Effets de mode dans les enquêtes ménages multimodes : état des lieux et illustrations

Patrick SILLARD

INSEE-DMCSI

Juin 2023

Plan

Introduction

A l'Insee, depuis les années 2010:

- des documents de travail:
  - Razafindranovona (2015): *La collecte multimode et le paradigme de l'erreur d'enquête totale*, DT INSEE M2015/01.
  - Razafindranovona (2016): *Exploitation de l'enquête expérimentale Vols, violence et sécurité*, DT INSEE M2016/03.
  - Razafindranovona (2016): *Exploitation de l'enquête expérimentale Logement internet/papier*, DT INSEE M2016/08.
  - Razafindranovona (2017): *Exploitation de l'enquête expérimentale Qualité de vie au travail*, DT INSEE M2017/01 .
- des papiers aux JMS:
  - Legleye et al. (2015): L'utilisation des historiques d'appels pour redresser une enquête téléphonique : une étude par simulation à partir de l'enquête Fecond.
  - Legleye et al. (2018): Correction des effets de mode et biais de sélection : Les apports d'une expérimentation de l'enquête TIC en face-à-face.
  - Legleye (2018): Agréger les échantillons d'une enquête multimode en limitant l'effet de mesure : une proposition d'imputation raisonnable et pragmatique

A l'Insee, depuis les années 2010:

- et récemment, encore des documents de travail:
  - Castel et Sillard (2021): *Le traitement du biais de sélection endogène dans les enquêtes auprès des ménages par modèle de Heckman*, DT INSEE M2021/02.
  - Castell et al. (2023): *Redressements de la première vague de l'enquête EpiCov : un exemple de correction des effets de sélection dans les enquêtes multimodes*, DT INSEE M2023/02.
  - Castell et Clerc (2023): *Victimations déclarées et effets de mode: enseignements de l'expérimentation panel multimode de l'enquête Cadre de vie et sécurité*, DT INSEE à paraître.

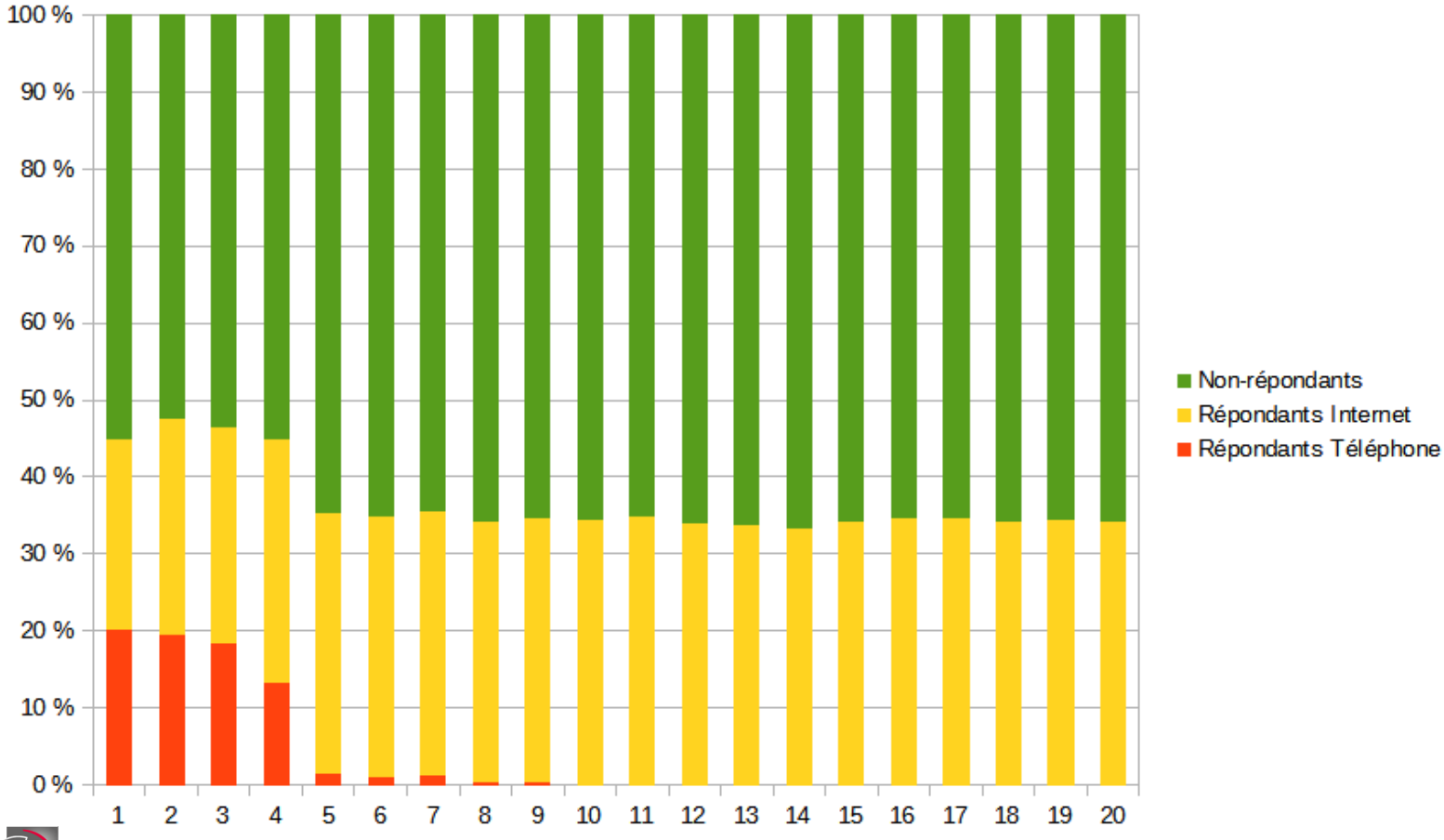
En synthèse, ces travaux proposent:

