Medidas de concordancia: el índice de Kappa

Autores: López de Ullibarri Galparsoro I, Pita Fernández, S.
Unidad de Epidemiología Clínica y Bioestadística. Complexo Hospitalario-Universitario Juan Canalejo. A Coruña (España)
Cad Aten Primaria 1999; 6: 169-171.
Actualización 24/09/2001.

En cualquier estudio de investigación una cuestión clave es la fiabilidad de los procedimientos de medida empleados. Como señala Fleiss [1] en el contexto de los estudios clínicos, ni el más elegante de los diseños sería capaz de paliar el daño causado por un sistema de medida poco fiable.

Tradicionalmente se ha reconocido una fuente importante de error de medida en la variabilidad entre observadores [1,2]. Consecuentemente, un objetivo de los estudios de fiabilidad debe consistir en estimar el grado de dicha variabilidad.

En este sentido, dos aspectos distintos entran a formar parte típicamente del estudio de fiabilidad: de una parte, el ***sesgo* entre observadores** –dicho con menos rigor, la tendencia de un observador a dar consistentemente valores mayores que otro– y de otra, la ***concordancia* entre observadores** –es decir, hasta qué punto los observadores coinciden en su medición–.

Ciñéndonos a este segundo aspecto, la manera concreta de abordar el problema depende estrechamente de la naturaleza de los datos: si éstos son de tipo continuo es habitual la utilización de estimadores del coeficiente de correlación intraclase, mientras que cuando se trata de datos de tipo categórico el estadístico más empleado es el índice kappa, al que dedicamos el resto de este artículo.

El índice kappa

Supongamos que dos observadores distintos clasifican independientemente una muestra de **n** ítems en un mismo conjunto de C categorías nominales. El resultado de esta clasificación se puede resumir en una tabla como la tabla 1, en la que cada valor x_{ij} representa el número de ítems que han sido clasificados por el observador 1 en la categoría i y por el observador 2 en la categoría j.

Tabla 1. Formato de los datos en un estudio de concordancia

	Observador 2				
Observador 1	1	2	...	C	**Total**
1	X_{11}	X_{12}	...	X_{1C}	$\mathbf{X_{1.}}$
2	X_{21}	X_{22}	...	X_{2C}	$\mathbf{X_{2.}}$
.	.			.	.
.	.			.	.
.	.			.	.
C	X_{C1}	X_{C2}	...	X_{CC}	$\mathbf{X_{C.}}$
Total	$\mathbf{X_{.1}}$	$\mathbf{X_{.2}}$	**...**	$\mathbf{X_{.C}}$	**n**

Por ejemplo, podemos pensar en dos radiólogos enfrentados a la tarea de categorizar una muestra de radiografías mediante la escala: "anormal, "dudosa", "normal". La tabla 2 muestra un conjunto de datos hipotéticos para este ejemplo, dispuesto de acuerdo con el esquema de la tabla 1.

Tabla 2. Datos hipotéticos de clasificación de una muestra de 100 radiografías por dos radiólogos.				
	Radiólogo 2			
Radiólogo 1	**Anormal**	**Dudosa**	**Normal**	**Total**
Anormal	18	4	3	**25**
Dudosa	1	10	5	**16**
Normal	2	4	53	**59**
Total	**21**	**18**	**61**	**100**

Desde un punto de vista típicamente estadístico es más adecuado liberarnos de la muestra concreta (los **n** ítems que son clasificados por los dos observadores) y pensar en términos de la población de la que se supone que ha sido extraída dicha muestra. La consecuencia práctica de este cambio de marco es que debemos modificar el esquema de la tabla 1 para sustituir los valores x_{ij} de cada celda por las probabilidades conjuntas, que denotaremos por π_{ij} (tabla 3).

