

POUR TOUT SAVOIR OU PRESQUE SUR LE COEFFICIENT KAPPA...

I. BERGERI, R. MICHEL, J-P. BOUTIN

Med Trop 2002 ; **62** : 634-636

RESUME • Il est classique d'évaluer la qualité d'une méthode diagnostique au moyen de la sensibilité, de la spécificité et des valeurs prédictives. Ceci nécessite toutefois d'avoir une méthode de référence. Comment faire lorsque cette condition n'est pas remplie ? C'est tout le problème connu sous le nom de mesure de l'accord entre deux juges. Les auteurs présentent à partir d'un exemple tiré de la pratique africaine, l'intérêt et le calcul du coefficient Kappa qui est l'outil le plus utilisé dans ces problèmes de concordance et de reproductibilité de mesures.

MOTS-CLES • Concordance, reproductibilité, qualitatif, accord, coefficient Kappa.

ABSTRACT • EVERYTHING (OR ALMOST) ABOUT THE KAPPA COEFFICIENT...

ABSTRACT • Sensitivity, specificity, and predictive value are the standard parameters used to evaluate the efficacy of diagnostic tests. However all three parameters assume the existence of a gold standard test. Thus the problem arises as to what method to use in the absence of a benchmark. The solution involves assessment of agreement between two raters. This article based on an example drawn from an African setting describes the value and computation of the most widely used tool for assessing agreement and reproducibility of measures, i.e., the kappa coefficient.

KEY WORDS • Agreement - Reproducibility - Qualitative - Kappa coefficient.

Si nos jugements reflètent notre pensée, ils sont plus rarement en accord avec ceux d'autrui.

Cette variabilité inter-individuelle bénéfique pour l'homme, est cependant pénalisante dans de nombreuses disciplines scientifiques, où il est souvent nécessaire d'évaluer et d'améliorer l'accord entre des informations appliquées au même sujet.

Le test non paramétrique Kappa de Cohen (1) permet de chiffrer l'accord entre deux ou plusieurs observateurs ou techniques lorsque les jugements sont qualitatifs.

Pour illustrer nous prendrons un exemple tiré de la revue *Médecine Tropicale* et relatif de la culture du sang périphérique comme moyen diagnostique de la leishmaniose viscérale (la technique diagnostique habituelle étant jusqu'à présent la culture de moelle osseuse) (2). Prenons le cas où ces deux techniques, tentant de poser un diagnostic chez le même patient, proposent des résultats divergents en matière de leishmaniose viscérale. Cette multiplication des avis n'apporte pas la sécurité attendue d'un parfait accord diagnostique pour le médecin traitant et le patient.

• Travail du Service de Médecine des collectivités (IB, Interne de pharmacie ; RM, Assistant des hôpitaux des armées ; JPB Professeur agrégé du SSA), Institut de Médecine Tropicale du Service de Santé des Armées, Marseille, France.

• Correspondance : I. BERGERI, Service de Médecine des collectivités, Institut de Médecine Tropicale du Service de Santé des Armées, Marseille, France
• Fax: +33 (0) 4 91 52 26 07 e-mail : medco.imtssa@wanadoo.fr •

• Article sollicité.

Une solution consiste ici à réaliser une séance de «concordance» entre les techniques pour estimer leur taux d'accord par le coefficient Kappa et d'étudier leurs désaccords pour y remédier.

Plus généralement, le test statistique Kappa est utilisé dans les études de reproductibilité.

DEFINITION DE L'ACCORD

L'accord entre des jugements est défini comme la conformité de deux ou plusieurs informations qui se rapportent au même objet. Le taux d'accord ou de «concordance» est donc estimé par le coefficient Kappa proposé par Cohen (1).

Ici on souhaite évaluer le degré d'accord entre les diagnostics positifs et négatifs fournis par deux techniques A (culture de sang périphérique) et B (culture de moelle

Tableau 1 - Résultats de la recherche concomitante des leishmanies dans le sang périphériques (A) et la moelle osseuse (B) chez 49 patients (d'après 2).

