De l'épidémiologie à la sociologie de l'éducation : comparaisons de proportions et *odds ratios*

Jacques Faisant, Denis Lanier, Jean Lejeune, Rémy Morello, Didier Trotoux
IREM de Caen Normandie
septembre 2019

En épidémiologie, lorsqu'une variation du pourcentage d'apparition d'une pathologie, d'une maladie ou d'un décès, semble associée à l'exposition ou non de l'individu à un facteur de risque, l'apport de l'outil statistique a été de pouvoir décider si cette variation était « significative » (dans un sens à préciser) pour inférer l'influence ou non de ce facteur. Le premier scientifique à avoir tenté dépasser la simple constatation d'une différence entre les pourcentages a été Jules Gavarret, dans son ouvrage « *Principes généraux de la Statistique médicale* » [Gavarret, 1840] et Richard Doll et Bradford Hill ont été les premiers à avoir utilisé à une grande échelle un outil statistique (la statistique du chi-deux) pour l'étude d'une association entre la pratique tabagique et l'apparition du cancer du poumon [Doll et Hill, 1950]. L'objet de cet article est de continuer l'exploration des outils utilisés pour la comparaison de pourcentages et pour inférer l'existence ou non d'une cause pouvant expliquer les variations de ces pourcentages. Aujourd'hui, un outil utilisé en épidémiologie est appelé *odds ratio*. Il apparait également souvent en sociologie et en sociologie de l'éducation. En particulier, cette dernière discipline peut s'intéresser à l'association entre l'événement « obtenir le baccalauréat » et l'origine sociale des parents. Cet outil est un indicateur de l'association entre un facteur (en épidémiologie : fumeur/non fumeur, en sociologie de l'éducation : enfant de cadres/enfant d'ouvriers) et un événement (en épidémiologie : contracter/ne pas contracter un cancer du poumon, en sociologie de l'éducation : obtenir le baccalauréat/ne pas obtenir le baccalauréat) qui reste cependant aléatoire.

1 L'exemple des inégalités scolaires d'après P. Mercklé.

Cet exemple est tiré d'un billet de Pierre Mercklé paru dans le supplément Science & Techno du journal *Le Monde* du 7/06/2012, intitulé « Les inégalités scolaires diminuent-elles ? » [Mercklé, 2012]. Il présente la comparaison des pourcentages d'obtention du baccalauréat pour les enfants de cadres et d'ouvriers en 1960 et en 2010,

	1960	2010
Enfants de cadres	45%	90%
Enfants d'ouvriers	5%	45%

et pose la question : les inégalités ont-elles augmenté ou diminué ?

Plus précisément, la classe sociale d'origine et l'obtention du bac a-t-elle ou non augmenté ?

Par obtention du baccalauréat, il faut entendre accès à l'enseignement supérieur.

Les données du tableau précédent proviennent des enquêtes *Formation et Qualification Professionnelle* et des enquêtes *Emploi* de l'INSEE [Thélot et Vallet, 2000].

1.1 Mesures de l'inégalité : des résultats contradictoires.

Si on considère que les pourcentages indiqués sont issus, non pas d'un échantillon, mais de la population toute entière, on peut les confondre avec la probabilité d'obtention du baccalauréat et on pose, pour les données de chaque année :

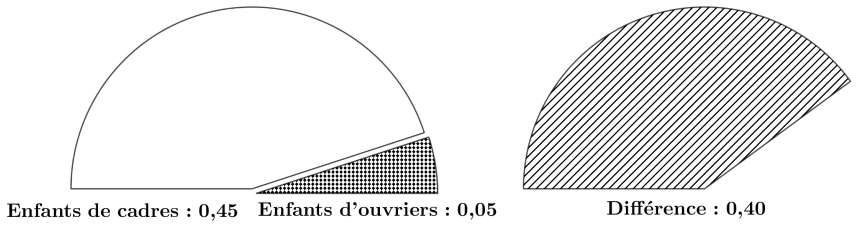
- $p_1 =$ probabilité d'obtention du baccalauréat pour les enfants de cadres ;
- $p_2 =$ probabilité d'obtention du baccalauréat pour les enfants d'ouvriers.

L'inégalité de probabilité **d'obtention du baccalauréat** entre les enfants de cadres et les enfants d'ouvriers peut, dans un premier temps, se mesurer par deux indices :

la différence de probabilité $dp = p_1 - p_2$ et **le rapport de probabilité** $rp = \dfrac{p_1}{p_2}$.

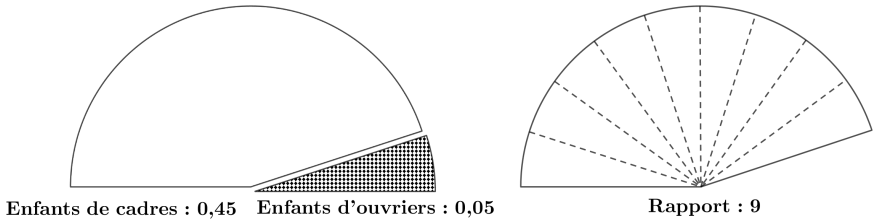
En **1960**, suivant l'indice utilisé, l'inégalité **d'obtention du baccalauréat** peut être mesurée par :

- **La différence de probabilité** : $dp_{1960} = 0{,}45 - 0{,}05 = 0{,}40$.



Les enfants de cadres ont une probabilité d'avoir obtenu le baccalauréat supérieure de $0{,}40$ à celle des enfants d'ouvriers.

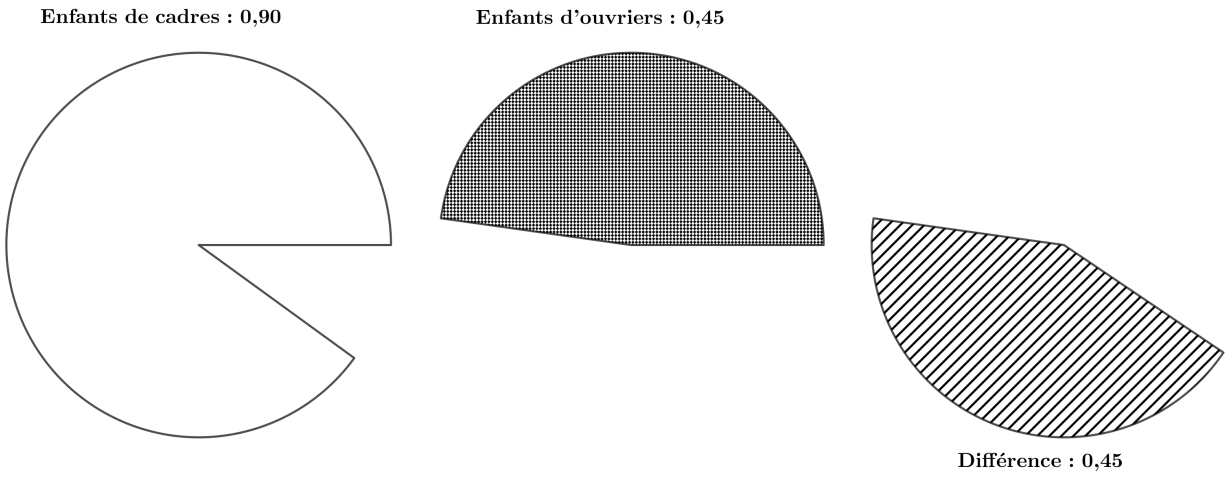
- **Le rapport de probabilité** : $rp_{1960} = \dfrac{0{,}45}{0{,}05} = 9$.



Les enfants de cadres ont 9 fois plus de chances d'avoir obtenu le baccalauréat que les enfants d'ouvriers.

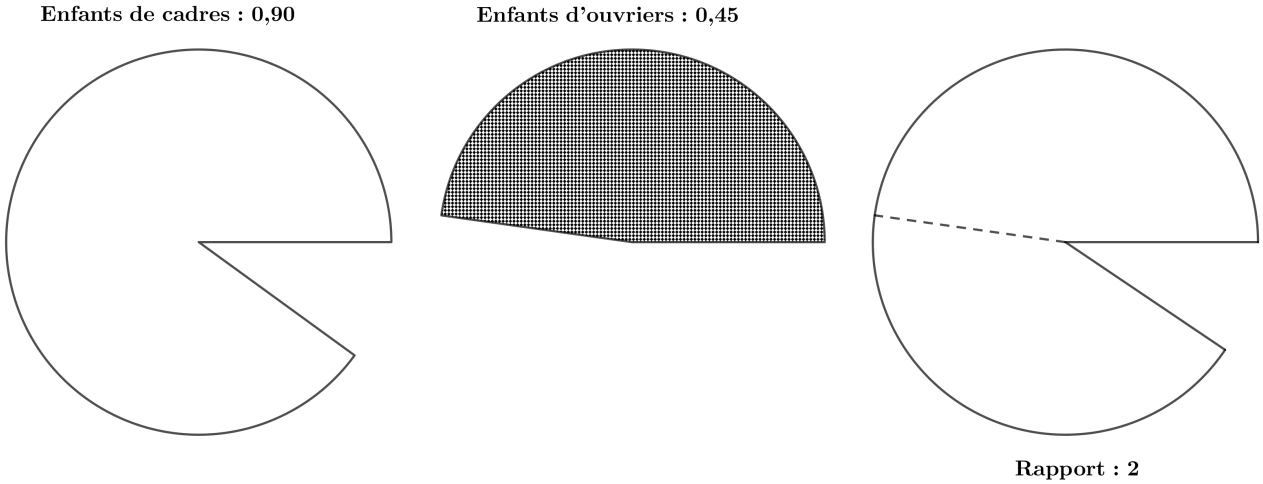
En **2010**, suivant l'indice utilisé, l'inégalité **d'obtention du baccalauréat** peut être mesurée par :

- **La différence de probabilité** : $dp_{2010} = 0{,}90 - 0{,}45 = 0{,}45$.



Les enfants de cadres ont une probabilité d'avoir obtenu le baccalauréat supérieure de $0{,}45$ à celle des enfants d'ouvriers.

- **Le rapport de probabilité** : $rp_{2010} = \dfrac{0{,}90}{0{,}45} = 2.$



Les enfants de cadres ont 2 fois plus de chances d'avoir obtenu le baccalauréat que les enfants d'ouvriers.

Conclusions obtenues quant à l'**inégalité d'obtention** du baccalauréat :

1. Si l'indice utilisé est la **différence de probabilité**, puisque $0{,}40 = dp_{1960} < dp_{2010} = 0{,}45$, on observe une **augmentation** de l'inégalité.
2. Si l'indice utilisé est le **rapport de probabilité**, puisque $9 = rp_{1960} > rp_{2010} = 2$, on observe une **diminution** de l'inégalité.

On peut aussi adopter un autre point de vue, à savoir celui de la **non-obtention du baccalauréat**. Les données sont alors les suivantes :

	1960	2010
Enfants de cadres	55%	10%
Enfants d'ouvriers	95%	55%

et poser :

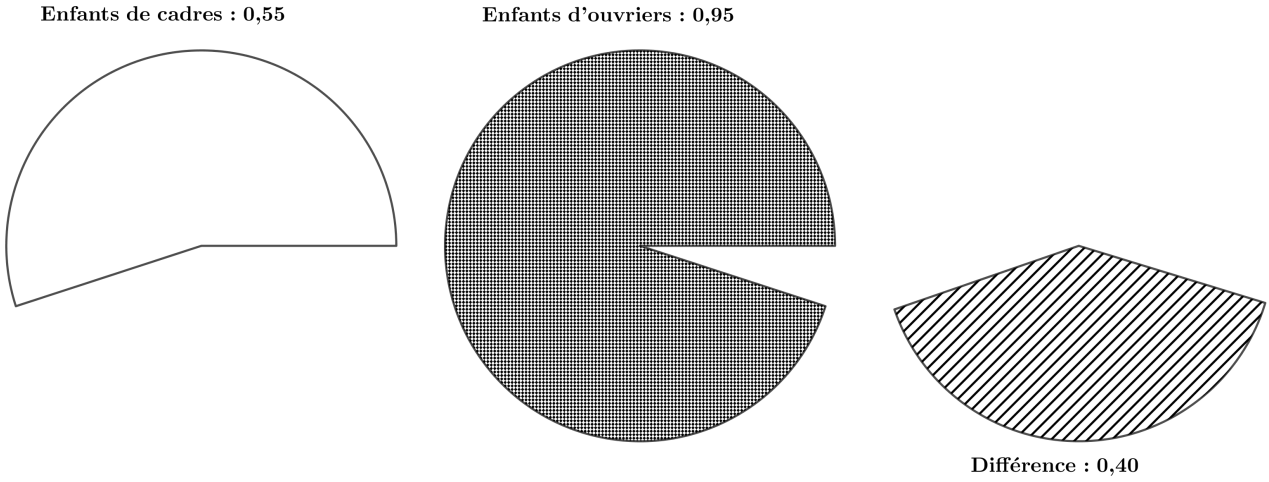
- $q_1 = 1 - p_1 =$ probabilité de non-obtention du baccalauréat pour les enfants de cadres ;
- $q_2 = 1 - p_2 =$ probabilité de non-obtention du baccalauréat pour les enfants d'ouvriers.

L'inégalité de probabilité **de non-obtention du baccalauréat** entre les enfants de cadres et les enfants d'ouvriers peut, comme précédemment, se mesurer par les deux indices :

la différence de probabilité $dq = q_2 - q_1$ et **le rapport de probabilité** $rq = \dfrac{q_2}{q_1}$.

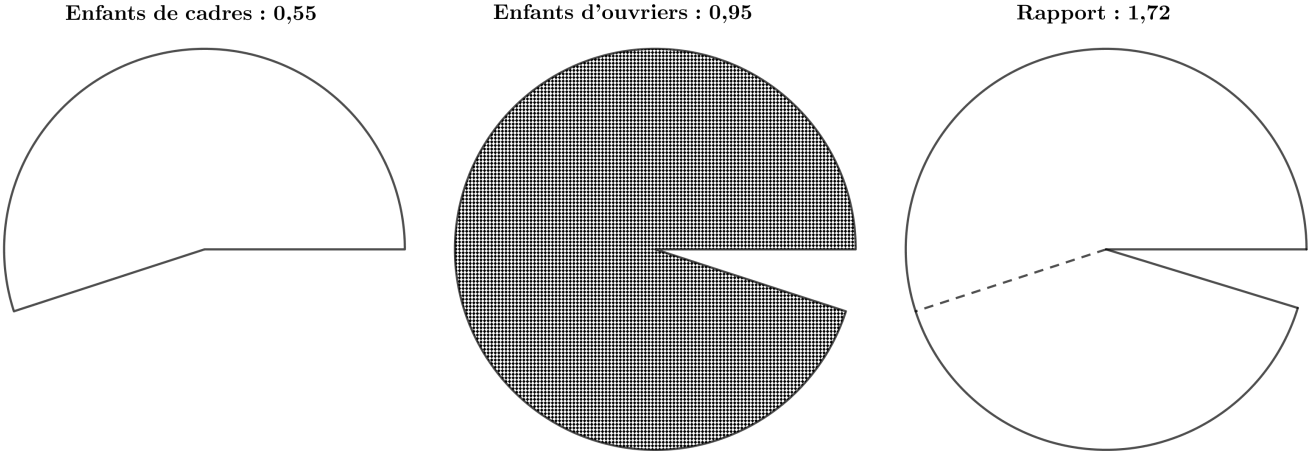
En **1960**, suivant l'indice utilisé, l'inégalité de **non-obtention du baccalauréat** peut être mesurée par :

- **La différence de probabilité** : $dq_{1960} = 0{,}95 - 0{,}55 = 0{,}40$.



Les enfants d'ouvriers ont une probabilité de ne pas avoir obtenu le baccalauréat supérieure de $0{,}40$ à celle des enfants de cadres.

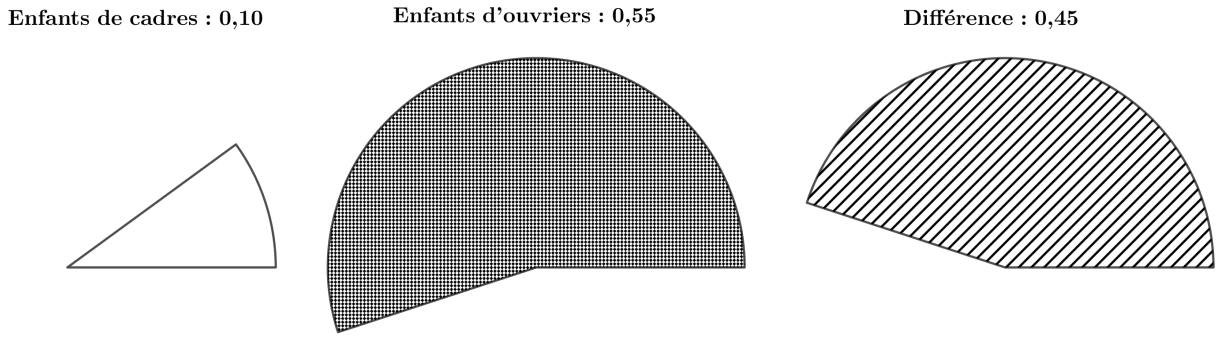
- **Le rapport de probabilité** : $rq_{1960} = \dfrac{0{,}95}{0{,}55} = 1{,}72$.



Les enfants d'ouvriers ont $1{,}72$ fois plus de « chances » de ne pas avoir obtenu le baccalauréat que les enfants de cadres.

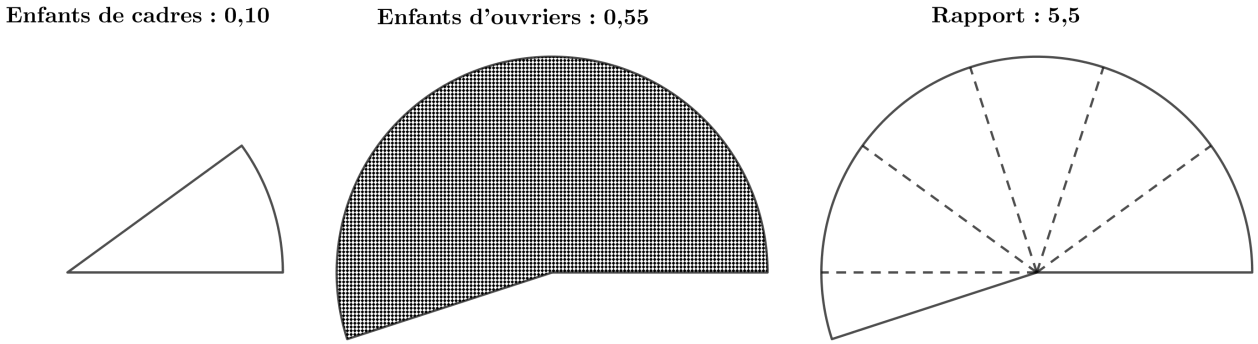
En **2010**, suivant l'indice utilisé, l'inégalité **de non-obtention du baccalauréat** peut être mesurée par :

- **La différence de probabilité** : $dq_{2010} = 0{,}55 - 0{,}10 = 0{,}45$.



Les enfants d'ouvriers ont une probabilité de ne pas avoir obtenu le baccalauréat supérieure de $0{,}45$ à celle des enfants de cadres .

- **Le rapport de probabilité** : $rq_{2010} = \dfrac{0{,}55}{0{,}10} = 5{,}5$.



Les enfants d'ouvriers ont $5{,}5$ fois plus de « chances » de ne pas avoir obtenu le baccalauréat que les enfants de cadres.

Conclusions obtenues quant à l'**inégalité de non-obtention** du baccalauréat :

1. Si l'indice utilisé est la **différence de probabilité**, puisque $0{,}40 = dq_{1960} < dq_{2010} = 0{,}45$, on observe une **augmentation** de l'inégalité.
2. Si l'indice utilisé est le **rapport de probabilité**, puisque $1{,}72 = rq_{1960} < rq_{2010} = 5{,}5$, on observe une **augmentation** de l'inégalité.

Les résultats sont résumés ci-dessous suivant le point de vue adopté et l'indice utilisé :

	Obtention	Non-obtention
Différence	Augmentation	Augmentation
Rapport	Diminution	Augmentation

Sur cet exemple, apparaissent deux contradictions :

1. si on se place du point de vue de l'obtention du baccalauréat, l'indice basé sur la différence conclut à une augmentation de l'inégalité alors que celui basé sur le rapport conclut à une diminution de l'inégalité.

2. l'indice basé sur le rapport amène à conclure à une diminution de l'inégalité si on se place du point de vue de l'obtention et à une augmentation de l'inégalité si on se place du point de vue de la non-obtention.

L'indice basé sur la différence et celui sur le rapport ont en commun d'associer la classe sociale de l'enfant à l'obtention ou non du baccalauréat. Le problème est de mesurer la force de cette association et ses variations dans le temps. Les deux indices présentés ci-dessus donnent des résultats contradictoires. Dans le cas général, la situation peut être modélisée par un tableau 2×2 comme les deux tableaux qui peuvent être construits à partir des données de Mercklé, le premier, pour l'année 1960, et l'autre, pour l'année 2010.

