Représentativité et méthodes de pondération

Thomas Merly-Alpa
thomas.merly-alpa@ined.fr

Séminaire de méthodologie de Sciences Po (MetSem) - 04/04/24

Présentation

Responsable du Service des Enquêtes et Sondages de l'Ined `https://ses.site.ined.fr/` :

- Service d'appui à la recherche
- Vingtaine d'ingénieur.e.s
- Conception : questionnaires, protocole, plan de sondage
- Réalisation : logistique, information, suivi, marchés, formation
- Diffusion : apurement, documentation, mise à disposition
- Valorisation des travaux

Auparavant (2015/2019) : responsable de la section Échantillonnage au Département Méthodes Statistiques de l'Insee

Sommaire I

Chapitre 1

Pondérations

Concept

Qu'est-ce que l'échantillonnage / l'estimation par sondage ?

- Une population de grande taille
- Compter ou interroger est coûteux
- On sélectionne quelques individus qui répondent "pour tout le monde"

Idée cruciale : sélectionner **aléatoirement** ces individus.

Échantillon représentatif

Un "échantillon représentatif" :

- On entend souvent cette formule

Échantillon représentatif

Un "échantillon représentatif" :

- On entend souvent cette formule
- Quel est son sens ? "Village" de 100 habitants

Échantillon représentatif

Un "échantillon représentatif" :

- On entend souvent cette formule
- Quel est son sens ? "Village" de 100 habitants
- Souvent associé à la méthodes des quotas.

Méthodes des quotas

La méthode des quotas est une méthode de sondage non probabiliste souvent utilisée par les instituts de sondage :

- On fixe le quota que l'on souhaite atteindre pour chaque catégorie : femmes ; cadres ; habitants en Bretagne . . .

Méthodes des quotas

La méthode des quotas est une méthode de sondage non probabiliste souvent utilisée par les instituts de sondage :

- On fixe le quota que l'on souhaite atteindre pour chaque catégorie : femmes ; cadres ; habitants en Bretagne . . .
- On appelle de nombreuses personnes : une personne souhaitant participer ne peut le faire que si elle rentre dans le quota.

Méthodes des quotas

La méthode des quotas est une méthode de sondage non probabiliste souvent utilisée par les instituts de sondage :

- On fixe le quota que l'on souhaite atteindre pour chaque catégorie : femmes ; cadres ; habitants en Bretagne . . .
- On appelle de nombreuses personnes : une personne souhaitant participer ne peut le faire que si elle rentre dans le quota.
- On obtient à la fin un échantillon avec la structure souhaitée.

Méthodes des quotas

Cela pose de nombreux problèmes :

- Les premiers à répondre ont plus de chance d'être dans les quotas ;
- Comment trouver les derniers répondants ? Il faut une femme de 15 à 30 ans à Paris qui soit agricultrice
- Quid de ceux qui n'ont jamais répondu ?
- Pire : l'utilisation des *access panel*, c'est à dire des panélistes auto-recrutés en ligne, perturbe encore plus ces étapes.

$\Rightarrow$ *La singulière fabrique des sondages d'opinion*, M. Lejeune.

L'estimation naïve

Pour l'estimation du total et de la moyenne d'une variable Y, l'estimateur « naïf » est :

- Pour le total, la somme des valeurs Y des individus de l'échantillon.
- Pour la moyenne, la moyenne des valeurs Y des individus de l'échantillon.

En général, l'estimation naïve est fausse (*biaisée*), surtout quand l'échantillon est choisi de façon complexe.

L'estimateur naïf

Rappel : pour l'estimation du total et de la moyenne d'une variable Y, l'estimateur ≪ naïf ≫ s'écrit :

$$\hat{T}(Y)_{naif} = \sum_{k \in s} y_k$$

$$\hat{\bar{y}}_{naif} = \frac{1}{n} \sum_{k \in s} y_k$$

L'estimateur naïf

En général, l'estimation naïve est biaisée :

$$\mathbb{E}(\hat{\Phi}_{naif}) = \sum_s p(s) \cdot \hat{\Phi}(s)$$
$$\neq \Phi$$

$\mathbb{E}(\hat{\Phi})$ est la valeur moyenne de $\hat{\Phi}$ obtenue avec le plan de sondage considéré **sur tous les échantillons possibles**.

Plan de sondage - définition

On note $\mathcal{S}$ l'ensemble des parties de $\mathcal{U}$.
Le plan de sondage p est une loi de probabilité sur $\mathcal{S}$ telle que :

$$\forall s \in \mathcal{S},\ p(s) \geq 0$$

$$\sum_{s \in \mathcal{S}} p(s) = 1$$

Plan de sondage - exemple

Soit $\mathcal{U} = \{1, 2, 3\}$. On a alors :
$\mathcal{S} =$

Plan de sondage - exemple

Soit $\mathcal{U} = \{1, 2, 3\}$. On a alors :
$\mathcal{S} = \{\{1\}, \{2\}, \{3\}, \{1, 2\}, \{1, 3\}, \{2, 3\}, \{1, 2, 3\}\}$

Plan de sondage - exemple

Soit $\mathcal{U} = \{1, 2, 3\}$. On a alors :
$\mathcal{S} = \{\{1\}, \{2\}, \{3\}, \{1, 2\}, \{1, 3\}, \{2, 3\}, \{1, 2, 3\}\}$

On peut définir un plan de sondage p par :

$$p(\{1\}) = 0 \quad p(\{1, 2\}) = \frac{1}{2} \quad p(\{1, 2, 3\}) = 0$$

$$p(\{2\}) = 0 \quad p(\{1, 3\}) = \frac{1}{3}$$

$$p(\{3\}) = 0 \quad p(\{2, 3\}) = \frac{1}{6}$$

Pondérer ?

Pour éviter d'utiliser l'estimateur naïf, on utilise généralement ce qu'on appelle des poids, qu'on note w (pour *weight* en anglais).

Le poids d'un individu correspond au nombre d'individus que l'individu de l'échantillon représente dans la population. Si l'on interroge 1 individu sur 100, le poids est alors de 100.

