Capítulo 8

ESTIMACIÓN DE PARÁMETROS DE UN TEST DIAGNÓSTICO BINARIO BAJO UN MUESTREO TRANSVERSAL

José Antonio Roldán-Nofuentes
Bioestadística, Facultad de Medicina, Universidad de Granada, 18071, España
E-mail: jaroldan@ugr.es

RESUMEN

Un test diagnóstico es una prueba médica que se aplica a un individuo para determinar si tiene o no una cierta enfermedad, y su uso es fundamental en la práctica clínica. En este trabajo se realiza una revisión de los parámetros de un test diagnóstico binario. Se analizan las propiedades de los parámetros y se estiman estos parámetros cuando el test diagnóstico binario y el gold estándar se aplican a todos los individuos de una muestra aleatoria. Los resultados se han aplicado al diagnóstico de la enfermedad coronaria.

Palabras Clave: Coeficiente kappa ponderado, Coeficiente kappa promedio, Especificidad, Razones de verosimilitud, Sensibilidad, Test diagnóstico binario, Valores predictivos.

ABSTRACT

A diagnostic test is a medical test that is applied to a patient in order to determine the presence or absence of a specific disease, and its use is essential in clinical practice. In this paper a review of the parameters of a binary diagnostic test is performed. The properties of the parameters are analysed and these parameters are estimated when the binary diagnostic test and the gold standard are applied to all of the individuals in a random sample. The results have been applied to the diagnosis of coronary artery disease.

Key Words: Weighted kappa coefficient, Average kappa coefficient, Specificity, Likelihood ratios, Sensitivity, Binary diagnostic test, Predictive values.

1. INTRODUCCIÓN

Un test diagnóstico es una prueba médica que se aplica a un paciente para determinar la presencia o ausencia de una cierta enfermedad. La mamografía para el diagnóstico del cáncer de mama o la ecocardiografía para el diagnóstico de la enfermedad coronaria son dos ejemplos de tests diagnósticos. En la práctica de la Medicina moderna el uso de los tests diagnósticos es fundamental, y aunque tal uso representa un bajo porcentaje en los costes en salud, sus resultados tienen una gran repercusión en la toma de decisiones médicas.

Los tests diagnósticos se clasifican en:

a). Binarios: dan lugar a dos posibles resultados, positivo (indica la presencia provisional de la enfermedad) o negativo (indica la ausencia provisional de la enfermedad). Ejemplo: ecocardiografía con dobutamina para el diagnóstico de la estenosis coronaria.

b). Cuantitativos o continuos: dan lugar a valores numéricos. Ejemplo: concentración de glucosa en sangre en ayunas para el diagnóstico de la diabetes.

c). Ordinales: dan lugar a valores ordinales. Por ejemplo, una clasificación de la presencia de la enfermedad en definitivamente no, probablemente no, probablemente sí, definitivamente sí.

El problema que presentan los tests diagnósticos es que estos se pueden equivocar en el diagnóstico de la enfermedad, por lo que es necesario determinar su exactitud, la cual se expresa en términos de probabilidades o de funciones de probabilidades. Una revisión sobre la estimación de parámetros de tests diagnósticos se puede ver en los libros de Zhou [1] y de Pepe [2]. Para evaluar la exactitud de un test diagnóstico es necesario disponer de un estimador insesgado de dicha exactitud, para lo cual se necesita determinar el estado de enfermedad (enfermedad presente o enfermedad ausente) de cada individuo, independientemente del resultado del test diagnóstico. El procedimiento mediante el cual se conoce el verdadero estado de enfermedad de un individuo se denomina gold estándar. Por tanto, un gold estándar es una prueba médica que determina de forma objetiva si un individuo tiene o no una cierta enfermedad. Biopsia para el diagnóstico del cáncer de mama o la angiografía coronaria para el diagnóstico de la enfermedad coronaria son dos ejemplos de gold estándar.

Por tanto existen dos métodos que permiten diagnosticar la enfermedad: el test diagnóstico y el gold estándar. El test diagnóstico puede cometer errores en el diagnóstico de la enfermedad, mientras que el gold estándar no.Entonces, ¿por qué no utilizar siempre el gold estándar? Las siguientes razones justifican el uso de los tests diagnósticos:

1. El test diagnóstico suele ser más económico que el gold estándar.
2. El uso de un gold estándar puede suponer un riesgo para el individuo. Por ejemplo, la angiografía coronaria puede provocar ictus, trombosis e incluso la muerte del paciente.
3. No siempre existe un gold estándar. Por ejemplo, algunas enfermedades psiquiátricas.

El objetivo que se persigue cuando se evalúa la exactitud de un test diagnóstico no es determinar si un individuo tiene o no una cierta enfermedad, sino estimar su exactitud.

En este trabajo se realiza una revisión de parámetros de interés de un test diagnóstico binario (por ser de los más comunes en la práctica clínica y por su interés desde la perspectiva de los datos categóricos, pues su análisis da lugar al estudio de tablas 2×2 o de mayor dimensión). En la Sección 2 se describen los parámetros de interés. En la Sección 3 se estiman dichos parámetros bajo un muestreo transversal. En la Sección 4, se presenta un ejemplo real, y en la sección 5 se concluye.

2. PARÁMETROS DE UN TEST BINARIO

Considérese una población de individuos que pueden tener o no una cierta enfermedad. Supóngase que se dispone de un gold estándar para el diagnóstico de dicha enfermedad, entonces se define D como la variable aleatoria binaria que modeliza el resultado del GE, de tal forma que $D=1$ (gold estándar positivo) cuando el individuo tiene la enfermedad y $D=0$ (gold estándar negativo) cuando el individuo no tiene la enfermedad. La probabilidad de que un individuo (elegido al azar en la población objeto de estudio) tenga la enfermedad se denomina prevalencia de la enfermedad (p), esto es

$$p = P\ \ D=1\ .$$

Considérese también un test diagnóstico binario cuyaexactitud es evaluada con respecto a un gold estándar. Sea T la variable aleatoria que modeliza el resultado del test diagnóstico, de tal forma que $T=1$ cuando el resultado del test es positivo (indicando la presencia provisional de la enfermedad) y $T=0$ cuando el resultado del test es negativo

(indicando la ausencia provisional de la enfermedad). A continuación se presentan parámetros del test diagnóstico binario.

2.1. Sensibilidad y especificidad

Los parámetros más importantes para evaluar la exactitud de un test binario son la sensibilidad (*Se*) y la especificidad (*Sp*). Ambos parámetros son las probabilidades condicionadas de clasificar correctamente a un individuo enfermo y a un individuo no enfermo respectivamente, y representan medidas de asociación o conformidad entre el test diagnóstico y el gold estándar.

La sensibilidad es la probabilidad de que el resultado del test binario sea positivo cuando el individuo está enfermo, es decir

$$Se = P\ \ T = 1 \big| D = 1\ \ .$$

Se verifica que

$$Se + P\ \ T = 0 \big| D = 1\ \ = 1,$$

siendo $P\ \ T = 0 \big| D = 1$ la probabilidad de un falso positivo (pues estando el individuo enfermo, el test ha dado un resultado erróneo).

La especificidad (*Sp*) es la probabilidad de que el resultado del test binario sea negativo cuando el individuo no está enfermo, es decir

$$Sp = P\ \ T = 0 \big| D = 0\ \ .$$

Se verifica que

$$Sp + P\ \ T = 1 \big| D = 0\ \ = 1,$$

siendo $P\ \ T = 1 \big| D = 0$ la probabilidad de un falso negativo (pues estando el individuo no enfermo, el test ha dado un resultado erróneo).

La sensibilidad y la especificidad del test binario dependen únicamente de la habilidad intrínseca del test diagnóstico para distinguir entre individuos enfermos e individuos no enfermos. Es decir, ambos parámetros dependen de las bases biológicos, físicas, químicas, etc…, con las que se ha desarrollado el test diagnóstico. Un test binario con una alta sensibilidad es útil para descartar la enfermedad ya que la probabilidad de un falso negativo es pequeña (y por tanto es útil para aplicar como procedimiento de rutina para el diagnóstico de la enfermedad). Un test binario con una alta especificidad es útil para confirmar la enfermedad ya que la probabilidad de un falso positivo es pequeña (y por tanto es útil para aplicarlo a individuos que son sospechosos de tener la enfermedad). A todo test binario hay que exigirle que

$$Se + Sp - 1 > 0\,.$$

Este parámetro $Se + Sp - 1$ se denomina índice de Youden, y sus valores varían entre -1 y 1. El índice de Youden presenta las siguientes propiedades:

a). Si el test binario no está relacionado con la enfermedad (aspecto que no es deseable), entonces la sensibilidad y la especificidad son complementarias y el índice de Youden es igual a cero. Esto es,

$$\text{si } P\ \ T = i \big| D = i\ \ = P\ \ T = i\ \ \Rightarrow \begin{cases} Se = P\ \ T = 1 \\ Sp = P\ \ T = 0 \end{cases} \Rightarrow Se + Sp = 1 \Rightarrow$$

$$Se = 1 - Sp \Rightarrow Y = 0.$$

b). Si el índice de Youden es menor que cero, entonces los resultados del tes binario se deben intercambiar; esto es, $T = 1$ debe ser un resultado negativo y $T = 0$ debe ser un resultado positivo.