1960	Bac obtenu	Bac non obtenu
Enfants de cadres	45	55
Enfants d'ouvriers	5	95

2010	Bac obtenu	Bac non obtenu
Enfants de cadres	90	10
Enfants d'ouvriers	45	55

P. Mercklé après avoir évoqué d'autres indices possibles signale qu'« un autre indice s'est imposé parmi les sociologues, qui consiste à comparer plutôt des *odds ratios*, ou "rapports des chances relatives", autrement dit les rapports entre le taux de réussite et le taux d'échec ».

La suite de cet article permettra à la fois de définir cet indice mesurant la force de l'association, d'en faire l'historique et d'en explorer quelques applications.

2 La recherche d'un indice d'association d'après Yule (1912)

2.1 Biographie de George Udny Yule

George Udny Yule est né en Ecosse en 1871. Il obtient un diplôme d'ingénieur en 1892 au *University College* de Londres. Après un séjour d'un an à Bonn, consacré à l'étude des ondes électromagnétiques, sous la direction d'Heinrich Hertz, il rentre à Londres en 1893 et accepte un poste de chargé de travaux pratiques que lui offre Karl Pearson. Ce dernier était alors professeur de mathématiques appliquées à *University College* et avait connu Yule en tant qu'étudiant. Yule découvre en Pearson un enseignant stimulant et va rapidement produire des contributions fondamentales en théorie de la statistique. En 1895, il publie son premier article à caractère statistique sur la corrélation et est élu membre de la *Royal Statistical Society*. Ses qualités

pédagogiques et son attirance vers les sciences sociales et économiques rendent sa collaboration avec Pearson bénéfique pour l'ensemble du laboratoire, cependant il démissionne du *University College* en 1899 à la recherche d'un meilleur salaire. Malgré ce départ, il poursuit sa recherche en statistique, publie de nombreux articles sur l'association et la corrélation (1900, 1901, 1903) et il donne entre 1902 et 1909 des conférences, les *Newmarch Lectures in Statistics*, dont le contenu sera rassemblé sous la forme d'un ouvrage, *Introduction to the Theory of Statistics*, dont la première édition est publiée en 1911 [Yule, 1911].

L'année 1912 marque un tournant dans sa carrière, avec son recrutement comme maître de conférences à l'université de Cambridge. Il va y rester jusqu'à la fin de sa carrière. L'interruption due à la première guerre mondiale ne l'empêche pas de publier avec son ami Major Greenwood deux articles où l'on peut trouver les prémices d'une formalisation de ce qui sera appelé plus tard processus stochastique.

La période 1920-1930 est la plus prolifique avec la publication d'articles dans le domaine des séries chronologiques (corrélogramme, fondements des processus auto-régressifs). En 1921, il est élu membre de la *Royal Society* et il préside la *Royal Statistical Society* de 1924 à 1926.

À partir de 1931, sa santé fragile met progressivement fin à sa carrière scientifique. Il s'éteint à Cambridge, le 26 juin 1951, victime d'une crise cardiaque.

2.2 L'article de 1912

Le 23 avril 1912, Yule présente à la *Royal Statistical Society* de Londres une communication intitulée : « *Sur les méthodes de mesure de l'association entre deux attributs* » [Yule, 1912]. Ce long texte décrit de manière très pédagogique la recherche d'une évaluation de l'association entre deux attributs. Il suppose qu'une population donnée est partagée en quatre classes suivant deux divisions successives. Pour fixer les idées, Yule prend l'exemple des guérisons ou décès dans une population affectée par une épidémie de petite vérole, certains patients ayant été vaccinés et d'autres non.

On peut présenter les données sous forme d'un tableau, comme le suivant, portant sur l'épidémie de petite vérole à Sheffield en 1887-88 :

Tableau I. - *Épidémie de petite vérole à Sheffield*, 1887-88

	Guérisons	Décès	Total
Vaccinés	3951	200	4151
Non vaccinés	278	274	552
Total	4229	474	4703

2.2.1 Première étape : comparer les différences de proportions

Dans un premier temps, on peut tester l'existence d'une association entre les caractères vaccination et guérison en calculant la proportion des guérisons parmi les vaccinés, ici $0{,}952$ et la proportion des guérisons parmi les non-vaccinés, ici $0{,}504$. Il y a donc manifestement une association positive très marquée entre vaccination et guérison.

On peut aussi voir cette association en calculant la proportion des vaccinés parmi les guérisons, ici $0{,}934$ et la proportion des vaccinés parmi les décès, ici $0{,}422$.

Le problème que veut résoudre Yule est d'évaluer la force de cette association, en particulier si on souhaite comparer les résultats du tableau ci-dessus avec ceux d'une autre région ou à une autre période.

Yule définit donc la situation générale par la donnée d'un tableau deux lignes deux colonnes avec des notations qui nous semblent un petit peu inhabituelles. Nous avons donc préféré utiliser des notations simples déjà présentées dans nos articles précédents.

Dans le cas général, on a un tableau que l'on peut présenter comme suit :

	Guérisons	Décès	Total
Vaccinés	a	b	$a+b$
Non vaccinés	c	d	$c+d$
Total	$a+c$	$b+d$	$N=a+b+c+d$

Yule introduit ensuite les quatre proportions, notées respectivement p_1, p_2, p_3 et p_4.

On a : $$p_1=\frac{a}{a+c}\,;\,p_2=\frac{b}{b+d}\,;\,p_3=\frac{a}{a+b}\,;\,p_4=\frac{c}{c+d}.$$

Yule présente ensuite deux autres tableaux donnant les résultats obtenus respectivement à Leicester pour l'épidémie de 1892-93 et à Homerton et Fulham entre 1873 et 1885.

Tableau II. - *Épidémie de petite vérole à Leicester*, 1892-93

	Guérisons	Décès	Total
Vaccinés	197	2	199
Non vaccinés	139	19	158
Total	336	21	357

Tableau III. - *Cas de petite vérole à l'hôpital de Homerton*, 1873-84 *et à l'hôpital de Fulham*, 1880-85 : les cas douteux étant exclus

	Guérisons	Décès	Total
Vaccinés	8207	692	8899
Non vaccinés	1424	1103	2527
Total	9631	1795	11426

Pour comparer ces trois tableaux de proportions, Yule calcule d'abord pour chacun les proportions p_3 et p_4 puis leur différence p_3-p_4 pour essayer d'évaluer les forces respectives de l'association guérison-vaccination, d'où le tableau :

District ou hôpital	Proportion de guérisons parmi		Différence p_3-p_4
	Vaccinés : p_3	Non vaccinés : p_4	
Sheffield	0,952	0,504	0,448
Leicester	0,990	0,880	0,110
Homerton et Fulham	0,922	0,564	0,358

Activité 1

Le classement, pour la force de l'association, donne donc en tête Sheffield, suivi de Homerton-Fulham et enfin Leicester. Yule remarque qu'à Leicester, l'épidémie a été relativement douce et que, même si tous les vaccinés avaient guéri, la différence entre les taux de guérison des vaccinés et des non-vaccinés n'aurait pas dépassé $0{,}120$, reléguant Leicester en queue de peloton dans tous les cas.

Yule s'interroge sur la validité de ce critère et propose alors de s'intéresser aux différences $p_1 - p_2$ dont on a vu qu'elles mesuraient aussi l'association et de comparer les résultats avec ceux obtenus avec $p_3 - p_4$. Après calculs, Yule trouve que c'est toujours Sheffield qui est en tête mais suivi cette fois par Leicester, Homerton-Fulham fermant la marche.

Activité 2

2.2.2 Deuxième étape : comparer les valeurs réelles et celles obtenues sous l'hypothèse d'indépendance

Yule propose ensuite un autre indice d'association sous la forme de la différence δ entre la valeur réelle a et la valeur qu'on aurait dans le cas de l'indépendance.

Si les deux attributs étudiés sont indépendants, on doit avoir un tableau ayant les mêmes sommes marginales mais des effectifs différents dans le tableau. Notons-les comme Yule : a_0, b_0, c_0 et d_0.

Dans le cas d'indépendance, on a : $a_0 = \dfrac{(a_0 + b_0)(a_0 + c_0)}{N}$ et, puisque les sommes marginales sont les mêmes on a : $a_0 = \dfrac{(a + b)(a + c)}{N}$.

Yule propose d'étudier le paramètre : $\delta = a - a_0 = a - \dfrac{(a + b)(a + c)}{N}$.

En fait, Yule calcule $\dfrac{\delta}{N}$.

Dans les trois cas ci-dessus, Yule trouve les résultats suivants :

District ou hôpital	Valeur de $\dfrac{\delta}{N}$
Sheffield	$0{,}046$
Leicester	$0{,}027$
Homerton et Fulham	$0{,}062$

Activité 3

Cette fois, c'est Homerton-Fulham qui gagne devant Sheffield, suivi de Leicester.

Conclusion de Yule : « *les trois indices différents que nous avons essayés ont placé les districts dans trois ordres différents* » !

2.2.3 Troisième étape : déterminer l'indice par ses caractéristiques principales

Yule écrit :

> « *These illustrations suffice to show [...] that the choice of a measure or index of association is not quite a simple and straightforward matter : that the fundamental quantities which would serve quite well if all tables showed the same ratios for* $\frac{a+b}{N}$ *and* $\frac{a+c}{N}$ *give conflicting results when this condition fails to hold as it invariably fails in practice, and that consequently a useful purpose may be served by an index or "coefficient of association" of somewhat more complex form.* »
>
> « Ces exemples illustrent à eux seuls [...] que le choix d'une mesure ou d'un indice d'association n'est pas un problème simple et direct : que les grandeurs fondamentales qui pourraient être utiles pour tous les tableaux ayant les mêmes rapports $\frac{a+b}{N}$ et $\frac{a+c}{N}$, donnent des résultats contradictoires lorsque cette condition n'est pas remplie, comme c'est systématiquement le cas en pratique, et que, par conséquent, un indice ou "coefficient d'association" d'une forme un peu plus complexe serait utile. »
>
> traduit par nos soins

Il recherche alors un indice plus sophistiqué pour mesurer l'association et commence par énoncer les propriétés essentielles que cet indice doit satisfaire :

- il doit être égal à 0 si et seulement si les deux attributs sont indépendants ;
- pour des raisons de commodité, il doit être compris entre -1 et $+1$;
- il ne doit être égal à 1 ou -1 que si l'un au moins des quatre nombres a, b, c ou d est égal à 0 ;
- enfin, il doit être une fonction continue croissante de δ dans le cas où l'effectif total de la population N et les sommes marginales $a+b$ et $a+c$ sont fixés.

Avant d'examiner des propositions pour cet indice, Yule commence par donner quelques propriétés du coefficient δ.

a) Yule attribue à Karl Pearson la relation : $\delta = \dfrac{ad-bc}{N}$.

Activité 4

b) Yule indique qu'avec « un peu d'algèbre », on trouve :

$$p_1 - p_2 = \frac{N\delta}{(a+c)(b+d)} ;$$

$$p_3 - p_4 = \frac{N\delta}{(a+b)(c+d)}.$$

Activité 5

c) Yule en conclut que la plus grande valeur positive que δ peut prendre est soit $\frac{(a+b)(c+d)}{N}$, soit $\frac{(a+c)(b+d)}{N}$ et en déterminant la valeur négative la plus grande en valeur absolue que δ peut prendre, il en déduit que ces valeurs absolues extrêmes ne sont égales que si $a = d$ ou $b = c$, ce qui ne remplit donc pas les propriétés imposées plus haut. Il faut donc abandonner δ en tant que critère unique de la force de l'association.

Activité 6

2.2.4 Construction de Q et κ

Yule suggère ensuite de construire des coefficients qui ne soient pas affectés par la sélection de l'un ou l'autre des attributs. Pour cela il propose d'étudier un coefficient qu'il avait décrit dans son mémoire de 1900 [Yule, 1900] et noté Q, en hommage à Quételet.

À savoir :

$$\boldsymbol{Q = \frac{ad - bc}{ad + bc}}.$$

Ce coefficient ne provient pas de considérations extérieures, mais c'est une formule empirique qui remplit les quatre conditions énoncées ci-dessus :

- $Q = 0$ si et seulement si $\delta = 0$ (même numérateur) ce qui correspond au cas de l'indépendance ;
- quand $b = 0$ ou $c = 0$ alors $Q = +1$; quand $a = 0$ ou $d = 0$ alors $Q = -1$;
- enfin, Q est une fonction continue croissante de δ dans le cas où l'effectif total de la population N et les sommes marginales $a + b$ et $a + c$ sont fixés grâce à la démonstration suivante :

Démonstration. Soit $\kappa = \frac{bc}{ad} = \frac{(b_0 - \delta)(c_0 - \delta)}{(a_0 + \delta)(d_0 + \delta)}$, où a_0 est la valeur prise par a dans le cas de l'indépendance, valeur qui est constante pour des valeurs constantes de N, $a + b$ et $a + c$; et de même pour les autres.

On a alors : $Q = \frac{1 - \kappa}{1 + \kappa}$, et donc la dérivée de Q par rapport à κ est négative. Mais celle de κ par rapport à δ est aussi négative, donc la dérivée de Q par rapport à δ est positive.

□

Activité 7

Yule remarque ensuite que son coefficient Q , véritable premier « coefficient d'association » a été repris par Lipps en 1905 [Lipps, 1905]. Il a le mérite d'avoir une forme extrêmement simple et donc une grande rapidité de calcul. En revanche, il a le défaut de ne pas avoir la même simplicité d'interprétation.

Yule note que Q peut être exprimé en fonction des seuls p_1 et p_2 , ou des seuls p_3 et p_4.

À savoir :

$$Q = \frac{p_1(1 - p_2) - p_2(1 - p_1)}{p_1(1 - p_2) + p_2(1 - p_1)};$$

$$Q = \frac{p_3(1-p_4) - p_4(1-p_3)}{p_3(1-p_4) + p_3(1-p_4)}.$$

Même chose pour κ :

$$\kappa = \frac{p_2(1-p_1)}{p_1(1-p_2)};$$

$$\kappa = \frac{p_4(1-p_3)}{p_3(1-p_4)}.$$

Donc Q comme κ s'exprime par la même fonction de p_1 et p_2, d'une part, et de p_3 et p_4, d'autre part.

Ces coefficients lèvent donc le conflit relevé entre les résultats obtenus à partir soit de p_1 et p_2 seuls, soit de p_3 et p_4 seuls et remplissent l'objectif de Yule, de construire des coefficients non affectés par la sélection d'un attribut.

Activité 8

2.2.5 Construction de tableaux d'association réduits

Yule remarque d'abord qu'il est clair que les proportions marginales p_1 et p_2 ne changent pas si l'on multiplie, ou divise, l'une ou les deux colonnes du tableau par un facteur arbitraire. Le coefficient Q du tableau reste le même (puisque pouvant s'écrire uniquement en fonction de p_1 et p_2) bien que les valeurs de p_3 et p_4 sont changées ! Le même raisonnement conduit au fait que Q reste inchangé si l'on multiplie les lignes du tableau par un facteur arbitraire. Yule insiste sur le fait que cette propriété de Q est « la plus importante » car elle n'est pas simplement une propriété de manipulation des données et formules, elle a aussi une signification statistique essentielle : en effet, la proportion marginale d'un des attributs peut dépendre de circonstances purement arbitraires.
Par exemple, le nombre de vaccinés dépend de la volonté des autorités locales d'appliquer les lois sur la vaccination. De même, le nombre de guérisons peut être affecté dans une région où on expérimente un nouveau traitement, par exemple. Le coefficient Q, qui alors ne change pas, mesure donc la force réelle de l'association entre vaccination et guérison.

Pour comparer un tableau d'association avec un tableau de référence, Yule suggère donc de manipuler le deuxième tableau par multiplications des lignes et colonnes de manière à obtenir dans les deux tableaux les mêmes proportions marginales.

Considérons ainsi le tableau :

a	b
c	d

Si l'on multiplie la première ligne par x et la première colonne par y, les rapports des sommes marginales des deux premières lignes et des deux premières colonnes sont : $\frac{xya + xb}{yc + d}$ et $\frac{xya + yc}{xb + d}$. Si l'on veut que ces deux rapports soient ceux connus m et n du tableau de référence, on obtient les deux équations :

$$\begin{cases} xya + xb &= m(yc + d) \\ xya + yc &= n(xb + d) \end{cases}$$

La résolution de ce système conduit à une équation du second degré dont la résolution fournit les valeurs de x et y.

Ainsi si on veut réduire le tableau de Leicester en prenant comme référence celui de Sheffield, on trouve $x = 8{,}98$ et $y = 0{,}17$ et le tableau réduit de Leicester obtenu est le suivant :

Leicester	Guérisons	Décès	Total
Vaccinés	305,63	17,95	323,58
Non vaccinés	24,03	19	43,03
Total	329,66	36,95	366,61

Yule remarquant que le tableau réduit n'a pas le même nombre total d'observations que le tableau initial, construit, par proportionnalité, un tableau réduit ayant un nombre total d'observations de $10\,000$ individus :

Leicester	Guérisons	Décès	Total
Vaccinés	8 336	490	8 826
Non vaccinés	655	519	1 174
Total	8 991	1 009	10 000

qu'il peut comparer facilement à celui de Sheffield ramené, lui aussi, à un effectif total de $10\,000$ individus où les effectifs sont arrondis à l'entier le plus proche.

Sheffield	Guérisons	Décès	Total
Vaccinés	8 401	425	8 826
Non vaccinés	591	583	1 174
Total	8 992	1 008	10 000

On peut ainsi, par une observation visuelle, se rendre compte directement que l'association est plus forte à Sheffield qu'à Leicester. C'est ce qu'avaient indiqué les calculs des coefficients Q respectifs.

Activité 9

2.2.6 Construction de tableaux d'association symétriques

Enfin Yule remarque que dans l'exemple traité il n'y a aucune raison de choisir Sheffield comme tableau de référence plutôt qu'un autre. Il propose donc de réduire tous les tableaux à la forme « naturelle » : celle d'un tableau symétrique où toutes les proportions marginales sont égales deux à deux. C'est-à-dire qu'on a alors un tableau fictif où les nombres de vaccinés et de non-vaccinés sont égaux ainsi que les nombres de décès et de guérisons.

Par exemple, pour la comparaison des résultats de la vaccination, cela permet de compare les taux de guérison chez les vaccinés en ramenant pour chacun des cas les taux de décès et la proportion de non-vaccinés à 50%. Comme le remarque Yule, ce nouveau tableau est fictif mais garde les mêmes proportions marginales p_1 , p_2 , p_3 et p_4 et donc la même valeur de Q. Tout se passe comme « *si un démon tout puissant avec un mauvais caractère (sans relation avec le copain de Maxwell) avait visité Sheffield, Leicester et les deux hôpitaux londoniens et avait fait monter le taux de décès et la proportion de non-vaccinés chacun à* 50% *sans rien changer d'autre* ».

Dans le cas général, on part d'un tableau :

a	b
c	d

Un tableau symétrique équivalent est :

d	$\sqrt{\frac{bcd}{a}}$
$\sqrt{\frac{bcd}{a}}$	d

Activité 10

Dans ce nouveau tableau, on note p_0 et q_0 les proportions marginales (qui sont évidemment les mêmes suivant les lignes ou les colonnes) et on a :

$$p_0 = \frac{2d}{N'} \text{ et } q_0 = \frac{2\sqrt{\frac{bcd}{a}}}{N'}$$

où N' est l'effectif total de la population du nouveau tableau.

Yule introduit alors le coefficient $\boldsymbol{\omega = p_0 - q_0}$ et cherche à exprimer Q en fonction de ω.

On a : $p_0 = \frac{1}{1+\sqrt{\kappa}}$, $q_0 = \frac{\sqrt{\kappa}}{1+\sqrt{\kappa}}$ et donc $Q = \frac{1-\kappa}{1+\kappa} = \frac{2\omega}{1+\omega^2}$ soit $\omega = \frac{1-\sqrt{1-Q^2}}{Q}$.

Donc Q est une fonction simple de la différence entre les proportions marginales dans le tableau symétrique équivalent.
Yule pose alors le problème : pourquoi ne pas choisir ω comme coefficient d'association à la place de Q ?

Bien sûr ω vérifie comme Q, les conditions essentielles pour être désigné ; par exemple, parce que $\omega = \frac{1-\sqrt{\kappa}}{1+\sqrt{\kappa}}$. D'autre part, comme on peut le vérifier sur les trois tableaux précédents, Q et ω donnent des valeurs cohérentes.

Activité 11

Yule démontre ensuite que ω n'est autre que le coefficient de corrélation du tableau symétrique équivalent.

Activité 12

Tous ces résultats permettent à Yule de conclure que ce coefficient ω mérite d'être adopté et il propose de le nommer « *coefficient de colligation* ».

Activité 13

3 La controverse Pearson Yule.

C'est en 1900 que Pearson publia l'article où est défini le test du chi-deux (χ^2) [Pearson, 1900b]. Une extension de ce test permet de déterminer statistiquement si deux variables discrètes sont ou non associées.

De son côté, Yule cherchait à mesurer la force d'une telle association lorsqu'elle est avérée.