L'estimateur pondéré du total est alors la somme des $w_i y_i$ sur l'échantillon.

Probabilité d'inclusion π_k

Pour améliorer l'estimateur naïf, il faut utiliser une pondération. On va calculer celle-ci à l'aide des **probabilités d'inclusion**.

La probabilité d'inclusion simple d'un individu k est la probabilité que cet individu soit dans l'échantillon. Ainsi, pour $k \in \mathcal{U}$,

$$\pi_k = \mathbb{P}(k \in s) = \mathbb{P}(\delta_k = 1) = \sum_{s \ni k} p(s)$$

où δ_k est l'indicatrice d'appartenance de k à $\mathcal{S}$, appelée aussi variable de Cornfield.

Poids dans quelques exemples simples

Dans un sondage aléatoire simple, le poids de tous les individus est le même :

$$w_k = \frac{N}{n}$$

Dans un plan de sondage stratifié, le poids varie selon les strates.
Dans un sondage aléatoire simple stratifié, nous avons :

$$\forall k \in S_h \ , \ w_k = \frac{N_h}{n_h}$$

Estimateur de Horvitz-Thompson

Définition

L'estimateur d'Horvitz-Thompson (ou π-estimateur) est défini :

$$\textit{pour un total : } \hat{T}_{y\pi} = \sum_{k \in s} \frac{y_k}{\pi_k}$$

$$\textit{pour une moyenne : } \hat{\bar{y}}_{\pi} = \frac{1}{N} \sum_{k \in s} \frac{y_k}{\pi_k}$$

C'est donc un **estimateur pondéré** *utilisant les poids* $w_k = \frac{1}{\pi_k}$

Estimation sans biais

Théorème

Si $\forall k \in \mathcal{U}, \pi_k > 0$, *alors l'estimateur d'Horvitz-Thompson est* **sans biais** *pour le total et la moyenne.*

La condition signifie que toutes les unités de la population ont une chance non nulle d'être dans l'échantillon.

Estimation sans biais

Démonstration.

$$\begin{aligned}
\mathbb{E}[\hat{T}_{y\pi}] &= \mathbb{E}\left[\sum_{k\in s} \frac{y_k}{\pi_k}\right] \\
&= \mathbb{E}\left[\sum_{k\in \mathcal{U}} \frac{y_k \delta_k}{\pi_k}\right] \\
&= \sum_{k\in \mathcal{U}} \frac{y_k \mathbb{E}[\delta_k]}{\pi_k} \\
&= \sum_{k\in \mathcal{U}} y_k \\
&= T(y)
\end{aligned}$$

Biais

Espérance :

$$\mathbb{E}(\hat{\Phi}) = \sum_{s} p(s) \cdot \hat{\Phi}(s)$$

C'est la valeur moyenne de $\hat{\Phi}$ obtenue avec le plan de sondage considéré **sur tous les échantillons possibles**.

Biais

Biais :

$$B(\hat{\Phi}) = \mathbb{E}(\hat{\Phi}) - \Phi$$

Si $B(\hat{\Phi}) = 0$, alors on parle **d'estimateur sans biais**.

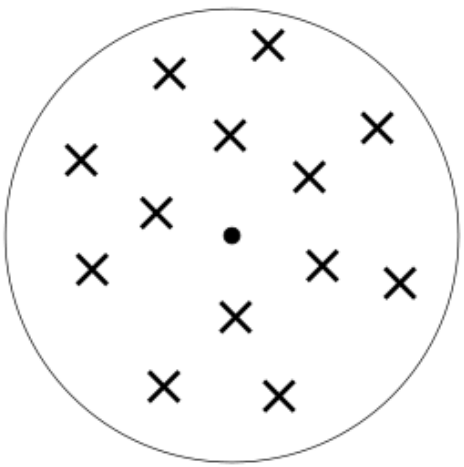
Variance / Précision

Il reste de l'incertitude - variance :

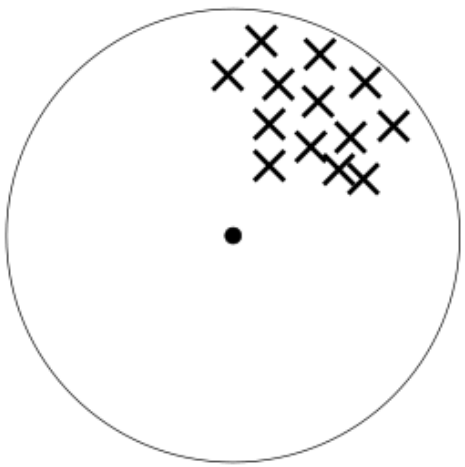
$$\mathrm{Var}(\hat{\Phi}) = \sum_{\mathrm{s}} \mathrm{p(s)} \cdot \left[\mathbb{E}(\hat{\Phi}) - \hat{\Phi}(\mathrm{s})\right]^2$$

C'est une mesure de la dispersion des valeurs $\hat{\Phi}(s)$ autour de leur moyenne.

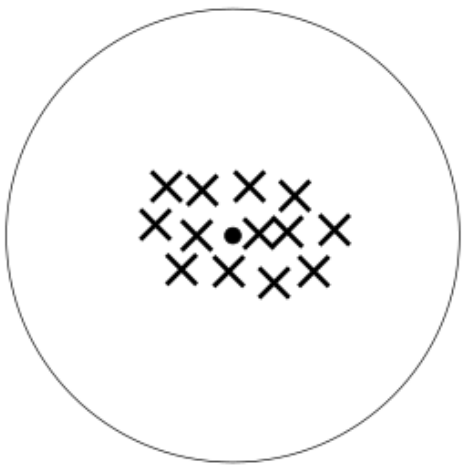
Schéma



Cas 1



Cas 2



Cas 3

Construction d'un intervalle de confiance

Comment relier à l'incertitude autour d'un résultat ?
Quantités liées :

$$\sigma(\hat{\Phi}) = \sqrt{\mathrm{Var}(\hat{\Phi})}, \text{écart-type}$$

Construction d'un intervalle de confiance

On fait **l'hypothèse** : $\hat{\Phi}(s) \sim \mathcal{N}(\Phi, \mathrm{Var}(\Phi))$

