Modelización de fenómenos del Medio Ambiente, Salud y Desarrollo Humano: estudios medio ambientales. *pp*. *125-144* García, Bouza y Covarrubias Eds. Universidades de Guerrero (Mex) y de La Habana (Cuba) ISBN: 978-607-7760-60-3

EL MODELO DELTA PARA EVALUAR EL GRADO DE ACUERDO ENTRE DOS OBSERVADORES

Pedro Femia Marzo[1]

Antonio Martín Andrés[2]

María Álvarez Hernández[3]

[1,2,3] *Bioestadística. Facultad de Medicina. Universidad de Granada. 18071 Granada (Spain).*

E-mail: [1]pfemia@ugr.es, [2] amartina@ugr.es, [3]mariaalvarez@ugr.es

Resumen

Este trabajo revisa el modelo Delta que, desarrollado originalmente por Martín y Femia (2004), permite evaluar el nivel de acuerdo clasificatorio (escala nominal) entre dos observadores. El modelo fue propuesto como una alternativa al clásico coeficiente Kappa, pues el mismo no siempre caracteriza bien el nivel de acuerdo (especialmente si las distribuciones marginales son muy asimétricas). Además de la medida de acuerdo global, el modelo propuesto proporciona medidas de acuerdo parcial, lo que permite realizar un análisis mas detallado para cada categoría.

Abstract

This article is a review of the Delta model, originally proposed by Martin and Femia (2004) to assess the degree of agreement between the nominal responses of two raters. The model was proposed as an alternative to the classical measure of agreement kappa, which performs poorly specially when the marginal distributions are very assymetric. In addition to a measure of overall agreement, the proposed model provides measures of partial agreement, which allows a more detailed analysis for each category.

1. Introducción

Cuando un observador debe clasificar un objeto como perteneciente o no a una categoría dada, el azar puede estar presente en su decisión dependiendo de la intensidad con la que haya sido reconocido el objeto. Cabe esperar que si el objeto es perfectamente conocido, entonces sea clasificado sin error. Sin embargo, en la medida en que haya incertidumbre, en la decisión intervendrá el azar, siendo su influencia

máxima cuando el objeto no sea reconocido en absoluto por el observador (que entonces lo clasificará de forma puramente arbitraria). Consecuentemente, dos observadores que actúen clasificando objetos simplemente al azar podrán coincidir, por azar, en alguna de sus clasificaciones, dando la apariencia de que presentan un cierto grado de acuerdo.

Tabla 1. Presentación de los datos en un problema de acuerdo nominal. En el interior de la tabla aparecen las frecuencias observadas cuando dos observadores R (en filas) y C (en columnas) clasifican *n* objetos en *K* categorías.

		Respuesta del observador C					
		C_1	…	C_j	…	C_K	*Total*
Respuesta del observador R	R_1	x_{11}	…	x_{1j}	…	x_{1K}	r_1
	⋮	⋮		⋮		⋮	⋮
	R_i	x_{i1}	…	x_{ij}	…	x_{iK}	r_j
	⋮	⋮		⋮		⋮	⋮
	R_K	x_{K1}	…	x_{Kj}	…	x_{KK}	r_K
	Total	c_1	…	c_j	…	c_K	n

El problema de determinar hasta qué punto dos observadores están realmente de acuerdo (no por azar) al clasificar una serie de objetos en un conjunto predefinido de categorías, se presenta prácticamente en la vertiente aplicada de cualquier disciplina científica, siendo la medida tradicionalmente utilizada para ello el coeficiente *Kappa* de Cohen (Cohen, 1960). Por ejemplo, en el ámbito de esta monografía, Monserud y Leemans (1994) consideraron *Kappa* para comparar mapas de vegetación con el fin de determinar el impacto del cambio climático. Otros estudios han considerado también este índice para evaluar la precisión de modelos predictivos sobre la distribución de especies (e.g. Allouche et al., 2006; Freeman et al., 2008, Hanspach, 2010), o para calibrar modelos explicativos de la interpretación cartográfica (Luoto et al., 2010; van Vliet et al., 2011). Del mismo modo que ha ocurrido con muchas otras medidas estadísticas, la formulación de *Kappa* se hizo desde el ámbito de la Psicología, constituyendo actualmente un índice clásico en estudios de validez y de fiabilidad (eg. Weiner *et al*., 2003; Shrout, 1998). Pero este tipo de estudios es de trascendental importancia en otros contextos, como la Epidemiología (Kottner *et al.,* 2011) o la Medicina (eg. McGinn *et al.,* 2008; Fisher et al., 2011), en donde el diagnóstico precoz

de determinadas enfermedades es fundamental para poder combatirlas; por ejemplo, este es el caso de enfermedades de etiología parasitaria o el cáncer. Aquí es del máximo interés desarrollar nuevos métodos de detección que sean rápidos y fáciles de aplicar al tiempo que económicos. *Kappa* es tradicionalmente utilizado como indicador de fiabilidad global de estos nuevos métodos. Por ejemplo, Buppan *et al.* (2010) lo consideran en la evaluación de un método de detección de malaria, Wastling *et al*. (2010) en el diagnóstico de la tripanosomiasis, y es muy abundante la literatura que hace uso de *Kappa* para evaluar métodos de detección del cáncer (eg. Baines et al 1990; Linsell e*t al*., 2010; Webster et al. 2010).

Una de las ventajas de *Kappa* es la sencillez de sus cálculos. Si se presentan los datos tal y como aparecen en la Tabla 1, el índice observado de acuerdos es $\mathrm{I_o} = \sum x_{ii}$, siendo de esperar que algunos de ellos se deban al azar. Si los dos observadores R y C actúan de forma independiente, las frecuencias esperadas en cada casilla bajo el modelo de independencia son las clásicas $r_i c_j / n$, de manera que el número esperado de acuerdos viene dado por $\mathrm{I_e} = \sum r_i c_i / n$. El exceso de acuerdos respecto a los obtenidos por azar será entonces $\mathrm{I_o}$–$\mathrm{I_e}$. Considerando que el máximo posible es n–$\mathrm{I_e}$, el cociente entre ambas cantidades es el índice *Kappa*:

$$\hat{\kappa} = (\mathrm{I_o} - \mathrm{I_e}) / (n - \mathrm{I_e}), \qquad (1)$$

índice que caracteriza el exceso de acuerdos sobre el azar de entre el máximo posible de acuerdos no debidos al azar (Fleiss, 1975; Martín y Femia, 2004).

