Accord entre observateurs : indice kappa de Cohen

Bernard BRANGER – Réseau « Sécurité Naissance – Naître ensemble »
des Pays de la Loire - 2, rue de la Loire – 44200 NANTES. Tél 02 40 48 55 81 –
Courriel : bernard.branger@naitre-ensemble-ploire.org
Octobre 2009

Tiré de P. Bonnardel : http://kappa.chez-alice.fr/kappa_intro.htm

Introduction

La variabilité inter-individuelle est une constante biologique, et est sans doute bénéfique pour l'Homme. Elle est cependant pénalisante dans de nombreuses disciplines scientifiques, où il est souvent nécessaire d'évaluer et d'améliorer l'accord entre des informations de même nature appliquées au même objet dans une exigence de contrôle de la qualité ou d'assurance qualité.

Le test non paramétrique **Kappa** (K) de Cohen[1] permet de chiffrer l'accord entre deux ou plusieurs observateurs ou techniques lorsque les jugements sont qualitatifs, contrairement au coefficient u de Kendall[3] par exemple, qui évalue le degré d'accord entre des jugements quantitatifs.

Prenons le cas dans le domaine médical où deux ou plusieurs praticiens examinant le même patient proposent des diagnostics différents ou des décisions thérapeutiques différentes. En l'absence d'une référence, cette multiplication des avis n'apporte pas la sécurité attendue d'un parfait accord diagnostique ou thérapeutique pour le médecin traitant et le patient. Il est donc important que l'accord dans une équipe de travail ou entre plusieurs équipes soit le meilleur possible pour garantir la qualité et la continuité des soins.

Une solution consiste ici à réaliser une séance de «concordance» entre les médecins 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** qui nécessitent d'estimer l'agrément entre deux ou plusieurs cotations lorsqu'on étudie une variable discontinue.

Définition de l'accord

L'accord entre des jugements est défini comme la conformité de deux ou plusieurs informations qui se rapportent au même objet. Cette notion implique l'existence d'une liaison entre les variables, exige des variables de même nature et un appariement des jugements.

Dans le cas de jugements catégoriels, c'est à dire que la variable aléatoire est discrète et mesurée sur une échelle à *r* niveaux, le taux d'accord ou de «concordance» est estimé par le coefficient Kappa proposé par Cohen[1] en 1960.

Il faut opposer à la notion d'accord ou d'agrément, la notion d'association qui ne tient pas compte du sens de la liaison et qui n'exige pas des variables de même nature. Il existe différentes statistiques d'association, en particulier le coefficient f de Kendall[2] et le coefficient C de Cramer[3].

Accord entre 2 juges

On souhaite évaluer le degré d'accord entre les réponses positives et négatives fournies par deux tests biologiques A et B appliqués aux mêmes échantillons sériques. L'étude porte sur 200 échantillons et les résultats sont présentés dans le tableau IV.

Tableau I - ***Résultats des tests A et B appliqués aux mêmes échantillons***

		Résultat du test A		
	Réponses	+	-	Total
Résultat du test B	+	**72**	**16**	88
	-	**25**	**87**	112
	Total	97	103	200

La présentation des résultats sous la forme d'un tableau de contingence montre que les deux tests sont en accord pour 159 échantillons avec 72 réponses positives concordantes et 87 réponses négatives concordantes.

La proportion d'accord observé et la proportion d'accord aléatoire sont :

$$P_o = \frac{1}{200} \times (72 + 87) = 0{,}795$$

$$P_e = \frac{1}{200^2} \times (88 \times 97 + 112 \times 103) = 0{,}5018$$

$$K = \frac{0{,}795 - 0{,}5018}{1 - 0{,}5018} = 0{,}5885 \approx 0{,}59$$

Cette valeur indique un accord modéré entre les deux tests.

Interprétation du résultat

- **< 0 Désaccord**
- **0.00 — 0.20 : Accord très faible**
- **0.21 — 0.40 : Accord faible**
- **0.41 — 0.60 : Accord modéré**
- **0.61 — 0.80 : Accord fort**
- **0.81 — 1.00 : Accord presque parfait**

Dans le cas particulier où deux catégories de réponses sont proposées, la formule de Kappa peut s'écrire :

$$K = \frac{2(ad - bc)}{n_{1.}n_{.2} + n_{2.}n_{.1}}$$

Tableau I – Détail d'un tableau pour le calcul de kappa

		Résultat du test A		
	Catégorie	+	-	Total
Résultat du test B	+	***a***	***b***	$n_{1.}$
	-	***c***	***d***	$n_{2.}$
	Total	$n_{.1}$	$n_{.2}$	***n***

➔ **La valeur du coefficient Kappa est indépendante de la taille de l'échantillon étudié.** Par exemple, si nous multiplions par 10 chacun des effectifs des cases du tableau de contingence présenté dans le tableau IV, soit un effectif total de l'échantillon égal à 2000, nous obtenons alors le même coefficient Kappa que pour notre échantillon égal à 200, mais la signification statistique de la valeur du coefficient Kappa sous l'hypothèse nulle sera plus grande.

