La statistique du chi-deux : son usage à partir de l'article de Doll et Hill sur l'association entre cancer et tabac.

Jacques Faisant, Denis Lanier, Jean Lejeune, Rémy Morello, Didier Trotoux
IREM de Basse-Normandie
juin 2016

La question de l'introduction de méthodes numériques en médecine est au cœur des préoccupations de notre groupe de recherche StatMed au sein du Cercle d'histoire des Sciences de l'I.R.E.M. de Basse-Normandie. Cet article prolonge notre article de 2012, *Statistique inférentielle au fil de l'ouvrage de Jules Gavarret* [LLT12] où nous analysions l'ouvrage *Principes généraux de Statistique Médicale ou développement des règles qui doivent présider à son emploi* [Gav40]. Cet ouvrage, publié en 1840, nous semblait porter en germe de nombreux outils de la statistique inférentielle développés principalement dans le milieu scientifique anglo-saxon du début du 20^{e} siècle et nous nous étions étonnés de l'absence de successeurs à Gavarret parmi les savants français entre 1850 et 1950, date à laquelle l'école d'épidémiologie française est apparue sous l'impulsion de Daniel Schwartz. Nous nous sommes alors intéressés aux outils utilisés dans les premières enquêtes épidémiologiques de taille importante et particulièrement, dans celle de R. Doll et A. B. Hill, *Smoking and Carcinoma of the Lung* [DH50], consacrée à la liaison entre tabac et cancer. Le travail de notre groupe a commencé en octobre 2012 et la question des relations entre cancer et statistique a rebondi récemment à propos de la polémique engendrée par la publication le 2 janvier 2015 de l'article de Cristian Tomasetti et Bert Vogelstein, *Variations in cancer risk among tissues can be explained by the number of stem divisions* [TV15], dans la revue *Science*.

Le titre de l'étude est assez clair : *La variation dans les risques de cancer entre différents organes peut être expliquée par le nombre de divisions des cellules souches*. La confusion s'est accrue avec une phrase de l'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*." Cette phrase déformée, simplifiée d'abord par les éditeurs de la revue, puis par des journalistes spécialisés est devenue : « deux tiers des cancers sont dus à la malchance, et non à la pollution, aux industries agroalimentaires, chimiques ou cigarettières ».

L'étude est évidemment fondée sur des études statistiques et comme le hasard, la malchance sont invoqués, le rôle des mathématiques en épidémiologie est vite mis en cause.

Par exemple Sandrine Cabut dans le journal *Le Monde* du même jour, intitule son article : *Cancer : le rôle du hasard réévalué* et insiste : " Une chose est sûre, les sciences mathématiques occupent une place croissante en cancérologie ".

En réponse à cet article, la sociologue Annie Thébaud-Mony publie dans *Le Monde* du 7 janvier, l'article *Non, le cancer n'est pas le fruit du hasard !* , où elle remet en cause fortement le rôle des statistiques dans ce genre d'étude : « Dans le champ de l'épidémiologie, des chercheurs s'obstinent à produire des modèles statistiques dénués de sens par rapport à la réalité dramatique du cancer. L'outil mathématique utilisé pour cette production de l'incertitude donne à la démarche l'apparence de la rigueur, de l'objectivité, pour tout dire de la science ».

Bien sûr la polémique ne s'arrête pas là, citons par exemple les titres de deux articles des jours suivants dans *Le Monde* : *Le cancer joue-t-il aux dés ?* d'Édouard Hannezo et *Méfiez-vous du hasard* de Stéphane Foucart.

C'est donc l'étude de la relation entre deux "événements" à savoir le fait d'être malade, ici : « Contracter le cancer du poumon », et l'exposition à un facteur de risque, ici : « Être et/ou avoir été fumeur », qui est en jeu.

Nous nous intéresserons dans un premier temps à la nature possible de cette relation et aux deux grands types d'études qui continuent d'être utilisées pour tenter de la quantifier. Nous rappellerons ensuite les méthodes statistiques de comparaison de proportions, puis, après avoir donné les diverses expressions que peut prendre une statistique dite du chi-deux, nous verrons l'usage qu'en font Doll et Hill et les conclusions tirées de leur étude. Cet article sera complété par quelques rappels sur les prolongements que cet article a eu, les polémiques qu'il a suscitées et par les controverses que provoquent aujourd'hui d'autres études portant sur la santé publique.

1 Formalisation de la relation entre deux événements

1.1 Relation logique ?

Les événements dont il s'agit, sont les suivants :

— E_1 : « Être et/ou avoir été fumeur », E_0 : « Ni avoir été fumeur, ni l'être »
(Exposition ou non au facteur de risque)

— M_1 : « Contracter le cancer du poumon », M_0 : « Ne pas contracter le cancer du poumon »
(Malade ou non)

Ils peuvent être considérés comme des propositions qui sont vraies ou fausses. Dans ce cas, il est légitime de se poser la question de l'implication "logique" entre ces propositions.

— A-t-on $E_1 \Longrightarrow M_1$?
Nous connaissons, heureusement, des personnes qui sont ou qui ont été fumeurs mais qui n'ont pas le cancer du poumon.
Cette implication est donc fausse.

— A-t-on $E_0 \Longrightarrow M_0$?
Nous connaissons, malheureusement, des personnes qui ne sont pas fumeurs et qui ne l'ont jamais été mais qui ont le cancer du poumon.
Cette implication est donc fausse.

Si on considère maintenant ces événements comme aléatoires, il est possible de modéliser la relation en termes probabilistes.

1.2 Les deux types d'étude : l'étude de cohorte et l'étude cas-témoins.

Notations : dans ce chapitre, les nombres qui sont connus mais aléatoires, car issus de l'expérience, sont notés en rouge (idem pour les variables aléatoires) et ceux qui sont certains, mais inconnus car il faudrait connaître toute la population pour les connaître, sont notés en bleu. Ceux qui sont connus et certains sont notés en noir.

Étude de cohorte : Dans le cas de maladies qui n'apparaissent pas immédiatement après que le sujet ait été exposé à un facteur de risque, il paraît naturel d'effectuer une étude **prospective** c'est-à-dire d'examiner d'abord à une date t_0 à la fois des sujets exposés au facteur de risque et des sujets non exposés à ce facteur de risque et ensuite, à une date $t_0 + D$, le devenir de ces sujets quant au fait de contracter ou non la maladie. C'est cette étude qui est appelée **étude de cohorte**.

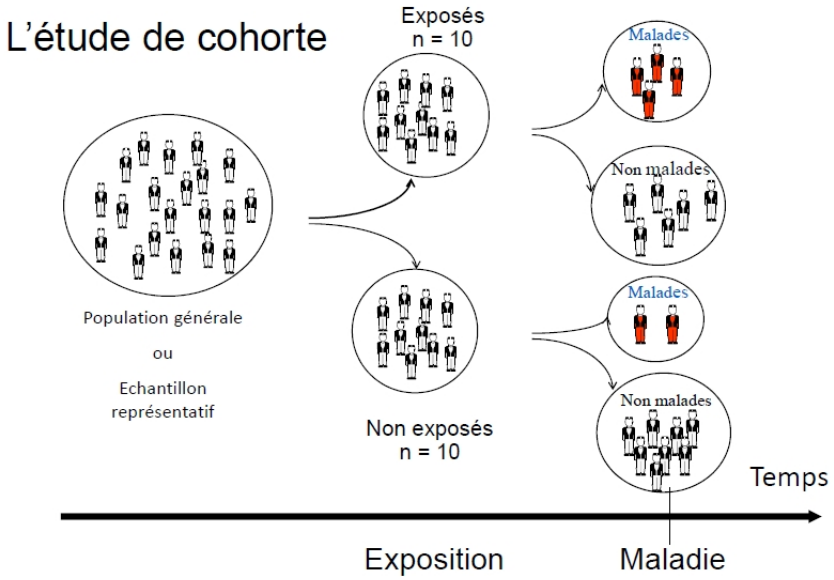
— L_1 sujets se sont trouvés exposés au facteur de risque. Parmi ceux-ci, a ont été malades et b n'ont pas été malades ($a + b = L_1$).

— L_0 sujets ne se sont pas trouvés exposés au facteur de risque. Parmi ceux-ci, c ont été malades et d n'ont pas été malades ($c + d = L_0$).

En tout, $T = L_0 + L_1 = a + b + c + d$ sujets ont été observés.
Les résultats sont transcrits dans le tableau ci-dessous :

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

Le schéma ci-dessous illustre la démarche de ce type d'enquête.



Le tableau ci-dessous condense les résultats :

	Malades	Non malades	Total
Exposés	4	6	10
Non exposés	2	8	10
Total	6	14	20

Dans ce cas, les événements 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 les événements M_1 et M_0 sont eux aléatoires. Il s'agit donc de comparer des probabilités conditionnelles inconnues :

— $p_1 = \mathbb{P}_{E_1}(M_1)$ = probabilité de contracter le cancer du poumon sachant qu'il est exposé fumeur. Cette probabilité conditionnelle est appelée risque r_1 chez les épidémiologistes.
— $p_0 = \mathbb{P}_{E_0}(M_1)$ = probabilité de contracter le cancer du poumon sachant qu'il n'est pas fumeur. Cette probabilité conditionnelle est appelée risque r_0 chez les épidémiologistes.

Le problème : comparer p_0 et p_1.
Si $p_0 = p_1$, E_1 n'est pas un facteur de risque.

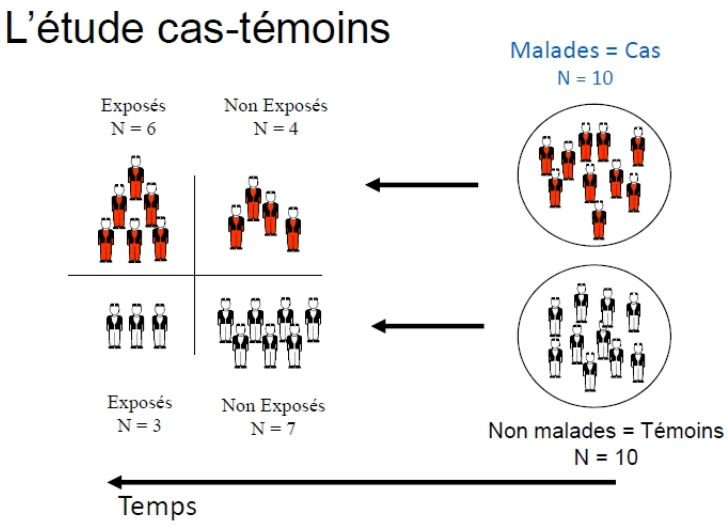
Étude cas-témoins : La durée D d'une étude de cohorte pouvant être importante, cette étude est souvent longue et coûteuse et les épidémiologistes peuvent utiliser une étude **rétrospective** qui consiste à examiner d'abord à une date t_0 à la fois des sujets malades et des sujets non malades, et à les interroger pour savoir si ils ont été exposés ou non au facteur de risque auparavant. C'est cette étude qui est appelée **étude cas-témoins**.

— C_1 sujets malades (cas) ont été examinés. Parmi ceux-ci, a se sont trouvés exposés au facteur de risque et c ne se sont pas trouvés exposés ($a + c = C_1$).
— C_0 sujets non malades (témoins) ont été examinés. Parmi ceux-ci, b se sont trouvés exposés au facteur de risque et d ne se sont pas trouvés exposés au facteur de risque ($b + d = C_0$).

En tout, $T = C_0 + C_1 = a + b + c + d$ sujets ont été observés dans l'étude.

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

Le schéma ci-dessous illustre la démarche de ce type d'enquête.



Le tableau ci-dessous condense les résultats :

	Cas	Témoins	Total
Exposés	6	3	9
Non exposés	4	7	11
	10	10	20

Dans ce cas, les événements M_0 et M_1 sont connus car des sujets ont été choisis parmi ceux qui sont malades et d'autres parmi ceux qui ne sont pas malades, mais les événements E_1 et E_0 sont eux aléatoires. Il s'agit donc de comparer des probabilités conditionnelles inconnues :

— $p'_1 = \mathbb{P}_{M_1}(E_1)$ = probabilité pour quelqu'un qui a le cancer du poumon d'être fumeur
— $p'_0 = \mathbb{P}_{M_0}(E_1)$ = probabilité pour quelqu'un qui n'a pas le cancer du poumon d'être fumeur

Le problème : comparer p'_0 et p'_1.

Même si, dans les deux cas, les résultats a, b, c et d sont aléatoires, la différence essentielle entre les études 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.

2 Rappel sur la comparaison de deux proportions

2.1 L'exemple fictif d'une étude de cohorte portant sur 1000 sujets

	Malades	Non malades	Total
Exposés	125	375	500
Non exposés	100	400	500
Total	225	775	1000

La "méthode numérique" popularisée par Pierre Charles Alexandre Louis [1], si elle avait été appliquée à cet ensemble de données aurait simplement consisté à comparer les proportions observées $f_1 = \frac{a}{L_1}$ et $f_0 = \frac{c}{L_0}$.

Puisque $f_1 = \frac{a}{L_1} = \frac{125}{500} = 0{,}25$, la fréquence d'apparition de la maladie chez les exposés est supérieure à $f_0 = \frac{c}{L_0} = \frac{100}{500} = 0{,}20$, fréquence d'apparition de la maladie chez les non-exposés, on en conclurait que l'exposition au facteur de risque accroit la probabilité de contracter la maladie. Gavarret dans son traité [Gav40, p. 145-146] a objecté que cette comparaison ne suffisait pas et qu'il fallait en réalité comparer la quantité $d = |f_1 - f_0|$ avec "la *limite* compatible avec l'*invariabilité des causes*" notée l. En paraphrasant Gavarret, nous dirions que si la valeur absolue de la *différence* est supérieure à la *limite* calculée, on doit nécessairement en conclure que l'exposition au facteur de risque modifie la probabilité de contracter la maladie et, si la *différence* entre les fréquences est inférieure à la *limite* calculée, "tout porte à croire que" l'exposition au facteur de risque ne modifie pas la probabilité de contracter la maladie.

Gavarret utilise [LLT12, p. 23-24] une formule établie par Poisson pour calculer la limite l qui, ici, est égale à $0{,}07456$. Puisque la différence des fréquences $d = 0{,}25 - 0{,}20 = 0{,}05$ est inférieure à cette limite, "tout porte à croire que" l'exposition au facteur de risque ne modifie pas la probabilité de contracter la maladie.

La différence $f_1 - f_0$ est appelée "réduction absolue des risques" ou aussi "différence des risques" par les épidémiologistes. Dans les logiciels actuels, le test de comparaison de proportions, appelé **test *Z*** , utilise une formule légèrement différente de celle de Poisson [LLT12, p. 25]. Nous indiquons ci-après la formule utilisée actuellement.

Le test Z

Soit p_0 (resp. p_1) la probabilité de contracter la maladie pour les personnes non exposées (resp. exposées) au facteur de risque.

Notons $f^* = \dfrac{a+c}{T}$, la proportion de malades dans les deux échantillons réunis.

Puisque sous l'hypothèse nulle $H_0 : p_1 = p_0$, $Z_{obs} = \dfrac{f_1 - f_0}{\sqrt{\frac{1}{L_1} + \frac{1}{L_0}}\sqrt{f^*(1-f^*)}}$ suit approximativement la loi normale réduite centrée, alors $\mathbb{P}(-z_\alpha < Z_{obs} < z_\alpha) \simeq 1 - \alpha$ ou encore $\mathbb{P}(|Z_{obs}| > z_\alpha) \simeq \alpha$.

En particulier, si l'on prend comme risque de première espèce (c.-à-d. ici le risque de déclarer à tort que l'exposition au facteur de risque modifie la probabilité de contracter la maladie) $\alpha = 0{,}05$ ou encore 5%, alors $z_\alpha = 1{,}96$. On décide de rejeter l'hypothèse nulle si et seulement si $|Z_{obs}| > 1{,}96$, ce qui est équivalent à $|f_1 - f_0| > l$ où $l = z_\alpha \sqrt{\dfrac{1}{L_1} + \dfrac{1}{L_0}}\sqrt{f^*(1-f^*)}$.

En appliquant ce test Z aux données fictives ci-dessus, $|Z_{obs}| = 1{,}8932$ et donc l'hypothèse nulle n'est pas rejetée. Les données ne permettent pas de conclure, que l'exposition au facteur de risque modifie la probabilité d'être malade.

1. P. C. A. Louis (1850-1890) était un médecin qui, le premier, utilisa les pourcentages pour comparer les effets [LLT12, p. 2].

2.2 Une autre façon d'aborder le problème de comparaison de proportions.

C'est par l'étude de Doll et Hill [DH50], publiée le 30 septembre 1950 dans le *British Medical Journal*, que nous aborderons l'étude de l'influence du facteur de risque qu'est le tabagisme sur la probabilité de contracter le cancer du poumon.

Les conclusions de l'article de Doll et Hill étant la plupart du temps appuyées sur la statistique dite du chi-deux, il est utile de rappeler ce qu'est cette statistique.

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

Rappelons que si la probabilité de réalisation d'un événement est p et que 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 que la probabilité d'être malade quand on est exposé est la même que la probabilité d'être malade quand on n'est pas exposé, l'idée fondamentale est de comparer le tableau des résultats observés avec le tableau des effectifs espérés souvent appelés aussi effectifs théoriques ou calculés.

Une difficulté provient du fait que cette hypothèse se traduit par $p_0 = p_1$, donc qu'il existe une probabilité p^* telle que $p^* = p_0 = p_1$, mais que cette probabilité n'est pas connue. Elle sera estimée par $f^* = \dfrac{a+c}{T}$. De même, $1 - p^*$ sera estimée par $1 - f^* = \dfrac{b+d}{T}$. Puisque les probabilités en question sont seulement estimées, les effectifs espérés sont aussi eux-mêmes estimés.

Un calcul d'estimation d'un effectif espéré à partir de l'exemple fictif du paragraphe précédent.

Si l'hypothèse nulle était vraie, c'est-à-dire si la probabilité d'être malade était la même, que l'on soit exposé ou non, on pourrait estimer cette probabilité p^* d'être malade par $f^* = \dfrac{a+c}{T} = \dfrac{125+100}{1000} = 0{,}225$.

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 : $Np^* = L_1 \times 0{,}225 = 500 \times 0{,}225 = 112{,}5$

Le tableau des effectifs espérés et l'application à l'exemple fictif.

On peut donc construire le tableau des effectifs espérés :

	Malades	Non malades	Total
Exposés	a'	b'	L_1
Non exposés	c'	d'	L_0
Total	$a+c$	$b+d$	T

où

$$a' = L_1 \times \frac{a+c}{T} \qquad c' = L_0 \times \frac{a+c}{T}$$

$$b' = L_1 \times \frac{b+d}{T} \qquad d' = L_0 \times \frac{b+d}{T}.$$

Pour l'exemple fictif, le tableau des effectifs espérés est donc :

	Malades	Non malades	Total
Exposés	112,5	387,5	500
Non exposés	112,5	387,5	500
Total	225	775	1000

La statistique du chi-deux de Pearson.

A partir du tableau des effectifs observés et celui des effectifs espérés, on construit une statistique de test, dite **statistique du chi-deux de Pearson**, et dont la valeur est donnée par :

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

On rejette $H_0 : p_1 = p_0$ si et seulement si $\chi^2_{obs} > 3{,}84$.

Exemple : En appliquant ce test aux données fictives, on obtient $\chi^2_{obs} = 3{,}5842$. Puisque $3{,}5842 < 3{,}84$, on ne rejette pas H_0. Les données ne permettent pas de conclure que le fait d'être exposé au facteur de risque modifie la probabilité d'être malade. C'est la même conclusion qu'avec le test Z utilisant la loi normale.

On peut vérifier que : ${Z_{obs}}^2 = 1{,}8932^2 = 3{,}5842 = \chi^2_{obs}$ et que $1{,}96^2 = 3{,}84$. Ici, $|Z_{obs}| > 1{,}96$ (resp. $|Z_{obs}| < 1{,}96$) est donc équivalent à $\chi^2_{obs} > 3{,}84$ (resp. $\chi^2_{obs} < 3{,}84$).

Quel que soit le test utilisé, la conclusion s'énonce souvent, par abus de langage, en terme d'indépendance entre l'exposition ou non au facteur de risque et l'apparition de la maladie. Par abus de langage, on dirait ici qu'on ne peut pas conclure à une dépendance entre l'exposition au facteur de risque et le fait de contracter ou non la maladie.

Remarque : Pour appliquer ce test, l'usage veut que la condition suivante soit vérifiée : chacun des effectifs **espérés** du tableau doit être supérieur à 5. Dans le cas contraire, un test dit "test exact de Fischer" peut être réalisé si le nombre de données n'est pas trop grand.

Activité 1

2.3 L'origine de la statistique du chi-deux

C'est Karl Pearson qui le premier, proposa en 1900 dans l'article *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* [Pea00], un test statistique d'ajustement d'une distribution observée à une distribution attendue. Depuis 1892, il avait collaboré avec le zoologiste Weldon qui souhaitait, pour améliorer la compréhension sur l'évolution des espèces, déterminer une preuve empirique de la sélection naturelle à partir de données nombreuses qui ne suivaient pas nécessairement des lois normales. A la suite de ces travaux, il avait écrit en 1896 à Galton qu'il souhaitait développer un test d'ajustement à des distributions asymétriques à l'attention des biologistes et des économistes.

Quelle est la grande nouveauté introduite dans cette publication ? À la suite d'une expérience dont on suppose que les résultats sont conformes à des hypothèses probabilistes, une distribution des données observées est obtenue. Cette distribution est comparée à celle des données espérées si les hypothèses probabilistes étaient vérifiées. Un calcul de la statistique de test, qui est en quelque sorte une mesure de la distance entre les données observées et les données espérées, permet alors de calculer ce que nous appelons le chi-deux observé : χ^2_{obs}. La loi de probabilité de cette statistique de test dépend d'un nombre n de "degrés de liberté". Pearson donne alors explicitement les formules qui permettent de calculer la probabilité que la statistique du chi-deux soit supérieure ou égale à ce χ^2_{obs} et si cette probabilité est faible (resp. forte), il conclut à la non adéquation (resp. l'adéquation) des données observées avec hypothèses probabilistes.

Il calcule donc, pour χ^2_{obs}, noté ici χ^2, la probabilité $P = \mathbb{P}(Y \geqslant \chi^2)$ où Y suit une loi du chi-deux à n degrés de liberté.

