Les enjeux méthodologiques liés à l'usage de bases de sondage imparfaites

12es Journées de Méthodologie Statistique





Olivier Sautory
Département des méthodes statistiques
Insee

Insee
Mesurer pour comprendre

Plan

- définition, contenu d'une base de sondage
- imperfections d'une base de sondage
  - ✓ sous-couverture
  - ✓ sur-couverture
  - ✓ doublons
- utilisation d'informations auxiliaires
- bases de sondage multiples
- erreurs sur les variables de la base de sondage
- absence de base de sondage

Définition d'une base de sondage

- ✓ ***Population-cible*** = population sur laquelle on cherche à obtenir de l'information et estimer des paramètres
  (= *champ* de l'enquête)

- ✓ ***Base de sondage*** = liste d'unités permettant d'identifier les éléments de la population-cible, à partir de laquelle un échantillon est sélectionné selon une méthode probabiliste

  Idéalement : population-cible et base de sondage coïncident

Contenu d'une base de sondage

Pour chaque unité :

- ✓ un identifiant
- ✓ des données permettant de la repérer ou de la contacter (adresse, n° de téléphone…)
- ✓ des informations auxiliaires utilisables lors de la définition du plan de sondage et/ou lors de la phase d'estimation

Défauts d'une base de sondage

- Sous-couverture
- Sur-couverture
- Présence de répétitions (doublons)
- Erreurs sur les variables contenues dans la base
- Manque de fraîcheur
- Absence de base

1. Sous-couverture

Des unités de la population cible sont absentes de la base de sondage.

Causes possibles :

- Délai entre la date de sélection de l'échantillon dans la base, qui peut elle-même avoir une certaine « ancienneté », et la collecte : des unités ont pu « naître » dans le champ de l'enquête.
- Une partie de la population-cible est absente de la seule base de sondage disponible.

Si on ignore la sous-couverture… (1)

U = population-cible, taille N

U_F = population de la base de sondage, taille N_F

U_0 = population omise de la base de sondage, taille N_0

$$U = U_F \bigcup U_0 \qquad N = N_F + N_0$$

Variable d'intérêt ***Y***

Dans U : $\bar{Y}$ = moyenne Y = total

Dans U_F : $\bar{Y}_F$ = moyenne Y_F = total

Dans U_0 : $\bar{Y}_0$ = moyenne Y_0 = total

$$Y = Y_F + Y_0$$

Si on ignore la sous-couverture… (2)

On note : $r = \dfrac{\bar{Y}_0}{\bar{Y}_F}$ $\tau_0 = \dfrac{N_0}{N}$ = taux d'omission

- Biais relatif pour un estimateur sans biais du total Y_F

$$\frac{Y_F - Y}{Y} = \frac{-r\tau_0}{r\tau_0 + (1-\tau_0)} < 0$$

$\rightarrow$ biais petit dès que *r* ou τ_0 est petit

- Biais relatif pour un estimateur sans biais de la moyenne $\bar{Y}_F$

$$\frac{\bar{Y}_F - \bar{Y}}{\bar{Y}} = \frac{\tau_0(1-r)}{r\tau_0 + (1-\tau_0)}$$

$\rightarrow$ biais nul si r = 1, négligeable si r $\cong$ 1 et τ_0 petit

Évaluation de la sous-couverture (1)

Comparaison avec des sources externes indépendantes

1. Au niveau agrégé : calcul de taux de couverture globaux, ou par sous-populations

   Exemple : comparaison de structures par sexe et âge

2. Appariement individuel : *Dual system estimation* (Wolter, 1983)

   S_{ext} = échantillon aréolaire couvrant l'ensemble de la population

   On apparie les unités de S_{ext} avec celles de U_F sur les aires de l'échantillon.

Évaluation de la sous-couverture (2)

	Présent U (S_{ext})	
Présent U_F	$\hat{N}_{11}$	N_F
Absent U_F	$\hat{N}_{21}$	
Total	$\hat{N}_{+1}$	N

Estimation du taux de couverture N_F / N de U_F par : $\hat{N}_{11} / \hat{N}_{+1}$

Calcul fait par post-strate

Évaluation de la sous-couverture (3)

Et si des unités sont présentes dans U_F et absentes de S_{ext}…

	Présent U (S_{ext})	Absent U (S_{ext})	
Présent U_F	$\hat{N}_{11}$	$\hat{N}_{12}$	N_F
Absent U_F	$\hat{N}_{21}$	N_{22}	N_{2+}
	$\hat{N}_{+1}$	N_{+2}	N

Estimation de la taille N de la population U par :

$$\hat{N}_{11} + \hat{N}_{12} + \hat{N}_{21} + (\hat{N}_{12} \times \hat{N}_{21})/\hat{N}_{11}$$

Estimateur connu sous le nom d'estimateur de Peterson (1896), de Chandrasekar & Deming (1949) ; ou méthode de capture-recapture.

