Utilisation des probabilités d'inclusion exactes pour le sondage indirect en population asymétrique

Henri Bodet(Pise) - Arnaud Fizzala (Division Sondages)

INSEE

Journées de Méthodologie Statistique - 31 mars 2022

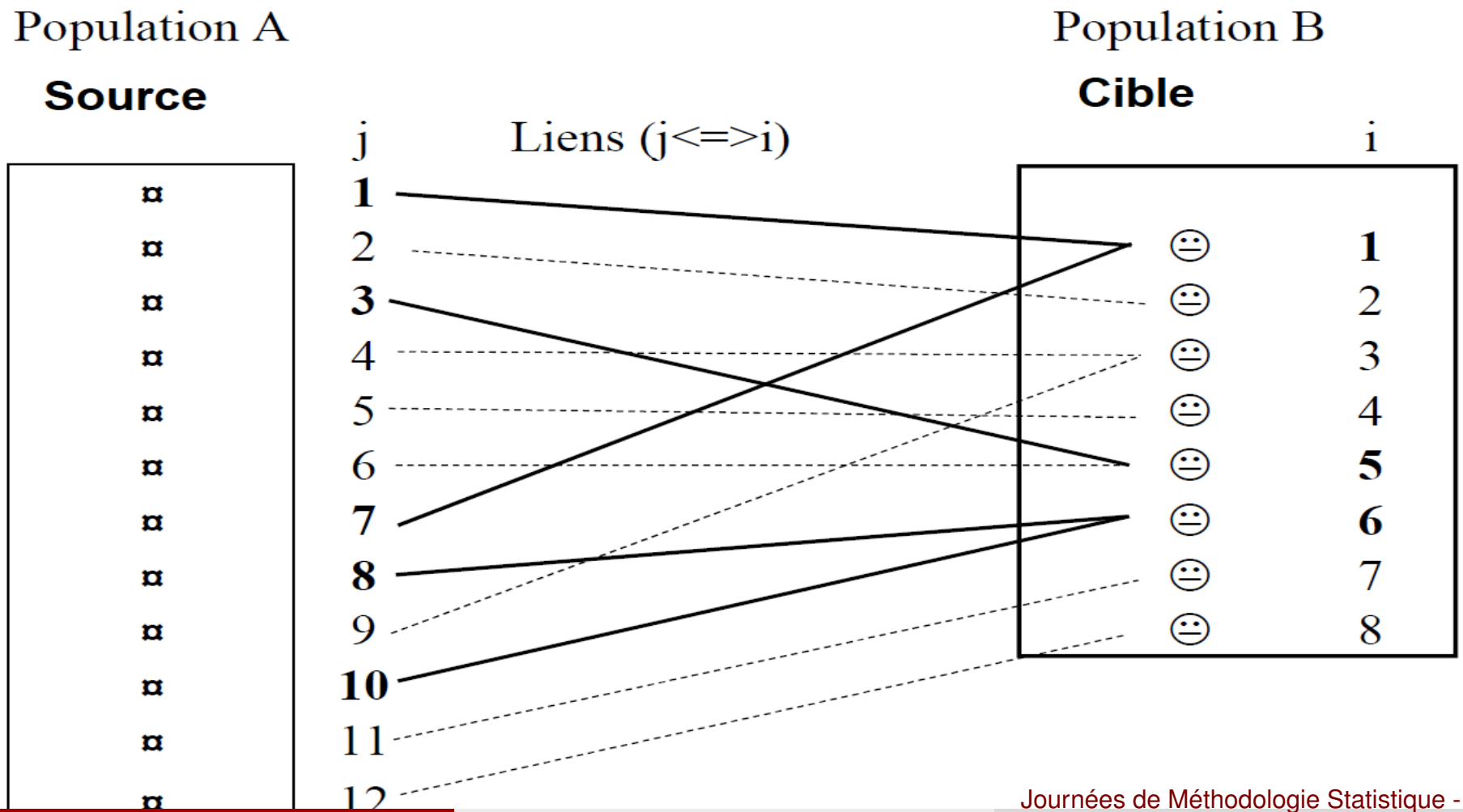
Plan

Plan

Le sondage indirect

Sondage indirect : situation où sélectionne des unités qui ne sont pas "directement" les unités interrogées.

