Optimisation d'une allocation mixte

Thomas Merly-Alpa [1] & Antoine Rebecq [2]

[1] *INSEE, 18, bd Adolphe Pinard. 75014 Paris. thomas.merly-alpa@insee.fr*
[2] *INSEE, 18, bd Adolphe Pinard. 75014 Paris. antoine.rebecq@insee.fr*

Résumé. Cet article vise à étudier les allocations mixtes, c'est à dire mélangeant deux allocations classiques en théorie des sondages. L'utilisation de la moyenne arithmétique d'une allocation proportionnelle et d'une allocation de Neyman est un choix usuel dans le cadre des enquêtes effectuées auprès des entreprises de l'INSEE (*Institut National de la Statistique et des Études Économiques*). Nous allons nous interésser ici à la détermination d'un meilleur choix de pondération des deux allocations que (1/2, 1/2) en réalisant un programme de minimisation de la dispersion des poids sous contrainte de proximité à l'allocation de Neyman, en suivant la méthode développée pour l'algorithme CURIOS (*Curios Uses Representativity Indicators to Optimize Samples*).

Mots-clés. Échantillonnage, Calcul d'allocation, Optimisation, Neyman.

1 Introduction

Le choix d'une allocation dans le cadre d'un sondage stratifié permet de répondre à un but précis, selon la règle « un échantillon = un objectif »[7]. En effet, chacune d'entre elles vise à répondre à un besoin relatif aux caractéristiques de l'échantillon obtenu. La plus classique est l'allocation proportionnelle, qui permet d'assurer un poids équivalent à tous les individus, et donc d'assurer une bonne robustesse des résultats lors de l'analyse de plusieurs variables simultanément [6]. L'allocation de Neyman [5] est utilisée lorsqu'une variable d'intérêt principale a été identifiée, et permet de minimiser la variance de l'estimateur d'Horvitz-Thompson associé à cette variable.

D'autres allocations, moins classiques, peuvent être utilisées. Dans le cadre d'un sondage visant à interroger des salariés d'un établissement, si l'on souhaite que ceux-ci aient des poids de sondage au deuxième degré peu dispersés, il peut être utile de réaliser une allocation proportionnelle à l'effectif des établissements de chaque strate. Dans un autre registre, il est possible d'adapter l'allocation de Neyman afin d'assurer une précision minimale dans chacun des domaines de diffusion de l'enquête [3].

Néanmoins, les enquêtes ont souvent plusieurs objectifs distincts. Usuellement, ces deux objectifs sont une bonne précision pour une variable d'intérêt, mais une dispersion des poids limitée afin de garantir une bonne qualité des estimations pour d'autres variables de l'enquête. Dans ce cas, une solution consiste à prendre la moyenne arithmétique de deux allocations, i.e :

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

Cette méthode permet de combiner les bénéfices des deux méthodes à faible coût. Cependant, on peut s'interroger sur le choix du facteur 1/2 pour la moyenne des allocations. Ce papier vise à étudier une méthode basée sur un programme de minimisation faisant intervenir la dispersion des poids ainsi que la distance à l'allocation de Neyman pour choisir un paramètre α tel que l'allocation mixte « optimale » entre allocation proportionnelle et allocation de Neyman soit :

$$n_{mixte}^{opt} = \alpha n_p + (1 - \alpha) n_{Neyman} \tag{1}$$

2 Programme d'optimisation

2.1 Forme du programme

Le programme d'optimisation utilisé dans le cadre de ce papier est inspiré de celui utilisé dans l'algorithme CURIOS (*Curios Uses Representativity Indicators to Optimize Samples*) [4]. L'algorithme CURIOS réalise un arbitrage entre dispersion des poids corrigés de la non-réponse (objectif dit d'équivalence) et distance à l'échantillon initialement déterminé par l'allocation de Neyman (objectif dit de spécificité) afin de réaliser une opération de priorisation de la collecte d'enquêtes en face-à-face, en déterminant une allocation de deuxième vague.

