Chapitre 5

Statistique d'ordre des variables qualitatives

Devant l'importance de cette statistique en géographie, il me semble intéressant de développer quelques idées pour éviter tout quiproquo sur la statistique d'ordre.

La statistique d'ordre est le cœur de la géographie humaine. De manière annuelle, mensuelle, voire hebdomadaire, un certain nombre de classements est opéré en utilisant des objets géographiques. Leur objectif commun est de montrer quelle entité a descendu, stagné ou monté dans le classement. Ainsi, un lien entre ordination et variable quantitative s'effectue spontanément dans toute étude géographique. Par exemple, les manuels du secondaire de géographie sont truffés de ce genre de tableaux qui, malgré leur apparence basique, peuvent subir des traitements statistiques relativement complexes.

5.1 Les règles générales des statistiques d'ordre

L'ordre qui doit être privilégié est l'**ordre croissant** (ou l'ordre naturel). Il existe des exceptions en géographie telles que la loi dite rang-taille. L'ordination permet de rechercher les valeurs aberrantes, trop grandes ou trop petites, d'une série d'observations. De plus, elle offre la possibilité d'étudier la loi de la plus grande valeur d'une série d'observations.

Soit une suite finie d'observations indépendantes (X_i), avec $i \in [1, n]$, classées par ordre croissant. On désigne par :

— $X_{(1)}$ la plus petite valeur observée, c'est-à-dire la plus petite des valeurs X_i ;

— $X_{(k)}$ la valeur de rang k, et ainsi de suite jusqu'à la plus grande valeurs (X_n).

Approximations de...	Fonction de répartition empirique
Haazen (1930)	$F(X) = \frac{i-0,5}{n}$
Weibull (1939)	$F(X) = \frac{i}{n+1}$
Chegodaev (1955)	$F(X) = \frac{i-0,3}{n+0,4}$
Tukey (1962)	$F(X) = \frac{i-\frac{1}{3}}{n-\frac{1}{3}}$

TABLE 5.1 – Les approximations de la fonction de répartition F d'un rang

On écrit cette suite d'observations sous la forme :

$$X_{(1)} \leq \ldots \leq X_{(n)} \tag{5.1}$$

La suite ordonnée des $X_{(i)}$ est appelée **statistique d'ordre associée à la série des observations** (X_i).

Remarque 1. On aurait dû écrire $X_{(i,n)}$, car le rang d'une observation dépend du nombre n des observations.

Remarque 2. Si la loi X est une loi continue, on peut se limiter à des inégalités strictes :

$$X_{(1)} < \ldots < X_{(n)} \tag{5.2}$$

car l'événement $X = k$ est un événement de probabilité nulle $\Pr(X = k) = 0$.

La quantité $X_{(n)} - X_{(1)}$ est l'étendue de l'échantillon.

Toute statistique d'ordre possède une fonction de répartition $F(x_i)$ dont les valeurs varient de 0 à 1 dont il faut calculer son approximation (Tab. 5.1).

Andrej Dmitrievich Chegodaev (1905-1994)

L'approximation de A. D. Chegodaev reste la meilleure formule d'approximation. L'erreur maximale demeure inférieure à 1 % quelle que soit la taille de l'échantillon n. La fonction de répartition étant la dérivée de la distribution, il est par la suite facile de déduire la distribution statistique en présence dans le cas étudié.

Soit $R_n(x)$ le nombre de répétitions de l'événement $(X < x)$ au cours de n épreuves indépendantes. Par définition, la fonction de répartition est :

$$F(x) = \Pr(X < x) \tag{5.3}$$

Pour x fixé, la probabilité est constante au cours des n épreuves. La variable $R_n(x)$ suit alors la loi binomiale $\beta[n, F(x)]$, d'où :

$$\Pr(R_n(x) = h) = C_n^h [F(x)]^h [1 - F(x)]^{n-h} \tag{5.4}$$

La réalisation de l'événement $X_{(k)} < a$ implique que :

1. k valeurs de la variable X, au moins, soient inférieurs à x ;
2. on peut en avoir $k+1, k+2, \ldots, n$.

