La statistique du chi-deux : son usage dans l'étude britannique de 1950 concernant l'association entre cancer du poumon et tabac.

Jacques Faisant, Denis Lanier, Jean Lejeune,
Rémy Morello et Didier Trotoux(*)

Introduction

En publiant cet article, nous espérons atteindre trois objectifs :

– Faire connaître le remarquable travail des deux auteurs de cette étude, Richard Doll et Austin Bradford Hill [DH50].
– Illustrer concrètement l'usage du test statistique qu'ils ont abondamment utilisé, le test du chi-deux de Karl Pearson.
– Éclairer, par une mise en perspective historique, les fréquentes controverses concernant l'usage de la statistique inférentielle, notamment en médecine.

En effet, récemment, la question de l'emploi de la statistique en cancérologie a rebondi lors de la publication, dans la revue Science, de l'article intitulé « La variation dans les risques de cancer entre différents organes peut être expliquée par le nombre de divisions des cellules souches » [TV15]. La polémique s'est développée à partir d'une phrase de cet article : « Ces résultats suggèrent qu'un tiers seulement de la variation dans les risques de cancer entre différents organes peut être attribuée à des facteurs environnementaux ou des prédispositions héréditaires. La majorité est due à la malchance. »

Ce qui est en jeu ici, c'est l'étude de la relation entre deux événements, à savoir « Contracter le cancer du poumon » et « Être et/ou avoir été fumeur ».

1. Qui étaient Richard Doll et Austin Bradford Hill ?

1.1. Austin Bradford Hill

Sir Austin Bradford Hill (1897-1991), fils d'un physiologiste renommé, se destinait à des études de médecine. Engagé comme pilote dans la Royal Navy en 1916, il contracta la tuberculose. Contre toute attente, il guérit mais il dut abandonner l'idée de travailler dans la médecine. Il s'orienta alors vers la statistique médicale, au *Medical Research Council* (MRC). En 1937, il publia une série d'articles dans *The Lancet* regroupés dans un ouvrage, *Principles of Medical Statistics* qui a connu, jusqu'en 1991, onze rééditions successives. Il fut professeur de statistique médicale à la *London School of Hygiene and Tropical Medicine* de 1945 à 1961 et président

(*) I.R.E.M. de Basse-Normandie.

de la *Royal Statistical Society* de 1950 à 1952. Hill pensait qu'un professionnel de santé devait avoir suffisamment de connaissances en techniques statistiques pour pouvoir planifier des expériences ou pour savoir en interpréter les résultats. Alors en retraite, il énonça en 1965 des conditions nécessaires pour fournir la preuve d'une relation causale entre deux événements. Ces conditions sont appelées critères de causalité de Hill.

1.2. Richard Doll

De mère pianiste concertiste et de père médecin, Sir William Richard Shaboe Doll (1912-2005) naquit aussi au sein d'une famille anglaise aisée. Il entreprit des études de mathématiques, mais il dut les abandonner pour se tourner vers la médecine qu'il étudia au *King's College* de Londres, où il continua à travailler jusqu'à la guerre après l'obtention de son diplôme en 1937. Membre important de la *Socialist Medical Association*, il fut incorporé, à la déclaration de guerre, dans le *Royal Army Medical Corps*. En 1945, il retourna à l'hôpital du *King's College*. En 1948, suite à des incompatibilités de vue avec ses collègues médecins, il collabora avec l'unité de recherche statistique du MRC et y fit la connaissance de Hill.

En 1969, il fut nommé professeur à l'université d'Oxford, et anobli en 1971. Il a été lauréat de la *Royal Medal* de la *Royal Society* en 1986 et du prix Shaw en sciences de la vie et médecine en 2004.

1.3. L'étude de Doll et Hill

L'étude qui nous intéresse et qui a marqué leur carrière a pour origine l'augmentation importante entre les années 1920 et 1930 du taux de cancer du poumon. Deux facteurs retinrent d'abord leur attention, le bitume et les gaz d'échappements des voitures, puis ils centrèrent leur étude sur le tabagisme.

Ce travail, qui marqua l'histoire scientifique médicale du XX^{e} siècle, fut publié dans le *British Medical Journal* en 1950. Après le décès de Hill, Richard Doll continua à travailler sur la mise en évidence de facteurs susceptibles d'être à l'origine de cancers, contribuant ainsi, parallèlement, à renforcer la dimension scientifique de l'épidémiologie.

2. Formalisation de la relation entre deux événements

2.1. Relation logique ?

Les faits dont nous nous occupons sont les suivants :

E_1 : « Être et/ou avoir été fumeur », E_0 : « Ni avoir été fumeur, ni l'être ».

