Optimisation d'une allocation mixte

Thomas Merly-Alpa, Antoine Rebecq (INSEE)

9ème Colloque Francophone sur les Sondages - Gatineau

14 octobre 2016

Sommaire

Allocation(s)

En sondages, une allocation est une ventilation d'une taille totale d'échantillon n entre plusieurs strates $h \in H$. Les allocations obtenues n_h sont choisies pour atteindre certains objectifs :

- Équipondération
- Précision maximale
- Précision sur des domaines de diffusion...

Allocation mixte

On peut souhaiter combiner plusieurs de ces objectifs. L'allocation mixte est souvent utilisée dans ce cadre. On a :

$$n_{\text{mixte}} = \frac{1}{2} n_1 + \frac{1}{2} n_2 \tag{1}$$

Ici, on utilisera l'allocation mixte entre Neyman et proportionnelle.

Allocations mixtes

On étudie dans cette présentation les allocations définies ici :

$$n_\alpha = \alpha \cdot n_{\text{Prop}} + (1 - \alpha) \cdot n_{\text{Neyman}} \tag{2}$$

On souhaite choisir α de telle sorte que l'allocation soit optimale.

Le programme d'optimisation

Programme d'optimisation

On étudie les allocations définies ici :

$$n_\alpha = \alpha \cdot n_{\text{Prop}} + (1 - \alpha) \cdot n_{\text{Neyman}} \tag{3}$$

Le programme de minimisation est le suivant :

$$\min_{\alpha \in [0,1]} \quad \underbrace{\text{Disp}(\text{Poids})}_{\text{Équivalence}} + \lambda \ \underbrace{\text{Dist}((\text{n}_\alpha), (\text{n}_{\text{Neyman}}))}_{\text{Spécificité}} \tag{4}$$

où $\lambda \in [0, +\infty[$ est à spécifier.

Respecter le plan de sondage initial

L'objectif de **Spécificité** consiste à respecter le plan de sondage spécifiquement créé pour l'enquête.

Ici, il s'agit de l'allocation de Neyman utilisée pour optimiser la précision.

En pratique, on veut minimiser l'écart de l'échantillon au plan de sondage initial.

Minimiser la dispersion des poids

L'objectif d'**Équivalence** consiste à minimiser la dispersion des poids (corrigés de la non-réponse) :

- Estimations multiples
- Robustesse
- Estimations économétriques
- Correction de la non-réponse et calage

On utilise ici les poids initiaux.

Programme d'optimisation

Dans le cadre d'un sondage aléatoire simple stratifié à H strates, on réécrit :

$$\min_{\alpha \in [0,1]} \quad \sum_{h=1}^{H} n_\alpha^h \left(\frac{N^h}{n_\alpha^h} - \bar{p} \right)^2 + \lambda \, \mathrm{Dist}((\mathrm{n}_\alpha), (\mathrm{n}_{\mathrm{Neyman}})) \qquad (5)$$

avec :

$$\bar{p} = \frac{\sum_{h=1}^{H} n_\alpha^h \frac{N^h}{n_\alpha^h}}{n} = \frac{N}{n}$$

Programme d'optimisation

On utilise la distance définie de la façon suivante :

$$Dist_m(x, y) = \max_{h \leq H} |x_h - y_h|$$

Le second terme se réécrit comme un polynôme en α, ici :

$$\min_{\alpha \in [0,1]} \quad \sum_{h=1}^{H} n_\alpha^h \left(\frac{N^h}{n_\alpha^h} - \frac{N}{n} \right)^2 + \lambda \cdot \alpha \tag{6}$$

Comment choisir λ ?

Le terme λ de l'équation (8) est choisi pour limiter la variance de l'estimation d'Horvitz-Thompson de X :

