LES STATISTICIENS PEUVENT T-ILS S'ACCORDER SUR UNE DÉFINITION DU CONCEPT D'ÉCHANTILLON EXTRAPOLABLE ?

Léo Gerville-Réache

LES STATISTICIENS PEUVENT T-ILS S'ACCORDER SUR UNE DEFINITION DU CONCEPT D'ECHANTILLON EXTRAPOLABLE ?

Léo Gerville-Réache[1]

[1]*Université de Bordeaux, CNRS, UMR 5251, France, leo.gerville@u-bordeaux2.fr*

Résumé : La théorie des sondages se concentre sur l'estimation statistique de quantités dans une population finie à partir de la sélection et de l'observation d'une partie de cette population. Tillé (2001) propose la définition suivante d'un sondage : « *On appelle sondage, toute étude partielle d'une population donnée en vue de son extrapolation à la totalité de celle-ci* ». Qu'entend-on par extrapolation ? Comment nommer, sélectionner et observer cette partie de la population qui permet cette extrapolation ? Partant de quelques citations marquantes sur ces questions, cette communication tente de poser précisément le problème ainsi que la solution qui en découle.

Mots-clés. Plan de sondage, représentativité, extrapolation, estimateur sans biais linéairement invariants.

Abstract. Sampling theory focuses on the statistical estimation of finite population quantities from the selection and observation of a subset of this population. Tillé (2001) proposes the following definition of a survey: "We call survey, any partial study of a population for extrapolation to the whole of it." What is extrapolation? How to name, select and watch this part of the population that makes this extrapolation? From a few notable quotes on these issues, this paper attempts to ask precisely the problem and the solution that follows.

Keywords. Sampling, representativeness, extrapolation, unbiased estimator linearly invariant.

1 Echantillon représentatif ou échantillon extrapolable ?

Dans la littérature statistique, le mot extrapolable, associé à celui d'échantillon bien moins fréquent que celui de représentatif. Pour certain auteur, le concept de représentativité est pourtant intimement lié à celui d'extrapolation statistique.

Pour J. Neyman, en 1934, « *Si nous nous intéressons à un caractère collectif X d'une population π et utilisons des méthodes d'échantillonnage et d'estimation nous permettant d'attribuer à chaque échantillon Σ possible un intervalle de confiance $X1(\Sigma)$, $X2(\Sigma)$ tel que la fréquence d'erreur dans l'affirmation que $X1(\Sigma) \leq X \leq X2(\Sigma)$ ne dépasse pas la limite $1 - \varepsilon$ déterminée à l'avance, quelles que soient les propriétés de la population, j'appelle cette méthode d'échantillonnage représentative.*»

Selon Neyman, c'est bien la possibilité d'extrapolé à la population, avec une précision calculable, les résultats de l'échantillon qui confère à la méthode d'échantillonnage sont caractère représentatif.

Pour H. Chernoff (1960) : « *On dira qu'un échantillon est représentatif d'une population si la*

distribution des données est précisément celle d'un échantillon pris au hasard avec remplacement dans la population ».

Selon Chernoff, un échantillon représentatif est un échantillon aléatoire simple. Il est clair que cette définition est plus restrictive que celle de Neyman. Elle permet néanmoins, comme Neyman le sous entend, l'extrapolation.

Pour Philippe Dutarte (2005) « *Voilà une expression qui, si elle n'est pas précisée, peut signifier à peu près n'importe quoi. Un échantillon constitué selon la méthode des quotas est évidemment « représentatif » des critères correspondants aux quotas (sexe, âge, catégorie socioprofessionnelle, région, taille de la commune...) selon lesquels il a été fabriqué. Mais on n'a aucun moyen de savoir jusqu'à quel point il est « représentatif » de ce pour quoi il a été prélevé, c'est-à-dire le sujet du sondage, l'opinion, le pourcentage que l'on cherche à évaluer. L'expression « représentatif de la population française », que l'on lit souvent dans la presse, prête évidemment à confusion. On a l'impression que l'échantillon est « représentatif» de tout ce que l'on veut. En statistique, on désigne plutôt par « échantillon représentatif », un échantillon où le hasard permet d'éviter les biais inconnus et d'appliquer le calcul des probabilités. La méthode optimale pour obtenir un échantillon « représentatif » est celle du sondage aléatoire stratifié optimal.* ».