M_1 : « Contracter le cancer du poumon » , M_0 : « Ne pas contracter le cancer du poumon ».

A-t-on $E_1 \Rightarrow M_1$? Heureusement, certaines personnes, qui sont ou qui ont été fumeurs, n'ont pas le cancer du poumon. Cette implication est donc fausse.

A-t-on $E_0 \Rightarrow M_0$? Malheureusement, il existe des personnes qui n'ont jamais fumé, mais qui ont le cancer du poumon. Cette implication est donc fausse. Considérons ces événements comme aléatoires. Il est alors possible de modéliser la relation en termes non plus logiques, mais probabilistes.

2.2. Le principe des tests d'hypothèses

Dans une population, on étudie une ou des caractéristiques numériques qui sont un ou des paramètres d'une variable aléatoire. On definit une hypothèse les concernant, appelée hypothèse nulle et notée H_0 et une hypothèse contradictoire H_1, puis on recueille, sur un échantillon d'individus de la population tirés au sort de manière indépendante, les données correspondant à la variable aléatoire.

Pour déterminer si ces données permttent ou non d'affirmer raisonnablement que H_1 est vraie, on utilise une statistique, c'est-à-dire une variable aléatoire basée sur les données recueillies, dont la loi de probabilité est connue dans le cas où H_0 serait vraie et dont la valeur apporte, si possible, des renseignements optimaux sur H_0 et H_1. Cela nous permet d'instaurer une règle de décision, qui peut conduire à deux types d'erreurs : rejeter H_0 alors qu'elle est vraie, ou erreur de première espèce ; ne pas rejeter H_0 alors qu'elle est fausse, ou erreur de deuxième espèce. On appelle risque de première espèce d'un test et on note α la probabilité, calculée en considérant H_0 comme vraie, que la règle de décision ci-dessus conduise cependant à rejeter H_0. D'après les propriétés de la statistique ci-dessus, il est possible de choisir α et on prend souvent $\alpha = 5\%$.

Pour compléter cette brève description, voir le *Bulletin Vert* nº 500, p. 482-487.

2.3 Deux types d'études médicales

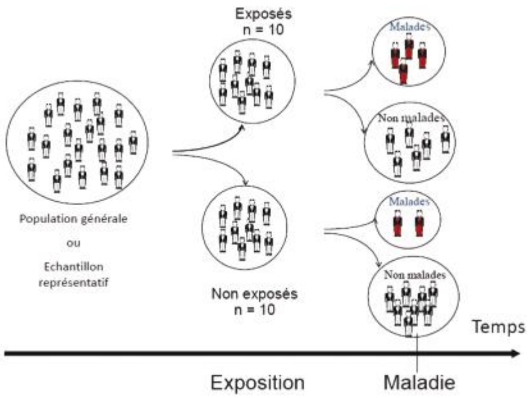
L'étude de cohorte

Il paraît naturel d'effectuer une étude prospective, c'est-à-dire

– de choisir aléatoirement deux groupes de sujets n'ayant pas développé la maladie, un premier groupe constitué de personnes exposées au facteur de risque (ici, le tabagisme) et un deuxième groupe, constitué de personnes non exposées à ce facteur de risque
– et, après un délai adapté, de déterminer le devenir de ces sujets quant au fait de contracter ou non la maladie.

C'est cette étude qui est appelée étude de cohorte.

La démarche de ce type d'enquête peut être illustrée par le schéma suivant :



Dans ce cas, E_0 et E_1 sont connus car des sujets ont été choisis parmi ceux qui sont exposés et d'autres parmi ceux qui n'ont pas été exposés, mais M_0 et M_1 sont eux des événements aléatoires.

Le problème consiste à comparer les probabilités inconnues suivantes :

$p_1 = P_{E_1}(M_1)$ = probabilité pour un fumeur de contracter le cancer du poumon, appelée risque r_1 par les épidémiologistes.

$p_0 = P_{E_0}(M_1)$ = probabilité pour un non-fumeur de contracter le cancer du poumon, appelée risque r_0 par les épidémiologistes.

Si $p_0 = p_1$, E_1 n'est pas un facteur de risque.

Tableau décrivant les caractéristiques d'une étude de cohorte :

	Malades	Non malades	Total
Exposés	a	b	L_1 (fixé)
Non exposés	c	d	L_0 (fixé)
Total	$a+c$	$b+d$	T (fixé)

Il est important de remarquer que les nombres L_0, L_1 et T sont déterminés à l'avance : ils n'ont pas de caractère aléatoire. Seules les données recueillies, a, c, b et d ont ce caractère, b et d n'apportant cependant aucune information supplémentaire par rapport à a et c.