On en déduit la fonction de répartition $H_{(k)}(x)$ de la variable aléatoire $X_{(k)}$:

$$H_{(k)}(x) = \Pr\left(X_{(k)} < x\right) = \sum_{h=k}^{n} C_n^h \left[F(x)\right]^h \left[1 - F(x)\right]^{n-h} \tag{5.5}$$

La densité de $X_{(k)}$ peut être obtenue à partir de la définition :

$$h_{(k)}\mathrm{d}x = \Pr\left(x \leq X_{(k)} < x + \mathrm{d}x\right) \tag{5.6}$$

La réalisation d'un événement implique que :

1. au moins une des valeurs x_i appartienne à l'intervalle $[x, x + \mathrm{d}x]$; la probabilité de réalisation de cet événement est $nf(x)\,\mathrm{d}x$, car il existe n choix possibles pour la valeur x_i ;
2. $(k-1)$ valeurs des x_i soient inférieures à x_i ; la probabilité de réalisation de cet événement est $\left[F(x)\right]^{k-1}$;
3. $(n-k)$ des valeurs x_i soient supérieures à x ; la probabilité de réalisation de cet événement est $\left[1 - F(x)\right]^{n-k}$; k nombre de réalisations possibles de cet événement est $C_{n-1}^{n-k} = C_{n-1}^{k-1}$.

La **densité de probabilité** de la variable aléatoire $X(k)$ est de fait égale à :

$$h_{(k)}(x) = nC_{n-1}^{k-1} \left[F(x)\right]^{k-1} \left[1 - F(x)\right]^{n-k} f(x) \tag{5.7}$$

Remarque 1. La fonction de répartition $H_{(k)}(x)$ ne dépend que $F(x)$, fonction de répartition de la variable X, et non de la nature de cette variable.

Remarque 2. Si X est une variable continue, la densité de la loi de probabilité de la variable $X_{(k)}$ peut être obtenue en dérivant la fonction de répartition $H_{(k)}(x)$.

Remarque 3. Si X est une variable discrète, la densité de la loi de probabilité de $H_{(k)}(x)$ est égale à :

$$H_{(k)}(x) = H_{(k)}(x+1) - H_{(k)}(x) \tag{5.8}$$

Remarque 4. Il existe une relation mathématique simple entre $H_{(k)}(x)$ et la fonction bêta incomplète :

$$H_{(k)}(x) = I_{F(x)}(k, n-k+1) \tag{5.9}$$

où la fonction bêta incomplète est définie par :

$$I_u(p,q) = \int_0^u t^{p-1}(1-t)^{q-1}\,\mathrm{d}t \tag{5.10}$$

La fonction de répartition $H_{(k)}(x)$ est égale à l'intégrale bêta incomplète, tronquée en $F(x)$.

Objet	Objet A	Objet B	...	Objet Z
Classement n°1	u_1	u_2	...	u_n
Classement n°2	v_1	v_2	...	v_n

TABLE 5.2 – Classements de n objets réalisés par deux individus

Les lois d'ordre sont aussi fréquentes en géographie humaine qu'en géographie physique. En géographie physique, elles servent notamment à étudier la hauteur maximale des crues d'un cours d'eau, l'intensité du plus fort tremblement de terre dans une zone sismique donnée, *etc*. En géographie humaine, leur utilisation découle du fait de l'apparition plus ou moins spontanée de hiérarchies au sein des sociétés et des espaces étudiés.

5.2 La corrélation des rangs

Tout classement dispose d'une part d'arbitraire due à celui qui l'élabore, et ce même en constituant des nomenclatures très détaillées. Il faut par conséquent être en mesure de proposer des analyses statistiques des différents classements possibles. Le problème se formule de manière relativement simple : « Les classements opérés sont-ils identiques ? ». Pour répondre à cette question, il existe deux tests : celui de C. Spearman et celui de M. G. Kendall.

Charles Spearman (1863-1945)

Maurice G. Kendall (1907-1983)

Soient X et Y deux variables ordinales prises dans un ensemble de n individus (ou objets) qui ont été soumis à deux classements différents. Ainsi, on obtient deux classements, dont les rangs sont représentés par deux variables aléatoires U et V.

