Cahiers québécois de démographie

Application des tests statistiques à des populations exhaustives : l'exemple de la mortalité

Norbert Robitaille

Volume 9, numéro 1, avril 1980

URI : https://id.erudit.org/iderudit/600811ar
DOI : https://doi.org/10.7202/600811ar

Aller au sommaire du numéro

Éditeur(s)

Association des démographes du Québec

ISSN

0380-1721 (imprimé)
1705-1495 (numérique)

Découvrir la revue

Citer cet article

Robitaille, N. (1980). Application des tests statistiques à des populations exhaustives : l'exemple de la mortalité. *Cahiers québécois de démographie*, *9*(1), 105–116. https://doi.org/10.7202/600811ar

Résumé de l'article

Le présent article est une réflexion concernant la pertinence d'appliquer des tests statistiques à des mesures fondées sur des populations exhaustives. Il appuie cette réflexion sur l'ensemble de Sainte-Lucie, petite île des Antilles pour laquelle on a observé de 1971 à 1975 la série de décès et de taux de mortalité suivante : 1971 : 796 (7,75%), 1972 : 947 (9,12%), 1973 : 840 (8,01%), 1974 : 829 (7,82%) et 1975 : 819 (7,65%) décès. La question posée par cet article est de savoir si les fluctuations observées dans les décès et les taux sont le fait du hasard.

Cahiers québécois de démographie
Vol. 9, no 1, avril 1980

Norbert ROBITAILLE*: APPLICATION DES TESTS STATISTIQUES À DES POPULATIONS EXHAUSTIVES: L'EXEMPLE DE LA MORTALITÉ

RÉSUMÉ

Le présent article est une réflexion concernant la pertinence d'appliquer des tests statistiques à des mesures fondées sur des populations exhaustives. Il appuie cette réflexion sur l'ensemble de Sainte-Lucie, petite île des Antilles pour laquelle on a observé de 1971 à 1975 la série de décès et de taux de mortalité suivante: 1971: 796 (7,75‰), 1972: 947 (9,12‰), 1973: 840 (8,01‰), 1974: 829 (7,82‰) et 1975: 819 (7,65‰) décès. La question posée par cet article est de savoir si les fluctuations observées dans les décès et les taux sont le fait du hasard.

* Département de démographie, Université de Montréal, C.P. 6128, Succursale A, Montréal H3C 3J7.

Cahiers québécois de démographie
Vol. 9, no 1, avril 1980

APPLICATION DES TESTS STATISTIQUES À DES POPULATIONS EXHAUSTIVES : L'EXEMPLE DE LA MORTALITÉ

Par Norbert ROBITAILLE*

La réflexion qui va suivre tire son origine de la définition du concept de mortalité. Le problème qui se pose est de savoir si la mortalité, phénomène démographique, est le risque inhérent à une situation ou à la réalisation particulière de celui-ci.

Lorsqu'on travaille sur des populations importantes, le problème s'estompe car on peut considérer que la réalisation particulière, nombre de décès, taux de mortalité, espérance de vie, que l'on peut mesurer à un moment donné, est le reflet du risque auquel la population était soumise. Cependant, ce problème, un peu théorique lorsqu'on considère une population importante, devient très concret quand on s'intéresse à une petite population. En effet, si le taux annuel de mortalité passe de

* Département de démographie, Université de Montréal, C.P. 6128, Succursale A, Montréal H3C 3J7.

1% à 2% dans une population de 100 personnes, peut-on dire que la mortalité a doublé? Si on considère que le phénomène mortalité est la réalisation elle-même, on est bien obligé de répondre oui. Il se peut cependant que les conditions de vie se soient, dans cette population, grandement améliorées et que le doublement des décès soit purement accidentel. Il serait alors un peu gênant de dire que la mortalité a doublé, alors que si une population importante avait été soumise au même risque durant les mêmes deux périodes, il y aurait eu une diminution de la mortalité.

Dans ce cas, il serait préférable d'adopter l'autre perspective qui considère l'indice de mortalité enregistré en un lieu et à un moment comme une réalisation du phénomène lui-même. Dans cette perspective, le phénomène serait le risque qui ne serait qu'estimé par sa réalisation qui, elle seule, peut être mesurée.

Réfléchissons à partir d'un exemple concret. Si, par exemple, dans une population d'environ 100 000 personnes[1] on observe, pour cinq années successives, 1971: 796 (7,75‰), 1972: 947 (9,12‰), 1973: 840 (8,01‰), 1974: 829 (7,82‰) et 1975: 819 (7,65‰) décès[2], est-il raisonnable de dire que la mortalité est demeurée constante et que les variations dans les décès sont dues à des phénomènes aléatoires?

