El índice kappa

V. Abraira

Unidad de Bioestadística Clínica. Hospital Ramón y Cajal. Madrid.

La medición es una actividad omnipresente tanto en la práctica como en la investigación clínica. Como ejemplos se pueden citar desde actividades relativamente simples, como registrar la presión arterial mediante un esfigmomanómetro, hasta actividades más complejas, como determinar la carga viral mediante una sofisticada técnica de laboratorio, pasando por la evaluación de la calidad de vida mediante un cuestionario diseñado al efecto. Estos procesos de medición están siempre amenazados por diversos errores que condicionan la calidad tanto de la investigación como de las decisiones clínicas que se apoyan en dichas mediciones[1]. Por ello es aconsejable que el clínico conozca algunos fundamentos de la teoría de la medida, en particular los índices usados en la evaluación de los errores de medición[2].

Básicamente hay que considerar dos tipos de errores: el error debido a la precisión limitada del instrumento, que atenta a la reproducibilidad de la medición introduciendo un error aleatorio, y el debido a la validez, también limitada, que introduce un error sistemático. De modo esquemático se puede decir que la validez tiene que ver con la cuestión de si el instrumento mide lo que debe medir, mientras que la precisión tiene que ver con cuánto se aproxima la medida al valor real de la magnitud. En ambos casos es siempre una cuestión de grado, pues no existen instrumentos infinitamente precisos y válidos: hay sólo instrumentos más precisos y/o válidos que otros.

En cuanto a la reproducibilidad, llamada también concordancia, se distingue entre la reproducibilidad del mismo instrumento en dos instantes de tiempo diferentes y se habla de concordancia o consistencia interna o intraobservador (p. ej., un radiólogo ¿clasifica igual la misma radiografía estudiada hoy y 2 meses después?), y la reproducibilidad del mismo instrumento usado en diferentes condiciones (p. ej., dos radiólogos diferentes ¿clasifican del mismo modo la misma radiografía?), se habla entonces de concordancia o consistencia externa o interobservador. Este ejemplo es útil también para resaltar que en clínica el término "instrumento de medida" se suele usar en sentido amplio; aquí no es sólo el aparato de rayos usado para obtener la imagen, sino el conjunto formado por el aparato y el observador que la interpreta.

Tabla 1.

	Radiólogo A		
Radiólogo B	**Neumonía**	**No neumonía**	**Total**
Neumonía	4 (a)	6 (b)	r = a + b 10
No	10 (c)	80 (d)	s = c+ d 90
Total	14 t = a + c	u = b + d 86	N = a + b + c + d 100

El procedimiento para evaluar la reproducibilidad de un instrumento consiste en comparar entre sí distintas medidas de un mismo objeto y evaluar su grado de acuerdo (cuanto más se parezcan estas medidas entre sí, más preciso es el instrumento). En el ejemplo anterior habría que comparar los resultados de la evaluación de una serie de radiografías por el mismo radiólogo en dos instantes de tiempo (concordancia interna) o por dos radiólogos diferentes (concordancia externa). La manera de expresar los resultados de esta comparación depende del tipo de variable implicada; en el caso de una variable binaria (tipo sí o no; p. ej., enfermo o no enfermo) el índice más sencillo es la proporción de acuerdos observados. Supongamos que en un estudio para evaluar la concordancia entre dos radiólogos que interpretan radiografías de tórax, clasificando cada una como neumonía sí o no, ofrece los resultados de la tabla 1. La proporción de acuerdo observado es P_o = (80 + 4)/100 = 0,84. Este índice es muy intuitivo y fácilmente interpretable: tomará valores entre 0 (total desacuerdo) y 1 (máximo acuerdo). Sin embargo, como indicador de reproducibilidad tiene el inconveniente de que, aun en el caso de que los dos observadores clasifiquen con criterios independientes (p. ej., un radiólogo con todo su leal saber y entender y el otro tirando un dado al aire), se produciría un cierto grado de acuerdo por azar. Puede haber coincidencia en el resultado sin que exista nada más que el puro azar, no el mismo criterio en la decisión. Es deseable que un índice de concordancia tenga en cuenta este hecho y que, de algún modo, indique el grado de

Correspondencia: Dr. V. Abraira.
Unidad de Bioestadística Clínica. Hospital Ramón y Cajal.
Ctra. Colmenar, km 9,100. 28034 Madrid.
Correo electrónico: victor.abraira@hrc.es

SEMERGEN: 2000; 27: 247-249.

Puntos clave

- El *índice kappa* (κ) se usa para evaluar la *concordancia* o *reproducibilidad* de instrumentos de medida cuyo resultado es categórico (2 o más categorías).
- El *índice kappa* (κ) representa la proporción de acuerdos observados más allá del azar respecto del máximo acuerdo posible más allá del azar.
- En la interpretación del *índice kappa* (κ) hay que tener en cuenta que el índice depende del acuerdo observado, pero también de la prevalencia del carácter estudiado y de la simetría de los totales marginales.

acuerdo que existe por encima del esperado por azar. En este sentido Cohen[3] propuso el denominado índice kappa (κ), que definió como:

$$\kappa = \frac{P_o - P_e}{1 - P_e}$$

siendo P_o la proporción de acuerdos observados y P_e la proporción de acuerdos esperados en la hipótesis de independencia entre los observadores, es decir, de acuerdos por azar. A partir de la tabla 1, $P_o = (a + d)/N$ y $P_e = (rt + su)/N^2$. La interpretación de este índice se facilita mediante su representación gráfica[4]. En la figura 1 se observa que el índice κ representa la proporción de concordancia observada más allá del azar, respecto de la máxima concordancia posible más allá del azar.

