Stratification et calcul d'allocations dans les enquêtes auprès des entreprises

Ronan Le Gleut

***Résumé* — L'objectif de cette note méthodologique est de décrire de façon rapide les méthodes usuelles de stratification et de calcul d'allocations dans les enquêtes auprès des entreprises.**

I. La base de sondage

Pour construire leurs bases de sondage, les statisticiens de l'Insee disposent du répertoire statistique Sirus « Système d'Identification au Répertoire des Unités Statistiques ». Ce dernier recense les entreprises, unités légales et établissements français et quelques-unes de leurs caractéristiques : localisation géographique, activité principale exercée, effectif salarié et chiffre d'affaires annuel déclarés à l'administration, probabilité d'existence, etc.

Ce répertoire statistique comporte quelques différences avec le répertoire administratif Sirene « Système Informatisé du Répertoire national des ENtreprises et des Établissements » qui a longtemps constitué le socle des bases de sondage des enquêtes auprès des entreprises réalisées à l'Insee. Ainsi, le répertoire Sirus identifie en plus les entreprises, qui ont un sens économique, alors que Sirene identifie les unités légales, qui ont un sens juridique.

II. Stratification

La population U est dite stratifiée quand les unités peuvent être partitionnées en H sous-populations disjointes U_1,...,U_H appelées strates (voir schéma en Figure 1). On doit pour cela disposer d'information auxiliaire sur l'ensemble de la population.

Le plan de sondage est dit stratifié quand des échantillons indépendants sont sélectionnés dans chaque strate. On tire ainsi un échantillon S_h de taille n_h dans chaque strate U_h de taille N_h. On parle de sondage aléatoire simple stratifié si des échantillons aléatoires simples sont sélectionnés dans chaque strate.

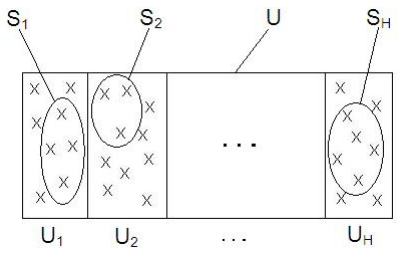

Fig. 1. La population U est dite stratifiée quand les unités peuvent être partitionnées en H sous-populations disjointes $U_1, U_2, ..., U_H$ appelées strates.

A. Quels critères de stratification ?

Les échantillons des enquêtes auprès des entreprises sont tirés selon des plans de sondages aléatoires simples stratifiés [1]. Le plus souvent, la population d'entreprises correspondant au champ de l'enquête est stratifiée en croisant deux critères [2] :

- un critère d'activité utilisant des niveaux plus ou moins fin de la nomenclature d'activités française (NAF)
- un critère de taille (utilisant des tranches d'effectifs salariés et/ou des tranches de chiffres d'affaires).

Par exemple (voir [3]), l'enquête sur les technologies de l'information et de la communication (TIC) est tirée en stratifiant selon :

- le secteur d'activité avec des niveaux d'agrégation très divers (de la classe au regroupement de sections de la NAF) ;
- la tranche d'effectif de l'entreprise (10-19, 20-49, 50-249, 250-499, 500 et +) ;
- le chiffre d'affaires ;

avec un seuil d'exhaustivité pour les plus grandes tranches d'effectif (500 salariés et +) et les plus gros chiffres d'affaires (dernier critère de stratification).

B. Comment définir les strates ?

Se pose alors la question du nombre de strates à construire qui revient ici à choisir un niveau de détail pour nos deux critères (secteur d'activité et tranche d'effectifs par exemple).

Tout d'abord, rappelons que les unités exhaustives appartiennent à une strate à part (dite « strate exhaustive ») où toutes les unités sont interrogées. Afin de définir ces strates, des seuils d'exhaustivité (en termes d'effectifs ou de chiffres d'affaires) sont souvent définis afin de forcer les plus grosses unités dans l'échantillon [3]. Des méthodes de « cut-off sampling » [2] permettent par exemple d'inclure d'office dans l'échantillon toutes les plus grosses unités permettant de couvrir un certain taux (de chiffres d'affaires par exemple) de la population.

Dans les enquêtes auprès des entreprises, il est courant que la moitié de l'échantillon soit concernée par ces seuils d'exhasutivité.

1. La nouvelle méthode de coordination des échantillons mise en oeuvre à l'Insee [1] conduit à devoir sélectionner les échantillons selon un sondage aléatoire simple stratifié.

