Formation calage sur marges - Aspects théoriques du calage sur marges

Emmanuel Gros, Antoine Rebecq
emmanuel.gros@insee.fr, antoine.rebecq@insee.fr

INSEE - Division Sondages

29 avril 2015

Sommaire I

Sommaire II

Chapitre 1

Introduction

Partie 1

Généralités

Principe

La technique du calage sur marges généralise les méthodes de redressements vues auparavant. 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 $\mathcal{U}$.

Ceci permet d'améliorer la précision des estimations.

Principe

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

Exemple - calage sur les marges de variables catégorielles

- $X =$ catégorie socioprofessionnelle
- $Y =$ âge

On note les effectifs estimés sur l'échantillon et les effectifs connus sur la population dans un tableau.

$$\hat{N}_{ij} = \sum_{k \in s, X=i, Y=j} \frac{N}{n} = \frac{N}{n} n_{ij}$$

$$\hat{N}_{i+} = \frac{N}{n} \sum_{j} n_{ij}$$

$$\hat{N}_{j+} = \frac{N}{n} \sum_{i} n_{ij}$$

Exemple

	15-24 ans	...	35-44 ans	...	Plus de 75 ans	Marges
Agriculteurs						$\hat{N}_{1+}/N_{1+}$
...						
Cadres supérieurs			$\hat{N}_{ij}/N_{ij}$			$\hat{N}_{1+}/N_{1+}$
...						
Indépendants						$\hat{N}_{I+}/N_{I+}$
Marges	$\hat{N}_{+1}/N_{+1}$		$\hat{N}_{+j}/N_{+j}$		$\hat{N}_{+J}/N_{+J}$	$\hat{N}/N$

Principe

On cale l'échantillon sur les distributions marginales des variables dans la population ; on utilise comme information auxiliaire les valeurs $N_{i+}, \ldots, N_{+j}$, i.e. les marges du tableau de contingence croisant les deux variables. D'où le nom de **calage sur marges**.

Par extension, on parle de calage sur marges dans le cas où l'on cale sur les totaux / les effectifs dans la population d'un nombre quelconque de variables quantitatives / catégorielles.

Partie 2

Problème - objectif

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 $G(\frac{w_k}{d_k})$ 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. $G(\frac{w_k}{d_k})$ 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 :

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

où : $x_k = (x_{1k} \dots x_{jk} \dots x_{Jk})'$ et $T_X = (T_{X_{1k}} \dots T_{X_{jk}} \dots T_{X_{Jk}})'$

Partie 3

Résolution théorique

Résolution théorique

Résolution : Le lagrangien s'écrit (avec λ le vecteur des multiplicateurs de Lagrange) :

$$\mathcal{L} = \sum_{k \in s} d_k G(\frac{w_k}{d_k}) - \lambda' \left(\sum_{k \in s} w_k x_k - T_X \right)$$

Résolution théorique

Les conditions du premier ordre donnent :

$$w_k = d_k F(x_k' \lambda)$$
$$\text{où : } F = (G')^{-1}$$
$$\sum_{k \in s} d_k F(x_k' \lambda) x_k = T_X$$

La dernière équation (équation de calage) permet de déterminer λ.

Résolution théorique

Le système d'équations se résout par la méthode de Newton. La convergence est obtenue lorsque :

$$\max_{k \in s} \left| \frac{w_k^{(i+1)}}{d_k} - \frac{w_k^{(i)}}{d_k} \right| < \epsilon$$

Partie 4

Colinéarités

Équations redondantes

Reprenons l'exemple de l'exercice 1. Les équations de calage s'écrivaient :

$$\begin{cases} \sum_{k|x_k=X_1} w_k = N_{1+} = 20 & (1) \\ \sum_{k|x_k=X_2} w_k = N_{2+} = 40 & (2) \\ \sum_{k|y_k=Y_1} w_k = N_{+1} = 40 & (3) \\ \sum_{k|y_k=Y_2} w_k = N_{+2} = 20 & (4) \end{cases}$$

Équations redondantes

La somme des équations 1 et 2 donne :

$$\left\{ \sum_{k \in s} w_k = N = 60 \right. \tag{5}$$

L'équation 4 s'obtient en faisant la différence entre 5 et 3, elle est donc redondante par rapport aux équations 1, 2 et 3.

Équations redondantes

De façon générale, l'équation de calage relative à la dernière modalité de chacune des variables catégorielles numéros 2, 3, ... est redondante : elle ne fait qu'assurer le calage de l'échantillon sur le total de la population, déjà réalisé grâce à la variable catégorielle numéro 1.

On peut donc supprimer ces équations redondantes.

Chapitre 2

Les fonctions de pseudo-distance G usuelles

Les fonctions de pseudo-distance G usuelles

On indique dans la suite les fonctions $G(r)$ $(r = \frac{w_k}{d_k})$ et $F(u)$ $(u = x_k'\lambda)$.

