Société de Calcul Mathématique, S. A.
Algorithmes et Optimisation

Méthodologie probabiliste des études épidémiologiques

- *Evaluation critique et essai de définition de bonnes pratiques* -

Société de Calcul Mathématique SA

rédaction : Bernard Beauzamy et Manon Baradat

juillet 2009

Siège social et bureaux : 111, Faubourg Saint Honoré, 75008 Paris. Tel : 01 42 89 10 89. Fax : 01 42 89 10 69. www.scmsa.com
Société Anonyme au capital de 56 200 Euros. RCS : Paris B 399 991 041. SIRET : 399 991 041 00035. APE : 7219Z

Introduction

En 1904, la Chambre Criminelle de la Cour de Cassation a chargé le mathématicien Henri Poincaré d'une expertise concernant le "Système Bertillon", système probabiliste de graphologie qui était à la base de l'accusation portée contre Alfred Dreyfus. Voici ce qu'écrit Henri Poincaré dans son rapport [Poincaré] : *« le calcul des probabilités ne devrait pas empêcher les savants d'avoir du bon sens ».*

La Chambre Criminelle de la Cour de Cassation réhabilite Dreyfus et ridiculise les experts [Dreyfus] :

"La reconstitution du Bordereau effectuée par Bertillon est fausse ; ces planches sont le résultat d'un traitement compliqué, infligé au document primitif, et d'où celui-ci est sorti altéré, après avoir subi une série d'agrandissements et de réductions photographiques, et même de calcages, recalcages, découpages, collages, gouachages, badigeonnages et retouches. "

La question qui nous est posée est de nous prononcer sur la valeur méthodologique de certaines études épidémiologiques : les lignes à haute tension favorisent-elles l'apparition de certaines maladies ? Nous avions déjà procédé à de telles expertises, à la demande du CEA, à propos des études relatives à l'effet des rayonnements ionisants sur la santé (2007-2008).

Ces questions, en elles-mêmes, sont légitimes et raisonnables. Mais les études que nous avons examinées s'appuient sur des données brutes (des nombres de morts, dans certaines populations et dans des populations de référence) et l'examen immédiat de ces données permet de conclure immédiatement à l'inverse des auteurs : il y a moins de morts dans ces zones "à risque", proportionnellement à la population, qu'ailleurs.

Les auteurs ont-ils du bon sens ? Se sont-ils laissé abuser par ces logiciels statistiques modernes qui permettent d'obtenir n'importe quel résultat, pourvu qu'on presse le bouton approprié ? Sans doute, et l'on peut se demander, comme le fait André Aurengo [Aurengo] si l'épidémiologie est encore une science. Ou bien, comme disait plaisamment Laurent Schwartz, "quand on veut démontrer quelque chose, on y arrive toujours, même si c'est faux !".

Le principal reproche que l'on peut faire à ces études va cependant bien au-delà, et ce n'est pas seulement de manquer de bon sens. Ce n'est pas non plus la précision des calculs, la justesse des données, qui posent problème. Le problème majeur tient aux fautes de logique qui sont commises : le problème n'est pas correctement posé, et les données qui sont fournies ne peuvent permettre d'y répondre, quand bien même il y aurait dix fois plus de morts dans les zones à risque !

Henri Poincaré fait précéder son analyse critique d'un rappel des notions fondamentales des probabilités : nous suivrons son exemple – il est de plus mauvais maîtres !

Nous commençons donc par un exposé détaillé des principes probabilistes qui doivent être utilisés en épidémiologie : il s'agit de principes logiques. Après quoi, nous passons chacune des études au crible des principes logiques qui viennent d'être énoncés, et les fautes apparaissent clairement.

La qualité d'une méthodologie

Ce qu'on demande au mathématicien, c'est de se prononcer sur la qualité d'une méthodologie : nous n'avons pas à nous prononcer sur les résultats. Que les lignes HT ou les faibles doses de rayonnement soient ou non dangereuses pour la santé, ce n'est pas ce que nous examinons ici.

Les mathématiques sont de portée universelle ; elles n'obéissent qu'aux seules lois de la logique. Il n'y a pas des "mathématiques électriques" et des "mathématiques des faibles doses", il y a des mathématiques tout court, et ce sont les mêmes à Tombouctou et à Paris. Il n'y a pas non plus, comme le rappelle André Aurengo, de "mathématiques citoyennes". Si vos données au départ sont justes et si vos raisonnements sont corrects, vos conclusions auront une valeur.

La question n'est pas de savoir qui signe l'étude : une étude sur la fiabilité des chaudières, signée Landru, payée par les fabricants de chaudières, sera correcte si elle est correctement menée ; à l'inverse, une étude menée par les chercheurs de Yale, patronnée par Sainte Thérèse d'Avila, peut être sans valeur si elle comporte des fautes de logique. La question n'est pas non plus de savoir qui la finance ; ces questions : qui signe ? qui finance ? sont précisément des fautes de logique.

On nous dit souvent : mais enfin, toute une profession agit de telle manière, utilise tels tests au quotidien. Peut-être, mais cela ne lui donne pas raison pour autant. La logique n'a rien à voir avec le consensus. Des milliards de gens, tous les jours, prennent des décisions sur la foi de convictions, de croyances, que rien ne vient étayer.

Si une faute de logique est décelée, les conclusions sont sans valeur scientifique ; elles parviendront peut-être, ici ou là, à émouvoir des âmes convaincues d'avance, des journaux, des politiciens à la recherche de soutiens, tout une frange de l'humanité qui se nourrit de pseudo-science et s'y complaît, mais, encore une fois, elles sont sans valeur. Ce n'est plus notre problème ; comme disait Von Neumann, le mathématicien n'est pas responsable de l'obscurantisme dans le monde.

S'agissant de données réelles et de raisonnements en situation réelle, la notion de "donnée correcte" et de "raisonnement juste" doit être abordée avec précaution : nous ne sommes pas en présence de situations académiques, avec des tableaux bien faits et un cadre axiomatique bien défini. Toute donnée de la vie réelle est entachée d'incertitude, tout raisonnement réel fait nécessairement un certain nombre d'hypothèses, explicites ou non, qui peuvent se révéler douteuses. Et enfin, personne, et surtout pas la SCM, n'est "l'arbitre des élégances" ; nos propres raisonnements peuvent être entachés d'erreurs.

La meilleure façon de se prémunir contre ces erreurs est de tout présenter, de tout rédiger : chacun peut ainsi vérifier. Une étude scientifique n'a de valeur que si elle est intégralement vérifiable (ce qui exige que les données soient présentes). A l'inverse, un travail qui vous dit "nous sommes partis de telles données (non présentées) et nous avons fait tel test (on ne sait pas comment)" est absolument sans valeur.

C'est pourquoi, avant de présenter notre analyse critique, nous commençons par définir les règles auxquelles toute étude statistique (et notamment épidémiologique) doit se soumettre. C'est ce que nous appelons les "bonnes pratiques" ; nous les définissons de manière si explicite, si détaillée, que chacun pourra ainsi former son jugement.

Un outil très souvent utilisé est le "modèle de Cox" ; nous traitons un exemple simple (Annexe 3) : deux populations de même importance, soumises à deux traitements différents. L'un favorise la mortalité à court terme, mais garantit une certaine longévité aux survivants, et l'autre fait l'inverse, l'espérance de vie globale étant la même pour les deux. Le modèle de Cox affirme

brutalement que l'une est meilleure que l'autre, ce qui est faux. Bien entendu, les hypothèses d'emploi du modèle ne sont pas respectées, mais les tests statistiques préalables à l'emploi sont cependant positifs, sans réserve !

L'absence de danger

On est frappé de voir que les études épidémiologiques sont incapables de conclure à l'absence de danger. On lit souvent "cette étude n'a pas pu mettre le danger en évidence" ; on ne lit jamais "cette étude montre qu'il n'y a pas de danger".

Cette incapacité est choquante à deux titres. Tout d'abord, un test, quel qu'il soit, a vocation à départager clairement entre deux situations ; on n'admettrait pas un test de grossesse qui dirait : 0 = je ne sais pas ; 1 = vous êtes enceinte.

Ensuite, on se dit, comme Poincaré, qu'un seul Bertillon manipulant un seul bordereau, ce n'est pas là le cadre adéquat pour le calcul des probabilités. Mais, s'agissant de dizaines de millions de personnes, exposées pendant des dizaines d'années, si le risque existe, on doit le voir, et s'il n'existe pas, on doit l'écrire : nous sommes là dans un cadre typique d'application de la loi des grands nombres, qui relève des sciences exactes !

En vérité, les méthodes mathématiques qui permettent de mettre en évidence le risque ou son absence existent et sont fort simples ; elles reposent d'abord sur la comparaison des cas recensés, à proportion des populations : c'est du bon sens. Nous les détaillons au chapitre II.

Si, dans la zone à risque, il y a moins de cas (en proportion de la population) que dans la zone de référence, il faut avoir le courage de conclure : "il n'y a rien à voir", et ne pas se lancer dans l'utilisation désordonnée de tests statistiques que le premier mathématicien venu tournera en ridicule.

Mais, à l'inverse, si cette comparaison montre qu'il y a quelque chose à voir, c'est là qu'intervient l'art de l'épidémiologiste : essayer de voir si le risque est fonction de l'âge, du poids, de l'exposition au tabac, de l'usage de l'alcool, etc. Toutes ces questions sont, à l'évidence, extrêmement difficiles. Elles vont conduire à une réduction des populations prises en considération (par exemple, se limiter aux non-fumeurs, au sein de la population test et de la population de référence).

Si l'on souhaite mettre en évidence un mécanisme d'action (par exemple un champ magnétique) et pas seulement la présence d'un danger, il faut faire intervenir la physique de ce mécanisme. Un champ magnétique dépend de l'intensité du courant qui le crée et de la distance à l'observateur ; s'il est source de risque, à distance égale, il doit donc y avoir plus de cas de maladie à proximité des lignes à forte intensité que des lignes à faible intensité. Il faut avoir le courage de faire toutes les vérifications qu'implique la physique du problème et le courage de rejeter l'hypothèse faite si ces vérifications sont négatives.

Nous espérons ici que les bonnes pratiques que nous cherchons à définir permettront aux épidémiologistes une vigilance accrue quant aux outils qu'ils emploient.

Remerciements

Nous avons rédigé et diffusé un premier document de travail ; la présente version tient compte :

- Des commentaires du Professeur Aurengo (Hôpital de la Pitié-Salpétrière) ;
- Des documents issus du Laboratoire de Zététique, transmis par M. Aurengo ;
- Du document "Revue critique des études épidémiologiques", IRSN 2008, transmis par M. Jacques Repussard ;
- Des commentaires de M. Jean-Claude Barescut, IRSN.

Nous remercions chacun pour sa contribution.

Chapitre I

Les bonnes pratiques probabilistes en épidémiologie

Contrairement à ce que l'on croit souvent, les probabilités sont une science exacte, mais elles ont, comme toute science, un domaine d'application qu'il convient de respecter.

Une étude épidémiologique consiste généralement en la comparaison entre une "population-test" et une "population de référence", du point de vue de l'exposition à un risque. Ces "populations" peuvent être des zones, des tranches d'âge, des catégories socio-professionnelles, etc.

I. Le phénomène que l'on souhaite mettre en évidence doit être convenablement défini

Le phénomène est en général caractérisé par un danger, par exemple un surcroît de mortalité, dû à telle maladie. Mais la population test et la population de référence dépendent du choix qui est fait. Voici trois exemples ; ces trois questions sont parfaitement légitimes, mais elles conduisent à des études distinctes :

Première question : "les lignes HT sont-elles dangereuses pour la santé ?" (sans autre précision, sans savoir de quelle manière)

Pour répondre à cette question, on comparera l'espérance de vie à la naissance des gens qui vivent à proximité des lignes et celle du pays tout entier.

Seconde question : "favorisent-elles l'apparition de la maladie d'Alzheimer ? ".

Pour répondre à cette question, on comptera le nombre d'Alzheimer, pour 1 000 habitants, au voisinage des lignes, et on le comparera au nombre d'Alzheimer d'une population de référence de même âge.

Troisième question : les champs magnétiques sont-ils dangereux pour la santé ?

Pour répondre à cette question, on évaluera l'espérance de vie à la naissance des gens qui vivent en présence d'un fort champ magnétique (en particulier au voisinage proche d'une ligne HT de forte intensité) et on la comparera à l'espérance de vie dans l'ensemble de la population.

Comme on voit, les trois études sont différentes, et la population de référence ne sera pas la même.

II. La population de référence doit être aussi vaste que possible

Supposons, comme c'est le cas ici, que l'on cherche à tester la dangerosité des lignes HT. On comparera la population vivant "près" des lignes HT à la population vivant "loin ". Comme la population vivant près est minoritaire, il est légitime de prendre pour population de référence l'ensemble de la population du pays (ou de la région, ou du continent, etc.). On aura donc d'une part une "population test", qui est constituée de gens vivant près des lignes, et une "population de référence", qui est l'ensemble du pays.