2. Un troisième critère de localisation géographique est parfois utilisé pour le tirage mais il n'est, en général, pas pris en compte lors de l'optimisation du plan de sondage.

3. Il est également possible de forcer d'autres unités dans l'exhaustif dont on sait qu'elles ont un comportement atypique (e.g. des restructurations, des unités atypiques à dire d'expert, etc.)

Ensuite, l'objectif de la stratification est avant tout de définir des strates au sein desquelles le comportement des unités est homogène au sens de la variable d'intérêt. Pour cela, il est nécessaire que les variables servant à la stratification soient liées à la variable d'intérêt.

En d'autres termes, il s'agit de minimiser la dispersion intra de la variable d'intérêt (i.e. la dispersion à l'intérieur des strates), ou de maximiser la dispersion inter (i.e. la dispersion entre les strates).

La décomposition de la variance d'une variable d'intérêt y peut s'écrire :

$$\begin{aligned} S_y^2 &= \frac{1}{N-1}\sum_{k\in U}(y_k-\mu_y)^2 \\ &= \underbrace{\sum_{h=1}^{H}\frac{N_h-1}{N-1}S_{yh}^2}_{S_{y,intra}^2} + \underbrace{\sum_{h=1}^{H}\frac{N_h}{N-1}(\mu_{yh}-\mu_y)^2}_{S_{y,inter}^2} \end{aligned}$$

avec $S_{yh}^2 = \frac{1}{N_h-1}\sum_{k\in U_h}(y_k-\mu_{yh})^2$, $\mu_y = \frac{1}{N}\sum_{k\in U} y_k = \sum_{h=1}^{H}\frac{N_h}{N}\mu_{yh}$ et $\mu_{yh} = \frac{1}{N_h}\sum_{k\in U_h} y_k$.

Des méthodes de découpage optimal afin de définir des bornes de stratification permettant de minimiser la variance (de la variable de stratification[4]) existent. On peut par exemple citer la méthode de Dalenius [5], la méthode géométrique proposée par Gunning et Horgan [6] ou encore la méthode de Lavallée-Hidiroglou [7], où les strates sont définies par les valeurs d'une variable quantitative bien corrélée à la variable d'intérêt. Cette dernière méthode permet d'ailleurs de définir un seuil optimal à partir duquel toutes les unités peuvent être considérées comme exhaustives.

En pratique, les strates sont souvent définies « à dire d'expert » selon les niveaux auxquels on souhaite publier les résultats (domaines de diffusion, e.g. niveau section ou division de la NAF, regroupements de tranches d'effectifs). Ainsi, les strates fines de tirage (définies dans la Section II-A) doivent être incluses dans des strates agrégées dites « d'optimisation » (e.g. niveau agrégé de la NAF croisé avec des tranches d'effectifs), elles-mêmes incluses dans des domaines de diffusion.

Deux niveaux de stratification sont donc généralement retenus. Le premier niveau correspond, à quelques aménagements près, aux croisements des domaines sur lesquels on souhaite diffuser des résultats. Comme on le verra dans la partie suivante, ce niveau sert souvent au calcul de taux de sondage t_h assurant une certaine précision dans chaque domaine de diffusion envisagé. Le fait que ces strates d'optimisation soient relativement agrégées permet des estimations de dispersions (et donc des calculs de précisions anticipés) robustes, puisque basées sur un nombre important d'unités.

Le second niveau qui est utilisé pour le tirage est plus fin que le premier. Plus précisément, chaque strate de tirage t est incluse dans une strate d'optimisation h. On calcule le nombre d'unités à tirer n_t dans la strate de tirage t en y appliquant le taux de sondage t_h de la strate d'optimisation correspondante :

$$n_t = t_h \times N_t$$

Cette procédure, basée sur les propriétés des allocations proportionnelles au nombre d'unités (voir Section III-A), permet d'améliorer la précision des estimations à venir si les critères de stratification sont liés aux paramètres que l'on souhaite mesurer[5].

Les strates de tirage correspondront ainsi au niveau de détail le plus fin possible[6] compte tenu de l'étendue du champ de l'enquête et de la taille d'échantillon envisagée. De cette façon, on espère obtenir des estimations au moins aussi précises que si l'on effectuait le tirage au niveau des strates d'optimisation.

