Psicothema 2008. Vol. 20, nº 4, pp. 918-923
www.psicothema.com

ISSN 0214 - 9915 CODEN PSOTEG

Un índice de sesgo entre observadores basado en modelos mixtura

Manuel Ato García, Juan José López García y Ana Benavente Reche
Universidad de Murcia

Los modelos mixtura son procedimientos apropiados para la valoración del acuerdo entre dos (o más) observadores que asumen que los objetos a clasificar se extraen de una población que constituye una mezcla de dos subpoblaciones finitas, la primera de las cuales representa acuerdo sistemático y la segunda acuerdo aleatorio y desacuerdo. Una generalización del modelo mixtura básico a cuatro subpoblaciones que representan dos variables latentes con dos clases cada una permite preservar su naturaleza (el ajuste del modelo y la subpoblación de acuerdo sistemático son iguales) y distinguir además una subpoblación para el acuerdo aleatorio y dos subpoblaciones para el desacuerdo (una para el triángulo superior y otra para el triángulo inferior de la tabla de contingencia). En este contexto es posible definir una medida de sesgo entre observadores basada en modelos mixtura similar al índice descriptivo propuesto por Ludbrook.

A mixture model-based rater bias index. Mixture models are outstanding procedures to evaluate rater agreement that assume that the objects to be classified by two observers are extracted from a population that is a mixture of two finite subpopulations, the first one representing systematic agreement and the second one random agreement and disagreement. A generalization of the basic mixture model to include four subpopulations representing two latent variables with two classes allows us to preserve its nature (the fit of the model and the systematic subpopulation are the same) and to distinguish a subpopulation for random agreement and two subpopulations for disagreement (one for the upper triangle and the other for the lower triangle of contingency table). In this context, it is possible to define a new rater bias measure based on a mixture model, which is similar to the descriptive index proposed by Ludbrook.

Dada una tabla de contingencia obtenida mediante la clasificación de un conjunto de objetos por dos observadores sobre una escala de respuesta categórica nominal u ordinal, el interés de muchos investigadores de Psicología y disciplinas afines se concentra en la evaluación del acuerdo existente entre los observadores y la fiabilidad de la medida utilizada. La mayoría se decanta por utilizar alguna medida descriptiva (particularmente, el índice κ de Cohen, 1960). Menos frecuente es el empleo de procedimientos alternativos fundamentados en la formulación y el ajuste de modelos estadísticos (Ato y otros, 2005), y concretamente los modelos loglineales, los modelos de clase latente y los modelos con mezcla de distribuciones o modelos mixtura ('mixture models'). Estos procedimientos superan algunos de los problemas de sesgo y prevalencia de los índices descriptivos que están bien documentados en la literatura (véase Brennan y Prediger, 1981; Byrt, Bishop y Carlin, 1993; Feinstein y Cichetti, 1990; Hoehler, 2000; Hsu y Field, 2003; y Lantz y Nebenzahl, 1996, entre otros), y permiten profundizar en cuestiones de detalle sobre la estructura del acuerdo y del desacuerdo que no es posible con medidas simples de tipo descriptivo. Una revisión comparativa de estos enfoques puede consultarse en Ato, Benavente y López (2006).

Sea cual sea el enfoque utilizado, en la práctica los investigadores limitan la evaluación del acuerdo entre observadores o bien a la utilización de un índice descriptivo (por ejemplo, el índice κ de Cohen, 1960) o más raramente al empleo de una medida basada en modelos (por ejemplo, la medida de acuerdo sistemático μ de Schuster, 2002, y la medida de acuerdo Δ de Martín y Femia, 2004). Ludbrook (2002, 2004) propuso considerar también un índice descriptivo (BI, bias index) para evaluar el desacuerdo diferencial de los observadores ante la clasificación de los objetos en las categorías de respuesta.

