REVUE DE STATISTIQUE APPLIQUÉE

J.-M. TRICOT
Y. LEPAGE

Tests de l'incertitude des observateurs dans une analyse de concordance sur une échelle nominale

Revue de statistique appliquée, tome 40, n° 3 (1992), p. 35-45
<http://www.numdam.org/item?id=RSA_1992__40_3_35_0>

Rev. Statistique Appliquée, 1992, **XXXX** (3), 35-45

TESTS DE L'INCERTITUDE DES OBSERVATEURS DANS UNE ANALYSE DE CONCORDANCE SUR UNE ÉCHELLE NOMINALE

J.-M. Tricot*, Y. Lepage**

* *Groupe d'études et de recherche en analyse des décisions (GERAD), HEC, Université de Montréal, 5255 ave Decelles, Montréal, Québec H3T 1V6 (Canada).*
** *Département de mathématiques et de statistique, Université de Montréal, 5620 rue Darlington, CP 6128 Succ. A, Montréal, Québec H3C 3J7 (Canada).*

RÉSUMÉ

La notion d'incertitude des observateurs est définie dans un contexte d'analyse de concordance sur une échelle nominale. Cette incertitude est traduite sous forme d'hypothèses et testée, ensuite, à l'aide de tests d'ajustement. Les tests proposés sont liés à une mesure d'accord. Une technique d'agrégation sur les sujets permet d'améliorer les conditions d'utilisation des tests.

Mots-clés : *Statistique kappa, Mesure d'accord, Accord interobservateur, Echelle nominale, Statistique d'ordre, Test d'ajustement.*

ABSTRACT

The concept of uncertainty of multiple raters is defined in a context of agreement analysis on a nominal scale. Goodness-of-fit tests are used to test some hypothesis concerning uncertainty. These tests are based on an agreement measure. An agregation of subjects technic is applied in order to relax the conditions of utilisation of the tests.

Key-words : *Kappa statistic, Agreement measure, Multiple raters reliability, Nominal scale, Order statistics, Goodness-of-fit test.*

1. Introduction

Plusieurs auteurs ont présenté et étudié des mesures de l'accord entre deux ou plusieurs observateurs classant des sujets sur une échelle qualitative nominale. Une partie importante de la littérature repose sur la mesure kappa introduite par Cohen (1960, 1968). On trouve dans Landis et Koch (1975a, 1975b) une revue intéressante des premiers travaux sur ce sujet. D'autres auteurs ont continué à exploiter ce problème et citons, par exemple : Huber (1977), Fleiss, Nee et Landis (1979),

Kraemer (1980), Schouten (1982, 1986), James (1983), Zwick (1988), Bloch et Kraemer (1989), Jolayemi (1990) et Tricot (1991).

L'approche de Cohen (1960, 1968) consiste à utiliser la mesure kappa non seulement comme une statistique descriptive mais aussi comme base de tests paramétriques sur l'absence d'accord. Toutefois, comme le signalent Bloch et Kraemer (1989), cette approche peut entraîner, dans certains cas, des problèmes conceptuels puique la mesure kappa n'estime pas toujours explicitement un paramètre de population.

Dans le présent travail, l'idée de Cohen (1960) est reprise mais en modifiant la définition de l'absence d'accord, et en se dégageant de toute référence paramétrique. La problématique se situe alors dans un contexte d'ajustement. Deux statistiques différentes sont proposées, chacune pouvant être utilisée pour tester l'absence d'accord de plusieurs observateurs classant un sujet donné sur une échelle qualitative nominale. Chacune des statistiques mène à une statistique globale servant à tester l'absence d'accord de plusieurs observateurs classant plusieurs sujets sur l'échelle en question.

