Universidad de Granada

Departamento de Estadística e Investigación Operativa

Tesis Doctoral

Programa de Doctorado en Estadística e Investigación Operativa

Coeficiente kappa promedio: un nuevo parámetro para evaluar y comparar el rendimiento de test diagnósticos binarios

Mª Carmen Olvera Porcel

Granada, 2015

Editorial: Universidad de Granada. Tesis Doctorales
Autora: María del Carmen Olvera Porcel
ISBN: 978-84-9125-187-3
URI: http://hdl.handle.net/10481/40533

Universidad de Granada

Departamento de Estadística e Investigación Operativa

Coeficiente kappa promedio: Un nuevo parámetro para evaluar y comparar el rendimiento de test diagnósticos binarios

Memoria de TESIS DOCTORAL realizada bajo la dirección del Doctor José Antonio Roldán Nofuentes, del Departamento de Estadística e Investigación Operativa de la Universidad de Granada, que presenta la Licenciada Mª Carmen Olvera Porcel para optar al grado de Doctora en Estadística

Fdo: Mª Carmen Olvera Porcel

Vº Bº del director

Fdo: José Antonio Roldán Nofuentes

La doctorando María del Carmen Olvera Porcel y el director de la tesis José Antonio Roldán Nofuentes, garantizamos, al firmar esta tesis doctoral, que el trabajo ha sido realizado por la doctorando bajo la dirección del director de la tesis y hasta donde nuestro conocimiento alcanza, en la realización del trabajo, se han respetado los derechos de otros autores a ser citados, cuando se han utilizado sus resultados o publicaciones.

Granada, a 12 de Mayo de 2015

Director de la Tesis Doctorando

Fdo.: José A. Roldán Nofuentes Fdo.: María del Carmen Olvera Porcel

La presente Tesis Doctoral está avalada hasta la fecha de su lectura por el artículo:

Roldán Nofuentes J.A., Olvera Porcel M.C., (2015). 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,* 85(8):1601-1620.

DOI:10.1080/00949655.2014.881816

AGRADECIMIENTOS

A mi director de tesis José Antonio Roldán Nofuentes, por su gran ayuda, compromiso, seriedad, conocimientos aportados, paciencia y tiempo dedicado para que este trabajo saliese adelante.

A todos y cada uno de los miembros del Departamento de Medicina Preventiva y Salud Pública por todo los años que he compartido con ellos, por el apoyo y cariño. En especial a Don Ramón Gálvez Vargas que con este trabajo estará orgulloso de mí allá donde esté. Descanse en paz. A Rocío Olmedo Requena, Carmen Amezcua Prieto y Eladio Jiménez Mejías, Anne-Marie Lewis Mikhail, Virginia Martinez Ruiz, Elena Espigares y Elena Moreno e Isabel Salazar porque son unos grandes compañeros y amigos. A Aurora Cavanillas porque siempre me ha estado animando para acabar este trabajo.

A mis padres, mi hermano, mi cuñada y a mi bonita sobrina Marisol. Gracias por el apoyo y comprensión diaria.

A María José Segura, María Bolívar y a Anawa porque me han demostrado que no hace falta mucho tiempo para tener una buena amistad.

A Inma, Silvia, Chus, Yolanda, Lola, Sandra, Klejdha, Francisco Gutiérrez, Santos, Azi, Pedro y a tantas personas a las que estoy muy agradecida y que desde hace poco o mucho tiempo caminan junto a mí.

A mi enfermedad que me ha dado una buena lección sobre la vida y como vivirla.

Índice

Resumen

La incesante evolución de la Medicina en estos últimos tiempos ha hecho necesario que la Estadística desarrolle modelos para resolver los nuevos problemas que se han ido planteando en todos los ámbitos de la Medicina. En este contexto, el diagnóstico médico ocupa un lugar privilegiado. Desde las primeras investigaciones sobre curvas *ROC* en las décadas de los 50 y 60, no se ha cesado de investigar nuevos Métodos Estadísticos de estimación y de test de hipótesis sobre parámetros de test diagnósticos. La presente Tesis Doctoral pretende ser una contribución a la investigación de nuevos parámetros que permitan evaluar y comparar el rendimiento de test diagnósticos. Esta Tesis está centrada en el estudio de los test diagnósticos binarios, cuya evaluación con respecto a un gold estándar da lugar al estudio de tablas 2×2 cuando se trata de un único test diagnóstico, o de tablas de mayor dimensión cuando se trata de dos o más test diagnósticos. En todas las situaciones se asume que el estado de enfermedad de todos los

individuos de la muestra es conocido. Esta Tesis Doctoral está estructurada en tres Capítulos.

En el Capítulo 1 se definen los principales parámetros de un test diagnóstico binario, la sensibilidad y especificidad, el índice de Youden, las razones de verosimilitud, los valores predictivos y el coeficiente kappa ponderado, y se realiza una aportación, el coeficiente kappa promedio. Este nuevo parámetro, que es un promedio de coeficientes kappa ponderados, pretende resolver el problema de la elección del valor del índice de ponderación para el coeficiente kappa ponderado, y se define como una medida del acuerdo promedio más allá del azar entre el test diagnóstico y el gold estándar. El coeficiente kappa promedio presenta unas propiedades que lo validan como medida para evaluar el rendimiento de un test diagnóstico binario.

En el Capítulo 2 se estudian las estimaciones de los parámetros presentados en el Capítulo 1 cuando el muestreo es de tipo transversal. Este tipo de muestreo consiste en la aplicación del test diagnóstico binario y del gold estándar a todos los individuos de una muestra aleatoria, y es el tipo de muestreo más utilizado en la práctica clínica. La aportación que se realiza en este Capítulo es la estimación del coeficiente kappa promedio. Se han estudiado varios intervalos de confianza para este parámetro, se han

realizado experimentos de simulación Monte Carlo para estudiar la cobertura asintótica de estos intervalos y se ha escrito un programa en *R* para estimar este parámetro. Finalmente, los resultados obtenidos se han aplicado a dos ejemplos reales.

En el Capítulo 3 se aborda el problema de la comparación de parámetros de dos (y en algunos casos de más de dos) test diagnósticos binarios bajo un diseño apareado. El diseño apareado consiste en aplicar los dos test diagnósticos binarios y el gold estándar a todos los individuos de una muestra aleatoria, y es el muestreo más utilizado cuando se comparan dos test diagnósticos binarios. Se explican los test de hipótesis e intervalos de confianza para comparar los parámetros estudiados en el Capítulo 1. La aportación en este Capítulo es la comparación de los coeficientes kappa promedio de dos (y de más de dos) test diagnósticos binarios. Se han estudiado varios test de hipótesis para comparar estos parámetros y se han realizado experimentos de simulación Monte Carlo para estudiar el error tipo I y la potencia de estos test de hipótesis. Se ha escrito un programa en *R* para resolver el problema planteado y, finalmente los resultados se han aplicado a un ejemplo real.

Granada, mayo de 2015

Capítulo 1

Medidas de un test diagnóstico binario

Los test diagnósticos son indispensables en la Medicina moderna. Un test diagnóstico (*TD*), también denominado método de diagnóstico, es una prueba médica que se utiliza para diagnosticar la presencia o ausencia de una cierta enfermedad. La ecocardiografía para el diagnóstico de la enfermedad coronaria, el antígeno carbohidrato 19.9 para el diagnóstico del cáncer de colon o la concentración de *PCR* en líquido cefarrolaquídeo para el diagnóstico de la meningitis, son ejemplos de *TDs*. Debido al desarrollo científico que han experimentado las ciencias médicas en estos últimos años, se han ido desarrollando novedosos y sofisticados *TDs*, provocando la

necesidad de investigar nuevos métodos estadísticos para la estimación de la calidad de estos nuevos *TDs*.

La aplicación de un *TD* tiene varios propósitos (McNeil y Adelstein, 1976; Sox et al., 1989; Zhou et al., 2002):

a) Proporcionar información fiable sobre el estado de enfermedad de un individuo (enfermo o no enfermo).

b) Intervenir en la planificación del tratamiento de un individuo

c) Intuir, mediante la investigación, el mecanismo y la naturaleza de la enfermedad.

Asimismo, el resultado de un *TD* depende de varios factores:

a) de la precisión intrínseca del propio test para distinguir entre individuos enfermos y no enfermos (exactitud discriminatoria).

b) de factores externos (por ejemplo, ingesta de medicamentos, alcohol, etc.).

c) de las características de cada individuo (por ejemplo, estado fisiológico anormal del individuo que pueda interferir en la medida del test).

Por ello, la aplicación de un *TD* puede dar lugar a errores, por lo que su exactitud se mide en términos de probabilidades (o de funciones de probabilidades). Para obtener unos estimadores insesgados de esas probabilidades (o de sus funciones), es necesario evaluar el *TD* respecto a un gold estándar. Un gold estándar (*GE*) es una prueba médica que determina de forma objetiva si un individuo tiene o no una enfermedad. La angiografía coronaria para el diagnóstico de la enfermedad coronaria, una biopsia para el diagnóstico del cáncer de colon o un cultivo para el diagnóstico de la meningitis son ejemplos de *GEs*. Existen diferentes tipos de *TDs*:

a) Test diagnósticos binarios: dan lugar a dos resultados: positivo (indicando la presencia provisional de la enfermedad) y negativo (indicando la ausencia provisional de la enfermedad). Por ejemplo, la ecocardiografía para el diagnóstico de la enfermedad coronaria.

b) Test diagnósticos cuantitativos: dan lugar a valores numéricos. Por ejemplo, la concentración de PCR en líquido cefarrolaquídeo para el diagnóstico de la meningitis.

c) Test diagnósticos ordinales: dan lugar a distintos valores con una estructura jerárquica. Por ejemplo, la clasificación de la presencia de

la enfermedad en definitivamente sí, probablemente sí, probablemente no y definitivamente no.

En esta Tesis se estudian los test diagnósticos binarios (*TDBs*), la estimación de sus parámetros y test de hipótesis para comparar sus parámetros, por ser los más frecuentes en la práctica clínica y porque su estudio da lugar al análisis de tablas de contingencia muy usuales en Estadística, como son las tablas 2×2 (en el caso de un único *TDB*) o las tablas 2×4 (en el caso de dos *TDBs*).

Considérese una enfermedad que puede estar presente o no en los individuos de una población y supóngase que se dispone de un *GE* para el diagnóstico de dicha enfermedad. Sea *D* una variable aleatoria que modeliza el resultado del *GE*, de tal forma que si $D=1$ el individuo tiene la enfermedad y si $D=0$ el individuo no tiene la enfermedad. La probabilidad de que un individuo de la población elegido al azar tenga la enfermedad se denominada prevalencia de la enfermedad (p), y es la probabilidad de que $D=1$, esto es $p=P(D=1)$. Considérese un *TDB* cuyo rendimiento es evaluado con respecto a un *GE*. Sea *T* la variable aleatoria que modeliza el resultado del *TDB*: $T=1$ cuando el resultado del test diagnóstico es positivo y $T=0$ cuando el resultado es negativo. Como el *TDB* puede dar lugar a

errores en el diagnóstico de la enfermedad, la evaluación de este con respecto a un *GE* da lugar a cuatro sucesos diferentes: cuando en un individuo enfermo el *TDB* es positivo, se está ante un acierto denominado verdadero positivo (*VP*); cuando en un individuo enfermo el *TDB* es negativo, se está ante un error denominado falso negativo (*FN*); cuando en un individuo no enfermo el *TDB* es positivo, se está ante un error denominado falso positivo (*FP*); y cuando en un individuo no enfermo el *TDB* es negativo, se está ante un acierto denominado verdadero negativo (*VN*). En la Tabla 1.1 se resumen estos cuatro posibles resultados.

Tabla 1.1. Sucesos asociados a la aplicación de un *TDB*.

	$T=1$	$T=0$
$D=1$	*VP*	*FN*
$D=0$	*FP*	*VN*

A continuación se estudian las medidas clásicas del rendimiento de un *TDB*. En primer lugar se definen las medidas de calidad que no dependen de la prevalencia de la enfermedad (sensibilidad, especificidad, índice de Youden y razones de verosimilitud), y a continuación las que sí dependen de la prevalencia (valores predictivos y coeficiente kappa ponderado). Por

último, y como aportación, se define una nueva medida para estudiar el rendimiento de un test diagnóstico binario: el coeficiente kappa promedio.

1.1. Sensibilidad y Especificidad

La sensibilidad (Se) es la probabilidad de que el *TDB* de un resultado positivo cuando el individuo está enfermo, esto es $Se = P(T = 1 \mid D = 1)$. La sensibilidad hace referencia a la probabilidad de un verdadero positivo. La especificidad (Sp) es la probabilidad de que el *TDB* de un resultado negativo cuando el individuo no tiene la enfermedad, esto es $Sp = P(T = 0 \mid D = 0)$. La especificidad hace referencia a la probabilidad de un verdadero negativo. La suma de las probabilidades de un verdadero positivo y de un falso negativo es igual a la unidad,

$$P(T = 1 \mid D = 1) + P(T = 0 \mid D = 1) = 1, \qquad (1.1)$$

y de forma similar, la suma de las probabilidades de un verdadero negativo y de un falso positivo es igual a la unidad,

$$P(T = 0 \mid D = 0) + P(T = 1 \mid D = 0) = 1, \qquad (1.2)$$

siendo $P(T=0 \mid D=1)$ la probabilidad de un falso negativo y $P(T=1 \mid D=0)$ la probabilidad de un falso positivo.

La sensibilidad y la especificidad de un *TDB* dependen únicamente de la habilidad intrínseca del test para distinguir individuos enfermos e individuos no enfermos. Estos parámetros dependen de las bases físicas, químicas o biológicas con las que se han desarrollado el *TDB*.

Un *TDB* con una alta sensibilidad es útil para descartar la enfermedad, pues la probabilidad de un falso negativo es pequeña; y un *TDB* con una alta especificidad es útil para confirmar la enfermedad, pues la probabilidad de un falso positivo es pequeña.

Una de las principales utilidades clínicas de un *TD* es el screening. Un screening 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. Ejemplos muy conocidos son el screening del cáncer de mama, el screening del cáncer de próstata, etc. Para que un *TDB* se pueda utilizar como test de screening, se deben verificar las siguientes características:

a) De la población:

- La prevalencia de la enfermedad sea suficiente grande.
- Sea susceptible de aplicación de diferentes pruebas médicas y tratamientos.

b) De la enfermedad:

- Morbilidad y mortalidad significativas.
- Que tenga un tratamiento eficaz y aceptable.
- Periodo presintomático detectable.
- Mejora con un tratamiento precoz.

c) Del *TD*:

- Alta sensibilidad y especificidad.
- Bajo coste.
- Su aplicación suponga un bajo riesgo para el individuo.
- Exista un *GE*.

1.2. Índice de Youden

El índice de Youden (Youden, 1950) mide la diferencia entre la proporción de los verdaderos positivos y de los falsos positivos. Se define como

$$Y = Se + Sp - 1 = Se - (1 - Sp)\,. \qquad (1.3)$$

Este parámetro toma valores entre -1 y 1 y tiene de las siguientes propiedades:

a) Si la sensibilidad y la especificidad son complementarios $(Se = 1 - Sp)$ entonces el índice de Youden vale 0 y el *TDB* es no informativo. En esta el *TDB* no está relacionado con la enfermedad y el diagnóstico de la misma se puede realizar lanzado una moneda cuya probabilidad de cara sea igual a ½.

b) Si el índice de Youden es menor que 0, entonces $T = 1$ debe ser un resultado negativo y $T = 0$ un resultado positivo. Por tanto si $Y < 0$, los resultados del *TDB* se deben intercambiar.

c) A todo *TDB* hay que exigirle que su índice de Youden sea mayor que 0, es decir que $Se + Sp > 1$.

El inconveniente que presenta el índice de Youden es que no es una medida válida para comparar el rendimiento de *TDBs*, ya que dos *TDBs* con un mismo índice de Youden pueden tener diferente exactitud. Por ejemplo, si un *TDB* tiene una sensibilidad del 90% y una especificidad del 60% (el *TDB* es muy útil para descartar la enfermedad y poco útil para confirmarla), su índice de Youden es 0.5; otro *TDB* con una sensibilidad del 60% y una especificidad del 90%, tiene un índice de Youden igual al del test anterior y sin embargo su utilidad clínica es muy diferente (el *TDB* es muy útil para confirmar la enfermedad y poco útil para descartarla).

1.3. Razones de verosimilitud

Las razones de verosimilitud (*LRs*) son otros parámetros que se utilizan para evaluar la exactitud de un *TDB*. La razón de verosimilitud es un cociente de dos probabilidades definido como

$$LR = \frac{P\left(T = i \mid D = 1\right)}{P\left(T = i \mid D = 0\right)}\ , i = 0,1\,. \tag{1.4}$$

La razón de verosimilitud representa el cociente entre la probabilidad de un resultado, positivo o negativo, del *TDB* en individuos enfermos y la

probabilidad del mismo resultado del *TDB* en individuos no enfermos. Cuando el resultado del *TDB* es positivo, la *LR*, denominada *LR* positiva, es el cociente entre la sensibilidad y uno menos la especificidad, esto es,

$$LR(+) = \frac{P(T=1 \mid D=1)}{P(T=1 \mid D=0)} = \frac{Se}{1-Sp}. \tag{1.5}$$

Cuando el resultado del *TDB* es negativo, la *LR*, denominada *LR* negativa, es el cociente entre uno menos la sensibilidad y la especificidad, es decir,

$$LR(-) = \frac{P(T=0 \mid D=1)}{P(T=0 \mid D=0)} = \frac{1-Se}{Sp}. \tag{1.6}$$

Si la *LR* (positiva o negativa) es igual a 1, el resultado del *TDB* es igualmente probable en individuos enfermos y en individuos no enfermos, es decir, el *TDB* y el *GE* son independientes. Si la *LR* positiva es mayor que 1, un resultado positivo del *TDB* es más probable en los individuos enfermos que en los no enfermos. Si la *LR* negativa es menor que 1, un resultado negativo del *TDB* es más probable en los individuos no enfermos que en los enfermos. Si $LR(+) = \infty$ y $LR(-) = 0$, el *TDB* clasifica correctamente a todos los individuos, (enfermos y no enfermos) y por tanto el *TDB* es considerado como un *GE*.

Las *LRs* son particularmente útiles como medidas para evaluar la exactitud de un *TDB*, ya que permiten comprender con que fuerza un resultado positivo (o negativo) del *TDB* indica la presencia (o ausencia) de la enfermedad. Antes de aplicar el *TDB*, la probabilidad de que un individuo tenga la enfermedad es la prevalencia de la enfermedad, siendo la odds pre-test

$$\text{odds pre-test} = \frac{\text{probabilidad pre-test}}{1 - \text{probabilidad pre-test}}. \tag{1.7}$$

Cuando se aplica el *TDB*, la odds de que un individuo tenga la enfermedad, denominada odds post-test, es

$$\text{odds post-test}\left(T = i\right) = \frac{P\left(D = 1 \middle| T = i\right)}{P\left(D = 0 \middle| T = i\right)}\ ,\ i = 0,1. \tag{1.8}$$

Las *LRs* relacionan la odds pre-test con las odds post-test, esto es

$$\begin{aligned} &\text{odds post-test } \left(T = 1\right) = LR(+) \times \text{odds pre-test} \\ &\text{odds post-test } \left(T = 0\right) = LR(-) \times \text{odds pre-test,} \end{aligned} \tag{1.9}$$

y por tanto las *LRs* cuantifican el cambio en las odds pre-test una vez aplicado el *TDB*.