Pour n pair, la formule établie par Pearson est la suivante :

$$P = \frac{\int_\chi^\infty e^{-\frac{1}{2}\chi^2}\chi d\chi + e^{-\frac{1}{2}\chi^2}\left\{\frac{\chi^2}{2} + \frac{\chi^4}{2.4} + \frac{\chi^6}{2.4.6} + \ldots + \frac{\chi^{n-2}}{2.4.6\ldots n-2}\right\}}{\int_0^\infty e^{-\frac{1}{2}\chi^2}\chi d\chi}$$

$$= e^{-\frac{1}{2}\chi^2}\left(1 + \frac{\chi^2}{2} + \frac{\chi^4}{2.4} + \frac{\chi^6}{2.4.6} + \ldots + \frac{\chi^{n-2}}{2.4.6\ldots n-2}\right). \quad \text{(vi.)}$$

alors que pour n impair, la formule est :

$$P = \sqrt{\frac{2}{\pi}}\int_\chi^\infty e^{-\frac{1}{2}\chi^2}d\chi + \sqrt{\frac{2}{\pi}}\,e^{-\frac{1}{2}\chi^2}\left(\frac{\chi}{1} + \frac{\chi^3}{1.3} + \frac{\chi^5}{1.3.5} + \ldots + \frac{\chi^{n-2}}{1.3.5\ldots n-2}\right).$$

Pour faciliter la tâche des utilisateurs, Pearson construit une table dont nous donnons un extrait :

	n' = 3.	4.	5.	6.
1	·606,531	·801,253	·909,796	·962,566
2	·367,879	·572,407	·735,759	·849,146
3	·223,130	·391,633	·557,825	·699,994
4	·135,335	·261,470	·406,006	·549,422
5	·082,085	·171,799	·287,298	·415,882
6	·049,787	·111,611	·199,148	·306,220
7	·030,197	·071,888	·135,888	·220,631
8	·018,316	·046,012	·091,578	·156,236
9	·011,109	·029,291	·061,099	·109,064
10	·006,738	·018,567	·040,428	·075,236
15	·000,553	·001,817	·004,701	·010,363
20	·000,045	·000,170	·000,499	·001,250

(χ^2)

Lecture de la table de Pearson

Soit Y suit une loi du chi-deux à n degrés de liberté avec $n = 3$. On cherche $P = \mathbb{P}(Y \geqslant 6)$.

On calcule $n' = n + 1$. Ici $n' = 3 + 1 = 4$. P se trouve au croisement de la colonne $n' = 4$ et de la ligne $\chi^2 = 6$. D'où $P = \mathbb{P}(Y \geqslant 6) = 0{,}111611$. La justification de l'emploi de $n' = n + 1$ dans cette table plutôt que celui de n est exposée dans le paragraphe 1.1 de l'Annexe II.

La lourdeur des calculs en l'absence d'outils numériques (dans le cas de l'utilisation directe des formules) et la limitation à des valeurs entières de χ^2 (dans le cas de l'utilisation de la table) ont certainement amené les successeurs de Pearson à préférer l'utilisation d'autres tables qui permettent, non pas de calculer explicitement P, mais de l'encadrer.

Cette probabilité, aujourd'hui présente dans la plupart des publications, est souvent appelée "p-value" et notée p ou P. Elle dépend bien sûr aussi du nombre n de "degrés de liberté". Dans le cas d'un tableau à 2 lignes et 2 colonnes comme ceux étudiés précédemment, $n = 1$.

Actuellement, la définition de la loi d'une variable qui suit une loi du chi-deux à n degrés de liberté est celle de la somme des carrés de n variables aléatoires indépendantes de loi normale réduite centrée (*cf* Annexe I).

À la suite de Pearson, Ronald A. Fisher a établi dans son livre *Statistical Methods for Research Workers* [Fis25], une table qui donne les quantiles $\chi^2_P(n)$, c'est-à-dire les nombres positifs vérifiant : $\mathbb{P}(Y \geqslant \chi^2_P(n)) = P$ où Y suit une loi du chi-deux à n degrés de liberté,

— pour $n = 1\ ;\ 2\ ;\ .....\ ;\ 29\ ;\ 30$,

— pour $P = 0{,}01\ ;\ 0{,}02\ ;\ 0{,}05\ ;\ 0{,}10\ ;\ 0{,}20\ ;\ 0{,}30\ ;\ 0{,}50\ ;\ 0{,}70\ ;\ 0{,}80\ ;\ 0{,}90\ ;\ 0{,}95\ ;\ 0{,}98\ ;\ 0{,}99$.

Nous reproduisons, en fin d'article à la page 28, cette table telle qu'elle apparaît dans l'ouvrage de Hill, *Principles of Medical Statistics* [Hil37, p. 308-309].

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

Lecture de la table

Exemple 1

Soit Y une statistique de test qui suit une loi du chi-deux à $n = 1$ degré de liberté.
Au croisement de la ligne $n = 1$ et de la colonne $P = 0{,}05$, on lit $\chi^2_{0{,}05}(1) = 3{,}841$.
On a donc $\mathbb{P}(Y \geqslant 3{,}841) = 0{,}05$.

Exemple 2

Soit Y une statistique de test qui suit une loi du chi-deux à $n = 3$ degrés de liberté.
Au croisement de la ligne $n = 3$ et de la colonne $P = 0{,}01$, on lit $\chi^2_{0{,}01}(3) = 11{,}341$.
On a donc $\mathbb{P}(Y \geqslant 11{,}341) = 0{,}01$.

Exemple 3

De même, pour Y une statistique de test qui suit une loi du chi-deux à $n = 14$ degrés de liberté, $\chi^2_{0{,}90}(14) = 7{,}790$ et on a $\mathbb{P}(Y \geqslant 7{,}790) = 0{,}90$.

Exemple 4

Pour Y une statistique de test qui suit une loi du chi-deux à n degrés de liberté $1 \leqslant n \leqslant 30$, cette table ne permet pas de calculer explicitement $\mathbb{P}(Y \geqslant x)$, mais elle permet un encadrement de cette probabilité.

Soit Y une statistique de test qui suit une loi du chi-deux à $n = 5$ degrés de liberté, on cherche à encadrer $P = \mathbb{P}(Y \geqslant 9{,}8)$.
Sur la ligne $n = 5$ de la table, $9{,}8$ est compris entre $9{,}236$ qui correspond à la colonne $P = 0{,}10$ et $11{,}070$ qui correspond à la colonne $P = 0{,}05$, donc $0{,}05 \leqslant \mathbb{P}(Y \geqslant 9{,}8) \leqslant 0{,}10$ ou encore $0{,}05 \leqslant P \leqslant 0{,}10$.

Activité 2

Pour $n = 3$ degrés de liberté, la table de Fisher donne, au croisement de la ligne $n = 3$ et de la colonne $P = 0{,}01$ (resp. $P = 0{,}05$, $P = 0{,}1$), la valeur du fractile $\chi^2_{0{,}01}(3) = 11{,}341$ (resp. $\chi^2_{0{,}05}(3) = 7{,}815$, $\chi^2_{0{,}10}(3) = 6{,}251$).

Retrouver, par interpolation, approximativement ces valeurs à partir de l'extrait de la table de Pearson donné plus haut.

Utilisation pour tester une hypothèse probabiliste

Il s'agit de tester :

H_0 : (l'hypothèse probabiliste est vraie) contre H_1 : (l'hypothèse probabiliste n'est pas vraie) avec le risque α, c'est-à-dire avec la probabilité α de décider que hypothèse probabiliste n'est pas vraie alors qu'en réalité elle est vraie.

Cette hypothèse probabiliste H_0 peut prendre diverses formes qui sont données en Annexe II. Dans tous les cas, les effectifs observés sont contenus dans un tableau de données avec r lignes et m colonnes et il s'agit de les comparer à un tableau de effectifs espérés (ou effectifs calculés) de même taille, obtenu en supposant que l'hypothèse probabiliste H_0 est vraie.

Pour ces divers problèmes, il existe une statistique du chi-deux que l'on calcule à partir des effectifs observés et qui mesure d'une certaine façon, la "distance" entre les effectifs observés et les effectifs espérés. Cette statistique est donc un indicateur de la "distance" entre l'hypothèse nulle et l'hypothèse alternative. Si l'hypothèse nulle est vraie, il y a "peu" de chance d'obtenir une valeur de cette statistique qui soit "trop grande".

L'idée du test est donc de rejeter l'hypothèse nulle quand la valeur de la statistique est "trop grande". Mais même dans le cas où l'hypothèse nulle est vraie, la valeur de la statistique augmente avec le nombre de degrés de liberté de cette statistique. Pour prendre une décision, il faut donc prendre en compte ce nombre de degrés de liberté.

Remarque : Pour appliquer ce test, l'usage veut que la condition suivante soit vérifiée : chacun des effectifs espérés du tableau doit être supérieur à 5. Dans le cas contraire, on peut essayer de regrouper des données, par ligne ou/et par colonne. Si le tableau de données a $r = 2$ lignes et $m = 2$ colonnes, et si le nombre total de données n'est pas trop élevé, on peut effectuer un test dit "test exact de Fisher" (voir littérature sur ce sujet).

Réalisation du test.

Calcul de la valeur χ^2_{obs} (notée la plupart du temps χ^2 dans les articles de recherche) de la statistique de test à partir des données de l'expérience.
Déterminer le nombre de degrés de liberté $n = (r-1)(m-1)$.

Deux méthodes sont équivalentes pour prendre une décision :

1ère méthode Si $\chi^2_{obs} > \chi^2_\alpha(n)$, rejet de H_0

2ème méthode Calcul de la p-value : $P = \mathbb{P}(Y \geqslant \chi^2_{obs})$ où Y suit une loi du chi-deux à n degrés de liberté. On peut obtenir un encadrement de cette "p-value" à l'aide de la table de Fisher.

Si $P < \alpha$, rejet de H_0.

Exemple : Dans l'exemple fictif du 1.1, il s'agissait de tester avec le risque $\alpha = 0{,}05$ si la probabilité d'être malade quand on est exposé au facteur de risque était la même que la probabilité d'être malade quand on n'est pas exposé. $\chi^2_{obs} = 3{,}5842$ et $n = 1$

1ère méthode : $\chi^2_\alpha(n) = \chi^2_{0{,}05}(1) = 3{,}84$. Puisque $3{,}5842 < 3{,}84$, on ne rejette pas H_0. Les données ne permettent pas de conclure que le fait d'être exposé au facteur de risque modifie la probabilité d'être malade.

2ème méthode : On utilise la ligne $n = 1$ de la table. Puisque $2{,}706 < 3{,}5842 < 3{,}843$, alors : $0{,}05 < \mathbb{P}(Y \geqslant 3{,}5842) < 0{,}10$ où Y est une variable aléatoire qui suit une loi du chi-deux à $n = 1$ degré de liberté. Donc $0{,}05 < \mathbb{P}(Y \geqslant \chi^2_{obs}) < 0{,}10$ ou encore $0{,}05 < P < 0{,}10$. On ne rejette pas H_0.

2.4 Une formule simple pour calculer le chi-deux dans le cas d'un tableau 2×2

Il s'agit encore de tester, avec le risque $\alpha = 0{,}05$,
$H_0 : p_1 = p_0$ contre $H_1 : p_1 \neq p_0$
Les T données sont encore présentées dans le tableau ($T = a + b + c + d$) :

	Malades	Non malades
Exposés	a	b
Non exposés	c	d

La formule de calcul de la statistique de test, dite statistique du chi-deux, figure dans *Principles of Medical Statistics* [Hil37], ouvrage qui a certainement servi de référence statistique à toute l'étude de Doll et Hill. Cette statistique y est notée χ^2 [Hil37, p. 139].

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

On rejette $H_0 : p_1 = p_0$ si et seulement si $\chi^2_{obs} > 3{,}84$.

Exemple : En appliquant ce test aux données fictives ci-dessus, on obtient $\chi^2_{obs} = 3{,}5842$.
Même conclusion que plus haut.

Activité 3

2.5 Le 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 $H_0 : p'_0 = p'_1$ contre $H_1 : p'_0 \neq p'_1$, avec le risque $\alpha = 0{,}05$.
La difficulté provient encore du fait que cette hypothèse se traduit par $p'_0 = p'_1$, donc qu'il existe une probabilité p^* telle que $p^* = p'_0 = p'_1$, mais que cette probabilité n'est pas connue.

Elle sera estimée par $f^* = \dfrac{a + b}{T}$, proportion de personnes exposées dans les deux échantillons réunis.

De même, $1 - p^*$ sera estimée par $1 - f^* = \dfrac{c + d}{T}$, proportion de personnes non exposées dans les deux échantillons réunis.
Puisque les probabilités en question sont seulement estimées, les effectifs espérés sont aussi eux-mêmes estimés.

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

Et le chi-deux observé est égal à : $\chi^2_{obs} = \dfrac{(a - a')^2}{a'} + \dfrac{(b - b')^2}{b'} + \dfrac{(c - c')^2}{c'} + \dfrac{(d - d')^2}{d'}$

On rejette $H_0 : p'_1 = p'_0$ si et seulement si $\chi^2_{obs} > 3{,}84$

Remarque : Comme précédemment, pour appliquer ce test, l'usage veut que chacun des effectifs **espérés** du tableau soit supérieur à 5. Dans le cas contraire, le "test exact de Fisher" peut être réalisé.

On peut montrer que, si les données de l'exemple étaient issues d'une étude cas-témoins, on aurait obtenu exactement le même chi-deux observé. En effet, dans le calcul des effectifs espérés, les valeurs de b et c sont échangées et comme la formule est symétrique en b et c, le résultat du chi-deux observé est le même.

2.6 Généralisation aux tableaux à r lignes et m colonnes.

Dans une étude cas-témoins publiée en 1950, Wynder et Graham ont interrogé 605 hommes ayant le cancer du poumon et 780 hommes sans cancer du poumon. Chacun devait se situer dans une des 6 catégories suivantes correspondant à ses habitudes tabagiques :

Groupe 0 Non fumeur (moins d'une cigarette par jour pendant plus de 20 ans)

Groupe 1 Fumeur léger (de 1 à 9 cigarettes par jour pendant plus de 20 ans)

Groupe 2 Fumeur modérément lourd (de 10 à 15 cigarettes par jour pendant plus de 20 ans)

Groupe 3 Fumeur lourd (de 16 à 20 cigarettes par jour pendant plus de 20 ans)

Groupe 4 Fumeur excessif (de 21 à 34 cigarettes par jour pendant pendant plus 20 ans)

Groupe 5 Fumeur dépendant (35 cigarettes ou plus par jour pendant au moins 20 ans)

En reconstituant les résultats donnés dans la figure 3 de leur article [WG50, p. 2990], les résultats sont les suivants :

	Groupe 0	Groupe 1	Groupe 2	Groupe 3	Groupe 4	Groupe 5	Total
Avec cancer du poumon	8	14	61	213	187	122	605
Sans cancer du poumon	114	90	148	278	90	60	780
Total	122	104	209	591	277	182	1385

La question posée est celle de savoir si on rejette l'hypothèse suivant laquelle les deux populations, celles des hommes ayant le cancer du poumon (AC) et celle des hommes n'ayant pas le cancer du poumon (SC) sont homogènes au regard de leur habitude tabagique.

Dire que ces deux populations (AC) et (SC) sont homogènes au regard de leur habitude tabagique signifie que, pour chaque groupe caractérisé par une habitude tabagique, la probabilité pour un homme atteint d'un cancer du poumon, d'appartenir à ce groupe est la même que celle pour un homme sans cancer du poumon. Cette probabilité est inconnue.

Si l'hypothèse nulle est vraie, il existe 6 probabilités inconnues et ceci se traduit par la suite d'égalités qui constitue l'hypothèse nulle H_0 :

- $p_1^* = \mathbb{P}_{\text{AC}}(\text{Groupe 0}) = \mathbb{P}_{\text{SC}}(\text{Groupe 0})$ c'est-à-dire la probabilité pour un homme atteint d'un cancer du poumon d'être non-fumeur est la même que celle d'un homme sans cancer du poumon
- $p_2^* = \mathbb{P}_{\text{AC}}(\text{Groupe 1}) = \mathbb{P}_{\text{SC}}(\text{Groupe 1})$ c'est-à-dire la probabilité pour un homme atteint d'un cancer du poumon d'être un fumeur léger est la même que celle d'un homme sans cancer du poumon
- $p_3^* = \mathbb{P}_{\text{AC}}(\text{Groupe 2}) = \mathbb{P}_{\text{SC}}(\text{Groupe 2})$ c'est-à-dire la probabilité pour un homme atteint d'un cancer du poumon d'être un fumeur modérément lourd est la même que celle d'un homme sans cancer du poumon
- $p_4^* = \mathbb{P}_{\text{AC}}(\text{Groupe 3}) = \mathbb{P}_{\text{SC}}(\text{Groupe 3})$ c'est-à-dire la probabilité pour un homme atteint d'un cancer du poumon d'être un fumeur lourd est la même que celle d'un homme sans cancer du poumon
- $p_5^* = \mathbb{P}_{\text{AC}}(\text{Groupe 4}) = \mathbb{P}_{\text{SC}}(\text{Groupe 4})$ c'est-à-dire la probabilité pour un homme atteint d'un cancer du poumon d'être un fumeur excessif est la même que celle d'un homme sans cancer du poumon

— $p_6^* = \mathbb{P}_{AC}(\text{Groupe 5}) = \mathbb{P}_{SC}(\text{Groupe 5})$ c'est-à-dire la probabilité pour un homme atteint d'un cancer du poumon d'être un fumeur dépendant est la même que celle d'un homme sans cancer du poumon
avec bien sûr : $p_1^* + p_2^* + \cdots + p_6^* = 1$

Pour chaque groupe, chacune de ces probabilités peut être estimée par le rapport entre le nombre total d'observations appartenant au groupe et le nombre total d'observations.

Ainsi p_1^* est estimée par $f_1^* = \dfrac{8+114}{1385} = \dfrac{122}{1385} = 0{,}08808664$.

De même : $f_2^* = \dfrac{104}{1385}$ $f_3^* = \dfrac{209}{1385}$ $f_4^* = \dfrac{491}{1385}$ $f_5^* = \dfrac{277}{1385}$ $f_6^* = \dfrac{182}{1385}$

On peut alors, que ce soit pour les hommes atteints d'un cancer du poumon ou pour ceux sans cancer du poumon, calculer les effectifs **espérés** dans chaque groupe correspondant aux habitudes tabagiques.

Trois exemples :

1. Pour les 605 hommes atteints du cancer du poumon, l'effectif **espéré** de ceux qui sont dans le groupe 0 est $605 \times p_1^*$. Il est estimé par $605 \times f_1^* = 605 \times \dfrac{122}{1385} = 53{,}29242$
2. Pour les 605 hommes atteints du cancer du poumon, l'effectif **espéré** de ceux qui sont dans le groupe 4 est $605 \times p_5^*$. Il est estimé par $605 \times f_5^* = 605 \times \dfrac{277}{1385} = 121{,}0000$
3. Pour les 780 hommes sans cancer du poumon, l'effectif **espéré** de ceux qui sont dans le groupe 1 est $780 \times p_2^*$. Il est estimé par $780 \times f_2^* = 780 \times \dfrac{104}{1385} = 58{,}57040$

Le tableau des effectifs **espérés** est alors :

	Groupe 0	Groupe 1	Groupe 2	Groupe 3	Groupe 4	Groupe 5	Total
Avec cancer du poumon	53,29242	45,42960	91,29603	214,48014	121,00000	79,50181	605
Sans cancer du poumon	68,70758	58,57040	117,70397	276,51986	156,00000	102,49819	780
Total	122	104	209	591	277	182	1385

Pour obtenir la statistique du chi-deux, il faut d'abord, pour chaque case du tableau des observations, diviser le carré de la différence entre les effectifs observés et les effectifs espérés par les effectifs espérés.

Un exemple :

Pour les hommes atteints d'un cancer du poumon et fumeurs modérément lourds (Groupe 2), ceci donne le résultat :

$$\frac{\left(61 - \dfrac{209 \times 605}{1385}\right)^2}{\dfrac{209 \times 605}{1385}} = \frac{(61 - 91{,}29603)^2}{91{,}29603} = 10{,}05355191$$

Le tableau devient :

	Groupe 0	Groupe 1	Groupe 2	Groupe 3	Groupe 4	Groupe 5	Total
Avec cancer du poumon	38,49334	21,74397	10,05355	0,01021	36,00000	22,71768	129,0188
Sans cancer du poumon	29,85701	16,86551	7,79794	0,00792	27,92307	17,62076	100,0722
Total	68,35035	38,60948	17,85149	0,01813	63,92307	40,33844	229,091

En effectuant la somme des résultats, on obtient la valeur de la statistique de test.

La valeur du chi-deux observé est : $\chi^2_{obs} = 229{,}091$.

Remarque : il n'est pas nécessaire de faire le total par colonnes si le total par lignes a déjà été effectué puisque le total général est le même.

Rappel : le nombre de degrés de liberté à considérer pour un tableau à r lignes et m colonnes est égal à : $n = (r-1)(m-1)$

Ici le tableau est à $r = 2$ lignes et $m = 6$ colonnes car il y a $r = 2$ populations et $m = 6$ modalités, donc le nombre de degrés de liberté à considérer est égal à $n = (r-1)(m-1) = (2-1)(6-1) = 5$.

On cherche à encadrer la "p-value" $P = \mathbb{P}(Y \geqslant 229{,}091)$ où Y suit une loi du chi-deux à 5 degrés de liberté. Sur la ligne $n = 5$ de la table, $229{,}091$ est supérieur à $15{,}086$ qui correspond à la colonne $P = 0{,}01$ donc $P < 0{,}01$.

En prenant le risque $\alpha = 0{,}05$, puisque $P < 0{,}01 < 0{,}05$, l'hypothèse nulle est rejetée. Les populations ne sont pas homogènes au regard de leur habitude tabagique.

Remarque : l'usage veut que les effectifs **espérés** soient supérieurs à 5. Dans le cas contraire, on effectue un regroupement de données.

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

3.1 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 à l'accroissement, en Angleterre et au Pays de Galles, du taux de mortalité dû au cancer du poumon, entre 1927 et 1947. Le nombre de décès est effectivement passé de 612 à 9287, accroissement sans commune mesure avec l'accroissement de la population.

Pour certains auteurs, cet accroissement a été mis sur le compte d'une amélioration des techniques de diagnostic mais il est vite apparu que cette amélioration ne pouvait rendre compte à elle seule de l'accroissement des décès dus au cancer du poumon puisqu'on observait le même accroissement dans les villes que dans les campagnes alors que dans ces dernières les techniques de diagnostic était beaucoup moins évoluées. Ce sont d'autres causes qu'il fallait alors rechercher.

Deux causes étaient mises en avant :

— une pollution atmosphérique générale provenant des échappements des automobiles, des poussières de revêtements des routes, des usines à gaz, des usines industrielles et des centrales à charbon,

— l'augmentation de la consommation de tabac.

Doll et Hill étaient-ils les premiers à s'intéresser à la relation entre la consommation de tabac et le cancer du poumon ?

Non, deux exemples :

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

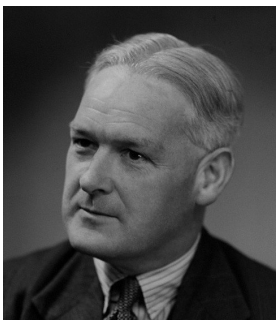
Ernest L. Wynder, étudiant à l'Université de Washington et Evarts A. Graham son professeur, publient en mai 1950 un article suite à une étude portant sur 605 patients atteints du cancer du poumon et 780 non atteints [WG50].

La comparaison entre les non fumeurs (moins d'une cigarette par jour pendant plus de 20 ans : Groupe 0 du tableau p. 12) et les fumeurs (plus de 20 cigarettes par jour pendant au moins 20 ans : regroupements des Groupe 4 et Groupe 5 du tableau p. 12) est éclairante (les sujets ayant une consommation de tabac intermédiaire des Groupes 1, 2 et 3 ne figurent pas dans le tableau ci-dessous).

