

## CALAGE SUR MARGES ALÉATOIRES : UNE AVENTURE HASARDEUSE ?

Maxime Vitot<sup>1</sup>

### RÉSUMÉ

Quel que soit le domaine d'application des sondages, le redressement d'échantillon est devenu depuis longtemps pratique courante, en France comme ailleurs. Il vise à mettre à profit toute l'information auxiliaire disponible à l'étape finale de l'estimation, afin d'améliorer les résultats. Plus concrètement, il aide surtout à crédibiliser les instituts, qui peuvent présenter des estimations cohérentes avec les données que le marché connaît par ailleurs. Utiliser à bon escient l'information auxiliaire est censé améliorer la qualité des estimations, sauf si cette information provient elle-même d'une enquête par sondage de trop faible taille, auquel cas il peut s'avérer que la précision des résultats en soit si altérée qu'il soit éventuellement préférable de ne pas effectuer de redressement. L'objet de cette communication est de rappeler l'ampleur de cette perte de précision pour plusieurs estimateurs de redressement simples, et de fournir des estimations bootstrap sur des exemples, notamment pour le cas du calage sur marges.

### 1. INTRODUCTION

La mise en œuvre d'un sondage est une opération certes courante aujourd'hui mais néanmoins complexe à mener à bien. La méthodologie d'enquête comporte en effet de nombreuses étapes qu'il est indispensable de réaliser avec soin et rigueur si l'on veut avoir une chance d'obtenir des résultats fiables. De plus chaque situation a ses particularités et ses écueils spécifiques (cf. par exemple Verger, 2004). Malgré les efforts fournis lors de la réalisation de l'enquête, on va observer en pratique un certain nombre de biais plus ou moins importants sur les variables pour lesquelles on dispose de données externes de référence (recensements, autres enquêtes). Ces écarts résultent de diverses causes comme les fluctuations d'échantillonnage ou la non réponse<sup>2</sup>. Parce qu'il peut être gênant pour l'institut de produire des estimations différentes de celles dont dispose le marché, ce dernier va employer la plupart du temps une technique de redressement afin de caler les structures de son échantillon sur ces sources externes. Pour le cas des études en continu (notamment les panels), le redressement permet aux instituts de stabiliser les tailles de cibles.

Cette procédure est ainsi censée améliorer la précision des estimations, pour les variables d'intérêt fortement corrélées aux variables redressées.

Même si l'on sait qu'il est totalement illusoire de croire que l'on va éliminer tous les biais en redressant un échantillon sur un certain nombre de variables, il est en général admis qu'un redressement améliore la précision des résultats (à condition que celui-ci soit réalisé dans de bonnes conditions : identité des concepts observés et théoriques, coïncidence temporelle, proximité des structures par rapport à celles de référence, corrélation forte entre les variables de redressement et les variables d'intérêt, etc.). La confiance accordée au redressement repose d'abord sur le fait que l'on introduit dans les estimations une information exhaustive, *a priori* non entachée d'erreur. Lorsque cette donnée provient elle-même d'une enquête par sondage, le redressement introduit dans les résultats une incertitude supplémentaire correspondant à la variabilité de l'information auxiliaire. Ce cas de figure se produit en fait très souvent : il est en effet fréquent de sonder une population bien spécifique pour laquelle on ne dispose pas

---

<sup>1</sup> CESP (Centre d'Étude des Supports de Publicité), 136 boulevard Haussmann, 75008 Paris, France ([mvitot@cesp.org](mailto:mvitot@cesp.org)).

<sup>2</sup> A noter que l'institut a pu mettre en œuvre un plan de sondage à probabilités inégales pour sur-représenter délibérément certaines catégories de population intéressantes vis-à-vis des variables à étudier. Les biais évoqués ici sont observés après correction des taux de sondage différenciés, i.e. après rétablissement des structures déformées volontairement.

des structures de cadrage adéquates dans les recensements. Un terme a même été créé pour l'occasion : on parle « d'enquête de calage ou de cadrage » pour dénommer la source auxiliaire.