Lorsque le nombre de strates de tirage est important, des méthodes de contrôle des arrondis des nombres d'unités à tirer sont mises en oeuvre afin de ne pas trop s'éloigner de la taille d'échantillon initialement visée. On utilise pour cela le logiciel τ-argus (servant initialement à l'anonymisation des données, voir [8]) ou encore la méthode de Cox [9].

III. Calcul d'allocations

On suppose que la taille globale d'échantillon n est fixée, et que les strates ont été définies. On doit choisir les tailles $n_1, n_2, ..., n_H$ des sous-échantillons à sélectionner dans chaque strate.

Dans le cas d'une strate exhaustive h, l'allocation n_h est égale à la taille de la strate N_h (toutes les unités sont d'office sélectionnées dans l'échantillon).

A. Les allocations proportionnelles

Avec une allocation proportionnelle **au nombre d'unités**, le taux de sondage est le même dans chaque strate :

$$f_h = \frac{n_h}{N_h} = \frac{n}{N} = f$$

On peut le réécrire sous la forme :

$$n_h = n\frac{N_h}{N}$$

Autrement dit, plus la strate est grande, plus l'échantillon sélectionné dedans est grand.

Chaque unité de la population possède la même probabilité d'inclusion $\pi_k = n/N$. Cette allocation conduit donc à un plan de sondage *auto-pondéré* où tous les individus possèdent le même poids $d_k = N/n$. Cela permet d'assurer une bonne robustesse des résultats lors de l'analyse de plusieurs variables simultanément, en particulier si ces variables sont qualitatives.

4. Il existe également des méthodes qui tiennent compte des divergences entre les variables de stratification et la variable d'intérêt (voir [4]).

5. Même lorsque les critères de stratification ne s'avèrent pas liés aux paramètres que l'on souhaite mesurer, cette procédure ne dégrade pas (sauf cas très particuliers) les estimations à venir.

6. En pratique, pour faciliter les traitements post-enquêtes (notamment les calculs de précision), les responsables d'enquêtes souhaitent généralement imposer un nombre minimum d'unités tirées dans chaque strate de tirage.

La variance de l'estimateur stratifié du total d'une variable d'intérêt y avec allocation proportionnelle est donnée par :

$$\mathbb{V}_p[\hat{t}_{y\pi}] = \sum_{h=1}^{H} \mathbb{V}_p[\hat{t}_{yh\pi}] = \sum_{h=1}^{H} N_h^2 \frac{1-f_h}{n_h} S_{yh}^2$$
$$= \frac{1-f}{n} N^2 \sum_{h=1}^{H} \frac{N_h}{N} S_{yh}^2 \approx N^2 \frac{1-f}{n} S_{y,intra}^2$$

On constate ainsi que le sondage aléatoire simple stratifié à allocation proportionnelle est (presque) toujours plus efficace que le sondage aléatoire simple (dont la formule de variance est identique en remplaçant $S_{y,intra}^2$ par S_y^2). On retrouve ici le fait que la stratification doit être choisie de façon à ce que la dispersion à l'intérieur des strates soit minimisée (voir Section II-B).

D'autres allocations proportionnelles **à une variable auxiliaire** x sont possibles, et conduisent à de meilleurs résultats que l'allocation proportionnelle au nombre d'unités si la variable auxiliaire x est corrélée positivement à la variable d'intérêt y. Pour cela, les totaux de x doivent être connus sur chaque strate :

$$n_h = n \frac{t_{xh}}{t_x} = n \frac{\sum_{k \in U_h} x_k}{\sum_{k \in U} x_k}$$

On retrouve le cas de l'allocation proportionnelle au nombre d'unités si $x_k = 1 \ \forall k \in U$.

Dans le cadre d'un sondage visant par exemple à interroger le même nombre de salariés dans chaque 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 au premier degré.

Les allocations proportionnelles à des quantités économiques x correspondent à des approximations d'allocations de Neyman (voir Section III-B), en faisant l'hypothèse que le coefficient de variation empirique de la variable x (S_{xh}/μ_{xh}) est le même dans chaque strate.

B. Les allocations de Neyman

L'allocation de Neyman selon une variable d'intérêt, très largement documentée dans la théorie des sondages et régulièrement utilisée à l'Insee dans les plans de sondage des enquêtes auprès d'entreprises, optimise la précision de l'estimateur du total **de cette variable d'intérêt au niveau de l'ensemble de la population**.