Il n'est absolument pas légitime (comme le fait l'une des études) de prendre pour population de référence une population de même taille, mais située ailleurs, même si elle est extraite aléatoirement. En effet, la population test est en très petit nombre, et une extraction aléatoire peut donner des résultats faux. Une extraction aléatoire doit systématiquement être proscrite lorsque la loi du phénomène est inconnue, comme c'est le cas ici, même si la taille de l'échantillon est importante. Imaginons un explorateur qui veut s'enquérir des habitudes alimentaires des Chinois ; il interroge cent millions de personnes sur la côte et il aura toujours la même réponse : poisson.

Retenons cette règle :

Règle 1. – *La population de référence doit systématiquement être aussi vaste que possible. L'extraction aléatoire d'une population de référence représente une faute de logique, puisqu'on a extrait selon une loi de probabilité définie a priori (factice), alors que l'on ne connaît pas la vraie loi.*

A propos de cette règle, nous recevons le commentaire qui suit (fait à partir du document de travail du 09/06/2009) :

Le fait que la population vivant près des lignes soit minoritaire n'a rien à voir – strictement aucun lien d'implication, contrairement à ce qui est écrit – avec la légitimité qu'il y aurait à prendre pour population de référence l'ensemble de la population du pays. La population est certes minoritaire mais les maladies (de cette population minoritaire) ne le sont peut-être pas !

Il y a ici à mon avis une confusion de domaine, confusion qui peut ne pas trop se remarquer dans le cas où ce que l'on recherche (ici une maladie) dans ladite population minoritaire est également minoritaire à l'intérieur de cette population.

La population test étant incluse dans la population totale, le problème est faussé d'entrée de jeu.

Pour être clair, un petit exemple chiffré : Imaginons un petit sous-groupe de personnes X représentant 1/1000 d'une population de 10 millions de personnes (donc 10.000 personnes X) mais ayant un taux d'incidence d'une maladie quelconque de 50% (donc 5000 cas de cette maladie – provoquée par ce que vous voulez, peu importe – chez les personnes X) alors que dans la population générale (population autre que les X) le taux soit de 0,1% (donc 9990 cas).

Moralité : les cas de maladie chez les X représentent plus de la moitié des cas totaux dans la population (hors X) et il serait, selon BB, (puisque les personnes X sont minoritaires dans la population)... légitime de prendre comme population de référence l'ensemble de la population, c'est-à-dire les 10 millions de personnes avec les... 14990 cas de maladie !

Ce commentaire est très intéressant et mérite d'être analysé. L'auteur commet en effet plusieurs erreurs.

Tout d'abord, il n'y a évidemment rien qui interdise, sur le plan de la logique, de regarder une population de référence égale à la France entière et une population test réduite à Paris ; le fait que Paris soit en France n'invalide pas une étude. De même, on a parfaitement le droit de se demander si l'espérance de vie au sud de la Loire est supérieure ou non à celle de la France entière. On ne voit pas au nom de quel axiome universitaire les deux ensembles (référence et test) devraient être disjoints !

Que les maladies de la population test soient ou non minoritaires dans l'ensemble des maladies est absolument sans rapport avec la question. Dans l'exemple choisi par notre interlocuteur, la population totale est effectivement de 10 millions de personnes, avec 14 990 cas de maladies, et la population test de 10 000 personnes avec 5000 malades ; la comparaison des quotients $\frac{14\,990}{10\,000\,000}$ et $\frac{5\,000}{10\,000}$ montre bien qu'il y a un problème dans la population test.

Notons N_{ref}, N_{test} les populations de référence et de test, et n_{ref}, n_{test} les nombres de malades. Bien entendu, si ces nombres sont connus avec exactitude, le problème ne se pose pas : dans la population "hors test" (complémentaire de la population test, par rapport à la population de référence), la population est $N_{ref} - N_{test}$ et le nombre de cas est $n_{ref} - n_{test}$. On peut comparer la population test à la population de référence, ou au complémentaire de la population test : les deux sont légitimes et le passage de l'un à l'autre est immédiat.

Mais, en pratique, il y a des "incertitudes de frontière" : par exemple, pour Alzheimer, on n'est pas sûr d'avoir recensé correctement les décès vivant à proximité des lignes, tandis que, pour la population tout entière, l'incertitude est moins grande. A cause de ces "incertitudes de frontière", notre règle : "prendre une population de référence la plus large possible" est particulièrement nécessaire. En effet, si vous vous trompez de quelques unités pour les décès au sein de la population test, vous vous trompez d'autant, en sens contraire, dans la population "hors test" : l'erreur est en quelque sorte multipliée par deux. Pour bien faire comprendre ceci, prenons un exemple.

Admettons que la population test soit de 20 000 personnes, et la population hors test aussi (donc $N_{ref} = 40\,000$). Nous avons un total $n_{ref} = 50$ décès.

Admettons d'abord que 25 décès soient dans la population test, et donc 25 en dehors. Alors la probabilité que la zone test soit plus dangereuse que la zone de référence est évidemment 0.5.

Imaginons maintenant que, par suite des incertitudes de dénombrement, nous ayons attribué seulement 22 décès à la zone test. Nous comparons à la population de référence, et nous comparons le rapport $\frac{22}{20\,000}$ au rapport $\frac{50}{40\,000}$; la théorie générale (voir Appendice 2) montre alors que la probabilité que la zone test soit plus dangereuse que la référence est 0.32.

Comparons maintenant la zone test à son complémentaire, la zone hors test ; nous comparons alors le rapport $\frac{22}{20\,000}$ au rapport $\frac{28}{20\,000}$, et la théorie générale montre que la probabilité que la première soit plus dangereuse que la seconde est 0.20 : le résultat est complètement différent.

Notre règle : "prendre une population de référence la plus large possible" répond donc à trois impératifs :

- C'est en général pour le pays tout entier que les chiffres sont connus ;

- Plus la population est large, et plus les conclusions sont stables ;

- Cela correspond à la véritable attente des usagers (qui se demandent par exemple : suis-je soumis à un danger si je suis au voisinage des lignes HT ? sous entendu par rapport à un individu moyen).

III. Prise en compte de la mort naturelle

Les épidémiologistes ont souvent tendance à l'oublier : les êtres humains finissent par mourir, aussi bien dans la population test que dans la population de référence. Par conséquent, la question : "combien de morts ?" a pour réponse : "tout le monde", dans les deux populations. On posera donc plutôt les questions : "de quoi", et "quand ?".

La question "de quoi ?" est naturelle : s'il apparaît qu'il y a un surcroît d'Alzheimer dans la population test, cela peut indiquer que cette population est soumise à une influence dangereuse. Il s'agit cependant d'une erreur de logique, comme nous allons le voir.

Dans notre étude pour le CEA (2007-2008), nous avons observé que le nombre de cancers en France augmentait globalement, mais diminuait pour chaque tranche d'âge, par exemple 0 – 80 ans (phénomène bien connu ; nous ne sommes pas les premiers à faire cette observation !). Ceci signifie que l'on vit de plus en plus vieux en France, et on finit par mourir d'un cancer alors que l'on mourait d'autre chose avant.

Prenons une zone dite "à risque", comme par exemple la région à proximité de lignes HT.

- Dire : "elle est dangereuse parce qu'il y a plus de morts" constitue une faute de logique : il faut au préalable avoir vérifié que la zone en question ne contenait pas, tout simplement, plus de vieux (ou moins de jeunes) que la population de référence.

- Dire "elle est dangereuse parce qu'il y a plus d'Alzheimer" constitue aussi une faute de logique, parce que si la proportion de vieux est plus élevée, il y aura mécaniquement plus d'Alzheimer, lignes HT ou pas.

Règle 2. - *Dans le cas de maladies liées à l'âge, analyser le nombre de morts ou de cas dans la zone test et dans la population de référence constitue une faute de logique, si l'on n'analyse pas en même temps la pyramide des âges de chaque zone.*

Sans aller jusqu'à l'analyse de toutes les tranches d'âge, il faudrait au moins vérifier que dans la population test et dans la population de référence la proportion des plus de 60 ans est la même. L'étude [Huss], dont nous parlons plus loin, commet cette faute de logique.

La question de la pyramide des âges, dans une zone donnée, n'est nullement une évidence. Les populations humaines se répartissent très diversement du point de vue de la pyramide des âges. Tout d'abord, d'un pays à l'autre (et même entre pays voisins, comme la France et l'Allemagne), il y a de fortes disparités dans les taux de naissance. A l'intérieur d'un même pays, il y a des régions plus jeunes et d'autres moins, en particulier selon le type d'emplois. Et surtout, localement, à l'intérieur d'une même ville, il y a de fortes différences : certains quartiers ont des équipements qui attirent les ménages ayant beaucoup d'enfants, ou l'inverse ; le montant des loyers, les moyens de transport, etc., sont autant de critères qui influent sur la diversité.

Ne mentionnons pas (simple question d'honnêteté intellectuelle !) le fait qu'une zone donnée peut précisément accueillir, de manière absolument factuelle, un surcroît de maladie, s'il s'y trouve précisément un hôpital, une maison de retraite, etc. L'étude [Huss] prend explicitement cette précaution.

IV. Quelle question poser ?

La bonne question à poser, pour tester la dangerosité d'une zone, n'est pas "de quoi l'on meurt ?", mais "quand ?". Si je peux dire : l'espérance de vie des filles en France est 80 ans à la naissance, mais seulement 75 pour cette zone-là, je puis légitimement dire qu'il y a un problème. En fait, comme nous le verrons, le bon indicateur n'est pas l'espérance de vie (qui est trop globale, trop grossière), mais la probabilité de mourir à un âge donné.

La bonne question est celle de la probabilité, à la naissance, de vivre au moins 10, 20, 30,… années, comme nous allons maintenant l'expliquer, et toute autre présentation doit en définitive se ramener à celle-là.

1. La présentation des résultats

Règle 3. - *La présentation des résultats qui permet la comparaison entre la population test et la population de référence est nécessairement de la forme suivante :*

Tranche d'âge	population test	population de référence
T_1	p_1	p'_1
T_2	p_2	p'_2
⋮	⋮	⋮
T_K	p_K	p'_K

Tableau 1 : la présentation des résultats

où $T_1, T_2, ..., T_K$ sont des tranches d'âge (par exemple : 0-10 ans, $10-20$,…, $110-120$) et p_k la probabilité à la naissance, pour un membre de la population test, de mourir dans cette tranche-là. Par exemple, p_2 est la probabilité qu'un nouveau-né de la population test ait une durée de vie entre 10 et 20 ans.

Les probabilités p'_k sont définies de manière identique, mais pour la population de référence. Bien entendu, $\sum_{k=1}^{K} p_k = \sum_{k=1}^{K} p'_k = 1$: on finit par mourir un jour ou l'autre.

2. Utilisation des résultats du tableau

Pour utiliser ces résultats, il faut tout d'abord passer aux cumuls (en probabilité : fonction de répartition), car la comparaison entre les p_k et les p'_k n'est pas possible directement.

Le cumul $p_k + \cdots + p_K$ représente, pour tout k, la probabilité de décéder au cours de l'une des tranches d'âge $T_k, ..., T_K$, soit, en langage commun, la probabilité de dépasser la tranche $k-1$: dépasser un âge donné, et non plus mourir dans une tranche d'âge. On constitue donc le tableau suivant :

Tranche d'âge	population test	population de référence
T_1	P_1	P'_1
T_2	P_2	P'_2
$\vdots$	$\vdots$	$\vdots$
T_K	P_K	P'_K

Tableau 2 : les cumuls

où $P_1 = p_1 + \cdots + p_K$, $P_2 = p_2 + \cdots + p_K, ...,$ $P_j = p_j + \cdots + p_K$, et de même pour les P'_j. Le cumul P_j s'entend donc comme la probabilité de dépasser une borne. Si la première tranche commence à la naissance, on a $P_1 = 1$ et si K est la dernière tranche de vie possible, $P_{K+1} = 0$.