Or on constate que *l'incertitude* qui lui est associée est presque automatiquement ignorée en pratique (cf. par exemple Lebart, 2004). Cela paraît sans doute légitime si l'enquête de calage est de grande taille devant celle à redresser : les préconisations habituelles sont d'utiliser une enquête comme source auxiliaire lorsque sa taille dépasse 10 fois celle de l'enquête à caler (cf. par exemple Caron, 2002). On en est bien loin dans de nombreux cas ; il arrive même souvent que l'on redresse l'échantillon à partir d'une enquête de taille bien inférieure. Dans cette situation, on va préférer au *redressement* le terme d'*ajustement*. On se trouve ainsi dans un cas de « dérive » par rapport au cadre idéal d'application de la théorie des sondages (cf. Rancourt, 2001 pour un point sur tous ces comportements).

Sans vouloir y mettre un frein, il semble néanmoins raisonnable d'évaluer systématiquement la perte de précision de résultats *ajustés* par rapport à des résultats *redressés*, afin de savoir s'il est préférable de ne pas effectuer l'ajustement : alors que le redressement est censé améliorer la fiabilité des résultats, il se peut en effet que la précision de l'estimateur *ajusté* soit plus faible que celle du  $\pi$ -estimateur. Cet exposé s'attache à mettre en valeur ce phénomène par quelques illustrations.

## 2. LE REDRESSEMENT D'UNE ENQUÊTE SUR UNE AUTRE

### 2.1 Rappel de la théorie du redressement

Il existe plusieurs méthodes de redressement, qui font toutes partie du cadre général de la théorie du calage<sup>3</sup>. Pour fonctionner, cette théorie fait intervenir le choix d'une fonction de distance entre les poids de sondage et les poids de calage. Cinq fonctions ont retenu l'attention des statisticiens : la méthode *linéaire* (régression généralisée) dont les cas particuliers sont l'estimateur par la régression, par le quotient et par la différence ; la méthode *exponentielle* (raking-ratio) dont le calage sur marges et la post-stratification<sup>4</sup> sont des cas particuliers ; leurs variantes tronquées i.e. les méthodes *linéaire tronquée* et *exponentielle tronquée* (méthode logit ou logistique) qui permettent un bornage des poids ; et dernièrement la méthode du *sinus hyperbolique* (cf. Roy *et al.*, 2001).

Dans le cas où l'information auxiliaire est issue d'un sondage<sup>5</sup>, examinons ce que donnent les estimateurs de redressement classiques par la différence, par le quotient et par la régression, puis l'estimateur post-stratifié et enfin le cas du calage sur marges (aléatoires). Pour simplifier, on se place dans l'hypothèse d'un échantillon  $S$  (de taille  $n$ ) et d'une enquête de calage  $S_0$  (de taille  $n_0$ ) indépendants et tirés par sondage aléatoire simple dans une même population  $U$ , et on s'intéresse à l'estimation d'une moyenne.

Soient  $\hat{y}_\pi$  et  $\hat{x}_\pi$  les estimateurs d'Horvitz-Thompson des moyennes  $\bar{y}$  et  $\bar{x}$ , où  $y$  est la variable d'intérêt et  $x$  la variable auxiliaire. On distingue l'estimateur *redressé*  $\hat{y}(S)$  calculé en supposant  $\bar{x}$  connu, et l'estimateur *ajusté*  $\tilde{y}(S, S_0)$  obtenu quand  $\bar{x}$  est estimé par  $\hat{x}_\pi(S_0)$ . On note  $CV$  le coefficient de variation,  $S_x^2$ ,  $S_y^2$  et  $S_{xy}$  respectivement les variances et covariances entre  $x$  et  $y$ , on appelle  $\rho$  leur coefficient de corrélation, et on pose  $R = \bar{y}/\bar{x}$  et  $\gamma = n/n_0$ . On remarque (cf. Tableau 1 page suivante) que la variance de l'estimateur *ajusté*  $Var[\tilde{y}(S, S_0)]$  s'obtient en ajoutant à l'expression de la variance de l'estimateur *redressé* (où l'on considère  $\bar{x}$  connu) un terme positif dépendant de l'estimateur choisi<sup>6</sup>. Les conditions dans lesquelles la variance de l'estimateur *ajusté* est plus grande que celle du  $\pi$ -estimateur ( $Var[\hat{y}_\pi(S)] = S_y^2/n$ ) se lisent directement dans la colonne de droite.

