3%)

Le volet employeurs

- Les individus travaillant dans des établissements de plus de 20 salariés donnent le nom et l'adresse de leur employeur, et éventuellement le SIRET (N=4 634)
- Un appariement est réalisé avec le répertoire SIRENE (Insee) pour récupérer des informations supplémentaires pouvant faciliter la collecte
- Envoi d'un questionnaire auto-administré de 8 pages pour une enquête postale ou internet auprès de ces employeurs
- Si non-réponse : relances (équipe INED)

Les non-réponses au volet employeurs

- Méthodologie :
  - 2 niveaux d'observations possibles :
    - Niveau individu : N=4 634
    - Niveau établissement (on supprime les répétitions si plusieurs salariés de l'établissement ont été enquêtés) : N=4 197
- Sources de données :
  - Le fichier individus
  - Les informations issues de l'appariement avec SIRENE
- On peut décrire et modéliser la non-réponse totale

Les non-réponses au volet employeurs

- Au niveau individu : 3 065 répondants sur 4 476, soit un taux de réponse de **68,5%**
- Au niveau établissement : 2 703 répondants sur 4 041, soit un taux de réponse de **66,9%**
- Variations des taux de réponse selon :
  - La taille : taux de réponse plus faible pour les petits établissements (20-49 salariés) : 63%
  - Le statut de l'établissement (déclaré par le salarié) : collectivités locales & hôpitaux : 73%, Etat : 67%, entreprises privées : 66%
  - La région : 59% pour l'Ile de France vs 69% pour les autres régions
  - L'âge de l'établissement (variable SIRENE) : 60% si 0-4 ans, 74% si 20 ans ou plus
  - le fait d'être « siège » de l'entreprise : 74% si siège, 62% sinon

Les non-réponses au volet employeurs

- Résultats d'une analyse multivariée parmi les établissements parfaitement appariés avec SIRENE (n=1 285) :
  - variables significatives : l'âge de l'établissement (taux de réponse augmente avec l'âge), le fait d'être le « siège » de l'entreprise (favorise la réponse), la région (les établissements d'Ile de France répondent moins)
  - variables non significatives (seuil 10%) : la taille, le secteur d'activité, l'existence d'autres établissements dans l'entreprise, la présence d'1 crèche , la présence d'une direction des ressources humaines (DRH)

Evaluation des biais dus à l'enchaînement des non-réponses

- Principe :
  - Calcul de pondération pour les individus, en redressant ou non
  - Utilisation de la méthode du partage des poids (Lavallée, 2002 et 2007) pour passer de la pondération des individus à celle des établissements
  - au niveau établissement, correction de la non-réponse ou non
  - Puis comparer, au niveau établissement, les répartitions des variables d'intérêt selon le type de pondération utilisé à chaque volet, le poids de référence étant celle redressée dans les deux volets

Pondération des individus

- Poids de base : tient compte du mode de tirage des ménages, de l'existence des deux vagues d' enquête, et du mode de sélection des individus 20-49 ans dans le ménage
- Redressements :
  - Correction de la non-réponse au niveau ménage puis individus (calcul de coefficients de redressement par groupes homogènes de non-répondants – obtenus par régression logistique)
  - calage sur les données de l'enquête Emploi en continu. 7 variables de calage : sexe*âge, diplôme le plus élevé, statut d'activité, nationalité, taille du ménage, tranche d'unité urbaine, zone d'étude et d'aménagement du territoire (ZEAT)

Pondération des établissements

- Poids de base : obtenu par la méthode du partage des poids

Poids établissement=

(∑(poids individus enquêtés dans l'établissement)/(effectif de 20-49 ans)

- Imputation de l'effectif de 20-49 ans pour 16% des établissements
- Correction de la non-réponse : calcul de coefficients de correction au sein de groupes homogènes de non-répondants, créés par régression logistique
- Pas de calage car pas de source externe directement comparable

Résumé des pondérations disponibles pour les comparaisons

Pondération individus	Pondération employeurs (après partage des poids)	
	Sans correction de non-réponse	*Avec correction de non-réponse*
Aucune correction (poids de base)	P00	P01
Correction non-réponse puis calage sur 7 variables	P10	P11

Pondération salariés=Poids établissement*nb de salariés

Distribution des poids établissements

Poids	Moyenne	CV (%)	Min	Max	Max/ Min	Nb de poids <1
P00	26,0	131,8	0,2	426,4	2 715,9	120
P01	38,5	139,2	0,2	841,7	3 627,8	58
P10	41,2	140,3	0,2	588,3	3 197,0	63
P11	61,3	149,7	0,3	1 424,4	5 652,2	28