Una de las principales utilidades clínicas de un test diagnóstico es el screening. Un screening médico es un protocolo que se utiliza para detectar una enfermedad en individuos asintomáticos. El screening permite identificar individuos enfermos de forma temprana y su objetivo es reducir los efectos de la enfermedad en el individuo. Por ejemplo, screening del cáncer de mama, el screening del cáncer de próstata, etc…Para que un test diagnóstico se pueda utilizar como test de screening, se deben verificar las siguientes características:

1. Características de la población:
   1.1. La prevalencia de la enfermedad sea suficiente grande.
   1.2. Sea susceptible de aplicación de diferentes pruebas médicas y tratamientos.
2. Características de la enfermedad:
   2.1. Morbilidad y mortalidad significativas.
   2.2. Que tenga un tratamiento eficaz y aceptable.
   2.3. Periodo presintomático detectable.
   2.4. Mejora con un tratamiento precoz.
3. Características del test diagnóstico:
   3.1. Alta sensibilidad y especificidad.
   3.2. Bajo coste.
   3.3. Su aplicación suponga un bajo riesgo para el individuo.
   3.4. Exista un gold estándar.

2.2. Razones de verosimilitud

Para un resultado cada resultado del test binario, la razón de verosimilitud es un cociente de dos probabilidades definido como

$$LR = \frac{P\ \ T = i \mid D = 1}{P\ \ T = i \mid D = 0},\ \ i = 0,1.$$

Si el resultado del test diagnóstico es positivo $i = 1$, la razón de verosimilitud, denominada razón de verosimilitud positiva, es el cociente entre la sensibilidad y uno menos la especificidad, esto es

$$LR^{+} = \frac{Se}{1 - Sp},$$

y representa el cociente entre la probabilidad de un resultado positivo del test diagnóstico en un individuo enfermo y la probabilidad de un resultado positivo del test diagnóstico en un individuo no enfermo. Cuando el resultado del test es negativo $i = 0$, la razón de verosimilitud, denominada razón de verosimilitud negativa, es

$$LR^{-} = \frac{1 - Se}{Sp},$$

y representa el cociente entre la probabilidad de un resultado negativo del test diagnóstico en un individuo enfermo y la probabilidad de un resultado negativo del test diagnóstico en un individuo no enfermo. Las razones de verosimilitud solamente dependen de la sensibilidad y especificidad del test diagnóstico, y sus valores varían

entre cero e infinito. Cuando el test binario y el gold estándar son independientes, entonces $LR^{+} = LR^{-} = 1$. Si el test diagnóstico clasifica correctamente a todos los individuos (enfermos o no enfermos) entonces $LR^{+} = \infty$ y $LR^{-} = 0$. Un valor $LR^{+} > 1$ indica que un resultado positivo del test diagnóstico es más probable en un individuo enfermo que en un individuo no enfermo (como es de esperar), y un valor $LR^{-} < 1$ indica que un resultado negativo del test diagnóstico es más probable en un individuo no enfermo que en uno enfermo (como también es de esperar). Las razones de verosimilitud cuantifican el incremento sobre el conocimiento de la presencia de la enfermedad mediante la aplicación del test binario. Antes de aplicar el test binario, la odds de que un individuo tenga la enfermedad es

$$\text{pre-test odds} = \frac{p}{1-p},$$

y después de aplicar el test, la odds de que un individuo tenga la enfermedad es

$$\text{post-test odds} = \frac{P\ \ D=1 \middle| T=i}{P\ \ D=0 \middle| T=i}, \quad i = 0,1.$$

Por consiguiente, las razones de verosimilitud relacionan la pre-test odds con las post-tests odds, esto es

$$\text{Post-test odds}\ \ T=1\ \ = LR^{+} \times \text{pre-test odds}$$
$$\text{Post-test odds}\ \ T=0\ \ = LR^{-} \times \text{pre-test odds},$$

y por tanto las razones de verosimilitud cuantifican el cambio en la odds de la enfermedad obtenido por el conocimiento de la aplicación del test diagnóstico binario.

2.3. Valores predictivos

La sensibilidad (especificidad)es la probabilidad de que el test diagnóstico dé un resultado positivo (negativo)cuando el individuo tiene (no tiene) la enfermedad. Sin embargo, cuando a un individuo se le aplica un test diagnóstico binario, el clínico o médico carece de información a priori sobre el verdadero estado de enfermedad (enfermedad presente o ausente) del individuo, por lo que desde el punto de vista práctico la sensibilidad y la especificidad no son parámetros que realmente interesen tanto al clínico como al individuo. Es decir, tanto el individuo como el clínico están interesados en determinar cuál es el estado de enfermedad sabiendo el resultado del test diagnóstico. Los parámetros que aportan esta información son el valor predictivo positivo y el valor predictivo negativo, que son los parámetros que se utilizan para evaluar la exactitud clínica de un test binario. El valor predictivo positivo (*VPP*) es la probabilidad de que un individuo tenga la enfermedad cuando el resultado del test es positivo, y el valor predictivo negativo (*VPN*) es la probabilidad de que un individuo no tenga la enfermedad cuando el resultado del test es negativo. Los valores predictivos dependen de la sensibilidad y especificidad del test diagnóstico y también de la prevalencia de la enfermedad. Estos parámetros se calculan aplicando el Teorema de Bayes, esto es,

$$VPP = \frac{p \times Se}{p \times Se + \ 1 - p \ \times \ 1 - Sp} \quad y \quad VPN = \frac{1 - p \ \times Sp}{p \times \ 1 - Se \ + \ 1 - p \ \times Sp}.$$

Los valores predictivos cuantifican el valor clínico del test diagnóstico, ya que tanto el individuo como el clínico están más interesados en conocer cómo de probable es tener la enfermedad dado un resultado del test diagnóstico. Un valor predictivo positivo alto indica que el test diagnóstico es útil para confirma la enfermedad en la población cuya prevalencia de la enfermedad es *p*.Un valor predictivo negativo alto indica que el test diagnóstico es útil para descartar la enfermedad en la población cuya prevalencia de la enfermedad es *p*. El valor predictivo positivo es una función creciente de la prevalencia de la enfermedad, y el valor predictivo negativo es una función decreciente de la prevalencia.

2.4. Coeficiente kappa ponderado

Sean L y L' las pérdidas asociadas a una clasificación errónea con el test diagnóstico:*L* es la pérdida que se comete cuando en un individuo enfermo el test diagnóstico es negativo, y L' es la pérdida que se comete cuando en un individuo no enfermo el test diagnóstico es positivo. Las pérdidas L y L' son cero cuando un individuo (enfermo o no) es clasificado correctamente con el test diagnóstico. En la Tabla I se muestran las probabilidades y las pérdidas asociadas a la evaluación de un test diagnóstico binario con respecto a un gold estándar. En términos de las probabilidades y pérdidas de la Tabla I, la pérdida esperada al aplicar el test diagnóstico es

$$p \ \ 1 - Se \ \ L + q \ \ 1 - Sp \ \ L'$$

y la pérdida aleatoria es

$$p \ \ p \ \ 1 - Se \ + \ 1 - p \ \ Sp \ \ L + \ 1 - p \quad pSe + \ 1 - p \quad 1 - Sp \quad L'.$$