Le calage sur marges est la méthode la plus utilisée : elle est employée lorsque l'information auxiliaire est disponible sous forme de plusieurs variables qualitatives non croisées (critères de redressement marginaux). Le cas qui nous intéresse (où l'information auxiliaire est estimée par enquête) peut être qualifié de calage sur marges aléatoires.

<sup>3</sup> Voir par exemple Tillé (2001) pour un cours clair et récent sur les méthodes de redressement dans le cas où l'information auxiliaire est issue d'un recensement.

<sup>4</sup> Post-stratification = calage sur une seule marge.

<sup>5</sup> Remarquons que les poids de calage sont ici *doublement* aléatoires (fonctions de  $S$  et  $S_0$ ).

<sup>6</sup> On retrouve les variances des estimateurs *redressés* en prenant  $\gamma = 0$  (c'est-à-dire quand  $n$  est négligeable devant  $n_0$ ).

L'expression de la variance de l'estimateur par calage sur marges étant assez délicate (voir par exemple Tillé, 2001), seules des estimations bootstrap ont été calculées sur des exemples (partie 2.2).

**Tableau 1 – Quelques estimateurs ajustés et leurs variances**

	Estimateur $\tilde{y}(S, S_0)$	Variance <sup>7</sup> $Var[\tilde{y}(S, S_0)]$	Conditions pour que : $Var[\tilde{y}(S, S_0)] > Var[\hat{y}_\pi(S)]$
<b>Différence</b>	$\hat{y}_\pi(S) + \hat{x}_\pi(S_0) - \hat{x}_\pi(S)$	$Var[\hat{y}_{DIFF}(S)] + Var[\hat{x}_\pi(S_0)]$ $= \frac{S_y^2 + S_x^2(1+\gamma) - 2S_{xy}}{n}$	$S_x \cdot (1+\gamma) > 2\rho \cdot S_y$
<b>Quotient</b>	$\hat{y}_\pi(S) \times \frac{\hat{x}_\pi(S_0)}{\hat{x}_\pi(S)}$	$Var[\hat{y}_{QUOT}(S)] + R^2 \times Var[\hat{x}_\pi(S_0)]$ $= \frac{S_y^2 + R^2 S_x^2(1+\gamma) - 2.R.S_{xy}}{n}$	$CV_x \cdot (1+\gamma) > 2\rho \cdot CV_y$
<b>Régression</b>	$\hat{y}_\pi(S) + \frac{S_{xy}}{S_x^2} \times [\hat{x}_\pi(S_0) - \hat{x}_\pi(S)]$	$Var[\hat{y}_{REGR}(S)] + \rho^2 \cdot \gamma \times Var[\hat{y}_\pi(S)]$ $= \frac{S_y^2}{n} [1 - \rho^2(1-\gamma)]$	$\gamma > 1$ (si $\rho \neq 0$ )
<b>Post-stratification</b>	$\sum_{h=1}^H \frac{\hat{N}_h}{N} \hat{y}_{sh}(S) = \sum_{h=1}^H \frac{n_{0h}}{n_0} \hat{y}_{sh}(S)$	$Var[\hat{y}_{POST}(S)] + \frac{1}{n_0} \sum_{h=1}^H \frac{N_h}{N} (\bar{y}_h - \bar{y})^2$ $\approx \frac{1}{n} \sum_{h=1}^H \frac{N_h}{N} S_{yh}^2 + \frac{1}{n_0} \sum_{h=1}^H \frac{N_h}{N} (\bar{y}_h - \bar{y})^2$	$\gamma > 1$

