ESTIMATION D'UN QUOTIENT À L'AIDE DE DONNÉES AGRÉGÉES

La recherche des liens entre analyses aux niveaux agrégé et individuel est un thème d'intérêt commun en sciences humaines et, en particulier, en démographie. En effet le passage lors de l'interprétation de la corrélation macroscopique à la causalité individuelle est contestable. D'ailleurs si l'inférence d'hypothétiques comportements individuels faite à partir de résultats observés au niveau agrégé est remise en cause, des travaux récents en démographie (Courgeau, 1994) ont mis en évidence les liens formels qui existent entre les deux analyses et les apparentes contradictions qui résultent selon qu'elles sont conduites à ces deux niveaux.

L'objet de la présente note est d'apporter dans ce cadre théorique des précisions sur les différences d'estimation d'un quotient supposé constant par période selon que le mode d'observation du phénomène étudié est individuel ou agrégé.

Soit m un quotient instantané constant sur l'intervalle $(0, h)$ et une population soumise au risque dont la taille N est suffisamment importante pour que la théorie des grands nombres s'applique.

---

(30) « Les magistrats ont souvent rappelé aux coroners que le suicide devait être prouvé de manière incontestable et que s'il subsistait un doute raisonnable le coroner ou le jury devait rendre un « Verdict Ouvert » Stone, E., (1986), *op.cit.*, Chap 21.16, page 145.

Supposons d'abord que l'on dispose d'observations individuelles donnant, pour chaque individu soumis au risque, l'occurrence de l'événement observé M ($M = 0,1$) et la durée de séjour avant de connaître l'événement ou de sortir d'observation, Y. L'estimateur du maximum de vraisemblance du quotient, noté m_{ind}, est (Courgeau et Lelièvre, 1989) :

$$\hat{m}_{ind} = \frac{\sum_{\alpha=1}^{N} M_\alpha}{\sum_{\alpha=1} Y_\alpha} = \frac{M}{Y} \qquad [1]$$

où M est le nombre total d'événements et Y la somme des durées de séjour de tous les individus. La variance de cette estimation s'écrit :

$$\text{var}(\hat{m}_{ind}) = \frac{M}{Y^2} = \frac{(\hat{m}_{ind})^2}{M} \qquad [2]$$

Lorsqu'on ne dispose que des observations agrégées sur la période $(0, h)$, on ne connaît plus des durées de séjour individuelles Y_α et seulement le nombre total d'événements observés M. Il est cependant possible d'obtenir une estimation du maximum de vraisemblance du quotient instantané.

En effet, si $S(t)$ est la fonction de séjour, on peut écrire la probabilité pour qu'un individu connaisse l'événement au cours de la période $(0, h)$:

$$\frac{S(0) - S(h)}{S(0)} = 1 - \exp(-mh)$$

La probabilité pour qu'il ne le connaisse pas est égale à :

$$\exp(-mh)$$

La vraisemblance s'écrit alors :

$$L(m) = [1 - \exp(-mh)]^M [\exp(-mh)]^{N-M} \qquad [3]$$

et le logarithme de cette vraisemblance :

$$\log[L(m)] = M \log[1 - \exp(-mh)] - (N - M)mh \qquad [4]$$

L'estimateur du quotient instantané agrégé est alors obtenu en annulant la dérivée de ce logarithme par rapport à m. C'est l'estimateur du maximum de vraisemblance à partir de données agrégées.

$$\frac{\partial \log(L(m))}{\partial m} = \frac{Mh \exp(-\hat{m}_{agr} h)}{1 - \exp(-\hat{m}_{agr} h)} - h(N - M) = 0 \qquad [5]$$

ce qui conduit à la valeur :

$$\hat{m}_{agr} = -\frac{1}{h} \log\left(1 - \frac{M}{N}\right) \qquad [6]$$

De façon semblable on obtient la variance de cette estimation comme l'inverse de l'opposé de la dérivée seconde du logarithme de la vraisemblance :

$$var(\hat{m}_{agr}) = \frac{[1 - \exp(-\hat{m}_{agr} h)]^2}{Mh^2 \exp(-\hat{m}_{agr} h)} = \frac{M}{Nh^2(N - M)} \qquad [7]$$

On peut dès lors comparer l'efficacité des estimations faites à l'aide de données agrégées et individuelles à partir de :

$$Eff = \frac{\text{var}(\hat{m}_{ind})}{\text{var}(\hat{m}_{agr})} = \frac{Nh^2(N-M)}{Y^2} \qquad [8]$$

Pour tester les valeurs prises par *Eff*, nous avons effectué des simulations sur une population de 1 000 individus afin d'estimer cette efficacité en fonction du quotient m et de la durée d'observation h. Les résultats sont portés sur la figure 1, pour une durée d'observation de 1 an et de 5 ans. Comme on peut s'y attendre lorsqu'il s'agit d'un événement très rare, l'efficacité est voisine de 1 dans les deux cas : il n'y a alors aucun problème à utiliser des données agrégées. Celle-ci décroît ensuite très lentement : lorsque la durée d'observation est de un an elle reste proche de l'unité quelle que soit la valeur du quotient. Lorsque la durée d'observation est de cinq ans, la décroissance initialement lente s'accélère pour les valeurs plus élevées du quotient : lorsque le quotient est égal à 0,1, l'efficacité est de 97,9 %, et lorsqu'il est égal à 0,5, ce qui est une valeur très élevée, celle-ci ne tombe qu'à 60,0 %.