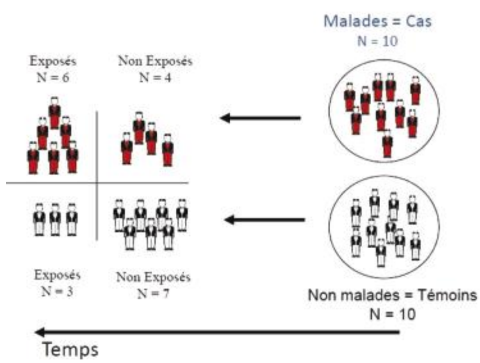
L'étude cas-témoins.

La durée d'une étude de cohorte pouvant être importante lorsque la maladie étudiée se déclare tardivement, cette étude est souvent longue et coûteuse. A contrario, une étude rétrospective consiste

– premièrement à déterminer un groupe de sujets malades et un groupe de sujets non malades,
– puis à les interroger pour savoir si, auparavant, ils ont été exposés ou non au facteur de risque.

Un telle étude est généralement appelée étude cas-témoins et c'est cette forme d'étude que Doll et Hill ont réalisée en 1950.

La démarche de ce type d'enquête peut être illustrée par le schéma suivant :



Ici, M_0 et M_1 sont connus d'après la détermination des groupes mais E_0 et E_1 sont, eux, des événements aléatoires. Il s'agit donc de comparer les probabilités inconnues :

$p'_1 = P_{M_1}(E_1)$ = probabilité, pour quelqu'un atteint du cancer du poumon, d'être fumeur.

$p'_0 = P_{M_0}(E_1)$ = probabilité, pour quelqu'un qui n'a pas le cancer du poumon, d'être fumeur.

Le problème consiste à comparer p'_1 et p'_0.

Tableau décrivant les caractéristiques d'une étude cas-témoins :

	Cas	Témoins	Total
Exposés	a	b	$a+b$
Non exposés	c	d	$c+d$
Total	C_1 (fixé)	C_0 (fixé)	T (fixé)

Les éléments aléatoires sont toujours a, c, b et d ; la différence essentielle entre les deux types d'étude réside dans le fait que, pour l'étude cas-témoins, dans chaque colonne, la somme des résultats est fixée d'avance alors que pour l'étude de cohorte, c'est dans chaque ligne que la somme des résultats est fixée d'avance.

3. La comparaison de deux proportions

3.1. Un exemple fictif d'étude de cohorte

Exemple fictif	Malades	Non malades	Total
Exposés	$a = 125$	$b = 375$	$L_1 = 500$ (fixé)
Non exposés	$c = 100$	$d = 400$	$L_0 = 500$ (fixé)
Total	$a + c = 225$	$b + d = 775$	T=1000 (fixé)

Appliquée à cet ensemble de données, la méthode numérique, popularisée à partir de 1823 par le médecin Pierre Louis (1787-1872) [Lou23], aurait simplement consisté à comparer les proportions observées $f_1 = \dfrac{a}{L_1}$ et $f_0 = \dfrac{c}{L_0}$.

Puisque $f_1 = \dfrac{a}{L_1} = \dfrac{125}{500} = 0{,}25$ et $f_0 = \dfrac{c}{L_0} = \dfrac{100}{500} = 0{,}20$, la fréquence d'apparition de la maladie chez les exposés est supérieure à la fréquence d'apparition de la maladie chez les non-exposés, Louis en aurait conclu que l'exposition au facteur de risque accroit la probabilité de contracter la maladie.

Jules Gavarret(1809-1890), dans son traité [Gav40, p. 145-146], a objecté qu'il fallait en réalité comparer la quantité $d = |f_1 - f_0|$, appelée, de nos jours, variation absolue des risques, avec l, la « limite compatible avec l'invariabilité des causes », qu'il obtenait à l'aide d'une formule due à Poisson. Si la valeur absolue de la différence est supérieure à l, on en conclut que l'exposition au facteur de risque modifie la probabilité de contracter la maladie et, sinon, « tout porte à croire » que l'exposition au facteur de risque ne modifie pas la probabilité de contracter la maladie.

Donc, de notre point de vue, ce médecin est à l'origine du principe des tests d'hypothèses que nous avons rappelé au § 2.2.

3.2. Le test du chi-deux de Pearson : une façon d'aborder le problème de comparaison de proportions.

Les conclusions de l'article de 1950 de Doll et Hill s'appuyant la plupart du temps sur un test du chi-deux, décrivons ici ce test statistique.

L'origine de ce test

C'est Karl Pearson qui le premier, proposa en 1900 [Pea00], un test statistique d'ajustement d'une distribution observée à une distribution attendue, ceci lors de sa collaboration avec un zoologiste qui recherchait une démonstration de la sélection naturelle à partir de données nombreuses ne suivant pas nécessairement des lois normales.

