REVUE DE STATISTIQUE APPLIQUÉE

D. SCHWARTZ
PH. LAZAR

Taux de mortalité par une cause donnée de décès en tenant compte des autres causes de décès ou de disparition

Revue de statistique appliquée, tome 12, n° 3 (1964), p. 15-28
<http://www.numdam.org/item?id=RSA_1964__12_3_15_0>

TAUX DE MORTALITÉ PAR UNE CAUSE DONNÉE DE DÉCÈS EN TENANT COMPTE DES AUTRES CAUSES DE DÉCÈS OU DE DISPARITION (*)

par D. SCHWARTZ et Ph LAZAR
Unité de Recherches Statistiques de l'Institut National d'Hygiène
(Institut Gustave Roussy)

INTRODUCTION

Objet de l'étude

La mortalité que peut entrainer une cause donnée de décès est troublée dans son expression par les autres causes de disparition. Par exemple :

- en irradiant des souris, on a observé un nombre de leucémies décroissant en fonction de la dose ; c'est que l'irradiation, tuant directement une proportion élevée des animaux, les soustrayait au risque leucémie [6] ;

- la diminution de mortalité par maladies infectieuses chez les personnes âgées fait apparaître un nombre plus élevé de cancers, qui peut faire croire à une augmentation propre de ce danger ;

- l'évaluation de la mortalité totale d'un groupe de malades par la méthode actuarielle [2] nécessite le calcul des quotients de mortalité pour les intervalles annuels successifs ; toute une catégorie de sujets, encore vivants, mais de recul insuffisant, "disparaissent" au cours de chaque intervalle annuel par manque de recul.

Dans chacun de ces cas, le besoin se fait sentir de calculer les taux "purifiés" qu'on aurait obtenus si les causes de décès ou de disparition "parasites" n'avaient pas existé.

On doit d'emblée distinguer deux cas :

a) *La cause du décès n'est pas connue* ; dans ces circonstances où même le taux de mortalité *brut* pour une cause donnée ne peut être évalué, il n'est pas question en général d'établir un taux corrigé. C'est cependant possible dans un cas particulier important : quand on étudie un groupe de sujets atteints d'une maladie déterminée, si l'on dispose de tables de mortalité pour la population générale, et si la surmortalité du groupe étudié peut valablement être imputée à la maladie en cause, on peut calculer la mortalité qui lui est due en propre [2]. Nous n'examinerons pas cette éventualité.

b) *La cause de décès est connue pour chaque sujet* (éventualité fréquente surtout chez les animaux, grâce à des autopsies systèmatiques). C'est ce cas seulement que nous étudierons. La littérature propose di-

---

(*) Revue Inst. Int. de Stat. 29 : 3 (1961).

verses expressions du "taux corrigé", qui souvent diffèrent formellement les unes des autres. Notre objet est de confronter les formules les plus employées, afin de permettre un choix rationnel. Cette confrontation conduit à une formulation très générale de la solution du problème.

Hypothèse de base

Soit 1 la cause étudiée, 2 l'ensemble des autres causes de décès ou de disparition. *On devra supposer essentiellement que ces deux causes agissent indépendamment,* l'énoncé rigoureux de cette hypothèse étant le suivant : dans un modèle où l'atteinte par 1 ou 2 n'entraînerait pas la disparition du sujet et le laisserait exposé à l'autre cause, ces deux atteintes seraient des événements statistiquement indépendants.

Cette hypothèse est tout-à-fait raisonnable si la cause 2 paraît sans rapport avec la cause 1 (disparition par recul insuffisant par exemple) ; elle mérite d'être discutée s'il s'agit de deux maladies différentes, et surtout de deux formes voisines d'une même maladie.

Plan de l'étude

A. Formules classiques.
B. Présentation d'une formule générale.
C. Confrontation des différentes formules.
D. Etude plus complète dans le cas où la répartition des décès provoqués par la cause 1 agissant seule serait uniforme.
E. Un exemple.
F. Conclusion, Formulaire.
Bibliographie.

Notations principales

Toute l'étude sera établie pour *l'intervalle de temps (0,1),* ce qui diminue pas la généralité.

Nous emploierons indifféremment les termes "disparition" ou "décès". On appellera 1 la cause étudiée, 2 la cause (ou l'ensemble des causes) compétitive.