Pour le cas de deux variables continues, Pearson considérait qu'il avait déjà tranché la question en 1896 : pour lui, le coefficient de corrélation linéaire ρ est **la** mesure de la force de l'association entre de telles variables et on doit donc utiliser son estimation r.

Yule souhaitait traiter spécifiquement le cas de deux variables discrètes, A et B, ayant chacune exactement deux modalités notées, si nécessaire, 0 et 1. Il publia, aussi en 1900, son premier article à ce sujet : « *On the association of attributes in statistics* » [Yule, 1900]. Il y présente les données dans le tableau à double entrée suivant des variables indicatrices A et B des événements respectifs « Être malade » et « Être exposé au facteur de risque ».

	$B=1$	$B=0$
$A=1$	a	b
$A=0$	c	d

Tableau IV.

Pearson publia cette même année 1900, un article définissant le coefficient r_{hk} [Pearson, 1900a], qui est aujourd'hui noté r_T et appelé coefficient de corrélation tétrachorique, et qui est obtenu comme estimation du coefficient de corrélation linéaire entre deux variables latentes (c'est-à-dire inconnues) X et Y, continues, binormales, dont les variables de Bernoulli A et B découleraient par la donnée de deux seuils, h' et k', tels que :

$$\Big((A=1) \iff (X \geq h')\Big) \text{ et } \Big((B=1) \iff (Y \geq k')\Big)$$

.

C'est dans cet article que Pearson commença à critiquer publiquement le travail de Yule en exhibant trois nouveaux indicateurs différents de Q, qui vérifient effectivement les conditions de Yule et en montrant que ce n'était pas la valeur de Q qui était la meilleure approximation de son « étalon-or », le coefficient de corrélation tétrachorique. Selon son collègue Burton H. Camp (1933), Pearson pensait que le coefficient de corrélation polychorique (poly : généralisation de tétra) était une de ses plus importantes contributions à la statistique. Cependant, la popularité de ce coefficient souffrait de sa difficulté de calcul. C'est pourquoi Pearson popularisa des tables numériques destinées à contourner cette difficulté.

Yule réagit en 1906 en critiquant le principe de la corrélation tétrachorique en raison de son manque de cohérence interne. Répondant à son tour en 1907, Pearson écrivit : « *Je ne peux donc pas accepter le test de M. Yule comme très susceptible d'être utile ...* »

En 1911, Yule publia son manuel *Introduction to the Theory of Statistics* [Yule, 1911] dans lequel le coefficient de corrélation tétrachorique de Pearson n'était cité que dans une note de bas de page.

La controverse Yule-Pearson, à laquelle se joignit David Heron, confrère de Pearson, atteignit alors son apogée. Au sujet du manuel de Yule, David Heron écrivit en 1911 :

« *La théorie statistique a beaucoup souffert, dans le passé, de l'application illégitime de processus qui, lorsqu'ils sont appliqués à des données appropriées, sont parfaitement sains. Mais l'introduction, sans un seul mot d'avertissement, de méthodes qui, en aucun cas, ne peuvent donner des résultats corrects est*

beaucoup plus dangereuse. C'est particulièrement le cas lorsque ces méthodes prétendent raccourcir le travail de calcul des constantes statistiques, puisqu'elles sont invariablement adoptées par ceux qui, ne pouvant ou ne voulant pas examiner de façon critique leurs prétentions à la validité, sont dépendants de toute formule qui leur sera offerte. Un manuel de théorie statistique devrait avant tout être exempt de telles bévues ... » [Heron, 1911]

Dans la dernière partie de son article de 1912, Yule commence par donner son opinion sur le coefficient de corrélation tétrachorique de Pearson :

« *Le coefficient normal* [i.e. tétrachorique] *a obtenu sa réputation uniquement à partir de la croyance qu'il donnait la vraie corrélation entre les variables continues que les variables qualitatives étaient supposées représenter ; on l'appelle, en général, « la corrélation » sans qualificatif et la « méthode du tableau à quatre cases du professeur Pearson pour déterminer la corrélation », ou une expression équivalente. Il est vrai que l'auteur de la méthode a donné plusieurs avertissements sur son manque de fiabilité dans son mémoire original, mais ceux-ci semblent avoir été presque immédiatement oubliés, même par lui-même.* [...] *En conclusion, je peux peut-être ajouter que, si* A *et* B *sont, en fait, des variables continues et si rien n'est connu sur elles à l'exception des données contenues dans le tableau* 2×2*, rechercher un coefficient qui donnerait une approximation fiable du coefficient de corrélation* [linéaire] *entre les variables revient à rechercher quelque chose qui n'existe pas.* » (§ 54).

Puis, en dix pages, il réfute point par point les critiques de Heron.

Nous donnons en annexe un exemple de calcul du coefficient tétrachorique de Pearson et sa comparaison avec les coefficients de Yule.

En 1913, Pearson et Heron répondirent à cette critique de Yule par un article de plus de 150 pages, publié dans Biometrika, revue fondée et éditée par Pearson. Dans une partie de cet article [Pearson et Heron, 1913], où le manuel de Yule était à nouveau critiqué, on peut lire : « *Si les vues de M. Yule sont acceptées, des dommages irréparables seront infligés à la croissance de la théorie statistique moderne ...* [Le coefficient Q de Yule] *n'a jamais été et ne sera jamais utilisé dans aucun travail effectué sous la supervision* [de Pearson] [...] *Nous regrettons d'avoir à attirer l'attention sur la façon dont M. Yule s'est égaré à chaque étape de son traitement de l'association, mais la critique de ses méthodes s'est imposée à nous non seulement parce que nous avons récemment été agressés par lui, mais aussi à cause des éloges irréfléchis qui ont été accordés à un manuel qui, sur bien des points, ne peut qu'égarer les étudiants en statistique*. ».

Plus généralement, Pearson et Heron attaquèrent les « *notions à moitié au point* » de Yule et « *son raisonnement spécieux* » et soutinrent que Yule aurait à retirer ses idées « *s'il voulait conserver sa réputation de statisticien* ».

Rétrospectivement, on voit que les idées de Pearson et de Yule comportaient toutes deux des points positifs.

D'une part, il est clair que beaucoup de variables qualitatives n'ont pas de lois de probabilité continues sous-jacentes. Comme Yule l'a énoncé : « *Ceux qui sont non-vaccinés sont tous pareillement non-vaccinés et, de même, tous ceux qui sont morts de la variole sont tous pareillement morts.* »

Mais cependant, dans certains cas, l'existence de variables latentes ne peut pas être niée.

Enfin, on constate qu'aujourd'hui des chercheurs tendent à établir un lien théorique entre les visions, au départ si différentes, de Yule et de Pearson, par exemple Leo Goodman, qui écrit : « *Le lecteur sera surpris de voir qu'une réconciliation est possible* » [Goodman, 1981], ou Joakim Ekström : « *Il s'agit d'une tentative pour conduire le débat Pearson-Yule vers une réconciliation en montrant que leurs mesures d'association sont en fait plus semblables que différentes ...* » [Ekström, 2011].

4 L'approche de Fisher (1935) et les jumeaux criminels.

L'exemple présent dans l'article de Ronald A. Fisher [Fisher, 1935] (Partie I, p. 48, Exemple 1.), traite d'une étude portant sur 30 individus ayant un frère jumeau déjà condamné pour crime. Le frère jumeau de celui qui a été condamné aura-t-il la même probabilité d'être lui aussi condamné pour crime selon que les jumeaux sont monozygotes ou dizygotes ?

Les données recueillies sont reportées dans le tableau suivant :

Convictions of Like-sex Twins of Criminals.

	Convicted.	Not Convicted.	Total.
Monozygotic	10	3	13
Dizygotic	2	15	17
Total	12	18	30

Posons :

- $p =$ « probabilité pour un individu dont le frère jumeau a déjà été condamné pour crime, d'être lui aussi condamné pour crime, sachant que ce sont des jumeaux monozygotes » ;
- $1 - p = q =$ « probabilité pour un individu dont le frère jumeau a déjà été condamné pour crime, de ne pas être condamné pour crime, sachant que ce sont des jumeaux monozygotes » ;
- $p' =$ « probabilité pour un individu dont le frère jumeau a déjà été condamné pour crime, d'être lui aussi condamné pour crime, sachant que ce sont des jumeaux dizygotes » ;
- $1 - p^{'} = q^{'} =$ « probabilité pour un individu dont le frère jumeau a déjà été condamné pour crime, de ne pas être condamné pour crime, sachant que ce sont des jumeaux dizygotes » .

et

- $D =$ « nombre d'individus condamnés pour crime parmi les 17 dont le frère jumeau a déjà été condamné pour crime, sachant que ce sont des jumeaux dizygotes » ;
- $M =$ « nombre d'individus condamnés pour crime parmi les 13 dont le frère jumeau a déjà été condamné pour crime, sachant que ce sont des jumeaux monozygotes » ;
- $T =$ « nombre d'individus condamnés pour crime parmi les 30 dont le frère jumeau a été condamné pour crime » .

Le problème de test semble être celui-ci :

Tester avec le risque α l'hypothèse nulle $H_0 : p = p^{'}$ contre l'hypothèse alternative $H_1 : p > p^{'}$ (α sera choisi égal à 1% puis à 5%).

Rappel : une des méthodes pour effectuer ce test est la suivante :

- on fixe le risque α de se tromper en rejetant l'hypothèse nulle ;

- on construit, à l'aide d'une fonction (appelée statistique de test), un événement dépendant des résultats de l'expérience (donc aléatoire) qui a « peu de chance » de se produire si l'hypothèse nulle est vérifiée ;
- on calcule la probabilité, sous l'hypothèse nulle, d'obtenir cet événement.

Si, sous l'hypothèse nulle, cette probabilité est « trop » petite (ici, inférieure à α), on rejette l'hypothèse nulle.

Dans le cas de l'exemple, pour construire l'événement qui sert à effectuer le test, remarquons que connaissant la valeur t de T, le tableau est alors parfaitement déterminé par la valeur d prise par D. On peut alors donner une estimation des probabilités inconnues et avoir l'intuition que cet événement a ou non « peu de chance » de se produire si l'hypothèse nulle est vérifiée.

Si $t = 12$, on peut, par exemple, avoir les configurations suivantes

Pour $d = 1$:

	Condamné	Non condamné	Total
Monozygote	11	2	13
Dizygote	1	16	17
Total	12	18	30

Dans ce cas, parmi tous les individus ayant un frère jumeau condamné pour crime, la proportion observée d'individus condamnés pour crime chez les dizygotes est : $\frac{1}{17} = 0{,}06$ mais de $\frac{11}{13} = 0{,}84$ chez les monozygotes. On aura donc tendance à rejeter l'hypothèse nulle.

Pour $d = 7$:

	condamné	non condamné	Total
Monozygote	5	8	13
Dizygote	7	10	17
Total	12	18	30

Dans ce cas, parmi tous les individus ayant un frère jumeau condamné pour crime, la proportion observée d'individus condamnés pour crime chez les dizygotes est : $\frac{7}{17} = 0{,}41$ mais de $\frac{5}{13} = 0{,}38$ chez les monozygotes. On aura donc tendance à ne pas rejeter l'hypothèse nulle.

et, de façon générale :

	condamné	non condamné	Total
Monozygote	$12-d$	$d+1$	13
Dizygote	d	$17-d$	17
Total	12	18	30

Intuitivement, plus la valeur d prise par D sera petite, plus on aura tendance à rejeter : $H_0 : p = p^{'}$.
En conséquence, puisque $d = 2$, l'événement qui sera utilisé pour le test est : $(D \leq 2)$. Ici, la statistique de test est donc D.
Remarque sur les notations : Pour simplifier les notations, on admettra que, dans la suite du texte, toutes les probabilités seront calculées sachant que l'événement $(T = 12)$ est réalisé. $\mathbb{P}_{(p,p^{'})}$ désigne la loi de probabilité de D, qui dépend bien sûr de p et de $p^{'}$.

On rejettera donc H_0 si et seulement si, sous l'hypothèse $H_0 : p = p^{'}$, $\mathbb{P}_{(p,p^{'})}(D \leq 2) < \alpha$.

Après un calcul assez long, Fisher trouve $\mathbb{P}_{(p,p^{'})}(D \leq 2) = \dfrac{3095}{6653325} = 0{,}000465$.

Que ce soit avec le risque $\alpha = 5\%$ ou avec le risque $\alpha = 1\%$, il rejette l'hypothèse selon laquelle un individu ayant un frère jumeau condamné pour crime a la même probabilité d'être lui aussi condamné pour crime, que les jumeaux soient monozygotes ou dizygotes.

Plus généralement, il élargit son travail au cas où l'hypothèse sur les paramètres n'est pas obligatoirement $H_0 : p = p^{'}$ en établissant une formule qui permet de relier $\mathbb{P}_{(p,p^{'})}(D \leq 2)$ et un nombre réel positif, noté $\psi = \dfrac{p^{'}q}{pq^{'}}$ dans son article.

Cette formule est la suivante :

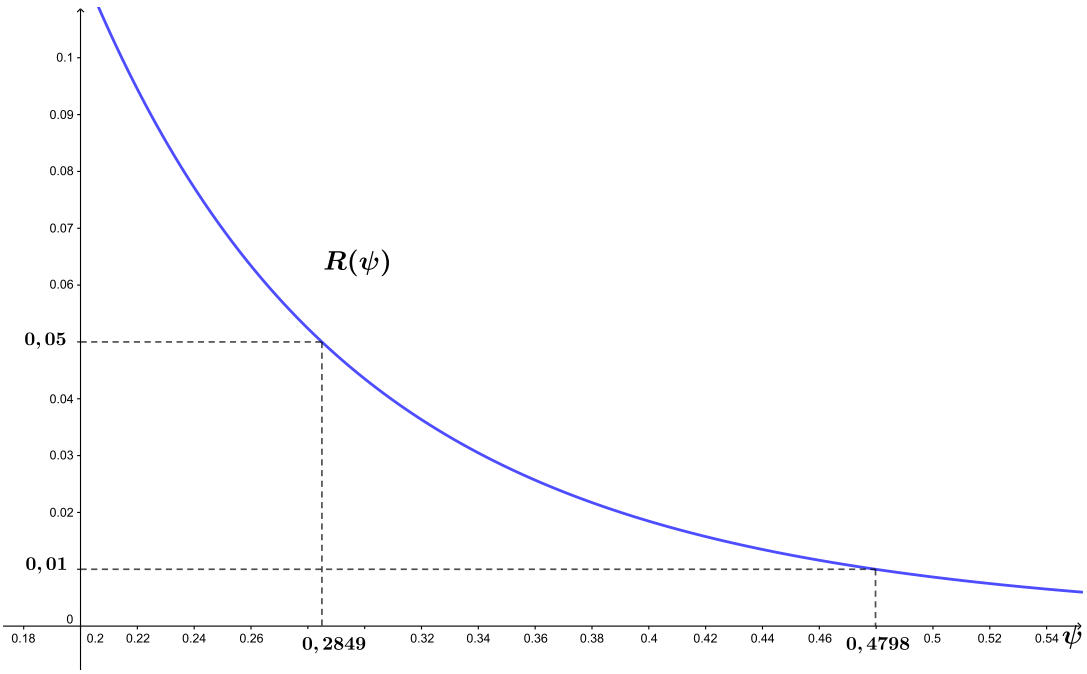
$$\mathbb{P}_{(p,p^{'})}(D \leq 2) = R(\psi) = \frac{A(\psi)}{B(\psi)}$$

où $A(\psi) = 1 + 102\psi + 2992\psi^2$

et $B(\psi) = 1 + 102\psi + 2992\psi^2 + 37400\psi^3 + 235620\psi^4 + 816816\psi^5 + 1633632\psi^6 + 1925352\psi^7 + 1337050\psi^8 + 534820\psi^9 + 116688\psi^{10} + 12376\psi^{11} + 476\psi^{12}$

Activité 14

Le graphe de la fonction R est donné ci-dessous :



Pour un α donné, on dira que les observations sont contradictoires avec les hypothèses sur p et $p^{'}$ au niveau de signification α, si et seulement si $\mathbb{P}_{(p,p')}(D \leq 2) < \alpha$.

Fisher examine d'abord le cas où $\alpha = 0{,}01$.

Puisque $R(0{,}4798) = 0{,}01$ et que $\psi \geq 0{,}4798 \iff R(\psi) \leq 0{,}01$, les hypothèses sont contradictoires avec les observations au niveau de signification $\alpha = 1\%$, si et seulement si $\psi \geq 0{,}4798$

Inversement, une hypothèse sur p et $p^{'}$ n'est pas contradictoire avec les données au niveau de signification $\alpha = 1\%$, si et seulement si $\mathbb{P}_{(p,p')}(D \leq 2) > 0{,}01$ et donc si $\psi \leq 0{,}4798$.

Mais $\psi \leq 0{,}4798$ entraîne $\dfrac{1}{\psi} \geq 2{,}084$ et donc $\dfrac{p}{q} \geq 2{,}084\dfrac{p^{'}}{q^{'}}$.

C'est ce qui permet à Fisher d'écrire :

> « *That is to say, that any hypothesis, which is not contradicted by the data at this level of signifiance, must make the ratio of criminals to non-criminals at least* 2.084 *times as high among the monozygotic as among the dizygotic cases.* »
>
> « Autrement dit, toute hypothèse qui n'est pas contradictoire avec les données à ce niveau de signification, doit aboutir à un rapport entre les criminels et les non-criminels $2{,}084$ fois plus élevé pour les jumeaux monozygotes que pour les jumeaux dizygotes.»
>
> traduit par nos soins

De même pour le cas où $\alpha = 5\%$.

Puisque $R(0{,}28496) = 0{,}05$ une hypothèse sur p et $p^{'}$ n'est pas contradictoire avec les données au niveau de signification $\alpha = 5\%$, si et seulement si $\mathbb{P}_{(p,p')}(D \leq 2) \geq 0{,}05$ et donc si $\psi \leq 0{,}28496$ ou encore $\dfrac{1}{\psi} \geq 3{,}5$ et donc $\dfrac{p}{q} \geq 3{,}5\dfrac{p^{'}}{q^{'}}$.

D'où la conclusion de Fisher :

> « *that any hypothesis, which is not contradicted by the data at the 5 per cent. level of significance, must make the ratio of criminals to non-criminals at least three and a half times as high among the monozygotic as among the dizygotic.* »
>
> « toute hypothèse qui n'est pas contradictoire avec les données à un niveau de signification de 5 pour cent, doit aboutir à un rapport entre les criminels et les non-criminels trois fois et demi plus élevé pour les jumeaux monozygotes que pour les dizygotes. »
>
> traduit par nos soins

Remarques :

1. Ce qui est remarquable dans le texte de Fisher est l'intérêt qu'il porte aux rapports $\frac{p}{q}$ et $\frac{p^{'}}{q^{'}}$ et à la relation entre ces deux rapports, en disant qu'un des deux est x fois plus grand que l'autre. Cette manière d'exprimer une comparaison est très parlante et couramment utilisée aujourd'hui.

2. Lorsqu'il s'agit de tester l'hypothèse nulle $H_0 : p = p^{'}$ contre l'hypothèse alternative $H_1 : p > p^{'}$ au niveau α, il suffit de remarquer que $H_0 : p = p^{'}$ est équivalent à $H_0 : \psi = 1$. Dans ce cas, la valeur de $\mathbb{P}_{(p,p')}(D \leq 2) = R(1)$ sera égale à la **p-value** et on rejette H_0 si et seulement si **p-value** $< \alpha$. On retrouve ce qui est noté dans la littérature comme le **test exact de Fisher**. Dans le cas présent, $R(1) = 0{,}0004652$ ce qui correspond approximativement à 1 cas sur 2 150 suivant le calcul de Fisher, d'où le rejet de l'hypothèse nulle au niveau 1%.

3. Dans beaucoup de problèmes de statistique, l'ensemble des valeurs d'un paramètre non contradictoires avec les données pour le niveau de signification α constitue un intervalle de confiance de niveau $1 - \alpha$. Ce qui joue ici le rôle d'intervalle de confiance de $\frac{1}{\psi}$ pour le niveau de confiance $1 - \alpha$ est $[\ 2{,}084\ ;\ +\infty\ [$ pour $\alpha = 0{,}01$ et $[\ 3{,}5\ ;\ +\infty\ [$ pour $\alpha = 0{,}05$.

5 Versions contemporaines de l'*odds ratio*.

5.1 Origines du concept et du mot.

Selon l'ouvrage de Herbert A. David et Anthony W. F. Edwards, *Annotated Readings in the History of Statistics* publié chez Springer [David et Edwards, 2001], le terme *odds ratio* apparaît pour la première fois dans l'article de David R. Cox, *The Regression Analysis of Binary Sequences* [Cox, 1958] publié dans le *Journal of the Royal Statistical Society*. Cox fait référence à l'article de Fisher [Fisher, 1935] étudié dans le paragraphe précédent et reprend sa notation ψ de ce paramètre.