L'intervalle de confiance à 95% est défini par :

$$IC_{95\%} = \left[\hat{\Phi} - 2\sigma(\hat{\Phi}); \hat{\Phi} + 2\sigma(\hat{\Phi})\right]$$

Formules de variance

Pour les plans de sondages usuels :

$$\hat{\mathrm{Var}}(\bar{y}) = (1-f)\frac{s^2}{n}$$

$$\hat{V}(\hat{T}_{SAS-str}(Y)) = \sum_{h=1}^{H} N_h^2 (1-f_h)\frac{s_h^2}{n_h}$$

Cas particulier :

$$\hat{\mathrm{Var}}(\bar{y}_{SAS-str-prop}) \approx (1-f)\frac{s_{intra}^2}{n}$$

Chapitre 2

Non-réponse et repondération

Partie 1

Définition

Définition

Non répondants : on aurait souhaité que ces individus nous donnent des informations car ils nous intéressent dans le cadre de l'enquête, mais ce n'est pas le cas, pour différentes raisons.

On peut distinguer deux types de non-réponse :

- Partielle : il manque une ou plusieurs réponses, mais pas toutes.
- Totale : il manque toutes ou quasiment toutes les réponses.

Ignorabilité

On caractérise la non-réponse :

- Ignorable : les répondants et les non-répondants diffèrent en structure mais leur comportement de réponse est le même conditionnellement à leur âge, sexe...(dit *Missing at Random*, MAR)
- Non ignorable : les non-répondants sont différents des répondants ; cette différence explique pourquoi ils n'ont pas répondu. (dit Non MAR, NMAR)

Le type de non-réponse vaut pour une variable Y et un paramètre précis (moyenne, ...).

Ignorabilité - Exemples

Exemples :

- Ignorable : les jeunes répondent moins souvent à l'enquête que les personnes âgées.
- Non ignorable : on pose la question : "faites-vous confiance aux institutions de votre pays ?" $\rightarrow$ ne pas répondre à une enquête officielle semble un bon indicateur que non...

Conséquences

Que penser des conséquences ?

- *Baisse de la taille de l'échantillon exploitable :* correspond à une perte en précision ; on peut anticiper.
- *Différences de structure entre échantillon et répondants :* risque de biais, corrigeable.
- *Différences entre le comportement des répondants et des non-répondants :* risque de biais, corrigeable ?

Pistes de solution

Il existe deux visions de ce problème :

- Les poids de sondage ne sont plus égaux au nombre d'unités de la population que chaque répondant représente : il faut ajuster les poids pour que les poids des répondants absorbent les poids des non-répondants $\Rightarrow$ **Méthodes de repondération**.
- L'information dont nous avons besoin est manquante pour certaines observations de l'échantillon : nous devons boucher les trous $\Rightarrow$ **Méthodes d'imputation**.

Imputation

L'imputation consiste à remplacer une donnée manquante par une donnée « plausible » déduite ou calculée en fonction des renseignements obtenus pour l'unité défaillante et/ou pour les unités qui lui sont proches.
Les méthodes d'imputation ont pour but :

1. de réduire le biais de non-réponse ;
2. de produire des tableaux de données « rectangulaires » sans trous.

Partie 2

Repondération

La non-réponse comme phase supplémentaire du plan de sondage 1/3

Plan de sondage en deux phases :

1. PREMIÈRE PHASE : le plan de sondage initial détermine les probabilités d'inclusion π_i et les poids de sondage $1/\pi_i$
2. SECONDE PHASE : sélection aléatoire des répondants selon un plan de sondage inconnu $\Pi_{/\mathcal{S}}$ qui détermine des probabilités d'inclusion ρ_i

La non-réponse comme phase supplémentaire du plan de sondage 2/3

Si les probabilités d'inclusion de seconde phase étaient connues :

- l'estimateur HT $\hat{Y}_{HT} = \sum_{i \in \mathcal{R}} \frac{y_i}{\pi_i}$ est biaisé
- il serait remplacé par l'estimateur sans biais selon le plan de sondage $\hat{Y}_{NR} = \sum_{i \in \mathcal{R}} \frac{y_i}{\pi_i \rho_i}$

Cet estimateur est appelé **estimateur corrigé de la non-réponse** ou **ajusté de la propension à répondre** *propensity score adjusted estimator*

La non-réponse comme phase supplémentaire du plan de sondage 3/3

$\Rightarrow$ objectif de la repondération : estimer ρ_i, parfois appélés score de propension à répondre ou simplement score de propension (*propensity score*)

ρ_i sont inconnus $\Rightarrow$ ils doivent être remplacés par des estimateurs $\hat{\rho}_i$ (mais comment spécifier le modèle ?)

Les variables auxiliaires

Rappels : les variables auxiliaires sont de trois types

1. Variable **connue pour toutes les unités de la population** (*e.g.* : variables de la base de sondage)
2. Variable **connue pour toutes les unités de l'échantillon**, répondantes ou non (*e.g.* : paradonnées, réponse au questionnaire de première phase dans une enquête en deux phases)
3. Variable **connue pour les répondants uniquement dont le total est connu par une autre source**

Que faire ?