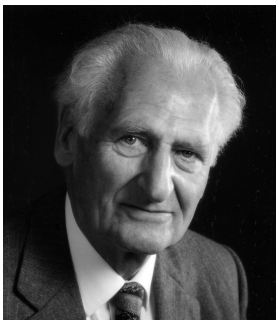
	Non fumeurs	Fumeurs	Effectif total
Ayant le Cancer	$8\ (1,3\ \%)$	$309\ (51,1\ \%)$	605
Sans le cancer	$114\ (14,6\ \%)$	$150\ (19,2\ \%)$	780

L'étude de Doll et Hill de 1950 porte sur les éventuels liens entre la consommation de tabac et l'existence d'un cancer du poumon et elle semble être la première à utiliser l'outil statistique qu'est le test du chi-deux pour inférer sur l'existence ou non de ce lien.

Qui étaient Richard Doll et Austin Bradford Hill ?



Sir Austin Bradford Hill (1897-1991), fils d'un physiologiste renommé, se destinait à des études de médecine lorsque la première guerre mondiale fut déclarée. Il s'engagea comme pilote dans la Royal Navy et ayant contracté la tuberculose alors qu'il faisait route vers les Dardanelles, il fut renvoyé chez lui pour y mourir. Contre toute attente, il guérit mais sa longue convalescence lui fit abandonner l'idée de travailler dans les domaines nécessitant des efforts physiques comme la médecine ou la science. Conseillé et encouragé par Major Greenwood, élève de Karl Pearson et ami proche de la famille, il s'orienta alors vers une carrière dans les statistiques médicales, d'abord au *Medical Research Council* (MRC) et, par la suite, à l'Ecole d'Hygiène et de Médecine Tropicale de Londres. En 1937, il publia une série de dix-sept articles dans le journal The Lancet. Ces articles furent regroupés comme chapitres d'un ouvrage, *Principles of Medical Statistics*, publié la même année. Durant les cinquante-cinq années qui ont suivi, ce manuel a eu onze rééditions successives (la 12^{e} édition date de 1991) et a été rapidement connu des médecins, épidémiologistes et statisticiens du domaine médical du monde entier. Hill pensait qu'il était dangereux de fonder des conclusions à partir de données insuffisantes d'un point de vue statistique et il eut la clairvoyance de proposer d'éviter ce risque, non pas en mettant en avant la nécessité de prendre conseil auprès d'un spécialiste en statistique mais en insistant sur le fait qu'un professionnel du domaine médical, aussi bien en médecine clinique qu'en médecine préventive, devait avoir lui-même des connaissances en techniques statistiques aussi bien pour planifier des expériences que pour interpréter des chiffres. En 1946, il persuada deux comités du MRC d'adopter une méthode aléatoire dite randomisation dans la conduite d'essais cliniques contrôlés à grande échelle (en médecine préventive pour évaluer la valeur d'un vaccin contre la coqueluche et en médecine clinique, pour tester l'efficacité de la streptomycine dans le traitement de la tuberculose). Il énonça enfin, dans un article de 1965, *The Environment and Disease : Association or Causation*, des conditions minimales pour fournir une preuve adéquate d'une relation causale entre deux événements. Ces conditions sont appelées critères de causalité de Hill. Ils sont couramment utilisés comme aide dans l'évaluation de la nature causale d'une association malgré leurs nombreuses limites et l'impossibilité d'aboutir à une conclusion formelle.



De mère pianiste concertiste et de père médecin, Sir William Richard Shaboe Doll (1912-2005) naquit, tout comme Sir Austin Bradford Hill, au sein d'une famille anglaise aisée. Son père fut contraint d'interrompre sa vie professionnelle en raison de la survenue d'une sclérose en plaques. Doll a eu un parcours académique à l'image de celui de Hill. Alors que ce dernier souhaitait faire médecine, il fut contraint de suivre une autre voie et devint statisticien tandis que Doll avait entrepris des études de mathématiques qu'il abandonna pour se tourner vers la médecine qu'il étudia à la Faculté de l'hôpital St Thomas, King College de Londres où il continua à travailler jusqu'à la guerre après l'obtention de son diplôme en 1937. Il s'impliqua très tôt dans les mouvements sociaux de son époque ; inscrit au parti communiste anglais, il le quitta en 1939. Il devint un membre important de la *Socialist Medical Association*. À la déclaration de guerre, il fut incorporé dans le *Royal Army medical Corps*, qu'il ne quitta qu'en 1945 pour retourner à l'hôpital St Thomas. En 1948, suite à des incompatibilités de vue avec ses collègues médecins, il rejoignit au Middlesex Central Hospital, l'équipe de recherche du Dr Francis Avery-Jones dont les travaux portant sur l'ulcère gastroduodénal furent réalisés avec l'unité de recherche statistique du MRC. C'est là qu'il fit la connaissance de Hill avec lequel il allait travailler. L'étude qui a marqué leurs carrières communes a pour origine l'augmentation importante dans les années 20 et 30 du taux de cancer du poumon et s'intéresse aux patients atteints de cancer du poumon dans 20 hôpitaux de Londres. Pour expliquer l'origine de cette hausse majeure, deux facteurs retenaient leur attention : le macadam recouvrant les chaussées et les gaz d'échappements des voitures. Ce n'est que plus tard qu'ils

centrèrent leur étude sur le tabagisme qui s'avéra être le seul dénominateur commun au sein de leur population d'étude. Ce premier travail qui marqua l'histoire scientifique médicale du XXème siècle, fut publié dans le *British Medical Journal* en 1950. Par la suite, 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. 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.

3.2 La collecte des données.

Dans 20 hôpitaux londoniens, une enquête était faite sur les malades atteints d'une des quatre pathologies suivantes : cancer du poumon, de l'estomac, du côlon ou du rectum.

Pour chaque malade atteint du cancer du poumon, était recherché dans le même hôpital ou un hôpital voisin, un patient de même sexe et dans la même tranche d'âge (parmi 5 tranches d'âge) non atteint du cancer du poumon. C'est ce patient qui était appelé par les auteurs "contrôle" (témoin). Dans le cas où plusieurs malades correspondaient à ces critères, c'est le premier de la liste qui était choisi.

L'article de Doll et Hill explicite longuement les méthodes utilisées pour obtenir un diagnostic précis de la maladie afin de diminuer le biais qu'un diagnostic erroné pourrait entraîner sur les résultats de l'enquête. Entre avril 1948 et octobre 1949, les notifications de cas de cancer pour les quatre zones étudiées étaient au nombre de 2370. Cependant, beaucoup d'entre eux ont été éliminés de l'étude pour différentes raisons (trop âgés, incapables physiquement de répondre à l'enquêteur, diagnostic modifié en cours d'enquête,...)

En définitive, ce sont 709 patients atteints du cancer du poumon et 709 patients non cancéreux qui ont été interrogés et qui feront partie d'une étude qui est typiquement une étude cas-témoins.

Remarque : dans la suite du texte, les patients atteints du cancer du poumon sont les **cas**, et les patients appelés par les auteurs "contrôle" sont les **témoins**.

3.2.1 Les deux groupes sont-ils homogènes quant à leur classe sociale et leur lieu de résidence ?

Le lieu de résidence et la classe sociale (pour les hommes) ont été renseignés et transcrits dans le tableau II [DH50, p. 741].

TABLE II.—*Comparison Between Lung-carcinoma Patients and Non-cancer Patients Selected as Controls, With Regard to Sex, Age, Social Class, and Place of Residence*

Age	No. of Lung-carcinoma Patients		No. of Non-cancer Control Patients		Social Class (Registrar-General's Categories. Men Only)	No. of Lung-carcinoma Patients	No. of Non-cancer Patients
	M	F	M	F			
25– ..	2	1	2	1	I and II ..	77	87
30– ..	6	0	6	0	III	388	396
35– ..	18	3	18	3	IV and V ..	184	166
40– ..	36	4	36	4			
45– ..	87	10	87	10	All classes ..	649	649
50– ..	130	11	130	11			
55– ..	145	9	145	9	*Place of residence*		
60– ..	109	9	109	9	County of London	330	377
65– ..	88	9	89*	9	Outer London	203	231
70–74 ..	28	4	27*	4	Other county borough ..	23	16
					Urban district	95	54
					Rural district ..	43	27
					Abroad or in Services ..	15	4
All ages	649	60	649	60	Total (M + F)..	709	709

* One control patient was selected, in error, from the wrong age group.

Doll et Hill se sont d'abord légitimement demandés si, pour les hommes, le groupe des cas et le groupe des témoins étaient homogènes quant à leur classe sociale.

Après avoir calculé un chi-deux sur un tableau à 3 lignes et 2 colonnes, donc à $(3-1)(2-1) = 2$ degrés de liberté, ils en concluaient que les différences observées pouvaient être dues au hasard.

Activité 4

Doll et Hill se sont aussi demandés si le groupe des cas et le groupe des témoins étaient homogènes quant à leur lieu de résidence. Ils en concluaient que ce n'était pas le cas.

Activité 5

L'utilisation de ces deux tests montre le souci des auteurs de vérifier une homogénéité ou non entre les cas et les témoins quant à certains facteurs afin de tenter d'éliminer l'influence éventuelle d'autres facteurs que le facteur de risque étudié.

Ce type d'approche utilisant des outils statistiques pour l'élimination d'éventuels autres facteurs est très important et il fait partie encore aujourd'hui de la méthodologie utilisée dans toute enquête épidémiologique.

3.3 Qu'est-ce qu'un fumeur ?

Un homme qui a été "fumeur léger" peut devenir "fumeur lourd" et réciproquement un "fumeur lourd" peut réduire sa consommation ou même l'arrêter pour différentes raisons (survenue d'une maladie respiratoire aiguë, augmentation du prix du tabac,...)

Pour approfondir l'"histoire" personnelle de la consommation de tabac de chaque patient (cas et témoins), les patients étaient interrogés sur les questions suivantes :

1. Avaient-ils été fumeur à une période de leur vie ?
2. À quel âge avaient-ils commencé et, dans le cas d'un arrêt, à quel âge avaient-ils arrêté ?
3. Quelle quantité de tabac avaient-ils l'habitude de fumer avant le début de la maladie qui les avait conduit à l'hôpital ?
4. Quelles avaient été les principaux changements dans leur consommation de tabac et la quantité maximum de tabac consommée pendant toute la période au cours de laquelle ils avaient été fumeurs ?
5. Dans le cas où ils avaient été fumeurs de pipe et de cigarette, quelle avait été la répartition entre l'usage de la pipe et celle de la cigarette ?
6. Avalaient-ils ou non la fumée ?

Dans les résultats de l'enquête, un fumeur est défini comme celui qui avait fumé au moins une cigarette par jour depuis un an.

Pour juger de la fiabilité des récits, 50 témoins ont été interrogés une seconde fois six mois plus tard sur la même question : "Quelle quantité de tabac avez-vous fumé avant le déclenchement de votre maladie ?". Ces résultats sont transcrits dans le tableau III [DH50, p. 742] de l'article de Doll et Hill mais n'ont cependant pas fait l'objet d'un test car le test "classique" du chi-deux était inadapté au traitement de résultats appariés. Les auteurs, au vu des résultats admettent que dans l'ensemble, les réponses indiquent la même tendance générale. Seule, une généralisation du test dit de Mac-Nemar, apparue ultérieurement dans la littérature statistique, permet de résoudre ce problème.

3.4 Un premier résultat

C'est bien sûr le résultat principal de l'étude où ont été séparées dans le tableau IV [DH50, p. 742], 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	

Doll et Hill affirment l'existence d'une association entre le fait d'avoir été fumeur et 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 a le cancer du poumon est supérieure à celle d'avoir été fumeur quand on ne l'a pas.

Activité 6 **Activité 7**

3.5 L'approfondissement de l'étude

Les tableaux ci-dessus ne différencient pas les "fumeurs lourds" des "fumeurs légers". Le tableau V ci-dessous [DH50, p. 742] permet cette distinction. Les quantités de cigarettes indiquées dans le tableau sont celles fumées immédiatement avant le début de la maladie. S'ils avaient abandonné avant le début de la maladie, c'est la quantité de cigarettes immédiatement avant l'abandon qui est prise en compte.

TABLE V.—*Most Recent Amount of Tobacco* Consumed Regularly by Smokers Before the Onset of Present Illness; Lung-carcinoma Patients and Control Patients with Diseases Other Than Cancer*

Disease Group	No. Smoking Daily					Probability Test
	1 Cig.–*	5 Cigs.–	15 Cigs.–	25 Cigs.–	50 Cigs. +	
Males: Lung-carcinoma patients (647)	33 (5·1%)	250 (38·6%)	196 (30·3%)	136 (21·0%)	32 (5·0%)	$\chi^2 = 36{\cdot}95$; n = 4; P < 0·001
Control patients with diseases other than cancer (622) ..	55 (8·8%)	293 (47·1%)	190 (30·5%)	71 (11·4%)	13 (2·1%)	
Females: Lung-carcinoma patients (41) ..	7 (17·1%)	19 (46·3%)	9 (22·0%)	6 (14·6%)	0 (0·0%)	$\chi^2 = 5{\cdot}72$; n = 2; 0·05 < P < 0·10 (Women smoking 15 or more cigarettes a day grouped together)
Control patients with diseases other than cancer (28) ..	12 (42·9%)	10 (35·7%)	6 (21·4%)	0 (0·0%)	0 (0·0%)	

* Ounces of tobacco have been expressed as being equivalent to so many cigarettes. There is 1 oz. of tobacco in 26·5 normal-size cigarettes, so that the conversion factor has been taken as: 1 oz. of tobacco a week = 4 cigarettes a day.

Activité 8

3.5.1 Historique de la consommation

Pour affiner l'hypothèse par laquelle on trouve plus de "fumeurs lourds" chez les malades atteints du cancer du poumon que dans les autres catégories de malades, Doll et Hill font remarquer que la quantité de cigarettes fumée par jour ne donne pas nécessairement une juste représentation de l'historique de la consommation de tabac, certains pouvant peut être avoir fumé davantage (ou non) à une autre période de leur vie.

Ainsi, parmi les hommes atteints du cancer du poumon, si 32 déclaraient fumer plus de 50 cigarettes par jour juste avant le début de la maladie (ou avant leur arrêt de consommation de tabac), 45 déclaraient avoir fumé plus de 50 cigarettes dans une journée, au moins une fois. Le tableau VI de l'article de Doll et Hill donne les résultats de l'enquête.

Un autre indicateur est l'estimation du nombre total de cigarettes fumées avant l'admission à l'hôpital. Si par exemple, le patient a fumé à une période de sa vie 15 cigarettes par jour pendant 1000 jours et à une autre période de sa vie 25 cigarettes pendant 2000 jours, on estime sa consommation totale à 65000. Cet indicateur permet de prendre en compte à la fois si le patient a été, à une période de sa vie, un fumeur lourd et le temps pendant lequel il a été ce fumeur lourd. Le tableau VII de l'article de Doll et Hill donne les résultats de l'enquête.

Doll et Hill font remarquer qu'on aurait pu penser qu'en affinant les questions, le lien entre la consommation de tabac et le cancer du poumon, d'une certaine mesure quantifié par la valeur du chi-deux correspondant, aurait été encore plus significatif. Mais ce qui est semble être gagné en tentant de mieux décrire l'historique de la consommation de tabac est contrebalancé par une certaine imprécision des réponses qui font appel à une mémoire souvent lointaine. Puisque les résultats issus des tableaux VI et VII [DH50, p. 743] vont dans le même sens que ceux du tableau V, Doll et Hill adoptent le critère des "plus récentes quantités fumées" pour décrire le type de consommation de tabac.

Dans l'étude de l'historique de la consommation de tabac d'un patient, d'autres éléments peuvent être pertinents comme par exemple l'âge de début de la consommation de tabac, le nombre d'années pendant lesquelles il a été fumeur, le nombre d'années d'arrêt de la consommation de tabac. Ils ont été étudiés par Doll et Hill et les résultats sont transcrits dans le tableau VIII [DH50, p. 743]. Curieusement, seule l'association entre le nombre d'années d'arrêt de la consommation de tabac et le cancer du poumon est significative $(0{,}01 < P < 0{,}02)$ alors que celle entre le nombre d'années pendant lesquelles on a été fumeur et le cancer du poumon ne l'est pas $(0{,}05 < P < 0{,}10)$.

3.5.2 Cigarettes et pipes.

Aucune distinction n'a été faite entre les fumeurs de cigarettes et les fumeurs de pipes. Il importait de savoir si le mode de consommation (cigarettes ou pipe) influait sur une liaison entre la consommation de tabac et le cancer du poumon. Doll et Hill font remarquer qu'une des difficultés provient du fait qu'un fumeur de pipe au moment de l'interrogatoire peut avoir fumé des cigarettes jusqu'à un temps court avant celui-ci, et réciproquement, un fumeur de cigarette peut avoir substitué la cigarette à la pipe qui était son mode de consommation habituel. Pour éviter cette difficulté, les seules personnes retenues pour une étude portant sur la liaison entre le cancer du poumon et le mode de consommation ont été celles qui n'avaient pratiqué qu'un seul de ces deux modes pendant toute la période où ils ont consommé du tabac.

Ainsi sur les 525 personnes atteintes du cancer du poumon qui avaient fumé pipe ou cigarette mais pas les deux, $5{,}7\%$ étaient des fumeurs de pipe et $94{,}3\%$ des fumeurs de cigarettes et sur les 507 patients ayant une autre maladie $9,7\%$ étaient des fumeurs de pipe et $90{,}3\%$ des fumeurs de cigarettes.

A la suite d'un test statistique, Doll et Hill en concluaient qu'il était peu probable que la plus forte proportion de fumeurs de cigarettes parmi les malades atteints du cancer du poumon soit due au hasard.

Les auteurs n'en concluent pas que le facteur décisif est la pratique de la cigarette plutôt que celui de la pipe. En effet, le fait qu'un fumeur de pipe absorbe en moyenne moins de tabac que le fumeur de cigarette

peut être l'explication. Cependant, une étude effectuée sur les purs fumeurs de pipe confirme les résultats du tableau V (qui portaient sur l'ensemble des fumeurs) avec un risque plus important de cancer du poumon chez ceux qui fumaient une grande quantité de tabac.

Ils dégagent de cette étude que la façon dont le tabac est fumé est importante quant au risque de cancer du poumon mais qu'il leur est impossible de quantifier la différence des risques à l'aide des données dont ils disposent [DH50, p. 744].

3.5.3 Inhalation de la fumée.

Sur 688 malades atteints du cancer du poumon, $61{,}6\%$ inhalaient la fumée tandis que sur 650 malades témoins, ils étaient $67{,}2\%$. Là encore, un test statistique permet aux auteurs de dire que les malades atteints du cancer du poumon inhalent légèrement moins souvent la fumée que les autres patients.

Activité 9

3.6 Tentative de démonstration de l'absence de biais.

Doll et Hill souhaitent prendre toutes les précautions possibles quant à cette interprétation des résultats. Ainsi écrivent-ils : bien que les tableaux précédents ne laissent aucun doute sur l'association entre le tabac et le cancer du poumon, il est nécessaire d'envisager d'autres explications pour ces résultats [DH50, p. 744].

Trois questions sont alors posées :

— les échantillons interrogés (cas et témoins) sont-ils représentatifs de la population ?
— les malades qui viennent d'apprendre qu'ils souffraient d'une maladie des voies respiratoires n'ont-ils pas tendance à modifier leurs réponses aux questions concernant leurs habitudes de consommation du tabac ?
— les enquêteurs n'ont-ils pas eu tendance à choisir comme témoins des "fumeurs légers" ?

3.6.1 Représentativité des échantillons

Il avait été noté un moins grand nombre de patients atteints par le cancer du poumon résidant à Londres.

Une étude est alors restreinte aux malades enquêtés (98 cancers du poumons et 98 contrôles) dans les hôpitaux du district de Londres, étude dont les résultats confirment que le lien entre le tabac et le cancer du poumon existe quel que soit le lieu d'investigation.

Les témoins choisis à l'hôpital pouvaient présenter 5 catégories de maladies autres que le cancer du poumon. On aurait pu penser qu'il existait des différences éventuelles quant à leur habitudes de consommation de tabac.

L'étude dont les résultats sont notés dans le tableau X [DH50, p. 744] montre qu'il n'en n'est rien et que le fait que les témoins soient atteints de cinq maladies différentes n'est pas susceptible de troubler les résultats obtenus.

Une dernière question concernant le choix par les enquêteurs des témoins a été étudiée par Doll et Hill. Les enquêteurs n'ont-ils pas eu tendance à choisir dans les malades recensés par les hôpitaux comme ceux ayant des maladies autres que le cancer du poumon, un nombre disproportionné de "fumeurs légers" ?

Une comparaison a donc été effectuée entre les "témoins" et les malades recensés par les hôpitaux comme ayant des maladies autre que le cancer du poumon.

Les résultats consignés dans le tableau XI [DH50, p. 745] montrent que, concernant leurs habitudes de consommation de tabac, les témoins sont représentatifs de l'ensemble des malades recensés par les hôpitaux comme ceux ayant des maladies autre que le cancer du poumon.

3.6.2 Biais possibles dus aux récits des malades atteints d'une maladie respiratoire.

On peut se poser la question «les malades qui viennent d'apprendre qu'ils souffraient d'une maladie des voies respiratoires n'ont-ils pas tendance à modifier leurs réponses aux questions concernant leurs habitudes de consommation du tabac ? ». Il est difficile d'y répondre mais les auteurs remarquent, dans le tableau X, que les malades atteints de maladies respiratoires autres que le cancer du poumon n'ont pas des habitudes de consommation de tabac statistiquement différentes des autres malades.

3.6.3 Biais possible provenant de la croyance, à juste raison ou à tort, à l'existence d'un cancer du poumon chez le malade interrogé.

Les interviewers connaissaient la maladie dont chaque patient souffrait et cela aurait peut-être pu influencer les données obtenues. Or, chez certains patients, un cancer du poumon avait été diagnostiqué à tort à l'époque des interviews ; plus tard, ces diagnostics ont été corrigés. On a étudié les données correspondantes et on a constaté qu'elles étaient effectivement analogues à celles des patients diagnostiqués comme n'ayant pas de cancer du poumon, ce qui a permis de lever toute ambigüité.

3.7 En conclusion

« 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].

Les auteurs ajoutent que le tableau X ne révèle aucune association entre les habitudes de consommation de tabac et le fait d'avoir une maladie respiratoire autre que le cancer. De même, on ne constate aucune association entre les habitudes de consommation de tabac et le fait d'avoir un cancer sur une autre localisation que celle du poumon. Pour eux, « l'association semble être spécifique au cancer du poumon » [DH50, p. 746].

C'est certainement le point faible des conclusions de leur étude, point qui servira d'appui à de violentes critiques contre leur étude par d'autres auteurs. Ils font bien sûr remarquer que parler d'association entre deux événements pourrait laisser entendre que la réalisation d'un des événements a de fortes chances d'entraîner la réalisation de l'autre et ici ce n'est pas le cas. En effet, il est peu probable que ce soit l'apparition de la maladie qui entraîne la consommation de tabac !

L'étude est complétée par une enquête auprès des habitants du Grand Londres. Les auteurs calculent pour chaque tranche d'âge et pour une quantité de tabac donnée le quotient suivant :

$$\frac{\text{Nombre de personnes atteintes d'un cancer du poumon fumant la quantité donnée}}{\text{Nombre de personnes fumant la quantité donnée}}$$

Les résultats sont consignés dans le tableau XIV [DH50, p. 746].