Tabla 3. Modificación del esquema de la Tabla 1 cuando se consideran las probabilidades de cada resultado					
	Observador 2				
Observador1	1	2	...	C	**Marginal**
1	π_{11}	π_{12}	...	π_{1}	$\boldsymbol{\pi_{1}}$
2	π_{12}	π_{22}	...	π_{2C}	$\boldsymbol{\pi_{2}}$
.	.			.	.
.	.			.	.
.	.			.	.
C	π_{C1}	π_{C2}	...	π_{CC}	$\boldsymbol{\pi_{C}}$
Marginal	$\boldsymbol{\pi_{\cdot 1}}$	$\boldsymbol{\pi_{\cdot 2}}$	**...**	$\boldsymbol{\pi_{\cdot C}}$	**1**

Con el tipo de esquematización que hemos propuesto en las tablas 1 ó 3 es evidente que las respuestas que indican concordancia son las que se sitúan sobre la diagonal principal. En efecto, si un dato se sitúa sobre dicha diagonal, ello significa que ambos observadores han clasificado el ítem en la misma categoría del sistema de clasificación. De esta observación surge naturalmente la más simple de las medidas de concordancia que consideraremos: la suma de las probabilidades a lo largo de la diagonal principal. En símbolos, si denotamos dicha medida por π_0, será

$\pi_0 = \sum \pi_{ii}$ donde los índices del sumatorio van desde i = 1 hasta i = C.

Como es obvio, se cumple que $0 \leq \pi_0 \leq 1$

correspondiendo el valor 0 a la mínima concordancia posible y el 1 a la máxima.

Aunque este sencillo índice ha sido propuesto en alguna ocasión [3] como medida de concordancia de elección, su interpretación no está exenta de problemas. La tabla 4 ilustra el tipo de dificultades que pueden surgir. En el caso A, $\pi_0 = 0.2$, luego la concordancia es mucho menor que en el caso B, donde $\pi_0 = 0.8$. Sin embargo, *condicionando por las distribuciones marginales* se observa que en el caso A la concordancia es la máxima posible, mientras que en el B es la mínima.

Tabla 4. Ejemplos de concordancia.							
A				**B**			
	Observador 2				**Observador 2**		
Observador 1	1	2	**Marginal**	**Observador 1**	1	2	**Marginal**
1	0.1	0.8	**0.9**	1	0.8	0.1	**0.9**
2	0	0.1	**0.1**	2	0.1	0	**0.1**
Marginal	**0.1**	**0.9**	**1**	**Marginal**	**0.9**	**0.1**	**1**

Por lo tanto, parece claro que la búsqueda se debe orientar hacia nuevas medidas de concordancia que tengan en cuenta las distribuciones marginales, con el fin de distinguir entre dos aspectos distintos de la concordancia, a los que podríamos aludir informalmente como concordancia absoluta o relativa [4]. El índice **kappa** representa una aportación en esta dirección, básicamente mediante la incorporación en su fórmula de una corrección que excluye la concordancia debida exclusivamente al azar –corrección que, como veremos, está relacionada con las distribuciones marginales–.

Con la notación ya empleada en la tabla 3, el índice kappa, κ , se define como

$$\kappa = \frac{\sum \pi_{ii} - \sum \pi_i \sum \pi_i}{1 - \sum \pi_i \pi_i} \qquad [1]$$

donde los índices del sumatorio van desde i = 1 hasta i = C.

Es instructivo analizar la expresión anterior. Observemos en primer lugar que si suponemos la independencia de las variables aleatorias que representan la clasificación de un mismo ítem por los dos observadores, entonces la probabilidad de que un ítem sea clasificado por los dos en la misma categoría i es $\pi_{i.}\pi_{.i}$. Por lo tanto, si extendemos el sumatorio a todas las categorías, $\Sigma\, \pi_{i.}\pi_{.i}$ es precisamente la probabilidad de que los dos observadores concuerden por razones exclusivamente atribuibles al azar. En consecuencia, el valor de κ simplemente es la razón entre el exceso de concordancia observado más allá del atribuible al azar ($\Sigma\, \pi_{ii} - \Sigma\, \pi_{i.}\pi_{.i}$) y el máximo exceso posible ($1 - \Sigma\, \pi_{i.}\pi_{.i}$) [5].

La máxima concordancia posible corresponde a $\kappa = 1$. El valor $\kappa = 0$ se obtiene cuando la concordancia observada es precisamente la que se espera a causa exclusivamente del azar. Si la concordancia es mayor que la esperada simplemente a causa del azar, $\kappa > 0$, mientras que si es menor, $\kappa < 0$. El mínimo valor de κ depende de las distribuciones marginales.

