6.6

Rev. Psiquiatría Fac. Med. Barna., 14,5,258-264 (1987)

Fiabilidad y coeficientes de acuerdo en el diagnóstico psiquiátrico

A. Aluja*

RESUMEN

En la literatura actual sobre investigación clínica psiquiátrica se impone la utilización de estadísticos que informen sobre el grado de acuerdo en la evaluación diagnóstica. Este trabajo revisa los estadísticos de uso más común analizando las diferencias y analogías entre ellos en función de la tasa de base y prevalencia de la condición clínica estudiada.

Palabras clave: Kappa, tasa de base, fiabilidad, coeficiente de acuerdo, sesgo.

SUMMARY

Current literature on psychiatric research imposes the use of statistics which reveal degree of agreement in clinical evaluation. This article reviews the most commonly used statistics, analysing the differences and similarities between them with regard to the base rates and prevalence of the clinical condition under study.

Key words: Kappa, base rates, reliability, coefficients of agreement, bias.

Introducción

Hasta los años sesenta no existe una clara preocupación en incorporar el concepto de fiabilidad al diagnóstico en psicopatología (Spitzer, Fleiss, Endicott y Cohen, 1967). Este interés despertó con la convicción de que si una clasificación psiquiátrica no es fiable, el desarrollo de una ciencia que trate de los trastornos mentales no será posible.

El diagnóstico o sistema de clasificación tiene dos propiedades primarias que son la validez y la fiabilidad. Se entiende por validez la utilidad del sistema para sus diversos fines y por fiabilidad, la consistencia o compatibilidad con la que los sujetos son clasificados para sus diversos fines (Blashfield, 1984). En el diagnóstico psiquiátrico los objetivos de la clasificación son la comunicación de rasgos clínicos, etiología, curso de la enfermedad y tratamiento. La fiabilidad es necesaria en este sistema hasta el punto de que no hay ninguna garantía de que un sistema no fiable sea válido, pero con toda seguridad un sistema no fiable será inválido (Spitzer y Fleiss, 1974).

A efectos de aumentar la fiabilidad del diagnóstico se propusieron criterios explícitos (Feighner, 1972), enfocados a aumentar el acuerdo interclínico y asignar categorías de diagnóstico. Con el objeto también de superar la vaguedad y la subjetividad del diagnóstico se han creado entrevistas estructuradas y sistemas computerizados de evaluación. Las entrevistas estructuradas han demostrado que son al-

* Escuela Profesional de Psicología Clínica, Facultad de Medicina, Universidad de Barcelona.

tamente superiores a la entrevista clínica tradicional, alcanzándose mediante el uso de las primeras coeficientes de fiabilidad mucho mayores (Endicott y Spitzer, 1972).

Para la medida de la fiabilidad o grado de acuerdo en el diagnóstico se han venido usando diversos estadísticos, tales como el coeficiente de correlación producto-momento de Pearson y otros. Esta prueba estadística ha demostrado serios inconvenientes cuando se usa como prueba de fiabilidad. Uno de estos inconvenientes es que no es aplicable cuando los evaluadores utilizados para evaluar a un paciente son diferentes a los utilizados para evaluar a otro. El segundo inconveniente es que no es adaptable al caso de dos o más evaluadores evaluando el mismo caso. No permite tampoco separar los análisis de las diferentes fuentes de no fiabilidad (error de medida debido al azar, diferencias entre los clasificadores, etc.). De estas mismas limitaciones adolecen otros estadísticos como el Ji Cuadrado o el coeficiente de Contingencias.

Los primeros pasos hacia la consecución de un estadístico que midiera la fiabilidad de forma óptima fueron dados por Ebel (1951), Barko (1966), Cohen (1960), Spitzer y Cohen (1968) y Fleiss (1966).

Cohen (1960), toma como punto de partida un modelo estadístico relativamente sencillo en cada una de las fuentes de variación que se encuentran bajo investigación en un trabajo de fiabilidad tiene asociado un componente de la varianza. Cuando se aplica un análisis de la varianza a los datos resultantes cada uno de estos componentes puede ser estimado por separado y luego combinado con un coeficiente de correlación interclase. La proporción del componente de la varianza asociado con su auténtica variabilidad sujeto a sujeto resulta ser la suma de todos los componentes de la varianza. Este razonamiento matemático se cristalizó con una fórmula estadística llamada «kappa».

