# Stratification a posteriori

Myriam Maumy-Bertrand[1]

[1]IRMA, Université de Strasbourg
Strasbourg, France

Master 1ère Année 13-11-2014

## Référence

Ce chapitre s'appuie essentiellement sur l'ouvrage :

« Méthodes statistiques des sondages »,
de Jean-Marie Grosbras,
aux éditions Economica, 1987.

# Sommaire

## Remarque

C'est la deuxième méthode (après le sondage stratifié) qui va utiliser une variable auxiliaire car il est rare que nous ne disposons pas d'une variable quantitative ou qualitative dont la valeur/modalité est connue pour chacun des individus de la population.

## Principe fondamental

Lorsque nous disposons d'une information auxiliaire, il faut chercher à l'utiliser dans le but d'obtenir des estimateurs plus précis que les estimateurs simples de la moyenne ou du total qui apparaissent dans le cadre du sondage à PESR ou à PISR.

L'information auxiliaire peut être utilisée au niveau de la construction de l'échantillon (stratification, tirage proportionnel à un critère de taille,...) ou au niveau de l'expression de l'estimateur (techniques de redressement/calage).

Si plusieurs variables auxiliaires sont utilisées, nous pouvons recourir à une technique mixte dans laquelle certaines variables servent à améliorer le tirage de l'échantillon, et les autres à améliorer l'estimateur.

## Définition

*La stratification a posteriori est une méthode de redressement d'échantillon sur une variable qualitative.*

## C

ette méthode fait partie des méthodes de calage aux marges. Parmi les méthodes de calage aux marges, nous citons :

- post-stratification (ce chapitre)
- estimation par quotient (chapitre 5)
- estimation par régression (chapitre 7)
- estimation par régression multiple (non traité)

# Sommaire

## Principe

Nous étudions un caractère $X$ sur une population. Nous connaissons un autre caractère $Y$ sur cette **même** population et surtout sa distribution.

L'échantillon n'est **pas stratifié a priori sur** $Y$ mais pour chacune des unités échantillonnées on relève le couple $(x_i; y_i)$.

Nous définissons, à posteriori, des strates selon les valeurs de $Y$.

Nous repondérons les données par les poids véritables des strates définies sur $Y$.

Si ce critère $Y$ est corrélé avec $X$, c'est-à-dire si la variabilité de $X$ s'explique en partie par des différences entre les classes de $Y$, le calage de l'échantillon lui donne alors une représentativité plus fidèle et conduit à des résultats plus fiables.

C'est pourquoi les questionnaires comportent souvent en plus des questions qui abordent le thème de l'étude, des éléments de description de l'unité interrogée comme par exemple, le nombre de personnes du ménage, le nombre d'enfants, la CSP des adultes, les caractéristiques du logement...

Ces éléments permettent de juger de la qualité de l'échantillon et de suggérer des calages éventuels.

# Sommaire

## Exemple

Un échantillon de 1 000 personnes interrogées sur la question

« Allez-vous au cinéma au moins une fois par mois ? »

Nous avons croisé cette question avec une autre question

« Avez-vous une télévision ? »

## Voici la répartition des réponses obtenues

| | | Cinéma | | |
|---|---|---|---|---|
| | | oui | non | total |
| Télé | oui | 20 | 680 | 700 |
| | non | 80 | 220 | 300 |
| | total | 100 | 900 | 1 000 |

100 personnes répondent « oui » à la première question, ce qui nous permet d'estimer le pourcentage à 10%.
Le calcul que nous avons fait s'écrit de la façon suivante

$$\widehat{\pi} = \frac{20}{700} \times \frac{700}{1\,000} + \frac{80}{300} \times \frac{300}{1\,000} = 0,10.$$

## Remarque

Dans l'échantillon, il y a une sous représentation des possesseurs de télévisions. Comment le savons-nous ? Par d'autres sources qui nous indiquent qu'il y a 80% de gens qui possèdent une télévision.

## Conséquence

L'estimation du pourcentage ne se calcule plus de la même façon ! Rectifions le calcul

$$\widehat{\pi} = \frac{20}{700} \times \frac{800}{1\,000} + \frac{80}{300} \times \frac{200}{1\,000} = 0,076$$

ou encore $\widehat{\pi}$ est égal à $7,6\%$. Que faut-il en conclure ?

# Sommaire

## Définition

*L'estimateur d'une moyenne $\mu$ de la population U est défini par*

$$\widehat{\mu}_{post} = \sum_{h=1}^{k} \frac{N_h}{N} \widehat{\mu}_h,$$

*où $N_h$ représente l'effectif des strates a posteriori et*

$$\widehat{\mu}_h = \frac{1}{n_h} \sum_{i=1}^{n_h} x_{hi}.$$

## Remarque

C'est la même formule que la moyenne $\widehat{\mu}_{st}$ d'un échantillon stratifié a priori. Mais c'est seulement une apparence !

