Groupe STATMED
Histoire des probabilités et de la statistique
I.R.E.M. de Caen Normandie

Commission inter-I.R.E.M. Épistémologie et Histoire des Mathématiques

23 mars 2019

Composition du groupe

- Didier Trotoux
- Denis Lanier
- Jean Lejeune
- Rémy Morello
- Jacques Faisant

Thématiques de travail du groupe

- Jusqu'en 2010 :
  Travail sur les textes fondateurs du calcul des probabilités

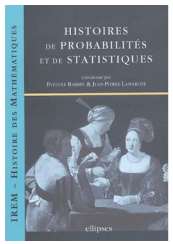
- Depuis 2010, dans le prolongement du XIV^{e} colloque inter IREM d'Epistémologie et Histoire des Mathématiques, *Histoire des probabilités et des statistiques*, Orléans, 2002 :
  Recherches sur les origines des indicateurs en Statistique et leur utilisation dans les sciences du vivant

Stages de formation continue

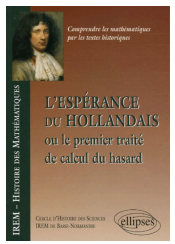
- Stage CAFPEN 1994-1995 : La création du calcul des probabilités et la loi des grands nombres, de Pascal à Poisson.
- Stage CAFPEN 1998-1999 : Histoire des probabilités par les problèmes : *Du calcul dans les Jeux de Hasard* à la *Doctrine des Chances*.
- Stage PAF 2010-2011 : Une histoire des probabilités et des statistiques,
  Session 1 : https ://irem.unicaen.fr/spip.php?article104
  Session 2 : https ://irem.unicaen.fr/spip.php?article113
- Stage PAF 2011-2012 : Aux origines du calcul des probabilités et des statistiques,
  Session 1 : https ://irem.unicaen.fr/spip.php?article128
  Session 2 : https ://irem.unicaen.fr/spip.php?article132

Publications en histoire des probabilités

- *La loi des grands nombres : le théorème de De Moivre-Laplace*, Actes de l'UE de Besançon, 1995,
- *Huygens et ses lecteurs : le 5e exercice*, Histoires de probabilités et de statistiques, CIIÉHM, Ellipses, 2004, pp. 25-53,



- L'espérance du Hollandais ou le premier traité de calcul du hasard, Ellipses, 2005, 176 pages.



- *Le jeu de la baguette de Buffon*, Miroir des maths n°9, pp. 13-24, 2012.

Publications en histoire de la statistique

- *L'invention de la médiane*, https ://irem.unicaen.fr/IMG /pdf/inventionmediane.pdf, 2010,
- *Jules Gavarret, précurseur de la statistique inférentielle ?*, https ://irem.unicaen.fr/spip.php?article117, 2010,
- *La statistique du chi-deux : son usage à partir de l'article de Doll et Hill sur l'association entre cancer et tabac*, https ://irem.unicaen.fr/spip.php?article180, 2015.
- *De l'épidémiologie à la sociologie de l'éducation : comparaisons de proportions et odds-ratios*, article en cours de rédaction, 2019.

L'exemple des inégalités scolaires d'après Pierre Mercklé.

Comparaison des pourcentages d'obtention du baccalauréat pour les enfants de cadres et d'ouvriers en 1960 et 2010.

Pierre Mercklé, *Les inégalités scolaires diminuent-elles* ?, Supplément Science & Techno du journal *Le Monde* du 07.06.2012

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

Mercklé pose la question : les inégalités ont-elles diminué ?
Plus précisément, l'association entre la classe sociale d'origine et l'obtention du baccalauréat a-t-elle augmenté ou diminué ?

Mesures de l'inégalité d'obtention du baccalauréat.

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

Pour l'inégalité de probabilité **d'obtention du baccalauréat** entre les enfants de cadres et les enfants d'ouvriers, on a , tout d'abord, **deux coefficients d'association possibles :**

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

Calcul des valeurs prises par ces deux coefficients.

Différence de probabilité	$dp_{1960} = 0{,}45 - 0{,}05 = 0{,}40$
Quotient de probabilité	$rp_{1960} = \dfrac{0{,}45}{0{,}05} = 9$

Différence de probabilité	$dp_{2010} = 0{,}90 - 0{,}45 = 0{,}45$
Quotient de probabilité	$rp_{2010} = \dfrac{0{,}90}{0{,}45} = 2$

- Coefficient utilisé : **différence de probabilité**,
  $0{,}40 = dp_{1960} < dp_{2010} = 0{,}45 \Longrightarrow$ **augmentation de l'inégalité.**
- Coefficient utilisé : **rapport de probabilité**
  $9 = rp_{1960} > rp_{2010} = 2 \Longrightarrow$ **diminution de l'inégalité.**

Mesures de l'inégalité de non-obtention du baccalauréat.