3. Exploitation des résultats

On représente les deux fonctions de répartition sur un même graphique, et on compare. Voici quatre situations parmi toutes les situations possibles :

cas 1 : la fonction de répartition F_1 pour la population test est toujours au dessus de la fonction F_2 pour la population de référence

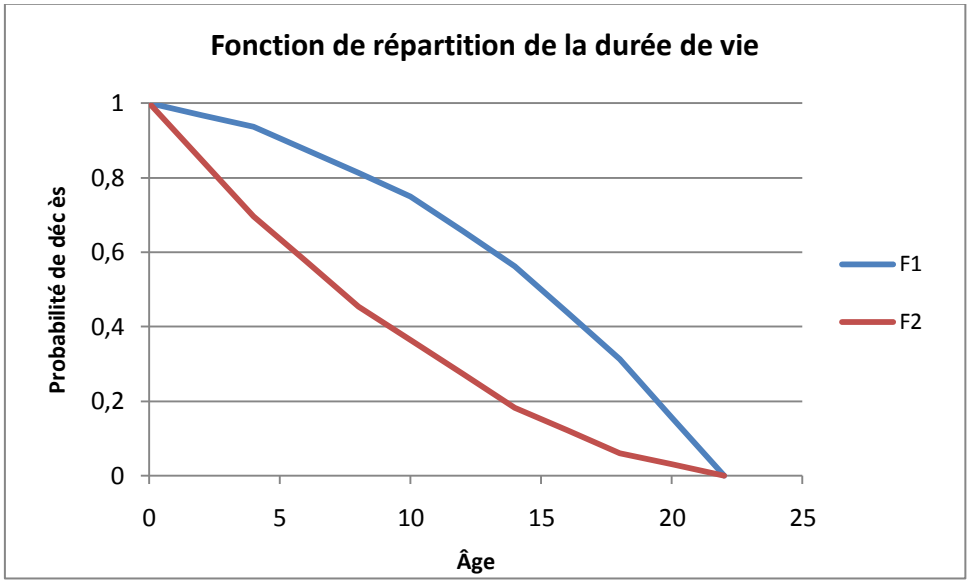

Figure 1 : Fonctions de répartition de la durée de vie

cas 2 : F_1 est toujours au dessous de F_2

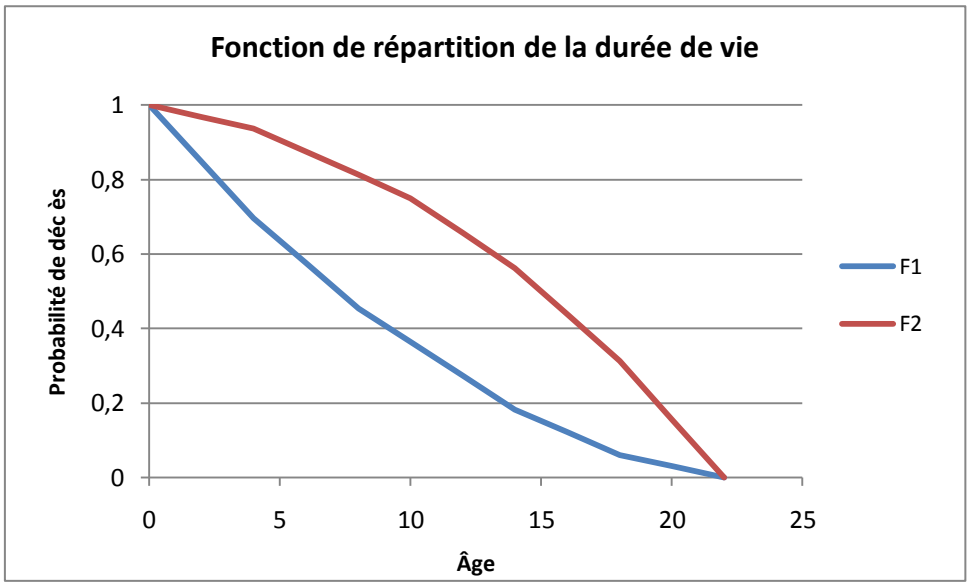

Figure 2 : Fonctions de répartition de la durée de vie

cas 3 : F_1 est d'abord au dessus de F_2, puis au dessous

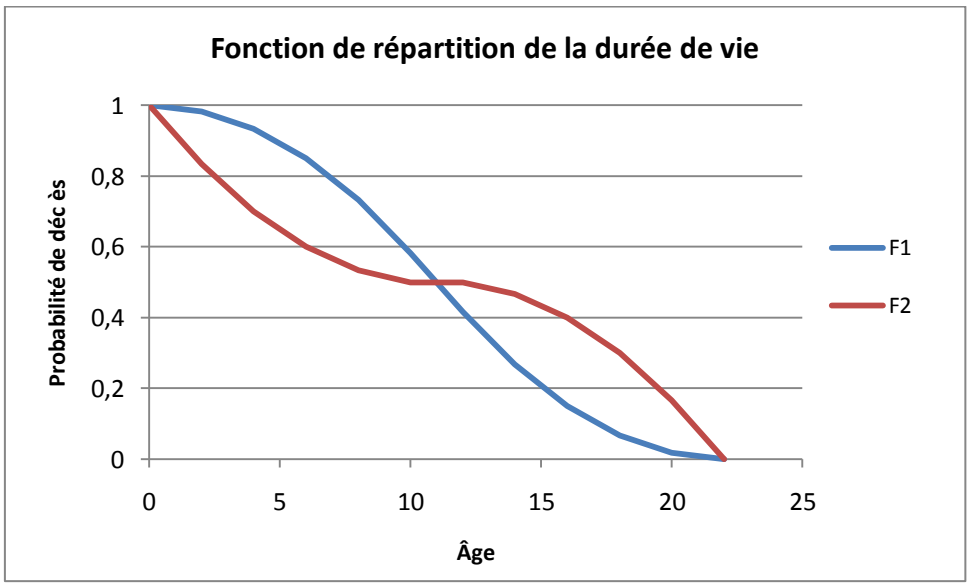

Figure 3 : Fonctions de répartition de la durée de vie

cas 4 : F_1 est d'abord au dessous de F_2, puis au dessus.

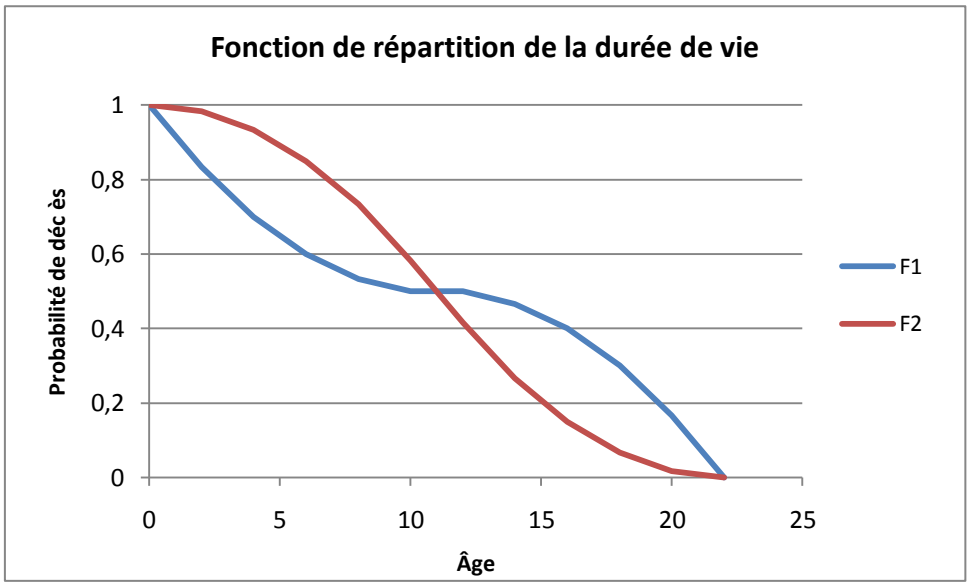

Figure 4 : Fonctions de répartition de la durée de vie

Dans le cas 1, la conclusion est claire : quel que soit l'âge k, on a moins de chances d'atteindre cet âge si l'on est dans la population test que dans la population de référence : la population test est soumise à un excès de mortalité.

Dans le cas 2, c'est l'inverse : quel que soit l'âge k, on a plus de chances de l'atteindre si on est dans la population test que dans la population de référence : la population test a une meilleure longévité.

Dans le cas 3, les choses sont moins simples. Les courbes se croisent en k_0 ; pour les âges inférieurs, la population test est au dessus, ce qui signifie qu'on a plus de chance de dépasser tout âge inférieur si l'on est dans la population test : la mortalité infantile y est plus faible. Par contre, c'est l'inverse pour les âges supérieurs à k_0 : il y a moins de vieillards dans la population test que dans la population de référence.

Le cas 4 est l'opposé : la population test comprend moins de bambins (forte mortalité infantile) mais plus de vieillards que la population de référence.

Bien entendu, des cas beaucoup plus complexes peuvent se rencontrer : la présentation décrite plus haut permet de différencier les populations selon les classes d'âge.

Cette présentation est parfaitement correcte et simple à mettre en œuvre. La seule critique qu'on puisse lui faire est la suivante : elle procède par "tranches d'âge", ce qui n'est pas neutre. Si par exemple on définit une tranche 20 – 40 ans, on compte de la même manière une personne de 21 ans et une personne de 39, ce qui n'est pas forcément légitime.

4. Espérance de vie

L'espérance de vie, souvent utilisée, est la moyenne des valeurs données au Tableau 1. On remplace chaque T_k par sa valeur centrale (le centre de la tranche), notée t_k et on calcule $E = \sum_k p_k t_k$. On procède de même pour la population de référence, avec $E' = \sum_k p'_k t_k$; on peut alors comparer les espérances de vie (à la naissance) pour la population test et la population de référence : une différence significative pourra alerter. Mais nous recommandons de ne pas se limiter au calcul de l'espérance de vie, qui est un indicateur grossier : mettre en évidence les nombres du Tableau 1 est bien préférable, puisqu'alors on comprend les raisons de la différence (zone favorisant la mortalité infantile, ou l'inverse, etc.).

5. Nombre de morts par tranche d'âge

Dans la pratique, déterminer la probabilité, à la naissance, de dépasser tel âge n'est pas facile, surtout pour une population de taille restreinte. On préfère donc compter le nombre de morts par tranche d'âge, ce qui est beaucoup plus facile, mais présente certains dangers.

On présente les résultats sous la forme suivante :

Tranche d'âge	population test	population de référence
T_1	n_1	n'_1
T_2	n_2	n'_2
⋮	⋮	⋮
T_K	n_K	n'_K

Tableau 3 : le nombre de morts par tranche d'âge

où n_k est le nombre de morts dont l'âge est dans la tranche T_k, pour 1 000 habitants dans la population test (et de même pour la population de référence).

Cette phrase, souvent utilisée, n'est pas claire en pratique. Voyons ce qu'elle signifie. Nous prenons toute la population (mettons N personnes), nous prenons une année quelconque, et nous comptons combien de décès, cette année-là, se sont produits dans la tranche T_k, par exemple combien de personnes avaient, au moment de leur décès, entre 20 et 30 ans. Ensuite, ce nombre est rapporté à 1 000 personnes : on le divise par N et on le multiplie par 1 000.

Il y a une différence fondamentale avec la présentation en termes de probabilités de survie : alors que $\sum_{k=1}^{K} p_k = 1$ (tout le monde finit par mourir), on n'a certainement pas $\sum_{k=1}^{K} n_k = 1\,000$ (la population tout entière ne meurt pas au cours d'une année donnée).

Le passage d'une présentation à l'autre est détaillé à l'annexe 1. Il est correct, formellement, si la population est stationnaire, c'est-à-dire si les effectifs de chaque tranche d'âge ne changent pas d'une année sur l'autre. En pratique, cette assertion est fausse : la population vieillit, et il faudra faire des corrections (empiriques) pour passer des décès par tranche d'âge aux probabilités à la naissance.

Mais dans la mesure où ces corrections seront les mêmes pour la population test et la population de référence, nous considérons que la présentation par nombre de décès par tranche d'âge (tableau 3 ci-dessus) est un moyen correct pour comparer les deux populations. A condition bien sûr que le tableau soit complet : il ne faut pas se contenter du nombre de décès pour l'une des tranches d'âge. Et il ne faut pas se contenter du nombre de décès par tranche d'âge pour la population à risque : il le faut aussi pour la population de référence.

Récapitulons :

Règle 4. - *La présentation par nombre de décès par tranche d'âge permet une comparaison acceptable des deux populations, si elles sont stationnaires. Si elles ne le sont pas, il faut s'assurer que les corrections nécessaires sont les mêmes pour les deux. Le résultat final doit impérativement être présenté en termes de probabilités à la naissance.*

Une bonne manière de savoir si les populations sont stationnaires ou non, et de calculer les corrections nécessaires, consiste à établir le tableau 3 pour plusieurs années : si la population est stationnaire, les nombres n_k sont sensiblement constants d'une année sur l'autre.

Remarque : cas des données censurées

Il peut arriver que certaines données de mortalité soient "censurées" : on perd contact avec la personne suivie, et on sait seulement que sa durée de vie est supérieure ou égale à une valeur (on connaît $P\{X \geq k\}$, mais non $P\{X = k\}$, où X est la variable aléatoire indiquant la durée de vie). Les méthodes à utiliser en ce cas sont détaillées dans notre article [BB2].

V. Facteurs externes

La comparaison de deux populations, si elle est faite correctement, peut évidemment montrer des probabilités de survie plus faibles chez l'une que chez l'autre ; encore faut-il savoir à quoi les attribuer. Il est possible, en particulier, que l'une des populations soit soumise à plusieurs risques simultanément ; savoir lequel est la cause principale de la moindre longévité n'est pas simple. En particulier, comme le fait remarquer André Aurengo [Aurengo], le rôle du tabac est essentiel dans certains cancers ; on ne peut mettre en évidence une autre cause (comme les radiations ionisantes) que si on a pris soin de relever l'exposition au tabac.

Il est possible aussi, comme le fait remarquer M. Barescut, qu'un biais existe, du fait de différences de catégories socio-professionnelles entre la population test et la population de référence. Typiquement, si une zone est considérée comme "à risque", ou bien souffrant d'une nuisance quelconque (réelle ou supposée), les loyers y sont plus bas ainsi que le prix des terrains. Les habitants seront alors souvent issus de catégories socio-professionnelles moins favorisées et certaines maladies pourront être plus répandues.

En tout état de cause, pour poursuivre l'étude il faudra se ramener à des sous-populations homogènes vis-à-vis de ces facteurs externes (par exemple : même exposition au tabac dans les deux cas) ; voir chapitre II.