Les unités interrogées sont liées à une (ou plusieurs) unités sélectionnées.



Le sondage indirect

Exemples enquêtes ménages

- Enquête auprès des bénéficiaires de repas chauds (une personne peut fréquenter plusieurs lieux)
- Enquête auprès des touristes en Bretagne (plusieurs lieux touristiques interrogés)

Le sondage indirect

Pour pondérer les unités "indirectement" interrogées, la solution standard est la méthode généralisée du partage des poids (Lavalée).

Le poids d'une unité "cible" est une moyenne des poids des unités "sources" qui lui sont liées et sont dans l'échantillon.

Sondage indirect et enquêtes auprès des entreprises : deux situations

Deux cas d'application du sondage indirect :

- Enquête auprès d'unités légales : reconstitution du contour d'une entreprise (d'un groupe)
- Enquête auprès des associations : bases de sondages avec des unités en commun

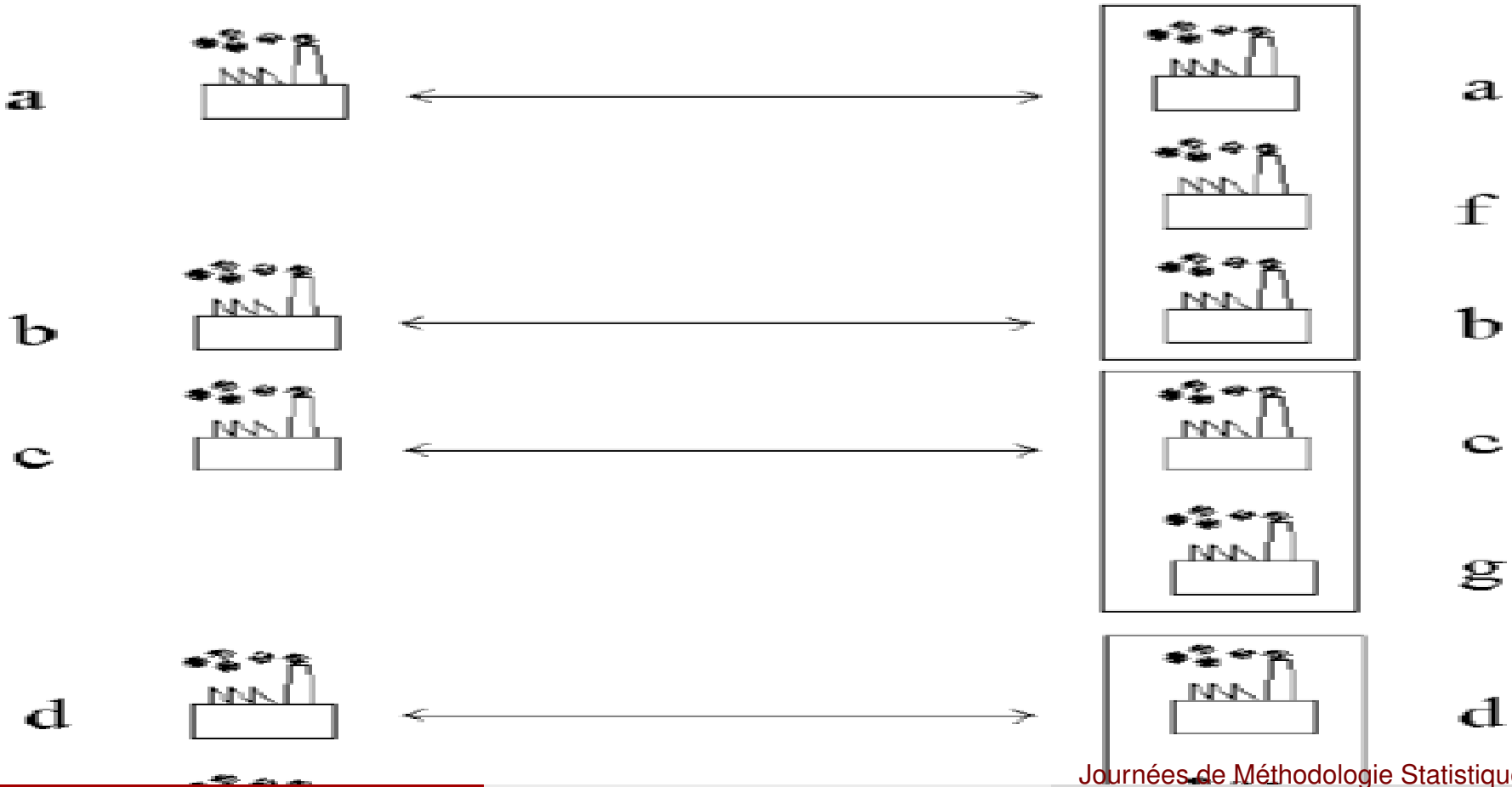
Sondages indirect et enquêtes auprès des entreprises : deux situations