La pérdida esperada es la pérdida promedio que se comete al clasificar erróneamente con el test diagnóstico a un individuo enfermo o no enfermo, y su rango de valores varía entre cero e infinito. La pérdida aleatoria es la pérdida que se comente cuando el test diagnóstico y el gold estándar son independientes, es decir cuando $P \ \ T = i \middle| D = j \ = P \ \ T = i$. En términos de la pérdida esperada y de la pérdida aleatoria, el coeficiente kappa ponderado de un test binario [3] se define como

$$\kappa = \frac{\text{Pérdida aleatoria} - \text{Pérdida esperada}}{\text{Pérdida aleatoria}}$$

y es por tanto una medida del acuerdo más allá del azar entre el test diagnóstico y el gold estándar cuando ambos se aplican a una misma cohorte de individuos. Los valores del coeficiente kappa ponderado varían entre -1 y 1. Sustituyendo en la ecuación anterior cada pérdida por su correspondiente expresión, se obtiene que el coeficiente kappa ponderado es

$$\kappa \ \ c \ = \frac{p \ \ 1 - p \ \ Y}{p \ \ 1 - Q \ \ c + \ 1 - p \ \ Q \ \ 1 - c},$$

donde $Q = pSe + 1-p \ 1-Sp$, $Y = Se + Sp - 1$ es el índice de Youden $c = L/\ L + L'$ es el índice de ponderación. Cuando la pérdida L es cero, entonces $c = 0$ and el coeficiente kappa ponderado es

$$\kappa\ 0\ = \frac{Sp - \ 1-Q}{Q} = \frac{VPP - p}{1-p}.$$

Cuando la pérdida L' es cero, entonces $c = 1$ y el coeficiente kappa ponderado es

$$\kappa\ 1\ = \frac{Se - Q}{1-Q} = \frac{VPN - \ 1-p}{p}.$$

Cuando $L = L'$, entonces $c = 0.5$ y el coeficiente kappa ponderado (denominado coeficiente kappa de Cohen) es

$$\kappa\ 0.5\ = \frac{p\ 1-p\ Y}{\frac{p+Q}{2} - pQ} = \frac{2}{\frac{1}{\kappa\ 0} + \frac{1}{\kappa\ 1}} = \frac{2\kappa\ 0\ \kappa\ 1}{\kappa\ 0\ + \kappa\ 1},$$

por lo que el coeficiente kappa de Cohen es la media armónica de $\kappa\ 0$ and $\kappa\ 1$. Las expresiones $\kappa\ 0\ = \ Sp - \ 1-Q\ \ /Q$ and $\kappa\ 1\ = \ Se - Q\ /\ 1-Q$ son la especificidad y la sensibilidad corregidas por azar según el modelo kappa, respectivamente.

El coeficiente kappa ponderado se puede escribir en términos de p, Q, $\kappa\ 0$ y $\kappa\ 1$ como

$$\kappa\ c\ = \frac{p\ 1-Q\ c\kappa\ 1\ + \ 1-p\ Q\ 1-c\ \kappa\ 0}{p\ 1-Q\ c + \ 1-p\ Q\ 1-c},$$

y por tanto el coeficiente kappa ponderado es una media ponderada de $\kappa\ 0$ y $\kappa\ 1$. También se puede escribir únicamente en términos de $\kappa\ 0$ y $\kappa\ 1$ como

$$\kappa\ c\ = \frac{\kappa\ 0\ \kappa\ 1}{c\kappa\ 0\ + \ 1-c\ \kappa\ 1} = \frac{\varphi^2}{c\kappa\ 0\ + \ 1-c\ \kappa\ 1}.$$

siendo $\varphi = \sqrt{\kappa\ 0\ \kappa\ 1}$ el coeficiente phi (coeficiente de correlación entre las variable aleatorias binarias *T* and *D* cuando estas se observan en todos los individuos).

El índice de ponderación *c* varía entre 0 y 1 y representa la pérdida relativa entre los falsos positivos y los falsos negativos. En la práctica el índice *c* es desconocido, pero sus valores se pueden intuir dependiendo del objetivo para el que se va a utilizar el test diagnóstico. Si el test diagnóstico se va a utilizar como paso previo a un tratamiento intensivo (es decir, como un test definitivo previo a un tratamiento de riesgo), hay más preocupación por los falsos positivos $L' > L$ and $0 < c < 0.5$ .Si el test diagnóstico se va a utilizar como un test de screening, hay una mayor preocupación por los falsos

negativos $L > L'$ y $0.5 < c < 1$.El índice *c* vale 0.5 cuando el test diagnóstico se utiliza para un diagnóstico simple (los falsos positivos y los falsos negativos tienen la misma importancia). El coeficiente kappa ponderado de un test binario presenta las siguientes propiedades:

a). Si el acuerdo clasificatorio entre el test binario y el gold estándar es perfecto $Se = Sp = 1$ entonces $\kappa\ c\ = 1$.

b). Si la sensibilidad y la especificidad son complementarias $Se = 1 - Sp$ entonces $\kappa\ c\ = 0$.

c). Si la pérdida aleatoria es mayor que la pérdida esperada entonces $\kappa\ c\ > 0$; y si la pérdida aleatoria es menor que la pérdida esperada entonces $\kappa\ c\ < 0$ y los resultados del test diagnóstico deben intercambiarse ($T = 1$ debe ser el resultado negativo y $T = 0$ debe ser el resultado positivo). Por tanto, el análisis se debe limitar a los valores positivos del coeficiente kappa ponderado, y sus valores se pueden clasificar en la siguiente escala [10]: de 0 a 0.20 el acuerdo es malo, de 0.21 a 0.40 el acuerdo es mediocre, de 0.41 a 0.60 el acuerdo es moderado, de 0.61 a 0.80 el acuerdo es bueno, y de 0.81 a 1 el acuerdo clasificatorio es muy bueno o casi perfecto.

d). El coeficiente kappa ponderado es una función del índice *c* que es creciente si $Q > p$, decreciente si $Q < p$, o constante igual al índice de Youden $Y = Se + Sp - 1$ si $Q = p$.

2.5. Coeficiente kappa promedio

En la práctica las pérdidas L and L' definidas en la Sección 2.4 no se pueden determinar, por lo que el clínico suele asignar valores al índice de ponderación dependiendo de su conocimiento sobre la importancia relativa entre los falsos positivos y los falsos negativos. Así por ejemplo, si el clínico considera que los falsos positivos son el doble de importantes que los falsos negativos, entonces asignará al índice de ponderación *c* el valor $1/3$. Sin embargo, en muchas situaciones prácticas el clínico no dispone de un criterio o conocimiento que le permita fijar el valor del índice de ponderación. En esta situación, el clínico puede asignar distintos valores al índice de ponderación y analizar los resultados con cada uno de los valores fijados. Incluso en una misma situación práctica, diferentes clínicos pueden asignar valores distintos al índice de ponderación, dependiendo de sus propios conocimientos sobre el problema. Para resolver este problema de asignación de valores al índice de ponderación, Roldán-Nofuentes y Olvera-Porcel [4] han propuesto una nueva medida: el coeficiente kappa promedio.

Para unos valores fijos de sensibilidad, especificidad y prevalencia, el coeficiente kappa ponderado es una función continua del índice de ponderación *c*. Si el clínico considera que $L' > L$ (por ejemplo, si el test diagnóstico se va a utilizar como un test definitivo previo a un tratamiento de riesgo), entonces $0 < c < 0.5$ y el coeficiente kappa promedio se define como

$$\kappa_1 = \frac{1}{0.5} \int_0^{0.5} \kappa\ c\ dc$$

es decir, el coeficiente kappa promedio κ_1 es el valor medio de la función $\kappa\ c$ cuando $0 < c < 0.5$. Si el clínico considera que $L > L'$ (por ejemplo, si el test diagnóstico se va a utilizar como un test de screening), entonces $0.5 < c < 1$ y el coeficiente kappa promedio se define como

$$\kappa_2 = \frac{1}{0.5}\int_{0.5}^{1} \kappa\ c\ dc,$$

por lo que el coeficiente kappa promedio κ_2 es el valor medio de la función $\kappa\ c$ cuando $0.5 < c < 1$. Resolviendo las integrales definidas se obtiene que

$$\kappa_1 = \begin{cases} \dfrac{2\kappa\ 0\ \kappa\ 1}{\kappa\ 0\ -\kappa\ 1}\ln\left[\dfrac{\kappa\ 0\ +\kappa\ 1}{2\kappa\ 1}\right], & p \neq Q \\ Y, & p = Q, \end{cases}$$

y

$$\kappa_2 = \begin{cases} \dfrac{2\kappa\ 0\ \kappa\ 1}{\kappa\ 0\ -\kappa\ 1}\ln\left[\dfrac{2\kappa\ 0}{\kappa\ 0\ +\kappa\ 1}\right], & p \neq Q \\ Y, & p = Q, \end{cases}$$

donde $\ln\ \cdot$ es el logaritmo neperiano. Cuando $p = Q$ el coeficiente kappa ponderado es siempre igual al índice de Youden Y para cualquier valor del índice de ponderación *c*, por lo que los dos coeficientes kappa promedio son también iguales al índice de Youden. Asimismo, para $p \neq Q$ los coeficientes kappa promedio se puede expresar en términos de $\kappa\ 0$, $\kappa\ 1$ and $\kappa\ c$ como

$$\kappa_1 = \frac{2\left[c\kappa\ 0\ +\ 1-c\ \kappa\ 1\right]\kappa\ c}{\kappa\ 0\ -\kappa\ 1}\ln\left[\frac{\kappa\ 0\ +\kappa\ 1}{2\kappa\ 1}\right]$$

y

$$\kappa_2 = \frac{2\left[c\kappa\ 0\ +\ 1-c\ \kappa\ 1\right]\kappa\ c}{\kappa\ 0\ -\kappa\ 1}\ln\left[\frac{2\kappa\ 0}{\kappa\ 0\ +\kappa\ 1}\right].$$