Cas usuel : nous avons de l'information auxiliaire conditionnellement à laquelle la réponse est MAR

- que faire quand les variables auxiliaires sont de type 1 ou 2, *i.e.* disponibles pour les répondants et les non-répondants ? $\Rightarrow$ **Groupes de réponse homogènes** et **Calage sur marges**
- que faire quand les variables auxiliaires sont de type 3 uniquement : disponibles pour les répondants mais dont le total est connu ? $\Rightarrow$ **Calage sur marges** uniquement

Partie 3

Groupes de réponse homogènes

Introduction

Idée de la méthode des GRH :

- quand le comportement de réponse est indépendant des caractéristiques des unités et que toutes les mêmes unités ont la même probabilité de réponse, on sait corriger facilement la non-réponse
- malheureusement, les comportements de réponse sont différents suivant les caractéristiques des unités de l'échantillon
- $\Rightarrow$ **se ramener au cas que l'on sait traiter** en cherchant à diviser la population en sous-parties à l'intérieur desquelles on pourra supposer que les comportements de réponse sont indépendants des caractéristiques des unités et corriger la non-réponse

Principe

Groupes de Réponse Homogène (GRH) = *Weighting classes* = *Response Homogeneous Groups (RHG)* = *Adjustment cells*

Hypothèses :

- l'échantillon peut être divisé en un ensemble de classes dans lesquelles toutes les unités ont la même propension à répondre
- $\Leftrightarrow$ la sélection des répondants (plan de sondage de deuxième phase) se fait selon un plan de sondage stratifié avec sondage de Bernoulli dans chaque strate
- l'ajustement pour non-réponse dans un GRH $\frac{1}{\hat{\pi}_{k/\mathcal{S}}}$ est égal à l'inverse de l'estimateur de la propension à répondre commune à toutes les unités du GRH

Analyse

- Les GRH doivent être constitués de façon à minimiser la covariance entre la propension à répondre et les réponses à l'enquête
- Une manière d'y arriver : les GRH doivent être des classes dans lesquelles les propensions à répondre et les réponses à l'enquête sont très homogènes (aussi proches que possible)
- Ainsi, les variables auxiliaires pertinentes pour constituer des GRH sont les variables **corrélées au comportement de réponse et aux variables d'intérêt de l'enquête**

Contraintes

Les GRH doivent respecter deux types de contraintes :

1. TAUX DE RÉPONSE MINIMAL : pour éviter les ajustements pour non-réponse trop importants
2. TAILLE MINIMALE : pour garantir que le taux de réponse commun soit estimé de manière assez robuste, les GRH doivent contenir un nombre minimal d'unités (en général 50)

Estimer la propension à répondre dans les GRH

Deux formules :

1. ESTIMATEUR STANDARD : (nombre de répondants dans le GRH) / (nombre de répondants + nombre de non-répondants)
2. ESTIMATEUR PONDÉRÉ : (somme des poids des répondants dans le GRH) / (somme des poids des répondants et des non-répondants)

Comment construire des GRH : quelles méthodes ?

Utilisation de l'information auxiliaire disponible sur les répondants et les non-répondants. Deux types de méthodes :

1. MÉTHODES PAR PARTITION : utilisation directe des variables auxiliaires pour construire les GRH
   - méthodes par croisement (*cross-classification method*)
   - méthodes d'arbres de classification
2. MÉTHODES DES SCORES : les variables auxiliaires sont d'abord résumées dans un estimateur du score de propension à répondre, qui est ensuite utilisé pour construire des GRH
   - Méthode des quantiles
   - Classification Ascendante Hiérarchique
   - Méthode d'Haziza et Beaumont

Partie 4

Calage

Exemple introductif

On cherche à estimer le nombre d'habitants d'une région comportant $N = 2\,536$ villages. On tire un échantillon de $n = 127$ villages par sondage aléatoire simple. On observe une taille moyenne de $\bar{y} = 377.2$ habitants sur l'échantillon.

Pour chacun des 2 536 villages, on connaît la population au dernier recensement, organisé trois ans auparavant.

Exemple introductif

Taille moyenne	Ensemble des villages de la région	Échantillon de villages
Au moment de l'enquête	?	377,2
Au moment du recensement	345,1	341,7

(extrait de *Manuel de sondages, Applications aux pays en développement*, R. Clairin et Ph. Brion, Documents et Manuels du CEPED numéro 3)

Principe

Au moment de l'enquête, on recueille deux types d'informations :

1. sur la variable d'intérêt Y
2. sur une variable auxiliaire X dont le total sur la population est connu.

Principe

Les estimateurs d'Horvitz-Thompson pour les totaux de X et Y sont :

$$\hat{T}_{Y\pi} = \sum_{k \in s} \frac{y_k}{\pi_k} = \sum_{k \in s} d_k y_k$$

$$\hat{T}_{X\pi} = \sum_{k \in s} \frac{x_k}{\pi_k} = \sum_{k \in s} d_k x_k$$

$$\text{où : } d_k = \frac{1}{\pi_k} = \text{poids de sondage}$$

Principe

En général, $\hat{T}_{X\pi}$ est différent du vrai total connu $T(X)$. On se sert de la connaissance de $T(X)$ pour modifier l'estimateur de $T(Y)$:

- On suppose que la variable Y est, même approximativement, proportionnelle à la variable X : $y_k \approx R \cdot x_k$
- On a donc également : $T(Y) \approx R \cdot T(X)$
- on utilise l'échantillon pour estimer R : $\hat{R} = \dfrac{\hat{T}_{Y\pi}}{\hat{T}_{X\pi}}$
- On estime $T(Y)$ par $\hat{T}_{Y,ratio} = \hat{R} \cdot T(X)$

Principe

La technique du calage sur marges généralise les méthodes de redressements. En effet, toutes ces méthodes peuvent êtres vues comme des cas particulier de calage sur marges.
Les méthodes de calage consistent à repondérer les unités de l'échantillon, i.e. à modifier les poids d'échantillonnage, de telle façon que les estimations :

- de totaux de variables numériques coïncident avec les vrais totaux connus, par une information externe, sur la population
- d'effectifs des modalités de variables catégorielles coïncident avec les vrais effectifs connus, par une information externe, sur $\mathcal{U}$.

Principe

Le calage vise à répondre à deux objectifs :

1. respecter la structure de la population $\rightarrow$ les résultats obtenus via l'enquête concordent avec ce qui est attendu : **objectif de diffusion**
2. améliorer la qualité des estimations $\rightarrow$ caler sur la structure des X réduit la volatilité de l'estimateur de Y : **objectif de précision**

Principe

Les macro CALMAR (CALage sur MARges) et CALMAR 2 sous SAS et le package ICARUS sous R permettent de mettre en œuvre ces méthodes proposées par J.-C. Deville et C.-E. Särndal (1992-1993).