Sous-couverture : que faire ?

Redéfinition de la population-cible, pour la faire coïncider avec la base de sondage !

Linking procedure : si on peut définir une règle d'association entre tous les éléments omis (U_0) et des éléments de la base de sondage U_F, on enquête également les éléments omis reliés à des éléments de l'échantillon s_F

Exemple : *half-open interval procedure* (Kish, 1965)

Bases de sondage multiples (voir plus loin)

Échantillonnages adaptatifs (voir plus loin)

2. Sur-couverture

Des unités de la base de sondage ne font pas partie de la population-cible.

Causes possibles :

- Délai entre la date de sélection de l'échantillon dans la base, qui peut elle-même avoir une certaine « ancienneté », et la collecte : des unités de la base ont pu « disparaître » du champ de l'enquête.

- La seule base de sondage disponible contient des éléments n'appartenant pas à la population-cible.

Si on ignore la sur-couverture… (1)

U_F = population de la base de sondage, taille N_F

U = population-cible, taille N

U_0 = population de la base de sondage hors champ, taille N_0

$$U_F = U \bigcup U_O \quad N_F = N + N_O$$

Variable d'intérêt ***Y***

Dans U_F : $\bar{Y}_F$ = moyenne Y_F = total

Dans U : $\bar{Y}$ = moyenne Y = total

Dans U_0 : $\bar{Y}_0$ = moyenne Y_0 = total

$$Y_F = Y + Y_O$$

Si on ignore la sur-couverture… (2)

On note : $r = \dfrac{\bar{Y}_0}{\bar{Y}}$ $\tau_0 = \dfrac{N_0}{N_F} =$ taux de sur-couverture

- Biais relatif pour un estimateur sans biais d'un total Y_F :

$$\frac{Y_F - Y}{Y} = \frac{r\tau_0}{1-\tau_0} > 0$$

$\rightarrow$ biais petit dès que *r* ou τ_0 est petit

- Biais relatif pour un estimateur sans biais d'une moyenne $\bar{Y}_F$:

$$\frac{\bar{Y}_F - \bar{Y}}{\bar{Y}} = \tau_0 (r-1)$$

$\rightarrow$ biais nul si r = 1, négligeable si r $\cong$ 1 et τ_0 petit

Évaluation de la sur-couverture

Comparaison avec des sources externes indépendantes

Au niveau agrégé : calcul de taux de couverture globaux, ou par sous-populations

Exemple : comparaison de structures par sexe et âge

Enquête spécifique sur le terrain (exemple : enquêtes post-censitaires)

Sur-couverture : que faire ?

Souvent, on peut identifier parmi les unités de l'échantillon celles qui sont dans le champ et celles qui sont hors champ

$\rightarrow$ U est un ***domaine*** de U_F

$\rightarrow$ estimation d'un total, d'une moyenne, sur un domaine

Mais : pour les unités non-répondantes de l'échantillon, il peut être difficile de savoir si les unité sont en réalité hors-champ, ou de véritables non répondantes.

Que faire ?

- apparier l'échantillon avec la version la plus récente de la base de sondage
- mobiliser des sources externes

3. Les doublons

Des unités apparaissent plus d'une fois dans la base.

Causes possibles :

- Pour éviter la sous-couverture, la base résulte de la fusion de plusieurs listes contenant pour partie les mêmes unités.

- Plusieurs identifiants désignent en réalité la même unité.

En général, les doublons détectés sans ambiguïté dans la base de sondage sont enlevés de la base !

Si on ignore les doublons… (1)

Population – cible $U = \{1 \ldots k \ldots N\}$

Base de sondage $U_F = \{1 \ldots i \ldots N_F\}$

On suppose : pas de sur-couverture

On note L_k le nombre d'unités de U_F associés à l'unité k de U

On suppose $L_k \geq 1$ pour tout k : pas de sous-couverture
Si $L_k > 1$ pour certains k : doublons

L_k prend la valeur a (= 1…A) pour N_a unités de U

$$N = \sum_{a=1}^{A} N_a \qquad N_F = \sum_{a=1}^{A} a\, N_a$$

Si on ignore les doublons… (2)

- Biais relatif pour un estimateur sans biais du total Y_F :

$$\frac{Y_F - Y}{Y} = \frac{\sum_{k \in U} (L_k - 1) Y_k}{\sum_{k \in U} Y_k} > 0$$