## 2.2 Exemples numériques

### Structure du programme informatique utilisé pour le calcul des variances par bootstrap

$$\text{Répéter pour } b = 1 \text{ à } B \left\{ \begin{array}{l} \bullet S_0 \rightarrow S_0^{(sb)} : \text{réplication bootstrap de l'enquête de calage} \\ \bullet S \rightarrow S^{(sb)} : \text{réplication bootstrap de l'échantillon} \\ \bullet \text{Redressement (ajustement) de } S^{(sb)} \text{ sur } S_0^{(sb)} \Rightarrow \text{poids de calage } w_k(S^{(sb)}, S_0^{(sb)}) \\ \bullet \text{Calcul de la moyenne d'intérêt : } \tilde{y}^{(sb)} = \sum_{k \in S^{(sb)}} w_k(S^{(sb)}, S_0^{(sb)}) \cdot y_k \end{array} \right.$$

$\Rightarrow$  La variance  $Var[\tilde{y}(S, S_0)]$  est estimée par la variance des  $\tilde{y}^{(sb)}$ .

Ce même programme a été utilisé dans chacun des exemples qui suivent (seule la méthode d'ajustement change selon les différents estimateurs).

### Exemple 1 : Estimateurs par différence, quotient et régression (B = 10000 réplifications)

<sup>7</sup> Voir par exemple l'ouvrage de Desabie (1966) pour le calcul des variances lorsque l'information auxiliaire est issue d'un sondage indépendant. Voir aussi Hidiroglou (2001) pour une présentation du cadre général de l'échantillonnage double. Les formules de variance du Tableau 1 se retrouvent également par exemple en conditionnant par rapport à  $S_0$  et en utilisant le théorème de la variance totale. Notons que les formules de variance présentées ici sont des approximations valables si l'on suppose  $n$  assez grand. On néglige en outre ici le taux de sondage  $f$  (la population est supposée grande devant les échantillons  $S_0$  et  $S$ ). Il n'apparaît donc pas de terme correctif de population finie.

Le Tableau 2 page suivante présente les estimations des moyennes et variances de  $\tilde{y}(S, S_0)$  obtenues « par formule » et par bootstrap<sup>8</sup>, pour le cas du redressement<sup>9</sup> d'une part et de l'ajustement d'autre part, en faisant varier la taille  $n_0$  de l'enquête de calage. La colonne de droite donne le rapport entre la variance de l'estimateur *ajusté* et celle du  $\pi$ -estimateur.

Pour cet exemple on étudie un échantillon de taille  $n = 1000$  tel que  $\hat{y}_\pi(S) = 0,46$ ,  $Var[\hat{y}_\pi(S)] = 25.10^{-5}$  et  $\rho = 0,53$ .

**Tableau 2 – Estimateurs par différence, quotient et régression**

$n_0$	Estimateur	Redressement				Ajustement				Var(ajust.) / Var(pi-est)
		Par formule		Par bootstrap		Par formule		Par bootstrap		
		Moyenne	Variance x 10 <sup>5</sup>	Moyenne	Variance x 10 <sup>5</sup>	Moyenne	Variance x 10 <sup>5</sup>	Moyenne	Variance x 10 <sup>5</sup>	
10000	Différence	0,46	23	0,46	23	0,46	26	0,46	26	1,0
	Quotient	0,46	22	0,46	21	0,46	24	0,46	24	1,0
	Régression	0,46	18	0,46	18	0,46	18	0,46	19	0,7
1000	Différence	0,45	23	0,45	24	0,45	48	0,45	48	1,9
	Quotient	0,45	22	0,45	21	0,45	43	0,45	42	1,7
	Régression	0,46	18	0,46	18	0,46	25	0,46	25	1,0
200	Différence	0,45	23	0,44	23	0,45	149	0,44	148	6,0
	Quotient	0,45	22	0,45	20	0,45	130	0,45	128	5,2
	Régression	0,45	18	0,45	18	0,45	53	0,45	53	2,1