Este trabajo se propone complementar la evaluación del acuerdo con la utilización de una medida de acuerdo sistemático que depende del modelo tomado de una familia de modelos (por ejemplo, la familia de modelos de cuasi-independencia) y un índice del sesgo (ε) entre observadores basado en modelos mixtura. Para ello se precisa ampliar el número de variables latentes del modelo mixtura básico pero preservando al mismo tiempo su naturaleza, tanto en lo relativo al ajuste del modelo como a la magnitud de la medida de acuerdo sistemático. En la segunda sección de este trabajo se fundamenta la ampliación del número de variables latentes, partiendo de un modelo mixtura básico con una variable latente y dos clases, a un modelo con dos variables latentes y dos clases cada una. En la tercera sección se formula el índice de sesgo basado en modelos mixtura y se ilustra en la práctica con un ejemplo tomado de la literatura (Fennig y otros, 1994). En la última sección se

Fecha recepción: 29-5-08 • Fecha aceptación: 4-6-08
Correspondencia: Manuel Ato García
Facultad de Psicología
Universidad de Murcia
30080 Murcia (Spain)
E-mail: matogar@um.es

comentan algunas consideraciones relativas al empleo conjunto de ambas medidas.

Enfoques alternativos para analizar el acuerdo entre observadores

Hay tres enfoques para la evaluación del acuerdo entre dos (o más) observadores cuando la clasificación de un conjunto de objetos se realiza sobre una escala de medida categórica, nominal u ordinal, y los resultados se representan mediante una tabla de contingencia (Shoukri, 2004; von Eye y Mun, 2005; Ato, Benavente y López, 2006). Por orden temporal de aparición son: el enfoque descriptivo, el enfoque loglineal y el enfoque mixtura.

El *enfoque descriptivo* se desarrolló durante la década de los 50, utiliza fórmulas para la corrección del azar calculadas a partir de los marginales de la tabla de contingencia con el objeto de obtener una medida de acuerdo no contaminada por la respuesta aleatoria. Dentro de esta categoría destacan las medidas de acuerdo σ de Bennet, Alpert y Goldstein (1954), π de Scott (1955) y κ de Cohen (1960). A pesar de su popularidad, estos índices suponen algún tipo de restricción sobre las frecuencias de una tabla de contingencia, presentan problemas de sesgo y de prevalencia y, en general, no permiten comprender la naturaleza del acuerdo y desacuerdo (véase, no obstante, von Eye y von Eye, 2005).

El *enfoque loglineal* se desarrolló durante la década de los 80 (Tanner y Young, 1985ab; Agresti, 1992) y utiliza los modelos loglineales para analizar la estructura del acuerdo y desacuerdo. Varias familias de modelos loglineales son posibles, en particular, la familia de modelos de cuasi-independencia y de cuasi-simetría (Ato y López, 1996). La característica más importante de este enfoque es que la interpretación del acuerdo depende del ajuste del modelo loglineal propuesto para explicar los datos empíricos.

El *enfoque mixtura* es una generalización del enfoque loglineal que asume que los objetos que los observadores deben clasificar se extraen de una población que representa una mezcla de dos subpoblaciones finitas. Mediante la utilización de una variable latente no observable con dos clases, cada clase (o subpoblación) identifica un conglomerado de objetos homogéneos, por ejemplo, la subpoblación que representa el acuerdo sistemático (X1), representado por todos los elementos diagonales de la tabla de contingencia, y la subpoblación que representa el acuerdo aleatorio y el desacuerdo (X2), que afecta similarmente a todas las casillas de la tabla de contingencia (Agresti, 1989, 1992; Guggenmoos-Holtzman y Vonk, 1998; Schuster, 2002; Schuster y Smith, 2002).

Una dificultad del enfoque mixtura concierne a la interpretación de la segunda clase latente, que agrupa de forma conjunta tanto acuerdo aleatorio como desacuerdo y representa por ello una subpoblación de difícil identificación. No hay hasta ahora ninguna solución para subsanar este problema. En este trabajo se propone una solución simple que consiste en ampliar el número de clases latentes del modelo mixtura, convirtiéndolo en un modelo con cuatro clases latentes para el que resulta posible derivar al mismo tiempo una medida apropiada del sesgo entre observadores.

Utilizaremos como ilustración un ejemplo tomado del trabajo de Fennig y otros (1994), quienes estudiaron el acuerdo existente entre las unidades de servicio público y las unidades de investigación para el diagnóstico inicial de 223 pacientes psiquiátricos en una de cuatro categorías: esquizofrenia, trastorno bipolar, depresión y otros diagnósticos (tabla 1; Schuster y Smith, 2002).