TABLE XIV.—*Ratios of Patients Interviewed With Carcinoma of Lung and with a Given Daily Consumption of Tobacco to the Estimated Populations in Greater London Smoking the Same Amounts (Male and Female Combined; Ratios per Million)*

Age	Daily Consumption of Tobacco						Total
	0	1–4 Cigs.	5–14 Cigs.	15–24 Cigs.	25–49 Cigs.	50 Cigs.+	
25– ..	*0**	*11*	*2*	*6*	*28*	—	4
35– ..	*2*	*9*	43	41	67	*77*	29
45– ..	*12*	34	178	241	429	667	147
55– ..	14	133	380	463	844	600	244
65–74 ..	21	110	300	510	1,063	*2,000*	186

* Ratios based on less than 5 cases of carcinoma of the lung are given in italics.

En faisant l'hypothèse que le "risque" d'avoir un cancer du poumon est proportionnel à la quantité de tabac fumé, les auteurs mettent ainsi en évidence que ce risque s'accroit avec la quantité fumée. C'est une des premières tentatives de quantifier le risque de cancer en fonction des quantités fumées même si aucun traitement statistique n'est effectué.

L'importance du genre dans l'étude de cette liaison est aussi explorée sans qu'elle mette en cause l'existence de cette liaison entre la consommation de tabac et le cancer du poumon.

Leur étude se termine enfin par une question plus biologique avec la recherche d'éventuels cancérigènes dans le tabac qui seraient donc cause de sa nocivité. La seule substance cancérigène qui avait été trouvée était l'arsenic. Or l'arsenic était aussi présent dans des insecticides qui ont de plus en plus été utilisés. Ceci pourrait expliquer pourquoi l'augmentation du nombre de cancers du poumon a été plus importante que celle de la consommation de tabac.

4 L'influence de l'étude et ses prolongements, ses détracteurs et les polémiques analogues aujourd'hui

4.1 L'influence de l'étude

Le retentissement qu'a eu l'article de Doll et Hill à la fois auprès des médecins, biologistes et statisticiens mais aussi de l'opinion publique et des décideurs en matière de santé publique a maintes fois été relaté dans de nombreuses publications. Citons entre autres Luc Berlivet [Ber03] et Élodie Giroux [Gir11].

L'étude de Doll et Hill a failli rester dans les tiroirs. En effet, « Impressionnés par ces premiers résultats, les dirigeants du *Medical Research Council* s'inquiètent néanmoins des conséquences que pourrait entraîner leur divulgation. C'est qu'ils saisissent fort bien les enjeux, en terme de santé publique, mais également aux plans économique, social, et finalement politique, d'une étude aboutissant à incriminer un produit dont la taxation représente environ $17,5\%$ du revenu fiscal du gouvernement. »[Ber03, p. 44]

Comme déjà signalé, cette étude avait été précédée de la publication en mai 1950 dans le journal *JAMA* par Ernst L. Wynder, étudiant à l'Université de Washington et Evarts A. Graham son professeur, d'un article *Tobacco Smoking as a Possible Etiologic Factor in Bronchiogenic Carcinoma* [WG50], où les deux auteurs évoquaient le rôle possible du tabagisme dans le cancer du poumon suite à une étude cas-témoins portant sur 605 "cas" et 780 "témoins". Ayant pris connaissance de cette étude, Doll et Hill décidèrent l'éditeur en chef du *British Medical Journal* de publier leurs résultats, ce qui fut fait en septembre 1950.

Un tournant ?

Même si c'est la conjonction de ces deux études qui créa l'"événement", l'étude de Doll et Hill se caractérise par l'utilisation répétée et quasi systématique d'un test statistique, le test du chi-deux, pour à la fois inférer d'une association entre le tabagisme et le cancer du poumon et tenter de neutraliser les biais possibles en testant l'homogénéité ou non des "témoins" et leur représentativité. Pas moins de 16 tests du chi-deux sont mis en œuvre dans cette étude, ce qui n'apparait pas clairement à la lecture de l'article de Wynder et Graham.

Aujourd'hui, bien que ce ne soit plus le test du chi-deux qui domine l'analyse des résultats, les conclusions des études épidémiologiques sont, depuis la parution de l'article de Doll et Hill, la plupart du temps appuyées sur des tests de signification. De même, la démarche très présente dans l'article de Doll et Hill consistant à vérifier l'homogénéité des populations observées afin de se prévenir d'un facteur « caché », est toujours d'actualité.

Cependant, ce que semblent retenir les historiens des sciences comme "tournant" est davantage l'introduction de la notion de rapport de risque et sa liaison avec la quantité de tabac fumée, esquissée à la fin de l'article de Doll et Hill. L'analyse de cette liaison ne pouvait pas faire à cette époque l'objet d'un test de

signification statistique, les outils nécessaires -modèles linéaires généralisés- pour le faire n'apparaissant que plus tard dans la littérature scientifique.

4.2 Les prolongements

Doll et Hill ont persévéré dans l'analyse de cette association entre le tabagisme et le cancer du poumon en effectuant, non plus une nouvelle étude cas-témoins, mais en lançant une étude de cohorte de grande ampleur auprès des médecins de Grande Bretagne. Des questionnaires les interrogeant sur leur état de santé et leur consommation de tabac étaient régulièrement envoyés. Plus du tiers des médecins (soit environ 40 000 personnes) ont répondu et les résultats furent publiés dans trois rapports successifs, en 1954, 1956 et 1964, qui allaient dans le même sens que le rapport de 1950. Le suivi de cette cohorte dura en réalité quarante ans et en 2002, Richard Doll y mettra fin en écrivant aux survivants.

L'article de Luc Berlivet évoque au moins deux études importantes qui ont prolongé celle de Doll et Hill.

« Outre-Atlantique, plusieurs équipes vont emprunter cette même voie. L'une d'entre elles met à profit la capacité de mobilisation de l'association américaine de lutte contre le cancer (*American Cancer Society*), qui jouit déjà à l'époque d'une très grande notoriété. Avec l'appui de milliers de volontaires chargés de recueillir des informations dans tout le pays, les statisticiens Cuyler Hammond et Daniel Horn parviennent ainsi à suivre 187 783 hommes durant quarante-quatre mois en moyenne, puis à analyser les taux de décès en fonction de la consommation de tabac. La seconde étude débute en 1954 à l'initiative de Harold Dorn, un sociologue recruté par les *National Institutes of Health* pour développer la recherche statistique. Deux cent mille vétérans de l'armée américaine, détenteurs d'une police d'assurance-vie gouvernementale, et donc faciles à contacter, seront suivis pendant trente deux mois » [Ber03, p. 44-45].

Les conséquences de ces études en terme de politique de santé publique sont remarquablement décrites dans le chapitre 6 du livre de J.P. Gaudillière, *Inventer la biomédecine, la France, l'Amérique et la production des savoirs du vivant* (1945-1965) [Gau02] et continuent d'avoir un impact important sur les décideurs.

4.3 Les détracteurs et leurs arguments

Les arguments des détracteurs de Doll et Hill ont été largement développés dans nombres de publications.

L'existence d'une corrélation entre le tabagisme et le cancer du poumon signifie bien que la probabilité d'avoir le cancer du poumon est affectée par le fait d'être ou non un gros fumeur. Cependant, pour certains détracteurs, cette modification de la probabilité d'avoir le cancer du poumon était attribuée à un "tiers facteur".

Évoquons-en quelques-uns :

La localisation géographique

> « Un spécialiste comme W. Hueper, le chef de la division de cancérigène chimique du NCI arguait ainsi (et avec lui nombre de médecins progressistes britanniques) que la classe ouvrière vivait dans des villes industrielles fortement polluées par des substances carcinogènes et que, par ailleurs, les hommes de cette classe sociale sont souvent de gros fumeurs. » [Gau02, p. 228]

L'existence d'un gène de prédisposition

> « De façon analogue, R.A. Fischer maniait un argument relevant de l'arsenal des généticiens des populations. La corrélation pouvait simplement trouver son origine dans le fait que, comme beaucoup de comportements, la forte consommation de tabac est favorisée par un gène de prédisposition et que ce facteur est, par un pur hasard de localisation chromosomique, situé à proximité d'un facteur génétique de prédisposition au cancer du poumon. » [Gau02, p. 228]

Le rôle cancérigène du papier à cigarette Doll et Hill avaient déjà remarqué dans leur article de 1950 que chez les malades atteints du cancer du poumon la probabilité d'être un fumeur de cigarettes est plus grande que la probabilité d'être un fumeur de pipe, et, s'appuyant sur les résultats d'un test du chi-deux que, ceci n'était probablement pas dû au hasard (*cf.* ci-dessus le paragraphe « Cigarettes et Pipes »).

Le généticien C.C. Little, président du Tobacco Industry Research Committee de 1954 à 1969, s'appuyait alors sur cette constatation pour mettre en cause le rôle cancérigène du papier à cigarette, ce qui n'a pu être prouvé par la suite.

Les différences de constitution entre fumeurs et non fumeurs

« L'Américain Joseph Berkson, pionnier de la statistique médicale dans l'entre-deux-guerres, fit aussi remarquer que les non-fumeurs différaient très probablement des fumeurs sur de nombreux points, qu'ils paraissaient généralement plus pondérés, qu'il s'agisse de leur alimentation, de leur hygiène de vie, etc. Cette série de petites variations pouvait fort bien être la cause réelle de leur plus faible mortalité. » [Ber03, p. 43]

Outre l'existence supposée d'un "tiers facteur" qui est pour lui une constitution de l'individu qui lui permettrait ou non de résister aussi bien aux "tentations" qu'aux maladies, J. Berkson mit en cause la spécificité de cette association en remarquant dans l'article *Smoking and Lung Cancer : Some Observations on Two Recent Reports* [Ber58] que d'autres causes de mort peuvent être attribuables au tabac. En utilisant les données de Doll et Hill dans leur étude de 1956, il considère 5 catégories de "cause de mort" : "Lung cancer", "Other cancer", "Other respiratory diseases", "Coronary thrombosis" et "Other causes". En étudiant les différences de taux de mortalité entre les gros fumeurs et les autres catégories de fumeurs (légers, moyens et non-fumeurs) pour chacune des causes de mort, il constate bien que pour chacune des "causes de mort", il existe une différence entre les gros fumeurs et les autres, mais cette différence est la plus importante, non pas pour la mort par cancer du poumon mais pour la mort par maladies coronariennes.

Il en conclut : « Pour ma part, je trouve incroyable que fumer puisse être la cause de toutes ces maladies. » [Ber58, p. 43].

4.4 Les polémiques analogues aujourd'hui

Gilles-Eric Séralini, professeur de biologie moléculaire à l'université de Caen, a publié avec son équipe un article intitulé « Long term toxicity of a Roundup herbicide and a Roundup-tolerant genetically modified maize » dans la revue *Food and Chemical Toxicology* en date du 19 septembre 2012 [Sér12].

Il y étudie la toxicité de deux produits du groupe Monsanto, l'herbicide Roundup et le maïs OGM NK603.

Aussitôt après la publication de cet article, un grand nombre de lettres ont été envoyées à l'éditeur par d'autres chercheurs, le plus souvent pour en critiquer aigrement le contenu.

Sans prendre parti [2] dans cette forte divergence d'opinion, intéressons-nous au commentaire écrit à l'époque par Louis Ollivier, ancien directeur de recherche à l'INRA [Oll13].

Ce chercheur met l'accent sur une méthode d'analyse qui, d'après lui, aurait pu être utilisée pour comparer les taux de mortalité des rats correspondant aux différents types d'alimentation décrits dans l'article de Séralini. Il s'agit du test du chi-deux avec la formule de Brandt et Snedecor pour les tableaux de taille $2 \times k$.

Dans l'étude dirigée par Séralini, tous les échantillons sont composés du même nombre de rats, à savoir $n = 10$, et le nombre de traitements différents est toujours égal à $k = 4$, le nombre de degrés de liberté étant

2. Il est, malgré tout, fort étonnant que la revue *Food and Chemical Toxicology* ait pris la décision de "retirer" l'article de Séralini et alii (jeudi 28 novembre 2013). Il est à espérer que d'autres travaux seront entrepris pour mettre au clair la question de la toxicité éventuelle du NK603 et du Roundup.

donc $(4-1).(2-1) = 3$. Voici, par exemple, le cas des rats mâles recevant une nourriture partiellement OGM :

	OGM Mâles		
Traitement	Morts (x_i)	Vivants ($n - x_i$)	Total
0% (contrôle)	3	7	**10**
11% OGM	5	5	**10**
22% OGM	1	9	**10**
33% OGM	1	9	**10**
Somme	10		
Somme des Carrés	36		
chi-deux	5,87		

Dans ce cas particulier, il est possible d'utiliser la formule de Brandt et Snedecor pour calculer la valeur du chi-deux :

$$\chi^2_{obs} = \frac{k.\sum x_i^2 - \left(\sum x_i\right)^2}{\sum x_i - \frac{\left(\sum x_i\right)^2}{n.k}}$$

Activité 10

Louis Ollivier fait remarquer que le calcul du chi-deux, facilité par cette formule, peut être fait "au coin d'un bureau". Il donne alors les résultats des calculs de ce type correspondant aux six échantillons de l'étude, sous la forme du tableau suivant :

	OGM		OGM+R		R	
	Mâles	Femelles	Mâles	Femelles	Mâles	Femelles
chi-deux	5,87	5,83	1,17	5,22	2,38	2,50

On peut donc voir qu'aucun des six résultats numériques ne dépasse le seuil de 7,82, nécessaire pour affirmer, au risque de première espèce de 5%, qu'une différence statistiquement significative entre les traitements a été décelée.

D'après Louis Ollivier, cela aurait dû conduire les auteurs à modérer les conclusions qu'ils tirent des taux de mortalité observés lors de leur étude.

Cependant, il ne faut pas oublier que le "non-rejet" de chacune des six hypothèses nulles H_0 n'en constitue pas une démonstration : il se peut que l'étude Séralini, par manque de moyens financiers, n'ait pas pu être conçue de manière à assurer, finalement, une puissance statistique suffisante.

Bien que les techniques utilisées par Séralini ou ses détracteurs n'utilisent pas majoritairement de tests du chi-deux, nous voyons qu'ils peuvent toujours jouer un rôle important.

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 donne effectivement la rigueur et l'objectivité et pas seulement son apparence.

D'autre part, tout le monde a remarqué 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.

TABLE OF χ^2

TABLE OF χ^2 (*contd.*)

n	P = ·99	·98	·95	·90	·80	·70	·50	·30	·20	·10	·05	·02	·01	*n*
1	·000157	·000628	·00393	·0158	·0642	·148	·455	1·074	1·642	2·706	3·841	5·412	6·635	1
2	·0201	·0404	·103	·211	·446	·713	1·386	2·408	3·219	4·605	5·991	7·824	9·210	2
3	·115	·185	·352	·584	1·005	1·424	2·366	3·665	4·642	6·251	7·815	9·837	11·341	3
4	·297	·429	·711	1·064	1·649	2·195	3·357	4·878	5·989	7·779	9·488	11·668	13·277	4
5	·554	·752	1·145	1·610	2·343	3·000	4·351	6·064	7·289	9·236	11·070	13·388	15·086	5
6	·872	1·134	1·635	2·204	3·070	3·828	5·348	7·231	8·558	10·645	12·592	15·033	16·812	6
7	1·239	1·564	2·167	2·833	3·822	4·671	6·346	8·383	9·803	12·017	14·067	16·622	18·475	7
8	1·646	2·032	2·733	3·490	4·594	5·527	7·344	9·524	11·030	13·362	15·507	18·168	20·090	8
9	2·088	2·532	3·325	4·168	5·380	6·393	8·343	10·656	12·242	14·684	16·919	19·679	21·666	9
10	2·558	3·059	3·940	4·865	6·179	7·267	9·342	11·781	13·442	15·987	18·307	21·161	23·209	10
11	3·053	3·609	4·575	5·578	6·989	8·148	10·341	12·899	14·631	17·275	19·675	22·618	24·725	11
12	3·571	4·178	5·226	6·304	7·807	9·034	11·340	14·011	15·812	18·549	21·026	24·054	26·217	12
13	4·107	4·765	5·892	7·042	8·634	9·926	12·340	15·119	16·985	19·812	22·362	25·472	27·688	13
14	4·660	5·368	6·571	7·790	9·467	10·821	13·339	16·222	18·151	21·064	23·685	26·873	29·141	14
15	5·229	5·985	7·261	8·547	10·307	11·721	14·339	17·322	19·311	22·307	24·996	28·259	30·578	15
16	5·812	6·614	7·962	9·312	11·152	12·624	15·338	18·418	20·465	23·542	26·296	29·633	32·000	16
17	6·408	7·255	8·672	10·085	12·002	13·531	16·338	19·511	21·615	24·769	27·587	30·995	33·409	17
18	7·015	7·906	9·390	10·865	12·857	14·440	17·338	20·601	22·760	25·989	28·869	32·346	34·805	18
19	7·633	8·567	10·117	11·651	13·716	15·352	18·338	21·689	23·900	27·204	30·144	33·687	36·191	19
20	8·260	9·237	10·851	12·443	14·578	16·266	19·337	22·775	25·038	28·412	31·410	35·020	37·566	20
21	8·897	9·915	11·591	13·240	15·445	17·182	20·337	23·858	26·171	29·615	32·671	36·343	38·932	21
22	9·542	10·600	12·338	14·041	16·314	18·101	21·337	24·939	27·301	30·813	33·924	37·659	40·289	22
23	10·196	11·293	13·091	14·848	17·187	19·021	22·337	26·018	28·429	32·007	35·172	38·968	41·638	23
24	10·856	11·992	13·848	15·659	18·062	19·943	23·337	27·096	29·553	33·196	36·415	40·270	42·980	24
25	11·524	12·697	14·611	16·473	18·940	20·867	24·337	28·172	30·675	34·382	37·652	41·566	44·314	25
26	12·198	13·409	15·379	17·292	19·820	21·792	25·336	29·246	31·795	35·563	38·885	42·856	45·642	26
27	12·879	14·125	16·151	18·114	20·703	22·719	26·336	30·319	32·912	36·741	40·113	44·140	46·963	27
28	13·565	14·847	16·928	18·939	21·588	23·647	27·336	31·391	34·027	37·916	41·337	45·419	48·278	28
29	14·256	15·574	17·708	19·768	22·475	24·577	28·336	32·461	35·139	39·087	42·557	46·693	49·588	29
30	14·953	16·306	18·493	20·599	23·364	25·508	29·336	33·530	36·250	40·256	43·773	47·962	50·892	30

This table is reproduced, by kind permission of the author and publishers, from *Statistical Methods for Research Workers*, by R. A. Fisher, Sc.D., F.R.S., 6th edition, 1936, Oliver & Boyd, Edinburgh and London.

Table du chi-deux

Références

[Ber58] J. Berkson. Smoking and Lung Cancer : Some Observations on two Recent Reports. *American Statistical Association Journal*, 1958. p. 28-38.

[Ber03] Luc Berlivet. La preuve par le tabac. *La Recherche*, 2003. Hors Série numéro 13, p. 42-46.

[DH50] Richard Doll and Austin B. Hill. Smoking and Carcinoma of the Lung. *British Medical Journal*, 1950.

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

[Gau02] Jean-Paul Gaudillière. *Inventer la biomédecine, la France, l'Amérique et la production des savoirs du vivant (1945-1965)*. La Découverte, 2002. Chapitre 6, p. 218-245.

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

[Gir11] Élodie Giroux. Contribution à l'histoire de l'épidémiologie des facteurs de risques. *Revue d'Histoire des Sciences*, 2011. Tome 64, p. 219-224.

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

[LLT12] Denis Lanier, Jean Lejeune, and 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 juin 2015.

[Oll13] Louis Ollivier. A comment on "Séralini Gilles et al. Long term toxicity of a Roundup herbicide and a Roundup-tolerant genetically modified maize. Food and Chem. Tox. (2012)". *Food and Chemical Toxicology*, 2013. Volume 53, p. 458.

[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.

[Sch63] Daniel Schwartz. *Méthodes statistiques à l'usage des médecins et des biologistes*. Editions médicales Flammarion, 1963.

[Sér12] Gilles Séralini. Long term toxicity of a Roundup herbicide and a Roundup-tolerant genetically modified maize. *Food and Chemical Toxicology*, 2012. Volume 50, Issue 11.

[TV15] Cristian Tomasetti and Bert Vogelstein. Variation in cancer risk among tissues can be explained by the number of stem cell divisions. *Science*, 2015.

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

Table des matières

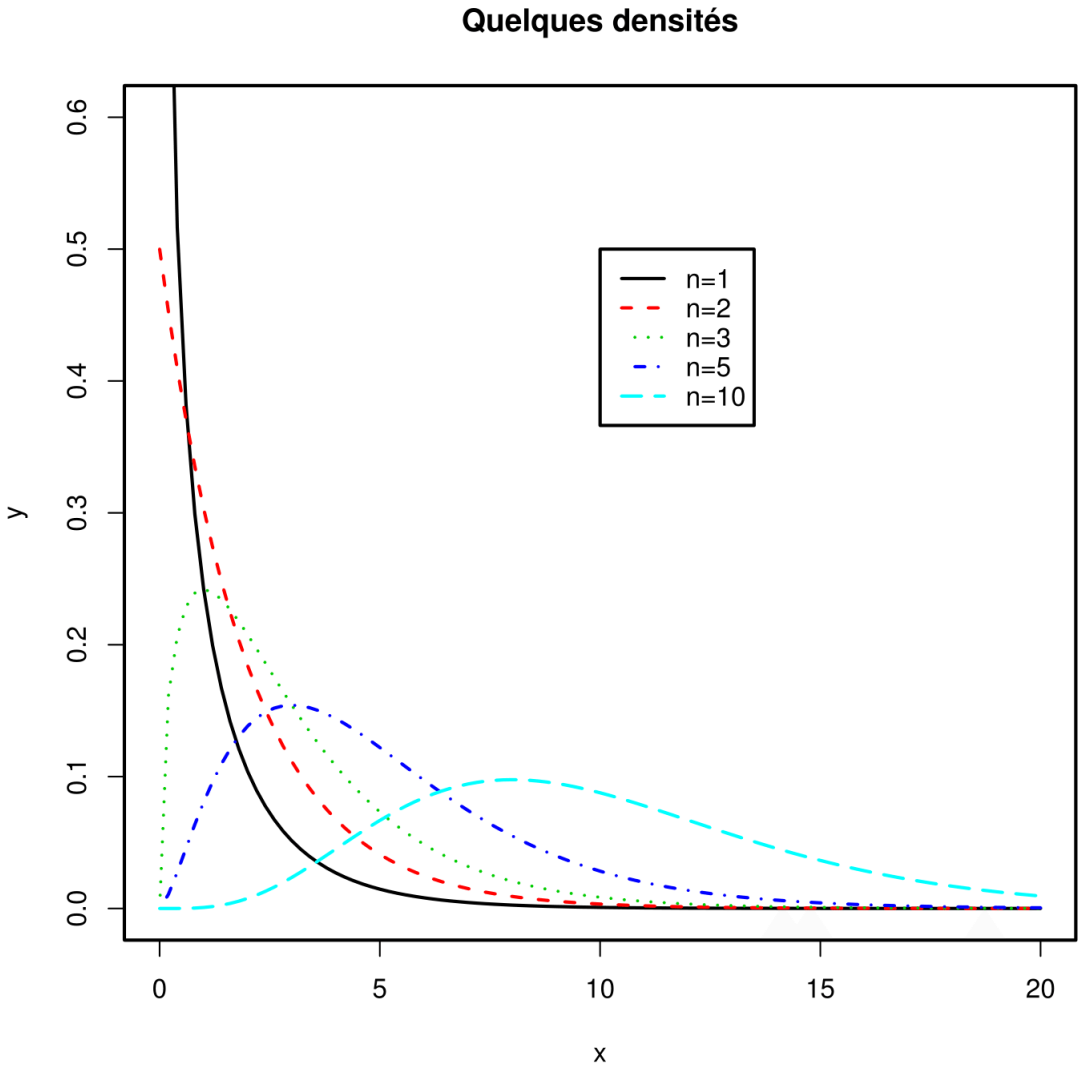
Annexe I
Définition actuelle de la loi du chi-deux

Définition :
Soient $Z_1, Z_2, ..., Z_n$, v.a. indépendantes avec $Z_j \sim \mathcal{N}(0,1)$ (loi normale centrée réduite).
$X = Z_1^2 + Z_2^2 + \cdots + Z_n^2$ suit une loi du chi-deux à n degrés de liberté.
On note $X \sim \chi^2(n)$.

