le cnam

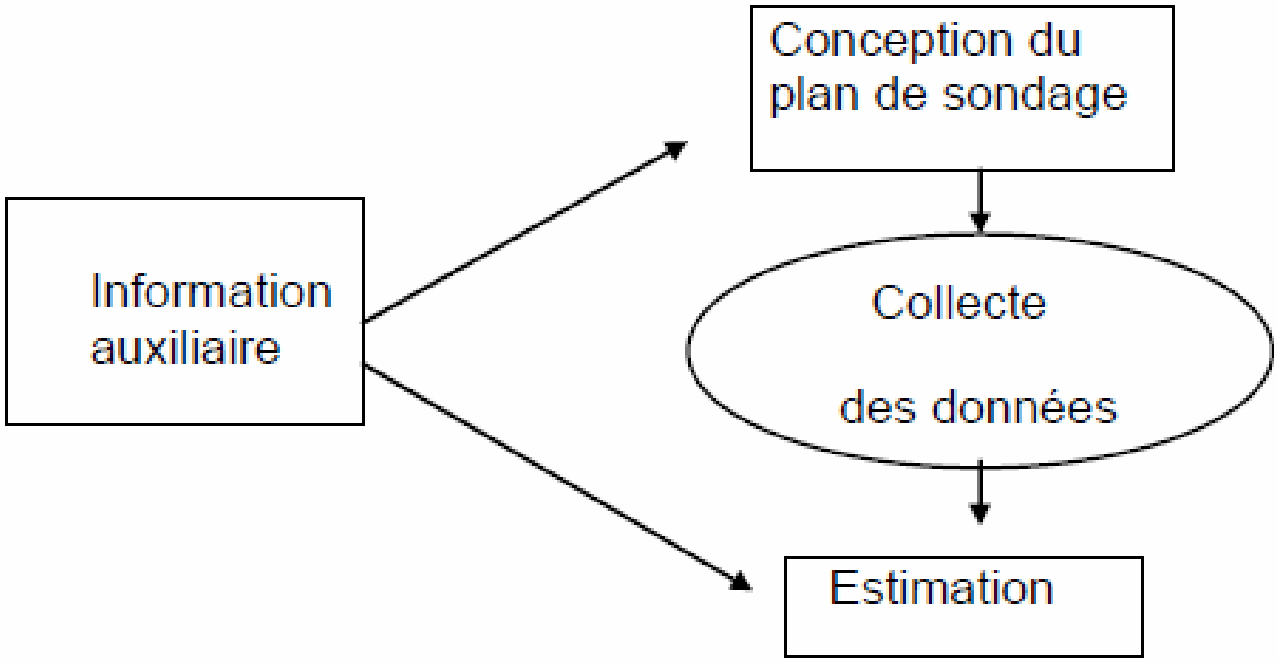
SONDAGES STRATIFIES

Philippe Périé & Gilbert Saporta

STA108, 26 octobre 2012

Information auxiliaire

FIG. 1 - *Les deux étapes de l'utilisation de l'information auxiliaire*



STRATIFICATION

- Idée :
  - S'il existe dans la base de sondage un critère permettant de distinguer a priori entre eux les individus, on aura tout à gagner à utiliser cette information pour répartir l'échantillon dans chaque sous-population.
  - C'est le principe de la stratification: découper la population en sous ensembles homogènes appelés strates et réaliser un sondage dans chacune d'elles.
- La stratification a pour objectifs de pour objectif de diminuer la variance, augmenter la précision

Intuition

Dans un sondage aléatoire simple, toutes les combinaisons de *n* éléments parmi N sont possibles avec la *même probabilité.*

Or, il arrive que certaines d'entre elles puissent s'avérer *a priori* indésirables

N=5
Variable d'intérêt Y {13 15 17 25 30} dépôt en k€
$Y_{moy} = 20$

Recensement des résultats possibles **n=2**

y_1	13	13	13	13	15	15	15	17	17	25
y_2	15	17	25	30	17	25	30	25	30	30
y_{moy}	14	15	19	21,5	16	20	22,5	21	23,5	27,5

Par exemple, parmi ces échantillons de 2 unités, on trouve les cas extrêmes (13, 15) et (25, 30) qui sont particulièrement « mauvais ».

S'il existe dans la base de sondage un critère permettant de distinguer *a priori* les catégories des petits et gros clients, on aura tout à gagner à utiliser cette information pour répartir l'échantillon dans chaque sous-population.