- Biais relatif pour un estimateur sans biais de la moyenne $\bar{Y}_F$

$$\frac{\bar{Y}_F - \bar{Y}}{\bar{Y}} = \frac{\sum_{a=1}^{A} \left(\frac{a N_a}{N_F} - \frac{N_a}{N} \right) \bar{Y}_a}{\sum_{a=1}^{A} \frac{N_a}{N} \bar{Y}_a} \qquad avec \quad \bar{Y}_a = \frac{\sum_{k : L_k = a} Y_k}{N_a}$$

$\rightarrow$ biais nul si les moyennes $\bar{Y}_a$ sont toutes égales

Les doublons : que faire ?

Sélection d'un échantillon s_F d'unités i, probabilités d'inclusion π_i

$\hat{Y}_F = \sum_{i \in s_F} y_i / \pi_i$ est un estimateur biaisé de Y

Deux cas de figure :

- ✓ La collecte ne permet pas de connaître la « multiplicité » des unités enquêtées :
  - → il existe des méthodes permettant d'obtenir des estimateurs sans biais, fondées uniquement sur l'observation de doublons dans l'échantillon

- ✓ Si la collecte permet de connaître la « multiplicité » des unités enquêtées, il existe plusieurs méthodes permettant de construire des estimateurs sans biais prenant en compte la multiplicité.

4. Utilisation d'information auxiliaire pour corriger les erreurs de couverture (1)

Ajustement par le ratio (sous-couverture)

On note s_F un échantillon tiré de U_F, avec des probabilités d'inclusion π_i

$\hat{Y}_F = \sum_{s_F} \frac{y_i}{\pi_i}$ est un estimateur biaisé négativement du total Y sur U

Si X_U, total sur U d'une variable ***X*** est connu, on peut estimer Y par un ajustement par le ratio :

$$\hat{Y}_r = X_U \frac{\sum_{s_F} y_i / \pi_i}{\sum_{s_F} x_i / \pi_i}$$

Biais relatif : $\frac{E(\hat{Y}_r) - Y}{Y} \approx \frac{R_{U_F}}{R_U} - 1,$ avec $R_{U_F} = \frac{Y_{U_F}}{X_{U_F}}$ et $R_U = \frac{Y_U}{X_U}$

négligeable si $R_{U_F} \cong R_U$

Utilisation d'information auxiliaire pour corriger les erreurs de couverture (2)

Särndal & Lundstrom, 2005

Ajustement par post-stratification

Cas d'un sondage aléatoire simple stratifié, avec non-réponse

N_h = taille de la post-strate h dans U

r_{Fh} = échantillon des répondants de s_F (dans le champ) dans la post-strate h

m_{Fh} = nombre de répondants de s_F (dans le champ) dans la post-strate h

$$\hat{Y}_{ps} = \sum_{h=1}^{H} \frac{N_h}{m_{F_h}} \sum_{r_{F_h}} y_i$$

sans biais si les éléments d'une post-strate h ont la même probabilité de réponse et la même "probabilité" d'appartenir à la base U_F.

Utilisation d'information auxiliaire pour corriger les erreurs de couverture (3)

Calage (généralisé) (Deville, 1998 ; Estevao & Särndal, 2000)

Calage sur un vecteur X_U de totaux connus sur U

$$\hat{Y}_{cal} = \sum_{r_F} w_i \, y_i = \sum_{r_F} v_i \, d_i \, y_i \quad \text{avec } d_i = 1/\pi_i$$

r_F = échantillon des répondants de s_F dans le champ

$$v_i = 1 + \left(X_U - \sum_{r_F} d_i \, x_i \right)' \left(\sum_{r_F} d_i \, z_i \, x_i' \right) x_i$$

z_i = vecteur de variables explicatives de la non-réponse et du défaut de couverture

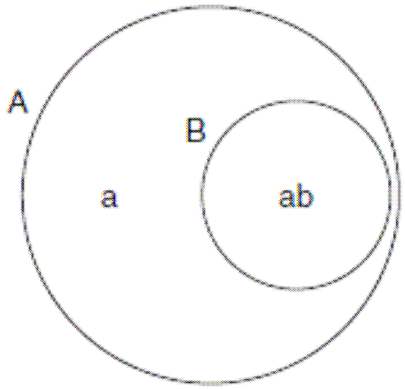
5. Bases de sondage multiples

Objectifs :

- Améliorer la couverture
- Coûts de collecte plus faible pour enquêter des populations rares
- Améliorer les taux de réponse