En effet

- Dans le calcul de $\widehat{\mu}_{st}$, les $\widehat{\mu}_h$ sont fondés sur des tailles $n_h$ fixées à l'avance.
- Dans le calcul de $\widehat{\mu}_{post}$, les $\widehat{\mu}_h$ sont fondés sur des tailles $n_h$ qui ne sont pas fixées à l'avance, mais qui sont des résultats constatés sur l'échantillon. Donc les tailles $n_h$ sont des quantités aléatoires.

## Comment faire dans les calculs si les $n_h$ sont aléatoires ?

La démarche se fait en deux étapes.

- Nous fixons d'abord les $n_h$.
- Puis nous introduisons l'aléatoire sur les $n_h$.

C'est cette démarche qui va nous permettre de calculer l'espérance de $\widehat{\mu}_{post}$ pour savoir si $\widehat{\mu}_{post}$ est un estimateur biaisé ou pas.

## Calcul de l'espérance de $\widehat{\mu}_{post}$

Nous avons par conditionnement

$$\mathbb{E}\left[\widehat{\mu}_{post}\right] = \mathbb{E}\left[\mathbb{E}\left[\widehat{\mu}_{post}|n_h\right]\right].$$

D'autre part, nous avons

$$\begin{aligned}
\mathbb{E}\left[\widehat{\mu}_{post}|n_h\right] &= \sum_{h=1}^{H} \frac{N_h}{N} \mathbb{E}\left[\widehat{\mu}_h|n_h\right] \\
&= \sum_{h=1}^{H} \frac{N_h}{N} \mu_h \\
&= \mu.
\end{aligned}$$

D'où, nous en concluons que

$$\begin{aligned}\mathbb{E}\left[\widehat{\mu}_{post}\right] &= \mathbb{E}\left[\mathbb{E}\left[\widehat{\mu}_{post}|n_h\right]\right] \\ &= \mathbb{E}\left[\mu\right] \quad \text{d'après ce que nous venons d'établir} \\ &= \mu.\end{aligned}$$

## Propriété

*Nous montrons, par calcul, que $\widehat{\mu}_{post}$ est un* **estimateur sans biais** *d'une moyenne $\mu$ de la population $U$.*

## Définition

*L'estimateur d'un total $T$ d'une population $U$ est défini par*

$$\widehat{T}_{post} = \sum_{h=1}^{H} N_h \widehat{\mu}_h,$$

*où $N_h$ représente l'effectif des strates a posteriori et*

$$\widehat{\mu}_h = \frac{1}{n_h} \sum_{i=1}^{n_h} x_{hi}.$$

## Propriété

*Nous montrons, par calcul comme précedemment, que* $\widehat{T}_{post}$ *est* **un estimateur sans biais** *d'un total* $T$ *d'une population* $U$, *i.e.*

$$\mathbb{E}\left[\widehat{T}_{post}\right] = \mathbb{E}\left[\sum_{h=1}^{k} N_h \widehat{\mu}_h\right] = T.$$

## Calcul de la variance de $\widehat{\mu}_{post}$

Nous procédons de la même manière que nous avons calculé l'espérance de cet estimateur, c'est à dire en conditionnant par $n_h$.
Par conséquent, nous obtenons

$$\mathrm{Var}\left[\widehat{\mu}_{post}\right] = \mathrm{Var}\left[\mathbb{E}\left[\widehat{\mu}_{post}|n_h\right]\right] + \mathbb{E}\left[\mathrm{Var}\left[\widehat{\mu}_{post}|n_h\right]\right].$$

Or nous avons montré précedemment que

$$\mathbb{E}\left[\widehat{\mu}_{post}|n_h\right] = \mu.$$

Par conséquent, nous avons

$$\mathrm{Var}\left[\mathbb{E}\left[\widehat{\mu}_{post}|n_h\right]\right] = \mathrm{Var}\left[\mu\right] = 0.$$

Reste plus qu'à calculer le second membre de l'équation de la variance.

$$\begin{aligned}
\mathrm{Var}\left[\widehat{\mu}_{post}|n_h\right] &= \sum_{h=1}^{k} \frac{N_h^2}{N^2} \mathrm{Var}\left[\widehat{\mu}_h|n_h\right] \\
&= \sum_{h=1}^{k} \frac{N_h^2}{N^2} \frac{N_h - n_h}{N_h n_h} S_{h,c}^2 \\
&= \sum_{h=1}^{k} \frac{N_h^2}{N^2} \frac{1}{n_h} S_{h,c}^2 - \frac{1}{N} \sum_{h=1}^{k} \frac{N_h}{N} S_{h,c}^2.
\end{aligned}$$