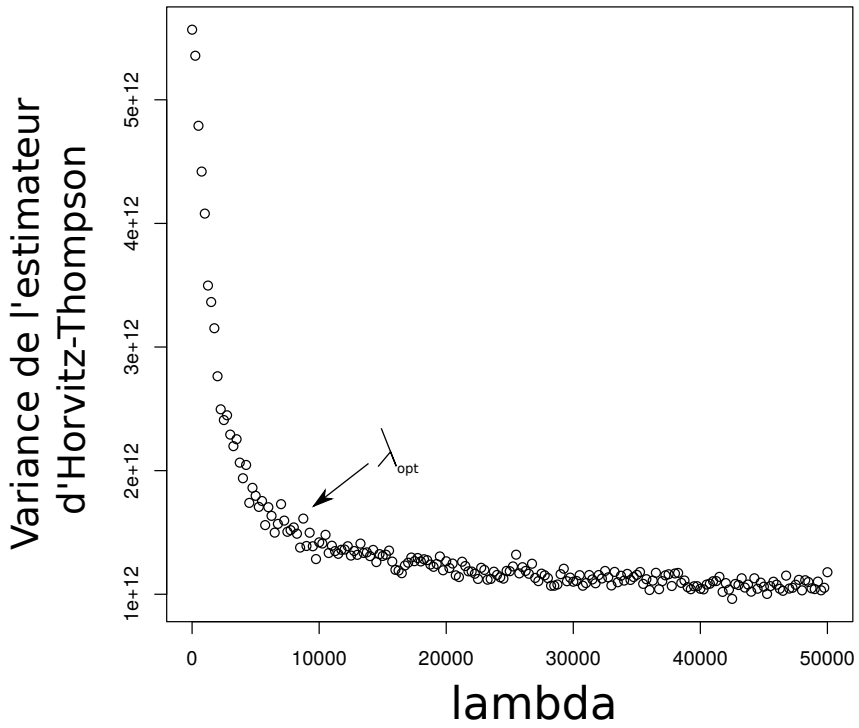
Théorème

Il existe un segment $S \subset [0, +\infty[$ *tel que :*

- $\alpha(S) = [0, 1]$, *où* $\alpha(\lambda)$ *associe à* λ *la solution du programme 4.*
- $V(\lambda)$, *la fonction de variance pour l'allocation associée à* $\alpha(\lambda)$, *est décroissante sur* S.
- *Sa dérivée seconde admet un maximum dans* S *qu'on appelle* **point de torsion**.

Forme du coude

La forme de la courbe est la suivante :



Exemple

On s'intéresse au tirage d'un échantillon de 1000 entreprises de l'industrie selon différents plans de sondages stratifiés afin de connaître le CA total du secteur. Le champ exact est défini comme suit :

- Entreprises actives situées en France.
- Entreprises dont l'effectif est compris entre 1 et 100.
- Entreprises dont le code APE commence par un code division entre 10 et 33 (sauf 12 et 19).

La population initiale est de 102 172 entreprises.

Exemple

Cette population est stratifiée selon deux critères :

1. L'APE, au niveau division (deux premiers chiffres).
2. La tranche d'effectif, de la façon suivante : 1 à 9 salariés ; 10 à 19 salariés ; 20 à 49 salariés ; 50 salariés ou plus.

ce qui constitue 88 strates.

On calcule alors les allocations proportionnelle et de Neyman relative à la dispersion du chiffre d'affaires dans chacune de ces strates, pour $n = 1000$.

Allocations initiales

Allocation	Min	Médiane	Max
Proportionnelle	1	3	278
Neyman	1	5	162

Allocation	Prop.	Neyman
Strate (10,1)	**278**	80
Strate (10,3)	18	**162**

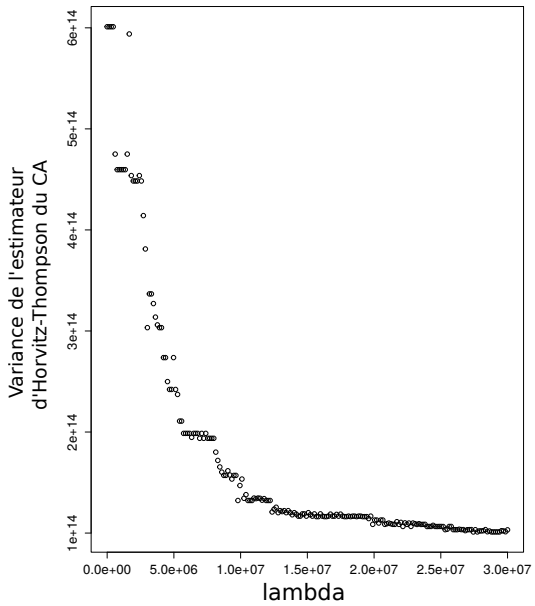
Méthode

On souhaite choisir l'allocation mixte optimale pour le problème présenté au paragraphe précédent. Pour cela on procède comme suit :

- Calculer pour différentes valeurs de λ la valeur de α solution du programme de minimisation.
- Pour chaque α, calculer l'allocation correspondante.
- Pour chacune des allocations, calculer analytiquement la variance de l'estimateur d'Horvitz-Thompson du total du chiffre d'affaire.

Résultats

On obtient finalement la courbe suivante :



On obtient $\lambda_{\text{coude}} = 1 \cdot 10^7$, ce qui donne $\alpha = 0.644$.

Allocation obtenue

Allocation	Min	Médiane	Max
Proportionnelle	1	3	278
Coude	1	4	208
Mixte	1	4	179
Neyman	1	3	162

Allocation	Prop.	Coude	Mixte	Neyman
Strate (10,1)	278	208	179	80
Strate (10,3)	18	69	90	162

Allocation obtenue

Allocation	Prop.	Coude	Mixte	Neyman
$\sigma(\hat{T}(CA)_{HT})$	24.7	12.5	11.4	9.8
Dispersion des poids	47.5	1929	3473	18585

Lorsque l'on compare l'allocation obtenue à la stratégie $\ll$ mixte $\gg$ utilisant le facteur $\alpha = \frac{1}{2}$, on remarque que la perte d'un facteur 1.1 en précision est compensée par le gain d'un facteur 1.8 en dispersion des poids.