Como el coeficiente kappa ponderado es una medida del acuerdo más allá del azar entre el test diagnóstico y el gold estándar, los coeficientes kappa promedio (que se calculan a partir de coeficientes kappa ponderados) son medidas del acuerdo promedio más allá del azar entre el test diagnóstico y el gold estándar, y no dependen del índice de ponderación *c*. Los coeficientes kappa promedio κ_1 and κ_2 presentan las siguientes propiedades:

a). Si $Se = Sp = 1$, entonces $\kappa_1 = \kappa_2 = 1$. Si $Se = 1 - Sp$, entonces $\kappa_1 = \kappa_2 = 0$. Por consiguiente, como la evaluación del test diagnóstico se debe limitar a los valores positivos del coeficiente kappa ponderado (propiedad (c) de la Sección 2.4), los valores de los coeficientes κ_1 and κ_2 son mayores que 0 and menores que 1.

b). El coeficiente κ_1 es mayor que κ_2 si $p > Q$, and κ_1 es menor que κ_2 si $Q > p$.

c). Para $p \neq Q$ las expresiones de κ_1 and κ_2 se pueden escribir como

$$\kappa_1 = 2\varphi^2\,\frac{log\left[\kappa\ 0.5\ \right] - log\left[\kappa\ 0\ \right]}{\kappa\ 1\ -\kappa\ 0}$$

y

$$\kappa_2 = 2\varphi^2 \frac{log\left[\kappa\ 1\ \right] - log\left[\kappa\ 0.5\ \right]}{\kappa\ 1\ - \kappa\ 0},$$

respectivamente. Por tanto, el coeficiente kappa promedio κ_1 es proporcional al término $\frac{log\left[\kappa\ 0.5\ \right] - log\left[\kappa\ 0\ \right]}{\kappa\ 1\ - \kappa\ 0}$, que es el cociente entre la máxima diferencia entre los coeficientes kappa ponderados (en logaritmos) cuando $L' > L$ y la máxima diferencia posible entre los coeficientes kappa ponderados. De forma similar, el coeficiente kappa promedio κ_2 es proporcional a $\frac{log\left[\kappa\ 1\ \right] - log\left[\kappa\ 0.5\ \right]}{\kappa\ 1\ - \kappa\ 0}$, que es el cociente entre la máxima diferencia entre los coeficientes kappa ponderados (en logaritmos) cuando $L > L'$ y la máxima diferencia posible entre los coeficientes kappa ponderados.

d). El coeficiente kappa ponderado es una función continua en el intervalo $0,1$, por lo que el coeficiente kappa promedio κ_1 κ_2 coincide con un valor del coeficiente kappa ponderado. De esta forma, una vez estimado el coeficiente kappa promedio, puede determinar el valor del índice de ponderación asociado al coeficiente kappa promedio estimado. Por tanto, la estimación del coeficiente kappa promedio permite estimar la pérdida relativa entre los falsos positivos y los falsos negativos asociada a este parámetro.

e). El coeficiente kappa promedio κ_1 minimiza la expresión $2\int_0^{0.5} \kappa\ c\ - x\ ^2\ dc$. Cuando $x = \kappa_1$ esta expresión es la varianza del coeficiente kappa ponderado en torno a κ_1. De forma similar, el coeficiente kappa ponderado κ_2 minimiza la expresión $2\int_{0.5}^{1} \kappa\ c\ - x\ ^2\ dc$. Cuando $x = \kappa_2$, esta expression es la varianza del coeficiente kappa ponderado en torno a κ_2.

3. ESTIMACIÓN DE LOS PARÁMETROS

La estimación de los parámetros se va a realizar bajo un muestreo transversal. Este tipo de muestreo consiste en aplicar el test diagnóstico binario y el gold estándar de forma independiente a cada uno de los individuos de una muestra aleatoria de tamaño *n*, y es uno de los tipos de muestreo que más se aplican en la práctica. En la Tabla 1 se muestran las frecuencias observadas bajo un muestreo transversal.

Tabla 1. Frecuencias observadas bajo un muestreo transversal.

	$T=1$	$T=0$	Total
$D=1$	s_1	s_0	s
$D=0$	r_1	r_0	r
Total	$s_1 + r_1$	$s_0 + r_0$	n

3.1. Sensibilidad y especificidad

Condicionando en los totales de las filas, es decir en el resultado del gold estándar (variable *D*), se obtiene que la frecuencia s_1 es la realización de la distribución binomial

$B\ s, Se$ y la frecuencia r_0 es la realización de la distribución binomial $B\ r, Sp$, por lo que los estimadores puntuales de la sensibilidad y de la especificidad son

$$\hat{Se} = \frac{s_1}{s} \quad \text{y} \quad \hat{Sp} = \frac{r_0}{r},$$

y son por tanto estimadores de proporciones binomiales, siendo sus varianzas estimadas

$$Var\ \hat{Se}\ = \frac{\hat{Se}\ 1-\hat{Se}}{s} \quad \text{y} \quad Var\ \hat{Sp}\ = \frac{\hat{Sp}\ 1-\hat{Sp}}{r}.$$

Un intervalo de confianza, denominado intervalo score modificado [5], para la sensibilidad es

$$Se \in 0.5 + \frac{s + \frac{z_{1-\alpha/2}^4}{53}}{s + z_{1-\alpha/2}^2}\ \hat{Se} - 0.5\ \pm \frac{z_{1-\alpha/2}}{s + z_{1-\alpha/2}^2} \sqrt{\hat{Se}\ 1-\hat{Se}\ s + \frac{z_{1-\alpha/2}^2}{4}},$$

donde $z_{1-\alpha/2}$ es el $100\ 1-\alpha$ % percentil de la distribución normal estándar. Este intervalo es válido para $s \geq 10$. De forma similar, el intervalo de confianza score modificado para la especificidad es

$$Sp \in 0.5 + \frac{r + \frac{z_{1-\alpha/2}^4}{53}}{r + z_{1-\alpha/2}^2}\ \hat{Sp} - 0.5\ \pm \frac{z_{1-\alpha/2}}{r + z_{1-\alpha/2}^2} \sqrt{\hat{Sp}\ 1-\hat{Sp}\ r + \frac{z_{1-\alpha/2}^2}{4}}$$

y es un intervalo válido para $r \geq 10$. El intervalo score modificado [5] es el intervalo de confianza para una proporción binomial que presenta un mejor rendimiento.

3.2. Razones de verosimilitud

Los estimadores máximo verosímiles de las *LRs* vienen dados por las expresiones

$$LR^+ = \frac{s_1 r}{r_1 s} \quad \text{and} \quad LR^- = \frac{s_0 r}{r_0 s}$$

y aplicando el método delta, sus varianzas asintóticas estimadas son

$$Var\ LR^+\ = \frac{LR^+}{1-\hat{Sp}} \left(\frac{1-\hat{Se}}{s} + \frac{\hat{Sp}}{r} LR^+ \right)$$

y

$$Var\ LR^-\ = \frac{LR^-}{\hat{Sp}} \left(\frac{\hat{Se}}{s} + \frac{1-\hat{Sp}}{r} LR^- \right).$$

Condicionando en los totales de filas de la Tabla I, las *LR*sson el ratio de dos proporciones binomiales independientes, por lo que las *LRs* se pueden estimar aplicando métodos para estimar el ratio de dos proporciones binomiales independientes. Martín-Andrés y Álvarez-Hernández[6]han estudiado intervalos de confianza para el ratio de dos proporciones binomiales independientes, recomendando utilizar los siguientes intervalos de confianza (adaptados a la notación de la Tabla I),

$$LR^+ \in \frac{n' s_1' r_1' + \frac{z_{1-\alpha/2}^2}{2}\ s' s_1' + r' r_1' - 2 s_1' r_1'\ \pm z_{1-\alpha/2} \sqrt{n'^2 s_1' r_1'\ s_1' + r_1' - n' \hat{p}_1' \hat{p}_2'\ + \frac{z_{1-\alpha/2}^2}{4}\ s' s_1' - r' r_1'^{\ 2}}}{r_1'\ n' s' \hat{p}_1' - z_{1-\alpha/2}^2\ s' - r_1'}$$

y

$$LR^- \in \frac{n's_0'r_0' + \frac{z_{1-\alpha/2}^2}{2}\ s's_0' + r'r_0' - 2s_0'r_0'\ \pm z_{1-\alpha/2}\sqrt{n'^2 s_0'r_0'\ \ s_0' + r_0' - n'\hat{p}_3'\hat{p}_4'\ + \frac{z_{1-\alpha/2}^2}{4}\ s's_0' - r'r_0'^{\ 2}}}{r_0'\ \ n's'\hat{p}_3' - z_{1-\alpha/2}^2\ \ s' - r_0'}$$