Par conséquent, nous avons

$$\begin{aligned}\mathbb{E}[\mathrm{Var}[\widehat{\mu}_{post}|n_h]] &= \mathbb{E}\left[\sum_{h=1}^{k}\frac{N_h^2}{N^2}\frac{1}{n_h}S_{h,c}^2 - \frac{1}{N}\sum_{h=1}^{k}\frac{N_h}{N}S_{h,c}^2\right] \\ &= \sum_{h=1}^{k}\frac{N_h^2}{N^2}S_{h,c}^2\mathbb{E}\left[\frac{1}{n_h}\right] - \frac{1}{N}\sum_{h=1}^{k}\frac{N_h}{N}S_{h,c}^2.\end{aligned}$$

Il ne reste plus qu'à calculer

$$\mathbb{E}\left[\frac{1}{n_h}\right].$$

Posons

$$\pi_h = \frac{N_h}{N} \quad \text{et} \quad \widehat{\pi}_h = \frac{n_h}{n}.$$

Remarquons que

$$\mathbb{E}[\widehat{\pi}_h] = \pi_h.$$

De plus, nous avons

$$\begin{aligned} n_h = n\frac{n_h}{n} = n\widehat{\pi}_h &= n(\widehat{\pi}_h - \pi_h + \pi_h) \\ &= n\pi_h\left(1 + \frac{\widehat{\pi}_h - \pi_h}{\pi_h}\right). \end{aligned}$$

Par conséquent nous en tirons que

$$\frac{1}{n_h} = \frac{1}{n\pi_h} \times \frac{1}{1 + \frac{\widehat{\pi}_h - \pi_h}{\pi_h}}.$$

Comme $\frac{\widehat{\pi}_h - \pi_h}{\pi_h}$ tend vers 0, nous pouvons faire un développement limité sur l'égalité ci-dessus et nous obtenons que :

$$\frac{1}{n_h} = \frac{1}{n\pi_h} \times \left(1 - \frac{\widehat{\pi}_h - \pi_h}{\pi_h} + \frac{(\widehat{\pi}_h - \pi_h)^2}{\pi_h^2} + o_{\mathbb{P}}\left(\frac{(\widehat{\pi}_h - \pi_h)^2}{\pi_h^2}\right)\right).$$

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{n_h}\right] &= \frac{1}{n\pi_h}\mathbb{E}\left[\left(1-\frac{\widehat{\pi}_h-\pi_h}{\pi_h}+\frac{(\widehat{\pi}_h-\pi_h)^2}{\pi_h^2}\right.\right.\\
&\quad \left.\left.+o_{\mathbb{P}}\left(\frac{(\widehat{\pi}_h-\pi_h)^2}{\pi_h^2}\right)\right)\right]\\
&= \frac{1}{n\pi_h}\left(1-0\right.\\
&\quad \left.+\mathbb{E}\left[\frac{(\widehat{\pi}_h-\pi_h)^2}{\pi_h^2}+o_{\mathbb{P}}\left(\frac{(\widehat{\pi}_h-\pi_h)^2}{\pi_h^2}\right)\right]\right).
\end{aligned}$$

Calculons maintenant

$$\mathbb{E}\left[\frac{(\widehat{\pi}_h - \pi_h)^2}{\pi_h^2} + o_{\mathbb{P}}\left(\frac{(\widehat{\pi}_h - \pi_h)^2}{\pi_h^2}\right)\right].$$

En remarquant que $\mathbb{E}[(\widehat{\pi}_h - \pi_h)^2]$ est égale à la variance de l'estimateur $\widehat{\pi}_h$ et que l'on est dans un cas de tirage à PESR, nous obtenons que

$$\mathbb{E}\left[\frac{(\widehat{\pi}_h - \pi_h)^2}{\pi_h^2} + o_{\mathbb{P}}\left(\frac{(\widehat{\pi}_h - \pi_h)^2}{\pi_h^2}\right)\right] \simeq \frac{N-n}{N-1}\frac{\pi_h(1-\pi_h)}{n} \times \frac{1}{\pi_h^2}.$$

Finalement, nous avons

$$\mathbb{E}\left[\frac{1}{n_h}\right] \simeq \frac{1}{n\pi_h}\left(1+\frac{N-n}{Nn}\frac{(1-\pi_h)}{\pi_h}\right).$$