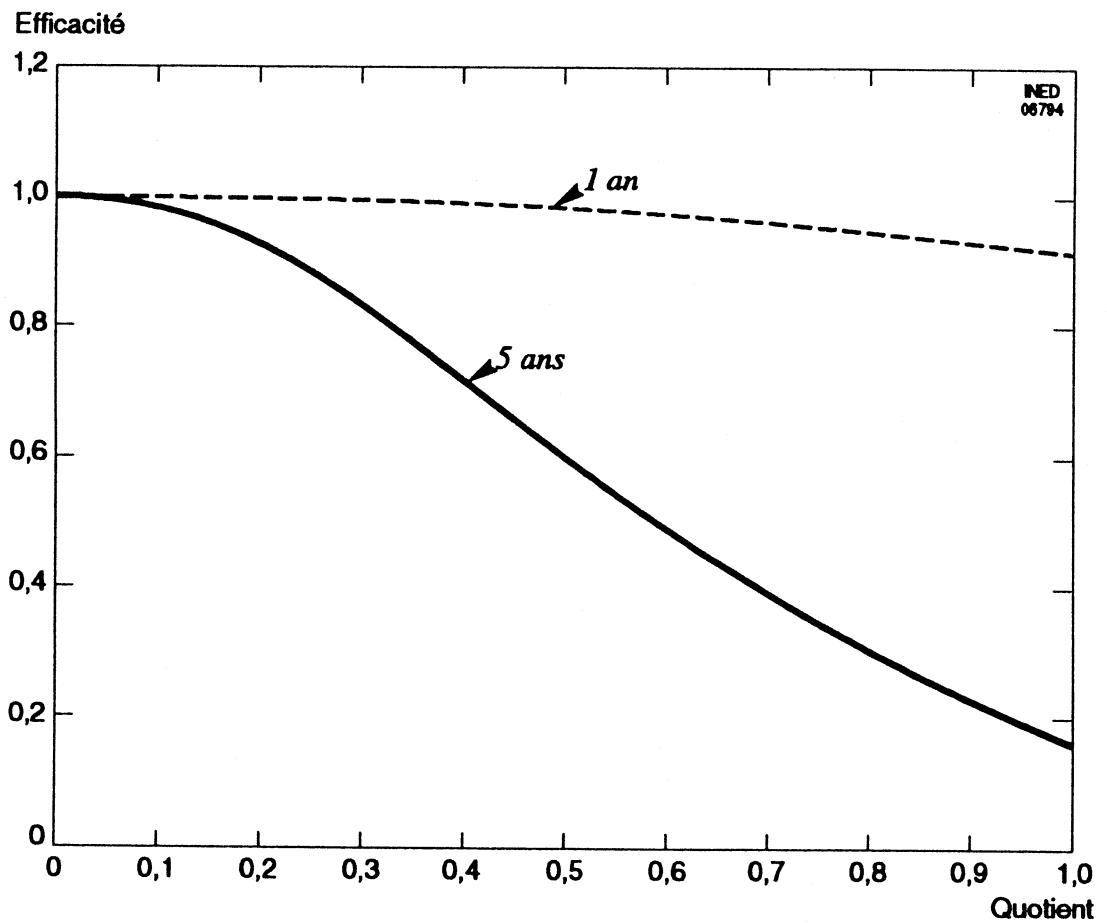

Figure 1. – Efficacité des variances individuelles et agrégées (obs. sur 1 ou 5 ans)

On peut donc conclure que lorsque l'on cherche à estimer un quotient constant à l'aide de données agrégées, l'efficacité de cette estimation par rapport à celle que l'on obtiendrait à partir de données individuelles reste toujours élevée pour les valeurs usuelles des quotients en démographie. Ainsi par exemple, les taux de mor-

talité annuels calculés à partir de données d'état civil ou de données individuelles sont généralement satisfaisants.

Des analyses semblables peuvent être faites dans le cas plus général où l'on a plusieurs événements en interaction : les migrations entre diverses régions d'un pays, les changements d'état matrimonial (célibataire, marié, veuf, divorcé), etc. Cependant les expressions matricielles qui en résultent deviennent si complexes que des conclusions générales, du type de celles données ici, ne peuvent être obtenues (Gill and Keilman, 1990) et seuls des cas plus particuliers peuvent être analysés.

Daniel COURGEAU*, Nico KEILMAN**, Eva LELIÈVRE*

BIBLIOGRAPHIE

COURGEAU D. et LELIÈVRE E. (1989).– *Analyse démographique des biographies*, P.U.F.-I.N.E.D., Paris, 268 p.

COURGEAU D. (1994).– « Du groupe à l'individu : les modèles de migration », dans ce numéro.

GILL R. and KEILMAN N. (1990).- « On the estimation of multidimensional demographic models with population registration data », *Mathematical Population Studies*, 2, 2, pp. 119-143.