On se place ici dans le cadre d'un tirage stratifié à H strates, et on négligera l'influence de la non-réponse [1]. Le programme de minimisation est le suivant :

$$\min_{n \in \mathbb{R}^H} \quad \text{Disp(Poids)} + \lambda \ \text{Dist}((\text{n}), (\text{n}_{\text{Neyman}})) \tag{2}$$

On s'intéresse ici à un jeu d'allocations (n_α) qui parcourent un segment entre l'allocation proportionnelle (n_p) et l'allocation de Neyman (n_{Neyman}), comme indiqué dans l'équation 1. On se limite donc à la réalisation du programme de minimisation suivant :

1. Il est possible d'intégrer facilement la non-réponse attendue par strate dans la plupart des allocations, mais cela complique inutilement l'analyse des phénomènes décrits ici.

$$\min_{\alpha\in[0,1]} \quad \text{Disp(Poids)} + \lambda \ \text{Dist}((\text{n}_\alpha), (\text{n}_{\text{Neyman}})) \tag{3}$$

où Disp est l'opérateur de dispersion autour de leur moyenne des poids de sondage, Dist une distance dans $\mathbb{R}^H$. Ce programme de maximisation dépend de la constante $\lambda \geq 0$ choisie. On remarque aisément que lorsque λ est suffisamment grand, le terme de distance devient prépondérant et on obtient $\alpha = 0$ et donc $(n_\alpha) = (n_{Neyman})$.

Étant donné que nous sommes dans le cadre d'un sondage aléatoire simple stratifié à H strates, il est possible de réécrire le programme de minimisation sous la forme suivante :

$$\min_{\alpha\in[0,1]} \quad \sum_{h=1}^{H} n_\alpha^h \left(\frac{N^h}{n_\alpha^h} - \bar{p}\right)^2 + \lambda \ \text{Dist}((\text{n}_\alpha), (\text{n}_{\text{Neyman}})) \tag{4}$$

avec :

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

le poids moyen des individus échantillonnés.

2.2 Choix de la distance

Il est possible d'utiliser plusieurs fonctions de distance pour le second terme du programme d'optimisation. Étant donné que nous sommes en dimension finie H, toutes les distances sont équivalentes : on peut donc s'attendre à obtenir des résultats assez proches pour toutes les distances. Il n'est néanmoins pas nécessaire de respecter la définition mathématique de distance, uniquement de disposer d'une mesure de la proximité. Regardons plus précisément deux cas possibles.

1. Distance euclidienne : On utilise pour Dist la distance euclidienne. Dans ce cas, le programme de minimisation peut se réécrire, en utilisant l'équation 1 :

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

2. Maximum des distances sur chacune des coordonnées : On utilise la distance définie de la façon suivate :

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

Dans ce cas, le programme de minimisation peut se réécrire :

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

Dans chacun de ces deux cas, on remarque le second terme se réécrit comme un polynôme en α, dont les coefficients sont fixés par le calcul des allocations initiales (proportionnelle et Neyman). La forme générale du problème se réécrit ici, avec $\mathcal{P}$ un polynôme :

$$\min_{\alpha\in[0,1]} \quad \sum_{h=1}^{H} n_\alpha^h \left(\frac{N^h}{n_\alpha^h} - \frac{N}{n}\right)^2 + \lambda\mathcal{P}(\alpha) \tag{7}$$

Dans la suite de cet article, on considèrera que $\mathcal{P} = X$ (cas 2, avec renormalisation de λ).

2.3 Choix du λ

Il nous faut également choisir une valeur de λ. Pour cela, nous allons nous intéresser à la variance de l'estimateur de Horvitz-Thompson du total d'une variable d'intérêt de l'enquête. En effet, l'idée est d'utiliser une propriété clef de l'allocation de Neyman, qui est son caractère plat (voir par exemple [1]). Cela signifie qu'au voisinage de l'allocation, la variance de l'estimateur du total de la variable d'intérêt de l'enquête est proche de sa valeur minimale, ce qui est satisfaisant tant d'un point de vue théorique qu'empirique. La question consiste à bien définir ce voisinage.