Les symboles suivants seront utilisés :

Données observées

N = Nombre de sujets au début de l'intervalle
D_1 = " " " disparaissant par effet de la cause 1
D_2 = " " " " " " 2
S = " " " en fin d'intervalle
$N = D_1 + D_2 + S$
t_i = date de disparition du ième sujet frappé par 1
t_j = " " " jème " " 2
T_1 = date moyenne de disparition par 1 $T_1 = \frac{1}{D_1} \sum_i t_i$
T_2 = " " " " 2 $T_2 = \frac{1}{D_2} \sum_j t_j$

Taux

Q_1 = Taux brut de disparition par 1. Estimation $\hat{Q}_1 = \frac{D_1}{N}$

Q_2 = Taux brut de disparition par 2. Estimation $\hat{Q}_2 = \frac{D_2}{N}$

$Q = Q_1 + Q_2$

q_1 = taux de disparition par la cause 1 en supposant qu'elle agisse seule

q_2 = " " " 2 " " "

Le problème est de trouver *une estimation* $\hat{q}_1$ de q_1 et si possible *la variance de cette estimation.*

A - FORMULES CLASSIQUES

Depuis *Neyman* [7] différentes formules ont été proposées, dont nous indiquons les plus courantes, sous le nom qui sert habituellement à les désigner dans la littérature médicale.

1 / *Berkson* [2], dans le calcul des taux de survie par la méthode actuarielle, admet que les D_2 sujets de recul insuffisant ont disparu en moyenne au milieu de l'intervalle ; ils ont donc été exposés au risque 1 pendant $\frac{1}{2}$ intervalle seulement. On estime alors q_1 comme le quotient du nombre de décès dus à 1 par le nombre de sujets exposés, les sujets ayant disparu par effet 2 comptant pour $\frac{1}{2}$.

$$\hat{q}_1 = \frac{D_1}{N - \frac{1}{2} D_2} \qquad (1)$$

Admettre qu'un sujet exposé pendant tout l'intervalle équivaut en moyenne à deux sujets exposés pendant $\frac{1}{2}$ intervalle implique les hypothèses suivantes :

1/ la cause 1 agissant seule donnerait des décès uniformément répartis sur l'intervalle ;

2/ q_1 est faible.

q_1 étant faible, on peut considérer que les disparitions par 2 se produisent en moyenne au milieu de l'intervalle si l'on peut supposer que :

3/ la cause 2 agissant seule donnerait des décès uniformément répartis sur l'intervalle.

2 / La méthode "*Sujet Année*" dispense de l'hypothèse 3/ ci-dessus en faisant intervenir *les temps exacts* pendant lesquels les sujets frappés par 2 ont été exposés au risque 1. On rapporte alors le nombre des décès dus à 1 au temps total d'exposition au risque 1 de tous les sujets :

$$\hat{q}_1 = \frac{D_1}{N - (1 - T_2) D_2} \qquad (2)$$

Le "sujet-unité de temps" est ici "l'unité" fondamentale.

3 / *Cornfield* [4], à partir du concept de *force de mortalité*, développe un raisonnement ne faisant appel à aucune hypothèse particulière. La force de mortalité par la cause i, soit $\mu_i(t)$ est définie par :

$$\mu_i(t) = \lim_{\Delta t \to 0} \frac{-\Delta_i\, l(t)}{l(t).\Delta t} \qquad (3)$$

où $l(t)$ est le nombre de survivants à l'instant t, et $-\Delta_i l(t)$ le nombre de décès dus à la cause i pendant l'intervalle de temps Δt. Si la cause i était la seule à agir, il y aurait, à l'instant t, $l_i(t)$ survivants et, pendant dt, on observerait $-dl_i(t)$ décès par i ; on peut donc écrire :

$$\mu_i(t) = \frac{-dl_i(t)}{l_i(t)\,dt}$$

soit :

$$l_i(t) = N\, e^{-\int_0^t \mu_i(u)\,du}$$

et

$$\boxed{q_1 = 1 - e^{-\int_0^t \mu_1(t)\,dt}} \qquad (4)$$

Le problème de la détermination de q se trouve ramené à celui de l'estimation d'une intégrale : $\int_0^1 \mu_1(t)\,dt$; on peut en chercher une *approximation numérique* si $\mu_1(t)$ est défini avec une précision suffisante par (3), ce qui pourra être le cas par exemple si l'on travaille sur une population très large (problèmes démographiques). On peut également poursuivre le calcul à partir de l'expression (4) si l'on peut trouver pour $\mu_1(t)$ un *modèle* mathématique qui s'adapte aux résultats observés ($\mu_1(t)$ constant, linéaire, homographique, etc.).

