JOURNAL DE LA SOCIÉTÉ STATISTIQUE DE PARIS

L. HENRY

Le problème du choix des communes : échantillons à éléments tous distincts

Journal de la société statistique de Paris, tome 92 (1951), p. 302-308

<http://www.numdam.org/item?id=JSFS_1951__92__302_0>

Le problème du choix des communes :

Échantillons à éléments tous distincts

Dans le cas du choix des communes, le problème du plan d'échantillonnage exposé dans le *Journal de la Société de Statistique de Paris* par M. P. Thionet (mars-avril 1948, p. 136-153) s'énonce comme suit :

Une région étant divisée en communes (1, 2... i, ...k) de populations respectives N_1, N_2 ... N_i ... N_k, on doit tirer au sort l de ces communes et à l'intérieur de celles-ci des éléments en nombre n_1, n_2 ... n_i ... n_k (2).

A partir de l'échantillon ainsi obtenu on veut estimer la valeur moyenne $\bar{u}$ d'une caractéristique u des éléments, la taille, le poids, le revenu, la production par exemple.

Pour le premier tirage, celui des communes, deux solutions se présentent naturellement à l'esprit : soit donner à chaque commune la même probabilité d'être tirée, soit lui donner une probabilité de sortie proportionnelle à sa population.

---

(1) La ligne de régression mutuelle ainsi obtenue est fréquemment appelée « ligne du meilleur ajustement » (line of best fit). C'est à tort semble-t-il que Mrs Helen Walker qualifie de la sorte la droite obtenue par la régression unilatérale en utilisant le procédé des moindres carrés.

Cf. Mrs Helen WALKER : *Mathematics essential for elementary Statistics*, 1934, p. 193 et suivantes.

(2) Le mot population est pris ici au sens large. Les éléments dont elle se compose peuvent être des personnes, des animaux, des bâtiments, des exploitations.

Dans les deux cas on tire les communes au moyen d'une loterie où pour un total de L billets la commune (i) est représentée par L_i d'entre eux.

Dans le premier cas $\frac{L_i}{L} = \frac{1}{k}$; en théorie L peut être différent de k, mais en pratique on se contente d'attribuer un seul billet à chaque commune et l'on a $L = k$.

Dans le second cas, on a $\frac{L_i}{L} = \frac{N_i}{N}$ avec $N = \Sigma N_i$.

Pour le second tirage, celui des éléments à extraire des communes désignées par le premier tirage, chaque élément de ces communes a la même probabilité d'être tiré. Mais si l'on veut, dans le cadre du mode de tirage adopté, et compte tenu des crédits disponibles, réduire au minimum la variance de l'estimation cherchée, le nombre d'éléments à tirer de chaque commune dépend du mode de tirage de ces communes.

Jusqu'ici nous ne nous sommes pas préoccupé de la précision relative des deux modes de tirage. Or, on a un gros intérêt à adopter le deuxième (à probabilité de sortie de chaque commune proportionnelle à sa population), car la variance de l'estimation obtenue est plus petite que pour le premier. C'est donc ce deuxième mode de tirage qui doit particulièrement retenir l'attention.

Mais dans ce mode de tirage, plusieurs communes sinon toutes, sont représentées par plus d'un billet; il peut donc arriver qu'une des communes sorte plusieurs fois. Que doit-on faire dans ce cas? La solution classique, à laquelle correspondent les formules données dans l'article de M. Thionet, consiste à faire pour chaque billet sorti un tirage d'éléments *indépendant* de tous les autres. Si une commune (i) sort h fois on procédera donc à h tirages indépendants de n_i éléments parmi les N_i que compte la commune.

Il arrivera donc, et ceci d'autant plus souvent que n_i sera plus grand par rapport à N_i, que le même élément figure dans plus d'un des h tirages effectués. Autrement dit, l'échantillon obtenu ne comprendra pas toujours des éléments tous distincts puisque certains peuvent y figurer jusqu'à l fois si l est plus petit que L_i ou jusqu'à L_i fois si l est plus grand que L_i.