Partie 1

La méthode linéaire

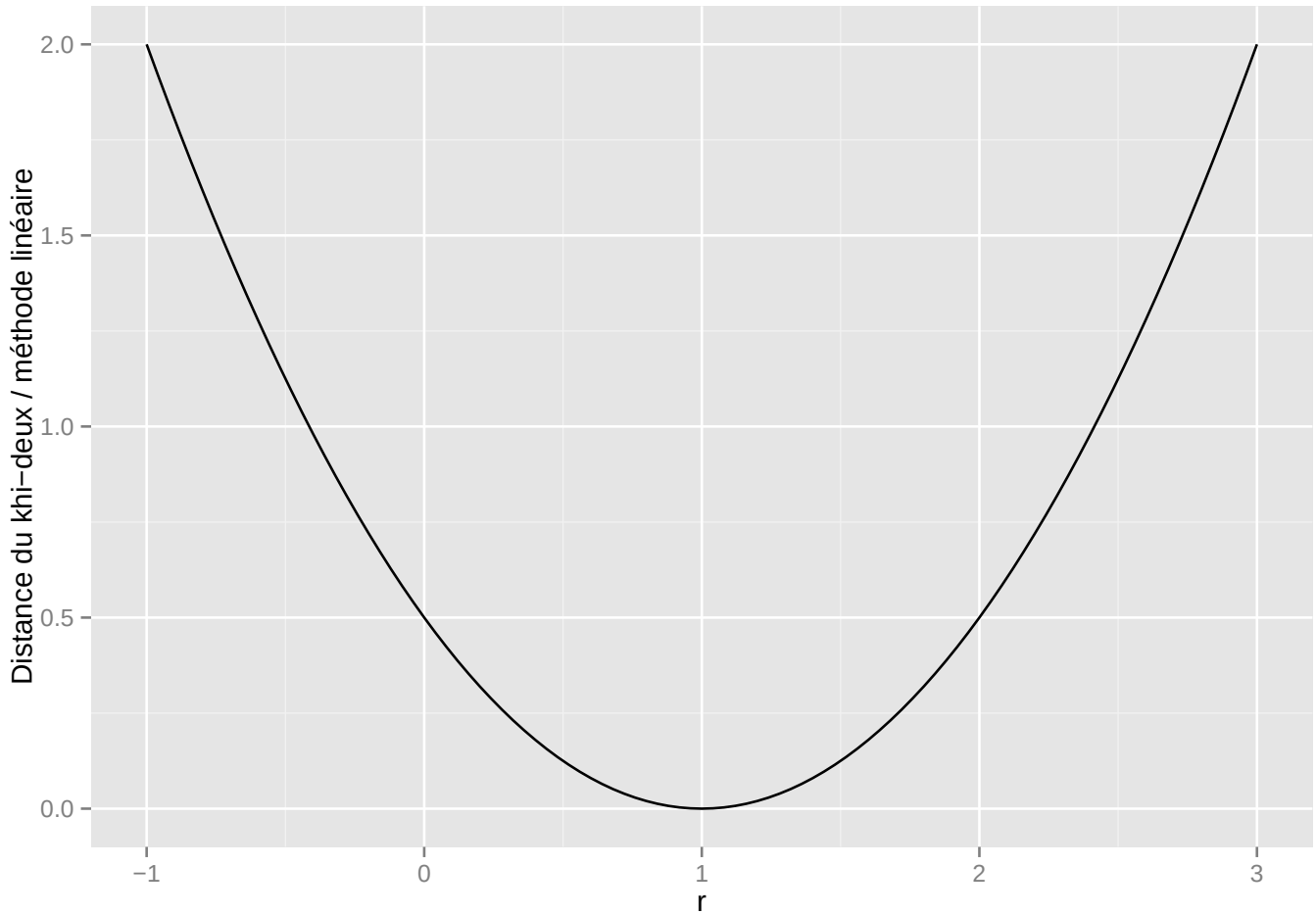
La méthode linéaire

Distance du khi-deux. Fonction réciproque de la dérivée : linéaire.

$$G(r) = \frac{1}{2}(r-1)^2 \qquad F(u) = 1 + u$$

Il y a convergence de l'algorithme de Newton dès la deuxième itération.

La méthode linéaire



La méthode linéaire

Théorème (Estimateur par la régression généralisée)

En utilisant la distance du khi-deux, on a :

$$\hat{T}_{Yw} = \hat{T}_{Y\pi} + \hat{b}_1(T_{X_1} - \hat{T}_{X_{1\pi}}) + \hat{b}_2(T_{X_2} - \hat{T}_{X_{2\pi}}) + \ldots + \hat{b}_J(T_{X_J} - \hat{T}_{X_{J\pi}})$$

où $\hat{b}_1, \hat{b}_2, \ldots, \hat{b}_J$ sont les coefficients de la régression (pondérée) de Y sur les X_j dans l'échantillon. $\hat{T}_{Yw}$ est donc **l'estimateur par la régression généralisée** *(GREG) du total T_Y*

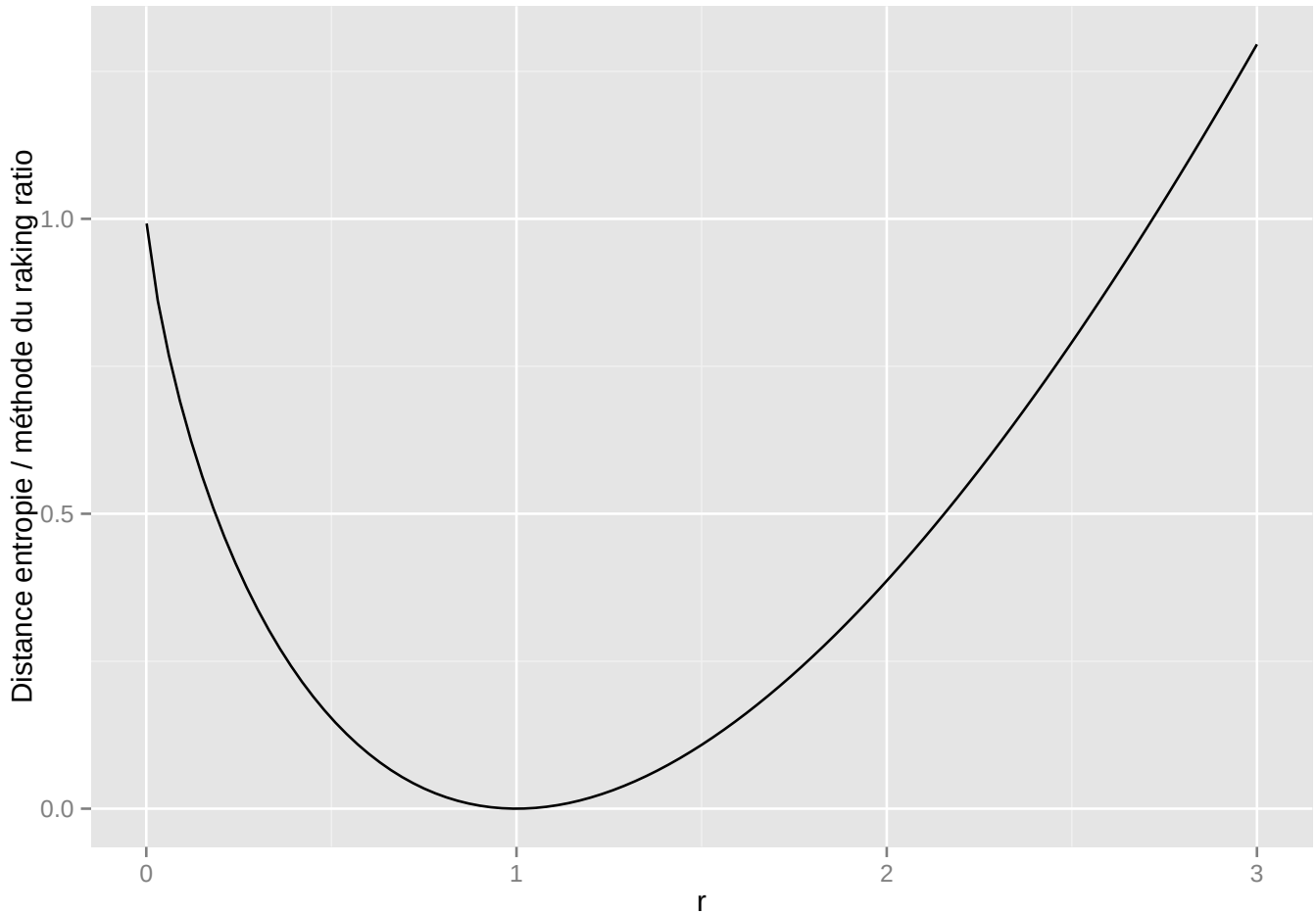
Partie 2

La méthode du raking ratio

La méthode du raking ratio

$$G(r) = r\log(r) - r + 1 \qquad F(u) = \exp(u)$$

La méthode du raking ratio



La méthode du raking ratio

Théorème (Coïncidence avec le redressement par raking ratio)

Dans le cas où les X_j sont uniquement des variables catégorielles, le calage sur marges utilisant la fonction exponentielle coïncide avec l'estimation par le raking ratio présentée au chapitre précédent.