Ce tableau illustre numériquement les formules présentées au Tableau 1. On observe ainsi dans cet exemple qu'une enquête de calage de taille 10 fois supérieure à celle de l'échantillon fournit une variance d'estimateur *ajusté* à peu près équivalente à celle de l'estimateur *redressé* (qui a lui-même une variance légèrement inférieure à celle du  $\pi$ -estimateur). Il n'en est pas de même lorsque  $n_0$  est inférieur ou égal à  $n$  : les précisions obtenues sont alors systématiquement plus faibles que celle du  $\pi$ -estimateur. Ainsi par exemple une taille cinq fois plus petite pour  $n_0$  entraîne ici un doublement de la variance pour l'estimateur ajusté par régression.

**Exemple 2 : Post-stratification (B = 10000 répliques)**

Pour cet exemple on post-stratifie un échantillon de taille  $n = 1000$  tel que  $\hat{y}_\pi(S) = 0,79$  et  $Var[\hat{y}_\pi(S)] = 17.10^{-5}$ , en faisant varier la taille  $n_0$  de l'enquête de calage.

**Tableau 3 – Post-stratification**

$n_0$	Redressement				Ajustement				Var(ajust.) / Var(pi-est)
	Par formule		Par bootstrap		Par formule		Par bootstrap		
	Moyenne	Variance x 10 <sup>5</sup>	Moyenne	Variance x 10 <sup>5</sup>	Moyenne	Variance x 10 <sup>5</sup>	Moyenne	Variance x 10 <sup>5</sup>	
10000	0,77	15	0,77	15	0,77	15	0,77	16	0,9
2000	0,76	15	0,76	17	0,76	17	0,76	18	1,0
1000	0,77	15	0,77	16	0,77	18	0,77	20	1,1
500	0,76	15	0,76	17	0,76	22	0,76	24	1,3
200	0,76	15	0,76	17	0,76	32	0,76	33	1,9

A nouveau, on observe une croissance de la variance à mesure que  $n_0$  diminue. Une taille cinq fois plus petite pour  $n_0$  entraîne aussi un doublement de la variance pour l'estimateur post-stratifié.

La méthode de rééchantillonnage par bootstrap s'étant avérée fructueuse<sup>10</sup> pour estimer les variances des estimateurs par différence, quotient, régression ainsi que pour l'estimateur post-stratifié, elle a également été mise en œuvre dans le cas du calage sur marges.

**Exemple 3 : Calage sur marges (B = 10000 répliques)**

<sup>8</sup> Cf. par exemple Ardilly (1994) pour une introduction à la méthode du bootstrap appliqué aux sondages.

<sup>9</sup> Calculs par bootstrap effectués dans ce cas sans rééchantillonner  $S_0$  i.e. en supposant ses marges fixées.

<sup>10</sup> Voir dans les Tableaux 2 et 3 la comparaison des résultats obtenus par bootstrap avec ceux provenant des formules.

On reprend le même exemple que pour la post-stratification mais en l'ajustant cette fois-ci respectivement sur 2 marges et sur 4 marges. Là encore on observe le même « envol » de la variance.

