Redressements

Sylvie Rousseau, Gilbert Saporta, novembre 2012

- Une opération indispensable pour améliorer les estimations dans une enquête par sondages
  - Egalement appelée: calibration, calage

- "Calibration is a procedure than can be used to incorporate auxiliary data. This procedure adjusts the sampling weights by multipliers known as calibration factors that make the estimates agree with known totals. The resulting weights are called calibration weights or final estimation weights. These calibration weights will generally result in estimates that are **design consistent**, and that have a **smaller variance** than the Horvitz Thompson estimator."

  Quality Guidelines (fourth edition) of Statistics Canada (2003)

- Trop souvent mal comprise et mal interprétée (*manipulations*…)
- Repondération des individus de l'échantillon
  - *L'expression de redressement des résultats est tout à fait impropre. On redresse l'échantillon, c'est à dire qu'on modifie le poids accordé à chaque enquêté sur la base de la conformité aux statistiques de référence*. (B.Riandey, SFdS)
- Mais non exempte de risque dans certains cas

- Principe :

  Utiliser *a posteriori* une information supplémentaire corrélée avec la variable à étudier

- Objectif:
  - accroître la précision de l'estimation
  - assurer la cohérence des résultats par rapport à l'information supplémentaire
- Information auxiliaire :

  Variables de contrôle dont on connaît :
  - des caractéristiques globales,
  - ou des caractéristiques par classes,
  - ou les valeurs pour chaque unité de la population

Plan

1. Estimateur par le quotient (ou ratio)
2. Estimateur par la régression
3. Estimateur post-stratifié
4. Estimateur du raking-ratio
5. Calage sur marges

1.Estimateur par le quotient

- Cadre :
  - La variable auxiliaire est quantitative
  - On connaît le total (ou la moyenne) de cette variable sur l'échantillon **et** sur la population
  - On va ajuster l'estimation sur cette grandeur connue
- *Exemple :*
  - *On veut estimer le CA moyen d'hypermarchés ($\overline{Y}$)*
  - *On a enquêté 80 hypermarchés*
  - *On sait que le nombre moyen de caisses dans la population des hypermarchés est* $\overline{X} = 28$
  - *On relève sur l'échantillon*

$$\hat{\overline{Y}} = 110{,}2 \ k€ \qquad \hat{\overline{X}} = 28{,}8$$

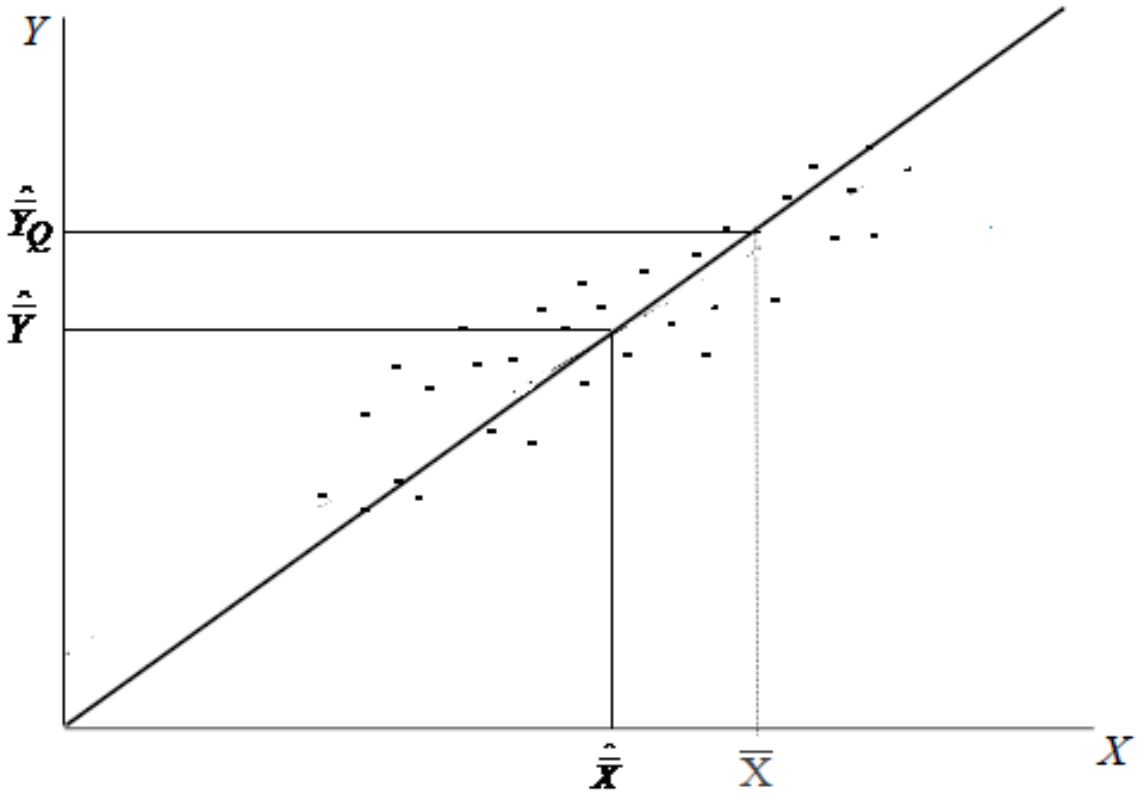
  - *L'estimateur par le quotient vaut :*

$$\hat{\overline{Y}}_Q = 110{,}2 \times \frac{28}{28{,}8} = 107{,}1 \ k€$$

Formule générale

- Principe : règle de 3
- Formule générale : $\hat{\bar{Y}}_Q = \hat{\bar{Y}} \times \frac{\bar{X}}{\hat{\bar{X}}}$
- Hypothèse de proportionnalité
- Biaisé mais négligeable si *n>1000*
- Gain de précision par rapport à un PESR de même taille pourvu que l'hypothèse de proportionnalité soit valide

Interprétation graphique



poids après redressement

- On a : $\hat{T}_{y_Q} = \left(\sum_{k \in s} \frac{Y_k}{\pi_k} \right) \frac{\overline{X}}{\hat{\overline{X}}}$ et $\hat{T}_y = \sum_{k \in S} \frac{Y_k}{\pi_k}$
- Le poids après redressement de *k* vaut $\frac{1}{\pi_k} \frac{\overline{X}}{\hat{\overline{X}}} = \frac{1}{\pi_k} \frac{T_X}{\hat{T}_X}$
- Le poids de sondage valait $\frac{1}{\pi_k}$

Espérance

- Cas général

$$E\left(\hat{\bar{Y}}_Q\right) \cong \bar{Y}\left[1+\frac{Var\left(\hat{\bar{X}}\right)}{\bar{X}^2}-\frac{Cov\left(\hat{\bar{X}},\hat{\bar{Y}}\right)}{\bar{X}\times\bar{Y}}\right]$$

