Estimation par le quotient

Myriam Maumy[1]

[1]IRMA, Université Louis Pasteur
Strasbourg, France

Master 1ère Année 26-11-2014

Ce chapitre s'appuie essentiellement sur le livre :

- « Méthodes statistiques des sondages »,

  Jean-Marie Grosbras,

  Economica.

Sommaire

Définition

L'estimation par le quotient est une méthode de redressement d'échantillon sur des variables quantitatives.

Sommaire

Exemple 1

- Un échantillon d'établissements industriels réalise un produit.
- La quantité moyenne produite par les établissements échantillonnés, pour un mois donné, est égale à :

$$\widehat{\mu}_y = 23\,550.$$

- **Hypothèse :** le volume de la production est lié à la consommation. La consommation moyenne, pour ce même mois, est égale à :

$$\widehat{\mu}_x = 3\,225.$$

D'autre part, nous savons par une autre source qu'en fait, la consommation moyenne du mois sur l'ensemble des établissements industriels est égale à

$$\mu_X = 3\,350.$$

Remarque

$\widehat{\mu}_x = 3\,225$ est légèrement inférieure à la vraie valeur μ_X.

Par conséquent nous sommes amenés à penser qu'il existe vraisemblablement une différence du même ordre entre $\overline{x}$ et $\overline{X}$.

- Nous redressons les résultats pour les caler sur la moyenne de la population, à savoir $\overline{Y} = 3\,350$.
- La production moyenne estimée après ce calage est égale à :

$$\overline{y}_Q = 23\,550 \times \frac{3\,350}{3\,225} = 24\,463.$$

Nous apportons donc une correction égale à

$$\frac{3\,350}{3\,225} \quad \text{soit encore} \quad 3,4\%.$$

Sommaire

But

Estimer une moyenne μ, un total T, une proportion π,...

Principe

Pour atteindre notre but, nous utilisons le **principe** suivant :

- L'échantillon fournit une estimation ponctuelle $\overline{x}$.
- D'autre part, nous connaissons la moyenne échantillonnée $\overline{x}$ d'une variable auxiliaire X.
- Or, nous connaissons également la vraie moyenne $\overline{Y}$ de cette variable auxiliaire Y.

Définition

L'estimateur de la moyenne $\overline{x}_Q$ par la méthode de l'estimation par le quotient est égal à :

$$\widehat{\mu}_{Q,y} = \mu_x \frac{\widehat{\mu}_y}{\widehat{\mu}_x}.$$

Remarques

- Dans le paragraphe suivant, nous montrerons que $\overline{x}_Q$ est en général un estimateur **biaisé**.
- Malgré tout, l'estimateur $\overline{x}_Q$ peut-être plus précis que l'estimateur $\overline{x}$.

Remarque

Il est important que les données portant sur $\overline{y}$ soient bien dans le même champ que celles sur $\overline{Y}$.
Autrement dit, les y_i sont parmi les Y_i.

Remarque

Une des difficultés techniques de **l'estimation par le quotient** réside dans la présence de l'estimateur $\overline{y}$ au dénominateur de l'estimateur $\overline{x}_Q$.

Pourquoi ?

Parce que l'estimateur $\overline{y}$ est une variable aléatoire.

Les méthodes présentées dans ce chapitre peuvent être utilisées dans les cas où les phénomènes étudiés sont caractérisés par des quantités aléatoires au dénominateur.

Sommaire

Remarque

Comme dans le chapitre 5, nous utilisons les développements limités pour les calculs. On remarque que

$$\begin{aligned}\overline{x}_Q &= \overline{Y}\,\frac{\overline{x}}{\overline{y}}\\ &= \overline{Y}\,\frac{\overline{x}-\overline{X}+\overline{X}}{\overline{y}-\overline{Y}+\overline{Y}}\\ &= \overline{X}\,\frac{1+\dfrac{\overline{x}-\overline{X}}{\overline{X}}}{1+\dfrac{\overline{y}-\overline{Y}}{\overline{Y}}}.\end{aligned}$$

- Si l'échantillon est correct, c'est-à-dire si la remarque évoquée au paragraphe précédent a bien été suivie, alors $\overline{y}$ est un estimateur sans biais de la moyenne $\overline{Y}$.
- Nous pouvons donc espérer que la quantité suivante

$$\frac{\overline{y} - \overline{Y}}{\overline{Y}}$$

est très petite afin de faire les développements limités.