Lohr, 2009, 2011

Bases imbriquées



La base B est un sous-ensemble de la base A

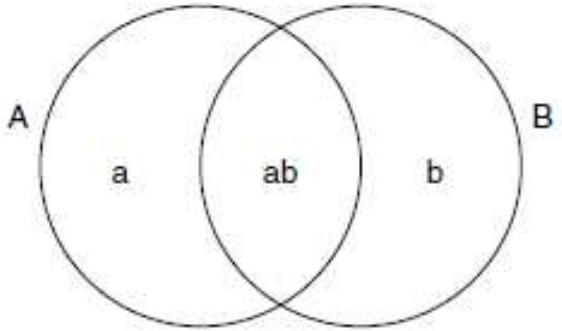
Exemple : A = base « générale », avec peu d'informations (ou base aréolaire)

B = liste plus adaptée à l'objectif de l'enquête, mais incomplète

Si identification possible dans A des éléments de B : on enlève les éléments de B de la base A avant la sélection des échantillons

→ cas particulier de sondage stratifié

Bases chevauchantes (1)



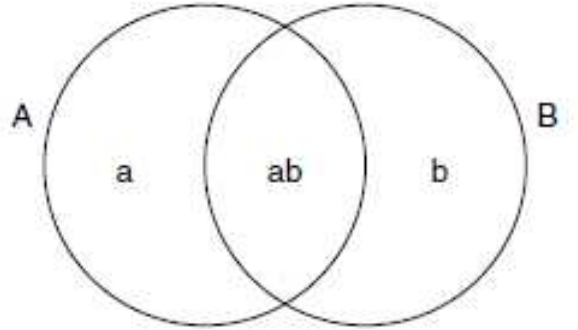
Trois domaines : a, b, et ab

Si appartenance aux domaines connue à l'avance : on supprime de la base B les éléments de ab. Mais rarement possible !

Après l'enquête, on peut enlever de l'échantillon les éléments de ab sélectionnés à partir de B $\rightarrow$ mais on perd des éléments !

Bases chevauchantes (2)

La situation la plus fréquente : on garde tous les éléments



Total $Y = Y_a + Y_b + Y_{ab}$

Échantillon s_A tiré dans A $\rightarrow$ estimateurs $\hat{Y}_a^A$ et $\hat{Y}_{ab}^A$

Échantillon s_B tiré dans B $\rightarrow$ estimateurs $\hat{Y}_b^B$ et $\hat{Y}_{ab}^B$

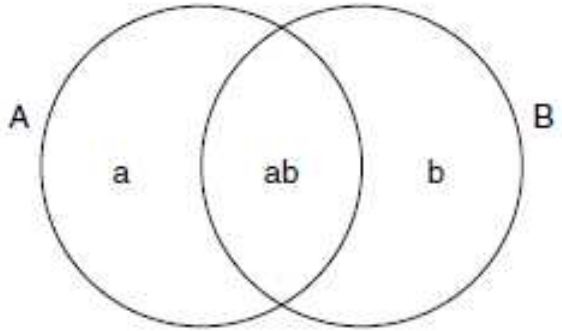
Question : comment combiner $\hat{Y}_{ab}^A$ et $\hat{Y}_{ab}^B$ pour estimer Y_{ab} ?

Plusieurs méthodes proposées.

Bases chevauchantes – estimation (1)

Moyenne pondérée des estimateurs
(Hartley, 1962)

$$\hat{Y}(\theta) = \hat{Y}_a^A + \theta \hat{Y}_{ab}^A + (1-\theta)\hat{Y}_{ab}^B + \hat{Y}_b^B$$



Interprétation en termes de poids :

Si i∈ A : $w_i^A \rightarrow \tilde{w}_i^A = m_i^A w_i^A$, avec $m_i^A = \begin{cases} 1 \ si \ i \in a \\ \theta \ si \ i \in ab \end{cases}$

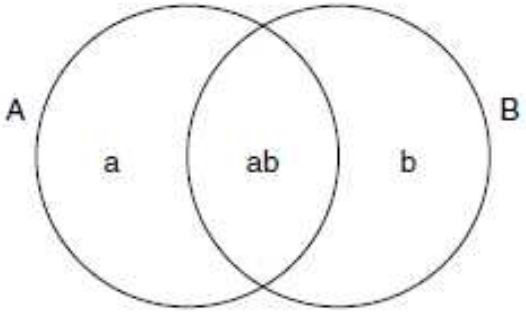
Hartley : choisir θ_{opt} qui minimise $V\left(\hat{Y}(\theta)\right)$