	Réponses	Résultat de A		Total
		+	-	
Résultat de B	+	19	10	29
	-	9	11	20
	Total	28	21	49

osseuse) à la recherche de leishmanies sur les mêmes patients (2). L'étude porte sur 49 patients et les résultats sont présentés dans le tableau I.

DÉFINITION DU COEFFICIENT KAPPA

L'accord observé entre un ou plusieurs jugements qualitatifs, résulte de la somme d'une composante «aléatoire» (hasard) et d'une composante d'accord «véritable» (réel). Mais la part du hasard gêne notre appréciation. Pour contrôler ce hasard, le coefficient Kappa (K) propose de chiffrer l'intensité ou la qualité de l'accord réel (3). C'est un indice qui permet de «retirer» la portion de hasard ou de subjectivité de l'accord entre les techniques.

$$K = \frac{\text{Proportion d'accords observés} - \text{proportions d'accords dus au hasard}}{100 - \text{proportions d'accords dus au hasard}}$$

$$= \frac{P_o - P_e}{1 - P_e}$$

avec P_o : la proportion d'accord observée ou concordance observée
 P_e : la proportion d'accord aléatoire ou concordance aléatoire.

La présentation des résultats sous la forme d'un tableau de contingence montre que les deux techniques sont en accord pour 30 patients avec 19 réponses positives concordantes et 11 réponses négatives concordantes (Tableau I).

La concordance observée P_o est la proportion des individus classés dans les cases diagonales concordantes (↖↗) du tableau de contingence, soit :

$$P_o = (19+11)/49 = 0,612$$

Et la concordance aléatoire P_e est égale à la somme des effectifs théoriques des 2 cases concordantes, divisée par la taille de l'échantillon (N= 49) :

$$P_e = [(29*28)/49 + (20*21)/49] / 49 = 0,513$$

NB : Les effectifs théoriques se calculent ici comme dans un test du Chi carré de Pearson.

D'où :

$$K = \frac{0,612 - 0,513}{1 - 0,513} = 0,20$$

INTERPRETATION

Selon les auteurs de l'article nous ayant servi d'exemple (2), une valeur de 0,20 indique un bon accord entre les deux techniques lorsqu'on a pris en compte (enlevé) le hasard qui pouvait les mettre d'accord.

En effet, le coefficient Kappa est un nombre réel, sans dimension, compris entre -1 et +1. L'accord sera d'autant plus élevé que la valeur de Kappa est proche de +1 et l'accord maximal est atteint si $K = 1$. Lorsqu'il y a indépendance des jugements, le coefficient Kappa est égal à zéro ($P_o = P_e$), et dans le cas d'un désaccord total entre les juges, le coefficient Kappa prend la valeur -1.

Suivant le classement de Landis et Koch (5) qui est fréquemment utilisé en biologie, le Kappa de notre exemple aurait été considéré comme mauvais (Tableau II).

Tableau II - Degré d'accord et valeur de Kappa proposé par Landis et Koch (5).

Accord	Kappa
Excellent	0,81
Bon	0,80 - 0,61
Modéré	0,60 - 0,21
Mauvais	0,20 - 0,0
Très mauvais	< 0,0

Les limites de ce classement sont donc arbitraires et peuvent varier selon l'étude réalisée ; par exemple en psychiatrie où la part d'incertitude est grande, un accord «modéré» dans l'échelle proposée ci-dessus pourrait être considéré comme très satisfaisant, tel n'étant pas le cas devant un résultat d'analyse biologique. Dans tous les cas, le classement devra être défini avec des experts avant la réalisation de l'étude.

SIGNIFICATION STATISTIQUE

En général, on accompagne le calcul du coefficient K de son degré de signification p afin de pouvoir interpréter le résultat avec certitude (ce qui aurait dû être fait dans notre exemple).

Pour tester l'hypothèse nulle H_0 «indépendance des jugements» (d'où $K = 0$) contre l'hypothèse alternative H_1 ($K > 0$), on utilise :

$$Z = \frac{K}{(\text{Var } K)}$$

qui, sous H_0 , suit approximativement une loi normale centrée réduite.

On remarquera que la formulation de H_1 est toujours unilatérale, en effet un K négatif signifierait que les observateurs sont plus souvent en désaccord que sous l'effet du hasard, ce qui n'est pas vraisemblable dans la pratique courante.

