Sondage stratifié

Myriam Maumy[1]

[1]IRMA, Université Louis Pasteur
Strasbourg, France

Master 1ère Année 13-03-2006

Ce chapitre s'appuie essentiellement sur 2 ouvrages :

1. "Les sondages : Principes et méthodes"
   de Anne-Marie Dussaix et Jean-Marie Grosbras,
   aux éditions Que sais-je ? .

2. "Manuel de Sondages"
   de Rémy Clairin et Philippe Brion,
   téléchargeable sur
   http ://ce-ped.cirad.fr/activite/publi/integral/html/manuels/pdf/
   puis "manuels cpd 03.pdf"

Sommaire

Dans un sondage aléatoire simple

toutes les combinaisons de n unités de l'échantillon parmi N éléments de la population U ont la même probabilité.

Remarque

Mais certaines d'entre elles peuvent être indésirables.

Exemple

Soit une population de 5 éléments. On relève sur ces 5 individus la variable d'intérêt "salaire annuel" (en milliers d'euros) :

$$13, 15, 17, 25, 30.$$

Parmi les échantillons à 2 unités, on a 2 cas extrêmes

$$(13, 15) \quad \text{et} \quad (25, 30)$$

qui se révèlent "mauvais" s'il s'agit d'estimer la moyenne

$$\mu = \frac{13 + 15 + 17 + 25 + 30}{5} = 20.$$

Remarque

Il y a plusieurs types de classes dans cette population :

- des classes d'individus "à salaires modestes"
- des classes d'individus "à salaires élevés".

Remarque

Il serait malencontreux que :

- les hasards de l'échantillonnage conduisent à n'interroger que des individus appartenant à une seule de ces catégories
- ou l'échantillon soit trop déséquilibré en faveur de l'une d'elles.

Le but du jeu :

Exclure les échantillons extrêmes et améliorer la précision des estimateurs du chapitre précédent.

Remarque

On a constaté qu'à taille égale un échantillon est plus efficace dans une population homogène que dans une population hétérogène.

Remarque

Plus précisément, l'erreur type d'estimation est liée à la variance du caractère étudié dans la population.

Le but du jeu :

Découper la population en sous-ensembles, appelés des **strates**, les plus homogènes possibles.

Chaque sondage partiel s'effectue de façon efficace et l'assemblage des sondages partiels précis donnera des résultats plus fiables qu'un sondage de même taille effectué "en vrac".

Quelques exemples :

- Les échantillons de ménages ou d'invidus, dans les enquêtes usuelles, sont stratifiés par région croisée par type d'habitat (taille des communes).
- Les échantillons d'entreprises sont stratifiés par secteur et par taille, exprimée en effectifs salariés ou chiffre d'affaires.
- Les échantillons d'exploitations agricoles sont stratifiés par tranches de surface.
- Les échantillons de jeunes sortis de l'enseignement supérieur sont stratifiés par discipline,
- etc...

Retour à l'exemple :

Pour une étude sur le "salaire annuel", il sera pertinent d'utiliser des critères liés :

- à l'âge,
- au niveau d'études,
- éventuellement au sexe,

c'est-à-dire à des facteurs susceptibles d'expliquer les différences de comportement au niveau des salaires.

Définition

Stratifier correspond souvent à un objectif de réduction des coûts d'enquête ou d'optimisation de sa gestion.

Remarque

C'est en particulier le cas :

- lorsque l'on utilise un critère de découpage géographique comme la région,
- ou, dans les échantillons d'entreprises, un critère sectoriel,

ce qui permet alors, de spécialiser les enquêteurs.

Retour à l'exemple :

Supposons que l'on sache, *a priori*, que les 3 premiers individus forment une catégorie de "petits" salaires et que les 2 derniers soient catalogués "gros" salaires.

- On décide alors que l'**échantillon de** 2 **individus** doit être constitué d'un **représentant de chaque strate**.
- Les échantillons possibles sont dans ce cas au nombre de 6. Chacun des 3 individus de la première strate pouvant être associé à l'un des 2 autres de la seconde strate.