Plus $V\left(\hat{Y}_{ab}^B\right)/V\left(\hat{Y}_{ab}^A\right)$ est élevé, plus θ_{opt} est élevé

Bases chevauchantes – estimation (2)

Variante proposée par Fuller et Burmeister (1972)

$$\hat{Y}_{FB}(\beta) = \hat{Y}_a^A + \beta_1 \hat{Y}_{ab}^A + (1-\beta_1)\hat{Y}_{ab}^B + \hat{Y}_b^B + \beta_2 (\hat{N}_{ab}^A - \hat{N}_{ab}^B)$$



choisir $\beta=(\beta_1, \beta_2)_{opt}$ qui minimise $V\left(\hat{Y}_{FB}(\beta)\right)$

$\rightarrow$ poids aléatoires et dépendent de la variable d'intérêt

Estimateur du pseudo-maximum de vraisemblance (Skinner et Rao, 1996)

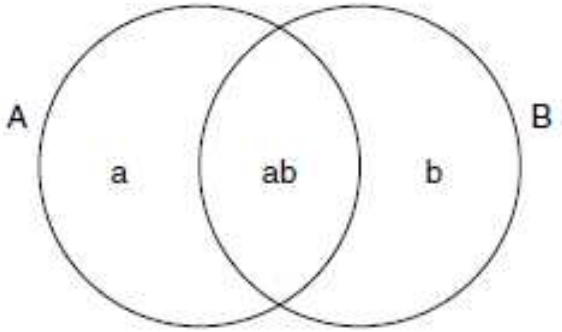
Estimateur modifiant l'estimateur de FB, fondé sur N_A, N_B, et un estimateur spécifique de N_{ab}

Poids aléatoires, mais ne dépendent pas de la variable d'intérêt.

(Lohr & Rao, 2000, 2006)

Bases chevauchantes – estimation (3)

Estimateurs à base de sondage unique
(Bankier, 1986 ; Kalton et Anderson, 1986)



On traite tous les éléments comme s'ils provenaient d'une base de sondage unique, en ajustant les poids pour les éléments du domaine *ab* :

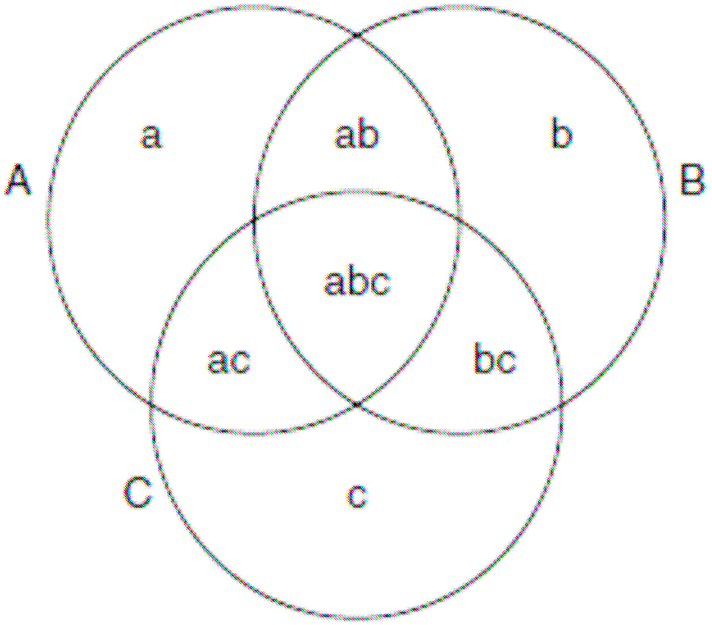
$$\text{Si } i \in A : w_i^A \rightarrow \tilde{w}_i^A = m_i^A \, w_i^A, \text{ avec } \quad m_i^A = \begin{cases} 1 & \text{si } i \in a \\ w_i^B / (w_i^A + w_i^B) & \text{si } i \in ab \end{cases}$$

$$\text{Par exemple, pour } i \in ab, \text{ si } \quad w_i^A = 1/\pi_i^A \quad \text{et} \quad w_i^B = 1/\pi_i^B :$$

$$\tilde{w}_i = 1/(\pi_i^A + \pi_i^B)$$

Poids ne dépendent pas de la variable d'intérêt.

Cas de 3 bases de sondage et plus (1)



Q bases de sondage A_1 ... A_q ... A_Q ; Q échantillons s_1 ... s_q ... s_Q

$\rightarrow$ D (= $2^Q - 1$) domaines d

Les estimateurs précédents se généralisent au cas de Q bases.