Nous allons nous intéresser à la variance de l'estimateur obtenu lorsque λ varie. Plus précisément, pour une valeur de λ fixée, on peut résoudre le programme d'optimisation de l'équation 3 et obtenir une valeur $\alpha(\lambda)$, et donc une allocation $n_{\alpha(\lambda)}$. À partir de cette allocation, il est possible d'étudier la variance de l'estimateur d'Horvitz-Thompson du total d'une variable d'intérêt, et en particulier de repérer le voisinage qui nous intéresse. Le théorème suivant permet de définir ce voisinage :

Théorème 1. *Soit $V(\lambda)$ la fonction de variance d'un estimateur du total de X pour l'allocation obtenu pour le α solution du programme de minimisation de l'équation 3 pour un tel λ. Alors, il existe un segment $S \subset [0, +\infty[$ tel que :*

- *$\alpha(S) = [0, 1]$, où $\alpha(\lambda)$ associe à λ la solution du programme 3.*
- *$V(\lambda)$ est décroissante sur S.*
- *Sa dérivée seconde admet un maximum dans S qu'on appelle **point de torsion**.*

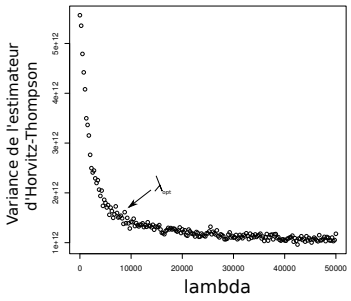

FIGURE 1 – Exemple d'un point de torsion de la fonction $V(\lambda)$ pour un plan de sondage explicité dans [4]

On veut donc prendre λ au point de torsion de la courbe, qui est aussi un point d'inflexion de sa dérivée ; en effet, cela permet d'être suffisamment proche du plateau de variance dû à la proximité de l'allocation de Neyman, qui est un optimum plat, tout en limitant au maximum la valeur de λ et donc la dispersion des poids corrigés de la non-réponse. La Figure 1 illustre ce choix.

3 Résolution du programme

3.1 Résolution analytique

Dans les cas les plus simples, c'est à dire dès lors que toute l'information est disponible, et que le tirage s'effectue à un seul degré, sans prendre en compte d'autres paramètres, la solution la plus simple pour déterminer le λ consiste à tracer la courbe de la Figure 2 en utilisant les formules analytiques classiques de calcul de variance de Y dans un plan de sondage stratifié.

Il convient ensuite de déterminer le coude de cette courbe, en suivant la logique expliquée à la fin de la partie 2.3. On aboutit donc rapidement à une valeur λ_{coude} acceptable.

3.2 Résolution approchée

Il n'est pas toujours possible de calculer analytiquement la variance, par exemple quand d'autres contraintes (regroupement de strates, etc.) rentrent en jeu. Il peut être souhaitable de rechercher une méthode *ad hoc* de calcul d'une valeur de $\lambda \ll$ acceptable $\gg$, au sens où celle-ci est à droite du coude, sur le plateau de variance. En effet, se trouver à gauche du coude induirait une variance de l'estimation du total de X bien supérieure, ce qui est à éviter, même pour gagner un peu en dispersion des poids.

3.2.1 Approche numérique

On définit λ_{num} de telle sorte que chacun des termes de l'équation 3 participe de façon égale au terme à minimiser, les deux composantes - dispersion des poids et écart à l'allocation de Neyman - étant également importantes dans le choix d'une nouvelle allocation. On écrit donc une procédure visant à égaliser les deux termes de l'équation 3. La conjecture suivante affirme que la valeur obtenue par la méthode numérique se situe à proximité du point de torsion. Il suffit ensuite de retenir une valeur de λ légèrement supérieure à λ_{num}, de manière à s'assurer que l'allocation se situe bien sur la partie plate de la courbe $V(\lambda)$ présentée en figure 1.

Conjecture 1. *On a :*

$$\lambda_{\text{num}} \sim \lambda_{\text{coude}}$$

Cette approche permet ainsi de donner une valeur de λ rapidement, via une méthode qu'il est possible d'automatiser afin de réaliser des simulations de tirage, par exemple.