On cherche donc à résoudre un problème de minimisation (de la variance) sous contraintes (taille globale d'échantillon n fixée) :

$$\begin{cases} \min\limits_{n_1,\ldots,n_H} \mathbb{V}_p[\hat{t}_{y\pi}] = \sum\limits_{h=1}^{H} N_h^2 \frac{1-f_h}{n_h} S_{y,h}^2 \\ s.c. \ \sum\limits_{h=1}^{H} n_h = n \end{cases}$$

L'allocation en sortie de ce programme de minimisation est la suivante :

$$n_h = n \frac{N_h S_{yh}}{\sum_{j=1}^{H} N_j S_{yj}}$$

L'allocation de Neyman indique qu'il faut sélectionner un échantillon plus grand :

— dans les grandes strates ;
— dans les strates présentant une forte dispersion.

L'allocation est optimale pour la variable d'intérêt y et quasi-optimale pour des variables corrélées positivement à y. Cependant, pour des variables corrélées négativement ou non corrélées à la variable d'intérêt, elle peut conduire à des résultats plus imprécis que l'allocation proportionnelle (voire que le sondage aléatoire simple).

L'allocation de Neyman peut conduire à des tailles d'échantillon supérieures aux tailles de strates, si ces dernières présentent une forte dispersion et/ou sont de grande taille. Dans ce cas :

— on effectue un recensement dans les strates concernées (on fixe $n_h = N_h$) ;
— on recalcule l'allocation dans les autres strates.

Un problème alternatif peut être d'optimiser la précision sous une contrainte de coût global C fixé :

$$C_0 + \sum_{h=1}^{H} C_h n_h = C$$

où C_0 donne le coût fixe de l'enquête, et C_h le coût associé à la collecte d'une unité de U_h.

Il est également possible d'intégrer des taux de réponse anticipés dans le calcul de l'allocation de Neyman. Il est alors usuel de récupérer les taux de réponse par strate issus d'une édition précédente de l'enquête ou d'une enquête portant sur le même champ.

Cette allocation de Neyman, dans sa forme « traditionnelle », ne répond en général qu'en partie aux objectifs d'une enquête car, comme nous l'avons vu dans la Section II-B, la publication des totaux de la variable est non seulement réalisée au niveau de l'ensemble de la population mais aussi à des niveaux intermédiaires appelés domaines de diffusion et correspondant à des sous-parties de la population (certaines activités seulement, certaines tailles d'entreprises seulement...).

Rien ne garantit que l'allocation de Neyman soit performante dans ces sous-parties. En particulier, les entreprises des secteurs correspondant à des montants peu importants (ou plus homogènes) relativement aux autres risquent d'être peu nombreuses dans l'échantillon et la précision des estimations limitées à ces entreprises risque de s'avérer insuffisante.

Ainsi, les taux de sondage correspondant aux enquêtes auprès d'entreprises réalisées par l'Insee sont de plus en plus basés sur une variante de l'allocation de Neyman introduisant **des contraintes de précision locales** [10]. Cette variante proposée par Koubi et Mathern (2009) optimise la précision de l'estimateur du total de la variable d'intérêt au niveau de l'ensemble de la population en garantissant une précision minimale dans chaque domaine de diffusion. Le cas des strates saturées ($n_h > N_h$) est également traité dans

cet algorithme.

Le programme de minimisation résolu par l'algorithme peut donc s'écrire :

$$\begin{cases} \min\limits_{n_1,\dots,n_H} \mathbb{V}_p[\hat{t}_{y\pi}] = \sum\limits_{h=1}^{H} N_h^2 \dfrac{1-f_h}{n_h} S_{y,h}^2 \\ s.c. \ \sum\limits_{h=1}^{H} n_h = n, \ n_h \leq N_h \\ s.c. \ \max\limits_{d\in D} CV_d \leq CV_{loc} \end{cases}$$

avec D l'ensemble des domaines de diffusion et CV_{loc} le coefficient de variation maximal attendu.

Pour utiliser cette méthode, les statisticiens de l'Insee doivent estimer les dispersions de la variable sur laquelle sera basée l'allocation de Neyman sous contraintes. Si l'enquête porte sur un thème nouveau, la pratique la plus courante est d'optimiser l'allocation sur une variable connue dans la base de sondage (chiffre d'affaires ou effectif salarié) et supposée liée aux variables d'intérêt de l'enquête. Lorsqu'il existe des éditions précédentes de l'enquête, les résultats de ces dernières sont généralement utilisés pour estimer les dispersions.