1.4. Valores predictivos

Los valores predictivos positivo y negativo representan la exactitud clínica de un *TDB*. El valor predictivo positivo (*VPP*) es la probabilidad de que un individuo esté enfermo cuando el resultado del *TDB* es positivo, esto es $VPP = P(D = 1 \mid T = 1)$. El valor predictivo negativo (*VPN*) es la probabilidad de que un individuo no esté enfermo cuando el resultado del *TDB* es negativo, es decir $VPN = P(D = 0 \mid T = 0)$. Los valores predictivos (*VPs*) dependen de la habilidad intrínseca del test diagnóstico (sensibilidad y especificidad) y de la prevalencia de la enfermedad (p). Aplicando el teorema de Bayes, los *VPs* se expresan como

$$VPP = \frac{p \times Se}{p \times Se + (1-p) \times (1-Sp)} \tag{1.10}$$

y

$$VPN = \frac{(1-p) \times Sp}{p \times (1-Se) + (1-p) \times Sp}. \tag{1.11}$$

Si un *VPP* de un *TDB* es elevado, el test es útil para confirmar la enfermedad; y si el *VPN* es elevado, el *TDB* es útil para descartarla.

1.5. Coeficiente kappa ponderado

El coeficiente kappa ponderado (Kraemer, 1992; Kraemer et al., 2002) es una medida del acuerdo más allá del azar entre el test diagnóstico binario y el gold estándar, y es un parámetro que tiene en cuenta las pérdidas asociadas a una clasificación errónea con el test diagnóstico. Sea L la pérdida que se comete cuando en un individuo enfermo el resultado del *TDB* es negativo (falso negativo), y sea L' la pérdida que se comete cuando en un individuo no enfermo el resultado del *TDB* es positivo (falso positivo). Por tanto, La pérdida L está asociada a un falso negativo y la pérdida L' lo está a un falso positivo. Las pérdidas L y L' son iguales a cero si todos los individuos sean clasificados correctamente por el *TDB*. En la Tabla 1.2 y la Tabla 1.3 se presentan las pérdidas y las probabilidades asociadas a la evaluación de un *TDB* con respecto a un *GE*. En términos de estas pérdidas y probabilidades se define el riesgo de error (Bloch, 1997) o pérdida esperada como la pérdida promedio que se comete cuando se clasifica erróneamente a un individuo, esto es,

$$p\left(1-Se\right)L+q\left(1-Sp\right)L', \tag{1.12}$$

y también se define la pérdida aleatoria como la pérdida que se comete cuando el *TDB* y el *GE* son independientes, es decir cuando $P\left(T=i\middle|D=j\right)=P\left(T=i\right)$,

$$p\left\{p\left(1-Se\right)+qSp\right\}L+q\left\{pSe+q\left(1-Sp\right)\right\}L', \tag{1.13}$$

siendo $q=1-p$.

Tabla 1.2. Pérdidas asociadas a la aplicación de un *TDB*.

	$T=1$	$T=0$	Total
$D=1$	0	L	L
$D=0$	L'	0	L'
Total	L'	L	$L+L'$

Tabla 1.3. Probabilidades asociadas a la aplicación de un *TDB* a una muestra aleatoria.

	$T=1$	$T=0$	Total
$D=1$	pSe	$p\left(1-Se\right)$	p
$D=0$	$\left(1-p\right)\left(1-Sp\right)$	$\left(1-p\right)Sp$	$1-p$
Total	$Q=pSe+\left(1-p\right)\left(1-Sp\right)$	$1-Q=p\left(1-Se\right)+\left(1-p\right)Sp$	1

En término de las pérdidas esperada y aleatoria, se define el coeficiente kappa ponderado como

$$\kappa = \frac{\text{Pérdida debida al azar} - \text{Pérdida esperada}}{\text{Pérdida debida al azar} - \text{Min}\left(\text{Pérdida esperada}\right)}. \qquad (1.14)$$

Como la pérdida esperada mínima es cero, el coeficiente kappa ponderado es

$$\kappa = \frac{\text{Pérdida debida al azar} - \text{Pérdida esperada}}{\text{Pérdida debida al azar}}. \qquad (1.15)$$

El coeficiente kappa ponderado es por tanto una medida de la discrepancia relativa entre la pérdida debida al azar y la pérdida esperada, y mide el acuerdo clasificatorio más allá del azar entre el test diagnóstico binario y el gold estándar. Sus valores varían entre -1 y 1. Al sustituir en la ecuación (1.15) las expresiones de la pérdida esperada (1.12) y e la pérdida aleatoria (1.13), el coeficiente kappa ponderado es

$$\kappa(c) = \frac{pqY}{p\left(1-Q\right)c + qQ\left(1-c\right)}, \qquad (1.16)$$

donde $Q = pSe + \left(1-p\right)\left(1-Sp\right)$, $q = 1-p$, Y es el índice de Youden y $c = L/\left(L+L'\right)$ es el índice de ponderación. El índice de ponderación representa la pérdida relativa entre los falsos positivos y los falsos negativos

y varía entre 0 y 1. Si $L=0$ entonces $c=0$ y el coeficiente kappa ponderado es

$$\kappa(0)=\frac{pY}{Q}=\frac{Sp-(1-Q)}{Q}=\frac{VPP-p}{1-p}. \tag{1.17}$$

Si $L'=0$ entonces $c=1$ y el coeficiente kappa ponderado es

$$\kappa(1)=\frac{qY}{1-Q}=\frac{Se-Q}{1-Q}=\frac{VPN-q}{p}. \tag{1.18}$$

Los coeficientes $\kappa(1)$ y $\kappa(0)$ son la sensibilidad corregida por azar y la especificidad corregida por azar respectivamente. Si $L=L'$ entonces $c=0.5$ y el coeficiente kappa ponderado (denominado coeficiente kappa de Cohen) es

$$\kappa(0.5)=\frac{pqY}{\frac{p+Q}{2}-pQ}=\frac{2}{\frac{1}{\kappa(0)}+\frac{1}{\kappa(1)}}=\frac{2\kappa(0)\kappa(1)}{\kappa(0)+\kappa(1)}, \tag{1.19}$$

por lo que el coeficiente kappa de Cohen es la media armónica de $\kappa(0)$ y $\kappa(1)$. En términos del coeficiente kappa de Cohen, el coeficiente kappa ponderado se escribe como

$$\kappa(c)=\frac{p(1-Q)+qQ}{2\{p(1-Q)c+qQ(1-c)\}}\kappa(0.5), \tag{1.20}$$

donde

$$\frac{p(1-Q)+qQ}{p(1-Q)c+qQ(1-c)} \tag{1.21}$$

es la inversa de una media ponderada de c y $(1-c)$. El coeficiente kappa ponderado se puede escribir en términos de p, Q, $\kappa(0)$ y $\kappa(1)$ como

$$\kappa(c)=\frac{p(1-Q)c\kappa(1)+qQ(1-c)\kappa(0)}{p(1-Q)c+qQ(1-c)}, \tag{1.22}$$

siendo por tanto una media ponderada de $\kappa(0)$ y $\kappa(1)$. También se puede escribir exclusivamente en términos de $\kappa(0)$ y $\kappa(1)$ como

$$\kappa(c)=\frac{\kappa(0)\kappa(1)}{c\kappa(0)+(1-c)\kappa(1)}, \tag{1.23}$$

y también en términos del coeficiente φ (medida de asociación entre el *TDB* y el *GE*) como

$$\kappa(c)=\frac{\varphi^2}{c\kappa(0)+(1-c)\kappa(1)}, \tag{1.24}$$

donde $\varphi=\sqrt{\kappa(0)\kappa(1)}$.

Por tanto, el coeficiente kappa ponderado depende no solamente de la sensibilidad y especificidad del *TDB* y de la prevalencia de la enfermedad, sino también del índice de ponderación. El índice de ponderación es una medida relativa entre las pérdidas L y L'. En la práctica, las dos pérdidas L y L' son desconocidas. Por ejemplo, considérese el diagnóstico del cáncer de mama utilizando como *TDB* una mamografía. Si la mamografía es positiva en una mujer que no tiene el cáncer (falso positivo), la mujer será sometida a una biopsia que finalmente descartará la presencia de la enfermedad. La pérdida L' será determinada por los costes económicos debidos a las pruebas médicas y también por el riesgo, estrés, ansiedad,…, causados a la mujer. Si la mamografía es negativa en una mujer que tiene el cáncer, la mujer será diagnosticada un tiempo más tarde, pero la enfermedad habrá avanzado y la posibilidad de éxito del tratamiento habrá disminuido. La pérdida L será determinada a partir de estas consideraciones. Por tanto, las pérdidas L y L' no solamente se miden en términos económicos, sino también en términos del riesgo, estrés,…, causados al individuo. Por ese motivo, en la práctica no es posible determinar los valores de las pérdidas y, por tanto, no se puede determinar el valor del índice de ponderación. Por este mismo motivo, la pérdida L se sustituye por "la importancia de los falsos negativos" y la pérdida L' por "la importancia de los falsos

positivos". Sin embargo, el valor del índice de ponderación *c* se puede fijar en función del objetivo para el que se va a utilizar el *TDB*. Si el *TDB* se va a utilizar como test definitivo previo a un tratamiento de riesgo, situación en la que los falsos positivos tienen más importancia que los falsos negativos, entonces $L' > L$ y $0 < c < 0.5$. Si el *TDB* se va a utilizar como un test de screening, situación en la que los falsos negativos tienen más importancia que los falsos positivos, entonces $L > L'$ y $0.5 < c < 1$. Si los falsos positivos y falsos negativos tienen la misma importancia entonces $c = 0.5$. Así por ejemplo, si el clínico va a utilizar el *TDB* como un test definitivo y considera que los falsos positivos son el doble de importantes que los falsos negativos, entonces asignará al índice de ponderación *c* el valor $1/3$.

El coeficiente kappa ponderado de un test binario presenta las siguientes propiedades:

a) Si el acuerdo clasificatorio entre el *TDB* y el *GE* es perfecto $(Se = Sp = 1)$ entonces la pérdida esperada es 0 y $\kappa(c) = 1$.

b) Si la sensibilidad y la especificidad son complementarias $(Se = 1 - Sp)$, el *TDB* es independiente del *GE* y por lo tanto $\kappa(c) = 0$.

c) Si la pérdida debida al azar es superior a la pérdida esperada entonces $\kappa(c) > 0$. Si la pérdida debida al azar es inferior a la pérdida esperada entonces $\kappa(c) < 0$ y los resultados del *TDB* se deben intercambiar, es decir, $T = 1$ debe ser el resultado negativo y $T = 0$ debe ser el resultado positivo. Por tanto, el análisis se debe limitar a valores positivos del coeficiente kappa ponderado. Sus valores se catalogan en la siguiente escala (Landis y Koch, 1977): de 0 a 0.20 el acuerdo es malo, de 0.21 a 0.40 mediocre, de 0.41 a 0.60 moderado, de 0.61 a 0.80 bueno y de 0.81 a 1 muy bueno. Otra clasificación basada en niveles de significación clínica (Cicchetti, 2001) es: < 0.40 malo, $0.40 - 0.59$ regular, $0.60 - 0.74$ bueno y $0.75 - 1$ muy bueno.

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

El coeficiente kappa ponderado es una medida válida para la evaluar y comparar el rendimiento de un *TDB* con respecto a un *GE*. El principal inconveniente en la aplicación práctica de este parámetro radica en conocer

el índice de ponderación *c*, cuyo valor es fijado por el investigador en función de su conocimiento sobre la importancia relativa entre los falsos negativos y los falsos positivos y sobre el fin para el que se vaya a utilizar el *TDB* (por ejemplo, test definitivo previo a un tratamiento de riesgo o test de screening). En la práctica, el clínico no siempre 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. En la Tabla 1.4 se muestran los valores del coeficiente kappa ponderado cuando $Se = 0.85$, $Sp = 0.70$ y $p = 0.10$, y para diferentes valores del índice de ponderación *c*. Si el *TDB* se va a utilizar como un test definitivo previo a un tratamiento de riesgo $(0 < c < 0.5)$, el acuerdo más allá del azar entre el *TDB* y el *GE* varía entre malo y mediocre dependiendo de la elección del valor de *c*; y si el *TDB* se va a utilizar como un test de screening $(0.5 < c < 1)$, el acuerdo más allá del azar entre el *TDB* y el *GE* varía entre mediocre, moderado y bueno.

Por consiguiente, el grado del acuerdo entre el *TDB* y el *GE* depende en gran medida del valor asignado al índice de ponderación *c*, incluso para una misma utilidad del *TDB* (test previo a un tratamiento de riesgo o test de screening). La asignación de distintos valores al índice de ponderación y el análisis de los resultados obtenidos es una forma de proceder común en la práctica (Roldán Nofuentes et al, 2009). El problema de la asignación de valores al índice de ponderación es la esencia de esta Tesis.

Tabla 1.4. Valores del coeficiente kappa ponderado.

$Se = 0.85 \;\; Sp = 0.70 \;\; p = 10\%$			
c	$\kappa(c)$	c	$k(c)$
0.05	0.16	0.55	0.28
0.10	0.17	0.60	0.30
0.15	0.18	0.65	0.32
0.20	0.18	0.70	0.35
0.25	0.19	0.75	0.39
0.30	0.20	0.80	0.43
0.35	0.21	0.85	0.48
0.40	0.23	0.90	0.55
0.45	0.24	0.95	0.64

A continuación se estudia una nueva medida de acuerdo más allá del azar entre el *TDB* y el *GE* que no depende del índice de ponderación *c*: el coeficiente kappa promedio.

1.6. Coeficiente kappa promedio

Para unos valores fijos de la sensibilidad, especificidad y prevalencia, el coeficiente kappa promedio es una función continua del índice de ponderación *c*. Si el clínico considera que la pérdida asociada a un falso positivo es mayor que la pérdida asociada a un falso negativo (tal y como ocurre cuando el *TDB* se utiliza como test previo a un tratamiento de riesgo), esto si $L' > L$ y por tanto $0 < c < 0.5$, el coeficiente kappa promedio se define como

$$\kappa_1 = \frac{1}{0.5}\int_0^{0.5} \kappa(c)\,dc = 2\int_0^{0.5} \kappa(c)\,dc, \qquad (1.25)$$

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 la pérdida asociada a un falso negativo es mayor que la pérdida asociada a un falso positivo (tal y como ocurre cuando el *TDB* se utiliza como test de screening), esto si

$L > L'$ y por tanto $(0.5 < c < 1)$, el coeficiente kappa promedio se define como

$$\kappa_2 = \frac{1}{0.5}\int_{0.5}^{1} \kappa(c)\, dc = 2\int_{0.5}^{1} \kappa(c)\, dc\,, \tag{1.26}$$

por lo que el coeficiente kappa promedio (κ_2) es el valor medio de la función $\kappa(c)$ cuando $0.5 < c < 1$. Sustituyendo en las ecuaciones (1.25) y (1.26) el coeficiente kappa ponderado dado por la expresión (1.23) (o por (1.16), (1.20), (1.22) o (1.24)) y resolviendo las integrales definidas se obtiene que

$$\kappa_1 = \begin{cases} \dfrac{2\kappa(0)\kappa(1)}{\kappa(0)-\kappa(1)} log\left[\dfrac{\kappa(0)+\kappa(1)}{2\kappa(1)}\right], & p \neq Q \\ Y, & p = Q \end{cases} \tag{1.27}$$

y

$$\kappa_2 = \begin{cases} \dfrac{2\kappa(0)\kappa(1)}{\kappa(0)-\kappa(1)} log\left[\dfrac{2\kappa(0)}{\kappa(0)+\kappa(1)}\right], & p \neq Q \\ Y, & p = Q, \end{cases} \tag{1.28}$$

donde $log[\cdot]$ 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. Para $p \neq Q$, sustituyendo en las ecuaciones (1.27) y (1.28) $\kappa(0)$ y $\kappa(1)$ por sus expresiones dadas en las ecuaciones (1.17) y (1.18), y realizando las operaciones algebraicas, los coeficientes kappa promedio se escriben en términos de *p*, *Q* e *Y* como

$$\kappa_1 = \frac{2p(1-p)Y}{p-Q} log\left[\frac{\frac{p+Q}{2} - pQ}{(1-p)Q}\right] \tag{1.29}$$

y

$$\kappa_2 = \frac{2p(1-p)Y}{p-Q} log\left[\frac{p(1-Q)}{\frac{p+Q}{2} - pQ}\right]. \tag{1.30}$$

Las expresiones (1.29) y (1.30) también se pueden obtener considerando en las integrales (1.25) y (1.26) la expresión del coeficiente kappa ponderado dado por las ecuaciones (1.16), (1.22), (1.23) o (1.24). Asimismo, para $p \neq Q$ los coeficientes kappa promedio se puede expresar en términos de $\kappa(0)$, $\kappa(1)$ y $\kappa(c)$ como

$$\kappa_1 = \frac{2\left[c\kappa(0)+(1-c)\kappa(1)\right]\kappa(c)}{\kappa(0)-\kappa(1)} log\left[\frac{\kappa(0)+\kappa(1)}{2\kappa(1)}\right] \qquad (1.31)$$

y

$$\kappa_2 = \frac{2\left[c\kappa(0)+(1-c)\kappa(1)\right]\kappa(c)}{\kappa(0)-\kappa(1)} log\left[\frac{2\kappa(0)}{\kappa(0)+\kappa(1)}\right]. \qquad (1.32)$$

Como el coeficiente kappa ponderado es una medida del acuerdo más allá del azar entre el *TDB* y el *GE*, los dos coeficientes kappa promedio (que se calculan a partir de coeficientes kappa ponderados) son medidas del acuerdo promedio más allá del azar entre el *TDB* y el *GE*, y no dependen del índice de ponderación *c*. Los coeficientes kappa promedio κ_1 y 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 *TDB* se debe limitar a los valores positivos del coeficiente kappa ponderado (propiedad (c) del coeficiente kappa ponderado), los valores de los coeficientes κ_1 y κ_2 son mayores que 0 y menores que 1.

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

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

$$\kappa_1 = 2\varphi^2 \frac{log\left[\kappa(0.5)\right] - log\left[\kappa(0)\right]}{\kappa(1) - \kappa(0)} \tag{1.33}$$

y

$$\kappa_2 = 2\varphi^2 \frac{log\left[\kappa(1)\right] - log\left[\kappa(0.5)\right]}{\kappa(1) - \kappa(0)} \tag{1.34}$$

respectivamente, siendo $\varphi = \sqrt{\kappa(0)\kappa(1)}$. 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) Para valores fijos de la sensibilidad, especificidad y prevalencia (o de $\kappa(0)$ y $\kappa(1)$), el coeficiente kappa ponderado es una función del índice c 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 que tiene un determinado valor del índice de ponderación c. Esto es, $\kappa_i = \kappa(c)$ para algún valor de c, por lo que sustituyendo en la ecuación (1.16) (o en (1.22), (1.23) o (1.24)) $\kappa(c)$ por κ_i se determina el valor del índice de ponderación asociado al coeficiente kappa promedio κ_i. 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 $x=\kappa_1$ la expresión anterior es la varianza del coeficiente kappa ponderado entorno a κ_1. De forma

similar, El coeficiente kappa promedio κ_2 minimiza la expresión $2\int_{0.5}^{1}\{\kappa(c)-x\}^2\,dc$. Cuando $x=\kappa_2$ la expresión anterior es la varianza del coeficiente kappa ponderado entorno a κ_2.

Una vez definidos los coeficientes kappa promedios y expuestas sus propiedades, estos son unos parámetros que permiten evaluar y comparar el rendimiento de *TDBs* con respecto a un *GE*.

En el siguiente Capítulo se estudia la estimación de los parámetros estudiados en este primer Capítulo.

Capítulo 2

Estimación de los parámetros de un test diagnóstico binario

En el Capítulo 1 se han definido las principales medidas de un *TDB*, y en este Capítulo 2 se estudian las estimaciones de estas medidas bajo un muestreo transversal. Un muestreo o diseño transversal consiste en aplicar el *GE* y el *TDB* a todos los individuos de una muestra aleatoria de n individuos. Este diseño es el que más se utiliza en la práctica, ya que a partir de el se pueden estimar todas las medidas del *TDB*. La Figura 2.1 representa este diseño y en la Tabla 2.1 se presentan las frecuencias obtenidas bajo este diseño.

