Le kappa de Cohen : un outil de mesure de l'accord inter-juges sur des caractères qualitatifs.

Frédéric Santos
CNRS, UMR 5199 PACEA
Courriel : frederic.santos@u-bordeaux1.fr

12 mars 2015

1. Présentation générale

Le κ de Cohen est un coefficient destiné à mesurer l'accord entre deux variables qualitatives ayant les *mêmes modalités*. Classiquement, il est utilisé afin de mesurer le degré de concordance entre les stades attribués par deux juges. Il peut également être appliqué afin de mesurer un accord intra-observateur [Coh60].

Classiquement, on dispose de plusieurs stades à attribuer, si possible en relativement petit nombre si l'étude s'effectue sur un faible effectif. Les deux observateurs répartissent n unités statistiques en p catégories (avec donc n très supérieur à p, idéalement).

Exemple. — Voici les stades donnés par deux juges à 20 objets différents :

Juge 1 : B, B, C, A, C, C, C, A, A, B, C, B, B, A, C, A, B, C, C, A.
Juge 2 : B, B, B, A, C, C, B, A, A, C, C, B, B, A, C, B, C, C, C, A.

Les deux listes de stades attribués par les juges peuvent être assimilées à deux variables qualitatives à p modalités (ici $p = 3$) dont on peut dresser la table de contingence :

	A	B	C
A	5	1	0
B	0	4	2
C	0	2	6

Remarquons que si l'accord était parfait entre les deux juges, la table de contingence aurait été nulle hors de la diagonale. Intuitivement, la qualité de l'accord se mesure au « poids » que représente la diagonale par rapport au reste du tableau.

En appelant $N = (n_{ij})_{i,j=1,\cdots,p}$ la table de contingence et n l'effectif total, la proportion d'accords observée est :

$$P_a = \frac{1}{n} \sum_{i=1}^{p} n_{ii}$$

Si les deux variables étaient indépendantes (*i.e.*, si l'accord entre les deux juges était parfaitement aléatoire), la proportion théorique d'accords observés pourrait être estimée par :

$$P_e = \frac{1}{n^2} \sum_{i=1}^{p} n_{i\cdot} n_{\cdot i}$$

En effet, « si tout se passe aléatoirement », $\frac{n_{i\cdot}}{n} \times \frac{n_{\cdot i}}{n}$ représente bien la probabilité de se voir attribuer simultanément le i-ème stade de cotation par les deux observateurs [Sap06].

On définit alors le coefficient kappa comme le rapport suivant :

$$\kappa = \frac{P_a - P_e}{1 - P_e}$$

Plus ce rapport est proche de 1, et plus l'écart entre la proportion d'accords observée et la proportion théorique d'accords « aléatoires », se rapproche de l'écart entre l'accord parfait des deux observateurs et leur concordance aléatoire.

Interprétation. — Le coefficient κ est toujours compris entre -1 et 1 (accord maximal). Habituellement, on utilise le « barème » suivant pour interpréter la valeur κ obtenue :

< 0	Grand désaccord
0.00 – 0.20	Accord très faible
0.21 – 0.40	Accord faible
0.41 – 0.60	Accord moyen
0.61 – 0.80	Accord satisfaisant
0.81 – 1.00	Accord excellent

Il s'agit de la grille de lecture proposée par Landis et Koch [LK77], mais elle n'est pas universellement acceptée, en bonne partie car elle ne prend pas compte le fait que l'interprétation du kappa de Cohen doit être modulée par le nombre de stades possibles. En effet, le kappa aura toujours tendance à être plus faible pour un grand nombre de stades possibles, que pour seulement deux stades à attribuer : il est plus facile de se mettre d'accord sur « Plutôt Noir / Plutôt Blanc » que sur toute une palette comportant 5000 gris !

Ainsi, par exemple, un $\kappa = 0.40$ pourra être considéré comme très médiocre si deux juges avaient seulement à choisir entre deux scores A et B, mais pourra être perçu comme honorable s'ils devaient choisir entre 10 stades différents. La table de Landis et Koch ne fournit donc qu'un guide de lecture, à compléter par l'expertise du praticien [BQMR97].

Préacutions d'usage. — (i) Le kappa de Cohen n'est pas à proprement parler un test, même s'il existe un test de non-nullité — qui n'a pas d'intérêt en soi. Tout comme un coefficient de corrélation entre deux variables quantitaives, il ne fournit qu'un indicateur numérique de ce que l'on cherche à mesurer, et son interprétation reste subjective.

