

# Stratification a posteriori

Myriam Maumy<sup>1</sup>

<sup>1</sup>IRMA, Université Louis Pasteur  
Strasbourg, France

Master 1ère Année 27-03-2006

Ce chapitre s'appuie essentiellement sur le livre :

- « Méthodes statistiques des sondages »,  
Jean-Marie Grosbras,  
Economica.

# Sommaire

- 1 Définition
- 2 Exemple
- 3 Principe
- 4 Formules d'estimation d'une stratification a posteriori
- 5 Comparaison avec un SAS
- 6 Redressements sur critères multiples

## Définition

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

# Sommaire

- 1 Définition
- 2 Exemple**
- 3 Principe
- 4 Formules d'estimation d'une stratification a posteriori
- 5 Comparaison avec un SAS
- 6 Redressements sur critères multiples

## Exemple

Un échantillon de 1 000 personnes interrogées sur la question

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

On a 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

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

## Remarque

Dans l'échantillon, il y a une sous représentation des possesseurs de télévisions. Comment le sait-on ? 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

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

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



# Sommaire

- 1 Définition
- 2 Exemple
- 3 Principe**
- 4 Formules d'estimation d'une stratification a posteriori
- 5 Comparaison avec un SAS
- 6 Redressements sur critères multiples

## Principe

On étudie un caractère  $X$  sur une population. On connaît 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  $(x_i; y_i)$ .

On définit, à posteriori, des strates selon les valeurs de  $Y$ .

On repondère 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

- 1 Définition
- 2 Exemple
- 3 Principe
- 4 Formules d'estimation d'une stratification a posteriori**
- 5 Comparaison avec un SAS
- 6 Redressements sur critères multiples

## Définition

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

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

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

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

## Remarque

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

En effet

- Dans le calcul de  $\hat{\mu}_{st}$ , les  $\hat{\mu}_h$  sont fondés sur des tailles  $n_h$  fixées à l'avance.
- Dans le calcul de  $\hat{\mu}_{post}$ , les  $\hat{\mu}_h$  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.

- On fixe d'abord les  $n_h$ .
- Puis on introduit l'aléatoire sur les  $n_h$ .

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

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

On a par conditionnement

$$\mathbb{E} [\hat{\mu}_{post}] = \mathbb{E} [\mathbb{E} [\hat{\mu}_{post} | n_h]] .$$

D'autre part, on a

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



D'où

$$\begin{aligned}\mathbb{E} [\hat{\mu}_{post}] &= \mathbb{E} [\mathbb{E} [\hat{\mu}_{post} | n_h]] \\ &= \mathbb{E} [\mu] \quad \text{d'après ce que l'on vient d'établir} \\ &= \mu.\end{aligned}$$

## Propriété

*On montre, par calcul, que  $\hat{\mu}_{post}$  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*

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

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

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

## Propriété

On montre, par calcul comme précédemment, que  $\hat{T}_{post}$  est un **estimateur sans biais** d'un total  $T$  d'une population  $U$ , i.e.

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

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

On procède de la même manière que l'on a calculé l'espérance de cet estimateur, c'est à dire en conditionnant par  $n_h$ .

Par conséquent, on obtient

$$\text{Var} [\hat{\mu}_{post}] = \text{Var} [\mathbb{E} [\hat{\mu}_{post} | n_h]] + \mathbb{E} [\text{Var} [\hat{\mu}_{post} | n_h]].$$

Or on a montré précédemment que

$$\mathbb{E} [\hat{\mu}_{post} | n_h] = \mu.$$

Par conséquent, on a

$$\text{Var} [\mathbb{E} [\hat{\mu}_{post} | n_h]] = \text{Var} [\mu] = 0.$$

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

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

Par conséquent, on a

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

Il ne reste plus qu'à calculer

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

Posons

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

Remarquons que

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

De plus, on a

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

Par conséquent on en tire que

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

Comme  $\frac{\hat{\pi}_h - \pi_h}{\pi_h}$  tend vers 0, on peut faire un développement limité sur l'égalité ci-dessus et on obtient que :

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



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

Calculons maintenant

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

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

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

Finalement, on a

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

D'où

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

En développant et en réorganisant les termes, on obtient

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

Finalement, on obtient que

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

variance de  $\bar{x}_{post}$  + 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

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

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

# Sommaire

- 1 Définition
- 2 Exemple
- 3 Principe
- 4 Formules d'estimation d'une stratification a posteriori
- 5 Comparaison avec un SAS**
- 6 Redressements sur critères multiples



## Comparaison avec un SAS

On rappelle que

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

et

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

D'où

$$\frac{n}{(1-f)} (\text{Var} [\hat{\mu}] - \text{Var} [\hat{\mu}_{post}]) \\ \simeq \sum_{h=1}^k \frac{N_h}{N} (X_h - \mu)^2 - \frac{1}{n} \sum_{h=1}^k \frac{N - N_h}{N} S_{h,c}^2.$$

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  $1/n \rightarrow 0$  lorsque  $n \rightarrow +\infty$ . 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

- 1 Définition
- 2 Exemple
- 3 Principe
- 4 Formules d'estimation d'une stratification a posteriori
- 5 Comparaison avec un SAS
- 6 Redressements sur critères multiples**

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

On a le tableau suivant :

	B <sub>1</sub>	B <sub>2</sub>	total
A <sub>1</sub>	20	680	700
A <sub>2</sub>	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. On sait par d'autres sources, que la marge de B est en réalité  $(80, 920)$ .

## Problème

On veut 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)

On ne développera 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 colonne est  $a_{.j}$ , le total théorique est  $s_j$ .
- On commence par ajuster les totaux en ligne :  
$$a_{ij} \rightarrow \bar{a}_{ij} = a_{ij} * (r_i / a_{i.}).$$
- Puis on ajuste les totaux en colonne :  
$$a_{ij} \rightarrow \bar{a}_{ij} = \bar{a}_{ij} * (s_j / a_{.j}).$$
- 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 », on a

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

$$s = [170 \quad 150 \quad 190 \quad 90]$$

## 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 cherche un tableau  $(x_{ij})$ , proche de  $(a_{ij})$  tel que

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

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 on veut 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

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

## Suite des remarques

- On prend donc les  $\rho_{ij}$  représentatifs de ce que l'on peut 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.