Sondage stratifié

Myriam Maumy-Bertrand[1]

[1]IRMA, Université de Strasbourg
Strasbourg, France

Master 1ère Année 16-10-2014

Références

Ce chapitre s'appuie essentiellement sur deux ouvrages :

1. « Les sondages : Principes et méthodes »
   de A.-M. Dussaix et J.M. Grosbras,
   P.U.F., Collection Que sais-je ?, 1993.
2. « Manuel de Sondages »
   de R. Clairin et P. Brion,
   téléchargeable à :
   `http://www.ceped.org/cdrom/integral_publication_1988_2002/manuels/pdf/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 : salaire annuel

Soit une population de 5 éléments. Nous relevons 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, nous avons 2 cas extrêmes

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

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

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

Remarques

- Il y a plusieurs types de classes dans cette population :
  - des classes d'individus « à salaires modestes »
  - des classes d'individus « à salaires élevés ».
- 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.

Remarques

- Nous avons constaté qu'à taille égale, un échantillon est plus efficace dans une population homogène que dans une population hétérogène.
- 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.

Conséquence

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'individus, 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 du salaire annuel

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 nous utilisons 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 du salaire annuel

Supposons que nous sachons, *a priori*, que les 3 premiers individus forment une catégorie de « salaires modestes » et que les 2 derniers soient catalogués « salaires élevés ».

- Nous décidons 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.

Suite de l'exemple

- Notons y_1 et y_2 les valeurs obtenues dans l'échantillon. Nous ne pouvons plus, comme auparavant, faire la moyenne arithmétique.
- En effet, l'unité échantillonnée dans la 1ère strate est désignée pour en représenter 3, celle de la 2ième strate vaut pour 2.
- Il convient alors de *pondérer* chaque y_i par le poids de la strate dont y_i est issue. Si $\widehat{\mu}_{st}$ désigne l'estimation de la moyenne, alors nous avons :

$$\widehat{\mu}_{st} = \frac{3}{5} y_1 + \frac{2}{5} y_2.$$

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

échantillons avec stratification

y_1	13,0	13,0	15,0	15,0	17,0	17,0
y_2	25,0	30,0	25,0	30,0	25,0	30,0
$\widehat{\mu}_{st}$	17,8	19,8	19,0	21,0	20,2	22,2

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

Remarque

Nous remarquons surtout que la plage des estimations est beaucoup plus resserrée autour de la cible que dans le cas du sondage aléatoire simple à probabilités égales sans remise (PESR).

En effet :

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

Conclusion

Il y a moins de risque d'obtenir une « mauvaise » estimation de μ en réalisant un sondage stratifié plutôt qu'un sondage à PESR.

La stratification a permis, en utilisant de l'information auxiliaire (l'existence de deux sous-populations), d'améliorer la qualité de l'estimateur de μ.

Nous pouvons maintenant d'écrire la méthode générale.

Pour cela, nous allons avoir besoin d'introduire quelques notations.

Remarque

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

Au niveau de la population U, pour la strate U_h (nous avons H strates de la population) :

- L'effectif de la strate U_h dans la population est égal à N_h.
- La moyenne d'une variable d'intérêt Y est égale à :

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

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

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

Au niveau de la population U

- L'effectif de la population U est égal à $N = \sum_{h=1}^{H} N_h$.
- La moyenne d'une variable d'intérêt Y, notée μ_Y, est égale à :

$$\mu_Y = \frac{1}{N} \sum_{i \in U} y_i = \sum_{h=1}^{H} \frac{N_h}{N} \mu_h.$$

- La variance non corrigée d'une variable d'intérêt Y est égale à :

$$\sigma_Y^2 = \frac{1}{N} \sum_{i \in U} (y_i - \mu_Y)^2 = \frac{1}{N} \sum_{h=1}^{H} \sum_{i \in U_h} (y_i - \mu_Y)^2 .$$

Formule de la décomposition de la variance

Ainsi, nous avons :

$$\sigma_Y^2 = \sum_{h=1}^{H} \frac{N_h}{N} \sigma_h^2 + \sum_{h=1}^{H} \frac{N_h}{N} (\mu_h - \mu_Y)^2 ,$$

le premier terme est noté σ^2_{dans} et le second terme est noté σ^2_{entre}.

- $\sigma^2_{dans} = \sigma^2_{intra}$ est une mesure globale de la dispersion de Y au sein des strates,
- $\sigma^2_{entre} = \sigma^2_{inter}$ est une mesure globale de la dispersion de Y entre les strates.