Plus précisément, dans la section 2, on introduit le problème d'un point de vue probabiliste, et on établit la dualité dans l'analyse entre accord et absence d'accord. La section 3 comprend la formulation explicite de l'hypothèse de l'absence d'accord sur un sujet ou sur un ensemble de sujets, comme une incertitude des observateurs ; on introduit deux statistiques basées sur une statistique d'ajustement, pour tester asymptotiquement l'absence d'accord des observateurs, d'abord sur un sujet, puis sur un ensemble de sujets. Dans la section 4, on présente deux nouvelles statistiques construites à l'aide d'une méthode d'agrégation des sujets ; ces statistiques permettent d'améliorer les conditions d'utilisation des lois asymptotiques servant aux tests d'hypothèses. Enfin, dans la section 5, un exemple permet d'illustrer les différentes méthodes proposées.

2. Le problème

Soient un échantillon de n sujets et un échantillon de d observateurs extraits respectivement d'une population de sujets et d'une population d'observateurs. On suppose que chaque observateur évalue chacun des sujets selon une échelle qualitative nominale comprenant L catégories notées $\{1, ..., L\}$. Par convention, on ajoute la catégorie 0 afin que, pour chaque sujet, les d observateurs se distribuent toujours en au moins deux groupes représentant respectivement une catégorie distincte. Ainsi, dans le cas d'un accord parfait réalisé sur la catégorie k ($1 \leqslant k \leqslant L$), on admet, par convention, que les observateurs se sont distribués en deux groupes d'effectifs d et 0 tels que d observateurs ont classé le sujet dans la catégorie k et 0, dans la catégorie 0. En fait la catégorie 0 n'est utilisée que dans le cas de l'accord parfait. Cette convention permet d'effectuer, dans tous les cas, l'analyse de concordance dans un contexte de dualité, soit sous l'angle de l'accord soit sous l'angle de l'absence d'accord.

Soient $S = \{1, ..., n\}$ l'ensemble des sujets de l'échantillon, et $S_m (2 \leqslant m \leqslant d)$, le sous-ensemble de S tel que les sujets de S_m ont été classés dans m catégories distinctes par les observateurs de la population d'observateurs. Pour $i = 1, ..., n$, on

suppose que $i \in S_{m_i}$ et soit $\{i_1, \ldots, i_{m_i}\}$ le sous-ensemble de $\{0, ..., L\}$ tel que i a été classé dans la catégorie i_j, $j = 1, \ldots, m_i$, au moins une fois si $i_j \in \{1, \ldots L\}$ et 0 fois si $i_j = 0$. Pour $i = 1, ..., n$, et $j = 1, ..., m_i$, on désigne par $n_{i_j m_i} > 0$ le nombre d'observateurs parmi les d observateurs de l'échantillon, qui ont classé le sujet i dans la catégorie i_j : $n_{i_j m_i} > 0$ si i_j appartient à $\{1, ..., L\}$ et 0 si $i_j = 0$; on a évidemment, $\sum_{j=1}^{m_i} n_{i_j m_i} = d$, $i = 1, \ldots n$. Enfin, pour $i = 1, ..., n$, et pour $j = 1, ..., m_i$, $p_{i_j m_i}$ désigne la probabilité pour un observateur tiré aléatoirement de la population d'observateurs, de classer le sujet $i \in S_{m_i}$ dans la catégorie i_j; on a donc, $\sum_{j=1}^{m_i} p_{i_j m_i} = 1$, $i = 1, \ldots, n$. Un estimateur de $p_{i_j m_i}$ est donné par :

$$\widehat{p}_{i_j m_i} = n_{i_j m_i}/d,$$

$i = 1, ..., n$ et $j = 1, ..., m_i$.

La formulation de certaines hypothèses appropriées concernant les $p_{i_j m_i}$ lorsque $i = 1, ..., n$ et $j = 1, ..., m_i$, permet de définir une nouvelle notion d'absence d'accord. Dans les deux prochaines sections, différentes statistiques seront présentées afin d'analyser le degré d'accord interobservateur à partir d'un sujet ou d'un ensemble de sujets.