Information auxiliaire

J variables auxiliaires $X_1, \ldots, X_j, \ldots, X_J$ connues sur s, et dont on connaît les totaux sur la population $T_{X_j} = \sum_{k \in \mathcal{U}} x_{jk}$

Si l'information auxiliaire est relative à des variables catégorielles, cela signifie que l'on connaît les effectifs des modalités de ces variables, i.e. les totaux des variables indicatrices associées à ces modalités.

Objectif - Contraintes

Tenir compte de cette information pour améliorer l'estimateur de Horvitz-Thompson. On va chercher un nouvel estimateur de $T(Y)$:

$$\hat{T}_{Y,w} = \sum_{k \in s} w_k y_k$$

et où les nouveaux poids w_k :

- Sont "proches" des poids d_k
- vérifient les **équations de calage** :
  $\forall j \in [[1, J]], \sum_{k \in s} w_k x_{jk} = T_{X_j}$

Résolution théorique

On choisit une fonction G telle que $G\left(\frac{w_k}{d_k}\right)$ mesure la "distance" entre les nouveaux poids w_k et le poids initial (Horvitz-Thompson) d_k.
Conditions sur G (pseudo-distance) :

- $G(1) = 0$
- G positive et convexe. $G\left(\frac{w_k}{d_k}\right)$ est d'autant plus élevé que $\frac{w_k}{d_k}$ est éloigné de 1.

Résolution théorique

Les poids w_k sont solution du problème d'optimisation :

$$\begin{cases} \min\limits_{w_k} \sum\limits_{k \in s} d_k G\left(\dfrac{w_k}{d_k}\right) \\ \text{sous contrainte : } \sum\limits_{k \in s} w_k x_k = T_X \end{cases}$$

La méthode linéaire

La méthode linéaire :

- Est la plus rapide : converge en deux itérations
- Peut conduire à des poids w_k négatifs
- Poids non bornés

La méthode du raking ratio

La méthode du raking ratio :

- Poids toujours positifs
- Poids non bornés supérieurement, borne supérieure en générale plus élevée que pour la méthode linéaire

Méthodes logit et linéaire tronquée

Les méthodes logit et linéaire tronquée :

- Permettent de définir une borne inférieure L et une borne supérieure U pour $g_k = \frac{w_k}{d_k}$. Toutefois, toutes les valeurs de L et U ne sont pas possibles : il existe une valeur maximale pour L_{max} pour L et une valeur minimale U_{min} pour U.
- La détermination de L_{max} et U_{min} se fait en général par approximations successives.

En pratique

Il n'y a pas de critère purement statistique sans ambiguïté pour choisir l'une des méthodes de calage.

En pratique, il faut prendre en compte les nécessités de la diffusion : attention par exemple aux poids négatifs ou inférieurs à 1.

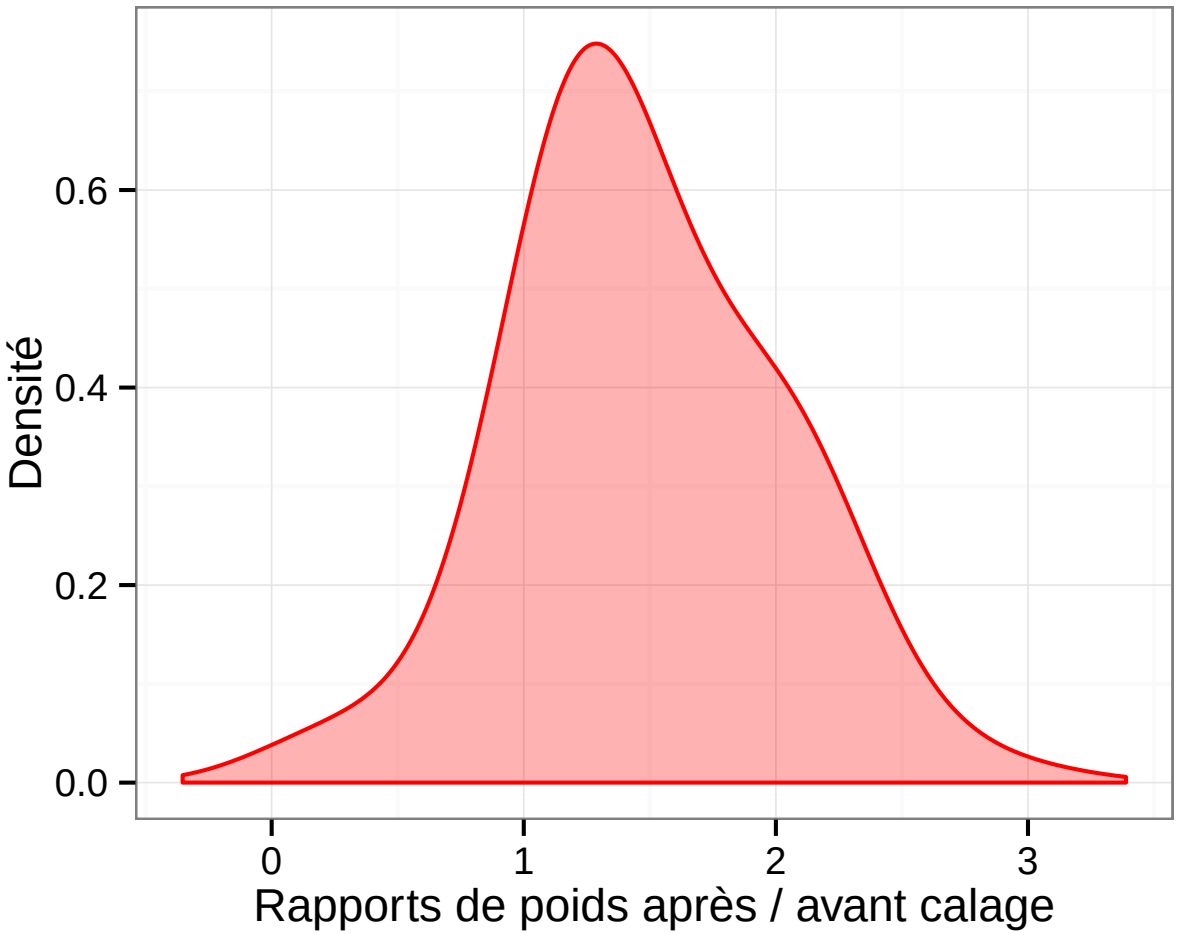
Pour des questions de robustesse, on souhaite en général limiter la dispersion des poids, et donc on privilégie les méthodes bornées. Toutefois, des bornes trop strictes conduisent à des accumulations de rapport de poids sur ces bornes.