Coeficiente de correlación intragrupo «kappa»

Antes de la aparición del estadístico «k» los estudios de fiabilidad caían fácilmente en algún error estadístico. Los más frecuentes eran los siguientes (Barko y Carpenter, 1976):

- Las fórmulas para calcular el porcentaje de acuerdo variaban según los autores por lo que se hacía difícil la comparación de los estudios.
- El uso de estadísticos de porcentaje de acuerdo no corregía la proporción de acuerdo debida al azar. Se ignoraban los índices base.
- Las propiedades de estas medidas con porcentajes de acuerdo eran desconocidas, por lo que no era posible realizar pruebas de significado.

Barko y Carpenter demostraron en una revisión (1976) que la «k» tenía la ventaja de superar los tres problemas mencionados porque tenía una fórmula bien definida, fue diseñada para corregir el azar y era indiferente a los índices base. Además, la «k» permite realizar test de significado y comparaciones.

El precursor del estadístico «k» fue Scott (1955), aunque la fórmula final fue propuesta por Cohen (1960). Spitzer y Fleiss tienen el mérito de haber difundido y generalizado su uso a partir de su conocido artículo «A re-analysis of the reliability of psychiatric diagnosis» (1974).

La fórmula de Cohen es la siguiente:

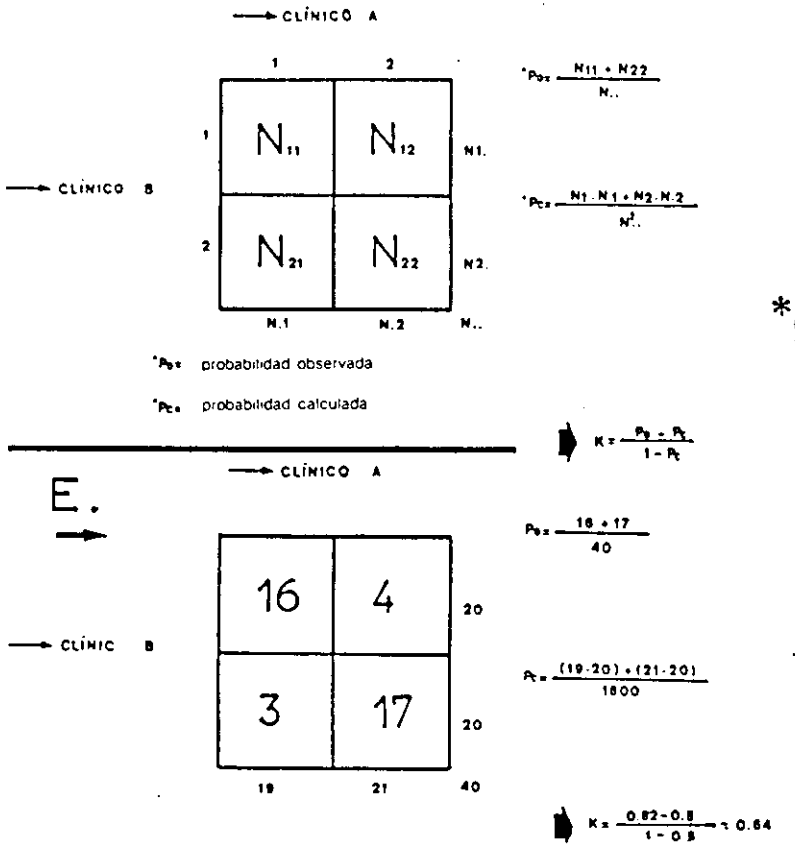
$$k = \frac{Po - Pc}{1 - Pc}$$

$$Po = \frac{A + D}{N} \qquad Pc = \frac{P1 \cdot Q1 + P2 \cdot Q2}{N\ (2)}$$

Po se define como la proporción de acuerdo observada entre los clínicos (acuerdo en la presencia del diagnóstico o ausencia de él). Pc es la proporción de acuerdo esperada por el azar (Cuadro I).

