

KLAUS KRICKEBERG

**Étude, du point de vue statistique, des rapports
sur « La Hague »**

Journal de la société française de statistique, tome 140, n° 3 (1999),
p. 45-51

http://www.numdam.org/item?id=JSFS_1999__140_3_45_0

© Société française de statistique, 1999, tous droits réservés.

L'accès aux archives de la revue « Journal de la société française de statistique » (<http://publications-sfds.math.cnrs.fr/index.php/J-SFdS>) implique l'accord avec les conditions générales d'utilisation (<http://www.numdam.org/conditions>). Toute utilisation commerciale ou impression systématique est constitutive d'une infraction pénale. Toute copie ou impression de ce fichier doit contenir la présente mention de copyright.

NUMDAM

Article numérisé dans le cadre du programme
Numérisation de documents anciens mathématiques
<http://www.numdam.org/>

ÉTUDE, DU POINT DE VUE STATISTIQUE, DES RAPPORTS SUR «LA HAGUE»

Klaus KRICKEBERG *

Les rapports sur La Hague dont il est question ici concernent un problème épidémiologique. L'épidémiologie se définit comme l'étude de l'influence de certains facteurs, dits «facteurs de risque», sur la répartition de telle ou elle condition liée à la santé, notamment d'une maladie, dans une population humaine qu'on appelle la «population cible». Son sujet est donc une répartition, notion purement statistique, et en ce sens l'épidémiologie est un sous-domaine de la statistique. Par conséquent elle utilise des méthodes mathématiques, mais l'essentiel est ailleurs : il s'agit surtout de poser les bonnes questions dans l'esprit de la définition que nous venons de rappeler. Dans une situation concrète, cela revient à donner un sens spécifique aux notions générales qui y figurent, qui sont les facteurs de risque, leur influence, et finalement l'étude. Ces notions sont traitées dans les autres exposés de ce volume, notamment celle de l'influence dans l'article de David Freedman. Nous verrons que le problème des rapports sur La Hague se situe précisément là, à savoir dans la précision des questions et des notions.

Quels sont ces rapports? Beaucoup a été écrit sur le sujet en question, sans doute trop, et je me suis demandé pourquoi l'Académie des Sciences m'avait proposé d'en rajouter; mais peut-être une réflexion qui place le problème dans un contexte épidémiologique et statistique plus général sera utile. En fait, il n'y a que très peu de rapports qui comptent : d'une part, un article de Jean-François Viel et Sylvia Richardson en 1990 et deux articles faisant suite en 1993 et 1995; d'autre part, un article de Jean-François Viel et Dominique Pobel en 1997. Tout le reste, y compris l'exposé actuel, n'est que discussion et commentaire, notamment le grand rapport de 1997 par le «Comité Scientifique» mis en place par le Ministère de l'Aménagement du Territoire et de l'Environnement et le Secrétariat d'Etat à la Santé. Une vue d'ensemble très intéressante des discussions diverses se trouve dans le livre récent de Viel (1998).

Essayons donc de déceler, dans les travaux mentionnés, l'interprétation précise des notions figurant dans la définition de l'épidémiologie. Commençons par l'élément fondamental, la population cible; après tout, une affirmation qui est vraie pour une population ne l'est pas forcément pour une autre. Or, lorsqu'on

* UFR de Mathématiques et d'Informatique et Laboratoire de Statistique Médicale (CNRS), Université de Paris V, 45 rue des Saints Pères, 75270 Paris Cedex 06.

e-mail . krik@math info univ-paris5.fr

Journal de la Société Française de Statistique, tome 140, n° 3, 1999

lit beaucoup de compte-rendus d'enquêtes épidémiologiques, on s'aperçoit que, le plus souvent, la population cible n'est pas définie explicitement mais seulement, d'une manière implicite et imprécise, par l'enquête elle-même, à savoir par l'échantillon de personnes tirées pour les besoins de l'enquête.