- Notons x_1 et x_2 les valeurs obtenues dans l'échantillon. On ne peut plus, comme dans le chapitre précédent, en faire la moyenne arithmétique simple.
- En effet, l'unité échantillonnée dans la première strate est désignée pour en représenter 3, celle de la deuxième strate vaut pour 2.
- Il convient alors de *pondérer* chaque valeur x_i par le poids de la strate dont la valeur x_i est issue. Si $\widehat{\mu_{st}}$ désigne le résultat, on a alors :

$$\widehat{\mu_{st}} = \frac{3}{5}x_1 + \frac{2}{5}x_2.$$

- Le tableau ci-dessous représente l'ensemble de tous les cas possibles.

Échantillons avec stratification

x_1	13	13	15	15	17	17
x_2	25	30	25	30	25	30
$\widehat{\mu}_{st}$	17,8	19,8	19	21	20,2	22,2

- D'autre part, on vérifie que la moyenne des 6 valeurs pour $\widehat{\mu}_{st}$ est $\mu = 20$. Cela signifie que la variable aléatoire $\widehat{\mu}_{st}$ a μ pour espérance mathématique. Donc $\widehat{\mu}_{st}$ est un estimateur sans biais pour μ.

Remarque

On remarque surtout que

- la plage des estimations est beaucoup plus resserée autour de la cible que dans le cas du sondage aléatoire simple.

En effet :

- les valeurs extrêmes sont moins éloignées,
- l'écart-type vaut $1,40$ au lieu de $3,95$.

Remarque

On peut maintenant décrire la méthode générale.

Pour cela, on va avoir besoin d'introduire quelques notations.

Remarque

Par la suite, on se placera dans le cas d'un **tirage aléatoire simple sans remise**, à l'intérieur de chaque strate.

Pour la strate h de la population U :

- L'effectif de la strate h est égal à N_h.
- La moyenne d'une variable d'intérêt X est égale à

$$\mu_h = \frac{1}{N_h} \sum_{k=1}^{N_h} X_k.$$

- La variance corrigée d'une variable d'intérêt X est égale à

$$S_{h,c}^2 = \frac{1}{N_h - 1} \sum_{k=1}^{N_h} (X_k - \mu_h)^2 .$$

Pour la strate h de l'échantillon $\mathcal{S}$:

- L'effectif de l'échantillon propre à la strate h est égal à n_h.
- Un estimateur de la moyenne dans la strate h est égal à

$$\widehat{\mu}_h = \frac{1}{n_h} \sum_{k=1}^{n_h} x_k .$$

- La variance corrigée dans la strate h est égale à

$$s^2_{h,c} = \frac{1}{n_h - 1} \sum_{k=1}^{n_h} (x_k - \widehat{\mu}_h)^2 .$$

Sommaire

Définition

L'estimateur de la moyenne μ d'une population U par sondage stratifié se définit par :

$$\widehat{\mu}_{st} = \sum_{h=1}^{H} \frac{N_h}{N} \widehat{\mu}_h.$$

Propriété

On montre, par calcul, que **cet estimateur est sans biais,** *i.e.*

$$\mathbb{E}\left[\widehat{\mu}_{st}\right] = \mu.$$

Définition

L'estimateur du total T d'une population U par un sondage stratifié se définit par :

$$\widehat{T}_{st} = \sum_{h=1}^{H} N_h \widehat{\mu}_h.$$

Propriété

On montre, par calcul, que **cet estimateur est sans biais,** *i.e.*

$$\mathbb{E}\left[\widehat{T}_{st}\right] = T.$$

Remarque

Cette formule peut aussi s'écrire sous la forme :

$$\widehat{T}_{st} = \sum_{h=1}^{H} N_h \left(\frac{1}{n_h} \sum_{k=1}^{n_h} x_k \right) = \sum_{h=1}^{H} \left(\sum_{k=1}^{n_h} \frac{N_h}{n_h} x_k \right).$$

Remarque

On remarque, dans la formule précédente, que x_k est pondérée par le coefficient $\frac{N_h}{n_h}$, appelé **coefficient d'extrapolation** (dont la valeur dépend de la strate h), afin d'extrapoler (ou "d'étendre") les résultats à la population U.