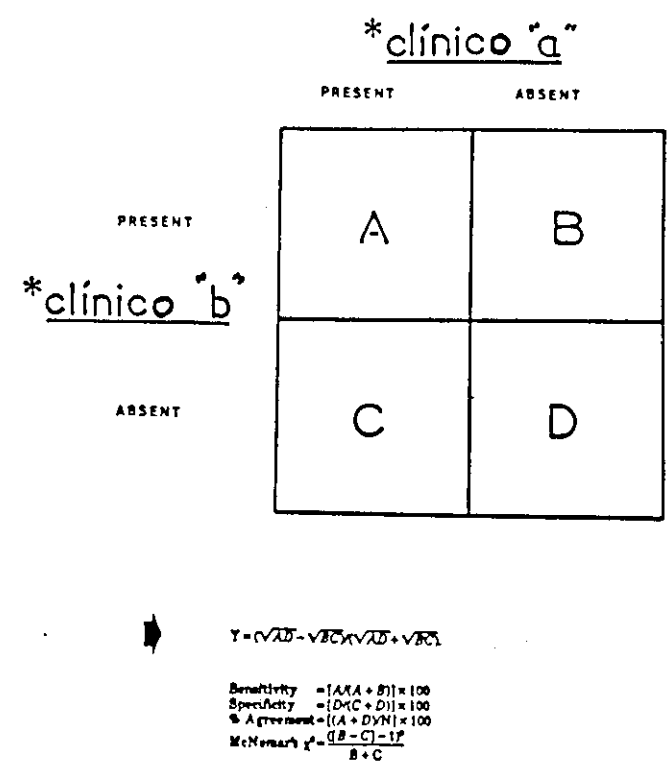
CUADRO I

ESTADÍSTICO "K" (Cohen)



CUADRO II

ESTADÍSTICO "Y" (Yule)



Co

E
Yul
rior
col
teri
de
tasa
mo
enf
rés
tad
pet
mc
dia
la
dic
dic
inv
U
cif
ba
ve
en
ec
er

La
do la
do a
+1(
yore
fiabi
dicar
lores
indic
D
virtu
renc
clini
raliz
la a
tien
cas
(197
su i
al f

La proporción resultante dada, aplicando la fórmula, oscila entre —1 (desacuerdo absoluto), 0 (acuerdo igual al azar) y +1 (acuerdo absoluto). Los valores mayores de 0,75 son indicativos de buena fiabilidad. Valores entre 0,50 y 0,75 indican una fiabilidad aceptable, y los valores por debajo de 0,50 son generalmente indicativos de baja fiabilidad.

Durante los años setenta la «k» se convirtió en la medida estadística de preferencia para los estudios de acuerdo interclínicos, siendo actualmente de uso generalizado. A pesar de la popularidad y de la aceptación general de estadístico, éste tiene sus críticos y detractores. Las críticas más conocidas son las de Maxwell (1977) y Janes (1979). Estas críticas, por su interés, serán expuestas y comentadas al final.

Coeficiente de coligación de Yule «Y»

El estadístico «Y» fue propuesto por Yule (1912) a principios de siglo y posteriormente fue recuperado por Helzer y cols. (1985). Este estadístico posee características complementarias a la «k» y puede utilizarse conjuntamente cuando la tasa de datos es variable. Se entiende como tasa de base la tasa verdadera de una enfermedad o condición clínica de interés en una población determinada. Un estadístico de concordancia que sea independiente de la tasa de base tendrá el mismo valor para un nivel dado de acuerdo diagnóstico entre evaluadores medido por la especificidad y sensibilidad, independientemente de la prevalencia de la condición estudiada en la población que se investigue. La «k» muestra una curva de U invertida para una sensibilidad y especificidad dadas a medida que la tasa de base se mueve de abajo a arriba, de niveles inferiores a superiores.