On transforme l'expression

$$\overline{x}_Q = \overline{X}\,\frac{1 + \dfrac{\overline{x} - \overline{X}}{\overline{X}}}{1 + \dfrac{\overline{y} - \overline{Y}}{\overline{Y}}}$$

à l'aide d'un développement limité.

En se limitant de l'ordre 2, nous avons

$$\begin{aligned}\overline{x}_Q &\simeq \overline{X}\left(\left(1+\frac{\overline{x}-\overline{X}}{\overline{X}}\right)\left(1-\frac{\overline{y}-\overline{Y}}{\overline{Y}}+\frac{\left(\overline{y}-\overline{Y}\right)^2}{\overline{Y}^2}\right)\right)\\ &\simeq \overline{X}\left(1+\frac{\overline{x}-\overline{X}}{\overline{X}}-\frac{\overline{y}-\overline{Y}}{\overline{Y}}-\frac{\overline{x}-\overline{X}}{\overline{X}}\frac{\overline{y}-\overline{Y}}{\overline{Y}}+\frac{\left(\overline{y}-\overline{Y}\right)^2}{\overline{Y}^2}\right).\end{aligned}$$

Comme $\overline{x}$ est un estimateur sans biais de la moyenne $\overline{X}$, nous avons

$$\mathbb{E}\left[\overline{x}\right] = \overline{X}.$$

De même pour $\overline{y}$,

$$\mathbb{E}\left[\overline{y}\right] = \overline{Y}.$$

On a alors

$$\mathbb{E}\left[\overline{x}_Q\right] \simeq \overline{X}\left(1 + 0 - 0 - \frac{\mathrm{Cov}\left[\overline{x}, \overline{y}\right]}{\overline{X}\,\overline{Y}} + \frac{\mathrm{Var}\left[\overline{y}\right]}{\overline{Y}^2}\right).$$

Le biais de $\mathbb{E}[\overline{x}_Q] - \overline{X}$, noté $B(\overline{x}_Q)$ est à peu près égal à

$$B(\overline{x}_Q) \simeq \overline{X}\left(\frac{\mathrm{Var}[\overline{y}]}{\overline{Y}^2} - \frac{\mathrm{Cov}[\overline{x}, \overline{y}]}{\overline{X}\,\overline{Y}}\right).$$

Il ne reste plus maintenant qu'à préciser la méthode de tirage,

- Tirage à probabilités égales avec remise (PEAR).
- Tirage à probabilités égales sans remise (PESR)

afin d'aller "plus loin" dans les calculs ou voir tout simplement si il n'y a pas de simplifications !

Tirage à PEAR

$$\mathrm{Var}[\overline{y}] = \frac{\sigma_Y^2}{n} \quad \text{et} \quad \mathrm{Cov}(\overline{x}, \overline{y}) = \frac{\sigma_{XY}}{n},$$

où

$$\sigma_{XY} = \frac{1}{N} \sum_{i=1}^{N} (X_i - \overline{X})(Y_i - \overline{Y}).$$

Donc, dans le cas du tirage à PEAR, nous avons

$$B(\overline{x}_Q) \simeq \frac{\overline{X}}{n} \left(\frac{\sigma_Y^2}{\overline{Y}^2} - \frac{\sigma_{XY}}{\overline{X}\,\overline{Y}} \right).$$

Tirage à PESR

$$\mathrm{Var}[\overline{y}] = \frac{N-n}{Nn} S^2_{Y,c} \quad \text{et} \quad \mathrm{Cov}(\overline{x}, \overline{y}) = \frac{N-n}{Nn} S_{XY,c},$$

où

$$S_{XY,c} = \frac{1}{N-1} \sum_{i=1}^{N} (X_i - \overline{X})(Y_i - \overline{Y}).$$

Donc, dans le cas du tirage à PESR, nous avons

$$B(\overline{x}_Q) \simeq \frac{N-n}{Nn} \overline{X} \left(\frac{S^2_{Y,c}}{\overline{Y}^2} - \frac{S_{XY,c}}{\overline{X}\,\overline{Y}} \right).$$