Propriétés :
$\mathbb{E}(X) = n, \quad var(X) = 2n.$

Indications :
$\mathbb{E}(Z_j^2) = var(Z_j) = 1.$
$var(X) = \sum_{j=1}^{n} var(Z_j^2)$ (variables Z_j indépendantes) et $var(Z_j^2) = \mathbb{E}(Z_j^4) - (\mathbb{E}(Z_j^2)) = 3 - 1 = 2.$

Densité :
Pour $x \geqslant 0$, $f(x) = \frac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})} e^{-\frac{x}{2}} x^{\frac{n}{2}-1}$ où $\Gamma(p) = \int_0^{+\infty} e^{-x} x^{p-1} dx.$

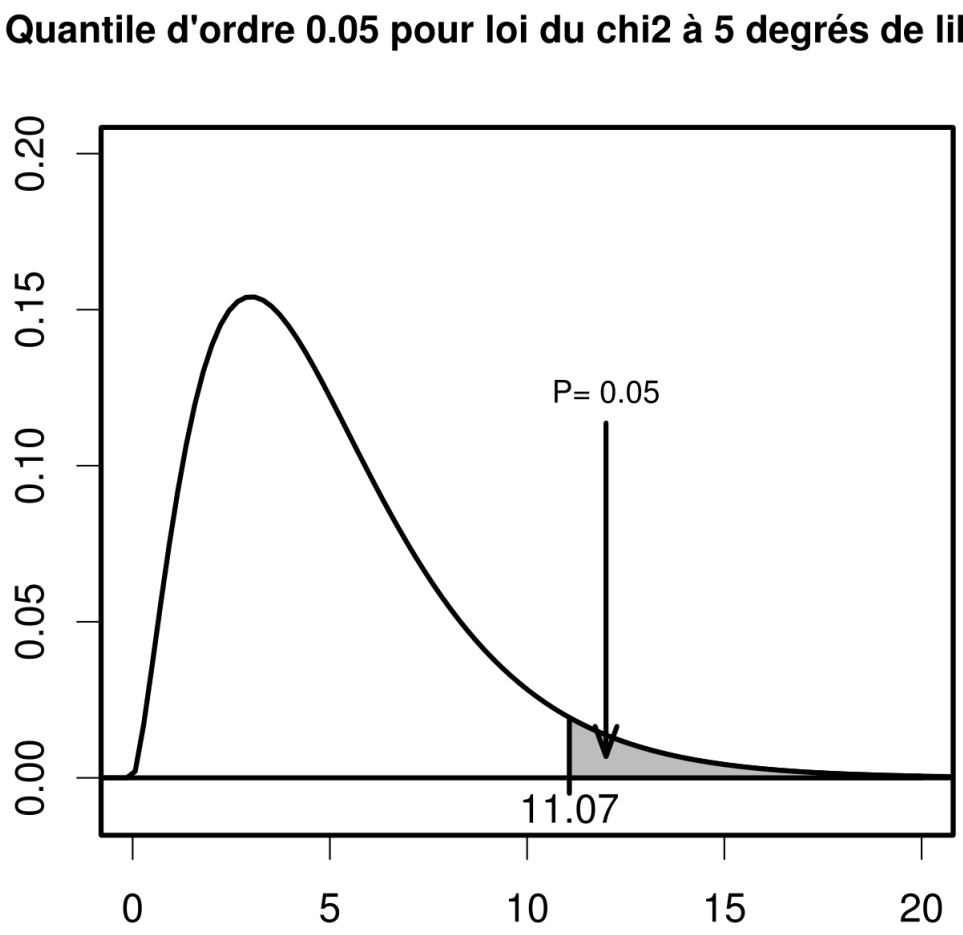


Quantile d'ordre P pour une loi du chi-deux à k degrés de liberté

Soit $0 < P < 1$ et $X \sim \chi^2(k)$.

On appelle **quantile d'ordre P pour une loi du chi-deux à k degrés de liberté** et on note $\chi^2_P(k)$ le réel tel que $\mathbb{P}(X > \chi^2_P(k)) = P$.

Exemple : Pour $k = 5$ et $P = 0{,}05$, $\mathbb{P}(X > 11{,}070) = 0{,}05$ donc $\chi^2_{0{,}05}(5) = 11{,}070$.



Retour à l'article

Annexe II
Les différentes utilisations du test du chi-deux
Définitions modernes illustrés par des exemples.

1 Test du chi-deux d'adéquation à une loi de référence appelé aussi test d'ajustement.

1.1 Cas où la loi de référence est entièrement connue.

Soit X une variable aléatoire à valeur dans un ensemble E et suivant une loi inconnue $\mathcal{L}$ et $\mathcal{L}_0$ une loi de référence complètement connue.

$E_1 \cup E_2 \cup \cdots \cup E_k$ est une partition de E.

$p_j = \mathbb{P}_{\mathcal{L}_0}(X \in E_j)$ probabilité qu'une observation appartienne à l'ensemble E_j si la loi de probabilité de X est la loi de référence $\mathcal{L}_0$.

$(X_1, X_2, ..., X_n)$ un n-échantillon de X.

On définit :

— n_j= Card $(i,\ X_i \in E_j)$ effectif observé dans E_j.

— $n_j^{'} = p_j n$ effectif espéré dans E_j.

$n_1 + n_2 + \cdots + n_k = n$ et $n_1^{'} + n_2^{'} + \cdots + n_k^{'} = n$.

La statistique de test est :

$$\chi^2_{obs} = \frac{(n_1 - n_1^{'})^2}{n_1^{'}} + \frac{(n_2 - n_2^{'})^2}{n_2^{'}} + \cdots + \frac{(n_k - n_k^{'})^2}{n_k^{'}}.$$

On teste $H_0 : \mathcal{L} = \mathcal{L}_0$ contre $H_1 : \mathcal{L} \neq \mathcal{L}_0$ au risque α.

Il y a rejet de H_0 si $\chi^2_{obs} > \chi^2_\alpha(k-1)$ (quantile de la loi du chi-deux à $k-1$ degrés de liberté, voir Annexe I), ou si la p-value définie par p-value $= P = \mathbb{P}(Y > \chi^2_{obs})$, avec $Y \sim \chi^2(k-1)$, est inférieure à α.

NB Dans l'extrait de la table de Pearson p. 8, le nombre n' correspond à la valeur k ci-dessus.

Retour à l'article

Exemple 1 tiré de l'ouvrage de Jules Gavarret [Gav40] et repris dans [LLT12, p. 30]

Travail publié en 1834 sur la marche et les effets du choléra dans le département de la Seine, relativement aux entrées des malades dans les hôpitaux de Paris [LLT12].

Jours de la semaine	Nombre de jours	Nombre de malades
Dimanches	27	1833
Lundis	27	2075
Mardis	27	1947
Mercredis	27	1978
Jeudis	27	2004
Vendredis	27	1971
Samedis	27	1969
	189	13 777

Soient $E = \{1, 2, 3, 4, 5, 6, 7\}$ les jours de la semaine. En posant $E_i = \{i\}$ pour $i = 1, 2, 3, 4, 5, 6, 7$, on obtient une partition de E en 7 classes.

La variable aléatoire X à considérer est définie ainsi : $X =$ "jour auquel arrive le malade".

L'étude porte sur un échantillon de $n = 13\,777$ malades et $X_i =$ jour auquel est arrivé le $i^{\text{ème}}$ malade $(1 \leqslant i \leqslant n)$.

L'hypothèse nulle se traduit par l'équiprobabilité du jour d'arrivée d'un malade et donc la loi de référence $\mathcal{L}_0$ est une loi uniforme définie par : $\forall j, \mathbb{P}_{\mathcal{L}_0}(X = j) = \frac{1}{7}$.

On peut construire le tableau des effectifs observés et des effectifs espérés.

j	1	2	3	4	5	6	7	Total
n_j	1833	2075	1947	1978	2004	1971	1969	13777
$n_j^{'}$	$\frac{13777}{7}$	$\frac{13777}{7}$	$\frac{13777}{7}$	$\frac{13777}{7}$	$\frac{13777}{7}$	$\frac{13777}{7}$	$\frac{13777}{7}$	13777

$\chi^2_{obs} = 16{,}01553$.

La loi de référence est entièrement connue.

E est partitionné en $k = 7$ classes donc le nombre de degrés de liberté est égal à $k - 1 = 7 - 1 = 6$.

On cherche à encadrer la p-value = P.

On a $\chi^2_{0,05}(6) = 12{,}592$ et $\chi^2_{0,01}(6) = 16{,}812$.

Puisque $12{,}592 < 16{,}01553 < 16{,}812$, on en déduit $0{,}01 < P < 0{,}05$.

Puisque $P < \alpha = 0{,}05$, on rejette l'hypothèse suivant laquelle les malades arrivent à l'hôpital avec la même probabilité quelque soit le jour de la semaine.

Cas particulier où la loi de référence est la loi de Bernoulli de paramètre p_0 connu

Soit X une variable aléatoire de loi de Bernoulli $\mathcal{B}(p)$ où $0 < p < 1$ est inconnu.

Il s'agit de tester avec le risque α :

$H_0 : p = p_0$ contre $H_1 : p \neq p_0$ ce qui équivalent à tester :

$H_0 : \mathcal{B}(p) = \mathcal{B}(p_0)$ contre $H_1 : \mathcal{B}(p) \neq \mathcal{B}(p_0)$. Ici, $\mathcal{B}(p_0) = \mathcal{L}_0$ est la loi de référence.

$p_0 = \mathbb{P}_{\mathcal{L}_0}(X = 1)$ avec $X \sim \mathcal{B}(p_0)$.

Soit $(X_1, X_2, ..., X_n)$ un n-échantillon de X.

On pose n_1= Card $(i, \ X_i = 1)$ et n_0= Card $(i, \ X_i = 0)$.

$n_1^{'} = p_0 n$ et $n_0^{'} = (1 - p_0)n = n - n_1^{'}$.

La statistique de test s'écrit : $\chi^2_{obs} = \dfrac{(n_0 - n_0^{'})^2}{n_0^{'}} + \dfrac{(n_1 - n_1^{'})^2}{n_1^{'}}$.

Puisque X suit une loi de Bernoulli, $E = \{0, 1\}$ et le nombre de classes qui partitionnent E est donc égal à $k = 2$.

Le nombre de degrés de liberté est égal à $k - 1 = 2 - 1 = 1$.

Rejet de H_0 si ${\chi_{obs}}^2 > \chi^2_\alpha(1)$ ou si la p-value est inférieure à α.

Remarque importante :

Si on pose $\widehat{f}_n = \dfrac{n_1}{n}$ = fréquence observée du nombre de 1 dans la série d'observations et

$$Z_{obs} = \frac{\widehat{f}_n - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

La p-value associée est $P = \mathbb{P}(|Z| > |Z_{obs}|)$ où $Z \sim \mathcal{N}(0, 1)$.

On peut montrer facilement que $Z^2_{obs} = \chi^2_{obs}$. Puisque $Z^2 \sim \chi^2(1)$, il en résulte que la p-value associée au test du chi-deux avec un degré de liberté est égale à celle associée au test Z.
Les conclusions seront les mêmes.

Exemple 2 inspiré de l'ouvrage de J. Gavarret [Gav40] et repris dans [LLT12, p. 29]

Parmi les 11500 cas de dysenterie constatés au Bengale de 1820 à 1825 pendant la saison chaude, 4500 l'ont été pendant la saison chaude et séche et 7000 pendant la saison chaude et humide.

En prenant le risque $\alpha = 0{,}47\%$ de se tromper, doit-on rejeter l'hypothèse par laquelle, lorsqu'on a contracté la dysenterie pendant la saison chaude, il y a la même probabilité de l'avoir contracté pendant la saison chaude et séche que pendant la saison chaude et humide ?

La variable à considérer est la suivante : $X = 1$ si la dysenterie a été contractée pendant la saison chaude et humide et $X = 0$ si la dysenterie a été contractée pendant la saison chaude et sèche.

Bien entendu $X \sim \mathcal{B}(p)$ où $0 < p < 1$ est inconnu.

Il s'agit de tester :

$H_0 : \mathcal{B}(p) = \mathcal{B}(0{,}50)$ contre $H_1 : \mathcal{B}(p) \neq \mathcal{B}(0{,}50)$ avec le risque $\alpha = 0{,}47\%$.

Ici $n_1 = 7000$ et $n_0 = 4500$.
$n_1^{'} = p_0 n = 0{,}50 \times 11500 = 5750$ et $n_0^{'} = (1 - p_0)n = n - n_1^{'} = 11500 - 5750 = 5750$.

$$\chi^2_{obs} = \frac{(4500 - 5750)^2}{5750} + \frac{(7000 - 5750)^2}{5750} = 543{,}47.$$

En examinant la 1ère ligne de la table du chi-deux, il est aisé de voir que la p-value $P \simeq 0$ donc on rejette l'hypothèse nulle. Gavarret arrive à la même conclusion.

1.2 Cas où l paramètres de la loi de référence sont inconnus.

Soit X une variable aléatoire à valeur dans un ensemble E et suivant une loi inconnue $\mathcal{L}$. $\mathcal{L}_0$ une loi de référence connue possédant l paramètres inconnus $\theta_1, \theta_2, ..., \theta_l$.

$E_1 \cup E_2 \cup \cdots \cup E_k$ est une partition de E.

$(X_1, X_2, ..., X_n)$ est un n-échantillon de X.

On définit :

— n_j= Card $(i,\ \ X_i \in E_j)$ effectif observé dans E_j.
— $\widehat{\theta_1}, \widehat{\theta_2}, ..., \widehat{\theta_l}$ estimation de $\theta_1, \theta_2, ..., \theta_l$ et $\widehat{\mathcal{L}_0}$ estimation de la loi de référence à partir de l'estimation des paramètres inconnus.
— $\widehat{p_j} = \mathbb{P}_{\widehat{\mathcal{L}_0}}(X \in E_j)$ probabilité estimée qu'une observation appartienne à l'ensemble E_j si la loi de probabilité de X est la loi de référence $\mathcal{L}_0$.
— $n_j^{'} = \widehat{p_j} n$ estimation de l'effectif espéré dans E_j.

$n_1 + n_2 + \cdots + n_k = n$ et $n_1^{'} + n_2^{'} + \cdots + n_k^{'} = n$.

La statistique de test s'écrit : $\chi^2_{obs} = \dfrac{(n_1 - n_1^{'})^2}{n_1^{'}} + \dfrac{(n_2 - n_2^{'})^2}{n_2^{'}} + \cdots + \dfrac{(n_k - n_k^{'})^2}{n_k^{'}}$.

On teste $H_0 : \mathcal{L} = \mathcal{L}_0$ contre $H_1 : \mathcal{L} \neq \mathcal{L}_0$ au risque α.

Il y a rejet de H_0 si $\chi^2_{obs} > \chi^2_\alpha(k - l - 1)$ ou si la p-value définie par p-value $= P = \mathbb{P}(Y > \chi^2_{obs})$, avec $Y \sim \chi^2(k - l - 1)$, est inférieure à α.

Exemple 3, très souvent trouvé dans la littérature ! ! !

Lors d'une journée, un vétérinaire peut avoir à traiter 0 cas grave, 1 cas grave, 2 cas graves,..., x cas graves, etc.

Soit X la variable aléatoire qui, à une journée prise au hasard, associe le nombre de cas graves traités cette journée.

Cette variable aléatoire est à valeur dans $E = \mathbb{N}$. On observe un échantillon de $n = 200$ journées dont la réalisation est une suite de x_i $i = 1, 2, ..., 200$ où $x_i \in \mathbb{N}$. En posant $E_1 = \{0\}$, $E_2 = \{1\}$, $E_3 = \{2\}$, $E_4 = \{3\}$, $E_5 = \{4\}$, $E_6 = \{5\}$, $E_7 = \{\geqslant 6\}$, on obtient une $k = 7$ partition de E.

Pour $j = 1, ..., 7$, on pose $n_j = Card(i,\ x_i \in E_j)$. Les résultats sont indiqués sur le tableau ci-dessous.

E_j	$\{0\}$	$\{1\}$	$\{2\}$	$\{3\}$	$\{4\}$	$\{5\}$	$\{\geqslant 6\}$
n_j	50	74	50	21	4	1	0

Lecture du tableau : Dans l'échantillon des 200 journées, il y a 74 journées avec 1 cas grave. Dans l'échantillon des 200 journées, il y a 1 journée avec 5 cas graves et aucune journée avec 6 ou plus de 6 cas graves.

$$\overline{x} = \frac{0 \times 50 + 1 \times 74 + 2 \times 50 + 3 \times 21 + 4 \times 4 + 5 \times 1}{200} = 1{,}29 \text{ et}$$

$$\overline{x^2} = \frac{0^2 \times 50 + 1^2 \times 74 + 2^2 \times 50 + 3^2 \times 21 + 4^2 \times 4 + 5^2 \times 1}{200} = 2{,}76$$

donc

$var(x) = \overline{x^2} - \overline{x}^2 = 1{,}09$

La variable X étant à valeur dans $\mathbb{N}$, ayant beaucoup de valeurs concentrées vers 0, 1, 2..., et la variance de l'échantillon étant proche de la moyenne, on soupçonne la variable X de suivre une loi de Poisson, c'est-à-dire :

Pour $x = 0, 1, 2, ..., \mathbb{P}(X = x) = \dfrac{e^{-\lambda}\lambda^x}{x!}$ avec $\lambda > 0$.

Puisque $\lambda = \mathbb{E}(X)$, le paramètre inconnu λ sera estimé par $\overline{x}$ moyenne des valeurs de l'échantillon.

Ici : $\overline{x} = 1{,}29$.

On peut calculer les probabilités estimées : $\widehat{p}_j = \widehat{\mathcal{L}_0}(X \in E_j) = \widehat{\mathbb{P}}(X \in E_j)$ avec $\widehat{\mathbb{P}}(X = x) = \dfrac{e^{-1{,}29}1{,}29^x}{x!}$:

$$\widehat{p}_1 = \widehat{\mathbb{P}}(X \in \{0\}) = \widehat{\mathbb{P}}(X = 0) = \frac{e^{-1{,}29}1{,}29^0}{0!} = 0{,}2753$$

$$\widehat{p}_2 = \widehat{\mathbb{P}}(X \in \{1\}) = \widehat{\mathbb{P}}(X = 1) = \frac{e^{-1{,}29}1{,}29^1}{1!} = 0{,}3551$$

$$\widehat{p}_3 = \widehat{\mathbb{P}}(X \in \{2\}) = \widehat{\mathbb{P}}(X = 2) = \frac{e^{-1{,}29}1{,}29^2}{2!} = 0{,}2290$$

$$\widehat{p}_4 = \widehat{\mathbb{P}}(X \in \{3\}) = \widehat{\mathbb{P}}(X = 3) = \frac{e^{-1{,}29}1{,}29^3}{3!} = 0{,}0985$$

$$\widehat{p}_5 = \widehat{\mathbb{P}}(X \in \{4\}) = \widehat{\mathbb{P}}(X = 4) = \frac{e^{-1{,}29}1{,}29^4}{4!} = 0{,}0318$$

$$\widehat{p}_6 = \widehat{\mathbb{P}}(X \in \{5\}) = \widehat{\mathbb{P}}(X = 5) = \frac{e^{-1{,}29}1{,}29^5}{5!} = 0{,}0082$$

$$\widehat{p}_7 = \widehat{\mathbb{P}}(X \in \{\geqslant 6\}) = 1 - \sum_{j=1}^{6} \widehat{p}_j = 0{,}0021$$

On calcule aussi l'estimation des effectifs espérés : $n_j^{'} = 200 \times \widehat{p}_j$ pour $j = 1, 2, ..., 7$

j	1	2	3	4	5	6	7	Total
n_j	50	74	50	21	4	1	0	200
$n_j^{'}$	55,06	71,02	45,80	19,70	6,36	1,64	0,42	200

Puisque les effectifs espérés sont inférieurs à 5 pour les deux dernières classes, on regroupe les classes 5, 6 et 7 pour obtenir le tableau suivant :

j	1	2	3	4	5	Total
n_j	50	74	50	21	5	200
$n_j^{'}$	55,06	71,02	45,80	19,70	8,42	200

La valeur de la statistique de test est :

$$\chi^2_{obs} = \frac{(50 - 55{,}06)^2}{55{,}06} + \frac{(74 - 71{,}02)^2}{71{,}02} + \frac{(50 - 45{,}80)^2}{45{,}80} + \frac{(21 - 19{,}70)^2}{19{,}70} + \frac{(5 - 8{,}42)^2}{8{,}42} = 2{,}45$$

Le nombre de classes qui partitionnent E est alors $k = 5$.

Le nombre de paramètres inconnus de la loi de référence est $l = 1$ puisqu'il a fallu estimer λ.

Le nombre de degrés de liberté est donc $k - l - 1 = 5 - 1 - 1 = 3$.

On cherche à encadrer la p-value = $P = \mathbb{P}(Y \geqslant 2{,}45)$ où Y suit une loi du chi-deux à 3 degrés de liberté. Sur la ligne $n = 3$ de la table, 2,45 est compris entre 2,366 qui correspond à la colonne $P = 0{,}50$ et 3,665 qui correspond à la colonne $P = 0{,}30$, donc $0{,}30 < P < 0{,}50$.

Décision : puisque $P > 0{,}05$, on ne rejette pas l'hypothèse suivant laquelle la loi du nombre de cas graves est une loi de Poisson.

2 Test du chi-deux de comparaison de deux populations ou plus, dénommé aussi test d'homogénéité.

Soient r populations données ($r \geqslant 2$). Soit E un ensemble de m modalités.

Pour chaque population l ($1 \leqslant l \leqslant r$), on considère une variable qualitative $X^{(l)}$ à valeur dans cet ensemble de m modalités dont la loi de probabilité est donnée par m nombres réels p_{jl} ($1 \leqslant j \leqslant m$) définis par $p_{jl} = \mathbb{P}(X^{(l)} = j)$.

On a bien sûr $0 \leqslant p_{jl} \leqslant 1$ et $\sum_{j=1}^{m} p_{jl} = 1$.

L'hypothèse nulle H_0 est celle d'une égalité des lois des r variables aléatoires $X^{(l)}$ ($1 \leqslant l \leqslant r$), ce qui se traduit par :

$\forall l,\ \forall l^{'},\ \forall j,\ 1 \leqslant j \leqslant m,\ p_{jl} = p_{jl'}$ et donc qu'il existe une suite de m nombres réels $p_{j.}$ inconnus tels que $\forall l,\ \forall l^{'},\ \forall j,\ p_{jl} = p_{jl'} = p_{j.}$, cette suite définissant la loi commune des r variables aléatoires $X^{(l)}$.

