Chapitre 6

Estimation d'un ratio

Dans ce chapitre, nous étudions l'estimation d'un ratio qui est une fonction non linéaire de deux totaux. L'estimation sur un domaine qui est un exemple d'application de l'estimation d'un ratio est détaillée.

6.1 Estimation d'un ratio

6.1.1 Introduction

– *Exemple 1.* Supposons une population U de ménages, y_k le revenu du ménage k et z_k le nombre de personnes composant le ménage. Le revenu moyen par tête dans cette population est :

$$R = \frac{\sum_U y_k}{\sum_U z_k} = \frac{t_y}{t_z}.$$

R est ce qu'on appelle un ratio, c'est-à-dire le rapport de deux totaux sur une même population.

– *Exemple 2.* La proportion d'électeurs qui, dans une élection présidentielle, choisissent un candidat particulier est le rapport :
Nombre de votants qui choisissent le candidat / Nombre de suffrages exprimés.
Cette proportion doit être estimée comme un ratio car la taille de la population, c'est-à-dire le nombre d'électeurs qui votent n'est pas connue.

6.1.2 Cadre général de l'estimation d'un ratio

On dispose d'un plan de sondage de probabilités d'inclusion, π_k et π_{kl}. Un échantillon s est obtenu par ce plan et on observe $y_k, z_k,\ k \in s$. On estime le ratio R par le quotient des estimateurs de H-T des totaux :

$$\hat{R} = \frac{\hat{t}_{y\pi}}{\hat{t}_{z\pi}}. \tag{6.1}$$

C'est un estimateur non linéaire et on ne peut donc pas calculer exactement son espérance mathématique. Nous en obtenons une expression approchée par une technique classique en sondages : la linéarisation.

Espérance mathématique et variance approchée de $\hat{R}$. Appelons f la fonction des totaux qui donne le ratio : $f(y, z) = y/z$ et écrivons le développement de Taylor à l'ordre 1 de f et au voisinage de $y_0 = t_y$ et $z_0 = t_z$. On obtient :

$$\hat{R} \simeq \frac{t_y}{t_z} - \frac{R}{t_z}(\hat{t}_{z\pi} - t_z) + \frac{1}{t_z}(\hat{t}_{y\pi} - t_y)$$

ou

$$\hat{R} \simeq R + \frac{1}{t_z} \sum_s \frac{y_k - Rz_k}{\pi_k}. \tag{6.2}$$

Prenant l'espérance mathématique des deux côtés de (6.2), on obtient : $\mathsf{E}(\hat{R}) \simeq R$. L'estimateur $\hat{R}$ est sans biais au premier ordre. La variable

$$\nu_k = \frac{1}{t_z}(y_k - Rz_k) \tag{6.3}$$

est appelée *linéarisée* de $R = t_y/t_z$. On voit sur (6.2) que la variance linéarisée de $\hat{R}$, c'est-à-dire la variance du côté droit de (6.2), n'est autre que la variance de $\sum_s \frac{\nu_k}{\pi_k}$, estimateur du total de la linéarisée. On peut donc appliquer les résultats obtenus pour l'estimation d'un total par les valeurs dilatées :

$$\mathsf{var}(\hat{R}) \simeq \mathsf{var}\left(\sum_s \frac{\nu_k}{\pi_k}\right) = \sum \sum_U \Delta_{kl}\, \check{\nu}_k \check{\nu}_l$$

où $\check{\nu}_k = \nu_k/\pi_k$. On ne connaît ni R ni t_z, on les remplace donc par $\widehat{R}$ et $\widehat{t}_z$ pour obtenir une estimation de variance :

$$\widehat{\mathsf{var}}(\hat{R}) = \sum \sum_U \frac{\Delta_{kl}}{\pi_{kl}} \check{\widehat{\nu}}_k \check{\widehat{\nu}}_l$$

où $\widehat{\nu}_k = (y_k - \widehat{R}z_k)/\widehat{t}_z$.

Note. La linéarisée est un outil classique pour approcher les variances d'estimateurs complexes. L'ouvrage de Tillé contient un développement général sur cette notion.
Si on a utilisé un plan de taille fixe on utilisera les expressions de variance correspondantes. Nous allons voir précisément la situation pour un plan SI.