Fondement du test : à la suite d'une expérience dont on suppose que les résultats sont conformes à une hypothèse probabiliste H_0 concernant la distribution de données, une distribution observée est obtenue. Cette distribution est comparée à celle des données espérées sous l'hypothèse considérée. La statistique de test, dont la valeur s'appelle χ^2_{obs} , ou chi-deux observé, est une mesure de la distance entre les données observées et les données espérées. Pearson a déterminé la loi de probabilité asymptotique de cette statistique dans le cadre de l'hypothèse H_0 : c'est la loi du χ^2 de Pearson à n « degrés de liberté »(1). Il a donné les formules qui permettent de calculer la probabilité qu'une variable aléatoire qui suit cette loi ait une valeur supérieure ou égale à χ^2_{obs} ; comme il s'agit d'une mesure de distance entre les données observées et les données espérées, si cette probabilité est faible (resp. forte), il conclut à la non adéquation (resp. l'adéquation) des données observées avec l'hypothèse probabiliste H_0.

Les effectifs espérés

Il s'agit donc ici de tester, avec le risque $\alpha = 0{,}05$, l'hypothèse $H_0 : p_1 = p_0$ contre l'hypothèse $H_1 : p_1 \neq p_0$.

Rappelons que, si la probabilité de réalisation d'un événement est p et si N expériences indépendantes sont effectuées, l'espérance du nombre de fois où cet événement est réalisé est Np. Traditionnellement, on dit que l'effectif espéré est Np, même si ce nombre n'est pas un entier.

Sous l'hypothèse H_0 étudiée : probabilité d'être malade quand on est exposé = probabilité d'être malade quand on n'est pas exposé, il existe une probabilité inconnue p^* avec $p^* = p_0 = p_1$;

(1) Cette loi n'est autre que celle du carré de la distance à l'origine d'un point M dont les n coordonnées sont des variables normales indépendantes

p^* est estimée par $f^* = \frac{a+c}{\mathrm{T}} = \frac{125+100}{1000} = 0{,}225.$

$1 - p^*$ est estimée par $1 - f^* = \frac{b+d}{\mathrm{T}} = 0{,}775.$

Le test du chi-deux de Pearson utilise les estimations *sous l'hypothèse* H_0 des effectifs espérés (souvent appelés aussi effectifs théoriques ou calculés) :

$$a' = \mathrm{L}_1 \times \frac{a+c}{\mathrm{T}},\, c' = \mathrm{L}_0 \times \frac{a+c}{\mathrm{T}},$$
$$b' = \mathrm{L}_1 \times \frac{b+d}{\mathrm{T}},\, d' = \mathrm{L}_0 \times \frac{b+d}{\mathrm{T}}.$$

Effectifs espérés	Malades	Non malades	Total
Exposés	$a' = 112{,}5$	$b' = 387{,}5$	L_1
Non exposés	$c' = 112{,}5$	$d' = 387{,}5$	L_0
Total	$a' + c' = 225$	$b' + d' = 775$	T

À titre d'exemple, détaillons le calcul de a' :

Considérons que l'hypothèse nulle est vraie : la probabilité d'être malade est la même que l'on soit exposé ou non ; on peut donc estimer cette probabilité p^* par $f^* = \frac{a+c}{\mathrm{T}} = \frac{125+100}{1000} = 0{,}225.$ Donc, pour estimer l'effectif espéré de malades parmi les N = L_1 = 500 exposés, il suffit de multiplier ce nombre par la probabilité p^* d'être malade. On obtient : $\mathrm{N}p^* = \mathrm{L}_1 \times 0{,}225 = 500 \times 0,225 = 112.5$.

Pour tester l'hypothèse que la probabilité d'être malade est la même qu'on soit exposé ou non, l'idée fondamentale est de comparer le tableau des résultats observés avec le tableau des effectifs espérés.

La statistique du chi-deux de Pearson.

À partir des tableaux des effectifs observés et des effectifs espérés, on construit la statistique de test, dont la valeur est donnée par :

$$\chi^2_{\text{obs}} = \frac{(a-a')^2}{a'} + \frac{(b-b')^2}{b'} + \frac{(c-c')^2}{c'} + \frac{(d-d')^2}{d'}.$$

Elle est dite statistique du chi-deux car, dans le cas où H_0 est vraie, elle suit, asymptotiquement, la loi de probabilité du chi-deux de Pearson à un degré de liberté.

Remarque 1 : On peut généraliser la formule ci-dessus dans le cas d'un tableau à r lignes et m colonnes. Mais, sous H_0, la statistique obtenue suit alors, asymptotiquement, la loi de probabilité du chi-deux de Pearson à n degrés de liberté avec $n = (r-1)(m-1)$. Lorsque $r = 2$ et $m = 2$, on obtient effectivement $n = 1$.