Varias familias de modelos de acuerdo pueden utilizarse en un enfoque mixtura con una variable latente con dos clases. La familia de modelos de cuasi-independencia (QI), tratada por Schuster y Smith (2002), y Ato, Benavente y López (2006), y está integrada entre otros por los seis modelos siguientes: modelo básico (QI), modelo QI constante (QIC), modelo QI homogéneo (QIH), modelo QI constante y homogéneo (QICH), modelo QI uniforme (QIU) y modelo QI homogéneo en las clases latentes (QIHX). Cada uno de ellos se caracteriza por satisfacer un conjunto de restricciones y se identifica con medidas de acuerdo ya existentes en la literatura. Así, QI se corresponde con la medida de acuerdo Δ propuesta por Martin y Femia (2004), QIC con la medida de acuerdo propuesta por Tanner y Young (1985) con modelos loglineales y con la medida α propuesta por Aickin (1990), QIH se corresponde con el modelo de observadores homogéneos de Schuster y Smith (2002) y con la medida λ (Ato, Benavente y López, 2006), QICH es equivalente al índice descriptivo π de Scott (1955), QIU es similar al índice descriptivo σ de Bennet y otros (1954) y el modelo QIHX se corresponde con el índice descriptivo κ de Cohen (1960). El ajuste de tales modelos, con sus restricciones características y las probabilidades de las dos clases latentes (X1, para acuerdo sistemático y X2 para acuerdo aleatorio y desacuerdo) se obtuvo con el programa LEM (Vermunt, 1997) y se muestra en el cuadro 1.

Tabla 1
Frecuencias empíricas del ejemplo de Fennig y otros (1994)

A: diagnóstico de los servicios públicos	**B: diagnóstico de los centros de investigación**				
	Esquizofrenia	**Trastorno bipolar**	**Depresión**	**Otro diagnóstico**	**Total**
Esquizofrenia	40	6	4	15	65
Trastorno bipolar	4	25	1	5	35
Depresión	4	2	21	9	36
Otro diagnóstico	17	13	12	45	87
Total	65	46	38	74	223

Cuadro 1
Ajuste de la familia QI de modelos mixtura con dos clases latentes

Modelo	**Restricciones utilizadas**	**Probabilidades latentes**	**Ajuste**
QI (Delta)	Independencia A-B Diagonal heterogénea	X1: 0.368 X2: 0.632	$L^2(5)= 1.56$; P= .91*
QIC (Alpha)	Independencia A-B Diagonal homogénea	X1: 0.444 X2: 0.556	$L^2(8)= 18.35$; P= .02
QIH (Lambda)	Homogeneidad marginal Diagonal heterogénea	X1: 0.362 X2: 0.638	$L^2(8)= 6.32$; P= .61*
QICH (Pi)	Homogeneidad marginal Diagonal homogénea	X1: 0.440 X2: 0.560	$L^2(11)= 22.94$; P= .02
QIU (Sigma)	Efecto nulo de A y B Diagonal heterogénea	X1: 0.450 X2: 0.550	$L^2(11)= 42.30$; P= .00
QIHX (Kappa)	Homogeneidad marginal Homogeneidad clases latentes Diagonal homogénea	X1: 0.436 X2: 0.564	$L^2(11)= 15.52$; P= .16*

Nota: los asteriscos representan los modelos mejor ajustados

Descomposición de la clase latente que incluye acuerdo aleatorio y desacuerdo

De las dos clases latentes que se utilizan en un modelo mixtura, la interpretación de la primera (X1), que representa la proporción de acuerdo sistemático es directa e inambigua y no plantea problema alguno. La interpretación de la segunda clase (X2), que representa la proporción de acuerdo aleatorio y desacuerdo, es ambigua y apenas tiene utilidad.

Una solución apropiada sería descomponer la clase X2 del modelo mixtura básico para aislar los componentes de acuerdo aleatorio y de desacuerdo. Tres condiciones deberían cumplirse para realizar adecuadamente tal descomposición: primera, el resultado no debe afectar al ajuste de los modelos de la familia; segunda, no debe ser afectada la proporción de acuerdo sistemático de la clase X1; y tercera, la suma de las proporciones de acuerdo sistemático y aleatorio debe ser igual a la proporción de acuerdo observado. Para preservar el equilibrio de los pesos utilizados se requiere ampliar el modelo original incluyendo dos variables latentes con dos clases cada una y seleccionar los pesos cuidadosamente con el objeto de lograr una solución satisfactoria. La primera variable latente (X) evalúa el acuerdo (donde la primera clase X1 representa el acuerdo sistemático y la segunda clase X2 el acuerdo aleatorio) y la segunda variable latente (Y) evalúa el desacuerdo. Puesto que asumimos que el desacuerdo es aleatorio, se optó por definir las dos clases de Y ponderando la magnitud de las frecuencias en los triángulos superior (clase Y1) e inferior (clase Y2) con la finalidad de promover el desarrollo de una medida de sesgo. El cuadro 2 resume el ajuste de la familia de modelos QI para un modelo mixtura con dos variables latentes con dos clases cada una que se obtuvo con LEM.