Biais

  - Dans le cas d'un PESR de *n* parmi *N* : $E\left(\hat{\bar{Y}}_Q\right) \cong \bar{Y}\left[1+\left(\frac{1}{n}-\frac{1}{N}\right)\left(\frac{S_x^2}{\bar{X}^2}-\frac{S_{x,y}}{\bar{X}\times\bar{Y}}\right)\right]$

- Biais en *1/n*
- Biais nul si *Y* et *X* sont proportionnelles (droite de régression passant par l'origine)

$$\frac{Cov\left(\hat{\bar{X}},\hat{\bar{Y}}\right)}{Var\left(\hat{\bar{X}}\right)}=\frac{\bar{Y}}{\bar{X}}=R \qquad \text{i.e.} \quad Y_k = RX_k + u_k$$

  - Dans le cas d'un PESR : $\frac{S_{x,y}}{S_x^2}=\frac{\bar{Y}}{\bar{X}}=R$

- Développement limité en 0 avec $\varepsilon = \dfrac{\hat{\bar{X}} - \bar{X}}{\bar{X}}$ soit $\hat{\bar{X}} = \bar{X}(1+\varepsilon)$

$$\hat{\bar{Y}}_Q - \bar{Y} = \hat{\bar{Y}}\frac{\bar{X}}{\hat{\bar{X}}} - \bar{Y} = \frac{\hat{\bar{Y}}.\bar{X} - \bar{Y}.\hat{\bar{X}}}{\hat{\bar{X}}} = \frac{\hat{\bar{Y}}.\bar{X} - \bar{Y}.\hat{\bar{X}}}{\bar{X}(1+\varepsilon)} = \frac{\hat{\bar{Y}} - \dfrac{\bar{Y}}{\bar{X}}\hat{\bar{X}}}{1+\varepsilon} = \frac{\hat{\bar{Y}} - R\hat{\bar{X}}}{1+\varepsilon} \qquad \left(\text{où } R = \frac{\bar{Y}}{\bar{X}}\right)$$

$$\cong \left(\hat{\bar{Y}} - R\hat{\bar{X}}\right)(1-\varepsilon) \cong \left(\hat{\bar{Y}} - R\hat{\bar{X}}\right)\left(1 - \frac{\hat{\bar{X}} - \bar{X}}{\bar{X}}\right)$$

$$E\left(\hat{\bar{Y}}_Q - \bar{Y}\right) \cong E\left[\left(\hat{\bar{Y}} - R\hat{\bar{X}}\right)\left(1 - \frac{\hat{\bar{X}} - \bar{X}}{\bar{X}}\right)\right] \cong -E\left[\left(\hat{\bar{Y}} - R\hat{\bar{X}}\right)\left(\frac{\hat{\bar{X}} - \bar{X}}{\bar{X}}\right)\right]$$

$$\cong -\frac{1}{\bar{X}}\left[E(\hat{\bar{Y}}.\hat{\bar{X}}) - \bar{Y}.\bar{X} - R.E(\hat{\bar{X}}^2) + R.\bar{X}^2\right]$$

$$\cong \frac{R.Var(\hat{\bar{X}}) - Cov(\bar{X}, \hat{\bar{Y}})}{\bar{X}}$$

Erreur quadratique moyenne

- Cas général : $EQM\left(\hat{\bar{Y}}_Q\right) = E\left(\hat{\bar{Y}}_Q - \bar{Y}\right)^2 \cong Var\left(\hat{\bar{Y}} - R\hat{\bar{X}}\right)$

$$\cong Var\left(\hat{\bar{Y}}\right) - 2RCov\left(\hat{\bar{X}}, \hat{\bar{Y}}\right) + R^2 Var\left(\hat{\bar{X}}\right)$$

  - Cas d'un PESR de taille *n* parmi *N :*

$$EQM\left(\hat{\bar{Y}}_Q\right) \cong \left(1 - \frac{n}{N}\right)\frac{S_y^2 - 2RS_{xy} + R^2 S_x^2}{n}$$

- Estimée par : $\hat{EQM}\left(\hat{\bar{Y}}_Q\right) \cong \hat{Var}\left(\hat{\bar{Y}}\right) - 2\hat{R}\hat{Cov}\left(\hat{\bar{X}}, \hat{\bar{Y}}\right) + \hat{R}^2\hat{Var}\left(\hat{\bar{X}}\right)$ avec $\hat{R} = \frac{\hat{\bar{Y}}}{\hat{\bar{X}}}$

  - Cas d'un PESR de taille *n* parmi *N :*

$$\hat{EQM}\left(\hat{\bar{Y}}_Q\right) \cong \left(1 - \frac{n}{N}\right)\frac{\hat{S}_y^2 - 2\hat{R}\hat{S}_{xy} + \hat{R}^2\hat{S}_x^2}{n}$$

Comparaison avec un pesr

- Cas général :

$$EQM\left(\hat{\bar{Y}}_Q\right) \leq Var\left(\hat{\bar{Y}}\right) \Leftrightarrow 2RCov\left(\hat{\bar{X}},\hat{\bar{Y}}\right) - R^2 Var\left(\hat{\bar{X}}\right) \geq 0$$

  - Cas d'un PESR de taille *n* parmi *N :*

$$EQM\left(\hat{\bar{Y}}_Q\right) \leq Var\left(\hat{\bar{Y}}\right) \Leftrightarrow 2RS_{xy} - R^2 S_x^2 \geq 0$$

$$\Leftrightarrow \frac{S_{xy}}{S_x^2} \geq \frac{R}{2} \qquad (X \text{ et } Y \text{ positives})$$

$$\Leftrightarrow b \geq \frac{1}{2}\frac{\bar{Y}}{\bar{X}} \qquad \left(b = \frac{S_{xy}}{S_x^2}\right.$$

pente de la droite de régression de Y sur X)

- L‘estimation par la méthode du ratio est efficace si les variables Y et X sont « à peu près » proportionnelles
- Mais:
  - N’améliore pas toujours la précision si la droite ne passe pas par l’origine
  - N’utilise pas les valeurs individuelles de la variable auxiliaire sur l’échantillon

