Psicothema
ISSN: 0214-9915
psicothema@cop.es
Universidad de Oviedo
España

Ato García, Manuel; Benavente, Ana; López, Juan J.
Análisis comparativo de tres enfoques para evaluar el acuerdo entre observadores
Psicothema, vol. 18, núm. 3, 2006, pp. 638-645
Universidad de Oviedo
Oviedo, España

Psicothema 2006. Vol. 18, nº 3, pp. 638-645
www.psicothema.com

ISSN 0214 - 9915 CODEN PSOTEG

METODOLOGÍA

Análisis comparativo de tres enfoques para evaluar el acuerdo entre observadores

Manuel Ato, Ana Benavente y Juan J. López
Universidad de Murcia

En este trabajo se realiza un análisis comparativo de tres enfoques generales para evaluar el acuerdo entre observadores que originan tres tipos diferentes de medidas: descriptivas (coeficientes σ, π y κ), basadas en modelos loglineales y basadas en modelos con mezcla (*mixture*) de distribuciones. Aunque los enfoques loglineal y mixture asumen un concepto de corrección del azar diferente, la flexibilidad de ambos enfoques permite utilizar sus parámetros para realizar un análisis en profundidad de las pautas de acuerdo y desacuerdo. Se demuestra que las medidas descriptivas pueden obtenerse introduciendo restricciones al modelo loglineal de cuasi-independencia. Y que aunque los enfoques loglineal y mixture tienen una interpretación similar cuando todos los parámetros diagonales son positivos, su interpretación difiere en caso contrario debido a la diferente escala de medida utilizada.

Comparative analysis of three approaches for rater agreement. This work is a comparative analysis of three general approaches for rater agreement measurement which produce three different statistical coefficients: descriptive (σ, π y κ coefficients), based on log-linear models and on mixture models. Although loglinear and mixture approaches assume a different concept of chance correction, their flexibility allow using its parameters to analyze agreement and disagreement patterns. We show that all descriptive measures can also be obtained by introducing constraints to a basic loglinear (or mixture) quasi-independence model. Due to the different measurement scale used, both loglinear and mixture approaches have a similar interpretation when diagonal parameters are positive but a different interpretation otherwise.

Son muchas las situaciones de investigación aplicada en las ciencias sociales y biológicas donde es necesario cuantificar el acuerdo existente entre las medidas reportadas por dos (o más) observadores (jueces o evaluadores) o entre dos (o más) instrumentos de medición de una determinada respuesta. Medidas altas de acuerdo indican la existencia de consenso en el proceso de clasificación de los observadores, de intercambiabilidad entre los instrumentos de medición o de perfecta reproductibilidad de la medida.

Se han propuesto muchos coeficientes de acuerdo bajo dos grandes situaciones prácticas, dependiendo de la escala de medida requerida por el instrumento de medición. Cuando la medida es cuantitativa, uno de los índices más utilizado es el *coeficiente de correlación de Pearson.* Es bien conocido que este coeficiente mide la asociación entre dos variables, pero no proporciona información válida acerca del acuerdo. Por ejemplo, el coeficiente de correlación entre una medida que es el doble de la primera es perfecto, pero el acuerdo entre ambas medidas es nulo. Más convenientes son el *coeficiente de correlación intraclase* (Shrout y Fleiss, 1979), un índice de fiabilidad que evalúa la intercambiabilidad u homogeneidad entre medidas cuantitativas, los procedimientos de la *teoría de la generalizabilidad* (Cronbach, Gleser y Rajaratnam, 1972) y el *coeficiente de concordancia* (Lin y otros, 2002), un índice de reproductibilidad que evalúa el acuerdo midiendo la desviación de los datos respecto de la línea de igualdad o concordancia (45º) a través del origen.

Cuando la medida es categórica, nominal u ordinal, el objeto es la clasificación de un conjunto de objetos en categorías bien definidas y los procedimientos estadísticos para la evaluación del acuerdo parten de una sugerencia de Scott (1955) de corregir del acuerdo observado la proporción de casos para los que el acuerdo tuvo lugar solo por azar (*chance correction*). Varias *medidas descriptivas* se han definido a partir de esta sugerencia (véase Zwick, 1988). Además del coeficiente π introducido por Scott (1955) y de

Fecha recepción: 2-8-05 • Fecha aceptación: 22-2-06
Correspondencia: Manuel Ato García
Facultad de Psicología
Universidad de Murcia
30080 Murcia (Spain)
E-mail: matogar@um.es

una propuesta original de Bennett y otros (1954), el más popular es el coeficiente κ presentado por Cohen (1960). Las medidas descriptivas son un enfoque sencillo y universalmente aceptado de medir el acuerdo, pero tienen el inconveniente de que no permiten comprender la naturaleza del acuerdo y del desacuerdo (véase, sin embargo, von Eye y Mun, 2005) y se basan en algún modelo estadístico que en su aplicación práctica se asume como válido.

Contra la simplicidad de las medidas descriptivas surgió un enfoque que utiliza *modelos loglineales* para descomponer los patrones de acuerdo y desacuerdo observados mediante la formulación de un conjunto específico de parámetros y del manejo adecuado de matrices de diseño, asociado al procedimiento de ajuste condicional de modelos. Cada uno de los modelos loglineales permite describir el acuerdo mediante una medida (λ) de naturaleza similar a las desarrolladas bajo el enfoque descriptivo, aunque basada en un concepto de corrección del azar diferente (Guggenmoos-Holtzman y Vonk, 1998).

El enfoque más general emplea *modelos con mezcla de distribuciones* (o *modelos mixture*), que descomponen un conjunto de objetos en dos clases latentes, una asociada al acuerdo de objetos de fácil clasificación (acuerdo sistemático), con probabilidad μ, y la otra al acuerdo de objetos de difícil clasificación (acuerdo aleatorio y desacuerdo), con probabilidad (1-μ). A su vez, la probabilidad μ de la clase latente para la distribución de acuerdo perfecto puede también considerarse como una medida de acuerdo basada en el mismo criterio de corrección del azar de los modelos loglineales, aunque utiliza una escala de medida diferente.