Con esta ampliación a cuatro clases, la primera clase (combinación X1Y1) representa la proporción de acuerdo sistemático, en la que los observadores clasifican del mismo modo los objetos para los que no existe duda alguna. La segunda clase (combinación X1Y2) representa la proporción de acuerdo aleatorio en la que los observadores coinciden por azar en la clasificación de los objetos dudosos. La tercera clase (combinación X2Y1) representa la proporción de desacuerdo esperado en el triángulo superior y la cuarta clase (combinación X2Y2) representa la proporción de desacuerdo esperado en el triángulo inferior. Nótese, en comparación con el cuadro 1, que, en cada uno de los modelos de la familia QI, la clase latente de acuerdo sistemático (X1Y1) produce el mismo resultado que la clase X1 del cuadro 1 y que el ajuste es también exactamente el mismo que se obtuvo con la familia de modelos definida para una variable latente con dos clases, ya que se trata esencialmente del mismo modelo. Además, la suma de las proporciones de acuerdo sistemático y acuerdo aleatorio es igual a la probabilidad empírica observada de acuerdo, que para los datos del ejemplo es igual a la proporción de los elementos diagonales (40+25+21+45 / 223= 0.587). Del mismo modo, la suma de las proporciones de desacuerdo en el triángulo superior e inferior es igual a la probabilidad empírica observada de desacuerdo, que por definición es el complemento de la probabilidad de acuerdo observado, o sea, 1 - 0.587= 0.413. Obviamente, para cada uno de los modelos ajustados la suma de las combinaciones X1Y2, X2Y1 y X2Y2 es igual al acuerdo aleatorio y desacuerdo que se obtiene en el modelo mixtura básico con dos clases. En consecuencia, se preserva la naturaleza del modelo pero se amplía el número de clases para distinguir el acuerdo aleatorio del desacuerdo.

Cuadro 2
Ajuste de la familia QI de modelos mixtura con cuatro clases latentes

Modelo	Proporción de la clase latente X1Y1	Proporción de la clase latente X1Y2	Proporción de la clase latente X2Y1	Proporción de la clase latente X2Y2	Ajuste
QI	0.368	0.219	0.181	0.232	$L^2(5)$= 1.56; P= .91*
QIC	0.444	0.143	0.182	0.231	$L^2(8)$= 18.35; P= .02
QIH	0.362	0.225	0.206	0.206	$L^2(8)$= 6.32; P= .61*
QICH	0.440	0.147	0.206	0.206	$L^2(11)$= 22.94; P= .02
QIU	0.450	0.138	0.206	0.206	$L^2(11)$= 42.30; P= .00
QIHX	0.436	0.158	0.203	0.203	$L^2(11)$= 15.52; P= .16*

Nota: los asteriscos representan los modelos ajustados

Un índice de sesgo basado en modelos mixtura

La definición del desacuerdo en el contexto de dos variables latentes, distinguiendo para la segunda variable latente una clase con los elementos del triángulo superior y otra clase con los elementos del triángulo inferior de la tabla de contingencia permite además desarrollar una medida de sesgo basada en modelos mixtura.

En una profunda revisión de la literatura hemos encontrado abundantes antecedentes sobre el sesgo sistemático entre observadores en modelos de acuerdo con datos cuantitativos (véase Ludbrook, 2002), pero son escasos los antecedentes que existen para datos categóricos (Ludbrook, 2004; Benavente, Ato y López, 2006). A destacar dos índices de sesgo desarrollados para tablas de contingencia 2×2, el índice de sesgo (BI, 'bias index') de Byrt, Bishop y Carlin (1993), el índice de simetría del desacuerdo (SDI, 'symmetry of disagreement index') de Lanz y Nebenzahl (1996) y la generalización del índice BI propuesta por Ludbrook (2004) para tablas de contingencia de cualquier dimensión. Todos los índices propuestos son de tipo descriptivo, y se obtienen aplicando una fórmula simple sobre las frecuencias de la tabla de contingencia.