VI. L'effet du hasard

L'effet du simple hasard est bien connu des probabilistes ; il l'est manifestement beaucoup moins des épidémiologistes… Si l'on admet que le cancer est dû au simple hasard, il est normal que toutes les villes de 10 000 habitants n'aient pas le même nombre de cancers : nous allons voir en quelles proportions elles peuvent différer. Cette variabilité due au seul hasard est très importante à prendre en compte : on ne peut affirmer qu'une ville est plus dangereuse qu'une autre (à cause d'un facteur spécifique) que si les différences sont supérieures à ce que le hasard seul donnerait.

1. Evaluations à probabilité fixée

– Un cas d'école

Commençons par un exemple d'école, pour faire comprendre les lois. Si je joue $n = 10\,000$ fois à pile ou face (proba $p = 1/2$), le nombre de "pile" ne sera pas exactement $\frac{n}{2} = 5\,000$. En général (95 fois sur cent), il sera compris dans l'intervalle $I = \left[np - 2\sigma\sqrt{n}, np + 2\sigma\sqrt{n} \right]$. Comme ici $\sigma = \sqrt{p(1-p)} = \frac{1}{2}$ (il s'agit ici, bien sûr, de la variance du jeu de pile ou face), cet intervalle de confiance à 95 % est :

$$I = \left[5000 - 100, 5000 + 100\right] = \left[4900, 5100\right]$$

Que signifie "intervalle de confiance à 95 %" ? Cela signifie que si je répète 100 fois mon "expérience" (laquelle consiste à jouer à pile ou face 10 000 fois), 95 fois sur 100 l'intervalle ci-dessus sera respecté.

On trouvera la théorie détaillée dans le livre [BB1].

Prenons maintenant (arbitrairement) un intervalle plus petit, comme :

$$I' = \left[np - \frac{1}{2}\sigma\sqrt{n}, np + \frac{1}{2}\sigma\sqrt{n} \right]$$

soit ici $\left[5000 - 25, 5000 + 25\right]$ et demandons-nous quelle est la probabilité de tomber dedans. D'après le théorème central limite, elle vaut :

$$\int_{-1/2}^{1/2} \exp(-t^2/2) \frac{dt}{\sqrt{2\pi}} \approx 0.38$$

Autrement dit, si vous répétez 100 fois votre expérience consistant à jouer 10 000 fois, pour 72 d'entre elles vous devez vous attendre à diverger de plus de 25 par rapport à la valeur attendue 5000.

- **Exemple en épidémiologie**

Admettons ([Draper]) un chiffre précis pour p : le taux de leucémies chez l'enfant est 42 par million, par an. Cela nous donne $p = 42 \times 10^{-6}$. Pour une population totale de $N = 400\,000$ personnes (population approximative à moins de 600 m des lignes, selon [Draper]), le nombre "attendu" est $Np \approx 16.8$ et l'intervalle de confiance à 95 % est :

$$\left[Np - 2\sqrt{Np(1-p)}, Np + 2\sqrt{Np(1-p)} \right] = [8, 25]$$

(dispersion de ± 8.6 autour de la valeur centrale).

Comme expliqué plus haut, si on prend l'intervalle plus petit :

$$\left[Np - \frac{1}{2}\sqrt{Np(1-p)}, Np + \frac{1}{2}\sqrt{Np(1-p)} \right] = [14.7, 18.8]$$

on a 72 chances sur 100 de n'être pas dedans : la déviation due au seul hasard est supérieure à ± 2 autour de la valeur centrale attendue 16.8.

Autrement dit, dans le cadre présent (population de 400 000 personnes, risque de 42 par million et par an), seules des déviations supérieures à 8 par rapport à la valeur attendue doivent requérir attention ; des déviations de l'ordre de 2 sont normales et nécessaires.

2. Evaluations à probabilité inconnue

Dans le cas du jeu de pile ou face, la valeur $p = 1/2$ est claire, indiscutable. Mais en épidémiologie, l'évaluation de p (par exemple l'incidence annuelle de la maladie) est difficile en elle-même.

On peut montrer (voir Annexe 2) que la loi de probabilité de p est :

$$f(\lambda) = c\lambda^n (1-\lambda)^{N_{tot}-n}$$

où c est une constante de normalisation, N_{tot} la taille de la population et n le nombre d'accidents (ici de morts).

Il en résulte (voir l'article [BB3]) qu'une estimation du taux de risque est donnée par l'intervalle :

$$I_R = [p - \eta, p + \eta]$$

avec $\eta = \sqrt{\dfrac{p(1-p)}{N_{tot}\varepsilon}}$.

Dans le cas de l'étude [Draper], N_{tot} = 9.7 millions, et $\varepsilon = 0.05$ si nous voulons un intervalle à 95 %. La valeur de η calculée par la formule précédente est de 9×10^{-6} si l'on veut un encadrement à 95 %, et dans ces conditions p peut être n'importe où entre $42-9=33$ et $42+9=51$ par million. La valeur basse conduit à l'encadrement $[5,20]$ et la valeur haute à l'encadrement $[11,29]$, si bien que l'encadrement final sera $[5,29]$, ce qui est très large.

La théorie développée dans le livre [BB1], chapitre 14, permet aussi la comparaison des taux de risque entre deux situations : sachant qu'une zone 1, avec population N_1, a enregistré n_1 décès par an et qu'une zone 2, avec population N_2, a enregistré n_2 décès par an, quelle est la probabilité que la zone 1 soit "plus dangereuse" que la zone 2 ? Voir Annexe 2.

Bien entendu, le mot "plus dangereux" ne fait pas référence à une cause précise : il signifie simplement, comme expliqué plus haut, que l'on y meurt plus souvent, et ce peut être de mort naturelle !

La conclusion générale de ce paragraphe est qu'avec un taux de risque de l'ordre de 40 par million, une population de 400 000 personnes est trop petite pour que l'on voie quoi que ce soit.

VII. Difficultés non prises en compte par la présente analyse

La comparaison de deux populations, du point de vue épidémiologique, comporte bien d'autres difficultés que celles que nous avons relevées :

- Une population n'est pas un ensemble clos : il y a des entrants et des sortants (migrations) ;

- Il est difficile de chiffrer correctement l'exposition à certains dangers. Par exemple, en ce qui concerne les lignes HT, le lieu de résidence est une chose, le temps que l'on y passe en est une autre.

- L'origine des décès n'est pas souvent correctement recensée ; en particulier, concernant la maladie d'Alzheimer, il semble que seulement la moitié des cas le soient. Cette affirmation se trouve à la page 28 dans la synthèse d'une expertise collective de l'INSERM (http://ist.inserm.fr/basisrapports/alzheimer/alzheimer_synthese.pdf)

  L'extrait est le suivant :

  La maladie d'Alzheimer demeure sous-diagnostiquée en France. Selon les données épidémiologiques disponibles, la moitié des patients est aujourd'hui identifiée. Cette insuffisance de diagnostic est liée à plusieurs facteurs, en particulier au fait que nombre de médecins ne sont pas encore convaincus de l'intérêt d'une médicalisation de la maladie d'Alzheimer ni de sa prise en charge thérapeutique. Ce sous-diagnostic est principalement observé chez les patients âgés, mais concerne également les sujets les plus jeunes. Quand le diagnostic est porté, il l'est souvent avec retard. C'est ainsi que le diagnostic de maladie d'Alzheimer n'est aujourd'hui porté qu'au stade de démence avérée.

VIII. Méthodes probabilistes versus tests statistiques

La présentation que nous avons donnée plus haut, sous forme de tableaux, relève des probabilités : nous avons construit, séparément pour la population test et la population de référence, ce qu'on appelle un histogramme puis une fonction de répartition (voir par exemple [BB1]). Cette façon de procéder est simple à mettre en œuvre et ne fait appel à aucune hypothèse complémentaire, à ceci près que la taille des tranches est arbitraire, comme nous l'avons déjà dit (mais on peut la faire varier, si on dispose de données en nombre suffisant).

Les méthodes à base de tests statistiques sont, au contraire, à prendre avec prudence. La statistique est une branche des probabilités, qui concerne les situations où la loi est connue et où l'on dispose d'un échantillon suffisant : aucune de ces deux conditions n'est réunie ici.

Par exemple, on peut, pour anticiper les résultats des élections présidentielles, réaliser un sondage auprès d'un panel bien choisi (par tranche d'âge, par catégorie socio-professionnelle, par région, etc.). Mais la constitution de ce panel a pris des années d'études : il faut connaître la distribution de probabilité de chacun des votes dans chacune des catégories. Nous-mêmes avons rencontré cette difficulté pour constituer un panel de consommateurs, pour Veolia Environnement, Région Ouest : il s'agissait d'anticiper les consommations d'eau potable. Il nous a fallu trois ans pour mettre au point un panel satisfaisant : les extractions aléatoires dans la base de données ne permettaient d'obtenir que des résultats très médiocres.

Règle 5. - *La création d'un panel-témoin est une opération illicite si la loi de probabilité n'est pas connue.*

Venons-en maintenant aux tests statistiques. Un test est une opération (souvent mécanique, c'est-à-dire effectuée par l'ordinateur) qui permet de répondre à une question, du type : deux quantités sont-elles significativement différentes ?

Les tests sont généralement basés sur le Théorème Central Limite (qui évidemment est correct !) qui dit la chose suivante : soit $X_1, X_2, \ldots$ une suite de variables aléatoires, indépendantes et de même loi, d'espérance m et d'écart-type σ ; posons :

$$Y_n = \frac{1}{n}\sum_{j=1}^{n} X_j \, .$$

Alors la variable "réduite" :

$$Z_n = \frac{Y_n - m}{\sigma / \sqrt{n}}$$

se comporte asymptotiquement (lorsque $n \to +\infty$) comme une variable gaussienne, d'espérance nulle et d'écart-type 1.

Ce théorème, pour être applicable, requiert que les variables soient indépendantes et de même loi. Et en sortie, il ne donne qu'une indication asymptotique, valable pour n grand (sans préciser à partir de quand). On trouvera dans [BB1] des exemples d'utilisation impropre de ce théorème.

A partir du Théorème Central Limite, on trouve quantité de tests, destinés par exemple à comparer la moyenne d'une population à une autre, vérifier une linéarité, évaluer une variance, etc. Ces tests sont en général assortis d'un "niveau de confiance", établi comme suit : après modification des variables, on se retrouve avec une gaussienne, dont les propriétés sont supposées connues (par exemple centrée réduite). On dispose par ailleurs d'un nombre, associé

aux résultats de l'expérience, et on vérifie si oui ou non le nombre tombe dans l'intervalle de confiance pour la gaussienne. Par exemple, à 95 %, l'intervalle $[-2,2]$ est approximativement un intervalle de confiance pour une gaussienne centrée réduite, par conséquent si le nombre est -1.32 on répond "oui" et s'il est $+2.01$ on répond "non".

Les tests statistiques sont si nombreux que l'on peut toujours en trouver un qui réponde "oui" à la question posée, quelle que soit la question. Comme disait Laurent Schwartz : "Lorsqu'on veut démontrer quelque chose, on y arrive toujours, même si c'est faux."

Règle 6. - *L'utilisation en aveugle de tests statistiques est formellement à proscrire. Il faut systématiquement s'assurer des conditions de validité du test.*

En particulier, la notion de "taux de confiance" associé à un test, souvent utilisée, est absolument trompeuse : elle n'a de sens qu'à l'intérieur du modèle que l'on a accepté ; si la variable est gaussienne, voici l'intervalle où elle doit se trouver 95 fois sur 100. Mais rien ne dit que la variable soit réellement gaussienne !

Règle 7. - *Toute étude statistique, et en particulier toute étude épidémiologique, dont les résultats sont basés sur des tests statistiques dont la validité n'a pas été vérifiée est absolument sans valeur.*

Cette remarque fort simple a récemment été faite par M. Marc Lavielle à propos des études concernant les OGM : il a observé que les études faites par les deux "camps" étaient sans valeur, car reposant toutes sur des tests statistiques factices ("Le Monde", 12.05.09). En 2007-2008, lors de notre étude pour le CEA concernant les dangers associés aux faibles doses de radiations ionisantes, nous étions parvenus aux mêmes conclusions.

IX. Modèles factices

Il se peut aussi qu'un auteur n'utilise pas un test, mais un modèle, c'est-à-dire une représentation de la réalité, et ce modèle peut être trompeur : il ne correspond pas à la réalité.

C'est le cas ici du "modèle de Cox", très répandu en épidémiologie, et utilisé par [Huss]. Le modèle de Cox [Cox] consiste en la représentation suivante :

On dispose d'une fonction du temps (mettons ici de l'âge), notée $f_0(k)$; elle représentera par exemple le "risque instantané de décès" à l'âge k, dans une population de référence. Ce risque instantané de décès se définit comme la probabilité de mourir à l'âge k, sachant que l'on a dépassé $k-1$; d'autres formulations sont possibles.

On cherche à étudier une population test, et on fait l'hypothèse que, pour cette population-là, le risque instantané sera donné par une formule du type :

$$f_1(k) = f_0(k)\, e^{\beta \cdot Z} \tag{1}$$

où $\beta \cdot Z = \beta_1 Z_1 + \cdots + \beta_n Z_n$ contient tous les paramètres que l'on cherche à prendre en considération (par exemple : distance aux lignes électriques, etc.).