D'où, nous en déduisons que

$$\mathrm{Var}\left[\widehat{\mu}_{post}\right] \simeq \sum_{h=1}^{k} \pi_h^2 S_{h,c}^2 \left( \frac{1}{n\pi_h} + \frac{N-n}{Nn^2} \frac{(1-\pi_h)}{\pi_h^2} \right) - \frac{1}{N} \sum_{h=1}^{k} \pi_h S_{h,c}^2.$$

En développant et en réorganisant les termes, nous obtenons

$$\begin{aligned}
\mathrm{Var}\left[\widehat{\mu}_{post}\right] &\simeq \frac{1}{n}\sum_{h=1}^{k}\pi_h S_{h,c}^2 - \frac{1}{N}\sum_{h=1}^{k}\pi_h S_{h,c}^2 \\
&\quad + \frac{1}{n}\frac{N-n}{Nn}\sum_{h=1}^{k}(1-\pi_h)S_{h,c}^2 \\
&\simeq \frac{N-n}{Nn}\sum_{h=1}^{k}\pi_h S_{h,c}^2 + \frac{1}{n}\frac{N-n}{Nn}\sum_{h=1}^{k}(1-\pi_h)S_{h,c}^2.
\end{aligned}$$

Finalement, nous obtenons que

$$\mathrm{Var}\left(\widehat{\mu}_{post}\right) \simeq \frac{(1-f)}{n}\sum_{h=1}^{H}\frac{N_h}{N}S_{h,c}^2 + \frac{(1-f)}{n^2}\sum_{h=1}^{H}\left(1-\frac{N_h}{N}\right)S_{h,c}^2$$

variance de $\widehat{\mu}_{post}$ + le prix à payer pour n'avoir pas tenu compte de la stratification dès le départ.

**Remarque**

Cette dernière quantité tend vers 0 lorsque $n$ devient grand.

## Propriété

*Nous montrons, par des calculs analogues à ceux développés pour l'estimateur de la moyenne, que*

$$\mathrm{Var}\left[\widehat{T}_{post}\right] \simeq N\left(\frac{(1-f)}{n}\sum_{h=1}^{k} N_h S_{h,c}^2 + \frac{(1-f)}{n^2}\sum_{h=1}^{k}(N-N_h)S_{h,c}^2\right).$$

# Sommaire

## Comparaison avec un SAS

Nous rappelons que

$$\begin{aligned} \mathrm{Var}[\widehat{\mu}] &= \frac{(1-f)}{n} S_c^2 \\ &\simeq \frac{(1-f)}{n} \left( \sum_{h=1}^{k} \frac{N_h}{N} S_{h,c}^2 + \sum_{h=1}^{k} \frac{N_h}{N} \left( \overline{X}_h - \mu \right)^2 \right) \end{aligned}$$

et

$$\mathrm{Var}\left[\widehat{\mu}_{post}\right] \simeq \frac{(1-f)}{n} \left( \sum_{h=1}^{k} \frac{N_h}{N} S_{h,c}^2 + \frac{1}{n} \sum_{h=1}^{k} \frac{N - N_h}{N} S_{h,c}^2 \right).$$

D'où, nous en déduisons que

$$\frac{n}{(1-f)}\left(\mathrm{Var}\left[\widehat{\mu}\right]-\mathrm{Var}\left[\widehat{\mu}_{post}\right]\right)$$
$$\simeq \sum_{h=1}^{k}\frac{N_h}{N}\left(\overline{X}_h-\mu\right)^2-\frac{1}{n}\sum_{h=1}^{k}\frac{N-N_h}{N}S_{h,c}^2.$$

La stratification a posteriori se justifie lorsque cette quantité est largement positive.

## Remarques

1. La variable étudiée doit-être corrélée avec le critère de stratification, c'est-à-dire avoir une valeur élevée du rapport de corrélation inter-strate.
2. $n$ doit être assez grand, puisque on se sert de $1/n \to 0 \quad$ lorsque $n \to +\infty$. Donc c'est inutile de repondérer les petits échantillons.
3. $(N - N_h)/N$ doit être très petit, puisque on se sert de cette hypothèse pour faire un développement limité. Il faut donc que $N_h/N$ doit être grand. Par conséquent, c'est inutile d'avoir beaucoup de petites strates.

# Sommaire

## Retour à l'exemple « Cinéma et Télévision »

Nous avons le tableau suivant :

| | $B_1$ | $B_2$ | total |
|---|---|---|---|
| $A_1$ | 20 | 680 | 700 |
| $A_2$ | 80 | 220 | 300 |
| total | 100 | 900 | 1 000 |

En réalité, la marge sur A est $(800, 200)$.