**Tableau 4 – Calage sur marges**

$n_0$	Calage sur 2 marges				Var(ajust) / Var(pi-est)	Calage sur 4 marges				Var(ajust) / Var(pi-est)
	Redressement		Ajustement			Redressement		Ajustement		
	Moyenne	Variance $\times 10^5$	Moyenne	Variance $\times 10^5$		Moyenne	Variance $\times 10^5$	Moyenne	Variance $\times 10^5$	
10000	0,77	16	0,77	16	1,0	0,77	15	0,77	16	0,9
2000	0,76	17	0,76	19	1,1	0,76	16	0,76	18	1,1
1000	0,77	16	0,77	20	1,2	0,77	15	0,77	19	1,2
500	0,76	17	0,76	24	1,5	0,76	16	0,76	25	1,5
200	0,76	17	0,76	36	2,2	0,75	17	0,75	44	2,7

### 3. CONCLUSION

Si une estimation *ajustée* est certes moins précise qu'une estimation *redressée*, il arrive même parfois qu'elle soit nettement moins précise que l'estimation *brute*. L'ampleur de cette perte de précision va dépendre de chaque situation. Il ressort toutefois qu'une taille d'enquête de calage inférieure à celle de l'échantillon pose problème. Il n'est bien sûr pas question de proscrire les ajustements : instituts et utilisateurs apprécieront toujours de se caler sur les sources externes dont ils disposent. Ils doivent simplement être conscients des conséquences qu'ils encourent.

On ne peut donc que recommander la prudence quant à cette pratique, notamment dans le domaine de la mesure d'audience où le souci de précision et de fiabilité est à la hauteur des enjeux financiers sous-jacents. Paradoxalement, cela n'empêche pas certains utilisateurs peu scrupuleux d'aller parfois au-delà des limites du raisonnable dans l'utilisation des données produites (cf. par exemple Brown, 2001), ni les instituts de contourner la théorie pour résoudre leurs difficultés (techniques ou financières). Il est à noter que l'ajustement d'une enquête sur une autre s'observe de plus en plus souvent (citons par exemple le cas des populations en évolution constante, comme les internautes, ou celui des populations de « niche »). Une réflexion d'ordre méthodologique devrait donc s'imposer dans chaque situation de ce type, afin de juger de la pertinence ou non de l'ajustement (même si cela ne permettra que de fournir une appréciation de l'erreur d'échantillonnage). Pour cela, l'approche par bootstrap paraît intéressante, dans la mesure où elle s'est avérée ici opérationnelle et plutôt satisfaisante (certes dans le cadre de sondages aléatoires simples indépendants). Mentionnons en outre que les moyens informatiques dont on dispose aujourd'hui ont assez facilement permis d'inclure la procédure d'ajustement dans chaque réplification. Il faudrait bien sûr prendre soin d'adapter convenablement la technique aux situations complexes (ajustement avec bornage des poids, plans à probabilités inégales, plans à plusieurs degrés, non réponse, etc.).

### RÉFÉRENCES

- Ardilly, P. (1994), *Les techniques de sondages*, Paris, Technip.
- Brown, M. (2001) « Précision et biais dans la mesure d'audience média : quelques problèmes et quelques solutions », in *Enquêtes, modèles et applications*, Paris, Dunod, p. 242-253.
- Caron, N. (2002), « Les problèmes de calage dans les enquêtes entreprises », présentation aux JMS 2002, Paris.
- Desabie, J. (1966), *Théorie et pratique des sondages*, Paris, Dunod.
- Hidiroglou, M. A. (2001), « Le double échantillonnage », in *Enquêtes, modèles et applications*, Paris, Dunod, p. 317-333
- Lebart, L. (2004), « Impact des nouvelles technologies », in *Echantillonnage et méthodes d'enquêtes*, Paris, Dunod, p. 249-259.
- Rancourt, E. (2001), « La régression étendue : un ensemble de pratiques d'estimation qui poussent constamment la théorie », in *Enquêtes, modèles et applications*, Paris, Dunod, p. 334-343.
- Roy, G. et Vanheuverzwyn, A. (2001), « Redressement par la macro CALMAR : applications et pistes d'amélioration », in *Traitements des fichiers d'enquêtes*, éditions PUG, p. 31-46.

Tillé, Y. (2001), *Théorie des sondages*, Paris, Dunod.

Verger, D. (2004), « La qualité dans les enquêtes auprès des ménages », in *Echantillonnage et méthodes d'enquêtes*, Paris, Dunod, p. 55-66.