---

(1) Cet exemple est tiré de: FLOR-LACHAPELLE, Lourdes: Analyse démographique de l'île de Sainte-Lucie: 1946-1970, Département de démographie, Université de Montréal, Mémoire de maîtrise 1977, non publié.

(2) Les populations utilisées pour le calcul des taux de mortalité ont été estimées à partir des résultats des projections faites dans le travail cité ci-dessus (p. 162).

Population observée au milieu de 1970: R^{1970} = 101 590

Population prévue au milieu de 1975 : R^{1975} = 107 110

$R^{1971} = R^{1970} + 1/5\,(R^{1975} - R^{1970})$ = 102 694

$R^{1972} = R^{1970} + 2/5\,(R^{1975} - R^{1970})$ = 103 798

R^{1973} $\vdots$ $\vdots$ = 104 902

R^{1974} $\vdots$ $\vdots$ = 106 006

R^{1975} $\vdots$ $\vdots$ = 107 110

I- La réponse à la question qui précède est non si on considère la mesure, donc la réalisation particulière, comme étant la mortalité. Cependant l'autre façon de considérer la mortalité, comme estimée par ses réalisations, permet de supposer que nous avons affaire à des échantillons d'environ 100 000 personnes parmi l'infinité de personnes que l'on pourrait théoriquement soumettre à cette mortalité. En considérant donc les personnes exposées chaque année comme des échantillons, on pourrait procéder de la façon suivante:

Soit un échantillon E_1 de N_1 individus pour la période 1971-1975 où l'on exclut 1972:

$$N_1 = \sum_{t=1971}^{1975} \left[R^t + \frac{D^t}{2}\right] - R^{1972} - \frac{D^{1972}}{2} = 422\ 354$$

$$\text{dont } D^{1971} + D^{1973} + D^{1974} + D^{1975} = 3\ 284 \text{ sont décédés}$$

R^t = population estimée au milieu de l'année t

D^t = décès, durant l'année t, d'individus de la population R^t

Soit un échantillon E_2 composé de N_2 individus de l'année 1972:

$$N_2 = R^{1972} + \frac{D^{1972}}{2} = 104\ 272$$

dont D^{1972} = 947 sont décédés

Dans E_1, la proportion des décédés est de $\frac{3\ 284}{422\ 354} = 7{,}775‰ = P_1$. Dans E_2, la proportion est de $9{,}082‰ = P_2$.

Considérons l'hypothèse nulle selon laquelle il n'y a pas de différence significative entre ces deux proportions ($H_0 : P_1 = P_2$), qui suppose donc la mortalité en 1972 identique à celle de 1971-1975.

On aura alors:

$$\mu_{P_1-P_2} = 0$$

$$\text{et } \sigma_{P_1-P_2} = \sqrt{p.q.(1/N_1 + 1/N_2)}$$

où p. est la proportion réelle des décédés dans la population dont est tiré l'échantillon E_1 (p. est donc la mortalité vraie dont P. est une estimation) et q. = 1 - p.

L'estimation de p. et de q. est faite de la façon suivante:

$$\hat{p}. = P. = \frac{\sum D}{\sum N} = \frac{3\,284 + 947}{422\,354 + 104\,272} = 8{,}034‰$$

$$\text{et } \hat{q}. = Q. = 1 - P. = 991{,}966‰.$$

En remplaçant p. et q. par leur estimation P. et Q. on aura:

$$\begin{aligned}\sigma_{P_1-P_2} &= \sqrt{(0{,}008034)(0{,}991966)(1/422\,354 - 1/104\,272)} \\ &= \sqrt{(0{,}008034)(0{,}991966)(0{,}000011958)} \\ &= 0{,}0003087\end{aligned}$$

On se poserait alors la question de savoir si l'on doit rejeter H_0. Pour ce, on calculerait:

$$\begin{aligned}Z &= \frac{(P_1 - P_2)}{\sigma_{P_1-P_2}} \\ &= \frac{(0{,}007775 - 0{,}009082)}{0{,}0003087} \\ &= -4{,}2339\end{aligned}$$

Comme $Z = -4{,}2339 < -1{,}96$, seuil d'un test à 0,05 de risque d'erreur, on doit rejeter l'hypothèse nulle et conclure qu'il ne semblerait pas raisonnable d'affirmer que les deux populations, dont E_1 et E_2 sont tirés, sont identiques. On devrait rejeter l'hypothèse nulle qui supposait la mortalité en 1972 (p_2 est estimé par P_2) identique à celle de 1971-1975 (sauf 1972) (p_1 est estimé par P_1).

II- Un autre modèle pourrait considérer que dans une population i, chaque individu est soumis annuellement à une épreuve dont la probabilité de "succès" est le taux de mortalité (p). La distribution de probabilité du nombre de décès observés dans cette population sera alors une binomiale de moyenne $N_i p_i$ et de variance $N_i p_i\ (1-p_i)$. Pour une population assez importante, cette distribution peut être approchée par une distribution normale de mêmes moyenne et variance.