- Des réflexions sur les biais de collecte d'enquête (satisficing, désirabilité sociale) ⟶ Comment rendre les questionnaires équivalents entre modes?
- Plusieurs "expérimentations" permettant de se familiariser avec les problèmes rencontrés et les données (et les paradonnées)
- Une formalisation du multimode mettant en évidence:
  - le **problème** du biais de mesure (systématisme d'un mode par rapport aux autres)
  - le **lien** entre identification du biais de mesure et correction de la non-réponse
  - le **rôle du design d'enquête** pour modéliser le lien précédent, avec l'usage de sous-échantillons collectés selon des compositions modales différentes

Des anomalies qui interrogent

EpiCov (1/2)

Enquête (INSERM, DREES, INSEE, SPF), 370 000 sélectionnés, 135 000 répondants, collectée en mai 2020

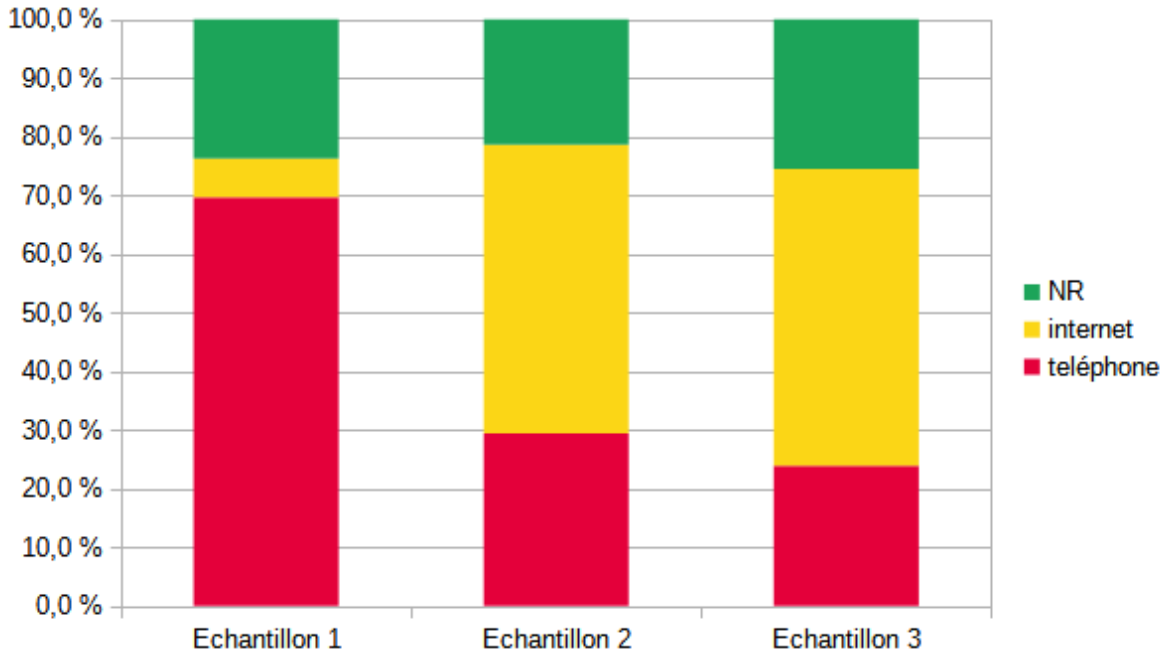


EpiCov (2/2)

	Lots monomodes	Lots multimodes
Variables socio-démographiques		
Âge moyen	48, 3 (0,19)	48, 6 (0,32)
Au chômage	6, 1 (0,10)	6, 2 (0,16)
Variables d'intérêt		
Fièvre*	8, 0 (0,10)	7, 0 (0,16)
Perte de goût*	2, 8 (0,06)	2, 5 (0,10)
Au moins 1 symptôme*	27, 2 (0,17)	23, 5 (0,27)

L'enquête expérimentale "CVS-Panel" (2019)

Interrogation int./tel. des répondants CVS2018



	CVS2019	CVS-Panel		
		Éch. 1	Éch. 2	Éch. 3
Vandalisme de la voiture (%)	3, 7 (0.2)	3, 8 (0.3)	6, 0 (0.4)	5, 7 (0.4)

De la sélection dans une enquête (multimode)

Contexte et notations

- Population $\mathscr{P}$, individus indicés sur $\{1, \ldots, N\}$
- Variable d'intérêt y ; moyenne μ sur $\mathscr{P}$
- Sélection aléatoire (sondage) d'un sous-ensemble $\mathscr{S}$
- Variable de sélection s : $\mathscr{S} = \{i \in \mathscr{P} | s_i = 1\}$
- Caractéristiques X (sur $\mathscr{S}$)
- Variable de réponse r : L'ensemble des répondants est $\mathscr{R} = \{i \in \mathscr{S} | r_i = 1\}$
- Variable de sélection au mode m. Répondants
  - au mode (1) : $m_i = 1$
  - au mode (0) : $m_i = 0$