A la réflexion, il ne s'impose nullement d'admettre cette répétition possible d'éléments. Puisqu'on est assuré que certaines communes peuvent sortir plusieurs fois, pourquoi ne tirerait-on pas de la commune (i), par exemple, un nombre d'éléments *distincts* égal au produit de n_i par le nombre h de sorties de (i) au premier tirage. Ce mode opératoire paraît bien naturel et il devrait, *a priori*, augmenter la précision puisqu'en l'adoptant on accroîtrait la représentation moyenne des grosses communes.

Ces remarques conduisent à étudier les échantillons à éléments tous distincts tirés comme suit :

On tire d'abord un échantillon de communes, chacune ayant une probabilité de sortie proportionnelle à sa population.

De chaque commune (i) ainsi désignée, on tire un nombre d'éléments *tous distincts* proportionnel au nombre h de sorties de (i) au cours du premier tirage.

Avant d'aborder les deux questions essentielles — formule d'estimation à

adopter, variance de l'estimation — examinons les particularités de ce mode de tirage.

On doit tirer $h\,n_i$ éléments de la commune (i) sortie h fois; mais on ne peut obtenir $h\,n_i$ éléments tous distincts que si le nombre maximum de billets que l'on risque de sortir pour la commune (i) est tel que son produit par n_i ne dépasse pas le nombre total d'éléments de la commune. Il faut donc organiser la loterie en conséquence.

Si L_i est plus grand que l, la commune (i) peut sortir au maximum l fois et l'on doit avoir :

$$N_i \geqslant l\,n_i$$

Si L_i est plus petit que l, la commune (i) peut sortir au maximum L_i fois et l'on doit avoir :

$$N_i \geqslant L_i\,n_i$$

ou

$$1 \geqslant n_i \frac{L_i}{N_i}.$$

Si cette dernière condition est remplie, l'inégalité

$$L_i \geqslant l$$

entraîne

$$L_i \geqslant l\,n_i \frac{L_i}{N_i} \text{ donc } N_i \geqslant l\,n_i.$$

Par conséquent, on peut donner à chaque commune un nombre de billets proportionnel à N_i pourvu que le coefficient de proportionnalité soit inférieur ou égal à la plus petite des fractions $\frac{1}{n_i}$. En pratique, le nombre des billets étant entier, la proportionnalité n'est pas rigoureuse; mais ce fait reste sans grande importance si N_i est grand et les n_i assez petits.

Formule d'estimation

Appelons u_{ij} la valeur de la caractéristique u pour l'élément (j) de la commune (i). Si la commune (i) sort h fois, la quantité $\mathrm{S}_j \frac{u_{ij}}{n_i}$ (où S désigne la somme des éléments tirés de la commune i) constitue une estimation de $h\,\bar{u}_i$ ($\bar{u}_i$ étant la valeur moyenne de la caractéristique u dans la commune (i)).

Comme on a $l = \mathrm{S}\,h$, il est naturel de calculer l'estimation de $\bar{u}$ par la formule :

$$\bar{x} = \frac{1}{l} \mathrm{S}_i \mathrm{S}_j \frac{u_{ij}}{n_i}$$

où S est la somme étendue aux communes distinctes figurant dans l'échantillon.

Appelons P_{ij} la probabilité de la double combinaison S S. L'espérance mathématique E $(\bar{x})$ de $\bar{x}$ est donnée par :

$$\mathrm{E}(\bar{x}) = \frac{1}{l} \sum_{i,j} P_{ij} \left(\mathrm{S}_i \mathrm{S}_j \frac{u_{ij}}{n_i} \right)$$

où le signe Σ s'étend à toutes les doubles combinaisons possibles.

Mais, pour une même combinaison de communes obtenue au premier tirage, on peut avoir toutes les combinaisons des éléments de ces communes de sorte que :

$$\mathrm{E}\,(\bar{x}) = \frac{1}{l}\,\mathrm{E}\,(\mathrm{S}_i\, h\,\bar{u}_i) = \frac{1}{l}\,\Sigma_i\,\mathrm{P}_i\,(\mathrm{S}_i\, h\,\bar{u}_i)$$

dans laquelle P_i est la probabilité de la combinaison simple S_i, le signe Σ_i étant étendu à toutes les combinaisons possibles.