En este trabajo se realiza un análisis comparativo de los tres enfoques citados definiendo un conjunto de seis medidas propuestas en la literatura para evaluar el acuerdo e ilustrando su utilización mediante un ejemplo tomado de la investigación psicológica. Para cada enfoque se analizan los procedimientos de corrección del azar, las restricciones que deben aplicarse a los modelos para reproducir las medidas descriptivas y la interpretación de sus parámetros. La siguiente sección introduce el ejemplo y algunas cuestiones básicas necesarias para proceder en las siguientes secciones con la definición de medidas descriptivas, basadas en modelos loglineales y modelos *mixture*. Finalmente se proponen algunas sugerencias a tener en cuenta por el investigador aplicado para abordar una más apropiada valoración del acuerdo.

Notación y ejemplo

Sea un conjunto de *K* observadores (jueces o evaluadores) que clasifican independientemente *N* sujetos (u objetos) sobre una escala categórica (nominal u ordinal) compuesta de *M* categorías. La notación general, para el caso de *K= 2* observadores, suele representar las frecuencias (n_{ij}) y/o las proporciones (p_{ij}) en una *tabla de acuerdo* como la que se reproduce en el cuadro 1.

Dos aspectos diferentes de la distribución conjunta n_{ij} de las respuestas son el grado de *acuerdo* y el grado de *asociación*. Para que exista acuerdo se requiere que exista asociación, pero es posible que exista un alto grado de asociación sin que exista un alto grado de acuerdo (Bloch y Kraemer, 1989). Por ejemplo, si el observador A valora los objetos sistemáticamente una categoría superior a la del observador B, el grado de acuerdo será bajo, pero el grado de asociación alto. En este trabajo nos interesa específicamente la evaluación del grado de acuerdo.

En el cuadro 2 se muestra un ejemplo, tomado de un trabajo de Dillon y Mullani (1984), en el que *K=2* observadores registraron un conjunto de 164 respuestas cognitivas elicitadas en un estudio de comunicación persuasiva sobre una escala con *M=3* categorías de respuesta («positiva», «neutral» y «negativa»). Cada casilla representa las frecuencias (en negrita) y las probabilidades (entre paréntesis).

Cabe considerar dos contextos (Martín y Femia, 2004; Agresti, 2002). Si uno de los observadores, por ejemplo A, es un experto medido sin error (*'gold standard'*), el objetivo es evaluar la clasificación realizada por el observador falible B y entonces se trata de un *estudio de concordancia (o de validez)*. En cambio, si ninguno de los dos observadores es experto (o sea, tienen una experiencia similar), el objetivo es evaluar su grado de acuerdo y entonces se trata de un *estudio de consistencia (o fiabilidad)*. A su vez, los estudios de fiabilidad pueden ser *inter-observadores* (A y B son ob-

Cuadro 1
Notación general

	Observador B						
Observador A	***1***	***2***	**...**	***j***	**...**	***M***	**Marginal A**
1	n_{11} (p_{11})	n_{12} (p_{12})	...	n_{1j} (p_{1j})	...	n_{1M} (p_{1M})	n_{1+} (p_{1+})
2	n_{21} (p_{21})	n_{22} (p_{22})	...	n_{2j} (p_{2j})	...	n_{2M} (p_{2M})	n_{2+} (p_{2+})
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
i	n_{i1} (p_{i1})	n_{i2} (p_{i2})	...	n_{ij} (p_{ij})	...	n_{iM} (p_{iM})	n_{i+} (p_{i+})
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
M	n_{M1} (p_{M1})	n_{M2} (p_{M2})	...	n_{Mj} (p_{Mj})	...	n_{MM} (p_{MM})	n_{M+} (p_{i+})
Marginal B	n_{+1} (p_{+1})	n_{+2} (p_{+2})		n_{+j} (p_{+j})		n_{+M} (p_{+M})	$n_{++}=N$ ($p_{++}=1$)

Nota: n_{ij} representan frecuencias de respuesta, p_{ij} representan probabilidades de respuesta.

Cuadro 2
Frecuencias (y probabilidades) del ejemplo de Dillon y Mullani (1984)

Observador A	**Observador B**			
	Positiva	***Neutral***	***Negativa***	***Total***
Positiva	**61** (0.372)	**26** (0.159)	**5** (0.030)	**92** (0.561)
Neutral	**4** (0.025)	**26** (0.159)	**3** (0.018)	**33** (0.201)
Negativa	**1** (0.006)	**7** (0.043)	**31** (0.189)	**39** (0.238)
Total	**66** (0.402)	**59** (0.360)	**39** (0.238)	**164** (1.000)

servadores diferentes) o *intra-observadores* (A y B representan al mismo observador que clasifica los objetos en dos o más ocasiones distintas). Aunque se puede extender a otros contextos, este trabajo analiza la consistencia o fiabilidad inter-observadores.

Es también conveniente distinguir dos diferentes tipos de muestreo. En el *muestreo tipo I* (muestreo multinomial) se prefija el total muestral (*N*) y los marginales de fila y columna se asumen aleatorios. En el *muestreo tipo II* (muestreo multinomial de producto), por el contrario, se prefijan los marginales de fila (o de columna). Asumimos en este trabajo muestreo tipo I.

Medidas descriptivas de acuerdo entre observadores

La probabilidad de acuerdo observada

La forma más directa de medir el acuerdo utiliza la *probabilidad de acuerdo observada* (p_0), que es simplemente la suma de las proporciones de la diagonal principal de la tabla de acuerdo. Para los datos del cuadro 2,

$$p_0 = \sum_i p_{ii} = 0.372 + 0.159 + 0.189 = 0.720$$

El principal problema de que adolece p_0 es que expresa el acuerdo bruto sin tener en cuenta que una parte del acuerdo observado puede ser debido al azar y ocurrir aún en el caso de que no haya ninguna tendencia sistemática por parte de los observadores para clasificar de forma similar los objetos.

