Medidas de concordancia para variables cualitativas

Preparado por Luis M. Molinero (Alce Ingeniería)
CorreoE: **bioestadistica@alceingenieria.net**

Artículo en formato PDF

Septiembre 2001
Actualizado Diciembre 2001

En el artículo anterior se estudió como medir la concordancia entre diferentes métodos cuyo resultado es una variable cuantitativa. Ahora se trata de medir el grado de acuerdo entre varios métodos o evaluadores que clasifican al paciente (o el resultado de una observación) según una serie de posibilidades (categorías) mutuamente excluyentes. El caso más sencillo se presenta cuando la variable cualitativa es dicotómica (dos posibilidades) y se está comparando dos métodos de clasficación (por ejemplo dos escalas clínicas). Esta situación se puede representar en una tabla de frecuencias:

		Método B		
		Positivo	Negativo	
Método A	Positivo	a	c	f1
	Negativo	b	d	f2
		c1	c2	n

La medida más simple de concordancia es la **proporción de coincidencias frente al total de sujetos**: (a + d) / n.

Pero resulta que aunque no existiera ninguna relación entre los dos métodos de clasificación, está claro que es previsible que encontremos algún grado de concordancia entre ellos por puro azar. Así, si el método A consiste en clasificar al paciente con resultado positivo si sale cara al lanzar una moneda al aire y cruz en el caso contrario, y hacemos lo mismo en el método B (con otra moneda diferente), es previsible encontrar en promedio del orden de un 50 % de coincidencias.

Supongamos que el sistema A es un método científico de diagnóstico y el método B es la opinión de un "vidente";también ahora es previsible encontrar un cierto grado de concordancia debido en parte al azar.

Con el fin de determinar hasta qué punto la concordancia observada es superior a la que es esperable obtener por puro azar, se define el **índice de concordancia kappa** de la siguiente manera:

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

donde Po es la *proporción de concordancia observada* (en tanto por 1) y Pe es la *proporción de concordancia esperada por puro azar*. En caso de acuerdo perfecto la proporción de concordancia será 1, por lo que 1−Pe representa el margen de acuerdo posible no atribuíble al azar. De ese margen nosotros observamos probablemente sólo una parte Po−Pe, salvo que haya acuerdo perfecto Po=1.

Así pues en caso de concordancia perfecta el valor de kappa es 1; si la concordancia observada es igual a la esperada kappa vale 0; y en el caso de que el acuerdo observado sea inferior al esperado el índice kappa es

menor que cero.

Para calcular Pe, la concordancia esperada, el razonamiento es el siguiente: de acuerdo con la tabla anterior la probabilidad de que el método A clasifique a un sujeto como positivo podemos estimarla como f1/n; mientras que la correspondiente probabilidad del método B la estimaremos como c1/n. Si consideramos que existe independencia entre ambos métodos de clasificación, la probabilidad de que coincidan clasificando al mismo sujeto como positivo será entonces el producto de las dos probabilidades (sucesos independientes). Aplicando el mismo razonamiento calculamos la probabilidad de que se produzca acuerdo entre los métodos al clasificar a un sujeto como negativo, y entonces la probabilidad de acuerdo cualquiera de las dos clasificaciones será la suma de ambos valores, esto es:

$$P_e = \frac{f_1 \cdot c_1 + f_2 \cdot c_2}{n^2}$$

El coeficiente kappa fue propuesto originalmente por Cohen (1960) para el caso de dos evaluadores o dos métodos, por lo que a menudo se le conoce como kappa de Cohen, y fue generalizado para el caso de más de dos evaluadores por Fleiss, por lo que a veces también se habla del índice kappa de Fleiss.

Landis y Koch propusieron unos márgenes para valorar el grado de acuerdo en función del índice kappa:

kappa	**grado de acuerdo**
< 0	sin acuerdo
0 − 0,2	insignificante
0,2 − 0,4	bajo
0,4 − 0,6	moderado
0,6 − 0,8	bueno
0,8 − 1	muy bueno

Este índice se puede generalizar para clasificaciones multinomiales (más de dos categorías) y para más de dos evaluadores, siendo similar su interpretación.

En el caso de más de dos categorías, además del índice de concordancia global puede ser interesante determinar el **grado de concordancia específico en alguna de las categorías** (o en todas), lo que equivale a convertir el resultado posible en dos únicas respuestas: se clasifica al paciente en la categoría de interés o se clasifica en alguna de las restantes. De esta manera para cada una de las categorías vamos convirtiendo la tabla original en tablas 2x2 y podemos entonces calcular el valor del correspondiente índice kappa como si de una variable dicotómica se tratara.

La gran utilización del índice de concordancia kappa en la literatura médica se debe probablemente tanto a la facilidad de cálculo, como a su clara intepretación; no obstante, tiene sus problemas y limitaciones que pueden consultarse por el lector interesado en la bibliografía que acompaña este artículo. El principal problema de esta medida de concordancia radica en que está pensada para **clasificaciones nominales**, en las que no existe un orden de graduación entre las diferentes categorías. Cuando esto no es así, pensemos por ejemplo en una clasificación del tipo *Muy grave − grave − leve − sin importancia*, donde no es lo mismo que el desacuerdo se produzca clasificando como *sin importancia* por un evaluador y *leve* por otro, a que uno de ellos clasifique como *sin importancia* y otro como *muy grave*. El índice kappa hasta ahora descrito únicamente tiene en consideración si hay o no acuerdo, esto es si se clasifica o no al sujeto en la misma categoría, por lo que a la hora de calcularlo pesan por igual las dos situaciones anteriormente descritas.

Si deseamos tener en cuenta el hecho de que estamos manejando **variables ordinales** para calcular una medida de concordancia, existen diferentes posibilidades. La más sencilla es calcular individualmente la concordancia en cada categoría, tal y como se comentó más arriba; pero de esta forma seguimos sin ponderar el nivel de desacuerdo global según esa clasificación ordinal.

Otro enfoque más global consiste en asignar un **peso** a las diferentes posibilidades de desacuerdo, de tal manera que se considere como más importante un desacuerdo entre categorías alejadas que entre las próximas. Este peso variará entre 0 (acuerdo, misma categoría) y 1 (desacuerdo con categorías extremas). El problema surge a la hora de determinar esos pesos, ya que el valor de concordancia obtenido será diferente según los pesos utilizados. En uno de los enlaces seleccionados se describen los pesos más habitualemente utilizados (lineales o bicuadrados) y que suelen proporcionar por defecto los programas de ordenador.

Enlaces de interés

- Statistical Methods for Rater Agreement
  La mejor página que he encontrado sobre este tema en Internet. Hay una detallada explicación de todos los análisis de concordancia entre evaluadores y dispone asimismo de un software gratuito.
- Calculadora on−line para determinar el índice kappa, y explicación del mismo
- Indice kappa con "pesos". V.Abraira

Bibliografía seleccionada

- **Statistical methods for rates and proportions**
  Joseph L. Fleis
  Ed.John Wiley
  New York 1981
- Fleiss JL (1971) **Measuring nominal scale agreement among many raters**. Psychol Bull Vol 76, nº 5, 378−382

Indice de artículos

Principio de la página