Su facilidad de cálculo y el hecho de no depender de ningún modelo probabilístico, hacen que *Kappa* sea la medida más utilizada para evaluar el acuerdo clasificatorio entre observadores cuando se considera una escala nominal. Sin embargo, también son muchos los autores que han criticado a este índice, siendo numerosos los argumentos según los cuáles *Kappa* puede resultar una medida de acuerdo muy deficiente. En Martín y Femia (2004) puede encontrarse una revisión exhaustiva de tales argumentos (véanse las referencias allí citadas). De ellos, quizá el más grave sea el hecho de que $\hat{\kappa}$ se ve excesivamente influido por las distribuciones marginales de los observadores, de manera que no logra el objetivo de medir el acuerdo entre los mismos cuando dichas distribuciones marginales son muy asimétricas.

Por otra parte, a menudo interesa profundizar en el análisis y estudiar cómo se da el acuerdo entre los observadores en cada categoría, pero *Kappa* es un índice global

(ómnibus) que no permite caracterizar los acuerdos parciales. Una estrategia habitual para estudiar el acuerdo en una clase dada es colapsar el resto de las clases y analizar la tabla 2×2 resultante (cf. Martín y Femia, 2005). Otras estrategias pueden verse en Roberts (2008), que extiende el concepto de *Kappa* para estudiar la heterogeneidad en el patrón de respuesta.

El objetivo de este artículo es hacer una revisión del modelo *Delta*, que fue propuesto originalmente por Martín y Femia (2004) para evaluar el grado de acuerdo entre dos observadores cuando se considera una escala nominal. Dicho modelo representa una alternativa preferible al índice *Kappa* –ya que no presenta sus limitaciones– y también preferible a otros modelos de acuerdo, tal y como ha puesto de manifiesto recientemente Ato (2011). Adicionalmente, el modelo *Delta* permite definir medidas de acuerdo parcial también corregidas por azar (Martín y Femia, 2005, 2008). Tales medidas constituyen una alternativa muy conveniente a algunos índices clásicos en los que no se considera dicha corrección; este es el caso de la sensibilidad, la especificidad y de los valores predictivos de un método de diagnóstico clínico.

Desde su formulación, *Delta* se ha utilizado como medida de fiabilidad por ejemplo en el ámbito de la Epidemiología (Dubois *et al*., 2007; Baranek, 2007, Guay *et al*., 2010), en el diagnóstico clínico (Fanshawe *et al*., 2008), y también en el ámbito de las Ciencias Ambientales: Ellis y Wang (2006) lo han utilizado en el análisis de mapas ecológicos y Prinzing *et al*. (2005) en el estudio de la distribución de especies invasoras en un ecosistema.

2. Modelo Delta de acuerdo nominal

2.1. Consideraciones previas

La formulación de *Delta* parte de la consideración previa del papel que juegan los observadores, así como del tipo del muestreo utilizado (cf. Martín y Femia, 2004, 2005). Si se resumen los datos tal y como se han presentado en la Tabla 1, R y C pueden representar al mismo observador en dos situaciones distintas o bien a dos observadores diferentes. Cuando se trata del mismo observador, una medida de acuerdo será una medida de *consistencia* que permite evaluar la *fiabilidad interna* de tal observador (se trataría de un *test-retest*). Cuando R y C son observadores distintos, uno de ellos –por ejemplo R– puede ser un estándar (en cuyo caso el acuerdo de C con R es una medida de *conformidad* que permite caracterizar la *validez* de C respecto al estándar R) o bien

puede no serlo (en cuyo caso el acuerdo entre C y R es una cuestión de *consistencia* que permite caracterizar la *fiabilidad entre observadores*). En adelante, para simplificar la exposición y mantener el criterio original del desarrollo del modelo *Delta*, asumiremos que si uno de los observadores es un estándar entonces se trata de aquel que aparece por filas al resumir los datos como en la Tabla 1, es decir, del observador R (un supuesto que no implica pérdida de generalidad).

Por otra parte, los datos de la Tabla 1 pueden obtenerse de acuerdo a dos tipos de muestreo. El muestreo es de *tipo I* cuando se dispone de un conjunto de *n* objetos cada uno de los cuales pertenece a una de las *K* categorías posibles, de manera que los dos observadores deben clasificarlos. En esta situación R puede ser un estándar o no. El muestreo de *tipo II* es aquel en el que se establece de antemano el número de objetos perteneciente a cada clase y un observador debe clasificarlos. Lo natural en este tipo de muestreo es que uno de los observadores –R– actúe como estándar y se trate de evaluar la conformidad del otro observador –C– con él, de manera que son los marginales r_i los que están fijados de antemano.

2.2. Formulación del modelo

Comencemos considerando la situación, intuitivamente mas sencilla, en que R es un estándar y los datos se obtienen bajo un muestreo de tipo II (en cuyo caso los totales $r_1,\ldots,r_K$ están fijados de antemano). La hipótesis subyacente al modelo *Delta* de acuerdo nominal es que el observador C reconoce un objeto de la categoría *i* (*i*=1,…, *K*) con una determinada intensidad que indicaremos por Δ_i (que va a ser característica de dicha categoría). Cuando C se enfrenta a un objeto de este tipo, si lo reconoce lo clasifica (correctamente) como tal, y si no lo clasifica al azar. El modelo supone que la probabilidad de clasificación al azar no es necesariamente uniforme, de manera que C tiene tendencia a elegir la categoría *j* con una cierta probabilidad π_j. De este modo, las probabilidades de clasificación propuestas por el modelo *Delta* vienen dadas por

$$\begin{aligned} p_{ii} &= \Pr(\mathrm{C}_i \mid \mathrm{R}_i) = \Delta_i + (1-\Delta_i)\pi_i \\ p_{ij} &= \Pr(\mathrm{C}_j \mid \mathrm{R}_i) = (1-\Delta_i)\pi_j \quad (\text{para } i \neq j). \end{aligned} \tag{2}$$

Obviamente, por tratarse de probabilidades, $0 \le \pi_j \le +1$ con $\sum \pi_j = +1$. Por otra parte $\Delta_i \le +1$ $(\forall i)$, pudiendo ser Δ_i negativo cuando un objeto se confunde con otro. En cualquier caso, $0 \le p_{ij} \le +1$ $(\forall i, j)$ con $\sum_j p_{ij} = +1$ $\forall i$, de manera que la matriz

cuadrada de elementos p_{ij} constituye una matriz estocástica. Obsérvese que, bajo el muestreo II, cada fila de datos de la Tabla 1 proviene de una distribución multinomial de tamaño r_i con parámetros $(p_{i1},\ldots,p_{iK})$, y que se hablará en términos de *modelo II* para evaluar la conformidad de C con R.