Figura 2.1. Diseño de un estudio transversal: T representa el resultado del *TDB* y D el resultado del *GE*.

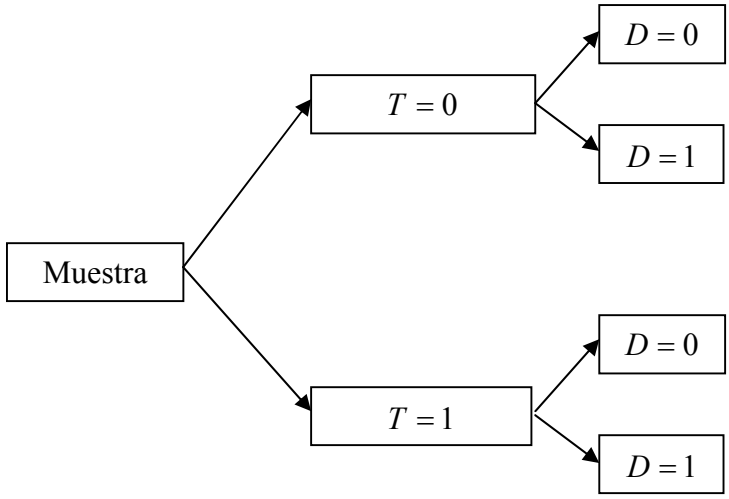

Tabla 2.1. Frecuencias observadas al evaluar un *TDB* bajo un diseño transversal.

	$T=1$	$T=0$	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

A continuación se estudia la estimación puntual y por intervalos de los parámetros de un *TDB* estudiados en el Capítulo 1.

2.1. Sensibilidad y especificidad

La sensibilidad y la especificidad de un *TDB* dependen de la habilidad intrínseca del *TDB* para distinguir entre enfermos y son los parámetros fundamentales de un *TDB*. Las frecuencias de la Tabla 2.1 se distribuyen como una distribución multinomial con probabilidades dadas en la Tabla 1.3 del Capítulo 1. Condicionando en los totales de las filas, es decir en el resultado del *GE* (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 máximo verosímiles de la sensibilidad y de la especificidad son

$$\hat{Se} = \frac{s_1}{s} \quad \text{y} \quad \hat{Sp} = \frac{r_0}{r}, \tag{2.1}$$

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

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

A continuación se presentan cuatro intervalos de confianza para la sensibilidad y especificidad, el intervalo exacto y tres intervalos aproximados.

2.1.1. Intervalo de confianza exacto de Clopper-Pearson

Clopper y Pearson (1934) estudiaron un intervalo de confianza exacto para una proporción binomial. Así, el intervalo de confianza exacto de Clopper y Pearson (1934) para la sensibilidad a la confianza $100\left(1-\alpha\right)\%$ es

$$Se\in\left(\frac{s_1}{s_1+\left(s_0+1\right)F_{\alpha/2}\left(2\left(s_0+1\right),2s_1\right)}\ ,\ \frac{\left(s_1+1\right)F_{\alpha/2}\left(2\left(s_1+1\right),2s_0\right)}{s_0+\left(s_1+1\right)F_{\alpha/2}\left(2\left(s_1+1\right),2s_0\right)}\right) \quad (2.2)$$

y el de la especificidad es

$$Sp\in\left(\frac{r_0}{r_0+\left(r_1+1\right)F_{\alpha/2}\left(2\left(r_1+1\right),2r_0\right)}\ ,\ \frac{\left(r_0+1\right)F_{\alpha/2}\left(2\left(r_0+1\right),2r_1\right)}{r_1+\left(r_0+1\right)F_{\alpha/2}\left(2\left(r_0+1\right),2r_1\right)}\right), \quad (2.3)$$

donde $F_{\alpha/2}\left(v_1,v_2\right)$ es el valor de una *F* de Snedecor con v_1 y v_2 grados de libertad que deja a su derecha un área $\alpha/2$.

2.1.2. Intervalo de confianza de Wilson

Este intervalo de confianza (Wilson, 1927), llamado también intervalo de confianza score, es un intervalo clásico en la literatura estadística para una proporción binominal. El intervalo de confianza de Wilson a la confianza del $100(1-\alpha)\%$ para la sensibilidad es

$$Se \in \frac{s_1 + \frac{z^2_{1-\alpha/2}}{2}}{s + z^2_{1-\alpha/2}} \pm \frac{z_{1-\alpha/2}}{s + z^2_{1-\alpha/2}} \sqrt{\frac{z^2_{1-\alpha/2}}{4} + \frac{s_1 s_0}{s}} \tag{2.4}$$

y para la especificidad

$$Sp \in \frac{r_0 + \frac{z^2_{1-\alpha/2}}{2}}{r + z^2_{1-\alpha/2}} \pm \frac{z_{1-\alpha/2}}{r + z^2_{1-\alpha/2}} \sqrt{\frac{z^2_{1-\alpha/2}}{4} + \frac{r_0 r_1}{r}}\,, \tag{2.5}$$

donde $z_{1-\alpha/2}$ es el percentil $100(1-\alpha/2)\%$ de la distribución normal estándar.

2.1.3. Intervalo de confianza de Agresti y Coull

El intervalo de Agresti y Coull (1998) es otro intervalo para una proporción binominal ampliamente conocido. El intervalo de Agresti y Coull (1998) para la sensibilidad es

$$Se \in \frac{s_1+2}{s+4} \pm \frac{z_{1-\alpha/2}}{s+4}\sqrt{\frac{(s_1+2)(s_0+2)}{s+4}} \tag{2.6}$$

y para la especificidad

$$Sp \in \frac{r_0+2}{r+4} \pm \frac{z_{1-\alpha/2}}{r+4}\sqrt{\frac{(r_0+2)(r_1+2)}{r+4}}\,. \tag{2.7}$$

2.1.4. Intervalo de confianza score modificado de Yu et al

Yu et al (2014) han propuesto un nuevo intervalo de confianza aproximado para una proporción binominal, denominado intervalo score modificado, basado en una modificación del punto medio del intervalo score de Wilson. Adaptando este intervalo a los datos de la Tabla 2.1, el intervalo score modificado para la sensibilidad es

$$Se \in 0.5 + \frac{s+\frac{z_{1-\alpha/2}^4}{53}}{s+z_{1-\alpha/2}^2}\left(\frac{s_1}{s}-0.5\right) \pm \frac{z_{1-\alpha/2}}{s+z_{1-\alpha/2}^2}\sqrt{\frac{z_{1-\alpha/2}^2}{4}+\frac{s_1 s_0}{s}}\,. \tag{2.8}$$

De forma similar, el intervalo de confianza score modificado para la especificidad es

$$Sp \in 0.5 + \frac{r + \frac{z_{1-\alpha/2}^4}{53}}{r + z_{1-\alpha/2}^2}\left(\frac{r_0}{r} - 0.5\right) \pm \frac{z_{1-\alpha/2}}{r + z_{1-\alpha/2}^2}\sqrt{\frac{z_{1-\alpha/2}^2}{4} + \frac{r_0 r_1}{r}}\,. \qquad (2.9)$$

2.1.5. Intervalo arcoseno de Martín-Andrés y Álvarez-Hernández

Martín-Andrés y Álvarez-Hernández (2014) han evaluado 29 intervalos de confianza aproximados (excluyendo el de Yu et al (2014)) para una proporción binomial, recomendando utilizar un intervalo basado en la transformación arcoseno incrementando los datos en 0.5. Aplicando este intervalo a la situación aquí analizada, se obtienen los siguientes intervalos para la sensibilidad y especificidad respectivamente,

$$Se \in \sin^2\left(\sin^{-1}\sqrt{\frac{s_1 + 0.5}{s+1}} \pm \frac{z_{1-\alpha/2}}{\sqrt{4(s+1)}}\right) \qquad (2.10)$$

y

$$Sp \in \sin^2\left(\sin^{-1}\sqrt{\frac{r_0 + 0.5}{r+1}} \pm \frac{z_{1-\alpha/2}}{\sqrt{4(r+1)}}\right) \qquad (2.11)$$

Martín-Andrés y Álvarez-Hernández (2015) han comparado el rendimiento de los intervalos anteriores, recomendando: (*i*) para $s \leq 80$ $\left(r \leq 80\right)$ y $\alpha = 1\%$ o 5% utilizar el intervalo de Yu et al (2014); (*ii*) para $s \geq 100$ $\left(r \geq 100\right)$ y $\alpha = 10\%$ utilizar el intervalo arcoseno de Martín-Andrés y Álvarez-Hernández (2014); (*iii*) en otras situaciones utilizar el intervalo de Agresti y Coull (1998).

2.2. Razones de verosimilitud

Las razones de verosimilitud son un cociente de la sensibilidad y la especificidad, por lo que es muy fácil obtener sus estimadores puntuales, dados por las siguientes expresiones

$$\hat{LR}(+) = \frac{\hat{Se}}{1-\hat{Sp}} = \frac{s_1 r}{r_1 s} \tag{2.12}$$

y

$$\hat{LR}(-) = \frac{1-\hat{Se}}{\hat{Sp}} = \frac{s_0 r}{r_0 s}\,. \tag{2.13}$$

La estimación de sus varianzas se obtiene aplicando el método delta (Serfling, 1980), siendo sus expresiones

$$\hat{V}ar\left(\hat{L}R(+)\right)=\frac{\hat{S}e\left(1-\hat{S}e\right)}{s\left(1-\hat{S}p\right)^{2}}+\frac{\hat{S}e^{2}\hat{S}p\left(1-\hat{S}p\right)}{r\left(1-\hat{S}p\right)^{4}} \tag{2.14}$$

y

$$\hat{V}ar\left(\hat{L}R(-)\right)=\frac{\hat{S}e\left(1-\hat{S}e\right)}{s\hat{S}p^{2}}+\frac{\left(1-\hat{S}e\right)^{2}\hat{S}p\left(1-\hat{S}p\right)}{r\hat{S}p^{4}}. \tag{2.15}$$

Condicionando en los totales de filas de la Tabla 2.1, las *LRs* son 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 (2014) han estudiado el intervalo de confianza para el ratio de dos proporciones binomiales independientes que se presenta a continuación.

2.2.1. Intervalo de confianza de Martín-Andrés y Álvarez-Hernández

Adaptados a la notación de la Tabla 2.1, el intervalo de Martín-Andrés y Álvarez-Hernández para la razón de verosimilitud positiva es

$$LR(+) \in \frac{n's_1'r_1' + \frac{z_{1-\alpha/2}^2}{2}\left(s's_1' + r'r_1' - 2s_1'r_1'\right) \pm z_{1-\alpha/2}\sqrt{n'^2 s_1'r_1'\left(s_1' + r_1' - n'\hat{p}_1'\hat{p}_2'\right) + \frac{z_{1-\alpha/2}^2}{4}\left(s's_1' - r'r_1'\right)^2}}{r_1'\left\{n's'\hat{p}_1' - z_{1-\alpha/2}^2\left(s' - r_1'\right)\right\}} \quad (2.16)$$

y para la razón de verosimilitud negativa

$$LR(-) \in \frac{n's_0'r_0' + \frac{z_{1-\alpha/2}^2}{2}\left(s's_0' + r'r_0' - 2s_0'r_0'\right) \pm z_{1-\alpha/2}\sqrt{n'^2 s_0'r_0'\left(s_0' + r_0' - n'\hat{p}_3'\hat{p}_4'\right) + \frac{z_{1-\alpha/2}^2}{4}\left(s's_0' - r'r_0'\right)^2}}{r_0'\left\{n's'\hat{p}_3' - z_{1-\alpha/2}^2\left(s' - r_0'\right)\right\}}, \quad (2.17)$$

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

$$\frac{1}{s'\left(\hat{p}_1'\right)^2 + z_{1-\alpha/2}^2}\left\{s_1'\hat{p}_1' + \frac{z_{1-\alpha/2}^2}{2} - z_{1-\alpha/2}\sqrt{\frac{z_{1-\alpha/2}^2}{4} + s_1'\left(\hat{p}_1' - \hat{p}_2'\right)}\right\} \quad (2.18)$$

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

$$\frac{1}{r'\left(\hat{p}_1'\right)^2}\left\{r_1'\hat{p}_2'+\frac{z_{1-\alpha/2}^2}{2}+z_{1-\alpha/2}\sqrt{\frac{z_{1-\alpha/2}^2}{4}+r_1'\left(\hat{p}_2'-\hat{p}_1'\right)}\right\}. \qquad (2.19)$$

Con respecto al IC para $LR(-)$, si el límite inferior de este intervalo es menor que $s_0'\Big/\left(n'-r_0'\right)$ o mayor que el estimador de $LR(-)$, entonces el límite inferior del intervalo es

$$\frac{1}{s'\left(\hat{p}_3'\right)^2+z_{1-\alpha/2}^2}\left\{s_0'\hat{p}_3'+\frac{z_{1-\alpha/2}^2}{2}-z_{1-\alpha/2}\sqrt{\frac{z_{1-\alpha/2}^2}{4}+s_0'\left(\hat{p}_3'-\hat{p}_4'\right)}\right\} \qquad (2.20)$$

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

$$\frac{1}{r'\left(\hat{p}_3'\right)^2}\left\{r_0'\hat{p}_4'+\frac{z_{1-\alpha/2}^2}{2}+z_{1-\alpha/2}\sqrt{\frac{z_{1-\alpha/2}^2}{4}+r_0'\left(\hat{p}_4'-\hat{p}_3'\right)}\right\}. \qquad (2.21)$$

2.3. Índice de Youden

El estimador máximo verosímil del índice de Youden es

$$\hat{Y}=\hat{S}e+\hat{S}p-1=\frac{s_1}{s}+\frac{r_0}{r}-1 \qquad (2.22)$$

y aplicando el método delta, el estimador de su varianza es

$$\hat{V}ar\left(\hat{Y}\right)=\frac{\hat{S}e\left(1-\hat{S}e\right)}{s}+\frac{\hat{S}p\left(1-\hat{S}p\right)}{r}. \tag{2.23}$$

Un intervalo de confianza tipo Wald para el índice de Youden es

$$Y\in\hat{Y}\pm z_{1-\alpha/2}\sqrt{\frac{\hat{S}e\left(1-\hat{S}e\right)}{s}+\frac{\hat{S}p\left(1-\hat{S}p\right)}{r}}. \tag{2.24}$$

2.4. Valores predictivos

Las razones de verosimilitud y el índice de Youden dependen únicamente de la sensibilidad y especificidad del *TDB*, sin embargo los valores predictivos dependen también de la prevalencia de la enfermedad. Bajo un diseño transversal, condicionando en los totales de las columnas de la Tabla 2.1, 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 $B\left(s_1+\mathrm{r}_1,VPP\right)$ y la frecuencia r_0 es la realización de la distribución binomial $B\left(s_0+r_0,VPN\right)$, por lo que los estimadores puntuales de los valores predictivos son

$$\hat{V}PP = \frac{s_1}{s_1 + r_1} \quad \text{y} \quad \hat{V}PN = \frac{r_0}{s_0 + r_0} \tag{2.25}$$

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

$$\hat{V}ar\left(\hat{V}PP\right) = \frac{\hat{V}PP\left(1-\hat{V}PP\right)}{s_1 + r_1} \quad \text{y} \quad \hat{V}ar\left(\hat{V}PN\right) = \frac{\hat{V}PN\left(1-\hat{V}PN\right)}{s_0 + r_0}.$$

Al igual que para la sensibilidad y especificidad, la estimación de los valores predictivos mediante intervalos de confianza se puede hacer mediante un método exacto y otro aproximado.

2.4.1. Intervalo de confianza exacto de Clopper-Pearson

El intervalo de confianza exacto de Clopper y Pearson (1934) para el valor predictivo positivo a la confianza $100(1-\alpha)\%$ es

$$VPP \in \left(\frac{s_1}{s_1 + (r_1+1) F_{\alpha/2}\left(2(r_1+1), 2s_1\right)} , \frac{(s_1+1) F_{\alpha/2}\left(2(s_1+1), 2r_1\right)}{r_1 + (s_1+1) F_{\alpha/2}\left(2(s_1+1), 2r_1\right)} \right) \tag{2.26}$$

y para el valor predictivo negativo

$$VPN \in \left(\frac{r_0}{r_0 + (s_0+1) F_{\alpha/2}\left(2(s_0+1), 2r_0\right)} , \frac{(r_0+1) F_{\alpha/2}\left(2(r_0+1), 2s_0\right)}{s_0 + (r_0+1) F_{\alpha/2}\left(2(r_0+1), 2s_0\right)} \right). \tag{2.27}$$

2.4.2. Intervalo de confianza de Wilson

El intervalo de confianza de Wilson a la confianza del $100(1-\alpha)\%$ para cada valor predictivo es

$$VPP \in \frac{s_1 + \frac{z_{1-\alpha/2}^2}{2}}{s_1 + r_1 + z_{1-\alpha/2}^2} \pm \frac{z_{1-\alpha/2}}{s_1 + r_1 + z_{1-\alpha/2}^2} \sqrt{\frac{z_{1-\alpha/2}^2}{4} + \frac{s_1 r_1}{s_1 + r_1}} \qquad (2.28)$$

y

$$VPN \in \frac{r_0 + \frac{z_{1-\alpha/2}^2}{2}}{s_0 + r_0 + z_{1-\alpha/2}^2} \pm \frac{z_{1-\alpha/2}}{s_0 + r_0 + z_{1-\alpha/2}^2} \sqrt{\frac{z_{1-\alpha/2}^2}{4} + \frac{s_0 r_0}{s_0 + r_0}}\,. \qquad (2.29)$$

2.4.3. Intervalo de confianza de Agresti y Coull

El intervalo de Agresti y Coull (1998) para cada valor predictivo es

$$VPP \in \frac{s_1 + 2}{s_1 + r_1 + 4} \pm \frac{z_{1-\alpha/2}}{s_1 + r_1 + 4} \sqrt{\frac{(s_1 + 2)(r_1 + 2)}{s_1 + r_1 + 4}} \qquad (2.30)$$

y

$$VPN \in \frac{r_0+2}{s_0+r_0+4} \pm \frac{z_{1-\alpha/2}}{s_0+r_0+4}\sqrt{\frac{(r_0+2)(s_0+2)}{s_0+r_0+4}}. \qquad (2.31)$$

2.4.4. Intervalo de confianza score modificado de Yu et al

El intervalo score modificado para el valor predictivo positivo es

$$VPP \in 0.5+\frac{s_1+r_1+\frac{z^4_{1-\alpha/2}}{53}}{s_1+r_1+z^2_{1-\alpha/2}}\left(\frac{s_1}{s_1+r_1}-0.5\right) \pm \frac{z_{1-\alpha/2}}{s_1+r_1+z^2_{1-\alpha/2}}\sqrt{\frac{z^2_{1-\alpha/2}}{4}+\frac{s_1 r_1}{s_1+r}} \qquad (2.32)$$

y el intervalo de confianza score modificado para el valor predictivo negativo es

$$VPN \in 0.5+\frac{s_0+r_0+\frac{z^4_{1-\alpha/2}}{53}}{s_0+r_0+z^2_{1-\alpha/2}}\left(\frac{r_0}{s_0+r_0}-0.5\right) \pm \frac{z_{1-\alpha/2}}{s_0+r_0+z^2_{1-\alpha/2}}\sqrt{\frac{z^2_{1-\alpha/2}}{4}+\frac{s_0 r_0}{s_0+r_0}}. \qquad (2.33)$$

2.4.5. Intervalo arcoseno de Martín-Andrés y Álvarez-Hernández

El intervalo arcoseno de Martín-Andrés y Álvarez-Hernández (2014) para cada valor predictivo es

$$VPP \in \sin^2\left(\sin^{-1}\sqrt{\frac{s_1+0.5}{s_1+r_1+1}} \pm \frac{z_{1-\alpha/2}}{\sqrt{4(s_1+r_1+1)}}\right) \qquad (2.34)$$

y

$$VPN \in \operatorname{sen}^2\left(\operatorname{sen}^{-1} \sqrt{\frac{r_0 + 0.5}{s_0 + r_0 + 1}} \pm \frac{z_{1-\alpha/2}}{\sqrt{4\left(s_0 + r_0 + 1\right)}} \right). \qquad (2.35)$$

En cuanto a la utilización de los intervalos para los valores predictivos, adaptando las recomendaciones de Martín-Andrés y Álvarez-Hernández (2015) a estos parámetros: (*i*) para $s_1 + r_1 \leq 80$ $\left(s_0 + r_0 \leq 80\right)$ y $\alpha = 1\%$ o 5% utilizar el intervalo de Yu et al (2014); (*ii*) para $s_1 + r_1 \geq 100$ $\left(s_0 + r_0 \geq 100\right)$ y $\alpha = 10\%$ utilizar el intervalo arcoseno de Martín-Andrés y Álvarez-Hernández (2014); (*iii*) en otras situaciones utilizar el intervalo de Agresti y Coull (1998).