6.1.3 Estimation d'un ratio dans un plan SI

Par un plan SI(N, n) qui donne un échantillon s dans une population U, on obtient $y_k, z_k,\ k \in s$. L'estimateur du ratio est

$$\hat{R} = \frac{\widehat{t}_y}{\widehat{t}_z} = \frac{\overline{y}_s}{\overline{z}_s} \tag{6.4}$$

On applique ensuite les formules spécifiques au plan SI pour l'estimation de la variance du total $\sum_U \widehat{\nu}_k$ de la linéarisée. On obtient ainsi

$$\mathsf{var}(\hat{R}) \simeq N^2 \left(\frac{1}{n} - \frac{1}{N}\right) S^2_{\nu,U} = \frac{1}{t^2_{zU}} N^2 \left(\frac{1}{n} - \frac{1}{N}\right) S^2_{y-Rz,U} \tag{6.5}$$

$$\widehat{\mathsf{var}}(\hat{R}) = N^2 \left(\frac{1}{n} - \frac{1}{N}\right) S^2_{\widehat{\nu}s} = \frac{N^2}{\widehat{t}^2_z} \left(\frac{1}{n} - \frac{1}{N}\right) S^2_{y-\hat{R}z,s}$$

$$\widehat{\mathsf{var}}(\hat{R}) = \frac{1}{\overline{z}^2_s} \left(\frac{1}{n} - \frac{1}{N}\right) S^2_{y-\hat{R}z,s} \tag{6.6}$$

avec

$$S^2_{\widehat{\nu}s} = \frac{1}{n-1} \sum_s (\widehat{\nu}_k - \overline{\widehat{\nu}})^2.$$

C'est la formule (6.6) qu'on utilise pour les calculs pratiques.

Si nous détaillons $S^2_{y-Rz,U}$ nous obtenons :

$$S^2_{y-Rz,U} = \frac{1}{N-1} \sum_U [(y_k - Rz_k) - (\overline{y} - R\overline{z})]^2 = S^2_{yU} - 2RS_{yz,U} + R^2 S^2_{zU} \tag{6.7}$$

où

$$S_{yz,U} = \frac{1}{N-1}\sum\nolimits_U (y_k - \overline{y})(z_k - \overline{z})$$

est la covariance entre y et z sur U. On a de même, en vue des calculs pratiques :

$$S^2_{y-\hat{R}z,s} = S^2_{ys} - 2\widehat{R}S_{yz,s} + \widehat{R}^2 S^2_{zs}.$$

6.2 Estimation sur un domaine

L'estimation sur un domaine est une question très étendue. L'exposé qui suit n'est qu'un traitement très élémentaire, mais qui montre une utilisation de l'estimation d'un ratio.

6.2.1 Introduction

On veut souvent, à l'occasion d'un sondage, estimer le total d'une variable d'intérêt, non seulement sur la population U sur laquelle le plan de sondage est défini mais aussi sur une ou des sous-populations de U *non prises en compte par le plan*. Dans le présent chapitre, la sous-population particulière à laquelle appartient chaque élément de l'échantillon est constatée après sondage. On appelle domaine et on note U_d, toute sous-population pour laquelle on veut une estimation séparée du total et de la moyenne et des intervalles de confiance associés. Si la sous-population d'intérêt représente une fraction assez importante de U, les techniques ordinaires qu'on va voir d'abord, donnent de bons résultats. Pour un petit domaine, c'est-à-dire pour une sous-population qui ne représente qu'une petite fraction de U, il se peut que l'échantillon prélevé par un plan sur U ne contienne que peu d'éléments du domaine. Les estimateurs usuels risquent d'avoir une forte erreur quadratique. On met en œuvre des estimateurs utilisant de l'information auxiliaire. Nous n'abordons pas cette question dans cette présentation purement introductive.

Exemples de domaines Un domaine est souvent une région géographique, (*Small area estimation* désigne l'ensemble des techniques pour des petits domaines définis géographiquement). L'unité est par exemple le ménage, le domaine un canton et on veut estimer le revenu moyen des ménages par canton.
Un domaine peut être une marque commerciale de voitures dans la population des voitures vendues une certaine année dans un pays. On veut estimer des parts de marché. L'information exhaustive est connue avec retard. Une étude par sondage peut fournir rapidement une information fiable.
Pour une région géographique donnée, un domaine peut être l'ensemble des habitants ayant eu une certaine maladie.
On est aussi amené à faire de l'estimation sur un domaine quand la base de sondage, c'est-à-dire l'organisation de la population contient strictement la population d'intérêt.