Dans ces deux cas :

Une unité que l'on veut observer peut-être "atteinte" par plusieurs unités de la population enquêtée.

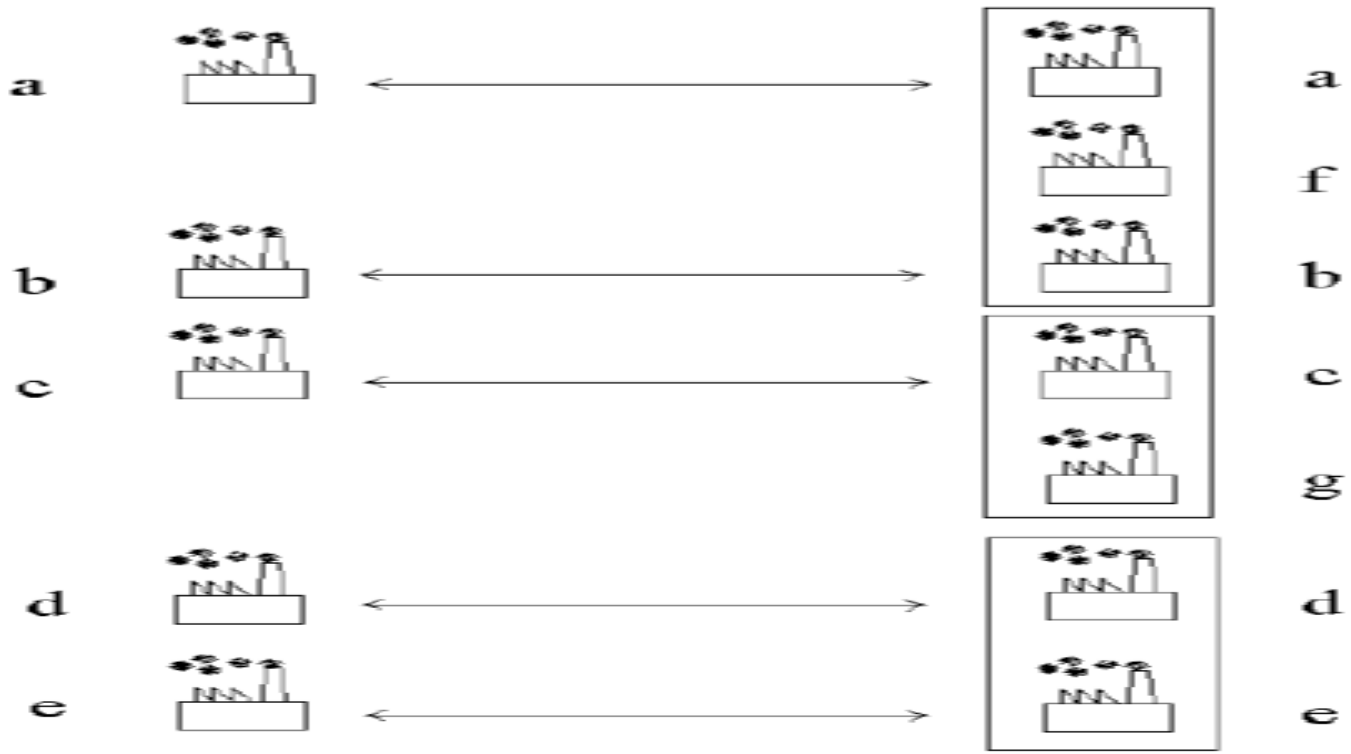
Cas concret I : Reconstitution du contour d'une entreprise (d'un groupe)

En matière de statistique d'entreprises, on connaît les "unités légales" mais on voudrait des résultats sur des "entreprises" qui sont toutes les unités légales liées entre elles.