Les problèmes de sélection: 1-le sondage (s)

Estimateur de μ par repondération (π_i) des observés :

$$\hat{\mu} = \frac{1}{N} \sum_{i \in \mathscr{S}} \frac{y_i s_i}{\pi_i}$$

On observe que :

$$\left\{ \begin{array}{l} y \perp\!\!\!\perp s \\ \pi_i = \Pr(s_i = 1) \end{array} \right\} \Rightarrow \left\{ \hat{\mu} \xrightarrow{\text{prob}} \mu \right\}_{(\#\mathscr{S}) \longrightarrow +\infty}$$

ou, de manière plus plausible si X joue dans s et dans y:

$$\left\{ \begin{array}{l} y \perp\!\!\!\perp s|X \\ \pi_i = \Pr(s_i = 1|X_i) \end{array} \right\} \Rightarrow \left\{ \hat{\mu} \xrightarrow{\text{prob}} \mu \right\}_{(\#\mathscr{S}) \longrightarrow +\infty}$$

Rq.1: il faut que $\mathbb{E}(y_i s_i)$ soit séparable :

$$y \perp\!\!\!\perp s|X \Rightarrow \mathbb{E}(y_i s_i) = \mathbb{E}\left[\mathbb{E}(y_i s_i|X_i)\right] = \mathbb{E}\left[\mathbb{E}(s_i|X_i)\mathbb{E}(y_i|X_i)\right]$$

Les problèmes de sélection: 2-la participation (r)

Estimateur de μ par repondération (π_i) des observés :

$$\hat{\mu} = \frac{1}{N} \sum_{i \in \mathscr{S}} \frac{y_i s_i r_i}{\pi_i}$$

$$\left\{ \begin{array}{l} y \perp\!\!\!\perp (s, r) | X \\ \pi_i = \Pr(s_i = 1 | X_i) \times \Pr(r_i = 1 | X_i) \end{array} \right\} \Rightarrow \left\{ \hat{\mu} \xrightarrow{\text{prob}} \mu \right\}_{(\#\mathscr{R}) \longrightarrow +\infty}$$

Rq.2: marche si l'espérance de $y_i s_i$ est séparable :

$$y \perp\!\!\!\perp (s, r) | X \Rightarrow \mathbb{E}(y_i s_i r_i) = \mathbb{E}\left[\mathbb{E}(y_i | X_i)\mathbb{E}(s_i | X_i)\mathbb{E}(r_i | X_i)\right]$$

Il faut en particulier que $y \perp\!\!\!\perp r | X$

$\longrightarrow$ Ce n'est pas forcément évident!!

Les problèmes de sélection: 3-le mode (m)

- Output potentiels selon que $m_i = (0, 1) : (y_i^1, y_i^0)$
- Observé sur $\mathscr{R}$: $y_i = m_i y_i^1 + (1 - m_i) y_i^0$.

On définit la moyenne *potentielle* mode-m de y :

$$\mu_{\mathscr{R}}^{m} = \frac{1}{\#\mathscr{R}} \sum_{i \in \mathscr{R}} y_i^m$$

On observe que :

$$\Delta_{\mathscr{R}}^{1/0} = \mu_{\mathscr{R}}^{1} - \mu_{\mathscr{R}}^{0}$$

est **l'erreur de mesure moyenne** des répondants au mode 1 par rapport à ceux du mode 0.

$\Delta_{\mathscr{R}}^{1/0}$ n'est pas observable car on observe y_i et non y_i^m

Les problèmes de sélection: 3-le mode (m)

Construire des estimateurs des $\mu_{\mathscr{R}}^{m=(0,1)}$:

$$\hat{\mu}_{\mathscr{R}}^{1} = \frac{1}{\#\mathscr{R}} \sum_{i \in \mathscr{R}} \frac{y_i m_i}{\pi_i^m} \quad \text{et} \quad \hat{\mu}_{\mathscr{R}}^{0} = \frac{1}{\#\mathscr{R}} \sum_{i \in \mathscr{R}} \frac{y_i (1 - m_i)}{1 - \pi_i^m}$$

$$\left\{ \begin{array}{l} y \perp\!\!\!\perp m | X, r = 1 \\ \pi_i^m = \Pr(m_i = 1 | X_i, r_i = 1) \end{array} \right\} \Rightarrow \left\{ \hat{\mu}_{\mathscr{R}}^{m} \xrightarrow{\text{prob}} \mu_{\mathscr{R}}^{m} \right\}$$

$$\Rightarrow \left\{ \hat{\mu}_{\mathscr{R}}^{1} - \hat{\mu}_{\mathscr{R}}^{0} \xrightarrow{\text{prob}} \Delta_{\mathscr{R}}^{1/0} \right\}_{(\#\mathscr{R}) \longrightarrow +\infty}$$

Il faut aussi que $y \perp\!\!\!\perp m | X, r = 1$

$\longrightarrow$ n'a rien d'évident!!!!