Au niveau de l'échantillon $\mathcal{S}$, pour la strate S_h

- L'effectif de la strate S_h de l'échantillon $\mathcal{S}$ est égal à n_h.
- L'estimateur de la moyenne, dans la strate S_h est égal à :

$$\widehat{\mu}_h = \frac{1}{n_h} \sum_{i \in S_h} Y_i = \frac{1}{n_h} \sum_{i=1}^{n_h} Y_i.$$

- L'estimateur de la variance corrigée, dans la strate S_h est égal à :

$$S^2_{h,c} = \frac{1}{n_h - 1} \sum_{i=1}^{n_h} (Y_i - \widehat{\mu}_h)^2 .$$

Au niveau de l'échantillon $\mathcal{S}$

- L'effectif de l'échantillon $\mathcal{S}$ est égal à $n = \sum_{h=1}^{H} n_h$.
- L'estimateur de la moyenne, dans l'échantillon $\mathcal{S}$ est égal à :

$$\widehat{\mu}_Y = \frac{1}{n} \sum_{i \in \mathcal{S}} Y_i = \sum_{h=1}^{H} \frac{n_h}{n} \widehat{\mu}_h.$$

- L'estimateur de la variance corrigée, dans l'échantillon $\mathcal{S}$ est égal à :

$$S^2_{n,c} = \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \widehat{\mu})^2 .$$

Remarque

Paragraphe 5.2 : détermination des tailles n_h où h varie de 1 à H, de manière à minimiser la variance de l'estimateur de la moyenne μ_Y sous la contrainte que la taille n de l'échantillon $\mathcal{S}$ est fixée *a priori*.

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

Nous montrons, par calcul, que **cet estimateur est sans biais :**

$$\mathbb{E}\left(\widehat{\mu}_{st}\right) = \mu.$$

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

Nous montrons, par calcul, que **cet estimateur est sans biais :**

$$\mathbb{E}\left(\widehat{T}_{st}\right) = T.$$

Remarque

Cette formule peut aussi s'écrire sous la forme :

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

Nous remarquons, dans la formule précédente, que Y_i est pondérée par le coefficient $\dfrac{N_h}{n_h}$, appelé **coefficient d'extrapolation** (dont la valeur dépend de la strate U_h), afin d'extrapoler (ou « d'étendre ») les résultats à la population.

Définition

L'estimateur d'une proportion π_A d'une population ayant la caractéristique A se fait, comme présenté dans un chapitre précédent, 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

Nous montrons, par calcul, que :

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

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

Propriété

Nous montrons, par calcul, que :

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

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

Remarques

- Comment démontrez-vous ces formules ?
- Ces formules posent un problème. Lequel ?

Pour répondre à la dernière question posée, nous définissons les deux quantités suivantes.

Définition

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

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

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 $\widehat{T}_{st}$ se définit par :

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

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

Remarque

Ces deux estimateurs de la variance permettent de calculer l'écart-type de chaque estimateur. Par conséquent, comme au chapitre 1, nous pouvons construire des intervalles de confiance au niveau de confiance égal à $1 - \alpha$ pour chacun des paramètres inconnus de la population.

Définition

L'intervalle de confiance asymptotique pour μ au niveau de confiance égal à $1-\alpha$ se définit par :

$$\left]\widehat{\mu}_{st} - z_{1-\alpha/2} \times \sqrt{\widehat{\mathrm{Var}}\left(\widehat{\mu}_{st}\right)}; \widehat{\mu}_{st} + z_{1-\alpha/2} \times \sqrt{\widehat{\mathrm{Var}}\left(\widehat{\mu}_{st}\right)}\right[,$$

où $z_{1-\alpha/2}$ est le quantile d'ordre $1-\alpha/2$ de la loi normale centrée et réduite.

Définition

L'intervalle de confiance asymptotique pour T au niveau de confiance égal à $1-\alpha$ se définit par :

$$\left]\ \widehat{T}_{st} - z_{1-\alpha/2} \times \sqrt{\widehat{\mathrm{Var}}\left(\widehat{T}_{st}\right)};\ \widehat{T}_{st} + z_{1-\alpha/2} \times \sqrt{\widehat{\mathrm{Var}}\left(\widehat{T}_{st}\right)}\ \right[,$$

où $z_{1-\alpha/2}$ est le quantile d'ordre $1-\alpha/2$ de la loi normale centrée et réduite.

Exemple bancaire

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 » clients
- 40 « gros » clients.

Nous nous intéressons 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.