L'hypothèse alternative H_1 est celle-ci : il existe au moins deux populations l et $l^{'}$ dont les variables aléatoires $X^{(l)}$ et $X^{(l^{'})}$ n'ont pas la même loi de probabilité.

On souhaite effectuer le test H_0 contre H_1 avec le risque α.

Pour chaque population l on tire un échantillon de taille $n_{.l}$.

n_{jl} = est le nombre d'observations pour lesquelles les $n_{.l}$ réalisations de la variable $X^{(l)}$ prennent la modalité j.

Soit $n_{..} = \sum_{l=1}^{r} n_{.l}$ le nombre total d'observations.

$n_{j.}$ = nombre total d'observations qui prennent la modalité j parmi les $n_{..}$.

$n_{..} = \sum_{l=1}^{r} n_{.l} = \sum_{j=1}^{m} n_{j.} = \sum_{l=1}^{r}\sum_{j=1}^{m} n_{jl}$.

Tableau des résultats :

	1	2	$\cdots$	j	$\cdots$	m	
1	n_{11}	n_{21}	$\cdots$	n_{j1}	$\cdots$	n_{m1}	$n_{.1}$
2	n_{12}	n_{22}	$\cdots$	n_{j2}	$\cdots$	n_{m2}	$n_{.2}$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
l	n_{1l}	n_{2l}	$\cdots$	n_{jl}	$\cdots$	n_{ml}	$n_{.l}$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
r	n_{1r}	n_{2r}	$\cdots$	n_{jr}	$\cdots$	n_{mr}	$n_{.r}$
	$n_{1.}$	$n_{2.}$	$\cdots$	$n_{j.}$	$\cdots$	$n_{m.}$	$n_{..}$

Les effectifs espérés sont définis par : $n^{'}_{jl} = \dfrac{n_{j.}n_{.l}}{n_{..}}$.

Le calcul de la statistique de test est alors : $\chi^2_{obs} = \sum_{j=1}^{m}\sum_{l=1}^{r} \dfrac{\left(n_{jl} - n^{'}_{jl}\right)^2}{n^{'}_{jl}}$.

On peut démontrer que le nombre de degrés de liberté à considérer pour un tableau à r lignes et m colonnes est égal à : $n = (m-1)(r-1)$.

Rejet de H_0 c'est-à-dire de l'égalité des lois de probabilité si $\chi^2_{obs} > \chi^2_{\alpha}\big((m-1)(r-1)\big)$ ou si la p-value définie par p-value $= P = \mathbb{P}(Y > \chi^2_{obs})$, avec $Y \sim \chi^2\big((m-1)(r-1)\big)$, est inférieure à α.

Remarque : Pour appliquer ce test, l'usage veut que la condition suivante soit vérifiée : chacun des effectifs **espérés** du tableau doit être supérieur à 5.

Exemple 4 de comparaison de deux populations tiré de l'article de Wynder et Graham

Les données sont celles qui figurent au paragraphe 2.6 du présent article.

Il est aisé de voir qu'il s'agit d'une comparaison de $r = 2$ populations, la population avec cancer du poumon (AC) et la population sans cancer du poumon (SC).

L'ensemble E est celui des 6 groupes de comportement tabagique (Groupe 0, Groupe 1, ..., Groupe 5).

La variable aléatoire $X^{(1)}$ (resp. $X^{(2)}$) est définie ainsi :

Pour $1 \leqslant j \leqslant 6$, $X^{(1)} = j$ (resp.$X^{(2)} = j$) si l'homme interrogé ayant un cancer du poumon (resp. sans cancer du poumon) se situe dans le Groupe $(j-1)$ quant à son habitude tabagique. Pour chacune des variables, il y a donc $m = 6$ modalités.

Dire alors que l'hypothèse nulle H_0 est vraie signifie que les lois de probabilités de $X^{(1)}$ et de $X^{(2)}$ sont les mêmes et donc que les deux populations (hommes ayant le cancer du poumon (AC) et hommes n'ayant pas le cancer du poumon(SC)) sont homogènes au regard de leur habitude tabagique.

Le tableau des effectifs observés est du type de tableau ci-dessus avec par exemple : $n_{31} = 61$ ou encore $n_{12} = 114$ mais aussi :

$n_{1.} = 122$, $n_{2.} = 104$, $n_{3.} = 209$, $n_{4.} = 591$, $n_{5.} = 277$, $n_{6.} = 182$ et $n_{.1} = 605$, $n_{.2} = 780$ et bien sûr $n_{..} = 1385$.

Le tableau des effectifs espérés utilise la formule :

$n'_{jl} = \dfrac{n_{j.}n_{.l}}{n_{..}}$.

Ainsi : $n'_{31} = \dfrac{n_{3.}n_{.1}}{n_{..}} = \dfrac{209 \times 605}{1385} = 91{,}29603$.

Le calcul de la statistique de test est alors :

$$\chi^2_{obs} = \sum_{j=1}^{6}\sum_{l=1}^{2} \frac{\left(n_{jl} - n'_{jl}\right)^2}{n'_{jl}} = \frac{\left(8 - 53{,}29242\right)^2}{53{,}29242} + \frac{\left(114 - 68{,}70758\right)^2}{68{,}70758} + \frac{\left(14 - 45{,}42960\right)^2}{45{,}42960} + \cdots + \frac{\left(60 - 102{,}49819\right)^2}{102{,}49819} = 229{,}091$$

Ici le nombre de degrés de liberté à considérer est égal à $n = (m-1)(r-1) = (6-1)(2-1) = 5$.

Nous rappelons la conclusion déjà donnée.

On cherche à évaluer la p-value $P = \mathbb{P}(Y \geqslant 229{,}091)$ = où Y suit une loi du chi-deux à 5 degrés de liberté. Sur la ligne $n = 5$ de la table, $229{,}091$ est supérieur à $15{,}086$ qui correspond à la colonne $P = 0{,}01$ donc $P < 0{,}01$.

En prenant le risque $\alpha = 0{,}05$, puisque $P < 0{,}01 < 0{,}05$, l'hypothèse nulle est rejetée. Les populations ne sont pas homogènes au regard de leur habitude tabagique.

Exemple 5 de comparaison de deux proportions : retour au texte de Gavarret repris dans [LLT, p. 23]

Gavarret compare 2 médications appliquées à 2 groupes de malades.

Sur 500 personnes à qui on a appliqué la 1ère médication, 400 ont été guéris.

Sur 500 autres personnes à qui on a appliqué la 2ème médication, 370 ont été guéris.

Soit p_1(resp. p_2) la probabilité d'être guéri avec la 1ère (resp. la 2ème) médication.

Il s'agit de tester avec le risque $\alpha = 0{,}47\%$:

$H_0 : p_1 = p_2$ contre $H_1 : p_1 \neq p_2$

La population 1 (resp. population 2) est celle à qui on a appliqué la première médication (resp. la seconde médication).

La variable aléatoire $X^{(1)}$ (resp. $X^{(2)}$) est définie ainsi :

— $X^{(1)} = 1$ (resp. $X^{(2)} = 1$) si la personne à qui on a appliqué la première médication (resp. la seconde médication) est guérie.

— $X^{(1)} = 0$ (resp. $X^{(2)} = 0$) si la personne à qui on a appliqué la première médication (resp. la seconde médication) est morte.

Dire alors que l'hypothèse nulle H_0 est vraie signifie que les deux populations sont homogènes au regard du résultat des deux médications et donc que les lois de probabilités de $X^{(1)}$ et de $X^{(2)}$ (lois de Bernoulli) sont les mêmes.

Le tableau de données est celui-ci :

	Morts	Guéris	Total
1ère médication	100	400	500
2ème médication	130	370	500
Total	230	770	1000

La valeur du chi-deux observé est :
$\chi^2_{obs} = 5{,}0819$

Il y a $r = 2$ populations (Médication 1/ Médication 2) et $m = 2$ modalités (Mort/Guéri), donc le nombre de degrés de liberté à considérer est égal à $n = (m-1)(r-1) = (2-1)(2-1) = 1$.

On cherche à évaluer la p-value $= P = \mathbb{P}(Y \geqslant 5{,}0819)$ où Y suit une loi du chi-deux à 1 degré de liberté.

Sur la ligne $n = 1$ de la table, $5{,}0819$ est compris entre $3{,}841$ qui correspond à la colonne $P = 0{,}05$ et $5{,}412$ qui correspond à la colonne $P = 0{,}02$, donc $0{,}02 < P < 0{,}05$.

Puisque $0{,}0047 < P$, l'hypothèse nulle n'est pas rejetée. On ne peut pas conclure à une différence entre les proportions de malades guéris entre les deux médications. C'est aussi la conclusion de Gavaret.

Exemple 6 : Comparaison de plusieurs proportions

On cherche à savoir si les cinq régions suivantes : Basse-Normandie (BN), Haute-Normandie (HN), Pays-de-Loire (PL), Bretagne (B) et Centre (C) sont homogènes quant à la prévalence du tabagisme des personnes agées de 17 ans en prenant le risque $\alpha = 0{,}05$.

Si on pose : p_R la proportion (prévalence) de jeunes de 17 ans habitant la région R qui fument quotidiennement , il s'agit de tester en prenant le risque $\alpha = 0{,}05$:

H_0 : $p_{BN} = p_{HN} = p_{PL} = p_B = p_C$ contre H_1 : "il existe au moins deux régions où la prévalence est différente"

Une enquête datant de 2002-2003 où ont été interrogés des jeunes de 17 ans sur le fait de savoir s'ils fumaient quotidiennement a donné les résultats suivants :

	Oui	Non	Nombre d'enquêtés	Prévalence observée
Basse-Normandie	268	342	610	44%
Haute-Normandie	218	314	532	41%
Pays-de-Loire	314	416	730	43%
Bretagne	329	356	685	48%
Centre	187	293	480	39%
Total des cinq régions	1316	1721	3037	43,3%

Remarque : Les données ne sont pas celles de l'enquête mais les prévalences observées sur les échantillons correspondent à celles fournies par le document
http ://www.fnors.org/fnors/ors/travaux/addictions.pdf

La valeur du chi-deux observé est : $\chi^2_{obs} = 11{,}2155$.

Il y a $r = 5$ populations (les 5 régions) et $m = 2$ modalités (Oui/Non) , donc le nombre de degrés de liberté à considérer est égal à $(m-1)(r-1) = (2-1)(5-1) = 4$.

On cherche à évaluer la p-value $= P = \mathbb{P}(Y \geqslant 11{,}2155)$ où Y suit une loi du chi-deux à 4 degrés de liberté.

Sur la ligne $n = 4$ de la table, $11{,}2155$ est compris entre $9{,}488$ qui correspond à la colonne $P = 0{,}05$ et $11{,}688$ qui correspond à la colonne $P = 0{,}02$, donc $0{,}02 < P < 0{,}05$.

Puisque $P < 0{,}05$, l'hypothèse nulle d'égalité des prévalences est rejetée. Les cinq régions ne sont pas homogènes quant à la prévalence du tabagisme des 17 ans.

3 Test du chi-deux d'indépendance.

Lorsque l'on cherche à savoir s'il existe une dépendance entre deux variables qualitatives X et Y, X ayant m modalités et Y ayant r modalités, la technique enseignée en France dans les U. F. R. de Médecine est une généralisation de la méthode utilisée ci-dessus.

On pose :

n_{jl} = nombre d'observations pour lesquelles la variable X prend la modalité j et la variable Y prend la modalité l.

$n_{j.}$ = nombre d'observations pour lesquelles la variable X prend la modalité j.

$n_{.l}$ = nombre d'observations pour lesquelles la variable Y prend la modalité l.

$n_{..}$ nombre total d'observations.

Tableau des observations :

	$(X=1)$	$(X=2)$	$\cdots$	$(X=j)$	$\cdots$	$(X=m)$	
$(Y=1)$	n_{11}	n_{21}	$\cdots$	n_{j1}	$\cdots$	n_{m1}	$n_{.1}$
$(Y=2)$	n_{12}	n_{22}	$\cdots$	n_{j2}	$\cdots$	n_{m2}	$n_{.2}$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
$(Y=l)$	n_{1l}	n_{2l}	$\cdots$	n_{jl}	$\cdots$	n_{ml}	$n_{.l}$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
$(Y=r)$	n_{1r}	n_{2r}	$\cdots$	n_{jr}	$\cdots$	n_{mr}	$n_{.r}$
	$n_{1.}$	$n_{2.}$	$\cdots$	$n_{j.}$	$\cdots$	$n_{m.}$	$n_{..}$

Les effectifs espérés sont définis par : $n'_{jl} = \dfrac{n_{j.}n_{.l}}{n_{..}}$,

et la statistique de test par : $\chi^2_{obs} = \displaystyle\sum_{j=1}^{m}\sum_{l=1}^{r} \frac{\left(n_{jl} - n'_{jl}\right)^2}{n'_{jl}}$.

Le nombre de degrés de liberté est : $n = (m-1)(r-1)$.

Soit W une variable aléatoire qui suit une loi du chi-deux à $(m-1)(r-1)$ degrés de liberté.

On définit la p-value par : p-value $= P = \mathbb{P}(W > {\chi^2}_{obs})$.

À l'aide de la table, il est possible d'encadrer cette p-value et de conclure.

Rejet de l'indépendance si $\chi^2_{obs} > \chi^2_\alpha\big((m-1)(r-1)\big)$ ou si p-value inférieure à α.

Remarque : Pour appliquer ce test, l'usage veut que la condition suivante soit vérifiée : chacun des effectifs **espérés** du tableau doit être supérieur à 5. Dans le cas contraire, on effectue des regroupements de classes.

Exemple 7 : 200 nouveaux-nés ont été observés. Un nouveau-né est issu d'une femme primipare si celle-ci accouche pour la première fois et il est issu d'une femme multipare si celle-ci a accouché plus d'une fois. Tester avec le risque $\alpha = 0{,}05$, l'indépendance entre la parité et le poids d'un nouveau-né à partir du tableau suivant tiré du livre *"Méthodes statistiques à l'usage des médecins et biologistes"* de D. Schwartz [Sch63], p. 81.

Poids	Primipares	Multipares	Total
$<$ 3 kg	26	20	46
Entre 3 et 4 kg	61	63	124
$>$ 4 kg	8	22	30
Total	95	105	200

Ici, il n'y a qu'un population observée, celle des nouveau-nés.

Les variables aléatoires X et Y sont définies ainsi :

— $X = 1$ si le nouveau-né est issu d'une femme primipare et $X = 2$ si le nouveau-né est issu d'une femme multipare.

— $Y = 1$ si son poids est inférieur à 3 kg, $Y = 2$ si son poids est compris entre 3 et 4 kg, $Y = 3$ si son poids est supérieur à 4 kg.

Avec les notations générales pour ce type de problème, on a par exemple : $n_{11} = 26$ = nombre de nouveau-nés de l'échantillon issus d'une mère primipare **et** pesant moins de 3 kg ou encore $n_{23} = 30$ = nombre de nouveau-nés de l'échantillon issus d'une mère multipare **et** pesant plus de 4 kg mais aussi :

$n_{1.} = 95$, $n_{2.} = 105$ et $n_{.1} = 46$, $n_{.2} = 124$, $n_{.3} = 30$ et bien sûr $n_{..} = 200$.

Le tableau des effectifs espérés utilise la formule :

$n'_{jl} = \dfrac{n_{j.}n_{.l}}{n_{..}}$.

Ainsi : $n'_{23} = \dfrac{n_{2.}n_{.3}}{n_{..}} = \dfrac{105 \times 30}{200} = 15{,}75$.

Le tableau complet des effectifs espérés est ainsi calculé :

Poids	Primipares	Multipares	Total
$<$ 3kg	21,85	24,15	46
Entre 3 et 4 kg	58,90	65,10	124
$>$ 4kg	14,25	15,75	30
Total	95	105	200

Le calcul de la statistique de test est alors :

$$\chi^2_{obs} = \sum_{j=1}^{2}\sum_{l=1}^{3} \frac{\left(n_{jl} - n'_{jl}\right)^2}{n'_{jl}} = \frac{\left(26 - 21{,}85\right)^2}{21{,}85} + \frac{\left(61 - 58{,}90\right)^2}{58{,}90} + \cdots + \frac{\left(22 - 15{,}75\right)^2}{15{,}75} = 6{,}865364$$

Ici le nombre de degrés de liberté à considérer est égal à $n = (m-1)(r-1) = (2-1)(3-1) = 2$.

On cherche à évaluer la p-value $= P = \mathbb{P}(Y \geqslant 6{,}865364)$ où Y suit une loi du chi-deux à 2 degrés de liberté.

Sur la ligne $n = 2$ de la table, $6{,}865364$ est compris entre $5{,}991$ qui correspond à la colonne $P = 0{,}05$ et $7{,}924$ qui correspond à la colonne $P = 0{,}02$, donc $0{,}02 < P < 0{,}05$.

Avec le risque $\alpha = 5\%$ de se tromper, on rejette l'hypothèse par laquelle les deux variables sont indépendantes.

Remarque importante

Si, pour les exemples 4, 5, 6 et 7,les calculs effectués semblent les mêmes, la collecte des données observées est différente. Alors que, dans les exemples 4, 5 et 6, seule une des variables était aléatoire, dans l'exemple 7, ce sont les deux variables qui sont aléatoires et les conclusions sont alors différentes.

Activité 1

Montrer, à l'aide d'un calcul algébrique, que de façon générale : $\chi^2_{obs} = Z^2_{obs}$.

Retour à l'article

Une solution de cette activité

Activité 2

Pour $n = 3$ degrés de liberté, la table de Fisher donne, au croisement de la ligne $n = 3$ et de la colonne $P = 0{,}01$ (resp. $P = 0{,}05$, $P = 0{,}1$), la valeur du fractile $\chi^2_{0,01}(3) = 11{,}341$ (resp. $\chi^2_{0,05}(3) = 7{,}815$, $\chi^2_{0,10}(3) = 6{,}251$).

Retrouver, par interpolation, approximativement ces valeurs à partir de l'extrait de la table de Pearson ci-dessous :

n'

χ^2	3.	4.	5.	6.
1	·606,531	·801,253	·909,796	·962,566
2	·367,879	·572,407	·735,759	·849,146
3	·223,130	·391,633	·557,825	·699,994
4	·135,335	·261,470	·406,006	·549,422
5	·082,085	·171,799	·287,298	·415,882
6	·049,787	·111,611	·199,148	·306,220
7	·030,197	·071,888	·135,888	·220,631
8	·018,316	·046,012	·091,578	·156,236
9	·011,109	·029,291	·061,099	·109,064
10	·006,738	·018,567	·040,428	·075,236
15	·000,553	·001,817	·004,701	·010,363
20	·000,045	·000,170	·000,499	·001,250

Retour à l'article

Une solution de cette activité

Activité 3

Montrer que, de façon générale, le chi-deux calculé avec la formule simple :

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

est le même que le chi-deux calculé avec la formule de Pearson.

Retour à l'article

Une solution de cette activité

Activité 4

Dans l'article de Doll et Hill, se trouve le tableau II [DH50, p. 741] où figurent les données relatives au sexe, à l'âge, à la classe sociale et au lieu de résidence des patients atteints du cancer et des patients témoins.

TABLE II.—*Comparison Between Lung-carcinoma Patients and Non-cancer Patients Selected as Controls, With Regard to Sex, Age, Social Class, and Place of Residence*

Age	No. of Lung-carcinoma Patients		No. of Non-cancer Control Patients	
	M	F	M	F
25– ..	2	1	2	1
30– ..	6	0	6	0
35– ..	18	3	18	3
40– ..	36	4	36	4
45– ..	87	10	87	10
50– ..	130	11	130	11
55– ..	145	9	145	9
60– ..	109	9	109	9
65– ..	88	9	89*	9
70–74 ..	28	4	27*	4
All ages	649	60	649	60

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

* One control patient was selected, in error, from the wrong age group.

Extraire de ce tableau les données relatives à la **classe sociale**, pour les **hommes**.
Représenter graphiquement les deux séries de données obtenues sous forme d'histogrammes.
Que remarquez-vous ?
Doll et Hill avaient conclu que les différences observées pouvaient être dues au hasard.
Calculer le chi-deux. Le résultat du test justifie-t-il cette conclusion ?

Retour à l'article

Une solution de cette activité

Activité 5

Dans l'article de Doll et Hill, se trouve le tableau II [DH50, p. 741] où figurent les données relatives au sexe, à l'âge, à la classe sociale et au lieu de résidence des patients atteints du cancer et des patients témoins.

TABLE II.—*Comparison Between Lung-carcinoma Patients and Non-cancer Patients Selected as Controls, With Regard to Sex, Age, Social Class, and Place of Residence*

Age	No. of Lung-carcinoma Patients		No. of Non-cancer Control Patients	
	M	F	M	F
25– ..	2	1	2	1
30– ..	6	0	6	0
35– ..	18	3	18	3
40– ..	36	4	36	4
45– ..	87	10	87	10
50– ..	130	11	130	11
55– ..	145	9	145	9
60– ..	109	9	109	9
65– ..	88	9	89*	9
70–74 ..	28	4	27*	4
All ages	649	60	649	60

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

* One control patient was selected, in error, from the wrong age group.

Extraire de ce tableau les données relatives au **lieu de résidence**, pour les **hommes**.
Représenter graphiquement les deux séries de données obtenues sous forme d'histogrammes.
Que remarquez-vous ?
Doll et Hill avaient conclu que les différences observées pouvaient être dues au hasard.
Calculer le chi-deux. Le résultat du test justifie-t-il cette conclusion ?

Retour à l'article

Une solution de cette activité

Activité 6

Dans l'article de Doll et Hill, se trouve le tableau IV [DH50, p. 742] où sont séparées les données sur les hommes et 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 $0{\cdot}01 < P < 0{\cdot}02$
Control patients with diseases other than cancer (60) ..	32 (53·3%)	28	

Calculer le χ^2_{obs} pour les **hommes**.

Peut-on en déduire, avec le risque $\alpha = 5\%$ de se tromper, que :

1. la probabilité d'avoir un cancer du poumon quand on est fumeur est différente de la probabilité d'avoir un cancer du poumon quand on est non-fumeur ?
2. la probabilité d'être fumeur quand on a un cancer du poumon est différente de la probabilité d'être fumeur quand on n'a pas de cancer du poumon ?
3. il existe un lien entre la consommation de tabac et le cancer du poumon ?

Retour à l'article

Une solution de cette activité

Activité 7

Dans l'article de Doll et Hill, se trouve le tableau IV [DH50, p. 742] où sont séparées les données sur les hommes et 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 $0{\cdot}01 < P < 0{\cdot}02$
Control patients with diseases other than cancer (60) ..	32 (53·3%)	28	

Calculer le χ^2_{obs} pour les **femmes**.

Peut-on en déduire, avec le risque $\alpha = 5\%$ de se tromper, que :

1. la probabilité d'avoir un cancer du poumon quand on est fumeur est différente de la probabilité d'avoir un cancer du poumon quand on est non-fumeur ?
2. la probabilité d'être fumeur quand on a un cancer du poumon est différente de la probabilité d'être fumeur quand on n'a pas de cancer du poumon ?
3. il existe un lien entre la consommation de tabac et le cancer du poumon ?