Définition

L'estimateur d'une proportion π_A d'une population ayant la caractéristique A se fait, comme présenté au chapitre 1 bis, par l'estimateur de la moyenne d'une variable d'intérêt qui vaut

- 1 *si l'unité a la caractéristique étudiée*
- 0 *si l'unité n'a pas la caractéristique étudiée.*

Propriété

On montre, par calcul, que

$$\mathrm{Var}\left[\,\widehat{\mu}_{st}\,\right] = \sum_{h=1}^{H} \left(\frac{N_h}{N}\right)^2 (1 - f_h)\,\frac{S_{h,c}^2}{n_h}$$

où $f_h = \dfrac{n_h}{N_h}$ *est le taux de sondage correspondant et* $S_{h,c}^2$ *est la variance corrigée définie auparavant.*

Propriété

On montre, par calcul, que

$$\mathrm{Var}\left[\widehat{T}_{st}\right] = \sum_{h=1}^{H} N_h^2 \left(1 - f_h\right) \frac{S_{h,c}^2}{n_h}.$$

Remarque

Comment démontrez-vous ces formules ?

Remarque

Ces formules posent un problème. Lequel ?

Pour répondre à la dernière question posée, on définit les deux quantités suivantes :

Définition

Un estimateur de la variance de $\widehat{\mu}_{st}$ se définit par

$$\widehat{\mathrm{Var}\left[\,\widehat{\mu}_{st}\,\right]} = \sum_{h=1}^{H} \left(\frac{N_h}{N}\right)^2 (1 - f_h)\, \frac{s_{h,c}^2}{n_h}$$

où f_h est le taux de sondage correspondant et $s_{h,c}^2$ est la variance corrigée définie auparavant.

Définition

Un estimateur de la variance de $\widehat{\mu}_{st}$ se définit par

$$\widehat{\mathrm{Var}\left[\widehat{T}_{st}\right]} = \sum_{h=1}^{H} N_h^2 \left(1 - f_h\right) \frac{s_{h,c}^2}{n_h}$$

Remarque

Ces deux estimations de la variance permettent de calculer l'écart-type de chaque estimateur. Par conséquent, comme au chapitre 1 bis, on peut construire des intervalles de confiance pour ces estimateurs.

Exemple

Une société bancaire compte 50 000 clients répartis en :

- 40 000 « petits »clients
- 10 000 « gros »clients.

Soit un sondage portant sur 200 clients répartis en :

- 160 « petits »
- 40 « gros ».

On s'intéresse au montant moyen μ des comptes au moment de l'enquête et à la proportion π des clients prêts à souscrire au nouveau produit financier.

Exemple

Le dépouillement du sondage donne les résultats suivants :

Statistiques	Strate 1	Strate 2
Effectif population	$N_1 = 40\ 000$	$N_2 = 10\ 000$
Effectif échantillon	$n_1 = 160$	$n_2 = 40$
Montant moyen	$\widehat{\mu}_1 = 12$	$\widehat{\mu}_2 = 58$
Variance observée	$s_1^2 = 85$	$s_2^2 = 930$
Écart-type	$s_1 = 9,22$	$s_2 = 30,50$
Clients favorables	$x_1 = 8$	$x_2 = 22$
Proportion	$\widehat{\pi}_1 = 5\%$	$\widehat{\pi}_2 = 55\%$

Estimation de μ :

- $\widehat{\mu}_{st} = \frac{40\,000}{50\,000} \times 12 + \frac{10\,000}{50\,000} \times 58 = 0,8 \times 12 + 0,2 \times 58 = 21,2$
- $\widehat{\mathrm{Var}[\,\widehat{\mu}_{st}\,]} = 0,64 \times 0,996 \times \frac{85}{159} + 0,04 \times 0,996 \times \frac{930}{39} \simeq 1,29$
- Écart-type $\simeq \sqrt{1,29} \simeq 1,14$
- Intervalle de confiance à 95% pour μ :

$$\mu \in [21,2 \pm 1,96 \times 1,14],$$

c'est-à-dire :

$$\mu \in [18,97; 23,43].$$

Estimation de π :

- $\widehat{\pi_{st}} = 0,8 \times 0,05 + 0,2 \times 0,55 = 15\%$
- $\widehat{\mathrm{Var}[\,\widehat{\pi_{st}}\,]} = 0,64 \times 0,996 \times \dfrac{0,05 \times 0,95}{159} + 0,04 \times 0,996 \times \dfrac{0,55 \times 0,45}{39} \simeq 4,433 \times 10^{-4}$
- Écart-type $\simeq \sqrt{4,433} \times 10^{-2} \simeq 2,11\%$
- Intervalle de confiance à 95% pour π :

$$\pi \in [10,87\%; 19,13\%].$$

Sommaire

Les formules ci-dessus sont valables quels que soient les nombres d'unités statistiques tirées par strate. Le taux de sondage f_h peut donc être variable d'une strate h à une autre.