La méthode du raking ratio

Démonstration.

Idée de la preuve : on montre par récurrence que l'algorithme de règles de trois itératives minimise l'entropie : voir Tillé, *Théorie des sondages* (Dunod, 2001). □

Partie 3

La méthode logit

La méthode logit

L et U sont deux constantes telles que $L < 1 < U$.

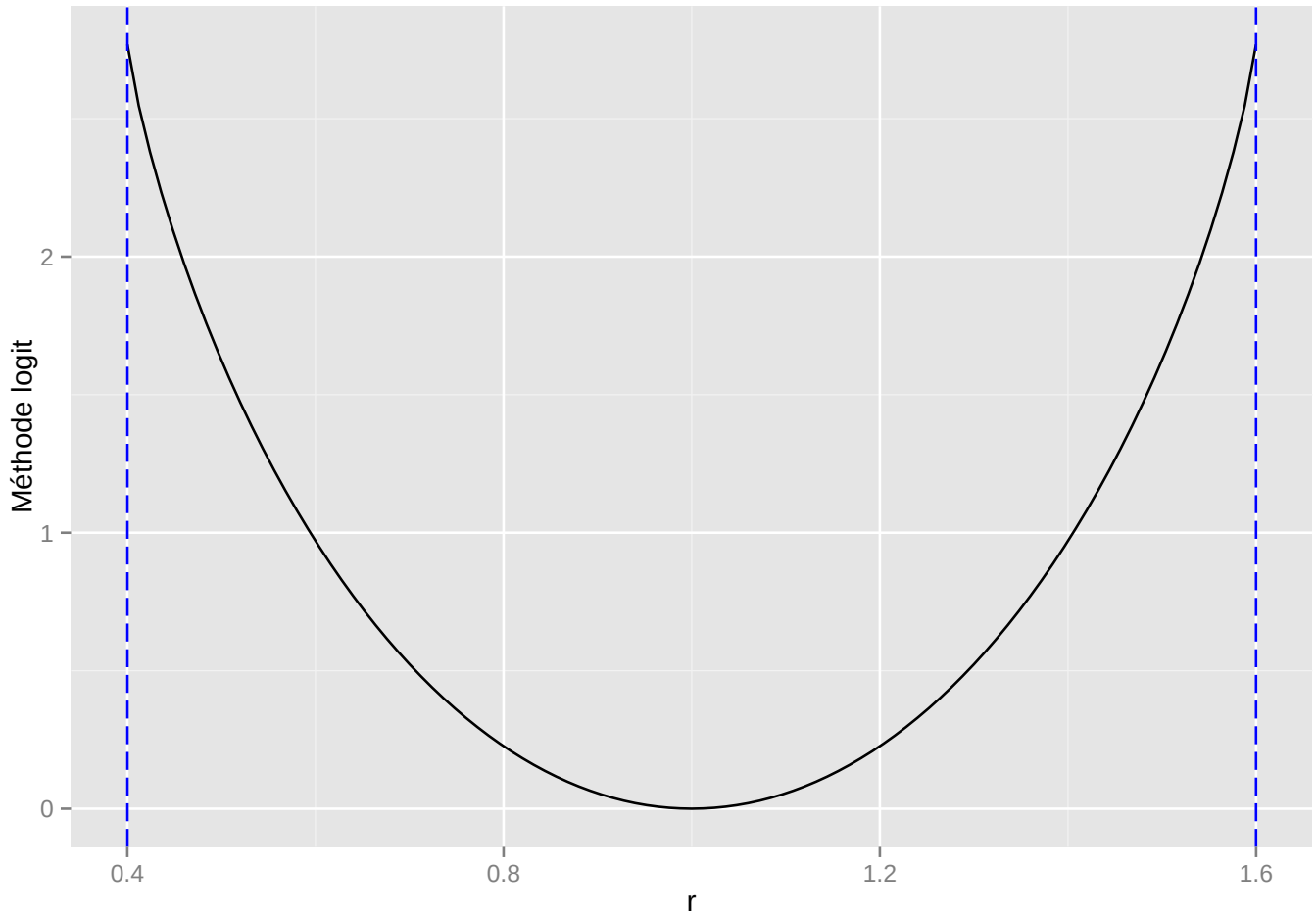
$$\begin{cases} G(r) & = \frac{1}{A}\left[(r-L)\log\frac{r-L}{1-L} + (U-R)\log\frac{U-r}{U-1}\right] \text{ si } r \in]L;U[\\ & = +\infty \text{ sinon} \end{cases}$$

où : $A = \dfrac{U-L}{(1-L)(U-1)}$

$$F(u) = \frac{L(U-1) + U(1-L)\exp(Au)}{(U-1) + (1-L)\exp(Au)}$$

et alors : $F(u) \in]L;U[$

La méthode logit



La méthode logit

C'est une méthode du raking ratio bornée : les rapports de poids après / avant calage sont compris entre L (< 1) et U (> 1).