C'est ainsi qu'on peut retrouver les résultats que Berkson et Elveback [1] obtiennent directement en supposant la *constance* des forces de mortalité $\mu_1(t) = \mu_1$ et $\mu_2(t) = \mu_2$.

Il en résulte en effet de la définition de la force de mortalité que pour l'ensemble des deux causes :

$$\mu = \mu_1 + \mu_2$$

donc :

$$Q = Q_1 + Q_2 = \int_0^1 e^{-\mu t}\,\mu\,dt = 1 - e^{-\mu} \qquad (5)$$

$$Q_1 = \int_0^1 e^{-\mu t}\,\mu_1\,dt = \frac{\mu_1}{\mu}(1 - e^{-\mu})$$

d'où :

$$Q_1 = \frac{\mu_1}{\mu} Q$$

Or d'après (4) et (5) $q_1 = 1 - e^{-\mu_1} = 1 - (1 - Q)^{\frac{\mu_1}{\mu}}$

d'où :

$$\boxed{q_1 = 1 - (1 - Q)^{\frac{q_1}{q}}} \qquad (6)$$

On peut également dans ce cas trouver la variance de $\hat{q}_1$.

Cette formule a été très complètement étudiée par Chin Long Chinag dans [3]. On peut facilement l'étendre au cas où l'on suppose seulement que le rapport $\frac{\mu_1(t)}{\mu_2(t)}$ est constant.

A première vue la formule de Cornfield paraît résoudre le problème que nous nous posons dans le cas le plus général. Elle peut cependant être d'un emploi difficile en dehors des cas cités.

Dans le paragraphe suivant on trouvera une formule plus concrète et aussi générale du taux corrigé, qui permettra de comparer les différentes formules présentées ci-dessus, dans de nombreux cas de calculer $\hat{q}_1$ et d'obtenir une estimation de sa variance.

Kimball [6] a proposé comme taux "purifié" :

$$q_1 = \frac{Q_1}{1 - Q_2} \tag{7}$$

soit :

$$\boxed{\hat{q}_1 = \frac{D_1}{N - D_2}} \tag{8}$$

L'avantage de cette formule est qu'elle ne fait appel à aucune hypothèse sur la distribution des deux causes de décès au cours de l'intervalle ; par contre elle répond à la question : "quelle est la probabilité que le sujet meure de 1 *s'il ne meurt pas de 2 ?*" et non à celle que nous nous posons : "probabilité que le sujet meure de 1 *si 2 n'existe pas ?* ". Elle revient en effet à supprimer les sujets morts de 2 comme s'ils n'avaient pas été exposés au risque 1, alors qu'en fait ils y ont été exposés jusqu'à leur décès.

B - PRESENTATION D'UNE FORMULE GENERALE (DITE "FORMULE G")

Nous allons supposer, pour le raisonnement, que les sujets frappés par l'une des causes de disparition continuent à courir le second risque après leur date de disparition. Dans ces conditions :

$q_1 (1 - q_2)$ est la probabilité qu'un sujet soit frappé par 1 et pas par 2
$q_2 (1 - q_1)$ " " " 2 " 1
$q_1 q_2$ " " " par 1 et par 2.

Appelons λ_1 la probabilité que 1 se produise avant 2 si le sujet est frappé par 1 et par 2, λ_2 la probabilité complèmentaire. On peut alors écrire :

$$Q_1 = q_1 (1 - q_2) + \lambda_1 q_1 q_2 = q_1 (1 - \lambda_2 q_2) \tag{9}$$

$$Q_2 = q_2 (1 - \lambda_1 q_1) \tag{10}$$

Les relations (9 - 10) fournissent $\hat{q}_1$ par résolution d'une équation du second degré (dont une seule racine, la plus petite, est comprise entre 0 et 1) :

$$\boxed{\hat{q}_1 = \frac{N + \lambda_1 D_1 - \lambda_2 D_2 - \sqrt{(N + \lambda_1 D_1 - \lambda_2 D_2)^2 - 4 \lambda_1 D_1 N}}{2 \lambda_1 N}} \tag{11}$$