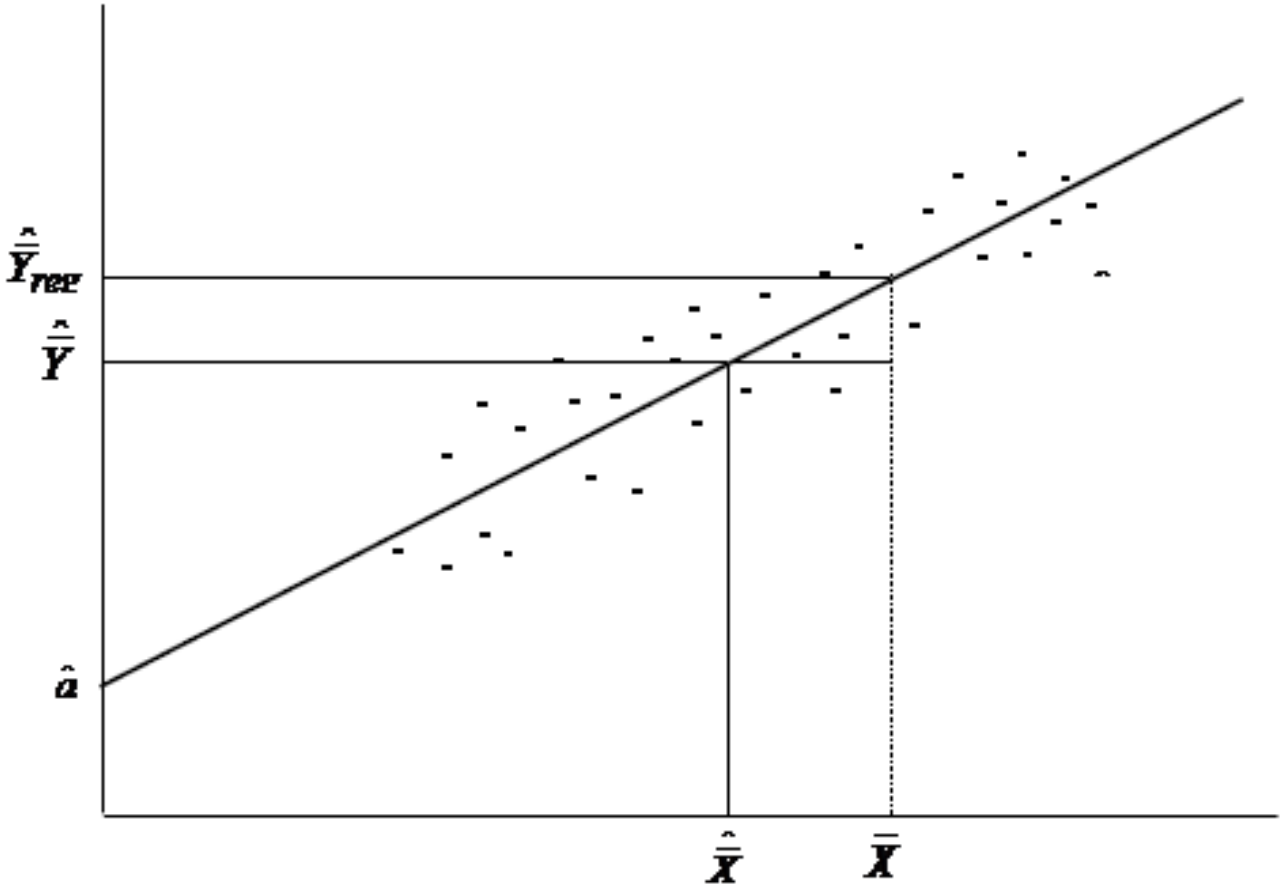
2.Estimateur par la régression

- Cadre :
  - La variable auxiliaire est quantitative
  - On l'observe pour chaque individu de l'échantillon et on en connaît la vraie moyenne sur la population
  - On va ajuster l'estimation sur cette grandeur connue
- Hypothèse : relation affine entre *Y* et *X* $y = a + bx$
- Formule générale :

$$\hat{\overline{Y}}_{reg} = \hat{\overline{Y}} + \hat{b}\left(\overline{X} - \hat{\overline{X}} \right)$$

avec $\hat{b} = \dfrac{\hat{S}_{xy}}{\hat{S}_x^2}$ pente estimée de la droite de régression de *Y* sur *X*

Interprétation graphique



Propriétés

- Biaisé mais biais négligeable pour *n* assez grand
- Erreur quadratique moyenne dans le cas d'un PESR

$$EQM\left(\hat{\bar{Y}}_{reg}\right) = \left(1 - \frac{n}{N}\right)\frac{S_y^2}{n}\left(1 - \rho^2\right) \qquad \text{avec} \quad \rho = \frac{S_{xy}}{S_x S_y}$$

- Estimée par :

$$\hat{EQM}\left(\hat{\bar{Y}}_{reg}\right) = \left(1 - \frac{n}{N}\right)\frac{\hat{S}_y^2}{n}\left(1 - \hat{\rho}^2\right) \qquad \text{avec} \quad \hat{\rho} = \frac{\hat{S}_{xy}}{\hat{S}_x \hat{S}_y}$$

Comparaisons

- Meilleur que l'estimateur d'Horvitz-Thompson

$$EQM\left(\hat{\bar{Y}}_{reg}\right) \leq Var\left(\bar{Y}_{HT}\right) \Leftrightarrow 1-\rho^2 \geq 0$$ *(toujours vrai)*

- Meilleur que l'estimateur par le quotient

$$EQM\left(\hat{\bar{Y}}_{reg}\right) \leq EQM\left(\hat{\bar{Y}}_{Q}\right) \Leftrightarrow S_y^2\left(1-\rho^2\right) \leq S_y^2 - 2RS_{xy} + R^2S_x^2$$

$$\Leftrightarrow R^2S_x^2 + \frac{S_{xy}^2}{S_x^2} - 2RS_{xy} \geq 0$$

$$\Leftrightarrow R^2S_x^4 - 2RS_x^2S_{xy} + S_{xy}^2 \geq 0$$

$$\Leftrightarrow \left(RS_x^2 - S_{xy}\right)^2 \geq 0$$ *(toujours vrai)*

- Si la relation entre X et Y est linéaire et non affine (ordonnée à l'origine nulle), alors l'estimateur par la régression est égal à l'estimateur par le quotient

- Un inconvénient:
  - certains poids de redressement peuvent être négatifs

$$w_i = \frac{1}{n} + \left(\bar{X} - \bar{x}\right)\frac{\left(x_i - \bar{x}\right)}{\sum_s \left(x_i - \bar{x}\right)^2}$$

3. Estimateur post-stratifié

- Cadre :
  - La variable auxiliaire est qualitative
  - On définit après l'enquête des groupes d'individus, appelés **post-strates**.
  - On observe les effectifs des post-strates sur l'échantillon
  - On connaît la répartition de la population selon ces post-strates
  - On va ajuster l'estimation sur cette répartition

- Remarques :
  - Les effectifs des post-strates dans l'échantillon ne sont connus qu'après enquête
  - Ils dépendent de l'échantillon choisi : ce sont des variables aléatoires