Les problèmes de sélection: 3-le mode (m)

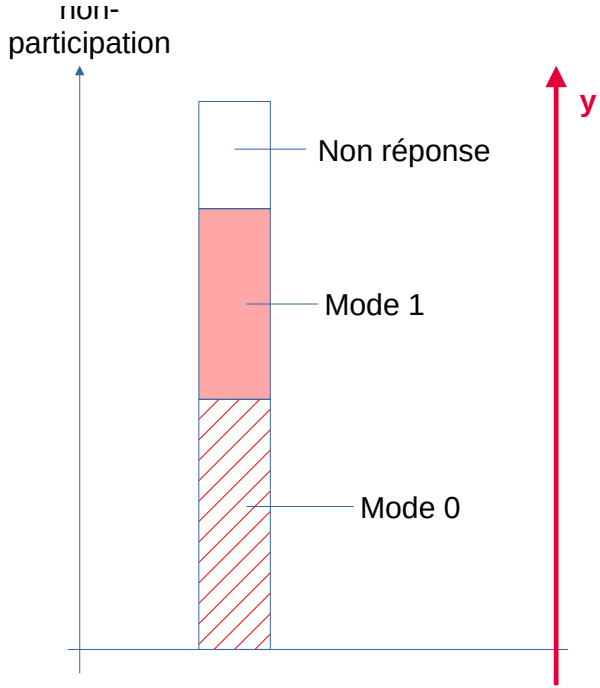
Commentaires:

- On travaille en réalité sur $\overline{\mathscr{R}} = \left\{ i \in \mathscr{R} | \pi_i^m \neq (0,1) \right\}$
- Sur ce champ, regardons deux protocoles “classiques”:
  - m_i est choisi par l’enquêté
  - m_i est sélectionné aléatoirement et imposé à l’enquêté

↪ Si sélection endogène ($r(y)$) + biais de mesure: les 2 effets se confondent avec ces protocoles “classiques”

Les problèmes de sélection: 3-le mode (m)

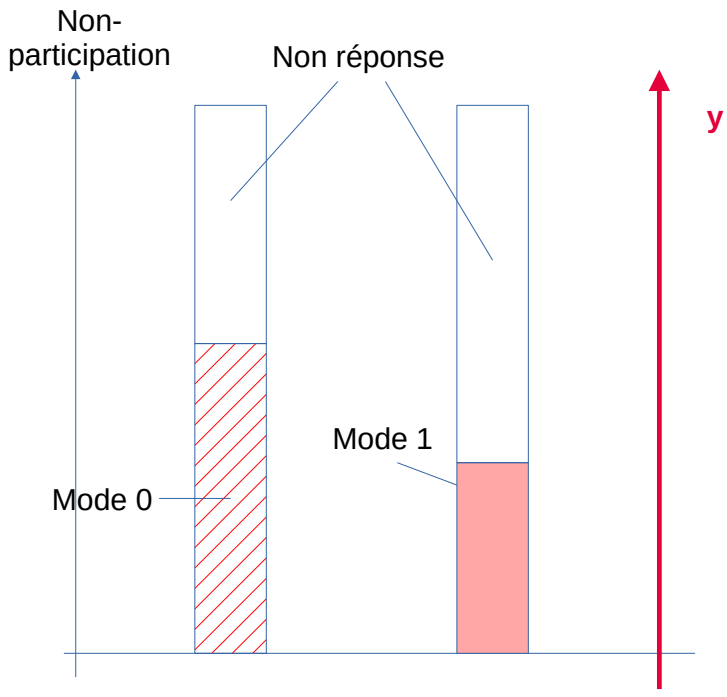
Si le mode est laissé au choix de l'enquêté:



Si $r(y)$, alors "mécaniquement", $m_i(y_i)$ et donc $y \perp\!\!\!\perp m$

Les problèmes de sélection: 3-le mode (m)

Si le mode est pré-sélectionné aléatoirement: 2 sous-échantillons



Alors la non-réponse est modale: r^m.

Si en outre $r(y)$
c'est-à-dire $r^m(y)$
alors $m_i(y_i^m)$
et donc $y \perp\!\!\!\perp m$

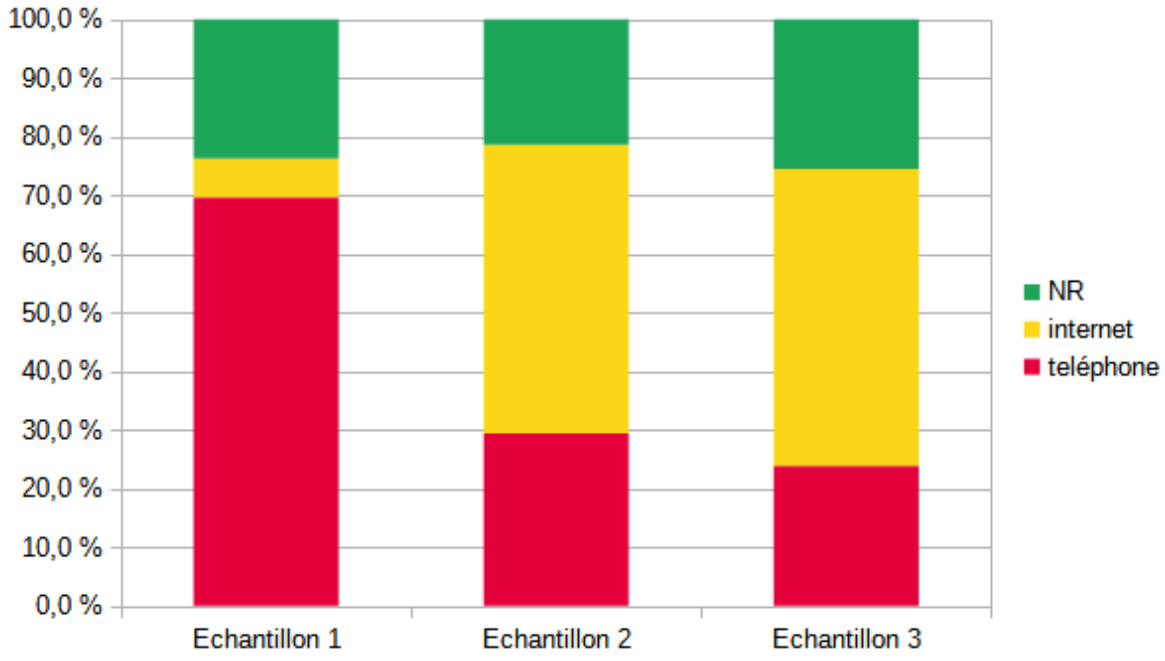
Les leçons des expériences CVS Panel et EpiCov