(ii) Il est nécessaire de bien insister sur le fait que les deux juges doivent travailler avec les *mêmes* stades : le kappa de Cohen ne pourra rien mesurer de l'accord entre un juge donnant 3 stades « Neuf / Bon état / Usagé », et un juge donnant 5 stades « Neuf / Excellent état / Très bon état / État moyen / Mauvais état ». Au minimum, l'ensemble des stades attribués par un juge doit être totalement inclus dans l'ensemble des stades utilisés par l'autre juge : par exemple, si le juge 1 a attribué les stades A, B et C, et le juge 2 les stades A et B, cela peut convenir... À la condition philosophique près que le juge 2 devait être au courant de l'existence du stade C, et avoir *choisi* de ne pas l'utiliser.

Extension. — Le κ de Cohen ne fonctionne que pour mesurer l'accord entre *deux* juges. Pour plus de deux juges, on peut utiliser le kappa de Fleiss, dont l'interprétation est strictement identique, et qui est très bien illustré sur Wikipedia :

`http://en.wikipedia.org/wiki/Fleiss'_kappa`

2. Le kappa pondéré

Lorsque l'échelle de cotation est constituée de stades *totalement ordonnés*, le coefficient κ présente un inconvénient : un désaccord entre le premier et le dernier stade (donc, très important) n'aura pas plus de poids qu'un désaccord entre deux stades contigus. Il conviendrait donc de donner plus d'importance aux désaccords graves qu'aux désaccords légers [Coh68, BG97].

Si les observateurs ont à choisir entre p stades de cotation, on définit une matrice W de poids $(w_{ij})_{i,j=1,\cdots,p}$, symétrique, dont chaque valeur w_{ij} reflète l'importance que l'on souhaite donner au désaccord entre le i-ème et le j-ème stade de cotation.

La valeur du κ pondéré est alors donnée par :

$$\kappa_w = 1 - \frac{\sum_{i=1}^{p}\sum_{j=1}^{p} w_{ij}a_{ij}}{\sum_{i=1}^{p}\sum_{j=1}^{p} w_{ij}e_{ij}} = 1 - \frac{\mathrm{tr}(AW)}{\mathrm{tr}(EW)}$$

où $A = (a_{ij})_{i,j=1,\cdots,p}$ et $E = (e_{ij})_{i,j=1,\cdots,p}$ sont respectivement les valeurs réellement observées et les valeurs théoriques (sous hypothèse de comportement aléatoire) de la table de contingence des cotations faites par les deux observateurs.

Les principaux schémas de pondération sont les suivants :

— la pondération *linéaire*, correspondant à des valeurs de poids $w_{ij} = |i - j|$. En d'autres termes, un écart de cotation d'un seul stade se voit attribuer un poids 1, un écart de 2 stades se voit attribuer un poids 2, etc. Il s'agit d'une pondération « modérée », qui conviendra à la plupart des situations.

— la pondération *quadratique*, correspondant à des valeurs de poids $w_{ij} = (i - j)^2$. En d'autres termes, un écart de cotation d'un seul stade se voit attribuer un poids 1, un écart de 2 stades se voit attribuer un poids 4, etc. Il s'agit d'une pondération « sévère », qui sanctionne très fortement les écarts importants, et qui sanctionne peu les écarts n'impliquant qu'un seul stade.

Remarque. — Par construction, le κ pondéré quadratiquement fournira une mesure d'accord très optimiste et favorable dans le cas de deux observateurs dont les cotations ne diffèrent jamais (ou quasiment jamais) de plus d'un stade.

3. Utilisation pratique

3.1. Outils en ligne

Quelques outils gratuits disponibles en ligne pour le calcul du kappa de Cohen :

— Une feuille Excel prenant en argument le tableau de contingence des scores attribués par les deux juges :

`http://www.er.uqam.ca/nobel/r30574/Calcul/Kappa.xlsx`

Ce tableau de contingence peut quant à lui être préalablement généré sous Statistica ou R.