Intuition

Le principe de la stratification :

Découper la population en sous-ensembles appelés *strates* et réaliser un sondage dans chacune d'elles : on espère ainsi exclure les échantillons extrêmes, et - plus généralement – améliorer la précision des estimateurs

(On a vu qu'à taille égale un échantillon est plus efficace dans une population homogène que dans une population hétérogène. Plus précisément, l'erreur type d'estimation est lié à la variance du caractère étudié dans la population.)

Chaque sondage partiel s'effectuera ainsi de façon plus efficace et l'assemblage de sondages partiels plus précis donnera des résultats plus fiables qu'un sondage de même taille effectué « en vrac »

La plupart des fois la stratification correspond par ailleurs à un objectif de réduction des coûts d'enquête ou d'optimisation de sa gestion

C'est en particulier le cas lorsque l'on utilise un critère de découpage géographique comme la région, ou, dans les échantillon d'entreprise, un critère sectoriel permettant de spécialiser les enquêteurs

Intuition

N=5
Variable d'intérêt Y {13 15 17 25 30} dépôt en k€
$Y_{moy} = 20$

Échantillons avec stratification **n=2 (un chez les petits, un chez les grands)**

$.y_1$	13	13	15	15	17	17
$.y_2$	25	30	25	30	25	30
y_{moy}	17,8	19,8	19	21	20,2	22,2

L'unité échantillonnée dans la première strate est désignée pour en représenter trois, celle de la deuxième strate vaut pour deux. **Il convient** donc **de pondérer chaque valeur par le poids de la strate dont elle est issue**

$$y_{moy} = \frac{3}{5} y_1 + \frac{2}{5} y_2$$

On peut vérifier que la moyenne des six valeurs réalisables pour y_{moy} est encore 20. Cela signifie que **la variable aléatoire y_{moy} a Y_{moy} pour espérance mathématique** et qu'**elle est donc un estimateur sans biais** pour ce paramètre.

Intuition

N=5
Variable d'intérêt Y {13 15 17 25 30} dépôt en k€
Y_{moy} = 20

Échantillons avec stratification **n=2 (un chez les petits, un chez les grands)**

$.y_1$	13	13	15	15	17	17
$.y_2$	25	30	25	30	25	30
y_{moy}	17,8	19,8	19	21	20,2	22,2

On remarque également que **la plage des estimations est beaucoup plus resserrée autour de la cible que dans le cas du SAS : les valeurs extrêmes sont moins éloignées**, l'erreur type (c'est-à-dire la racine carrée de la variance des six valeurs) vaut 1,40 au lieu de 3,95.

STRATIFICATION

Déterminer des strates les plus homogènes possibles, par rapport au sujet étudié.

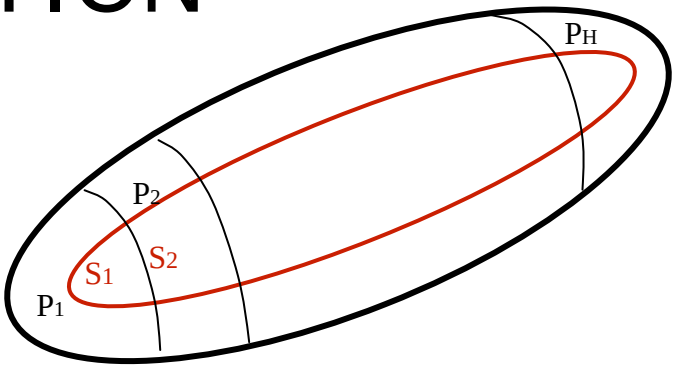
2 types de considérations vont conduire au choix des critères de stratification :

- 1. disponibilité des critères dans la base de sondage ;
- 2. pertinence des différents critères pour créer des strates homogènes.

Ceci nécessite une connaissance

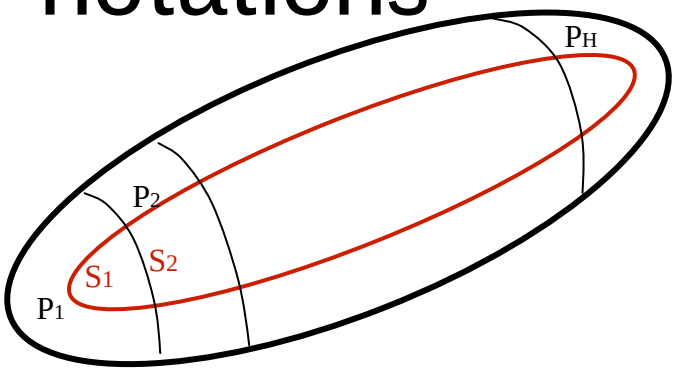
- soit intuitive,
- soit venant d'études réalisées antérieurement.