$h\,u_i$ est une fonction $f\,(h,\,i)$ de h et de i et l'on peut, d'une manière plus générale, chercher l'espérance mathématique $\mathrm{E}\,(y)$, d'une quantité y, donnée par :

$$\mathrm{E}\,(y) = \frac{1}{l}\,\Sigma_i\,\mathrm{P}_i\left[\mathrm{S}_i\, f\,(h,\,i)\right].$$

Si C est le nombre de combinaisons de L billets l à l on peut écrire :

$$\mathrm{E}\,(y) = \frac{1}{\mathrm{C}l}\,\Sigma'_i\left[\mathrm{S}_i\, f\,(h,\,i)\right],$$

Σ'_i étant étendu à toutes les combinaisons de communes distinctes ou non correspondant aux C combinaisons de billets.

Si, au lieu de considérer les sommes S_i, on considère leurs termes $f\,(h,\,i)$, on a :

$$\mathrm{E}\,(y) = \frac{1}{\mathrm{C}l}\,\Sigma'_{hi}\, f\,(h,\,i)$$

Σ'_{hi} étant étendu à toutes les valeurs $f\,(h,\,i)$ distinctes ou non provenant de l'éclatement des sommes S_i.

Le nombre de fois où l'on trouve $f\,(h,\,i)$ dans Σ'_{hi} est égal au nombre de fois où i sort h fois et l'on a donc :

$$\mathrm{E}\,(y) = \frac{1}{l}\,\Sigma_{hi}\, p_{hi}\, f\,(h,\,i)$$

où p_{hi} est la probabilité pour que la commune (i) soit tirée h fois et où Σ_{hi} est étendue à tous les termes $f\,(h,\,i)$ distincts.

Appliquons ces considérations à $\mathrm{E}\,(\bar{x})$; il vient :

$$\mathrm{E}\,(\bar{x}) = \frac{1}{l}\,\Sigma_{hi}\, p_{hi}\, h\,\bar{u}_i = \frac{1}{l}\,\Sigma_i\,\bar{u}_i\,\Sigma_h\, p_{hi}\, h.$$

Mais $\Sigma_h\, p_{hi}\, h$ est égal au nombre moyen de fois où la commune (i) figure dans le tirage des l billets. On a donc :

$$\Sigma_h\, p_{hi}\, h = l\,\frac{\mathrm{L}_i}{\mathrm{L}}.$$

Il en résulte :

$$\mathrm{E}\,(\bar{x}) = \frac{1}{l}\,\Sigma_i\,\bar{u}_i\,\frac{l\,\mathrm{L}_i}{\mathrm{L}} = \Sigma_i\,\bar{u}_i\,\frac{\mathrm{N}_i}{\mathrm{N}} = \bar{u}.$$

L'espérance mathématique de l'estimation est égale à la valeur moyenne

cherchée. Cette estimation $\bar{x}$ est donc sans biais. On remarque qu'elle a été obtenue par la même formule que dans le cas classique où certains éléments peuvent être répétés; ce qui provient du fait que pour (i) sortant h fois, $S_j u_{ij}$ a comme espérance mathématique $h\, n_i\, \bar{u}_i$, que les u_{ij} soient tous distincts ou non.

Variance de l'estimation.

Si on désigne par $V(\bar{x})$ la variance de l'estimation $\bar{x}$ on a :

$$V(\bar{x}) = \frac{1}{l^2} E\left(S_i S_j \frac{u_{ij}}{n_i} - l\,\bar{u}\right)^2.$$

On peut écrire :

$$S_i S_j \frac{u_{ij}}{n_i} = S_i S_j \frac{u_{ij} - \bar{u}_i}{n_i} + S_i S_j \frac{\bar{u}_i}{n_i} = S_i S_j \frac{u_{ij} - \bar{u}_i}{n_i} + S_i\, h\, \bar{u}_i$$

d'où

$$V(\bar{x}) = \frac{1}{l^2} E\left(S_i S_j \frac{u_{ij} - \bar{u}_i}{n_i}\right)^2 + \frac{1}{l^2} E\left(S_i\, h\, \bar{u}_i - l\,\bar{u}\right)^2$$

les termes rectangles disparaissant puisque à chaque combinaison S_i est associée une double combinaison $S_i S_j$ dont l'espérance mathématique est nulle.