Partie 4

La méthode linéaire tronquée

La méthode linéaire tronquée

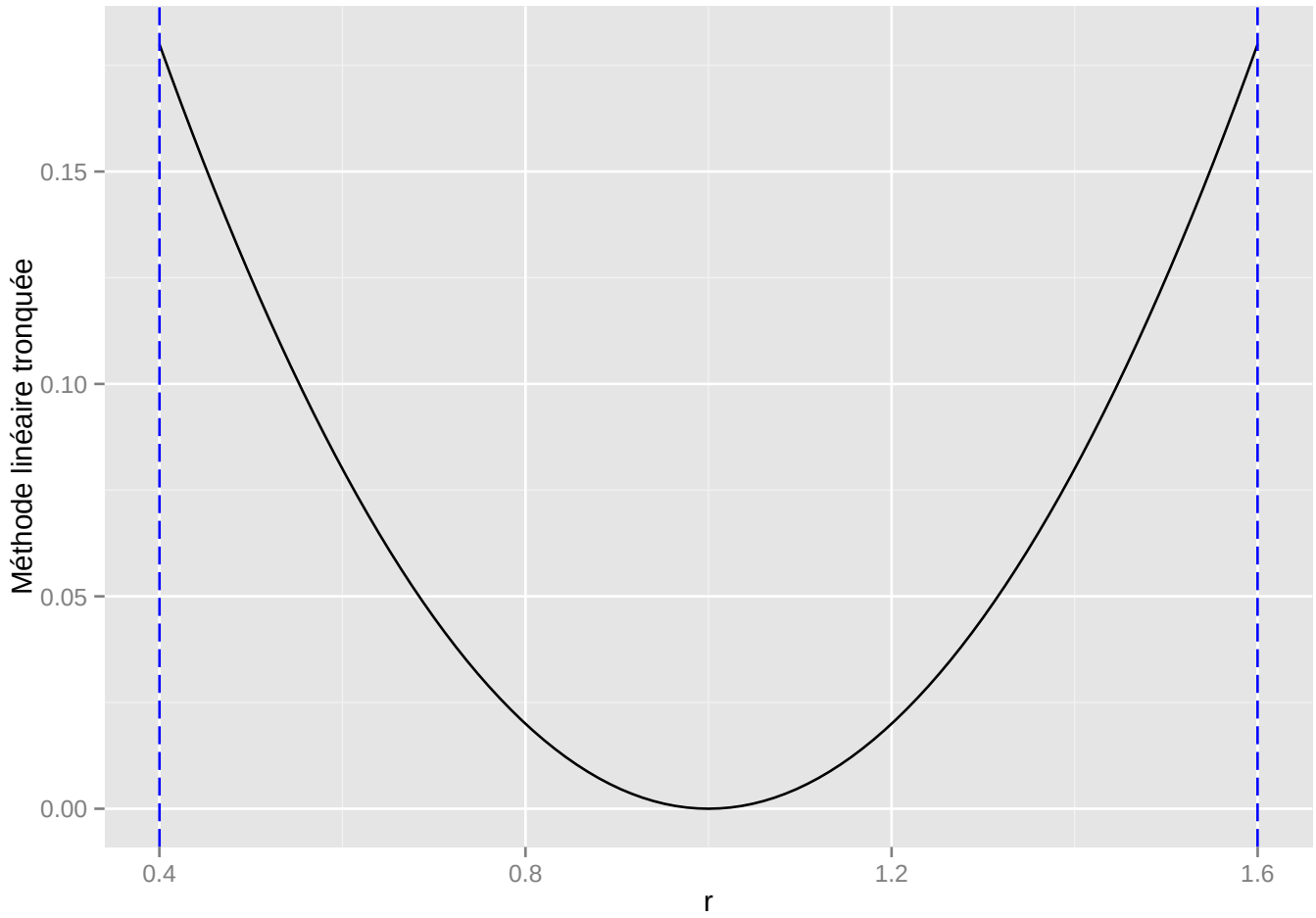
L et U sont deux constantes telles que $L < 1 < U$.

$$\begin{cases} G(r) & = \frac{1}{2}(r-1)^2 \text{ si } r \in]L;U[\\ & = +\infty \text{ sinon} \end{cases}$$

$$F(u) = 1 + u$$

et alors : $F(u) \in]L;U[$

La méthode linéaire tronquée



Chapitre 3

Propriétés

Partie 1

Estimateur pondéré et calé

Estimateur pondéré et calé

Théorème

L'estimateur par calage $\hat{T}_{Yw}$ est linéaire homogène et calé.

Estimateur pondéré et calé

Démonstration.

Par construction ! □

Partie 2

Biais

Estimateur asymptotiquement sans biais

Théorème (Estimateur asymptotiquement sans biais)

Quelle que soit la méthode utilisée, l'estimateur par calage $\hat{T}_{Yw}$ est asymptotiquement sans biais :

$$B(\hat{T}_{Yw}) \rightarrow_{n \rightarrow +\infty} 0$$

Estimateur asymptotiquement sans biais

Démonstration.

voir Deville & Särndal, *Calibration estimators in survey sampling* (1992) □

Partie 3

Variance

Rappel : variance de l'estimateur d'Horvitz-Thompson

Rappel : La variance de l'estimateur d'Horvitz-Thompson est de la forme :

$$\sum_{k \in \mathcal{U}} \sum_{l \in \mathcal{U}} \Delta_{kl} (d_k y_k)(d_l y_l)$$

Variance

Théorème

Quelle que soit la méthode utilisée, la variance de l'estimateur $\hat{T}_{Yw}$ *est de la forme :*

$$\sum_{k\in\mathcal{U}}\sum_{l\in\mathcal{U}}\Delta_{kl}(d_k E_k)(d_l E_l)$$

où $E_k = y_k - \tilde{b}' x_k$ *est le résidu de la régression de* Y *sur* $X_1 \ldots X_j \ldots X_J$ **dans** $\mathcal{U}$.

Variance

Quelle que soit la méthode utilisée, la variance de l'estimateur calé est approximativement égale à celle de l'estimateur par régression : toutes les méthodes sont (asymptotiquement) équivalentes. Cette variance de l'estimateur calé est fonction des résidus de la régression de Y sur $X_1 \dots X_j \dots X_J$: plus les résidus sont petits, plus faible est la variance.

Partie 4

Estimation de variance

Estimation de variance

On peut utiliser deux estimateurs de variance :

$$\hat{\mathrm{Var}}_1(\hat{T}_{Yw}) = \sum_{k\in s}\sum_{l\in s}\frac{\Delta_{kl}}{\pi_{kl}}(d_k e_k)(d_l e_l)$$

$$\hat{\mathrm{Var}}_1(\hat{T}_{Yw}) = \sum_{k\in s}\sum_{l\in s}\frac{\Delta_{kl}}{\pi_{kl}}(g_k d_k e_k)(g_l d_l e_l)$$

où : $g_k = \dfrac{w_k}{d_k}$ et $e_k = y_k - \hat{b}' x_k$, résidus dans la régression (pondérée par les d_k) de Y sur les $X_1 \ldots X_j \ldots X_J$ **dans l'échantillon** s.

La seconde formule est en général préférable.

