JOURNAL DE LA SOCIÉTÉ STATISTIQUE DE PARIS

P. THIONET

Sur une extension de l'estimation sans biais (théorie des sondages)

Journal de la société statistique de Paris, tome 110 (1969), p. 144-151

<http://www.numdam.org/item?id=JSFS_1969__110__144_0>

SUR UNE EXTENSION DE L'ESTIMATION SANS BIAIS
(Théorie des sondages)

Dans un article qui n'est plus très récent (1), DURBIN indique une extension très commode de la notion d'estimation sans biais, — tout en en attribuant à KENDALL la paternité (2). Transposant cette idée de la théorie statistique à la théorie des sondages, nous obtenons une généralisation de la condition classique d'estimation sans biais :

$$\mathcal{E}\, T(\mathcal{S}) = \theta(\Omega), \quad \text{ou} \quad \mathcal{E}\,[T(\mathcal{S}) - \theta] = 0 \tag{1}$$

T désignant le vecteur d'estimateurs du vecteur des paramètres θ,

$\mathcal{S}$ désignant l'échantillon, et Ω la population sondée,

$\mathcal{E}$ désignant l'opérateur : espérance mathématique.

Il s'agit simplement d'écrire une *équation d'estimation sans biais*, et non un estimateur sans biais à proprement parler.

I. — 1re EXTENSION : ÉQUATION LINÉAIRE D'ESTIMATION

Soit $T(\mathcal{S}) = P(\mathcal{S})/Q(\mathcal{S})$

un certain estimateur de $\theta = \frac{\xi}{\eta}$; par exemple $\begin{cases} \theta = \sum_{(\Omega)} x / \sum_{(\Omega)} y \\ T = \sum_{(\mathcal{S})} x / \sum_{(\mathcal{S})} y \end{cases}$

La condition (1) est souvent beaucoup moins maniable que la condition (2) :

ou
$$\left.\begin{aligned} \mathcal{E}\,[P - \theta \cdot Q] &= 0 \\ \mathcal{E}\,[Q(T - \theta)] &= 0 \end{aligned}\right\} \tag{2}$$

Supposons T un estimateur *cohérent* (*consistent*) de θ, c'est-à-dire

$$\theta = \frac{P(\Omega)}{Q(\Omega)} \tag{3}$$

ou
$$P(\Omega) - \theta\, Q(\Omega) = 0 \tag{3'}$$

La condition (2) s'écrit (car θ n'est pas aléatoire) :

$$\mathcal{E}\,[P(\mathcal{S})] - \theta\ \mathcal{E}\,[Q(\mathcal{S})] = 0 \tag{2'}$$

relation vérifiée si on a :

$$\mathcal{E}\,[P(\mathcal{S})] = P(\Omega), \quad \mathcal{E}\,[Q(\mathcal{S})] = Q(\Omega) \tag{4}$$

ou encore :

$$\mathcal{E}\,[P(\mathcal{S})] = k\,P(\Omega), \quad \mathcal{E}\,[Q(\mathcal{S})] = k\,Q(\Omega) \tag{4'}$$

(par exemple : $k = (n-1)/n$, n étant la taille d'un échantillon de Bernoulli).

1. DURBIN (J.), Estimation of parameters in time-series regression models, Journal of the Royal Statistical Society, Series B (1960), 22 N° 1, pp. 139-153.
2. KENDALL (M.), Biometrika 38 (1951), pp. 11-25.