Selon Dutarte, le terme représentatif est ambigu. La représentativité des certaines caractéristiques n'engendre pas nécessairement la représentativité des variables d'intérêt. Il conclut néanmoins par une définition qui sous entends la possibilité d'extrapolation via l'absence de biais et le calcul des probabilités.

Pour P. Ardilly (2006), « *On dit qu'un plan de sondage est représentatif d'une expression donnée et numériquement connue construite à partir d'une variable auxiliaire (un total le plus souvent) lorsque, pour la méthode d'échantillonnage choisie, l'estimateur estime parfaitement bien (c'est-à-dire avec un biais nul et une variance nulle) l'expression donnée en question. Ce terme ne s'applique pas de façon satisfaisante à un échantillon.* »

Selon Ardilly, le plan de sondage est représentatif d'une quantité si cette quantité est estimée sans erreur. L'idée de l'extrapolation est ici poussée à l'extrême. L'auteur conclu que cette définition ne s'applique pas de façon satisfaisante à un échantillon.

Pour finir cette revue de citation, qui n'est qu'une infime partie de la littérature, Tillé (2001) écrit, « *Le concept de représentativité est aujourd'hui à ce point galvaudé qu'il est désormais porteur de nombreuses ambivalences. Cette notion, d'ordre essentiellement intuitif, est non seulement sommaire mais encore fausse et, à bien des égards, invalidée par la théorie.* ».

Selon Tillé, le terme à conduit à tellement d'approximation et d'erreur qu'il ne peut plus faire partie du vocabulaire de la théorie des sondages. Ce point de vue est légitime mais, en l'absence de définition statistique, nous risquons de l'abandonner aux utilisateurs peu scrupuleux.

Ce n'est pourtant pas la première fois que ce problème se pose aux statisticiens : définir statistiquement un terme du langage courant dont tout le monde (ou presque), statisticien ou non, à une idée, à la fois, précise et floue. Prenons le terme « significatif ». Dans le langage courant, le Larousse propose : Significatif : « *qui exprime quelque chose nettement, sans ambiguïté* ». La communauté statistique s'est approprié le terme est lui à donnée une définition statistique que Wikipédia résume ainsi *« En statistiques, un résultat est dit statistiquement significatif lorsqu'il est improbable qu'il puisse être obtenu par un simple hasard.* ». C'est bien le principe fondateur de la théorie des tests.

Mais arrêtons ici le débat sur terme d'échantillon représentatif et concentrons-nous sur celui d'échantillon extrapolable. Ce terme étant actuellement bien moins usité, nous espérons qu'il ne focalisera pas les passions et permettra une réflexion apaisée.

2 Estimateur linéairement invariant en théorie des sondages

Comme nous l'avons vue précédemment, nous cherchons à définir le concept d'échantillon extrapolable. Cette extrapolation passe nécessairement par les propriétés statistiques des estimateurs qui seront construit à partir de l'échantillon.

Les notations et définitions de cette partie sont celles de Tillé (2001). Soit *U*, une population de *N* unités : $U = \{1,...,k,...,N\}$. Soit $\mathbf{y} = (y_1,...,y_k,...,y_N)'$, les valeurs prises par la variable d'intérêt *y* par les unités de *U*. Soit $\mathbf{Y} = (\mathbf{y}_1,...,\mathbf{y}_i,...,\mathbf{y}_I)$, l'ensemble des valeurs prises par l'ensemble des variables (d'intérêt ou pas) des unités de *U*.

Définition 1 : Un échantillon non-ordonné sans remise *s* est un n-uplet non-ordonné sans remise d'unité de *U*. *s* est un sous-ensemble (au sens ensembliste) de *U*. On note $\mathcal{S} = \{s | s \subset U\} \setminus \varnothing$.

Définition 2 : Un plan de sondage (non ordonné) sans remise $p(.)$ est une loi de probabilité sur $\mathcal{S}$ telle que :

$$p(s) \geq 0 \text{ pour tout } s \in \mathcal{S} \quad \text{et} \quad \sum_{s \in \mathcal{S}} p(s) = 1.$$

Soit *S*, la variable aléatoire prenant les valeurs $s \in \mathcal{S}$ avec probabilité $p(s) \geq 0$.