Précisons la valeur de λ_1 : appelons $q_1(t)$ (respectivement $q_2(t)$) la probabilité d'être frappé par 1 (respectivement 2) avant la date t, si la cause 1 (respectivement 2) est seule à agir :

$$\lambda_1 \, q_1 \, q_2 = \int_0^1 q_1(t) \, dq_2(t) \tag{12}$$

Posons :

$$q^*_1(t) = \frac{q_1(t)}{q_1} \quad \text{et} \quad q^*_2(t) = \frac{q_2(t)}{q_2}$$

$$(q^*_1(0) = q^*_2(0) = 0 \qquad q^*_1(1) = q^*_2(1) = 1)$$

on peut écrire :

$$\lambda_1 = \int_0^1 q^*_1(t) \, dq^*_2(t)$$

c'est-à-dire que λ_1 est *la valeur moyenne de* $q^*_1(t)$ pondéré par la distribution de q^*_2.

$$\boxed{\lambda_1 = \underset{q^*_2}{\text{Moy}} \, [q^*_1(t)]} \tag{13}$$

Il est intéressant de constater que la solution (11) est celle du *maximum de vraisemblance*. La fonction de vraisemblance Φ s'écrit en effet :

$$\Phi = [(1 - q_1)(1 - q_2)]^S \, Q_1^{D_1} \, Q_2^{D_2}$$

soit d'après (9) et (10) :

$$\Phi = (1 - q_1)^S (1 - q_2)^S \, q_1^{D_1} \, q_2^{D_2} \, (1 - \lambda_2 q_2)^{D_1} (1 - \lambda_1 q_1)^{D_2}$$

et l'équation de vraisemblance relative à q_1 s'écrit :

$$\frac{\partial \, \text{Log} \, \Phi}{\partial \, q_1} = \frac{D_1}{q_1} - \frac{S}{1 - q_1} - \frac{\lambda_1 \, D_2}{1 - \lambda_1 \, q_1} = 0 \tag{14}$$

dont la résolution conduit à (11).

Si λ_1 *est connu* ou *peut être déterminé par* (13), on peut poursuivre l'exploitation de la méthode du maximum de vraisemblance, qui fournit la *variance asymptotique de* $\hat{q}_1$: on a en effet :

$$\frac{1}{N \, \text{var} \, q_1} = - E \left(\frac{\partial_2 \, \text{Log} \, \Phi}{\partial \, q_1^2} \right)$$

d'où :

$$\boxed{\text{Var} \, \hat{q}_1 = \frac{\hat{q}_1 \, (1 - \hat{q}_1)}{N} \, \frac{1}{1 - \dfrac{1 - \lambda_1}{1 - \lambda_1 \, \hat{q}_1} \, \hat{q}_2}} \tag{15}$$

On reconnaît le terme habituel $\frac{q(1 - q)}{N}$ multiplié par un terme correctif qui prend la valeur 1 quand q_2 est nul, et dont on peut donner une

expression approchée en remplaçant $\hat{q}_2$ par $\hat{Q}_2$ (ce qui est certainement légitime pour un calcul de variance). On a alors :

$$\text{Var } \hat{q}_1 = \frac{\hat{q}_1 (1 - \hat{q}_1)}{N - \frac{1 - \lambda_1}{1 - \lambda_1 \hat{q}_1} D_2} \tag{16}$$

Or d'après (9) et (10) on peut écrire $\hat{q}_1$ sous la forme :

$$\hat{q}_1 = \frac{D_1}{N'} \text{ en posant } N' = N - \frac{1 - \lambda_1}{1 - \lambda_1 \hat{q}_1} D_2 \tag{17}$$

et alors, d'après (16) :

$$\text{Var } \hat{q}_1 = \frac{\hat{q}_1 (1 - \hat{q}_1)}{N'} \tag{17 bis}$$

c'est-à-dire qu'on *retrouve les formules binomiales en rapportant les décès à l'effectif corrigé N'*.

Cas particulier du dernier intervalle

Dans le dernier intervalle, on a $Q_1 + Q_2 = 1$

La formule G (11) s'écrit :

$$q_1 = \frac{\lambda_1 + Q_1 - |\lambda_1 - Q_1|}{2 \lambda_1}$$

et on a une formule analogue pour q_2.

$$\text{De : } \lambda_1 + \lambda_2 = 1$$
$$Q_1 + Q_2 = 1$$

il résulte que :

$$\lambda_1 - Q_1 = - (\lambda_2 - Q_2)$$

De deux choses l'une :

- ou $Q_1 < \lambda_1$ et alors $q_1 = \frac{Q_1}{\lambda_1}$ et $q_2 = 1$

- ou $Q_1 > \lambda_1$ et alors $q_1 = 1$ et $q_2 = \frac{Q_2}{\lambda_2}$

L'une des deux causes de mortalité au moins n'épargnerait aucun des sujets si elle agissait seule.