Del modelo propuesto se desprenden dos medidas de acuerdo, una parcial y otra global. Como Δ_i es la intensidad de reconocimiento de C para el objeto *i*, una medida de *conformidad global* de C respecto a R vendrá dada por la media ponderada $\Sigma\omega_i\Delta_i/\Sigma\omega_i$, en donde el factor de ponderación ω_i es la importancia relativa que se le asigna a cada acuerdo de C con R en la clase *i*. Si dicha importancia es el número total de objetos de la clase *i*, entonces la medida de acuerdo total será

$$\Delta_{\mathrm{II}} = \sum r_i\Delta_i/n\,. \tag{3}$$

En ella se hace explícito el hecho de que el modelo subyacente es el II. Esta medida global está construida como la suma de los acuerdos parciales en cada clase, que serán

$$\mathcal{A}_i = r_i\Delta_i/n\,. \tag{4}$$

Consideremos ahora la situación bajo un muestreo del tipo I. Ahora hay una sola muestra de tamaño *n* y tanto los marginales c_j como r_i son variables aleatorias. Condicionando en los marginales por filas, el modelo de respuesta (2) debe ser completado con

$$\Pr\left(\mathrm{R}_i \cap \mathrm{C}_j\right) = \Pr\left(\mathrm{R}_i\right) \times \Pr\left(\mathrm{C}_j|\mathrm{R}_i\right) = q_i p_{ij} = q_{ij}\,,\quad 0 \le q_i \le +1\,, \tag{5}$$

en donde q_i alude a la distribución marginal de filas. En esta situación, las x_{ij} son observaciones de una (única) distribución multinomial de parámetros q_{ij}, y las medidas de acuerdo global y parcial correspondientes a este *modelo I* serán, respectivamente

$$\Delta_{\mathrm{I}} = \sum q_i\Delta_i \tag{6}$$

y

$$\mathcal{A}_i = q_i\Delta_i\,. \tag{7}$$

En principio, bajo el modelo I se debe poder hablar en términos de *conformidad* (si hay estándar) o en términos de *consistencia* (si no hay estándar), pero la *consistencia* requiere que el modelo sea coherente en el sentido de que las medidas de acuerdo no cambien al intercambiar filas por columnas. Como demostraron Martín y Femia (2004),

esto siempre es así bajo el modelo descrito y por tanto, en ausencia de estándar, las medidas dadas en (6) y (7) son medidas de consistencia entre R y C.

Como el tipo de muestreo es una cuestión que debe abordarse de forma previa a cualquier estudio, el contexto va a dejar siempre suficientemente claro si se está considerando el modelo de tipo I o el de tipo II, así que lo habitual en la literatura es simplificar la notación y aludir a Δ en lugar de hacerlo a Δ_{I} o Δ_{II}.

2.3. Estimación de los parámetros

La desventaja de Δ es que su estimación no es tan sencilla como la del coeficiente κ tradicional. Como se expone a continuación, para estimar Δ se debe resolver una ecuación no lineal que (en general) carece de solución explícita, requiriendo por tanto del uso de un procedimiento numérico.

Según la formulación de Martín y Femia (2004), el modelo (independientemente de que se trate del de tipo I o II) implica a una constante desconocida *B* y su dificultad radica precisamente en estimar dicha constante, pues se trata de determinar el valor

$$B = n(1-\Delta) \geq B_0 = \max_i \left\{ \sqrt{c_i - x_{ii}} + \sqrt{r_i - x_{ii}} \right\}^2$$

tal que

$$y(B) = (K-2)B + \sum s(i) \sqrt{\left[B + (c_i - r_i)\right]^2 - 4B(c_i - x_{ii})} = 0 \qquad (8)$$

en donde $s(i) = -1$ ($\forall i$) salvo cuando ocurra $y(B_0) < 0$, en cuyo caso será $s(i) = +1$ en la clase *i* que proporciona ese valor de B_0. Una vez que se ha obtenido el valor de *B*, resulta inmediato estimar los parámetros del modelo en base a las expresiones siguientes (Martín y Femia, 2004):

$$\hat{\pi}_i = \frac{B + (c_i - r_i) + s(i) \sqrt{\left[B + (c_i - r_i)\right]^2 - 4B(c_i - x_{ii})}}{2B} \qquad (9)$$

$$\hat{\Delta}_i = \frac{x_{ii} - r_i \hat{\pi}_i}{r_i (1 - \hat{\pi}_i)} \qquad (10)$$

$$\hat{\Delta} = \sum r_i \hat{\Delta}_i \big/ n \qquad (11)$$

El acuerdo global $\hat{\Delta}$ y los acuerdos parciales $\hat{\mathcal{A}}_i = r_i \hat{\Delta}_i / n$ no dependen del tipo de muestreo, y tampoco cambian al intercambiar filas por columnas, por lo que son válidas tanto como medidas de conformidad como de consistencia.

2.4. Existencia y unicidad de la solución

En el anexo I de Martín y Femia (2004) se comprueba que la ecuación $y(B)=0$ –dada en (8)– tiene solución (y es única) siempre que no exista una clase h en la que ocurra

$$c_h + r_h - 2x_{hh} = n - \sum x_{ii} \; . \qquad (12)$$

Cuando se da tal situación, la ecuación (8) no tiene solución o tiene infinitas soluciones. Esto ocurre cuando solo hay dos clases (K=2) o bien cuando, habiendo más de dos clases (K>2), no existen discordancias (es decir $x_{ii} = c_i = r_i \quad \forall i$) o todas las discordancias se concentran en una fila o en una columna (es decir, si $x_{ii} = r_i$ o $x_{ii} = c_i$ respectivamente, en todas las clases menos en una).

Cuando K>2, el problema se solventa acudiendo a un recurso muy utilizado en estadística que es cambiar los valores x_{ij} por x_{ij}+0,5 (cf. por ejemplo Agresti, 2002; Martín y Luna, 2004). Procediendo así la ecuación (8) tiene solución única siempre.