Mais la formule (1) comporte une hypothèse très forte, à savoir que le quotient $\frac{f_0(k)}{f_1(k)}$ est constant au cours du temps (puisque le terme $\exp(\beta \cdot Z)$ ne contient pas k). Si cette hypothèse est réalisée, Cox indique comment calculer les coefficients β. Mais il faudrait s'assurer de manière très stricte que c'est le cas, et pour cela présenter en un tableau les nombres $f_0(k)$ et les nombres $f_1(k)$ et vérifier qu'ils sont bien proportionnels (un simple calcul du coefficient de corrélation suffirait ici). Mais cette vérification n'est jamais faite. Le danger provient du fait que les logiciels permettent dans tous les cas le calcul des coefficients β, même lorsque l'hypothèse de linéarité n'est pas satisfaite : il n'y a pas de mise en garde. Nous donnons en Annexe 3 un exemple simple, tout à fait explicite, où les hypothèses du modèle ne sont pas satisfaites ; le logiciel ne le détecte pas et donne une conclusion fausse !

De manière générale, beaucoup d'études reposent sur un modèle choisi a priori : par exemple on décide que le phénomène que l'on cherche à étudier suit une loi de Poisson, ou une loi exponentielle, une loi de Gumbel, etc. (voir en particulier le document [IRSN]). Mais toutes ces lois sont purement académiques. Un phénomène naturel ne suit jamais une loi de Poisson (en 2007, nous avons réalisé pour le CEA une analyse critique des méthodes statistiques en sismologie : nous avons montré que le décalage observé entre les modèles et la réalité provenait précisément d'une hypothèse factice, selon laquelle l'apparition des séismes suivait une loi de Poisson).

Règle 8. – *Toute étude statistique reposant sur un modèle choisi a priori est absolument sans valeur en ce qui concerne l'aide à la décision. Pour être utilisable dans l'aide à la décision, une étude statistique doit utiliser exclusivement des données brutes, sans aucune hypothèse sur les lois sous-jacentes.*

Bien entendu, il reste tout à fait licite, dans un but d'investigation, de faire toutes les hypothèses que l'on voudra ; notre règle ne porte que sur les études supposées aider à la décision.

Une règle absolue, qui relève de la simple honnêteté intellectuelle, est la suivante :

Règle 9. – *Il ne faut pas aborder une étude avec un présupposé idéologique quant au résultat. Il faut publier absolument tous les résultats, même ceux qui ne vont pas dans le sens que l'on espérait.*

La partie statistique d'une étude devrait être confiée à des statisticiens, compétents évidemment, mais surtout non impliqués dans le résultat. Il faudrait avoir le courage d'anonymiser les données (par exemple parler de maladie sans dire laquelle, parler de zone sans dire laquelle, etc.) de telle sorte que le traitement statistique soit absolument honnête.

Comme le fait remarquer M. Barescut, on peut faire varier les tests, varier les questions, déplacer la limite des zones, etc. : tout ceci est licite à condition de le dire ! Si on a conduit vingt tests différents, dont 19 n'ont rien révélé d'anormal et un seul semble indiquer une déviation par rapport à la normale, il faut avoir le courage de mentionner l'ensemble, et ne pas se contenter de montrer celui qui appuie certaines thèses.

Chapitre II

Les bonnes pratiques statistiques en épidémiologie

Nous récapitulons ici les règles décrites au chapitre précédent et nous présentons leur mise en pratique. Comme on le constatera, les outils mathématiques sont extrêmement simples et bien connus ; les difficultés viennent de la définition des risques, des données, etc.

Typiquement, on dispose d'une population test (taille N_{test}) et d'une population de référence (taille N_{ref}) ; dans la première on observe n_{test} cas d'un "aléa" quelconque (maladie, accident, etc.) et dans la seconde n_{ref} cas de ce même aléa. Peu importe que ce soit dans l'absolu, par seconde, par an, ou ce que l'on voudra (mais les unités doivent être les mêmes pour les deux, bien sûr !).

On forme tout d'abord les deux quotients :

$$q_{test} = \frac{n_{test}}{N_{test}} \text{ et } q_{ref} = \frac{n_{ref}}{N_{ref}} \tag{1}$$

Remarque

Selon la théorie générale (voir Annexe), les quotients "corrects" sont en fait :

$$q_{test} = \frac{n_{test}+1}{N_{test}+2} \text{ et } q_{ref} = \frac{n_{ref}+1}{N_{ref}+2} \tag{2}$$

Dans la pratique, cela ne fait aucune différence, sauf si les nombres n_{test} et/ou n_{ref} sont très petits (quelques unités). Le lecteur voudra bien se souvenir que, en ce cas, les rapports (2) doivent être utilisés.

On regarde si :

$$q_{test} > q_{ref}$$

Si ce n'est pas le cas, la conclusion est claire : il n'y a rien à voir, et il faut avoir le courage de le dire.

Plaçons-nous maintenant dans le cas où l'on a effectivement $q_{test} > q_{ref}$; alors il y a effectivement quelque chose à voir ! On peut déjà (voir Annexe 2) calculer la probabilité que la zone test soit plus dangereuse que la zone de référence, à partir des quatre nombres N_{test}, N_{ref},

n_{test}, n_{ref}. Si cette probabilité est élevée (par exemple ≥ 0.75), cela mérite assurément d'être signalé.

Maintenant, on va rechercher les causes de cette bizarrerie. Pour cela, il faut éliminer les causes naturelles et certaines causes artificielles. Tout ceci se fait au moyen de réductions successives de la population test et de la population de référence.

Par exemple, s'il s'agit d'une maladie (comme Alzheimer) qui frappe surtout les vieux, on s'efforcera de réduire les deux populations pour qu'elles aient la même pyramide des âges.

S'il s'avère que le poids est un facteur influent, on s'efforcera de travailler sur des populations homogènes de ce point de vue (par exemple, pour les deux, la tranche 70-80 kg, ou, pour les deux, la tranche 100-110, etc.).

S'il s'avère que le tabac joue un rôle, on s'efforcera de travailler sur des populations ayant la même exposition : soit qui ne fument pas (des deux côtés), soit qui fument de la même façon des deux côtés (ceci est évidemment très difficile à définir et à mesurer).

Enfin, il ne faut surtout pas oublier que, à la fin, le résultat doit être donné en termes d'espérance de vie : pas forcément à la naissance, mais au-delà d'un âge donné. Par exemple, un résultat du type suivant :

Une personne de 50 ans (avec telles caractéristiques de poids, d'exposition au tabac, etc.) a une espérance de vie de 20 ans en général, mais seulement de 16 ans au voisinage des lignes...

montrerait à l'évidence un danger au voisinage des lignes.

Cette identification des populations "à risque", par restriction successive, est extrêmement intéressante en soi et devrait être pratiquée systématiquement. Pour l'action des médicaments, en particulier, les laboratoires pharmaceutiques se contentent d'identification grossière des populations susceptibles de bénéficier des effets, et ne savent pas correctement identifier des sous-populations spécifiques, susceptibles de ressentir des effets secondaires.

Si l'on veut aller plus loin, et identifier la cause (et pas seulement un effet statistique), il faut s'appuyer sur la physique du phénomène. Pour les lignes HT, une étude purement statistique (comme nous venons d'expliquer) peut mettre en évidence un surcroît de mortalité au voisinage des lignes, mais si l'on veut pouvoir dire "ceci est lié au champ magnétique", alors il faut savoir que celui-ci est lié à l'intensité du courant (voir Annexe 4). Dans ces conditions, il faut se renseigner sur l'intensité des lignes, et constater que, pour une même distance, il y a plus de mortalité au voisinage des lignes à forte intensité qu'au voisinage des lignes à faible intensité. Ceci ne "prouvera" pas que le champ magnétique est responsable (bien d'autres phénomènes peuvent être liés à l'intensité du courant), mais montrera au moins que l'intensité joue un rôle.

Récapitulatif des règles de bonne pratique

Règle 1. – *La population de référence doit systématiquement être aussi vaste que possible. L'extraction aléatoire d'une population de référence représente une faute de logique, puisqu'on a extrait selon une loi de probabilité définie a priori (factice), alors que l'on ne connaît pas la vraie loi.*

Règle 2. - *Dans le cas de maladies liées à l'âge, analyser le nombre de morts ou de cas dans la zone test et dans la population de référence constitue une faute de logique, si l'on n'analyse pas en même temps la pyramide des âges de chaque zone.*

Règle 3. - *La présentation des résultats qui permet la comparaison entre la population test et la population de référence est nécessairement sous la forme d'un tableau, établissant pour chacune la probabilité de parvenir à un âge donné.*

Règle 4. - *La présentation par nombre de décès par tranche d'âge permet une comparaison acceptable des deux populations, si elles sont stationnaires. Si elles ne le sont pas, il faut s'assurer que les corrections nécessaires sont les mêmes pour les deux. Le résultat final doit impérativement être présenté en termes de probabilités à la naissance.*

Règle 5. - *La création d'un panel-témoin est une opération illicite si la loi de probabilité n'est pas connue.*

Règle 6. - *L'utilisation en aveugle de tests statistiques est formellement à proscrire. Il faut systématiquement s'assurer des conditions de validité du test.*

Règle 7. - *Toute étude statistique, et en particulier toute étude épidémiologique, dont les résultats sont basés sur des tests statistiques dont la validité n'a pas été vérifiée est absolument sans valeur.*

Règle 8. – *Toute étude statistique reposant sur un modèle choisi a priori est absolument sans valeur en ce qui concerne l'aide à la décision. Pour être utilisable dans l'aide à la décision, une étude statistique doit utiliser exclusivement des données brutes, sans aucune hypothèse sur les lois sous-jacentes.*

Règle 9. – *Il ne faut pas aborder une étude avec un présupposé idéologique quant au résultat. Il faut publier absolument tous les résultats, même ceux qui ne vont pas dans le sens que l'on espérait.*

Chapitre III

Analyse de l'Etude [Draper]

Cancer infantile en lien avec la distance aux lignes hautes tensions de distribution de l'électricité en Angleterre et au Pays de Galles : une étude cas – témoins.

Gerald Draper, *Directeur de recherche* (*honorary senior research fellow)*1, Tim Vincent, *Chargé de recherche (research officer)*1, Mary E Kroll, *Statisticien (statistician)*, John Swanson, *Conseiller scientifique* (*scientific adviser).*

I. Présentation générale

Le résultat principal de l'étude est : comparé aux enfants qui vivent à plus de 600 m d'une ligne à la naissance, les enfants vivant à moins de 200 m ont un risque relatif de leucémie de 1.69 et ceux entre 200 et 600 un risque relatif de 1.23.

(Le "risque relatif" est défini par rapport à la population de référence : il est mis à 1 pour celle-ci, et un risque de 1.6 signifie qu'il y a aura 1.6 fois plus de leucémies dans la population test que dans la population de référence, pour 1 000 habitants.)

Mais le résultat de l'étude est en contradiction avec les chiffres que celle-ci fournit : comme dit Poincaré, le calcul des probabilités ne devrait pas empêcher d'avoir du bon sens.

L'auteur dit en effet :

"L'incidence de la leucémie chez l'enfant, en Angleterre et au Pays de Galles, est de 42 par million et par an. Nous estimons qu'il y a 9.7 millions d'enfants, dont 80 000 à moins de 200 m d'une ligne et 320 000 entre 200 et 600 m. "

Dans ces conditions, le nombre normal de leucémies chez l'enfant, à proximité immédiate des lignes, devrait être, par an :

$$n_1 = 42 \times 10^{-6} \times 80\,000 \approx 3.36$$

et dans la bande 200 – 600 m :

$$n_2 = 42 \times 10^{-6} \times 320\,000 \approx 13.44$$

soit en tout, pour la population à risque : $n_1 + n_2 \approx 16.80$ leucémies par an.

L'étude concerne 33 années (1962 à 1995). On devrait donc s'attendre à un nombre total de leucémies, au voisinage des lignes, de l'ordre de $33 \times (n_1 + n_2) \approx 554$. Or les auteurs ne dénombrent que 322 cas de leucémies au voisinage des lignes (5+19+40+44+61+78+75).

Il y a donc un défaut de bon sens, que l'on peut constater dès le départ. Les auteurs mettent ensuite en œuvre tout un arsenal statistique, que nous allons maintenant analyser, pour obtenir le résultat qu'ils souhaitent ; quand on veut démontrer quelque chose, on y arrive toujours.

Avant d'entrer dans l'analyse statistique, arrêtons-nous un moment sur une remarque de bon sens. Les calculs ci-dessus peuvent avoir logiquement trois conclusions :

- ou bien les lignes HT protègent de la leucémie ;
- ou bien il y a moins d'enfants dans les zones considérées (pyramide des âges différente) ;
- ou bien le recensement n'est pas correct (on n'a pas enregistré tous les cas de leucémies).

De manière générale, dans cette étude, les données ne paraissent pas très fiables. On ne sait pas si les populations étaient les mêmes il y a 10, 20, 30 ans, si les lignes HT étaient en fonction, etc.