Appliquons ce second modèle aux données précédemment présentées. Calculons d'abord un taux de mortalité sur la base de l'ensemble des résultats enregistrés.

$$p = \frac{\sum_{t=1971}^{1975} D^t}{\sum_{t=1971}^{1975} N^t}$$

$$= \frac{4\ 231}{526\ 635}$$

$$= 0{,}008034$$

$$1 - p = q = 1 - 0{,}008034$$

$$= 0{,}991966$$

Vérifions maintenant si dans notre exemple, les 947 décès enregistrés en 1972 pour une population de 104 272 peuvent être le résultat de 104 272 épreuves dont la probabilité de succès est de 0,008034.

La moyenne de la distribution des résultats provenant de ces épreuves serait:

$$\mu = Np$$
$$= 104\ 272\ (0{,}008034)$$
$$= 838$$

et
$$\sigma = \sqrt{Np(1-p)}$$
$$= \sqrt{838\ (1 - 0{,}008034)}$$
$$= \sqrt{831{,}27}$$
$$= 28{,}83$$

Or, d'après les données présentées au début de ce travail, on trouve 947 décès pour 1972.

L'écart à la moyenne attendue est de 109 (947 - 838), ce qui, divisé par l'écart-type est estimé comme: $\frac{109}{28{,}83} = 3{,}78$. Un tel écart, si la mortalité répondait au modèle ci-haut présenté ne se produirait que dans moins d'un cas sur 1000, ce qui porte à croire que la probabilité de décéder était différente pour l'année 1972, ceci, évidemment, si les hypothèses du modèle sont exactes.

III- Mais nous savons pertinemment qu'une des hypothèses du modèle précédent est inexacte. Tous les individus n'ont pas une même probabilité de décéder, qui serait égale au quotient de mortalité théorique. En fait, comme le montrerait une table de mortalité, cette probabilité varie suivant l'âge de façon très importante. Le problème se pose de savoir si le fait d'avoir une mortalité hétérogène risque de rendre les variations aléatoires plus importantes dans le cas où la mortalité demeurerait constante. En d'autres mots, si au lieu d'appliquer à chaque individu de la population, quel que soit son âge, une même mortalité, on appliquait la mortalité réelle correspondant à son âge, l'écart-type de la distribution des résultats que l'on obtiendrait serait-il supérieur à celui obtenu dans le cas précédent?

Nous allons montrer que le fait d'appliquer à une population une seule probabilité de décéder augmente la dispersion des résultats due à des variations aléatoires. Si donc un test conclut au rejet de l'hypothèse nulle, en utilisant une hypothèse unique de mortalité, comme nous l'avons fait en II, la preuve sera faite que le test serait encore plus concluant sur la base d'une mortalité hétérogène.

D'une façon plus statistique on peut se demander si la variance de la distribution de la somme de variables binomiales est supérieure, inférieure ou égale à la variance de la distribution binomiale ayant pour moyenne la somme des moyennes des distributions des variables susmentionnées.

De façon plus formelle on peut écrire:

Soit trois variables $X.$, X_1, X_2, de distribution binomiale ayant respectivement comme paramètres $n.p.q.$, $n_1p_1q_1$ et $n_2p_2q_2$

si $\quad n.p. = n_1p_1 + n_2p_2 \quad (1)$

et $\quad n. = n_1 + n_2 \quad (2)$

peut-on dire que:

$$\mathrm{Var}\,(X.) > \mathrm{Var}\,(X_1 + X_2) \quad \text{(Question a, 1}^{\text{ère}}\text{ forme)}$$

Du théorème d'addition pour la distribution normale, on peut déduire, en supposant que les distributions binomiales peuvent s'assimiler à des distributions normales que la variance de la distribution de la variable constituée par la somme des variables X_1 et X_2 est égale à la somme des variances de X_1 et X_2, donc que Var $(X_1 + X_2)$ = Var (X_1) + Var (X_2) + Cov $(X_1 + X_2)$.

Supposons Cov $(X_1, X_2) = 0$, les deux variables X_1 et X_2 étant indépendantes, et la question a devient:

$$n.p.q. > n_1p_1q_1 + n_2p_2q_2 \quad \text{(Question a, 2}^{\text{ème}}\text{ forme)} \quad (3)$$

Evaluons le côté droit de l'inégalité (3) en fonction de n.p.q. afin de vérifier cette inégalité.