3.2.2 Approche par Monte Carlo

L'approche précédente repose sur une conjecture qui pourrait ne pas être vérifiée dans certains cas. Afin d'être assuré de la qualité de l'allocation finale obtenue, il peut être préférable d'employer une autre méthode, même si elle implique un temps de calcul plus long. Il s'agit d'estimer via Monte Carlo, c'est-à-dire via la multiplication des tirages, la variance de l'estimateur du total de la variable d'intérêt, pour plusieurs valeurs de λ. Le pas utilisé pour le choix des λ testés ainsi que le nombre de simulations doivent être choisis en prenant en compte le temps de calcul, qui peut être assez long selon la population d'origine, mais également afin de s'assurer que la variance due aux simulations n'est pas trop importante, ce qui invaliderait les résultats obtenus.

Une fois ces résultats obtenus pour différentes valeurs de λ, on trace la courbe de $\text{V}(\lambda)$, que l'on espère pas trop bruitée. On peut alors afficher la courbe et placer visuellement le coude, ce qui permet de choisir la valeur de λ_{MC} finale.

4 Application

4.1 Description du plan de sondage

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. [2]
— Entreprises dont le code APE (voir [2] sur la définition du code d'Activité Principale de l'Entreprise) commence par un code division entre 10 et 33 (sauf 12 [3] et 19 [4]).

La population initiale est de 102 172 entreprises.

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 $n = 1000$. Le tableau suivant résume les caractéristiques de ces deux allocations, ainsi que les strates où l'allocation est maximale, toutes deux dans la division 10 [5] :

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

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

TABLE 1 – Description des allocations et des strates avec allocation maximale.

4.2 Résultats

On souhaite choisir l'allocation mixte optimale pour le problème présenté au paragraphe précédent. Pour cela, nous allons appliquer la méthode définie au paragraphe 3.1 :

— Calculer pour différentes valeurs de λ la valeur de α solution du programme de minimisation de l'équation 3.
— Pour chaque α, calculer l'allocation correspondante.

2. De manière générale, les entreprises ayant un fort effectif, par exemple plus de 100, sont souvent enquêtées exhaustivement. On se limite ici à la partie non exhaustive d'une enquête.
3. Produits à base de tabac.
4. Produits de la cokéfaction et du raffinage.
5. Industries alimentaires.

— Pour chacune des allocations, calculer analytiquement la variance de l'estimateur d'Horvitz-Thompson du total du chiffre d'affaire.

On obtient finalement la courbe représentée en Figure 2. On remarque que sa forme correspond globalement à ce qui était attendu en appliquant le Théorème 1. On détermine visuellement le point de torsion, qui semble situé vers $1 \cdot 10^7$. On pose donc $\lambda_{\text{coude}} = 1 \cdot 10^7$, qui se situe légèrement à droite du coude, sur la partie plate de la courbe $V(\lambda)$.

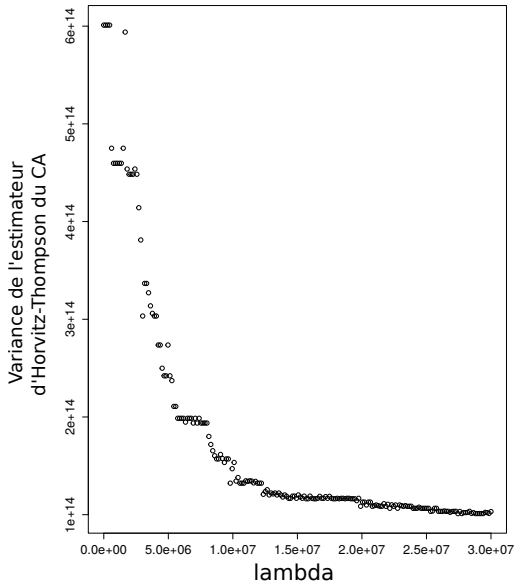

FIGURE 2 – Variance de l'estimateur d'Horvitz-Thompson du total du chiffre d'affaire.

On peut alors utiliser la valeur de λ_{coude} pour déterminer α_{coude}, à l'aide du programme d'optimisation de l'équation 3. Nous obtenons ici $\alpha_{\text{coude}} = 0.644$. Cette valeur de α obtenue peut être interprétée directement. Elle est assez proche de 0.5, ce qui montre que l'allocation finale est également assez proche de l'allocation qu'on appelle classiquement mixte, mais elle est supérieure à 0.5, ce qui montre que l'optimum du programme se rapproche sensiblement de l'allocation proportionnelle. L'allocation obtenue est décrite dans la Table 2, et comparée à l'allocation mixte usuelle utilisant la moyenne arithmétique entre les deux allocations initiales.