Dada una tabla de acuerdo entre 2 (o más) observadores, una medida general de acuerdo corregido del azar (*RCA*), que distingue la probabilidad de acuerdo observada (p_0) de la probabilidad de acuerdo esperada por azar (p_e), es

$$RCA = \frac{p_0 - p_e}{1 - p_e} \qquad \text{(Ec. 1)}$$

donde p_e es el acuerdo que se atribuye al azar y 1-p_e establece la magnitud máxima de acuerdo no atribuido al azar. Las diferentes opciones que se han planteado para definir una *RCA* han consistido en especificar alguna fórmula de corrección del azar para definir p_e. Tres de las opciones más comunes se contemplan en este trabajo. Otras opciones alternativas pueden consultarse en Dunn (1989) y Shoukri (2004).

El coeficiente σ de Bennet y otros

Una forma simple de corrección del azar originalmente propuesta en un trabajo de Bennet y otros (1954) es el coeficiente σ (*coeficiente Sigma*), que utiliza un valor de corrección del azar fijo, la inversa del número de categorías (*M*),

$$p_e^{\sigma} = \sum_i M\left(\frac{1}{M}\right)^2 = \frac{1}{M} \qquad \text{(Ec. 2)}$$

Esta forma de corrección asume que los observadores clasifican los objetos uniformemente entre las categorías de respuesta (supuesto de *uniformidad marginal*). σ ha sido reivindicado como una medida de acuerdo estable por Holley y Guilford (1964), que lo llamaron índice *G*, Janson y Vegelius (1979), que lo llamaron coeficiente *C*, Brennan y Prediger (1981), que lo llamaron índice κ_n, y Maxwell (1977), que lo llamó coeficiente RE (véase Zwick, 1988; Hsu y Field, 2003). Sustituyendo la Ec. 2 en la Ec. 1 se obtiene

$$\sigma = \frac{p_0 - (1/M)}{(M-1)/M} \qquad \text{(Ec. 3)}$$

Aplicando la Ec. 3 a los datos del cuadro 2, σ= 0.579. Para el caso dicotómico una fórmula abreviada para la Ec. 3 es: $\sigma = 2p_0 - 1$.

El coeficiente π de Scott

En un trabajo de Scott (1955) se propuso como corrección del azar el cuadrado de la suma de los promedios de los marginales de fila y columna para cada categoría

$$p_e^{\pi} = \sum_i \left(\frac{p_i + p_{+i}}{2}\right)^2 \qquad \text{(Ec. 4)}$$

Para los datos del cuadro 2 la probabilidad esperada por azar es

$$p_e^{\pi} = \left(\frac{0.561 + 0.402}{2}\right)^2 + \left(\frac{0.201 + 0.360}{2}\right)^2 + \left(\frac{0.238 + 0.238}{2}\right)^2 = 0.367$$

Esta corrección asume que la distribución de las probabilidades marginales es homogénea para ambos observadores (supuesto de *homogeneidad marginal*), se denomina *coeficiente Pi* y se define utilizando Ec. 1 como

$$\pi = \frac{p_0 - p_e^{\pi}}{1 - p_e^{\pi}} \qquad \text{(Ec. 5)}$$

Para los datos del cuadro 2, el coeficiente pi es π= 0.557.

El coeficiente κ de Cohen

Basándose en el trabajo de Scott (1955), Cohen (1960) propuso una fórmula de corrección del azar consistente en calcular el valor esperado de los elementos diagonales de la tabla de acuerdo mediante

$$p_e^{\kappa} = \sum_i p_{i+} + p_{+i} \qquad \text{(Ec. 6)}$$

que asume independencia entre observadores y denominó *coeficiente Kappa*

$$\kappa = \frac{p_0 - p_e^{\kappa}}{1 - p_e^{\kappa}} \qquad \text{(Ec. 7)}$$

Para los datos del cuadro 2,

p_e^{κ}= (.561)(.402)+(.201)(.360)+(.238)²= 0.355

y κ= 0.567 Kappa tiene propiedades estadísticas óptimas como medida de acuerdo. En primer lugar, cuando el acuerdo observado (p_0) es igual al acuerdo esperado por azar (p_e^κ) entonces κ= 0. En segundo lugar, κ tomará su valor máximo de 1 si y solo si el acuerdo es perfecto (o sea, p_0= 1 y p_e^κ= 0). Y, finalmente, κ nunca puede ser menor de –1. Sin embargo, los límites superior e inferior del índice son función de las probabilidades marginales. Así, κ toma el valor 1 si y solo si las probabilidades marginales son exactamente iguales y todas las casillas no diagonales son cero.

Desde su formulación, κ se convirtió en la medida de acuerdo más utilizada en las ciencias biológicas y sociales. Fleiss, Cohen y Everitt (1969) derivaron una fórmula asintótica para la desviación típica

$$S(\kappa) = \frac{(1)}{1 - p_e^\kappa \sqrt{n}} \sqrt{p_e^\kappa + \left(p_e^\kappa\right)^2 - \sum_i p_{i+} + p_{+i}\left(p_{i+} + p_{+i}\right)} \qquad \text{(Ec. 8)}$$

que permite probar la hipótesis nula de acuerdo nulo entre ambos observadores empleando $z = \kappa / S(\kappa)$.

La utilización de Kappa como medida de acuerdo ha recibido muchas críticas. Básicamente, se han detectado dos problemas que pueden explicarse en términos de los efectos de *sesgo* y *prevalencia*. El efecto de sesgo de un observador respecto de otro ocurre cuando sus probabilidades marginales son diferentes; el sesgo es mayor cuanto más discrepantes son sus respectivas probabilidades marginales y menor cuanto más similares son. El efecto de prevalencia ocurre en presencia de una proporción global extrema de resultados para una categoría. Ambos efectos se han demostrado en los trabajos de Spitznagel y Hazer (1985), Feinstein y Cicchetti (1990), Byrt, Bishop y Carlin (1993), Agresti, Ghosh y Bini (1995), Lantz y Nebenzahl (1996) y Hoehler (2000), entre otros. Un conjunto de 4 casos paradójicos para tablas 2 × 2 se muestra en el cuadro 3. En los dos primeros casos, siendo p_0= .85, las proporciones de casos discrepantes prácticamente idénticas y las distribuciones marginales muy similares, κ_2 es menos de la mitad de κ_1 y lo mismo sucede con π pero no con σ. Este efecto diferencial se atribuye a que la prevalencia de casos positivos es en el segundo caso de 0.80. En los dos últimos casos, siendo p_0= .60, las probabilidades marginales del observador A iguales pero las probabilidades marginales del observador B discrepantes, κ_4 es el doble de κ_3. Este efecto diferencial se atribuye a la discrepancia que existe entre las probabilidades marginales del observador B. Los mismos problemas afectan también al coeficiente π, pero no a σ.