➔ Nous avons vu que **la valeur maximale de Kappa est égal à 1 lorsque $P_o = 1$ et $P_e = 0,5$**. Ceci n'est vrai que dans le cas où les marginales sont égales ($p_{i.} = p_{.i}$) puisqu'il suffit de prendre les effectifs diagonaux (ceux qui expriment l'accord dans le tableau de contingence) égaux aux marginales et les effectifs non diagonaux égaux à 0. Pour des marginales données, Cohen[1] propose de déterminer la valeur maximale de Kappa (K_m) :

$$K_m = \frac{P_m - P_e}{1 - P_e}$$

avec

$$P_m = \sum_{i=1}^{r} \inf\left(p_{i.}, p_{.i}\right)$$

la proportion d'accord maximal.

Dans notre exemple des deux tests biologiques, les marginales ne sont pas égales d'où $K_m < 1$. Il est donc intéressant de connaître la valeur maximale de Kappa compte tenu des effectifs marginaux :

$$P_m = \frac{1}{200} \times (88 + 103) = 0,955$$

d'où

$$K_m = \frac{0,955 - 0,5018}{1 - 0,5018} = 0,9097 \approx 0,91$$

Ce qui nous permet de comparer le Kappa obtenu à K_m par le rapport :

$$\frac{0,59}{0,91} \times 100 \approx 65\ \%$$

En conclusion, l'accord obtenu entre les deux tests biologiques correspond à 65% de l'accord maximal qu'il pourrait atteindre.

Kappa pondéré

Certaines discordances entre les juges sont plus graves que d'autres. Cohen propose de donner à chacune des cases du tableau de contingence, **un poids w_{ij}** fixé a priori qui reflète l'importance que l'on attribue au désaccord.

Au tableau de contingence *r´r* représentant les résultats d'une étude d'accord, nous associons une matrice de poids *r´r* notée W définissant l'importance de chaque désaccord. Le tableau V défini la notation utilisée.

Tableau III - ***Matrice de poids W associée à un tableau de contingence r´ r***

		Juge A			
	Catégorie	1	2	...	*r*
Juge	1	w_{11}	w_{12}	...	w_{1r}
B	2	w_{21}	w_{22}	...	w_{2r}
	.				
	r	w_{r1}	w_{r2}	...	w_{rr}

On utilise le plus souvent des **poids de concordance** plutôt que des poids de discordance ; ceci peut varier de **1** pour les cases diagonales à **0** pour les cases qui correspondent au plus grand désaccord en considérant que l'échelle des catégories de jugements est ordonnée.

Le choix des poids du Kappa pondéré K_w peut être réalisé selon un **système de pondération linéaire** où chaque poids se calcule d'après :

$$w_{ij} = 1 - \frac{|i-j|}{r-1}$$

ou par un calcul de **type quadratique** :

$$w_{ij} = 1 - \frac{(i-j)^2}{(r-1)^2}$$

avec

- ***i*** : la i^{ème} colonne de la matrice des poids
- ***j*** : la j^{ème} ligne de la matrice des poids
- ***r*** : le nombre de modalités de jugement
- w_{ij} : le poids de la case *ij* du tableau de contingence

La matrice des poids sera choisie **symétrique** dans le cas d'une étude de reproductibilité et pour d'autres types d'étude elle peut être choisie **asymétrique** si l'on désire souligner une dissymétrie entre les juges. Prenons par exemple la situation d'apprentissage d'un étudiant A pour lequel on désire évaluer la conformité de ces jugements à ceux d'un expert B dans le domaine considéré. Deux modalités de jugement sont proposées (+ et -) et nous considérons en outre que le désaccord «A+ B-» est plus grave par ces conséquences que le désaccord «A- B+». Dans ces conditions, il est possible d'associer au tableau de contingence 2´2 une matrice asymétrique de poids de concordance qui pourrait être de la forme :

Tableau IV : Poids des concordances attribuées

		Etudiant (A)	
	Catégorie	**+**	**-**
Expert	**+**	1,0	**0,5**
(B)	**-**	**0,0**	1,0

Peu de travaux portent sur la façon dont on doit définir le système de poids et l'approche la plus logique semble être la définition du système par consensus entre experts.