STRATIFICATION



- Utilisation d'une information auxiliaire qualitative
- Toujours efficace

STRATIFICATION, notations



- **Strates:**

$N_1, N_2 \ldots\ldots N_h \ldots\ldots N_H$ $\quad N = \sum N_h$

$\bar{Y}_1, \bar{Y}_2 \ldots\ldots \bar{Y}_h \ldots\ldots \bar{Y}_H$ $\quad \bar{Y} = \sum \frac{N_h}{N} \bar{Y}_h$

$\sigma_1^2, \sigma_2^2 \ldots\ldots \sigma_h^2 \ldots\ldots \sigma_H^2$ $\quad \sigma^2 = \sum \frac{N_h}{N} \sigma_h^2 + \sum \frac{N_h}{N} \left(\bar{Y}_h - \bar{Y}\right)^2$

- **Échantillon:**

$n_1, n_2, \ldots\ldots n_h, \ldots\ldots n_H$ $\quad n = \sum n_h$

$\bar{y}_1, \bar{y}_2, \ldots\ldots \bar{y}_h, \ldots\ldots \bar{y}_H$ $\quad \bar{y} = \sum \frac{n_h}{n} \bar{y}_h$

$\hat{\sigma}_1^2, \hat{\sigma}_2^2, \ldots\ldots \hat{\sigma}_h^2 \ldots\ldots \hat{\sigma}_H^2$

STRATIFICATION

$$\sigma^2 = \sum \frac{N_h}{N} \sigma_h^2 + \sum \frac{N_h}{N} \left(\bar{Y}_h - \bar{Y} \right)^2 = \sigma^2_W + \sigma^2_B$$

- Variance totale=

  moyenne des variances (*variance intra*)

  +variance des moyennes (*variance inter*)

STRATIFICATION

- Pour la suite, on se placera dans le cas d'un **tirage aléatoire simple sans remise**, à l'intérieur de chaque strate.

STRATIFICATION

- Estimateur sans biais de $\bar{Y}$ (Horvitz Thomson)

$$\hat{\bar{Y}}_{str} = \sum \frac{N_h}{N} \bar{y}_h$$

- Variance:

$$V(\hat{\bar{Y}}_{str}) = \sum \left(\frac{N_h}{N} \right)^2 V(\bar{y}_h) = \sum_{h=1}^{H} \left(\frac{N_h}{N} \right)^2 \frac{\sigma_h^2}{n_h} \frac{N_h - n_h}{N_h - 1}$$

$$= \frac{1}{N^2} \sum_{h=1}^{H} N_h (N_h - n_h) \frac{S_h^2}{n_h}$$

STRATIFICATION, répartition proportionnelle

- Échantillon dit « représentatif »:

$$\frac{n_h}{n} = \frac{N_h}{N} \Rightarrow \tau_h = \frac{n_h}{N_h} = \frac{n}{N} = \tau$$

- Taux de sondage constant dans chaque strate

$$\hat{\bar{Y}}_{str} = \sum_{h=1}^{H} \frac{N_h}{N} \bar{y}_h = \sum_{h=1}^{H} \frac{n_h}{n} \bar{y}_h = \bar{y} = \hat{\bar{Y}}_{prop}$$

STRATIFICATION, répartition proportionnelle

- variance :

$$V(\hat{\bar{Y}}_{prop}) = \frac{1}{N^2}\sum_{h=1}^{H} N_h(N_h - n_h)\frac{S_h^{\,2}}{n_h} = \frac{1}{N^2}\sum_{h=1}^{H}\frac{N_h - n_h}{n_h} N_h S_h^{\,2}$$

$$= \frac{1}{N^2}\sum_{h=1}^{H}\left(\frac{N_h}{n_h} - 1\right)N_h S_h^{\,2} = \frac{1}{N^2}\sum_{h=1}^{H}\left(\frac{N}{n} - 1\right)N_h S_h^{\,2} = \frac{N-n}{nN}\sum_{h=1}^{H}\frac{N_h}{N} S_h^{\,2}$$

- Si N_h est grand:

$$V(\hat{\bar{Y}}_{prop}) = \frac{N-n}{nN}\sum_{h=1}^{H}\frac{N_h}{N} S_h^{\,2} \simeq \frac{N-n}{nN}\sum_{h=1}^{H}\frac{N_h}{N}\sigma_h^{\,2} = \frac{N-n}{N}\frac{\sigma^2_{\,w}}{n}$$

STRATIFICATION, répartition proportionnelle

- Variance de l'estimateur du SAS sans remise:

$$V(\bar{y}) = (1 - \frac{n}{N})\frac{S^2}{n} = \frac{N-n}{N}\frac{S^2}{n} \simeq \frac{N-n}{N}\frac{\sigma^2}{n}$$

- Avec les mêmes probabilités d'inclusion d'ordre 1, l'échantillon stratifié représentatif est plus efficace qu'un échantillon simple de même taille dès que les $\bar{Y}_h$ sont différents.