1er exemple

- *On veut estimer le taux de fréquentation des salles de cinéma*
- *On sait que cette activité est liée à la possession de TV*
- *On connait le taux d'équipement en TV : $p_{télé} = 80\%$*
- *On observe sur un échantillon de taille 1000 choisi par PESR :*

Cinéma / Télé	Oui	Non	Total		
Oui	20	680	700	70% au lieu de 80 %	×8/7
Non	80	220	300	30% au lieu de 20%	×2/3
Total	100	900	1000		

- Après redressement:

Cinéma / Télé	Oui	Non	Total
Oui	23	777	800
Non	53	147	200
Total	76	924	1000

- nouvelle estimation 7.6%
- coefficients de redressement: 1.14 pour télé, 0.67 pour non télé

2éme exemple

- *Enquête concernant les revenus : on observe X=classe d'âge et Y=revenu*
- *Résultats observés :*

Tranche d'âge	≤ 20	21 - 35	36 - 50	≥ 50
Proportion observée	15 %	30 %	30 %	25 %
Vraie proportion	20 %	35 %	30 %	15 %
Revenu moyen observé	6 000	9 000	15 000	12 000

- *Estimateur d'Horvitz-Thompson :*

$$\hat{\bar{Y}} = 6000 \times 0{,}15 + 9000 \times 0{,}3 + 15000 \times 0{,}3 + 12000 \times 0{,}25 = 11100$$

- *Estimateur post-stratifié*

$$\hat{\bar{Y}}_{post} = 6000 \times 0{,}2 + 9000 \times 0{,}35 + 15000 \times 0{,}3 + 12000 \times 0{,}15 = 10650$$

Formules

- Total et moyenne sur la population :

$$T_y = \sum_{k \in U} y_k = \sum_{h=1}^{H} \left(\sum_{k=1}^{N_h} y_k \right) = \sum_{h=1}^{H} T_{yh} = \sum_{h=1}^{H} N_h \bar{Y}_h \qquad \bar{Y} = \frac{T_y}{N} = \sum_{h=1}^{H} \frac{N_h}{N} \bar{Y}_h$$

- Estimateurs d'Horvitz-Thompson :

$$\hat{T}_y = \sum_{k \in S} \frac{y_k}{\pi_k} = \frac{N}{n} \sum_{k \in S} y_k = \frac{N}{n} \sum_{h=1}^{H} \sum_{k \in S_h} y_k = N \sum_{h=1}^{H} \frac{n_h}{n} \hat{\bar{y}}_h \qquad \hat{\bar{Y}} = \frac{\hat{T}_y}{N} = \frac{1}{n} \sum_{k \in S} y_k$$

- Estimateurs post-stratifiés

$$\hat{T}_{y\,post} = \sum_{h=1}^{H} N_h \hat{\bar{y}}_h = \sum_{h=1}^{H} N_h \left(\frac{1}{n_h} \sum_{k \in S_h} y_k \right)$$

$$\hat{\bar{Y}}_{post} = \sum_{h=1}^{H} \frac{N_h}{N} \hat{\bar{y}}_h = \sum_{h=1}^{H} \frac{N_h}{N} \left(\frac{1}{n_h} \sum_{k \in S_h} y_k \right)$$

Poids après redressement

- On a : $\hat{T}y_{post} = \sum_{h=1}^{H} N_h \left(\frac{1}{n_h} \sum_{k \in S_h} y_k \right)$ et $\hat{T}_y = \frac{N}{n} \sum_{k \in S} y_k$
- Le poids après redressement de *k* vaut $\frac{N_h}{n_h}$
- Le poids de sondage valait $\frac{N}{n}$

Espérance

$$E\left(\hat{T}_{y\,post}\right) = E\left[E\left(\hat{T}_{y\,post/n_h,h=1,...,H}\right)\right]$$

$$E\left(\hat{T}_{y\,post/n_h,h=1,...,H}\right) = \sum_{h=1}^{H} N_h E\left(\hat{\bar{Y}}_h / n_h, h=1,...,H\right) = \sum_{\substack{h=1 \\ n_h>0}}^{H} N_h \bar{Y}_h = T_y - \sum_{\substack{h=1 \\ n_h=0}}^{H} T_{yh}$$

$$E\left(\hat{T}_{y\,post}\right) = T_y - \sum_{h=1}^{H} T_{yh} P\left(n_h = 0\right)$$

car si n_h est fixé, le plan est un PESR

- Les effectifs n_h peuvent être nuls, d'où le léger biais de l'estimateur post-stratifié
- Pour l'éviter, définir les post-strates de sorte à vérifier :

$$n\frac{N_h}{N} \geq 30 \quad \forall h = 1,...,H$$

Variance

$$Var\left(\hat{T}_{y\,post}\right) = E\left[Var\left(\hat{T}_{y\,post/n_h,h=1,\ldots,H}\right)\right] + \underbrace{Var\left[E\left(\hat{T}_{y\,post/n_h,h=1,\ldots,H}\right)\right]}_{Var\left(T_y - \sum_{\substack{h=1 \\ n_h=0}}^{H} T_{yh}\right) \approx 0}$$

$$Var\left(\hat{T}_{y\,post}\right) \approx E\left[Var\left(\hat{T}_{y\,post/n_h,h=1,\ldots,H}\right)\right]$$

$$Var\left(\hat{T}_{y\,post/n_h,h=1,\ldots,H}\right) = \sum_{h=1}^{H} N_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{S_{yh}^2}{n_h}$$

$$Var\left(\hat{T}_{y\,post}\right) = E\left[\sum_{h=1}^{H} N_h^2 \left(\frac{1}{n_h} - \frac{1}{N_h}\right) S_{yh}^2\right] \approx \sum_{h=1}^{H} N_h \left[N_h E\left(\frac{1}{n_h}\right) - 1\right] S_{yh}^2$$

- Calcul de $E(1/n_h)$ par développement limité en 0 avec

$$\varepsilon = 1 - \frac{n_h}{E(n_h)} \qquad \text{soit} \qquad n_h = (1-\varepsilon)E(n_h) \Leftrightarrow \frac{1}{n_h} = \frac{1}{1-\varepsilon}\frac{1}{E(n_h)}$$

- D'où : $E\left(\frac{1}{n_h}\right) = E\left(\frac{1}{1-\varepsilon}\right)\frac{1}{E(n_h)} \approx E(1+\varepsilon+\varepsilon^2)\frac{1}{E(n_h)}$

- Or : $n_h \to H(n, N, N_h)$ i.e.

$$\begin{cases} E(n_h) = n\dfrac{N_h}{N} \\ Var(n_h) = n\dfrac{N_h}{N}\left(1-\dfrac{N_h}{N}\right)\dfrac{N-n}{N-1} \end{cases}$$