La concordance observée $P_{o(w)}$ du kappa pondéré en fonction de la matrice des poids de concordance est définie par :

$$P_{o(w)} = \sum_{i=1}^{r} \sum_{j=1}^{r} w_{ij}\, p_{ij}$$

et la concordance aléatoire $P_{e(w)}$ est :

$$P_{e(w)} = \sum_{i=1}^{r} \sum_{j=1}^{r} w_{ij}\, p_{i.}\, p_{.j}$$

avec

$p_{ij} = n_{ij} / n$

$p_{i.} = n_{i.} / n$

$p_{.j} = n_{.j} / n$

n étant le nombre total d'observations

Le **Kappa pondéré** est donné par :

$$K_w = \frac{P_{o(w)} - P_{e(w)}}{1 - P_{e(w)}}$$

Les formules exprimant le Kappa non pondéré sont une simplification des formules du Kappa pondéré. En effet, K est un cas particulier de K_w avec le système de pondération : w_{ij} = 1 " $i = j$ et w_{ij} = 0 " $i \neq j$.

Signification statistiques

- Erreur standard de la concordance aléatoire

Pour tester l'hypothèse nulle que les jugements sont indépendants (H_0 : K = 0), c'est à dire que la seule liaison entre les jugements est due au hasard, Fleiss, Cohen et Everitt[6], ont montré que l'erreur standard de la concordance aléatoire S_{K0} est estimée par :

$$S_{K_0} = \frac{1}{(1 - P_e)\sqrt{n}} \sqrt{P_e + P_e^2 - C}$$

avec

$$C = \sum_{i=1}^{r} p_{i.}\, p_{.i} \left(p_{i.} + p_{.i} \right)$$

Et pour le Kappa pondéré (confirmé par Hubert[7]) :

$$S_{Kw0} = \frac{1}{\left(1 - P_{e(w)}\right)\sqrt{n}} \sqrt{P_{e(w)} + P_{e(w)}^2 - C}$$

avec

$$C = \sum_{i=1}^{r} \sum_{j=1}^{r} p_{i.}\, p_{.j} \left[w_{ij} - \left(\overline{w}_{i.} + \overline{w}_{.j} \right) \right]^2$$

et

$$\overline{w}_{i.} = \sum_{j=1}^{r} w_{ij}\, p_{.j} \qquad \overline{w}_{.j} = \sum_{i=1}^{r} w_{ij}\, p_{i.}$$

Cette estimation de l'erreur standard ne requiert aucune hypothèse sur les marginales et suppose seulement *n* fixé. Pour tester l'hypothèse nulle : «indépendance des jugements» (d'où $K = 0$) contre l'hypothèse alternative $H_1 : K > 0$, on utilise la variable aléatoire centrée réduite du coefficient *K*, soit :

$$\boxed{Z = \frac{K - 0}{S_{K0}}}$$ qui sous H_0 suit approximativement une **loi normale centrée réduite**. Si $\mathbf{Z > Z_{1-a}}$, on rejette H_0 pour un risque a unilatéral. Les formules précédentes sont asymptotiquement exactes. Cicchetti[8] propose que la taille de l'échantillon de l'étude soit supérieur à $\mathbf{2r^2}$ avec *r* étant le nombre de modalités de jugement. D'après Fermanian[9], la taille minimale de l'échantillon devrait être **25** pour $r = 3$ et **30** pour $r = 4$ ou 5.

- ***Erreur standard du coefficient Kappa***

L'estimation asymptotique de l'erreur standard de Kappa a été formulée par Fleiss, Cohen et Everitt[6] :

$$\boxed{S_K = \frac{\sqrt{A + B - C}}{\left(1 - P_e\right)\sqrt{n}}}$$

avec

$$A = \sum_{i=1}^{r} p_{ii} \left[1 - \left(p_{i.} + p_{.i} \right) \left(1 - K \right) \right]^2$$

$$B = (1 - K)^2 \sum_{i=1}^{r} \sum_{j=1,\, i \neq j}^{r} p_{ij} \left(p_{.i} + p_{j.} \right)^2$$