Remarque 2 : Ce test n'étant qu'asymptotique, l'usage veut, pour l'appliquer, que la condition suivante soit vérifiée : chacun des effectifs espérés du tableau doit être supérieur ou égal à 5.

Utilisation de la table du chi-deux (Table de Fisher)

À la suite de Pearson, Ronald Fisher (1890-1962) a établi dans son livre *Statistical Methods for Research Workers* [Fis25], une table qui donne les quantiles de la loi du chi-deux pour différentes valeurs du nombre de degrés de liberté. Ci-contre, nous reproduisons une partie de cette table, publiée également dans l'ouvrage de Hill, *Principles of Medical Statistics* [Hil37, p. 308-309].

308 PRINCIPLES OF MEDICAL STATISTICS

TABLE OF χ^2

n	P = ·99	·98	·95	·90	·80	·70
1	·000157	·000628	·00393	·0158	·0642	·148
2	·0201	·0404	·103	·211	·446	·713
3	·115	·185	·352	·584	1·005	1·424
4	·297	·429	·711	1·064	1·649	2·195
5	·554	·752	1·145	1·610	2·343	3·000
■	■	■	■	■	■	■
28	13·565	14·847	16·928	18·939	21·588	23·647
29	14·256	15·574	17·708	19·768	22·475	24·577
30	14·953	16·306	18·493	20·599	23·364	25·508

Exemple de lecture de la table : X étant une variable aléatoire qui suit la loi du chi-deux à n degrés de liberté avec $1 \leq n \leq 30$, la table de Fisher permet de déterminer un encadrement de $P(X \geq x)$ où x est un réel positif : soit, par exemple, une statistique de test X qui suit une loi du chi-deux à $n = 5$ degrés de liberté ; cherchons à encadrer $P = P(Y \geq 9{,}8)$.

Sur la ligne $n = 5$ de la table, on voit que 9,8 est compris entre 9,236 et 11,070 correspondant respectivement aux colonnes P = 0,05 et P = 0,10. Donc $0{,}05 \leq P \leq 0{,}10$, une calculatrice nous donnant P = 0,081.

Remarquons déjà que, dans cette table, la case correspondant à la ligne $n = 1$ et à la colonne P = 0,05 contient 3,841 et que, pour $n = 2$, on lit 5,991.

Réalisation du test sur les données fictives précédentes : Revenons sur notre exemple ; il est possible d'utiliser un tableau pour réaliser le calcul de χ^2_{obs} :

Calcul de $\frac{(a-a')^2}{a'}\cdots$	Malades	Non malades	Total
Exposés	1,389	0,483	1,792
Non exposés	1,389	0,483	1,792
Total	2,778	0,806	3,584

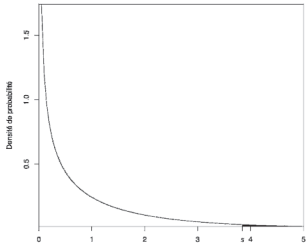
Mais, pour calculer le chi-deux dans le cas d'un tableau 2 × 2 on peut aussi utiliser une formule simple donnée par Hill dans son livre :

$$\chi^2_{\text{obs}} = \frac{(ad-bc)^2(a+b+c+d)}{(a+b)(c+d)(a+c)(b+d)}$$

$$= \frac{(125\cdot 400 - 375\cdot 100)^2(125+375+100+400)}{(125+375)(100+400)(125+100)(375+400)} \simeq 3{,}584.$$

Les deux méthodes de calcul nous ont donné le résultat $\chi^2_{\text{obs}} \simeq 3{,}584$.

Test au risque $\alpha = 0{,}05$ de l'hypothèse $H_0 : p_1 = p_0$, selon laquelle l'exposition ne modifie pas la probabilité de contracter la maladie, contre l'hypothèse contraire H_1.



La règle de décision consiste à rejeter H_0 si et seulement si χ^2_{obs} appartient à l'intervalle $[s\,;+\infty[$ où s est tel que $P(X > s) = 0{,}05$, X étant une variable aléatoire qui suit la loi du chi-deux à n = 1 degré de liberté. Cela donne $s \cong 3{,}841$.

On peut également donner cette règle de décision sous la forme équivalente : $P \leq 0{,}05$ où $P = P(X > \chi^2_{\text{obs}})$. Ce nombre P est appelé le degré de signification du test (ou, en anglais, *p-value*).