Cuadro 3
Conjunto de 4 casos paradójicos de Byrt y otros (1993)

		Observador B			
Caso 1		Sí	No	Total	Indice
Observador A	Sí	*0.40*	*0.09*	0.49	σ_1= .70
	No	*0.06*	*0.45*	0.51	π_1= .70
	Total	0.46	0.54	1.00	κ_1= .70
Caso 2		Sí	No	Total	
Observador A	Sí	*0.80*	*0.10*	0.90	σ_2= .70
	No	*0.05*	*0.05*	0.10	π_2= .32
	Total	0.85	0.15	1.00	κ_2= .32
Caso 3		Sí	No	Total	
Observador A	Sí	*0.45*	*0.15*	0.60	σ_3= .20
	No	*0.25*	*0.15*	0.40	π_3= .12
	Total	0.70	0.30	1.00	κ_3= .13
Caso 4		Sí	No	Total	
Observador A	Sí	*0.25*	*0.35*	0.60	σ_4= .20
	No	*0.05*	*0.35*	0.40	π_4= .19
	Total	0.30	0.70	1.00	κ_4= .26

Modelado estadístico del acuerdo entre observadores

Modelos loglineales

En lugar de describir el acuerdo mediante un índice, un enfoque alternativo desarrollado durante la década de los 80 (Tanner y Young, 1985a; Agresti,1992) consiste en analizar la estructura del acuerdo y desacuerdo existente en los datos con *modelos loglineales*. Dos rasgos cruciales de este enfoque son la posibilidad de probar el ajuste de los modelos y su capacidad de generalización a variables de respuesta ordinales (Tanner y Young, 1985a,b; Schuster y von Eye, 2001), al caso de más de dos observadores (Agresti, 1992) y a la inclusión de una o más covariantes (Graham, 1995).

Los modelos loglineales modelan el acuerdo observado en términos de componentes, tales como el acuerdo esperado por azar y el acuerdo no esperado por azar. Dado un conjunto de *N* objetos a clasificar por parte de dos observadores en *M* categorías, el *modelo de independencia*,

$$log\left(m_{ij}\right) = \lambda + \lambda_i^A + \lambda_j^B \qquad \text{(Ec. 9)}$$

donde m_{ij} es el valor esperado y λ_i^A y λ_j^B son efectos debidos a la categorización de los observadores A y B, es el modelo básico que representa el acuerdo esperado por azar y asume independencia estadística entre ambos observadores.

Parámetros adicionales pueden incorporarse al modelo básico de la Ec. 9 con la finalidad de probar hipótesis específicas relativas al acuerdo y desacuerdo, el más importante de los cuales es el conjunto de parámetros diagonales δ_{ij}. Puesto que el acuerdo entre observadores se concentra en las casillas de la diagonal principal, definiendo $\delta_{ij} = \delta_i I_i$ se obtiene el *modelo de cuasi-independencia* (QI)

$$log\left(m_{ij}\right) = \lambda + \lambda_i^A + \lambda_i^B + \delta_i I_i \qquad \text{(Ec. 10)}$$

donde I_i es una variable ficticia que es 1 cuando $i = j$ y 0 cuando $i \neq j$, y la transformación $\exp(\delta_i)$ representa el grado de acuerdo asociado a la categoría *i* (Guggenmoos-Holtzman y Vonk, 1998; von Eye y Mun, 2005). Un modelo más simple, que asume constancia del acuerdo sistemático entre categorías y se obtiene definiendo $\delta_{ij} = \delta I_i$, es el *modelo de cuasi-independencia constante* (QIC),

$$log\left(m_{ij}\right) = \lambda + \lambda_i^A + \lambda_i^B + \delta I_i \qquad \text{(Ec. 11)}$$

donde δ representa el acuerdo más allá del azar (independencia) que se asume constante para todas las categorías. Como consecuencia, las razones *odds* locales de los valores esperados m_{ij} se asumen iguales (Guggenmoos-Holtzman, 1996).

La transformación $\exp(\delta_{ij})$ se relaciona directamente con la capacidad de los observadores para distinguir entre las categorías de la tabla de acuerdo (Darroch y McCloud, 1986; Agresti, 2002).

Dado un modelo que asuma homogeneidad marginal, la razón de *odds* de que los observadores estén de acuerdo más que en desacuerdo sobre si un ítem debe asignarse a la categoría *i* en lugar de *j* es

$$\theta_{ij} = \frac{m_{ii}m_{ij}}{m_{ij}m_{ji}} = \exp\left(\delta_i + \delta_j\right)$$

donde m_{ij} son valores estimados del modelo para la fila *i* y columna *j* y δ_i y δ_j son los parámetros diagonales para las categorías *i* y *j*. θ_{ij} será mayor de 1 si la probabilidad de acuerdo de las casillas diagonales es mayor que la de las casillas no diagonales. El grado de *distinción entre categorías* se define entonces como

$$\pi_{ij} = 1 - \frac{1}{\theta_{ij}}$$

donde valores $\pi_{ij} \simeq 0$ se asocian con categorías indistinguibles, mientras que valores $\pi_{ij} \simeq 1$ indican categorías perfectamente distinguibles.