L'enquête expérimentale CVS Panel

↪ l'exemple d'un biais de mesure...

L'enquête expérimentale "CVS-Panel" (2019)

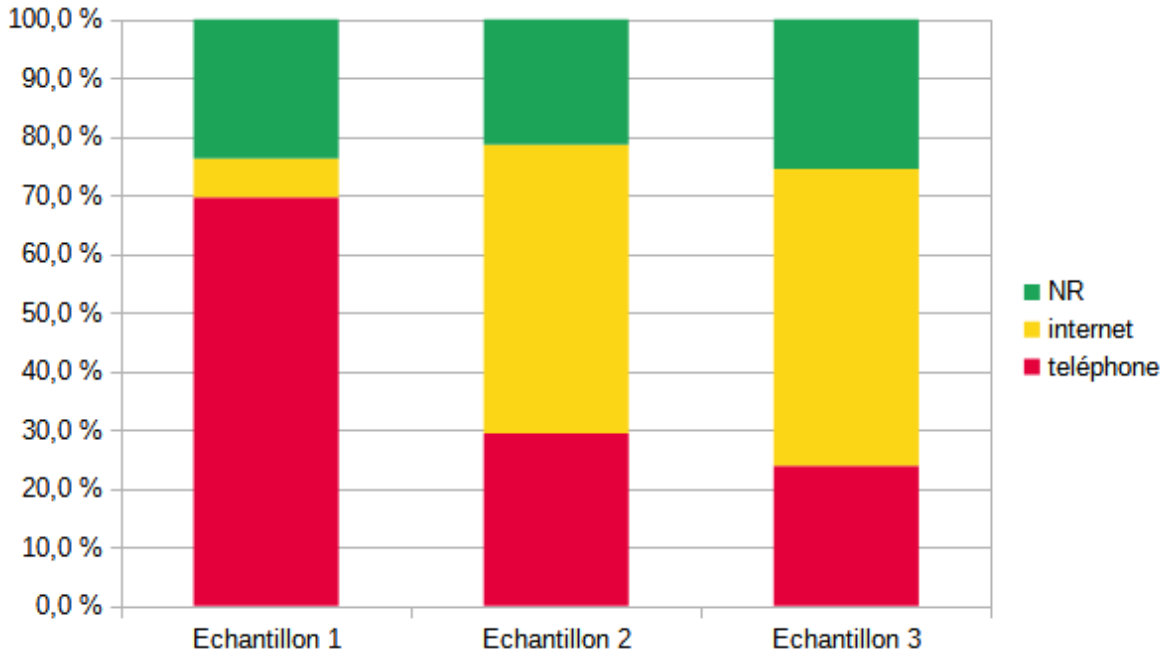
Interrogation int./tel. des répondants CVS2018



	CVS2019	CVS-Panel		
		Éch. 1	Éch. 2	Éch. 3
Vandalisme de la voiture (%)	3, 7 (0.2)	3, 8 (0.3)	6, 0 (0.4)	5, 7 (0.4)

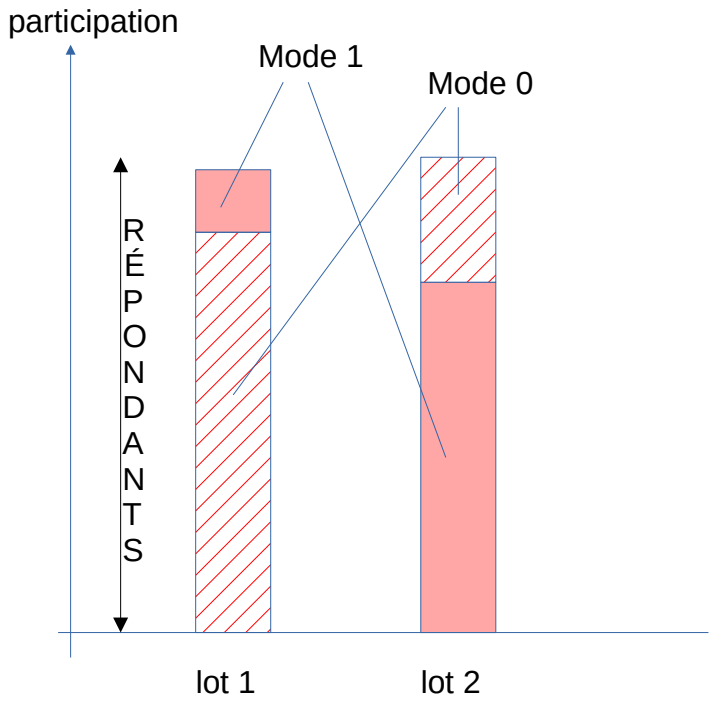
L'enquête expérimentale "CVS-Panel" (2019)