El caso en que solo hay dos categorías (K=2) merece una consideración especial. Por una parte, la presentación de datos en forma de tabla de contingencia 2×2 es muy habitual, ya que las variables dicotómicas aparecen en muchos problemas de decisión, bien sea de forma directa (por ejemplo para caracterizar la presencia/ausencia de un determinado rasgo), o bien por la dicotomización de variables cuantitativas (como sucede a menudo con los test de diagnóstico clínico). Pero, por otra parte, la limitación de sus grados de libertad hace que una tabla 2×2 constituya un problema difícil de abordar. En palabras de Guggenmoos-Holzmann (1993), "cualquier medida de acuerdo que se defina en una tabla 2×2 puede presentar serios problemas". En el apartado siguiente abordamos el enfoque del modelo *Delta* para abordar esta situación.

2.5. Estimación para K=2

Cuando solo hay dos clases, el modelo Delta –según la formulación dada en (2)– depende de tres parámetros: Δ_1, Δ_2 y, por ejemplo π_1 (ya que $\pi_2 = 1 - \pi_1$). Esto implica que hay tantos parámetros como variables bajo el muestreo I, y más parámetros que variables bajo el muestreo II (ya que estaríamos tratando con dos binomiales independientes). Como en estas condiciones el modelo no se puede estimar, la solución adoptada para por Martín y Femia (2004) fue la siguiente:

- Incrementar la tabla añadiendo una clase *extra* con $r_3=c_3=x_{33}$, siendo x_{33} un valor arbitrario. El valor que se asigne a x_{33} va a ser irrelevante, pues no va a afectar a la solución.

- Sustituir todos los valores x_{ij} por x_{ij}+c, siendo c una constante arbitraria c≤0,5
- Obtener los $\hat{\Delta}_i$ para esta nueva tabla y definir las medidas corregidas por azar solo para las clases originales i=1, 2

La tabla así obtenida tiene siempre solución (no puede ocurrir la situación expresada en (12)). Queda por concretar la elección de c, lo cual conduce a tres situaciones posibles:

- Si c=0,5 se obtiene el estimador propuesto por Martín y Femia (2004) en la formulación original del modelo.
- Si c→0, las soluciones $\hat{\Delta}_i$ convergen a

$$\hat{\Delta}_{i\mathrm{a}} = \frac{x_{ii} - \sqrt{x_{12}x_{21}}}{r_i} \tag{13}$$

obteniendo el acuerdo global asintótico

$$\hat{\Delta}_{\mathrm{a}} = \frac{x_{11} + x_{22} - 2\sqrt{x_{12}x_{21}}}{n} \tag{14}$$

- Si c→+1, entonces se obtiene la aproximación

$$\hat{\Delta}_{i\mathrm{a}(+1)} = \frac{x_{ii} + 1 - \sqrt{(x_{12}+1)(x_{21}+1)}}{r_i + 2} \tag{15}$$

y el acuerdo global vendrá dado por

$$\hat{\Delta}_{\mathrm{a}(+1)} = \frac{x_{11} + x_{22} + 2 - 2\sqrt{(x_{12}+1)(x_{21}+1)}}{n+4} \tag{16}$$

Realmente, el método apropiado es el primero, es decir con c=0,5, sin embargo esto supone que va a ser preciso resolver la ecuación (8) mediante un algoritmo iterativo, ya que en la práctica la situación es la de K=3. Las soluciones con c→0 y con c→+1 constituyen aproximaciones a la primera pero, en contrapartida, las dos resultan ser explícitas, de forma que son fácilmente calculables sin necesidad de programación. Martín y Femia (2008) han comprobado que la última de ellas, $\hat{\Delta}_{\mathrm{a}(+1)}$, es mejor aproximación a $\hat{\Delta}$ que $\hat{\Delta}_{\mathrm{a}}$ (y lo mismo para los acuerdos parciales).

2.6. Error estándar de los estimadores

Martín y Femia (2004), mediante la inversión de la matriz de información de Fisher, obtienen las expresiones que siguen para estimar el error estándar (S.E.) de los

estimadores de los parámetros del modelo. Considerando

$$u_i = \frac{r_i - x_{ii}}{\left(1-\hat{\pi}_i\right)^2},\ E_i = \frac{\hat{\pi}_i}{B-u_i} \text{ y } E = \sum E_i,$$

se define

$$U_{ii} = \frac{u_i x_{ii}}{r_i} + u_i^2 E_i\left(1-\frac{E_i}{E}\right),\ U_{ij} = -\frac{\left(u_i E_i\right)\left(u_j E_j\right)}{E}.$$

El S.E. de los $\hat{\Delta}_i$ responde a la misma expresión con independencia del modelo asumido:

$$\text{S.E.}\left(\hat{\Delta}_i\right) = \frac{\sqrt{U_{ii}}}{r_i} \tag{17}$$

Para el caso de los acuerdos parciales y del acuerdo total, si el muestreo es de tipo I:

$$\text{S.E.}(\hat{\mathcal{A}}_i) = \frac{\sqrt{U_{ii} + r_i\left(n-r_i\right)\hat{\Delta}_i^2/n}}{n} \text{ y } \text{S.E.}\left(\hat{\Delta}_{\text{I}}\right) = \frac{\sqrt{\sum\sum U_{ij} + \sum r_i\hat{\Delta}_i^2 - n\hat{\Delta}^2}}{n} \tag{18}$$

en tanto que si es del tipo II:

$$\text{S.E.}(\hat{\mathcal{A}}_i) = \frac{\sqrt{U_{ii}}}{n} \text{ y } \text{S.E.}\left(\hat{\Delta}_{\text{II}}\right) = \frac{\sqrt{\sum\sum U_{ij}}}{n} \tag{19}$$

Todas las expresiones son válidas si no existe ningún x_{ii}=0, c_i ó r_i. En el caso de darse algún x_{ii} de ese tipo, los valores de S.E. se obtienen a través de la tabla incrementada en +0,5, como es habitual en Estadística.

Se omiten aquí las expresiones que permiten obtener el error estándar de las soluciones asintóticas, los cuales pueden encontrarse en las referencias citadas de Martín y Femia (2008).