Erreur d'échantillonnage : nous pouvons retenir, pour mesurer les erreurs d'échantillonnage, le critère

$$\mathcal{E}\,[P\,(\mathcal{S}) - \theta\,Q\,(\mathcal{S})]^2 = W \tag{5}$$

c'est-à-dire (compte tenu de (3'))

$$\mathcal{V}\,P - 2\,\theta\,\mathrm{Cov}\,(P, Q) + \theta^2\,\mathcal{V}\,Q = W$$

Or il est bien connu que le calcul de la variance exacte d'un *ratio* P/Q est inaccessible en pratique des sondages, et qu'on lui substitue le calcul de :

$$\omega = \gamma^2 - 2\,\rho\,\gamma\,\gamma' + \gamma'^2$$

avec

$$\begin{aligned} \gamma^2 &= \mathcal{V}\,(P)/\xi^2, \quad \xi = \mathcal{E}\,P \\ \gamma'^2 &= \mathcal{V}\,(Q)/\eta^2, \quad \eta = \mathcal{E}\,Q \\ \rho\,\gamma\,\gamma' &= \mathrm{Cov}\,(P, Q)/\xi\,\eta \end{aligned}$$

On constate que :

$$W = \xi^2\,\omega$$

ce qui montre que les errements des praticiens sont en bon accord avec la méthode inspirée de DURBIN.

Remarque : Quand on écrit

$$\mathcal{E}\,(T - \theta) = 0$$

on exprime que $\mathcal{E}\,(T - \theta)^2$ est minimum si θ est estimé sans biais par T. Cette relation subsiste :

$$\mathcal{E}\,(P - \theta\,Q)^2 = W + (\mathcal{E}\,P - \theta\,\mathcal{E}\,Q)^2$$

a pour minimum

$$\mathcal{E}\,(P - \theta\,Q)^2 = \mathcal{V}\,(P - \theta\,Q) = W$$

si

$$\mathcal{E}\,(P - \theta\,Q) = \mathcal{E}\,P - \theta \cdot \mathcal{E}\,Q = 0$$

Applications :

1) *Estimation classique par ratio* (Sondage bernoullien) :

$$\begin{aligned} P &= \textstyle\sum_{(\mathcal{S})} x, \quad Q = \sum_{(\mathcal{S})} y \\ \mathcal{V}\,P &= \sigma^2/n, \quad \mathcal{V}\,Q = \sigma'^2/n, \quad \mathrm{Cov}\,(P, Q) = \rho\,\sigma\,\sigma'/n \end{aligned}$$

avec

$$\begin{aligned} &\mathcal{V}\,x = \sigma^2, \quad \mathcal{V}\,y = \sigma'^2, \quad \mathrm{Cov}\,(x, y) = \rho\,\sigma\,\sigma' \\ &\mathcal{E}\,x = \xi, \quad \mathcal{E}\,y = \eta \\ &W = \frac{1}{n}\left(\sigma^2 - 2\,\frac{\xi}{\eta}\,\rho\,\sigma\,\sigma' + \frac{\xi^2}{\eta^2}\,\sigma'^2\right) \end{aligned}$$

2) *Estimation stratifiée par ratio*

On peut définir la répartition optimale de l'échantillon n, — ou des ressources financières c, — entre les strates $h = 1,2 \ldots K$, — comme étant celle qui rend W minimum (pour n ou c donnés).

Soit $\sum_h c_h n_h = c$ le coût constant.

2.1) Pour une variable x, l'estimateur sans biais P est :

$$P = \sum_h \frac{N_h}{N} \bar{x}_h. \quad \bar{x}_h = \sum_{(s)} x_h / n_h$$

avec

$$\mathcal{V} P = \sum_h \left(\frac{N_h}{N}\right)^2 \cdot \frac{\sigma_h^2}{n_h} \cdot \frac{N_h - n_h}{N_h - 1}$$

qui est minimum pour : $n_h \div N_h \sigma_h / \sqrt{c_h}$ (théorème de Neyman-Yates).

Il s'agit au contraire ici de trouver les n_h qui rendent minimum :

$$W = \sum_h \left(\frac{N_h}{N}\right)^2 \frac{1}{n_h} \left(\sigma_h^2 - 2\frac{\xi}{\eta} \rho_h \sigma_h \sigma_{h'} + \frac{\xi^2}{\eta^2} \sigma_{h'}^2\right) \cdot \frac{N_h - n_h}{N_h - 1}$$