3. La statistique Q

D'un autre point de vue, pour $i = 1, ..., n$ et $j = 1, ..., m_i$, la probabilité $p_{i_j m_i}$ peut aussi être interprétée comme la probabilité, pour le sujet i, d'être classé dans la catégorie i_j. Or on sait que l'entropie de l'expérience consistant à classer le sujet i dans une catégorie prise parmi m_i catégories auquelles correspondent les m_i probabilités $p_{i_j m_i}$ est maximum lorsque $p_{i_j m_i} = 1/m_i$. Cette entropie mesure, d'après Yaglom et Yaglom (1969), le degré d'incertitude de l'expérience. Lorsqu'elle est maximum le sujet i est inclassable. Ainsi, pour ce sujet i, $i = 1, ..., n$, être qualifié d'inclassable se traduit par le constat de l'incertitude des observateurs relativement à i, c'est-à-dire que pour tout j appartenant à $\{1, ..., m_i\}$, $p_{i_j m_i} = 1/m_i$. On définit donc l'absence d'accord sur le sujet i, $i = 1, ..., n$, par l'incertitude des observateurs formulée au moyen de ces m_i égalités. Cohen (1960) ne prend pas la même approche puisque, dans son modèle, l'absence d'accord se traduit par l'incapacité des observateurs de distinguer les sujets les uns des autres. Une incertitude sur la spécificité des sujets est remplacée, dans la nouvelle approche, par une incertitude dans l'évaluation de chaque sujet. Cette nouvelle formulation de l'absence d'accord sur le sujet i, dépend du nombre m_i de catégories utilisées pour classer i, $i = 1, ..., n$. Dans cette optique, l'accord est obtenu lorsque l'échelle nominale permet d'évaluer chaque sujet d'une manière exacte, et par conséquent, une mesure d'accord devient une mesure de la validité de l'échelle des catégories. La vérification de l'hypothèse d'incertitude des observateurs sur un sujet, se ramène à effectuer un test d'ajustement. On s'écarte ainsi du contexte

paramétrique usuel qui peut présenter des inconvénients conceptuels comme le notent Bloch et Kraemer (1989).

Pour un sujet $i \in S_{m_i}$, on formule donc l'hypothèse nulle suivante :

$$H_{0i} : \text{ Pour tout } j \in \{1, \ldots, m_i\},\ p_{i_j m_i} = 1/m_i$$

Cette hypothèse de l'incertitude des observateurs concernant le sujet i peut être testée à l'aide de n'importe quelle statistique d'ajustement comme par exemple, la statistique du χ^2 suivante :

$$Q_i = dm_i \sum_{j=1}^{m_i} (n_{i_j m_i}/d - 1/m_i)^2.$$

Pour des valeurs élevées de Q_i, on rejette l'hypothèse de l'incertitude ou de l'absence d'accord des observateurs. Comme lorsque d tend vers l'infini, Q_i est distribuée sous H_{0i} selon une loi χ^2 à $m_i - 1$ degrés de liberté (Kendall et Stuart, 1979), les points critiques asymptotiques du test basé sur Q_i peuvent facilement être déterminés. Cette approche suppose donc un échantillon d'observateurs de grande taille.

On remarque que m_i étant fixé pour le sujet i, Q_i est une fonction Schur-convexe des $n_{i_j m_i}$ $j = 1, ..., m_i$ (Marshall et Olkin, 1979) puisque Q_i est une fonction linéaire de $\sum_{j=1}^{m_i} n_{i_j m_i}^2$:

$$Q_i = (m_i/d) \sum_{j=1}^{m_i} n_{i_j m_i}^2 - d.$$

Ainsi, Q_i est maximum lorsque tous les $n_{i_j m_i}$ valent 1 sauf un qui vaut $d - m_i + 1$, pour le cas où $m_i > 2$, et lorsque l'accord est parfait, pour le cas où $m_i = 2$. Les valeurs élevées de Q_i sont, en fait, associées à des situations d'accord qui peuvent être considérées comme les meilleures.