$$C = \left[K - P_e \left(1 - K \right) \right]^2$$

Et pour le Kappa pondéré :

$$S_{Kw} = \frac{1}{\left(1 - P_{e(w)}\right)\sqrt{n}}\sqrt{A - B}$$

avec

$$A = \sum_{i=1}^{r}\sum_{j=1}^{r} p_{ij}\left[w_{ij} - \left(\bar{w}_{i.} + \bar{w}_{.j}\right)\left(1 - K_w\right)\right]^2$$

$$B = \left[K_w - P_{e(w)}\left(1 - K_w\right)\right]^2$$

Les formules précédentes sont asymptotiquement exactes. Fleiss[10] conseille que la taille de l'échantillon de l'étude (*n*) soit supérieure ou égale à $3r^2$ pour comparer deux coefficients Kappa observés et $n \geq 16r^2$ pour déterminer l'intervalle de confiance du Kappa. La première estimation de S_k a été donnée par Cohen[1] :

$$S_K = \sqrt{\frac{P_o\left(1 - P_o\right)}{n\left(1 - P_e\right)^2}} = \sqrt{\frac{Var(P_o)}{\left(1 - P_e\right)^2}}$$

et en substituant P_o par P_e, il formule l'erreur standard de la concordance aléatoire :

$$S_{K0} = \sqrt{\frac{P_e\left(1 - P_e\right)}{n\left(1 - P_e\right)^2}} = \sqrt{\frac{Var(P_e)}{\left(1 - P_e\right)^2}}$$

Ce qui nous permet de tester l'indépendance entre les classements effectués par les deux juges ($p_{ij} = p_{i.}p_{.j}$ d'où $P_o = P_e$ et $K = 0$) en calculant la statistique :

$$Z = \frac{K - 0}{S_{K0}} = \frac{P_o - P_e}{1 - P_e}\sqrt{\frac{n\left(1 - P_e\right)^2}{P_e\left(1 - P_e\right)}} = \frac{P_o - P_e}{\sqrt{\frac{P_e\left(1 - P_e\right)}{n}}}$$

qui sous H_0 suit approximativement une loi normale centrée réduite.

Ces estimations tendaient à une surestimation de l'erreur standard de *K* en supposant que les p_{ij} suivaient des lois binomiales et que leurs marginales étaient égales.

Usages du test

Le test Kappa est un outil pratique, relativement simple et très utilisé en pratique dans le milieu médical :

- ✓ **Essai thérapeutique**
- ✓ **Evaluation de méthodes ou d'examens diagnostiques**
- ✓ Contrôle de la qualité des techniques et des soins.

Il permet d'estimer l'accord entre des jugements catégoriels fournis par deux ou plusieurs techniques ou observateurs en l'absence de référence et plus généralement, d'étudier la reproductibilité pour des variables aléatoires non continues.

Le coefficient Kappa peut révéler des désaccords cachés, une divergence systématique ou non entre des juges. Il permet de constater et de quantifier un désaccord pour permettre la mise en place d'une stratégie d'amélioration qui comporte 4 étapes selon Fermanian[9] :

1. Détection du désaccord par une séance de concordance ;
2. Discussion pour déterminer les causes du désaccord et standardiser les jugements ;
3. De nouvelles expérimentations courtes afin de vérifier l'efficacité de la discussion. Les étapes 2 et 3 peuvent être répétées jusqu'à obtenir un accord suffisant ;
4. Contrôle par une nouvelle séance de concordance portant sur les N sujets de la première séance de concordance, et comparaison des résultats.

Conclusion

Le test statistique Kappa de Cohen permet d'estimer, en prenant en compte la concordance aléatoire, l'accord entre des jugements catégoriels appliqués aux mêmes objets, fournis par deux ou plusieurs observateurs ou techniques dans le but de déceler et de quantifier les désaccords pour les corriger.

Ce test est un instrument précieux pour le contrôle de la qualité des techniques et des soins mais son interprétation exige une bonne connaissance de ces limites tout particulièrement sa dépendance vis-à-vis de la prévalence du signe recherché.

Le domaine d'application du coefficient Kappa dans les disciplines médicales est large, et des études d'accord, aussi bien entre des observateurs ou des examens, devraient être spontanément ou régulièrement réalisées dans le cadre de l'assurance qualité des soins et des techniques.