C. Les allocations mixtes

Les enquêtes ont souvent plusieurs objectifs distincts. Usuellement, ces deux objectifs sont une bonne précision pour une variable d'interêt, 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 :

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

Cette allocation permettant de combiner les bénéfices des deux méthodes à faible coût est évoquée dans le manuel de Cochran [11].

Si l'on reprend l'exemple de l'enquête TIC (voir [3]), le choix a été fait d'une allocation mixte correspondant à une moyenne entre :

— une allocation proportionnelle au nombre d'unités en garantissant, pour chaque activité, une demi-longueur de l'intervalle de confiance d'au plus 10 points pour l'estimation d'une proportion et en imposant un minimum de 10 unités tirées par strate ;
— une allocation proportionnelle au nombre de personnes occupées (en imposant dans chaque strate un nombre minimum d'unités à tirer).

L'allocation proportionnelle au nombre d'unités vise à répondre à un objectif de précision sur les variables de type proportion. Il s'agit d'un cas particulier d'allocation de Neyman sous contraintes locales présentée en Section III-B. L'allocation de Neyman est calculée sur une variable indicatrice dont la dispersion (ou écart-type empirique S_y) est estimée à 0.5 dans chaque strate [7], et les contraintes locales correspondent à une demi-longueur de l'intervalle de confiance de 10 points, par activité (domaine de diffusion), pour l'estimation de la proportion correspondant à cette variable.

L'allocation proportionnelle au nombre de personnes occupées vise à répondre à un objectif de précision relatif aux variables de type montants (en favorisant les strates contenant les entreprises de grande taille).

On peut cependant s'interroger sur le choix du facteur $1/2$ pour la moyenne des allocations. Le papier de Merly-Alpa et Rebecq [12] vise justement à étudier une methode 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 n_{prop} et allocation de Neyman n_{Neyman} soit :

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

7. Il s'agit là d'une majoration de la dispersion d'une variable indicatrice y : $S_y = \sqrt{\frac{N}{N-1}P(1-P)} \approx \sqrt{P(1-P)}$ avec $P = \frac{1}{N}\sum_{k\in U} y_k = 0.5$ (pas d'a priori sur la valeur de la proportion à estimer).

REFERENCES

[1] Gros, E., Merly-Alpa, T. (2016). La coordination d'échantillons. *Note de méthodologie du DMS - Insee*.

[2] Särndal, C. E., Swensson, B., Wretman, J. (2003). Model assisted survey sampling. *Springer Science & Business Media*, pp. 531-533.

[3] Demoly, E., Fizzala, A., Gros, E. (2014). Méthodes et pratiques des enquêtes entreprises à l'Insee. *Journal de la Société Française de Statistique*, vol. 155, No 4, pp. 134-159.

[4] Rivest, L.P. (2002). Une généralisation de l'algorithme de Lavallée et Hidiroglou pour la stratification dans les enquêtes auprès des entreprises. *Techniques d'enquête*, vol. 28, No 2, pp. 207-214.

[5] Dalenius, T., Hodges Jr, J.L. (1959). Minimum variance stratification. *Journal of the American Statistical Association*, vol. 54, No 285, pp. 88-101.

[6] Gunning, P., Horgan, J.M. (2004). A new algorithm for the construction of stratum boundaries in skewed populations. *Survey Methodology*, vol. 30, No 2, pp. 159-166.

[7] Lavallee, P., Hidiroglou, M.A. (1988). On the stratification of skewed populations. *Survey Methodology*, vol. 14, No 1, pp. 33-43.

[8] De Wolf, P.P., Hundepool, A., Giessing, S., Salazar, J.J., Castro, J. (2014). τ-argus User's manual. *Argus Open Source-project*, pp. 28-30.

[9] Cox, L.H. (1987). A constructive procedure for unbiased controlled rounding. *Journal of the American Statistical Association*, vol. 82, No 398, pp. 520-524.

[10] Koubi, M., Mathern, S. (2009). Résolution d'une des limites de l'allocation de Neyman. *Journées de Méthodologie Statistique, Paris*.

[11] Cochran, W.G. (1977). *Sampling Techniques, third edition*, pp. 119-120.

[12] Merly-Alpa, T., Rebecq, A. (2016). Optimisation d'une allocation mixte. *9ème colloque francophone sur les Sondages, Gatineau*.

Département des méthodes statistiques
Version n° 1, diffusée le 11 septembre 2017.