- On en déduit :

$$E\left(\frac{1}{n_h}\right) \approx \frac{N}{nN_h} E\left[1+\left(1-\frac{Nn_h}{nN_h}\right)+\left(1-\frac{Nn_h}{nN_h}\right)^2\right] \approx \frac{N}{nN_h}\left[1+0+\frac{N^2 Var(n_h)}{n^2 N_h^2}\right]$$

$$\approx \frac{N}{nN_h} + \frac{N(N-N_h)}{N_h^2}\frac{(N-n)}{n^2(N-1)}$$

- On a : $$Var\left(\hat{T}_{y\,post}\right) \approx \sum_{h=1}^{H} N_h \left[N_h E\left(\frac{1}{n_h}\right) - 1 \right] S_{yh}^2$$

avec $$E\left(\frac{1}{n_h}\right) \approx \frac{N}{nN_h} + \frac{N(N-N_h)}{N_h^2} \frac{(N-n)}{n^2(N-1)}$$

D'où la variance d'échantillonnage :

$$Var\left(\hat{T}_{y\,post}\right) \approx N^2 \left(1 - \frac{n}{N}\right) \frac{1}{n} \sum_{h=1}^{H} \frac{N_h}{N} S_{yh}^2 + N^2 \frac{N-n}{N-1} \frac{1}{n^2} \sum_{h=1}^{H} \left(1 - \frac{N_h}{N}\right) S_{yh}^2$$

Qu'on estime par :

$$\hat{Var}\left(\hat{T}_{y\,post}\right) \approx N^2 \left(1 - \frac{n}{N}\right) \frac{1}{n} \sum_{h=1}^{H} \frac{N_h}{N} \hat{S}_{yh}^2 + N^2 \frac{N-n}{N-1} \frac{1}{n^2} \sum_{h=1}^{H} \left(1 - \frac{N_h}{N}\right) \hat{S}_{yh}^2$$

Comparaisons

Avec un plan stratifié et des allocations proportionnelles

$$Var\left[\hat{T}_{y\,prop}\right] = N^2\left(1-\frac{n}{N}\right)\frac{1}{n}\sum_{h=1}^{H}\frac{N_h}{N}S_{yh}^2$$

$$Var\left(\hat{T}_{y\,post}\right) \approx N^2\left(1-\frac{n}{N}\right)\frac{1}{n}\sum_{h=1}^{H}\frac{N_h}{N}S_{yh}^2 + N^2\frac{N-n}{N-1}\frac{1}{n^2}\sum_{h=1}^{H}\left(1-\frac{N_h}{N}\right)S_{yh}^2$$

$$\frac{Var\left(\hat{T}_{y\,post}\right)}{Var\left(\hat{T}_{y\,prop}\right)} = 1+O\left(\frac{1}{n}\right)$$

En résumé: post-stratification

- Il vaut toujours mieux stratifier a priori que post-stratifier, mais lorsque que stratifier a priori n'est pas possible, la post-stratification peut être intéressante
- **Pour avoir une bonne post-stratification**
  - Variable auxiliaire bien corrélée avec Y
  - n grand
  - Grandes post-strates i.e. $(N-N_h)/N$ petit
  - Effectifs N_h ou poids des post-strates connus
- **Remarques :**
  - Les effectifs des post-strates dans l'échantillon ne sont connus qu'après enquête: ce sont des variables aléatoires
  - Mieux vaut éviter qu'ils soient nuls!

4. Redresser sur plusieurs critères

- Se caler sur les marges de plusieurs variables auxiliaires : age, PCS, sexe... mais pas seulement du socio-démo
- Fréquent dans les sondages par quota
- Nécessité d’algorithmes pour obtenir les poids de redressement

Le ratissage ou « raking-ratio »

– exemple à deux variables : sexe et PCS

1000 individus ont été interrogés. La répartition de l'échantillon par sexe et profession est la suivante

	P1	*P2*	*P3*	*Total*
H	300	100	200	*600*
F	100	150	150	*400*
Total	*400*	*250*	*350*	*1000*

Vraies marges 500 et 500 pour le sexe et 350,300, 350 pour la profession.

– Etape n°1: deux règles de 3 pour se caler en ligne
  - multiplier par 5/6 la première ligne
  - multiplier par 5/4 la deuxième ligne

	P1	*P2*	*P3*	*Total*
H	**250**	**83**	**167**	***500***
F	**125**	**187.5**	**187.5**	***500***
Total	***375***	***270.5***	***354.5***	***1000***

« Ratissage en ligne »

– Etape n°2: trois règles de 3 pour se caler en colonne

	P1	*P2*	*P3*	*Total*
H	**233**	**92**	**165**	***490***
F	**117**	**208**	**185**	***510***
Total	***350***	***300***	***350***	***1000***

« Ratissage en colonne»

– détruit le premier calage!

- on itère

- Convergence rapide:
  - étape n°4

	P1	***P2***	***P3***	**Total**
H	**236**	**95**	**168**	**499**
F	**114**	**205**	**182**	**501**
Total	***350***	***300***	***350***	**1000**

poids de redressement: les 300 H & P1 comptent pour 236.
Chacun a donc un poids de 0.0079 au lieu de 0.001 , et donc un **coefficient de redressement** de 0.79 (le minimum).
Le coefficient maximum est 1.37

5. Généralisation : calage sur marges

Objectifs :

- Améliorer la précision des estimateurs des paramètres d'intérêt d'une enquête
  - Pourvu que les critères de calage soient liés aux variables d'intérêt
- Assurer la cohérence des résultats avec des informations synthétiques connues par ailleurs. Ainsi, après calage, l'échantillon restitue :
  - les totaux de variables quantitatives connus sur la population
  - les effectifs de modalités de variables catégorielles connus sur la population

Principe

- Re-pondérer les individus échantillonnés en utilisant une information auxiliaire disponible sur un certain nombre de variables, appelées variables de calage
- Cas particuliers : les estimateurs par le ratio, par la régression, par le raking-ratio

Méthode

- Supposons connus les totaux sur la population de *J* variables auxiliaires $T_X = \left(T_{x_1},...,T_{x_j},...,T_{x_J}\right)$
  - Pour les caractères catégoriels, les totaux sont les effectifs de chaque modalité (= totaux des variables indicatrices associées à ces modalités)
- On va tenir compte de cette information pour améliorer l'estimateur d'Horvitz-Thompson

$$\hat{T}_{y\pi} = \sum_{k\in S} \frac{y_k}{\pi_k} = \sum_{k\in S} d_k y_k \qquad (d_k = \frac{1}{\pi_k} = \text{poids de sondage})$$