Suite de l'exemple bancaire

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 corrigée	$s_1^2 = 85$	$s_2^2 = 930$
Écart-type	$s_1 = 9,22$	$s_2 = 30,50$
Clients favorables	$n_{A,1} = 8$	$n_{A,2} = 22$
Proportion	$\widehat{\pi}_1 = 5\%$	$\widehat{\pi}_2 = 55\%$

Intervalle de confiance pour μ

- $\widehat{\mu}_{st} = \frac{40\,000}{50\,000} \times 12 + \frac{10\,000}{50\,000} \times 58 = 0,8 \times 12 + 0,2 \times 58 = 21,2$
- $\widehat{\mathrm{Var}(\widehat{\mu}_{st})} = 0,64 \times 0,996 \times \frac{85}{160} + 0,04 \times 0,996 \times \frac{930}{40}$
  $= 1,26492$
- Écart-type $= \sqrt{1,26492} \simeq 1,125$
- Intervalle de confiance à 95% pour μ :

$$\mu \in]21,2 \pm 1,96 \times 1,125[,$$

c'est-à-dire :

$$\mu \in]18,995; 23,405[.$$

Intervalle de confiance pour π

- $\widehat{\pi}_{st} = \dfrac{40\,000}{50\,000} \times 0,05 + \dfrac{10\,000}{50\,000} \times 0,55 = 15\%$
- $\widehat{\mathrm{Var}(\widehat{\pi}_{st})} = 0,64 \times 0,996 \times \dfrac{0,05 \times 0,95}{160} + 0,04 \times 0,996 \times \dfrac{0,55 \times 0,45}{40} = 4,3575 \times 10^{-4}$
- Écart-type $= \sqrt{4,3575} \times 10^{-2}\% \simeq 2,0875 \times 10^{-2}\%$
- Intervalle de confiance à 95% pour π :

$$\pi \in]10,90\%; 19,10\%[.$$

Remarque

Si dans le cas du sondage stratifié, nous avions estimé μ par

$$\widehat{\mu}_Y = \sum_{h=1}^{H} \frac{n_h}{n} \widehat{\mu}_h$$

au lieu de

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

alors, nous aurions un estimateur biaisé de μ.

Sommaire

Remarque

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

Un sondage est appelé un **sondage stratifié proportionnel** *quand le sondage stratifié est tel que les taux de sondage* $f_h = \frac{n_h}{N_h}$ *sont les mêmes dans toutes les strates. Ainsi, nous avons un taux de sondage global égal à* $f = \frac{n}{N} = f_h = \frac{n_h}{N_h}$.

Remarques

- 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.
- Dans l'application numérique du paragraphe précédent, nous avons 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 importantes de l'***échantillon stratifié proportionnel** *sont au nombre de 5 :*

- *Les probabilités d'inclusion d'ordre 1 sont égales pour tous les individus de la population et valent le taux de sondage unique* $f = n/N$.
- *L'estimateur de la moyenne* μ *vaut alors :*

$$\widehat{\mu}_{stp} = \frac{1}{n} \sum_{h=1}^{H} \left(\sum_{i=1}^{n_h} Y_i \right).$$

C'est donc la moyenne simple calculée sur l'échantillon qui permet d'estimer μ_Y. *Nous avons un sondage* **autopondéré**.

Troisième propriété

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

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

Remarque

L'expression de $\mathrm{Var}\left(\widehat{\mu}_{stp}\right)$ montre que la précision de $\widehat{\mu}_{stp}$ est liée à l'homogénéité/hétérogénéité des individus au sein des strates. Plus les strates sont homogènes (vis-à-vis de Y), plus $\sigma^2_{dans,c}$ est faible, plus $\mathrm{Var}\left(\widehat{\mu}_{stp}\right)$ est petite, plus $\widehat{\mu}_{stp}$ est précis.

Remarque

Si nous définissons : $\sigma^2_{dans,c} = \sum_{h=1}^{H} \frac{N_h}{N} \sigma^2_{h,c}$ nous avons alors

$$\text{Var}\left(\widehat{\mu}_{stp}\right) = \frac{(1-f)}{n} \sigma^2_{dans,c}.$$

Nous montrons que :

$$\sigma^2_{dans,c} = \sigma^2_{dans} + \frac{1}{N} \sum_{h=1}^{H} \sigma^2_{h,c}.$$

$\sigma^2_{dans,c}$ est, tout comme σ^2_{dans}, une mesure globale de la dispersion Y au sein des strates.

Remarque

Nous montrons que cette variance est liée à la variance de l'estimateur $\widehat{\mu}_Y$ issu du SAS à PE obtenu à partir du même nombre d'unités tirées. En effet, nous avons :

$$\mathrm{Var}\left(\widehat{\mu}_Y\right) = \mathrm{Var}\left(\widehat{\mu}_{stp}\right) + (1-f)\,\frac{1}{n}\sum_{h=1}^{H}\frac{N_h}{N}\left(\mu_h - \mu\right)^2 .$$

Que pouvez vous déduire de cette dernière égalité ?
Nous en déduisons 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** à PE.