Estimation de variance

Dans la pratique : Si l'on dispose d'un logiciel capable d'estimer la variance de $\hat{T}_{Y\pi}$, l'estimation de précision de l'estimateur calé s'obtient de la façon suivante :

- On effectue la régression pondérée par les d_k de Y sur les $X_1 \dots X_j \dots X_J$
- On récupère les résidus e_k de cette régression
- On crée la variable $g_k \cdot e_k$, où $g_k = \dfrac{w_k}{d_k}$
- On utilise cette variable en entrée du logiciel à la place des y_k

Chapitre 4

Choix des paramètres de calage

Partie 1

Les différentes méthodes

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.

Méthodes logit et linéaire tronquée

Comparaison des méthodes :

- Fortes corrélations entre les distributions de poids obtenues avec les différentes méthodes, mais les unités fortement impactées diffèrent
- Estimations pour des totaux ou distributions définies sur la population entière dépendant peu de la méthode de calage utilisée

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

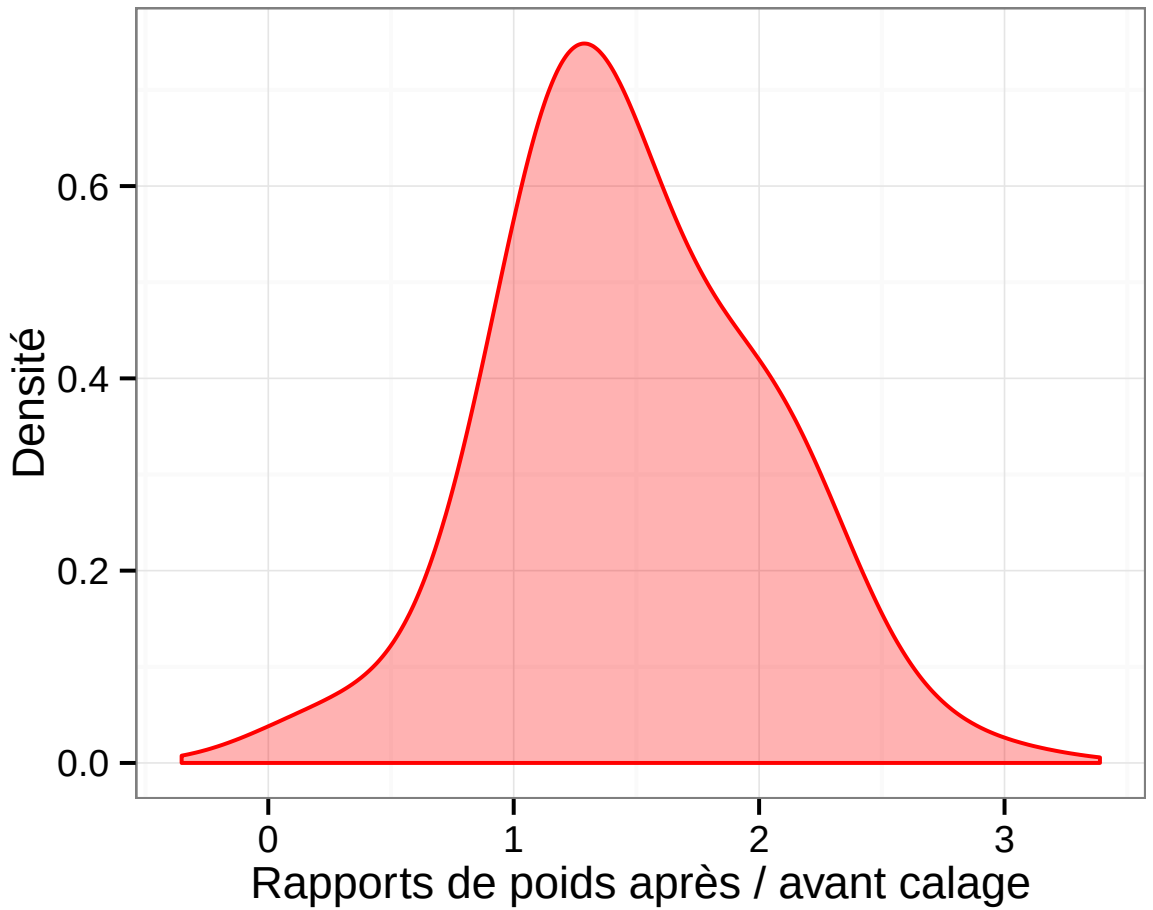
Le choix de la méthode s'effectue de manière empirique, en s'appuyant sur des critères tels que :