La «k» es útil para muestras clínicas en las cuales la proporción de personas con el diagnóstico que se estudia se sitúa entre el 20% y 60%. Sin embargo, en una muestra de la población general (estudios epidemiológicos), la tasa de base tiende a ser inferior al 10% y la especificidad tiende a ser mayor que la sensibilidad. Dentro de este rango los valores de la «k» son muy inestables, con valores que bajan bruscamente a medida que baja la tasa de base, incluso cuando la sensibilidad y especificidad permanecen constantes. Este es un problema que ha sido revisado recientemente (Spitznagel, 1986; Crove y cols., 1981).

El estadístico «Y» posee la peculiaridad de expresar el acuerdo entre evaluadores corrigiendo el azar, pero sin confundir la concordancia con la prevalencia del diagnóstico. El coeficiente de coligación de Yule se mantiene estable incluso con tasas de base inferiores al 2%. La «Y» tiene la misma gama de valores que «k» (—1 a +1), por lo que es comparable al valor «k» en el caso en que en ésta la prevalencia hubiera sido óptima. Los valores bajos de «Y» significan que existe una baja concordancia, en cambio los valores bajos en «k» pueden estar causados únicamente por tasas de prevalencia bajas. «Y» es de fácil uso y se basa como la «k» en una proporción de productos cruzados en una tabla de doble entrada (véase Cuadro II).

La única característica indeseable que tiene el estadístico de Yule es que si una única diagonal (A ó D) de la tabla es 0, «Y» toma el valor —1, y si una única diagonal (B ó C) es 0, «Y» toma el valor de +1. Para corregir este inconveniente se utiliza un método pseudobayesiano para estimar las probabilidades no 0 de las cuatro casillas de la tabla para que se pueda calcular la «Y» con más significado, incluso cuando una de las celdillas equivale a 0. Desafortunadamente, esta información no está disponible todavía (Spitznagel, 1986 —en prensa).

Con los mismos datos de la tabla de 2×2 se puede calcular la sensibilidad, especificidad y el sesgo inter-evaluadores (McNemar's).

Prueba de McNemar's para la significación de cambios

La prueba de McNemar's es una prueba pensada para diseños «pre-post» en los que se usa a cada persona como su propio control. Por tanto, sólo puede emplearse en muestras con datos dependientes. Para comprobar la significación de los cambios observados se confeccionó una tabla de doble entrada de frecuencias que represente al primero y al segundo conjunto de respuestas de los mismos individuos. Se usan los signos + y — para simbolizar las respuestas diferentes. Los cambios entre la primera y segunda clasificación (primer clínico y segundo clínico) aparecen en las celdillas B y C, puesto que en las celdillas A y D se registran las frecuencias no cambiantes (acuerdos).

Teniendo en cuenta que B + C son el número total de clasificaciones cambiantes (desacuerdos), se espera conforme a la hipótesis de nulidad que 1/2(B + C) frecuencias cambiarán en una dirección y 1/2(B + C) cambiarán en otra. Es decir, 1/2(B + C) es la frecuencia esperada conforme a la Ho en ambas celdillas B y C.

Teniendo en cuenta que la técnica Ji Cuadrado prueba si las frecuencias observadas están suficientemente próximas a las esperadas que podrían ocurrir conforme a Ho. La hipótesis de nulidad puede probarse mediante la fórmula:

$$\chi^2 = \sum_{i=1}^{k} \frac{(Oi - Ei)^2}{Ei}$$

Oi = al número observado de casos clasificados en la categoría de i.

Ei = al número de casos esperado en la categoría de i conforme a Ho.

En la prueba de McNemar's de significación de cambios sólo interesan los cambios de las celdillas B y C, por tanto es posible operar hasta obtener una corrección en los siguientes términos:

$$\chi^2 = \frac{([B - C] - 1)^2}{B + C} \text{ Con 1 g. l.}$$

Hay que restar 1 del valor absoluto de la diferencia B y C antes de elevar tal diferencia al cuadrado. Los valores obtenidos por la fórmula corregida se comparan, para determinar su significación, con la tabla de «valores críticos de Ji Cuadrado» que elaboraron Fisher y Yates (Siegel, 1975). La fórmula de McNemar's se ha usado para obtener información sobre el sesgo producido por dos evaluadores al evaluar de forma independiente a los mismos sujetos (Helzer, Robins y cols., 1985), sin tener en cuenta el grado de acuerdo global.