En pratique

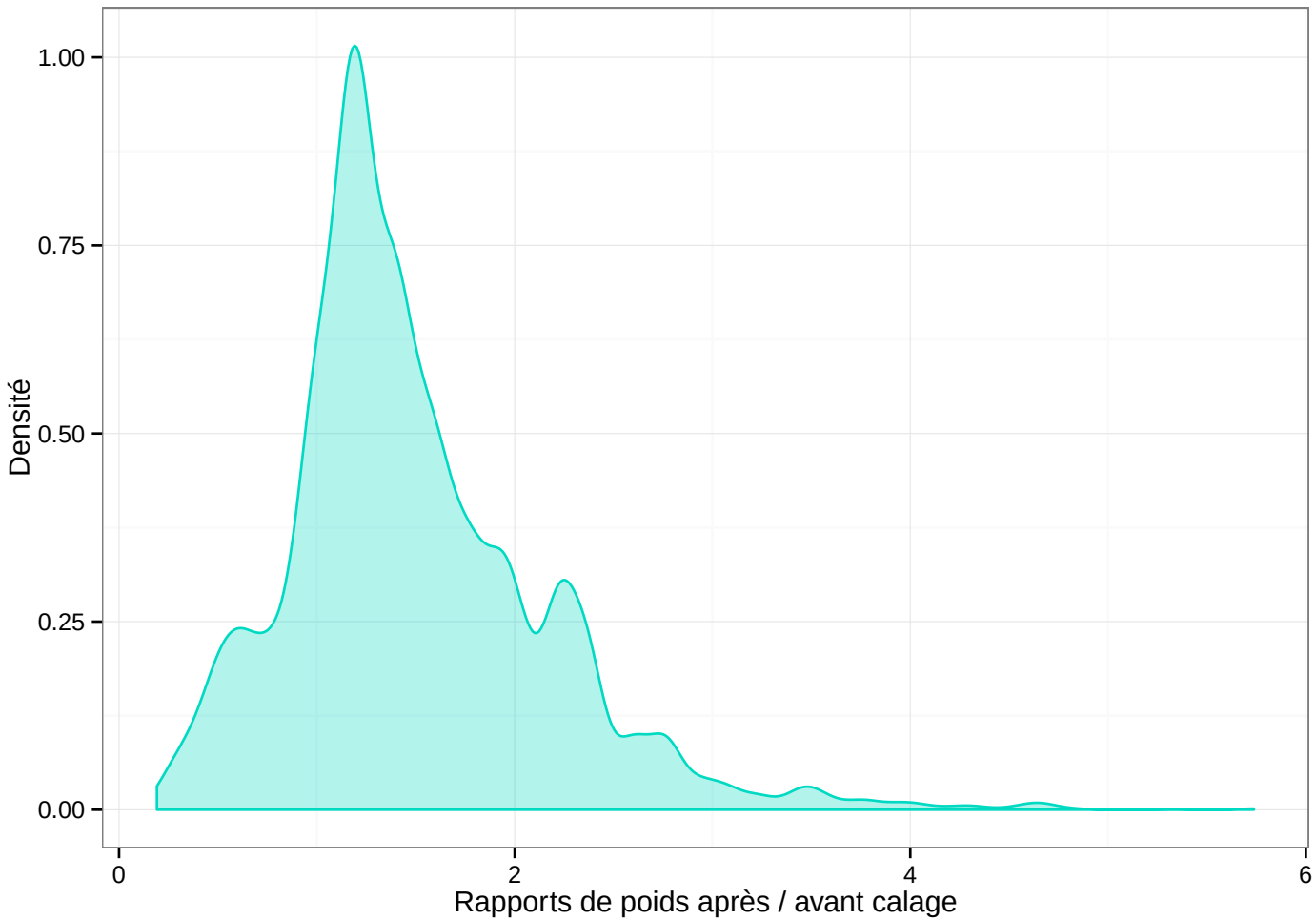
Le choix de la méthode s'effectue de manière empirique, en s'appuyant sur des critères concernant les rapports de poids tels que :

- La plus faible dispersion
- La plus faible étendue
- L'allure générale de la distribution