En el ejemplo:

$$P_e = \frac{14 \times 10 + 86 \times 90}{100^2} = 0{,}788$$

y por lo tanto

$$\kappa = \frac{0{,}84 - 0{,}788}{1 - 0{,}788} = 0{,}245$$

es decir, el grado de acuerdo, una vez corregido el debido al azar, es mucho más modesto (24,5%) que lo que indicaba el 84% de acuerdo "crudo". Landis y Koch[5] propusieron, y desde entonces ha sido ampliamente usada, la escala de valoración del índice κ que figura en la tabla 2.

*En la página 270 de este número, el autor del artículo al que se hace referencia contesta en una "Carta al director" a los comentarios del Dr. Abraira. Asimismo, en la página 272 la Dra. Pérez analiza la situación desde la sección "El rincón del autor y del lector".

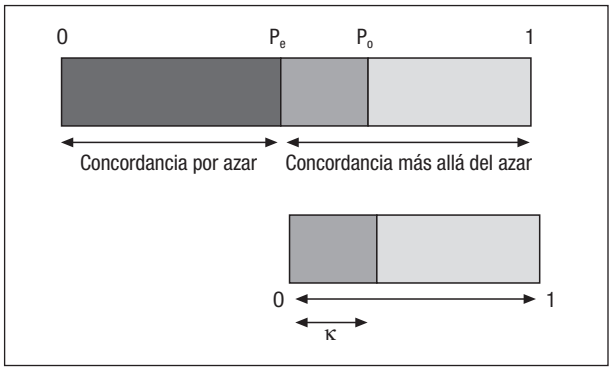

Figura 1. Representación gráfica del índice kappa.

Desde la propuesta inicial de Cohen[3] el índice κ ha sido progresivamente generalizado a clasificaciones multinomiales (con más de dos categorías), ordinales, a más de dos observadores, a diseños incompletos y a todas estas situaciones combinadas[6], generalizaciones que suponen una mayor complejidad en el cálculo pero que mantienen la misma interpretación. Esta interpretación está dificultada por algunos efectos poco intuitivos. En primer lugar, el índice κ depende de la prevalencia del carácter observado[7]: cuanto más cerca esté de 0 o de 1, menor es el índice κ para igual proporción de acuerdos observados. En segundo lugar, depende de la simetría de los totales marginales[7]: en igualdad de acuerdos observados, cuanto menor sea la diferencia entre las prevalencias observadas por cada observador, menor es el índice κ. El pequeño valor de κ para los datos de la tabla 1 se matiza a la luz de estos efectos: estamos en la peor de las situaciones posibles: baja prevalencia y similar para ambos observadores (0,14 para el radiólogo A y 0,10 para el B).

En un interesante artículo* recientemente publicado en esta Revista[8], se estudia la concordancia en el diagnóstico de nevus melanocíticos entre atención primaria (AP) y atención especializada (AE), y se encuentra un índice κ muy bajo, inferior al hallado en estudios similares, según los propios autores comentan. Aunque no se dan detalles de cómo se ha calculado el índice, la distribución de los diagnósticos alternativos (hay 25 juicios clínicos distintos en AP y 12 en AE) indica que en este estudio están presentes tanto el primer efecto comentado antes (prevalencias cercanas a 0, o incluso 0 si se han considerado todos los juicios clínicos para estimar el índice κ) como el segundo (prevalencias similares); en consecuencia, el índice κ estará fuertemente "penalizado" y podría ser ésta la causa de su bajo valor.

Tabla 2.

Kappa (κ)	Grado de acuerdo
< 0,00	Sin acuerdo
0,00-0,20	Insignificante
0,21-0,40	Mediano
0,41-0,60	Moderado
0,61-0,80	Sustancial
0,81-1,00	Casi perfecto

BIBLIOGRAFÍA

1. Sackett DL. A primer on the precision and accuracy of the clinical examination. JAMA 1992; 267: 2638-2644.
2. Abraira V. Errores en las mediciones y clasificaciones clínicas: precisión y validez. URL: http://www.hrc.es/bioest/Intro_errores.html [último acceso: 29 de enero de 2001].
3. Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas 1960; 20: 37-46.
4. Sackett DL, Hayes RJ, Guyatt G, Tugwell P. Epidemiología clínica. Ciencia básica para la medicina clínica (2.ª ed.). Buenos Aires: Editorial Médica Panamericana, 1994.
5. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977; 33: 159-174.
6. Abraira V, Pérez de Vargas A. Generalization of the kappa coefficient for ordinal categorical data, multiple observers and incomplete designs. Qüestiió 1999; 23: 561-571.
7. Feinstein AR, Cicchetti DV. High agreement but low kappa: I. The problems of two paradoxes. J Clin Epidemiol. 1990; 43: 543-549.
8. Rodríguez Caravaca G, Villar del Campo C, González Mosquera M, Úcar Corral E, González Piñeiro B, López Bran E. Concordancia diagnóstica entre atención primaria y atención especializada al evaluar nevus melanocíticos. SEMERGEN 2000; 26: 428-431.