Par exemple, dans l'article de Viel et Pobel (1995), le point de départ est l'ensemble des personnes chez lesquelles une leucémie fut diagnostiquée entre le 1er janvier 1978 et le 31 décembre 1992 et qui, au moment du diagnostic, avaient moins de 25 ans et habitaient ce que nous allons appeler la « région d'étude ». Celle-ci est constituée par les 10 cantons situés pour l'essentiel à moins de 35 km de distance de La Hague. On observe ainsi 25 cas de leucémies. On regarde ensuite une « région de référence », en l'occurrence le département du Calvados, où il existe un registre des cancers et où l'on peut donc également trouver le nombre de cas dans telle ou telle tranche d'âge. On concrétise l'idée de la répartition de la maladie en calculant, pour chacune des deux régions, le taux de l'incidence annuelle, défini comme nombre de cas par personne et par an et « standardisé » par tranches d'âge de 5 ans selon les règles de la démographie. L'objectif de l'étude est de comparer ces deux taux, et l'on a l'habitude de reformuler ce problème de la manière suivante : soit E le nombre de cas qui se produiraient dans la région d'étude si le taux y était égal au taux constaté dans la région de référence. Ce nombre s'appelle le « nombre attendu » ou, dans le langage des probabilistes, « espéré », de cas dans la région d'étude. Il s'élève ici à $E = 22,8$, à comparer avec le nombre observé, ou réalisé, $R = 25$.

Quelle était donc la population cible de cette étude ? C'était bien sûr la population dans laquelle on a calculé les taux d'incidence qui intervenaient, à savoir la tranche d'âge 0-24 ans dans les deux régions, et ceci année par année de 1978 à 1992. La population cible change donc dans le temps, ce qui est normal, mais surtout elle se limite aux habitants de la réunion des deux régions. Aucune conclusion ne peut être tirée de l'étude concernant une autre population, ce qui est trivial mais souvent négligé dans les discussions diverses.

Passons au facteur de risque. Qu'en était-il en fait ? C'était évidemment ce qu'on pourrait désigner par « facteur géographique » : son influence se manifeste par la différence entre les taux d'incidence dans les deux régions, ou encore par la différence entre $R = 25$ et $E = 22,8$. Nous nous intéressons maintenant à l'interprétation de cette différence.

Le nombre de cas observés auprès de La Hague, soit 25, est bien supérieur à ce qu'on attend à partir de la situation au Calvados, à savoir 22,8. L'un des buts d'une analyse statistique est de voir si cette différence est compatible avec les fluctuations aléatoires normales. Les méthodes statistiques traditionnelles nous fournissent pour le quotient $R/E = 1,1$ un intervalle de confiance $[0,7; 1,6]$ pour un coefficient de sécurité de 0,95. Cet intervalle contient le nombre 1. Par conséquent, en regardant les cas de leucémie comme résultant de mécanismes aléatoires, on ne peut pas affirmer que ces mécanismes soient différents dans les deux régions, même si l'on est prêt à se tromper avec une probabilité de 0,05. Il en est de même lorsqu'on se limite à certaines tranches

d'âge comme 0-4, 5-14 ou 15-24 ans, ou à un « anneau » autour de La Hague défini par les distances <10 km, 10-20 km, ou 20-35 km, ou aux deux à la fois. En particulier pour l'anneau le plus proche de La Hague, on découvre bien la valeur plus grande $R/E = 2,8$ mais toujours un intervalle de confiance qui inclut le nombre 1, à savoir ici [0,8 ; 7,2]. On ne peut donc pas affirmer (en ce sens) qu'il existe une grappe de cas autour de La Hague.

Ces méthodes statistiques traditionnelles reposent sur certaines hypothèses, en particulier sur celle que le nombre de cas qui se déclarent dans une zone et durant une période données suit une loi de Poisson. L'article de Viel et Pobel (1995) explore les mêmes données aussi par d'autres méthodes statistiques, en partie avec une notion de facteur géographique légèrement modifié. On tombe encore une fois sur certains éléments qui pourraient nous faire penser qu'une petite influence du facteur géographique existe dans le sens d'une augmentation du taux de l'incidence dans l'entourage de La Hague. Pourtant, aucune analyse ne permet de mettre en évidence une telle influence de manière « significative », c'est-à-dire avec un risque de se tromper inférieur à 0,05.