Définition

Quand on impose un taux de sondage

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

identique pour toutes les strates, alors le sondage est appelé **sondage stratifié proportionnel**.

Remarque

C'est ainsi que, dans un échantillon d'individus stratifié par sexe, les hommes et les femmes figurent au prorata de leur effectif dans la population étudiée.

Remarque

Dans l'application numérique du paragraphe précédent, on a considéré un **échantillon représentatif** de la population des petits clients et des gros clients.

Là encore, il faut prendre garde à la définition exacte des termes utilisés.

Définition

*Le terme "**représentatif**" signifie que l'échantillon a été dosé pour "représenter" une répartition d'effectifs dans la population.*

Remarque

Il ne signifie pas que le sondage soit parfait, sans erreurs, ni même que la répartition soit la meilleure possible ! Il est donc préférable, pour éviter les ambiguïtés, de parler d'**échantillon proportionnel**.

Propriétés

*Les propriétés de l'***échantillon proportionnel** *sont importantes :*

- *Les probabilités de sélection sont égales pour tous les éléments de la base de sondage. Elles valent le taux de sondage unique* $f = n/N$.
- *L'estimation de la moyenne* μ *vaut alors :* $\widehat{\mu}_{st} = \frac{1}{n}\sum_{h=1}^{H}\left(\sum_{k=1}^{n_h} x_k\right)$, *où n est la taille de l'échantillon. C'est donc la moyenne simple calculée sur l'échantillon qui permet d'estimer la moyenne sur la population. On a un sondage "***autopondéré***".*

Propriété

La variance de l'estimateur $\widehat{\mu}_{st}$ est égale à :

$$\mathrm{Var}\left[\widehat{\mu}_{st}\right] = \frac{(1-f)}{n}\left(\sum_{h=1}^{H}\frac{N_h}{N}S_{h,c}^2\right).$$

Remarque

Cette formule montre bien que plus les strates sont homogènes (variance intra-strates faible), plus la stratification est efficace.

Remarque

On montre que cette variance est liée à la variance de l'estimateur $\widehat{\mu}$ issu du SAS obtenu à partir du même nombre d'unités tirées. En effet, on a :

$$\mathrm{Var}\left[\widehat{\mu}\right] = \mathrm{Var}\left[\widehat{\mu}_{st}\right] + \frac{(1-f)}{n}\sum_{h=1}^{H}\frac{N_h}{N}\left(\overline{X}_h - \mu\right)^2.$$

Remarque

Que pouvez vous déduire de la dernière égalité qui porte sur les variances ?

Remarque

On en déduit que le **sondage stratifié représentatif** a une variance d'estimateur toujours plus petite ou égale à la variance de l'estimateur du **sondage aléatoire simple**.

Remarque

La variance de l'estimateur sera d'autant plus petite que les strates ont des moyennes différentes de μ.

On explique ce résultat en se rappelant que

Définition

Le **sondage stratifié** *est basé sur le principe de :*

- *forcer le hasard*
- *imposer à l'échantillon de représenter la population strate par strate.*

Retour à l'estimation du montant moyen des comptes des clients de la société bancaire.

On a :

$$\begin{aligned} \widehat{\mu}_{st} = \widehat{\mu} &= 21,2 \\ s^2_{intra} &= 0,8 \cdot 85 + 0,2 \cdot 930 = 254. \end{aligned}$$

Si les mêmes données étaient issues d'un sondage aléatoire simple, on aurait :

$$\mathrm{Var}[\,\widehat{\mu}_{st}\,] \simeq \frac{338,56 + 254}{200} = 2,96$$

mais elles sont issues d'un sondage stratifié proportionnel, donc :

$$\mathrm{Var}[\,\widehat{\mu}_{st}\,] \simeq \frac{254}{200} = 1,27.$$

La variance d'échantillonage a donc diminué de 57%.

Peut-on encore améliorer les résultats ?

Oui, on peut améliorer les résultats en faisant du **sondage stratifié proportionnel,** comme nous venons de le voir.

Sommaire

L'idée :

Déterminer des strates les plus homogènes possibles, par rapport au sujet étudié.

Deux types de considérations vont conduire au choix des critères de stratification :

1. disponibilité des critères dans la base de sondage ;
2. pertinence des différents critères pour créer des strates homogènes. Ceci nécessite une connaissance
   - soit intuitive,
   - soit venant d'études réalisées antérieurement.