Cas concret I : Reconstitution du contour d'une entreprise (d'un groupe)

Source $\mathbb{U}_A$: unités légales $\rightarrow$ Cible $\mathbb{U}_B$: entreprises (groupes)



Dans ce cas l'échantillon "d'entreprises" est liée à toutes les unités légales interrogées.

Cas concret II : Enquête Associations

Difficulté : pas de base de sondage "propre" car on utilise deux sources :

- le répertoire national des associations (RNA) géré par le ministère de l'Intérieur ;
- le répertoire des entreprises et des établissements (Sirene) géré par le ministère de l'Économie.

Base de sondage : On essaie de faire en sorte que chaque association ne soit présente qu'une seule fois...

En pratique : 3,5 % des associations interrogées sont présentes deux fois dans la base de sondage.

Problème : Les associations en doublon ont plus de chances d'être interrogées... => Biais ?

Solution :

- Demander à chaque unité enquêtée (au titre d'un répertoire) si elle est immatriculée dans l'autre et sous quel numéro.

Cas concret II : Enquête Associations

Population "cible" $\mathbb{U}_B$:Associations "réelles"

Population "source" $\mathbb{U}_A$:Immatriculations (une ou deux)

Plan

Comment pondérer les unités de s^B ?

Dans cette situation, la méthode généralisée du partage des poids est la solution traditionnelle. Les principaux arguments pour l'utiliser sont les suivants :

- les estimateurs obtenus sont sans biais
- les probabilités d'inclusion exactes sont en général trop complexes à calculer.

Comment pondérer les unités de s^B ?

Toutefois lors des enquêtes auprès des entreprises, les poids sont très dispersés. Des poids qui s'étendent de 1 à 60 sont fréquents.

Avec en général :

- des "petits poids" (proches voire égaux à l'unité) pour les grandes entreprises.
- Et des "grands poids" pour les petites entreprises.

La MGPP peut conduire à des résultats indésirables :