Remarques sur les deux dernières formules

- Dans les deux cas de tirage, le biais est en $1/n$, donc le biais devient négligeable pour des “gros” échantillons...
- Il est sans intérêt d'avoir une corrélation négative entre X et Y. Si le coefficient de variation de $\overline{y}$ est petit, c'est-à-dire s'il n'y a pas de distorsion de l'échantillon de la variable auxiliaire, le biais de $\overline{x}_Q$ est négligeable devant la variance de l'estimateur $\overline{y}$.
- Le biais $B(\overline{x}_Q)$ est nul ssi (cf. TD numéro 5, exercice 1)

$$\frac{\mathrm{Var}[\overline{y}]}{\mathrm{Cov}(\overline{x}, \overline{y})} = \frac{\overline{Y}}{\overline{X}} \quad \text{ou encore} \quad \frac{\sigma_x}{\mu_x} = \rho(X, Y)\frac{\sigma_y}{\mu_y}.$$

Sommaire

Remarque

Comme nous venons de le démontrer, **l'estimateur $\overline{x}_Q$ est biaisé.**

La précision

La précision est mesurée par l'erreur quadratique moyenne (EQM) qui se définit ainsi :

$$\text{EQM}\left(\widehat{\mu}_{Q,y}\right) = \mathbb{E}\left[\left(\overline{x}_Q - \overline{X}\right)^2\right].$$

Pour avoir une idée sur l'EQM($\overline{x}_Q$), calculons d'abord la quantité :

$$\overline{x}_Q - \overline{X}.$$

Ensuite, nous éléverons cette quantité au carré et nous en « prendrons » l'espérance.

Par définition de l'estimateur $\overline{x}_Q$, nous avons

$$\overline{x}_Q - \overline{X} = \frac{\overline{x}}{\overline{y}}\overline{Y} - \overline{X}.$$

D'après le développement limité déjà effectué et en se limitant toujours à l'ordre 2, nous avons

$$\overline{x}_Q - \overline{X} \simeq \overline{X}\left(\frac{\overline{x}-\overline{X}}{\overline{X}} - \frac{\overline{y}-\overline{Y}}{\overline{Y}} - \frac{\overline{x}-\overline{X}}{\overline{X}}\frac{\overline{y}-\overline{Y}}{\overline{Y}} + \frac{\left(\overline{x}-\overline{X}\right)^2}{\overline{X}^2}\right).$$

On élève maintenant cette quantité au carré toujours en se limitant à l'ordre 2. On a alors

$$(\overline{x}_Q - \overline{X})^2 \simeq \overline{X}^2 \left(\left(\frac{\overline{x} - \overline{X}}{\overline{X}} \right)^2 - 2\frac{\overline{x} - \overline{X}}{\overline{X}} \frac{\overline{y} - \overline{Y}}{\overline{Y}} + \left(\frac{\overline{y} - \overline{Y}}{\overline{Y}} \right)^2 \right)$$

$$\Rightarrow \mathbb{E}\left[(\overline{x}_Q - \overline{X})^2 \right] \simeq \mathrm{Var}[\overline{x}] - 2\frac{\overline{X}}{\overline{Y}} \mathrm{Cov}(\overline{x}, \overline{y}) + \frac{\overline{X}^2}{\overline{Y}^2} \mathrm{Var}[\overline{y}].$$

Tirage à PESR

Dans le cas d'un tirage à probabilités égales sans remise, la formule devient

$$\mathrm{EQM}(\overline{x}_Q) \simeq \frac{N-n}{Nn}\left(S^2_{X,c} - 2\,\frac{\overline{X}}{\overline{Y}}\,S_{XY,c} + \left(\frac{\overline{X}}{\overline{Y}}\right)^2 S^2_{Y,c}\right).$$

Tirage à PESR

Cette quantité peut-être estimée par

$$\widehat{\mathrm{EQM}}(\overline{x}_Q) \simeq \frac{N-n}{Nn}\left(s_X^2 - 2\,\frac{\overline{x}}{\overline{y}}\,s_{XY,c} + \left(\frac{\overline{x}}{\overline{y}}\right)^2 s_Y^2\right),$$

où

$$s_{XY,c} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}).$$

Tirage à PEAR

Dans le cas d'un tirage à probabilités égales avec remise, la formule devient

$$\mathrm{EQM}(\overline{x}_Q) \simeq \frac{N-n}{Nn}\left(S_X^2 - 2\,R\,S_{XY} + R^2\,S_Y^2\right),$$

où la quantité R se définit par

$$R = \frac{\overline{X}}{\overline{Y}}.$$

R est aussi ce que l'on appelle un ratio.

Sommaire

Question

La question naturelle qui se pose :

Est-ce que le redressement opéré est, malgré le biais introduit, avantageux par rapport à l'estimateur simple $\overline{x}$?