Ainsi les n_h devront être proportionnels aux $N_h \tau_h / \sqrt{c_h}$, avec :

$$\tau_h^2 = \left[\sigma_h^2 - 2\frac{\xi}{\eta} \rho_h \sigma_h \sigma_{h'} + \frac{\xi^2}{\eta^2} \sigma_{h'}^2\right] N_h/(N_h - 1)$$

2.2) Un problème différent se pose si l'on estime, disons $\overline{X}_h$, par ratio *dans chaque strate :*

$$\text{esti } \overline{X}_h = \frac{\bar{x}_h}{\bar{y}_h} \eta_h$$

et au total :

$$\text{esti } \overline{X} = \sum_h \frac{N_h}{N} \frac{\bar{x}_h}{\bar{y}_h} \eta_h$$

On peut choisir comme critère à minimer :

$$W = \sum_h \left(\frac{N_h}{N}\right)^2 \left(\frac{\eta_h}{\xi_h}\right)^2 W_h$$

qui est de la forme

$$W = \sum_h \left(\frac{N_h}{N}\right)^2 \lambda_h^2 W_h$$

La répartition optimale des ressources : $c = \sum_h n_h c_h$, est alors telle que les n_h soient proportionnels aux

$$N_h \lambda_h \dot{\tau}_h / \sqrt{c_h}$$

Or : $$\lambda_h^2 \dot{\tau}_h^2 = \left[\sigma_{h'}^2 - \frac{2}{\theta_h} \rho_h \sigma_h \sigma'_h + \frac{1}{\theta_h^2} \sigma_h^2\right] \frac{N_h}{N_h - 1}$$

Si l'on pose : $$\tau_h = \tau_h (x, y)$$

on a donc $$\lambda_h \dot{\tau}_h = \tau_h (y, x)$$

Pour choisir l'une ou l'autre de ces estimations, — donc l'une ou l'autre des répartitions de ressources, — on fera intervenir d'autres considérations, telles que commodité des calculs ou moindre biais (vraisemblablement).

3) *Estimation d'une pente de régression*

Soit $$\beta = \sum_{(\Omega)} (x_i - \overline{X})\,(y_i - \overline{Y}) \Big/ \sum_{(\Omega)} (x_i - \overline{X})^2$$

la pente de régression de Y en X; et sur un échantillon $\mathcal{S}$ de couples $(x_i\, y_i)$ considérons l'estimateur $b\ (\mathcal{S})$ consistant mais biaisé. Écrivons l'équation d'estimation :

$$\sum_{(\mathcal{S})} (x_i - \bar{x})\,(y_i - \bar{y}) - b \sum_{(\mathcal{S})} (x_i - \bar{x})^2 = 0$$

avec $$b = \text{esti}\ \beta$$

Posons

$$P\ (\mathcal{S}) = \sum_{(\mathcal{S})} (x_i - \bar{x})\,(y_i - \bar{y})/(n-1)$$
$$Q\ (\mathcal{S}) = \sum_{(\mathcal{S})} (y_i - \bar{y})^2/(n-1)$$

On a

$$\mathcal{E}\ P = \sum_{(\Omega)} (x_i - \overline{X})\,(y_i - \overline{Y})/N$$
$$\mathcal{E}\ Q = \sum_{(\Omega)} (x_i - \overline{X})^2/N$$

donc $$\mathcal{E}\ (P - \beta\ Q) = 0$$

On en déduit par exemple

$$W = \mathcal{V}\ P - 2\,\beta\ \text{Cov}\ (P, Q) + \beta^2\ \mathcal{V}\ Q$$

Remarque : Pour obtenir l'expression de $\mathcal{V}$ P, Cov (P, Q), $\mathcal{V}$ Q, on pourra utiliser les différences quadratiques moyennes

$$\sum\sum \left(\frac{x_i - x_j}{N}\right)^2, \sum\sum \left(\frac{x_i - x_j}{N}\right)\left(\frac{y_i - y_j}{N}\right), \sum\sum \left(\frac{y_i - y_j}{N}\right)^2.$$