Quant au concept d'*odds ratio*, il s'est construit progressivement grâce aux contributions successives de Yule (1903, 1912) (cf § 2), de Fisher (1935) (cf § 4), d'Edward Simpson [Simpson, 1951], de Jerome Cornfield (cf § 6.3.1) [Cornfield, 1951], de Barnet Woolf [Woolf, 1955], de John Haldane [Haldane, 1956].

Dans l'article *The Measure of Association in a* 2×2 *Table* [Edwards, 1963] publié dans le *Journal of the Royal Statistical Society*, A. Edwards montre que la mesure d'association dans un tableau 2×2 d'attributs associés doit être une fonction du *cross ratio* sans mentionner le terme d'*odds ratio* et analyse les contributions des auteurs cités ci-dessus.

5.2 Définitions et notations :

Un pathologie pouvant ou non se déclarer suite à l'exposition ou non à un facteur de risque, un indicateur important en épidémiologie est le rapport de deux probabilités :

- la probabilité d'être porteur de cette pathologie quand la personne est exposée au facteur de risque
- la probabilité d'être porteur de cette pathologie quand la personne n'est pas exposée au facteur de risque.

Ce rapport est dit « **rapport des risques** ou **risque relatif** ».
Lorsque les données sont issues d'un échantillon suite à une **étude de cohorte** [Faisant *et al.*, 2016] (p. 3), ce **rapport des risques peut être estimé directement** mais dans le cas d'une **étude cas-témoins** *ibid.* (p. 4), ce **rapport des risques ne peut être estimé directement**. Il est alors nécessaire de faire appel à un autre paramètre appelé l'*odds ratio*.

Notations :
Les événements qui peuvent être réalisés suite au tirage au hasard d'une personne sont notés ainsi :

- $D =$ « Malade » (de l'anglais *disease* : maladie) ;
- $\overline{D} =$ « Non malade » ;
- $E =$ « Exposé au facteur de risque » ;
- $\overline{E} =$ « Non exposé au facteur de risque ».

Les probabilités considérées dans l'étude sont alors :

- $r_1 = \mathbb{P}_E(D) =$ probabilité d'être **atteinte** par cette pathologie quand la personne **est exposée** au facteur de risque ;
- $r_2 = \mathbb{P}_{\overline{E}}(D) =$ probabilité d'être **atteinte** par cette pathologie quand la personne **n'est pas exposée** au facteur de risque ;
- $p_1 = \mathbb{P}_D(E) =$ probabilité d'**avoir été exposée** au facteur de risque sachant que la personne est **atteinte** par cette pathologie ;
- $p_2 = \mathbb{P}_{\overline{D}}(E) =$ probabilité d'**avoir été exposée** au facteur de risque sachant que la personne n'est pas **atteinte** par cette pathologie.

On introduit alors de nouveaux indicateurs, construits à partir du rapport entre la probabilité d'un événement et la probabilité de l'événement contraire, ce rapport étant nommé **cote** (**odds** en anglais) de l'événement.

La cote est exprimée généralement, non pas en terme de rapport de probabilités mais par un rapport de nombres entiers comme on peut le voir sur la photo suivante (cotes des paris sur le prénom du futur bébé de la famille royale britannique).

On peut alors considérer quatre sortes de cotes :

- **Cote de la maladie sous exposition au facteur de risque** : $o_E(D) = \dfrac{\mathbb{P}_E(D)}{\mathbb{P}_E(\overline{D})} = \dfrac{r_1}{1-r_1}$;
- **Cote de la maladie sous non-exposition au facteur de risque** : $o_{\overline{E}}(D) = \dfrac{\mathbb{P}_{\overline{E}}(D)}{\mathbb{P}_{\overline{E}}(\overline{D})} = \dfrac{r_2}{1-r_2}$;
- **Cote de l'exposition au facteur de risque chez les cas** (Malades) : $o_D(E) = \dfrac{\mathbb{P}_D(E)}{\mathbb{P}_D(\overline{E})} = \dfrac{p_1}{1-p_1}$;
- **Cote de l'exposition au facteur de risque chez les témoins** (Non malades) :

$$o_{\overline{D}}(E) = \frac{\mathbb{P}_{\overline{D}}(E)}{\mathbb{P}_{\overline{D}}(\overline{E})} = \frac{p_2}{1-p_2}.$$

Il est possible alors de définir deux rapports de cotes :

- **Rapport des cotes de la maladie**

$$or_D = \frac{o_E(D)}{o_{\overline{E}}(D)} = \frac{r_1(1-r_2)}{(1-r_1)r_2};$$

- **Rapport des cotes de l'exposition au facteur de risque**

$$or_E = \frac{o_D(E)}{o_{\overline{D}}(E)} = \frac{p_1(1-p_2)}{(1-p_1)p_2}.$$

On peut démontrer la proposition fondamentale suivante :

$$or_D = or_E$$

Activité 15

Définition
On appelle *odds ratio* :

$$\psi = or_D = or_E$$

Avec ces notations, le **rapport des risques** ou **risque relatif** est :

$$rr = \frac{r_1}{r_2}$$

L'absence d'influence du facteur de risque sur la maladie se traduit donc par $rr = 1$.

L'*odds ratio* possède les propriétés suivantes :

1. Si r_1 et r_2 sont « petits », alors
$$rr \simeq or_D = \psi$$
2. Si $\mathbb{P}(D)$ est « petit » alors
$$rr \simeq or_E = \psi$$

.

Activité 16

Lorsqu'on ne peut pas estimer directement le rapport des risques, pouvoir estimer l'*odds ratio* permet alors d'estimer le rapport des risques qui est l'objet d'intérêt en épidémiologie.

5.3 Estimation des indicateurs

Dans l'article *Smoking and Carcinoma of the Lung* [Doll et Hill, 1950], sont définis les deux grands types d'études rencontrées en épidémiologie, l'étude de cohorte et l'étude cas-témoins, études dont on rappelle ici la présentation des données.

5.3.1 Étude de cohorte

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

Dans une étude de cohorte :

- Le risque (probabilité) r_1 d'être **atteinte** par cette pathologie quand la personne **est exposée** au facteur de risque est estimé par $R_1 = \dfrac{a}{L_1}$;
- Le risque (probabilité) r_2 d'être **atteinte** par cette pathologie quand la personne **n'est pas exposée** au facteur de risque est estimé par $R_2 = \dfrac{c}{L_0}$;
- Le rapport de risques rr (risque relatif) est estimé par $RR = \dfrac{R_1}{R_2}$;

- La **cote de la maladie sous exposition au facteur de risque** $o_E(D)$ est estimée par $\dfrac{\dfrac{a}{L_1}}{1-\dfrac{a}{L_1}} = \dfrac{a}{b}$;

- La **cote de la maladie sous non-exposition au facteur de risque** $o_{\overline{E}}(D)$ est estimée par

$$\frac{\dfrac{c}{L_0}}{1-\dfrac{c}{L_0}} = \frac{c}{d};$$

- L'***odds ratio*** ψ est estimé par $OR = \dfrac{\dfrac{a}{b}}{\dfrac{c}{d}} = \dfrac{a \times d}{b \times c}$.

5.3.2 Étude cas-témoins

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

Dans une étude cas-témoins :

- Le risque (probabilité) r_1 d'être **atteinte** par cette pathologie quand la personne **est exposée** au facteur de risque ne peut être estimé correctement par $\dfrac{a}{a+b}$ car le nombre de personnes exposées $a+b$ n'est pas connu d'avance (il dépend en fait de C_0 et C_1) ;

- Le risque (probabilité) r_2 d'être **atteinte** par cette pathologie quand la personne **n'est pas exposée** au facteur de risque ne peut être correctement estimé par $\dfrac{c}{c+d}$ car le nombre de personnes non exposées $c+d$ n'est pas connu d'avance ;

- Puisque r_1 et r_2 ne peuvent être correctement estimés, le rapport de risque recherché rr ne peut être correctement estimé alors que c'est le cas dans une étude de cohorte ;

- La **cote de l'exposition au facteur de risque chez les cas** $o_D(E)$ estimée par $\dfrac{\dfrac{a}{C_1}}{1-\dfrac{a}{C_1}} = \dfrac{a}{c}$;

- La **cote de l'exposition au facteur de risque chez les témoins** $o_{\overline{D}(E)}$ estimée par $\dfrac{\dfrac{b}{C_0}}{1-\dfrac{b}{C_0}} = \dfrac{b}{d}$;

- Le **rapport des cotes pour l'exposition au facteur de risque** or_E est estimé par $OR = \dfrac{\frac{a}{c}}{\frac{b}{d}} = \dfrac{a \times d}{b \times c}$;

- Si $\mathbb{P}(D)$ est « petit » alors, puisque $rr \simeq or_E = \psi$, une **estimation du risque relatif** rr est celle de or_E c'est-à-dire $OR = \dfrac{a \times d}{b \times c}$.

Dans les deux cas, l'estimateur du rapport des cotes pour la maladie et l'estimateur du rapport des cotes pour l'exposition au facteur de risque a la même forme : $\dfrac{\frac{a}{c}}{\frac{b}{d}}$.

C'est pourquoi on nomme aussi ***odds ratio*** et on note OR ou Ψ la quantité : $OR = \dfrac{a \times d}{b \times c}$ qui n'est que l'estimateur d'un rapport des cotes.

5.4 Interprétation de OR dans une étude cas-témoins

Comme il a été vu précédement, si $\mathbb{P}(D)$ est « petit », un estimateur de rr est OR.

Si $OR = 1$ ou ne diffère pas significativement de 1 (ce qui revient à tester $H_0 : \psi = 1$ contre $H_1 : \psi \neq 1$ au risque $\alpha = 5\%$) , pas de différence entre être ou avoir été exposé ou non au facteur supposé à risque et le fait de voir survenir plus ou moins fréquemment la pathologie dont les cas souffrent.

Si $OR > 1$, diffère significativement de 1 et est égal à x, alors on estime que les sujets exposés au facteur de risque ont x fois plus de chance de voir survenir la pathologie présente chez les cas que chez les sujets exposés à ce même facteur : il y a significativement plus de cas dans le groupe exposé que dans le groupe non exposé. Le facteur est à risque délétère.

Si $OR < 1$, diffère significativement de 1 et est égal à x, alors on estime que les sujets non exposés au facteur de risque ont $\dfrac{1}{x}$ fois plus de chance de voir survenir la pathologie dont souffrent les cas que les sujets exposés à ce même facteur : il y a significativement moins de cas dans le groupe exposé que dans le groupe non exposé. Le facteur est protecteur de la maladie.

5.4.1 Outils pour l'interprétation

Test du chi-deux

Il permet de tester si OR diffère significativement de 1, c'est-à-dire de tester $H_0 : \psi = 1$ contre $H_1 : \psi \neq 1$ avec le risque $\alpha = 5\%$.

La statistique de test est :

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

Si $\chi^2_{obs} > 3{,}84$, OR est significativement différent de 1.

Construction de l'intervalle de confiance à 95% pour ψ

L'intervalle de confiance permet de construire un ensemble de valeurs pour le paramètre ψ tel que les données ne sont pas contradictoires avec les hypothèses sur ces valeurs de ψ au seuil de signification $\alpha = 0{,}05$.

Dans le cas d'une étude cas-témoins, si $\mathbb{P}(D)$ « petit », cet intervalle de confiance à 95%, sera approximativement aussi celui du rapport de risque rr.

Cet intervalle de confiance à 95% pour ψ permet aussi de tester $H_0 : \psi = 1$ contre $H_1 : \psi \neq 0$ avec le risque $\alpha = 0{,}05$. En effet, si 1 n'appartient pas à cet intervalle, on rejette H_0.

Deux intervalles sont souvent utilisés :

- **Méthode de Woolf ou semi-exacte**

$$\left[e^{lnOR - 1{,}96\sqrt{(\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d})}} \; ; \; e^{lnOR + 1{,}96\sqrt{(\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d})}} \right]$$

- **Méthode de Miettinen**

$$\left[OR^{1-\left(\frac{1{,}96}{\sqrt{\chi^2_{obs}}}\right)} \; ; \; OR^{1+\left(\frac{1{,}96}{\sqrt{\chi^2_{obs}}}\right)} \right]$$

La méthode de Cox

L'article de David Cox, *The Regression Analysis of Binary Sequences* [Cox, 1958], est adapté à une étude cas-témoins.

Soient n individus numérotés i ($i = 1, 2, ..., n$).

On note : $x_i = 1$, si le i-ème individu est malade (cas) et $x_i = 0$ sinon (témoin).

On définit n variables aléatoires $Y_1, Y_2, ..., Y_n$:

$Y_i = y_i$ où $y_i = 1$ si le i-ème individu a été exposé au facteur de risque et $y_i = 0$ sinon.

Suite à l'étude, on obtient donc une suite de n couples $(x_1, y_1), (x_2, y_2), ..., (x_i, y_i), ..., (x_n, y_n)$.

et le tableau de contingence :

	$x_i = 1$	$x_i = 0$	Total
$y_i = 1$	a	b	$a + b$
$y_i = 0$	c	d	$c + d$
	C_1 (fixé)	C_0 (fixé)	n

Réciproquement, la donnée d'un tel tableau de contingence permet de construire une suite de n couples (x_i, y_i).

Dans son article, David Cox étudie en particulier le rapport entre $\mathbb{P}(Y_i = 1)$ et x_i et propose la relation :

$$\ln\left(\frac{\mathbb{P}(Y_i=1)}{1-\mathbb{P}(Y_i=1)}\right)=\alpha+\beta x_i,$$

où α et β sont des paramètres inconnus.

Cette relation équivaut à $\mathbb{P}(Y_i=1)=\dfrac{e^{\alpha+\beta x_i}}{e^{\alpha+\beta x_i}+1}$.

Il s'agit donc d'effectuer une régression dite « logistique »

Si $\beta=0$, alors $\mathbb{P}(Y_i=1)$ ne dépend pas de la valeur de x_i, ce qui signifie que la probabilité d'avoir été exposé au facteur de risque est la même pour un cas que pour un témoin.

En reprenant la définition de ψ, $\psi=\dfrac{\mathbb{P}_D(E)}{1-\mathbb{P}_D(E)}\times\dfrac{1-\mathbb{P}_{\overline{D}}(E)}{\mathbb{P}_{\overline{D}}(E)}$ et en remarquant qu'ici :
$\mathbb{P}_D(E)=\mathbb{P}_{(x_i=1)}(Y_i=1)$, on retrouve la formule établie par Cox (page 221)

$$\psi=\frac{\text{pr}\,(Y_i=1\mid x_i=1)}{\text{pr}\,(Y_i=0\mid x_i=1)}\times\frac{\text{pr}\,(Y_i=0\mid x_i=0)}{\text{pr}\,(Y_i=1\mid x_i=0)}\ .$$

Cox en déduit la relation : $e^{\beta}=\psi$, relation qui montre que $\psi=1$ est équivalent à $\beta=0$.

Cela permet à Cox d'obtenir une méthode simple pour le calcul des limites de l'*odds ratio* qui est d'un intérêt fondamental : « *because the problem of getting a simple method of calculating limits for the odds ratio in a* 2×2 *table is of intrinsic interest* ». Il semble que ce soit la première apparition du terme *odds ratio* dans un texte mathématique.

Réciproquement, la construction d'un intervalle de confiance pour β permettra de construire un intervalle de confiance pour ψ et donc pour le rapport de risque.

Aujourd'hui, des logiciels permettent, à partir de la donnée d'une suite de couples (x_i,y_i), d'utiliser une procédure dite « GLM (*General Linear Model*) » pour donner une estimation et un intervalle de confiance de β et, par conséquent, une estimation et un intervalle de confiance de ψ, d'où un intervalle de confiance pour le rapport de risque rr.

6 Retour sur les situation précédentes.

Rappelons que l'*odds ratio* défini aujourd'hui est le rapport entre la cote avec facteur de risque et la cote sans le facteur de risque.

6.1 Application aux données de Yule issues de l'article publié en 1912

Sur un territoire donné, le risque étant ici le décès, on peut considérer que l'exposition au facteur de risque consiste dans le fait que la personne soit non-vaccinée. En conservant pour les données les mêmes notations que celles de Yule, mais en adoptant la disposition lignes-colonnes des épidémiologistes, le tableau se présente ainsi :

	Décès	Guérisons
Non vaccinés	d	c
Vaccinés	b	a

Par exemple, pour les données de Sheffield :

	Décès	Guérisons
Non vaccinés	274	278
Vaccinés	200	3951

L'*odds ratio* est :

$$OR = \frac{ad}{bc}$$

On constate que le coefficient $\kappa = \dfrac{bc}{ad}$ utilisé par Yule est relié à OR par la relation :

$$OR = \frac{1}{\kappa}$$

En appliquant ce résultat aux données de l'épidémie de petite vérole à Sheffield, on obtient :

$$\kappa = \frac{200 \times 278}{3951 \times 274} = 0{,}051359 \text{ et } OR = \frac{1}{0{,}051359} = 19{,}47.$$

En considérant que ces données sont issues d'une étude cas-témoins et en supposant la prévalence « petite », on peut dire qu'il y a $19{,}47$ fois plus de chances de décéder de la petite vérole chez les personnes non vaccinées que chez les personnes vaccinées.

Activité 17

6.2 Les jumeaux criminels.

Dans l'exemple de Fisher, on peut considérer que le fait d'être un jumeau monozygote constitue un facteur de risque d'être condamné pour crime quand son frère jumeau a déjà été condamné pour crime alors que le fait d'être un jumeau dizygote ne constitue pas un facteur de risque d'être condamné pour crime quand son frère jumeau a déjà été condamné pour crime.

Dans ce cas, la cote de condamnation pour crime pour un individu dont le frère jumeau est condamné pour crime, sachant que ce sont des jumeaux monozygotes, est : $\dfrac{p}{q}$ et la cote de condamnation pour crime pour un individu dont le frère jumeau est condamné pour crime, sachant que ce sont des jumeaux dizygotes, est : $\dfrac{p^{'}}{q^{'}}$. Le rapport des cotes est alors $\dfrac{pq^{'}}{qp^{'}}$. Quand Fisher établit que, pour un niveau de signification $\alpha = 5\%$, $\dfrac{p}{q} > 3{,}5\dfrac{p^{'}}{q^{'}}$, il sous-entend que pour des paramètres p et $p^{'}$ tels que $\dfrac{1}{\psi} \in [3{,}5\ ;\ +\infty[$, les données ne sont pas contradictoires avec ces hypothèses. Ce qui joue le rôle de l'*odds ratio* dans l'exemple de Fisher est en réalité $\dfrac{1}{\psi}$ où ψ est le paramètre défini par Fisher (qu'il ne nomme d'ailleurs pas *odds ratio*).

6.3 Application au cas de l'association entre pratique tabagique et apparition ou non d'un cancer du poumon.

6.3.1 La contribution de Cornfield

Dans son article [Cornfield, 1951], Jerome Cornfield utilise l'*odds ratio* (il semble que ce soit la première fois bien qu'il ne le nomme pas ainsi) pour calculer le rapport des risques lorsque la prévalence de la maladie est « petite ». Il utilise les données de Schrek datées de 1950 portant sur des hommes blancs âgés de 40 à 49 ans.

Reprenant les notations du paragraphe 5.1, les événements qui peuvent être réalisés sont notés ainsi :

- $D =$ « Atteint du cancer du poumon » ;
- $\overline{D} =$ « Non atteint du cancer du poumon » ;
- $E =$ « Fumeur » c'est-à-dire ayant fumé 10 cigarettes ou plus par jour ;
- $\overline{E} =$ « Non fumeur ».

Cornfield note X la prévalence $\mathbb{P}(D)$.

Remarque : les probabilités ci-dessous sont en réalité estimées par les proportions observées dans l'enquête.

- $p_1 = \mathbb{P}_D(E) = 0{,}77$ probabilité d'**avoir été fumeur** sachant que l'individu **a un cancer du poumon**.
- $p_2 = \mathbb{P}_{\overline{E}}(D) = 0{,}58$ probabilité d'**avoir un cancer du poumon** sachant que l'individu **n'a pas été fumeur**.

Il fait remarquer dans un premier temps que si X est « petit » alors le risque (probabilité) r_1, d'**avoir un cancer du poumon** quand on est **fumeur**, s'obtient par le calcul :

$$r_1 = \frac{p_1 X}{p_2}$$

et le risque (probabilité) r_2, d'**avoir un cancer du poumon** quand on est **non fumeur**, s'obtient par le calcul :

$$r_2 = \frac{(1-p_1)X}{1-p_2}$$

Cornfield énonce alors ce qui permettra de calculer le rapport des risques dans le cas d'une étude cas-témoins, proposition qui sera d'une grande importance dans l'analyse des données des études cas-témoins en épidémiologie.