Le problème posé consiste à comparer les deux classements, c'est-à-dire de répondre à la question suivante : « ces classements sont-ils identiques ou non ? ». Les tests de C. Spearman et de M. G. Kendall ont été conçus pour y répondre ce problème.

Le test de Spearman (1904)

Le psychologue C. Spearman proposa en 1904 un test sur les rangs *via* le coefficient de corrélation usuel rebaptisé Spearman r_s [Spearman, 1904].

$$r_s = \frac{\text{cov}\,(u, v)}{s_u s_v} \tag{5.11}$$

u et v sont des nombres entiers et s_u et s_v sont les écarts types empiriques. Chacun des classements n'est qu'une permutation de ces nombres. Par conséquent, moyenne et variance de U et V sont identiques. Par ailleurs, il est possible

d'identifier une distribution particulière applicable à ce cas d'espèce, la **loi uniforme discrète** qui est appliquée ici à un ensemble de nombres entiers allant de 1 à n. La moyenne et la variance de cette loi valent respectivement :

$$\langle U \rangle = \frac{n+1}{2} = \langle V \rangle \tag{5.12}$$

et

$$\mathbb{V}(U) = \frac{n^2-1}{12} = \mathbb{V}(V) \tag{5.13}$$

Ce cas montre l'importance de bien connaître les distributions statistiques afin de ne pas se tromper dans le calcul des moyennes.

En utilisant ces formules dans l'expression de r_s, on obtient la formule plus pratique :

$$r_s = 1 - \frac{6}{n\left(n^2-1\right)} \sum_{i=1}^{n} \left(u_i - v_i\right)^2 \tag{5.14}$$

Si le coefficient de corrélation de C. Spearman r_s vaut 1, c'est-à-dire que la différence entre u_i et v_i est nulle, les classements sont identiques. Si le coefficient de corrélation de C. Spearman r_s vaut -1, cela signifie que les classements sont strictement l'inverse l'un par rapport à l'autre. Si le coefficient de corrélation de C. Spearman r_s est nul, les classements sont indépendants.

Ainsi, le test de C. Spearman est un test non paramétrique qui permet d'établir deux hypothèses. D'une part, le coefficient de corrélation des rangs n'est pas significativement différent de zéro (H_0). D'autre part, le coefficient de corrélation des rangs est significativement différent de zéro (H_0). L'unique condition est que l'on considère que les rangs ont des permutations équiprobables.

Pour le classement de C. Spearman, il ne doit pas y avoir de rang *ex aequo*. S'il y en a, alors on a besoin d'une correction ξ. Dans ce cas,

$$r_s = \frac{\sum\left(r_1 \times r_2\right) - \xi}{\sqrt{\left(\sum {r_1}^2 - \xi\right)\left(\sum {r_2}^2 - \xi\right)}} \tag{5.15}$$

r_1 le rang 1, r_2 le rang 2, et ξ la correction valant :

$$\xi = \frac{1}{4} n \left(n+1\right)^2 \tag{5.16}$$

avec n le nombre de rangs.

Pour $n > 30$, la distribution du coefficient de corrélation de C. Spearman r_s peut être rapprochée d'une distribution normale qui possède une moyenne $\mathbb{E}(r_s) = 0$ et une variance $\mathbb{V}(r_s) = \frac{1}{n-1}$. Il est par conséquent possible de construire un intervalle de confiance avec une probabilité critique.

Objet	**D**	**C**	**...**	**B**
Classement n°1 Série (x_i)	1	2	...	n
Classement n°2 Série (y_i)	y_{i1}	y_{i2}	...	y_{in}

TABLE 5.3 – Classements des objets ordonnés selon l'ordre naturel du classement n° 1

En géographie urbaine, cela peut s'appliquer aux classements des villes par rapport à leur population. Les géographes sont en désaccord sur la définition des limites des villes, donc de la population contenue à l'intérieur. Ainsi, les classements peuvent être très différents, alors que le nom des agglomérations est identique. La méthode de C. Spearman peut permettre de rapprocher et de comparer les différents classements urbains.