Utilizando los parámetros diagonales exp(δ_{ij}) se pueden definir nuevas medidas de acuerdo y redefinir también algunas de las medidas descriptivas tratadas anteriormente. Una medida de acuerdo basada en un modelo loglineal utiliza una definición de acuerdo corregida del azar que difiere del caso descriptivo y se formula de forma general mediante (Guggenmoos-Holzman y Vonk, 1998):

$$\lambda = \left[\sum_i \hat{p}_{ii} - \frac{\hat{p}_{ii}}{\exp(\delta_i)}\right] \qquad \text{(Ec. 12)}$$

donde $\hat{p}_{ii}$ son probabilidades esperadas del modelo y, para la *i*-ésima fila/columna,

$$p_{ei}^{\lambda} = \frac{\hat{p}_{ii}}{\exp(\delta_i)} \qquad \text{(Ec. 13)}$$

es la corrección debida al azar utilizada con modelos loglineales. Puesto que los parámetros diagonales pueden ser negativos, exp(δ_i) pueden ser menores de 1 y la escala resultante puede adoptar valores dentro del rango –1/1, la misma escala utilizada por las medidas descriptivas.

Martín y Femia (2004) definieron una medida de acuerdo basado en el modelo QI (Ec. 10), que sus autores entroncan con la tradición de los tests de elección múltiple y llaman *Delta* (Δ). El modelo tiene $(M-1)^2 - M$ grados de libertad residuales, y por tanto no es posible su aplicación a tablas de acuerdo 2 × 2. Para los datos del cuadro 2, el vector de parámetros es exp(δ_i)= [11.745, 1.394, 26.083], el coeficiente de acuerdo resultante es, aplicando Ec. 12, Δ= 0.567 y el ajuste del modelo es óptimo: $L^2(1)$= .18; *p*= .67 (véase cuadro 4). Su interpretación se simplifica notablemente utilizando exp(δ_i): el grado de acuerdo entre observadores es proporcionalmente mayor para la tercera categoría, seguida de la primera y es prácticamente nulo para la segunda. Dado el ajuste del modelo, prácticamente toda la asociación de la tabla de acuerdo se concentra en la diagonal principal.

Un modelo más parsimonioso se obtiene asumiendo un parámetro diagonal constante. El resultado es el modelo QIC (Ec. 11), que deja $(M-1)^2 - 1$ grados de libertad, puede utilizarse con tablas de acuerdo 2 × 2 y es la versión loglineal del índice α (Aickin, 1990). Para los datos del cuadro 2, el parámetro diagonal constante del modelo QIC es exp(δ_i) y la medida de acuerdo es λ^{α}= .620, pero el ajuste no es aceptable: $L^2(3)$= 10.13; *p*= .02 (véase cuadro 4). Los datos empíricos no son compatibles con la hipótesis de que el parámetro diagonal sea constante.

Asumiendo homogeneidad marginal entre ambos observadores se obtiene el modelo QIH, que libera también $(M-1)^2 - 1$ grados de libertad y puede también utilizarse con tablas de acuerdo 2 × 2. Para los datos del cuadro 2, el vector de parámetros es exp(δ_i)= [6.778, 1.040, 31.000], la medida de acuerdo es λ= .570 pero el ajuste tampoco es aceptable: $L^2(3)$= 22.59; *p*= .00 (véase cuadro 4). Nótese que en este modelo vuelve a comprobarse que en la categoría 3 es proporcionalmente mayor el grado de acuerdo mientras que en la segunda categoría es prácticamente nulo.

Mediante la Ec. 12 pueden también obtenerse las versiones loglineales de los coeficientes σ de Bennett y π de Scott. Para el coeficiente σ se requiere un modelo (QIU) que asuma nulidad de los efectos de A y B e incluya los parámetros δ_i (lo que implica *uniformidad marginal*) y para el coeficiente π el modelo (QICH) debe incorporar una restricción de igualdad entre ambos observadores y un parámetro diagonal constante (lo que implica *homogeneidad marginal*). Para los datos del cuadro 2, λ^{a}= .579 y λ^{π}= .570 (véase cuadro 4). Sin embargo, debido a las restricciones requeridas, ninguno de los dos modelos se ajusta aceptablemente (véase cuadro 4). Y, por otra parte, el coeficiente κ no puede obtenerse mediante un modelo loglineal debido a la naturaleza de las restricciones que requiere (véase Guggenmoos y Holzman, 1998).

Modelos con mezcla de distribuciones (mixture models)

Una generalización del enfoque anterior consiste en incluir una o más variables latentes no observables y asumir que los objetos que los observadores deben clasificar se extraen de una población que representa una mezcla de dos (o más) subpoblaciones finitas (*modelo mixture*). Cada subpoblación identifica un conglomerado de ítems homogéneos, por ejemplo, la subpoblación que representa acuerdo sistemático (X1), que afecta únicamente a las casillas de la diagonal principal de la tabla de acuerdo, y la subpoblación que representa acuerdo aleatorio y desacuerdo (X2), que afecta por

Cuadro 4
Modelos loglineales y medidas de acuerdo para los datos del cuadro 2

Modelo	Restricciones	Parámetros diagonales exp(δ_i)	Medida de acuerdo	Ajuste del modelo
QI (Delta)	Independencia A y B δ_i heterogéneo	[11.75,1.39,26.08]	Δ= .567	$L^2(1)$= .18; *P*= .67
QIC	Independencia A y B δ_i constante	[7.23]	λ^{α}= .620	$L^2(3)$= 10.13; *P*= .02
QI	Igualdad A y B δ_i heterogéneo	[6.78,1.04,31.00]	λ= .506	$L^2(3)$= 22.59; *P*= .00
QICH	Igualdad A y B δ_i constante	[4.83]	λ^{π}= .570	$L^2(5)$= 40.06; *P*= .00
QIU	Efecto nulo A y B δ_i heterogéneo	[7.96,3.39,4.04]	λ^{σ}= .579	$L^2(5)$= 43.05; *P*= .00

igual a todas las casillas de la tabla. La distribución conjunta resultante es una mezcla de una distribución que asume acuerdo perfecto entre observadores, con probabilidad μ, y una distribución que asume independencia, con probabilidad 1–μ (Agresti, 1989, 2002). Así, el modelo loglineal QI con una variable latente X y dos observadores (A y B) requiere 3 dimensiones y se representa mediante

$$log\left(m_{hij}\right) = \lambda_h^X + \lambda_i^A + \lambda_j^B + \xi_i I_i \qquad \text{(Ec. 14)}$$