Coeficiente de error casual de acuerdo (RE)

Las críticas de Maxwell y Janes (1977, 1979), que se insinuaron más arriba, se basan en que la forma en que «k» asumía el azar no podía ser siempre plausible. Maxwell basa su razonamiento en los resultados hipotéticos: los diagnósticos de dos clínicos a 100 pacientes a los cuales había que atribuirles o no el diagnóstico de alcoholismo. (Tabla I).

En 4 casos ambos clínicos están de acuerdo en que los pacientes son alcohólicos y en 64 en que no lo son, por lo que el porcentaje de acuerdo es del 68%. Si consideramos el porcentaje específico de acuerdo según Helzer éste es sólo del 11%, puesto que se calcularía de la siguiente manera:

$\frac{4}{4 + 28 + 4}$, ya que los 64 casos negativos son ignorados como acuerdo.

En la estadística «k» también se calcula la proporción observada de forma similar al porcentaje clásico inicial, Po = $\frac{4 + 64}{100} = 68$. En cambio, el cálculo de la Pc consiste en sumar el producto de las proporciones marginales:

$$Pc = P1 \cdot Q1 + P2 \cdot Q2$$

P1 = casos considerados alcohólicos por el primer clínico
P2 = casos considerados como no alcohólicos por el primer clínico
Q1 = casos considerados alcohólicos por el segundo clínico
Q2 = casos considerados no alcohólicos por el segundo clínico

dica que no existe una homologación absoluta a la hora de aplicar los criterios de diagnóstico entre los dos evaluadores. Maxwell considera tal diferencia como un error informativo más que una diferencia debida al azar. Este autor sugirió un planteamiento alternativo para el cálculo del coeficiente de acuerdo:

Suponiendo que A1 es la proporción de pacientes en la que ambos clínicos es-

TABLA 1

		CLINICO A		
		Alcohólico	No alcohólico	T
CLINICO B	Alcohólico	4	4	8
	No alcohólico	28	64	92
	TOTAL	32	68	100

Según Maxwell existe un desacuerdo importante por parte de los dos clínicos en cuanto a la presencia del diagnóstico de alcoholismo, y el cálculo de la proporción debida al azar es también elevado (Pc = 65). Los índices base con los cuales los clínicos utilizan el diagnóstico dan como resultado un acuerdo al azar del 65%. Aplicando la fórmula de la «k» se obtiene un coeficiente de 0,8. Evidentemente, este resultado difiere del porcentaje clásico de acuerdo inicial del 68%. Maxwell (1977), se interesó por la forma en que la «k» asume implícitamente el azar para ser efectivo. El valor Pc de la estadística «k» depende de los índices base por lo cuales, como se ha dicho, los dos clínicos utilizan el diagnóstico. En el caso del ejemplo, los índices base eran muy diferentes (P1 = 32; Q1 = 08). Maxwell cree que si los diagnósticos efectuados por los dos clínicos fuesen realmente casuales, P1 sería igual que Q1. Esto in-

tán de acuerdo de forma «segura», en su condición de alcohólicos, y Ao es la proporción de sujetos no alcohólicos. Los casos restantes (G) son aquellos que los dos clínicos «han adivinado». Si Pa se define como la proporción observada de casos para los cuales los clínicos estuvieron de acuerdo, entonces, según Maxwell:

$$Pa = A1 + \frac{G}{4};\ Pa = A1 + \frac{1 - A1 - Ao}{4}$$

$$A1 + Ao + G = 1$$

Maxwell sugería que la proporción de casos de los cuales los dos clínicos están de acuerdo (4 en el ejemplo) es la suma de la proporción para los cuales están de acuerdo sin ninguna duda, más 1/4 de los casos cuestionables. De esta manera los casos cuestionables se reparten de igual forma las cuatro celdillas del acuerdo.