Conclusions et extensions

Conclusion

La méthode présentée ici décrit une allocation permettant de proposer une pondération peu dispersée (lié à la qualité des estimations multiples en sondages) tout en contrôlant la précision sur les principales statistiques d'intérêt.

Une extension logique de ce travail revient à étudier les allocations de Neyman avec prise en compte de contrainte de précision locale, c'est à dire lorsqu'on souhaite une précision minimale pour la variable d'intérêt sur ces domaines.

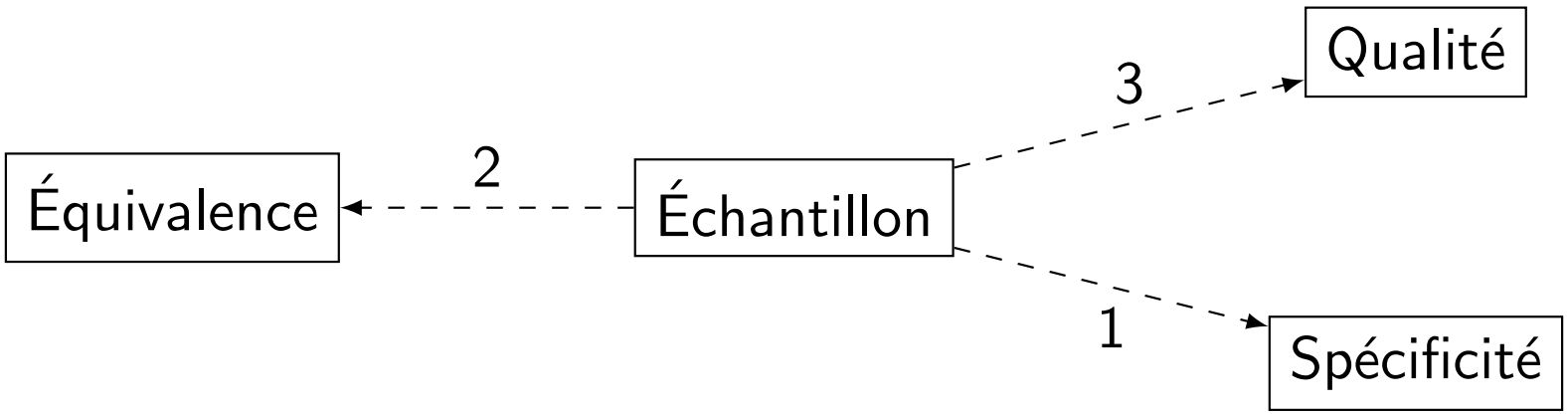
Conclusion

Merci pour votre attention !

Annexes

Principe de l'algorithme

L'algorithme CURIOS choisit l'échantillon selon des contraintes opposées :



Dans la pratique, on choisit des termes de **Qualité**, **Spécificité** et **Équivalence** qu'on intègre à un programme de minimisation.

Programme d'optimisation

On détermine l'échantillon en minimisant un programme de la forme :

$$\arg\min_{S} \ \mathbb{E}\left[\mathrm{Disp}(\mathrm{w_{CNR}}) + \lambda_1 \cdot \mathrm{Dist}(\mathrm{S}) + \lambda_2 \cdot \Gamma(\mathrm{S})\right] \qquad (7)$$

avec :

S = échantillon

w_{CNR} = poids corrigé de la non-réponse des unités de S

$\Gamma(S)$ = mesure de l'équilibre de l'échantillon de répondants

$\lambda_1, \lambda_2 \in [0, +\infty[$

Programme d'optimisation

On simplifie ici le programme :

$$\arg\min_{S} \ \mathbb{E}\left[\mathrm{Disp}(\mathrm{w_{CNR}}) + \lambda \cdot \mathrm{Dist}(\mathrm{S})\right] \tag{8}$$

avec :

$S =$ échantillon

$w_{CNR} =$ poids corrigé de la non-réponse des unités de S

$\lambda \in [0, +\infty[$

Comment choisir λ ?

Le terme λ de l'équation (8) doit être choisi pour limiter la variance de l'estimation d'Horvitz-Thompson de X.

Théorème

Soit $V(\lambda)$ la fonction de variance d'un estimateur du total de X pour les tailles d'échantillons $n_f^i(\lambda)$. Sous de bonnes hypothèses, $V(\lambda)$ est décroissante et sa dérivée seconde admet un maximum dans $]0, +\infty[$ qu'on appelle point de torsion de $V(\lambda)$.

Ce point de torsion est difficile à calculer.