donde $s_i' = s_i + 0.5$, $r_i' = r_i + 0.5$, $s' = s_1' + s_0'$, $r' = r_1' + r_0'$, $n' = s' + r'$, $\hat{p}_1' = r_1' / r'$, $\hat{p}_2' = s_1' / s'$, $\hat{p}_3' = r_0' / r'$ and $\hat{p}_4' = s_0' / s'$. Si el límite inferior del intervalo para LR^+ es menor que $s_1' / \ n' - r_1'$ o mayor que el estimador de LR^+, entonces el límite inferior del intervalo es

$$\frac{1}{s'\ \ \hat{p}_1'^{\ 2} + z_{1-\alpha/2}^2}\left\{ s_1'\hat{p}_1' + \frac{z_{1-\alpha/2}^2}{2} - z_{1-\alpha/2}\sqrt{\frac{z_{1-\alpha/2}^2}{4} + s_1'\ \ \hat{p}_1'\hat{p}_2'} \right\},$$

y si el límite superior de este intervalo es mayor que $n' - s_1' \ / r_1'$ o menor que el estimador de LR^+, entonces el límite superior del intervalo es

$$\frac{1}{r'\ \ \hat{p}_1'^{\ 2}}\left\{ r_1'\hat{p}_2' + \frac{z_{1-\alpha/2}^2}{2} + z_{1-\alpha/2}\sqrt{\frac{z_{1-\alpha/2}^2}{4} + r_1'\ \ \hat{p}_2'\hat{p}_1'} \right\}.$$

Con respecto al *CI* para LR^-, si el límite inferior de este intervalo es menor que $s_0' / \ n' - r_0'$ o mayor que el estimador de LR^-, entonces el límite inferior del intervalo es

$$\frac{1}{s'\ \ \hat{p}_3'^{\ 2} + z_{1-\alpha/2}^2}\left\{ s_0'\hat{p}_3' + \frac{z_{1-\alpha/2}^2}{2} - z_{1-\alpha/2}\sqrt{\frac{z_{1-\alpha/2}^2}{4} + s_0'\ \ \hat{p}_3'\hat{p}_4'} \right\},$$

y si el límite superior de este intervalo es mayor que $n' - s_0' \ / r_0'$ o menor que el estimador de LR^-,entonces el límite superior del intervalo es

$$\frac{1}{r'\ \ \hat{p}_3'^{\ 2}}\left\{ r_0'\hat{p}_4' + \frac{z_{1-\alpha/2}^2}{2} + z_{1-\alpha/2}\sqrt{\frac{z_{1-\alpha/2}^2}{4} + r_0'\ \ \hat{p}_4'\hat{p}_3'} \right\}.$$

3.3. Valores predictivos

Condicionando en los totales de las columna, es decir en el resultado del test diagnóstico binario (variable *T*), se obtiene que la frecuencia s_1 es la realización de la distribución binomial $B\ s_1 + r_1, VPP$ y la frecuencia r_0 es la realización de la distribución binomial $B\ s_0 + r_0, VPN$, por lo que los estimadores puntuales de los valores predictivos son

$$VPP = \frac{s_1}{s_1 + r_1} \quad \text{y} \quad VPN = \frac{r_0}{s_0 + r_0},$$

y son, al igual que la sensibilidad y la especificidad, estimadores de proporciones binomiales, siendo sus varianzas estimadas

$$Var\ VPP\ = \frac{VPP\ 1-VPP}{s_1 + r_1} \quad \text{y} \quad Var\ VPN\ = \frac{VPN\ 1-VPN}{s_0 + r_0}.$$

El intervalo score modificado [5] para el valorpredictivo positivo es

$$VPP \in 0.5 + \frac{s_1 + r_1 + \frac{z^4_{1-\alpha/2}}{53}}{s_1 + r_1 + z^2_{1-\alpha/2}}\ VPP - 0.5\ \pm \frac{z_{1-\alpha/2}}{s_1 + r_1 + z^2_{1-\alpha/2}} \sqrt{VPP\ 1-VPP\quad s_1 + r_1\ + \frac{z^2_{1-\alpha/2}}{4}},$$

que es válido para $s_1 + r_1 \geq 10$. De forma similar, el intervalo de confianza score modificado [5] para el valor predictivo negativo es

$$VPN \in 0.5 + \frac{s_0 + r_0 + \frac{z^4_{1-\alpha/2}}{53}}{s_0 + r_0 + z^2_{1-\alpha/2}}\ VPN - 0.5\ \pm \frac{z_{1-\alpha/2}}{s_0 + r_0 + z^2_{1-\alpha/2}} \sqrt{VPN\ 1-VPN\quad s_0 + r_0\ + \frac{z^2_{1-\alpha/2}}{4}},$$

expresión válida para $s_0 + r_0 \geq 10$.

3.4. Coeficiente kappa ponderado

Sustituyendo en la expresión del coeficiente kappa ponderado cada parámetro por su estimador, esto es

$$\hat{Se} = \frac{s_1}{s}, \quad \hat{Sp} = \frac{r_0}{r} \quad \text{y} \quad \hat{p} = \frac{s}{n},$$

el estimador máximo verosímil del coeficiente kappa ponderado es

$$\hat{\kappa}\ c\ = \frac{s_1 r_0 - s_0 r_1}{s\ s_0 + r_0\ c + r\ s_1 + r_1\quad 1-c},$$

con $0 < c < 1$. Aplicando el método delta la varianza asintótica estimada de $\hat{\kappa}\ c$ es

$$\hat{Var}\ \hat{\kappa}\ c\ = \frac{nr}{s\left[n^2\ 1-c\ r_1 + n\ cr_0 - 2\ 1-c\ r_1\ s_0 + n\ r_0 - 1-c\ r_1\ s_1 + s\ s_0 r_1 - s_1 r_0 \right]^4} \times$$

$$s_0 r_1 - s_1 r_0\ ^2 \left[2\ 1-c\ r_1 ns - 1-c\ r_1 n^2 + s\ c\ s_0 r_0 + 2 s_0 r_1 + s_1 r_1\ - r_1 s \right]^2 +$$

$$s_1 s_0 nr^3 \left[1-c\ r_1 n + s\ cr - r_1 \right]^2 + r_1 r_0 nsr^2 \left[s_1 r + c\ s^2 - s_1 n \right]^2 .$$

Roldán Nofuentes et al[7] han estudiado diferentes intervalos de confianza para el coeficiente kappa ponderado. Dependiendo del tamaño muestral se pueden utilizar los siguientes intervalos de confianza:

a) Intervalo de confianza tipo Wald

Asumiendo la normalidad asintótica de $\hat{\kappa}\ c$, el intervalo de confianza tipo Wald para el coeficiente kappa ponderado es

$$\hat{\kappa}\ c\ \pm z_{1-\alpha/2} \sqrt{\hat{Var}\ \hat{\kappa}\ c} .$$

Este intervalo de confianza tiene un buen rendimiento para muestras relativamente pequeñas $n = 100$.

b) Intervalo de confianza bootstrap

El intervalo de confianza bootstrap se calcula generando *K* muestras con reemplazamiento de la muestra que se dispone y calculando en cada una de ellas el estimador del coeficiente kappa ponderado. Como estimador del coeficiente kappa ponderado se utiliza la media de las *K* réplicas de tales estimadores, y el intervalo de confianza global se calcula empleando el intervalo de confianza corregido por el sesgo [8]. Este intervalo de confianza tiene un rendimiento muy similar al intervalo de confianza tipo Wald.

c). Intervalo de confianza logit

Asumiendo la normalidad asintótica de $\hat{\kappa}\ c$, la transformación logit de $\hat{\kappa}\ c$, $\ln\ \hat{\kappa}\ c\ /\ 1-\hat{\kappa}\ c$, sigue una distribución normal de media $\ln\ \kappa\ c\ /\ 1-\kappa\ c$. De esta forma el intervalo de confianza para el logit de $\kappa\ c$ es

$$\text{logit}\ \hat{\kappa}\ c\ \pm z_{1-\alpha/2} \sqrt{\hat{Var}\ \text{logit}\ \hat{\kappa}\ c} ,$$

donde el estimador de la varianza del logit de $\hat{\kappa}\ c$ es

$$\hat{V}ar\ \text{logit}\ \hat{\kappa}\ c\ = \frac{1}{\left[r_1 s - \ 1-c\ \ r_1 n - c\ \ s_0 r_0 + 2 s_0 r_1 + s_1 r_1\ \right]^2} \times$$

$$\left\{ \frac{s_1 s_0 r^2 \left[\ 1-c\ \ r_1 n + \ cr - r_1\ \ s\right]^2 + r_1 r_0 sr \left[s_1 n - s_1 s + c\ \ s^2 - s_1 n\ \right]^2}{s\ \ s_0 r_1 - s_1 r_0\ ^2} + \right.$$

$$\left. \frac{\left[2\ \ 1-c\ \ r_1 ns - \ 1-c\ \ r_1 n^2 + s\ \ c\ \ s_0 r_0 + 2 s_0 r_1 + s_1 r_1\ \ - r_1 s\ \right]}{nsr} \right\}.$$