Données de **non-obtention** :

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

On pose :

- $q_1 = 1 - p_1$ = 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.

Pour l'inégalité de probabilité **de non-obtention du baccalauréat**, on a aussi, a priori, **deux coefficients d'association possibles** :

Mesures de l'inégalité de non-obtention du baccalauréat.

Données de **non-obtention** :

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

On pose :

- $q_1 = 1 - p_1$ = 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.

Pour l'inégalité de probabilité **de non-obtention du baccalauréat**, on a aussi, a priori, **deux coefficients d'association possibles** :

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

Calcul des valeurs prises par les deux coefficients.

Différence de probabilité	$dq_{1960} = 0{,}95 - 0{,}55 = 0{,}40$
Quotient de probabilité	$rq_{1960} = \dfrac{0{,}95}{0{,}55} \approx 1{,}72$

Différence de probabilité	$dq_{2010} = 0{,}55 - 0{,}10 = 0{,}45$
Quotient de probabilité	$rq_{2010} = \dfrac{0{,}55}{0{,}10} = 5{,}5$

- Coefficient utilisé : **différence de probabilité**
  $0{,}40 = dq_{1960} < dq_{2010} = 0{,}45 \Longrightarrow$ **augmentation de l'inégalité.**

- Coefficient utilisé : **rapport de probabilité**,
  $1{,}72 = rq_{1960} < rq_{2010} = 5{,}5 \Longrightarrow$ **augmentation de l'inégalité.**

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

Les résultats obtenus sont résumés ci-dessous :

	Obtention	Non-obtention
Différence	Augmentation	Augmentation
Rapport	Diminution	Augmentation

Deux contradictions :

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

Les résultats obtenus sont résumés ci-dessous :

	Obtention	Non-obtention
Différence	Augmentation	Augmentation
Rapport	Diminution	Augmentation

Deux contradictions :

- si on se place du point de vue de l'obtention du baccalauréat, le coefficient 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.

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

Les résultats obtenus sont résumés ci-dessous :

	Obtention	Non-obtention
Différence	Augmentation	Augmentation
Rapport	Diminution	Augmentation

Deux contradictions :

- si on se place du point de vue de l'obtention du baccalauréat, le coefficient 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.
- le coefficient 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.

Recherche d'un coefficient d'association d'après Yule (1912)

Biographie de George Udny Yule (1871-1951)

- 1871 : Naissance en Écosse dans une famille intellectuelle.
- 1892 : Diplômé ingénieur *University College* Londres.
- 1893-1899 : Chargé de TP au *University College* recruté par Karl Pearson.

Biographie de George Udny Yule (1871-1951)

- 1871 : Naissance en Écosse dans une famille intellectuelle.
- 1892 : Diplômé ingénieur *University College* Londres.
- 1893-1899 : Chargé de TP au *University College* recruté par Karl Pearson.
- 1900-1909 : Publication d'articles sur l'association et la corrélation, conférences : *Newmarch Lectures in Statistics.*
- 1911 : Publication de *Introduction to the Theory of Statistics.*
- 1912 : Maître de conférence à l'Université de Cambridge.
- 1920-1930 : Publication d'articles sur les séries chronologiques. Élu à la *Royal Society* en 1921. Président de la *Royal Statistical Society* de 1924 à 1926.
- 1951 : Décès à Cambridge (crise cardiaque).

L'article de 1912

Yule présente le 23 avril à la *Royal Statistical Society* de Londres une communication intitulée :

"*Sur les méthodes de mesure de l'association entre deux attributs*".

"*On the methods of measuring association between two attributes*".

L'article de 1912

Yule présente le 23 avril à la *Royal Statistical Society* de Londres une communication intitulée :
"*Sur les méthodes de mesure de l'association entre deux attributs*".
"*On the methods of measuring association between two attributes*".

- Une population donnée est partagée en quatre classes suivant deux divisions successives.
  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**.
- L'objet est la recherche d'une évaluation de l'association entre deux attributs, ici, **guérison** et **vaccination**.

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

Existe-t-il une association entre les attributs **vaccination** et **guérison** ?

