OFFICE DE LA RECHERCHE SCIENTIFIQUE ET TECHNIQUE OUTRE - MER
(O.R.S.T.O.M.)

Section de Démographie

Document de travail n° 8

ESTIMATION DE L'EFFET DE GRAPPE SUR LA PRÉCISION DE L'ESTIMATION DES TAUX DÉMOGRAPHIQUES

Application à la Haute-Volta (1960-1961)

Rémy CLAIRIN
INSEE/IDP

juillet 1978

ESTIMATION DE L'EFFET DE GRAPPE SUR LA PRECISION DE L'ESTIMATION DES TAUX DEMOGRAPHIQUES

Application à la Haute-Volta (1960-1961)

I. L'EFFET DE GRAPPE

Bien que ce ne soit pas rigoureusement exact, on peut, en pratique, assimiler les taux bruts de natalité et de mortalité à des fréquences ou à des probabilités.

Si l'unité de tirage de l'échantillon était l'individu, la valeur de la variance serait donnée par la formule :

$$(1) \quad Vo\ (t) = \frac{t\ (1-t)}{P}$$

t étant le taux brut considéré (natalité ou mortalité) et P le nombre de personnes recensées dans l'échantillon.

Exemple : Sur un échantillon de 100.000 personnes, on a un taux brut de natalité de 40 ‰ (t = 0,04), dans l'hypothèse ci-dessous, on aurait :

$$Vo\ (t) = \frac{0,04 \times 0,96}{100.000} = 0,384 \times 10^{-6}$$

l'écart type est : $\sigma_t = \sqrt{Vo\ (t)} = 0,62 \times 10^{-3}$

l'écart type relatif : $\frac{\sigma_t}{t} = \frac{0,62 \times 10^{-3}}{0,04} = 1,55\ \%$

et l'intervalle de confiance à 95 % qui est approximativement défini par plus ou moins deux écarts types relatifs serait donc de $\pm\ 3,1\ \%$

si les individus avait été tirés isolément.

Mais en pratique, on tire non pas des individus isolés, mais des unités de sondage telles que des villages, des groupes de villages ou de hameaux, des quartiers, secteurs ou ilôts d'agglomérations, etc... qui constituent ce que l'on appelle des grappes d'individus.

Il n'y a évidemment pas indépendance au sens statistique du terme entre les caractéristiques des habitants ou des ménages d'une même grappe et on ne peut donc pas appliquer sans correction la formule (1).

Nous avons appelé Vo (t) la variance du taux dans l'hypothèse du tirage direct des individus, nous désignerons par V_G (t) la variance effective, compte tenu du fait que l'échantillonnage a porté sur des villages ou unités analogues. La différence entre Vo (t) et V_G (t) est due à l'effet de grappe.

La formule (1) montre que le calcul de Vo (t) est des plus rapide et qu'il fait intervenir uniquement les chiffres globaux. Au contraire, le calcul de V_G (t) est assez compliqué et il implique que l'on dispose des résultats détaillés par unité de sondage.

Mais il existe une relation simple entre les valeurs de V_G et Vo : (2) $V_G\ (t) = Vo\ (t)\ \left[1 + (\bar{m} - 1)\ \delta\right]$, $\bar{m}$ est la taille moyenne des unités de sondage.

δ s'appelle le coefficient de corrélation intragrappe. il traduit le degré de corrélation (ou d'homogénéité) entre les caractéristiques

des unités statistiques appartenant à une même grappe (par exemple, les habitants d'un village).

Il arrive que soit négatif et dans ce cas, l'effet de grappe a pour conséquence une diminution de la variance. Mais c'est l'inverse que l'on observe généralement.

Les valeurs extrêmes que peut prendre sont :

$$-\frac{1}{\bar{m}-1} \quad \text{et} + 1$$

De (1) et de (2) on déduit :

$$(3)\ V_G(t) = \frac{t(1-t)}{P}\ 1 + (\bar{m} - 1)$$

P population recensée dans l'échantillon est égal à $n.\bar{m}$ et on a donc :

$$V_G(t) = \frac{t(1-t)}{n\,\bar{m}}\ 1 + (\bar{m} - 1)$$

L'effectif global à recenser dans l'échantillon ou le nombre d'unités de sondage à y inclure pour obtenir une certaine variance effective $V_G(t)$ sont données par :

$$(4)\quad P = \frac{t(1-t)\ 1+(\bar{m}-1)}{V_G(t)} \qquad n = \frac{t(1-t)\ 1+(\bar{m}-1)}{\bar{m}\ V_G(t)}$$

En pratique, on a toujours au moins une idée de l'ordre de grandeur du taux t.

En ce qui concerne la valeur de $\bar{m}$, on a en général une assez grande latitude pour la fixer, bien qu'il puisse y avoir des contraintes pratiques. On verra plus loin des critères qui interviennent pour déterminer cette valeur et les chiffres auxquels on arrive généralement dans les enquêtes démographiques.