Par ailleurs, le test de C. Spearman permet également de comparer plusieurs variables rattachées à un objet d'étude. Par exemple, toujours en géographie urbaine, les villes ont souvent des indicateurs de fécondité, d'activités, *etc*. Il est possible de faire un classement sur un échantillon de plusieurs villes de la fécondité et de l'activité, par exemple. Dans ce cas, le coefficient de C. Spearman donne un lien de dépendance entre le classement des villes en fonction de la fécondité et le classement des villes en fonction des activités.

Pour effectuer un test, on distingue deux cas.

Auguste Bravais (1811-1863)

— Si $n \leq 30$, la table r de Bravais-Pearson est le plus souvent utilisée.

— Si $n > 30$, r_s peut se calculer avec un t de Student.

$$t = \left(\sqrt{n-2}\right) \frac{r_s}{\sqrt{1 - {r_s}^2}} \tag{5.17}$$

t est lu dans la table de Student avec un degré de liberté $ddl = n - 2$.

Le test de Kendall

M. G. Kendall inventa son test d'indépendance des rangs en 1938 [Kendall, 1938] [Rateau, 2001].

Soient deux classements correspondant respectivement à deux séries de valeurs x_i et de valeurs y_i correspondant à des valeurs allant de 1 à n. Ainsi, il est possible de former des couples de rang (x_i, y_i). Pour simplifier l'estimation du coefficient de M. G. Kendall, il est conseillé de classer par ordre naturel l'un des classements (Tab. 5.3).

À présent, il faut évaluer un score S_c grâce au tableau 5.3. On regarde tous les couples y_{i1} par rapport aux autres valeurs qui suivent $y_{i2}, \ldots, y_{in}$. Si l'ordre

naturel est respecté, on note $+1$, sinon -1. Dans le premier cas, les classements sont **concordants** ; dans le second, **discordant**. On effectue la même manipulation avec y_{i2}, et ainsi de suite jusque la valeur $y_{in(n-1)}$. Le score S_c correspond à la somme des valeurs $+1$ et -1 observée. En sachant que la somme totale S_T, c'est-à-dire le cas de concordance parfaite, vaut :

$$S_T = \frac{n\,(n-1)}{2} \tag{5.18}$$

Le coefficient de M. G. Kendall τ correspond au rapport entre S_c et S_T. Après quelques calculs, il s'estime plus simplement de la manière suivante :

$$\tau = \frac{2S_c}{n\,(n-1)} \tag{5.19}$$

Si le coefficient de M. G. Kendall τ vaut $+1$, les classements sont identiques. Si le coefficient de M. G. Kendall τ vaut -1, les classements sont inverses. Si le coefficient de M. G. Kendall τ vaut $+1$, les classements sont identiques.

Pour $n \geq 8$, la distribution du coefficient de M. G. Kendall est une distribution normale qui possède une moyenne $\mathbb{E}\,(\tau) = 0$ et une variance $\mathbb{V}\,(\tau) = \sqrt{\frac{2(2n+5)}{9n(n-1)}}$. Il est par conséquent possible de construire un intervalle de confiance avec une probabilité critique.

L'avantage du coefficient de M. G. Kendall τ réside dans sa capacité à se généraliser à k classements.

Exemple. Soient quatre objets A, B, C et D qui peuvent être classés de deux manières (Tab. 5.4).

Objet	A	B	C	D
Classement 1	3	4	2	1
Classement 2	3	1	4	2

TABLE 5.4 – Rangs obtenus pour chaque objet

1. Réarranger par ordre naturel du classement 1 (Tab. 5.5)

Objet	D	C	A	B
Classement 1	1	2	3	4
Classement 2	2	4	3	1

TABLE 5.5 – Rangs obtenus pour chaque objet par rapport au classement 1

2. Étudier les cas possibles. Au rang n° 1, la valeur 2 du classement 2 se combine avec les trois valeurs suivantes : $(2,4)\,,(2,3)\,,(2,1)$. Les deux premiers ont un ordre concordant, tandis que le dernier a un ordre discordant. Au rang n° 2, la valeur 4 du classement 2 se combine avec les deux valeurs suivantes : $(4,3)\,,(4,1)$. Il y a deux valeurs discordantes. Au rang n° 3, la valeur 3 du classement 2 se combine avec la valeur suivante : $(3,1)$, soit une valeur discordante. Cela se traduit d'après le test en :