2.5. Coeficiente kappa ponderado

El coeficiente kappa ponderado de un *TDB* se define como una medida del acuerdo más allá del azar entre el *TDB* y el *GE*, y es una medida que considera las pérdidas asociadas a una clasificación errónea con el *TDB*. Este parámetro depende de la sensibilidad, especificidad, de la prevalencia de la enfermedad y del índice de ponderación. Roldán-Nofuentes et al

(2009) han estudiado distintos intervalos de confianza para el coeficiente kappa ponderado bajo un diseño transversal. En esta situación, el estimador máximo verosímil del coeficiente kappa ponderado es

$$\hat{\kappa}(c) = \frac{s_1 r_0 - s_0 r_1}{s(s_0 + r_0)c + r(s_1 + r_1)(1-c)}, \tag{2.36}$$

con $0 < c < 1$, obtenida sustituyendo en la expresión de $\kappa(c)$ (ecuación (1.16)) cada parámetro (sensibilidad, especificidad y prevalencia) por su estimador puntual. Aplicando el método delta, la expresión del estimador de la varianza de $\hat{\kappa}(c)$ es

$$\begin{aligned}\hat{V}ar(\hat{\kappa}(c)) = {} & \frac{nr}{s\left[n^2(1-c)r_1 + n\left(cr_0 - 2(1-c)r_1\right)s_0 + n\left\{r_0 - (1-c)r_1\right\}s_1 + s\left(s_0 r_1 - s_1 r_0\right)\right]^4} \times \\ & \left\{\left(s_0 r_1 - s_1 r_0\right)^2\left[2(1-c)r_1 ns - (1-c)r_1 n^2 + s\left(c\left(s_0 r_0 + 2s_0 r_1 + s_1 r_1\right) - r_1 s\right)\right]^2 + \right. \\ & \left. s_1 s_0 n r^3\left[(1-c)r_1 n + s\left(cr - r_1\right)\right]^2 + r_1 r_0 n s r^2\left[s_1 r + c\left(s^2 - s_1 n\right)\right]^2\right\}.\end{aligned} \tag{2.37}$$

Roldán-Nofuentes et al (2009) han estudiado los siguientes intervalos de confianza para el coeficiente kappa ponderado:

2.5.1. Intervalo de confianza tipo Wald

Basado en la normalidad asintótica de $\left(\hat{\kappa}(c)-\kappa(c)\right)\Big/\sqrt{\hat{V}ar\left(\hat{\kappa}(c)\right)}$, el intervalo de confianza de Wald a nivel $100\left(1-\alpha\right)\%$ es

$$\kappa\left(c\right)\in\hat{\kappa}(c)\pm z_{1-\alpha/2}\sqrt{\hat{V}ar\left(\hat{\kappa}(c)\right)}\,. \tag{2.38}$$

Este intervalo también se puede obtener con una corrección por continuidad, es decir,

$$\kappa\left(c\right)\in\hat{\kappa}(c)\pm z_{1-\alpha/2}\sqrt{\hat{V}ar\left(\hat{\kappa}(c)\right)}\pm\frac{1}{2n}\,. \tag{2.39}$$

El intervalo de confianza de Wald tiene un buen rendimiento para muestras relativamente pequeñas $\left(n=100\right)$. El intervalo de Wald con corrección por continuidad no mejora el comportamiento asintótico del intervalo sin dicha corrección.

2.5.2. Intervalo de confianza logit

Basándose en la normalidad asintótica de $\hat{\kappa}\left(c\right)$, su transformación logit, es decir $\ln\left\{\hat{\kappa}(c)\big/\left(1-\hat{\kappa}(c)\right)\right\}$, se distribuye según una distribución normal de

media $\log\left\{\kappa(c)/\left(1-\kappa(c)\right)\right\}$. Entonces el intervalo de confianza para el logit del coeficiente kappa ponderado es

$$\operatorname{logit}\left(\hat{\kappa}(c)\right) \pm z_{1-\alpha/2}\sqrt{\hat{V}ar\left(\operatorname{logit}\left(\hat{\kappa}(c)\right)\right)}\,, \tag{2.40}$$

donde el estimador de la varianza del logit de $\hat{\kappa}(c)$, obtenida aplicando el método delta, es

$$\begin{aligned}\hat{V}ar\left[\operatorname{logit}\left(\hat{\kappa}(c)\right)\right] &= \frac{1}{\left[r_1 s-(1-c)r_1 n-c\left(s_0 r_0+2s_0 r_1+s_1 r_1\right)\right]^2}\times \\ &\left\{\frac{s_1 s_0 r^2\left[(1-c)r_1 n+(cr-r_1)s\right]^2+r_1 r_0 sr\left[s_1 n-s_1 s+c\left(s^2-s_1 n\right)\right]^2}{s\left(s_0 r_1-s_1 r_0\right)^2}+\right. \\ &\left.\frac{\left[2(1-c)r_1 ns-(1-c)r_1 n^2+s\left(c\left(s_0 r_0+2s_0 r_1+s_1 r_1\right)-r_1 s\right)\right]}{nsr}\right\}.\end{aligned} \tag{2.41}$$

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

$$\begin{aligned}&\kappa(c)\in \\ &\left(\frac{\exp\left\{\operatorname{logit}\left(\hat{\kappa}(c)\right)-z_{1-\alpha/2}\sqrt{\hat{V}ar\left(\operatorname{logit}\left(\hat{\kappa}(c)\right)\right)}\right\}}{1+\exp\left\{\operatorname{logit}\left(\hat{\kappa}(c)\right)-z_{1-\alpha/2}\sqrt{\hat{V}ar\left(\operatorname{logit}\left(\hat{\kappa}(c)\right)\right)}\right\}};\frac{\exp\left\{\operatorname{logit}\left(\hat{\kappa}(c)\right)+z_{1-\alpha/2}\sqrt{\hat{V}ar\left(\operatorname{logit}\left(\hat{\kappa}(c)\right)\right)}\right\}}{1+\exp\left\{\operatorname{logit}\left(\hat{\kappa}(c)\right)+z_{1-\alpha/2}\sqrt{\hat{V}ar\left(\operatorname{logit}\left(\hat{\kappa}(c)\right)\right)}\right\}}\right).\end{aligned} \tag{2.42}$$

Este intervalo tiene un buen rendimiento para muestras de tamaño igual o mayor a 200.

2.5.3. Intervalo de confianza bootstrap

El intervalo de confianza bootstrap se calcula generando B 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 B réplicas de tales estimadores, y el intervalo de confianza global se calcula empleando el intervalo de confianza corregido por el sesgo (Efron y Tibshirani, 1993). Este intervalo de confianza tiene un rendimiento muy similar al intervalo de confianza tipo Wald.

2.6. Coeficientes kappa promedios

En el Capítulo 1 se ha definido cada coeficiente kappa promedio como una medida del acuerdo promedio más allá del azar entre el *TDB* y el *GE*. A continuación se deducen sus estimadores puntuales y se estudian varios intervalos de confianza para estos parámetros.

2.6.1. Estimadores puntuales

Los coeficientes kappa promedios se expresan en términos de $\kappa(0)$ y $\kappa(1)$ como

$$\kappa_1 = \begin{cases} \dfrac{2\kappa(0)\kappa(1)}{\kappa(0)-\kappa(1)} \ln\left[\dfrac{\kappa(0)+\kappa(1)}{2\kappa(1)}\right], & p \neq Q \\ Y, & p = Q \end{cases} \tag{2.43}$$

y

$$\kappa_2 = \begin{cases} \dfrac{2\kappa(0)\kappa(1)}{\kappa(0)-\kappa(1)} \ln\left[\dfrac{2\kappa(0)}{\kappa(0)+\kappa(1)}\right], & p \neq Q \\ Y, & p = Q, \end{cases} \tag{2.44}$$

Como $\kappa(0)$ se puede expresar en función del valor predictivo positivo y de la prevalencia de la enfermedad, esto es

$$\kappa(0) = \frac{VPP - p}{1 - p} \tag{2.45}$$

y $\kappa(1)$ en función del valor predictivo negativo y de la prevalencia de la enfermedad,

$$\kappa(1) = \frac{VPN - (1-p)}{p}, \tag{2.46}$$

al sustituir los estimadores máximo verosímil de cada uno de ellos en las expresión de $\kappa(0)$ y $\kappa(1)$, se obtiene que sus estimadores máximo verosímil son

$$\hat{\kappa}(0) = \frac{s_1 r_0 - s_0 r_1}{n_1 r} \tag{2.47}$$

y

$$\hat{\kappa}(1) = \frac{s_1 r_0 - s_0 r_1}{n_0 s}. \tag{2.48}$$

Sustituyendo las ecuaciones (2.47) y (2.48) en la de los coeficientes kappa promedios, ecuaciones (2.43) y (2.44), los estimadores máximo verosímil de los coeficientes kappa promedios son

$$\hat{\kappa}_1 = \begin{cases} \dfrac{2(s_1 r_0 - s_0 r_1)}{n_0 s - n_1 r} \ln\left[\dfrac{n_1 r + n_0 s}{2 n_1 r}\right], & s_0 \neq r_1 \\ \dfrac{s_1 r_0 - s_0 r_1}{sr}, & s_0 = r_1 \end{cases} \tag{2.49}$$

y

$$\hat{\kappa}_2 = \begin{cases} \dfrac{2(s_1 r_0 - s_0 r_1)}{n_0 s - n_1 r} \ln\left[\dfrac{2 n_0 s}{n_1 r + n_0 s}\right], & s_0 \neq r_1 \\ \dfrac{s_1 r_0 - s_0 r_1}{sr}, & s_0 = r_1. \end{cases} \tag{2.50}$$

Se verifica que:

a) Si $s_0 = r_1 = 0$ el coeficiente kappa promedio, $\hat{\kappa}_i$, no se puede estimar.

b) Si $s_1 r_0 = s_0 r_1$ entonces $\hat{k}(c) = 0$ y, por tanto, $\hat{\kappa}_i = 0$.

c) Si $s_1 r_0 < s_0 r_1$ o si las frecuencia $s_1 = 0$ o $r_0 = 0$, entonces $\hat{Y} < 0$ y los resultados del *TDB* se ha de intercambiar, es decir el resultados positivo será negativo y el resultado negativo será positivo.

d) Si $\hat{Y}$ tiene un valor próximo a cero, con muestras pequeñas es posible que se obtengan resultados sesgados para $\hat{Y}$ y por lo tanto también lo serán para $\hat{\kappa}(c)$ y para $\hat{\kappa}_i$.

La expresión del estimador de la varianzas de $\hat{\kappa}_i$ se obtiene aplicando el método delta. Como los coeficientes kappa promedios dependen de $\kappa(0)$ y de $\kappa(1)$, para obtener las expresiones de sus varianzas estimadas es necesario determinar las expresiones de las varianzas de $\kappa(0)$ y de $\kappa(1)$. Los parámetros $\kappa(0)$ y $\kappa(1)$ dependen de la sensibilidad, especificidad y prevalencia de la enfermedad, por lo que sus varianzas asintóticas también

se estiman aplicando el método delta. Sean los vectores $\boldsymbol{\omega} = \left(Se, Sp, p\right)^T$ y $\boldsymbol{\psi} = \left(\kappa(0), \kappa(1)\right)^T$, y sea $\Sigma_{\boldsymbol{\omega}}$ la matriz de varianzas-covarianzas del vector $\boldsymbol{\omega}$, entonces la matriz de varianzas-covarianzas del vector $\boldsymbol{\psi}$ es

$$\Sigma_{\boldsymbol{\psi}} = \left(\frac{\partial \boldsymbol{\psi}}{\partial \boldsymbol{\omega}}\right) \Sigma_{\boldsymbol{\omega}} \left(\frac{\partial \boldsymbol{\psi}}{\partial \boldsymbol{\omega}}\right)^T . \tag{2.51}$$

Como la muestra observada es la realización de una distribución multinomial, la matriz de varianzas-covarianzas asintóticas de $\hat{\boldsymbol{\omega}}$ es

$$\hat{\Sigma}_{\hat{\boldsymbol{\omega}}} = \begin{pmatrix} \hat{Var}\left(\hat{Se}\right) & 0 & 0 \\ 0 & \hat{Var}\left(\hat{Sp}\right) & 0 \\ 0 & 0 & \hat{Var}\left(\hat{p}\right) \end{pmatrix}, \tag{2.52}$$

siendo $\hat{Var}\left(\hat{Se}\right) = \hat{Se}\left(1-\hat{Se}\right)/s$, $\hat{Var}\left(\hat{Sp}\right) = \hat{Sp}\left(1-\hat{Sp}\right)/r$ y $\hat{Var}\left(\hat{p}\right) = \hat{p}\left(1-\hat{p}\right)/n$. Realizando en la ecuación (2.51) las operaciones algebraicas y sustituyendo cada parámetro por su estimador se obtiene que las varianzas-covarianzas asintóticas estimadas de $\hat{\kappa}(0)$ y $\hat{\kappa}(1)$ son

$$\hat{Var}\left(\hat{\kappa}(0)\right) = \frac{\left(1-\hat{Sp}\right)^2 \hat{Y}^2 \hat{Var}\left(\hat{p}\right) + \hat{p}^2 \left[\left(1-\hat{Sp}\right)\hat{Var}\left(\hat{Se}\right) + \hat{Se}^2 \hat{Var}\left(\hat{Sp}\right)\right]}{\hat{Q}^4} , \tag{2.53}$$

$$\hat{V}ar\left(\hat{\kappa}(1)\right)=\frac{\left(1-\hat{S}e\right)^2\hat{Y}^2\hat{V}ar\left(\hat{p}\right)+\left(1-\hat{p}\right)^2\left[\hat{S}p\hat{V}ar\left(\hat{S}e\right)+\left(1-\hat{S}e\right)^2\hat{V}ar\left(\hat{S}p\right)\right]}{\left(1-\hat{Q}\right)^4} \quad (2.54)$$

y

$$\hat{C}ov\left(\hat{\kappa}(0),\hat{\kappa}(1)\right)=$$
$$\frac{(1-\hat{p})\,\hat{p}\left[\left(1-\hat{S}e\right)\hat{S}e\hat{V}ar\left(\hat{S}p\right)+\left(1-\hat{S}p\right)\hat{S}p\hat{V}ar\left(\hat{S}e\right)\right]-\left(1-\hat{S}e\right)\left(1-\hat{S}p\right)\hat{Y}^2\hat{V}ar\left(\hat{p}\right)}{\hat{Q}^2\left(1-\hat{Q}\right)^2}. \quad (2.55)$$

Una vez obtenidas las varianzas asintóticas de $\hat{\kappa}(0)$ y $\hat{\kappa}(1)$, las varianzas asintóticas de los coeficientes kappa promedios se aplicando el mismo método. Así, para $p\neq Q$ la varianza asintótica de κ_i es

$$\mathrm{Var}\left(\kappa_i\right)=$$
$$\left(\frac{\partial\kappa_i}{\partial\kappa(0)}\right)^2\mathrm{Var}\left(\kappa(0)\right)+\left(\frac{\partial\kappa_i}{\partial\kappa(1)}\right)^2\mathrm{Var}\left(\kappa(1)\right)+2\frac{\partial\kappa_1}{\partial\kappa(0)}\frac{\partial\kappa_1}{\partial\kappa(1)}\mathrm{Cov}\left(\kappa(0),\kappa(1)\right) \quad (2.56)$$

y cuando $p=Q$

$$Var\left(\kappa_i\right)=Var\left(Y\right)=\left(\frac{\partial\kappa_i}{\partial Se}\right)^2Var\left(Se\right)+\left(\frac{\partial\kappa_i}{\partial Sp}\right)^2Var\left(Sp\right). \quad (2.57)$$

Realizando las operaciones algebraicas y sustituyendo en las ecuaciones anteriores cada parámetro por su estimador, se obtienen las expresiones de $\hat{V}ar\left(\hat{\kappa}_1\right)$ y de $\hat{V}ar\left(\hat{\kappa}_2\right)$, esto es,

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

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

$$\frac{\left(1-\hat{S}p\right)^2\hat{Y}^2\hat{V}ar\left(\hat{p}\right)+\hat{p}^2\left[\left(1-\hat{S}p\right)\hat{V}ar\left(\hat{S}e\right)+\hat{S}e^2\hat{V}ar\left(\hat{S}p\right)\right]}{\hat{Q}^4} +$$

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

$$\frac{\left(1-\hat{S}e\right)^2\hat{Y}^2\hat{V}ar\left(\hat{p}\right)+\left(1-\hat{p}\right)^2\left[\hat{S}p\hat{V}ar\left(\hat{S}e\right)+\left(1-\hat{S}e\right)^2\hat{V}ar\left(\hat{S}p\right)\right]}{\left(1-\hat{Q}\right)^4} +$$

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

$$\left.\frac{\left(1-\hat{p}\right)\hat{p}\left[\left(1-\hat{S}e\right)\hat{S}e\hat{V}ar\left(\hat{S}p\right)+\left(1-\hat{S}p\right)\hat{S}p\hat{V}ar\left(\hat{S}e\right)\right]-\left(1-\hat{S}e\right)\left(1-\hat{S}p\right)\hat{Y}^2\hat{V}ar\left(\hat{p}\right)}{\hat{Q}^2\left(1-\hat{Q}\right)^2}\right\} \quad (2.58)$$

y

$$\hat{V}ar\left(\hat{\kappa}_2\right)=\frac{1}{\left[\hat{\kappa}(0)+\hat{\kappa}(1)\right]^2\left[\hat{\kappa}(0)-\hat{\kappa}(1)\right]^2}\times$$
$$\left\{\left\{\frac{\hat{\kappa}(1)\left[2\hat{\kappa}(0)\hat{\kappa}(1)-\left(\hat{\kappa}(0)+\hat{\kappa}(1)\right)\hat{\kappa}_2\right]}{\hat{\kappa}(0)}\right\}^2\times\right.$$
$$\frac{\left(1-\hat{S}p\right)^2\hat{Y}^2\hat{V}ar\left(\hat{p}\right)+\hat{p}^2\left[\left(1-\hat{S}p\right)\hat{V}ar\left(\hat{S}e\right)+\hat{S}e^2\hat{V}ar\left(\hat{S}p\right)\right]}{\hat{Q}^4}+$$
$$\left\{\frac{\hat{\kappa}(0)\left[\hat{\kappa}(0)+\hat{\kappa}(1)\right]\hat{\kappa}_2-2\hat{\kappa}(0)\hat{\kappa}(1)^2}{\hat{\kappa}(1)}\right\}^2\times$$
$$\frac{\left(1-\hat{S}e\right)^2\hat{Y}^2\hat{V}ar\left(\hat{p}\right)+\left(1-\hat{p}\right)^2\left[\hat{S}p\hat{V}ar\left(\hat{S}e\right)+\left(1-\hat{S}e\right)^2\hat{V}ar\left(\hat{S}p\right)\right]}{\left(1-\hat{Q}\right)^4}+$$
$$2\left\{\frac{\hat{\kappa}(1)\left[2\hat{\kappa}(0)\hat{\kappa}(1)-\left(\hat{\kappa}(0)+\hat{\kappa}(1)\right)\hat{\kappa}_2\right]}{\hat{\kappa}(0)}\right\}\left\{\frac{\hat{\kappa}(0)\left[\hat{\kappa}(0)+\hat{\kappa}(1)\right]\hat{\kappa}_2-2\hat{\kappa}(0)\hat{\kappa}(1)^2}{\hat{\kappa}(1)}\right\}\times$$
$$\left.\frac{\left(1-\hat{p}\right)\hat{p}\left[\left(1-\hat{S}e\right)\hat{S}e\hat{V}ar\left(\hat{S}p\right)+\left(1-\hat{S}p\right)\hat{S}p\hat{V}ar\left(\hat{S}e\right)\right]-\left(1-\hat{S}e\right)\left(1-\hat{S}p\right)\hat{Y}^2\hat{V}ar\left(\hat{p}\right)}{\hat{Q}^2\left(1-\hat{Q}\right)^2}\right\}. \quad (2.59)$$

Cuando $s_0=r_1$ las varianzas asintóticas estimadas son

$$\hat{V}ar\left(\hat{\kappa}_1\right)=\hat{V}ar\left(\hat{\kappa}_2\right)=\hat{V}ar\left(\hat{Y}\right)=\frac{\hat{S}e\left(1-\hat{S}e\right)}{s}+\frac{\hat{S}p\left(1-\hat{S}p\right)}{r}. \quad (2.60)$$

Para los coeficientes kappa ponderados se han estudiado tres intervalos de confianza aproximados. Estos se presentan a continuación.