4) *Remarque : Estimation d'un coefficient de corrélation.*

La formule

$$\rho = \frac{\sum (x_i - \overline{X})\,(y_i - \overline{Y})}{\sqrt{\sum (x_i - \overline{X})^2}\,\sqrt{\sum (y_i - \overline{Y})^2}}$$

récrite sous l'une des formes

$$\sum (x_i - \overline{X})\,(y_i - \overline{Y}) - \rho \sqrt{\sum (x_i - \overline{X})^2 \sum (y_i - \overline{Y})^2} = 0$$

ou bien

$$\left[\sum (x_i - \overline{X})\,(y_i - \overline{Y})\right]^2 - \rho^2 \sum (x_i - \overline{X})^2 \sum y_i - \overline{Y})^2 = 0$$

est encore linéaire (en ρ ou en ρ^2).

Toutefois on peut constater que la présente théorie ne s'applique ni à ρ, ni à ρ^2 :
En effet : On peut bien considérer les équations d'estimation :

soit $$P - \rho \sqrt{Q\,R} = 0$$

soit $$P^2 - \rho^2\ Q \cdot R = 0$$

R désignant l'expression en y homologue du Q (x) défini plus haut.

On aura : $\mathcal{E}\,P = \rho\,\sigma\,\sigma'$, $\mathcal{E}\,Q = \sigma^2$, $\mathcal{E}\,R = \sigma'^2$.

1er cas : $$\mathcal{E}\,(P - \rho\sqrt{Q\,R}) = \rho\,\sigma\,\sigma' - \rho\,\mathcal{E}\sqrt{Q\,R}$$

Il n'y a aucune raison que $\mathcal{E}\sqrt{Q\,R}$ soit égal à $\sigma\,\sigma'$.

Si l'on suppose les distributions de Q et R indépendantes, on aura seulement

$$\mathcal{E}\,(Q\,R) = \sigma^2\,\sigma'^2$$

2e cas : $$\mathcal{E}\,(P^2 - \rho^2\,Q\,R) = \mathcal{E}\,P^2 - \rho^2\,\mathcal{E}\,(Q\,R)$$

Si les distributions de Q et R peuvent être indépendantes, soit $\mathcal{E}\,(Q\,R) = \sigma^2\,\sigma'^2$, on a sans aucune doute :

$$\mathcal{E}\,P^2 - (\mathcal{E}\,P)^2 = \mathcal{E}\,(P - \mathcal{E}\,P)^2 > 0$$

donc $$\mathcal{E}\,P^2 > \rho^2\,\sigma^2\,\sigma'^2$$

Conclusion : La présente théorie ne s'applique donc pas au coefficient de corrélation de couples $(x_i\ y_i)$ tirés au sort.

Remarque : On envisage souvent le cas des x_i *choisis* (souvent équidistants) et des y_i *tirés au sort*, ce qui constitue un problème tout à fait différent. Il ne s'agit plus d'estimer la corrélation ρ dans une distribution double, mais la corrélation liée par le choix des x_i.

Cependant la présente théorie ne s'applique pas davantage, qu'on écrive

soit $$E\left[\frac{1}{\sqrt{Q}}\,P - \rho\sqrt{R}\right] = 0, \quad \text{soit } E\left[\frac{1}{Q}\,P^2 - \rho^2\,R\right] = 0;$$

puisqu'on a : $$E\sqrt{R} \neq \sigma'; \qquad E\,P^2 \neq (E\,P)^2$$

II. — 2e EXTENSION DE L'ESTIMATION SANS BIAIS

On vient de substituer à l'estimateur T de θ vérifiant

$$\mathcal{E}\,[T\,(\mathcal{S}) - \theta] = 0 \tag{1}$$

un estimateur P/Q de θ vérifiant

$$\mathcal{E}\,[P\,(\mathcal{S}) - \theta\,Q\,(\mathcal{S})] = 0 \tag{2}$$

Autrement dit, on a utilisé *l'équation linéaire* d'estimation :

$$P\,(\mathcal{S}) - \text{esti}\,\theta \cdot Q\,(\mathcal{S}) = 0$$

avec $$T = \text{esti}\,\theta$$

Or rien ne s'oppose *a priori* à ce qu'on utilise une équation d'estimation *quelconque*, ni un système d'équations d'estimations.