Estos razonamientos matemáticos no son compartidos, por poco claros, por muchos clínicos (Dewey, 1983), pero han servido de base a un nuevo estadístico denominado «coeficiente de error casual de acuerdo» (RE) (Maxwell, 1977). Su estadística aplicada al ejemplo hipotético es la siguiente:

$$RE = AI = Ao;\ RE = - \ .\ 12 + \ .\ 48;\ RE = \ .\ 36$$

El valor RE es más alto que «k» (RE = . 36); «k» = 0,8), por lo cual se otorga un mayor acuerdo clínico. Janes (1979) demostró que la diferencia más importante entre RE y «k» tiene lugar cuando los índices base de los clínicos son relativamente distintos. Cuando los índices base son casi iguales, RE y «k» representan índices comparables.

La kappa continúa siendo el estadístico de acuerdo interclínicos más usual. No obstante, la propuesta de Helzer y cols. (1985) nos parece plausible, pues la «Y» de Yule es de utilidad en los estudios de comunidad o epidemiológicos en los que la prevalencia del trastorno investigada tienda a ser al 20%. El test de McNemar's nos proporciona información de forma independiente al acuerdo global sobre la tendencia de los evaluadores a infra o supravalorar el diagnóstico.

Bibliografía

Barko, J. J. (1966) (citado por Spitzer y cols., 1982).

Barko, J. J. y Carpenter, W. T.: «On the methods and theory of reliability». *J. Nerv. Ment. Dis.*, 163: 307-317, 1976.

Blashfield, R. K.: *The Classification of Psychopatology. Neo-Kraepelinian and Quantitative Approches.* Plenum Press. N.Y., 1984.

Cohen, J.: «A Coefficient of agreement for nominal scales». *Educat. Psychol. Meas.*, 20: 37-46, 1960.

Crove, W. M.: «Reliability studies of psychiatric diagnosis». *Arch. Gen. Psychiat.*, 38: 408-413, 1981.

Ebel, R. L. (1951) (citado por Spitzer y cols., 1982).

Endicott, J. y Spitzer, R. L. (1972) (citado por Spitzer y cols., 1982).

Feighner, J. P.; Robins y cols.: «Diagnostic Criteria for use in psychiatric research». *Arch. Gen. Psychiat.*, 26: 56-63, 1972.

Fleiss, J. L. (1966 y 1971) (citado por Spitzer y cols., 1982).

Helzer, H. y cols.: «A comparison of clinical and diagnosis interview schedule diagnostic». *Arch. Gen. Psychiat.*, 42: 657-666, 1985.

Janes, C. L.: «Agreement measurement and the judgement process». *J. Nerv. Ment. Dis.*, 167: 343-347, 1979.

Maxwell, A. E.: «Coefficients of agreement between observers and their interpretation». *Brit. J. Psychiat.*, 130: 79-83, 1977.

Siegel, S.: *Non parametric statistics for the behavioral sciencies.* McGraw-Hill Book Company. N.Y., 1956.

Spitzer, R.; Fleiss, J.; Endicott, J. y Cohen, J.: «Quantification of agreement in psychiatric diagnosis». *Arch. Gen. Psychiat.*, 17: 83-87, 1967.

Spitzer, R. L. y Fleiss, J. L.: «A re-analysis the reliability of psychiatric diagnosis». *Brit. J. of Psychiat.*, 125: 341-347, 1974.

Spitzer, R. L. y Cohen, J. (1968) (citado por Spitzer y cols., 1982).

Spitzer, R. y cols.: «Problemas de clasificación, fiabilidad y validez». En: *Psicofarmacología. A los 30 años de progreso.* M. A. Lipton, A. Di Mascio y K. F. Killam (Eds.) Espaxs, Barcelona, 1975.

Spitznagel, E. L. y Helzer, J. E.: A proposed solution to the base problem in the kappa statistic. *Arch. Gen. Psychiat.* (en prensa).

Yule, G. U.: «On the methods of measuring association between two attributes». *J. Roy. Stat. Soc.*, 75: 518-642, 1912.