Interrogation int./tel. des répondants CVS2018



	CVS-Panel			
	Éch. 1	Éch. 2	Éch. 3	Éch. 2-3
Pr. de répondre par internet (%)	10, 5	59, 1	63, 8	61, 3

Identification de l'erreur de mesure par méthode instrumentale



estimateur de Wald:
$\Delta_{\overline{\mathscr{R}}}^{1/0}$ en 2SLS sur $\overline{\mathscr{R}}$

$$\begin{cases} y_i &= c^1 + \Delta_{\overline{\mathscr{R}}}^{1/0} . m_i + \epsilon_i \\ m_i &= c^2 + \beta . \mathbf{1}(i \in \text{lot}_1) + \nu_i \end{cases}$$

Composition Lot 1 ≠ Lot 2

$$\Rightarrow \begin{cases} \beta \neq 0 \\ \mathbf{1}(i \in \text{lot}_1) \perp\!\!\!\perp (\nu_i, \epsilon_i) \end{cases}$$

L'enquête expérimentale "CVS-Panel" (2019)

	CVS-Panel			
	Éch. 1	Éch. 2	Éch. 3	Éch. 2-3
Proba. de répondre par internet (%)	10, 5	59, 1	63, 8	61, 3
Impact internet sur V. de la v. (Wald– pts %)	traité	5, 3 (1,2)	4, 7 (1,1)	4, 9 (1,0)

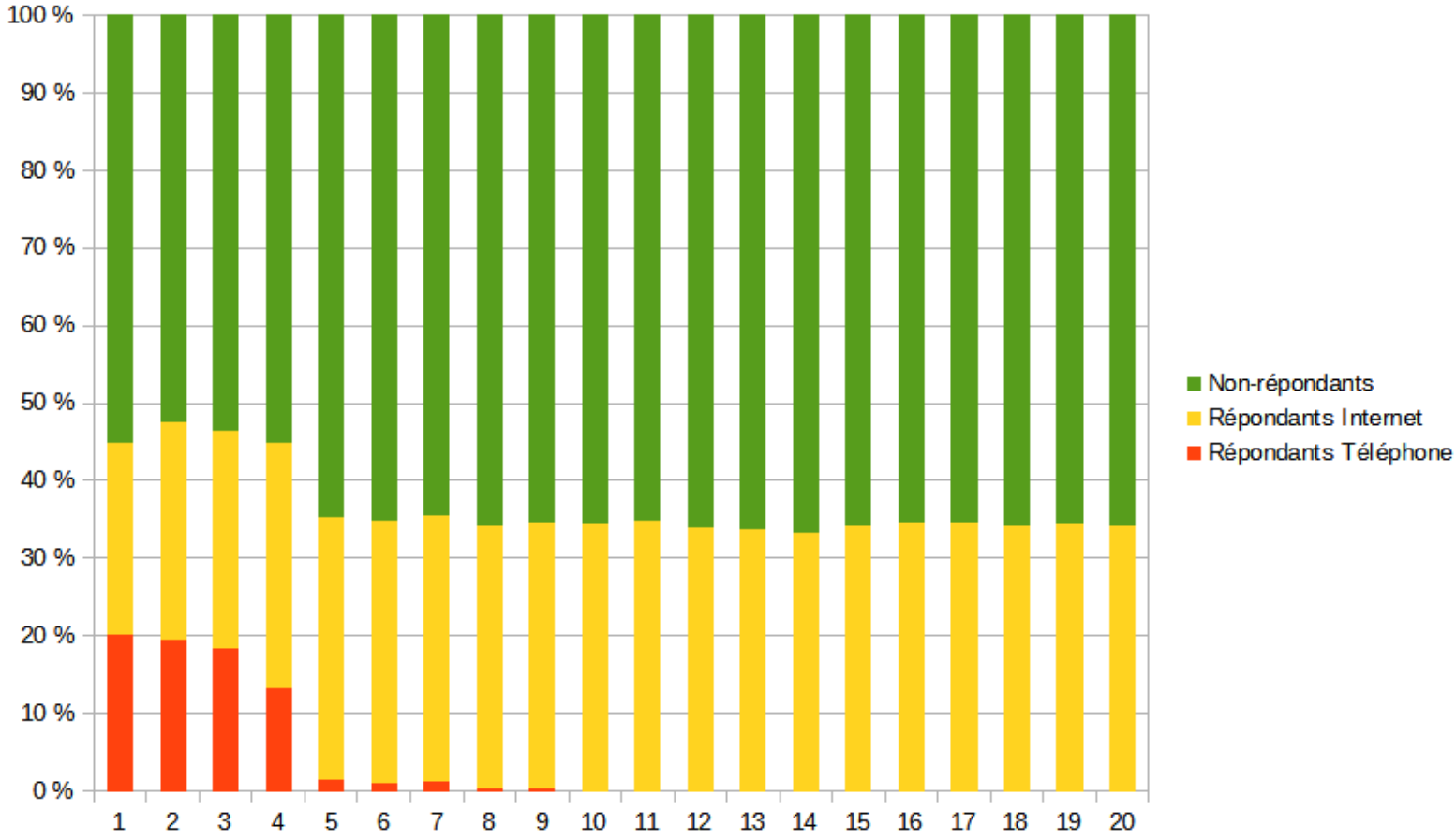
	CVS2019	CVS-Panel	
		non corrigé	corrigé
Vandalisme de la voiture (%)	3, 7 (0.2)	5, 2 (0.3)	3, 8 (0.3)

L’enquête EpiCov-vague 1

↪ (sans doute) Pas de biais de mesure... mais de la sélection endogène

EpiCov

370 000 sélectionnés, 135 000 répondants



Erreur de mesure?