> « *If one is interested only in knowing the relative amount by which the prevalence of disease is augmented by the possession of the attribute, one may calculate this whithout knowledge of* X*, since the ratio of the two rates is* $\dfrac{p_1(1-p_2)}{p_2(1-p_1)}$ *when* X *is small. One can thus conclude from the Schrek data alone that the prevalence of cancer of the lung among white males aged* $40-49$ *is* $2{,}4$ *times as high among those who smoke* 10 *or more cigarettes a day as among those who do not.* »

« Si l'on s'intéresse seulement à savoir de quelle quantité relative, la possession par l'individu de l'attribut en question augmente la prévalence, on peut la calculer sans connaissance de X, puisque le rapport des deux risques est $\frac{p_1(1-p_2)}{p_2(1-p_1)}$ quand X est petit. On peut alors conclure des seules données de Schrek que la prévalence du cancer du poumon parmi les hommes blancs agés de 40 à 49 ans est $2{,}4$ fois plus grande parmi ceux qui fument 10 cigarettes ou plus par jour que parmi ceux qui ne fument pas »

traduit par nos soins

Remarque :

Il est ici question « des seules données de Schrek ». En effet, en utilisant d'autres données, celles de Dorn, Cornfield estime X par $15{,}5$ pour 100000 et il peut calculer directement les risques r_1 et r_2.

6.3.2 Les données de Doll et Hill

Rappelons un des tableaux de données de l'article [Doll et Hill, 1950].

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	

En regroupant les femmes et les hommes, on obtient le tableau ci-dessous :

	Cancer du poumon	Pas de cancer du poumon	Total
Fumeurs	688	650	1338
Non fumeurs	21	59	80
Total	709	709	1418

La démarche est la suivante :

1. Calcul de OR

   $OR = \dfrac{688 \times 59}{650 \times 21} = 2{,}97$

2. On effectue un test du chi-deux pour déterminer si OR est significativement différent de 1.

   $\chi^2_{obs} = 19{,}2 > 3{,}84$ et $P = 0{,}000012 < 0{,}05$

   Le rapport de risque est significativement différent de 1.

3. On détermine l'intervalle de confiance de rr par la méthode de Woolf :

   $$[1{,}78 \ ; \ 4{,}94]$$

4. On détermine l'intervalle de confiance de rr par la méthode de Miettinen :

   $$[1{,}82 \ ; \ 4{,}83]$$

Utilisation de la méthode de Cox

Les résultats de cette régression logistique appliquée aux données de Doll et Hill sont les suivants :

- Estimation de $\beta = 1{,}089831$
- Intervalle de confiance de niveau 95% pour β : $[0{,}580\,;\ 1{,}60]$, ce qui signifie que l'on rejette l'hypothèse $H_0 : \beta = 0$ au risque 5%.
- Estimation de $\psi = e^{1{,}089831} = 2{,}97377$
- Intervalle de confiance de niveau 95% pour ψ donc pour rr : $[e^{0{,}580}; e^{1{,}60}] = [1{,}786721\,;\ 4{,}9494]$

Conclusion
Le risque de survenue d'un cancer du poumon est $2{,}97$ fois plus élevé chez les fumeurs que chez les non-fumeurs. Ce risque est « encadré » par l'un des trois intervalles de confiance calculés ci-dessus.

6.4 Les inégalités scolaires.

Dans l'exemple de Mercklé, le facteur de risque d'obtention du baccalauréat est celui d'être enfant de cadres plutôt qu'enfant d'ouvriers. Il s'agit de s'intéresser, non plus seulement à la probabilité d'obtention mais au rapport entre cette probabilité d'obtention et la probabilité de non-obtention du baccalauréat appelé aussi cote d'obtention. L'inégalité d'obtention entre enfants de cadres et enfants d'ouvriers est mesurée par le rapport des cotes, rapport nommé *odds ratio*. Pour conclure si les inégalités scolaires ont évolué entre le temps t_1 et le temps t_2, il suffit de comparer le rapport des cotes au temps t_1 avec le rapport des cotes au temps t_2.

6.4.1 Sous le point de vue de l'obtention du baccalauréat

En 1960

La cote d'obtention d'un enfant de cadres est égale à : $\frac{p_1}{1-p_1} = \frac{p_1}{q_1} = \frac{0{,}45}{0{,}55} = \frac{9}{11} = 0{,}98$.

On dit alors que la cote d'obtention d'un enfant de cadres est de 9 contre 11.

La cote d'obtention d'un enfant d'ouvriers est égale à : $\frac{p_2}{1-p_2} = \frac{p_2}{q_2} = \frac{0{,}05}{0{,}95} = \frac{5}{95} = 0{,}05$.

On dit alors que la cote d'obtention d'un enfant d'ouvriers est de 1 contre 19.

Le rapport des cotes d'obtention entre enfants de cadres et enfants d'ouvriers était égal à :

$$\frac{\frac{p_1}{q_1}}{\frac{p_2}{q_2}} = \frac{p_1 q_2}{q_1 p_2} = 15{,}54, \text{ ou, exprimé autrement,}$$

« la cote d'obtention des enfants de cadres est $15{,}54$ fois plus grande que celle des enfants d'ouvriers ».

En 2010

La cote d'obtention d'un enfant de cadres est égale à : $\frac{p_1}{1-p_1} = \frac{p_1}{q_1} = \frac{0{,}90}{0{,}10} = \frac{9}{1} = 9$.

On dit alors que la cote d'obtention des enfants de cadres est de 9 contre 1.

La cote d'obtention d'un enfant d'ouvriers est égale à : $\frac{p_2}{1-p_2} = \frac{p_2}{q_2} = \frac{0{,}45}{0{,}55} = \frac{9}{11} = 0{,}98$.

On dit alors que la cote d'obtention des enfants d'ouvriers est de 9 contre 11.

Le rapport des cotes d'obtention entre enfants de cadres et enfants d'ouvriers était de $\frac{\frac{p_2}{q_2}}{\frac{p_1}{q_1}} = \frac{p_1 q_2}{q_1 p_2} = 11$,

ou, exprimé autrement,

« la cote d'obtention des enfants de cadres est 11 fois plus grande que celle des enfants d'ouvriers ».

On est passé d'un *odds ratio* de $15{,}54$ en 1960 à un *odds ratio* de 11 en 2010.

On peut conclure qu'entre 1960 et 2010, l'inégalité d'obtention du baccalauréat a diminué.

6.4.2 Sous le point de vue de la non-obtention du baccalauréat

Il s'agit de s'intéresser, non plus seulement à la probabilité de non-obtention mais au rapport entre cette probabilité de non-obtention et la probabilité d'obtention appelée aussi cote de la non-obtention. L'inégalité de non-obtention entre enfants de cadres et enfants d'ouvriers est mesurée par le rapport des cotes.

En 1960

La cote de non-obtention d'un enfant de cadres est égale à : $\frac{q_1}{1-q_1} = \frac{q_1}{p_1} = \frac{0{,}55}{0{,}45} = \frac{11}{9} = 1{,}22$.

On dit alors que la cote de non-obtention d'un enfant de cadres est de 11 contre 9.

La cote de non-obtention d'un enfant d'ouvriers est égale à : $\dfrac{q_2}{1-q_2} = \dfrac{q_2}{p_2} = \dfrac{0{,}95}{0{,}05} = \dfrac{19}{1} = 19$.

On dit alors que la cote de non-obtention d'un enfant d'ouvriers est de 19 contre 1.

Le rapport des cotes de non-obtention du baccalauréat entre enfants de cadres et enfants d'ouvriers était égal à :

$\dfrac{\dfrac{q_1}{p_1}}{\dfrac{q_2}{p_2}} = \dfrac{q_1 p_2}{p_1 q_2} = 0{,}064$, ou, exprimé autrement : « la cote de non-obtention des enfants de cadres est $0{,}064$ fois plus grande que celle des enfants d'ouvriers », ou encore « la cote de non-obtention des enfants d'ouvriers est $\dfrac{1}{0{,}064} = 15{,}54$ fois plus grande que celle des enfants de cadres ».

En 2010

La cote de non-obtention d'un enfant de cadres est égale à : $\dfrac{q_1}{1-q_1} = \dfrac{q_1}{p_1} = \dfrac{0{,}10}{0{,}90} = \dfrac{1}{9} = 0{,}11$.

On dit alors que la cote de non-obtention d'un enfant de cadres est de 1 contre 9.

La cote de non-obtention d'un enfant d'ouvriers est égale à : $\dfrac{q_2}{1-q_2} = \dfrac{q_2}{p_2} = \dfrac{0{,}55}{0{,}45} = \dfrac{11}{9} = 1{,}22$.

On dit alors que la cote de non-obtention d'un enfant d'ouvriers est de 11 contre 9.

Le rapport des cotes de non-obtention entre enfants de cadres et enfants d'ouvriers était de :

$\dfrac{\dfrac{q_1}{p_1}}{\dfrac{q_2}{p_2}} = \dfrac{q_1 p_2}{p_1 q_2} = 0{,}09$, ou, exprimé autrement,

« la cote de non-obtention des enfants de cadres est $0{,}09$ fois plus grande que celle des enfants d'ouvriers », ou encore, « en 2010, la cote de non-obtention des enfants d'ouvriers est $\dfrac{1}{0{,}09} = 11$ fois plus grande que celle des enfants de cadres ».

On est passé d'un *odds ratio* de $15{,}54$ en 1960 à un *odds ratio* de 11 en 2010.
On peut conclure qu'entre 1960 et 2010, l'inégalité de non-obtention du baccalauréat a diminué.

L'utilisation de l'*odds ratio* semble avoir fait disparaître les contradictions entre les points de vue adoptés.

7 Conclusion

Le parcours que nous avons suivi concernant l'outil statistique appelé « *odds ratio* » ou « rapport des cotes » a montré une difficulté particulière pour un historien des sciences. En effet, on a vu que le 20e siècle a progressivement mis en place des méthodes pour étudier la force d'une association entre deux caractères. Nous avons comme principe de toujours lire et faire lire les textes fondateurs. Malheureusement, malgré nos efforts nous n'avons pas trouvé de tel texte fondateur : le langage et les méthodes sont au début très fluctuantes. On a vu qu'il s'agissait d'un travail collectif de l'épidémiologie anglo-saxonne et que l'idée s'est répandue partout dans la deuxième moitié du 20e siècle, y compris en s'appliquant à d'autres domaines comme la sociologie.

Un deuxième aspect de cette histoire est aussi surprenant, en tout cas pour des enseignants (dé)formés par une vision peut-être trop dogmatique des mathématiques. En effet, comme on l'a vu dans les travaux de Yule, la recherche s'est orientée non pas vers la découverte d'un indice universel justifié par une magnifique formule mais plutôt, d'une façon pragmatique très anglo-saxonne, vers la création d'un indice remplissant certaines conditions. Du coup, il peut y avoir plusieurs indices concurrents remplissant ces conditions et cela ouvre la porte éventuellement à des polémiques sur la validité respective de ces indices. Ce qui n'a pas manqué de se produire.

Nous avons essayé de montrer que l'épidémiologie, qui est le domaine d'émergence et de conceptualisation du rapport de cotes, a assuré son succès pour au moins deux raisons, d'une part la simplicité du calcul de cet outil et d'autre part la possibilité de remplacer les résultats qu'on aurait obtenus par une étude de cohorte par ceux issus d'une étude cas-témoins, cette dernière étant beaucoup moins coûteuse en moyens. Cette approximation est légitimée par la proposition que nous considérons comme fondamentale, énoncée et démontrée (*cf.* Activité 15 Chap. 5). C'est cette propriété qui a conduit à l'utilisation massive des rapports de cotes en épidémiologie (par exemple pour étudier les effets des vaccins). C'est devenu maintenant une technique automatique et largement répandue.

La question se complique quand on cherche à utiliser cet indice dans d'autres domaines comme la sociologie. Par exemple sur la question de la mesure des inégalités sociales à l'école (comme celle que nous avons présentée sur le baccalauréat), une intense polémique s'est développée entre sociologues à propos des résultats obtenus grâce au rapport des cotes. Le reproche principal est que l'*odds ratio* donnerait une « vision rose » de la situation. Nous n'avons bien sûr pas voulu rentrer plus avant dans ce débat dont certains termes nous paraissent abscons ou dérisoires.

Il ne faudrait pas oublier que, même en restant dans le domaine de l'épidémiologie, l'utilisation d'outils statistiques peut avoir des conséquences sociales et politiques importantes. On peut, par exemple, rappeler les mésaventures de la vaccination contre l'hépatite B. On peut consulter à ce sujet l'article de Jean-René Brunetière dans le numéro de novembre 2000 de la revue Pénombre. En résumé, la vaccination contre l'hépatite B a été largement pratiquée en France depuis 1994, en particulier, dans les collèges où elle a été systématique. Mais certains doutes ont émergé, non pas sur l'effet contre la maladie, mais sur l'apparition de cas de sclérose en plaques concomitants à la vaccination. Plusieurs études épidémiologiques ont été menées sous forme d'études cas-témoins avec calcul de l'*odds ratio* et de l'intervalle de confiance correspondant. Malheureusement les conclusions de ces études ne pouvaient ni confirmer, ni infirmer l'association entre vaccination et sclérose en plaques. Au vu de ces études, en 1998, Bernard Kouchner, alors secrétaire d'État à la santé, décidait de maintenir la vaccination pour les populations à risque, mais d'arrêter la vaccination systématique des collégiens. Le choix de la vaccination ou non des enfants est laissée à la famille en dialogue avec le médecin de famille ! C'est donc aux familles et aux médecins généralistes de se battre avec les l'*odds ratio* et les intervalles de confiance. D'autres études ultérieures n'ont pas apporté d'éléments nouveaux.

En inventant des outils complexes de plus en plus perfectionnés, la statistique s'assure des méthodes fiables pour produire des résultats. Mais il ne s'agit pas seulement d'utiliser des boîtes noires pour continuer à produire du chiffre, il faut aussi remettre en cause et démonter les dites boîtes noires pour porter dans l'espace public les modalités et les choix de la science. C'est ce que nous avons essayé de faire.

Annexe

Un exemple de calcul du coefficient r_{hk} de Pearson et des coefficients de Yule

En 1910, P. F. Everitt publia dans la revue Biometrika, un article destiné à rendre plus accessible le calcul de r_{hk} [Everitt, 1910]. Dix des quinze pages de cet article contiennent les résultats des 3000 calculs qu'il avait effectués pour constituer une table numérique spécialement conçue pour faciliter l'obtention de r_{hk}.

En tête de son article, il propose un exemple numérique explicatif dont les données figurent dans le tableau ci-dessous :

	$B = 1$	$B = 0$	Total
$A = 1$	$a = 1668$	$b = 131$	$a + b = 1799$
$A = 0$	$c = 137$	$d = 64$	$c + d = 201$
Total	$a + c = 1805$	$b + d = 195$	$N = a + b + c + d = 2000$

Tableau V.

Après avoir remarqué que les données sont organisées de manière que $\frac{b+d}{N} \leq \frac{1}{2}$ et que $\frac{c+d}{N} \leq \frac{1}{2}$, suivons Everitt sur la piste du nombre r_{hk}.

Le coefficient r_{hk} de Pearson

D'après Pearson, r_{hk} est une solution de l'équation d'inconnue r :

$$\frac{d}{N} = \frac{b+d}{N}.\frac{c+d}{N} + \sum_{n=1}^{+\infty}(\tau_n.\tau_n{}'.r^n)$$

où τ_n est l'image de $h = \frac{h'}{\sigma_X}$ par une certaine fonction numérique f_n, construite à partir de la fonction de répartition de la loi normale centrée réduite, Φ, et d'une suite récurrente d'ordre 2, $\tau_n{}'$ étant l'image de $k = \frac{k'}{\sigma_Y}$ par la même fonction. Or il n'est pas utile de rechercher h', k', σ_X et σ_Y car on peut estimer directement $\frac{h'}{\sigma_X}$ par $\Phi^{-1}\left(1 - \frac{b+d}{N}\right)$ et $\frac{k'}{\sigma_Y}$ par $\Phi^{-1}\left(1 - \frac{c+d}{N}\right)$.

Par ailleurs, Everitt considère la somme des termes en r de degré strictement supérieur à 6 comme négligeable et, donc, il s'agit de résoudre dans [0 ; 1] l'équation polynomiale :

$$\frac{d}{N} = \frac{b+d}{N}.\frac{c+d}{N} + \sum_{n=1}^{6}(\tau_n.\tau_n{}'.r^n)$$

Or les 3000 calculs effectués par Everitt lui ont donné les images par f_n, $1 \leq n \leq 6$, des nombres décimaux depuis 0,001 jusqu'à 0,500 par pas de 0,001. Nous avons :

$$\frac{b+d}{N} \approx 0,09750 \approx 0,098 \text{ et } \frac{c+d}{N} \approx 0,10050 \approx 0,101 \text{ et } \frac{d}{N} \approx 0,032000$$

En conséquence, Everitt recopie les lignes de sa table correspondant respectivement à $\frac{b+d}{N} \approx 0,098$ et à $\frac{c+d}{N} \approx 0,101$:

TABLE—(*continued*).

$\frac{1}{2}(1-\alpha)$	r_1	r_2	r_3	r_4	r_5	r_6	h
·098	·17292	·15811	+·04744	−·06061	−·06687	+·00897	1·29303
·101	·17678	·15948	+·04531	−·06317	−·06644	+·01153	1·27587

N. B. *Ne pas tenir compte du titre porté en première colonne* ; en référence à un article de W. F. Sheppard précédemment publié dans Biometrika, Everitt utilise la notation $\frac{1}{2}(1-\alpha)$ pour représenter $\frac{b+d}{N}$ ou $\frac{c+d}{N}$.)

Il multiplie entre eux les nombres situés dans une même colonne et obtient l'équation à résoudre

$$0,032000 = 0,009898 + 0,030569r + 0,025215r^2 + 0,002150r^3 + 0,003829r^4 + 0,004443r^5 + 0,000103r^6$$

Everitt résout cette équation par la méthode de Newton et trouve :

$$r_{hk} = 0,498$$

Qui ne serait impressionné par une telle débauche d'énergie ?

Activité 18

Activité 19

Passons maintenant au calcul des valeurs des différents coefficients définis par Yule pour les données de l'exemple d'Everitt ci-dessus.

Le coefficient Q de Yule

Le calcul est très simple :

$$Q = \frac{ad-bc}{ad+bc} = \frac{1668 \times 64 - 137 \times 131}{1668 \times 64 + 137 \times 131} \approx 0{,}712$$

Le coefficient ω (ou Y) de Yule

Dans son article de 1912, Yule a, de plus, défini le coefficient de colligation ω qui, aujourd'hui, est généralement noté Y. On a :

$$Y = \frac{\sqrt{ad}-\sqrt{bc}}{\sqrt{ad}+\sqrt{bc}} = \frac{\sqrt{1668 \times 64} - \sqrt{137 \times 131}}{\sqrt{1668 \times 64} + \sqrt{137 \times 131}} \approx 0{,}418$$

Le coefficient de corrélation linéaire

Enfin Yule s'est intéressé au coefficient de corrélation linéaire des variables A et B elles-mêmes. Prenant, comme plus haut, les valeurs conventionnelles 0 et 1, il appelle ce troisième coefficient le « *product sum coefficient* » et le note r. Aujourd'hui, il est souvent appelé le coefficient Phi.

Activité 20

Bibliographie

[Cornfield, 1951] CORNFIELD, J. (1951). A Method of Estimating Comparative Rates from Clinical Data. Applications to Cancer of the Lung, Breast and Cervix. *Journal of the National Cancer Institute*. 11, pp. 1269-1275.

[Cox, 1958] COX, D. R. (1958). The Regression Analysis of Binary Sequences. *Journal of the Royal Statistical Society*. Series B, vol. XX, N^{0}2., pp. 215-242.

[David et Edwards, 2001] DAVID, H. A. et EDWARDS, A. W. F. (2001). *Annotated Readings in the History of Statistics*. Springer.

[Doll et Hill, 1950] DOLL, R. et HILL, A. B. (1950). Smoking and Carcinoma of the Lung. *British Medical Journal*.

[Edwards, 1963] EDWARDS, A. W. F. (1963). The Measure of Association in a 2×2 Table. *Journal of the Royal Statistical Society*. Vol. 126, N^{0}1, pp. 109-114.

[Ekström, 2011] EKSTRÖM, J. (2011). The Phi-coefficient, the Tetrachoric Correlation Coefficient and the Pearson-Yule Debate. *UCLA, Departement of Statistics Papers*.

[Everitt, 1910] EVERITT, P. E. (1910). Tables of the Tetrachoric Functions for Fourfold Correlation Tables. *Biometrika*. Vol. 7, Issue 4, pp. 437-451.

[Faisant *et al.*, 2016] FAISANT, J., LANIER, D., LEJEUNE, J., MORELLO, R. et TROTOUX, D. (2016). La statistique du chi-deux : son usage à partir de l'article de doll et hill sur l'association entre cancer et tabac. *Irem de Caen Normandie*. Article en ligne : `https://irem.unicaen.fr/spip.php?article180`, consulté le 3 décembre 2018.

[Fisher, 1935] FISHER, R. A. (1935). The logic of inductive inference. *Journal of the Royal Statistical Society*. Vol 98, N^0 1, pp. 39-82.

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

[Goodman, 1981] GOODMAN, L. (1981). Association Models and the Bivariate Normal for the Contingency Tables with Ordered Categories. *Biometrika*. 68, 2, pp. 347-365.

[Gorroochurn, 2016] GORROOCHURN, P. (2016). *Classic Topics on the History of Modern Mathematical Statistics*. Wiley.