Reste la valeur de . Celle-ci ne peut être estimée qu'a posteriori à partir des résultats d'une enquête -tout au moins lorsqu'on étudie une population pour la première fois. Malheureusement, lorsqu'on cherche à analyser les résultats, on constate trop souvent que les données permettant de faire cette estimation ont disparu corps et biens, la seule solution consiste à s'inspirer des chiffres calculés dans des populations analogues.

En ce qui concerne l'Afrique Noire, un certain nombre d'estimations de la valeur de ont été faites et comparées par Christopher SCOTT (*).

Il en conclut que l'on peut retenir comme ordre de grandeur "plausible" de :

0.002 pour le taux brut de natalité
0.003 pour le taux brut de mortalité.

Bien entendu, ce ne sont que des moyennes ou des médianes, et, compte tenu des conditions très variables d'une région à l'autre, les valeurs

(*) Christopher SCOTT. L'enquête in Sources et analyse des données démographiques. 1ère partie, Paris, INED, INSEE, ORSTOM, SEAE 1973 (page 110).

effectuées risquent d'être assez dispersées. D'autre part, bien qu'en théorie, la valeur de δ soit indépendante de celle de $\bar{m}$ (effectif moyen de la grappe), il semble bien qu'en pratique, différents facteurs font qu'il y a une certaine corrélation entre $\bar{m}$ et les valeurs observées de δ .

Prenons un exemple : On veut organiser une enquête par sondage permettant de connaître le taux brut de mortalité avec une précision (intervalle de confiance à 95 %) de ± 4 %. Quel doit être approximativement l'effectif de l'échantillon à étudier, sachant que l'ordre de grandeur vraisemblable du taux est de 15 ‰ (0,015) ?

Comme l'intervalle de confiance est à peu près égal à deux écarts-types relatifs $\frac{\sigma_t}{t}$, on voit que $\frac{\sigma_t}{t} = 2\ \% = 0{,}02$

C'est à dire :

$$\sigma_t = t \times 0{,}02 = 0{,}015 \times 0{,}02 = 0{,}0003 = 0{,}3 \times 10^{-3}$$

donc :

$$V_G(t) = \sigma_t^2 = 0{,}09 \times 10^{-6}$$

et P est donné par :

$$(5)\quad P = \frac{0{,}015 \times 0{,}985 \left[1 + (\bar{m} - 1)\,\delta\right]}{0{,}09}$$

$$(6)\quad P = 164167 \left[1 + (\bar{m} - 1)\,\delta\right]$$

Ainsi l'ordre de grandeur à prévoir pour l'échantillon est fonction de $\bar{m}$ et δ .

Tableau 1. Effectif de l'échantillon qui donnera une précision de ± 4 % pour un taux brut de mortalité de l'ordre de 15 ‰ en fonction de $\bar{m}$ (taille moyenne de l'unité de sondage) et de δ coefficient de corrélation intragrappe.

$\bar{m}$ \ δ	0.002	0.003	0.004
100	197 000	213 000	230 000
200	230 000	263 000	296 000
300	263 000	312 000	361 000
400	296 000	361 000	427 000
500	328 000	410 000	493 000
1 000	493 000	657 000	821 000
2 000	821 000	1 149 000	1 478 000

δ est supposé donné, on ne peut donc pas changer sa valeur. Par contre, tout au moins en principe, on est libre de se fixer à l'avance la taille de la grappe (unité de tirage).

Voyons ce que cela donnerait en adoptant la valeur moyenne de (0.003).

Tableau 2. Composition de l'échantillon qui donnerait un intervalle de confiance de ± 4 % pour un taux brut de mortalité de l'ordre de 15 ‰ avec un coefficient de corrélation intragrappe égal à 0.003.

Taille moyenne de l'unité de sondage	Nombre d'unités à étudier	Population recensée
$\bar{m}$	n	$n \quad \bar{m} = P$
100	2 130	213 000
200	1 315	263 000
300	1 040	312 000
400	903	361 000
500	820	410 000
1 000	657	657 000
2 000	575	1 149 000

Le choix final peut dépendre de diverses considérations : dispersion de l'habitat, personnel disponible, moyens de transport, budget.

Supposons que la contrainte principale soit le nombre de journées de travail/enquêteur. Le temps de travail peut se décomposer en deux éléments :

a) la durée nécessaire à la mise en route du travail d'interrogations des ménages dans chaque unité : déplacement d'une unité à l'autre, installation des équipes, présentation de l'enquête à la population, reconnaissance des lieux, établissement d'un plan de la localité, délimitation des ilôts, etc...

Cette durée est à peu près proportionnelle au nombre d'unités, n. Nous désignerons par :

T_1 : le nombre de journées de travail nécessaires à la mise en route de l'interrogation proprement dit sur une unité.

b) le temps passé à l'interrogatoire proprement dit. Ce temps est sensiblement proportionnel au nombre de personnes recensées, $n.\bar{m}$

Donc, si l'on désigne par T_T le nombre total de journées de travail, on a :

$$(7) \qquad T_T = T_1 \times n + T_2 \times n.\bar{m}$$

Supposons que T_1 soit égal à 5 jours et que, lorsqu'il ne fait que cela, un enquêteur interroge 50 personnes par journ ce qui donne :

$$T_2 = \frac{1}{50} = 0,02$$

L'équation (7) devient :

(8) $T_T = 5 \times n + 0{,}02 \times n.\bar{m}$

Ce qui donne :

Tableau 3. Nombre d'unités et nombre de journées/enquêteur nécessaires pour obtenir un intervalle de confiance de ± 4 % pour un taux brut de mortalité de 15 ‰ avec un coefficient de corrélation intragrappe égal à 0,003.