$$+1;+1;-1;-1;-1;-1 \tag{5.20}$$

3. Calculer $S_C = 1+1-1-1-1-1 = -2$
   — On peut également poser que le nombre de concordances $N_a = 2$ et le nombre de discordances $N_d = 4$, ce qui fait que : $S_C = N_a - N_d = 2-4 = -2$.
4. Calculer $S_T = \frac{4\times 3}{2} = 6$
   — En utilisant le nombre de concordances N_a et le nombre de discordances N_d, on peut calculer : $S_T = N_a + N_d = 2+4 = 6$.
5. Calculer $\tau = \frac{S_C}{S_T} = \frac{-2}{6} = -\frac{1}{3}$

   Remarque. $\tau = \frac{S_C}{S_T} = \frac{S_C}{\frac{n(n-1)}{2}} = \frac{2S_C}{n(n-1)}$
6. On utilise la loi normale si $n \geq 8$, ce qui n'est pas le cas ici.

$$\begin{aligned}&\mathbb{E}(\tau) = 0\\ &\mathbb{V}(\tau) = \sqrt{\tfrac{2(2n+5)}{9n(n-1)}} = \sqrt{\tfrac{2(2\times 4+5)}{9\times 4(4-1)}} = \sqrt{\tfrac{26}{108}} \approx 0,4907\\ &\mathrm{SE} = \tfrac{0,4907}{\sqrt{4}} \approx \tfrac{0,4907}{2} \approx 0,2453\end{aligned} \tag{5.21}$$

   Pour construire un intervalle de confiance, on pourrait utiliser :
   — pour un risque $\alpha = 0,05$, $t_{0,95} = 3,182$
   — pour un risque $\alpha = 0,01$, $t_{0,99} = 51,075$

Les coefficients de corrélation des rangs sont très utiles pour tester l'indépendance des deux variables non normales, car le test du coefficient de corrélation linéaire ne s'applique pas dans ce cas. De plus, ils sont invariants par toute transformation monotone croissante des variables.

Objet	Classement 1	Classement 2	Concordant	Discordant	TOTAL
A	1	1	11	0	**11**
B	2	2	10	0	**10**
C	3	4	8	1	**9**
D	4	3	8	0	**8**
E	5	6	6	1	**7**
F	6	5	6	0	**6**
G	7	8	4	1	**5**
H	8	7	4	0	**4**
I	9	10	2	1	**3**
J	10	9	2	0	**2**
K	11	12	0	1	**1**
L	12	11	-	-	**-**
TOTAL			**61**	**5**	**66**

TABLE 5.6 – Concordances et discordances entre le classement 2 et le classement 1

Exercice type On souhaite comparer deux classements de 12 objets (A, B, C, D, E, F, G, H, I, J, K, L) (Tab 5.6). La concordance ou la discordance s'évalue en comparant l'ordre du classement n° 2 avec celui du classement n° 1.

Le classement n° 2 présente un ordre alternatif. Pour chaque individu de ce classement, on compare si l'ordre est naturel, donc **concordant**, ou si l'ordre est **discordant**. En première ligne, $1 = 1$, l'ordre est concordant ; on place dans la colonne « Concordant » $12 - 1 = 11$, et dans la colonne « Discordant » 0. En deuxième ligne, $2 = 2$, l'ordre est concordant ; on place dans la colonne « Concordant » $11 - 1 = 10$, et dans la colonne « Discordant » 0. En troisième ligne, $4 > 3$, l'ordre est discordant ; on place dans la colonne « Concordant » 8, et dans la colonne « Discordant » 1, de sorte que $8 + 1 = 9$. En quatrième ligne, $3 < 4$, l'ordre est concordant ; on place dans la colonne « Concordant » 8, et dans la colonne « Discordant » 0. En cinquième ligne, $6 > 5$, l'ordre est discordant ; on place dans la colonne « Concordant » 6, et dans la colonne « Discordant » 1, de sorte que $6 + 1 = 7$, et ainsi de suite. Une fois l'algorithme appliqué, on dénombre les couples concordants et les couples discordants. Plus on avance dans les lignes, plus le nombre de couples testés diminue, comme l'illustre la colonne « TOTAL ».

