

Agréger Les échantillons d'une enquête multimode en limitant l'effet de mesure : une proposition d'imputation raisonnable et pragmatique

Stéphane Legleye (*), Tiaray Razafindranovona (**), Gaël de Peretti (*)

() Insee, Division Recueil et traitement e l'information*

*(**) Insee, Division Emploi*

Plan

- Rappel des enjeux du multimodes
- Présentation de CVS-VVS
 - Représentativité, mesures des victimations
- Recherche des individus portant l'effet de mesure
- Imputations
- Mesures finales
- Discussion

La mode du multimode

- Les enquêtes multimodes pour :
 - Augmenter le taux de réponse +
 - Augmenter la représentativité +
 - Diminuer le coût +++
 - Réduire le biais de réponse ??
- auto-administré (Internet)
 - Diminution de la désirabilité sociale
 - Bénéfique pour les sujets sensibles
 - Augmentation du satisficing
 - Néfaste pour les enquêtes longues, les sujets peu motivants, les questions difficiles cognitivement

Les designs multimodes

- Designs très variables:
 - Séquentiels : un mode puis l'autre en cas de non-réponse
 - Du moins cher au plus cher (Internet, Téléphone, Face-à-face)
 - Concurrentiels: modes proposés au choix du répondant
 - Dès le contact
 - stratifié : plusieurs branches monomodes
 - Souvent pour des expérimentations

Effet de mode et de mesure

- Effet de mode=écart entre deux échantillons pour une variable cible Y.
- $EM = BS + BM$
 - Les deux échantillons diffèrent en composition (socio-démographique par exemple): biais de composition
 - L'effet du biais de composition sur Y est appelé Biais de sélection BS
 - BM est le biais de mesure pour Y
- Par définition:
 - $BM = \text{effet du mode sur Y pour un individu}$
 - Situation contrefactuelle presque jamais identifiable
 - $BM = \text{effet du mode sur Y pour un ensemble d'individus comparables (donc corrigé du biais de sélection)}$

Exemple de CVS-VVS

- Cadre de vie et sécurité (CVS) : annuelle depuis 2007, face-à-face + audio-CASI pour les questions les plus sensibles
 - Taux de réponse de 63% en 2013, n=14 585
- VVS : expérimentation ponctuelle Internet/papier
 - Taux de réponse 32% en 2013
 - Le papier et l'Internet seront confondus ici: n=12 896
- Deux tirages indépendants (sur fichiers fiscaux)
- Cadre de travail: design stratifié
 - Comparaison à CVS classique
 - On ne redresse pas VVS relativement à CVS (comme dans une expérimentation): les deux échantillons sont pris comme un tout

Les variables

- 8 Socio-démographiques (SD): variables de calage de CVS classique
 - Sexe, âge (6 tranches, de 14 ans à 70 ans et plus), taille d'unité urbaine (4 modalités), diplôme (5 modalités), type de ménage (4 modalités), la CSP (8 modalités), le type de logement (2 modalités) et le statut d'occupation du logement (2 modalités)
- 14 variables auxiliaires d'opinions (X), binaires ou Likert
 - Dont 4 Likert étudiées ici: opinion sur la police, sur sa présence; sentiment d'insécurité au domicile et dans le quartier.
- 6 variables cibles binaires (Y): (occurrence 24 derniers mois)
 - Vol avec violence; vol sans violence; violence physique; cambriolage; vol de véhicule, menaces.

Echantillons et représentativité

CVS calé seul=référence

VVS calé seul= à titre de comparaison (vu le design stratifié)

CVS-VVS calé=échantillon multimode d'étude

- Indicateur R (Bethlehem, Cobben, & Schouten, 2009)
 - Les 8 variables de calage
 - Sans interaction (R1), avec toutes les interactions bivariées (R2)
 - Population de référence= recensement (enquête annuelle 2013)
- $R1(CVS)=0,81$; $R1(VVS)=0,82$; $R1(CVS-VVS)=0,78$
- $R2(CVS)=0,76$; $R2(VVS)=0,63$; $R2(CVS-VVS)=0,69$
- ➔ A première vue: VVS seul comparable à CVS et CVS-VVS
- ➔ Si on est plus strict, CVS meilleur