Puisque $3{,}584 < 3{,}841$, H_0 n'est pas rejetée : les données ne permettent pas de conclure que le fait d'être exposé au facteur de risque modifie la probabilité d'être malade. On ne peut pas conclure à une relation entre l'exposition au facteur de risque et le fait de contracter ou non la maladie.

3.3. Le test du chi-deux dans une enquête cas-témoins

Ici, l'hypothèse à tester est différente.

Soit p'_0 (resp. p'_1) la probabilité pour un témoin (resp. un malade) d'avoir été exposé au facteur de risque.

Il s'agit de tester $H_0 : p'_1 = p'_0$ contre $H_1 : p'_1 \neq p'_0$ avec le risque $\alpha = 0;05$.

De manière analogue, H_0 se traduit par le fait qu'il existe une probabilité p^* inconnue telle que $p^* = p'_0 = p'_1$.

On obtient :

$$a' = C_1 \times \frac{a+b}{T}, b' = C_0 \times \frac{a+b}{T}, c' = C_1 \times \frac{c+d}{T} \text{ et } d' = C_0 \times \frac{c+d}{T}.$$

Et le chi-deux observé est égal à :

$$\chi^2_{\text{obs}} = \frac{(a-a')^2}{a'} + \frac{(b-b')^2}{b'} + \frac{(c-c')^2}{c'} + \frac{(d-d')^2}{d'}.$$

On rejette $H_0 : p'_1 = p'_0$ si et seulement si $\chi^2_{\text{obs}} > 3{,}841$.

Remarque : si les données de l'exemple provenaient d'une étude de cohorte, mathématiquement, les rôles de b' et c' auraient été respectivement joués par c et b. Mais on aurait obtenu exactement le même résultat pour χ^2_{obs}. En effet, la formule de Hill est symétrique en b et c.

4 La première étude à grande échelle de l'influence d'un facteur de risque

4.1 Pourquoi se poser la question : le tabac est-il un facteur de risque pour le cancer du poumon ?

L'étude de Doll et Hill [DH50], publiée dans le *British Medical Journal* le 30 septembre 1950, a été motivée par la recherche d'une explication de l'accroissement, en Angleterre et au Pays de Galles, du taux de mortalité dû au cancer du poumon, entre 1922 et 1947. Le nombre annuel de décès est effectivement passé de 612 à 9287. Cet accroissement de quinze fois était sans commune mesure avec l'accroissement de la population.

Deux causes étaient généralement mises en avant :

– la pollution atmosphérique générale provenant des échappements des automobiles, des poussières de revêtements des routes, des usines à gaz, des usines industrielles et des centrales au charbon,
– l'augmentation de la consommation de tabac.

Le fait que les femmes restaient relativement protégées de l'augmentation de la prévalence de ce cancer était un argument important accusant le tabagisme.

4.2 Doll et Hill n'étaient pas les premiers à s'intéresser à la relation entre la consommation de tabac et le cancer du poumon.

Franz H. Müller, médecin allemand, avait comparé en 1939, 86 cas de cancer du poumon masculins à 86 personnes saines du même âge que les cas atteints de cancer du poumon et avait constaté que les victimes du cancer du poumon avaient une probabilité six fois plus élevée d'être d'« extrêmement gros fumeurs ».

Ernest L. Wynder, étudiant à l'Université de Washington et Evarts A. Graham son professeur, avaient publié en mai 1950 un article, suite à une étude portant sur 605 patients atteints du cancer du poumon et 780 non atteints [WG50].

L'étude de Doll et Hill de 1950 semble, elle, être la première à utiliser le test du chi-deux pour inférer l'existence ou non de l'association tabagisme-cancer du poumon.

4.3 La collecte des données.

Dans 20 hôpitaux londoniens, une enquête a été effectuée sur les malades atteints d'un cancer ayant une des quatre localisations suivantes : poumon, estomac, côlon ou rectum.

Pour chaque malade atteint du cancer du poumon, on a recherché dans le même hôpital ou un hôpital voisin, un patient de même sexe et de même tranche d'âge (parmi cinq), atteint d'un type de pathologie différent.

Finalement, 709 patients atteints du cancer du poumon et 709 autres patients ont été interrogés sur leur usage du tabac, un fumeur étant défini comme ayant fumé au moins une cigarette par jour depuis un an. L'étude est donc typiquement une étude cas-témoins.

4.4 Le principal résultat de l'étude

Il figure dans le tableau IV [DH50, p. 742] où ont été séparées les données observées sur les hommes et sur les femmes.