Cas de 3 bases de sondage et plus (2)

Forme générale d'un estimateur $\hat{Y} = \sum_{q=1}^{Q} \sum_{i \in s_q} m_i^{(A_q, d)} w_i^{A_q} y_i$

Estimateur à poids fixes :

$m_i^{(A_q, d)} = m^{(A_q, d)}$ avec $\sum_{q=1}^{Q} m^{(A_q, d)} = 1$ pour tout domaine d

Estimateur fondé sur la multiplicité (Lavallée 2002, Mecatti 2007) :

$m^{(A_q, d)} = 1 / ($nombre de bases de sondage qui contiennent d$)$

Q = 2 $\rightarrow$ estimateur de Hartley avec θ = 1/2

$\tilde{w}_i^A = \frac{w_i^A}{2}$ et $\tilde{w}_i^B = \frac{w_i^B}{2}$ si i$\in$ ab

Spécificités des bases de sondage multiples

- Correction de la non-réponse : sur les échantillons individuels, ou après les avoir combinés ?

  Les non-réponses peuvent être très différentes d'un échantillon à l'autre → nécessité de modèles de correction distincts

- Calage sur les tailles des bases de sondage, sur les tailles de strates, sur des effectifs dans la population,…

- Estimations de variance : techniques de linéarisation de Taylor, jackknife, bootstrap

- Si modes ou protocoles de collecte différents selon la base de sondage : risque d'effets de mode, risque de mesurer des concepts différents

- Les estimateurs supposent connue l'appartenance de chaque unité de l'échantillon aux différents domaines (a, b, ab) : estimateurs biaisés en cas d'erreur de classification

6. Erreurs sur les variables de la base de sondage

- Erreurs sur les données d'identification ou de repérage : difficultés de repérage des enquêtés pendant la collecte

- Erreurs sur les variables auxiliaires : échantillonnage et méthodes de correction de la non-réponse et d'estimation (calage par exemple) moins efficaces.

  Manque de fraîcheur de la base : mises à jour pas assez fréquentes.

  Exemple : stratification des échantillons d'enquêtes-entreprises selon l'effectif salarié datant de l'année T-2

  → stratification moins efficace, i.e. moindre précision des estimations

  → problème des *strata jumpers* : unités influentes en raison de leur poids "inapproprié"

7. Que faire en l'absence de base de sondage ?

Solutions « classiques »

(a) Base de sondage trop coûteuse à construire :

sondage à deux degrés, avec constitution de bases de sondage sur les unités primaires sélectionnées (souvent des aires géographiques).

(b) Pas de base de sondage pour une population d'intérêt rare :

sondage en deux phases :

- 1ère phase (peu coûteuse…) : échantillon de grande taille servant à repérer l'appartenance de l'individu à la population rare ;
- 2ème phase : interrogation des individus de cette population rare par une collecte « classique ».

Population intermédiaire - sondage indirect

Pas de base de sondage correspondant à la population-cible

Mais il existe une base de sondage reliée indirectement à la population-cible :

U_1 = population de la base de sondage

U_2 = population-cible

Exemple : enquête auprès des sans-domicile

Bases intermédiaires = services d'hébergement

services de distribution de repas

On suppose qu'il existe des « liens » entre les unités i de U_1 et les unités j de U_2 :