6.2.2 Estimation sur un domaine - notions élémentaires

On appelle domaine une sous-population U_d de taille N_d, $U_d \subset U$ et le sondage porte sur U. On note π_k, π_{kl} les probabilités d'inclusion, Δ_{kl} les covariances des indicatrices d'inclusion et s l'échantillon sur U obtenu. On observe y_k ainsi que l'appartenance éventuelle au domaine, $k \in s$. Notons $s_d = s \bigcap U_d$, le sous-échantillon constaté appartenir à U_d. La taille n_d de s_d est aléatoire. On envisage l'estimation du total t_{y,U_d} d'une variable d'étude y sur U_d et de sa moyenne : y_{U_d}.

$$t_{y,U_d} = \sum_{U_d} y_k \qquad\qquad \overline{y}_{U_d} = \frac{t_{y,U_d}}{N_d}$$

Introduisons $z_{dk} = \begin{cases} 1 \text{ si} & k \in U_d \\ 0 \text{ sinon} & \end{cases}$.

On peut maintenant écrire :

$$t_{y,U_d} = \sum\nolimits_U y_k z_{dk}, \qquad N_d = \sum\nolimits_U z_{dk}.$$

L'estimation du total sur U_d est ainsi ramenée à un problème sur la population sur laquelle on a un plan de sondage. D'autre part la moyenne sur U_d se note :

$$\overline{y}_{U_d} = \frac{\sum_U y_k z_{dk}}{\sum_U z_{dk}}, \tag{6.8}$$

elle apparaît comme un ratio.

On peut maintenant écrire les estimateurs :

$$\widehat{t}_{y,U_d} = \sum\nolimits_s y_k z_{dk}/\pi_k = \sum\nolimits_{s_d} y_k/\pi_k. \tag{6.9}$$

et comme N_d est souvent inconnue, l'écriture de N_d comme un total, permet de définir :

$$\widehat{N}_d = \sum\nolimits_s \frac{z_{dk}}{\pi_k} = \sum\nolimits_{s_d} \frac{1}{\pi_k}. \tag{6.10}$$

Enfin on applique la technique d'estimation d'un ratio pour estimer la moyenne sur U_d.

1. L'estimateur de la moyenne retenu est le rapport des estimateurs des totaux des numérateur et dénominateur :

$$\widetilde{y}_{s_d} = \frac{\sum_s \frac{y_k z_{dk}}{\pi_k}}{\sum_s \frac{z_{dk}}{\pi_k}} = \frac{\widehat{t}_{y,U_d}}{\widehat{N}_d}. \tag{6.11}$$

2. La linéarisée est :

$$\nu_k = \frac{1}{N_d}(y_k z_{dk} - \overline{y}_{U_d} z_{dk}).$$

La variance approchée de $\widetilde{y}_{s_d}$ est donc :

$$\mathsf{var}_{\mathrm{app}}(\widetilde{y}_{s_d}) = \frac{1}{N_d^2} \sum \sum\nolimits_U \Delta_{kl} \frac{y_k z_{dk} - \overline{y}_{U_d} z_{dk}}{\pi_k} \; \frac{y_l z_{dl} - \overline{y}_{U_d} z_{dl}}{\pi_l}.$$

Comme $z_{dk} = 0$ si $k \notin U_d$, ceci se réduit à

$$\mathsf{var}_{\mathrm{app}}(\widetilde{y}_{s_d}) = \frac{1}{N_d^2} \sum \sum\nolimits_{U_d} \Delta_{kl} \frac{y_k - \overline{y}_{U_d}}{\pi_k} \; \frac{y_l - \overline{y}_{U_d}}{\pi_l}.$$

3. Enfin la variance approchée est estimée par :

$$\widehat{\mathsf{var}}(\widetilde{y}_{s_d}) = \frac{1}{\widehat{N}_d^2} \sum \sum\nolimits_{s_d} \frac{\Delta_{kl}}{\pi_{kl}} \frac{y_k - \widetilde{y}_{s_d}}{\pi_k} \; \frac{y_l - \widetilde{y}_{s_d}}{\pi_l}.$$