TABLE IV.—*Proportion of Smokers and Non-smokers in Lung-carcinoma Patients and in Control Patients with Diseases Other Than Cancer*

Disease Group	No. of Non-smokers	No. of Smokers	Probability Test
Males:			
Lung-carcinoma patients (649)	2 (0·3%)	647	P (exact method) = 0·00000064
Control patients with diseases other than cancer (649) ..	27 (4·2%)	622	
Females:			
Lung-carcinoma patients (60)	19 (31·7%)	41	$\chi^2 = 5{\cdot}76$; n = 1 0·01 < P < 0·02
Control patients with diseases other than cancer (60) ..	32 (53·3%)	28	

$$\text{Rappel : } \chi^2_{\text{obs}} = \frac{(ad-bc)^2(a+b+c+d)}{(a+b)(c+d)(a+c)(b+d)}.$$

$$\text{Calcul de } \chi^2_{\text{obs}} \text{ pour les hommes : } \chi^2_{\text{obs}} = \frac{(2\cdot 622 - 27\cdot 647)^2(2+647+27+622)}{(2+647)(27+622)(2+27)(647+622)}.$$

$$\chi^2_{\text{obs}} \cong 22{,}04 > 3{,}84 \text{ ; P} \cong 2{,}7\cdot 10^{-6}.$$

$$\text{Calcul de } \chi^2_{\text{obs}} \text{ pour les femmes : } \chi^2_{\text{obs}} = \frac{(19\cdot 28 - 41\cdot 32)^2(19+41+32+28)}{(19+41)(32+28)(19+32)(41+28)}.$$

$$\chi^2_{\text{obs}} \cong 5{,}76 > 3{,}84 \text{ ; P} \cong 0{,}016.$$

Doll et Hill affirment donc l'existence d'une association entre les faits d'avoir été fumeur et d'être atteint d'un cancer du poumon. Que ce soit pour les hommes ou pour les femmes, la probabilité d'avoir été fumeur quand on est atteint du cancer du poumon, est supérieure à celle d'avoir été fumeur quand on n'en est pas atteint.

4.5 L'approfondissement de l'étude

Doll et Hill se sont préoccupés de la cohérence de leur étude ; par construction, les deux groupes avaient des compositions identiques concernant le sexe et âge. Cependant, ils font remarquer que, par exemple, une différence de distribution des classes sociales entre le groupe des cas, d'une part, et le groupe des témoins, d'autre part, pourrait créer un biais sur les résultats de l'enquête. Le lieu de résidence et la classe sociale (pour les hommes) ont été renseignés et transcrits dans le tableau II ci-contre [DH50, p. 741].

-*Comparison Between Lung-carcinoma Patients and Non-cancer Patients Selected as Controls* TABLE II.

Social Class (Registrar-General's Categories. Men Only)	No. of Lung-carcinoma Patients	No. of Non-cancer Patients
I and II ..	77	87
III	388	396
IV and V ..	184	166
All classes ..	649	649
Place of residence		
County of London	330	377
Outer London	203	231
Other county borough ..	23	16
Urban district	95	54
Rural district ..	43	27
Abroad or in Services ..	15	4
Total (M + F) ..	709	709

Test du chi-deux pour la catégorie sociale (hommes seulement). Pour le calcul de χ^2_{obs} dans le cas de deux lignes et trois colonnes, la formule de Pearson est, avec (2 − 1)(3 − 1) = 2 degrés de liberté :

$$\chi^2_{\text{obs}} = \frac{(a-a')^2}{a'} + \frac{(b-b')^2}{b'} + \frac{(c-c')^2}{c'} + \frac{(d-d')^2}{d'} + \frac{(e-e')^2}{e'} + \frac{(f-f')^2}{f'}$$

avec $a' = \mathrm{L}_1 \times \dfrac{a+d}{\mathrm{T}}, d' = \mathrm{L}_0 \times \dfrac{a+d}{\mathrm{T}}$, etc.

Effectifs espérés :

Catégorie sociale	I et II	III	IV et V	Total
Cas	$a' = 82$	$b' = 392$	$c' = 175$	649
Témoins	$d' = 82$	$e' = 392$	$f' = 175$	649
Total	164	784	350	1298

$$\chi^2_{\text{obs}} = 2 \cdot \frac{5^2}{82} + 2 \cdot \frac{4^2}{392} + 2 \cdot \frac{9^2}{175} \cong 1,62 < 5,99 \text{ ; } \mathrm{P} \cong 0,445.$$

4.6. Conclusion des auteurs

Après avoir effectué, dans le même but, un grand nombre de tests du chi-deux de ce type, ils ont pu écrire : « En résumé, de notre point de vue, il n'est pas raisonnable d'attribuer les résultats à une sélection spéciale des cas ou à un biais dans la transmission des données. Autrement dit, il peut être conclu qu'il existe une réelle association entre le cancer du poumon et le fait de fumer ». [DH50, p. 746](traduit par nos soins).