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

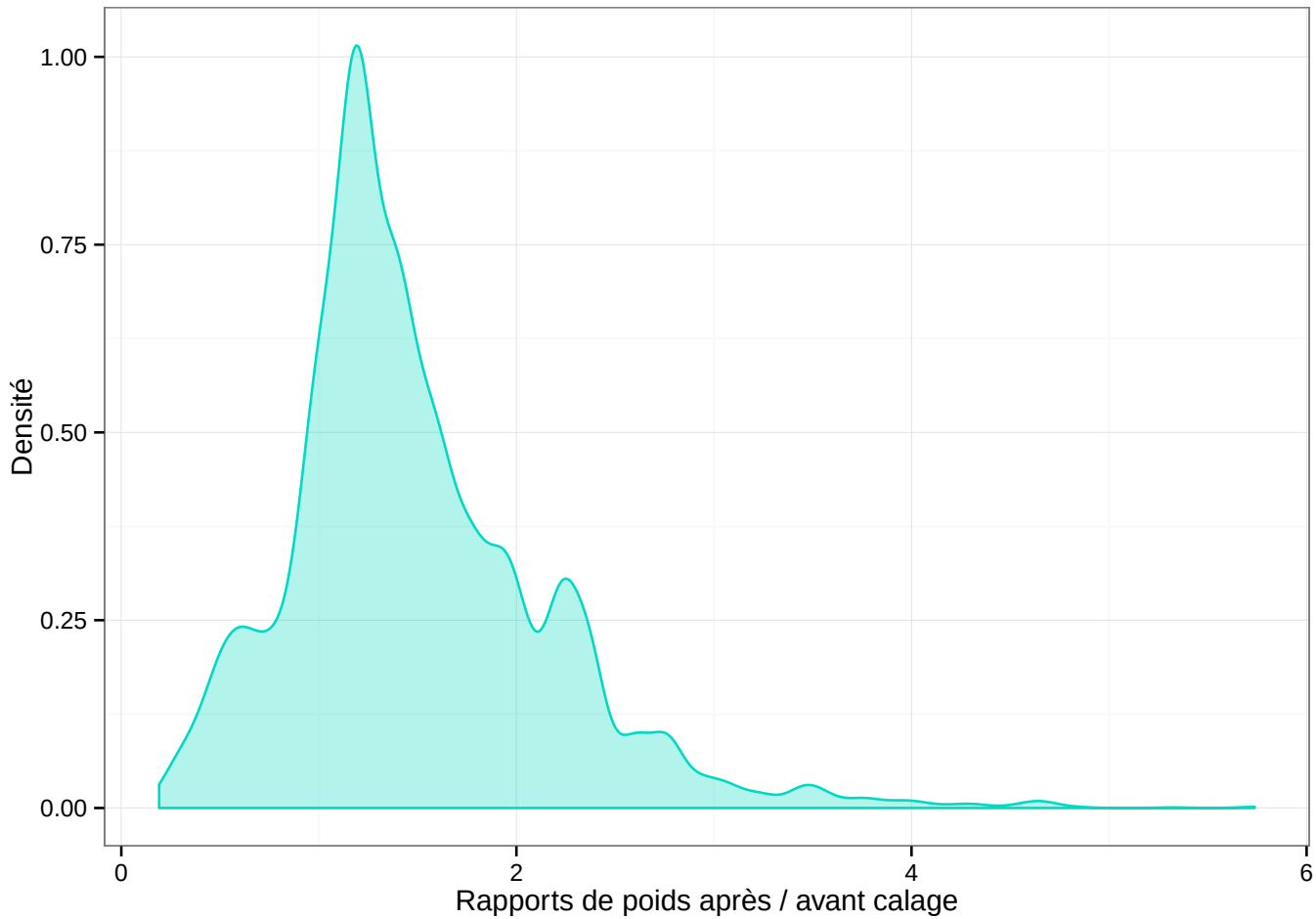
Partie 2

Considérations sur la forme de la distribution des rapports de poids

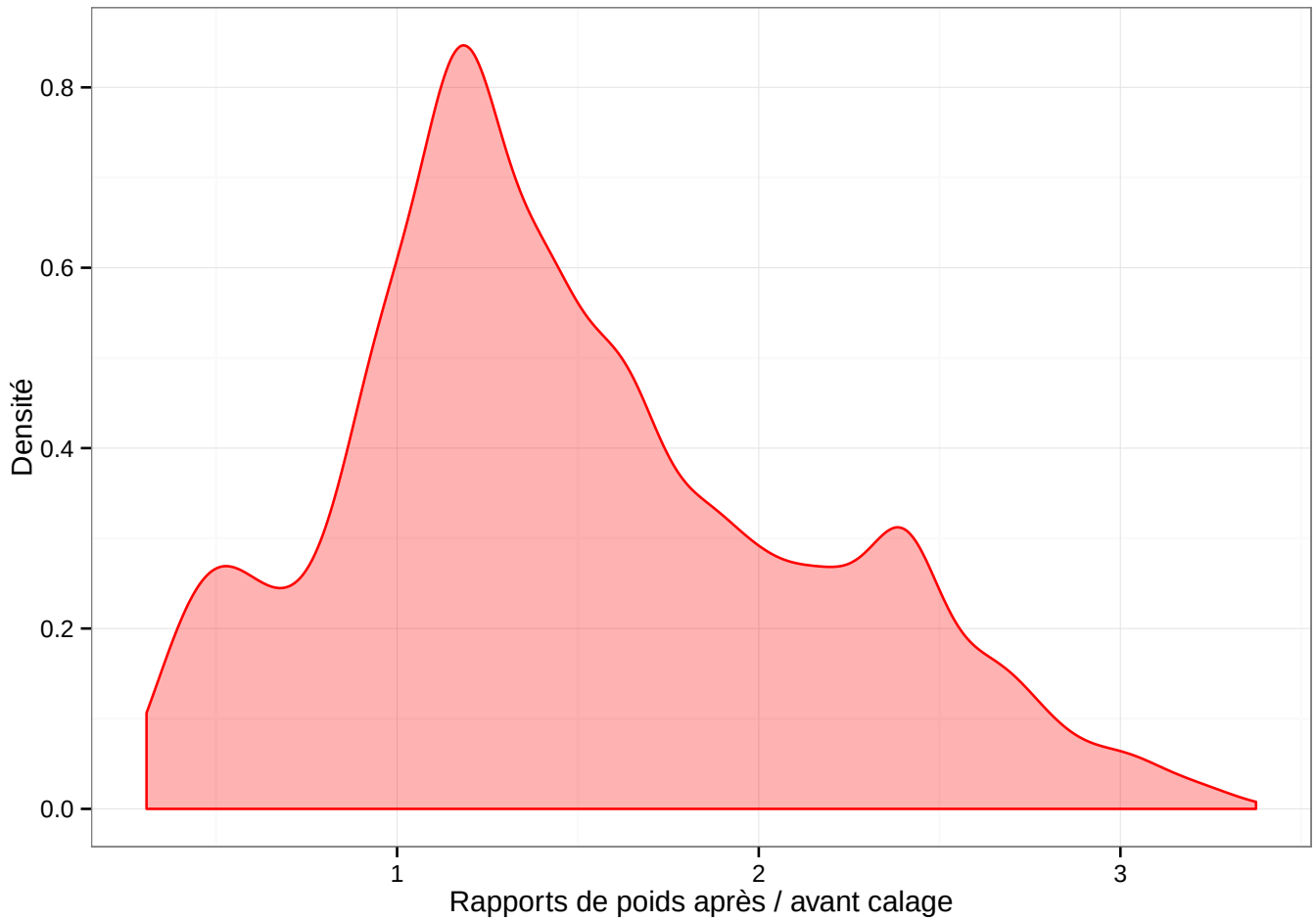
Choix des paramètres de calage - méthode linéaire