Finalmente, el intervalo de confianza logit para el coeficiente kappa ponderado es

$$\left(\frac{\exp\ \text{logit}\ \hat{\kappa}\ c\ \ - z_{1-\alpha/2} \sqrt{\hat{V}ar\ \text{logit}\ \hat{\kappa}\ c}}{1 + \exp\ \text{logit}\ \hat{\kappa}\ c\ \ - z_{1-\alpha/2} \sqrt{\hat{V}ar\ \text{logit}\ \hat{\kappa}\ c}} ; \frac{\exp\ \text{logit}\ \hat{\kappa}\ c\ \ + z_{1-\alpha/2} \sqrt{\hat{V}ar\ \text{logit}\ \hat{\kappa}\ c}}{1 + \exp\ \text{logit}\ \hat{\kappa}\ c\ \ + z_{1-\alpha/2} \sqrt{\hat{V}ar\ \text{logit}\ \hat{\kappa}\ c}} \right),$$

y que tiene un buen rendimiento para muestras de tamaño igual o mayor a 200.

3.5. Coeficiente kappa promedio

Roldán-Nofuentes y Olvera-Porcel [4] han deducido los estimadores máximo verosímiles de los coeficientes kappa promedios y han estudiado varios intervalos de confianza para estos parámetros. Los estimadores máximo verosímilesde los coeficientes kappa promedio κ_1 and κ_2 son

$$\hat{\kappa}_1 = \begin{cases} \dfrac{2\ \ s_1 r_0 - s_0 r_1}{n_0 s - n_1 r} log\left[\dfrac{n_1 r + n_0 s}{2 n_1 r}\right], & s_0 \neq r_1 \\ \dfrac{s_1 r_0 - s_0 r_1}{sr}, & s_0 = r_1, \end{cases}$$

y

$$\hat{\kappa}_2 = \begin{cases} \dfrac{2\ \ s_1 r_0 - s_0 r_1}{n_0 s - n_1 r} log\left[\dfrac{2 n_0 s}{n_1 r + n_0 s}\right], & s_0 \neq r_1 \\ \dfrac{s_1 r_0 - s_0 r_1}{sr}, & s_0 = r_1, \end{cases}$$

respectivamente. En las ecuaciones anteriores $\hat{p} \neq \hat{Q}$ $\hat{p} = \hat{Q}$ es equivalente a que $s_0 \neq r_1$ $s_0 = r_1$. Asimismo, aplicando el método delta, cuando $s_0 \neq r_1$ las varianzas asintóticas estimadas de $\hat{\kappa}_1$ and $\hat{\kappa}_2$ son

$$Var\ \hat{\kappa}_1 = \frac{1}{\left[\hat{\kappa}\ 0 + \hat{\kappa}\ 1\right]^2\left[\hat{\kappa}\ 0 - \hat{\kappa}\ 1\right]^2} \times$$

$$\left\{\left\{\frac{2\hat{\kappa}\ 0^{\ 2}\hat{\kappa}\ 1 - \hat{\kappa}\ 1\left[\hat{\kappa}\ 0 + \hat{\kappa}\ 1\right]\hat{\kappa}_1}{\hat{\kappa}\ 0}\right\}^2 \frac{1-\hat{S}p^{\ 2}\hat{Y}^2 Var\ \hat{p} + \hat{p}^2\left[1-\hat{S}p\ Var\ \hat{S}e + \hat{S}e^2 Var\ \hat{S}p\right]}{\hat{Q}^4} +$$

$$\left\{\frac{\hat{\kappa}\ 0\left[\hat{\kappa}\ 0 + \hat{\kappa}\ 1\ \hat{\kappa}_1 - 2\hat{\kappa}\ 0\ \hat{\kappa}\ 1\right]}{\hat{\kappa}\ 1}\right\}^2 \frac{1-\hat{S}e^{\ 2}\hat{Y}^2 Var\ \hat{p} + 1-\hat{p}^{\ 2}\left[\hat{S}p Var\ \hat{S}e + 1-\hat{S}e^{\ 2}\ Var\ \hat{S}p\right]}{\left(1-\hat{Q}\right)^4} +$$

$$2\left\{\frac{2\hat{\kappa}\ 0^{\ 2}\hat{\kappa}\ 1 - \hat{\kappa}\ 1\left[\hat{\kappa}\ 0 + \hat{\kappa}\ 1\right]\hat{\kappa}_1}{\hat{\kappa}\ 0}\right\}\left\{\frac{\hat{\kappa}\ 0\left[\hat{\kappa}\ 0 + \hat{\kappa}\ 1\right]\hat{\kappa}_1 - 2\hat{\kappa}\ 0\ \hat{\kappa}\ 1}{\hat{\kappa}\ 1}\right\} \times$$

$$\frac{\left(1-\hat{p}\right)\left[\left(1-\hat{S}e\right)\hat{S}e Var\ \hat{S}p + \left(1-\hat{S}p\right)\hat{S}p Var\ \hat{S}e\right] - \left(1-\hat{S}e\right)\left(1-\hat{S}p\right)^2 Var\ \hat{p}}{\hat{Q}^2\left(1-\hat{Q}\right)^2}\Bigg\}$$

y

$$Var\ \hat{\kappa}_2 = \frac{1}{\left[\hat{\kappa}\ 0 + \hat{\kappa}\ 1\right]^2\left[\hat{\kappa}\ 0 - \hat{\kappa}\ 1\right]^2} \times$$

$$\left\{\left\{\frac{\hat{\kappa}\ 1\left[2\hat{\kappa}\ 0\ \hat{\kappa}\ 1 - \hat{\kappa}\ 0 + \hat{\kappa}\ 1\ \hat{\kappa}_2\right]}{\hat{\kappa}\ 0}\right\}^2 \frac{1-\hat{S}p^{\ 2}\hat{Y}^2 Var\ \hat{p} + \hat{p}^2\left[1-\hat{S}p\ Var\ \hat{S}e + \hat{S}e^2 Var\ \hat{S}p\right]}{\hat{Q}^4} +$$

$$\left\{\frac{\hat{\kappa}\ 0\left[\hat{\kappa}\ 0 + \hat{\kappa}\ 1\right]\hat{\kappa}_2 - 2\hat{\kappa}\ 0\ \hat{\kappa}\ 1^{\ 2}}{\hat{\kappa}\ 1}\right\}^2 \frac{1-\hat{S}e^{\ 2}\hat{Y}^2 Var\ \hat{p} + 1-\hat{p}^{\ 2}\left[\hat{S}p Var\ \hat{S}e + 1-\hat{S}e^{\ 2}\ Var\ \hat{S}p\right]}{\left(1-\hat{Q}\right)^4} +$$

$$2\left\{\frac{\hat{\kappa}\ 1\left[2\hat{\kappa}\ 0\ \hat{\kappa}\ 1 - \hat{\kappa}\ 0 + \hat{\kappa}\ 1\right]}{\hat{\kappa}\ 0}\right\}\left\{\frac{\hat{\kappa}\ 0\left[\hat{\kappa}\ 0 + \hat{\kappa}\ 1\right]\hat{\kappa}_2 - 2\hat{\kappa}\ 0\ \hat{\kappa}\ 1^{\ 2}}{\hat{\kappa}\ 1}\right\} \times$$

$$\frac{\left(1-\hat{p}\right)\left[\left(1-\hat{S}e\right)\hat{S}e Var\ \hat{S}p + \left(1-\hat{S}p\right)\hat{S}p Var\ \hat{S}e\right] - \left(1-\hat{S}e\right)\left(1-\hat{S}p\right)^2 Var\ \hat{p}}{\hat{Q}^2\left(1-\hat{Q}\right)^2}\Bigg\},$$

respectivamente, siendo $\hat{S}e = s_1/s$, $\hat{S}p = r_0/r$, $\hat{p} = s/n$, $\hat{V}ar\ \hat{S}e = \hat{S}e\ 1-\hat{S}e\ /s$, $\hat{V}ar\ \hat{S}p = \hat{S}p\ 1-\hat{S}p\ /r$, $\hat{V}ar\ \hat{p} = \hat{p}\ 1-\hat{p}\ /n$, $\hat{Y} = s_1 r + r_0 s - sr\ /sr$ y $\hat{Q} = s_1 + r_1\ /n$. Cuando $s_0 = r_1$ las varianzas asintóticas estimadas son

$$Var\ \hat{\kappa}_1 = Var\ \hat{\kappa}_2 = Var\ \hat{Y} = \frac{\hat{S}e\ 1-\hat{S}e}{s} + \frac{\hat{S}p\ 1-\hat{S}p}{r}.$$