6.2.3 Cas d'un plan SI

Si le plan est SI(N, n) sur U alors :

$$\widehat{N}_d = n_d \frac{N}{n}, \qquad \widehat{t}_{y,U_d} = \frac{N}{n} \sum\nolimits_{s_d} y_k. \tag{6.12}$$

et

$$\widetilde{y}_{s_d} = \frac{\frac{N}{n} \sum_{s_d} y_k}{n_d \frac{N}{n}} = \overline{y}_{s_d}.$$

Posons $v_k = z_{dk}(y_k - \overline{y}_{s_d})$, on vérifie facilement que $\sum_s v_k = 0$. Nous estimons maintenant la variance de $\widetilde{y}_{s_d}$ à l'aide des résultats (6.4) à (6.6). L'estimation de la variance est :

$$\widehat{\mathsf{var}}(\widetilde{y}_{s_d}) = \frac{1}{\overline{z}_{ds}^2}\left(\frac{1}{n} - \frac{1}{N}\right) S^2_{yz_d - \overline{y}_{s_d} z_d, s} = \left(\frac{n}{n_d}\right)^2 \left(\frac{1}{n} - \frac{1}{N}\right) S^2_{vs}.$$

et

$$(n-1)S^2_{yz_d - \overline{y}_{s_d} z_d, s} = \sum\nolimits_s v_k^2 = \sum\nolimits_{s_d} v_k^2 + \sum\nolimits_{s - s_d} v_k^2$$

$$\sum\nolimits_{s_d} v_k^2 = (n_d - 1) S^2_{y, s_d}$$

$$\sum\nolimits_{s - s_d} v_k^2 = 0.$$

Finalement :

$$\widehat{\mathsf{var}}(\overline{y}_{s_d}) = \left(\frac{n}{n_d}\right)^2 \left(\frac{1}{n} - \frac{1}{N}\right) \frac{n_d - 1}{n - 1} S^2_{y, s_d} \simeq \frac{1 - f}{n_d} S^2_{y, s_d}$$

Remarques et compléments.

1. Dans l'estimation sur un domaine, il ne faut pas oublier que le plan porte sur U, une population qui contient strictement le domaine, d'où la nécessité d'introduire la variable z_d pour se ramener à U.
2. Observons que $\widehat{t}_{y, U_d}$ est basé sur un échantillon de taille aléatoire : $n_d = \text{card}(s) \cap U_d$. Donc, pratiquement, on n'attachera pas la même confiance à une telle estimation selon qu'elle est basée sur peu ou sur beaucoup d'observations. On peut cependant calculer la taille moyenne du sous-échantillon s_d :

   $$E(n_d) = \sum\nolimits_U z_{dk} \pi_k = \sum\nolimits_{U_d} \pi_k$$

   On peut ainsi savoir, avant tirage de l'échantillon, si le domaine sera bien représenté en moyenne. On peut de même calculer la variance de la taille.

Chapitre 7

Estimation par régression

Dans ce chapitre nous étudions l'estimation par régression.

7.1 Introduction

On dispose souvent, quand on doit faire un sondage, d'une *information auxiliaire* sous la forme d'une ou plusieurs variables x, connues pour tous les individus de la population U et corrélées avec la variable d'étude y. Les techniques d'estimation par régression servent à incorporer cette information auxiliaire dans les estimateurs par sondage. L'information auxiliaire peut être, par exemple : la variable d'étude connue à une époque antérieure, des variables proches de la variable d'étude mais peu coûteuses à observer. On verra que souvent, on n'utilise que la somme sur la population d'une variable auxiliaire et ses valeurs sur l'échantillon et non sur toute la population.

On a utilisé une telle information dans le plan stratifié et dans l'estimation post-stratifiée. En effet, définissons $x_{hk} = 1$ si $k \in U_h, = 0$ sinon, alors les effectifs des strates sont $N_h = \sum_U x_{hk}$. Dans le plan stratifié on utilise cette information pour définir le plan de sondage alors que dans la post-stratification on l'utilise pour corriger l'estimateur de H-T obtenu sans cette information. C'est cette dernière idée qui est mise en œuvre dans l'estimation par régression.

Dans le présent chapitre, on étudie d'abord l'estimation du total par ratio puis par différence. L'estimateur par différence dépend de paramètres rarement connus. On définit parallèlement un *modèle de superpopulation* ; c'est un modèle de régression qui suppose que la population finie U qui nous intéresse est elle-même tirée d'une population infinie. Ce cadre permet d'étendre l'estimateur par différence quand ses paramètres sont inconnus. De plus on verra que suivant le modèle de superpopulation adopté, on peut obtenir l'estimateur par ratio ou l'estimateur poststratifié.