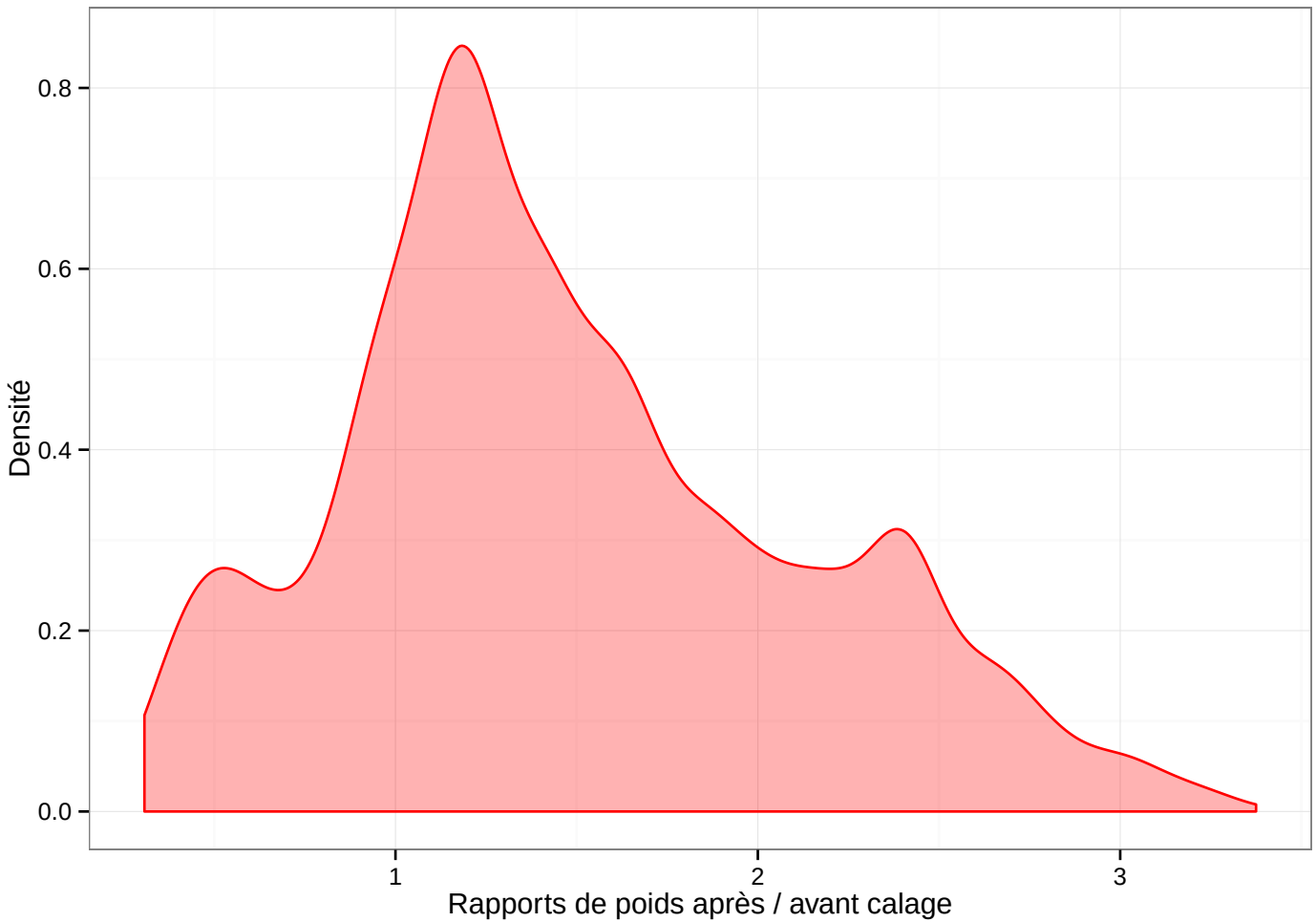
Distribution - méthode linéaire



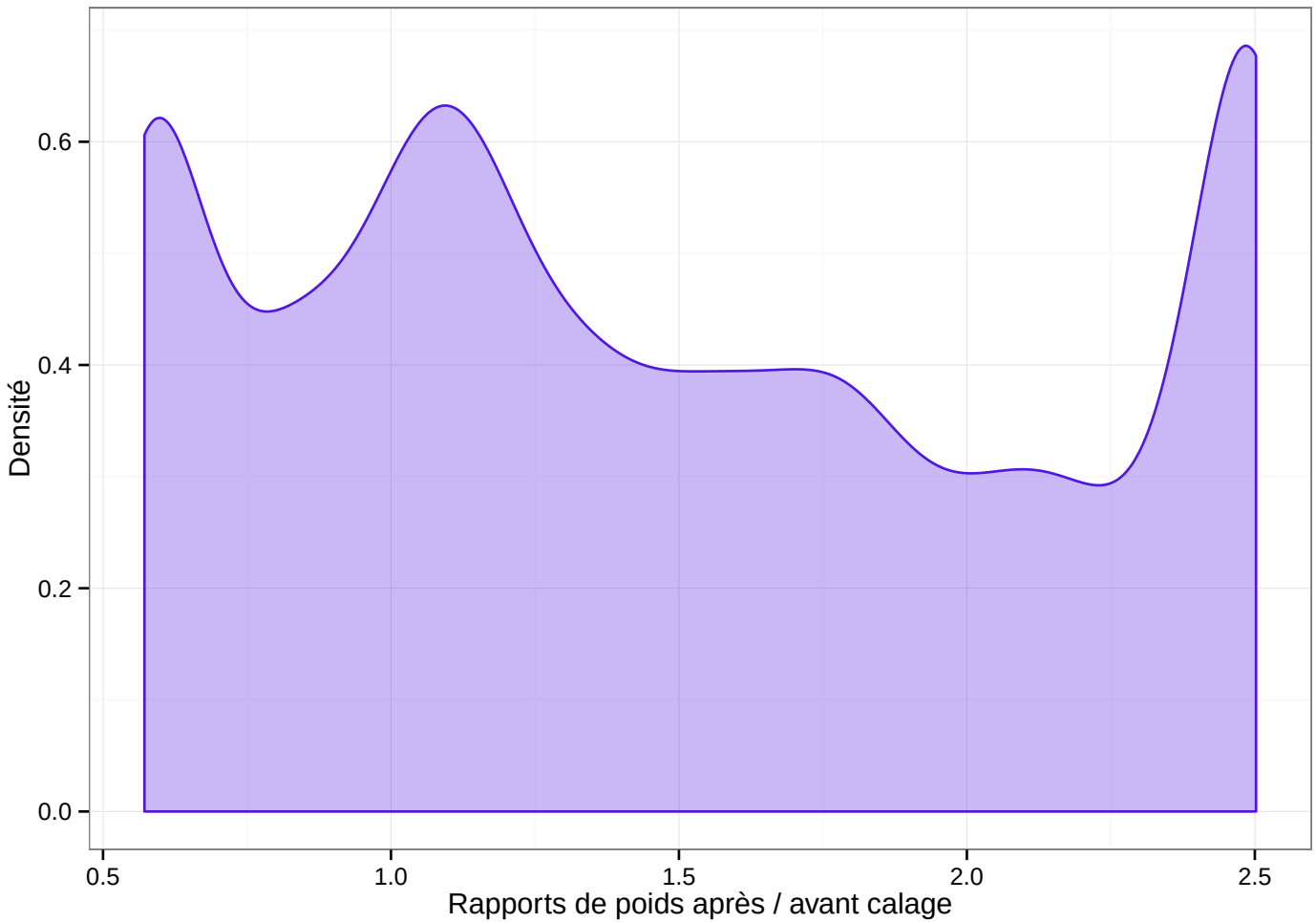
Distribution - raking ratio



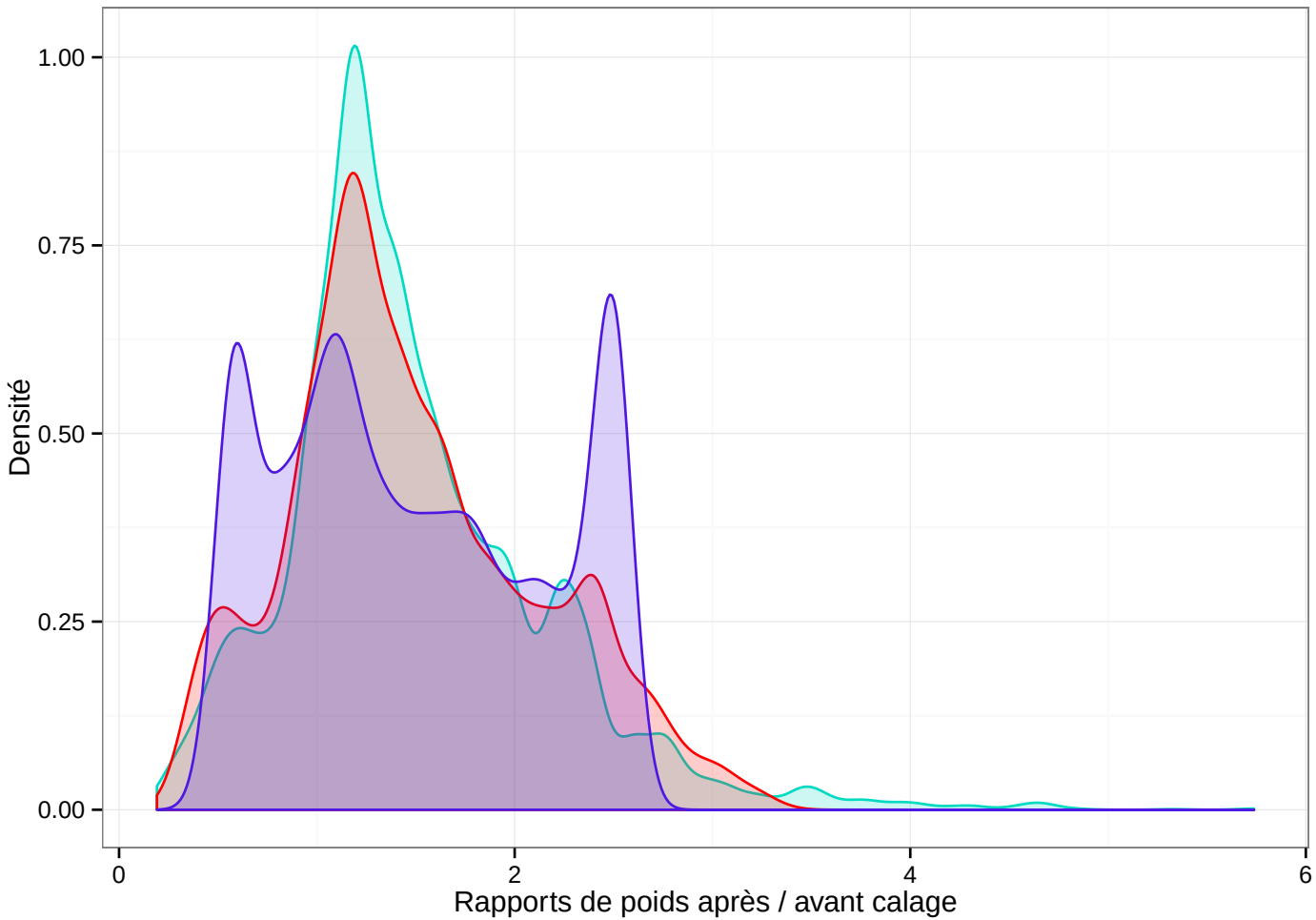
Distribution - méthode logit



Distribution - méthode logit



Distributions



Marges de calage

La source qui sert au calcul des marges de calage doit être "certaine" : source exhaustive (base de sondage) ou enquête de taille importante (RP, EEC).

Marges de calage

Les variables de calage doivent être cohérentes entre l'enquête et la source qui sert au calcul des marges : même concept, même date de référence, etc. Sinon, le calage peut dégrader les estimateurs !

Marges de calage

En résumé, une bonne variable de calage doit être :

- corrélée avec les variables d'intérêt de l'enquête ;
- correctement mesurée ;
- cohérente entre l'enquête et la source externe ;
- avec un total estimé avec précision (ou exact) dans la source externe

Calage sur marges et non-réponse

Si on utilise directement une méthode de calage sur un échantillon de répondants sans traitement préalable de la non-réponse, on peut montrer que ceci permet à la fois de corriger la non-réponse totale et d'améliorer la précision des estimations, sous la condition que les variables explicatives de la non-réponse soient incluses dans les variables de calage.

Calage sur marges et non-réponse

Possibilité de corriger de la non-réponse et de caler en une seule étape, en incluant les variables explicatives de la non-réponse dans le calage.

- Avantage : légèrement plus simple à mettre en œuvre et ne nécessite pas de connaître les X_i sur les non-répondants !
- Inconvénients : interprétation plus difficile et nécessité de connaître les totaux des X_i sur la population (ou à défaut sur l'échantillon).