On pose $N_a = 61$ et $N_d = 5$. Le coefficient τ vaut alors :

$$\tau = \frac{N_a - N_d}{N_a + N_d} = \frac{S_C}{S_T} = \frac{56}{66} \approx 0,85 \tag{5.22}$$

La concordance $\tau = 1$ est **parfaite** s'il n'y a aucune paire discordante. Si $\tau = -1$, alors les classements sont parfaitement **inverses**.

Il est possible de mesurer la significativité de τ en calculant z.

$$z = \frac{3\tau\sqrt{n\,(n-1)}}{\sqrt{2\,(2n+5)}} \tag{5.23}$$

ici $\tau \approx 0,85$ et $n = 12$ rangs, donc $z \approx 3,85$. z est une variable t de Student avec un degré de liberté $ddl = S_T - 2$. Il n'y a plus qu'à regarder la probabilité critique au risque voulu pour juger si τ est significatif. Le test d'hypothèse est :

— H_0 : la concordance entre les classements est due au hasard ;

— H_1 : la concordance entre les classements n'est pas due au hasard.

5.3 La concordance de p classements

La concordance de p classements est la généralisation du coefficient de corrélation des rangs de M. G. Kendall. Dans ce cas, n individus ont été classés selon p critères (Tab. 5.7).

	Individu 1	**Individu 2**	**...**	**Individu** n
Critère 1	r_{11}	r_{21}	...	r_{n1}
Critère 2	r_{12}	r_{22}	...	r_{n2}
...	...	...	...	...
Critère p	r_{1p}	r_{2p}	...	r_{np}
Total	$r_{1.}$	$r_{2.}$	...	$r_{n.}$

TABLE 5.7 – Classement de n individus selon p critères

Chaque ligne est une permutation des entiers de 1 à n, la somme des termes de n'importe quelle ligne est égale à $\frac{n(n-1)}{2}$. La somme des termes du tableau N est de fait égale à $N = \frac{np(n+1)}{2}$. Si les classements étaient rigoureusement identiques, une des colonnes aurait pour somme p, une autre $2p$, une autre $3p$, *etc*.

Pour étudier la concordance entre ces classements, on considère la statistique :

$$S = \sum_{i=1}^{n} \left(r_{i.} - \frac{N}{n} \right)^2 \tag{5.24}$$

Cette statistique est une mesure de la dispersion des sommes des colonnes par rapport à leur moyenne.

Si la concordance est parfaite, la statistique S est alors maximale. Elle vaut :

$$S_{\max} = \frac{np^2\,(n^2-1)}{12} \tag{5.25}$$

En partant de ce constat, afin d'étudier la concordance de p classements, D. G. Kendall a proposé le coefficient W :

$$W = \frac{12S}{np^2\left(n^2-1\right)} = \frac{S}{S_{\max}} \tag{5.26}$$

Ce coefficient est compris entre 0 et 1.

Propriété 1. Pour $W = 0$, les sommes de toutes les colonnes sont égales.

Propriété 2. Une faible valeur de W indique l'indépendance entre les classements.

Propriété 3. H_0 est rejetée si W est trop grand. Des tables donnent les valeurs critiques de W pour différentes valeurs de n et de p.

N.B. 1. Pour $n \geq 15$ et $p < 7$, la variable $\frac{(p-1)W}{1-W}$ est une variable de Fisher $F\left[n-1-\frac{2}{p}, (n-1)\left(n-1-\frac{2}{p}\right)\right]$.

N.B. 2. Pour $p \geq 7$, la variable $p\,(n-1)\,W$ suit une loi du χ^2 à $(n-1)$ degrés de liberté.