2.6.2. Intervalo de confianza tipo Wald

Basado en la normalidad asintótica de $\hat{\kappa}_1$ y $\hat{\kappa}_2$, el intervalo de confianza de Wald para cada uno de los coeficientes kappa promedio es

$$\kappa_i \in \hat{\kappa}_i \pm z_{1-\alpha/2}\sqrt{\hat{V}ar\left(\hat{\kappa}_i\right)}, \quad i=1,2\,. \tag{2.61}$$

2.6.3. Intervalo de confianza logit

Basándose en la normalidad asintótica de $\hat{\kappa}_1$ y $\hat{\kappa}_2$, sus transformaciones logit, $\text{logit}\left(\hat{\kappa}_i\right)=\ln\left\{\hat{\kappa}_1/\left(1-\hat{\kappa}_1\right)\right\}$ y $\text{logit}\left(\hat{\kappa}_2\right)=\ln\left\{\hat{\kappa}_2/\left(1-\hat{\kappa}_2\right)\right\}$, se distribuyen cada una según una distribución normal de media $\text{logit}\left(\kappa_i\right)=\ln\left\{\kappa_i/\left(1-\kappa_i\right)\right\}$. De esta forma, el intervalo de confianza logit a nivel $100\left(1-\alpha\right)\%$ para cada coeficiente kappa promedio es

$$\text{logit}\left(\hat{\kappa}_i\right) \pm z_{1-\alpha/2}\sqrt{\hat{V}ar\left(\text{logit}\left(\hat{\kappa}_i\right)\right)}, \quad i=1,2. \tag{2.62}$$

Las expresiones de las varianzas del logit de $\hat{\kappa}_1$ y $\hat{\kappa}_2$ se obtienen aplicando de nuevo el método delta. Para $p \neq Q$ la varianza asintótica de $\text{logit}\left(\kappa_i\right)$ es

$$Var\left(\text{logit}\left(\kappa_i\right)\right) = \left(\frac{\partial \text{logit}\left(\kappa_i\right)}{\partial \kappa(0)}\right)^2 Var\left(\kappa(0)\right) + \left(\frac{\partial \text{logit}\left(\kappa_i\right)}{\partial \kappa(1)}\right)^2 Var\left(\kappa(1)\right) + 2\frac{\partial \text{logit}\left(\kappa_i\right)}{\partial \kappa(0)}\frac{\partial \text{logit}\left(\kappa_i\right)}{\partial \kappa(1)} Cov\left(\kappa(0), \kappa(1)\right) \tag{2.63}$$

y cuando $p = Q$

$$Var\left(\text{logit}\left(\kappa_i\right)\right) = Var\left(\text{logit}\left(Y\right)\right) = \left(\frac{\partial \text{logit}\left(\kappa_i\right)}{\partial Se}\right)^2 Var\left(Se\right) + \left(\frac{\partial \text{logit}\left(\kappa_i\right)}{\partial Sp}\right)^2 Var\left(Sp\right). \tag{2.64}$$

Realizando las operaciones algebraicas y sustituyendo en las ecuaciones anteriores cada parámetro por su estimador, se obtienen las expresiones de $\hat{Var}\left(\text{logit}\left(\hat{\kappa}_1\right)\right)$ y de $\hat{Var}\left(\text{logit}\left(\hat{\kappa}_2\right)\right)$, esto es,

$$\hat{V}ar\left(\text{logit}\left(\hat{\kappa}_1\right)\right)=\frac{1}{\left(\hat{\kappa}(0)+\hat{\kappa}(1)\right)^2\left[\hat{\kappa}(0)-\hat{\kappa}(1)\right]^2\hat{\kappa}_1^2\left[1-\hat{\kappa}_1^2\right]^2}\times$$
$$\left\{\left\{\frac{\hat{\kappa}(1)\left[2\left(1-\hat{\kappa}_1\right)\hat{\kappa}(0)^2-\left(\hat{\kappa}(0)\left(1-2\hat{\kappa}(0)\right)+\hat{\kappa}(1)\right)\hat{\kappa}_1\right]}{\hat{\kappa}(0)}\right\}^2\times\right.$$
$$\frac{\left(1-\hat{S}p\right)^2\hat{Y}^2\hat{V}ar\left(\hat{p}\right)+\hat{p}^2\left[\left(1-\hat{S}p\right)\hat{V}ar\left(\hat{S}e\right)+\hat{S}e^2\hat{V}ar\left(\hat{S}p\right)\right]}{\hat{Q}^4}+$$
$$\left\{\frac{\hat{\kappa}(0)\left[\left(\hat{\kappa}(0)+\hat{\kappa}(1)\right)\hat{\kappa}_1-2\hat{\kappa}(0)\hat{\kappa}(1)\right]}{\hat{\kappa}(1)}\right\}^2\times$$
$$\frac{\left(1-\hat{S}e\right)^2\hat{Y}^2\hat{V}ar\left(\hat{p}\right)+\left(1-\hat{p}\right)^2\left[\hat{S}p\hat{V}ar\left(\hat{S}e\right)+\left(1-\hat{S}e\right)^2\hat{V}ar\left(\hat{S}p\right)\right]}{\left(1-\hat{Q}\right)^4}+$$
$$2\left\{\frac{\hat{\kappa}(1)\left[2\left(1-\hat{\kappa}_1\right)\hat{\kappa}(0)^2-\left(\hat{\kappa}(0)\left(1-2\hat{\kappa}(0)\right)+\hat{\kappa}(1)\right)\hat{\kappa}_1\right]}{\hat{\kappa}(0)}\right\}\times$$
$$\left\{\frac{\hat{\kappa}(0)\left[\left(\hat{\kappa}(0)+\hat{\kappa}(1)\right)\hat{\kappa}_1-2\hat{\kappa}(0)\hat{\kappa}(1)\right]}{\hat{\kappa}(1)}\right\}\times$$
$$\left.\frac{\left(1-\hat{p}\right)\hat{p}\left[\left(1-\hat{S}e\right)\hat{S}e\hat{V}ar\left(\hat{S}p\right)+\left(1-\hat{S}p\right)\hat{S}p\hat{V}ar\left(\hat{S}e\right)\right]-\left(1-\hat{S}e\right)\left(1-\hat{S}p\right)\hat{Y}^2\hat{V}ar\left(\hat{p}\right)}{\hat{Q}^2\left(1-\hat{Q}\right)^2}\right\} \qquad (2.65)$$

y

$$\hat{V}ar\left(\text{logit}\left(\hat{\kappa}_2\right)\right)=\frac{1}{\left[\hat{\kappa}(0)+\hat{\kappa}(1)\right]^2\left[\hat{\kappa}(0)-\hat{\kappa}(1)\right]^2\hat{\kappa}_2^2\left[1-\hat{\kappa}_2^2\right]^2}\times$$
$$\left\{\left\{\frac{\hat{\kappa}(1)\left[2\hat{\kappa}(0)\hat{\kappa}(1)-\left(\hat{\kappa}(0)+\hat{\kappa}(1)\right)\hat{\kappa}_2\right]}{\hat{\kappa}(0)}\right\}^2\times\right.$$
$$\frac{\left(1-\hat{S}p\right)^2\hat{Y}^2\hat{V}ar\left(\hat{p}\right)+\hat{p}^2\left[\left(1-\hat{S}p\right)\hat{V}ar\left(\hat{S}e\right)+\hat{S}e^2\hat{V}ar\left(\hat{S}p\right)\right]}{\hat{Q}^4}+$$
$$\left\{\frac{\hat{\kappa}(0)\left[\left(\hat{\kappa}(0)+\hat{\kappa}(1)\right)\hat{\kappa}_2-2\hat{\kappa}(1)^2\right]}{\hat{\kappa}(1)}\right\}^2\times$$
$$\frac{\left(1-\hat{S}e\right)^2\hat{Y}^2\hat{V}ar\left(\hat{p}\right)+\left(1-\hat{p}\right)^2\left[\hat{S}p\hat{V}ar\left(\hat{S}e\right)+\left(1-\hat{S}e\right)^2\hat{V}ar\left(\hat{S}p\right)\right]}{\left(1-\hat{Q}\right)^4}+$$
$$2\left\{\frac{\hat{\kappa}(1)\left[2\hat{\kappa}(0)\hat{\kappa}(1)-\left(\hat{\kappa}(0)+\hat{\kappa}(1)\right)\hat{\kappa}_2\right]}{\hat{\kappa}(0)}\right\}\left\{\frac{\hat{\kappa}(0)\left[\left(\hat{\kappa}(0)+\hat{\kappa}(1)\right)\hat{\kappa}_2-2\hat{\kappa}(1)^2\right]}{\hat{\kappa}(1)}\right\}\times$$
$$\left.\frac{\left(1-\hat{p}\right)\hat{p}\left[\left(1-\hat{S}e\right)\hat{S}e\hat{V}ar\left(\hat{S}p\right)+\left(1-\hat{S}p\right)\hat{S}p\hat{V}ar\left(\hat{S}e\right)\right]-\left(1-\hat{S}e\right)\left(1-\hat{S}p\right)\hat{Y}^2\hat{V}ar\left(\hat{p}\right)}{\hat{Q}^2\left(1-\hat{Q}\right)^2}\right\} \qquad (2.66)$$

para $s_0 \neq r_1$, y

$$\hat{V}ar\left(\text{logit}\left(\hat{\kappa}_1\right)\right)=\hat{V}ar\left(\text{logit}\left(\hat{\kappa}_2\right)\right)=\hat{V}ar\left(\text{logit}\left(\hat{Y}\right)\right)=$$
$$\frac{1}{\hat{Y}^2\left(1-\hat{Y}\right)^2}\left\{\frac{\hat{S}e\left(1-\hat{S}e\right)}{s}+\frac{\hat{S}p\left(1-\hat{S}p\right)}{r}\right\} \qquad (2.67)$$

para $s_0 = r_1$. Finalmente el intervalo de confianza logit a la confianza $100(1-\alpha)\%$ para cada coeficiente kappa promedio es

$$\kappa_i \in \left(\frac{\exp\left\{\text{logit}(\hat{\kappa}_i) - z_{1-\alpha/2}\sqrt{\hat{V}ar(\text{logit}(\hat{\kappa}_i))}\right\}}{1+\exp\left\{\text{logit}(\hat{\kappa}_i) - z_{1-\alpha/2}\sqrt{\hat{V}ar(\text{logit}(\hat{\kappa}_i))}\right\}}; \frac{\exp\left\{\text{logit}(\hat{\kappa}_i) + z_{1-\alpha/2}\sqrt{\hat{V}ar(\text{logit}(\hat{\kappa}_i))}\right\}}{1+\exp\left\{\text{logit}(\hat{\kappa}_i) + z_{1-\alpha/2}\sqrt{\hat{V}ar(\text{logit}(\hat{\kappa}_i))}\right\}} \right). \quad (2.68)$$

2.6.4. Intervalo de confianza bootstrap

La estimación de los coeficientes kappa promedios se puede realizar aplicando el método bootstrap, de forma similar a como se ha realizado para el coeficiente kappa ponderado. A partir de la muestra aleatoria observada se generan *B* muestras con reemplazamiento, y a partir de estas *B* muestras se calculan los intervalos de confianza. Por consiguiente, como estimador de cada coeficiente kappa promedio se propone el valor promedio obtenido con las *B* muestras con reemplazamiento, y a continuación se calculan a partir de las *B* muestras el intervalo de confianza corregido por el sesgo (Efron y Tibshirani, 1993) para cada uno de los coeficientes kappa promedio.

Una vez estudiados los intervalos de confianza aproximados para los coeficientes kappa ponderados, se han realizado experimentos de simulación para estudiar su comportamiento asintótico.

2.6.5. Experimentos de simulación

El estudio de la cobertura asintótica de los intervalos de confianza propuestos anteriormente se ha realizado mediante unos experimentos de simulación Monte Carlo. Estos experimentos han consistido en generar 5000 muestras aleatorias de distribuciones multinomiales con tamaños 100, 200, 300, 400, 500, 1000 y 2000, cuyas probabilidades se muestran en la Tabla 1.3 del Capítulo 1. Las probabilidades de las distribuciones multinomales se han calculado a partir de las ecuaciones (2.43) y (2.44) de la siguiente forma. Como valores de la prevalencia de la enfermedad se han tomado los valores $p = \{10\%,\ 20\%,...,\ 90\%\}$ y como valores de κ_1 y κ_2 se han tomado los valores $\{0.10,\ 0.20,...,\ 0.90\}$. Fijados los valores de κ_1 y κ_2 se ha usado el método de Newton-Raphson para la resolución del sistema formado por las ecuaciones (2.43) y (2.44) y así obtener los valores de $\kappa(0)$ y $\kappa(1)$, considerando solamente aquellos valores cuyas soluciones se han encontrado entre 0 y 1. Una vez obtenidos los valores de $\kappa(0)$ y $\kappa(1)$, como el valor de la prevalencia *p* se ha fijado previamente, se han calculado los valores de la sensibilidad (*Se*) y especificidad (*Sp*) resolviendo el sistema formado por las ecuaciones

$$\kappa(0)=\frac{Sp-(1-Q)}{Q} \quad \text{y} \quad \kappa(1)=\frac{Se-Q}{1-Q}, \tag{2.69}$$

siendo $Q=pSe+(1-p)(1-Sp)$, y a continuación se han calculado las probabilidades de las distribuciones multinomiales. Por lo tanto, los experimentos de simulación se han diseñado a partir de valores de los coeficientes kappa promedios κ_1 y κ_2, y no fijando los valores de la sensibilidad y especificidad del *TDB*. Por tanto, a cada pareja de valores de κ_1 y κ_2 le corresponde unos valores fijos de sensibilidad, especificidad y prevalencia, por lo que en los experimentos de simulación se estudia el efecto conjunto de estos parámetros (*Se*, *Sp* y *p*) sobre la cobertura de los intervalos de confianza. Los experimentos se han diseñado de tal forma que para cada una de las 5000 muestras aleatorias generadas, el índice de Youden estimado $\left(\hat{Y}\right)$ sea mayor de cero. Por lo tanto, aquellas muestras cuyo $\hat{Y}$ han sido menor o igual a cero se han descartado y, en su lugar, se ha generado otra muestra. Para $n=100$, el porcentaje medio de muestras descartadas ha sido inferior a 0.001%, y para $n\geq 200$ ninguna muestra ha obtenido un $\hat{Y}\leq 0$. Por ello esta cuestión no ha tenido ningún efecto relevante sobre los resultados obtenidos.

En cada una de las 5000 muestras aleatorias multinomiales se han calculado todos los intervalos de confianza propuestos en la Sección a un nivel de confianza del 95%, y se han calculado la cobertura promedio y la longitud promedio de cada intervalo. Con respecto a los intervalos de confianza obtenidos aplicando el método bootstrap, para cada una de las 5000 muestras aleatorias se han generado a su vez 2000 muestras con reemplazamiento, y a partir de estas se ha calculado el intervalo corregido por el sesgo para cada coeficiente kappa promedio. Una vez calculados los intervalos por bootstrap, se han calculado su cobertura y longitud promedio.

En las Tablas 2.2 a 2.5 se muestran los resultados obtenidos para κ_1 cuando este toma valores iguales a 0.2, 0.4, 0.6 y 0.8, para distintos valores de κ_2, $\kappa(0)$, $\kappa(1)$, prevalencia, sensibilidad y especificidad. De los resultados de estos experimentos se obtienen, en términos generales, las siguientes conclusiones:

e) Intervalo de confianza de Wald. El IC de Wald presenta una cobertura promedio que fluctúa en torno al 95%. Cuando el valor de κ_1 es elevado $(k_1 = 0.8)$ y el tamaño muestral es relativamente

pequeño $(n = 100\text{-}200)$ la cobertura promedio puede desbordar levemente al 95%.

f) Intervalo de confianza logit. El IC logit presenta una cobertura promedio que fluctúa en torno al 95% sobre todo para muestras con $n \geq 400\text{-}500$. Cuando el tamaño muestral es relativamente pequeño $(n = 100\text{-}200)$, el IC logit presenta una cobertura promedio que puede desbordar levemente al 95%. Para $n \geq 500$, el IC logit presenta, en términos generales, mejores fluctuaciones en torno al 95% que el intervalo de Wald.

g) Intervalo de confianza bootstrap (IC BC). El IC bootstrap corregido por el sesgo presenta una cobertura promedio que fluctúa en torno al 95% sobre todo cuando $n \geq 300\text{-}400$; si bien cuando el valor de κ_1 es elevado $(\kappa_1 = 0.8)$, se necesitan muestras de gran tamaño $(n \geq 1000)$ para que la cobertura fluctúe en torno al 95%. En algunas ocasiones, especialmente para muestras de tamaño relativamente pequeño $(n = 100 - 200)$, el IC bootstrap puede tener una cobertura promedio mucho menor que el 95% o desbordar al 95%.

En términos de amplitud promedio, los tres intervalos de confianza tienen una amplitud promedio muy similar, sobre todo para $n \geq 200$. Además, la sensibilidad, especificidad y prevalencia de la enfermedad, y por tanto los coeficientes $\kappa(0)$ y $\kappa(1)$ (que dependen de los anteriores), no tienen un importante efecto en la cobertura de los intervalos estudiados. Por consiguiente, la cobertura de los intervalos depende fuertemente del tamaño de la muestra y no de los valores de los parámetros con lo que la muestra aleatoria ha sido generada.

Tabla 2.2. Coberturas y amplitudes promedio de los intervalos de confianza para κ_1 igual a 0.2.