Pour le premier terme de $V(\bar{x})$ on écrit :

$$E\left(S_i S_j \frac{u_{ij} - \bar{u}_i}{n_i}\right)^2 = E\left[S_i \left(S_j \frac{u_{ij} - \bar{u}_i}{n_i}\right)^2\right].$$

Ici aussi les termes rectangles disparaissant en raison de l'indépendance des tirages des éléments de deux communes distinctes.

Comme à une même combinaison S correspondent toutes les combinaisons possibles des éléments des communes on a :

$$E\left[S_i \left(S_j \frac{u_{ij} - \bar{u}_i}{n_i}\right)^2\right] = E\left[S_i\, E\left(S_j \frac{u_{ij} - \bar{u}_i}{n_i}\right)^2\right].$$

Lorsque la commune (i) sort h fois, on a :

$$(a)\quad E\left(S_j \frac{u_{ij} - \bar{u}_i}{n_i}\right)^2 = \frac{1}{n_i^2} V\left(S_j \frac{u_{ij} - \bar{u}_i}{n_i}\right) = \frac{h}{n_i} \frac{N_i - h n_i}{N_i - 1} V_i$$

en désignant par V_i la variance de la caractéristique u dans la commune (i). On a ici :

$$f(h, i) = \frac{h}{n_i} \frac{N_i - h\, n_i}{N_i - 1} V_i,$$

d'où :

$$E\left[S_i \left(S_j \frac{u_{ij} - u_i}{n_i}\right)^2\right] = \Sigma_{hi}\, p_{hi} \frac{h}{n_i} \frac{N_i - h n_i}{N_i - 1} V_i = \Sigma_i \frac{V_i}{n_i (N_i - 1)} \Sigma_h\, p_{hi}\, h\, (N_i - h n_i)$$

avec

$$\Sigma_h\, p_{hi}\, h\, (N_i - h\, n_i) = \frac{l\, L_i}{L}\left[N_i - n_i - n_i\, (l-1) \frac{L_i - 1}{L - 1}\right]$$

Passons au second terme :

$$E\,(S\,h\,\bar{u}_i - l\,\bar{u})^2.$$

C'est la variance de la somme des valeurs moyennes $\bar{u}_i$ représentées par les l billets tirés parmi les L existants. On a donc :

$$E\,(\underset{i}{S}\,h\,\bar{u}_i - l\,\bar{u})^2 = l\,\frac{L-l}{L-1}\,V\,(\bar{u}_i),$$

en désignant par $V\,(\bar{u}_i)$ la quantité

$$\sum_i \frac{L_i\,(\bar{u}_i - \bar{u})^2}{L}.$$

On a donc finalement :

$$V\,(\bar{x}) = \frac{1}{l}\,\frac{L-l}{L-1}\,V\,(\bar{u}_i) + \frac{1}{l}\sum_i \frac{N_i}{N}\,\frac{N_i - n_i}{N_i - 1}\,\frac{V_i}{n_i} - \frac{1}{l}\sum_i \frac{N_i}{N}\,\frac{L_i - 1}{L-1}\,\frac{l-1}{N_i - 1}\,V_i.$$

La variance $V\,(\bar{x})$ que nous venons d'obtenir est plus petite que la variance $V'(\bar{x})$ correspondant au procédé classique (1). On a en effet :

$$V\,(\bar{x}) = V'\,(\bar{x}) - \frac{1}{l}\sum_i \frac{N_i}{N}\,\frac{L_i - 1}{L-1}\,\frac{l-1}{N_i - 1}\,V_i.$$

L'échantillon à éléments tous distincts est donc plus précis que l'échantillon classique; le terme correctif ne dépend pas des n_i, il est donc d'autant plus petit par rapport au second terme, $\frac{1}{l}\sum_i \frac{N_i}{N}\,\frac{N_i - n_i}{N_i - 1}\,\frac{V_i}{n_i}$ que les n_i sont plus petits. Dans ce procédé classique on a, en effet, d'autant moins de chances d'avoir des répétitions d'éléments que le rapport du nombre des éléments tirés par billet à la population de la commune est plus petit.