Ludbrook (2004, pp. 114-115) propuso una generalización del índice BI mediante la diferencia en valores absolutos entre la suma de las frecuencias del triángulo superior (y la suma de las frecuencias del triángulo inferior (ΣTI) dividida por el N total, de la tabla de contingencia,

$$\mathrm{BI} = \frac{\left|\sum TS - \sum TI\right|}{N}$$

En el ejemplo de la tabla 1, siendo ΣTS= (6 + 4 + 15 + 1 + 5 + 9)= 40 y ΣTI= (4 + 4 + 2 + 17 + 13 + 12)= 52, el índice es BI= | 40 - 52 | / 223= 0.054. BI varía teóricamente entre los límites 0 (ausencia de sesgo) y 1 (sesgo extremo). Aunque se obtiene aplicando una fórmula muy sencilla, plantea, sin embargo, dos serios inconvenientes: primero, se obtiene directamente a partir de los valores empíricos de la tabla de contingencia, y, por tanto, es un índice descriptivo, y segundo, al considerar en la fórmula la suma de los valores triangulares superior e inferior, sea cual sea la posición que ocupan tales valores, no se tiene en cuenta la existencia de simetría o asimetría en la tabla de contingencia.

Para los modelos de la familia QI que no tienen restricciones de homogeneidad o uniformidad marginal (en concreto, para el modelo QI, que se asocia con la medida de acuerdo Δ de Martín y Femia, 2006; y QIC, que se asocia con la medida de acuerdo α de Aickin, 1990) puede también definirse un índice de sesgo de similares características a la generalización del índice BI propuesta por Ludbrook (2004), pero obtenido mediante el ajuste de modelos mixtura. El índice de sesgo (ε) que proponemos aquí puede obtenerse, una vez conocidas las proporciones de las cuatro clases del modelo mixtura ampliado a dos variables latentes, calculando también la diferencia en valores absolutos entre las proporciones de las dos clases latentes que evalúan el desacuerdo (clases X2Y1 y X2Y2).

El cuadro 3 presenta cuatro tablas de contingencia en las que se ajustan los modelos QI y QIC de la familia QI sin alterar el total muestral (N= 223), con el objeto de valorar adecuadamente las diferencias entre ambos índices. La primera tabla corresponde a los datos empíricos de la tabla 1 y representa un caso de sesgo moderado. La segunda tabla es una modificación de la primera obtenida con los mismos datos, pero la práctica totalidad de las frecuencias del triángulo superior se han trasladado al triángulo inferior con el objeto de representar una situación con un alto grado de sesgo. La tercera tabla utiliza los mismos datos empíricos pero calculando un promedio entre los elementos equivalentes del triángulo superior e inferior de la tabla para representar una situación donde se cumpla simetría y homogeneidad marginal con sesgo nulo. La cuarta tabla es similar a la anterior pero se ha realizado una permutación de los elementos del triángulo inferior para no hacerlos coincidir simétricamente con los elementos del triángulo superior, forzando así que no se cumpla simetría y homogeneidad marginal con presencia de sesgo no nulo.

Varias ventajas pueden derivarse de la utilización del índice de sesgo ε que proponemos aquí, en comparación con el índice descriptivo BI de Ludbrook (2004). En primer lugar, a diferencia del índice BI, ε es un índice basado en un modelo mixtura con dos variables latentes obtenido a partir de los valores esperados y, por tanto, su interpretación depende del ajuste del modelo subyacente. Desde esta perspectiva, si para los datos de una tabla de contingencia no se ajustara ninguno de los dos modelos QI o QIC de la familia QI, no podría de hecho definirse una medida de sesgo apropiada. En los datos del ejemplo de la tabla 1, el modelo mixtura en el que se basa la formulación del coeficiente κ de Cohen (modelo QIHX) se ajusta aceptablemente y, por tanto, la interpretación del valor $\hat{\kappa}$= .432 es totalmente apropiada en el contexto del modelado estadístico (el modelo mixtura estima un valor concreto de $\hat{\kappa}$= .436; véase cuadro 1). Como se observa también en el cuadro 1, el modelo QIHX debe satisfacer los restrictivos supuestos de homogeneidad marginal y homogeneidad de las clases latentes y postula valores homogéneos para la diagonal principal. Sin embargo, si el modelo QIHX no se ajustara aceptablemente, como es usual, su interpretación no sería apropiada. Obviamente, para los modelos que requieren satisfacer homogeneidad marginal no tiene sentido definir medidas de sesgo. Por esta razón, el cálculo de las medidas de sesgo se realiza exclusivamente con los modelos QI y QIC de la familia de modelos de cuasi-independencia.