2.7. Bondad del ajuste

La medida de acuerdo *Delta* depende de la validez del modelo subyacente (cosa que no ocurre con *Kappa*, que no depende de modelo alguno), de modo que la validez de las inferencias depende de que el modelo sea adecuado para analizar un problema dado. El test apropiado para ello es el clásico test χ^2 de bondad de ajuste. Las frecuencias esperadas por el modelo son $\text{E}_{ii} = x_{ii}$ y $\text{E}_{ij} = r_i(1-\hat{\Delta}_i)\hat{\pi}_j = (r_i - x_{ii})\hat{\pi}_j/(1-\hat{\pi}_i)$ (M&F, 2004) de modo que el estadístico de contraste viene dado por la expresión

$$\chi^2_{\text{exp}} = \sum_{i\neq j}\left(x_{ij} - \text{E}_{ij}\right)^2 / \text{E}_{ij} \tag{20}$$

que, bajo la hipótesis de que el modelo es adecuado, seguirá una distribución χ^2 con $(K-1)(K-2)-1$ grados de libertad (d.f.). Estos d.f. provienen de que en el muestreo II hay $K(K-1)$ casillas libres de tomar valores y $2K-1$ parámetros estimados (similarmente para el muestreo I). La significación del test indica que el modelo de respuesta Delta no es adecuado para analizar el problema.

3. Medidas de validez y de fiabilidad corregidas por azar a través del modelo Delta

Con frecuencia interesa determinar algo más que el nivel de acuerdo global. Por ejemplo, cuando se estudia la eficacia de un método de clasificación interesan medidas tales como la *sensibilidad* (*especificidad)* del método, es decir, la probabilidad de que dicho método clasifique como poseedor (no poseedor) de la característica de interés a aquellos casos que realmente la poseen (no la poseen); o sus *valor predictivo positivo* (*negativo*), es decir, la probabilidad de que los casos que han sido clasificados como poseedores (no poseedores) de la característica de interés, realmente la tengan (no la tengan). Tales medidas son de interés primordial en Medicina y en Epidemiología, puesto que constituyen los índices de precisión de los métodos de diagnóstico clínico. Sin embargo, su definición tradicional no contempla la corrección por azar. Martín y Femia (2005, 2008) han considerado la corrección por azar de este tipo de medidas a partir del modelo de acuerdo *Delta* descrito en la sección anterior.

Tabla 2. Presentación de los datos en un problema de clasificación binaria. Se asume que R es un estándar y C el método de clasificación cuyas propiedades se desean evaluar. Se indica con +/– la presencia/ausencia de la característica de interés.

		Clasificación de C		
		+	–	*Total*
Clasificación de R	+	x_{11}	x_{12}	r_1
	–	x_{21}	x_{22}	r_2
	Total	c_1	c_2	n

Para centrar ideas, supongamos un test de diagnóstico binario (C) cuyos resultados se verifican con un estándar (R). Según se ha asumido previamente, el estándar aparecerá dispuesto por filas al resumir los datos como en la Tabla 1, que ahora

será una tabla 2×2 al tratarse de un test binario tal y como se representa en la Tabla 2. Como es tradicional, el muestreo puede ser de tipo I o de tipo II.

Las propiedades intrínsecas de C como método de clasificación vienen dadas en términos de su sensibilidad (*S*) y de su especificidad (*E*) que, a partir de los datos observados, se estiman tradicionalmente como $\hat{S} = x_{11}/r_1$ y $\hat{E} = x_{22}/r_2$ respectivamente (con independencia del tipo de muestreo asumido). Por otra parte, la utilidad práctica de C como método de clasificación suele caracterizarse en términos de sus valores predictivos positivo (*PPV*) y negativo (*NPV*). Para estimarlos, suele considerarse el muestreo I (cf. con Pepe, 2003) de manera que $\widehat{PPV} = x_{11}/c_1$ y $\widehat{NPV} = x_{22}/c_2$.

Por su parte, el modelo *Delta* permite expresar los acuerdos observados en una categoría (x_{ii}) como la suma de dos componentes: aquellos acuerdos que no son debidos al azar $(r_i\hat{\Delta}_i)$ mas los acuerdos que si lo son $(x_{ii} - r_i\hat{\Delta}_i)$. De manera que sustituyendo x_{ii} por $r_i\hat{\Delta}_i$ en los índices anteriores, se obtiene su versión corregida por azar: $\hat{\Delta}_1$ y $\hat{\Delta}_2$ para la sensibilidad y la especificidad respectivamente (con independencia del tipo de muestreo) y $\hat{\Delta}_1 r_1/c_1$, $\hat{\Delta}_2 r_2/c_2$ para los valores predictivos positivo y negativo respectivamente (cuando se considera que el muestreo es de tipo I).

La situación descrita es generalizable al caso en que la clasificación no sea binaria (*K*>2) y también cuando no exista estándar (Martín y Femia, 2005). En presencia de estándar y para un valor arbitrario de *K*, los índices tradicionales del tipo x_{ii}/r_i son estimadores de los *índices de conformidad* en la clase *i*; ahora, el estimador corregido por azar según el modelo *Delta*, viene dado por

$$\hat{\mathcal{F}}_i = \hat{\Delta}_i . \tag{21}$$

Análogamente, bajo el muestreo de tipo I, los índices del tipo x_{ii}/c_i, constituyen estimadores de los *índices de predictividad* en la clase *i*, de manera que ahora la estimación corregida por azar vendrá dada por

$$\hat{\mathcal{P}}_i = r_i\hat{\Delta}_i/c_i \tag{22}$$

Cuando no existe estándar (asumimos que entonces el muestreo es de tipo I), no tiene sentido considerar los índices individuales de conformidad o de predictividad, debiendo considerar la mezcla de ambos (M&F, 2005) para dar lugar al índice de acuerdos i de Cicchetti y Feinstein (1990), o índice de *Consistencia* en la categoría *i*, que viene dado

por $\hat{S}_i = 2x_{ii}/(r_i + c_i)$ y no está corregido por azar. En paralelo con lo anterior, la versión corregida por azar de este índice será:

$$\hat{\mathcal{S}}_i = 2r_i\hat{\Delta}_i/(r_i + c_i) \qquad (23)$$

En Martín y Femia (2005) pueden encontrarse las expresiones para estimar la varianza de estos índices corregidos por azar.