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

Existe-t-il une association entre les attributs **vaccination** et **guérison** ?

Proportion des guérisons parmi les vaccinés : $\frac{3951}{4151} \approx 0{,}952$.
Proportion des guérisons parmi les non-vaccinés : $\frac{278}{552} \approx 0{,}504$.

Conclusion :
Association positive très marquée entre vaccination et guérison.

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

Existe-t-il une association entre les attributs **vaccination** et **guérison** ?

Proportion des guérisons parmi les vaccinés : $\frac{3951}{4151} \approx 0{,}952$.
Proportion des guérisons parmi les non-vaccinés : $\frac{278}{552} \approx 0{,}504$.

Conclusion :
Association positive très marquée entre vaccination et guérison.

NB : même conclusion en calculant la proportion des vaccinés parmi les guérisons (0,934) et celle des vaccinés parmi les décès (0,422).

Évaluation de la force de l'association

L'objectif de Yule est d'évaluer la force de cette association, en particulier pour comparer les résultats du tableau précédent avec ceux d'une autre région ou à une autre période.

Cas général :

	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 les quatre proportions :

Évaluation de la force de l'association

L'objectif de Yule est d'évaluer la force de cette association, en particulier pour comparer les résultats du tableau précédent avec ceux d'une autre région ou à une autre période.

Cas général :

	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 les quatre proportions :

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

Présentation par Yule de deux autres sources de données

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

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

Forces respectives de l'association guérison-vaccination des trois tableaux

- **Méthode 1** : Calcul des proportions p_3 et p_4 et de $p_3 - p_4$

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

Classement selon la force de l'association :

1. Sheffield ; 2. Homerton-Fulham ; 3. Leicester.

Forces respectives de l'association guérison-vaccination des trois tableaux

- **Méthode 1** : Calcul des proportions p_3 et p_4 et de $p_3 - p_4$

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

Classement selon la force de l'association :

1. Sheffield ; 2. Homerton-Fulham ; 3. Leicester.

- **Méthode 2** : Calcul des proportions p_1 et p_2 et de $p_1 - p_2$

Classement selon la force de l'association :

1. Sheffield ; 2. Leicester ; 3. Homerton-Fulham.

- **Méthode 3** : Calcul de $\delta = a - a_0$ où a_0 est la valeur qu'on aurait dans le cas de l'indépendance des attributs.

$$\delta = a - a_0 = a - \frac{(a+b)(a+c)}{N} = \frac{ad-bc}{N}.$$

En fait, Yule calcule le coefficient $\frac{\delta}{N}$.

- **Méthode 3** : Calcul de $\delta = a - a_0$ où a_0 est la valeur qu'on aurait dans le cas de l'indépendance des attributs.

$$\delta = a - a_0 = a - \frac{(a+b)(a+c)}{N} = \frac{ad-bc}{N}.$$

En fait, Yule calcule le coefficient $\frac{\delta}{N}$.

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

Classement selon la force de l'association :

1. Homerton-Fulham ; 2. Sheffield ; 3. Leicester.

Yule remarque : « *Les trois coefficients différents essayés ont placé les districts dans trois ordres différents !* »

Détermination d'un coefficient plus sophistiqué pour mesurer la force de l'association.

Depuis 1900 et son premier article sur le même sujet, Yule a en tête certaines propriétés qu'il juge essentielles à satisfaire pour ce coefficient. Voici les deux premières :

- il doit être égal à 0 si et seulement si les deux attributs sont indépendants ;
- il doit être compris entre -1 et $+1$.

δ ne convient pas car il ne remplit pas la 2ème propriété précédente. Yule propose alors de réétudier le coefficient nommé Q, en hommage à Quételet, qu'il avait introduit en 1900 :

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

Ce coefficient Q, formule empirique, satisfait les propriétés voulues.
("*On the association of attributes in statistics...*", 1900.)

Conclusions

Yule montre que Q est une fonction 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.

- Pour faciliter cette démonstration, il note que l'on peut écrire $Q = \dfrac{1-\kappa}{1+\kappa}$, en posant $\kappa = \dfrac{bc}{ad}$.

Conclusions

Yule montre que Q est une fonction 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.

- Pour faciliter cette démonstration, il note que l'on peut écrire $Q = \dfrac{1-\kappa}{1+\kappa}$, en posant $\kappa = \dfrac{bc}{ad}$.