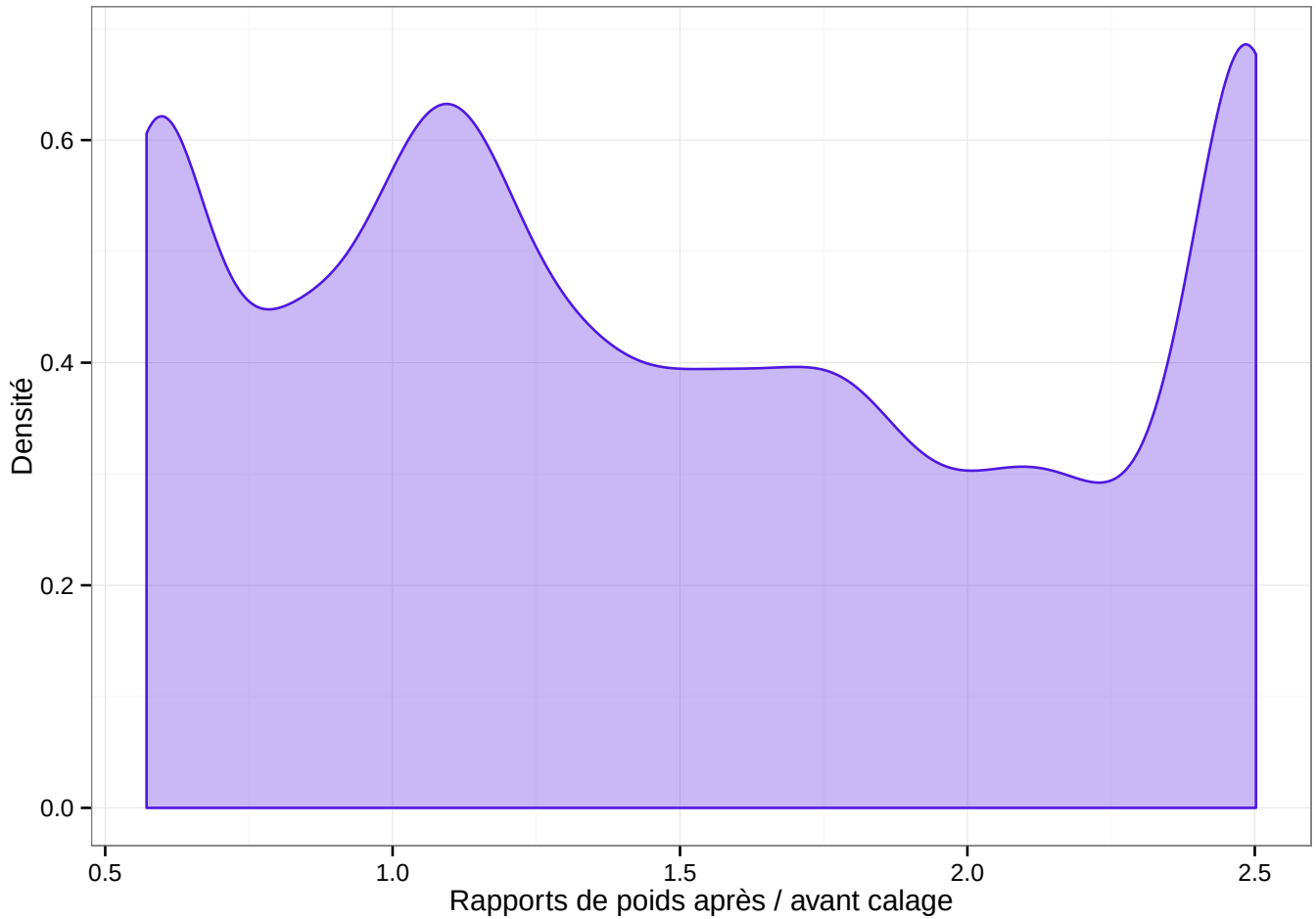
Choix des paramètres de calage - raking ratio



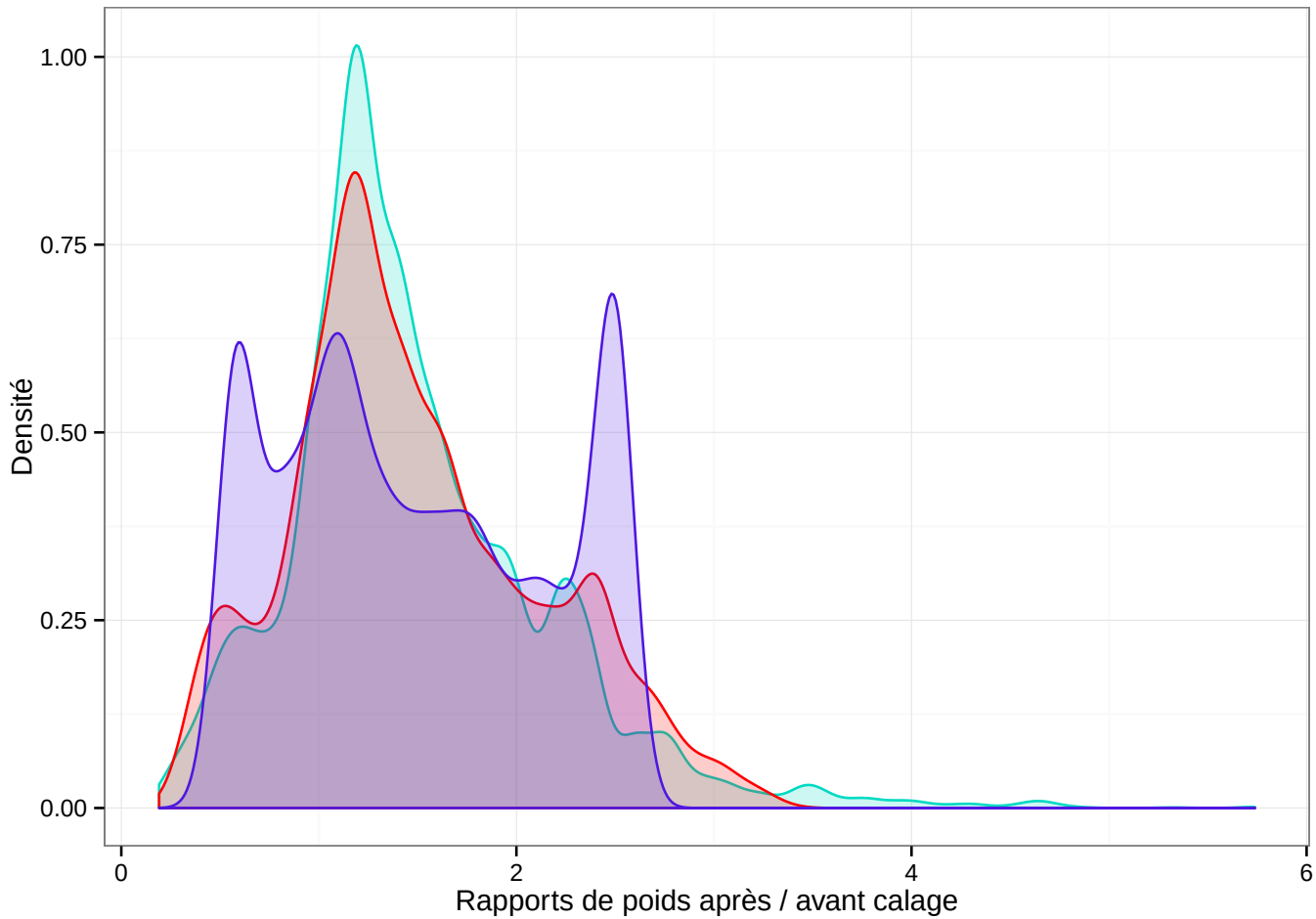
Choix des paramètres de calage - méthode logit



Choix des paramètres de calage - méthode logit



Choix des paramètres de calage



Partie 3

Remarques sur le choix des variables auxiliaires et des marges

Choix des paramètres 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).