$V\,(\bar{x})$ dépend des n_i. Dans un échantillon la somme des n_i, c'est-à-dire le nombre total d'éléments, est variable puisqu'elle dépend de la combinaison de billets qui est sortie. Sa valeur moyenne est égale à $l\sum_i \frac{N_i}{N}\,n_i$. Il est intéressant de chercher quelle valeur il faut donner aux n_i pour que la variance soit minimum lorsque le nombre moyen n d'éléments par échantillon est fixé à l'avance. On doit alors rendre minimum $\sum_i \frac{N_i}{N}\,\frac{N_i - n_i}{N_i - 1}\,\frac{V_i}{n_i}$ compte tenu de la condition $\frac{N_i}{N}\sum_i n_i = \frac{n}{l}$.

C'est un problème classique d'extremum lié qui conduit à écrire :

$$\lambda\,\frac{N_i}{N} - \frac{N^2}{N\,(N_i - 1)}\,\frac{V_i}{n_i^2} = 0$$

---

(1) Dans l'article de M. Thionet $V'\,(\bar{x})$ est obtenu en supposant indépendants les tirages d'éléments correspondant à chaque billet sorti du premier tirage. On peut aussi obtenir $V'\,(\bar{x})$ en partant de nos calculs. Il suffit dans la formule (a) qui donne $E\left(S_j \frac{\bar{u}_{ij} - \bar{u}_i}{n_i}\right)^2$ d'écrire différemment $V\left(S_j \frac{\bar{u}_{ij} - \bar{u}_i}{n_i}\right)$ c'est-à-dire de lui donner la valeur $h\,n_i\,\frac{N_i - n_i}{N_i - 1}\,V'$ au lieu de $h\,n_i\,\frac{N_i - h\,n_i}{N_i - 1}\,V_i$.

soit :

$$n_i \# \sqrt{\frac{V_i}{\lambda}}.$$

Autrement dit on doit prendre les n_i proportionnels aux écarts-types de chaque commune. Si ces écarts-types sont peu différents, les n_i seront pris tous égaux. On sera également amené à prendre des n_i égaux lorsque les variances par commune sont inconnues et que l'on n'a pas de raisons de les supposer très différentes. Ce cas se présentera souvent et mérite donc un examen particulier; on écrit alors :

$$V(\bar{x}) = \frac{1}{l}\frac{L-l}{L-1} V(\bar{u}_i) + \frac{1}{ln}\sum_i \frac{N_i}{N}\frac{N_i-n}{N_i-1} V_i - \frac{1}{l}\sum_i \frac{N_i}{N}\frac{L_i-1}{L-1}\frac{l-1}{N_i-1} V_i,$$

et l'on a d'après ce qui a été dit au début :

$$\frac{1}{n} \geqslant \frac{L_i}{N_i}.$$

Qu'advient-il de $V(\bar{x})$ lorsqu'on a l'égalité, c'est-à-dire lorsque :

$$N_i = n\, L_i$$

On a alors :

$$N_i - n - \frac{n(l-1)(L_i-1)}{L-1} = (N_i - n)\frac{L-l}{L-1}$$

d'où :

$$V(\bar{x}) = \frac{1}{l}\frac{L-l}{L-1}\left[V(u_i) + \frac{1}{n}\sum_i \frac{N_i}{N}\frac{N_i-n}{N_i-1} V_i\right]$$
$$= \frac{1}{l}\frac{L-l}{L-1}\sum_i \frac{N_i}{N}\left[(\bar{u}_i - \bar{u})^2 + \frac{1}{n}\frac{N_i-n}{N_i-1} V_i\right].$$

La quantité sous le signe Σ représente la variance de la moyenne des combinaisons de n éléments de la même commune que le tirage permet de réaliser.

$V(\bar{x})$ a donc la même forme que dans un tirage a 1 degré où l'on sort l boules d'une urne qui en contient L.

Ce résultat n'a rien de surprenant; en effet quand on a $N_i = l\, N_i$ on peut supposer les tirages effectués dans l'ordre inverse :

a) Dans chaque commune (i) on tire successivement au hasard les N_i éléments qu'elle contient. On affecte les n premiers à un billet, les n suivants à un autre et ainsi de suite jusqu'à épuisement des N_i éléments.

b) On tire les billets.

On a alors procédé à un tirage à 1 degré après fixation au hasard des éléments représentés par les billets. La variance trouvée plus haut est la variance moyenne de tous les tirages à 1 degré ainsi réalisables.

L. Henry.