Distribution des poids salariés

Poids	Moyenne	CV (%)	Min	Max	Max/min
P00	2 494,3	77,6	565,0	49 548,8	87,7
P01	3 698,1	83,3	823,5	76 682,6	93,1
P10	3 920,5	81,2	600,4	65 971,7	109,9
P11	5 843,1	90,1	948,7	102 099,1	107,6

Interprétation en termes d'estimation d'effectifs

- Par rapport à la référence :
  - La non-réponse totale des ménages et des individus amène à sous-estimer l'effectif total d'établissements et des salariés de 37%
  - La non-réponse totale des établissements, elle, amène à sous-estimer l'effectif total d'établissements et de salariés de 33%
  - Le cumul des non-réponses dans les 2 volets amène à sous-estimer ces effectifs totaux de 57%

Répartition des établissements selon type de propriété

	P00	P01	P10	P11
Privé non lucratif	15,8	16,2	15,0	15,1
Privé lucratif	57,7	57,7	60,2	60,3
F.P d'Etat	12,4	12,5	11,5	11,5
F.P hospitalière	2,8	2,5	2,4	2,2
F.P territoriale	7,5	6,9	7,3	6,6
Secteur public ou nationalisé	3,8	4,3	3,7	4,2

Répartition des établissements selon activité (NES16)

nes16	P00	P01	P10	P11
Agriculture, sylviculture, pêche	1,4	1,2	1,3	1,1
Industrie automobile	0,6	0,6	0,7	0,7
Transports	3,5	3,8	3,8	4,2
Activités financières	1,6	1,8	1,7	1,9
Services aux entreprises	11,5	13,0	11,5	13,0
Services aux particuliers	4,0	4,1	4,4	4,5
Éducation, santé, action sociale	23,9	23,6	21,9	21,5
Administration	13,6	13,5	13,2	13,0

Répartition des établissements par région

Région	P00	P01	P10	P11
Ile de France	14,4	16,5	15,8	18,1
Autre région	85,6	83,5	84,2	81,9

Répartition des établissements selon le % de femmes parmi le personnel

% de femmes	P00	P01	P10	P11
Inférieur ou égal à 25%	25,6	25,6	27,5	27,5
26% à 75%	49,8	50,6	49,4	50,1
76% ou +	17,7	17,1	16,2	15,7
Vide	6,8	6,7	6,9	6,7

Répartition des établissements selon présence d'une crèche

Présence d'une crèche	P00	P01	P10	P11
Oui	1,7	1,7	1,7	1,6
Non	93,4	93,4	93,7	93,7
Ne sait pas	4,8	4,9	4,6	4,7

Observations

- Les écarts à la référence les plus importants proviennent de non-réponses différentielles qui jouent dans le même sens. Exemples : sous-représentation des établissements en Ile de France ou du secteur privé lucratif
- Pour d'autres variables d'intérêt, les non-réponses différentielles jouent en sens contraire, et entraînent une estimation plus proche de la référence. Exemples : % d'établissements du secteur privé non lucratif, ou public nationalisé
- Pour la répartition par région, c'est la non-réponse des établissements qui est la principale source de biais

Conclusion

- 1er effet de la non-réponse dans Famille-employeurs: une sous-estimation des totaux de 57%. La non-réponse des ménages et des individus entraîne une sous-estimation de 37%, la non-réponse des établissements de 33%
- 2e effet : un biais lors de l'estimation de certains pourcentages
- Pour la plupart des variables étudiées, seule la non-réponse cumulée induit des biais significatifs: la non-réponse différentielle à un seul volet ne joue guère. Exception : la répartition par région

Limites et perspectives

- La mesure des biais n'est pas facile en l'absence de données externes de référence. On fait l'hypothèse que les pondérations (sans calage en volet 2) les éliminent
- Tester la robustesse des méthodes de correction de non-réponses utilisées
- Prise en compte des non-réponses partielles
- Prise en compte des variances
- Biais sur les analyses multivariées ?

Références

- Lavallée, P. (2002) : *Le sondage indirect, ou la méthode généralisée du partage des poids*, Bruxelles, Presses de l'université libre de Bruxelles et Paris, Ellipses, 242 pages.
- Lavallée, P. (2007) : *Indirect sampling*, New York, Springer (Springer series in statistics), 245 pages.
- Ouvrage présentant les résultats de l'enquête à paraître bientôt (dir A. Pailhé et A. Solaz)

Enquête Familles et employeurs

Merci de votre attention !