Afin de tester globalement, sur un ensemble de sujets, l'incertitude des observateurs, on considère la conjonction des H_{0i}, c'est-à-dire l'hypothèse $H_0 = H_{01} \wedge \ldots \wedge H_{0n}$. Lorsque les procédures d'évaluation des sujets sont indépendantes d'un sujet à l'autre, les statistiques Q_i sont indépendantes, et il est donc possible de tester l'hypothèse H_0 à l'aide de la statistique :

$$Q = \sum_{i=1}^{n} Q_i.$$

Pour des valeurs élevées de Q, on rejette l'hypothèse H_0. Comme lorsque d tend vers l'infini, Q est alors distribuée sous H_0 selon une loi χ^2 à $\sum_{i=1}^{n} m_i - n$

degrés de liberté, les points critiques asymptotiques du test basé sur Q peuvent être aisément déterminés.

Pour des valeurs de $\sum_{i=1}^{n} m_i - n$ élevées (supérieures à 30 conventionnellement), la statistique de Fisher $\sqrt{2Q} - \sqrt{2\sum_{i=1}^{n} m_i - 2n - 1}$ ou celle de Wilson-Hilferty $\{(9Q/q)^{1/3} - 1 + 2/q\}/\sqrt{2/q}$, où $q = 9\sum_{i=1}^{n} m_i - 9n$ (Tassi, 1989), procurent des lois asymptotiquement normales, centrées et réduites lorsque d tend vers l'infini. Les points critiques asymptotiques du test basé sur Q peuvent aussi être obtenus à l'aide d'une loi normale, centrée et réduite.

Les conditions d'approximation des lois des statistiques Q_i et Q sont difficilement atteintes en pratique puisque la taille de l'échantillon des observateurs doit être grande. Pour pallier à ce problème, on introduit maintenant deux nouvelles statistiques généralisant les précédentes, afin de tester l'incertitude des observateurs avec un niveau critique adéquat pour toute valeur de d.

4. La statistique Q_T

Rappelons qu'à tout sujet i est associé un nombre m_i de catégories distinctes i_j , $j = 1, ..., m_i$, auxquelles correspondent des probabilités $p_{i_j m_i}$. Il s'ensuit que Q_i est une mesure pondérée des écarts entre les $\widehat{p}_{i_j m_i}$ et la probabilité théorique $1/m_i$, $i = 1, ..., n$. Ainsi Q_i ne dépend pas de la spécificité des catégories i_j mais d'un ensemble de probabilités estimées indépendant de toute permutation sur ces probabilités. La statistique Q_i s'inscrit dans un contexte d'analyse de concordance tel que déjà présenté par Bloch et Kraemer (1989) lorsqu'il distinguent les notions d'accord et d'association.

D'une manière générale, soient s et s' deux sujets dans S_m associés (de par leurs classements dans des catégories distinctes) respectivement aux deux sous-ensembles de catégories $\{s_1, \ldots, s_m\}$ pour s et $\{s'_1, \ldots, s'_m\}$ pour s' auxquels correspondent respectivement deux distributions de probabilités données par $p_{s_j m}$ et $p_{s'_j m}$, $j = 1, ..., m$, où $p_{s_j m}$ (resp. $p_{s'_j m}$) est la probabilité pour le sujet s (resp. s') d'être classé dans la catégorie s_j (resp. s'_j). Une comparaison de ces probabilités à $1/m$ peut être effectuée globalement et indépendamment de toute spécification des catégories, en comparant à $1/m$ une nouvelle probabilité h_{jm}, $j = 1, ..., m$, qui soit, par exemple, à distance euclidienne minimum des deux précédentes pour une bonne représentativité de celles-ci. Pour ce faire, on ordonne préalablement les $p_{s_j m}$ (resp. les $p_{s'_j m}$) de la plus petite à la plus grande ce qui n'apporte pas de restriction au problème en l'absence de spécification des catégories. Les $p_{s_j m}$ (resp. les $p_{s'_j m}$) ordonnées sont renotées $p_{s_{(j)} m}$ (resp. $p_{s'_{(j)} m}$). L'équation fondamentale de l'analyse de la variance (Scheffé, 1959) montre que la valeur minimum de