C - CONFRONTATION DES DIFFERENTES FORMULES

La formule G va permettre de comparer les principales formules présentées ci-dessus.

1. Méthode de Berkson

Dans la mesure où l'on admet les hypothèses 1/ et 3/ de Berkson, on a :

$$q_1(t) = q_1 t \text{ et } q_2(t) = q_2 t$$

soit $q^*_1(t) = t$

$q^*_2(t) = t$

et par conséquent :

$$\lambda_1 = \underset{q^*_2}{\text{Moy}} [q^*_1(t)] = \int_0^1 t\, dt = \frac{1}{2}$$

La formule G fournit alors d'après (17) :

$$\hat{q}_1 = \frac{D_1}{N - \frac{1}{2 - \hat{q}_1} D_2} \qquad (18)$$

qui n'est autre que la "formule exacte" de Berkson [2 bis].

Notons que cette formule est valable dans un cas théoriquement beaucoup plus général que celui de la double linéarité de $q_1(t)$ et $q_2(t)$, qui est suffisante, mais nullement nécessaire ; il suffit en effet que $\lambda_1 = \frac{1}{2}$.

La formule *approchée de Berkson* s'obtient, *conformément à son hypothèse* 2/ en négligeant q_1 dans le terme $\frac{1}{2 - q_1}$. Nous allons évaluer *l'erreur relative par défaut* que cela entraîne sur q_1.

On peut en effet écrire, à partir de (9) et (10) :

$$q_1 = \frac{Q_1}{1 - \frac{1}{2 - q_1} Q_2} = \frac{Q_1}{1 - \frac{Q_2}{2}} \left(1 + \frac{q_1}{Q_1} \frac{q_1 q_2}{4}\right)$$

L'erreur est donc $\frac{q_1}{Q_1} \frac{q_1 q_2}{4}$. Mais d'après (9) (où $\lambda_1 = \lambda_2 = \frac{1}{2}$), on a :

$$Q_1 \geqslant \frac{1}{2} q_1$$

de même :

$$Q_2 \geqslant \frac{1}{2} q_2$$

L'erreur est donc certainement bornée par $2 Q_1 Q_2$. Elle sera en fait en général de *l'ordre de* $\frac{Q_1 Q_2}{4}$ si q_1 et q_2 sont suffisamment faibles.

Ces remarques permettent de définir le domaine d'emploi de la formule de Berkson.

Notons qu'en posant :

$$\boxed{N' = N - \frac{D_2}{2}} \qquad (19)$$

on peut écrire d'après (17) et (17 bis) :

$$\boxed{\hat{q}_1 = \frac{D_1}{N'}} \quad (20) \qquad \text{et} \qquad \boxed{\text{Var } \hat{q}_1 = \frac{\hat{q}_1 (1 - \hat{q}_1)}{N'}}$$

2. Méthode "Sujet-Année"

On a, d'après (9) et (10) :

$$q_1 = \frac{Q_1}{1 - \lambda_2 q_2}$$

et, d'après (2) :

$$\hat{q}_1 = \frac{\hat{Q}_1}{1 - (1 - T_2)\hat{Q}_2}$$

En employant (2), on commet une *erreur relative par défaut* $\frac{q_1 - \hat{q}_1}{q_1}$ qu'on peut estimer par :

$$\Delta = \frac{q_1}{Q_1}[\lambda_2 q_2 - Q_2 (1 - T_2)]$$

Estimation de l'expression entre crochets

$$\lambda_2 q_2 = \int_0^1 (1 - t)\, dq_2(t)$$

$$E(Q_2 T_2) = \frac{1}{N} E\left(\sum_j t_j\right) = \hat{Q}_2 E(t_j) \sim \int_0^1 t (1 - q_1 t)\, dq_2(t)$$

$$Q_2 = \int_0^1 (1 - q_1 t)\, dq_2(t)$$

soit :

$$[\lambda_2 q_2 - Q_2 (1 - T_2)] \sim q_1 \int_0^1 t (1 - t)\, dq_2(t)$$

L'intégrale est positive si q_2 n'est pas nul (ou concentré au début ou à la fin de l'intervalle) ; elle est bornée par $\frac{1}{4} q_2$.