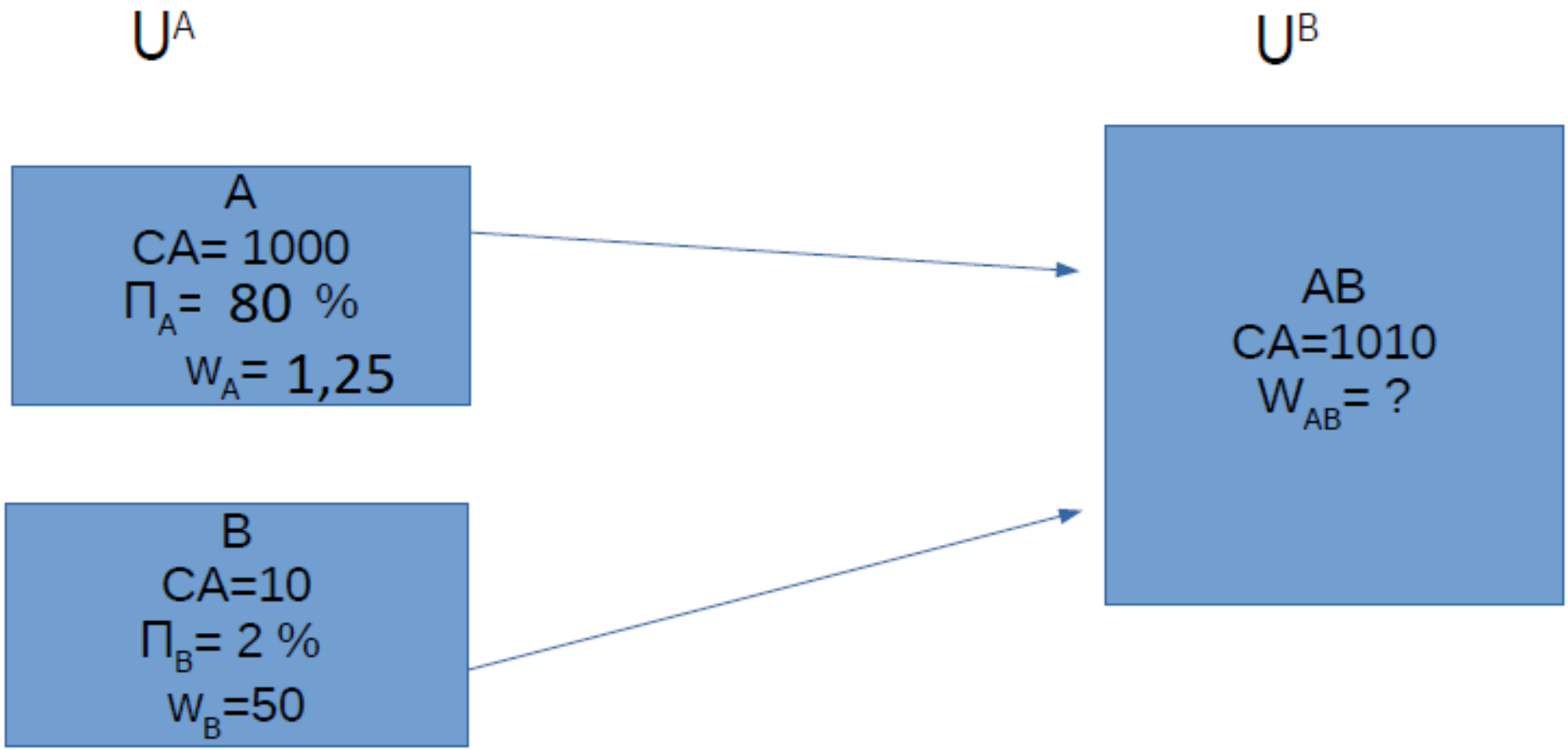
- poids élevés pour les grandes entreprises
- poids inférieurs à l'unité

Comment pondérer les unités de s^B ?

Ces inconvénients sont limités par l'emploi d'une version pondérée de la méthode de partage des poids.

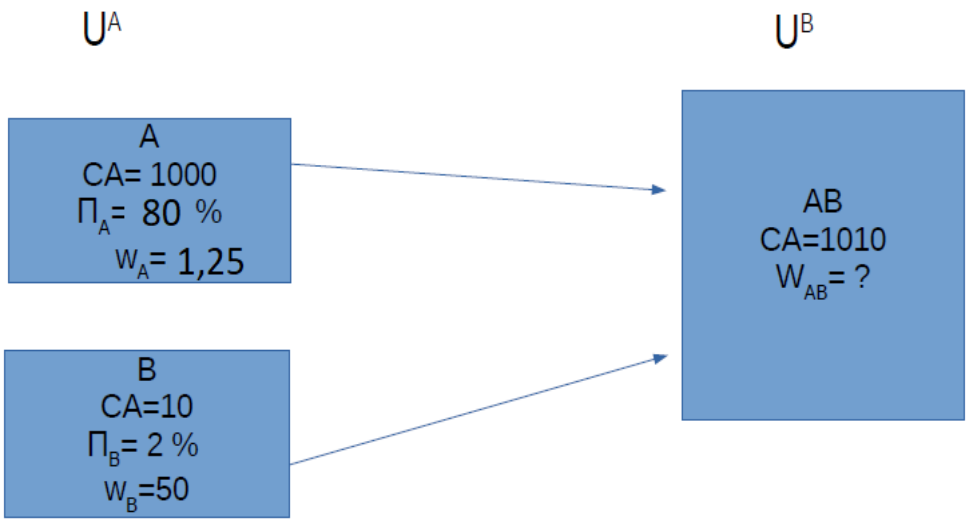
Toutefois, dans certains cas (Associations), il n'existe pas de variable de pondération.