Examinons ensuite l'étude de Pobel et Viel (1997) sous le même angle. Y figurent d'abord les 25 cas qui sont apparus de 1978 à 1992, plus les deux cas qui se sont produits dans la région d'étude en 1993. Pour chaque cas, on choisit un ensemble de « témoins » non atteints de leucémie. Il s'agit donc de ce qu'on appelle une « étude cas-témoins avec appariement ». Encore une fois, aucune population-cible n'avait été définie préalablement, et indépendamment de l'enquête ; le lecteur du rapport doit la reconstituer à partir du mécanisme du choix des témoins ce qui, cette fois, n'est pas tout à fait évident. Le choix des témoins s'effectue ainsi : pour chaque cas, on demande aux médecins généralistes qui l'avaient pris en charge, d'identifier parmi leurs autres patients les parents d'enfants qu'on peut « apparier » avec le cas ; ces enfants sont les témoins pour le cas en cause. Les critères d'appariement utilisés ici sont les suivants : même sexe ; même âge dans la limite de deux ans ; même lieu de naissance dans un sens précis ; lieu de résidence proche au moment du diagnostic du cas, également selon un critère précis.

Quelle est donc la population présomptive que l'on peut en déduire, population dont cet échantillon de cas et de témoins est issu d'une manière « représentative », par exemple dans le sens de la théorie des sondages ? Elle est bien sûr constituée de sujets qui, à une certaine date, avaient moins de 25 ans et habitaient la région d'étude, mais ils sont soumis à d'autres contraintes : un des parents des témoins a consulté un médecin généraliste à tel ou tel moment entre 1978 et 1993. On peut penser que pratiquement tout le monde l'a fait, mais tout de même il n'est pas tout à fait exclu que les témoins proviennent d'une population dont les parents sont plus enclins à se rendre chez un médecin généraliste que les gens moyens. En outre, il ne s'agit pas de n'importe quels généralistes : ils ont eu contact avec un cas de leucémie.

Parfois, l'absence d'une population-cible bien définie dans une enquête n'exclut pas de tirer certaines conclusions utiles, mais dans le contexte de la discussion autour du « problème de La Hague », elle me paraît un peu gênante.

Passons à l'analyse statistique d'une étude de cas-témoins. Elle a un sens pratique seulement pour les maladies rares, mais heureusement la leucémie est rare. Sa philosophie se présente ainsi : pour le facteur de risque à étudier, nous cherchons, par une méthode mathématique appropriée, à mettre en évidence une différence éventuelle entre la répartition de ce facteur parmi les cas d'une part, et les témoins d'autre part. Si, par exemple, tous les cas avaient des yeux bleus et tous les témoins des yeux bruns, nous soupçonnerions que le facteur « couleur des yeux » a une influence sur l'apparition d'une leucémie. Ce raisonnement suppose que le choix des témoins se fait indépendamment du facteur ; dans notre exemple, il ne faut pas privilégier les sujets aux yeux bruns lorsqu'on constitue l'ensemble des témoins.

Par l'appariement on poursuit le but de réduire les fluctuations entre cas et témoins qui ne sont pas directement dues aux différences mentionnées entre les valeurs du facteur de risque en cause. En premier lieu on cherche à augmenter ainsi la « puissance statistique » de l'analyse, c'est-à-dire la probabilité de découvrir et de confirmer l'influence du facteur de risque lorsqu'elle existe. On y réussit notamment lorsqu'il n'y a aucune association statistique entre les critères qui ont défini l'appariement et le facteur de risque. Dans le cas contraire, l'appariement sert à éliminer un « effet troublant » éventuel de ces critères sur l'analyse de l'influence du facteur. Nous allons expliquer cela dans le contexte actuel en portant notre attention, à titre d'exemple, sur un facteur de risque particulier parmi ceux qui font l'objet de l'étude, à savoir la fréquentation des plages locales par le sujet (cas ou témoin).