Remarque

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

Nous expliquons ce résultat en se rappelant que

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

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

Deux dernières propriétes

- L'estimateur de la variance de l'estimateur est égal à :

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

- L'intervalle de confiance asymptotique pour μ au niveau de confiance égal à $1-\alpha$ se définit par :

$$\left]\widehat{\mu}_{stp} - z_{1-\alpha/2}\sqrt{\widehat{\mathrm{Var}\left(\widehat{\mu}_{stp}\right)}}; \widehat{\mu}_{stp} + z_{1-\alpha/2}\sqrt{\widehat{\mathrm{Var}\left(\widehat{\mu}_{stp}\right)}}\right[$$

où $z_{1-\alpha/2}$ est le quantile d'ordre $1-\alpha$ de la loi normale centrée et réduite.

Remarque

Nous avons établi les dernières propriétés pour l'estimateur stratifié de μ_Y, mais nous pouvons bien sûr établir les mêmes propriétés pour l'estimateur de T_Y.

Propriété

Nous allons comparer le sondage stratifié proportionnel au sondage aléatoire simple à PESR.

$$\begin{aligned} D(STP/PESR) &= \frac{\mathrm{Var}\left(\widehat{\mu}_{stp}\right)}{\mathrm{Var}\left(\widehat{\mu}_{Y}\right)} \\ &= \frac{(1-f)}{n}\sigma^2_{dans,c} \Big/ \frac{(1-f)}{n}\sigma^2_{c} \\ &= \frac{\sigma^2_{dans,c}}{\sigma^2_{c}}. \end{aligned}$$

Remarque

Si la taille N de la population est grande, alors $\sigma^2_{dans,c} \simeq \sigma^2_{dans}$ et $\sigma^2_c \simeq \sigma^2$. Par conséquent, nous avons :

$$D(STP/PESR) \simeq \frac{\sigma^2_{dans}}{\sigma^2} \leqslant 1.$$

Donc quand la taille N de la population est grande, le sondage stratifié proportionnel est plus efficace que le sondage aléatoire simple à PESR.

Retour à l'exemple bancaire

Calculons $s^2_{dans,c}$:

$$s^2_{dans,c} = s^2_{intra} = \frac{40\,000}{50\,000} \times 85 + \frac{10\,000}{50\,000} \times 930 = 254.$$

Puis estimons l'effet de sondage :

$$\begin{aligned} D(STP/PESR) &= \frac{\mathrm{Var}\left(\widehat{\mu}_{stp}\right)}{\mathrm{Var}\left(\widehat{\mu}_{Y}\right)} \\ &\simeq \frac{\frac{254}{200}}{\frac{338{,}56+254}{200}} = \frac{254}{592,56}. \end{aligned}$$

La variance d'échantillonnage a donc diminué de 43%.

Remarque

Pouvons-nous améliorer ces résultats ?

Oui, nous pouvons améliorer ces résultats comme nous le verrons dans le paragraphe suivant.

Sommaire

Définition

La répartition optimale à taille fixe n consiste à respecter l'égalité :

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

Remarques

- La théorie montre que cette répartition est celle qui fournit la variance la plus faible une fois les strates déterminées.
- Plus une strate est hétérogène vis-à-vis de Y, plus nous utilisons un taux de sondage f important.
- L'application de la formule pour calculer **la répartition optimale** suppose connues *a priori* les valeurs $\sigma_{h,c}$. Ce peut être le cas à partir d'études antérieures au sondage, mais en général il n'en est pas ainsi.

Deux dernières remarques

- Lorsque le critère de stratification est la taille des unités, nous constatons 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 que nous utilisons pour calculer la répartition des individus entre les strates.
- En pratique, nous utilisons **la répartition de Neyman** quand le phénomène étudié a une distribution très dissymétrique.
  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 bancaire

Nous tirons un échantillon de 200 clients de la société bancaire. Nous avons le choix entre :

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

Remarque

Remarquons que l'échantillon de Neyman dépend du caractère que nous voulons estimer en priorité. C'est pour ce caractère que nous prendrons 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 bancaire

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.

Conclusion

- Ainsi, nous perdons en simplicité des calculs du cas « proportionnel » puisque l'échantillon n'est plus autopondéré, mais nous gagnons en précision.
- 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.

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.

Nous prendrons généralement comme critères :

- des critères relevant d'une typologie,

Exemple

La catégorie sociale

- des critères de taille,

Exemple

Le nombre de personnes du ménage

souvent en les croisant ensemble.

Au niveau des **unités de sondage « géographiques »**

Nous séparons souvent milieu rural et milieu urbain.

Exemple

Pour les villes, la stratification selon la région, l'activité dominante des localités.

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

Utilisation 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 nous multiplions 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.