Para los dos coeficientes kappa promedio se han estudiado los siguientes intervalos de confianza asintóticos (similares a los del coeficiente kappa ponderado).

a). Intervalo de confianza tipo Wald

Asumiendo la normalidad asintótica de $\hat{\kappa}_1$ and $\hat{\kappa}_2$, el intervalo de confianza cada coeficiente kappa promedio es

$$\hat{\kappa} \pm z_{1-\alpha/2}\sqrt{Var\ \hat{\kappa}}\ ,$$

donde $\hat{\kappa}$ es $\hat{\kappa}_1$ o $\hat{\kappa}_2$.

b). Intervalo de confianza bootstrap
La estimación de los coeficientes kappa promedio se puede realizar aplicando el método bootstrap, de forma similar a como se ha explicado para el coeficiente kappa ponderado. A partir de la muestra aleatoria observada se generan *K* muestran con reemplazamiento, y a partir de estas *K* muestras se calculan los intervalos de confianza. Por consiguiente, como estimador de cada coeficiente kappa promedio se propone el valor promedio obtenido con las *K* muestras aleatorias, y a continuación se calculan a partir de las *K* muestras el intervalo de confianza corregido por el sesgo [8] para cada uno de los coeficientes kappa promedio.

c). Intervalo de confianza logit
Como la evaluación del coeficiente kappa ponderado de un test diagnóstico se limita a los valores positivos (entre 0 and 1) de este parámetro, los valores de los coeficientes kappa promedio también están entre 0 y 1, pudiéndose realizar la transformación logit de estos parámetros. Así, asumiendo la normalidad asintótica de $\hat{\kappa}$, la transformación logit de $\hat{\kappa}$, $\ln\left[\hat{\kappa}/\ 1-\hat{\kappa}\ \right]$, sigue una distribución normal de media $\ln\left[\kappa/\ 1-\kappa\ \right]$. De esta forma intervalo de confianza para el logit de κ es

$$\text{logit}\ \hat{\kappa}\ \pm z_{1-\alpha/2}\sqrt{Var\ \text{logit}\ \hat{\kappa}}\ ,$$

siendo las varianzas asintóticas estimadas cuando $s_0 \neq r_1$

$$Var\ \text{logit}\ \hat{\kappa}_1\ = \frac{1}{\hat{\kappa}\ 0\ +\hat{\kappa}\ 1\ ^2\left[\hat{\kappa}\ 0\ -\hat{\kappa}\ 1\ \right]^2\hat{\kappa}_1^2\left[1-\hat{\kappa}_1^2\right]^2}\times$$

$$\left\{\left\{\frac{\hat{\kappa}\ 1\left[2\ 1-\hat{\kappa}_1\ \hat{\kappa}\ 0\ ^2-\ \hat{\kappa}\ 0\ \ 1-2\hat{\kappa}\ 0\ \ +\hat{\kappa}\ 1\ \ \hat{\kappa}_1\right]}{\hat{\kappa}\ 0}\right\}^2\frac{1-\hat{S}p\ ^2\hat{Y}^2Var\ \ \hat{p}\ +\hat{p}^2\left[\ 1-\hat{S}p\ Var\ \ \hat{S}e\ +\hat{S}e^2Var\ \ \hat{S}p\ \right]}{\hat{Q}^4}+\right.$$

$$\left\{\frac{\hat{\kappa}\ 0\left[\ \hat{\kappa}\ 0\ +\hat{\kappa}\ 1\ \ \hat{\kappa}_1-2\hat{\kappa}\ 0\ \hat{\kappa}\ 1\ \right]}{\hat{\kappa}\ 1}\right\}^2\frac{1-\hat{S}e\ ^2\hat{Y}^2Var\left(\hat{p}\right)\left(1-\hat{p}\right)^3\left[\hat{S}pVar\left(\hat{S}e\right)+\left(1-\hat{S}e\right)^2Var\left(\hat{S}p\right)\right]}{\left(1-\hat{Q}\right)^4}+$$

$$2\left\{\frac{\hat{\kappa}\left(1\right)\left[2\left(1-\hat{\kappa}_1\right)\left(0\right)^3-\left(\hat{\kappa}\ 0\right)\left(1-2\hat{\kappa}\ 0\right)\left(\hat{\kappa}\ 1\right)\hat{\kappa}_1\right]}{\hat{\kappa}\left(0\right)}\right\}\left\{\frac{\hat{\kappa}\left(0\right)\left[\left(\hat{\kappa}\ 0\right)+\hat{\kappa}\left(1\right)\hat{\kappa}_1-2\hat{\kappa}\left(0\right)\left(1\right)\right]}{\hat{\kappa}\left(1\right)}\right\}\times$$

$$\left.\frac{\left(1-\hat{p}\right)\left[\left(1-\hat{S}e\right)\hat{S}eVar\left(\hat{S}p\right)+\left(1-\hat{S}p\right)\hat{S}pVar\left(\hat{S}e\right)\right]+\left(1-\hat{S}e\right)\left(1-\hat{S}p\right)\hat{Y}^2Var\left(\hat{p}\right)}{\hat{Q}^2\left(1-\hat{Q}\right)^3}\right\},$$

y

$$Var\ \text{logit}\ \hat{\kappa}_2 = \frac{1}{\left[\hat{\kappa}_0 + \hat{\kappa}_1\right]^2\left[\hat{\kappa}_0 - \hat{\kappa}_1\right]^2 \hat{\kappa}_2^2\left[1-\hat{\kappa}_2^2\right]^2} \times$$

$$\left\{\left\{\frac{\hat{\kappa}_1\left[2\hat{\kappa}_0\hat{\kappa}_1 - \hat{\kappa}_0 + \hat{\kappa}_1\hat{\kappa}_2\right]}{\hat{\kappa}_0}\right\}^2 \frac{1-\hat{Sp}^2\hat{Y}^2 Var\ \hat{p} + \hat{p}^2\left[1-\hat{Sp}\ Var\ \hat{Se} + \hat{Se}^2 Var\ \hat{Sp}\right]}{\hat{Q}^4} + \right.$$

$$\left\{\frac{\hat{\kappa}_0\left[\hat{\kappa}_0 + \hat{\kappa}_1\hat{\kappa}_2 - 2\hat{\kappa}_1^2\right]}{\hat{\kappa}_1}\right\}^2 \frac{1-\hat{Se}^2\hat{Y}^2 Var\ \hat{p} + (1-\hat{p})^2\left[\hat{Sp}\,Var\,\hat{Se} + (1-\hat{Se})^2 Var\,\hat{Sp}\right]}{(1-\hat{Q})^4} +$$

$$2\left\{\frac{\hat{\kappa}_1\left[2\hat{\kappa}_0\hat{\kappa}_1 - \hat{\kappa}_0 + \hat{\kappa}_1\hat{\kappa}_2\right]}{\hat{\kappa}_0}\right\}\left\{\frac{\hat{\kappa}_0\left[\hat{\kappa}_0 + \hat{\kappa}_1\hat{\kappa}_2 - 2\hat{\kappa}_1^2\right]}{\hat{\kappa}_1}\right\} \times$$

$$\left.\frac{(1-\hat{p})\left[(1-\hat{Se})\hat{Se}\,Var\,\hat{Sp} + (1-\hat{Sp})\hat{Sp}\,Var\,\hat{Se}\right] - (1-\hat{Se})(1-\hat{Sp})\hat{Y}^2 Var\,\hat{p}}{\hat{Q}^2(1-\hat{Q})^2}\right\},$$

respectivamente, y cuando $s_0 = r_1$ las varianzas asintóticas estimadas son

$$Var\ \log\ \hat{\kappa}_1 = Var\ \log\ \hat{\kappa}_2 = Var\ \log\ \hat{Y} = \frac{1}{\hat{Y}\ 1-\hat{Y}}\left(\frac{\hat{Se}\ 1-\hat{Se}}{s} + \frac{\hat{Sp}\ 1-\hat{Sp}}{r}\right).$$

Finalmente, elintervalo de confianza logit para cada coeficiente kappa promedio es

$$\left(\frac{\exp\ \text{logit}\ \hat{\kappa} - z_{1-\alpha/2}\sqrt{Var\ \text{logit}\ \hat{\kappa}}}{1+\exp\ \text{logit}\ \hat{\kappa} - z_{1-\alpha/2}\sqrt{Var\ \text{logit}\ \hat{\kappa}}},\ \frac{\exp\ \text{logit}\ \hat{\kappa} + z_{1-\alpha/2}\sqrt{Var\ \text{logit}\ \hat{\kappa}}}{1+\exp\ \text{logit}\ \hat{\kappa} + z_{1-\alpha/2}\sqrt{Var\ \text{logit}\ \hat{\kappa}}}\right)$$

donde $\hat{\kappa}$ es $\hat{\kappa}_1$ o $\hat{\kappa}_2$.