Partage des poids : exemple entreprise



Comment pondérer *AB* ?

Exemple entreprises



Comment pondérer *AB* ?

Situation	MGPP	MGPP Pondérée	Probas exactes
A et $B \in s_A$	25,63	1,73	1,24
$A \in s_A$ et $B \notin s_A$	0,63	1,23	1,24
$A \notin s_A$ et $B \in s_A$	25	0,50	1,24

Comment pondérer les unités de s^B ?

Dans les enquêtes auprès des entreprises, les plans de sondages sont très simples : il s'agit plans stratifiés à un degré.

$\Longrightarrow$ Il est possible de calculer les probabilités exactes

Plan

Le calcul des probabilités exactes

Illustrons ce calcul par un cas simple : si deux unités sont liées.

C'est le cas de l'enquête Associations où les unités peuvent être présentes dans deux sources.

Le calcul des probabilités exactes

L'idée est de calculer simplement

$$\mathbb{P}\left(i \text{ ou } j \in s^A\right) = \mathbb{P}\left(i \in s^A\right) + \mathbb{P}\left(j \in s^A\right) - \mathbb{P}\left(i \text{ et } j \in s^A\right) \tag{1}$$

Le seul terme qui manque est : $\mathbb{P}\left(i \text{ et } j \in s^A\right) = \pi_{ij}$.

En fait, on le connaît la plupart du temps : c'est la probabilité d'inclusion d'ordre 2.

Le calcul des probabilités exactes

Pour les plans stratifiés, ce terme vaut :

- $\pi_{ij} = \dfrac{n_h(n_h-1)}{N_h(N_h-1)}$ si i et j sont dans la même strate h
- $\pi_{ij} = \dfrac{n_h}{N_h}\dfrac{n_k}{N_k}$ si elles sont dans deux strates h et k

Le calcul des probabilités exactes

Et on peut en déduire les probabilité d'inclusion de la "cible".

- si les unités sont dans la même strate

$$\mathbb{P}\left(i \text{ ou } j \in s^A\right) = \frac{n_h}{N_h}\left(2 - \frac{n_h - 1}{N_h - 1}\right)$$

- si les unités sont dans la deux strates différentes

$$\mathbb{P}\left(i \text{ ou } j \in s^A\right) = \frac{n_h}{N_h} + \frac{n_k}{N_k} - \frac{n_h}{N_h}\frac{n_k}{N_k}$$

Donc des calculs assez simples..

Le calcul des probabilités exactes

En particulier, dans le cas fréquent (type enquête Associations) où les unités sont dans deux strates différentes. si on note α l'unité "cible", w_i et w_j les poids de sondage des deux unités qui la représentent, on trouve :

$$w_\alpha = \frac{w_i w_j}{w_i + w_j - 1}$$

On peut donc calculer le poids "correct" de l'unité présente deux fois à partir des poids des deux unités qui la représentent.

Le calcul des probabilités exactes

Le poids $w_\alpha = \dfrac{w_i w_j}{w_i + w_j - 1}$ vérifie les propriétés suivantes :

- Il fournit un estimateur sans biais
- $w_\alpha \leq \min(w_i, w_j)$
- Si $w_i << w_j$, alors $w_\alpha \approx w_i$
- Si w_i ou $w_j = 1$, alors $w_\alpha = 1$ (Les unités exhaustives restent exhaustives..)
- $w_\alpha \geq 1$

Le calcul des probabilités exactes

Dans le cas où il y a plus d'unités liées, le calcul se complexifie bien sûr mais reste tout à fait faisable et repose sur les mêmes idées.

S'agissant d'un sondage stratifié, c'est un simple problème de dénombrement.