Victimations

	Référence		Avant correction				Ratio / CVS
	CVS seul (calé)		VVS seul (calé)		Agrégation (calée)		
	%	Stderr	%	Stderr	%	Stderr	
vavn	0.99%	0.08%	2.88%	0.15%	1.63%	0.08%	1.64
vsvn	2.78%	0.14%	6.44%	0.22%	4.16%	0.12%	1.50
vphyn	2.20%	0.12%	2.90%	0.15%	2.28%	0.09%	1.04
vlogn	3.74%	0.16%	6.44%	0.22%	4.89%	0.13%	1.31
vvehn	4.52%	0.17%	5.89%	0.21%	5.06%	0.13%	1.12
men	4.53%	0.17%	7.68%	0.23%	5.48%	0.14%	1.21
Moyenne	3.13%		5.37%		3.92%		1.25

Très fortes différences entre CVS et VVS

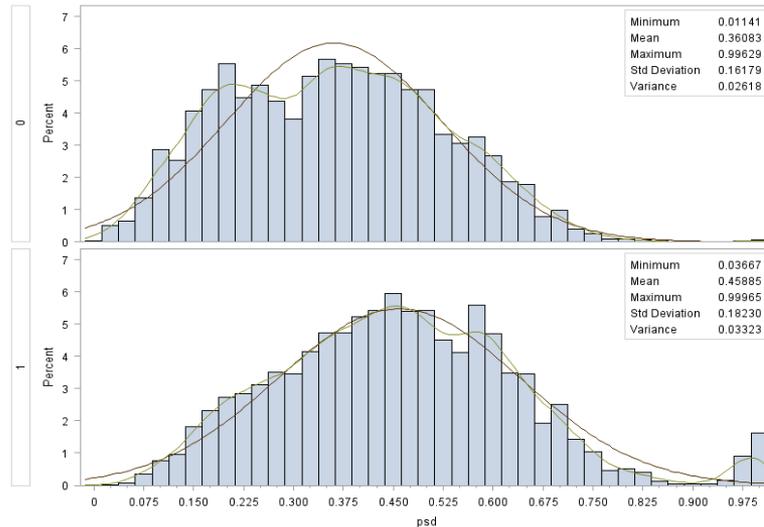
L'échantillon agrégé présente des proportions bien plus élevées (25% en moyenne)

Stratégie

- Identifier les individus qui portent le biais de mesure BM: corriger une partie des Y des répondants de VVS
 - Moyen: appariement sur score de propension
 - Sélectionner les variables liées aux Y et à l'échantillon CVS ou VVS
 - Contrainte: ne pas introduire de variable affectée par un effet de mesure (Rubin, 2007; Vannieuwenhuyze & Loosveldt, 2013)
- ➔ On se restreint par prudence aux 8 variables SD (le sexe est la variable la moins corrélée aux Y)

Modélisation du score de propension et appariement

- Régression logistique; 8 variables SD plus leurs interactions bivariées (44 effets)



- Appariement: sur le score plus contrainte d'identité de chaque variable SD. Appariement 1-1 sans remise.

Paires et imputations

- 7253 paires de jumeaux SD dont 31,5% ont au moins une différence de Y (sur les 6 étudiées).
- Imputation déterministe: on corrige au sein des paires en prenant CVS comme référence
- Imputation stochastique (5 imputations): on corrige dans l'échantillon de jumeaux sans se soucier de l'appariement. Permet de calculer une variance.
 - Truc: on sélectionne la proportion de valeurs à imputer en prenant comme référence celle de CVS au sein des paires pour chaque variable cible Y.

Nombre de valeurs à imputer

Imputations déterministes: 17,7% d'individus concernés

Imputations stochastiques: 8,0% d'individus concernés

Détail par variable	Imputations déterministes		Imputations stochastiques	
	Fréquence	%	Fréquence	%
Vavn	211	1.6	121	0.9
Vsvn	583	4.5	366	2.8
Vphyn	283	2.2	76	0.6
Vlogn	703	5.5	297	2.3
Vvehn	582	4.5	56	0.4
Men	741	5.7	361	2.8
Total	3103		1277	
Taux (valeurs)		4.0		1.7
N	12896 × 6		12896 × 6	

Résultats des imputations: déclarations de victimations

	Agrégation initiale		Agrégation déterministe			Agrégation stochastique		
	%	Ratio/ CVS	%	Ratio / initial	Ratio/ CVS	%	Ratio / initial	Ratio / CVS
vavn	1.63%	1.64	1.33%	0.82	1.34	1.33%	0.82	1.34
vsvn	4.16%	1.50	3.35%	0.80	1.21	3.32%	0.80	1.20
vphyn	2.28%	1.04	2.20%	0.96	1.00	2.17%	0.95	0.99
vlogn	4.89%	1.31	4.17%	0.85	1.11	4.31%	0.88	1.15
vvehn	5.06%	1.12	4.86%	0.96	1.08	4.97%	0.98	1.10
men	5.48%	1.21	4.88%	0.89	1.08	4.79%	0.87	1.06
Moyenne	3.92%	1.25	3.46%	0.88	1.11	3.48%	0.89	1.11