5. L'influence de l'étude, ses détracteurs, ses prolongements.

Elle a fini par avoir un impact très important, aussi bien auprès des médecins, biologistes et statisticiens qu'auprès de l'opinion publique et des décideurs en matière de santé publique.

Cependant les détracteurs ont été nombreux et puissants.

Parmi les arguments des détracteurs, citons par exemple :

– l'existence possible d'un « tiers-facteur » : la pollution plus importante dans les villes ouvrières où résident les « plus gros fumeurs » (W. Hueper),
– l'existence d'un gène de prédisposition (R. A. Fisher),
– la constitution de l'individu qui lui permettrait ou non de résister aussi bien aux « tentations » qu'aux maladies (J. Berkson),
– le rôle du papier à cigarette (Clarence Cook Little).

Concernant C. C. Little, il est utile de savoir qu'il a dirigé, de 1954 à 1969, le *Tobacco Industry Research Committee*, organisme « contre-feu » fondé, en 1953, notamment par le président de l'*American Tobacco Company*. Voir [Pro14].

Doll et Hill ont alors prolongé leur article par une étude de cohorte de grande ampleur auprès de médecins en Grande-Bretagne, la « *British Doctors Study* », dont les rapports de 1954, 1956, 1964 et suivants vont dans le même sens que celui de 1950, ainsi que les études conduites aux États-Unis par les statisticiens C. Hammond et D. Horn portant sur 187 783 hommes, et par le sociologue H. Dorn portants sur 200 000 vétérans.

6. Conclusion

Cet article met en lumière deux aspects qui nous semblent importants. D'abord, le succès du travail de Doll et Hill n'est pas simplement dû au résultat obtenu. Il l'est aussi par l'extraordinaire luxe de précautions prises par les auteurs pour essayer d'éliminer tous les biais et par l'utilisation intensive d'un outil mathématique, le test du chi-deux. C'est donc l'outil statistique qui apporte ici, effectivement, la rigueur et l'objectivité et pas seulement leur apparence.

D'autre part, il est important de noter que Doll et Hill, ainsi que leurs successeurs, ne parlent pas de causalité mais d'association. On dirait aujourd'hui de corrélation. L'étude statistique permet de quantifier des relations, des augmentations (ou diminutions) des risques. Mais le risque reste toujours du domaine de la probabilité. Même si on évite en épidémiologie de parler de nombre de chances (d'avoir une maladie !) ou de cas favorables (au décès d'un malade !), il reste qu'on parle toujours en termes de probabilité, donc de hasard. La statistique médicale n'a pas pour objet d'éliminer toute forme de hasard pour découvrir les vraies causes (sociales, environnementales, industrielles, etc.) d'une maladie, mais d'encadrer ce hasard pour mieux le comprendre et éventuellement le conjurer. Que des scientifiques et des journalistes spécialisés tombent encore dans ces travers, comme on l'a vu en introduction, cela montre que nous avons beaucoup à faire dans l'enseignement de ces notions.

Références

[DH50] Richard DOLL & Austin B. HILL. Smoking and Carcinoma of the Lung. *British Medical Journal*, 1950.

[Fis25] Ronald A. FISHER. *Statistical Methods for Research Workers*. Oliver & Boyd, Edinburgh, 1925.

[Gav40] Jules GAVARRET. *Principes généraux de Statistique médicale ou développement des règles qui doivent présider à son emploi*. Béchet jeune et Labé, 1840.

[Hil37] Austin B. HILL. *Principles of Medical Statistics*. Oxford University Press, 1937.

[LLT12] Denis LANIER, Jean LEJEUNE et Didier TROTOUX. Statistique inférentielle au fil de l'ouvrage de Jules Gavarret. Irem de Basse-Normandie, 2012. Article en ligne : http ://www.math.unicaen.fr/irem/spip.php ?article117, consulté le 25 mars 2016.

[Lou23] Pierre LOUIS. *Recherches sur les effets de la saignée dans quelques maladies inflammatoires, et sur l'action de l'émétique et des [...]*, chez Jean Baptiste Baillière, Libraire de l'Académie Royale de Médecine, 1823.

[Pea00] Karl PEARSON. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine*, 1900.

[Pro14] Robert N. PROCTOR. *Golden Holocaust - La conspiration des industriels du tabac*, Éditions des Équateurs, 2014.

[TV15] Cristian TOMASETTI & Bert VOGELSTEIN. Variation in cancer risk among tissues can be explained by the number of stem cell divisions. *Science*, Vol. 347, Issue 6217, janvier 2015.

[WG50] Ernst L. WYNDER & Evarts A. GRAHAM. Tobacco Smoking as a Possible Factor in Bronchiogenic Carcinoma. *Journal of the American Medical Association*, 1950.