Sachant que q. = 1-p., on peut écrire d'après (3):

$$n_1p_1q_1 + n_2p_2q_2 = n_1p_1\ (1-p_1) + n_2p_2\ (1-p_2)$$

Remplaçons n_2 par sa valeur en (2)

$$n_1p_1q_1 + n_2p_2q_2 = n_1p_1\ (1-p_1) + (n. - n_1)\ p_2\ (1-p_2)$$

Remplaçons p_2 par sa valeur en (1), $p_2 = \dfrac{n.p. - n_1p_1}{n. - n_1}$

$$n_1p_1q_1 + n_2p_2q_2 = n_1p_1\ (1-p_1) + (n.-n_1)\left(\frac{n.p.-n_1p_1}{n.-n_1}\right)\left[1-\left(\frac{n.p.-n_1p_1}{n.-n_1}\right)\right] \quad (4)$$

Posons $\quad p_1 = p.k \qquad (4.1)$

$\qquad\qquad n_1 = n.m \qquad (4.2)$

et remplaçons p_1 et n_1 par leur valeur dans la partie droite de l'équation (4)

$$n_1p_1q_1 + n_2p_2q_2 = n.mp.k(1-p.k)+(n.-nm)\left(\frac{n.p.-n.mp.k}{n.-n.m}\right)\left[1-\left(\frac{n.p.-n.mp.k}{n.-n.m}\right)\right]$$

$$n_1p_1q_1 + n_2p_2q_2 = n.p.\left[mk(1-p.k)+(1-mk)\left(\frac{1-m-p.+p.m.k}{1-m}\right)\right] \quad (5)$$

$$= n.p.\left(\frac{mk-mpk^2-m^2k+m^2k^2p.+1-m-p.+p.mk-mk+m^2k+mkp.-p.m^2k^2}{1-m}\right)$$

$$= n.p.\frac{(-mp.k^2+1-m-p+2p.mk)}{(1-m)}$$

$$= n.p.\left[\left(\frac{1-m}{1-m}\right) -p.\left(\frac{mk^2+1-2mk}{1-m}\right)\right]$$

$$= n.p.\left(1-p.\frac{(mk^2+1-2mk)}{1-m}\right)$$

$$= n.p.\left[1-p.\left(\frac{1-m(-k^2+2k)}{1-m}\right)\right] \quad (6)$$

Or, d'après (1) et (2) p. est une moyenne pondérée de p_1 et p_2 de telle sorte que p_1 peut être posé comme inférieur à p. de telle sorte que $0 < k = \frac{p_1}{p.} < 1$.

De ceci, il découle que dans (6), $2k-k^2 < 1$, et donc que

$$\alpha = \frac{1-m(-k^2+2k)}{1-m} > 1$$

Et si on replace le tout dans l'équation (3) en remplaçant la partie de droite par sa valeur en (6) on a

$$n.p.(1-p.) > n.p.\left[1-p(\alpha)\right] \qquad \alpha > 1 \quad \text{(Question a, 3}^{\text{ème}}\text{ forme)}$$

Présentée de cette façon, l'inégalité est évidente et on peut affirmer que

$$\boxed{n.p.q. > n_1p_1q_1 + n_2p_2q_2}$$

On pourrait généraliser cette preuve et montrer que si:

$$n.p. = n_1p_1 + n_2p_2 + \ldots + n_ip_i$$

et $$n. = n_1 + n_2 + \ldots + n_i$$

alors $$\boxed{n.p.q. > n_1p_1q_1 + n_2p_2q_2 + \ldots + n_ip_iq_i}$$

Ce qui précède montre que si l'on tient compte de l'hétérogénéité de la population quant au risque de décéder, les variations aléatoires seront moindres dans cette population que celles que l'on trouverait dans une population à mortalité homogène ayant même taux de mortalité.

IV- Pour conclure, rappelons que les deux types de modèles présentés d'une part en I et d'autre part en II et III sont évidemment des abstractions qui ne pourront être testées. Dans le premier cas, la population théorique très nombreuse dont les populations réelles plus réduites de diverses années seraient des échantillons, n'existe pas. Dans le second cas, le modèle qui présente chaque individu soumis à une épreuve dont la probabilité de succès est p ne peut lui non plus être vérifié. En effet, on ne saura jamais avec certitude si les variations des décès d'année en année sont le fait de variations du p ou de variations aléatoires. Un point semble cependant certain, comme nous l'avons montré en III, c'est que si l'on tient compte de l'hétérogénéité de la mortalité, suivant le modèle binomial, la dispersion des décès diminuerait.

Cependant, même si les modèles ne peuvent être vérifiés avec certitude, il nous semble que ceux-ci sont plausibles, surtout celui présenté en III, et qu'ils offrent un point de comparaison. Face à une modification du taux de mortalité, ils aident tout de même à répondre à la question: si la mortalité était conforme à tel ou tel modèle plausible, serait-il raisonnable de considérer qu'il n'y a pas eu de variation de celle-ci, à la lumière des fluctuations observées du taux de mortalité?

Novembre 1979