Logiciels utilisés : - Epi-Info DOS 6, - SPSS (tableaux croisés), - MedCalc, - Autres d'accès libre par internet

Références

1. Cohen J. : A coefficient of agreement for nominal scales., Educ. Psychol. Meas., 1960, 20, 27-46.
2. Kendall M.G. : Rank correlation methods, Hafner Pub.Co, New-York.
3. Siegel S., Castellan N.J. Jr. : Nonparametric Statistics for the Behavioral Sciences, McGraw-Hill International Editions, 1988, 2nd ed..
4. Grenier : Décision médicale, Masson, 1993.
5. Landis J.R., Koch G.G. : The Measurement of Observer Agreement for Categorical Data, *Biometrics*, 1977a, 33, 159-174.
6. Fleiss J.L., Cohen J., and Everitt B.S. : Large sample standard errors of kappa and weighted kappa, *Psychol. Bull*., 1969, 72, 323-327.
7. Hubert J.L. : A general formula for the variance of Cohen's weighted kappa, *Psychol. Bull*., 1978, 85, 183-184.
8. Cicchetti D.V., Fleiss J.L. : Comparaison of the null distributions of weighted Kappa and the C ordinal statistic, *Appl. Psychol. Meas*., 1977, 1, 195-201.
9. Fermanian J. : Mesure de l'accord entre deux juges. Cas qualitatif*, Rev. Epidém. et Santé Publ*., 1984, 32, 140-147.
10. Fleiss J.L. : Inference about weighted Kappa in the non-null case*, Appl. Psychol. Meas*., 1978, 1, 113-117.
11. Fleiss J.L. : Statistical Methods for Rates and Proportions, John Wiley and Sons, New York, 1981.
12. Landis J.R., Koch G.G. : A one-way components of variance model for categorical data, , *Biometrics*, 1977b, 33, 671-679.
13. Fleiss J.L., Cuzick J. : The reliability of dichotomous judgments : Unequal numbers of judges per subject. *Appl. Psychol. Meas*., 1979, 3, 537-542.
14. Feinstein A.R., Cicchetti D.V. : High agreement but low kappa : I. The problems of Two Paradoxes, *J. Clin. Epidemiol*., 1990, 43, 543-548.
15. Scott W.A. : Reliability of content analysis : The case of nominal scale coding, *Public Opinion Q*, 1955, 19, 321-325.
16. Bennet E.M., Alpert R., Goldstein A.C. : Communications through limited response questioning, *Public Opinion Q*, 1954, 18, 303-308.
17. Cicchetti D.V., Feinstein A.R. : High agreement but low kappa : II. Resolving the paradoxes, *J. Clin. Epidemiol*., 1990, 43, 551-558.
18. Byrt T., Bishop J., Carlin J.B. : Biais, Prevalence and Kappa*, J. Clin. Epidemiol*., 1993, 46, 423-429.
19. Holley J.W., Guilford J.P. : A note on the G index of agreement*, Educ. Psychol. Bull*., 1964, 32, 281-288.
20. Hui S.L., Walter S.D. : Estimating the error rates of diagnostic tests, *Biometrics*, 1980, 36, 167-171.
21. Walter S.D. : Measuring the reliability of clinical data : the case for using three observers, *Rev. Epidém. et Santé Publ*., 1984, 32, 206-211.
22. Walter S.D., Irwig L.M. : Estimation of test error rates, disease prevalence and relative risk from misclassified data : a review, *J. Clin. Epidemiol*., 1988, 41, 923-937.
23. Bertrand P., Benichou J., Chastang C. : Estimation par la méthode de Hui et Walter de la sensibilité et la spécificité d'un test diagnostique en l'abscence d'un test de référence : résultats d'une étude de simulation*, Rev. Epidém. et santé Publ*., 1994, 42, 502-511.
24. Reed III J.F., Reed J.J. : Cohen's weighted kappa with Turbo Pascal (FORTRAN), *Computer Methods and Programs in Biomedecine*, 1992, 38, 153-165.
25. Boushka W.M., Marinez Y.N., Prihoda T.J., Dunford R., Barnwell G.M. : A computer program for calculating kappa : application to interexaminer agreement in periodontal research*, Computer Methods and Programs in Biomedecine*, 1990, 33, 35-41.
26. Landis J.R., Koch G.G. : A one-way components of variance model for categorical data, , *Biometrics*, 1977b, 33, 671-679.
27. Haley S.M., Osberg J.S. : Kappa Coefficient Calculation Using Multiple Ratings Per Subjects : A Special Communication, *Phys. Ther*., 1989, 69, 970-974.