Finalement $\Delta < \frac{q_1}{Q_1} \frac{q_1 q_2}{4}$, qui est de *l'ordre de* $\frac{Q_1 Q_2}{4}$ si q_1 et q_2 sont suffisamment faibles.

3. Méthode de Cornfield

La formule G et celle de Cornfield, qui ne font appel à aucune hypothèse particulière, doivent être équivalentes. On peut en effet démontrer les formules (9), (10) et (12) à partir de la formule (4).

$$q_1(t) = \int_0^1 \mu_1(t)\ e^{-\int_0^t \mu_1(x)\,dx}\,dt$$

$$Q_1 = \int_0^1 \mu_1(t)\ e^{-\int_0^t \mu(x)\,dx}\ dt$$

donc :

$$q_1 - Q_1 = \int_0^1 (1 - e^{-\int_0^t \mu_2(x)\,dx})\ \mu_1(t)\ e^{-\int_0^t \mu_1(x)\,dx}\ dt$$

$$= \int_0^1 q_2(t)\ dq_1(t)$$

D. ETUDE PLUS COMPLETE DU CAS OU LA REPARTITION DES DECES DUS A LA CAUSE 1 SERAIT UNIFORME SI ELLE AGISSAIT SEULE

(soit $q_1(t) = q_1 t$)

C'est l'hypothèse 1/, déjà invoquée dans les formules de Berkson et Sujet-Année ; elle correspond à un cas pratiquement fort important car elle sera très souvent valable à condition de découper la période étudiée en intervalles suffisamment courts. Par contre il est intéressant *de ne pas fixer la loi* $q_2(t)$, et en particulier de ne pas se limiter au cas ou elle est linéaire ; les causes de disparition qui sont groupées sous le vocable "cause 2" peuvent en effet être multiples et obéir à des lois compliquées (accidents, cannibalisme et sacrifices s'il s'agit d'animaux - disparitions, affections diverses, etc.).

Première méthode

La "formule G" permet d'écrire (d'après 13) :

$$\lambda_1 = \int_0^1 t\ dq^*_2(t)$$

λ_1 est donc *la date moyenne de disparition* des sujets frappés par la cause 2 *agissant seule*. Mais cette date moyenne n'est pas directement accessible par l'observation. Faisons alors intervenir l'espérance mathématique de la date de disparition par la cause 2 *sans élimination de l'action de la cause 1*. C'est :

$$E(t_j) = \frac{1}{Q_2}\int_0^1 t(1 - q_1 t)\ dq_2(t) \qquad (21)$$

On trouve de même pour la cause 1 :

$$E(t_i) = \frac{1}{Q_1}\int_0^1 t\,[1 - q_2(t)]\ q_1\ dt \qquad (22)$$

De (21) et (22) on déduit la relation :

$$2\,Q_1\,E(t_i) + Q_2\,E(t_i) = q_1 + q_2 - q_1\,q_2 - \lambda_2\,q_2$$

et comme :

$$q_1 + q_2 - q_1\,q_2 = Q_1 + Q_2$$

on a :

$$\lambda_2 \, q_2 = Q_1 [1 - 2 \, E(t_i)] + Q_2 [1 - E(t_j)]$$

On peut estimer $E(t_i)$ par T_1 et $E(t_j)$ par T_2, T_1 et T_2 étant, rappelons-le, les dates moyennes *observées* de disparition par 1 et par 2.

Or, (10) peut s'écrire :

$$q_1 = \frac{Q_1}{1 - \lambda_2 \, q_2}$$

On a donc :

$$\boxed{\hat{q}_1 = \frac{D_1}{N - (1 - 2 \, T_1) \, D_1 - (1 - T_2) \, D_2}} \qquad (23)$$

soit encore :

$$\boxed{\hat{q}_1 = \frac{D_1}{S + 2 \sum_i t_i + \sum_j t_j}} \qquad (23 \text{ bis})$$

On peut remarquer que (23) se rapproche d'autant plus de la formule sujet-année que le terme $(1 - 2 \, T_1) \, D_1$ est plus voisin de 0 (c'est-à-dire que D_1 est faible, ou T_1 proche de $\frac{1}{2}$, de la formule de Berkson que $(1 - 2T_1) \, D_1$ est plus voisin de 0 et T_2 plus voisin de $\frac{1}{2}$. On peut ainsi apprécier l'approximation correspondant à chacune de ces deux formules.