$\sum_{j=1}^{m}[(h_{jm}-p_{s_{(j)}m})^2+(h_{jm}-p_{s'_{(j)}m})^2]$ est atteinte pour $h_{jm}=(p_{s_{(j)}m}+p_{s'_{(j)}m})/2$, $j=1,...,m$, valeur que l'on retient donc pour une comparaison à $1/m$. D'un autre point de vue, on effectue ainsi une moyenne de distributions, en supprimant les effets de mode par permutation préalable des probabilités initiales (Titterington, Smith et Makov, 1985).

En général, lorsque $n_m = \text{Card}\, S_m$, la comparaison globale de toutes les probabilités $p_{s_j m}$ pour $s \in S_m$, à $1/m$, est effectuée en faisant la comparaison de $f_{jm} = (1/n_m) \sum_{s\in S_m} p_{s_{(j)}m}$ à $1/m$, pour $j = 1,...,m$. Si maintenant $n_{s_j m}$ est le nombre d'observateurs qui ont classé le sujet $s \in S_m$ dans la catégorie s_j, un estimateur $\widehat{f}_{jm}$ de la probabilité théorique $f_{jm} = (1/n_m) \sum_{s\in S_m} p_{s_{(j)}m}$ est :

$$\widehat{f}_{jm} = (1/n_m) \sum_{s\in S_m} n_{s_{(j)}m/d},$$

où $n_{s_{(1)}m} \leqslant \ldots \leqslant n_{s_{(m)}m}$ sont les statistiques d'ordre associées au vecteur $(n_{s_1 m}, \ldots, n_{s_m m})$.

La quantité f_{jm} représente une probabilité moyenne pour un sujet de S_m d'être classé dans la j^{e} catégorie d'un ensemble de m catégories sans spécificité. L'hypothèse de l'incertitude des observateurs pour un sujet de S_m s'exprime alors de la façon suivante :

$$H_m : \text{ Pour tout } j \in \{1,...,m\},\ f_{jm} = 1/m.$$

Par analogie avec les tests de H_{0i} précédents, un test pour H_m peut être basé sur la statistique :

$$Q_{mT} = n_m dm \sum_{j=1}^{m} (\widehat{f}_{jm} - 1/m)^2.$$

L'hypothèse H_m est rejetée pour des grandes valeurs de Q_{mT}. Puisque sous H_m, cette dernière statistique suit asymptotiquement une loi χ^2 à $m-1$ degrés de liberté lorsque $n_m d$ tend vers l'infini, les points critiques asymptotiques du test peuvent facilement être déterminés. Il s'ensuit que pour tester la conjonction $H = H_1 \wedge \ldots \wedge H_m$, on peut utiliser la statistique :

$$Q_T = \sum_{m\in M} n_m dm \sum_{j=1}^{m} (\widehat{f}_{jm} - 1/m)^2,$$

où $M = \{m : S_m \neq \varnothing\}$, l'hypothèse de l'incertitude des observateurs étant rejetée pour de grandes valeurs de Q_T. L'indépendance des différentes classes de sujets nous assure que, sous l'hypothèse H, la statistique Q_T suit asymptotiquement,

lorsque $n_m d$ tend vers l'infini, pour chaque $m \in M$, une loi χ^2 à $\sum_{m \in M} (m-1)$ degrés de liberté ; et ainsi, les points critiques asymptotiques du test basé sur Q_T peuvent aisément être déterminés.

Tout comme pour le test basé sur Q, les statistiques de Fisher ou de Wilson-Hilferty peuvent être utilisées lorsque $\sum_{m \in M} n_m d > 30$, pour déterminer les points critiques asymptotiques du test de l'hypothèse H.