$\ell_{j,i}$ = 1 si lien entre j et i

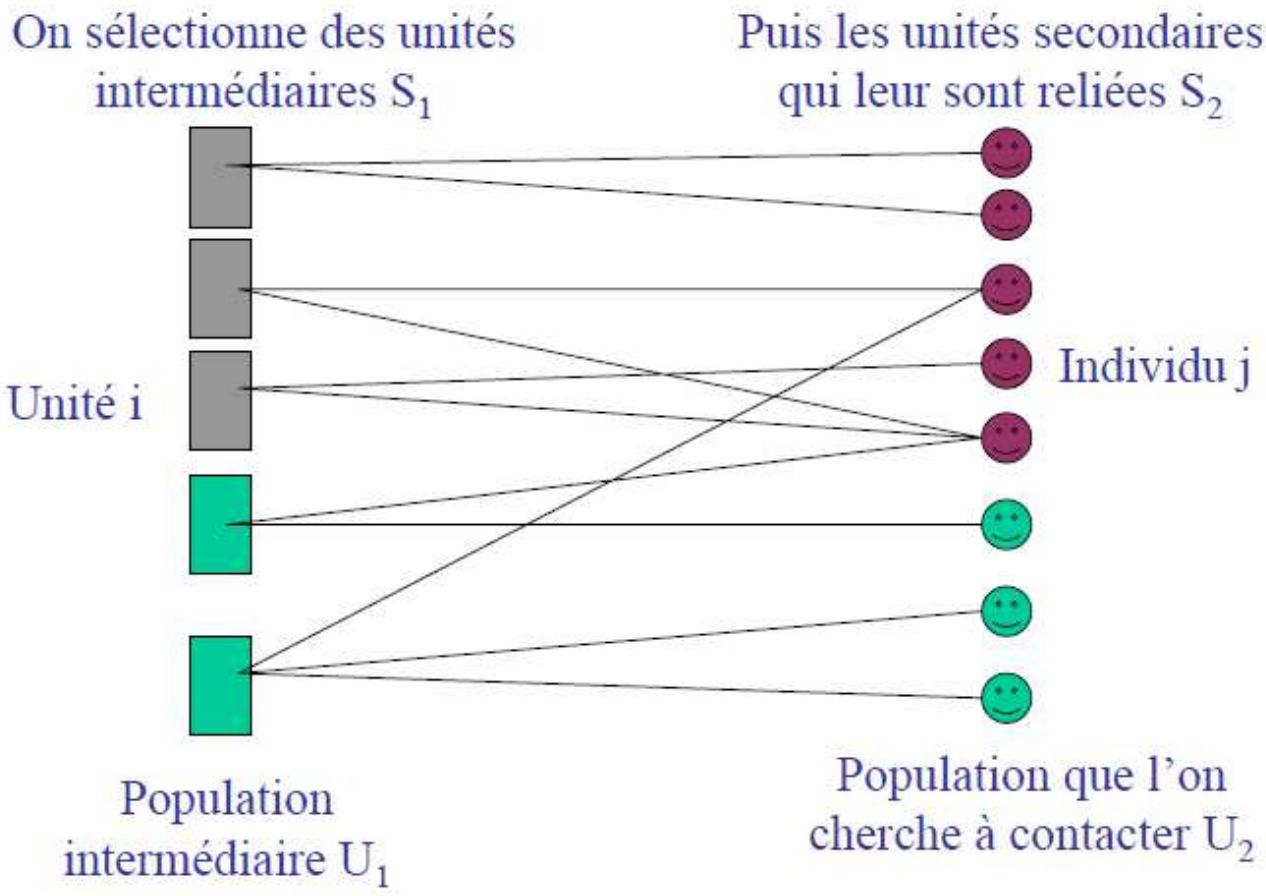
= 0 sinon

Sondage indirect

On suppose que toute unité j de U_2 possède au moins un lien avec une unité i de U_1.

Principe : sélectionner un échantillon s_1 dans U_1 afin de produire une estimation sur U_2 en utilisant les liens entre U_1 et U_2.

s_2 = échantillon de U_2 constitué de toutes les unités j en lien avec les unités i de s_1.



Sondage indirect – partage des poids

Pondération de l'unité i de s_1 : $w_i = 1/\pi_i$

Y = variable intérêt mesurée mesurée pour toute unité j de s_2

Le total de ***Y*** est estimé sans biais par :