Définition 3 : Un plan de sondage est de taille fixe *n* si

$$\mathcal{S} = \{s \subset U | \# s = n\}$$

Définition 4 : La probabilité d'inclusion π_k est la probabilité pour une unité d'appartenir à l'échantillon. Ces probabilités découlent du plan de sondage :

$$\pi_k = \sum_{s \supset k} p(s), \text{ pour tout } k \in U$$

Dans la suite, on supposera que $\pi_k > 0$ pour tout $k \in U$.

Définition 5 : La probabilité d'inclusion π_{kl} est la probabilité pour un couple d'unités d'appartenir à l'échantillon. Ces probabilités découlent du plan de sondage :

$$\pi_{kl} = \sum_{s \supset k, s \supset l} p(s), \text{ pour tout } k,l \in U, k \neq l$$

Définition 6 : Un estimateur d'une moyenne est dit linéaire s'il peut s'écrire sous la forme

$$\hat{\bar{y}}_L = w_0(S) + \sum_{k \in S} w_k(S) y_k$$

Définition 7 : Un estimateur d'une moyenne est dit linéaire homogène s'il peut s'écrire sous la forme

$$\hat{\bar{y}}_{LH} = \sum_{k \in S} w_k(S) y_k$$

Définition 8 : Un estimateur d'une moyenne est dit linéairement invariant si pour tout changement d'origine et d'unité du vecteur $\mathbf{y} = (y_1,...,y_k,...,y_N)'$ en $\mathbf{y}^* = (a + by_1,...,a + by_k,...,a + by_N)'$ alors,

$$\hat{\bar{y}}^*_{LI} = a + b\hat{\bar{y}}_{LI}$$

Dans la suite on se place dans le cadre des plans de sondages sans remise de taille fixe (selon les définitions 2 et 3).

3 Echantillon extrapolable : définition et propriétés

Avant de proposer une définition, il est nécessaire de s'accorder un point de départ. Nous devons convenir qu'un échantillon aléatoire simple, (issu d'un plan de sondage où chaque échantillon à la même probabilité de sélection) est un échantillon extrapolable. En effet, il est difficile d'accepter la théorie des sondages, et de ne pas supposé qu'a minima, un échantillon aléatoire simple n'est pas extrapolable.

On se doit alors de remarquer que la propriété fondamentale d'un plan simple est que la fonction de répartition théorique de la variable d'intérêt dans l'échantillon est égale à sa fonction de répartition empirique dans la population. Cette propriété est le fils conducteur de la définition qui suit.

Définition d'un échantillon extrapolable : Un échantillon *s* est extrapolable à la population *U* si et seulement si, il est issu d'un plan d'échantillonnage permettant de construire une deuxième phase au dit plan (produisant un échantillon d'une unité, noté S') tel que, quelque soit la variable d'intérêt *y*, on a

$$P\left(y(S')=y_0\right)=\frac{1}{N}\sum_{k=1}^{N}1_{\{y=y_0\}}\ ,\ \forall\ y_0\in\left(y_1,...,y_k,...,y_N\right).$$

Ici, on note $y(S')$ la variable aléatoire généré par le plan de sondage en deux phases.

Cette définition n'impose pas que la fonction de répartition théorique de la variable d'intérêt dans l'échantillon soit égale à sa fonction de répartition empirique dans la population, elle impose, d'une certaine manière, la possibilité de la retrouver.

Corolaire 1 : Un échantillon *s* est extrapolable à la population *U* si et seulement si :

$$P(k\in S')=\pi_k E\left[P(k\in S'|S)\right]=\frac{1}{N},\ \text{pour tout } k\in U$$

Preuve : Si est clair que si $P(k\in S')=\frac{1}{N},$ pour tout $k\in U$ alors, $P\left(y(S')=y_0\right)=\frac{1}{N}\sum_{k=1}^{N}1_{\{y=y_0\}}$, $\forall\ y_0\in\left(y_1,...,y_k,...,y_N\right)$. Si $P\left(y(S')=y_0\right)=\frac{1}{N}\sum_{k=1}^{N}1_{\{y=y_0\}}$, le cas le plus générale est celui où les valeurs de $y(S')$ sont toutes différentes. Dans ce cas $P\left(y(S')=y_0\right)=\frac{1}{N},\ \forall\ y_0\in\left(y_1,...,y_k,...,y_N\right)$ et, comme dans ce cas, $P\left(y(S')=y_0\right)=P(k\in S')$, on a bien $P(k\in S')=\frac{1}{N},$ pour tout $k\in U$.