En segundo lugar, a diferencia del índice BI, que se obtiene mediante la diferencia absoluta de las proporciones empíricas de los triángulos superior e inferior de la tabla de contingencia, el índice ε se obtiene mediante la diferencia absoluta de las proporciones esperadas de las dos clases latentes que valoran el desacuerdo, y, como consecuencia, representan magnitudes corregidas de efectos no controlados. En general, el índice ε produce valores de sesgo más pequeños que el índice BI. Así, en la primera tabla de contingencia del cuadro 3 se muestran los datos del ejemplo de la tabla 1, que presenta un grado leve de sesgo, donde BI= 0.054 mientras que ε_1= 0.051 para el modelo QI y ε_2= 0.049 para el modelo QIC (aunque este modelo no es interpretable porque no obtiene un ajuste aceptable). En la segunda tabla se muestra un caso con grado ex-

Cuadro 3
Análisis comparativo de cuatro diferentes situaciones de sesgo

Datos empíricos				Modelo	Proporciones de clase latente	Índice BI	Índice ε	Ajuste del modelo
40	6	4	15	QI	11: 0.368; 12: 0.219	0.054	0.051	$L^2(5)$= 1.56; P= .91*
4	25	1	5		21: 0.181; 22: 0.232			
5	2	21	9	QIC	11: 0.444; 12: 0.143	0.054	0.049	$L^2(8)$= 18.35; P= .02
17	13	12	45		21: 0.182; 22: 0.231			
40	1	0	0	QI	11: 0.543; 12: 0.045	0.386	0.367	$L^2(5)$= 7.33; P= .20*
9	25	1	0		21: 0.023; 22: 0.390			
8	2	21	1	QIC	11: 0.531; 12: 0.057	0.386	0.350	$L^2(8)$= 11.75; P= .16*
32	18	20	45		21: 0.031; 22: 0.381			
40	5	5	16	QI	11: 0.372; 12: 0.215	0.000	0.000	$L^2(5)$= 2.49; P= .83*
5	25	1	9		21: 0.206; 22: 0.206			
5	1	21	10	QIC	11: 0.440; 12: 0.147	0.000	0.000	$L^2(8)$= 18.47; P= .00
16	9	10	45		21: 0.206; 22: 0.206			
40	5	5	16	QI	11: 0.466; 12: 0.122	0.000	0.043	$L^2(5)$= 9.56; P= .10*
9	25	1	9		21: 0.228; 22: 0.185			
10	16	21	10	QIC	11: 0.466; 12: 0.122	0.000	0.050	$L^2(8)$= 10.21; P= .14*
5	5	1	45		21: 0.231; 22: 0.181			

Nota: los asteriscos representan, para cada tabla de contingencia, los modelos bien ajustados

tremo de sesgo, donde BI= 0.386 mientras que ε_1= 0.367 para el modelo QI y ε_2= 0.350 para el modelo QIC. Ambos modelos son interpretables porque el ajuste que se obtiene es aceptable.

En tercer lugar, una característica indeseable del índice BI es que tanto el acuerdo como el desacuerdo (y por ende el índice de sesgo) es invariante ante una permutación de los elementos dentro de su triángulo (superior o inferior). En cambio, el índice ε no es invariante ante una permutación de los elementos, que de hecho puede cambiar tanto la proporción de acuerdo sistemático como el índice de sesgo. En la tercera tabla de contingencia del cuadro 3, que se obtiene forzando la igualdad de los triángulos superior e inferior para representar simetría y homogeneidad marginal perfecta, el índice BI es cero y ε_1 y ε_2 para los diferentes modelos es también cero, pero la proporción de acuerdo sistemático es 0.372 para el modelo QI y 0.440 para el modelo QIC (aunque este modelo tampoco es interpretable). Nótese además que la suma de las proporciones correspondientes a las clases X1Y1 y X1Y2 es en ambos casos igual a P= 0.587). Por el contrario, en la cuarta tabla de contingencia, que es una simple permutación de los datos de la tabla anterior que afecta únicamente a las frecuencias del triángulo inferior, el índice BI es también cero, mientras que ε_1= 0.043 para el modelo QI y ε_2= 0.050 para el modelo QIC, y ambos modelos son interpretables. Nótese además que en los dos modelos la proporción de acuerdo sistemático es igual a 0.466 debido al equilibrio obtenido en la permutación de los elementos.