— Dans le même esprit, une plateforme de calcul en ligne prenant également en argument un tableau de contingence :

`http://faculty.vassar.edu/lowry/kappa.html`

— Un logiciel DOS très simple et documenté pour le calcul du kappa :

`http://kappa.chez-alice.fr/Kappa_cohen.htm`

3.2. Commandes R

Le logiciel R [R D11, CGH+12] (multiplateforme, libre et gratuit) dispose de nombreuses fonctions pour calculer les kappa de Cohen ou de Fleiss :

— la fonction `ckappa` du package `psy`, qui prend en argument un tableau à deux colonnes contenant les données brutes concernant les observations des juges (donc, la liste des stades attribués par chacun d'entre eux) ;
— la fonction `cohen.kappa` du package `psych`, qui prend en argument soit un tableau à deux colonnes soit une table de contingence comme en p. 1 ;
— la fonction `kappa2` du package `irr`, qui fonctionne de la même manière que la précédente ;
— la fonction `kappam.fleiss` du package `irr`, qui fonctionne encore de la même manière mais avec plus de deux colonnes, pour le cas du kappa de Fleiss.

Nous renvoyons ici à l'aide incluse dans R pour plus de précisions sur ces fonctions.

3.3. Un package R avec interface graphique : KappaGUI

Le package `KappaGUI`offre une interface graphique simple mais complète pour calculer les valeurs κ de Cohen ou de Fleiss [San13].

Pour l'installer, taper la commande suivante dans une console R (sur un ordinateur connecté à Internet) : `install.packages("KappaGUI", dep=TRUE)`.

Pour l'utiliser, charger le package *via* la commande usuelle `library(KappaGUI)`, puis ensuite taper simplement la commande `StartKappa()` et se laisser guider par l'interface graphique.

On notera simplement que :

(i) Le fichier d'entrée doit impérativement être au format CSV avec le point-virgule comme séparateur de champ.

(ii) S'il y a K observateurs ($K = 2$ pour le kappa de Cohen, $K \geqslant 3$ pour le kappa de Fleiss) ayant coté q variables différentes sur n individus, le fichier d'entrée doit être un tableau à n lignes et $K \times q$ colonnes, organisé comme l'exemple en table 1.

(iii) L'utilisateur reçoit en sortie un fichier CSV à 1 ligne et q colonnes, donnant la valeur du coefficient κ pour chacune des q variables observées.

(iv) Au moins lorsque l'on introduit une pondération, seuls des stades « réellement observés » doivent figurer dans le fichier d'entrée. Par exemple, il peut arriver qu'un des observateurs ait jugé un caractère comme étant « non observable » sur un individu, tandis que d'autres observateurs ont réussi à coter ce caractère. Il est alors prudent de laisser vides les K cases concernées pour l'individu et la variable en question, afin d'exclure l'individu de l'analyse.

Remarque. — Il n'existe pas de pondération pour le kappa de Fleiss, c'est pourquoi le choix du schéma de pondération n'est proposé que dans le cas de deux observateurs.

Individu	Var1_A	Var1_B	···	Var1_K	Var2_A	···	Var2_K	···	Var$q$_A	···	Var$q$_K
1											
2											
⋮											

TABLE 1 – Exemple de tableau d'entrée pour le package KappaGUI : les K premières colonnes correspondent aux cotations effectuées par les K observateurs sur la première variable, les K suivantes correspondent aux cotations effectuées par les K observateurs sur la deuxième variable, et ainsi de suite jusqu'aux K dernières colonnes correspondant à la dernière variable.

Références

[BG97] R. Bakeman and J.M. Gottman. *Observing interaction : An introduction to sequential analysis*. Cambridge University Press, 2nd edition, 1997.

[BQMR97] R. Bakeman, V. Quera, D. McArthur, and B.F. Robinson. Detecting sequential patterns and determining their reliability with fallible observers. *Psychological Methods*, (2) :357–370, 1997.

[CGH+12] P.A. Cornillon, A. Guyader, F. Husson, N. Jégou, J. Josse, M. Kloareg, E. Matzner-Løber, and L. Rouvière. *Statistiques avec R*. Presses universitaires de Rennes, 3e edition, 2012.

[Coh60] J. Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, (20) :37–46, 1960.

[Coh68] J. Cohen. Weighted kappa : Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, (70) :213–220, 1968.

[LK77] J.R. Landis and G.G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, (33) :159–174, 1977.

[R D11] R Development Core Team. R : A language and environment for statistical computing. `http://www.R-project.org/`, 2011. ISBN 3-900051-07-0.

[San13] Frédéric Santos. KappaGUI : GUI for Cohen's and Fleiss' Kappa. `http://CRAN.R-project.org/package=KappaGUI`, 2013. R package version 1.1.

[Sap06] G. Saporta. *Probabilité, statistique et analyse de données*. Technip, 2e edition, 2006.