II. Analyse statistique de l'étude Draper

L'erreur méthodologique commise par l'étude Draper se trouve dans la définition de la population de référence, obtenue par "appariements" ; nous avons expliqué plus haut (voir Règle 1, chapitre I) en quoi cette approche était erronée. Les "cas-témoin" ne sont pas statistiquement représentatifs de l'ensemble de la population de référence (population générale) ; les auteurs s'en seraient aperçus s'ils avaient fait les remarques de bon sens développées plus haut.

III. Autres erreurs méthodologiques de l'étude Draper

Le choix du domicile à la naissance ne semble pas réellement pertinent pour juger si oui ou non les gens ont habité (et pendant combien de temps ?) au voisinage d'une ligne HT. Il suppose que les sujets sont restés à leur domicile 24 h sur 24 pendant 14 ans. Or, la plupart des enfants passent leur journée à la crèche ou à l'école. Le calcul est donc correct pour l'exposition nocturne mais pas pour la journée. De plus, l'adresse prise en compte est l'adresse de naissance mais rien ne prouve que les sujets vivent leurs 14 premières années à leur domicile de naissance. L'étude [Huss], au contraire, prend soin d'évaluer le temps de résidence.

Pourquoi les enfants adoptés sont-ils éliminés ? Ils sont exposés comme les autres !

Les informations de distance, reposant sur un code postal, sont douteuses. Les auteurs disent "la moitié des enfants atteints de leucémie dans cette étude ont la même adresse de résidence à la naissance et au moment du diagnostic", ce qui signifie que la moitié a changé d'adresse ! Mais n'oublions pas que le taux de risque est seulement de 42 par million : c'est très faible, les nombres seront très petits (quelques unités, comme on l'a vu), et une variation de $\pm 50\,\%$ est considérable.

Revenons aux chiffres donnés par [Draper] ; nous en extrayons le tableau :

distance à la ligne HT(m)	nb de cas de leucémie
0 – 100	24
100 – 200	40
200 – 300	44
300 – 400	61
400 – 500	78
500 – 600	75

Tableau 1 : chiffres de leucémies en fonction de la distance à la ligne HT

Il s'agit du nombre de cas enregistrés sur 33 ans. Les populations concernées ne sont pas mentionnées exactement : ceci est inacceptable, s'agissant d'une publication, car ces chiffres sont évidemment essentiels. Il peut y avoir plus ou moins de leucémies parce qu'il y a plus ou moins de gens !
Pour les deux premières tranches, 0 – 100 m, et 100 – 200 m, les auteurs donnent une estimation totale de 80 000 personnes (pour les deux tranches) et une estimation totale de 320 000 personnes pour la somme des quatre dernières tranches.

Prenons les deux premières tranches : nous avons 64 cas pour 80 000 personnes en 33 ans, soit une incidence par an de 24 pour un million : nous sommes très au-dessous de la moyenne nationale (42 cas par million et par an).

Prenons maintenant la troisième tranche, 200 – 300 m, et attribuons-lui le quart de la population des quatre dernières tranches, soit à nouveau 80 000 personnes. Nous avons 44 cas pour 80 000 personnes en 33 ans, soit une incidence de 17 par million et par an : nous sommes encore très au dessous de la moyenne nationale !

Et on constate, contrairement aux affirmations de l'étude, que ce sont les zones les plus voisines des lignes qui ont le plus faible taux de leucémies.

Prenons la dernière tranche, 500 – 600 m ; le champ généré par les lignes HT (500 A à 500 m) est le même que celui généré par une installation de cuisine (1 A à 1 m). Le taux d'incidence des leucémies de l'enfant est cependant de 28 par million et par an : il reste un effet protecteur des lignes HT (puisque la moyenne est 42 par million et par an), bien que nous soyons très loin !

Concernant les cancers autres que la leucémie, le rapport conclut : "aucun excès de risque en lien avec la proximité des lignes n'a été trouvé pour les autres cancers infantiles". Pourtant, aux abords des lignes, le risque relatif est 0.44 et il augmente significativement lorsqu'on s'éloigne. On peut donc conclure plus vigoureusement, si l'on en croit l'étude : il y a une diminution du risque de tumeurs du système nerveux central et du cerveau avec la proximité à la ligne, ou encore : les lignes THT protègent des tumeurs du système nerveux central et des tumeurs du cerveau.

IV. Que retenir de cette étude ?

Si les chiffres rappelés dans le tableau ci-dessus étaient fiables, on pourrait en déduire que, contrairement à ce qu'annoncent les auteurs, les lignes HT protègent contre la leucémie de l'enfant. Cette hypothèse n'est en rien plus absurde que l'hypothèse contraire, selon laquelle les lignes HT favorisent cette leucémie. Les champs magnétiques et électriques sont utilisés de manière thérapeutique, et rien ne dit que, dans certaines conditions, ils ne puissent avoir un effet bénéfique.

Malheureusement, notre conclusion est ici que rien ne permet de l'affirmer : ces données ne sont pas pertinentes. Nous recensons 24 cas de leucémies, dont le domicile à la naissance est à moins de 100 m des lignes, mais combien de temps ces enfants sont-ils restés exposés ? Le domicile à la naissance n'a été conservé que pour la moitié des sujets, disent les auteurs. A l'inverse, bien des enfants, nés n'importe où, ont pu s'établir au voisinage des lignes ; ceux parmi eux qui ont eu une leucémie ne sont pas recensés. Ne disposant d'aucune évaluation de la population soumise à l'influence des lignes HT, nous ne retenons de cette étude aucune conclusion scientifiquement fondée.

Chapitre IV

Analyse de l'Etude [Huss]

Residence Near Power Lines and Mortality From Neurodegenerative Diseases: Longitudinal Study of the Swiss Population

Anke Huss, Adrian Spoerri, Matthias Egger, and Martin Röösli

American Journal of Epidemiology Advance Access published November 5, 2008

I. Présentation générale

Cette étude concerne la maladie d'Alzheimer et la démence sénile : sont-elles plus fréquentes au sein de la population suisse vivant à proximité des lignes HT ?

Du point de vue de la présentation des données, l'étude est très bien faite (à la différence de [Draper]) : les auteurs prennent soin en effet d'évaluer la population "à risque", en fonction de la durée d'exposition. Cependant, si l'on prend en compte les affirmations de l'INSERM (citées plus haut) selon lesquelles la moitié seulement des cas d'Alzheimer est recensée, disons clairement que cette étude n'aurait jamais dû être lancée. Mais c'est là une réserve qui concerne avant tout l'épidémiologiste. Dans notre critique mathématique, nous ferons comme si les données étaient les bonnes.

Les auteurs concluent à un excès de risque au voisinage des lignes : "Overall, the adjusted hazard ratio for Alzheimer's disease in persons living within 50 m of a 220–380 kV power line was 1.24 (95% confidence interval (CI): 0.80, 1.92) compared with persons who lived at a distance of 600 m or more."

II. Un défaut de bon sens

Malheureusement, les chiffres bruts qu'ils utilisent contredisent leur conclusion ; celle-ci, une fois encore, repose sur une utilisation inappropriée de modèles statistiques (en l'occurrence le modèle de Cox). On ne peut que citer à nouveau Henri Poincaré : il suffit de regarder les chiffres pour avoir du bon sens.

Reproduisons en effet le tableau 2 issu de [Huss] :

Cause of Death	Distance to 220–380 kV Power Line, m	No. Of cases	No. Of Person-Years	Proba
Entire study population				
Alzheimer's disease	0–<50	20	58 423	0,000342
	50–<200	130	363 460	0,000358
	200–<600	572	1 688 323	0,000339
	>= 600	8 506	20 711 618	0,000411
Senile dementia	0–<50	60	58 423	0,001027
	50–<200	371	363 460	0,001021
	200–<600	1 702	1 688 323	0,001008
	>=600	26 155	20 711 618	0,001263
Individuals living at least 15 years at the identical place of residence				
Alzheimer's disease	0–<50	15	22 320	0,000672
	50–<200	63	145 148	0,000434
	200–<600	259	641 017	0,000404
	>= 600	3 861	7 698 419	0,000502
Senile dementia	0–<50	33	22 320	0,001478
	50–<200	169	145 148	0,001164
	200–<600	819	641 017	0,001278
	>= 600	11 930	7 698 419	0,001550

Tableau 1 : Données [Huss]

Nous avons simplement ajouté à droite une colonne "proba" qui est calculée de la manière suivante : c'est le nombre de la colonne 3 (nombre de cas) divisé par le nombre de la colonne 4 (nombre de personnes × années).

C'est en effet la probabilité de mourir de la maladie considérée, sachant que l'on est dans la catégorie correspondante (en termes de distance à la ligne). Pour la première ligne, par exemple (Alzheimer, distance inférieure à 50 m), nous avons 58 423 personnes × années d'exposition, et 20 cas recensés : cela nous fait $\frac{20}{58\,423} = 0{,}000342331$ cas par personne × année d'exposition.

Eh bien, on constate que, pour la maladie d'Alzheimer, toutes ces probabilités, pour toutes les distances à la ligne, sont inférieures à la probabilité de référence (personnes à plus de 600 m), qui est $0{,}000410687$! Il en est de même de la démence sénile.

Si maintenant on considère les populations ayant vécu 15 années à la même place, on constate que pour la démence sénile toutes les probabilités concernant les populations proches des lignes sont inférieures à la probabilité de référence.

Il en est de même pour la maladie d'Alzheimer, pour ces mêmes populations sédentaires, sauf pour celles à proximité immédiate des lignes : ici nous avons une probabilité de $0{,}000672043$

pour les personnes vivant au voisinage immédiat des lignes, et de $0{,}000501532$ pour la population de référence.

Prenons donc une probabilité de référence :

$$p_R = 0{,}0005$$

et voyons à quoi on peut s'attendre pour une population à risque constituée de $N = 22\,320$ personnes. L'intervalle de confiance à 95 %, vu plus haut, nous donne :

$$I = \left[Np_R - 2\sqrt{Np_R(1-p_R)}, Np_R - 2\sqrt{Np_R(1-p_R)} \right]$$

soit avec les valeurs présentes :

$$I = \left[4.5, 17.88\right]$$

La valeur observée, ici 15, est compatible avec cet intervalle et peut s'expliquer par le seul fait du hasard. Pour toutes les autres distances, la probabilité observée est inférieure à la probabilité de référence.

III. Calcul de la probabilité qu'une zone test soit plus dangereuse que la référence

Comme nous l'avons expliqué, les méthodes probabilistes développées dans [BB1] permettent de répondre à la question suivante : étant donnés une population et un nombre d'accidents dans une zone test et une zone de référence, quelle est la probabilité que la zone test soit plus dangereuse que la zone de référence ? Voici les résultats (dernière colonne du tableau) :

Cause of Death	Distance to 220–380 kV Power Line, m	No. Of cases	No. Of Person-Years	Proba	Proba + dgrx que référence
Entire study population					
Alzheimer's disease	0–<50	20	58 423	0,000342	0,24314
	50–<200	130	363 460	0,000358	0,06122
	200–<600	572	1 688 323	0,000339	0,00000
	>= 600	8 506	20 711 618	0,000411	
Senile dementia	0–<50	60	58 423	0,001027	0,05760
	50–<200	371	363 460	0,001021	0,00001
	200–<600	1 702	1 688 323	0,001008	0,00000
	>=600	26 155	20 711 618	0,001263	
Individuals living at least 15 years at the identical place of residence					
Alzheimer's disease	0–<50	15	22 320	0,000672	0,89622
	50–<200	63	145 148	0,000434	0,13872
	200–<600	259	641 017	0,000404	0,00029
	>= 600	3 861	7 698 419	0,000502	
Senile dementia	0–<50	33	22 320	0,001478	0,43720
	50–<200	169	145 148	0,001164	0,00006
	200–<600	819	641 017	0,001278	0,00000
	>= 600	11 930	7 698 419	0,001550	

Tableau 2 : probabilité que la zone test soit plus dangereuse

Dans chacun des quatre cas, la zone de référence est la zone située à plus de 600 m des lignes.

IV. Données globales de l'étude [Huss]

Travaillons maintenant sur données globales, pour Alzheimer, et non plus par tranche de distance. Nous faisons la somme des trois premières lignes du tableau 1. Nous obtenons :

Nombre de cas : 20 + 130 + 572 = 722

Nombre de personnes × années : 58 423 + 363 460 + 1 688 323 = 2 110 206

Si on prend pour référence la zone à plus de 600 m, elle comporte 8 506 cas pour 20 711 618 personnnes × années, soit une probabilité de $4{,}1\times10^{-4}$. Pour 2 110 206 personnes × années, cela devrait nous faire 866 cas, alors que nous n'en comptons que 722 !

V. Erreurs méthodologiques de l'étude [Huss]

L'erreur commise tient à l'utilisation du modèle de Cox, qui réclame la proportionnalité des données, d'une situation à l'autre. Les auteurs affirment avoir testé cette proportionnalité : "We tested our models successfully for the proportionality assumption using Nelson-Aalen survivor functions and statistical tests based on Schoenfeld residuals". Mais manifestement le test de proportionnalité était défecteux !

Nous donnons en Annexe 3 un exemple très simple qui montre que le modèle de Cox, utilisé hors de ses hypothèses, conduit à des conclusions absurdes. Mais dans le cas présenté en Annexe, les données "passent" le test d'utilisation du modèle : le test statistique affirme que Cox peut être utilisé, alors que ce n'est pas le cas !