Retour à l'article

Une solution de cette activité

Activité 8

Dans l'article de Doll et Hill, se trouve le tableau V [DH50, p. 742] qui permet de distinguer les fumeurs lourds des fumeur légers. Les quantités de cigarettes indiquées dans le tableau sont celles fumées immédiatement avant le début de la maladie.

TABLE V.—*Most Recent Amount of Tobacco* Consumed Regularly by Smokers Before the Onset of Present Illness; Lung-carcinoma Patients and Control Patients with Diseases Other Than Cancer*

Disease Group	No. Smoking Daily					Probability Test
	1 Cig.–*	5 Cigs.–	15 Cigs.–	25 Cigs.–	50 Cigs. +	
Males:						
Lung-carcinoma patients (647)	33 (5·1%)	250 (38·6%)	196 (30·3%)	136 (21·0%)	32 (5·0%)	$\chi^2=36{\cdot}95$; $n=4$; $P<0{\cdot}001$
Control patients with diseases other than cancer (622) . .	55 (8·8%)	293 (47·1%)	190 (30·5%)	71 (11·4%)	13 (2·1%)	
Females:						
Lung-carcinoma patients (41) . .	7 (17·1%)	19 (46·3%)	9 (22·0%)	6 (14·6%)	0 (0·0%)	$\chi^2=5{\cdot}72$; $n=2$; $0{\cdot}05<P<0{\cdot}10$ (Women smoking 15 or more cigarettes a day grouped together)
Control patients with diseases other than cancer (28) . .	12 (42·9%)	10 (35·7%)	6 (21·4%)	0 (0·0%)	0 (0·0%)	

* Ounces of tobacco have been expressed as being equivalent to so many cigarettes. There is 1 oz. of tobacco in 26·5 normal-size cigarettes, so that the conversion factor has been taken as: 1 oz. of tobacco a week = 4 cigarettes a day.

1. Vérifier les valeurs des chi-deux et les encadrements obtenus pour P dans le tableau V.
2. Quelles conclusions peut-on en tirer ?

Retour à l'article

Une solution de cette activité

Activité 9

Sur 688 malades atteints du cancer du poumon, $61{,}6\%$ inhalaient la fumée tandis que sur 650 malades témoins, ils étaient $67{,}2\%$. Un test statistique avait permis à Doll et Hill d'écrire que les malades atteints du cancer du poumon inhalaient légèrement moins souvent la fumée que les autres patients.

À partir des pourcentages annoncés, effectuer ce test et vérifier les conclusions de Doll et Hill.

Retour à l'article

Une solution de cette activité

Activité 10

Dans l'étude dirigée par Séralini, tous les échantillons sont composés du même nombre de rats, à savoir $n = 10$, et le nombre de traitements différents est toujours égal à $k = 4$. Voici, par exemple, le cas des rats mâles recevant une nourriture partiellement OGM :

	OGM Mâles		
Traitement	Morts (x_i)	Vivants ($n - x_i$)	Total
0% (contrôle)	3	7	**10**
11% OGM	5	5	**10**
22% OGM	1	9	**10**
33% OGM	1	9	**10**
Somme	10		
Somme des Carrés	36		
chi-deux	5,87		

1. Vérifier le résultat du chi-deux du tableau ci-dessus avec la formule usuelle.

Dans ce cas particulier, il est possible d'utiliser la formule de Brandt et Snedecor pour calculer la valeur du chi-deux :

$$\chi^2_{obs} = \frac{k.\sum x_i^2 - \left(\sum x_i\right)^2}{\sum x_i - \frac{\left(\sum x_i\right)^2}{n.k}}$$

2. Vérifier le résultat du chi-deux du tableau avec la formule de Brandt et Snedecor.

3. Démontrer la formule de Brandt et Snedecor.

Retour à l'article

Une solution de cette activité

Une solution de l'activité 1

La statistique du χ^2 a été définie par :

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

Il s'agit de prouver l'égalité :

$$\boxed{\chi^2_{obs} = Z^2_{obs} = \left(\frac{f_1 - f_0}{\sqrt{\frac{1}{L_1} + \frac{1}{L_0}}\sqrt{f^*(1-f^*)}} \right)^2}$$

Notons comme précédemment, $L_1 = a + b$, $L_0 = c + d$ et $T = a + b + c + d = L_0 + L_1$.

Les valeurs a', b', c' et d' ont été définies par :

$$a' = \frac{L_1(a+c)}{T}, b' = \frac{L_1(b+d)}{T}, c' = \frac{L_0(a+c)}{T} \text{ et } d' = \frac{L_0(b+d)}{T},$$

et les valeurs f_0, f_1 et f^* par : $f_1 = \frac{a}{L_1}, f_0 = \frac{c}{L_0}$ et $f^* = \frac{a+c}{T}$.

Nous avons : $a - a' = a - \frac{L_1(a+c)}{T} = \frac{aT - L_1(a+c)}{T} = \frac{a(L_0+L_1) - L_1(a+c)}{T} = \frac{aL_0 - cL_1}{T}$,

c'est-à-dire, $a - a' = \frac{f_1 L_1 L_0 - f_0 L_1 L_0}{T} = \frac{(f_1 - f_0)L_1 L_0}{T} = (f_1 - f_0)\frac{L_1 L_0}{L_1 + L_0} = \frac{f_1 - f_0}{\frac{1}{L_1} + \frac{1}{L_0}}$.

Nous obtenons de manière analogue :

$$b - b' = \frac{f_0 - f_1}{\frac{1}{L_1} + \frac{1}{L_0}}, c - c' = \frac{f_0 - f_1}{\frac{1}{L_1} + \frac{1}{L_0}} \text{ et } d - d' = \frac{f_1 - f_0}{\frac{1}{L_1} + \frac{1}{L_0}}.$$

Nous pouvons maintenant calculer χ^2_{obs}.

$$\begin{aligned}
\chi^2_{obs} &= \frac{(a-a')^2}{a'} + \frac{(b-b')^2}{b'} + \frac{(c-c')^2}{c'} + \frac{(d-d')^2}{d'} \\
&= \frac{(f_1-f_0)^2}{\left(\frac{1}{L_1} + \frac{1}{L_0}\right)^2}\left[\frac{1}{a'} + \frac{1}{b'} + \frac{1}{c'} + \frac{1}{d'}\right] \\
&= \frac{(f_1-f_0)^2}{\left(\frac{1}{L_1} + \frac{1}{L_0}\right)^2}\left[\frac{T}{L_1(a+c)} + \frac{T}{L_1(b+d)} + \frac{T}{L_0(a+c)} + \frac{T}{L_0(b+d)}\right] \\
&= \frac{(f_1-f_0)^2}{\left(\frac{1}{L_1} + \frac{1}{L_0}\right)^2}\left[\frac{TL_0(b+d) + TL_0(a+c) + TL_1(b+d) + TL_1(a+c)}{L_1L_0(a+c)(b+d)}\right]
\end{aligned}$$

... /...

Soit,

$$\begin{aligned}
\chi^2_{obs} &= \frac{(f_1-f_0)^2}{\left(\frac{1}{L_1}+\frac{1}{L_0}\right)^2}\left[\frac{TL_0(b+d)+TL_0(a+c)+TL_1(b+d)+TL_1(a+c)}{L_1L_0(a+c)(b+d)}\right] \\
&= \frac{(f_1-f_0)^2}{\left(\frac{1}{L_1}+\frac{1}{L_0}\right)^2}\times\frac{T(L_0+L_1)\times(a+b+c+d)}{L_1L_0(a+c)(b+d)} = \frac{(f_1-f_0)^2}{\left(\frac{1}{L_1}+\frac{1}{L_0}\right)^2}\times\frac{T^3}{L_1L_0.Tf^*.T(1-f^*)} \\
&= \frac{(f_1-f_0)^2}{\left(\frac{1}{L_1}+\frac{1}{L_0}\right)^2}\times\frac{T}{L_1L_0f^*(1-f^*)} = \frac{(f_1-f_0)^2}{\left(\frac{1}{L_1}+\frac{1}{L_0}\right)^2}\times\frac{L_0+L_1}{L_1L_0}\times\frac{1}{f^*(1-f^*)} \\
&= \frac{(f_1-f_0)^2}{\left(\frac{1}{L_1}+\frac{1}{L_0}\right)f^*(1-f^*)}
\end{aligned}$$

Ce qui prouve que $\chi^2_{obs} = Z^2_{obs}$.

Retour à l'article

Retour à l'énoncé de cette activité

Une solution de l'activité 2

Il faut consulter la table de Pearson pour $n^{'} = n + 1 = 4$ car le nombre de degrés de liberté est $n = 3$.

n'

χ^2	3.	4.	5.	6.
1	·606,531	·801,253	·909,796	·962,566
2	·367,879	·572,407	·735,759	·849,146
3	·223,130	·391,633	·557,825	·699,994
4	·135,335	·261,470	·406,006	·549,422
5	·082,085	·171,799	·287,298	·415,882
6	·049,787	·111,611	·199,148	·306,220
7	·030,197	·071,888	·135,888	·220,631
8	·018,316	·046,012	·091,578	·156,236
9	·011,109	·029,291	·061,099	·109,064
10	·006,738	·018,567	·040,428	·075,236
15	·000,553	·001,817	·004,701	·010,363
20	·000,045	·000,170	·000,499	·001,250

Cette table donne, pour $\chi^2 = 1; 2; 3; ...; 10; 15; 20; ...; 60; 70$, les valeurs de $\mathbb{P}(Y > \chi^2)$ où Y est une variable aléaloire qui suit une loi du chi-deux à 3 degrés de liberté.

La fonction g définie par $g(y) = \mathbb{P}(Y > y)$ est décroissante. En effet, $g(y) = 1 - F(y)$ où F est la fonction de répartition de la loi du chi-deux à 3 degrés de liberté qui est croissante.

Calcul de $\chi^2_{0,01}(3)$

Pour trouver par exemple le fractile d'ordre $0{,}01$ c'est-à-dire $\chi^2_{0,01}(3)$, il faut chercher y tel que $g(y) = \mathbb{P}(Y > y) = 0{,}01$.

En observant les valeurs indiquées dans la colonne $n^{'} = 4$ de la table de Pearson, on trouve :

$g(10) = 0{,}018567, \quad g(15) = 0{,}001817$. Puisque $0{,}01$ est compris entre $0{,}001817 = g(15)$ et $0{,}018567 = g(10)$, y est compris entre 10 et 15. Il suffit de faire une interpolation linéaire pour obtenir une approximation de y.

Posons $y = 15 - h$. Nous avons $\frac{h}{15-10} = \frac{0{,}01-0{,}001817}{0{,}018567-0{,}001817}$.

Nous en déduisons $h = 2{,}442686$ et $y = 12{,}557313$, valeur différente de celle de la table de Fisher où $\chi^2_{0,01}(3) = 11{,}341$.

Calcul de $\chi^2_{0,05}(3)$

Puisque $0{,}05$ est compris entre $0{,}046012 = g(8)$ et $0{,}071888 = g(7)$, y est compris entre 7 et 8.

Posons $y = 8 - h$. Nous avons $\frac{h}{8-7} = \frac{0{,}05-0{,}046012}{0{,}071888-0{,}046012}$.

Nous en déduisons $h = 0{,}1581911$ et $y = 7{,}8418088$, valeur différente de celle de la table de Fisher où $\chi^2_{0,05}(3) = 7{,}815$.

Calcul de $\chi^2_{0,10}(3)$

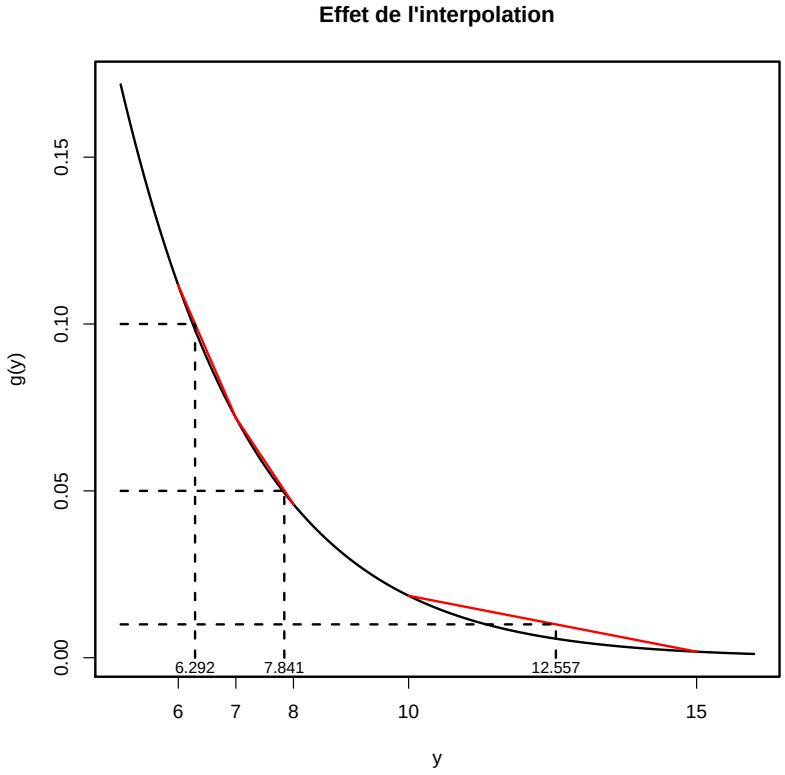
Puisque $0{,}10$ est compris entre $0{,}071888 = g(7)$ et $0{,}111611 = g(6)$, y est compris entre 6 et 7.

Posons $y = 7 - h$. Nous avons $\frac{h}{7-6} = \frac{0{,}10-0{,}071888}{0{,}111611-0{,}071888}$.

Nous en déduisons $h = 0{,}7077$ et $y = 6{,}2922992$, valeur différente de celle de la table de Fisher où $\chi^2_{0,10}(3) = 6{,}251$.

Les différences avec la table de Fisher s'expliquent par la convexité de la fonction g, mais elles s'atténuent quand les valeurs qui "encadrent" y sont plus rapprochées comme on peut le voir sur le graphique de la page suivante. Sur la table de Pearson, les valeurs de g pour $10 < y < 15$ sont absentes. Cette table a dû être complétée par Fischer.

... /...



Retour à l’article

Retour à l’énoncé de cette activité

Une solution de l'activité 3

La statistique du χ^2 a été définie par :

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

Il s'agit de prouver l'égalité :

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

Notons comme précédemment, $L_1 = a+b$, $L_0 = c+d$, $C_1 = a+c$, $C_0 = b+d$ et $T = a+b+c+d$.
Nous avons donc $T = L_0 + L_1 = C_0 + C_1$. Posons enfin $\Delta = ad - bc$.
Les valeurs a', b', c' et d' ont été définies par : $a' = \frac{L_1 C_1}{T}, b' = \frac{L_1 C_0}{T}, c' = \frac{L_0 C_1}{T}$ et $d' = \frac{L_0 C_0}{T}$.

$$a - a' = a - \frac{L_1 C_1}{T} = \frac{aT - L_1 C_1}{T} = \frac{a(L_0+L_1) - L_1(a+c)}{T} = \frac{aL_0 - cL_1}{T} = \frac{ad-bc}{T} = \frac{\Delta}{T}.$$

Nous obtenons de manière analogue :
$b - b' = \frac{bL_0 - dL_1}{T} = -\frac{\Delta}{T}$, $c - c' = \frac{cL_1 - aL_0}{T} = -\frac{\Delta}{T}$ et $d - d' = \frac{dL_1 - bL_0}{T} = \frac{\Delta}{T}$.
Nous pouvons alors calculer χ^2_{obs}.

$$\begin{aligned}
\chi^2_{obs} &= \frac{(a-a')^2}{a'} + \frac{(b-b')^2}{b'} + \frac{(c-c')^2}{c'} + \frac{(d-d')^2}{d'} \\
&= \frac{\Delta^2}{T^2}\left[\frac{1}{a'} + \frac{1}{b'} + \frac{1}{c'} + \frac{1}{d'}\right] \\
&= \frac{\Delta^2}{T^2}\left[\frac{T}{L_1 C_1} + \frac{T}{L_1 C_0} + \frac{T}{L_0 C_1} + \frac{T}{L_0 C_0}\right] \\
&= \frac{\Delta^2}{T}\left[\frac{L_0 C_0 + L_0 C_1 + L_1 C_0 + L_1 C_1}{L_1 L_0 C_1 C_0}\right] \\
&= \frac{\Delta^2}{T}\left[\frac{(L_0+L_1) \times (C_0+C_1)}{L_1 L_0 C_1 C_0}\right] \\
&= \frac{\Delta^2}{T}\left[\frac{T^2}{L_1 L_0 C_1 C_0}\right] \\
&= T \times \frac{\Delta^2}{L_1 L_0 C_1 C_0}
\end{aligned}$$

c'est-à-dire :

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

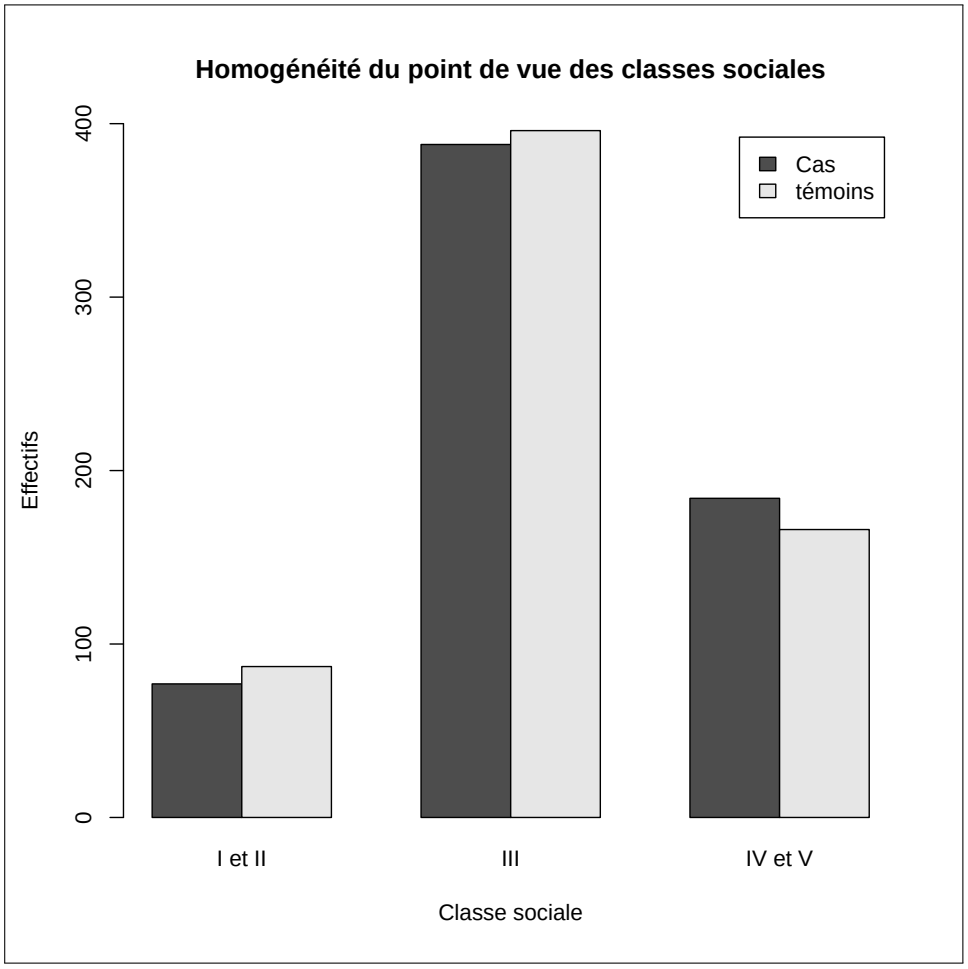
Retour à l'article Retour à l'énoncé de cette activité

Une solution de l'activité 4

Voici le tableau des données :

Classe sociale	Nombre de patients atteints de carcinome pulmonaire	Nombre de patients non atteints	Total
I et II	77	87	164
III	388	396	784
IV et V	184	166	350
Total	649	649	1298

Représentation des données sous forme de diagrammes en bâtons :



Conclusion : nous remarquons ... très peu de différences.

... /...

Calcul de χ^2_{obs} :

Effectifs espérés			
Classe sociale	Cas	Témoins	Total
I et II	82	82	164
III	392	392	784
IV et V	175	175	350
Total	649	649	1298

Calcul de χ^2_{obs}		
Cas	Témoins	Total
0,3049	$0,3049$	0,6098
0,0408	0,0408	0,0816
0,4629	0,4629	0,9257
0,8086	0,8086	1,6171

Le tableau des données comportant trois lignes et deux colonnes, le nombre de degrés de liberté est égal à $(3-1)\times(2-1)=2$. Le résultat obtenu, $1{,}6171$ environ, justifie effectivement la conclusion de Doll et Hill (voir leur commentaire ci-dessous), car il correspond à une valeur de la P-valeur égale à $0{,}4455$, qui *ne permet pas de rejeter* l'hypothèse H_0 d'*homogénéité* des classes sociales.

> The difference in social class distribution is small and is no more than might easily be due to chance ($\chi^2=1.61$; $n=2$; $0.30<P<0.50$).

Retour à l'article

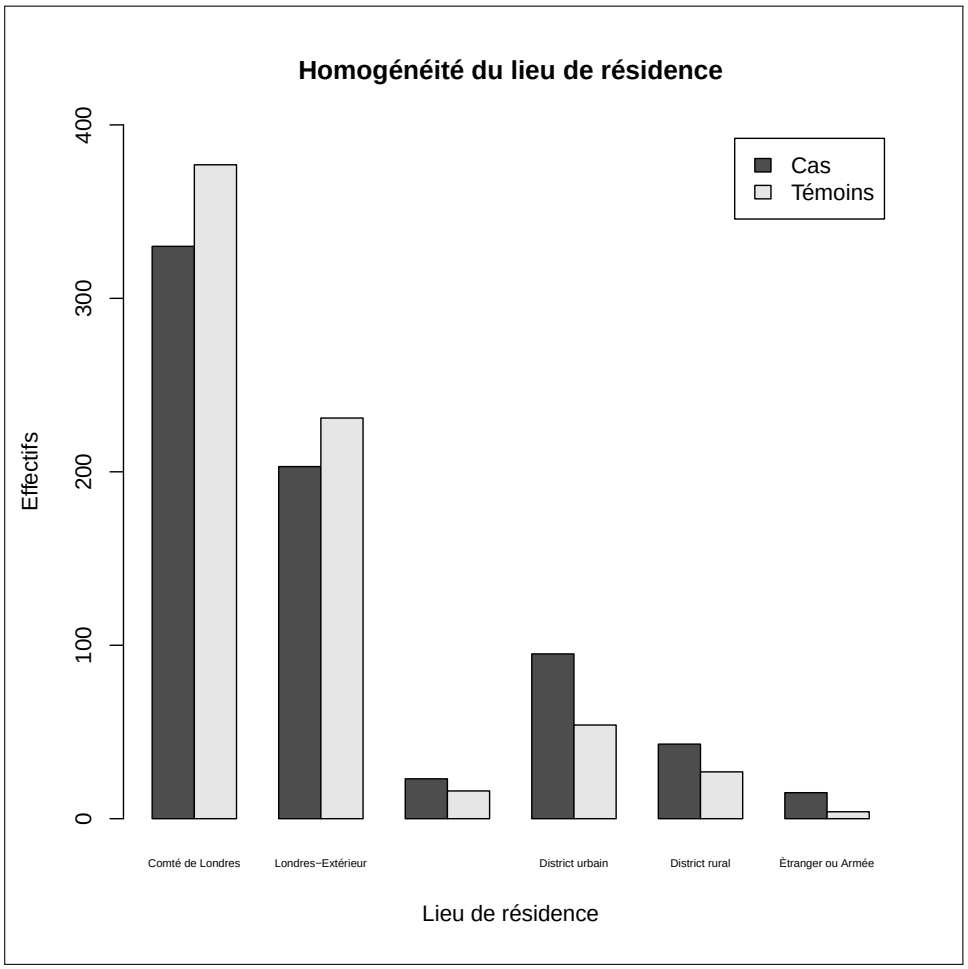
Retour à l'énoncé de cette activité

Une solution de l'activité 5

Voici le tableau des données :

Lieu de résidence	Nombre de patients atteints de carcinome pulmonaire	Nombre de patients non atteints	Total
Comté de Londres	330	377	707
Londres-Extérieur	203	231	434
Autre arrondissement du comté	23	16	39
District urbain	95	54	149
District rural	43	27	70
À l'étranger ou à l'armée	15	4	19
Total	709	709	1418

Représentation des données sous forme de diagrammes en bâtons :



Conclusion :

D'après ce diagramme, les cas et les témoins ne sont pas répartis d'une même manière en ce qui concerne leur lieu de résidence.