Como criterios de aplicación se puede establecer, en términos generales, la siguiente regla:

a). Para tamaños muestrales de entre 100 y 500 individuos, aplicar el intervalo tipo Wald.

b). Para tamaños muestrales superiores a 500, se puede aplicar cualquiera de los tres intervalos, si bien el intervalo bootstrap requiere un mayor esfuerzo computacional.

3. EJEMPLO

Los resultados obtenidos se han aplicado al estudio de Weiner et al [9] sobre el diagnóstico de la enfermedad de la arteria coronaria, utilizando como test diagnóstico un test de ejercicio y como gold estándar la arteriografía coronaria. En la Tabla2 se muestran los resultados obtenidos por Weiner et al para individuos con angina de pecho, y donde la variable *T* modeliza el resultado del test de ejercicio y la variable *D* el resultado de la angiografía coronaria.

Tabla 2. Datos del estudio de Weiner et al (1979).

Frecuencias observadas			
	$T=1$	$T=0$	Total
$D=1$	473	81	554
$D=0$	22	44	66
Total	495	125	620

En la Tabla 3 se muestran los valores estimados de cada parámetro y los intervalos de confianza al 95%. La interpretación de los resultados (en términos de las estimaciones puntuales) es la siguiente:

a). Sensibilidad y especificidad. La sensibilidad estimada toma un valor moderadamente alto, por lo que el test diagnóstico (test de ejercicio) es útil para descartar la enfermedad. La especificidad estimada toma un valor moderado, por lo que el test de ejercicio es moderadamente útil para confirmar la enfermedad en individuos que son sospechosos de tenerla.

b). Razones de verosimilitud. La probabilidad de que el test de ejercicio sea positivo es 2.561 veces mayor en los individuos con la enfermedad de la arteria coronaria que en los individuos sin esta enfermedad. La probabilidad de que el test de ejercicio sea negativo es $4.566 = 1/0.219$ veces mayor en los individuos sin la enfermedad de la arteria coronaria que en los individuos con esta enfermedad.

c). Valores predictivos. El valor predictivo positivo estimado toma un valor muy alto, por lo que el test de ejercicio es muy bueno para confirmar la enfermedad de la arteria coronaria en los individuos pertenecientes a la población objeto de estudio. El valor predictivo negativo estimado toma un valor bajo, por lo que el test de ejercicio no es un test adecuado para descartar la enfermedad en los individuos pertenecientes a la población objeto de estudio.

Tabla 3. Resultados del estudio de Weiner et al (1979).

Sensibilidad y especificidad		
	Estimación	IC al 95%
Se	0.854	0.822 , 0.881
Sp	0.667	0.547 , 0.769
Razones de verosimilitud		
	Estimación	IC al 95%
LR^+	2.561	1.871 , 3.665
LR^-	0.219	0.172 , 0.290
Valores predictivos		
	Estimación	IC al 95%
VPP	0.956	0.934 , 0.971
VPN	0.352	0.274 , 0.439
Coeficiente kappa promedio		
	Estimación	IC logit al 95%
κ_1	0.463	0.360 , 0.568
κ_2	0.319	0.239 , 0.412
Coeficiente kappa ponderado		

c	$\hat{\kappa}\ c$	IC logit al 95%
0.1	0.527	0.408 , 0.639
0.2	0.462	0.370 , 0.584
0.3	0.412	0.338 , 0.539
0.4	0.371	0.310 , 0.502
0.5	0.338	0.285 , 0.471
0.6	0.310	0.264 , 0.444
0.7	0.287	0.245 , 0.420
0.8	0.267	0.229 , 0.399
0.9	0.249	0.214 , 0.380

Coeficiente kappa promedio		
	Estimación	IC logit al 95%
κ_1	0.463	0.360 , 0.568
κ_2	0.319	0.239 , 0.412

d). Coeficiente kappa ponderado. Si el test de ejercicio se utiliza como test previo a un tratamiento de riesgo $0 < c < 0.5$, el coeficiente kappa ponderado estimado toma un valor intermedio para cada valor de c y el acuerdo más allá del azar entre el test de ejercicio y la angiografía coronaria (gold estándar) es principalmente moderado. Si el test de ejercicio se utiliza para un diagnóstico simple $c = 0.5$, el acuerdo más allá del azar entre el test de ejercicio y la angiografía coronaria varía entre mediocre y moderado; y si el test de ejercicio se utiliza como un test de screening $0.5 < c < 1$, el acuerdo más allá del azar entre el test de ejercicio y la angiografía coronaria es principalmente mediocre.

e). Coeficiente kappa promedio. Si el clínico considera que la pérdida asociada a un falso positivo L' es mayor que la pérdida asociada a un falso negativo L , el valor estimado del coeficiente kappa promedio es moderado, y por tanto el acuerdo promedio más allá del azar entre el test de ejercicio y la angiografía es moderada en la población objeto de estudio. Si el clínico considera que la pérdida asociada a un falso negativo L es mayor que la pérdida asociada a un falso positivo L' , el valor estimado del coeficiente kappa promedio es mediocre, y por tanto el acuerdo promedio más allá del azar entre el test de ejercicio y la angiografía es mediocre en la población objeto de estudio. Por tanto, considerando las pérdidas, el test de ejercicio no debería de aplicarse como test de screening en la población objeto de estudio, y si se puede aplicar como test previo a un tratamiento de riesgo en la población objeto de estudio.

4. CONCLUSIONES

La utilización de los tests diagnósticos se ha convertido en una herramienta fundamental en la práctica clínica, por lo que la Estadística ha ido desarrollando métodos para ir resolviendo los problemas que la aplicación práctica de los tests diagnósticos ha dado lugar. En este trabajo se ha realizado una revisión de los parámetros más usuales (sensibilidad, especificidad, razones de verosimilitud y valores predictivos) y de otros

más novedosos (coeficiente kappa ponderado y coeficiente kappa promedio) de un test diagnóstico binario, exponiéndose los mejores intervalos de confianza para cada uno de ellos bajo un muestreo transversal. Todos estos parámetros no son alternativos unos a otros, sino que son parámetros complementarios que, conjuntamente, ayudan a entender el mecanismo de clasificación de un test diagnóstico binario. El presente trabajo tiene como principal objetivo orientar al investigar aplicado (clínico, epidemiólogo,…) sobre la interpretación y aplicación práctica de estos parámetros, y también a aquellos otros investigadores (estadísticos o matemáticos) que deseen introducirse en este ámbito de la Bioestadística.

AGRADECIMIENTOS

Este trabajo ha sido financiado por la Subdirección General de Proyectos de Investigación del Ministerio de Economía y Competitividad, España, Proyecto MTM2012-35591. También quiero dar las gracias al profesor Carlos Bouza y al resto de editores, sin cuyo esfuerzo este libro no sería posible.

REFERENCIAS

1. ZHOU, X.H., Obuchowski, N., McClish, D.,(2002):Statistical Methods in Diagnostic Medicine. John Wiley and Sons: New York.
2. PEPE, M.S.,(2003):The Statistical Evaluation of Medical Tests for Classification and Prediction. Oxford University Press: New York.
3. KRAEMER, H.C., (1992): Evaluating medical tests. SAGE Publications, Newburry Park.
4. ROLDÁN-NOFUENTES, J.A., OLVERA-PORCEL, C., (2014):Average kappa coefficient: a new measure to assess a binary test considering the losses associated with an erroneous classification. **Journal of Statistical Computation and Simulation**, in press, doi: 10.1080/00949655.2014.881816.
5. YU, W., GOU, X., XU, W., (2014):An improved score interval with a modified midpoint for a binomial proportion. **Journal of Statistical Computation and Simulation** 84, 1022-1038.
6. MARTÍN-ANDRÉS, A., ÁLVAREZ-HERNÁNDEZ, M., (2014):Two-tailed approximate confidence intervals for the ratio of proportions**. Statistics and Computing**,24, 65-75.
7. ROLDÁN NOFUENTES, J.A., LUNA DEL CASTILLO, J.D. and MONTERO ALONSO, M.A., (2009):Confidence intervals of weighted kappa coefficient of a binary diagnostic test. **Communications in Statistics - Simulation and Computation** 38, 1562-1578.
8. EFRON, B.,TIBSHIRANI, R.J., (1993): An Introduction to the Bootstrap. Chapman and Hall: New York.
9. WEINER, D. A., RYAN, T. J., MCCABE, C. H., KENNEDY, J. W., SCHLOSS, M., TRISTANI, F., CHAITMAN, B. R., FISHER, L. D., (1979): Exercise stress testing. Correlations among history of angina, ST-segmant response and prevalence of coronary-artery disease in the coronary artery surgery study (CASS). **The New England Journal of Medicine** 301, 230-235.