La formule (23 bis) montre que, dans le calcul de l'effectif corrigé N' auquel il faut rapporter D_1, on doit tenir compte des survivants pendant tout l'intervalle, des disparus par 2 jusqu'à leur date de disparition (comme on le fait pour la formule sujet-année), des disparus par 1 jusqu'au *double de leur date de disparition* et non jusqu'à la fin de l'intervalle.

L'inconvénient majeur de cette formule est de ne pas fournir d'estimation de la variance. Par analogie avec (17) on peut penser que, si l'on pose :

$$\hat{q}_1 = \frac{D_1}{N'} \text{ avec } N' = N - (1 - 2 \, T_1) \, D_1 - (1 - T_2) \, D_2$$

le terme $\frac{\hat{q}_1 \, (1 - \hat{q}_1)}{N'}$ fournit une valeur approchée de la variance.

Il resterait à le justifier.

Deuxième méthode

Il est possible de donner une solution plus rigoureuse en reprenant le problème par la méthode du maximum de vraisemblance en tenant compte des dates de disparition par 2. Nous ne la donnons que pour mémoire, car elle est d'un emploi peu commode.

La fonction de vraisemblance Φ peut s'écrire :

$$\Phi = (1 - q_1)^s \, (1 - q_2)^s \, \Pi_i \, \{q_1 [1 - q_2(t_i)]\} \, \Pi_j \left[\frac{dq_2}{dt}(t_j) \, (1 - q_1 \, t_j) \right]$$

L'équation de vraisemblance relative à q_1 s'écrit donc :

$$\boxed{\frac{\partial \text{ Log } \Phi}{\partial q_1} = \frac{D_1}{q_1} - \frac{S}{1 - q_1} - \sum_j \frac{t_j}{1 - q_1 t_j} = 0} \qquad (24)$$

La résolution de cette équation se fera par les procédés habituels d'approximations successives.

On peut trouver une estimation de la variance asymptotique de $\hat{q}_1$.

On a en effet :

$$(\text{Var } q_1)^{-1} = N E \frac{\partial_2 \text{ Log } \Phi}{\partial q^2_1}$$

d'où :

$$(\text{Var } \hat{q}_1)^{-1} = \frac{N}{q_1}\left[\frac{1 - \hat{q}_2}{1 - \hat{q}_1} - \frac{\hat{q}_2}{\hat{q}_1^2}(\hat{q}_1 + M)\right]$$

avec :

$$M = \int_0^1 \frac{dq^*_2(t)}{1 - q_1 t}$$

On peut prendre comme valeur approchée de M celle qui correspond au cas simple où $q_2(t) = q_2 t$ (premier terme du développement de $q_2(t)$).

On a alors $M = \text{Log } (1 - q_1)$ soit :

$$\text{Var } \hat{q}_1 = \frac{\hat{q}_1 (1 - \hat{q}_1)}{N} \cdot \frac{1}{1 - \frac{\hat{q}_2}{\hat{q}_1}\left[1 + \frac{1 - \hat{q}_1}{\hat{q}_1} \text{Log } (1 - \hat{q}_1)\right]} \qquad (25)$$

E - ETUDE D'UN EXEMPLE

Date de décès		0,05	0,15	0,25	0,35	0,45	0,55	0,65	0,75	0,85	0,95	Total
Nombre de décès	cause 1	3	2	2	1	2	2	1	0	2	1	16
	cause 2	20	16	18	14	8	4	4	0	4	0	88

$N = 200$ $D_1 = 16$ $D_2 = 88$ $Q_1 = 8\,\%$ $Q_2 = 44\,\%$

$\int_0^1 \mu_1(t)\,dt$ estimé par 0,123 $\sum t_i = T_1\ D_1 = 6,6$ $\sum t_j = T_2\ D_2 = 24,6$

Taux brut : $Q_1 = 8\,\%$

Calcul du taux corrigé :

Méthode ⟶	Berkson		Sujet-Année	Cornfield		Hyp $q_1(t) = q_1 t$	
	approché	exact		estimation numérique de μ_1	μ_i constant (Elveback)	Form. (23)	Max. Vrai. (24)
$\hat{q}_1$	10,3 %	10,5 %	11,7 %	11,6 %	10,7 %	12,0 %	11,9 %

Les formules de Cornfield, "Sujet-Année", (23) et (24) (hypothèse ; $q_1(t) = q_1 t$) donnent pratiquement le même résultat : une correction de l'ordre de 4 % (soit 50 % en valeur relative). L'emploi des formules de Berkson ou Elveback est ici moins justifié, compte tenu de l'allure des courbes de mortalité ; la correction qu'elles fournissent est sous-estimée.