A sa manière, ce corolaire nous dit qu'un échantillon est extrapolable s'il est « équivalent » à un échantillon aléatoire simple.

NB : Il est clair qu'une stratification proportionnelle aux effectifs produit des échantillons extrapolables.

Corollaire 2 : Un échantillon *s* est extrapolable à la population *U* si et seulement si il existe des $w_k(s)\in]0;1]$ pour tout $k\in U$ et tout s $\subset\mathcal{S}$ tels que:

$$\sum_{s\supset k}w_k(s)\,p(s)=1/N\ \text{pour tout } k\in U\ \text{ avec }\sum_{k\in s}w_k(s)=1,\ \text{pour tout } s\in\mathcal{S}$$

Preuve : Posons $w_k(s) = P(k \in S' | s)$, alors on a clairement $\sum_{k \in s} w_k(s) = 1,$ pour tout $s \in \mathcal{S}$ et on a également $E\left[P(k \in S' | S) \right] = \sum_{s \supset k} w_k(s) p(s) = 1/N$ pour tout $k \in U$.

Ce corolaire nous montre que l'extrapolation est conditionnée par les probabilités des échantillons et non par les seules probabilités d'inclusion des individus.

Propriété 1 : Si un échantillon *s* est extrapolable à la population *U* alors existe des $w_k(s) \in]0;1]$ pour tout $k \in U$ et tout $s \subset \mathcal{S}$, avec $\sum_{k \in s} w_k(s) = 1,$ pour tout $s \in \mathcal{S}$ tels que:

$$\sum_{k \in U} \sum_{s \supset k} w_k(s) p(s) = \sum_{s \in \mathcal{S}} \sum_{k \in s} w_k(s) p(s)$$

Preuve : Si $\sum_{s \supset k} w_k(s) p(s) = 1/N$ pour tout $k \in U$, alors $\sum_{k \in U} \sum_{s \supset k} w_k(s) p(s) = \sum_{k \in U} 1/N = 1$ et comme $\sum_{k \in s} w_k(s) = 1,$ pour tout $s \in \mathcal{S}$, alors $\sum_{s \in \mathcal{S}} \sum_{k \in s} w_k(s) p(s) = \sum_{s \in \mathcal{S}} p(s) = 1$.

Cette propriété nous dit que, si l'échantillon est extrapolable, les unités *k* et les échantillons *s* jouent des rôles « symétriques ».

Propriété 2 : Si l'échantillon *s* est extrapolable à la population *U* alors il existe un estimateur sans biais linéairement invariant pour la moyenne de la variable d'intérêt.

Preuve : soit $\hat{\bar{y}}_R$ l'estimateur linéaire homogène suivant :

$$\hat{\bar{y}}_R = \sum_{k \in S} w_k(S) y_k ,$$

de la condition $\sum_{s \supset k} w_k(s) p(s) = 1/N$ pour tout $k \in U$, on déduit que $\hat{\bar{y}}_R$ est sans biais :

$$E\left(\hat{\bar{y}}_R\right) = E\left(\sum_{k \in S} w_k(S) y_k \right) = \sum_{s \subset \mathcal{S}} p(s) \sum_{k \in s} w_k(s) y_k = \sum_{s \subset \mathcal{S}} \sum_{k=1}^{N} p(s) w_k(s) y_k 1_{\{k \in s\}}$$
$$= \sum_{k=1}^{N} \sum_{s \subset \mathcal{S}} p(s) w_k(s) y_k 1_{\{k \in s\}} = \sum_{k=1}^{N} y_k \sum_{s \subset \mathcal{S}} p(s) w_k(s) 1_{\{k \in s\}} = \sum_{k=1}^{N} y_k \sum_{s \supset k} w_k(s) p(s) = \sum_{k=1}^{N} y_k / N$$