Taille moyenne de l'unité de sondage $\bar{m}$	Nombre d'unités à étudier n	Population recensée $n.\bar{m}$	Nombre total de journées/enquêteur
100	2 130	213 000	14 910
200	1 315	263 000	11 835
300	1 040	312 000	11 440
400	903	361 000	11 735
500	820	410 000	12 300
1 000	657	657 000	16 425
2 000	575	1 149 000	25 855

On constate que du point de vue de la durée du travail, c'est la grappe de 300 personnes qui est la plus économique.

La formule qui donne exactement l'optimum de taille est la suivante :

(a) $\bar{m}\ \text{opt.} = \sqrt{\frac{1-\delta}{\delta} \cdot \frac{T1}{T_2}}$

soit, avec les chiffres ci-dessous

(10) $\bar{m}\ \text{opt.} = \sqrt{\frac{0{,}997}{0{,}003} \cdot \frac{5}{0{,}02}} = 288$

Cet ordre de grandeur de l'optimum de taille de grappe (disons entre 200 et 500) est celui que l'on a observé en général lorsqu'on a pu faire les calculs.

II. APPLICATION A LA HAUTE-VOLTA

En Haute-Volta, on a calculé la précision (intervalle de confiance) des taux bruts de natalité et de mortalité à partir des résultats par village, c'est-à-dire en tenant compte de l'effet de grappe.

Malheureusement, les chiffres de base et les feuilles de calcul ne sont plus disponibles. Ils auraient notamment permis l'estimation du coefficient de corrélation intragrappe par strate, ce qui aurait présenté un grand intérêt. Ceci illustre une fois de plus le regrettable gaspillage d'information qui s'est trop souvent produit à l'occasion de dépouillement d'enquêtes démographiques.

Globalement, les chiffres sont les suivants :

- fraction de sondage : 1/49
- effectif total recensé : 88 317
- nombre d'unités de sondage (villages ou parties de villages) : 235
- taux brut de natalité : 49,6 ‰
- taux brut de mortalité : 31,7 ‰

D'après le rapport d'enquête, les intervalles de confiance à 95 % sont : . ± 4,2 % pour le taux brut de natalité
. ± 6,6 % pour le taux brut de mortalité

Nous allons estimer ce que seraient ces intervalles en l'absence d'effet de grappe.

La variance serait alors donnée par la formule

$$(11)\quad V_o(t) = (1 - f)\, t\, \frac{(1 - t)}{P}$$

f est la fraction de sondage, le facteur $(1 - f)$ intervient parce que les unités ont été tirées de façon exhaustive (c'est-à-dire qu'une même unité en peut être tirée plus d'une fois), ce qui donne :

Taux brut de natalité, b

$$V_o(b) = \left(1 - \frac{1}{49}\right) \frac{0{,}0496 \times 0{,}9504}{88317} = 0{,}523 \times 10^{-6}$$

$$\sigma_{o\,b} = \sqrt{V_o(b)} = 0{,}723 \times 10^{-3}$$

$$\frac{\sigma_{\partial\, b}}{b} = 1{,}46\ \%$$

Intervalle de confiance à 95 % = ± 2,9 %

Taux brut de mortalité, m

$$V_o(m) = \left(1 - \frac{1}{49}\right) \frac{0{,}0317 \times 0{,}9683}{88317} = 0{,}340 \times 10^{-6}$$

$$\sigma_{o\,m} : \sqrt{V_o(m)} = 0{,}583 \times 10^{-3}$$

$$\frac{\sigma_{\partial m}}{m} = 1{,}84\ \%$$

Intervalle de confiance à 95 % = ± 3,6 %

Valeurs du coefficient de corrélation intragrappe

Le rapport entre les intervalles de confiance est égal à la racine carrée du rapport entre les variances

$$V_G/V_o = 1 + (\bar{m} - 1)\,\delta \qquad \delta = \frac{V_G}{V_o} - 1 \,/\, \bar{m} - 1$$

$$\bar{m} = \frac{88317}{235} = 375{,}8$$

Taux brut de natalité

Rapport entre les intervalles de confiance : 4,2/2,9 = 1,45

Rapport entre les variances : 2,10

$$\delta = \frac{2,10 - 1}{374,8} = 0,003$$

Taux brut de mortalité

Rapport entre les intervalles de confiance : 6,6/3,6 = 1,83

Rapport entre les variances : 3,36

$$\delta = \frac{3,36 - 1}{374,8} = 0,006$$

Ces chiffres sont voisins des ordres de grandeur indiqués par Christopher SCOTT.