Conclusiones

En este trabajo se propone un índice de sesgo para valorar el desacuerdo entre observadores en el contexto de los modelos mixtura. Además de la medida de acuerdo sistemático, que evalúa la proporción del acuerdo observado corregido del azar, y ha sido propuesta anteriormente (en particular, Agresti, 1989, Schuster, 2002; Schuster y Smith, 2004; Martín y Femia, 2004; véase una revisión de estos trabajos en Ato, Benavente y López, 2006), se define un índice de sesgo ε, que evalúa la simetría de la respuesta de ambos observadores a cada una de las categorías de una tabla de contingencia, y se propone aquí como alternativa al índice de sesgo BI propuesto por Ludbrook (2004).

La utilización conjunta de un índice de acuerdo sistemático y un índice de sesgo entre observadores tiene gran interés para analizar con detalle la estructura del acuerdo y el desacuerdo, en particular si se articula en el contexto de las cuatro proporciones latentes que el ajuste del modelo mixture ampliado a dos variables latentes proporciona, ya que en general la presencia de un alto grado de sesgo puede afectar a la proporción del acuerdo sistemático en menoscabo de la proporción de acuerdo aleatorio. A título de ejemplo, la situación que presenta máximo grado de sesgo (segunda tabla de contingencia del cuadro 3) produce con el modelo QI una medida de acuerdo sistemático de μ= 0.543 y un índice de sesgo de ε_1= 0.386, lo que permite identificar con claridad que se concede importancia mínima al acuerdo aleatorio dando máxima relevancia al acuerdo sistemático en presencia de un grado extremo de sesgo (la suma de ambas medidas es 0.929, cercano al máximo posible). Por el contrario, la situación de los datos originales (primera tabla de contingencia del cuadro 3) produce con el modelo QI una medida de acuerdo sistemático de μ= 0.368 junto con un índice de sesgo de ε_1= 0.051.

Por otra parte, la ampliación de los modelos de la familia QI a dos variables latentes no plantea dificultades técnicas en lo relativo a la estimación por máxima verosimilitud. Todos los modelos se han ajustado con el programa LEM (Vermunt, 1997). Se requiere no obstante una cuidadosa selección de los pesos a utilizar para distinguir el acuerdo sistemático del acuerdo aleatorio y el triángulo superior del triángulo inferior de una tabla de contingencia. El flujo del programa utilizado puede solicitarse por correo electrónico al primero de los autores.

Agradecimientos

Esta investigación se ha financiado con fondos del proyecto de investigación del MEC SEJ2006-09025/PSIC.

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.

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., y López, J.J. (1996). *Análisis estadístico para datos categóricos*. Madrid: Síntesis.

Ato, M., Benavente, A., y López, J.J. (2006). Análisis comparativo de tres enfoques para evaluar el acuerdo entre observadores. *Psicothema, 18*(3), 638-645.

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.

Ato, M., Losilla, J.M., Navarro, J.B., Palmer, A., y Rodrigo, M.F. (2005). *Modelo lineal generalizado*. Girona: EAP, S.L.

Benavente, A., Ato, M., y López, J.J. (2006). Procedimientos para detectar y medir el sesgo entre observadores. *Anales de Psicología, 22*(1), 161-167.

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

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.

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

Fennig, S., Craig, T.J., Tanenberg-Karant, M., y Bromet, E.J. (1994). Comparison of facility and research diagnoses in first-admission psychotic patients. *American Journal of Psychiatry, 151,* 1423-1429.

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.

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

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

Ludbrook, J. (2002). Statistical techniques for comparing measurers and methos of measurement: A critical review. *Clinical and Experimental Pharmacology and Physiology, 19,* 527-536.

Ludbrook, J. (2004). Detecting systematic bias between two raters. *Clinical and Experimental Pharmacology and Physiology, 31,* 113-115.

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.

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

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.

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.

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