Lot	Taux de réponse (%)	% de répondants tel.
lot 1	45,4	45,4
lot 2	47,6	40,7
lot 3	46,5	39,5
lot 4	45,0	29,3

Variables d'intérêt	Prévalences -lots		*WALD* (1/4)
	monomodes	multimodes	(*tel./int*)
Fièvre	8,0 (0,10)	7,0 (0,16)	−0.04 (2,2)
Perte de goût	2,8 (0,06)	2,5 (0,10)	0,68 (1,5)
Au moins 1 symptôme	27,2 (0,17)	23,5 (0,27)	0,30 (4,0)

Le traitement de la participation endogène (1)

La participation dépend de la variable d'intérêt (MNAR):

$$y \not\perp\!\!\!\perp r|X$$

Pour construire un estimateur convergent:

$$\hat{\mu} = \frac{1}{N} \sum_{i \in \mathscr{S}} \frac{y_i s_i r_i}{\pi_i}$$

En conditionnant sur X seulement, on postulait:

$$\left\{ \begin{array}{l} y \perp\!\!\!\perp (s, r)|X \\ \pi_i = \Pr(s_i = 1|X_i) \times \Pr(r_i = 1|X_i) \end{array} \right\} \Rightarrow \hat{\mu} \xrightarrow{\text{prob}} \mu$$

En conditionnant sur y aussi, on a:

$$\left\{ \begin{array}{l} y \perp\!\!\!\perp s|X \text{ mais } y \not\perp\!\!\!\perp r|X \\ \pi_i = \Pr(s_i = 1|X_i) \times \Pr(\mathbf{r_i = 1|X_i, y_i}) \end{array} \right\} \Rightarrow \hat{\mu} \xrightarrow{\text{prob}} \mu$$

Le traitement de la participation endogène (2)

- Il faut donc modéliser $\Pr(r_i = 1|X_i, y_i)$.
- Difficulté: y_i n'est observé que pour $r_i = 1$, donc modèle pas identifiable en général.
- Problème de sélection étudiés dans la littérature économétrique.
- Solution "classique" : modèle d'Heckman

Modèle d'Heckman

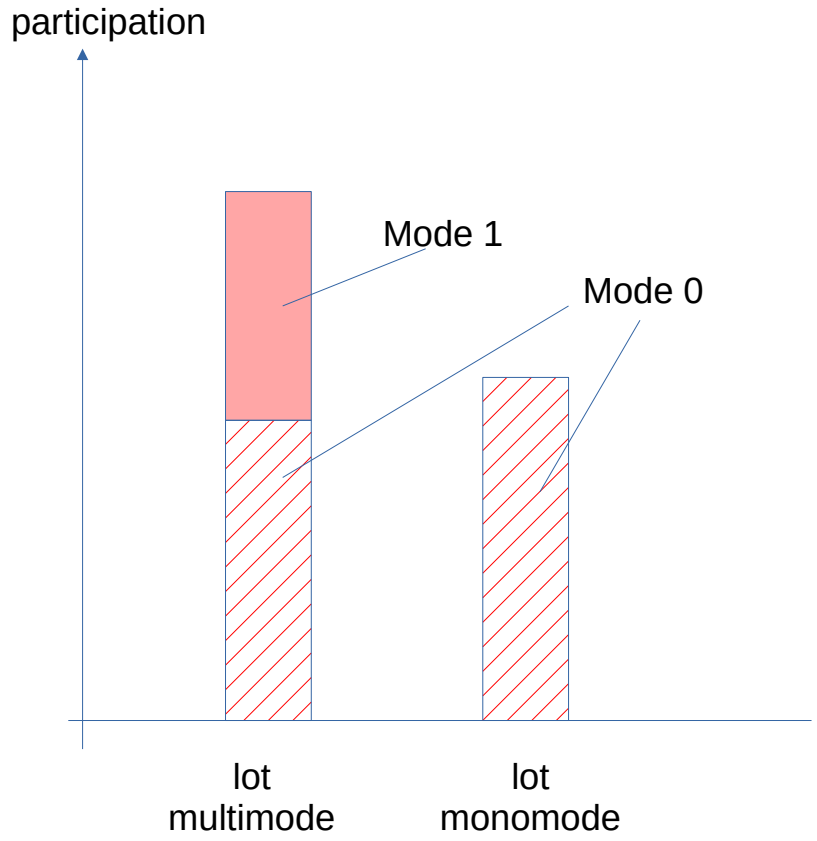
$$\begin{cases} (0) & y_i & = & \mathbf{1}(y_i^* \geqslant 0) \\ (i) & y_i^* & = & c^1 + \mathbf{X}_i\chi + \epsilon_i^1 \\ (ii) & r_i^* & = & c^0 + \mathbf{X}_i\beta + \mathbf{w}_i\psi + \epsilon_i^0 \\ (iii) & r_i & = & \mathbf{1}(r_i^* \geqslant 0) \end{cases}$$