donde, para cada clase latente h (para h= 1,2) de X, se asume independencia local entre ambos observadores. Los parámetros ξ_i de los modelos *mixture* se relacionan estrechamente con los parámetros δ_i de los modelos loglineales ya que (Guggenmoos-Holtzman, 1993),

$$\exp(\xi_i) = \exp(\delta_i) \pm 1 \qquad \text{(Ec. 15)}$$

Con modelos *mixture* las medidas de acuerdo son proporciones latentes para la subpoblación de acuerdo sistemático (μ). La diferencia esencial entre modelos *mixture* y loglineal es que en los primeros las medidas de acuerdo son proporciones (y por tanto se miden en escala 0–1), mientras que las medidas descriptivas y las basadas en modelos loglineales se miden en escala –1/1. El ajuste de los modelos y sus parámetros son similares al caso loglineal en el caso de parámetros exp(ξ_i)≥0, pero difieren en el caso de valores menores de cero. A diferencia del concepto de corrección del azar del enfoque descriptivo, que afecta a todos los casos de la tabla de acuerdo, los modelos *mixture* asumen que la corrección del azar afecta únicamente a la subpoblación de acuerdo/desacuerdo aleatorio.

Asumiendo un modelo QI, la ecuación que descompone las probabilidades esperadas para cada una de las casillas de la tabla de acuerdo (Schuster, 2002; Ato y otros, 2004) es

$$\hat{p}_{ij} = I_i \mu \phi_i + (1-\mu)\psi_i^A \psi_j^B \qquad \text{(Ec. 16)}$$

donde I_i es un indicador que selecciona los elementos diagonales, μ es la proporción de la clase latente para la subpoblación de acuerdo sistemático (en adelante, X1) y 1 – μ para la subpoblación de acuerdo/desacuerdo aleatorio (en adelante, X2), ϕ_i es la probabilidad marginal de X1 (que es igual para ambos observadores por tratarse de una tabla diagonal) y ψ_i^A y ψ_j^B son las probabilidades marginales de X2. El modelo resultante se llama también *modelo de observadores heterogéneos* (Schuster y Smith, 2002). La medida de acuerdo sistemático μ es equivalente a la Δ de Martín y Femia (2004) cuando los parámetros son positivos. Nótese que μ puede obtenerse también, utilizando las Ecs. 12 y 15, mediante

$$\mu = \sum_i \left[\hat{p}_{ii} - \frac{\hat{p}_{ii}}{\exp(\xi_i)+1}\right] \qquad \text{(Ec. 17)}$$

donde el denominador será siempre 1 en el caso de valores exp(ξ_i)= 0, en cuyo caso probabilidad observada y esperada por azar son iguales y el acuerdo será nulo.

Una medida de acuerdo para modelos *mixture* más restrictiva es el índice α propuesto por Aickin (1990) que se basa en el modelo loglineal QIC (Ec. 10). A diferencia de μ, el índice α puede aplicarse incluso a tablas de acuerdo 2 × 2, aunque el modelo resulta saturado. La ecuación que descompone las probabilidades esperadas es similar a la Ec. 16, pero los parámetros mantienen una constante de proporcionalidad, definida como la razón entre las probabilidades marginales latentes $\phi_i/(\psi_i^A\psi_j^B)$, que se asume igual para todas las categorías. Debido a esta restricción, se denomina también al modelo QIC como *modelo de probabilidad predictiva constante*.

Más restringido es el modelo que Schuster y Smith (2002) denominan *modelo de observadores homogéneos* (modelo QIH), cuyas probabilidades esperadas se obtienen (siendo $\psi_i^A = \psi_j^B$) mediante

$$\hat{p}_{ij} = I_i \alpha \phi_i + (1-\alpha)\psi_i^2 \qquad \text{(Ec. 18)}$$

Este modelo asume que las probabilidades latentes para la subpoblación de acuerdo aleatorio son iguales para ambos observadores (supuesto de homogeneidad marginal).

Y de forma similar a los modelos loglineales pueden definirse también modelos *mixture* para reproducir los coeficientes π y σ. El modelo para π (QICH) asume la existencia de una constante de proporcionalidad, pero a diferencia del modelo QIC las probabilidades condicionales para X2 se asumen iguales en ambos observadores (supuesto de homogeneidad marginal). El modelo para ϕ (QIU) asume que las probabilidades condicionales para X2 son iguales para todo i (supuesto de uniformidad marginal). Ambos modelos liberan un total de $M^2 - M - 1$ grados de libertad residuales, pero su ajuste no es aceptable (cuadro 5).

Finalmente, el coeficiente κ puede también definirse como modelo *mixture* (véase Agresti, 1989), siendo $\phi_i = \psi_i^A = \psi_j^B$, mediante

$$\hat{p}_{ij} = I_i \mu \phi_i + (1-\mu)\psi_i \qquad \text{(Ec. 18)}$$

donde simultáneamente se asume, además de constancia de los parámetros diagonales, igualdad entre observadores (homogeneidad marginal) e igualdad entre clases (homogeneidad de clases latentes). El modelo resultante (QIHX) libera también un total de $M^2 - M - 1$ grados de libertad residuales, pero el ajuste tampoco es aceptable.

El cuadro 5 presenta un resumen de las medidas de acuerdo definidas con modelos *mixture*. Los modelos se representan en un continuo desde el mínimo (QI) al máximo grado de restricción (QIHX), junto con sus respectivas probabilidades latentes, índices de acuerdo y bondad de ajuste. Por ejemplo, para el único modelo interpretable (modelo QI), aplicando la Ec. 16 puede obtenerse la siguiente descomposición de las probabilidades latentes para la casilla 11 en la suma de sus correspondientes probabilidades condicionales latentes:

$$.372 = (.567)(.600) + (.433)(.510)(.144) = .340 + .032$$

lo cual implica que más del 91% representa acuerdo sistemático (o sea, corresponde a la distribución que asume acuerdo perfecto entre observadores) y el restante 9% acuerdo aleatorio y desacuerdo (o sea, corresponde a la distribución que asume independencia entre observadores), mientras que para la casilla 22 la descomposición resulta:

$$.159 = (.567)(.079) + (.433)(.361)(.727) = .045 + .114$$

y, por tanto, el acuerdo sistemático representa poco más de un 28% y el acuerdo aleatorio y desacuerdo el restante 72%.