Comme nous l'avons montré au début de ce chapitre, la moyenne calée sur A se calcule par

$$\frac{1}{1\,000}\left[\frac{800}{700}\sum_{i\in A_1} y_i + \frac{200}{300}\sum_{i\in A_2} y_i\right].$$

Les observations de $A_1$ sont redressées par $800/700$ et celles de $A_2$ par $200/300$.

Imaginons que l'échantillon soit déformé par rapport à B. Nous savons par d'autres sources, que la marge de B est en réalité $(80, 920)$.

## Problème

Nous voulons caler l'échantillon sur les deux critères simultanément.

## Solution idéale

Connaître les effectifs théoriques croisés mais en réalité on ne dispose que des marges.

## Problème

Estimer les coefficients de redressement par case, respectant les conditions à la marge.

## Quatre Solutions

- La méthode RAS
- La méthode ASAM
- L'ajustement par l'analyse des données
- La méthode de Lemel (1976)

Nous ne développerons pas les deux dernières méthodes, mais nous renvoyons au livre de Jean-Marie Grosbras pour de plus amples renseignements sur ces deux méthodes.

## La méthode RAS : Le principe

- Le tableau à ajuster est $A = (a_{ij})$.
- Le total de ligne est $a_{i.}$, le total théorique est $r_i$.
- Le total de ligne est $a_{.j}$, le total théorique est $s_j$.
- On commence par ajuster les totaux en ligne : $a_{ij} \rightarrow a_{ij} = a_{ij} * (r_i / a_{i.})$.
- Puis on ajuste les totaux en colonne : $a_{ij} \rightarrow a_{ij} = a_{ij} * (s_j / a_{.j})$.
- En ajustant les totaux en colonne, on a détruit l'ajustement des totaux en ligne. On recommence...
- On itère le processus jusqu'à convergence.

Avec les données de l'exemple « Cinéma-Télévision », nous avons

$$A = \begin{bmatrix} 15 & 20 & 45 & 12 \\ 45 & 67 & 23 & 12 \\ 67 & 23 & 91 & 15 \\ 77 & 33 & 91 & 35 \end{bmatrix} \quad r = \begin{bmatrix} 100 \\ 150 \\ 150 \\ 200 \end{bmatrix}$$

$$s = \begin{bmatrix} 170 & 150 & 190 & 90 \end{bmatrix}$$

## La méthode RAS donne

| | | | | |
|---|---|---|---|---|
| 16 | 24 | 42 | 18 | 100 |
| 41 | 73 | 20 | 16 | 150 |
| 51 | 21 | 62 | 16 | 150 |
| 62 | 32 | 66 | 40 | 200 |
| 170 | 150 | 190 | 90 | 600 |

## Ajustement Statistique et Algébrique d'une Matrice (ASAM)

Cette méthode est plus générale et englobe comme cas particulier la méthode RAS.

## Idée

Si l'échantillon n'est pas trop mauvais, la structure croisée observée doit avoir des similitudes avec la « vraie » structure.

- On a un tableau $(a_{ij})$ tel que $T = \sum_i \sum_j a_{ij}$.
- On chercher un tableau $(x_{ij})$, proche de $(a_{ij})$ tel que

$$\sum_j x_{ij} = r_i, \quad \sum_i x_{ij} = s_j, \quad \sum\sum x_{ij} = T.$$

La méthode ASAM consiste en la résolution du programme suivant

$$\min_{x_{ij}} \sum_i \sum_j \frac{1}{\rho_{ij}} (x_{ij} - a_{ij})^2$$

avec

$$\sum_j x_{ij} = r_i, \quad \sum_i x_{ij} = s_j, \quad \sum_i \sum_j x_{ij} = T.$$

Les $\rho_{ij}$ sont à choisir si nous voulons moduler l'importance de chaque case.

## Remarques

- La méthode ASAM est une méthode des moindres carrés pondérés et contraints.
  Il existe des programmes traitant ce genre de problème.
- Le choix optimal pour les $\rho_{ij}$ est de les prendre proportionnels aux variances des $a_{ij}$, considérés comme des variables aléatoires

$$\rho_{ij} = c\mathrm{Var}[a_{ij}].$$

## Suite des remarques

- Nous prenons dons les $\rho_{ij}$ représentatifs de ce que nous pouvons connaître des variances des effectifs $a_{ij}$.
- La méthode RAS est un cas particulier de la méthode ASAM dans le cas où les $a_{ij}$ sont proportionnels à leur variance.
- La méthode ASAM est plus satisfaisante puisqu'elle recherche une similitude de structure.
  Elle est évidemment plus coûteuse.