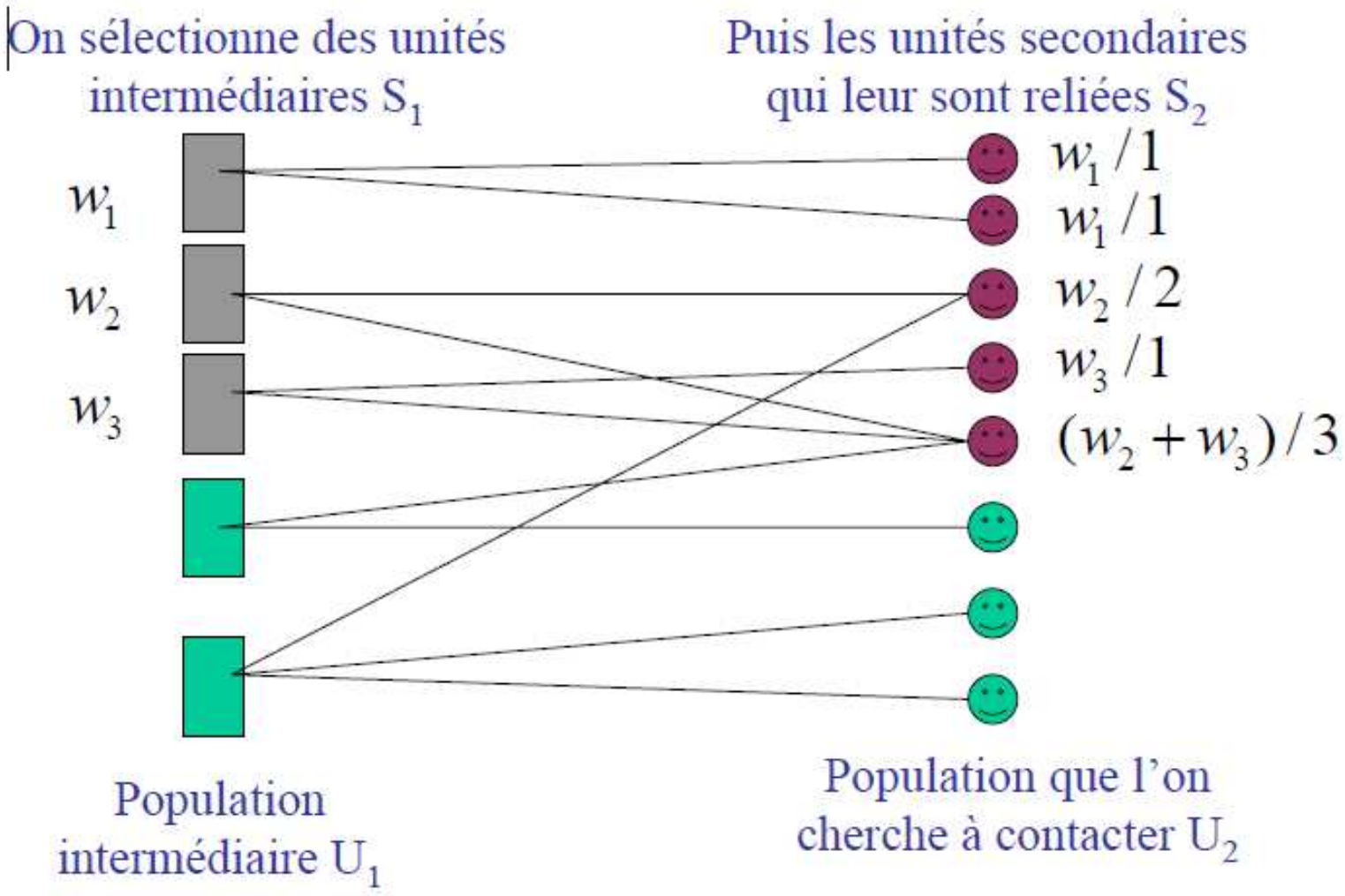
$$\hat{Y} = \sum_{j \in s_2} w_j \, y_j$$

où la pondération de l'unité j de s_2 vaut :

$$w_j = \frac{\sum_{i \in s_1} \ell_{j,i} \, w_i}{L_j}$$

avec $L_j = \sum_{i \in s_1} \ell_{j,i}$ = nombre total de liens de j avec U_1

Sondage indirect – partage des poids



Échantillonnage boule de neige

On fait l'hypothèse que les individus de la population rare P se connaissent les uns les autres.

On identifie quelques individus appartenant à P, on les enquête en leur demandant s'ils connaissent d'autres individus de P, que l'on enquête à leur tour…

L'échantillon prend de l'ampleur comme une boule de neige, on arrête lorsque l'échantillon est de taille suffisante.

Méthode non probabiliste.

Certains individus peuvent n'avoir aucune chance être sélectionnés. Il faut de fortes hypothèses de modélisation (en général non vérifiées…) pour faire de l'inférence statistique.

Permet d'apprendre de l'information sur P.

Échantillonnage déterminé selon les répondants

Heckathorn (1997)

Adaptation de l'échantillonnage « boule de neige ».

Procédure de « recrutement » des unités de l'échantillon par un système de coupons, qui permet de reconstituer les schémas de recrutement.

Modélisation mathématique fondée sur la théorie des chaînes de Markov, qui permet de compenser le caractère non aléatoire de la sélection, pour produire, sous certaines hypothèses, des estimations non biaisées sur la population d'intérêt.

Échantillonnage en grappes adaptatif

Thompson (1990)

Étude d'une population rare P constituée de personnes géographiquement regroupées

Grappes = aires géographiques - Sélection aléatoire de grappes

Chaque fois que l'on observe pour une ou plusieurs unités d'une grappe échantillonnée une valeur de la variable d'intérêt « intéressante », les grappes « voisines » sont ajoutées à l'échantillon, et on recommence.

Peut être utilisé aussi dans un contexte de réseau : réseau social, personnes souffrant de maladies épidémiques

On sait construire des estimateurs sans biais, sous certaines hypothèses.

Merci de votre attention !