Le facteur admet six « niveaux », ou valeurs possibles : jamais, seulement pendant les vacances, moins d'une fois par mois, entre une fois par mois et une fois par semaine, entre une fois par semaine et une fois par jour, presque chaque jour. De cette façon, sont définies six catégories de personnes numérotées de 1 à 6. Pour aucun des critères d'appariement (rappelons-les : lien fourni par le médecin généraliste impliqué, sexe, âge, lieu de naissance, lieu de résidence) on ne peut exclure qu'il soit associé à la fréquentation des plages locales. Supposons, par exemple, que le lieu de résidence présente une association statistique forte avec la fréquentation des plages (ce qui est sans doute vrai) et qu'il est lui-même un facteur de risque de leucémie en étant lié à un facteur tiers qui agit directement dans un sens causal et étiologique. Dans ce cas, une étude sans appariement pour le critère « lieu de résidence », et sans toute autre forme de « contrôle » pour le lieu de résidence, pourrait nous amener à affirmer que, en vérité, l'effet en cause était dû à ce facteur causal tiers. C'est l'appariement selon le critère « lieu de résidence » qui empêche une telle affirmation erronée.

Or, lorsqu'on apparie selon beaucoup de critères qui sont associés avec le facteur de risque en cause, il y aura finalement très peu de différences entre les valeurs de ce facteur pour les cas et pour les témoins d'où une perte de puissance statistique : même si le facteur a une influence réelle sur l'apparition des leucémies, avec une probabilité élevée on ne pourra pas la mettre en évidence. Dans ce cas, on parle d'un « sur-appariement ».

Quels sont au juste les résultats de l'analyse statistique de cette étude? Premièrement, lorsqu'on compare le taux de l'incidence des leucémies dans n'importe laquelle des catégories 2-6 avec celui dans la catégorie 1, au niveau de signification 0,05 on ne constate aucune différence. Pourtant, en comparant la réunion des catégories 1-3 avec celle des catégories 4-6, on peut affirmer, avec la même probabilité de se tromper, que le taux dans la deuxième est supérieur au taux dans la première. En fait, pour leur quotient, dit « risque relatif », on trouve l'intervalle de confiance (1,05 ; 8,72) qui ne contient plus le nombre 1 ; l'estimation de ce risque relatif donne la valeur 2,87.

Deuxièmement, les estimations des cinq risques relatifs précités fournissent les valeurs 1,49 ; 1,21 ; 2,28, 4,99 et 6,59 et un test statistique approprié permet d'affirmer, cette fois au niveau de signification de seulement 0,01, qu'ils suivent effectivement une « tendance » croissante.

Ce sont ces résultats, et des résultats similaires pour les facteurs « fréquentation des plages locales par la mère durant la grossesse » et « consommation de poissons et crustacés locaux par le sujet », qui ont déclenché les débats homériques autour du « problème de La Hague ». Ajoutons quelques commentaires.

D'une part, vu le nombre relativement petit de sujets, à savoir 27 cas et en moyenne 7,11 témoins par cas, il importe de souligner que les tests utilisés ici sont « exacts », c'est-à-dire qu'ils ne se fondent pas sur des approximations normales des lois de certaines variables aléatoires impliquées ; il n'y a donc pas d'erreur à craindre de ce côté. En outre, la probabilité d'accepter à tort l'hypothèse « au moins un des trois facteurs mentionnés ci-dessus a une influence » est sans doute bien inférieure à 0,05 mais ne peut pas être calculée parce que ces facteurs sont liés entre eux. Il est difficile de dire s'il y a eu un sur-appariement qui a peut-être « masqué » un effet existant.

D'autre part, comme dans la plupart des études cas-témoins, il a fallu « mesurer » les valeurs des facteurs de risque pour les cas et les témoins de façon rétrospective, à l'aide de questionnaires et d'entretiens, par exemple en posant à la mère d'un cas qui s'est déclaré il y a 13 ans, la question suivante : avant de tomber malade, votre enfant a-t-il fréquenté les plages à tel ou tel rythme ? On ne se souvient pas toujours bien et, en particulier, je pense à l'instar d'autres commentateurs et contrairement aux auteurs de l'étude, que les parents des cas pourraient être tentés d'indiquer une fréquence plus élevée, ce qui conduirait, dans l'analyse statistique, à une surestimation du risque relatif.