Il faut introduire pour chaque unité α de la population $\mathbb{U}_B$ le nombre $m_{\alpha,h}$ d'unités dans la strate h de la population $\mathbb{U}_A$ qui sont liées à α

$$\mathbb{P}\left(\alpha \text{ sélectionnée}\right) = 1 - \prod_{h \text{ tel que } m_{\alpha,h}>0} \prod_{l=0}^{m_{\alpha,h}-1} \frac{N_h - n_h - l}{N_h - l}$$

Le poids w_α est l'inverse de cette probabilité - son expression générale n'a pas d'intérêt.

Le calcul des probabilités exactes

Le poids w_α vérifie les propriétés suivantes :

- Il fournit un estimateur sans biais
- $w_\alpha \leq w_i$, pour tout i lié à α
- $w_\alpha \geq 1$
- Si pour un i lié à α, $w_i = 1$, alors $w_\alpha = 1$ (Les unités exhaustives restent exhaustives..)

Le calcul des probabilités exactes : probabilités d'inclusion doubles

Pour calculer la variance (ou pour l'estimer), il est nécessaire de disposer des probabilités d'inclusion d'ordre 2.

C'est-à-dire pour des unités α et β de $\mathbb{U}_B$, il faut déterminer $\pi_{\alpha\beta} = \mathbb{P}\left(\alpha \in s_B \text{ et } \beta \in s_B\right)$.

Le calcul des probabilités exactes : probabilités d'inclusion doubles

On peut écrire $\pi_{\alpha\beta}$ ainsi :

$$\pi_{\alpha\beta} := \mathbb{P}\left(\alpha \in s_B \text{ et } \beta \in s_B\right) = \mathbb{P}\left(\alpha \in s_B\right) + \mathbb{P}\left(\beta \in s_B\right) - \mathbb{P}\left(\alpha \in s_B \text{ ou } \beta \in s_B\right)$$

On connaît $\mathbb{P}\left(\alpha \in s_B\right)$ et $\mathbb{P}\left(\beta \in s_B\right)$.

Il suffit donc de déterminer $\mathbb{P}\left(\alpha \in s_B \text{ ou } \beta \in s_B\right)$

Le calcul des probabilités exactes : probabilités d'inclusion doubles

$\mathbb{P}(\alpha \in s_B \text{ ou } \beta \in s_B)$?

Imaginons :

- α liée à A,B et C
- β liée à D et E

On sélectionne α ou β

$\Longleftrightarrow$

On sélectionne γ "fictive' liée à A,B,C,D et E

Le calcul des probabilités exactes : probabilités d'inclusion doubles

Pour déterminer $\mathbb{P}\left(\alpha \in s_B \text{ ou } \beta \in s_B\right)$

- l'unité α est liée à une liste l_α d'unités de $\mathbb{U}_A$
- l'unité β à une liste l_β d'unités de $\mathbb{U}_A$

Il suffit de faire une observation très simple :
Sélectionner α ou β équivaut à sélectionner au moins une unité de la réunion des listes l_α et l_β.

Donc $\mathbb{P}\left(\alpha \in s_B \text{ ou } \beta \in s_B\right)$ se détermine avec la formule qui donne les probabilités d'inclusion d'ordre 1 appliquée à l'unité fictive liée à la réunion des deux listes.

Le calcul des probabilités exactes : probabilités d'inclusion doubles

On peut ensuite déterminer $\pi_{\alpha\beta}$ grâce à la relation :

$$\pi_{\alpha\beta} := \mathbb{P}\left(\alpha \in s_B \text{ et } \beta \in s_B\right) = \mathbb{P}\left(\alpha \in s_B\right) + \mathbb{P}\left(\beta \in s_B\right) - \mathbb{P}\left(\alpha \in s_B \text{ ou } \beta \in s_B\right)$$

L'expression générale ne présente pas d'intérêt.