4. Ejemplos

En la Tabla 3 se consideran dos ejemplos que permiten ganar apreciación sobre el modelo expuesto: en (a) se presentan datos tomados de Blackman y Koval (2000) y en (b) de Nelson y Pepe (2000). Puede observarse que $\hat{\kappa} \approx \hat{\Delta}$ cuando los marginales no están muy desequilibrados (como en el primer caso), pero que $\hat{\kappa}$ es muy inadecuado y muy diferente de $\hat{\Delta}$ cuando los marginales están muy desequilibrados (como en el segundo caso). La interpretación de $\hat{\Delta}$ en el primero de ellos es que hay un 71.2% de acuerdos no debidos al azar (el 28.8% restante son desacuerdos o acuerdos asignables al azar), mientras que aquí $\hat{\kappa}$ indica que los observadores concuerdan el 70.3% del máximo posible de acuerdos no debidos al azar.

Tabla 3. Ejemplos de clasificación binaria

(a)

R \ C	+	−	*Total*
+	297	40	337
−	39	181	220
Total	336	221	557

$\hat{\kappa} = 0,703\,;\, \hat{\Delta} = 0,712$

(b)

R \ C	+	−	*Total*
+	80	10	90
−	10	0	10
Total	90	10	100

$\hat{\kappa} = -0,11\,;\, \hat{\Delta} = +0,60$

Si consideramos los datos de la Tabla 3(a) como un problema de diagnóstico bajo el muestreo de tipo I, las medidas clásicas de especificidad y sensibilidad, corregidas por azar en base al modelo Delta, serán $\hat{\mathcal{E}}_1$=0,761 (=$\hat{\Delta}_1$ o conformidad en la clase 1) y $\hat{\mathcal{E}}_2$=0,639 ($\hat{\Delta}_2$ o conformidad en la clase 2) respectivamente, cuyos valores contrastan con los índices clásicos no corregidos de $\hat{S}$=297/337=0,881 y $\hat{E}$= 0,823 respectivamente. Del mismo modo, los valores predictivos tradicionales resultan ser de

$\widehat{PPV}$ = 297/336=0,884 para el positivo y $\widehat{NPV}$ =181/221=0,819 para el negativo, que contrastan a su vez con los valores de predictividad corregidos por azar, que respectivamente son $\hat{\mathcal{P}}_1$=0,763 y $\hat{\mathcal{P}}_2$=0,636.

Tabla 4. Medidas de acuerdo correspondientes a los datos de la Tabla 3(a) obtenidas por los procedimientos de añadir una clase extra y sumar a todos los valores: (*i*) c→0,5; (*ii*) c→0; (*iii*) c→1. Obsérvese cómo la solución asintótica (*iii*) es mejor aproximación a la (i) que la proporcionada por (*ii*). Los resultados se han obtenido con el programa Delta desarrollado por los autores.

Acuerdo global

Estimación	$\hat{\Delta}$	±S.E.
(*i*)	0,712	±0,030
(*ii*)	0,716	±0,030
(*iii*)	0,711	±0,030

Medidas de acuerdo parcial y medidas de fiabilidad y validez

Estimación	*i*	$\hat{\Delta}_i$	$\hat{\pi}_i$	$\hat{\mathcal{A}}_i$	±S.E.	$\hat{\mathcal{F}}_i$	±S.E.	$\hat{\mathcal{P}}_i$	±S.E.
(*i*)	1	0,761	0,494	0,460	±0,104	0,761	±0,170	0,763	±0,171
	2	0,639	0,500	0,253	±0,104	0,639	±0,260	0,636	±0,259
(*ii*)	1	0,764	0,497	0,462	±0,025	0,764	±0,028	0,766	±0,028
	2	0,643	0,503	0,254	±0,023	0,643	±0,042	0,640	±0,042
(*iii*)	1	0,760	0,497	0,059	±0,025	0,760	±0,028	0,762	±0,028
	2	0,637	0,503	0,252	±0,023	0,637	±0,042	0,635	±0,042

En la Tabla 4 se reproducen los resultados correspondientes al análisis completo de los datos de la Tabla 3(a). En ella se dan las medidas de acuerdo global y por categorías, así como los índices de fiabilidad y de validez descritos en el apartado anterior (además de los estimadores de las probabilidades de azar $\hat{\pi}_i$). Como se indicó en el apartado 2.5, no existe solución única para la tabla 2×2 constituida por los datos originales, de manera que la solución se obtiene añadiendo una clase extra y sumando c=0,5 a cada categoría (solución (*i*) de la Tabla 4). No obstante, según se ha visto, es posible obtener aproximaciones a esta solución que son sencillas de calcular (soluciones

(*ii*) e (*iii*) de la Tabla 4, en las que puede observarse que, aunque las estimaciones puntuales son fiables, sus valores de SE no lo son).

Tabla 5. Clasificación de 100 objetos en K=3 categorías con distribuciones marginales muy desequilibradas (datos de Martín y Femia, 2004).

		C			
		a	*b*	*c*	*Total*
R	*a*	1	1	2	4
	b	1	1	2	4
	c	0	0	92	92
	Total	2	2	96	100

$\hat{\kappa}(\pm\text{S.E.}) = 0,479(\pm 0,146)$

$\hat{\Delta}(\pm\text{S.E.}) = 0,920(\pm 0,040)$

La Tabla 5 reproduce una situación similar a la de la Tabla 3(b). El muestreo asumido es de tipo I, con R actuando como estándar. La distribuciones marginales son muy asimétricas y ello provoca que $\hat{\kappa}$ proporcione una estimación muy deficiente de lo que intuitivamente debe constituir un acuerdo mucho mayor, tal y como pone de manifiesto $\hat{\Delta}$. Así mismo es destacable la mayor precisión con que se estima $\hat{\Delta}$ frente a la de $\hat{\kappa}$.