... /...

Calcul de χ^2_{obs} :

Effectifs espérés			
Lieu de résidence	Cas	Témoins	Total
Comté de Londres	353,5	353,5	707
Londres-Extérieur	217	217	434
Autre arrondissement du comté	19,5	19,5	39
District urbain	74,5	74,5	149
District rural	35	35	70
À l'étranger ou à l'armée	9,5	9,5	19
Total	709	709	1418

Calcul de χ^2_{obs}		
Cas	Témoins	Total
1,5622	1,5622	3,1245
0,9032	0,9032	1,8065
0,6282	0,6282	1,2564
5,6409	5,6409	11,2819
1,8286	1,8286	3,6571
3,1842	3,1842	6,3684
13,7474	13,7474	27,4948

Le tableau des données comportant six lignes et deux colonnes, le nombre de degrés de liberté est égal à $(6-1)\times(2-1)=5$. Le résultat obtenu, $27{,}4948$ (environ), justifie effectivement la conclusion de Doll et Hill car il correspond à une valeur de la P-valeur égale à $4{,}6.10^{-5}$, qui *permet de rejeter* l'hypothèse H_0 d'*homogénéité* du lieu de résidence.

N. B. Nous pouvons remarquer que la valeur du χ^2_{obs} que nous avons calculée diffère de la valeur $31{,}49$ figurant dans les commentaires du tableau II de l'article de Doll et Hill (voir ci-dessous) mais cette différence (que nous n'expliquons pas) n'en modifie heureusement pas la conclusion.

The difference in place of residence is, however, large ($\chi^2 = 31.49$; n = 5 ; P<0.001), and Table II shows that a higher proportion of the lung patients were resident outside London at the time of their admission to hospital.

Retour à l'article

Retour à l'énoncé de cette activité

Une solution de l'activité 6

Du tableau IV, nous pouvons extraire le tableau 2×2 suivant, concernant uniquement les **hommes** :

Groupe de maladies	Non fumeurs	Fumeurs	Total
Patients atteints du cancer du poumon	2	647	649
Patients témoins n'ayant pas le cancer du poumon	27	622	649
Total	29	1269	1298

Tableau des effectifs observés

En présence d'un tableau 2×2, le nombre de degrés de liberté est égal à 1 et pour calculer le χ^2_{obs}, nous pouvons utiliser la formule :

$$\chi^2_{obs} = \frac{(2\times622-27\times647)^2\times1298}{649\times649\times29\times1269} \simeq 22{,}044.$$

D'après la table de Fisher ($n = 1$), la p-value P vérifie : $P < 0{,}01$.

En utilisant un tableur, on peut obtenir une valeur approchée de P :

$$P = \text{LOI.KHIDEUX}(22{,}044\ ;\ 1) \simeq 0{,}0000027.$$

On peut lire dans la colonne de droite du tableau IV que la p-value a été obtenue en utilisant le test exact de Fisher et qu'elle vaut $P = 0{,}00000064$. Comme indiqué dans l'article, l'utilisation du test du chi-deux nécessite que les effectifs **espérés** soit supérieurs à 5 et, dans notre tableau, il y a un effectif **observé** égal à 2. C'est peut-être la raison de l'utilisation, dans ce cas, du test exact de Fisher, ce qui n'est pas nécessaire, comme on peut le voir en calculant le tableau des effectifs observés.

Groupe de maladies	Non fumeurs	Fumeurs	Total
Patients atteints du cancer du poumon	14,5	634,5	649
Patients témoins n'ayant pas le cancer du poumon	14,5	634,5	649
Total	29	1269	1298

Tableau des effectifs espérés

Les questions posées à la suite de ce calcul sont des questions pièges !

Dans cette étude, ce qui est fixé est le fait d'être fumeur ou de ne pas l'être et ce qui est aléatoire, c'est d'être atteint ou non du cancer du poumon. On ne peut donc pas répondre à la question 1. Pour pouvoir le faire, il aurait fallu que les données soient issues d'une étude de cohorte et non d'une étude cas-témoins.

En revanche, on peut répondre aux questions 2. et 3. :

D'après la valeur du χ^2_{obs} calculée, on peut affirmer, avec le risque $\alpha = 5\%$ de se tromper, que, chez les **hommes**, la probabilité d'être fumeur quand on a le cancer du poumon, est différente(*) de celle d'être fumeur quand on ne l'a pas et qu'il existe un lien entre la consommation de tabac et le cancer du poumon.

(*) L'observation des proportions observées amène à conclure que, chez les **hommes**, la probabilité d'être fumeur quand on a le cancer du poumon, est supérieure à celle d'être fumeur quand on ne l'a pas.

Retour à l'article Retour à l'énoncé de cette activité

Une solution de l'activité 7

Du tableau IV, nous pouvons extraire le tableau 2×2 suivant, concernant uniquement les **femmes** :

Groupe de maladies	Non fumeurs	Fumeurs	Total
Patientes atteints du cancer du poumon	19	41	60
Patientes témoins n'ayant pas le cancer du poumon	32	28	60
Total	51	69	120

Tableau des effectifs observés

En présence d'un tableau 2×2, le nombre de degrés de liberté est égal à 1 et pour calculer le χ^2_{obs}, nous pouvons utiliser la formule :

$$\chi^2_{obs} = \frac{(19\times28-41\times32)^2\times120}{60\times60\times51\times69} \simeq 5{,}763.$$

Nous obtenons exactement la même valeur du χ^2_{obs} que celle du tableau IV de Doll et Hill.
D'après la table de Fisher ($n = 1$), la p-valeur P vérifie : $0{,}01 < P < 0{,}02$.

En utilisant un tableur, on peut obtenir une valeur approchée de P :

$$P = \text{LOI.KHIDEUX}(5{,}763\ ;\ 1) \simeq 0{,}016.$$

Les questions posées à la suite de ce calcul sont des questions pièges !

Dans cette étude, ce qui est fixé est le fait d'être fumeur ou de ne pas l'être et ce qui est aléatoire, c'est d'être atteint ou non du cancer du poumon. On ne peut donc pas répondre à la question 1. Pour pouvoir le faire, il aurait fallu que les données soient issues d'une étude de cohorte et non d'une étude cas-témoins.

En revanche, on peut répondre aux questions 2. et 3. :

D'après la valeur du χ^2_{obs} calculée, on peut affirmer, avec le risque $\alpha = 5\%$ de se tromper, que, chez les **femmes**, la probabilité d'être fumeur quand on a le cancer du poumon, est différente(*) de celle d'être fumeur quand on ne l'a pas et qu'il existe un lien entre la consommation de tabac et le cancer du poumon.

(*) L'observation des proportions observées amène à conclure que, chez les **femmes**, la probabilité d'être fumeur quand on a le cancer du poumon, est supérieure à celle d'être fumeur quand on ne l'a pas.

Retour à l'article

Retour à l'énoncé de cette activité

Une solution de l'activité 8

1. Vérifier les valeurs des chi-deux et les encadrements obtenus pour P dans le tableau V.

1.1 Test du chi-deux pour la quantité de tabac consommée par les hommes.

Dans le cas d'un tableau de deux lignes et cinq colonnes, le nombre de degrés de liberté est égal à $(2-1)\times(5-1)=4$ et le calcul du chi-deux par la formule de Pearson s'obtient de la manière suivante :

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

où $a' = L_1 \times \frac{a+f}{T}$, $f' = L_0 \times \frac{a+f}{T}$, etc.

Effectifs observés :

Quantité de tabac	1 Cig.-	5 Cig.-	15 Cig.-	25 Cig.-	50 Cig.+	Total
Patients atteints du cancer du poumon	$a = 33$	$b = 250$	$c = 196$	$d = 136$	$e = 32$	$L_1 = 647$
Patients indemnes de cancer du poumon	$f = 55$	$g = 293$	$h = 190$	$i = 71$	$j = 13$	$L_0 = 622$
Total	$a+f = 88$	$b+g = 543$	$c+h = 386$	$d+i = 207$	$e+j = 45$	1269

Effectifs espérés :

Quantité de tabac	1 Cig.-	5 Cig.-	15 Cig.-	25 Cig.-	50 Cig.+	Total
Patients atteints du cancer du poumon	$a' = 44{,}87$	$b' = 276{,}85$	$c' = 196{,}80$	$d' = 105{,}54$	$e' = 22{,}94$	$L_1 = 647$
Patients indemnes de cancer du poumon	$f' = 43{,}13$	$g' = 266{,}15$	$h' = 189{,}20$	$i' = 101{,}46$	$j' = 22{,}06$	$L_0 = 622$
Total	$a'+f' = 88$	$b'+g' = 543$	$c'+h' = 386$	$d'+i' = 207$	$e'+j' = 45$	1269

Pour calculer χ^2_{obs} selon la formule indiquée plus haut, il est commode de grouper les calculs des différentes fractions dans un tableau à 4 lignes et 7 colonnes ; cela facilite aussi leur sommation.

Quantité de tabac	1 Cig.-	5 Cig.-	15 Cig.-	25 Cig.-	50 Cig.+	Total
Patients atteints du cancer du poumon	$\frac{(a-a')^2}{a'} \simeq 3{,}14$	$\frac{(b-b')^2}{b'} \simeq 2{,}60$	$\frac{(c-c')^2}{c'} \simeq 0{,}003$	$\frac{(d-d')^2}{d'} \simeq 8{,}79$	$\frac{(e-e')^2}{e'} \simeq 3{,}58$	18,11
Patients indemnes de cancer du poumon	$\frac{(f-f')^2}{f'} \simeq 3{,}26$	$\frac{(g-g')^2}{g'} \simeq 2{,}71$	$\frac{(h-h')^2}{h'} \simeq 0{,}003$	$\frac{(i-i')^2}{i'} \simeq 9{,}15$	$\frac{(j-j')^2}{j'} \simeq 3{,}72$	18,84
Total	6,40	5,31	0,007	17,94	7,29	36,95

Nous obtenons exactement la valeur figurant dans l'article de Doll et Hill : $\chi^2_{obs} = 36{,}95$.
D'après la table de Fisher ($n = 4$), la valeur de P vérifie : $P < 0{,}001$.
De plus, il est possible d'obtenir une valeur approchée de P en utilisant, par exemple, un tableur :

$$P = \text{LOI.KHIDEUX}(36{,}95\ ;\ 4) \simeq 0{,}000000184.$$

... /...

1.2 Test du chi-deux pour la quantité de tabac consommée par les femmes.

Nous pouvons remarquer ici que les fortes consommations sont rares. Or le test du chi-deux est asymptotique donc approximatif et l'approximation est considérée comme bonne lorsque toutes les valeurs espérées a', b', etc., sont supérieures ou égales à 5, ce qui n'est pas le cas ici. La pratique recommandée dans ce cas consiste à regrouper judicieusement des colonnes.

Nous regroupons donc les colonnes 15 Cig.-, 25 Cig.- et 50 Cig.+, ce qui donne un nombre de degrés de liberté égal à $(2-1)\times(3-1)=2$ et un calcul du χ^2_{obs} suivant :

Effectifs observés :

Quantité de tabac	1 Cig.-	5 Cig.-	15 Cig.+	Total
Patientes atteints du cancer du poumon	$a=7$	$b=19$	$c=15$	$L_1=41$
Patientes indemnes de cancer du poumon	$d=12$	$e=10$	$f=6$	$L_0=28$
Total	$a+d=19$	$b+e=29$	$c+f=21$	69

Effectifs espérés :

Quantité de tabac	1 Cig.-	5 Cig.-	15 Cig.+	Total
Patientes atteints du cancer du poumon	$a'=11{,}29$	$b'=17{,}23$	$c'=12{,}48$	$L_1=41$
Patientes indemnes de cancer du poumon	$d'=7{,}71$	$e'=11{,}77$	$f'=8{,}52$	$L_0=28$
Total	$a'+d'=19$	$b'+e'=29$	$c'+f'=21$	69

Calcul de χ^2_{obs} présenté comme précédemment :

Quantité de tabac	1 Cig.-	5 Cig.-	15 Cig.+	Total
Patientes atteints du cancer du poumon	$\frac{(a-a')^2}{a'}\simeq 1{,}63$	$\frac{(b-b')^2}{b'}\simeq 0{,}18$	$\frac{(c-c')^2}{c'}\simeq 0{,}51$	$2{,}32$
Patientes indemnes de cancer du poumon	$\frac{(d-d')^2}{d'}\simeq 2{,}39$	$\frac{(e-e')^2}{e'}\simeq 0{,}27$	$\frac{(f-f')^2}{f'}\simeq 0{,}75$	$3{,}40$
Total	$4{,}02$	$0{,}45$	$1{,}26$	$5{,}72$

Nous obtenons exactement la valeur figurant dans l'article : χ^2_{obs} = 5,72.
D'après la table de Fisher ($n=2$), la valeur de P vérifie : 0,05 < P < 0,10.
Valeur approchée de P obtenue en utilisant un tableur :

$$P = \text{LOI.KHIDEUX}(5{,}72\ ;\ 2) \simeq 0{,}057275.$$

TABLE V.—*Most Recent Amount of Tobacco* Consumed Regularly by Smokers Before the Onset of Present Illness; Lung-carcinoma Patients and Control Patients with Diseases Other Than Cancer*

Disease Group	No. Smoking Daily					Probability Test
	1 Cig.–*	5 Cigs.–	15 Cigs.–	25 Cigs.–	50 Cigs. +	
Males: Lung-carcinoma patients (647)	33 (5·1%)	250 (38·6%)	196 (30·3%)	136 (21·0%)	32 (5·0%)	$\chi^2=36{\cdot}95$; n=4; P<0·001
Control patients with diseases other than cancer (622)	55 (8·8%)	293 (47·1%)	190 (30·5%)	71 (11·4%)	13 (2·1%)	
Females: Lung-carcinoma patients (41)	7 (17·1%)	19 (46·3%)	9 (22·0%)	6 (14·6%)	0 (0·0%)	$\chi^2=5{\cdot}72$; n=2; 0·05<P<0·10 (Women smoking 15 or more cigarettes a day grouped together)
Control patients with diseases other than cancer (28)	12 (42·9%)	10 (35·7%)	6 (21·4%)	0 (0·0%)	0 (0·0%)	

* Ounces of tobacco have been expressed as being equivalent to so many cigarettes. There is 1 oz. of tobacco in 26·5 normal-size cigarettes, so that the conversion factor has been taken as: 1 oz. of tobacco a week = 4 cigarettes a day.

$P \simeq 1{,}8.10^{-7}$

$P \simeq 0{,}057$

... /...

2. Quelles conclusions peut-on en tirer ?

Pour ce qui concerne les hommes :
au risque de première espèce de 5%, on peut affirmer que la quantité de tabac consommée par un individu est associée à l'apparition d'un carcinome pulmonaire.

Pour ce qui concerne les femmes :
au risque de première espèce de 5%, on ne peut pas affirmer la même association. Pour le faire, il aurait fallu accepter, au départ de l'étude, un risque de première espèce de 10%, par exemple.

On peut cependant penser que c'est le manque de puissance du test, dû au faible effectif féminin, qui a produit ce résultat décevant pour les auteurs.

Retour à l'article Retour à l'énoncé de cette activité

Une solution de l'activité 9

À partir des valeurs figurant dans l'énoncé, nous pouvons obtenir le tableau 2×2 suivant :

Quantité de tabac	Inhalant la fumée	N'inhalant pas la fumée	Total
Patients atteints du cancer du poumon	424	264	688
Patients indemnes de cancer du poumon	437	213	650
Total	861	477	1338

En présence d'un tableau 2×2, le nombre de degré de liberté est égal à 1 et pour calculer le χ^2_{obs}, nous pouvons utiliser la formule :

$$\chi^2_{obs} = \frac{(424\times213-264\times437)^2\times1338}{861\times477\times688\times650} \simeq 4{,}57.$$

D'après la table de Fisher ($n = 1$), la valeur de P vérifie : $0{,}02 < P < 0{,}05$.

En utilisant un tableur, nous pouvons obtenir une valeur approchée de P :

$$P = \text{LOI.KHIDEUX}(4{,}57\ ;\ 1) \simeq 0{,}032537.$$

Nous obtenons le même résultat que Doll et Hill et comme ils l'écrivent dans leur article (voir ci-dessous), « il semble que les patients atteints du cancer du poumon inhalent la fumée légèrement moins souvent que les autres que les autres patients ($\chi^2 = 4.58\ ;\ n = 1\ ;\ 0.02 < P < 0.05$). Toutefois, la différence n'est pas grande et si l'on compare les patients atteints du cancer du poumon avec tous les autres patients qui ont été interrogés et qu'on prend en compte le sexe et l'âge, cette différence devient insignifiante ($\chi^2 = 0.19\ ;\ n = 1\ ;\ 0.50 < P < 0.70$) ».

Inhaling

Another difference between smokers is that some inhale and others do not. All patients who smoked were asked whether or not they inhaled, and the answers given by the lung-carcinoma and non-cancer control patients were as follows : of the 688 lung-carcinoma patients who smoked (men and women) 61.6% said they inhaled and 38.4% said they did not ; the corresponding figures for the 650 patients with other diseases were 67.2% inhalers and 32.8% non-inhalers. It would appear that lung-carcinoma patients inhale slightly less often than other patients (χ^2=4.58 ; n=1 ; 0.02<P<0.05). However, the difference is not large, and if the lung-carcinoma patients are compared with all the other patients interviewed, and the necessary allowance is made for sex and age, the difference becomes insignificant (χ^2=0.19 ; n=1 ; 0.50<P<0.70).

Retour à l'article

Retour à l'énoncé de cette activité

Une solution de l'activité 10

1. Calcul de χ^2_{obs} :

Effectifs observés			
Traitement	Morts (x_i)	Vivants ($n - x_i$)	Total
0% (contrôle)	3	7	**10**
11% OGM	5	5	**10**
22% OGM	1	9	**10**
33% OGM	1	9	**10**
Total	10	30	40

Effectifs espérés			
Traitement	Morts	Vivants	Total
0% (contrôle)	2,5	7,5	10
11% OGM	2,5	7,5	10
22% OGM	2,5	7,5	10
33% OGM	2,5	7,5	10
Total	10	30	40

Calcul de χ^2_{obs}		
Morts	Vivants	Total
0,1000	0,0333	0,1333
2,5000	0,8333	3,3333
0,9000	0,3000	1,2000
0,9000	0,3000	1,2000
4,4000	1,4667	5,8667

Le résultat obtenu, $5{,}87$ environ, est bien égal à celui que donne Louis Ollivier.

2. Application de la formule de Brandt-Snedecor :

$$\chi^2_{obs} = \frac{k.\sum x_i^2 - (\sum x_i)^2}{\sum x_i - \frac{(\sum x_i)^2}{n.k}} = \frac{4{,}36 - (10)^2}{10 - \frac{(10)^2}{10{,}4}} = \frac{44}{7.5} \simeq 5{,}866667$$

3. Démonstration de la formule de Brandt-Snedecor :

Nous considérons ici un test du chi-deux sur k échantillons **de même effectif** n et correspondant à un tableau $(e_{i,j})$ ayant **deux** colonnes et k lignes.

Il s'agit de démontrer que la valeur de χ^2_{obs} peut être obtenue par la formule :

$$\chi^2_{obs} = \frac{k.\sum x_i^2 - (\sum x_i)^2}{\sum x_i - \frac{(\sum x_i)^2}{n.k}}.$$

Les éléments du tableau sont notés $e_{i,j}$, i variant de 1 à k, j variant de 1 à 2 ; selon les notations de Louis Ollivier, on a : $e_{i,1} = x_i$ pour tout i compris entre 1 et k.

Or $\chi^2_{obs} = \sum_{j=1}^{2} \sum_{i=1}^{k} \frac{\left(e_{i,j} - e'_{i,j}\right)^2}{e'_{i,j}}$ où $e'_{i,j}$ sont les effectifs espérés.

Notons $\sum_{i=1}^{k} e_{i,1} = s$. Alors $\sum_{i=1}^{k} e_{i,2} = k.n - s$ et pour tout i, $e'_{i,1} = \frac{n.s}{k.n} = \frac{s}{k}$ et $e'_{i,2} = \frac{n.(k.n-s)}{k.n} = n - \frac{s}{k}$ avec, bien sûr, $e_{i,2} = n - e_{i,1}$.

... /...

Ainsi :

$\chi^2_{obs} = \sum_{j=1}^{2}\sum_{i=1}^{k}\frac{\left(e_{i,j}-e'_{i,j}\right)^2}{e'_{i,j}} = \sum_{i=1}^{k}\frac{\left(e_{i,1}-\frac{s}{k}\right)^2}{\frac{s}{k}} + \sum_{i=1}^{k}\frac{\left(n-e_{i,1}-n+\frac{s}{k}\right)^2}{n-\frac{s}{k}} = \sum_{i=1}^{k}\left(e_{i,1}-\frac{s}{k}\right)^2.\left(\frac{1}{\frac{s}{k}}+\frac{1}{n-\frac{s}{k}}\right).$

Or $\sum_{i=1}^{k}\left(e_{i,1}-\frac{s}{k}\right)^2 = \sum_{i=1}^{k}e_{i,1}^2 - 2.\frac{s^2}{k} + k.\frac{s^2}{k^2} = \sum_{i=1}^{k}e_{i,1}^2 - \frac{s^2}{k}$ et $\frac{1}{\frac{s}{k}}+\frac{1}{n-\frac{s}{k}} = \frac{n}{\frac{n.s}{k}-\frac{s^2}{k^2}} = \frac{k}{s-\frac{s^2}{n.k}}$.

Donc $\chi^2_{obs} = \left(\sum_{i=1}^{k}e_{i,1}^2 - \frac{s^2}{k}\right).\frac{k}{s-\frac{s^2}{n.k}} = \frac{k.\sum x_i^2-\left(\sum x_i\right)^2}{\sum x_i-\frac{\left(\sum x_i\right)^2}{n.k}}$.

Retour à l'article Retour à l'énoncé de cette activité