- En formant un nouvel estimateur

  où les nouveaux poids à rechercher : $\hat{T}_{y,calé} = \sum_{k\in S} w_k y_k$

  - sont « proches » des poids initiaux
  - vérifient les équations de calage : $\hat{T}_{x_{j,calé}} = \sum_{k\in S} w_k x_{j,k} = T_{x_j} \quad \forall \text{j}$

- On choisit une fonction de distance entre le poids initial et le poids final : $G(w_k, d_k)$

- Les poids cherchés sont solutions du problème d'optimisation :

$$\min_{w_k} \sum_{k \in S} w_k G\left[\frac{w_k}{d_k}\right] \quad \text{avec} \quad \sum_{k \in S} w_k x_{j,k} = T_{x_j} \quad \forall \text{j}$$

- Résolution du système non linéaire $\sum_{k \in S} d_k F\left(x'_k \lambda\right) = \mathrm{T_X}$
  - où F est la fonction réciproque de la dérivée de la fonction G
  - et λ un vecteur de multiplicateurs de Lagrange

- Ce système d'équations peut être résolu par la méthode itérative de Newton

- En pratique, macro SAS CALMAR de l'Insee

fonctions de distance

G	$F = G'^{-1}$	Type de distance
$\frac{1}{2}(x-1)^2$	$1+u$	*Khi-deux* Méthode linéaire (1) i.e. estimateur par la régression
$x\log x - x + 1$	$\exp u$	Entropie Méthode du raking -ratio (2)
$\frac{1}{A}\left[(x-L)\log\left(\frac{x-L}{1-L}\right)+(U-x)\log\left(\frac{U-x}{U-1}\right)\right]$ $= \frac{U-L}{(1-L)(U-1)}$; $x \in [L,U]$, $(\infty \sin on)$	$\frac{L(U-1)+U(1-L)\exp u}{(U-1)+(1-L)\exp u}$ $\in]L,U[$	Logistique Méthode du raking ratio tronquée (3)
$\frac{1}{2}(x-1)^2$ si $x \in [L,U]$ ∞ sinon	$1+q_i u$ $\in [L,U]$	*Khi-deux tronquée* Méthode linéaire tronquée (3)

- Méthode linéaire
  - converge toujours en 2 étapes
  - redonne l'estimateur par régression
  - peut donner des poids négatifs
  - rapports de poids non bornés supérieurement
- Méthode exponentielle
  - poids positifs
  - redonne l'estimateur du raking-ratio
  - rapports de poids non bornés supérieurement, en général supérieurs à la méthode linéaire
- Méthodes logit, linéaire tronquée
  - poids positifs
  - contrôle des rapports de poids

Propriétés

- Espérance

  Quelle que soit la méthode utilisée, l'estimateur calé est approximativement sans biais

- 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

Calage sur marges
macro CALMAR

- Insee, 1993
- Macro SAS
- Disponible sur www.insee.fr
- Syntaxe *(paramètres obligatoires)*

```
%CALMAR (data    =,
         poids   =,
         ident   =,
         datamar =,
         M =,          LO=,    UP=,
         datapoi =,
         poidsfin=);
```

Calage sur marges exemple

- **1. les données individuelles**

```
DATA echant;
INPUT nom $ x $ y $ z pond;
CARDS;
A  1  f   1   10
B  1  h   2    0
C  1  h   3    .
D  5  f   1   11
E  5  f   3   13
F  5  h   2    7
H  1  h   2    8
G  5  h   2    8
I  5  f   2    9
J  .  h   2   10
K  5  h   2   14
;
RUN;
```

- **2. la table des marges**

```
DATA marges;
INPUT var $ n mar1 mar2;
CARDS;
X  2    20    60
Y  2    30    50
Z  0   140     .
;
RUN ;
```

- **3. lancement de Calmar**

```
%CALMAR(DATA    = echant,POIDS = pond,
        IDENT   = nom,
        DATAMAR = marges,
        M       = 2,    OBSELI = oui,
        DATAPOI = sortie,
        POIDSFIN = pondfin,
       LABELPOI = poids raking ratio);
```

- Avant calage

VARIABLE	MODALITÉ	MARGE ÉCHANTILLON	MARGE POPULATION	POURCENTAGE ÉCHANTILLON	POURCENTAGE POPULATION
X	1	18	20	22.50	25.00
	5	62	60	77.50	75.00
Y	f	43	30	53.75	37.50
	h	37	50	46.25	62.50
Z		152	140	.	.

- Après calage

Variable	Modalité	Marge échantillon	Marge population	Pourcentage échantillon	Pourcentage population
X	1	20.000	20	25.00	25.00
	5	60.000	60	75.00	75.00
Y	f	30.000	30	37.50	37.50
	h	50.000	50	62.50	62.50
Z		140.000	140	.	.

Méthode : raking ratio
Premier tableau récapitulatif de l'algorithme :
la valeur du critère d'arrêt et le nombre de poids négatifs après chaque itération

Itération	Critère d'arrêt	Poids négatifs
1	0.56651	0
2	0.17766	0
3	0.04198	0
4	0.00322	0
5	0.00002	0

Un petit exemple commenté de calage sur marges

Méthode : raking ratio
Deuxième tableau récapitulatif de l'algorithme :
les coefficients du vecteur lambda de multiplicateurs de Lagrange après chaque itération

Variable	Modalité	LAMBDA1	LAMBDA2	LAMBDA3	LAMBDA4	LAMBDA5
X	1	1.20511	1.70361	1.87331	1.88687	1.88695
X	5	1.32247	1.81959	1.99270	2.00648	2.00656
Y	f	-0.73974	-0.94297	-1.02331	-1.02984	-1.02987
Y	h	.	.	.	.	.
Z		-0.47287	-0.74661	-0.83348	-0.84035	-0.84039

Un petit exemple commenté de calage sur marges

Méthode : raking ratio
Statistiques sur les rapports de poids (= pondérations finales / pondérations initiales)
et sur les pondérations finales

Univariate Procedure

Variable=_F_ Rapport de poids

Moments

N	8	Sum Wgts	8
Mean	1.031891	Sum	8.255131
Std Dev	0.444812	Variance	0.197858
Skewness	-1.21649	Kurtosis	0.18399
USS	9.903406	CSS	1.385006
CV	43.10651	Std Mean	0.157265
T:Mean=0	6.561485	Pr>|T|	0.0003
Num ¬ =0	8	Num > 0	8
M(Sign)	4	Pr>=|M|	0.0078
Sgn Rank	18	Pr>=|S|	0.0078
W:Normal	0.811594	Pr<W	0.0394

Quantiles(Def=5)

100% Max	1.385113	99%	1.385113
75% Q3	1.385113	95%	1.385113
50% Med	1.187493	90%	1.385113
25% Q1	0.755692	10%	0.213423
0% Min	0.213423	5%	0.213423
		1%	0.213423
Range	1.17169		
Q3-Q1	0.629421		
Mode	1.385113		