C'est ce qu'on fait déjà quand on écrit, soit les équations des moindres *carrés*, soit les équations du maximum de vraisemblance.

Par exemple, avec une loi continue de densité $f\,(x; \theta)$, on écrit le système d'équations

$$\sum \frac{\delta}{\delta\,\theta_j}\,\text{Log}\,f\,(x_i; \theta) = 0 \mid j = 1, 2, \ldots K$$

qui correspond à

$$\mathcal{E}\left[\frac{\delta}{\delta\theta_j}\operatorname{Log} f(x;\theta)\right] = \int f \frac{1}{f}\frac{\delta f}{\delta\theta_j}\,\mathrm{d}x = \int \frac{\delta f}{\delta\theta_j}\,\mathrm{d}x = 0$$

conséquence de $\int f\,\mathrm{d}x = 1$

L'estimation du maximum de vraisemblance est donc fournie par une équation (ou un système d'équations) d'estimation *sans biais.*

Le critère de qualité de l'estimation W n'est autre que l'information (ou la matrice d'information) de FISHER. Ces diverses remarques se trouvent dans l'article de Durbin mais aussi déjà dans notre thèse (1958). Nous avons (malheureusement) rarement (en pratique) l'occasion d'utiliser l'estimation du maximum de vraisemblance, avec la méthode des sondages (sauf si par exemple l'on cherche à ajuster après coup quelque loi de Pareto ou de Galton-Gibrat sur les revenus d'un échantillon de ménages).

Estimation d'une moyenne généralisée

Rappel : Toute relation de la forme :

$$\sum_{(\Omega)} f(x_i) = \mathrm{N} f(\zeta)$$

définit la moyenne généralisée ζ estimée sur échantillon par z tel que :

$$\sum_{(S)} f(x_i) = \mathrm{n} f(z)$$

On a $\mathcal{E} f(z) - f(\zeta) = 0$

Le critère $\mathrm{W} = \mathcal{E}[f(z) - f(\zeta)]^2$

mesure les erreurs d'échantillonnage. Si f dépend de certains paramètres, à estimer, on peut convenir que les estimations rendant W minimum sont optimales.

On appliquera le procédé à des couples, $x\,y$ (et plus généralement des K-tuples) de variables.

III. — IMPERFECTION DE CETTE THÉORIE

Rappels concernant la perte d'information

Étant donné un estimateur sans biais, on sait que sa variance a les propriétés d'une perte d'information (voir par exemple notre Étude théorique n° 7, I. N. S. E. E., 1957).

Quand on remplace la condition

$$\mathcal{E}\,\mathrm{T} - \theta = 0$$

par une condition telle que

$$\mathcal{E}(\mathrm{P} - \theta\,\mathrm{Q}) = 0$$

l'expression

$$\mathrm{W} = \mathcal{V}(\mathrm{P}) - 2\theta\,\mathrm{Cov}(\mathrm{P}\cdot\mathrm{Q}) + \theta^2\,\mathcal{V}(\mathrm{Q})$$

est une combinaison linéaire de 3 pertes d'information

$$\mathcal{V}(\mathrm{P}),\ \mathrm{Cov}(\mathrm{P}\,\mathrm{Q}),\ \mathcal{V}(\mathrm{Q}).$$

Si θ et θ^2 étaient *constants*, W serait, elle aussi, une perte d'information. Mais quand on passe d'une population de taille ν à un grand échantillon de taille N puis à un petit échantillon de taille n, l'expression θ se modifie. Montrons-le dans le cas d'un ratio.