Comparaison lors d'un tirage PESR

On rappelle que

$$\mathrm{Var}\left[\,\overline{x}_{SR}\,\right] = \frac{N-n}{Nn} S^2_{X,c}$$

et

$$\mathrm{EQM}(\overline{x}_Q) \simeq \frac{N-n}{Nn}\left(S^2_{X,c} - 2\,R\,S_{XY,c} + R^2\,S^2_{Y,c}\right).$$

D'où

$$\mathrm{Var}\left[\overline{x}_{PESR}\right] - \mathrm{EQM}(\overline{x}_Q) > 0 \quad \Leftrightarrow \quad 2RS_{XY,c} - R^2 S^2_{Y,c} > 0.$$

Le redressement par le quotient est donc profitable si cette inégalité est vérifiée.

En d'autres termes, $\overline{x}_Q$ est préférable à l'estimateur d'un SAS, si entre X et Y il existe une relation qui est dans le voisinage de la proportionnalité.

Concrètement, nous calculons

$$\overline{x}, \quad \overline{y}, \quad s_X^2, \quad s_Y^2, \quad s_{XY}.$$

Si

$$\frac{s_{XY}}{s_Y^2} > \frac{1}{2}\frac{\overline{x}}{\overline{y}},$$

alors nous faisons le redressement.

Sinon, nous faisons un sondage aléatoire simple.

Sommaire

Dans l'expression de l'estimateur par le quotient $\overline{x}_Q$, nous faisons intervenir l'estimateur $\overline{x}$.

Il existe plusieurs méthodes de sondage que vous connaissez maintenant qui permettent de construire un estimateur de la moyenne d'une population.

Par exemple, nous pouvons donc regarder ce qui va se passer si nous décidons de choisir comme estimateur de la moyenne, l'estimateur de la moyenne construit par stratification (cf le chapitre correspondant).

Rappel

$$\overline{X} = \sum_h \frac{N_h}{N} \overline{X}_h.$$

Nous allons adopter la méthode suivante :

$$\overline{x}_{QSt} = \sum_h \frac{N_h}{N} \overline{x}_{Qh},$$

où

$$\overline{x}_{Qh} = \frac{\overline{x}_h}{\overline{y}_h} \overline{Y}_h,$$

c'est-à-dire nous pouvons faire les redressements par strate.

$$\begin{aligned}\mathbb{E}\left[\overline{x}_{QSt}\right] &= \sum_h \frac{N_h}{N}\mathbb{E}\left[\overline{x}_{Qh}\right] \\ &= \sum_h \frac{N_h}{N}\overline{X}_h + \sum \frac{N_h}{N} B_h.\end{aligned}$$

Le biais est donc la moyenne pondérée des biais locaux.

$$\text{EQM}\left[\overline{x}_{QSt}\right] = \sum_h \frac{N_h^2}{N^2}\text{EQM}\left[\overline{x}_{Qh}\right].$$

Dans le cas d'un sondage à probabilités égales sans remise, nous avons

$$\text{EQM}\left[\overline{x}_{QSt}\right] \simeq \sum_h \frac{N_h^2}{N^2}\frac{N_h - n_h}{N_h n_h}\left(S_{X,h}^2 - 2R_h S_{XY,h} + R_h^2 S_{Y,h}^2\right).$$

Un autre cas que l'on peut envisager : calculer les estimateurs stratifiés $\overline{x}_{St}$ et $\overline{y}_{St}$.

$$\overline{x}_{St} = \sum_h \frac{N_h}{N} \overline{x}_h \quad \text{et} \quad \overline{y}_{St} = \sum_h \frac{N_h}{N} \overline{y}_h.$$

Ensuite opérer un redressement global :

$$\overline{x}_{QSt} = \frac{\overline{x}_{St}}{\overline{y}_{St}} \overline{Y}.$$

Sondage PESR

Dans le cas d'un sondage PESR, nous avons

$$\mathrm{EQM}[\overline{x}_{QSt}] \simeq \sum_{h} \frac{N_h^2}{N^2} \frac{N_h - n_h}{N_h n_h} \left(S_{X,h}^2 - 2 R_h S_{XY,h} + R_h^2 S_{Y,h}^2 \right).$$

La première méthode

La première méthode est à employer si les R_h sont plutôt variables d'une strate à l'autre et si les effectifs n_h sont assez grands pour que les biais locaux soient négligeables.

La seconde méthode

La seconde méthode est à employer si les R_h sont stables ou si les échantillons par strate sont d'effectif faible.