$\kappa_1 = 0.2$

	$\kappa(0)=0.1588$ $\kappa(1)=0.6725$ $p=10\%$ $Se=0.7773$ $Sp=0.7308$ $\kappa_2=0.3$						$\kappa(0)=0.1588$ $\kappa(1)=0.6725$ $p=30\%$ $Se=0.8837$ $Sp=0.4575$ $\kappa_2=0.4$					
n	**IC Wald**		**IC Logit**		**IC BC**		**IC Wald**		**IC Logit**		**IC BC**	
	Cob.	Amp.	Cob.	Amp.	Cob.	Amp.	Cob.	Amp.	Cob.	Amp.	Cob.	Amp.
100	0.938	0.298	0.971	0.309	0.936	0.288	0.954	0.231	0.969	0.233	0.962	0.228
200	0.940	0.210	0.963	0.212	0.939	0.210	0.951	0.163	0.960	0.164	0.953	0.163
300	0.950	0.172	0.960	0.173	0.951	0.172	0.952	0.132	0.957	0.133	0.950	0.133
400	0.943	0.149	0.954	0.149	0.942	0.149	0.946	0.115	0.952	0.115	0.945	0.115
500	0.945	0.133	0.954	0.133	0.946	0.133	0.956	0.103	0.960	0.103	0.957	0.103
1000	0.943	0.094	0.953	0.094	0.945	0.094	0.944	0.073	0.947	0.073	0.946	0.073
2000	0.954	0.067	0.954	0.067	0.954	0.067	0.951	0.051	0.952	0.051	0.950	0.051
	$\kappa(0)=0.1704$ $\kappa(1)=0.3878$ $p=50\%$ $Se=0.8131$ $Sp=0.4237$ $\kappa_2=0.3$						$\kappa(0)=0.1588$ $\kappa(1)=0.6725$ $p=70\%$ $Se=0.9699$ $Sp=0.2360$ $\kappa_2=0.4$					
n	**IC Wald**		**IC Logit**		**IC BC**		**IC Wald**		**IC Logit**		**IC BC**	
	Cob.	Amp.	Cob.	Amp.	Cob.	Amp.	Cob.	Amp.	Cob.	Amp.	Cob.	Amp.
100	0.956	0.305	0.955	0.323	0.955	0.294	0.922	0.305	0.973	0.322	0.950	0.293
200	0.943	0.216	0.962	0.220	0.955	0.215	0.945	0.216	0.972	0.218	0.943	0.215
300	0.948	0.176	0.964	0.177	0.951	0.176	0.944	0.177	0.955	0.178	0.940	0.177
400	0.944	0.153	0.961	0.153	0.942	0.153	0.944	0.154	0.960	0.154	0.948	0.154
500	0.954	0.137	0.967	0.137	0.952	0.137	0.944	0.137	0.960	0.138	0.943	0.137
1000	0.946	0.097	0.958	0.097	0.948	0.097	0.946	0.097	0.950	0.098	0.945	0.098
2000	0.948	0.068	0.949	0.068	0.948	0.068	0.955	0.069	0.954	0.069	0.952	0.069

Cob.: cobertura promedio. Amp.: amplitud promedio.

Tabla 2.3. Coberturas y amplitudes promedio de los intervalos de confianza para κ_1 igual a 0.4.

$\kappa_1 = 0.4$

	$\kappa(0)=0.1588 \quad \kappa(1)=0.6725$ $p=10\% \quad Se=0.1803 \quad Sp=0.9916 \quad \kappa_2=0.2$						$\kappa(0)=0.2657 \quad \kappa(1)=0.4742$ $p=30\% \quad Se=0.4078 \quad Sp=0.8982 \quad \kappa_2=0.3$					
n	**IC Wald**		**IC Logit**		**IC BC**		**IC Wald**		**IC Logit**		**IC BC**	
	Cob.	Amp.	Cob.	Amp.	Cob.	Amp.	Cob.	Amp.	Cob.	Amp.	Cob.	Amp.
100	0.986	0.762	0.997	0.665	1.000	0.524	0.937	0.448	0.966	0.427	0.886	0.438
200	0.935	0.556	0.977	0.524	0.996	0.472	0.943	0.319	0.964	0.309	0.895	0.319
300	0.929	0.456	0.977	0.437	0.968	0.433	0.950	0.261	0.963	0.255	0.952	0.261
400	0.929	0.398	0.972	0.383	0.940	0.393	0.948	0.226	0.956	0.222	0.945	0.226
500	0.939	0.357	0.970	0.345	0.938	0.358	0.933	0.203	0.942	0.200	0.939	0.203
1000	0.937	0.254	0.951	0.249	0.942	0.256	0.948	0.144	0.951	0.143	0.948	0.144
2000	0.937	0.180	0.944	0.178	0.936	0.181	0.945	0.102	0.946	0.101	0.946	0.102
	$\kappa(0)=0.5677 \quad \kappa(1)=0.3627$ $p=50\% \quad Se=0.8315 \quad Sp=0.6111 \quad \kappa_2=0.5$						$\kappa(0)=0.7756 \quad \kappa(1)=0.3409$ $p=70\% \quad Se=0.9644 \quad Sp=0.4453 \quad \kappa_2=0.6$					
n	**IC Wald**		**IC Logit**		**IC BC**		**IC Wald**		**IC Logit**		**IC BC**	
	Cob.	Amp.	Cob.	Amp.	Cob.	Amp.	Cob.	Amp.	Cob.	Amp.	Cob.	Amp.
100	0.943	0.332	0.964	0.322	0.901	0.332	0.935	0.363	0.970	0.349	0.795	0.363
200	0.940	0.236	0.956	0.232	0.942	0.236	0.952	0.257	0.959	0.252	0.949	0.257
300	0.945	0.193	0.954	0.191	0.947	0.193	0.945	0.210	0.953	0.207	0.943	0.211
400	0.942	0.167	0.946	0.165	0.938	0.167	0.941	0.182	0.951	0.180	0.943	0.182
500	0.949	0.149	0.951	0.148	0.947	0.150	0.949	0.163	0.953	0.162	0.949	0.163
1000	0.957	0.106	0.959	0.105	0.956	0.106	0.953	0.115	0.954	0.115	0.952	0.115
2000	0.949	0.075	0.953	0.075	0.948	0.075	0.955	0.081	0.956	0.081	0.954	0.081

Cob.: cobertura promedio. Amp.: amplitud promedio.

Tabla 2.4. Coberturas y amplitudes promedio de los intervalos de confianza para κ_1 igual a 0.6.

$\kappa_1 = 0.6$

	$\kappa(0)=0.7756 \quad \kappa(1)=0.3409$ $p=10\% \quad Se=0.3716 \quad Sp=0.9896 \quad \kappa_2=0.4$						$\kappa(0)=0.4607 \quad \kappa(1)=0.6640$ $p=30\% \quad Se=0.5843 \quad Sp=0.9230 \quad \kappa_2=0.5$					
n	**IC Wald**		**IC Logit**		**IC BC**		**IC Wald**		**IC Logit**		**IC BC**	
	Cob.	Amp.	Cob.	Amp.	Cob.	Amp.	Cob.	Amp.	Cob.	Amp.	Cob.	Amp.
100	0.954	0.672	0.999	0.589	0.956	0.561	0.933	0.378	0.956	0.362	0.939	0.377
200	0.944	0.462	0.985	0.432	0.927	0.471	0.934	0.270	0.952	0.264	0.846	0.270
300	0.943	0.373	0.964	0.356	0.954	0.386	0.946	0.220	0.959	0.217	0.948	0.221
400	0.943	0.323	0.960	0.312	0.963	0.330	0.950	0.192	0.955	0.190	0.933	0.192
500	0.941	0.289	0.948	0.281	0.952	0.293	0.949	0.172	0.956	0.170	0.948	0.172
1000	0.940	0.204	0.943	0.201	0.940	0.205	0.950	0.121	0.952	0.121	0.948	0.121
2000	0.942	0.144	0.946	0.143	0.943	0.145	0.946	0.086	0.946	0.086	0.944	0.086
	$\kappa(0)=0.5592 \quad \kappa(1)=0.7616$ $p=50\% \quad Se=0.8991 \quad Sp=0.7458 \quad \kappa_2=0.7$						$\kappa(0)=0.9483 \quad \kappa(1)=0.5313$ $p=70\% \quad Se=0.9900 \quad Sp=0.6221 \quad \kappa_2=0.8$					
n	**IC Wald**		**IC Logit**		**IC BC**		**IC Wald**		**IC Logit**		**IC BC**	
	Cob.	Amp.	Cob.	Amp.	Cob.	Amp.	Cob.	Amp.	Cob.	Amp.	Cob.	Amp.
100	0.942	0.315	0.958	0.305	0.952	0.313	0.933	0.356	0.958	0.342	0.925	0.356
200	0.937	0.223	0.950	0.220	0.884	0.223	0.944	0.251	0.957	0.246	0.941	0.252
300	0.943	0.183	0.944	0.181	0.938	0.182	0.949	0.205	0.954	0.203	0.947	0.206
400	0.946	0.158	0.950	0.157	0.944	0.158	0.947	0.178	0.951	0.176	0.947	0.178
500	0.950	0.142	0.953	0.141	0.950	0.142	0.944	0.159	0.951	0.158	0.946	0.159
1000	0.948	0.100	0.948	0.100	0.946	0.100	0.949	0.112	0.954	0.112	0.949	0.112
2000	0.946	0.071	0.946	0.071	0.946	0.071	0.953	0.080	0.953	0.079	0.948	0.080

Cob.: cobertura promedio. Amp.: amplitud promedio.

Tabla 2.5. Coberturas y amplitudes promedio de los intervalos de confianza para κ_1 igual a 0.8.

$\kappa_1 = 0.8$

	$\kappa(0)=0.7572 \quad \kappa(1)=0.9586$ $p=10\% \quad Se=0.9637 \quad Sp=0.9701 \quad \kappa_2=0.9$						$\kappa(0)=0.8599 \quad \kappa(1)=0.6581$ $p=30\% \quad Se=0.7425 \quad Sp=0.9654 \quad \kappa_2=0.7$					
n	**IC Wald**		**IC Logit**		**IC BC**		**IC Wald**		**IC Logit**		**IC BC**	
	Cob.	Amp.	Cob.	Amp.	Cob.	Amp.	Cob.	Amp.	Cob.	Amp.	Cob.	Amp.
100	0.978	0.446	0.952	0.430	0.834	0.476	0.963	0.277	0.980	0.277	0.971	0.275
200	0.951	0.299	0.969	0.299	0.973	0.304	0.933	0.193	0.945	0.193	0.927	0.190
300	0.940	0.241	0.959	0.242	0.969	0.242	0.932	0.156	0.949	0.157	0.947	0.155
400	0.938	0.209	0.947	0.210	0.949	0.210	0.942	0.136	0.955	0.137	0.934	0.136
500	0.942	0.187	0.958	0.188	0.953	0.188	0.936	0.122	0.951	0.122	0.937	0.121
1000	0.944	0.132	0.954	0.133	0.945	0.133	0.943	0.086	0.947	0.087	0.941	0.086
2000	0.952	0.094	0.954	0.094	0.951	0.094	0.953	0.061	0.956	0.061	0.951	0.061
	$\kappa(0)=0.9483 \quad \kappa(1)=0.5313$ $p=50\% \quad Se=0.6996 \quad Sp=0.9814 \quad \kappa_2=0.6$						$\kappa(0)=0.9483 \quad \kappa(1)=0.5313$ $p=70\% \quad Se=0.7969 \quad Sp=0.9707 \quad \kappa_2=0.6$					
n	**IC Wald**		**IC Logit**		**IC BC**		**IC Wald**		**IC Logit**		**IC BC**	
	Cob.	Amp.	Cob.	Amp.	Cob.	Amp.	Cob.	Amp.	Cob.	Amp.	Cob.	Amp.
100	0.977	0.232	0.967	0.231	0.857	0.235	0.982	0.242	0.965	0.241	0.846	0.247
200	0.971	0.154	0.965	0.154	0.957	0.153	0.962	0.159	0.960	0.159	0.940	0.159
300	0.954	0.122	0.956	0.122	0.960	0.121	0.960	0.127	0.962	0.127	0.961	0.126
400	0.952	0.106	0.954	0.106	0.965	0.105	0.944	0.109	0.945	0.109	0.958	0.108
500	0.949	0.094	0.958	0.094	0.965	0.094	0.943	0.098	0.946	0.098	0.956	0.097
1000	0.946	0.067	0.946	0.067	0.945	0.067	0.942	0.069	0.943	0.069	0.945	0.069
2000	0.956	0.047	0.956	0.047	0.954	0.047	0.949	0.049	0.949	0.049	0.948	0.049

Cob.: cobertura promedio. Amp.: amplitud promedio.

En las Tablas 2.6 a 2.9 se muestran los resultados obtenidos para κ_2 cuando este toma también valores iguales a 0.2, 0.4, 0.6 y 0.8, para distintos valores del resto de parámetros. A partir de estos resultados, en términos generales, se obtienen unas conclusiones muy similares a las obtenidas para los intervalos de confianza de κ_1, aunque el IC bootstrap (IC BC) puede desbordar el 95% de cobertura o tener una cobertura promedio mucho menor que el 95% especialmente cuanto el tamaño de la muestra es relativamente pequeño $(n = 100 - 200)$.

Una vez analizados los resultados de los experimentos de simulación, se puede establecer, en términos generales, la siguiente regla de aplicación:

a) Para muestras de tamaño 100 a 500, utilizar el intervalo de confianza de Wald.

b) Para muestras de tamaño superior a 500 es posible usar cualquier IC, aunque el IC BC requiere un gran esfuerzo computacional.

Tabla 2.6. Coberturas y amplitudes promedio de los intervalos de confianza para κ_2 igual a 0.2.

$\kappa_2 = 0.2$

	$\kappa(0)=0.3878 \quad \kappa(1)=0.1704$ $p=10\% \quad Se=0.2091 \quad Sp=0.9715 \quad \kappa_1=0.3$						$\kappa(0)=0.6725 \quad \kappa(1)=0.1588$ $p=30\% \quad Se=0.2360 \quad Sp=0.9699 \quad \kappa_1=0.4$					
n	**IC Wald**		**IC Logit**		**IC BC**		**IC Wald**		**IC Logit**		**IC BC**	
	Cob.	Amp.	Cob.	Amp.	Cob.	Amp.	Cob.	Amp.	Cob.	Amp.	Cob.	Amp.
100	0.966	0.540	0.959	0.566	0.996	0.487	0.932	0.307	0.973	0.318	0.955	0.296
200	0.916	0.379	0.959	0.409	0.984	0.361	0.934	0.216	0.965	0.218	0.935	0.215
300	0.917	0.308	0.966	0.325	0.967	0.299	0.944	0.177	0.958	0.178	0.946	0.178
400	0.934	0.269	0.971	0.277	0.953	0.264	0.947	0.153	0.963	0.154	0.947	0.154
500	0.940	0.241	0.970	0.244	0.947	0.239	0.944	0.137	0.961	0.137	0.946	0.137
1000	0.938	0.172	0.955	0.173	0.935	0.172	0.953	0.097	0.957	0.097	0.952	0.097
2000	0.942	0.122	0.951	0.122	0.941	0.122	0.945	0.069	0.948	0.069	0.943	0.069
	$\kappa(0)=0.3878 \quad \kappa(1)=0.1704$ $p=50\% \quad Se=0.4237 \quad Sp=0.8131 \quad \kappa_1=0.3$						$\kappa(0)=0.6725 \quad \kappa(1)=0.1588$ $p=70\% \quad Se=0.4575 \quad Sp=0.8837 \quad \kappa_1=0.4$					
n	**IC Wald**		**IC Logit**		**IC BC**		**IC Wald**		**IC Logit**		**IC BC**	
	Cob.	Amp.	Cob.	Amp.	Cob.	Amp.	Cob.	Amp.	Cob.	Amp.	Cob.	Amp.
100	0.951	0.305	0.953	0.324	0.958	0.295	0.945	0.231	0.968	0.234	0.956	0.227
200	0.950	0.216	0.962	0.219	0.957	0.215	0.950	0.163	0.959	0.163	0.951	0.163
300	0.947	0.176	0.962	0.178	0.946	0.176	0.952	0.132	0.958	0.133	0.950	0.132
400	0.946	0.153	0.965	0.153	0.946	0.153	0.946	0.115	0.947	0.115	0.945	0.115
500	0.955	0.137	0.964	0.137	0.956	0.137	0.944	0.103	0.952	0.103	0.946	0.103
1000	0.950	0.097	0.959	0.097	0.953	0.097	0.941	0.073	0.944	0.073	0.943	0.073
2000	0.947	0.068	0.950	0.068	0.946	0.068	0.943	0.051	0.946	0.051	0.942	0.051

Cob.: cobertura promedio. Amp.: amplitud promedio.

Tabla 2.7. Coberturas y amplitudes promedio de los intervalos de confianza para κ_2 igual a 0.4.

$\kappa_2 = 0.4$

	$\kappa(0)=0.1588 \quad \kappa(1)=0.6725$ $p=10\% \quad Se=0.7773 \quad Sp=0.7308 \quad \kappa_1=0.2$						$\kappa(0)=0.2657 \quad \kappa(1)=0.4742$ $p=30\% \quad Se=0.7021 \quad Sp=0.6817 \quad \kappa_1=0.3$					
n	**IC Wald**		**IC Logit**		**IC BC**		**IC Wald**		**IC Logit**		**IC BC**	
	Cob.	Amp.	Cob.	Amp.	Cob.	Amp.	Cob.	Amp.	Cob.	Amp.	Cob.	Amp.
100	0.947	0.464	0.972	0.444	0.966	0.442	0.943	0.405	0.970	0.389	0.757	0.400
200	0.938	0.327	0.960	0.318	0.957	0.326	0.936	0.288	0.953	0.281	0.930	0.288
300	0.944	0.269	0.957	0.263	0.952	0.269	0.943	0.236	0.952	0.232	0.941	0.236
400	0.940	0.233	0.948	0.229	0.947	0.233	0.939	0.205	0.947	0.202	0.944	0.205
500	0.946	0.209	0.951	0.207	0.947	0.209	0.950	0.183	0.954	0.181	0.953	0.183
1000	0.944	0.148	0.946	0.147	0.945	0.148	0.949	0.130	0.954	0.129	0.952	0.130
2000	0.951	0.105	0.953	0.104	0.952	0.105	0.952	0.092	0.952	0.092	0.950	0.092
	$\kappa(0)=0.5677 \quad \kappa(1)=0.3627$ $p=50\% \quad Se=0.6111 \quad Sp=0.8315 \quad \kappa_1=0.5$						$\kappa(0)=0.7756 \quad \kappa(1)=0.3409$ $p=70\% \quad Se=0.6746 \quad Sp=0.8864 \quad \kappa_1=0.6$					
n	**IC Wald**		**IC Logit**		**IC BC**		**IC Wald**		**IC Logit**		**IC BC**	
	Cob.	Amp.	Cob.	Amp.	Cob.	Amp.	Cob.	Amp.	Cob.	Amp.	Cob.	Amp.
100	0.947	0.332	0.970	0.321	0.903	0.332	0.948	0.303	0.963	0.295	0.926	0.303
200	0.955	0.236	0.962	0.232	0.952	0.236	0.943	0.214	0.950	0.211	0.941	0.215
300	0.949	0.193	0.957	0.190	0.948	0.193	0.948	0.175	0.956	0.174	0.949	0.176
400	0.936	0.167	0.947	0.166	0.939	0.167	0.946	0.152	0.952	0.151	0.945	0.152
500	0.940	0.150	0.942	0.148	0.939	0.149	0.947	0.136	0.950	0.135	0.946	0.136
1000	0.949	0.106	0.950	0.105	0.947	0.106	0.953	0.096	0.954	0.096	0.951	0.096
2000	0.949	0.075	0.951	0.075	0.948	0.075	0.953	0.068	0.955	0.068	0.953	0.068

Cob.: cobertura promedio. Amp.: amplitud promedio.

Tabla 2.8. Coberturas y amplitudes promedio de los intervalos de confianza para κ_2 igual a 0.6.