Extremes

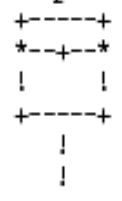
Lowest	ID	Highest	ID
0.213423	(E)	1.14602	(D)
0.494557	(I)	1.228966	(H)
1.016827	(A)	1.385113	(F)
1.14602	(D)	1.385113	(G)
1.228966	(H)	1.385113	(K)

Stem Leaf

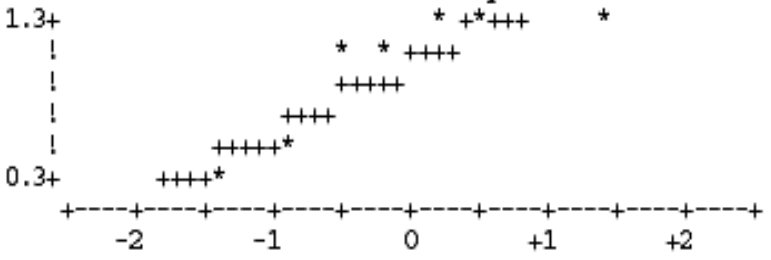
```
  12 3999                  4
  10 25                    2
   8                       
   6                       
   4 9                     1
   2 1                     1
     ----+----+----+----+
```

Multiply Stem.Leaf by 10**-1

Boxplot



Normal Probability Plot



Méthode : raking ratio
Statistiques sur les rapports de poids (= pondérations finales / pondérations initiales)
et sur les pondérations finales

Univariate Procedure

Variable=__WFIN Pondération finale

Moments

N	8	Sum Wgts	8
Mean	10	Sum	80
Std Dev	5.061209	Variance	25.61584
Skewness	0.439584	Kurtosis	1.109319
USS	979.3109	CSS	179.3109
CV	50.61209	Std Mean	1.789408
T:Mean=0	5.588441	Pr>|T|	0.0008
Num ¬ =0	8	Num > 0	8
M(Sign)	4	Pr>=|M|	0.0078
Sgn Rank	18	Pr>=|S|	0.0078
W:Normal	0.926636	Pr<W	0.4908

Quantiles (Def=5)

100% Max	19.39158	99%	19.39158
75% Q3	11.84356	95%	19.39158
50% Med	10	90%	19.39158
25% Q1	7.073401	10%	2.774494
0% Min	2.774494	5%	2.774494
		1%	2.774494
Range	16.61709		
Q3-Q1	4.770161		
Mode	2.774494		

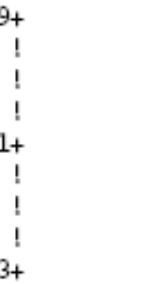
Extremes

Lowest	ID	Highest	ID
2.774494	(E)	9.831729	(H)
4.451013	(I)	10.16827	(A)
9.695789	(F)	11.0809	(G)
9.831729	(H)	12.60622	(D)
10.16827	(A)	19.39158	(K)

Stem Leaf

```
  18 4                    1
  16
  14
  12 6                    1
  10 21                   2
   8 78                   2
   6
   4 5                    1
   2 8                    1
     ----+----+----+----+
```

Boxplot



Normal Probability Plot

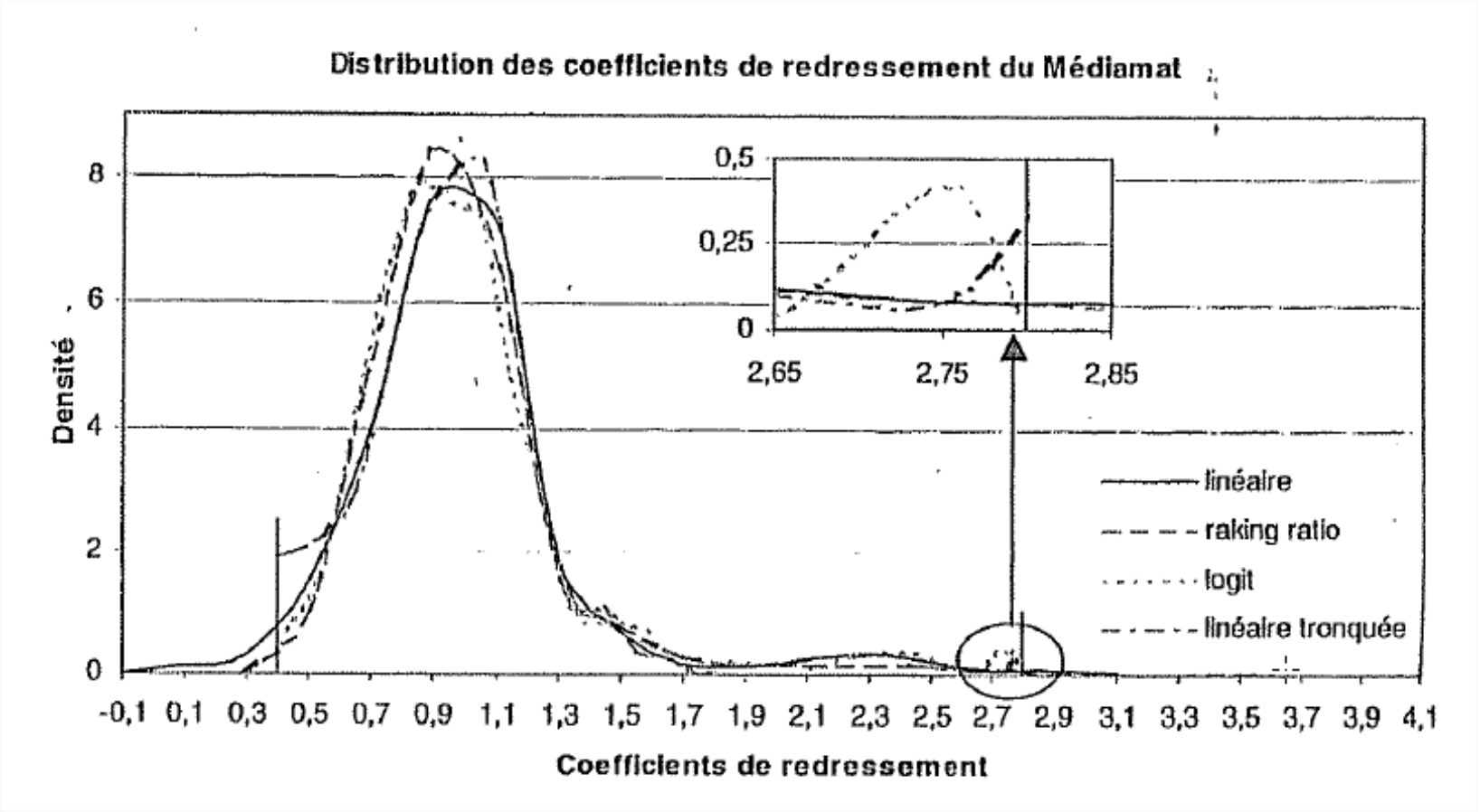
```
                ********************
                ***    BILAN     ***
                ********************
*
*    DATE : 16 JUIN 2000               HEURE : 14:03
*
*    ***********************************
*    TABLE EN ENTRÉE : DON
*    ***********************************
*
*    NOMBRE D'OBSERVATIONS DANS LA TABLE EN ENTRÉE  : 11
*    NOMBRE D'OBSERVATIONS ÉLIMINÉES                : 3
*    NOMBRE D'OBSERVATIONS CONSERVÉES               : 8
*
*    VARIABLE DE PONDÉRATION : POND
*
*    NOMBRE DE VARIABLES CATÉGORIELLES : 2
*    LISTE DES VARIABLES CATÉGORIELLES ET DE LEURS NOMBRES DE MODALITÉS :
       X (2) y (2)
*
*    TAILLE DE L'ÉCHANTILLON (PONDÉRÉ) :              80
*    TAILLE DE LA POPULATION           :              80
*
*    NOMBRE DE VARIABLES NUMÉRIQUES : 1
*    LISTE DES VARIABLES NUMÉRIQUES :
       z
*
*
*    MÉTHODE UTILISÉE : RAKING RATIO
*    LE CALAGE A ÉTÉ RÉALISÉ EN 5 ITÉRATIONS
*    LES POIDS ONT ÉTÉ STOCKÉS DANS LA VARIABLE PONDFIN DE LA TABLE SORTIE
```