Le calcul des probabilités exactes : probabilités d'inclusion doubles

Il suffit de :

$\Longrightarrow$ Programmer le calcul des probabilités d'inclusion simple pour un nombre quelconque d'unités

Le calcul des probabilités exactes : probabilités d'inclusion doubles

On peut également déterminer $\Delta^B_{\alpha,\beta} = \pi^B_{\alpha,\beta} - \pi^B_\alpha \pi^B_\beta$.
Le fait de disposer de ces probabilités d'inclusion doubles permet d'estimer sans biais la variance de l'estimateur avec :

$$\hat{\mathbb{V}}\left(\hat{Y}^B\right) = \sum_{\alpha,\beta \in s^B} \Delta^B_{\alpha,\beta} \frac{y_\alpha}{\pi^B_\alpha} \frac{y_\beta}{\pi^B_\beta} \frac{1}{\pi^B_{\alpha,\beta}}$$

Plan

Simulations : Génération des données ($\mathbb{U}_A$, $\mathbb{U}_B$, CA_j, $L_{i,j}$, VA_i)

On a simulé une population de données "type entreprise" et simulé l'existence d'un lien entre unités légales. L'objectif était d'estimer la valeur ajoutée de la population liée.

Simulations : Génération des données ($\mathbb{U}_A$, $\mathbb{U}_B$, CA_j, $L_{i,j}$, VA_i)

Table 1 – Paramètres de simulations

Strate	N^A	n^A	**Loi** CA_j
1	1000	30	$N(100, 50)$
2	500	50	$N(1000, 200)$
3	100	100	$N(2000, 500)$

Résultats

Table 2 – CV (en %) des différents estimateurs de la valeur ajoutée - données simulées

	MGPP classique		**Probabilité**
type d'estimateur	**standard**		**exacte**
Application directe des pondérations	8.3		5.0

Résultats

Table 3 – CV (en %) des différents estimateurs de la valeur ajoutée - données simulées

		MGPP classique	MGPP pondérée
Probabilités exactes			
Application directe des pondérations	8.3	4.3	5.0

Autre simulation : avec des taux de sondage homogènes

Mêmes simulations avec

- taux de sondage 1/5 dans les 3 strates
- $CA_j \rightsquigarrow \mathcal{N}(100, 50)$ dans les trois strates

CV (en %) des différents estimateurs de la valeur ajoutée - données simulées - taux de sondage homogènes

	MGPP classique	**MGPP pondérée**	**Probabilités exactes**
Application directe des pondérations	2.2	3.1	3.0

Plan

Conclusion

Des résultats très encourageants

- Estimateurs plus précis dans le cadre de la stat. d'entreprises (pop. asymétriques)
- Poids final majoré par le plus petits des poids des unités en cause
- Poids supérieur à l'unité

Si les poids de sondage sont homogènes : la MGPP fonctionne mieux.

Suite

- Confirmer ces résultats sur données réelles
- Mieux comprendre ces résultats : à partir de quand est-ce que ça marche ?
- Quid de la non-réponse ?

Conclusion

Le code R des fonctions permettant de calculer les poids :

- MGGP
- MGGP pondérée par une variable auxiliaire
- probabilités exactes

Est disponible sur Github (package tbis) dans le dépôt suivant :
https ://github.com/arnaudfi/tbis/tree/master

- Pierre Lavallée. *Indirect sampling* Springer Series in Statistics, 2007.
- Pierre Lavallée et Sébastien Labelle-Blanchet. *Le sondage indirect appliqué aux populations asymétriques* Techniques d'enquête, Vol. 39, No 1, pp. 207-241, juin 2013.
- Arnaud Fizzala. *La gestion par partage des poids des changements de contour des entreprises dans l'Enquête Sectorielle Annuelle*. Acte des Journées de Méthodologie Statistique de l'Insee 2018.
- Ronan Le Gleut et Thomas Merly-Alpa. *L'impact du profilage sur la refonte du plan de sondage des Enquêtes Sectorielles Annuelles*. Acte des Journées de Méthodologie Statistique de l'Insee 2018.
- Camilia Coga *Cours de sondages dispensé à l'université de Besançon*. http ://goga.perso.math.cnrs.fr/
- Pascal Ardilly *Les techniques de Sondage*. Editions TECHNIP, 2006.
- *Les entreprises en France* INSEE Références, Edition 2019.

Origine des illustrations

Diapositive 4 :Lucie Léon, Prise en compte de la fréquentation multiple des lieux d'enquête (www.epiter.org)

Diapositives 9, 10 : extrait du livre de Pierre Lavallée, Indirect Sampling