VI. Que retenir de cette étude ?

Cette étude est faite avec beaucoup de sérieux et les données (nombre de décès et population concernée) sont manifestement recueillies avec grand soin, dans chacun des cas qui sont traités.

Ces données montrent un déficit de mortalité dans les zones à risque, comme nous l'avons vu, et contrairement à ce qu'affirment les auteurs de l'étude. Ce déficit de mortalité, dans la mesure où les informations ont été recueillies avec grand soin, peut être considéré comme acquis.

Peut-on en déduire que les lignes HT protègent contre certaines maladies ? Évidemment non, et nous avons expliqué pourquoi au cours du premier chapitre : qu'il y ait moins de morts, soit, mais ce peut être dû à bien d'autres causes, et en premier lieu, tout simplement, parce que la population est plus jeune !

L'étude [Huss] ne fournit aucune indication sur la pyramide des âges des populations concernées, mais seulement sur l'âge au moment du décès. Nous voyons ici une illustration particulière de ce que nous avons présenté plus haut : il s'agit d'une faute de logique ; les auteurs ont tout simplement oublié que la population, dans son ensemble, n'était pas immortelle.

Références

[Aurengo] André Aurengo : L'épidémiologie environnementale est-elle encore une science? Journée RNI de la SFRP, LaTronche, 7 octobre 2008.

[BB1] Bernard Beauzamy : Méthodes Probabilistes pour l'étude des phénomènes réels. Ouvrage édité et commercialisé par la *Société de Calcul Mathématique SA*, ISBN 2-9521458-0-6. Mars 2004.

[BB2] Bernard Beauzamy : A probabilistic approach for censored data, 2009 (disponible sur le site web de la SCM, www.scmsa.com).

[BB3] Beauzamy, Bernard : The information associated with a sample, May 2009 (disponible sur le site web de la SCM).

[Cox] D. R. Cox : Regression Models and Life-Tables, *Journal of the Royal Statistical Society. Series B (Methodological)*, Vol. 34, No. 2. (1972), pp. 187-220.

[Dreyfus] Arrêt de la Cour de cassation du 12 juillet 1906, Affaire Dreyfus, Présidence de M. Ballot-Beaupré, Premier président.

[INSERM] " Maladie d'Alzheimer : enjeux scientifiques, médicaux et sociétaux". Les éditions de l'Inserm, 2007. Voir : http://ist.inserm.fr/basisrapports/alzheimer/alzheimer_synthese.pdf.

[IRSN] Les études épidémiologiques des leucémies autour des installations nucléaires chez l'enfant et le jeune adulte : revue critique. Rapport DRPH/SRBE n° 2008-001.

[Poincaré] Examen critique des divers systèmes ou études graphologiques auxquels a donné lieu le bordereau. Rapport de MM. Darboux, Appell et Poincaré, 1904.

Annexe 1

Nombre de morts par an et probabilités de vie à la naissance

Dans ce paragraphe, nous faisons le lien entre les deux concepts suivants :

- La probabilité à la naissance, pour une personne, de décéder dans sa k-ème année ;
- Le nombre de décès pour 1 000 habitants, dans une tranche d'âge donnée.

On suppose la population stationnaire. Soit B (birth) le nombre de naissances par an. Les naissances sont comptabilisées au 1er janvier et les décès sont comptabilisés au 31 décembre suivant.

On introduit :

p_k, $k=1,2,...,K$: probabilité pour une personne de décéder au cours de sa k-ème année ;

n_k : nombre de décès dans la k-ème tranche d'âge, pour 1000 habitants.

Ces deux concepts sont liés par la relation suivante :

Proposition. – *Pour tout* $k=1,2,...,K$,

$$n_k = \frac{1000\,p_k}{\sum_{j\geq 1}(1-p_1)...(1-p_j)}$$

et inversement :

$$p_k = \frac{n_k}{\sum_{j=1}^{K} n_j}$$

Démonstration :

Plaçons-nous au 31 décembre, minuit. Le nombre de décès de personnes âgées d'un an est $p_1 B$. Il nous reste donc $(1-p_1)B$ personnes âgées d'un an. C'était déjà le cas au 31 décembre précédent (hypothèse de stationnarité) et, pendant l'année, $p_2 B$ parmi les personnes de deux ans sont mortes. Il nous reste donc $(1-p_1)(1-p_2)B$ personnes âgées de deux ans.

Réitérant ce raisonnement, nous constatons que, au 31 décembre minuit, nous avons $v_k = (1-p_1)\cdots(1-p_k)B$ personnes âgées de k années, $k=1,2,...,K$.

La population totale au 31 décembre minuit est donc :

$$N_{tot} = \sum_{k\geq 1} \nu_k = \sum_{k\geq 1} (1-p_1)\cdots(1-p_k)B \tag{1}$$

Si n_k est le nombre de décès de personnes âgées de k années, pour 1 000 personnes, le nombre de décès dans cette tranche d'âge, parmi toute la population, sera :

$$n'_k = \frac{n_k N_{tot}}{1000} \tag{2}$$

mais ceci, par définition, est égal à $p_k B$. Nous avons donc :

$$n_k = \frac{1000\, p_k B}{N_{tot}} \tag{3}$$

et ceci prouve la première partie de la proposition.

Inversement, on peut calculer les p_k en fonction des n_k. Posons :

$$C = \sum_{k=1}^{K} (1-p_1)\cdots(1-p_k) \tag{4}$$

On a d'après (1) :

$$N_{tot} = CB \tag{5}$$

et :

$$n_k = \frac{1000\, p_k}{C},\ k = 1,...,K \tag{6}$$

et donc :

$$\sum_{k=1}^{K} n_k = \frac{1000}{C} \tag{7}$$

D'où :

$$C = \frac{1000}{\sum_{k=1}^{K} n_k} \tag{8}$$

et :

$$p_k = \frac{n_k C}{1000} \tag{9}$$

et finalement :

$$p_k = \frac{n_k}{\sum_{j=1}^{K} n_j} \qquad (10)$$

comme annoncé.

Il résulte de (5) et (8) que :

$$\sum_{k=1}^{K} n_k = \frac{1000B}{N_{tot}}.$$

Annexe 2

Outils mathématiques pour la comparaison des taux de risque

Prenons une population de taille N au sein de laquelle il se produit n accidents (dans l'absolu, ou par unité de temps). Il s'agit dévaluer la probabilité pour une personne d'avoir un accident : c'est ce que l'on appelle un "taux de risque".

La théorie est connue depuis longtemps ; elle a été développée en particulier par le Laboratoire de Statistiques de l'Université de Paris 6 et par la SCM, dans le cadre d'un contrat européen, sous la direction de Paul Deheuvels (de l'Académie des Sciences). Elle est décrite très en détail dans le livre de B. Beauzamy [BB1] et l'implémentation informatique a été réalisée par la SCM sous la forme du logiciel EvalRisk. Cette implémentation informatique est extrême, ment simple à réaliser si les nombres sont petits ; elle est très délicate lorsqu'ils sont grands.

La théorie dit que si n accidents se sont produits au sein d'une population de taille N, le taux de risque suit une loi de probabilité dont la densité est :

$$f_{n,N}(x) = c\, x^n (1-x)^{N-n}$$

où c est une constante de normalisation, qui vaut précisément :

$$c = \frac{n!(N-n)!}{(N+1)!}$$

Si maintenant nous disposons de deux populations, de tailles respectives N_1 et N_2, et que respectivement n_1 et n_2 accidents se soient produits, les taux de risques pour chacune sont :

$$f_{n_1,N_1}(x) = c_1\, x^{n_1} (1-x)^{N_1-n_1}$$

$$g_{n_2,N_2}(y) = c_2\, y^{n_2} (1-y)^{N_2-n_2}$$

La probabilité que le risque soit plus élevé pour la population 1 que pour la population 2 est :

$$\int\int_{x \geq y} f_{n_1,N_1}(x) g_{n_2,N_2}(y) dxdy$$

Le calcul pratique de cet intégrale est difficile lorsque les nombres sont élevés. Le logiciel "Evalrisk" réalise ce calcul, et a été testé jusqu'à des valeurs égales à 10 milliards, pour chacun des quatre nombres intervenant.

Le test correspondant est très précis. Comme le remarque un commentaire fait sur la version préliminaire du présent rapport, il y a une différence considérable entre 12 et 15 cas d'Alzheimer sur 20 000 personnes (population vivant plus de 15 ans à proximité des lignes). La proba-

bilité que la seconde situation soit plus dangereuse que la première est 0.71 : c'est tout à fait significatif. Le problème est que l'on ne sait pas exactement si le nombre d'Alzheimer est 12, 15, ou autre chose : l'incertitude, comme nous l'avons déjà dit, ne tient pas aux outils mathématiques mais aux données elles-mêmes.

Annexe 3

La physique du problème : lignes HT et champ magnétique

Les auteurs des études statistiques que nous analysons ici semblent faire reposer leurs raisonnements uniquement sur les statistiques (qu'ils maîtrisent mal), et non sur la physique, qu'ils ne maîtrisent pas du tout. La question de savoir si les lignes HT sont dangereuses est tout à fait légitime, encore faut-il qu'elle soit abordée avec les connaissances appropriées.

Une ligne électrique crée un champ électrique et un champ magnétique. Les auteurs semblent vouloir se limiter aux effets du champ magnétique. Pour celui-ci, les faits physiques sont les suivants :

L'intensité du champ magnétique (mesurée en Tesla) est proportionnelle à l'intensité du courant dans la ligne (et non à la tension !!) ; pour une ligne rectiligne de grandes dimensions (cas usuel), l'intensité du champ magnétique s'exprime par la formule :

$$B = c\frac{I}{d} \tag{1}$$

où c est une constante, I l'intensité du courant dans la ligne (en Ampères) et d la distance à la ligne (en mètres).

La formule est démontrée plus bas.

Par conséquent, si vous êtes à 100 m d'une ligne transportant 500 Ampères, vous recevez le même champ magnétique que si vous êtes à 1 m du câble alimentant une cuisinière (5 Ampères). La fréquence des deux champs est la même (50 Hz).

Dans l'étude [Draper], diverses hypothèses sont faites en ce qui concerne la dépendance du champ par rapport à la distance : on y rencontre $1/d$, $1/d^2$, $1/d^3$; cette ignorance du phénomène physique en cause est tout de même étonnante, car enfin si l'effet était en $1/d^3$, à 100 m il n'en reste plus que le millionième !

Démontrons la formule (1). Elle est claire intuitivement, car dans le cas d'un conducteur rectiligne infini (ou simplement de grande longueur par rapport à la distance à l'observateur) le champ a nécessairement une symétrie cylindrique. La quantité reçue dans une portion de l'espace à distance r et d'épaisseur dr est proportionnelle au périmètre du cercle, soit $2\pi r dr$.

Donnons aussi une démonstration complète, issue de la formule de Biot et Savart :

$$B(M) = \frac{\mu_0}{4\pi}\int_C I\frac{\overrightarrow{dl}\wedge\overrightarrow{SM}}{\left\|SM\right\|^3}$$

où $B(M)$ est le champ magnétique en M, μ_0 la perméabilité magnétique du vide, C le conducteur (ici l'axe des x), $\overrightarrow{dl}$ l'élément de longueur sur l'axe Ox, et S le point courant.

Mettant le point M sur l'axe Oy avec l'ordonnée d, nous obtenons :

$$B(M) = \frac{\mu_0 I}{4\pi} \int_{-\infty}^{+\infty} \frac{(x^2 + d^2)^{1/2} \sin(\vartheta) dx}{(x^2 + d^2)^{3/2}}$$

où ϑ désigne l'angle (Ox, SM), et donc $\sin(\vartheta) = \frac{|x|}{\sqrt{x^2 + d^2}}$. Reportant dans l'expression précédente, nous obtenons :

$$B(M) = \frac{\mu_0 I}{2\pi} \int_{0}^{+\infty} \frac{x dx}{(x^2 + d^2)^{3/2}}$$

Le changement de variable $x = d \cdot y$ donne :

$$B(M) = \frac{\mu_0 I}{2\pi d} \int_{0}^{+\infty} \frac{y dy}{(1^2 + y^2)^{3/2}} = \frac{\mu_0 I}{2\pi d},$$

comme annoncé.

Annexe 4

Utilisation inappropriée du modèle de Cox : un exemple simple

I. Description de la population

Prenons deux populations, chacune de 60 personnes, atteintes d'une même maladie et soumises à des traitements différents A et B. Chaque mois, des individus décèdent dans chacune des populations. Après 110 mois, il n'y a plus d'individus en vie.