Variance d'imputation : <0,5%

Résultats des imputations: estimations causales

	Pré imputation			Post imputation							
	%	StdErr	P	Déterministe				Stochastique			
%				StdErr	P	Ratio	%	StdErr	P	Ratio	
vavn	1.53%	0.16%	0.000	0.81%	0.15%	0.000	0.53	0.02%	0.16%	0.897	0.01
vsvn	3.55%	0.25%	0.000	1.62%	0.23%	0.000	0.46	0.11%	0.27%	0.679	0.03
vphyn	0.65%	0.19%	0.000	0.49%	0.19%	0.011	0.75	0.23%	0.22%	0.294	0.36
vlogn	3.05%	0.27%	0.000	0.91%	0.26%	0.001	0.30	0.46%	0.34%	0.176	0.15
vvehn	1.51%	0.26%	0.000	0.50%	0.28%	0.073	0.33	0.34%	0.35%	0.333	0.22
men	2.67%	0.29%	0.000	1.54%	0.28%	0.000	0.58	0.01%	0.34%	0.978	0.00
Moyenne	2.16%			0.98%			0.45	0.20%			0.09

Analyse par régression : ajustement sur le décile de score de propension, les 8 SD et les quatre X les plus stables.

Variance d'imputation : <0,5%

Discussion (1)

- Méthode utilisable pour tout design dès qu'il existe un échantillon recouvrant (qui peut être petit)
- Méthode oscillant entre une définition individuelle et populationnelle du biais de mesure
- Repose la définition du support commun
 - Plus on est strict, plus on réduit le support commun; moins la méthode est efficace
 - Différentes méthodes et paramétrages
 - Le choix 1-1 sans remise réduit le support commun qui est défini au niveau individuel
 - On peut envisager plusieurs appariements successifs pour étendre le nombre d'observations traitées

Discussion (2)

- Ne traite pas le biais de sélection (hors support commun)
 - Il peut être lui-aussi affecté d'un biais de mesure...
 - On peut le diminuer si on est plus laxé dans la définition du support commun (ou si on réitère l'appariement sur icelui)
- Méthode parcimonieuse
- Compatible avec une repondération visant à corriger le biais de composition
- Quel jeu d'imputation donner ? Un ? Tous?

Discussion (3): que décider?

- Évaluer la représentativité de l'échantillon alternatif (VVS)
- Soupçonner un effet de sélection non-ignorable
 - Examiner les variables auxiliaires liées aux victimations
Exemple : trop de réponses des quartiers « riches », etc.
Calcul d'un R-indicateur idoine.
 - Peut-on le corriger avec une repondération (enrichie)?
 - Si oui, la méthode s'applique et corrige l'effet de mesure résiduel
 - Si non, la méthode s'applique et corrige l'effet de mesure et de sélection non-ignorable
- Evaluer la qualité de la mesure dans le mode alternatif
 - Examen de la littérature!
 - Si biais de mesure favorable, ne rien faire est une option, voire imputer le mode de référence historique

Discussion (4) : déontologie et long terme

- Imputer implique:
 - ne pas croire les répondants
 - être tributaire d'un modèle de données/collecte qui peut devenir obsolète
 - Par exemple: la part de l'Internet relativement au face-à-face augmente d'une enquête à l'autre
Le doute radical et la rupture de séries sont inévitables

En conséquence, prudence !

Un échantillon de contrôle faciliter les estimations?

Enrichir au maximum les bases de sondages

Merci pour votre
attention

Références

- Bethlehem, J., Cobben, F., & schouten, B. (2009). Des indicateurs de la représentativité aux enquêtes. *Techniques d'Enquêtes*(Recueil du symposium 2008 de Statistique Canada), 1-10.
- Razafindranovona, T. (2016). *Exploitation de l'enquête expérimentale Vols, violence et sécurité*. Retrieved from Paris: www.insee.fr/fr/statistiques/2022146
- Imbens, G. W., & Rubin, D. B. (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge, MA: Cambridge University Press
- Vannieuwenhuyze, J., & Loosveldt, G. (2013). Evaluating Relative Mode Effects in Mixed-Mode Surveys: Three Methods to Disentangle Selection and Measurement Effects. *Sociological Methods and Research*, 42(1).