[Haldane, 1956] HALDANE, J. B. S. (1956). The Estimation and Significance of the Logarithm of a Ratio of Frequencies. *Annals of Human Genetics*. 20, pp. 309-311.

[Heron, 1911] HERON, D. (1911). The Danger of Certain Formulæ Suggested as Substitutes for the Correlation Coefficient. *Biometrika*. Vol. 8, pp. 109-122.

[Lipps, 1905] LIPPS, G. F. (1905). Die Bestimmung der Abhängigkeit zwischen den Merkmalen eines Gegenstandes. *Berichte über die Vorhandlungen der Königlich Sächsischen Gesellschaft der Wissenschaften zu Leipzig, Mathematisch-Physische Klasse*. Vol. 57, pp. 1-32.

[Mackenzie, 1981] MACKENZIE, D. (1981). *Statistics in Britain 1865-1930 : The Social Construction of Scientific Knowledge*. University Press, Edinburgh. trad. Dominique Erbnöther, in Les Scientifiques et leurs alliés, Michel Callon et Bruno Latour, Pandore, Paris, 1985, pp. 200-260.

[Mercklé, 2012] MERCKLÉ, P. (2012). Les inégalités scolaires diminuent-elles ? *Le Monde*. Supplément Science et Techno du 7 juin 2012.

[Pearson, 1900a] PEARSON, K. (1900a). Mathematical Contribution to the Theory of Evolution. - vii. On the Correlation of Characters not Quantitavely Mesurable. *Philosophical Transactions*, pages 1–47.

[Pearson, 1900b] PEARSON, K. (1900b). 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*.

[Pearson et Heron, 1913] PEARSON, K. et HERON, D. (1913). On Theories of Association. *Biometrika*. Vol. 9, Issue 1-2, pp. 159-315.

[Simpson, 1951] SIMPSON, E. H. (1951). The Interpretation of Interaction in Contingency Tables. *Journal of the Royal Statistical Society*. Series B (Methodological), Vol. 13, N^{0}2, pp. 238-241.

[Thélot et Vallet, 2000] THÉLOT, C. et VALLET, A. (2000). La réduction des inégalités sociales devant l'école depuis le début du siècle. *Économie et Statistique*. n° 334, pp. 3-32.

[Woolf, 1955] WOOLF, B. (1955). On Estimating the Relation Between Blood Group and Disease. *Annals of Human Genetics*. 19, pp. 251-253.

[Yule, 1900] YULE, G. U. (1900). On the Association of Attributes in Statistics. *In G. U. Yule Statistical Papers*, pages 7–69. Alan Stuart and Maurice G. Kendall, Griffin, London.

[Yule, 1911] YULE, G. U. (1911). *Introduction to the Theory of Statistics*. Griffin, London.

[Yule, 1912] YULE, G. U. (1912). On the Methods of Measuring the Association Between Two Attributes. *In G. U. Yule Statistical Papers*, pages 107–170. Alan Stuart and Maurice G. Kendall, Griffin, London.

Table des matières

Activité 1

Vérifier les résultats de Yule.

Retour à l'article

Une solution de cette activité

Activité 2

Faire les calculs avec p_1 et p_2 et vérifier le classement indiqué par Yule.

Retour à l'article

Une solution de cette activité

Activité 3

1. Démontrer que dans le cas de l'indépendance, on a : $a_0 = \dfrac{(a_0 + b_0)(a_0 + c_0)}{N}$.
   Vérifier les calculs dans les trois cas.
2. Calculer aussi b_0, c_0 et d_0. Que trouverait-on en prenant comme définition de δ :
   - $\delta = b - b_0$?
   - $\delta = c - c_0$?
   - $\delta = d - d_0$?

Retour à l'article

Une solution de cette activité

Activité 4

Démontrer la relation : $\delta = \dfrac{ad - bc}{N}$.

Retour à l'article

Une solution de cette activité

Activité 5

Prouver que :

$$p_1 - p_2 = \frac{N\delta}{(a+c)(b+d)};$$

$$p_3 - p_4 = \frac{N\delta}{(a+b)(c+d)}.$$

Retour à l'article

Une solution de cette activité

Activité 6

1. Montrer que la plus grande valeur positive que δ peut prendre est soit $\dfrac{(a+b)(b+d)}{N}$, soit $\dfrac{(a+c)(c+d)}{N}$.
2. Déterminer la valeur négative la plus grande en valeur absolue que peut prendre δ.
3. Pour quelles valeurs de a, b, c et d, ces valeurs absolues extrêmes sont-elles égales ?

Retour à l'article

Une solution de cette activité

Activité 7

Démontrer que la dérivée de κ par rapport à δ est négative.

Retour à l'article

Une solution de cette activité

Activité 8

Calculer le coefficient Q pour les trois districts des tableaux I, II et III.

Retour à l'article

Une solution de cette activité

Activité 9

Vérifier les calculs de Yule. Utiliser la même méthode pour réduire le tableau de Homerton et Fulham à la forme de Sheffield. Construire le tableau réduit de Homerton et Fulham à la forme de Sheffield avec un effectif total de $10\,000$ individus. Vérifier enfin que l'association est plus forte à Leicester qu'à Homerton-Fulham.

Retour à l'article

Une solution de cette activité

Activité 10

Montrer que le tableau symétrique :

d	$\sqrt{\frac{bcd}{a}}$
$\sqrt{\frac{bcd}{a}}$	d

peut être obtenu à partir du tableau suivant :

a	b
c	d

en multipliant la première ligne par un facteur et la première colonne par un autre facteur, les deux facteurs étant choisis de manière que les sommes marginales soient toutes égales.

Retour à l'article

Une solution de cette activité

Activité 11

Dresser un tableau donnant pour les trois districts les valeurs de Q et ω. Conclusion ?

Retour à l'article

Une solution de cette activité

Activité 12

Vérifier que le coefficient de corrélation du tableau symétrique est égal au coefficient de colligation ω.

Indication
Pour calculer le coefficient de corrélation du tableau symétrique, on peut ramener celui-ci au tableau du couple de variables aléatoires de Bernoulli, $(\mathbb{1}_{\text{Être guéri}}, \mathbb{1}_{\text{Être vacciné}})$, suivant :

	Première variable		
Deuxième variable	0	1	Total
0	$\frac{p_0}{2}$	$\frac{q_0}{2}$	$\frac{1}{2}$
1	$\frac{q_0}{2}$	$\frac{p_0}{2}$	$\frac{1}{2}$
Total	$\frac{1}{2}$	$\frac{1}{2}$	1

Retour à l'article

Une solution de cette activité

Activité 13

Yule a proposé d'utiliser ω à la place de Q car Q, facile à calculer, était difficile à interpréter en termes statistiques.

ω est-il meilleur de ce point de vue ?

Retour à l'article

Une solution de cette activité

Activité 14

Démontrer la formule donnant la fonction R.

Retour à l'article

Une solution de cette activité

Activité 15

Démontrer la proposition fondamentale :

$$or_D = or_E$$

Retour à l'article

Une solution de cette activité

Activité 16

Démontrer les deux propriétés de l'*odds ratio* suivantes :

1. Si r_1 et r_2 sont « petits », alors
$$rr \simeq or_D = \psi.$$

2. Si $\mathbb{P}(D)$ est « petit » alors
$$rr \simeq or_E = \psi.$$

Retour à l'article

Une solution de cette activité

Activité 17

Calculer l'*odds ratio* sur les données des hôpitaux de Leicester et celles de Homerton et Fulham. Comparer alors les trois hôpitaux du point de vue de l'*odds ratio*.

Retour à l'article

Une solution de cette activité

Activité 18

Utiliser un tableur (ou un autre logiciel de votre choix) pour résoudre l'équation ci-dessous :

$$0,032000 = 0,009898 + 0,030569r + 0,025215r^2 + 0,002150r^3 + 0,003829r^4 + 0,004443r^5 + 0,000103r^6$$

.

Retour à l'article

Une solution de cette activité

Activité 19

Après avoir lu les trois premières pages de l'article [Everitt, 1910] (accessible à l'adresse `https://www.jstor.org/stable/2345377`), reproduites ci-dessous, utiliser un tableur (ou un autre logiciel de votre choix) pour vérifier les deux lignes ci-dessous, extraites de la table numérique d'Everitt.

TABLE—(*continued*).

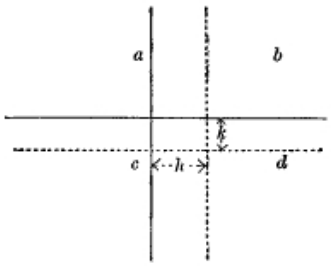
$\frac{1}{2}(1-\alpha)$	τ_1	τ_2	τ_3	τ_4	τ_5	τ_6	h
·098	·17292	·15811	+·04744	−·06061	−·06687	+·00897	1·29303
·101	·17678	·15948	+·04531	−·06317	−·06644	+·01153	1·27587

TABLES OF THE TETRACHORIC FUNCTIONS FOR FOURFOLD CORRELATION TABLES.

By P. F. EVERITT, B.Sc.

Explanation of the Tables.

The method of the fourfold table for determining correlation was described by Pearson in *Phil. Trans.* A. vol. 195, pp. 1—47.



a	b	$a+b$
c	d	$c+d$
$a+c$	$b+d$	N

Following his nomenclature, the normal correlation surface is divided into four parts by two planes at right angles to the axes of x and y at distances h', k' from the origin and these distances h' and k', when measured in terms of the standard deviations of their respective variables, are called h and k. The volumes or frequencies in the four divisions are represented by a, b, c, d in the manner shown in the plan and it will be seen, that $b+d$ and $c+d$, owing to the position given to the point of intersection of the traces of the dividing planes, cannot exceed $\frac{1}{2}N$.

In the paper referred to, the correlation coefficient r is determined from the equation

$$\frac{d}{N} = \frac{b+d}{N} \cdot \frac{c+d}{N} + \sum_1^\infty \left(\frac{r^n}{n!} HK \bar{v}_{n-1} \bar{w}_{n-1} \right),$$

where H, K are the ordinates of the normal curve of area N corresponding to the abscissae h and k and consequently dividing the curve into areas, of which the proportions to the whole are $\frac{b+d}{N}$ and $\frac{c+d}{N}$ respectively, while $\bar{v}$ and $\bar{w}$ are given by

$$\bar{v}_n = h\bar{v}_{n-1} - (n-1)\bar{v}_{n-2}, \qquad \bar{w}_n = k\bar{w}_{n-1} - (n-1)\bar{w}_{n-2},$$
$$\bar{v}_0 = 1, \quad \bar{v}_1 = h, \qquad \bar{w}_0 = 1, \quad \bar{w}_1 = k.$$

Page 437

Now let

$$\tau_n = \frac{H\overline{v}_{n-1}}{\sqrt{n!}} \text{ and } \tau_n' = \frac{K\overline{w}_{n-1}}{\sqrt{n!}},$$

then the equation becomes

$$\frac{d}{N} = \frac{b+d}{N} \cdot \frac{c+d}{N} + \sum_1^\infty (\tau_n \tau_n' r^n).$$

It is clear from the above, that τ_n' is the same function of $\frac{c+d}{N}$ as τ_n is of $\frac{b+d}{N}$ and that one table of these functions will serve for both, if we enter the table with $\frac{b+d}{N}$ as argument for the latter, and with $\frac{c+d}{N}$ for the former. It should be noted that these quantities $\frac{b+d}{N}$, $\frac{c+d}{N}$ are identical with $\frac{1}{2}(1-\alpha)$, where $\frac{1}{2}(1+\alpha)$ and α are used as arguments by Sheppard* in his published tables and to avoid ambiguity have been tabulated under that heading.

In the present tables the values of the first six τ functions, henceforth to be termed *tetrachoric* functions, have been computed for values of $\frac{1}{2}(1-\alpha)$ from ·001 to ·500 by successive increments of ·001; the last column contains the values of h† (or k) corresponding to the value of $\frac{1}{2}(1-\alpha)$ given in the first column, and required for computing the functions of higher order than the sixth as well as for the probable error of r.

In the auxiliary table are given the values of p_n and q_n required to compute the functions of the seventh to twelfth orders by means of the difference formula

$$\tau_n = hp_n\tau_{n-1} - q_n\tau_{n-2}.$$

Illustrations of the Use of the Tables.

(*a*) With interpolation.

Consider the hypothetical table given below:

1668	131	1799
137	64	201
1805	195	2000

Here $\frac{b+d}{N} = ·09750$, $\frac{c+d}{N} = ·10050$, $\frac{d}{N} = ·032000$.

* W. F. Sheppard, "New Tables of the Probability Integral," *Biometrika*, Vol. II. p. 174.

† x of Sheppard's Tables.

Now enter the table twice using interpolation, once with $\frac{1}{2}(1-\alpha)$ equal to ·09750 and a second time with it equal to ·10050, and we have

$\frac{1}{2}(1-\alpha)$	τ_1	τ_2	τ_3	τ_4	τ_5	τ_6
·09750	+ ·17228	+ ·15787	+ ·04779	− ·06018	− ·06693	+ ·00854
·10050	+ ·17614	+ ·15926	+ ·04567	− ·06275	− ·06652	+ ·01111

Multiplying the numbers in each column together, we find the equation

$$·032000 = ·009799 + ·030345r + ·025142r^2 + ·002183r^3 + ·003776r^4 + ·004452r^5 + ·000095r^6;$$

whence, solving by Newton's method,

$$r = ·501.$$

An illustration of the calculation of the probable error of r when found as above is given in Pearson's paper on p. 36.

(*b*) Without interpolation.

Using the same table as in example (*a*), we have as before

$$\frac{b+d}{N} = ·09750, \quad \frac{c+d}{N} = ·10050, \quad \frac{d}{N} = ·032000.$$

Entering the table with $\frac{1}{2}(1-\alpha)$ equal to ·098 and ·101 we have

$\frac{1}{2}(1-\alpha)$	τ_1	τ_2	τ_3	τ_4	τ_5	τ_6
·098	+ ·17292	+ ·15811	+ ·04744	− ·06061	− ·06687	+ ·00897
·101	+ ·17678	+ ·15948	+ ·04531	− ·06317	− ·06644	+ ·01153

Multiplying out as before

$$·032000 = ·009898 + ·030569r + ·025215r^2 + ·002150r^3 + ·003829r^4 + ·004443r^5 + ·000103r^6;$$

whence, solving by Newton's method,

$$r = ·498.$$

This particular illustration, chosen so that the true values of $\frac{1}{2}(1-\alpha)$ fall midway between tabulated values, shows the maximum error caused by working without interpolation for values of $\frac{1}{2}(1-\alpha)$ of ·100, and in practically every case this error will be negligible when compared with the probable error of r.

A study of the differences of the functions shows that for values of $\frac{1}{2}(1-\alpha)$ greater than ·100 the error introduced by working without interpolation will not be appreciably greater than in the example given and may quite easily be less; for values of $\frac{1}{2}(1-\alpha)$ less than ·100 it will be desirable to use interpolation if the greatest accuracy attainable is desired, but even in very unfavourable cases such errors will rarely become as large as the probable error of the result.

Activité 20

1. En suivant les notations du tableau IV. de la page 14, démontrer que le coefficient de corrélation linéaire des variables A et B vaut : $r = \dfrac{ad - bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}$.
2. Démontrer que $\chi^2 = Nr^2$.
3. Calculer une valeur approchée de r correspondant aux données de l'exemple de Everitt.

Retour à l'article

Une solution de cette activité

Une solution de l'activité 1

Pour Sheffield, $p_3 = \dfrac{3951}{4151} \simeq 0{,}952$ et $p_4 = \dfrac{278}{552} \simeq 0{,}504$ soit $p_3 - p_4 \simeq 0{,}448$.

Pour Leicester, $p_3 = \dfrac{197}{199} \simeq 0{,}990$ et $p_4 = \dfrac{139}{158} \simeq 0{,}880$ soit $p_3 - p_4 \simeq 0{,}110$.

Enfin, pour Homerton et Fulham, $p_3 = \dfrac{8207}{8899} \simeq 0{,}922$ et $p_4 = \dfrac{1424}{2527} \simeq 0{,}564$ soit $p_3 - p_4 \simeq 0{,}358$.

Comme $0{,}448 > 0{,}358 > 0{,}110$, le classement pour la force de l'association est bien celui annoncé par Yule, à savoir :

1. Sheffield
2. Homerton et Fulham
3. Leicester

Retour à l'article

Retour à l'énoncé de cette activité

Une solution de l'activité 2

Pour Sheffield, $p_1 = \dfrac{3951}{4229} \simeq 0{,}934$ et $p_2 = \dfrac{200}{474} \simeq 0{,}422$ soit $p_1 - p_2 \simeq 0{,}512$.

Pour Leicester, $p_1 = \dfrac{197}{336} \simeq 0{,}586$ et $p_2 = \dfrac{2}{21} \simeq 0{,}095$ soit $p_1 - p_2 \simeq 0{,}491$.

Enfin, pour Homerton et Fulham, $p_1 = \dfrac{8207}{9631} \simeq 0{,}852$ et $p_2 = \dfrac{692}{1795} \simeq 0{,}386$ soit $p_1 - p_2 \simeq 0{,}466$.

On obtient le tableau :

District ou hôpital	Proportion de guérisons parmi		Différence $p_1 - p_2$
	Vaccinés : p_1	Non vaccinés : p_2	
Sheffield	$0{,}934$	$0{,}422$	$0{,}512$
Leicester	$0{,}586$	$0{,}095$	$0,491$
Homerton et Fulham	$0{,}852$	$0{,}386$	$0{,}466$

et le classement suivant pour la force de l'association :

1. Sheffield
2. Leicester
3. Homerton et Fulham

Retour à l'article

Retour à l'énoncé de cette activité

Une solution de l'activité 3

1. Sous l'hypothèse d'indépendance, la proportion de guérisons parmi les individus vaccinés est égale à la proportion de guérisons parmi les individus non vaccinés.
   Nous avons donc $p_3 = p_4$, soit $\dfrac{a_0}{a_0+b_0} = \dfrac{c_0}{c_0+d_0}$.

   Or $\dfrac{a_0}{a_0+b_0} = \dfrac{c_0}{c_0+d_0} = \dfrac{a_0+c_0}{a_0+b_0+c_0+d_0} = \dfrac{a_0+c_0}{N}$, nous obtenons : $\boxed{a_0 = \dfrac{(a_0+b_0)(a_0+c_0)}{N}}$.

   Mais $\delta = a - a_0 = a - \dfrac{(a_0+b_0)(a_0+c_0)}{N} = a - \dfrac{(a+b)(a+c)}{N}$ les sommes marginales étant égales.

   Pour Sheffield, $\delta = 3951 - \dfrac{4229 \times 4151}{4703} \simeq 218{,}366$ et $\dfrac{\delta}{N} \simeq \dfrac{218{,}366}{4703} \simeq 0{,}046$.

   Pour Leicester, $\delta = 197 - \dfrac{199 \times 336}{357} \simeq 9{,}706$ et $\dfrac{\delta}{N} \simeq \dfrac{9{,}706}{357} \simeq 0{,}027$.

   Pour Homerton et Fulham, $\delta = 8207 - \dfrac{8899 \times 9631}{11426} \simeq 706{,}014$ et $\dfrac{\delta}{N} \simeq \dfrac{706{,}014}{11426} \simeq 0{,}062$.

   En prenant $\dfrac{\delta}{N}$ comme indice pour mesurer la force de l'association, on obtient le classement suivant :

   1. Homerton et Fulham
   2. Sheffield
   3. Leicester

2. Le même raisonnement qu'en 1. permet d'obtenir : $\boxed{c_0 = \dfrac{(a+c)(c+d)}{N}}$.

   De même, sous l'hypothèse d'indépendance, la proportion d'individus vaccinés parmi les individus guéris est égale à la proportion d'individus vaccinés parmi les individus décédés.
   Nous avons donc $p_1 = p_2$ soit $\dfrac{a_0}{a_0+c_0} = \dfrac{b_0}{b_0+d_0}$.

   Or $\dfrac{a_0}{a_0+c_0} = \dfrac{b_0}{b_0+d_0} = \dfrac{a_0+b_0}{a_0+c_0+b_0+d_0} = \dfrac{a_0+b_0}{N}$, nous obtenons donc :

   $$\boxed{b_0 = \dfrac{(a_0+b_0)(b_0+d_0)}{N} = \dfrac{(a+b)(b+d)}{N}}.$$

   Enfin, $\boxed{d_0 = \dfrac{b_0 c_0}{a_0} = \dfrac{(b+d)(c+d)}{N}}$.

   $a - a_0 = d - d_0 = \dfrac{ad-bc}{N}$ et $b - b_0 = c - c_0 = \dfrac{bc-ad}{N}$.
   Si nous posions $\delta = d - d_0$, nous obtiendrions la même valeur qu'avec $a - a_0$ mais si nous choisissions $\delta = b - b_0$ ou $\delta = c - c_0$, nous obtiendrions l'opposé de la valeur obtenue avec $a - a_0$.