Tabla 6. Diagnóstico de 100 individuos con trastorno psiquiátrico por parte de dos jueces (muestreo de tipo I, ninguno de los jueces es un estándar) (datos de Fleiss, 1981)

		C			
		Sicótico	Neurótico	Orgánico	*Total*
R	Sicótico	75	1	4	80
	Neurótico	5	4	1	10
	Orgánico	0	0	10	10
	Total	80	5	96	100

$\hat{\Delta}$=0.687±0.110

Medidas de acuerdo parcial y medidas de fiabilidad y validez

Clase	$\hat{\Delta}_i$	$\hat{\pi}_i$	$\hat{\mathcal{A}}_i$	±S.E.	$\hat{\mathcal{S}}_i$	±S.E.
Sicótico	0,687	0,80	0,550	±0,118	0,687	±0,144
Neurótico	0,375	0,04	0,038	±0,022	0,500	±0,206
Orgánico	1,000	0,16	0,100	±0,028	0,800	±0,108

Finalmente, en la Tabla 6 se presentan un ejemplo clásico (Fleiss, 1981) en el que dos jueces deben diagnosticar a n=100 pacientes en K=3 categorías. El nivel de acuerdo

global estimado por el modelo *Delta* es del 68,7% (muy similar en esta ocasión al valor de $\hat{\kappa}$=0,676). Como el muestreo asumido es de tipo I y ninguno de los jueces actúa como estándar, las medidas de conformidad y de predictividad no tienen sentido, siendo lo apropiado evaluar la consistencia en cada clase, que indica la proporción de acuerdos entre todos los individuos clasificados en cada categoría por cualquiera de los dos jueces. La consistencia entre los diagnósticos de ambos jueces es importante en la clase Orgánico (0,800), moderada en la clase Sicótico (0,687) y regular en la clase Neurótico (0,500), siendo especialmente en esta última en donde deben homogeneizar sus criterios de clasificación.

Todos los resultados expuestos se han obtenidos mediante la versión 4.1 del programa ejecutable Delta.exe, elaborado por los autores y que puede descargarse libremente en la dirección http://www.ugr.es/local/bioest/Delta.exe.

5. Conclusiones

En el presente trabajo se ha hecho una revisión sucinta del modelo propuesto por Martín y Femia (2004) para evaluar el grado de acuerdo entre dos observadores, cuando se considera una escala nominal. Este modelo ofrece numerosas ventajas respecto al clásico *Kappa* de Cohen (1960), el cual puede contemplarse como una aproximación a Delta solo cuando los marginales no están muy desequilibrados. Adicionalmente, el método Delta no solo ofrece una medida de acuerdo global, sino que permite evaluar el nivel de concordancia en cada categoría. Por otra parte, Δ tiene una interpretación más fácil e intuitiva que κ, ya que un valor $\hat{\Delta}$=0.80 indica que un 80% de las respuestas son concordantes no por causa del azar (el otro 20% son discordantes ó son concordantes por causa del azar), mientras que $\hat{\kappa}$=0.80 lo que indica es que los observadores están de acuerdo en una fracción 0.80 del máximo número de acuerdos no debidos al azar.

La ventaja de $\hat{\kappa}$ frente a $\hat{\Delta}$ es que no depende de ningún modelo –por tanto es válida siempre– y que es fácil de cálcular. Sin embargo, el primero de estos argumentos a favor de *Kappa* también tiene sus detractores, Uebersax (1987) discrepa de que *Kappa* corrija realmente por azar, ya que para ello es necesario disponer de un modelo que explique de qué manera el azar afecta a las decisiones de los observadores. Precisamente esto es lo que considera *Delta.* Por otra parte, es cierto que en general la resolución de (8) requiere utilizar un método numérico de determinación de las raíces de una ecuación no lineal, lo que requiere de cierta programación. Sin embargo, mas allá

del impedimento de no poder utilizar para ello una simple calculadora de bolsillo (al menos de forma eficiente), métodos iterativos bien conocidos –como es el caso del método de Newton-Raphson– convergen rápidamente a la única solución de (8), una vez tomadas las medidas pertinentes para garantizar que no se da la condición descrita en (12). Un programa para Windows disponible en la web para aplicar el modelo Delta puede descargarse de la dirección http://www.ugr.es/local/bioest/Delta.exe.

En el ámbito específico de evaluar un método de clasificación, el modelo *Delta* permite definir medidas corregidas por azar que deberían considerarse en lugar de los índices clásicos no corregidos, como son la sensibilidad y la especificidad o los valores predictivos de un método de diagnóstico binario. Estos índices son del máximo interés en las Ciencias de la Salud y, en su definición actual, suponen una contradicción metodológica, ya que es frecuente utilizar Kappa para tener en cuenta el efecto del azar al evaluar el acuerdo global del método con el estándar y, sin embargo, al estudiar sus medidas de eficacia se ignora dicho efecto. La necesidad de utilizar estos índices corregidos por azar es aún mas manifiesta si de lo que se trata es de comparar el valor de uno de estos índices en dos experiencias distintas.

En la introducción de este trabajo se han revisado algunos estudios que han considerado Delta desde el punto de vista aplicado. Desde el punto de vista formal este modelo también ha sido objeto de atención; cf. Ato *et al*. (2006, 2008), Park *et al*. (2007) y Abar y Loken (2010).

6. Agradecimientos / *Acknowledgements.*

Este trabajo ha sido financiado por el Ministerio Español de Educación y Ciencia, proyecto número MTM2009-08886 (y cofinanciado por el Fondo Europeo de Desarrollo Regional).

This research was supported by the Spanish Ministry of Education and Science, grant number MTM2009-08886 (co-financed by the European Regional Development Fund).

Referencias

Agresti, A. (2002) Categorical Data Analysis, 2nd Ed. Wiley.

Abar, B. y Loken, E. (2010) Peirce's i and Cohen's κ for 2×2 Measures of Rater Reliability. *Journal of Probability and Statistics,* 2010,1

Allouche, O.; Tsoar, A. and Kadmon, R. (2006) Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS). *Journal of Applied Ecology,* **43:**1223–1232

Ato, M., Benavente, A., López, J.J. (2006) Comparative analysis of three approaches for rater agreement. *Psicothema*, 18(3):638-645

Ato, M.; López, J.J.; Benavente, A. (2008) Un índice de sesgo entre observadores basado en modelos mixtura. *Psicothema*, 20(4):918-923

Ato, M.; López, J.J.; Benavente, A. (2011) A simulation study of rater agreement measures with 2×2 contingency table. *Psicológica,* **32**(2):385-402

Baines CJ, McFarlane DV, Miller AB. (1990) The role of the reference radiologist: estimates of inter-observer agreement and potential delay in cancer detection in the national breast screening study. *Invest Radiol,* **25**:971-976

Baranek, GT; Boyd BA; Poe MD; David FJ; and Watson LR (2007) Hyperresponsive Sensory Patterns in Young Children with Autism, Developmental Delay, and Typical Development. *American Journal on Mental Retardation*, 112(4): 233–245.