Nótese que para los datos del cuadro 2 la medida y el ajuste de los modelos *mixture* son similares a los del caso loglineal (cuadro 4), aunque la naturaleza de los modelos difiere sustancialmente en ambos casos. Como ilustración de este argumento, el cuadro 6 presenta las medidas descriptivas, los modelos loglineales y los modelos *mixture* para una tabla de acuerdo modificada donde respecto del cuadro 2 se han cambiado las frecuencias de las tres casillas diagonales por valores más bajos (en concreto, las casillas diagonales 11, 22 y 33 cambian sus frecuencias de 61, 26 y 31 por 5) con el objeto de forzar el carácter negativo de las medidas descriptivas y loglineales. A diferencia de las medidas representativas de ambos enfoques, los modelos *mixture* utilizan medidas que implican acuerdo nulo o no significativo, pero nunca acuerdo negativo, un resultado difícil de justificar desde perspectivas tanto metodológicas como sustantivas para una «medida de acuerdo».

La interpretación de los parámetros $\exp(\xi_i)$ y de las probabilidades latentes, marginales y condicionales permiten comprender cabalmente las pautas de acuerdo y desacuerdo y representan un antídoto contra la resistente tradición de emplear medidas descriptivas (y particularmente, el popular coeficiente κ) como medida universal de acuerdo entre observadores.

Cuadro 5
Modelos mixture y medidas de acuerdo para los datos del cuadro 2

Modelo	Restricciones	Probabilidades latentes	Medida de acuerdo	Bondad de ajuste
QI	Independencia A y B ζ_i heterogéneo	X1= [.600 .079, .321] AX2= [.51 .361 .129] BX2= [.144, .727, .129]	μ^{Δ}= .567	$L^2(1)$= .18; P= .67
QIC (Alpha)	Independencia A y B ζ_i constante	X1= [.518, .25, .232] AX2= [.633 .122 .247] BX2= [.215, .539, .247]	α= .620	$L^2(3)$= 10.13; P= .02
QIH	Homogeneidad marginal ζ_i heterogéneo	X1= [.627, .012, .361] X2= [.333 .556 .111]	μ^{λ}= .506	$L^2(3)$= 22.59; P= .00
QICH	Homogeneidad marginal ζ_i constante	X1= [.524, .264, .212] X2= [.426 .303 .271]	μ^{π}= .570	$L^2(5)$= 40.06; P= .00
QIU	Efecto nulo A y B ζ_i heterogéneo	X1= [.561, .193, .246] X2= [.333 .333 .333]	μ^{σ}= .579	$L^2(6)$= 43.05; P= .00
QIHX	Homogeneidad marginal Homogeneidad de clases latentes ζ_i constante	X1= [.482, .300, .218] X2= [.482, .300, .218]	μ^{κ}= .559	$L^2(5)$= 37.61; P= .00

Nota: X1 representa probabilidades latentes para la clase 1, que se asumen iguales para A y B; X2 representa probabilidades latentes para la clase 2.

Debido a su generalidad y a la riqueza interpretativa de sus parámetros es en el contexto de los modelos *mixture* donde pueden evaluarse óptimamente las medidas de acuerdo. En este sentido, los modelos loglineales representan un puente de unión entre enfoque descriptivo y enfoque *mixture*. Las medidas descriptivas clásicas representan situaciones muy restrictivas que solo obtienen un ajuste aceptable en el caso de que se cumplan las restricciones que asumen. En el caso del popular coeficiente κ, por ejemplo, los supuestos de constancia de los parámetros diagonales, homogeneidad marginal y homogeneidad de clases latentes representan restricciones que se cumplen en una limitada proporción de las situaciones de investigación aplicada en las que se valora el acuerdo entre observadores.

Software

En este trabajo, los modelos loglineales y *mixture* se estimaron utilizando estimación por máxima verosimilitud y se ajustaron con la ayuda del programa LEM (Vermunt, 1987). El flujo del programa utilizado para estimar los modelos citados puede solicitarse a la dirección de correo electrónico del primero de los autores.

Nota

Este trabajo ha sido financiado con fondos de un proyecto de investigación y desarrollo tecnológico concedido por el Ministerio de Educación y Ciencia (proyecto BSO 2002-02513).

Cuadro 6
Comparativa de los tres enfoques para evaluar el grado de acuerdo con datos del cuadro 2 modificados en las casillas diagonales ($n_{11} = n_{22} = n_{33} = 5$)

Modelo	Medidas descriptivas	Modelos loglineales	Modelos mixture
QI		Δ= -.165 δ_i= [.963,.268,4.207] $L^2(1)$= .18; P= .67	μ^{Δ}= .063 ζ_i= [.000,.000,3.207] $L^2(1)$= .18; P= .10
QIC		λ^{α}= -.035 δ_i= [.875] $L^2(3)$= 6.56; P= .09	α= .000 ζ_i= [.000] $L^2(3)$= 6.56; P= .09
QIH		λ= -.328 δ_i= [.556,.200,5.000] $L^2(3)$= 22.59; P= .00	μ= .066 ζ_i= [.000,.000,4.000] $L^2(3)$= 22.59; P= .00
QICH	$\hat{\pi}$= -.170	λ^{π}= -.182 δ_i= [.574] $L^2(5)$= 32.94; P= .00	μ^{π}= .000 ζ_i= [.000] $L^2(5)$= 32.94; P= .00
QIU	$\hat{\sigma}$= -.131	λ^{σ}= -.131 δ_i= [.652,.652,.652] $L^2(5)$= 43.05; P= .00	μ^{σ}= .000 ζ_i= [.000,.000,.000] $L^2(5)$= 43.05; P= .00
QIHX	$\hat{\kappa}$= -.026	No se puede estimar	μ^{κ}= .000 ζ_i= [.000] $L^2(5)$= 36.52; P= .00

Referencias

Agresti, A. (1989). An agreement model with kappa as parameter. *Statistics and Probability Letters, 7*, 271-273.