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

Une solution de l'activité 4

Démonstration. Sous l'hypothèse d'indépendance, la proportion de guérisons parmi les individus vaccinés est égale à la proportion de guérisons parmi les individus non vaccinés.
Nous avons donc $p_3 = p_4$, soit $\dfrac{a}{a+b} = \dfrac{c}{a+c}$.

Comme $\dfrac{a}{a+b} = \dfrac{c}{c+d} = \dfrac{a+c}{a+b+c+d} = \dfrac{a+c}{N}$, nous obtenons :

$$\boxed{a_0 = \frac{(a+b)(a+c)}{N}}.$$

$$\delta = a - a_0 = \frac{aN - (a+b)(a+c)}{N} = \frac{(\cancel{a^2} + \cancel{ab} + \cancel{ac} + ad) - (\cancel{a^2} + \cancel{ac} + \cancel{ab} + bc)}{N} = \frac{ad - bc}{N}.$$

Ainsi,

$$\boxed{\delta = \frac{ad - bc}{N}}.$$

□

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

Une solution de l'activité 5

Démonstration. Nous savons que $\delta = \dfrac{ad-bc}{N}$ (voir l'activité 4).

Nous en déduisons que : $ad - bc = N\delta$.

$$p_1 - p_2 = \frac{a}{a+c} - \frac{b}{b+d} = \frac{a(b+d) - b(a+c)}{(a+c)(b+d)} = \frac{(\cancel{ab} + ad) - (\cancel{ab} + bc)}{(a+c)(b+d)} = \frac{ad-bc}{(a+c)(b+d)}.$$

D'où

$$\boxed{p_1 - p_2 = \frac{N\delta}{(a+c)(b+d)}}.$$

De manière analogue,

$$p_3 - p_4 = \frac{a}{a+b} - \frac{c}{c+d} = \frac{a(c+d) - c(a+b)}{(a+b)(c+d)} = \frac{(\cancel{ac} + ad) - (\cancel{ac} + bc)}{(a+b)(c+d)} = \frac{ad-bc}{(a+b)(c+d)}.$$

D'où

$$\boxed{p_3 - p_4 = \frac{N\delta}{(a+b)(c+d)}}.$$

□

Retour à l'article

Retour à l'énoncé de cette activité

Une solution de l'activité 6

1. Cas des valeurs positives de δ : calcul du maximum de $ad - bc$.

   Multiplions l'inéquation $0 \leqslant a + b + c + d$ par b. Nous obtenons : $0 \leqslant ab + b^2 + bc + bd$.
   Ajoutons ensuite $ad - bc$ à chacun des deux membres.

   $ad - bc \leqslant ab + b^2 + bd + ad$ soit $ad - bc \leqslant (a + b)(b + d)$.

   Enfin, divisons chaque membre de cette inéquation par N.

   Nous obtenons : $\dfrac{ad - bc}{N} \leqslant \dfrac{(a + b)(b + d)}{N}$ soit $\boxed{\delta \leqslant \dfrac{(a + b)(b + d)}{N}}$.

   De manière analogue, en multipliant l'inéquation $0 \leqslant a + b + c + d$ par c, puis en ajoutant à chaque membre $ad - bc$ nous obtenons :
   $ad - bc \leqslant ac + ad + c^2 + cd$ soit $ad - bc \leqslant (a + c)(c + d)$.

   Divisons par N pour obtenir : $\boxed{\delta \leqslant \dfrac{(a + c)(c + d)}{N}}$.

   Si $b < c$ alors $a + b < a + c$ et $b + d < c + d$ d'où $(a + b)(b + d) < (a + c)(c + d)$.

   Conclusion : Si $b < c$ alors $\delta_{max} = \dfrac{(a + c)(c + d)}{N}$ sinon $\delta_{max} = \dfrac{(a + b)(b + d)}{N}$.

2. Cas des valeurs négatives de δ : calcul du maximum de $bd - ac$ (on supposera $a \neq 0$ et $d \neq 0$) .

   $(0 \leqslant a + b + c + d) \Leftrightarrow (0 \leqslant a^2 + ab + ac + ad) \Leftrightarrow (bc - ad \leqslant a^2 + ab + ac + bc)$

   $\Leftrightarrow (bc - ad \leqslant (a + b)(a + c)) \Leftrightarrow \left(\dfrac{bc - ad}{N} \leqslant \dfrac{(a + b)(a + c)}{N}\right)$

   Nous avons donc : $\boxed{|\delta| \leqslant \dfrac{(a + b)(a + c)}{N}}$.

   En multipliant par d, on obtient de manière analoque : $\boxed{|\delta| \leqslant \dfrac{(b + d)(c + d)}{N}}$.

   Conclusion : Si $a < d$ alors $|\delta|_{max} = \dfrac{(b + d)(c + d)}{N}$ sinon $|\delta|_{max} = \dfrac{(a + b)(a + c)}{N}$.

3. Ces valeurs sont en général différentes. Mais si $a = d$ et $b = c$, alors elles sont égales.

   Prenons, par exemple, $a = d = 3$ et $b = c = 1$. On a $N = 8$

   On a dans ce cas $\delta = \dfrac{3 - 1 \times 1}{8} = 1$.

   Cet exemple montre que δ ne pouvait pas être retenu par Yule comme indice de mesure de la force de l'association car il ne satisfaisait pas la troisième propriété essentielle qu'il avait énoncée.

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

Une solution de l'activité 7

Démonstration. Nous avons $\kappa = \dfrac{bc}{ad} = \dfrac{(b_0-\delta)(c_0-\delta)}{(a_0+\delta)(d_0+\delta)}$ (voir l'activité 3) , soit $\kappa = \dfrac{\delta^2-(b_0+c_0)\delta+b_0c_0}{\delta^2+(a_0+b_0)\delta+a_0d_0}$.

Calculons la dérivée de κ par rapport à δ.

$$\frac{\mathrm{d}\kappa}{\mathrm{d}\delta} = \frac{(2\delta-(b_0+c_0))(\delta^2+(a_0+b_0)\delta+a_0d_0)-(2\delta+(a_0+d_0))(\delta^2-(b_0+c_0)\delta+b_0c_0)}{(\delta^2+(a_0+b_0)\delta+a_0d_0)^2} = \frac{f(\delta)}{g(\delta)}$$

Le dénominateur $g(\delta)$ de ce quotient étant positif, il suffit d'étudier le signe du numérateur $f(\delta)$ pour déterminer le signe de la dérivée de κ par rapport à δ.

$$\begin{aligned}
f(\delta) &= (\cancel{2\delta^3} - (b_0+c_0)\delta^2 + 2(a_0+d_0)\delta^2 - \cancel{(a_0+d_0)(b_0+c_0)\delta} + 2a_0d_0\delta - (b_0+c_0)a_0d_0) \\
&\quad - (\cancel{2\delta^3} + (a_0+d_0)\delta^2 - 2(b_0+c_0)\delta^2 - \cancel{(a_0+d_0)(b_0+c_0)\delta} + 2b_0c_0\delta + (a_0+d_0)b_0c_0) \\
&= (a_0+b_0+c_0+d_0)\delta^2 + 2(a_0d_0-b_0c_0)\delta - (b_0+c_0)a_0d_0 - (a_0+d_0)b_0c_0
\end{aligned}$$

Mais dans le cas de l'indépendance, nous avons $p_3 = p_4$, soit $\dfrac{a_0}{a_0+b_0} = \dfrac{c_0}{c_0+d_0}$ ce qui entraîne $a_0d_0 = b_0c_0$. Nous pouvons alors simplifier l'expression de $f(\delta)$ en y supprimant le monôme en δ et en remplaçant b_0c_0 par a_0d_0.

Nous avons alors :

$$\begin{aligned}
f(\delta) &= N\delta^2 - a_0d_0(b_0+c_0+a_0+d_0) \\
&= N(\delta^2 - a_0d_0) \\
&= N\left[\left(\frac{ad-bc}{N}\right)^2 - \left(\frac{(a+b)(a+c)(b+d)(c+d}{N^2}\right)\right] \\
&= \frac{1}{N}\left((a^2d^2 - 2abcd + b^2c^2) - (a^2+ab+ac+bc)(bc+bd+cd+d^2)\right) \\
&= \frac{1}{N}\left((\cancel{a^2d^2} - 2abcd + \cancel{b^2c^2}) - (a^2bc + a^2bd + a^2cd + \cancel{a^2d^2} + ab^2c + ab^2d + abcd + abd^2 + abc^2)\right) \\
&\quad - \frac{1}{N}\left(abcd + ac^2d + acd^2 + \cancel{b^2c^2} + b^2cd + bc^2d + bcd^2\right) \\
&= \frac{-1}{N}\left(4abcd + a^2bc + a^2bd + a^2cd + ab^2c + ab^2d + abd^2 + abc^2 + ac^2d + acd^2 + b^2cd + bc^2d + bcd^2\right)
\end{aligned}$$

Les nombres a, b, c et d, étant des effectifs, sont positifs et par conséquent, $f(\delta)$ et, par la suite, la dérivée de κ par rapport à δ est négative.

□

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

Une solution de l'activité 8

Pour le district de Sheffield :
$Q_{\text{I}} = \dfrac{3951 \times 274 - 200 \times 278}{3951 \times 274 + 200 \times 268} = 0{,}902$

Pour le district de Leicester :
$Q_{\text{II}} = \dfrac{197 \times 19 - 2 \times 139}{197 \times 19 + 2 \times 139} = 0{,}862$

Pour le district de Homerton et Fulham :
$Q_{\text{III}} = \dfrac{8207 \times 1103 - 692 \times 1424}{8207 \times 1103 + 692 \times 1424} = 0{,}804$

Ce qui donne le classement pour la force de l'association :

1. Sheffield
2. Leicester
3. Homerton et Fulham

Retour à l'article

Retour à l'énoncé de cette activité

Une solution de l'activité 9

Le système

$$\begin{cases} xya + xb &= m(yc + d) \\ xya + yc &= n(xb + d) \end{cases}$$

est équivalent au système suivant :

$$\begin{cases} (n+1)abx^2 + (ad(n-m) + bc(1-nm))x - (n+1)mcd &= 0 \\ y &= n\dfrac{xb+d}{xa+c} \end{cases}$$

À partir les données de Sheffield (Tableau I.), nous pouvons calculer :

$m = \dfrac{a_I + b_I}{c_I + d_I} = \dfrac{3\,951 + 200}{278 + 274} \approx 7{,}520$ et $n = \dfrac{a_I + c_I}{b_I + d_I} = \dfrac{3\,951 + 278}{200 + 274} \approx 8{,}922.$

Données de Leicester (Tableau II.) $a_{II} = 197$, $b_{II} = 2$, $c_{II} = 139$ et $d_{II} = 19$.

La résolution du deuxième système donne : $\boxed{x = 8{,}974}$ et $\boxed{y = 0{,}173}$.

En utilisant ces valeurs, on obtient le tableau réduit de Leicester :

Leicester	Guérisons	Décès	Total
Vaccinés	305,63	17,95	323,58
Non vaccinés	24,03	19	43,03
Total	329,66	36,95	366,61

que l'on ramène, par proportionnalité, au tableau suivant d'effectif total $10\,000$.

Leicester	Guérisons	Décès	Total
Vaccinés	8 337	490	8 827
Non vaccinés	655	518	1 173
Total	8 992	1 008	10 000

Les calculs ayant été effectués avec 10 décimales et les effectifs étant arrondis à l'entier le plus proche, certains effectifs diffèrent d'une unité de ceux produits par Yule.

Données de Homerton et Fulham (Tableau III.) : $a_{III} = 8\,207$, $b_{III} = 692$, $c_{III} = 1\,424$ et $d_{III} = 1\,103$.

La résolution du deuxième système donne : $\boxed{x = 1{,}981}$ et $\boxed{y = 1{,}248}$.

En utilisant ces valeurs, on obtient le tableau réduit de Homerton et Fulham à la forme de Sheffield (voir page suivante) :

Homerton et Fulham	Guérisons	Décès	Total
Vaccinés	20 291,54	1 370,58	21 662,12
Non vaccinés	1 777,63	1 103	2 880,63
Total	22 069,17	2 473,58	24 542,75

que l'on ramène, par proportionnalité, au tableau suivant d'effectif total 10 000.

Homerton et Fulham	Guérisons	Décès	Total
Vaccinés	8 268	559	8 827
Non vaccinés	724	449	1 173
Total	8 992	1 008	10 000

Même remarque que précédemment concernant les arrondis entiers.

Comparons maintenant les deux tableaux d'effectifs total 10 000 de Leicester et de Homerton & Fulham réduits à la forme de Sheffield avec celui de Sheffield.

Sheffield	Guérisons	Décès	Total
Vaccinés	8 401	425	8 826
Non vaccinés	591	583	1 174
Total	8 992	1 008	10 000

Nous voyons par simple lecture ($8401 > 8337 > 8268$) que l'association entre guérison et vaccination est plus forte à Sheffield puis à Leicester et enfin à Homerton et Fulham.

Voici le tableau récapitulatif tel qu'il apparaît dans l'article [Yule, 1912] p. 117.

TABLE V.—*The data of Tables I, III and IV;* (a) *reduced to proportions per* 10,000 *observations;* (b) *reduced to proportions of vaccinated and recoveries in Sheffield data per* 10,000 *observations;* (c) *reduced to* 50 *per cent. of vaccinated and* 50 *per cent. of recoveries per* 10,000 *observations.*

Row and column totals		Sheffield.			Leicester.			Homerton and Fulham.		
		Recoveries.	Deaths.	Total.	Recoveries.	Deaths.	Total.	Recoveries.	Deaths.	Total.
(*a*) Of original data	Vaccinated	8,401	425	*8,826*	5,518	56	*5,574*	7,183	606	*7,789*
	Unvaccinated	591	583	*1,174*	3,894	532	*4,426*	1,246	965	*2,211*
	Total	*8,992*	*1,008*	*10,000*	*9,412*	*588*	*10,000*	*8,429*	*1,571*	*10,000*
(*b*) Reduced to Sheffield form	Vaccinated	8,401	425	*8,826*	8,336	490	*8,826*	8,268	559	*8,827*
	Unvaccinated	591	583	*1,174*	655	519	*1,174*	724	450	*1,174*
	Total	*8,992*	*1,008*	*10,000*	*8,991*	*1,009*	*10,000*	*8,992*	*1,009*	*10,001*
(*c*) Reduced to symmetrical form	Vaccinated	4,076	924	*5,000*	3,929	1,071	*5,000*	3,760	1,240	*5,000*
	Unvaccinated	924	4,076	*5,000*	1,071	3,929	*5,000*	1,240	3,760	*5,000*
	Total	*5,000*	*5,000*	*10,000*	*5,000*	*5,000*	*10,000*	*5,000*	*5,000*	*10,000*

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

Une solution de l'activité 10

Démonstration. On supposera qu'aucune des valeurs a, b et c n'est nulle.
Multiplions la première ligne par un facteur p et la première colonne par un facteur q pour obtenir le tableau symétrique souhaité.

Le tableau initial :

a	b
c	d

est transformé en un nouveau tableau :

apq	bp
cq	d

Les facteurs p et q doivent vérifier le système d'équations suivant :

$$\begin{cases} apq &= d \\ bp &= cq \end{cases}$$

En multipliant la première ligne par bc, il vient :

$$\begin{cases} abpcq &= bcd \\ bp &= cq \end{cases}$$

D'où, en divisant la première ligne par a et en remplaçant cq par bp :

$$\begin{cases} (bp)^2 &= \dfrac{bcd}{a} \\ bp &= cq \end{cases}$$

Les coefficients p et q cherchés sont donc :

$$\begin{cases} p &= \sqrt{\dfrac{cd}{ab}} \\ q &= \sqrt{\dfrac{bd}{ac}} \end{cases}$$

Réciproquement, si l'on prend pour p et q les valeurs précédentes, on obtient bien le tableau symétrique :

d	$\sqrt{\dfrac{bcd}{a}}$
$\sqrt{\dfrac{bcd}{a}}$	d

□

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

Une solution de l'activité 11

On a regroupé dans le tableau suivant les valeurs, pour chacun des trois districts, des indices Q et ω :

District	Q	ω
Sheffield	0,902	0,630
Leicester	0,862	0,572
Homerton et Fulham	0,824	0,526

On peut remarquer, sur cet exemple, que les valeurs de Q et de ω sont dans le même ordre.
Cette propriété est toujours vraie car ω est une fonction croissante de Q.

Retour à l'article

Retour à l'énoncé de cette activité

Une solution de l'activité 12

Démonstration. Considérons les deux variables aléatoires X et Y dont la loi de probabilité est donnée par le tableau ci-dessous :

	X		
Y	0	1	Total
0	$\frac{p_0}{2}$	$\frac{q_0}{2}$	$\frac{1}{2}$
1	$\frac{q_0}{2}$	$\frac{p_0}{2}$	$\frac{1}{2}$
Total	$\frac{1}{2}$	$\frac{1}{2}$	1

Le coefficient de corrélation linéaire de deux variables aléatoires X et Y est égal à :

$$Cor(X,Y) = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$$

où $Cov(X,Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$.

Ici, $X = \mathbb{1}_{\text{Être guéri}}$ et $Y = \mathbb{1}_{\text{Être vacciné}}$ sont des variables de Bernoulli de paramètre $\frac{1}{2}$.
Par conséquent, $\mathbb{E}(X) = \mathbb{E}(Y) = \frac{1}{2}$ et $\sigma(X) = \sigma(Y) = \frac{1}{4}$.
$\mathbb{E}(XY) = 0 \times \left(\frac{p_0}{2} + \frac{q_0}{2} + \frac{q_0}{2}\right) + 1 \times \left(\frac{p_0}{2}\right) = \frac{p_0}{2}$.
On a donc $Cov(X,Y) = \frac{p_0}{2} - \frac{1}{4}$.

D'où $Cor(X,Y) = \dfrac{\frac{p_0}{2} - \frac{1}{4}}{\frac{1}{4}} = 2p_0 - 1 = 2p_0 - (p_0 + q_0) = p_0 - q_0 = \omega$

Conclusion Le coefficient de corrélation du tableau symétrique est égal au coefficient ω.

□

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

Une solution de l'activité 13

Si comme nous l'avons vu dans l'activité précédente (voir l'activité 12), le coefficient de colligation ω peut être interprété en termes statistiques comme coefficient de corrélation du tableau symétrique, l'interprétation statistique de ce tableau symétrique est tout autant « artificielle » que peut l'être le coefficient Q.

Retour à l'article

Retour à l'énoncé de cette activité

Une solution de l'activité 14

Démonstration. Dans un premier temps, calculons : $\mathbb{P}_{(T=12)}(D=d)$.

Par définition : $\mathbb{P}_{(T=12)}(D=d) = \dfrac{\mathbb{P}\big((T=12)\cap(D=d)\big)}{\mathbb{P}(T=12)}$

Mais, puisque $T=M+D$, $(T=12)\cap(D=d) = (M=12-d)\cap(D=d)$.
De plus les variables aléatoires D et M sont indépendantes, $D\sim\mathcal{B}(17,p')$ et $M\sim\mathcal{B}(13,p)$.
Nous avons donc :

$$\begin{aligned}\mathbb{P}\big((T=12)\cap(D=d)\big) &= \mathbb{P}(D=d)\times\mathbb{P}(M=12-d)\\ &= \binom{17}{d}(p')^d(q')^{17-d}\binom{13}{12-d}p^{12-d}q^{d+1}\\ &= \binom{17}{d}\binom{13}{12-d}\left(\frac{p'q}{q'p}\right)^d p^{12}(q')^{17}q\end{aligned}$$

Par ailleurs,

$$\begin{aligned}\mathbb{P}(T=12) &= \mathbb{P}\Big(\bigcup_{z=0}^{12}\big((M=12-z)\cap(D=z)\big)\Big)\\ &= \sum_{z=0}^{12}\mathbb{P}(M=12-z)\,\mathbb{P}(D=z)\\ &= \sum_{z=0}^{12}\binom{17}{z}\binom{13}{12-z}\left(\frac{p'q}{q'p}\right)^z p^{12}(q')^{17}q\end{aligned}$$

Enfin, $\mathbb{P}_{(T=12)}(D\leq 2) = \mathbb{P}_{T=12}(D=0)+\mathbb{P}_{T=12}(D=1)+\mathbb{P}_{T=12}(D=2)$

En remplaçant $\dfrac{p'q}{q'p}$ par ψ, le numérateur de $\mathbb{P}_{(T=12)}(D\leq 2)$ peut s'écrire :
$\Big[\binom{17}{0}\binom{13}{12}+\binom{17}{1}\binom{13}{11}\psi+\binom{17}{2}\binom{13}{10}\psi^2\Big]p^{12}(q')^{17}q = (1+102\psi+2\,992\psi^2)13p^{12}(q')^{17}q$

et le dénominateur de $\mathbb{P}_{(T=12)}(D\leq 2)$: $\Big[\sum_{z=0}^{12}\binom{17}{z}\binom{13}{12-z}\psi^z\Big]p^{12}(q')^{17}q = S(\psi)p^{12}(q')^{17}q$.

$$\begin{aligned}S(\psi) &= \binom{17}{0}\binom{13}{12}+\binom{17}{1}\binom{13}{11}\psi+\binom{17}{2}\binom{13}{10}\psi^2+\binom{17}{3}\binom{13}{9}\psi^3+\binom{17}{4}\binom{13}{8}\psi^4\\ &+ \binom{17}{5}\binom{13}{7}\psi^5+\binom{17}{6}\binom{13}{6}\psi^6+\binom{17}{7}\binom{13}{5}\psi^7+\binom{17}{8}\binom{13}{4}\psi^8\\ &+ \binom{17}{9}\binom{13}{3}\psi^9+\binom{17}{10}\binom{13}{2}\psi^{10}+\binom{17}{11}\binom{13}{1}\psi^{11}+\binom{17}{12}\binom{13}{0}\psi^{12}\end{aligned}$$

Ce qui donne :

$$\begin{aligned} S(\psi) &= 13\big[1 + 102\psi + 2\,992\psi^2 + 37\,400\psi^3 + 235\,620\psi^4 + 816\,816\psi^5 + 1\,633\,632\psi^6 \\ &+ 1\,925\,352\psi^7 + 1\,337\,050\psi^8 + 534\,820\psi^9 + 116\,688\psi^{10} + 12\,376\psi^{11} + 476\psi^{12}\big] \end{aligned}$$

Finalement, après simplification des numérateur et dénominateur par $13p^{12}(q')^{17}q$, on obtient :

$$\boxed{\mathbb{P}_{(T=12)}(D \leq 2) = R(\psi) = \frac{A(\psi)}{B(\psi)}}$$

où

$$\boxed{A(\psi) = 1 + 102\psi + 2\,992\psi^2}$$

et

$$\boxed{\begin{aligned} B(\psi) &= 1 + 102\psi + 2\,992\psi^2 + 37\,400\psi^3 + 235\,620\psi^4 + 816\,816\psi^5 + 1\,633\,632\psi^6 \\ &+ 1\,925\,352\psi^7 + 1\,337\,050\psi^8 + 534\,820\psi^9 + 116\,688\psi^{10} + 12\,376\psi^{11} + 476\psi^{12} \end{aligned}}$$

□

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

Une solution de l'activité 15

Démonstration. Rappelons la définition du rapport des cotes de l'exposition au facteur de risque or_E :
$or_E = \dfrac{o_D(E)}{o_{\overline{D}}(E)} = \dfrac{p_1(1-p_2)}{(1-p_1)p_2}$ et notons X la prévalence $\mathbb{P}(D)$.