Dans toute la discussion précédente, le mot « rayonnement » n'apparaît nulle part, et pour cause : ce facteur n'a pas été étudié. Néanmoins, dans la partie « Conclusions » du résumé de l'article (1997), il est écrit : *there is some convincing evidence in childhood leukaemia of a causal role for environmental radiation exposure from recreational activities on beaches*. Cette affirmation reflète un procédé fréquent en épidémiologie, à savoir la substitution d'un facteur « succédané » à un autre, en postulant un lien étroit entre les deux. Or, pour les deux facteurs « exposition aux rayonnements » et « fréquentation des plages » dans le région d'étude, ce lien n'est pas encore très bien étudié.

Terminons ces remarques en jetant un coup d'œil dans le rapport du comité scientifique. Il est plein de répétitions, mais deux choses retiennent l'attention.

La première est une tentative pour établir une contradiction entre les résultats des articles (1995) et (1997) par le raisonnement suivant lié à ce qu'il est convenu d'appeler, en épidémiologie, le « risque attribué » ou « risque étiologique ». Le taux de l'incidence dans les catégories 4-6 réunies est 2,87 fois le taux dans les catégories 1-3, par conséquent le nombre espéré de cas dans les catégories 4-6 est 2,87 fois le nombre qu'on y espère en appliquant le taux de la région de référence (le Calvados). Dans l'échantillon constitué par les témoins, la proportion des catégories 4-6 est de 0,3. En supposant que la même proportion vaut dans la population-cible, cela nous mène, dans cette population, à un nombre de cas espéré de 41 tandis qu'on en observe seulement 25.

Ce raisonnement comporte au moins deux étapes fallacieuses. D'abord, mis à part le vague général dans la définition des populations impliquées, les réunions de certaines catégories sont des sous-populations particulières auxquelles on ne peut pas appliquer les taux de la population de référence. Ensuite, on fait comme si les taux et proportions qu'on utilise étaient les vraies valeurs; or, ce sont des estimations soumises à des erreurs importantes. Par exemple, si l'on travaille, au lieu du risque relatif 2,87, avec l'extrémité gauche de son intervalle de confiance, à savoir 1,05, on obtient un résultat tout à fait différent.

Le deuxième élément intéressant du rapport, très positif celui-ci, est une recommandation que je vais citer sous une forme condensée : « Mise en place d'une structure chargée de recueillir des informations sur différents indicateurs de santé et d'exposition (facteurs de risque) à l'échelle de petites unités géographiques et analyse de leurs variations géographiques et temporelles ». D'autres recommandations concernent plus spécifiquement les leucémies de l'enfant, notamment au niveau national. Un tel travail systématique et à long terme, qui s'appuie sur les méthodes modernes de la statistique spatiale et temporelle, est effectivement indispensable pour comprendre les divers problèmes dont nous avons parlé et les problèmes similaires concernant d'autres maladies, et pour en tirer des conclusions portant sur des mesures préventives.

RÉFÉRENCES

- [1] VIEL J.-F., RICHARDSON S.T. (1990), Childhood Leucaemia Around the La Hague Nuclear Waste Reprocessing Plant, *Brit. Med. J.*, 300, 580-581.
- [2] VIEL J.-F., RICHARDSON S.T. (1993), Childhood Leucaemia Incidence in the Vicinity of the La Hague Nuclear Waste Reprocessing Facility, *Cancer Causes and Control*, 4, 341-343
- [3] VIEL J.-F., POBEL D. (1995), Incidence of Leucaemia on Young People Around the La Hague Nuclear Waste Reprocessing Plant : a Sensitivity Analysis, *Statist. in Med.*, 14, 2459-2472.

ÉTUDE DES RAPPORTS SUR « LA HAGUE »

- [4] POBEL D., VIEL J.-F. (1997), Case Control Study of Leucaemia Among Young People Near the La Hague Nuclear Waste Reprocessing Plant : the Environmental Hypothesis Revisited, Brit. Med. J , 314, 101-106.
- [5] Comité scientifique pour une nouvelle étude épidémiologique des leucémies dans le Nord-Cotentin : rapport remis à Madame le Ministre de l'Aménagement du Territoire et de l'Environnement et Monsieur le Secrétaire d'Etat à la Santé, 1^{er} juillet 1997.
- [6] VIEL J.-F. (1998), La santé publique atomisée, Ed. La Découverte, Paris.