STRATIFICATION optimale

- Répartition optimale:

$$V(\widehat{\overline{Y}}_{str}) = \frac{1}{N^2}\sum \frac{N_h\left(N_h - n_h\right)}{n_h} S_h^{\,2}$$

avec $S_h^{\,2} = \frac{N_h}{N_h - 1}\sigma_h^{\,2}$

c_h – coût unitaire d'une observation

$$\begin{cases} \min \sum \frac{N_h\left(N_h - n_h\right)}{n_h} S_h^{\,2} \\ \sum n_h c_h = c_0 \end{cases}$$

$$\sum \frac{N_h^{\,2}}{n_h} S_h^{\,2} - \underbrace{\sum N_h S_h^{\,2}}_{fixe}$$

STRATIFICATION optimale

- Solution:

$$\frac{N_h^2 S_h^2}{n_h^2}$$ proportionnel à c_h

$$\frac{n_h}{N_h} \propto \frac{S_h}{\sqrt{c_h}}$$

Si c_h constant:

$$n_h = n\frac{N_h S_h}{\sum N_h S_h}$$ - Répartition de Neyman

STRATIFICATION optimale

- Cette répartition utilise un taux de sondage *f* proportionnel à la dispersion S_h de *X* étudiée dans chaque strate.
- Plus une strate est hétérogène vis-à-vis de *la* variable étudiée, plus on utilise un taux de sondage important.
- La théorie montre que cette répartition est celle qui fournit la variance la plus faible une fois les strates déterminées.

STRATIFICATION optimale

- Remarquons que l'échantillon de Neyman dépend du caractère que l'on veut estimer en priorité. C'est pour ce caractère que l'on prendra la variance en considération.
- En général, celle-ci ne sera pas connue *a priori*. Elle pourra être estimée à partir d'une enquête antérieure ou d'études limitées.

STRATIFICATION

- Exemple n° 1: présondage de 155 unités

Strates	**1**	**2**	**3**	**4**	
N_h	3750	3272	1387	2475	10 884
n_h	50	45	30	30	155
$\overline{y}_h$	12.6	14.5	18.6	13.8	
$\hat{\sigma}_h^2$	2.8	2.9	4.8	3.2	

STRATIFICATION

- Exemple n° 1:

$$\widehat{\overline{Y}} = \sum \left(\frac{N_h}{N} \right) \overline{y}_h = \frac{3750 \text{x} 12.6 + ... + 2475 \text{x} 13.8}{10884} = 14.21$$

$$\hat{V}(\widehat{\overline{Y}}) \simeq \sum \left(\frac{N_h}{N} \right)^2 \frac{\widehat{\sigma}_h^{\,2}}{n_h} = 0.02059 = (0.14)^2$$

Intervalle de confiance à 95% pour $\overline{Y}$:

$14.21 \pm 2 \text{x} 0.14$ soit: $\left[13.93 < \overline{Y} < 14.49 \right]$

Pour T: 154662 ± 3047

STRATIFICATION

- Exemple n° 1:

$$\sigma^2 = \sum \frac{N_h}{N} \sigma^2{}_h + \sum \frac{N_h}{N} \left(Y_h - Y \right)^2$$

On estime: σ^2_h par $\frac{n_h}{n_{h-1}} \hat{\bar{Y}}_{str2} \sigma_h$

$\overline{Y}_h$ par $\overline{y}_h$

$\overline{\overline{Y}}$ par $\hat{\bar{Y}}_{str}$

$$\hat{\sigma}^2 = 6.06 = \left(2.46 \right)^2$$

STRATIFICATION

- Suite: Répartition de Neyman pour n=1000:

$N_1S_1 = 6275$ $n_1 = 1000 \times 6275/19\,312 = 325$

$N_2S_2 = 5572$ $n_2 = 288$

$N_3S_3 = 3038$ $n_3 = 157$

$N_4S_4 = 4427$ $n_4 = 229$

$19\,312$

Variance: $\frac{1}{N^2}\sum\frac{N_h\left(N_h - n_h\right)}{n_h}S_h^{\,2} = 0.0029 = \left(0.0542\right)^2$

$\overline{Y}$ connu à $\pm$ 2 x 0.0542 soit $\pm$ 0.108

T connu à $\pm$ 1179

STRATIFICATION

- Échantillon simple à 1000:

$$\frac{\sigma^2}{n} \times \frac{N-n}{N-1} = 0.0055 = (0.0742)^2$$

$\overline{Y}$ connu à $\pm$ 0.15; T connu à $\pm$ 1615

- Échantillon stratifié représentatif:

$n_1 = 345$

$n_2 = 301$

$n_3 = 127$

$n_4 = 227$

STRATIFICATION

- Estimation d'une proportion p
- Même démarche: une proportion est une moyenne particulière

$$\hat{p}_{str} = \sum_{h=1}^{H} \frac{N_h}{N} f_h$$

$$V(\hat{p}_{str}) = \sum_{h=1}^{H} \left(\frac{N_h}{N} \right)^2 \frac{p_h(1-p_h)}{n_h} \frac{N_h - n_h}{N_h - 1}$$

$$\hat{V}(\hat{p}_{str}) \simeq \sum_{h=1}^{H} \left(\frac{N_h}{N} \right)^2 \frac{f_h(1-f_h)}{n_h} \left(1 - \frac{n_h}{N_h} \right)$$

STRATIFICATION

- Comment stratifier?
  - Remarque préalable: dans un sondage à probabilité inégale π_i proportionnel à Y_i annule la variance.
  - Nombre de strates: le maximum mais...
  - Limites de strates optimales:

    méthode de Dalenius et Hodges. Regrouper des classes selon le cumul de la racine des effectifs

STRATIFICATION

- Répartition dans les strates:
  - Si Sh inconnu : répartition proportionnelle
  - Si Sh connu: Neyman
  - Sinon, hypothèse fréquente $\frac{S_h}{\overline{Y}_h} = c$ d'où n_h proportionnel à la somme de la variable étudiée ou d'une variable corrélée.
  - Exemple: échantillon d'entreprises proportionnel au CA ou à l'effectif de la strate.

STRATIFICATION

- Variable de stratification: en théorie Y; sinon, variable bien corrélée avec Y.

- En pratique quand il y a plusieurs variables d'intérêt et une variable de stratification, on utilise la répartition proportionnelle

Exemples

Enquêtes INSEE auprès des entreprises, sondages B2B en institut.

« Le plan de sondage des enquêtes de l'INSEE auprès des entreprises est en général un plan de sondage stratifié avec un sondage aléatoire simple sans remise dans chaque strate. »

Exemples

Indice des prix

http://www.insee.fr/fr/methodes/default.asp?page=sources/ope-ipc.htm

« Le plan de sondage est stratifié selon trois types de critères :

- critère géographique : les relevés sont effectués dans 96 agglomérations de plus de 2 000 habitants dispersées sur le territoire métropolitain et de toute taille ainsi que 10 agglomérations dans les DOM ;

- type de produit : un échantillon d'un peu plus de 1000 familles de produits, appelées "variétés" est défini pour tenir compte de l'hétérogénéité des produits au sein des postes. La variété est le niveau de base pour le suivi des produits et le calcul de l'indice. La liste des variétés reste confidentielle et l'IPC n'est pas diffusé à ce niveau ;

- type de point de vente : un échantillon de 27 000 points de vente, stratifié par forme de vente, a été constitué pour représenter la diversité des produits et modes d'achat des consommateurs et prendre en compte des variations de prix différenciées selon les formes de vente.

Le croisement de ces différents critères aboutit à suivre un peu plus de 140 000 séries (produits précis dans un point de vente donné) donnant lieu à près de 160 000 relevés mensuels. »

Taille des strates
Autres considérations

- Dans la pratique, d'autres considérations que la précision optimale peuvent guider l'allocation dans les strates, comme la nécessité d'avoir des bases de lectures suffisantes sur chaque strate

- L'étude d'audience de la presse 'Audipresse ONE' part d'une répartition géographique proportionnelle, à partir de laquelle on impose des seuils minimaux dans chaque département.

- Citons aussi le type d'abonnement pour le secteur des télécoms, les classes d'ancienneté, les canaux de recrutement des client pour les études de satisfaction, ...