Les conditions d'approximation de la loi de Q_T sont meilleures que celles de la loi de Q puisque $n_m d \geqslant d$. La statistique Q_T tient compte, contrairement à Q, d'une forme d'agrégation sur les sujets et en ce sens Q_T généralise Q.

5. Exemple

Fleiss (1971) présente un exemple qui se réfère à des données médicales. Cet exemple a d'ailleurs été repris successivement par Landis et Koch (1977) et Schouten (1986). L'expérience consiste à analyser l'accord entre 6 psychiatres ($d = 6$ observateurs) qui ont classé 30 sujets ($n = 30$) selon une échelle à 5 catégories : 1) la dépression, 2) les troubles de la personnalité, 3) la schizophrénie, 4) la névrose et 5) autre. Le Tableau 1 présente les différentes valeurs des $n_{i_j m_i}$ pour $i = 1, ..., 30$. Les catégories ont été notées de 1 à 5. La catégorie 0 a été ajoutée selon la convention précisée au début, pour le cas de l'accord parfait.

Comme $d = 6$, la validité du test de l'incertitude des observateurs basé sur Q est sujette à caution. On reporte, à titre indicatif, dans le Tableau 2, les valeurs de Q_i associées aux huit premiers sujets de l'échantillon.

Puisque $\sum_{i=1}^{n} m_i - n = 38 > 30$ plutôt que de calculer Q, on calcule les approximations de Fisher, $\sqrt{2Q} - \sqrt{2\sum_{i=1}^{n} m_i - 2n - 1} = 2,944$ et de Wilson-Hilferty, $\{(9Q/q)^{1/3} - 1 + 2/q\}/\sqrt{2/q} = 2,823$. On rejette l'hypothèse de l'incertitude des observateurs au niveau de signification de $0,0016$ dans le premier cas et de $0,0024$ dans le deuxième cas.

D'autre part, concernant Q_T, on obtient les fréquences empiriques $\widehat{f}_{jm}$: $\widehat{f}_{12} = 0,227$; $\widehat{f}_{22} = 0,773$; $\widehat{f}_{13} = 0,167$; $\widehat{f}_{23} = 0,271$; $\widehat{f}_{33} = 0,562$. A l'aide de ces valeurs, on reporte dans le Tableau 3 suivant, les différentes statistiques Q_{mT}.

Pour le test basé sur Q_T, on rejette donc aussi l'hypothèse de l'incertitude des observateurs mais à un seuil de signification inférieur à $1/1000$ et donc inférieur au précédent.

Tableau 1
Effectifs des regroupements des observateurs pour chaque catégorie et chaque sujet

sujet	catégorie					
	0	1	2	3	4	5
1	0				6	
2			3			3
3			1	4		1
4	0					6
5			3		3	
6		2		4		
7				4		2
8		2		3	1	
9		2			4	
10	0					6
11		1			5	
12		1	1		4	
13			3	3		
14		1			5	
15			2		3	1
16				5		1
17		3			1	2
18		5	1			
19			2		4	
20		1		2		3
21	0					6
22			1		5	
23			2		1	3
24		2			4	
25		1			4	1
26			5		1	
27		4				2
28			2		4	
29		1		5		
30	0					6

Tableau 2
Les statistiques Q_i associées aux sujets numérotés de 1 à 8

i	m_i	$\sum_{j=1}^{m_i} n_{ij\,m_i}^2$	Q_i	dl
1	2	36	6,00	1
2	2	18	0,00	1
3	3	18	3,00	2
4	2	36	6,00	1
5	2	18	0,00	1
6	2	20	0,67	1
7	2	20	0,67	1
8	3	14	1,00	2

Tableau 3
Les statistiques Q_{mT} du test de l'incertitude des observateurs sur chaque sous-échantillon S_m

	n_m	m	Q_{mT}	dl
	22	2	39.35	1
	8	3	12.07	2
	0	4	–	–
	0	5	–	–
Total :	$n = 30$		$Q_T = 51.42$	3

6. Conclusion

La présente approche permet l'utilisation de statistiques d'ajustement afin de tester l'absence d'accord de plusieurs observateurs sur un sujet ou sur un ensemble de sujets. Cette approche est basée sur l'analogie entre incertitude des observateurs et absence de validité d'une échelle de catégories. D'autres statistiques d'ajustement peuvent être aussi utilisées pour tester les différentes hypothèses.