On ne donnera pas ici le calcul de la variance (Var K) qui est assez complexe (1).

Dans notre exemple :

$$Z = \frac{0,20}{(0,0044)} = 3,08$$

L'interprétation de ce test est la suivante : si $Z > 1,64$, on rejette H_0 pour un risque $\alpha = 5\%$ en unilatéral. Les deux techniques de notre exemple concordent significativement plus que la chance seule ne pourrait l'expliquer ($p < 0,5\%$). La conclusion de Belhadj *et Coll* était donc valable (2). La culture du sang périphérique est bien dans le contexte tunisien une méthode valide pour le diagnostic de la leishmaniose viscérale.

USAGES EN PRATIQUE

Dans la pratique, le Kappa va servir pour déterminer :

- **La fiabilité d'un instrument de mesure.** Est-ce que le dosage d'un échantillon standard de référence de chloroquine donne bien le résultat attendu avec un nouvel appareil ?

- **L'accorder-observateurs.** Est-ce que deux biologistes qui parcourent le frottis sanguin d'une même personne à la recherche de *Plasmodium falciparum* vont obtenir le même résultat ?

- **L'accord intra-observateur.** Est-ce qu'un biologiste qui revoit le frottis de la même personne le lendemain va obtenir le même résultat que la veille ?

- N'opposez pas le coefficient Kappa avec la sensibilité (Se), la spécificité (Sp), les valeurs prédictives (VPP, VPN) ! Ces dernières ne sont pas des indices « concurrentiels » du coefficient Kappa mais parfaitement complémentaires (4). Le K est utilisé essentiellement dans les études où il n'y a pas de méthode de référence disponible.

CONCLUSION

Le coefficient Kappa de Cohen permet d'estimer, en prenant en compte le hasard, l'accord entre des jugements qualitatifs appliqués aux mêmes objets, fournis par deux observateurs ou techniques, dans le but de déceler et de quantifier les désaccords pour les corriger ou les interpréter. On ajoutera que l'extension de la méthode à plusieurs observateurs ou techniques est possible (1, 6).

Mais encore :

- de nombreux sites sur Internet proposent gratuitement des petits programmes téléchargeables de calculs de la

variance du coefficient Kappa et du Kappa lui-même, et ceci entre deux ou plus de deux jugements, il est donc inutile d'aller créer votre propre programme (6) ;

- le calcul du coefficient Kappa (entre deux jugements seulement) est aussi disponible sur le logiciel EPI INFO 6.04 fr qui est gratuit, dans le module EPITABLE. Le chemin d'accès est le suivant :

PROGRAMMES □ EPITABLE □ COMPARE □ PROPORTION □ CONCORDANCE ENTRE OBSERVATEURS ;

- sachez enfin que pour étudier la concordance entre des réponses quantitatives, on peut utiliser le coefficient de détermination R^2 qui est simplement le carré du classique coefficient de corrélation « r » à condition qu'il y ait linéarité du lien entre les 2 jugements.

Vous savez maintenant « dompter » le hasard, à vous de jouer, entrez vos données...

POUR EN SAVOIR PLUS

- 1 - COHEN J - A coefficient of agreement for nominal scales. *Educ Psychol Meas* 1960 ; **20** : 27-46.
- 2 - BELHADJ S, TOUMI NH, DAKHLIA H *et Coll* - La culture du sang périphérique comme moyen diagnostique de la leishmaniose viscérale : a propos de 61 cas. *Med Trop* 2002 ; **62** : 155-157
- 3 - GRENIER A - Décision médicale. Collection Sciences et Médecine. Masson ed, Paris, 1993.
- 4 - LAPLANCHE A, COM-NOUGUE C, FLAMANT R - Méthodes statistiques appliquées à la recherche clinique. Flammarion Médecine-Sciences ed, Paris, 1987.
- 5 - LANDIS JR, KOCH GG - The measurement of observer agreement for categorical data. *Biometrics* 1977a ; **33** : 159-174.
- 6 - <http://perso.worldonline.fr/kappa>.