En el ejemplo de la tabla 4, κ vale 0.024 en el caso A y -0.0216 en el B, lo que sugiere una interpretación de la concordancia opuesta a la que sugiere el índice π_0 (*vide supra*). Para comprender resultados paradójicos como éstos [6], conviene recordar los comentarios que hacíamos más arriba acerca de las limitaciones del índice π_0.

A la hora de interpretar el valor de κ es útil disponer de una escala como la siguiente [7], a pesar de su arbitrariedad:

Valoración del Índice Kappa	
Valor de k	**Fuerza de la concordancia**
< 0.20	Pobre
0.21 – 0.40	Débil
0.41 – 0.60	Moderada
0.61 – 0.80	Buena
0.81 – 1.00	Muy buena

A partir de una muestra se puede obtener una estimación, k, del índice kappa simplemente reemplazando en la expresión [1] las probabilidades por las proporciones muestrales correspondientes:

$$k = \frac{\sum\left(X_{ii}/n\right)\sum\left(X_{i}/n\right)\left(X_{i}/n\right)}{1 - \Sigma\left(X_{i}/n\right)\left(X_{i}/n\right)} = \frac{n\Sigma X_{ii} - X_i X_i}{n^2 - \Sigma X_i X_i} \qquad [2]$$

Con los datos de la tabla 2 se obtiene aplicando esta fórmula un valor de k = 0.66, que según nuestra convención anterior calificaríamos como una buena concordancia.

Contrastes de hipótesis e intervalos de confianza.

La obtención de una simple estimación puntual del valor de κ no nos proporciona ninguna indicación de la precisión de dicha estimación. Desde el punto de vista de la Estadística Inferencial es esencial conocer la variabilidad de los estimadores y emplear ese conocimiento en la formulación de contrastes de hipótesis y en la construcción de intervalos de confianza.

Fleiss, Cohen y Everitt [8] dan la expresión de la varianza asintótica –es decir, para muestras infinitamente grandes– del estimador k, cuando el verdadero valor de κ es cero:

$$\sigma_0^2(k) = \frac{\sum \pi_i \pi_i + \left(\sum \pi_i \pi_i\right)^2 - \sum \pi_i \pi_i (\pi_i + \pi_i)}{(1 - \sum \pi_i \pi_i)^2 n} \qquad [3]$$

Reemplazando las probabilidades teóricas, que desconocemos, por las proporciones muestrales, obtenemos un estimador de $\sigma_0^2(k)$ que denotaremos por $s_0^2(k)$:

$$S_0^2(k) = \frac{n\sum x_i x_i + \left(\sum x_i x_i\right)^2 / n - \sum x_i x_i (x_i + x_i)}{\left(n^2 - \sum x_i x_i\right)^2} \qquad [4]$$

Podemos emplear este resultado para contrastar la hipótesis nula de que κ es cero frente a la alternativa de que no lo es, utilizando como estadístico del contraste el cociente

$$\frac{|k|}{S_0(k)} \qquad [5]$$

(|k| denota el valor absoluto de k) y comparando su valor con los cuantiles de la distribución normal estándar. Con los datos de la tabla 2, k = 0.6600 y $s_0^2(k)$= 0.0738, luego |k|/ $s_0(k)$= 8.9441 y como $z_{0.975}$ = 1.96, concluimos que, al nivel de significación α = 0.05, el valor de k es significativo y nos lleva a rechazar que κ sea cero.

Es discutible la utilidad del contraste de hipótesis anterior, ya que como en general es razonable esperar cierto grado de concordancia más allá del azar, nos encontraremos trivialmente con un resultado significativo. Para poder realizar contrastes de hipótesis más interesantes es necesario conocer la expresión de la varianza asintótica cuando no se supone que κ es cero. La expresión es sensiblemente más compleja que la [3] [4]:

$$\sigma^2(k) = \frac{T_1(1-T_1)(1-T_2)^2 + 2(1-T_1)(1-T_2)(2T_1T_2 - T_3) + (1-T_1)^2(T_4 - 4T_2^2)}{(1-T_2)^4 n} \qquad [6]$$

donde: $T_1 = \Sigma\, \pi_{ii}$,
$T_2 = \Sigma\, \pi_{i.}\pi_{.i}$,
$T_3 = \Sigma\, \pi_{ii}(\pi_{i.} + \pi_{.i})$,
$T_4 = \Sigma\Sigma\, \pi_{ij}(\pi_{j.} + \pi_{.i})^2$.