5.4 Le coefficient Γ de Goodman-Krusdal

Le coefficient de Goodman-Krusdal est noté Γ ou g. Il se base sur la différence entre les paires concordantes (N_a) et les paires discordantes (N_d) [Goodman et Kruskal, 1954] [Goodman et Kruskal, 1959] [Goodman et Kruskal, 1963] [Goodman et Kruskal, 1972].

William Henry Krusdal (1919-2005)

Leo A. Goodman (1928-2020)

$$\Gamma = \frac{N_a - N_d}{N_a + N_d} = \frac{S_C}{S_T} \tag{5.27}$$

Γ calcule le « surplus » de paires concordantes par rapport aux paires discordantes. Il s'agit d'une proportion.

Γ varie entre -1 et $+1$. Si $\Gamma = 0$, les paires sont indépendantes.

Attention ! Γ peut être nul même s'il n'y a pas d'indépendance statistique dans le cas où $S_C = 0$ par exemple.

Γ s'interprète comme ρ et τ.

On peut faire un test de Student

$$t \approx \Gamma\sqrt{\frac{S_C}{n\left(1-\Gamma^2\right)}} \tag{5.28}$$

avec $n \neq S_C$

5.5 Le coefficient Q d'association de Yule

George Udny Yule (1871-1851)

Le coefficient Q d'association de G. U. Yule est un cas particulier du coefficient de Goodman-Krusdal. Il est appliqué dans le cas des matrices 2×2. Il faut construire la table de contingence qui évalue la fréquence des événements.

	Oui	**Non**	**TOTAL**
Positif	a	b	$a+b$
Négatif	c	d	$c+d$
TOTAL	$a+c$	$b+d$	n

TABLE 5.8 – Tableau de contingence

À partir du tableau 5.8, le coefficient Q vaut :

$$Q = \frac{ad - bc}{ad + bc} \tag{5.29}$$

Le signe de Q dépend de l'appariement que l'analyse effectuée considère être concordant, mais les couples restent symétriques. Le choix de l'appariement n'affecte pas l'ampleur de Q.

Q varie entre -1 et $+1$.

— $Q = -1$ signifie que l'association est négative totale.

— $Q = 0$ signifie qu'il existe aucune association.

— $Q = +1$ signifie que l'association est positive parfaite.

En termes de rapport de cotes [1] (ou de chances) OU, Q est donné par :

$$Q = \frac{\mathrm{OU} - 1}{\mathrm{OU} + 1} \tag{5.30}$$

avec OU, le connecteur logique.

Le Q de Yule et le Y de Yule sont liés :

$$Q = \frac{2Y}{1 + Y^2} \tag{5.31}$$

et

$$Y = \frac{1 - \sqrt{1 - Q^2}}{Q} \tag{5.32}$$

1. En anglais : *odds ratio*

Bibliographie

[Goodman et Kruskal, 1954] GOODMAN, L. A. et KRUSKAL, W. H. (1954). Measures of association for cross classification. Journal of the American Statistical Association, 49(268):732–764.

[Goodman et Kruskal, 1959] GOODMAN, L. A. et KRUSKAL, W. H. (1959). Measures of association for cross classification ii : Further discussion and references. Journal of the American Statistical Association, 54(285):123–163.

[Goodman et Kruskal, 1963] GOODMAN, L. A. et KRUSKAL, W. H. (1963). Measures of association for cross classification iii : Approximate sampling theory. Journal of the American Statistical Association, 58(302):310–364.

[Goodman et Kruskal, 1972] GOODMAN, L. A. et KRUSKAL, W. H. (1972). Measures of association for cross classification iv : Simplification of asymptotic variances. Journal of the American Statistical Association, 67(338):415–421.

[Kendall, 1938] KENDALL, M. G. (1938). A new measure of rank correlation. Biometrika, 30(1-2):81–93.

[Rateau, 2001] RATEAU, P. (2001). Méthode et statistiques expérimentales en sciences humaines. Université. Ellipses, Paris.

[Spearman, 1904] SPEARMAN, C. (1904). The proof and measurement of association between two things. The American Journal of Psychology, 15(1):72–101.