$\kappa_2 = 0.6$

	$\kappa(0)=0.3409 \quad \kappa(1)=0.7756$ $p=10\% \quad Se=0.8209 \quad Sp=0.8670 \quad \kappa_1=0.4$						$\kappa(0)=0.4607 \quad \kappa(1)=0.6640$ $p=30\% \quad Se=0.7923 \quad Sp=0.7940 \quad \kappa_1=0.5$					
n	**IC Wald**		**IC Logit**		**IC BC**		**IC Wald**		**IC Logit**		**IC BC**	
	Cob.	Amp.	Cob.	Amp.	Cob.	Amp.	Cob.	Amp.	Cob.	Amp.	Cob.	Amp.
100	0.969	0.475	0.978	0.453	0.915	0.483	0.944	0.351	0.965	0.338	0.938	0.349
200	0.947	0.326	0.963	0.315	0.957	0.331	0.942	0.251	0.951	0.246	0.936	0.251
300	0.939	0.266	0.944	0.260	0.951	0.268	0.940	0.205	0.947	0.202	0.941	0.205
400	0.947	0.232	0.952	0.228	0.949	0.233	0.946	0.178	0.955	0.176	0.950	0.178
500	0.941	0.207	0.951	0.204	0.953	0.208	0.945	0.159	0.949	0.158	0.947	0.159
1000	0.946	0.147	0.948	0.146	0.946	0.147	0.950	0.113	0.955	0.112	0.952	0.113
2000	0.944	0.104	0.943	0.103	0.941	0.104	0.950	0.080	0.950	0.080	0.951	0.080
	$\kappa(0)=0.7616 \quad \kappa(1)=0.5592$ $p=50\% \quad Se=0.7458 \quad Sp=0.8991 \quad \kappa_1=0.7$						$\kappa(0)=0.9483 \quad \kappa(1)=0.5315$ $p=70\% \quad Se=0.7969 \quad Sp=0.9707 \quad \kappa_1=0.8$					
n	**IC Wald**		**IC Logit**		**IC BC**		**IC Wald**		**IC Logit**		**IC BC**	
	Cob.	Amp.	Cob.	Amp.	Cob.	Amp.	Cob.	Amp.	Cob.	Amp.	Cob.	Amp.
100	0.937	0.315	0.956	0.305	0.945	0.313	0.952	0.313	0.962	0.304	0.950	0.312
200	0.940	0.223	0.947	0.220	0.882	0.223	0.947	0.221	0.952	0.218	0.950	0.220
300	0.937	0.183	0.944	0.181	0.939	0.182	0.949	0.180	0.958	0.178	0.948	0.180
400	0.952	0.158	0.955	0.157	0.954	0.158	0.938	0.156	0.944	0.155	0.940	0.156
500	0.948	0.142	0.953	0.141	0.951	0.141	0.947	0.140	0.952	0.139	0.945	0.140
1000	0.943	0.100	0.946	0.100	0.942	0.100	0.943	0.099	0.944	0.099	0.941	0.099
2000	0.950	0.071	0.950	0.071	0.950	0.071	0.949	0.070	0.950	0.070	0.947	0.070

Cob.: cobertura promedio. Amp.: amplitud promedio.

Tabla 2.9. Coberturas y amplitudes promedio de los intervalos de confianza para κ_2 igual a 0.8.

$\kappa_2 = 0.8$

	$\kappa(0)=0.9586 \quad \kappa(1)=0.7768$ $p=10\% \quad Se=0.7768 \quad Sp=0.9967 \quad \kappa_1=0.9$						$\kappa(0)=0.6581 \quad \kappa(1)=0.8599$ $p=30\% \quad Se=0.9102 \quad Sp=0.8773 \quad \kappa_1=0.7$					
n	**IC Wald**		**IC Logit**		**IC BC**		**IC Wald**		**IC Logit**		**IC BC**	
	Cob.	Amp.	Cob.	Amp.	Cob.	Amp.	Cob.	Amp.	Cob.	Amp.	Cob.	Amp.
100	0.973	0.508	0.943	0.478	0.658	0.529	0.961	0.260	0.979	0.260	0.979	0.256
200	0.958	0.332	0.959	0.331	0.953	0.343	0.937	0.182	0.953	0.182	0.948	0.179
300	0.935	0.266	0.958	0.267	0.967	0.269	0.950	0.149	0.950	0.150	0.954	0.148
400	0.941	0.231	0.954	0.232	0.961	0.232	0.939	0.128	0.950	0.128	0.940	0.128
500	0.934	0.206	0.953	0.206	0.949	0.207	0.935	0.115	0.941	0.115	0.936	0.115
1000	0.942	0.146	0.955	0.146	0.947	0.146	0.943	0.082	0.947	0.082	0.942	0.082
2000	0.948	0.103	0.948	0.103	0.949	0.103	0.947	0.058	0.954	0.058	0.948	0.058
	$\kappa(0)=0.5313 \quad \kappa(1)=0.9483$ $p=50\% \quad Se=0.9814 \quad Sp=0.6996 \quad \kappa_1=0.6$						$\kappa(0)=0.9586 \quad \kappa(1)=0.7572$ $p=70\% \quad Se=0.9146 \quad Sp=0.9732 \quad \kappa_1=0.9$					
n	**IC Wald**		**IC Logit**		**IC BC**		**IC Wald**		**IC Logit**		**IC BC**	
	Cob.	Amp.	Cob.	Amp.	Cob.	Amp.	Cob.	Amp.	Cob.	Amp.	Cob.	Amp.
100	0.977	0.233	0.967	0.232	0.844	0.236	0.952	0.276	0.964	0.278	0.971	0.273
200	0.962	0.154	0.959	0.154	0.947	0.153	0.943	0.194	0.960	0.195	0.953	0.193
300	0.960	0.123	0.958	0.123	0.956	0.122	0.945	0.159	0.954	0.159	0.950	0.159
400	0.947	0.106	0.954	0.106	0.962	0.105	0.943	0.138	0.953	0.138	0.949	0.137
500	0.949	0.095	0.954	0.095	0.961	0.094	0.938	0.123	0.947	0.123	0.941	0.123
1000	0.950	0.067	0.950	0.067	0.953	0.067	0.951	0.087	0.950	0.087	0.948	0.087
2000	0.948	0.047	0.949	0.047	0.950	0.047	0.946	0.062	0.949	0.062	0.947	0.062

Cob.: cobertura promedio. Amp.: amplitud promedio.

2.6.6. El programa "akcbdt"

Se ha escrito un programa en *R* para estimar los coeficientes kappa promedios de un *TDB* cuando este y el *GE* se aplican a todos los individuos de una muestra aleatoria. Este programa, denominado "*akcbdt*" (average kappa coefficient of a binary diagnostic test) se ejecuta con el comando

$$\text{akcbdt}\left(s_1, s_0, r_1, r_0\right)$$

cuando el nivel de confianza de los intervalos es el 95% y se generan 2000 muestras con reemplazamiento. Para un nivel de confianza *conflevel* y *B* muestras con reemplazamiento el comando es

$$\text{akcbdt}\left(s_1, s_0, r_1, r_0, conflevel, B\right).$$

El programa proporciona las estimaciones de los coeficientes kappa promedios y los tres intervalos de confianza estudiados, para $0 < c < 0.5$ y para $0.5 < c < 1$. En el caso de que las frecuencias observadas s_0 y r_1 sean iguales, el programa estima un único valor del coeficiente kappa promedio (ya que en esta situación se verifica que $\kappa_1 = \kappa_2$). El programa también proporciona las estimaciones de la sensibilidad y especificidad (y el intervalo score modificado para estos parámetros), prevalencia de la enfermedad y de la sensibilidad y especificidad corregidas por azar, y sus

correspondientes errores estándares. Con respecto al intervalo de confianza bootstrap, el programa genera *B* muestras con reemplazamiento de tal forma que en todas ellas se pueda estimar el cada coeficiente kappa promedio. Los resultados obtenidos al ejecutar el programa se graban en un fichero denominado "Results.txt" en la misma carpeta desde donde se ejecuta el programa. El código del programa se puede ver en el Apéndice I.

2.7. Ejemplo

Los resultados de las Secciones anteriores se han aplicado al estudio de Weiner et al (1979) sobre el diagnóstico de la enfermedad coronaria y al estudio de Yee et al (2001) sobre el diagnóstico de la neoplasia colorectal. Los resultados relativos a los coeficientes kappa promedios se han obtenido utilizando el programa "*akcbdt"*.

2.7.1. Estudio de Weiner et al

Weiner et al (1979) han estudiado el diagnóstico de la enfermedad de la arteria coronaria utilizando como *TDB* un test de ejercicio (variable *T*) y como *GE* la arteriografía coronaria (variable *D*). En la Tabla 2.10 se

muestran los resultados obtenidos por estos autores al aplicar ambas pruebas a una muestra de 1465 hombres. En la Tabla 2.11 se muestran las estimaciones de los parámetros. Para la sensibilidad, especificidad y los valores predictivos se han calculado los intervalos de Yu et al (2014); para el resto de parámetros se han calculado los intervalos presentados en sus correspondientes Secciones. Todos los intervalos de confianza se han calculado al 95% de confianza.

Tabla 2.10. Datos del estudio de Weiner et al.

	$T=1$	$T=0$	Total
$D=1$	815	208	1023
$D=0$	115	327	442
Total	930	535	1465

Tabla 2.11. Estimaciones de los parámetros en el estudio de Weiner et al.

	Estimación	Error estándar	IC al 95%
Se	0.797	0.013	$(0.771\ ,\ 0.820)$
Sp	0.740	0.021	$(0.697\ ,\ 0.778)$
$LR(+)$	3.062	0.250	$(2.618\ ,\ 3.606)$
$LR(-)$	0.275	0.019	$(0.240\ ,\ 0.314)$
VPP	0.876	0.011	$(0.854\ ,\ 0.896)$
VPN	0.611	0.021	$(0.569\ ,\ 0.652)$
p	0.698	0.012	—
$\kappa(0)$	0.590	0.030	—
$\kappa(1)$	0.443	0.025	—

Coeficiente kappa ponderado

c	$\hat{\kappa}(c)$	IC Wald	IC Logit
0.1	0.571	$(0.518\ ,\ 0.625)$	$(0.517\ ,\ 0.624)$
0.2	0.553	$(0.503\ ,\ 0.604)$	$(0.503\ ,\ 0.603)$
0.3	0.537	$(0.489\ ,\ 0.585)$	$(0.489\ ,\ 0.584)$
0.4	0.521	$(0.474\ ,\ 0.568)$	$(0.474\ ,\ 0.568)$
0.5	0.506	$(0.460\ ,\ 0.553)$	$(0.460\ ,\ 0.552)$
0.6	0.492	$(0.446\ ,\ 0.539)$	$(0.446\ ,\ 0.539)$
0.7	0.479	$(0.432\ ,\ 0.526)$	$(0.432\ ,\ 0.526)$
0.8	0.466	$(0.419\ ,\ 0.514)$	$(0.419\ ,\ 0.514)$
0.9	0.455	$(0.406\ ,\ 0.503)$	$(0.407\ ,\ 0.503)$

Coeficientes kappa promedios

	κ_1 $(0 < c < 0.5)$	κ_2 $(0.5 < c < 1)$
Estimación	0.546	0.473
Error estándar	0.025	0.024
IC Wald	$(0.497\ ,\ 0.595)$	$(0.426\ ,\ 0.520)$
IC Logit	$(0.497\ ,\ 0.595)$	$(0.427\ ,\ 0.520)$
IC *BC*	$(0.500\ ,\ 0.597)$	$(0.426\ ,\ 0.523)$

A partir de estos resultados se obtienen las siguientes conclusiones:

a) Sensibilidad y especificidad. La sensibilidad del test de ejercicio toma un valor moderadamente alto (un valor entre el 77.1% y el 82.0% a la confianza del 95%) y la especificidad también (un valor entre el 69.7% y el 77.8% a la confianza del 95%), por lo que este test se puede utilizar tanto para descartar la enfermedad como para confirmarla.

b) Razones de verosimilitud. La probabilidad de que el test de esfuerzo sea positivo en los individuos con la enfermedad de la arteria coronaria es, con una confianza del 95%, un valor entre 2.618 y 3.606 veces mayor que la probabilidad de que el test de esfuerzo sea positivo en los individuos sin dicha enfermedad. La probabilidad de que el test de esfuerzo sea negativo en los individuos sin la enfermedad de la arteria coronaria es, con una confianza del 95%, un valor entre 3.185 $\left(=1/0.314\right)$ y 4.167 $\left(=1/0.240\right)$ veces mayor que la probabilidad de que esta prueba sea negativa en los individuos con esta enfermedad coronaria.

c) Valores predictivos. Cuando la prueba de esfuerzo se aplica a los individuos de la población objeto de estudio, el valor predictivo

positivo toma un valor alto (entre el 85.4% y el 89.6% a la confianza del 95%) mientras que el valor predictivo negativo toma un valor moderado (entre el 56.9% y el 65.2% a la confianza del 95%). Por tanto, el test de esfuerzo es una prueba muy útil para confirmar la enfermedad en los individuos de esta población y es moderadamente útil para descartar la enfermedad en estos mismos individuos.

d) Coeficiente kappa ponderado. Independientemente del valor del índice de ponderación c, el acuerdo entre el test de ejercicio y la angiografía coronaria es siempre moderado (tanto en términos de las estimaciones puntuales como de los intervalos de confianza). Por tanto, cuando la prueba de esfuerzo se aplica a los individuos de la población objeto de estudio, esta prueba es moderadamente buena para el diagnóstico de la neoplasia colorectal.

e) Coeficientes kappa promedios. En términos de los valores de los estimadores puntuales, si el clínico considera que $L' > L$ $\left(0 < c < 0.5\right)$, el acuerdo promedio más allá del azar entre el test de ejercicio y la angiografía es moderada $\left(\hat{\kappa}_1 = 0.546\right)$, y en términos de los intervalos de confianza, el acuerdo promedio más allá del azar es también moderado (95% confianza). Sustituyendo en la ecuación

$$\kappa(c) = \frac{\kappa(0)\kappa(1)}{c\kappa(0) + (1-c)\kappa(1)}, \tag{2.70}$$

cada parámetro por su valor estimado se obtiene que $0.546 = \frac{0.590 \times 0.443}{0.590 + 0.443(1-c)}$, y despejando c se obtiene que la pérdida relativa estimada entre los falsos positivos y los falsos negativos es $\hat{c} = 0.243$. Como $c = L/(L+L') = (L/L')/\{1+(L/L')\}$, una vez estimado el índice *c*, es posible calcular cuánto de grande es una pérdida con respecto a la otra es decir, el cociente L/L' o L'/L. De esta forma, la pérdida asociada a los falsos positivos (L') es 3.12 veces mayor que la pérdida asociada a los falsos negativos (L). Por tanto, si el clínico considera que $L' > L$ (como en la situación de un test definitivo previo a un tratamiento que suponga algún riesgo para el individuo, por ejemplo una operación quirúrgica), el acuerdo promedio más allá del azar entre el test de ejercicio y la angiografía coronaria es moderada $(\hat{\kappa}_1 = 0.546)$, y la pérdida cometida al clasificar erróneamente con el test de ejercicio a un individuo no enfermo es 3.12 veces mayor que la pérdida cometida al clasificar erróneamente con el test de ejercicio a un individuo enfermo.

Si el clínico considera que $L > L'$ $\left(0.5 < c < 1\right)$, el acuerdo promedio más allá del azar entre el test de ejercicio y la angiografía es moderado $\left(\hat{\kappa}_2 = 0.473\right)$, y la pérdida relativa estimada entre los falsos positivos y los falsos negativos es 0.745, por lo que la pérdida asociada a los falsos negativos $\left(L\right)$ es 2.92 veces mayor que la pérdida asociada a los falsos positivos $\left(L'\right)$. Por tanto, si $L > L'$ la pérdida cometida al clasificar erróneamente con el test de ejercicio a un individuo enfermo es 2.92 $\left(= 0.745/\left(1 - 0.745\right)\right)$ veces mayor que la pérdida cometida al clasificar erróneamente con el test de ejercicio a un individuo no enfermo.

Analizados los resultados en términos de los coeficientes kappa promedios, el test de ejercicio puede ser utilizado si $L' > L$ o si $L > L'$, siendo en ambos casos el acuerdo promedio más allá del azar entre el test de ejercicio y la angiografía moderado.

2.7.2. Estudio de Yee et al

Yee et al (2001) han evaluado el rendimiento de la colonografía por tomografía computarizada en el diagnóstico de la neoplasia colorectal utilizando como *GE* la colonoscopia. En la Tabla 2.12 se muestran los resultados obtenidos por estos autores al aplicar la colonografía por tomografía computarizada (variable *T*) y la colonoscopia (variable *D*) a una muestra de 300 pacientes, y en la Tabla 2.13 se muestran las estimaciones de los parámetros.

Tabla 2.12. Datos del estudio de Yee et al.

	$T=1$	$T=0$	Total
$D=1$	164	18	182
$D=0$	33	85	118
Total	197	103	300

Tabla 2.13. Estimación de los parámetros en el estudio de Yee et al.

	Estimación	Error estándar	IC al 95%
Se	0.901	0.022	$(0.848\ ,\ 0.937)$
Sp	0.720	0.041	$(0.633\ ,\ 0.793)$
$LR(+)$	3.222	0.483	$(2.433\ ,\ 4.364)$
$LR(-)$	0.137	0.032	$(0.085\ ,\ 0.212)$
VPP	0.832	0.027	$(0.775\ ,\ 0.879)$
VPN	0.825	0.037	$(0.739\ ,\ 0.887)$
p	0.607	0.028	—
$\kappa(0)$	0.574	0.054	—
$\kappa(1)$	0.712	0.057	—
Coeficiente kappa ponderado			
c	$\hat{\kappa}(c)$	IC Wald	IC Logit
0.1	0.585	$(0.483\ ,\ 0.688)$	$(0.481\ ,\ 0.683)$
0.2	0.597	$(0.499\ ,\ 0.696)$	$(0.496\ ,\ 0.691)$
0.3	0.610	$(0.515\ ,\ 0.704)$	$(0.512\ ,\ 0.699)$
0.4	0.622	$(0.530\ ,\ 0.714)$	$(0.527\ ,\ 0.709)$
0.5	0.636	$(0.546\ ,\ 0.726)$	$(0.542\ ,\ 0.720)$
0.6	0.650	$(0.560\ ,\ 0.739)$	$(0.556\ ,\ 0.733)$
0.7	0.664	$(0.573\ ,\ 0.755)$	$(0.568\ ,\ 0.749)$
0.8	0.679	$(0.584\ ,\ 0.774)$	$(0.578\ ,\ 0.766)$
0.9	0.695	$(0.593\ ,\ 0.797)$	$(0.585\ ,\ 0.787)$
Coeficientes kappa promedios			
	κ_1 $(0 < c < 0.5)$		κ_2 $(0.5 < c < 1)$
Estimación	0.604		0.672
Error estándar	0.049		0.048
IC Wald	$(0.508\ ,\ 0.700)$		$(0.579\ ,\ 0.766)$
IC Logit	$(0.505\ ,\ 0.695)$		$(0.573\ ,\ 0.758)$
IC *BC*	$(0.513\ ,\ 0.700)$		$(0.587\ ,\ 0.768)$

A partir de los resultados de la Tabla 2.13 se obtienen las siguientes conclusiones:

a) Sensibilidad y especificidad. La sensibilidad del test de ejercicio toma un valor muy alto (un valor entre el 84.8% y el 93.7% a la confianza del 95%) y la especificidad toma un valor más moderado (un valor entre el 63.3% y el 79.3% a la confianza del 95%), por lo que la colonoscopia por tomografía computarizada es muy útil para descartar la enfermedad pero moderadamente útil para confirmarla.

b) Razones de verosimilitud. La probabilidad de que la colonoscopia por tomografía computarizada sea positiva en los individuos con la neoplasia colorectal es, con una confianza del 95%, un valor entre 2.433 y 4.364 veces mayor que la probabilidad de que la colonoscopia por tomografía computarizada sea positiva en los individuos sin la neoplasia colorectal. La probabilidad de que la colonoscopia por tomografía computarizada sea negativa en los individuos sin la neoplasia colorectal es, con una confianza del 95%, un valor entre 4.717 $\left(=1/0.212\right)$ y 11.765 $\left(=1/0.085\right)$ veces mayor que la probabilidad de que dicha prueba sea negativa en los individuos con la neoplasia colorectal.

c) Valores predictivos. Cuando la colonoscopia por tomografía computarizada se aplica a los individuos de la población objeto de estudio, el valor predictivo positivo toma un valor alto (entre el 77.5% y el 87.9% a la confianza del 95%) y también el valor predictivo negativo (entre el 73.9% y el 88.7% a la confianza del 95%). Por tanto, cuando la colonoscopia por tomografía computarizada se aplica a los individuos de la población estudiada, es una prueba muy útil tanto para confirmar la neoplasia colorectal como para descartarla.

d) Coeficiente kappa ponderado. Independientemente del valor del índice de ponderación c, el acuerdo más allá del azar entre la colonoscopia por tomografía computarizada y la colonoscopia varía entre moderado y bueno (tanto en términos de las estimaciones puntuales como de los intervalos de confianza). Por tanto, cuando la colonoscopia por tomografía computarizada se aplica a los individuos de la población objeto de estudio, esta prueba es razonablemente buena para el diagnóstico de la neoplasia colorectal.

e) Coeficiente kappa promedio. Para $L' > L$, en términos de los estimadores puntuales, el acuerdo promedio más allá del azar entre

la colonografía por tomografía computarizada y la colonoscopia es moderado $\left(\hat{\kappa}_1 = 0.604\right)$ (aunque muy próximo a bueno). En términos de los intervalos de confianza, el acuerdo promedio más allá del azar es un valor comprendido entre moderado y bueno (al 95% confianza). Asimismo, la pérdida relativa estimada entre los falsos positivos y los falsos negativos es $\hat{c} = 0.256$. Por tanto, si $L' > L$ (como el en caso de un test previo a un tratamiento de riesgo), el acuerdo promedio entre la colonografía por tomografía computarizada y la colonoscopia varía entre moderado y bueno $\left(\hat{\kappa}_1 = 0.604\right)$, en cuyo caso L' es 2.91 veces mayor que L.