On peut écrire :
$o_E(D) = \dfrac{r_1}{1-r_1} = \dfrac{\mathbb{P}_E(D)}{\mathbb{P}_E(\overline{D})}$ mais puisque $\mathbb{P}_E(D) = \dfrac{\mathbb{P}(E \cap D)}{\mathbb{P}(E)}$ et $\mathbb{P}_E(\overline{D}) = \dfrac{\mathbb{P}(E \cap \overline{D})}{\mathbb{P}(E)}$

$$o_E(D) = \frac{\dfrac{\mathbb{P}(E \cap D)}{\mathbb{P}(E)}}{\dfrac{\mathbb{P}(E \cap \overline{D})}{\mathbb{P}(E)}} = \frac{\mathbb{P}(E \cap D)}{\mathbb{P}(E \cap \overline{D})}$$

Or, $\mathbb{P}(E \cap D) = \mathbb{P}_D(E)\mathbb{P}(D) = p_1 X$ et $\mathbb{P}(E \cap \overline{D}) = \mathbb{P}_{\overline{D}}(E)\mathbb{P}(\overline{D}) = p_2(1-X)$.

On en conclut que : $\boxed{o_E(D) = \dfrac{p_1 X}{p_2(1-X)}}$.

De même :

$$o_{\overline{E}}(D) = \frac{\dfrac{\mathbb{P}(\overline{E} \cap D)}{\mathbb{P}(\overline{E})}}{\dfrac{\mathbb{P}(\overline{E} \cap \overline{D})}{\mathbb{P}(\overline{E})}} = \frac{\mathbb{P}(\overline{E} \cap D)}{\mathbb{P}(\overline{E} \cap \overline{D})}$$

Or, $\mathbb{P}(\overline{E} \cap D) = \mathbb{P}_D(\overline{E})\mathbb{P}(D) = (1-p_1)X$ et $\mathbb{P}(\overline{E} \cap \overline{D}) = \mathbb{P}_{\overline{D}}(\overline{E})\mathbb{P}(\overline{D}) = (1-p_2)(1-X)$.

On a donc : $\boxed{o_{\overline{E}}(D) = \dfrac{(1-p_1)X}{(1-p_2)(1-X)}}$.

Le rapport des cotes de la maladie s'écrit alors :

$$or_D = \frac{\dfrac{p_1 X}{p_2(1-X)}}{\dfrac{(1-p_1)X}{(1-p_2)(1-X)}} = \frac{p_1(1-p_2)}{(1-p_1)p_2}$$

On reconnait l'expression de or_E d'où $\boxed{or_D = or_E}$. □

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

Une solution de l'activité 16

1. Montrons que si r_1 et r_2 sont « petits », alors $rr \simeq or_D = \psi$.

   Par définition : $or_D = \dfrac{r_1(1-r_2)}{(1-r_1)r_2}$.

   Si r_1 et r_2 tendent vers 0, alors $\dfrac{1-r_2}{1-r_1}$ tend vers 1 et donc or_D tend vers $\dfrac{r_1}{r_2}$ et $\boxed{rr \simeq or_D = \psi}$.

2. Montrons que si $\mathbb{P}(D)$ est « petit » alors $rr \simeq or_E = \psi$.

   Rappelons que : $r_1 = \mathbb{P}_E(D) = \dfrac{\mathbb{P}(E \cap D)}{\mathbb{P}(E)}$ et notons X la prévalence $\mathbb{P}(D)$.

   D'une part $\mathbb{P}(E \cap D) = \mathbb{P}_D(E)\mathbb{P}(D) = p_1 X$ et d'autre part, $\mathbb{P}(E) = \mathbb{P}(E \cap D) + \mathbb{P}(E \cap \overline{D})$.
   D'où $\mathbb{P}(E) = \mathbb{P}(D)\mathbb{P}_D(E) + \mathbb{P}_{\overline{D}}(E)\mathbb{P}(\overline{D}) = p_1 X + p_2(1-X)$.

   Nous pouvons donc écrire : $r_1 = \dfrac{p_1 X}{p_1 X + p_2(1-X)} = g(X) \times \dfrac{p_1 X}{p_2}$ où $g(X) = \dfrac{\dfrac{p_1 X}{p_1 X + p_2(1-X)}}{\dfrac{p_1 X}{p_2}}$.

   Après simplification, nous avons : $g(X) = \dfrac{p_2}{p_1 X + p_2(1-X)}$ avec $\lim\limits_{X \to 0} g(X) = 1$.

   On démontre de même que : $r_2 = \dfrac{(1-p_1)X}{(1-p_1)X + (1-p_2)(1-X)} = h(X) \times \dfrac{(1-p_1)X}{(1-p_2)}$
   où $h(X) = \dfrac{\dfrac{(1-p_1)X}{(1-p_1)X + (1-p_2)(1-X)}}{\dfrac{(1-p_1)X}{(1-p_2)}}$.

   Après simplification, nous avons : $h(X) = \dfrac{(1-p_2)}{(1-p_1)X + (1-p_2)(1-X)}$ avec $\lim\limits_{X \to 0} h(X) = 1$.

   Évaluons alors $\dfrac{r_1}{r_2}$:

$$\frac{r_1}{r_2} = \frac{g(X) \times \dfrac{p_1 X}{p_2}}{h(X) \times \dfrac{(1-p_1)X}{(1-p_2)}} = \frac{g(X) \times \dfrac{p_1}{p_2}}{h(X) \times \dfrac{(1-p_1)}{(1-p_2)}} = \frac{g(X)}{h(X)} \times \frac{p_1(1-p_2)}{p_2(1-p_1)}.$$

   Puisque $g(X)$ tend vers 1 et $h(X)$ tend vers 1 quand X tend vers 0,

$$\frac{r_1}{r_2} \text{ tend vers } \frac{p_1(1-p_2)}{p_2(1-p_1)} = or_E \text{ donc } \boxed{rr \simeq or_E = \psi}.$$

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

Une solution de l'activité 17

Pour le district de Leicester, l'*odds ratio* vaut :

$$OR_{\text{II}} = \frac{197 \times 19}{2 \times 139} \simeq 13{,}46.$$

Pour le district d'Homerton et Fulham, l'*odds ratio* vaut :

$$OR_{\text{III}} = \frac{8\,207 \times 1\,103}{692 \times 1\,424} \simeq 9{,}19.$$

D'où le tableau récapitulatif suivant :

District	OR
Sheffield	19,47
Leicester	13,46
Homerton et Fulham	9,19

Il semble que cette fois-ci, Leicester dépasse Homerton et Fulham quant au risque de décéder de la petite vérole quand on n'est pas vacciné !

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

Une solution de l'activité 18

1. On peut utiliser un solveur en ligne comme : https ://www.dcode.fr/solveur-equation

$$0{,}000103x^6 + 0{,}004443x^5 + 0{,}003829x^4 + 0{,}00215x^3 + 0{,}025215x^2 + 0{,}030569x + 0{,}009898 = 0{,}032$$
$$x \in \mathbb{R}^+$$
$$\Downarrow$$
$$\boxed{x \simeq 0{,}497777}$$

2. On peut aussi utiliser un logiciel comme Scilab.

```
function y=f(x)
    y=0.000103*x^6+0.004443*x^5+0.003829*x^4
    +0.00215*x^3+0.025215*x^2+0.030569*x+0.009898-0.032
endfunction

disp(fsolve(0,f))
```

   Ce programme retourne la valeur : $\boxed{0.4977773}$.

3. En utilisant un tableur : il suffit de mettre « 0 » (ou toute autre valeur) dans la cellule $A1$ puis « $= 0{,}000103.A1^6 + 0{,}004443.A1^5 + 0{,}003829.A1^4 + 0,00215.A1^3 + 0{,}025215.A1^2 + 0{,}030569.A1 + 0,009898$ » dans la cellule $A2$ et d'utiliser le solveur.

   On fixe l'objectif à définir comme étant la cellule $\$A\$2$ à la valeur $0{,}032$ et on indique que la cellule variable est la cellule $\$A\$1$.

   Aussi bien avec l'algorithme évolutionniste *DEPS (Differential Evolution & Particle Swarm Optimization)* du solveur de Libre Office qu'avec l'algorithme *GRG (Generalized Reduced Gradient)* d'Excel, le résultat retourné en $\$A\$1$ est $\boxed{0{,}497777257}$.

Retour à l'article

Retour à l'énoncé de cette activité

Une solution de l'activité 19

Utilisons le tableur Calc de Libre Office.

- h est l'image de $\dfrac{b+d}{N}$ par la fonction inverse de la fonction de répartition d'une variable aléatoire normale $\mathcal{N}(0,1)$, d'où la formule : « D1 = LOI.NORMALE.STANDARD.INVERSE(1 - B1) ».
- La suite $(v_n)_{n\in\mathbb{N}}$ est définie par $v_0 = 1$, $v_1 = h$ et la relation de récurrence $v_n = hv_{n-1} - (n-1)v_{n-2}$, d'où la formule : « E4 = $D$1*E3-(D4-1)*E2 ».
- H est l'image de h par la densité de loi normale $\mathcal{N}(0,1)$, d'où la formule : « F2=LOI.NORMALE.STANDARD(D1 ;0) ».
- $\tau_n = \dfrac{Hv_{n-1}}{\sqrt{n!}}$, d'où la formule : « G3 = ARRONDI($F$2*E2/RACINE(FAC(D3)) ;5) », pour obtenir un arrondi à 5 décimales comme dans les tables d'Everitt.

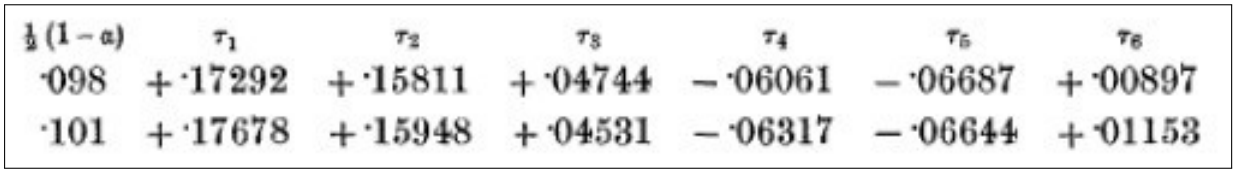
Nous obtenons le tableau suivant :

	A	B	C	D	E	F	G
1	$\frac{b+d}{N}$	0,098	h	=LOI.NORMALE.STANDARD.INVERSE(1-B1)	v_n ou w_n	H ou K	τ_n ou ${\tau_n}'$
2			n	0	1	=LOI.NORMALE.STANDARD(D1;0)	
3				1	=$D$1		=ARRONDI($F$2*E2/RACINE(FACT(D3));5)
4				2	=$D$1*E3-(D4-1)*E2		=ARRONDI($F$2*E3/RACINE(FACT(D4));5)
5				3	=$D$1*E4-(D5-1)*E3		=ARRONDI($F$2*E4/RACINE(FACT(D5));5)
6				4	=$D$1*E5-(D6-1)*E4		=ARRONDI($F$2*E5/RACINE(FACT(D6));5)
7				5	=$D$1*E6-(D7-1)*E5		=ARRONDI($F$2*E6/RACINE(FACT(D7));5)
8				6	=$D$1*E7-(D8-1)*E6		=ARRONDI($F$2*E7/RACINE(FACT(D8));5)
9							
10	$\frac{c+d}{N}$	0,101	h	=LOI.NORMALE.STANDARD.INVERSE(1-B10)	v_n ou w_n	H ou K	τ_n ou ${\tau_n}'$
11			n	0	1	=LOI.NORMALE.STANDARD(D10;0)	
12				1	=$D$10		=ARRONDI($F$11*E11/RACINE(FACT(D12));5)
13				2	=$D$10*E12-(D13-1)*E11		=ARRONDI($F$11*E12/RACINE(FACT(D13));5)
14				3	=$D$10*E13-(D14-1)*E12		=ARRONDI($F$11*E13/RACINE(FACT(D14));5)
15				4	=$D$10*E14-(D15-1)*E13		=ARRONDI($F$11*E14/RACINE(FACT(D15));5)
16				5	=$D$10*E15-(D16-1)*E14		=ARRONDI($F$11*E15/RACINE(FACT(D16));5)
17				6	=$D$10*E16-(D17-1)*E15		=ARRONDI($F$11*E16/RACINE(FACT(D17));5)

Après calcul, ce tableau devient le suivant :

	A	B	C	D	E	F	G
1	$\frac{b+d}{N}$	0,098	h	1,29303197614424	v_n ou w_n	H ou K	τ_n ou ${\tau_n}'$
2			n	0	1	0,172923776668043	
3				1	1,29303197614424		0,17292
4				2	0,671931691331485		0,15811
5				3	-1,71723478961219		0,04744
6				4	-4,23623456751035		-0,06061
7				5	1,39135240421031		-0,06687
8				6	22,9802359862808		0,00897
9							
10	$\frac{c+d}{N}$	0,101	h	1,27587417914913	v_n ou w_n	H ou K	τ_n ou ${\tau_n}'$
11			n	0	1	0,176777041370141	
12				1	1,27587417914913		0,17678
13				2	0,627854921019465		0,15948
14				3	-1,75068447631781		0,04531
15				4	-4,1172178822295		-0,06317
16				5	1,74968591940354		-0,06644
17				6	22,8184684973353		0,01153

où nous pouvons contrôler que les valeurs calculées de τ_n sont les mêmes que celles figurant dans les tables d'Everitt.



$\frac{1}{2}(1-\alpha)$	τ_1	τ_2	τ_3	τ_4	τ_5	τ_6
·098	+ ·17292	+ ·15811	+ ·04744	− ·06061	− ·06687	+ ·00897
·101	+ ·17678	+ ·15948	+ ·04531	− ·06317	− ·06644	+ ·01153

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

Une solution de l'activité 20

1. Formule donnant le coefficient de corrélation linéaire.

   Démonstration. Les variables aléatoires A et B sont définies par le tableau IV. de la page 14 de cet article.

	$B=1$	$B=0$
$A=1$	a	b
$A=0$	c	d

   Le coefficient de corrélation linéaire r des deux variables aléatoires A et B est égal à :

$$r = \text{Cor}(A,B) = \frac{\text{Cov}(A,B)}{\sigma_A \sigma_B}$$

   où $\text{Cov}(A,B) = \mathbb{E}(AB) - \mathbb{E}(A) \times \mathbb{E}(B)$.

   Les variables aléatoires A et B sont des variables de Bernoulli de paramètres respectifs $\frac{a+b}{N}$ et $\frac{a+c}{N}$ où $N = a+b+c+d$ désigne l'effectif total.

   On donc $\mathbb{E}(A) = \frac{a+b}{N}$ et $\mathbb{E}(B) = \frac{a+c}{N}$.

   ${\sigma_A}^2 = \mathbb{E}(A^2) - \mathbb{E}(A)^2 = \frac{a+b}{N} - \left(\frac{a+b}{N}\right)^2 = \frac{(a+b)[N-(a+b)]}{N^2} = \frac{(a+b)(c+d)}{N^2}$.

   D'où $\boxed{\sigma_A = \frac{\sqrt{(a+b)(c+d)}}{N}}$ et de manière analoque, on obtient $\boxed{\sigma_B = \frac{\sqrt{(a+c)(b+d)}}{N}}$.

   $\text{Cov}(A,B) = \frac{a}{N} - \frac{a+b}{N}\frac{a+c}{N} = \frac{Na-(a+b)(a+c)}{N^2} = \frac{(\cancel{a^2}+\cancel{ab}+\cancel{ac}+ad)-(\cancel{a^2}+\cancel{ac}+\cancel{ba}+bc)}{N^2}$.

   Soit, $\boxed{\text{Cov}(A,B) = \frac{ad-bc}{N^2}}$.

   Nous obtenons finalement :

$$\boxed{r = \frac{\frac{ad-bc}{N^2}}{\frac{\sqrt{(a+b)(c+d)}}{N} \times \frac{\sqrt{(a+c)(b+d)}}{N}} = \frac{ad-bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}}$$

   □

2. Formule liant r et χ^2.

Démonstration. La statistique du χ^2 est définie par :

$$\boxed{\chi^2 = \frac{(a-a_0)^2}{a_0} + \frac{(b-b_0)^2}{b_0} + \frac{(c-c_0)^2}{c_0} + \frac{(d-d_0)^2}{d_0}}$$

Notons comme précédemment $N = a+b+c+d$ et posons $\Delta = ad - bc$.
Les valeurs a_0, b_0, c_0 et d_0 ont été déterminées précédemment (voir l'activité 3).

$$a_0 = \frac{(a+b)(a+c)}{N},\ b_0 = \frac{(a+b)(b+d)}{N},\ c_0 = \frac{(c+d)(a+c)}{N} \text{ et } d_0 = \frac{(c+d)(b+d)}{N}.$$

et nous avions obtenu :

$$a - a_0 = d - d_0 = \frac{ad-bc}{N} = \frac{\Delta}{N} \text{ et } b - b_0 = c - c_0 = \frac{bc-ad}{N} = -\frac{\Delta}{N}.$$

Nous pouvons alors calculer χ^2.

$$\begin{aligned}
\chi^2 &= \frac{(a-a_0)^2}{a_0} + \frac{(b-b_0)^2}{b_0} + \frac{(c-c_0)^2}{c_0} + \frac{(d-d_0)^2}{d_0} \\
&= \frac{\Delta^2}{N^2}\left[\frac{1}{a_0} + \frac{1}{b_0} + \frac{1}{c_0} + \frac{1}{d_0}\right] \\
&= \frac{\Delta^2}{N^2}\left[\frac{N}{(a+b)(a+c)} + \frac{N}{(a+b)(b+d)} + \frac{N}{(c+d)(a+c)} + \frac{N}{(c+d)(b+d)}\right] \\
&= \frac{\Delta^2}{N}\left[\frac{(c+d)(b+d) + (c+d)(a+c) + (a+b)(b+d) + (a+b)(a+c)}{(a+b)(c+d)(a+c)(b+d)}\right] \\
&= \frac{\Delta^2}{N}\left[\frac{[(c+d)+(a+b)] \times [(b+d)+(a+b)]}{(a+b)(c+d)(a+c)(b+d)}\right] \\
&= \frac{\Delta^2}{N}\left[\frac{N^2}{(a+b)(c+d)(a+c)(b+d)}\right] \\
&= N \times \frac{\Delta^2}{(a+b)(c+d)(a+c)(b+d)} \\
&= N \times \frac{(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)}
\end{aligned}$$

c'est-à-dire :

$$\boxed{\chi^2 = Nr^2}.$$

□

3. Application numérique : $a = 1\,668$, $b = 131$, $c = 137$ et $d = 64$.

$$r = \frac{88\,805}{\sqrt{127\,273\,808\,025}} \simeq 0{,}000\,000\,7.$$

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