Un petit exemple commenté de calage sur marges

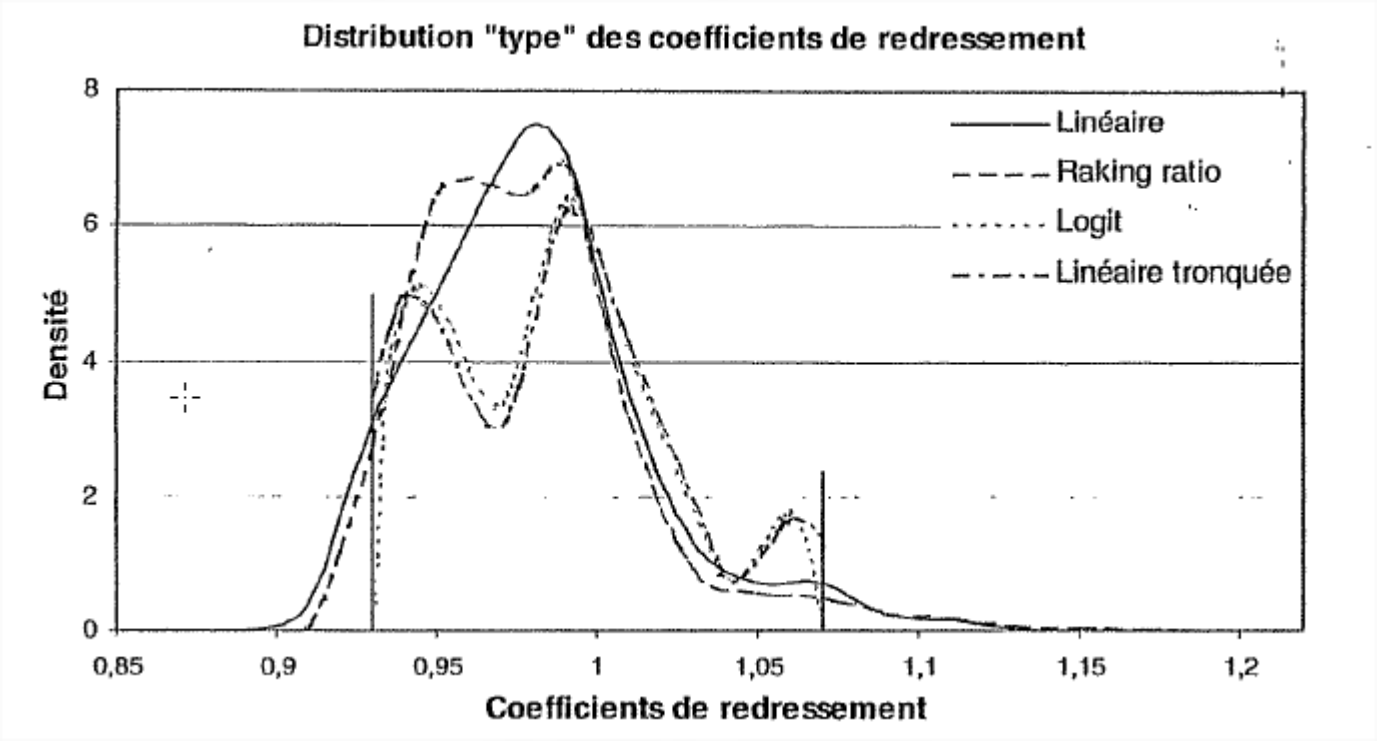
Liste des observations éliminées

Obs	nom	X	y	z	pond	__UN
1	B	1	h	2	0	1
2	C	1	h	3	.	1
3	J		h	2	10	1

- Médiamétrie
  - Redressement Médiamat



– 75000 Radio, redressement sur activité, age et type d’habitat



Références

- Deming, W.E. & Stephan, F.F. (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *Ann. Math. Statist.*, 11, 427–444.
- Deville, J.-C., Särndal, C.-E. et Sautory, O. (1993). « Generalized raking procedures in survey sampling », Journal of the American Statistical Association, vol 88, n°423, pp. 1013-1020.
- Deville, J.C. & Särndal, C.E. (1992). Calibration estimators in survey sampling. *J. Amer. Statist. Assoc.*, 87, 376–382
- Deville, J.C., Särndal, C.E. & Sautory, O. (1993). Generalized raking procedures in survey sampling. *J. Amer. Statist. Assoc.*, 88, 1013–1020
- Dupont, F. (1996). « Calage et redressement de la non-réponse totale ». Actes des journées de méthodologie statistique, 15 et 16 décembre 1993, INSEE-Méthodes n°56-57-58.
- Roy, G., et Vanheuverzwyn, A. (2001). « Redressement par la macro CALMAR : applications et pistes d'amélioration », Traitements des fichiers d'enquête, pp. 31-46. Presses Universitaires de Grenoble.
- Sautory O. (1993). « Redressement d'un échantillon par calage sur marges », Document de travail de la DSDS n°F9310,, www.insee.fr