7.2 Estimation par ratio

7.2.1 Définition

On s'intéresse à l'estimation du total t_y d'une variable d'étude y. On suppose que l'on dispose d'une variable auxiliaire x pour laquelle on connait le total t_x sur toute la population et qui est bien corrélée avec la variable d'étude. On définit l'estimateur par ratio du total $\widehat{t}_{y\mathtt{ra}}$ par :

$$\widehat{t}_{y\mathtt{ra}} = \hat{t}_{y\pi} \times \frac{t_x}{\hat{t}_{x\pi}}$$

On peut interpréter cette définition comme "une règle de trois" qui permet d'ajuster l'estimateur par les valeurs dilatés $\hat{t}_{y\pi}$ d'un coefficient multiplicatif qui tient compte de la qualité de l'estimation du total

par l'estimateur de H-T pour la variable x pour l'échantillon tiré. On peut aussi écrire :

$$\widehat{t}_{y\mathtt{ra}} = \widehat{R} \times t_x.$$

Cette dernière égalité définit l'estimateur par ratio comme l'estimateur d'un ratio multiplié par une constante t_x (non aléatoire). A partir des formules de variances de $\widehat{R}$ du chapitre 6, on déduit facilement la variance de l'estimateur par ratio.

7.2.2 Propriétés de l'estimateur par ratio

1. $\widehat{t}_{y\mathtt{ra}}$ est approximativement sans biais pour t_y
2. d'après les résultats (5) et (6) du chapitre 6, dans le cas d'un plan SI, une approximation de sa variance est
$$\mathsf{var}(\widehat{t}_{y\mathtt{ra}}) = t_{xU}^2 \mathsf{var}(\widehat{R}) \simeq N^2 \left(\frac{1}{n} - \frac{1}{N}\right) S^2_{y-Rx,U}. \tag{7.1}$$
mais
$$\widehat{\mathsf{var}}(\hat{R}) = \frac{1}{\overline{x}_s^2} \left(\frac{1}{n} - \frac{1}{N}\right) S^2_{y-\hat{R}x,s}.$$
3. On estime $\mathsf{var}(\widehat{t}_{y\mathtt{ra}})$ par
$$\widehat{\mathsf{var}}(\widehat{t}_{y\mathtt{ra}}) = N^2 \frac{\overline{x}_U^2}{\overline{x}_s^2} \left(\frac{1}{n} - \frac{1}{N}\right) S^2_{y-\hat{R}x,s}. \tag{7.2}$$

Bien noter le facteur $\dfrac{\overline{x}_U^2}{\overline{x}_s^2}$ qui arrive quand on passe de l'estimateur à l'estimation. On rencontre parfois l'estimateur de variance :

$$N^2 \left(\frac{1}{n} - \frac{1}{N}\right) S^2_{y-\hat{R}x,s} \tag{7.3}$$

obtenu par une subsitution directe dans (7.1). Si $\overline{x}_U \simeq \overline{x}_s$, les deux estimateurs sont proches. Observons que (7.3) est l'estimation de la variance du total des résidus $y_k - \widehat{R}x_k$.

7.2.3 Efficacité de l'estimateur par ratio pour le plan SI

Examinons le rapport : variance de l'estimateur par ratio sur variance de l'estimateur par les valeurs dilatées dans le plan SI :

$$\frac{S^2_{yU} - 2\widehat{R}\, S_{yx,U} + R^2\, S^2_{xU}}{S^2_{yU}} = 1 - 2R\, \frac{S_{yx,U}}{S^2_{yU}} + R^2\, \frac{S^2_{xU}}{S^2_{yU}}.$$

Dans cette expression $R = t_y/t_x = \overline{y}_U/\overline{x}_U$. Introduisons les coefficients de variation : $\mathrm{cv}(yU) = S_{t=yU}/\overline{y}_U$ et $\mathrm{cv}(xU)$, alors on voit que ce rapport est < 1 si le coefficient de corrélation entre y et x, ρ vérifie

$$\rho \geq \frac{1}{2} \frac{\mathrm{cv}(xU)}{\mathrm{cv}(yU)}$$

L'amélioration de l'estimation du total quand on passe à l'estimation par ratio, dépend de la situation du coefficient de corrélation par rapport au rapport des variabilités relatives de x et y.