Choix des paramètres 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 !

Choix des paramètres de calage

Se méfier des différents liens entre les sources. En particulier, ne pas caler sur des variables de la base de sondage actualisées via les résultats d'enquête.

Choix des paramètres de calage

En résumé, une bonne variable de calage doit être corrélée avec les variables d'intérêt de l'enquête et correctement mesurée.
Cohérence entre enquête et source externe et estimation précise du total dans source externe.

Chapitre 5

Calage sur marges et non-réponse

Partie 1

Non-réponse : définition et conséquences

Non-réponse : définition et conséquences

Non-réponse : incapacité d'obtenir des réponses utilisables, pour tout ou partie des variables d'intérêt.

On distingue 2 types de non-réponse :

- **la non-réponse totale :** on ne dispose d'aucune information sur l'unité sélectionnée autre que celles présentes dans la base de sondage
- **la non-réponse partielle :** l'unité sélectionnée répond seulement à une partie de l'enquête mais pas à l'ensemble des variables d'intérêt.

Non-réponse : définition et conséquences

En pratique, la non-réponse entraîne pour les estimateurs relatifs aux variables d'intérêt :

- l'introduction d'un biais
- une diminution de la précision.

Non-réponse : définition et conséquences

Principal problème = biais.

Ne rien faire revient à supposer que les non-répondants ont un comportement identique à celui des répondants.

Or les refus se répartissent rarement de manière aléatoire dans la population, et les répondants présentent généralement des caractéristiques différentes de celles des non-répondants $\rightarrow$ estimateur biaisé.

Non-réponse : définition et conséquences

Également, perte de précision de l'estimation :

- en effet, le fichier exploitable in fine est de taille plus faible que le fichier tiré
- possibilité de pallier cette perte de précision en anticipant le taux de non-réponse et en gonflant la taille de l'échantillon

Non-réponse : définition et conséquences

Il est donc impératif de traiter la non-réponse :

- En amont, méthodes pour minimiser le phénomène
- En aval, méthodes de correction de la non-réponse

Non-réponse : définition et conséquences

Toute la théorie du calage sur marges présentée auparavant est valide en l'absence de non-réponse... cadre théorique qu'on rencontre assez rarement en pratique !

Non-réponse : définition et conséquences

Pour procéder à un calage sur une enquête en présence de non-réponse, deux approches possibles :

- soit on corrige de la non-réponse – par repondération ou imputation – avant de procéder au calage, ce qui nous ramène à un contexte similaire au précédent et on peut procéder au calage sur données corrigées de la non-réponse de manière classique
- soit on cale directement l'échantillon de répondants sans traitement préalable de la non-réponse, et dans ce cas le calage peut, sous certaines conditions, corriger de la non-réponse par repondération.

Partie 2

Correction de la non-réponse par repondération

Correction de la non-réponse par repondération

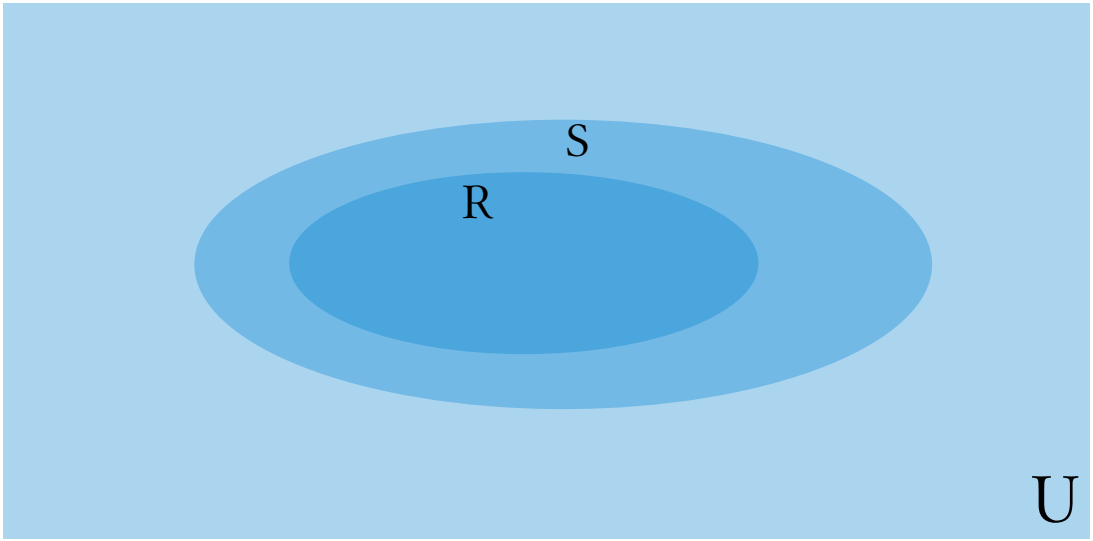
La repondération est une technique qui consiste à augmenter les poids des unités répondantes pour compenser les unités défaillantes. C'est une méthode utilisée pour corriger la **non-réponse totale**.

Rappel de théorie des sondages : plan en deux phases

Un plan en deux phases est un échantillon issu d'un échantillon de la population.

La première phase consiste à tirer un échantillon S1 dans la population U, et la deuxième consiste à tirer un échantillon S2 au sein de l'échantillon de première phase S1.

Rappel de théorie des sondages : plan en deux phases



$$\mathcal{U} \xrightarrow[\pi_k]{} S \xrightarrow[p_k]{} R$$

Rappel de théorie des sondages : plan en deux phases

$$\hat{T}_{Y,2\phi} = \sum_{k \in R} \frac{y_k}{\pi_k p_k}$$

est un estimateur sans biais du total $T(Y)$.

Rappel de théorie des sondages : plan en deux phases

La non-réponse peut-être considérée comme un tirage en deux phases.

Problème : Les p_k sont inconnues.

Principe de la repondération

On cherche à estimer le total : $T(Y)$.

En l'absence de non-réponse, on utilise l'estimateur d'Horvitz-Thompson : $\hat{T}_{Y\pi}$

En présence de non-réponse, un estimateur sans biais est donné par : $\sum_{k \in R} \frac{y_k}{\pi_k p_k}$, avec p_k estimées par un modèle.

Remarque : Comme on a $0 < p_k < 1$, cela conduit bien à augmenter les poids initiaux (de Horvitz-Thompson $\frac{1}{\pi_k}$)

Partie 3

Retour au calage sur marges

Retour au calage sur marges

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.

Retour au calage sur marges

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).