Yule remarque enfin que Q peut être exprimé en fonction des seuls p_1 et p_2 , ou des seuls p_3 et p_4 :

$$Q = \frac{p_1(1-p_2) - p_2(1-p_1)}{p_1(1-p_2) + p_2(1-p_1)} = \frac{p_3(1-p_4) - p_4(1-p_3)}{p_3(1-p_4) + p_3(1-p_4)}.$$

Conclusions

Yule montre que Q est une fonction 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.

- Pour faciliter cette démonstration, il note que l'on peut écrire $Q = \dfrac{1-\kappa}{1+\kappa}$, en posant $\kappa = \dfrac{bc}{ad}$.

Yule remarque enfin que Q peut être exprimé en fonction des seuls p_1 et p_2 , ou des seuls p_3 et p_4 :

$$Q = \frac{p_1(1-p_2) - p_2(1-p_1)}{p_1(1-p_2) + p_2(1-p_1)} = \frac{p_3(1-p_4) - p_4(1-p_3)}{p_3(1-p_4) + p_3(1-p_4)}.$$

- Les coefficients Q et κ ne sont donc pas affectés par le choix de l'ordre utilisé pour les deux attributs et 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. ◂ Voir

Propriétés du coefficient ω de Yule

L'interprétation de Q en termes statistiques étant difficile, Yule propose de prendre comme nouveau coefficient :

$$\omega = \frac{1-\sqrt{\kappa}}{1+\sqrt{\kappa}} = \frac{\sqrt{ad}-\sqrt{bc}}{\sqrt{ad}+\sqrt{bc}}$$

ceci après avoir démontré que ω est, en fait, un coefficient de corrélation linéaire.
Il propose de le nommer "*coefficient de colligation*".

On a alors :

$$Q = \frac{1-\kappa}{1+\kappa} = \frac{2\omega}{1+\omega^2} \quad \text{et} \quad \omega = \frac{1-\sqrt{1-Q^2}}{Q}.$$

- Calcul de Q et ω

$$Q = \frac{ad - bc}{ad + bc} \qquad \omega = \frac{\sqrt{ad} - \sqrt{bc}}{\sqrt{ad} + \sqrt{bc}}$$

District ou hôpital	Valeur de Q	Valeur de ω
Sheffield	0,902	0,630
Leceister	0,862	0,572
Homerton et Fulham	0,804	0,504

Classement obtenu :

1. Sheffield ; 2. Leicester ; 3. Homerton-Fulham.

Ce n'est pas l'ordre obtenu avec $\frac{\delta}{N}$ ◂ Voir

Retour aux inégalités scolaires

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

$$\boldsymbol{Q = 0{,}88} \text{ et } \boldsymbol{\omega = 0{,}60}$$

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

$$\boldsymbol{Q = 0{,}83} \text{ et } \boldsymbol{\omega = 0{,}54}$$

La diminution des valeurs des coefficients utilisés ici, Q et ω, indique une diminution des inégalités.

Une conclusion

L'utilisation des coefficients d'association proposés par Yule, Q et ω, permet donc de ne pas subir les incohérences découlant de l'emploi,

- soit de la différence de probabilité,
- soit du quotient de probabilité.

Une conclusion

L'utilisation des coefficients d'association proposés par Yule, Q et ω, permet donc de ne pas subir les incohérences découlant de l'emploi,

- soit de la différence de probabilité,
- soit du quotient de probabilité.

Mais c'est, finalement, le rapport des cotes, ou *odds ratio*

$$\frac{\frac{a}{b}}{\frac{c}{d}} = \frac{\frac{a}{c}}{\frac{b}{d}} = \frac{ad}{bc} = \frac{1}{\kappa}$$

qu'on emploie fréquemment aujourd'hui pour mesurer la force de l'association entre deux caractères qualitatifs.

Une conclusion

L'utilisation des coefficients d'association proposés par Yule, Q et ω, permet donc de ne pas subir les incohérences découlant de l'emploi,

- soit de la différence de probabilité,
- soit du quotient de probabilité.

Mais c'est, finalement, le rapport des cotes, ou *odds ratio*

$$\frac{\frac{a}{b}}{\frac{c}{d}} = \frac{\frac{a}{c}}{\frac{b}{d}} = \frac{ad}{bc} = \frac{1}{\kappa}$$

qu'on emploie fréquemment aujourd'hui pour mesurer la force de l'association entre deux caractères qualitatifs.
Ce rapport des cotes est souvent dénommé *rapport des chances relatives* par les sociologues (par exemple par Claude Thélot).