Calcul de la variance :

Dans le cadre de l'emploi de la formule G, on a justifié le calcul de la variance par la *formule binomiale* appliquée à l'effectif corrigé $N' = \frac{D_1}{q_1}$ (17 bis).

En admettant que cette formule est encore valable en première approximation dans les autres cas, on trouve des variances allant de 6 à 8.10^{-4} selon la méthode employée (pour la formule (24) on connaît la variance par (25) : on trouve 7.10^{-4} contre 8.10^{-4} par la méthode approchée).

F - CONCLUSION, FORMULAIRE

Lorsque leur emploi est licite, les différentes formules proposées dans la littérature sont à peu près équivalentes quant au taux corrigé qu'elles fournissent. Par contre les hypothèses qu'elles nécessitent sont différentes et, partant, les domaines d'emploi. On trouvera ci-dessous un tableau des formules et hypothèses correspondantes.

Hypothèse	Nom	Taux	Variance
aucune	Cornfield	$\hat{q}_1 = 1 - e^{-\int_0^1 \hat{\mu}_1(t)\,dt}$	cf. bibliographie
μ_i constant	Elveback	$\hat{q}_1 = 1 - (1 - \hat{Q}) \frac{\hat{Q}_1}{\hat{Q}}$	
aucune	Formule G	$\hat{q}_1 = \frac{D_1}{N'}$ avec $N' = N - \frac{1 - \lambda_1}{1 - \lambda_1 \hat{q}_1} D_2$ et $\lambda_1 = \frac{\text{Moy}}{q_2^*}\,[q_1^*(t)]$	$\sim \frac{\hat{q}_1 (1 - \hat{q}_1)}{N'}$
la cause 1 agissant seule donnerait des décès répartis uniformément	Max. Vrais. (24)	pour mémoire (cf. texte)	cf. texte
- id -	Formule (23)	$\hat{q}_1 = \frac{D_1}{N'}$ avec $N' = S + 2\sum_i t_i + \sum_j t_j$ $= N - (1 - 2T_1) D_1 - (1 - T_2) D_2$	
- id + q_1 suffisant faible	Sujet-Année	$\hat{q}_1 = \frac{D_1}{N'}$ avec $N' = S + D_1 + \sum_j t_j$ $= N - (1 - T_2) D_2$	
$\left(\text{erreur relative} \sim \frac{Q_1 Q_2}{4}\right)$			
chacune des 2 causes agissant seule donnerait des décès répartis uniformément	Berkson "exact"	$\hat{q}_1 = \frac{D_1}{N'}$ avec $N' = N - \frac{1}{2 - q_1} D_2$	$\sim \frac{\hat{q}_1 (1 - \hat{q}_1)}{N'}$
- id + q_1 suffisant faible			
$\left(\text{erreur relative} \sim \frac{Q_1 Q_2}{4}\right)$	Berkson "approché"	$N' = N - \frac{D_2}{2}$	- id -

Il importe de ne pas perdre de vue *l'hypothèse fondamentale*, soulignée dans l'introduction : toutes ces formules reposent sur l'"*indépendance*" des *causes de disparition*.

BIBLIOGRAPHIE

[1] BERKSON, J. and L. ELVEBACK - Competing exponential risk. J. Amer. Statist. Assoc., 55 : 291, 1960. p. 415-428.

[2] BERKSON, J. and R. GAGE - Calculations of survival rates for cancer. Procedings of the Staff meetings of the Mayo Clinic Rochester. 25 : 11, May, 1950. p. 270-286.

[2 bis] Note de correction ajoutée en 1953 à l'article ci-dessus.

[3] CHIN LONG CHIANG - The follow-up study with the considération of competing risks. Biometrics. 17 : 1, 1961. p. 57-78.

[4] CORNFIELD, J. - The estimation of the probability of developing a disease in the presence of competing risks. Amer. J. Public Health, 47 : 5, May 57.

[5] ELVEBACK, L. - Actuarial estimation of survivorship in chronic disease. J. Amer. Statist. Ass., 53 : 282, 1958, p. 420-440.

[6] KIMBALL, A - Dans : Bulletin de l'Institut International de Statistique. 36 : 3, 1958. p. 193-203.

[7] NEYMAN, J. FIRST - Course in probability and Statistics. New York, Holt, 1950.