Se puede demostrar que cuando κ es cero la expresión [6] se reduce a la [3]. Para contrastar la hipótesis nula de que κ es igual a un valor dado κ_0 frente a una alternativa bilateral, procedemos como en el caso κ = 0, sólo que empleando como estadístico del contraste:

$$\frac{|k - k_0|}{s(k)} \qquad [7]$$

donde s(k) ahora es la raíz cuadrada de $s^2(k)$, el estimador de $\sigma^2(k)$ obtenido sustituyendo en [6] probabilidades por proporciones muestrales. Es obvio que el caso $\kappa = 0$ que explicábamos con anterioridad no es más que un caso particular de este contraste, con una mejor estimación del error estándar.

Volviendo al ejemplo de la tabla 2, para contrastar la hipótesis de que el verdadero valor de κ es $\kappa_0 = 0.7$, como k = 0.6600 y s(k) = 0.0677, calculamos $|k - \kappa_0|/s(k) = 0.5908 < z_{0.975} = 1.96$. Por tanto, al nivel de significación $\alpha = 0.05$, no hay suficiente evidencia para rechazar la hipótesis nula.

Desde el punto de vista inferencial, un enfoque más versátil que el del contraste de hipótesis consiste en dar intervalos de confianza para el verdadero valor de κ . Tomados simultáneamente, k y el intervalo de confianza nos dan, además de la mejor estimación de κ , una medida del error que podemos cometer con esa estimación. Un intervalo de confianza aproximado del $(1-\alpha)100\%$, construido por el método estándar, es de la forma:

$$\left[k - Z_{1-\alpha/2}\, s(k),\ k + Z_{1-\alpha/2}\, s(k)\right]$$

donde $z_{1-\alpha/2}$ es el percentil de orden $(1-\alpha/2)100$ de la distribución normal estándar. Con los datos de la tabla 2, nuestro intervalo de confianza del 95% para κ sería [0.5273 , 0.7927]. Se observa como los valores 0 y 0.7 que considerábamos en los contrastes anteriores, quedan respectivamente fuera y dentro del intervalo, un hecho que ilustra la equivalencia entre los dos enfoques: contraste de hipótesis y estimación por intervalos.

Aunque el lector más interesado en los aspectos prácticos, aquél que se limita exclusivamente a usar un programa estadístico para analizar sus datos, quizás piense que todos estos detalles son algo prolijos, consideramos que son importantes para interpretar y explotar óptimamente los resultados que le brinda el programa. Por ejemplo, un programa ampliamente difundido como el SPSS, muestra solamente el valor de k (expresión [2]), su error estándar calculado a partir del estimador de [6], y el valor del estadístico [5]. Las explicaciones de este epígrafe muestran cómo utilizar estos valores para obtener intervalos de confianza y realizar otros contrastes de hipótesis.

Bibliografía

1. Fleiss JL. The design and analysis of clinical experiments. New York: Wiley; 1986.
2. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977; 33: 159-174. [Medline]
3. Holley WJ, Guilford JP. A note on the G index of agreement. Educ Psychol Meas 1964; 32: 281-288.
4. Bishop YMM, Fienberg SE, Holland PW. Discrete multivariate analysis: theory and practice. Cambridge, Massachussetts: MIT Press; 1977.
5. Fleiss JL. Statistical methods for rates and proportions, 2nd edition. New York: Wiley; 2000.
6. Feinstein AR, Cicchetti DV. High agreement but low kappa: I. The problems of two paradoxes. J Clin Epidemiol 1990; 43: 543-549. [Medline]
7. Altman DG. Practical statistics for medical research. New York: Chapman and Hall; 1991.
8. Fleiss JL, Cohen J, Everitt BS. Large sample standard errors of kappa and weighted kappa. Psychol Bull 1969; 72: 323-327.