Enfin, les différentes statistiques d'ajustement peuvent éventuellement être adaptées au cas où l'on considère un système de pondérations sur les catégories et au cas multidimensionnel où les sujets sont classés simultanément sur plusieurs échelles de catégories.

Remerciements : Ce travail a été partiellement subventionné par le Conseil National de Recherches en Sciences Naturelles et en Génie du Canada (subvention A-8555) et par le Fond National Suisse de la Recherche Scientifique.

Références

Bloch, D.A. and Kraemer, H.C. (1989), 2 × 2 Kappa coefficients : measures of agreement or association, *Biometrics* 45, 269-287.

Cohen, J. (1960), A coefficient of agreement for nominal scales, *Educ. and psychol. Measurement* 20, 37-46.

Cohen, J. (1968), Weighted kappa : nominal scale agreement with provision for scaled disagreement or partial credit, *Psychological Bulletin* 70, 213-220.

Fleiss, J.L. (1971), Measuring nominal scale agreement among many raters, *Psychological Bulletin* 76, N° 5, 378-382.

Fleiss, J.L., Nee, J.C.M. and Landis, J.R. (1979), Large sample variance of kappa in the case of different sets of raters, *Psychological Bulletin* 86, 974-977.

Huber, L. (1977), Kappa revisited, *Psychological Bulletin* 84, 289-297.

James, I.R. (1983), Analysis of nonagreements among multiple raters, *Biometrics* 39, 651-657.

Jolayemi, E.T. (1990), On the measure of agreement between two raters, *Biometrical Journal* 32, 87-93.

Kendall, J.R. and Stuart, A. (1979), *The Advanced Theory of Statistics*, 4th ed., Griffin, London.

Kraemer, H.C. (1980), Extension of the kappa coefficient, *Biometrics* 36, 207-216.

Landis, J.R. and Koch, G.G. (1975a), A review of statistical methods in the analysis of data arising from observer reliability studies (Part I), *Statistica Neerlandica* 29, 101-123.

Landis, J.R. and Koch, G.G. (1975b), A review of statistical methods in the analysis of data arising from observer reliability studies (Part II), *Statistica Neerlandica* 29, 151-161.

Landis, J.R. and Koch, G.G. (1977), A one-way components of variance model for categorical data, *Biometrics* 33, 671- 679.

Marshall, A.W. and Olkin, I. (1979), *Inequalities : Theory of Majorization and Its Applications*, Mathematics in Science and Engineering, Vol 143, Academic Press Inc., New York.

Scheffé, H. (1959), *The Analysis of Variance*, John Wiley, New York.

Schouten, H.J.A. (1982), Measuring pairwise agreement among many observers. II. Some improvements and additions, *Biometrical Journal* 24, N° 5, 497-504.

Schouten, H.J.A. (1986), Nominal scale agreement among observers, *Psychometrika* 51, 453-466.

Tassi, P. (1989),Méthodes Statistiques, 2e ed., Economica, Paris.

Titterington, D., Smith, A. and Makov, U. (1985), *Statistical Analysis of Finite Mixture Distributions*, John Wiley, New York.

Tricot, J.M. (1991), Un modèle d'accord entre observateurs sur une échelle nominale, *Comptes Rendus Mathématiques de l'Académie des Sciences du Canada, Vol XIII*, 4, 146-150.

Yaglom, A.M. et Yaglom, I.M. (1969), Probabilité et Information, Dunod, Paris.

Zwick, R. (1988), Another look at interrater agreement, *Psychological Bulletin* 103, 374-378.