Agresti, A. (1992). Modelling patterns of agreement and disagreement. *Statistical Methods in Medical Research, 1*, 201-218.

Agresti, A. (2002). *Categorical data analysis*. 2nd edition. Hoboken, NJ: Wiley.

Agresti, A., Ghosh, A. y Bini, M. (1995). Raking kappa: describing potential impact of marginal distributions on measure of agreement. *Biometrical Journal, 37*, 811-820.

Aickin, M. (1990). Maximum likelihood estimation of agreement in the constant predictive probability model, and its relation to Cohen's kappa. *Biometrics, 46*, 293-302.

Ato, M., Benavente, A., Rabadán, R. y López, J.J. (2004). Modelos con mezcla de distribuciones para evaluar el acuerdo entre observadores. *Metodología de las Ciencias del Comportamiento, V. Especial 2004*, 47-54.

Bennet, E.M., Alpert, R. y Goldstein, A.C. (1954). Communications through limited response questioning. *Public Opinion Quarterly, 18*, 303-308.

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

Brennan, R.L. y Prediger, D. (1981). Coefficient kappa: somes uses, misuses and alternatives. *Educational and Psychological Measurement, 41*, 687-699.

Byrt, T., Bishop, J. y Carlin, J.B. (1993). Bias, prevalence and kappa. *Journal of Clinical Epidemiology, 46*, 423-429.

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

Cronbach, L.J., Gleser, G.C. y Rajaratnam, J. (1972). *The dependability of behavioral measurements*. New York, NY: Wiley.

Darroch, J.M. y McCloud, P.I. (1986). Category distinguishability and observer agreement. *Australian Journal of Statistics, 28*, 371-388.

Dillon, W.R. y Mullani, N. (1984). A probabilistic latent class model for assessing inter-judge reliability. *Multivariate Behavioral Research, 19*, 438-458.

Dunn, C. (1989). *Design and analysis of reliability studies: the statistical evaluation of measurement errors*. Cambridge, UK: Cambridge University Press.

Feinstein, A. y Cichetti, D. (1990). High agreement but low kappa: I. The problem of two paradoxes. *Journal of Clinical Epidemiology, 43*, 543-549.

Fleiss, J.L., Cohen, J. y Everitt, B.S. (1969). Large sample standard errors of kappa and weighted kappa. *Psychological Bulletin, 72*, 323-327.

Graham, P. (1995). Modelling covariante effects in observer agreement studies: the case of nominal agreement. *Statistics in Medicine, 14*, 299-310.

Guggenmoos-Holtzmann, I. (1993). How reliable are change-corrected measures of agreement. *Statistics in Medicine, 12*, 2.191-2.205.

Guggenmoos-Holtzmann, I. (1996). The meaning of kappa: probabilistic concepts of reliability and validity revisited. *Journal of Clinical Epidemiology, 49*, 775-782.

Guggenmoos-Holtzmann, I. y Vonk, R. (1998). Kappa-like indices of observer agreement viewed from a latent class perspective. *Statistics in Medicine, 17*, 797-812.

Hoehler, F.K. (2000). Bias and prevalence effects on kappa viewed in terms of sensitivity and specificity. *Journal of Clinical Epidemiology, 53*, 499-503.

Holley, W. y Guilford, J.P. (1964). A note on the G-index of agreement. *Educational and Psychological Measuerement, 24*, 749-753.

Hsu, L.M. y Field, R. (2003). Interrater agreement measures: comments on $kappa_n$, Cohen's kappa, Scott's π and Aickin's α. *Understanding Statistics, 2*, 205-219.

Janson, S. y Vegelius, J. (1979). On generalizations of the G-index and the phi coefficient to nominal scales. *Multivariate Behavioral Research, 14*, 255-269.

Lantz, C.A. y Nebenzahl, E. (1996). Behavior and interpretation of the κ statistics: resolution of the two paradoxes. *Journal of Clinical Epidemiology, 49*, 431-434.

Lin, L., Hedayat, A.S., Sinha, B. y Yang, M. (2002). Statistical methods in assessing agreement: models, issues and tools. *Journal of the American Statistical Association, 97*, 257-270.

Martín, A. y Femia, P. (2004). Delta: a new measure of agreement between two raters. *British Journal of Mathematical and Statistical Psychology, 57*, 1-19.

Maxwell, A.E. (1977). Coefficients of agreement between observers and their interpretation. *British Journal of Psychiatry, 116*, 651-655.

Schuster, C. (2002). A mixture model approach to indexing rater agreement. *British Journal of Mathematical and Statistical Psychology, 55*, 289-303.

Schuster, C. y von Eye, A. (2001). Models for ordinal agreement data. *Biometrical Journal, 43*, 795-808.

Schuster, C. y Smith, D.A. (2002). Indexing systematic rater agreement with a latent-class model. *Psychological Methods, 7*, 384-395.

Scott, W.A. (1955). Reliability of content analysis: the case of nominal scale coding. *Public Opinion Quarterly, 19*, 321-325.

Shoukri, M.M. (2004). *Measures of Interobserver Agreement*. Boca Raton, Fl. CRC Press.

Shrout, P.E, y Fleiss, J.L. (1973). Intraclass correlations: uses in assessing rater reliability. *Psychological Bulletin, 2*, 420-428.

Spitznagel, E.I. y Helzer, J.E. (1985). A proposed solution to the base rate problem in the kappa statistics. *Archives of General Psychiatry, 42*, 725-728.

Tanner, M.A. y Young, M.A. (1985a). Modeling agreement among raters. *Journal of the American Psychological Association, 80*, 175-180.

Tanner, M.A. y Young, M.A. (1985b). Modeling ordinal scale disagreement. *Psychological Bulletin, 98*, 408-415.

Vermunt, J.K. (1997). *LEM: a general program for the analysis of categorical data*. Tilburg: University of Tilburg.

Von Eye, A. y Mun, E.Y. (2005). *Analyzing Rater Agreement*. Mahwah, NJ: Lawrence Erlbaum Associates.

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