Durée de traitement (en mois)	Nombre de décès parmi la population qui suit le traitement A	Nombre de décès parmi la population qui suit le traitement B
entre 0 et 10 mois	1	10
entre 10 et 20 mois	3	8
entre 20 et 30 mois	5	6
entre 30 et 40 mois	7	4
entre 40 et 50 mois	9	2
entre 50 et 60 mois	10	0
entre 60 et 70 mois	9	2
entre 70 et 80 mois	7	4
entre 80 et 90 mois	5	6
entre 90 et 100 mois	3	8
entre 100 et 110 mois	1	10

Tableau 1 : Nombre de décès, pour les personnes ayant suivi les traitements A ou B par intervalle de temps

La figure ci-dessous représente le nombre de décès par population en fonction du nombre de mois :

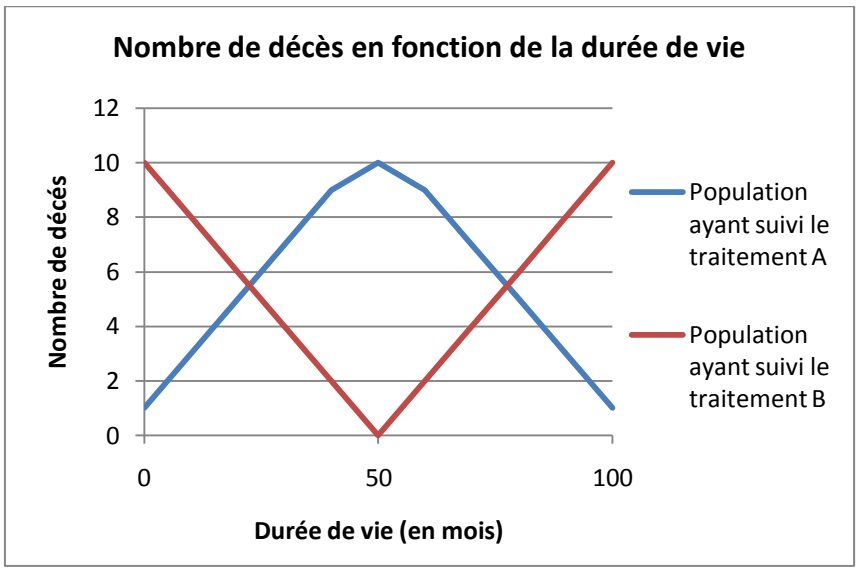

Figure 2 : Loi de probabilité des décès pour la population ayant suivi le traitement A

Les deux figures ci-dessous illustrent les lois de probabilité correspondantes.

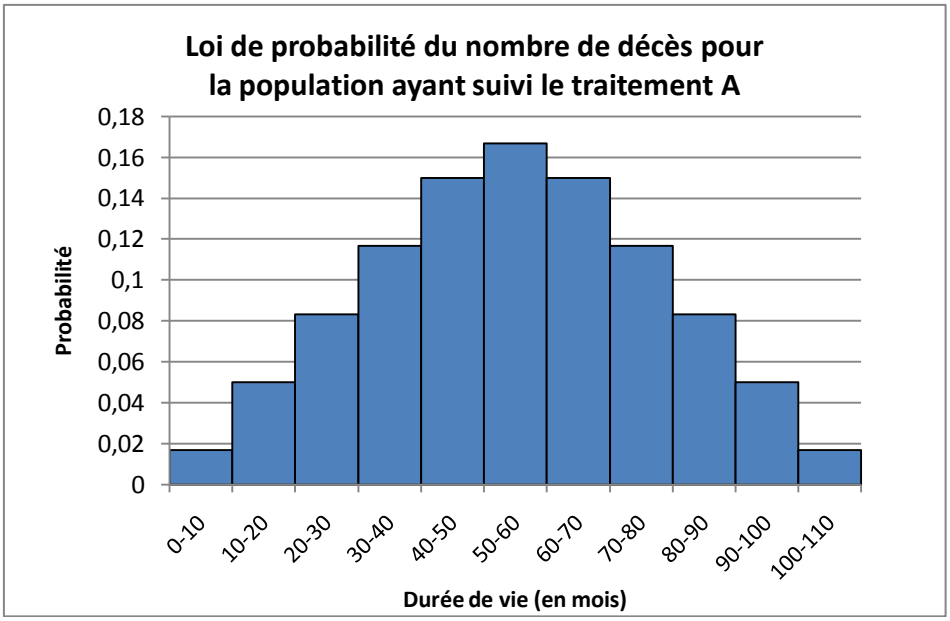

Figure 3 : Loi de probabilité des décès pour la population ayant suivi le traitement A

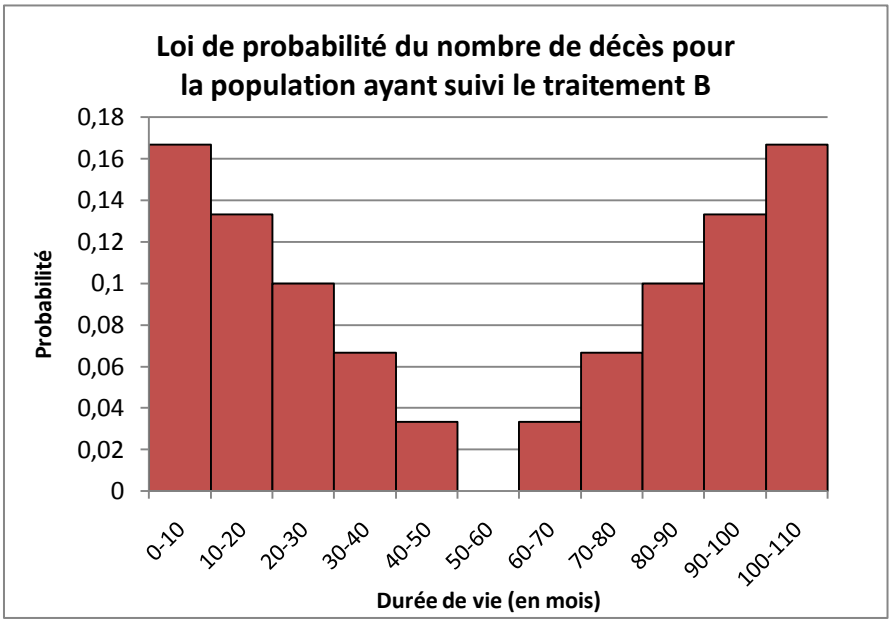

Figure 4 : Loi de probabilité des décès pour la population ayant suivi le traitement B

II. Utilisation du modèle de Cox

On regarde s'il existe une influence du type de traitement sur la durée de vie des individus. Pour cela, on regarde l'influence du type de traitement B par rapport au type de traitement A. Ainsi, les résultats fournis par le modèle correspondent au traitement B.

On utilise le logiciel R, qui est est un langage de programmation et un environnement mathématique utilisé pour le traitement de données et l'analyse statistique. Il permet notamment de simuler le modèle de Cox.

On fournit en entrée du logiciel trois vecteurs de taille 120 (une valeur par individu) :

- La durée de vie : 0, 0 …10, 10, 10…20, 20, 20… 100, 100, 100 ;
- Le type de traitement suivi par l'individu : on affecte 1 si l'individu a suivi le traitement B et 0 si l'individu a suivi le traitement A ;
- L'indicateur de décès : on affecte 1 à un individu mort et 0 à un individu perdu de vue. Dans notre cas, tous les individus sont morts.

En appliquant le modèle de Cox aux données, le logiciel R donnes les résultats suivants :

e^{β}	Intervalle de confiance	Degré de signification du test
0.66	[0,44 ; 0,96]	0.033 (< 0.05)

Tableau 5 : résultats

Le coefficient e^{β} et son intervalle de confiance sont strictement inférieurs à 1. De plus, le degré de signification du test est bien inférieur à 5%. Ainsi, le modèle de Cox conclut que le traitement *B* a une influence néfaste certaine sur la durée de vie et que, inversement, le traitement *A* a une influence bénéfique sur la durée de vie : mais ceci n'est pas correct.

III. Utilisation des probabilités conditionnelles

Pour évaluer l'impact d'un traitement sur la durée de survie d'un individu, nous utilisons une méthode probabiliste. Cette méthode consiste à tracer la fonction de répartition de la durée de survie dans deux cas de figure : pour la population ayant suivi le traitement *A*, et pour la population ayant suivi le traitement *B*.

La durée de survie est découpée en intervalles. Nous déterminons l'effectif cumulé de chaque intervalle, c'est-à-dire le nombre d'individus dont la durée de survie est supérieure à chaque borne.

En divisant l'effectif cumulé d'un intervalle par l'effectif du cas de figure (traitement *A* ou traitement *B*), nous obtenons le pourcentage d'individus dont la durée de vie dépasse le seuil fixé.

Les courbes ci-dessous sont obtenues en traçant les pourcentages correspondant à chaque intervalle en fonction du type de traitement.

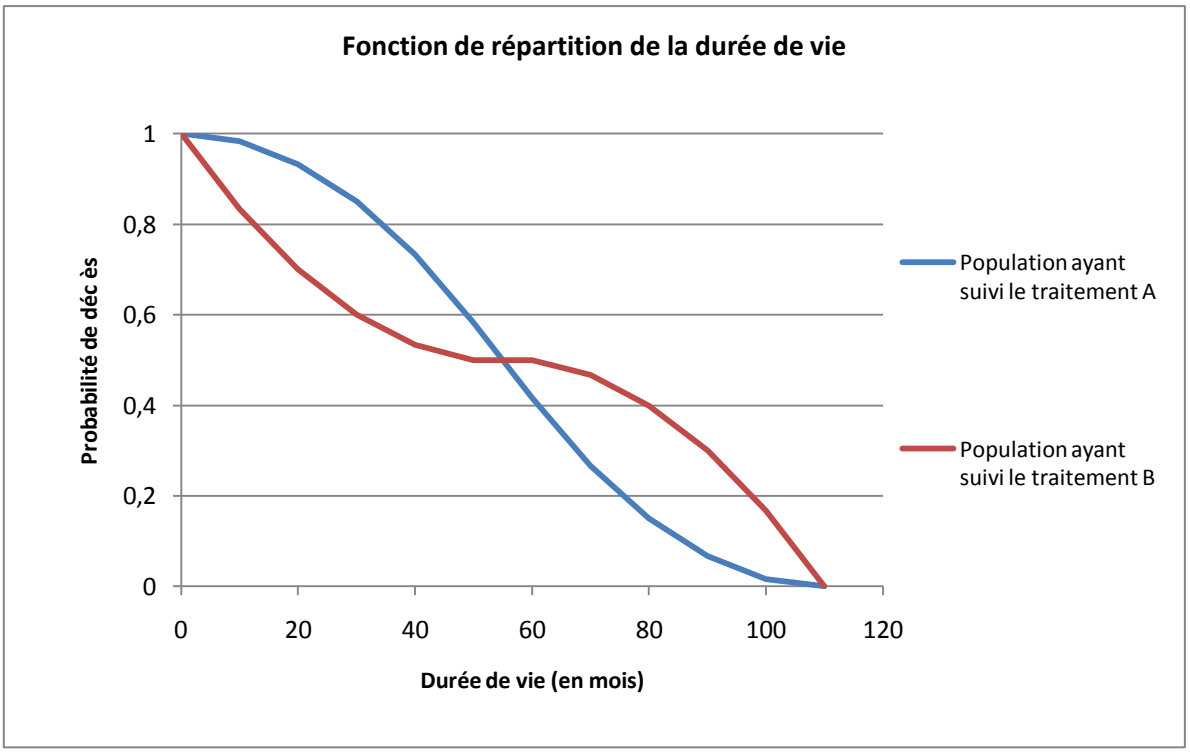

Figure 6 : Fonctions de répartition du de la durée de survie des individus sachant qu'ils ont une fonction rénale normale ou sachant qu'ils ont une fonction rénale anormale

La courbe rouge correspond à la fonction de répartition de la durée de vie pour les individus qui ont suivi le traitement B ; la courbe bleue correspond à la fonction de répartition de la durée de vie pour les individus qui suivi le traitement A.

Considérons une durée de vie de 20 mois, le pourcentage associé à la courbe rouge est 70 %, le pourcentage associé à la courbe bleue est 96 %. Ceci signifie que 96 % des individus ayant suivi le traitement A vivent plus de 20 mois, alors que seulement 70 % des individus ayant suivi le traitement B dépassent cette durée de vie.

Considérons maintenant une durée de vie de 80 mois, le pourcentage associé à la courbe rouge est 40 %, le pourcentage associé à la courbe bleue est 12 %. Ceci signifie que 12 % des individus ayant suivi le traitement A vivent plus de 80 mois, alors que 40 % des individus ayant suivi le traitement B dépassent cette durée de vie.

Ainsi, nous observons deux tendances :

- Le traitement A a un effet protecteur dans un premier temps, puis il a un effet néfaste dans un second temps ;

- Le traitement B a un effet néfaste dans un premier temps, puis il a un effet protecteur dans un second temps ;

IV. Comparaison des résultats des deux méthodes

Pour notre exemple, les résultats des deux méthodes sont différents : la méthode des probabilités conditionnelles montre qu'il y a deux situations distinctes, que le modèle de Cox n'identifie pas.

V. Autre remarque sur le modèle de Cox

Les résultats du modèle de Cox dépendent de l'échelle de temps choisie. Dans l'exemple développé ci-dessus, les données sont exprimées en mois : 0, 10, 20…, 110, 120 et le modèle de Cox conclut à une influence néfaste du traitement B. Si on prend les mêmes données en changeant l'échelle de temps, par exemple en prenant 0, 1, 2…,11, 12, le modèle de Cox ne conclut pas.

Les résultats sont donc différents si on exprime les données en jour, en mois, en années.

En revanche, dans notre modèle utilisant les lois de probabilités conditionnelles, les résultats sont indépendants de l'échelle de temps.

Table des matières