En termes de répartition de l'allocation dans les strates, on peut constater que l'on obtient un maximum pour la même strate que l'allocation proportionnelle (10,1), mais avec une allocation plus faible. D'autre part, l'allocation de la strate (10,3) qui est importante pour l'allocation de Neyman est bien augmentée par rapport à l'allocation proportionnelle,

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

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

TABLE 2 – Description de l'allocation obtenue.

mais reste également bien inférieure à l'allocation de Neyman. On voit bien l'apparition d'un compromis entre les allocations, comme dans le cas de l'allocation mixte usuelle.

Il reste cependant à s'intéresser aux deux critères qui motivent cette analyse, c'est à dire l'écart-type [6] de l'estimateur d'Horvitz-Thompson du total du chiffre d'affaires, ainsi que la dispersion des poids, exposés dans la Table 3.

Allocation	Proportionnelle	Coude	Mixte	Neyman
Écart-type de $\hat{T}(CA)_{HT}$	24.7	12.5	11.4	9.8
Dispersion des poids	47.5	1929	3473	18585

TABLE 3 – Gains en dispersion des poids et en variance

On remarque ici que l'allocation obtenue à l'aide de λ_{coude} a une précision assez proche de l'allocation de Neyman, alors que l'allocation proportionnelle entraîne un écart-type de l'estimateur de Horvitz-Thompson bien plus grand. Or, cette légère perte de précision est très largement contrebalancée par le gain en dispersion des poids par rapport à l'allocation de Neyman [7]. Lorsque l'on compare l'allocation obtenue à la stratégie « mixte » 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. L'allocation finale satisfait bien à nos contraintes, et répond à notre demande : avoir une bonne précision et une faible dispersion des poids.

6. À un facteur 10^6 près.

7. La dispersion des poids n'est pas nulle dans le cadre de l'allocation proportionnelle à cause des arrondis.

5 Conclusion et extensions

La méthode proposée dans ce papier permet de déterminer une allocation combinant plusieurs objectifs tels qu'une dispersion des poids minimale et une précision maximale pour l'estimation d'une variable d'intérêt. Cette optimisation permet ainsi d'obtenir un bon compromis entre ces objectifs. La méthode est facilement adaptable à des cas plus complexes tels qu'un tirage en deux degrés de salariés au sein d'entreprises, lorsqu'on souhaite minimiser la dispersion des poids attribués aux salariés interrogés.

On peut également se poser la question du lien entre la méthode proposée ici et celle proposée par [3]. En effet, la plupart des enquêtes auprès des entreprises ont également pour objectif de garantir une précision minimale dans des strates de diffusion plus agrégées. L'utilisation d'une allocation mixte entre l'allocation de Neyman sous contrainte de précision locale et une allocation proportionnelle reste à étudier, car les deux méthodes utilisent le caractère plat de l'allocation de Neyman, et il est possible que les contraintes imposées impliquent des choix incompatibles.

Références

[1] Pascal Ardilly. *Les techniques de sondage*. Editions Technip, 2006.

[2] Pascal Bihari. Module de détermination de l'activité principale exercée par une entreprise. 2011.

[3] Malik Koubi and S Mathern. Résolution d'une des limites de l'allocation de Neyman. *JMS*, 2009 :1, 2009.

[4] Thomas Merly-Alpa and Antoine Rebecq. L'algorithme CURIOS pour l'optimisation du plan de sondage en fonction de la non-réponse. *Journees de la Statistique de la SFdS, Lille*, 2015.

[5] Jerzy Neyman. On the two different aspects of the representative method : the method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, pages 558–625, 1934.

[6] Antoine Rebecq and Thomas Merly-Alpa. Pourquoi minimiser la dispersion des poids en sondage ? *preprint*.

[7] Yves Tillé. Théorie des sondages. *Dunod, Paris*, 2001.