- Estimation par MV des coefs $(c^0, c^1, \beta, \chi, \psi)$ et $\rho = \mathrm{cor}(\epsilon_i^0, \epsilon_i^1)$. L'endogénéité passe par ρ...
- On en déduit que:

$$\Pr(r_i = 1|\mathbf{X}_i, \mathbf{w}_i, y_i = 1) = \frac{\Phi_2(c^0 + \mathbf{X}_i\beta + \mathbf{w}_i\psi\ ,\ c^1 + \mathbf{X}_i\chi\ ;\ \rho)}{\Phi(c^1 + \mathbf{X}_i\chi)}$$

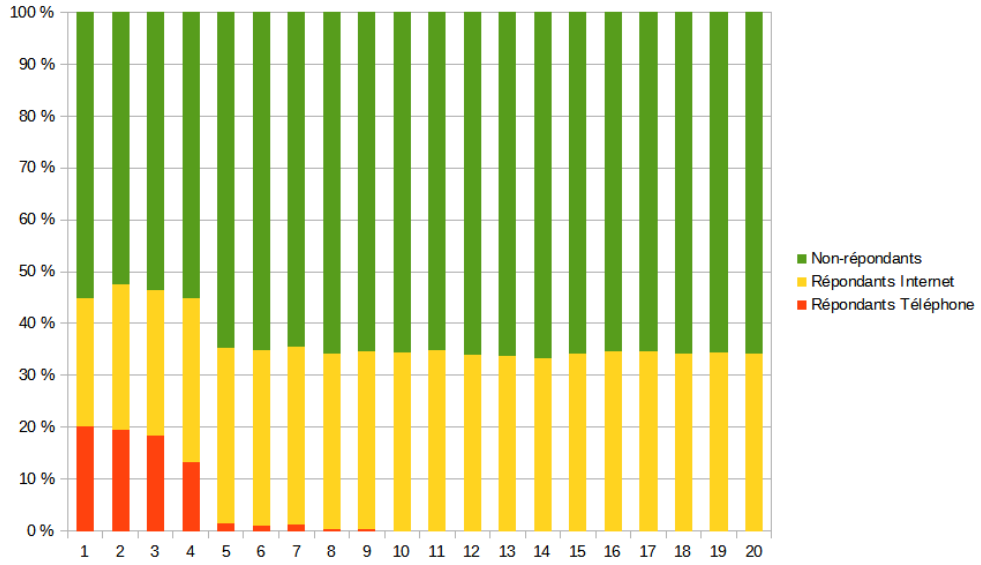
et, de même, $\Pr(r_i = 1|\mathbf{X}_i, \mathbf{w}_i, y_i = 0)$.

Générer un instrument de la participation grâce au multimode



Noter que $r_i = \mathbf{1}(r_i^* \geqslant 0)$ implique la monotonie de l'instrument, vérifiée ici.

Retour sur l'enquête EpiCov



Taux de réponse:

- 4 lots MMode ≈ 46%
- 16 lots internet ≈ 35%

⇒ l'indicatrice de lot multimode est un instrument

Sous l'hypothèse d'absence d'erreur de mesure (pur effet de sélection)

Prévalences (en %):

	CNR sur observabes		CNR	$\hat{\rho}$
	Lots monomode	Lots multimode	Heck.	
Fièvre	8, 0 (0,10)	7, 0 (0,16)	5, 0 (0,40)	0, 30 (0,05)
Perte de goût	2, 8 (0,06)	2, 5 (0,10)	1, 8 (0,30)	0, 18 (0,08)
Au moins 1 symptôme	27, 2 (0,17)	23, 5 (0,27)	14, 2 (0,55)	0, 53 (0,03)

Synthèse et conclusion

Après l'identification, la correction

- Biais de mesure (ε=espérance d'observations):
  - correction par imputation des observations biaisées
  - correction de l'estimateur d'HT, en moyenne pour ce qui concerne la contribution des observations biaisées
- Sélection endogène (ε=probabilités d'inclusion):
  - correction par repondération avec un modèle de $\Pr(r_i = 1|\mathbf{X}_i, \mathbf{w}_i, y_i = 1)$ et de $\Pr(r_i = 0|\mathbf{X}_i, \mathbf{w}_i, y_i = 1)$
- Si ni biais ni sélection endogène:
  - retour au traitement classique (Cond^t sur observables+GRH+calage) sur l'ensemble des sous-échantillons
  - Peu de perte d'efficacité par rapport à un échantillon unique de même taille

Sur le design d'enquête: dépasser l'approche mono-échantillon

- Viser un design permettant de générer des instruments de la participation et du choix du mode pour les répondants
- Multimode = moyen élégant de générer, par le design, un instrument de la participation en "empilant" les modes, donc régler des situations MNAR
- Mais nécessité de traiter les biais de mesure: le design peut grandement aider.
- Numériquement,
  - impact limité des biais de mesure sur les estimateurs d'ensemble car ne concernent qu'une sous-population;
  - impact possible délétère de la sélection endogène (MNAR) car effet amplificateur de la % de NR ⟶ s'en préoccuper!

  Le développement du multimode est une belle occasion pour cela!!

Encore du travail pour...

Mettre au point des designs qui permettent de tout traiter en même temps:

- aussi optimisés et pratiques que possible
- avec des modèles de traitement appropriés

On va apprendre beaucoup des enquêtes multimodes récentes et prochaines dont les designs permettent de générer des instruments. C'est le cas de l'ENL par exemple ou de l'enquête méthodologique VCV à l'Insee en 2023.