Et de la condition $\sum_{k \in s} w_k(s) = 1,$ pour tout $s \in \mathcal{S}$, on déduit que $\hat{\bar{y}}_R$ est linéairement invariant :

$$\hat{\bar{y}}^*{}_R = \sum_{k \in S} w_k(S)(a + b y_k) = \sum_{k \in S} a w_k(S) + b w_k(S) y_k$$
$$= \sum_{k \in S} a w_k(S) + \sum_{k \in S} b w_k(S) y_k = a \sum_{k \in S} w_k(S) + b \sum_{k \in S} w_k(S) y_k = a + b \hat{\bar{y}}_R$$

En 1977 Patel et Dharmadhikari ont établi les conditions nécessaires et suffisantes (CNS), sur les probabilités d'inclusion de premier et second ordre, de l'existence d'un estimateur sans biais linéairement invariant pour une moyenne ou une somme. Ils montrent en particulier, comme le sous entend l'expression la variance de l'estimateur d'Horvitz-Thompson, que la connaissance des probabilités des échantillons n'est pas nécessaire, celle des probabilités de premier et deuxième ordres est nécessaire et suffisante. On a ainsi le corolaire suivant.

Corollaire 3 : Un échantillon *s* est extrapolable à la population *U* si et seulement si les probabilités

d'inclusion de premier ordre π_k et de second ordre $\pi_{k,l}$ pour tout $k,l \in U, k \neq l$ sont telles que la matrice C_{NxN}, défini par $c_{kk} = (n-1)\pi_k / n$ et $c_{kl} = -\pi_{kl} / n$ est de rang N-1.

NB : si l'échantillon est stratifié en m classes, alors l'échantillon sera extrapolable si et seulement si le rang de C est N-m.

4 Conclusion

La définition proposée exprime la volonté « d'équivalence » entre la fonction de répartition théorique de la variable d'intérêt dans l'échantillon et sa fonction de répartition empirique dans la population. Nous avons traduit cette « équivalence » par la l'existence théorique d'une deuxième phase de sondage conduisant, au final, à un échantillonnage aléatoire simple de taille un.

Le concept d'échantillon extrapolable à une population n'est pas une question philosophique. C'est une question mathématique dont les contraintes peuvent être formalisées et qui produit un espace de développement théorique propre. Il est notable que cette traduction « intuitive » de la notion d'échantillon extrapolable conduise à une condition suffisante d'existence d'un estimateur sans biais linéairement invariant pour une moyenne (ou une somme).

Comme souvent en théorie des sondages, l'estimateur sans biais linéairement invariant qui découle de la définition d'un échantillon extrapolable n'a pas de forme explicite (sauf dans le cas d'un plan simple). Pour autant, le fait même de son existence nous semble une raison suffisante pour explorer plus avant ce domaine.

Bibliographie

[1] Ardilly P. (2006), Les techniques de sondage, Edition TECHNIP.

[2] Chernoff H (1960), A compromise between bias and variance in the use of non representative samples, Contributions to Probability and Statistics / Essays in Honor of H. Hotelling, pp. 153-167.

[3] Cochran W.G. (1977), Sampling techniques, 3rd edition, Wiley & Sons, NY.

[4] Dutarte P. (2005), L'induction statistique au lycée (ed : Didier).

[5] Kruskal W., Mosteller F. (1979) Representative Sampling, III: The Current Statistical Literature. International Statistical Review Vol. 47, No. 3, pp. 245-265

[6] Neyman J. (1934) 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, Vol. 97, No. 4, pp. 558-625.

[7] Patel H.C., Dharmadhikari S.W. (1977), On linear invariant unbiased estimators in survey sampling, *Sankhyā: The Indian Journal of Statistics*, Volume 39, Series C, Pt. 1, pp. 21-27

[8] Patel J.A., Patel H.C. (1993). On Balanced Sampling Designs, *Sankhyā: The Indian Journal of Statistics*, Series B, Vol. 55, No. 2, 283-287.

[9] Ross S.M. (1999), Introduction to Probability and Statistics, Elsevier

[10] Senat (Le) (2010), Proposition de loi du 14 févier 2010 relative à la publication et à la diffusion de certains sondages d'opinion.

[11] Shende, P.S., Ajgonkar, S.G. Prabhu (2002). A note on linear unbiased invariant estimator for some classes of estimators. J. Indian Soc. Agric. Stat. 55, No. 2, Article No. 1, 153-157.

[12] Tillé Y. (2001), Théorie des sondages, Edition DUNOD.