Blackman, N. J.-M. and Koval, J. J. (2000). Interval Estimation for Cohen's Kappa as a Measure of Agreement. *Statistics in Medicine 19*, 723-741.

Buppan, P.; Putaporntip, C.; Pattanawong, U.; Seethamchai, S. and Jongwutiwes, S.(2010). Comparative detection of *Plasmodium vivax* and *P. falciparum* DNA in saliva and urine samples from symptomatic malaria patients in a low endemic area. *Malaria Journal*, **9**:72

Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement,* **20:** 37-46.

Dubois, M.F.; Raîche, M.; Hébert, R. and Gueye, N.R. (2007) Assisted self-report of health-services use showed excellent reliability in a longitudinal study of older adults. *Journal of Clinical Epidemiology*, **60**(10): 1040-1045

Fisher, M.; Storfer-Isser, A.; Shaw, R. J.; Bernard, R. S.; Drury, S.; Ularntinon, S. and Horwitz, S. M. (2011) Inter-rater reliability of the Pediatric Transplant Rating Instrument (P-TRI): Challenges to reliably identifying adherence risk factors during pediatric pre-transplant evaluations, *Pediatric Transplantation*, 15(2):142–147.

Ellis, E. C. and Wang, H. (2006) Estimating area errors for fine-scale feature-based ecological mapping. *International Journal of Remote Sensing*, **27**(21):4731-4749.

Fanshawe, T.R., Lynch, A.G., Ellis, I.O.,Green, A.R., Hanka, R. (2008) Assessing agreement between multiple raters with missing rating information, applied to breast cancer tumour grading. *PLoS ONE*, **3**(8), (art. no. e2925)

Fleiss, J.L. (1975) Measuring agreement between two judges on the presence or absence of a trait. *Biometrics*, **31**: 651–659

Fleiss, J.L. (1981) *Statistical Methods for Rates and Proportions.* Wiley (NY)

Freeman, E. A. and Moisen, G. G. (2008) A comparison of the performance of threshold criteria for binary classification in terms of predicted prevalence and kappa. *Ecological Modelling,* 217:48–58

Guay, M; Desrosiers, J. and Dubois, M.F. (2010) Does the clinical context affect the validity of bathroom recommendations made by home health aides?. *International Journal of Industrial Ergonomics* 40: 82–89

Guggenmoos-Holzmann, I. (1993). How reliable are chance-corrected measures of agreement? *Statistics in Medicine* **12**, 2191-2205

Hanspach, J.; I.Kuhn,I.; Pompe, S. and Klotz, S. (2010) Predictive performance of plant species distribution models depends on species traits. *Perspectives in Plant Ecology, Evolution and Systematics*, 12:219–225

Kottner, J. *et al.* (2011) Guidelines for Reporting Reliability and Agreement Studies (GRRAS) were proponed. *Journal of Clinical Epidemiology,* **64**(1):96-106

Linsell,L; Forbes, L.J.L.; Burgess,C.; Kapari, M.; Thurnham, A. and Ramirez, A.J. (2010) Validation of a measurement tool to assess awareness of breast cancer. *European Journal of Cancer*, 46:1374–138

Luoto,M.; Marmion, M. and Hjort, J. (2010) Assessing spatial uncertainty in predictive geomorphological mapping: A multi-modelling approach. *Computers & Geosciences,* **36**:355–361

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

Martín Andrés, A. and Femia Marzo, P. (2005). Chance-corrected measures of reliability and validity in K×K tables. *Statistical Methods in Medical Research, 14*, 473-492.

Martín Andrés, A. and Femia Marzo, P. (2008) Chance-Corrected Measures of Reliability and Validity in 2 × 2 Tables. *Communications in Statistics—Theory and Methods*, **37**: 760–772.

Martín Andrés, A y Luna del Castillo (2004) *Bioestadística para las Ciencias de la Salud*. Ediciones Norma (Madrid, España).

McGinn,T; Guyatt,G.; Cook,R.;Korenstein, D and Meade, M.O. (2008) Advanced Topics in Diagnosis. Measuring Agreement Beyond Chance. In *User's guides to the medical literature. A manual for evidence-based clinical practice*. (Guyatt et al. Eds) *JAMA* (17.3:481–489)

Monserud, R.A. and Leemans, R. (1992). Comparing global vegetation maps with the Kappa statistic. *Ecological Modelling* **62**, 275–293.

Nelson, J.C. y Pepe, M.S. (2000). Statistical description of inter-rater variability in ordinal rating. *Statistical Methods in Medical Research 9*, 475-496.

Park, M.H. y Park, Y.G. (2007) A New Measure of Agreement to Resolve the Two Paradoxes of Cohen's Kappa, *The Korean Journal of Applied Statistics,* **20**(1):117-132

Pepe, M. (2003) *The Statistical Evaluation of Medical Test for Classification and Prediction*. Oxford University Press (N.Y.)

Prinzing, A.; Durka, W.; Klotz, S. and Brandl, R. (2005) How to characterize and predict alien species? A response to Pysek et al. (2004) *Diversity and Distributions*, **11:**121-123.

Roberts, C. (2008) Modelling patterns of agreement for nominal scales. *Statistics in Medicine*, **27**:810–830

Shrout, P.E. (1998) Measurement reliability and agreement in psychiatry. *Statistical Methods in Medical Research*, 7: 301-317

Uebersax, JS. (1987) Diversity of decision-making models and the measurement of interrater agreement. *Psychological Bulletin*, 1987, 101, 140-146.

van Vliet, J.; Bregtb, A. K. and Hagen-Zankerc, A. (2011) Revisiting Kappa to account for change in the accuracy assessment of land-use change models. *Ecological Modelling*, **222**:1367-1375.

Wastling, S.L.; Picozzi, K.; Kakembo,A.S.L. and Welburn,S.C. (2010). LAMP for Human African Trypanosomiasis: A Comparative Study of Detection Formats. *Plos, Neglected Tropical Diseases*, 4(11) e865

Webster, AC; Supramaniam, R; O'Connell, DL; Chapman, JR and Craig, JC. (2010) Validity of registry data: agreement between cancer records in an end-stage kidney disease registry (voluntary reporting) and a cancer register (statutory reporting). *Nephrology*, 15(4):491-501

Weiner, Shinka & Velicer (Eds.) (2003) *Handbook of Psychology. Vol. 2: Research Methods in Psychology*. Wiley.