Posons

$$t = \sum_{(n)} x / \sum_{(n)} y = \bar{x}/\bar{y}$$
$$T = \sum_{(N)} x / \sum_{(N)} y = \overline{X}/\bar{y}$$

et enfin

$$\theta = \sum_{(\nu)} x / \sum_{(\nu)} y = \xi/\eta$$

Supposons qu'il s'agisse de sondages exhaustifs. Rappelons d'abord les formules classiques :

$$\mathcal{V}\left[\sum_{(n)} x/n\right] = \frac{\nu\,\sigma^2}{\nu - 1}\left(\frac{1}{n} - \frac{1}{\nu}\right)$$
$$\mathcal{V}\left[\sum_{(N)} x/N\right] = \frac{\nu\,\sigma^2}{\nu - 1}\left(\frac{1}{N} - \frac{1}{\nu}\right)$$

donc

$$\mathcal{V}\left[\sum_{(n)} x/n\right] - \mathcal{V}\left[\sum_{(N)} x/N\right] = \frac{\nu\,\sigma^2}{\nu - 1}\left(\frac{1}{n} - \frac{1}{N}\right)$$

la variance du sondage du petit échantillon tiré du grand est

$$V\left[\sum_{n} x/n\right] = \frac{N\,s^2}{N - 1}\left(\frac{1}{n} - \frac{1}{N}\right)$$

où s^2 est la variance du grand échantillon. On a enfin

$$\mathcal{E}\left[\frac{N\,s^2}{N - 1}\right] = \frac{\nu\,\sigma^2}{\nu - 1}$$

d'où la relation caractérisant la perte d'information, écrite pour abréger :

$$\mathcal{V}(n) = \mathcal{V}(N) + \mathcal{E}\,V(n)$$

Passons au 2e terme : il n'y a pas de raison qu'on ait

$$\theta\,\mathrm{Cov}(n) = \theta\,\mathrm{Cov}(N) + \mathcal{E}[T - \mathrm{Cov}(n)]$$

c'est-à-dire :

$$\theta \sum_{(\nu)} \frac{(x - \xi)(y - \eta)}{\nu - 1} = \mathcal{E}\left[T \sum_{(N)} \frac{(x - \overline{X})(y - \overline{Y})}{N - 1}\right]$$

Alors qu'on est assuré d'avoir

$$\sum_{(\nu)} \frac{(x - \xi)(y - \eta)}{\nu - 1} = \mathcal{E}\left[\sum_{(N)} \frac{(x - \overline{X})(y - \overline{Y})}{N - 1}\right]$$

De même pour le 3e terme : Il n'y a pas de raison qu'on ait

$$\mathcal{E}\left[T^2 \sum_{(N)} \frac{(y - \overline{Y})^2}{N - 1}\right] = \theta^2 \sum_{(\nu)} \frac{(y - \eta)^2}{\nu - 1}$$

alors qu'on a certainement

$$\mathcal{E}\left[\sum_{(N)} \frac{(y - \overline{Y})^2}{N - 1}\right] = \sum_{(\nu)} \frac{(y - \eta)^2}{\nu - 1}$$

Conclusion : W n'est pas une perte d'information; alors que

$$\mathcal{V}(P) - 2\lambda \operatorname{Cov}(P, Q) + \lambda^2 \mathcal{V}(Q)$$

en est une [1] quel que soit le paramètre λ *indépendant du sondage.*

Ainsi la présente théorie de l'estimation par une équation d'estimation sans biais ne nous donne-t-elle pas pleinement satisfaction, puisqu'elle ne conserve ni ne transpose l'une des propriétés fondamentales de l'estimation sans biais.

Aurait-il fallu associer à $\mathcal{E}(P - \theta Q) = 0$ un critère mieux choisi que $W = \mathcal{E}(P - \theta Q)^2$? c'est assez vraisemblable.

Il y a cinq ans que nous nous posons la question et nous l'avons déjà posé quelques fois à nos étudiants de 3e cycle. Nous souhaitons qu'un lecteur plus imaginatif que nous suggère quelque solution à ce problème (s'il en possède).

P. THIONET

1. Plus généralement la forme quadratique $(u\ v)\ \mathcal{E}\begin{pmatrix} P^2 & PQ \\ PQ & Q^2 \end{pmatrix}\begin{pmatrix} u \\ v \end{pmatrix}$.