On prendra généralement comme critères :

- des critères relevant d'une typologie (par exemple la catégorie sociale) ;
- des critères de taille (prenant par exemple en compte le nombre de personnes du ménage) ;

souvent en les croisant ensemble.

Au niveau des **unités de sondage "géographiques"** :

Exemple : Pour les villes stratification selon la région, l'activité dominante des localités. On sépare souvent milieu rural et milieu urbain.

Au niveau des **ménages** ou des **individus :**

Utilsation des critères qui peuvent être en corrélation avec le sujet d'étude.
Exemple : la CSP, le niveau d'étude, la taille du ménage, le type d'habitation, etc...

Une **stratification** peut être :

- très efficace pour l'étude d'un phénomène, par exemple la mortalité,
- très peu efficace pour l'étude d'autres phénomènes, par exemple l'activité économique.

Cette situation se présente avec une acuité particulière lorsqu'un échantillon est destiné à des études à objectifs multiples.

Attention :

Plus on multiplie les strates, plus le gain d'efficacité devient faible.

De plus, les résultats calculés au niveau de chaque strate ne sont plus significatifs en raison de la petite taille de l'échantillon.

Sommaire

Remarque

La répartition représentative ou encore appelée allocation proportionnelle a déjà été présentée au paragraphe 3 de ce chapitre.

Définition

La répartition représentative consiste à utiliser le même taux de sondage f pour toutes les strates.

Définition

La répartition de Neyman ou encore appelée allocation optimale consiste à respecter l'égalité :

$$\frac{n_h}{N_h S_{h,c}} = constante = \frac{n}{\sum_{h=1}^{H} N_h S_{h,c}}.$$

Remarque

Cette répartiton utilise un taux de sondage f proportionnel à la dispersion $S_{h,c}$ de X étudiée dans chaque strate.

Remarque

Plus une strate est hétérogène vis-à-vis de X, plus on utilise un taux de sondage f important.

Remarque

La théorie montre que cette répartition est celle qui fournit la variance la plus faible une fois les strates déterminées.

Remarque

L'application de la formule pour calculer **la répartition de Neyman** suppose connues *a priori* les valeurs $S_{h,c}$. Ce peut être le cas à partir d'études antérieures au sondage, mais en général il n'en est pas ainsi.

Remarque

Lorsque le critère de stratification est la taille des unités, on constate que l'écart-type est sensiblement proportionnel à la taille moyenne des unités de la strate. C'est un ordre de grandeur de cette taille moyenne qu'on utilise pour calculer la répartition des individus entre les strates.

Remarque

En pratique, on utilise **la répartition de Neyman** quand le phénomène étudié a une distribution très dissymétrique.

Remarque

Par contre, si ce phénomène a une distribution symétrique par rapport à sa moyenne, un **sondage stratifié proportionnel** fournit des résultats d'une qualité suffisante.

Exemple

On tire un échantillon de 200 clients de la société bancaire. On a le choix entre :

- une répartition proportionnelle (les calculs ont déjà été faits) et
- la répartition de Neyman.

Remarque

Remarquons que l'échantillon de Neyman dépend du caractère que l'on veut estimer en priorité. C'est pour ce caractère que l'on prendra la variance en considération. En général, celle-ci ne sera pas connue *a priori*. Elle pourra être estimée à partir d'une enquête antérieure ou d'études limitées.

Retour à l'exemple

L'échantillon de Neyman est composé de :

- 110 « petits »clients contre 160
- et de 90 « gros »clients contre 40, 90 pour tenir compte de la plus grande variance de ces derniers.

Le calcul montre que la variance d'échantillonnage aurait été égale à $0,91$ au lieu de $1,27$, soit un gain de 28% par rapport à la répartition proportionnelle.

Remarque

Ainsi, on perd en simplicité des calculs du cas « proportionnel »puisque l'échantillon n'est plus autopondéré, mais on gagne en précision.

Remarque

C'est en vertu de considérations de cet ordre que, par exemple, les échantillons d'entreprises stratifiées par tranches de taille (moins de 10 salariés, de 10 à 50 salariés, etc.) sont répartis, non pas au prorata du nombre d'entreprises des tranches, mais au prorata du nombre total de salariés ou du chiffre d'affaires total.