Si $L > L'$ el acuerdo promedio más allá del azar entre la colonografía por tomografía computarizada y la colonoscopia es bueno $\left(\hat{\kappa}_2 = 0.672\right)$ (entre moderado y bueno, en términos de los intervalos de confianza al 95%), y la pérdida relativa estimada entre los falsos positivos y los falsos negativos es $\hat{c} = 0.752$. Por tanto, si $L > L'$ (como en el caso de que la colonografía se utilice como un test de screening) el acuerdo promedio más allá del azar entre la colonografía por tomografía computarizada y la colonoscopia es un valor entre moderado y bueno, y L es 3.03 veces mayor que L'.

Analizados los resultados, la colonografía por tomografía computarizada se puede utilizar tanto si $L' > L$ o si $L > L'$, siendo el acuerdo promedio más allá del azar entre moderado y bueno respectivamente. Por tanto, cuando la colonoscopia por tomografía computarizada se aplica a los individuos de la población objeto de estudio, esta prueba es razonablemente buena para el diagnóstico de la neoplasia colorectal.

Capítulo 3

Comparación de parámetros de dos test diagnósticos binarios bajo un diseño pareado

En los Capítulos anteriores se han definido las principales medidas de los *TDBs* y se han estudiado sus estimaciones. En este Capítulo, y debido a la relevancia que supone conocer cuál es el *TDB* más preciso para el diagnóstico de una determinada enfermedad, se presentan los métodos estadísticos para comparar parámetros de dos *TDBs* bajo un diseño apareado. Un diseño apareado consiste en la aplicación de dos *TDBs* y del *GE* a todos los individuos de una muestra aleatoria. Los parámetros que se van a comparar son las sensibilidades, especificidades, razones de

verosimilitud, valores predictivos, coeficiente kappa ponderado y coeficientes kappa promedios. También se presentan, en los casos en los que han sido estudiados, los métodos para comparar los parámetros anteriores en la situación de más de dos *TDBs*. Con respecto a las sensibilidades, especificidades, razones de verosimilitud y valores predictivos, se presentan los test de hipótesis y los intervalos de confianza que tienen un mejor comportamiento asintótico en términos de error tipo I y potencia (para los test de hipótesis) y de cobertura (para los intervalos de confianza). Con respecto al coeficiente kappa ponderado, se presentan los únicos métodos que han sido estudiados. Por último, la comparación los coeficientes kappa promedios es la aportación realizada en este Capítulo.

Las Tablas 3.1 y 3.2 muestran las frecuencias observadas y las probabilidades teóricas al aplicar dos *TDBs,* test 1 y test 2, a una muestra de tamaño *n*, donde la variable T_i modeliza el resultado del i-ésimo *TDB* ($T_i = 1$ si el resultado es positivo y $T_i = 0$ si el resultado es negativo), y D modeliza el resultado del *GE* ($D = 1$ si el individuo tiene la enfermedad y $D = 0$ si no la tiene).

Tabla 3.1. Frecuencias observadas al aplicar dos *TDBs* a una muestra aleatoria.

	$T_1=1$		$T_1=0$		
	$T_2=1$	$T_2=0$	$T_2=1$	$T_2=0$	Total
$D=1$	s_{11}	s_{10}	s_{01}	s_{00}	s
$D=0$	r_{11}	r_{10}	r_{01}	r_{00}	r
Total	n_{11}	n_{10}	n_{01}	n_{00}	n

Tabla 3.2. Probabilidades obtenidas al aplicar dos *TDBs* a una muestra aleatoria.

	$T_1=1$		$T_1=0$		
	$T_2=1$	$T_2=0$	$T_2=1$	$T_2=0$	Total
$D=1$	p_{11}	p_{10}	p_{01}	p_{00}	p
$D=0$	q_{11}	q_{10}	q_{01}	q_{00}	q
Total	$p_{11}+q_{11}$	$p_{10}+q_{10}$	$p_{01}+q_{01}$	$p_{00}+q_{00}$	1

3.1. Comparación de las sensibilidades y de las especificidades

En esta Sección se presentan los contrastes de hipótesis para comparar individualmente las sensibilidades y las especificidades de dos *TDBs* y los intervalos de confianza para la diferencia de las dos sensibilidades y de las

dos especificidades. Asimismo, se presenta un contraste de hipótesis global para comparar simultáneamente las dos sensibilidades y las dos especificidades.

A partir de la probabilidades dadas en la Tabla 3.2 las sensibilidades de los dos *TDBs* se escriben como

$$Se_1 = \frac{p_{10} + p_{11}}{p} \tag{3.1}$$

y

$$Se_2 = \frac{p_{01} + p_{11}}{p}, \tag{3.2}$$

y las especificidades como

$$Sp_1 = \frac{q_{00} + q_{01}}{q} \tag{3.3}$$

y

$$Sp_2 = \frac{q_{00} + q_{10}}{q}. \tag{3.4}$$

3.1.1. Comparación individual de las sensibilidades y de las especificidades

El test de hipótesis de comparación de las dos sensibilidades de dos *TDBs* es

$$H_0 : Se_1 = Se_2 \quad \text{vs} \quad H_1 : Se_1 \neq Se_2 ,$$

que es equivalente al test de hipótesis

$$H_0 : \frac{p_{10} + p_{11}}{p} = \frac{p_{01} + p_{11}}{p} \quad \text{vs} \quad H_1 : \frac{p_{10} + p_{11}}{p} \neq \frac{p_{01} + p_{11}}{p},$$

es decir

$$H_0 : p_{01} = p_{10} \quad \text{vs} \quad H_1 : p_{01} \neq p_{10} .$$

Condicionando en los individuos que tienen la enfermedad, este test de hipótesis consiste en la comparación de dos proporciones apareadas y se resuelve aplicando el test de McNemar, cuyo estadístico con corrección por continuidad es

$$z_{\exp} = \frac{|s_{01} - s_{10}| - 0.5}{\sqrt{s_{01} + s_{10}}} \xrightarrow[s_{01}+s_{10}>10]{} N(0,1) . \qquad (3.5)$$

En el caso de que $s_{01}+s_{10}\le 10$, el test de hipótesis se resuelve aplicando el test exacto de comparación de dos proporciones apareadas, cuyo *p*-valor viene dado por la expresión

$$p-valor = 2\times\sum_{j=0}^{k}\binom{s_{01}+s_{10}}{j}\left(\frac{1}{2}\right)^{s_{01}+s_{10}}, \qquad (3.6)$$

siendo $k=\min\left(s_{01},s_{10}\right)$.

El test de hipótesis para comparar las dos especificidades se define de forma similar al caso de las dos sensibilidades, esto es

$$H_0: Sp_1 = Sp_2 \quad \text{vs} \quad H_1: Sp_1 \neq Sp_2,$$

o equivalentemente

$$H_0: q_{01} = q_{10} \quad \text{vs} \quad H_1: q_{01} \neq q_{10}.$$

Condicionando en los individuos que no tienen la enfermedad, el estadístico de contraste es

$$z_{\exp} = \frac{\left|r_{01}-r_{10}\right|-0,5}{\sqrt{r_{01}+r_{10}}} \xrightarrow[r_{01}+r_{10}>10]{} N(0,1). \qquad (3.7)$$

Si $r_{01}+r_{10}\le 10$, entonces el p-valor exacto del test de hipótesis es

$$p-valor = 2\times\sum_{j=0}^{k}\binom{r_{01}+r_{10}}{j}\left(\frac{1}{2}\right)^{r_{01}+r_{10}}, \tag{3.8}$$

siendo $k=\min\left(r_{01},r_{10}\right)$.

Por otra parte, la estimación por intervalo de confianza de la diferencia de las dos sensibilidades (especificidades) consiste en la estimación por intervalo de la diferencia de dos proporciones binomiales apareadas. En la literatura Estadística se han propuesto varios intervalos de confianza para la diferencia de dos proporciones binomiales apareadas, siendo el que mejor rendimiento tiene el propuesto por Agresti y Min (2005), denominado intervalo de confianza $Wald+2$. Por tanto, el intervalo de confianza $Wald+2$ para la diferencia de las dos sensibilidades al nivel de confianza del $100(1-\alpha)\%$ es

$$Se_1 - Se_2 \in \frac{s_{10}-s_{01}}{s+2} \pm z_{1-\alpha/2}\frac{\sqrt{\left(s_{10}+s_{01}+1\right)-\frac{\left(s_{10}-s_{01}\right)^2}{s+2}}}{s+2} \tag{3.9}$$

y para la diferencia de las especificidades es

$$Sp_1 - Sp_2 \in \frac{r_{01}-r_{10}}{r+2} \pm z_{1-\alpha/2}\frac{\sqrt{\left(r_{01}+r_{10}+1\right)-\frac{\left(r_{01}-r_{10}\right)^2}{r+2}}}{r+2}. \tag{3.10}$$

3.1.2. Comparación simultánea de las sensibilidades y especificidades

El contraste de hipótesis global para la comparación de las sensibilidades y de las especificidades de dos *TDBs* (Lachembruch et al, 1998) es

$$H_0 : Se_1 = Se_2 \cap Sp_1 = Sp_2 \quad \text{vs} \quad H_1 : Se_1 \neq Se_2 \cup Sp_1 \neq Sp_2 ,$$

o equivalentemente

$$H_0 : p_{10} = p_{01} \cap q_{01} = q_{10} \quad \text{vs} \quad H_1 : p_{10} \neq p_{01} \cup q_{01} \neq q_{10} .$$

El estadístico para este contraste de hipótesis viene dado por la siguiente expresión

$$Q^2_{\text{exp}} = \frac{\left(s_{10} - s_{01}\right)^2}{s_{10} + s_{01}} + \frac{\left(r_{01} - r_{10}\right)^2}{r_{01} + r_{10}} \tag{3.11}$$

y que se distribuye según una chi-cuadrado de dos grados de libertad cuando la hipótesis nula es cierta. Una alternativa a este estadístico de contraste es el obtenido mediante el test de la razón de verosimilitud, cuya expresión es

$$Q^2_{\text{exp}} = 2 \times \left[s_{10} \ln\left\{\frac{2s_{10}}{s_{10} + s_{01}}\right\} + s_{01} \ln\left\{\frac{2s_{01}}{s_{01} + s_{10}}\right\} + r_{10} \ln\left\{\frac{2r_{10}}{r_{10} + r_{01}}\right\} + r_{01} \ln\left\{\frac{2r_{01}}{r_{01} + r_{10}}\right\} \right] \tag{3.12}$$

y que también se distribuye según una chi-cuadrado con dos grados de libertad cuando la hipótesis nula es cierta.

3.2. Comparación de las razones de verosimilitud

La comparación de las razones de verosimilitud de dos *TDBs* bajo un diseño apareado ha sido objeto de varios estudios. Leisenring y Pepe (1998) han estudiado la estimación y comparación de las razones de verosimilitud de dos *TDBs* mediante modelos *GEE*. Roldán Nofuentes y Luna del Castillo (2007) han estudiado la comparación individual y conjunta de las razones de verosimilitud mediante el método de máxima verosimilitud, y han comparado sus resultados con los de Leisenring y Pepe, obteniendo que ambos métodos proporcionan resultados (en términos de error tipo I y potencia) muy similares. Dolgun et al (2012) han ampliado los resultados de Leisenring y Pepe y han propuesto un modelo *GEE* para comparar simultáneamente las razones de verosimilitud de los dos *TDBs*, y han comparado su modelo con el modelo conjunto propuesto por Roldán Nofuentes y Luna del Castillo, obteniéndose que ambos modelos tiene un error tipo I y una potencia muy similares. A continuación se presenta el método de Roldán-Nofuentes y Luna del Castillo (2007).

Sean las probabilidades $\theta_{ij} = P\left(D=1 \middle| T_1 = i, T_2 = j\right)$ y $\eta_{ij} = \left(T_1 = i, T_2 = j\right)$ con $i, j = 0,1$ y $\eta_{11} = 1 - \eta_{00} - \eta_{01} - \eta_{10}$, y $\boldsymbol{\theta} = \left(\theta_{00}, \theta_{01}, \theta_{10}, \theta_{11}\right)^T$ y

$\boldsymbol{\eta} = \left(\eta_{00}, \eta_{01}, \eta_{10}\right)^T$ los vectores de dichas probabilidades. A partir de estas probabilidades las razones de verosimilitud positivas de los dos *TDBs* se expresan como

$$LR_1(+) = \left(\frac{1 - \sum_{i,j=0}^{1} \theta_{ij}\eta_{ij}}{\sum_{i,j=0}^{1} \theta_{ij}\eta_{ij}} \right) \left(\frac{\sum_{j=0}^{1} \theta_{1j}\eta_{1j}}{\sum_{j=0}^{1} \left(1 - \theta_{1j}\right)\eta_{1j}} \right) \tag{3.13}$$

y

$$LR_2(+) = \left(\frac{1 - \sum_{i,j=0}^{1} \theta_{ij}\eta_{ij}}{\sum_{i,j=0}^{1} \theta_{ij}\eta_{ij}} \right) \left(\frac{\sum_{i=0}^{1} \theta_{i1}\eta_{i1}}{\sum_{i=0}^{1} \left(1 - \theta_{i1}\right)\eta_{i1}} \right), \tag{3.14}$$

y las razones de verosimilitud negativas como

$$LR_1(-) = \left(\frac{1 - \sum_{i,j=0}^{1} \theta_{ij}\eta_{ij}}{\sum_{i,j=0}^{1} \theta_{ij}\eta_{ij}} \right) \left(\frac{\sum_{j=0}^{1} \theta_{0j}\eta_{0j}}{\sum_{j=0}^{1} \left(1 - \theta_{0j}\right)\eta_{0j}} \right) \tag{3.15}$$

y

$$LR_2(-) = \left(\frac{1 - \sum_{i,j=0}^{1} \theta_{ij}\eta_{ij}}{\sum_{i,j=0}^{1} \theta_{ij}\eta_{ij}} \right) \left(\frac{\sum_{i=0}^{1} \theta_{i0}\eta_{i0}}{\sum_{j=0}^{1} \left(1 - \theta_{i0}\right)\eta_{i0}} \right). \tag{3.16}$$

El logaritmo de la función de verosimilitud de los datos de la Tabla 3.1 es

$$\begin{aligned} l(\boldsymbol{\theta},\boldsymbol{\eta}) &= \sum_{i,j=0}^{1} s_{ij} \ln\left\{P(D=1,T_1=i,T_2=j)\right\} + \sum_{i,j=0}^{1} r_{ij} \ln\left\{P(D=0,T_1=i,T_2=j)\right\} \\ &= \sum_{i,j=0}^{1} s_{ij} \ln\left\{\theta_{ij}\eta_{ij}\right\} + \sum_{i,j=0}^{1} r_{ij} \ln\left\{(1-\theta_{ij})\eta_{ij}\right\} = \\ &= \sum_{i,j=0}^{1} \left\{s_{ij} \ln\theta_{ij} + r_{ij} \ln(1-\theta_{ij})\right\} + \sum_{i,j=0}^{1} n_{ij} \ln(\eta_{ij}) \\ &= l(\boldsymbol{\theta}) + l(\boldsymbol{\eta}), \end{aligned} \tag{3.17}$$

con $n_{ij} = s_{ij} + r_{ij}$. Maximizando esta función, los estimadores máximo verosímiles de θ_{ij} y de η_{ij} son

$$\hat{\theta}_{ij} = \frac{s_{ij}}{n_{ij}} \quad \text{y} \quad \hat{\eta}_{ij} = \frac{n_{ij}}{n}, \tag{3.18}$$

con $i, j = 0,1$, y la matriz de información de Fisher de la función $l(\boldsymbol{\theta},\boldsymbol{\eta})$ es

$$I_{(\boldsymbol{\theta},\boldsymbol{\eta})} = diag\left\{I_{\boldsymbol{\theta}}, I_{\boldsymbol{\eta}}\right\}, \tag{3.19}$$

donde $I_{\boldsymbol{\theta}}$ y $I_{\boldsymbol{\eta}}$ son las matrices de información de Fisher de las funciones $l(\boldsymbol{\theta})$ y $l(\boldsymbol{\eta})$ y sus expresiones son (Zhou, 1998)

$$I_{\theta}^{-1} = diag\left\{\frac{\theta_{ij}^2(1-\theta_{ij})^2}{s_{ij}(1-\theta_{ij})^2 + r_{ij}\theta_{ij}^2}\right\}, \quad i,j = 0,1 \tag{3.20}$$

y

$$I_{\eta}^{-1} = diag\left(\frac{n_{00}^2}{n_{00}}, \frac{n_{01}^2}{n_{01}}, \frac{n_{10}^2}{n_{10}}\right) - \frac{1}{\sum_{i,j=0}^{1} \frac{\eta_{ij}^2}{n_{ij}}} \left(\frac{n_{00}^2}{n_{00}}, \frac{n_{01}^2}{n_{01}}, \frac{n_{10}^2}{n_{10}}\right)^T \left(\frac{n_{00}^2}{n_{00}}, \frac{n_{01}^2}{n_{01}}, \frac{n_{10}^2}{n_{10}}\right). \quad (3.21)$$

Los estimadores máximo verosímiles de las razones de verosimilitud positivas viene dadas por las expresiones

$$\hat{LR}_1(+) = \frac{(s_{10} + s_{11})/s}{(r_{10} + r_{11})/r} \quad (3.22)$$

y

$$\hat{LR}_2(+) = \frac{(s_{01} + s_{11})/s}{(r_{01} + r_{11})/r}, \quad (3.23)$$

y los de las razones de verosimilitud negativos son

$$\hat{LR}_1(-) = \frac{(s_{00} + s_{01})/s}{(r_{00} + r_{01})/r} \quad (3.24)$$

y

$$\hat{LR}_2(+) = \frac{(s_{00} + s_{10})/s}{(r_{00} + r_{10})/r}. \quad (3.25)$$

3.2.1. Comparación individual de las razones de verosimilitud

El contraste de hipótesis para la comparación de las razones de verosimilitud, positivas y negativas, han sido estudiadas por Roldán Nofuentes y Luna del Castillo (2007). El contraste de hipótesis para comparar las razones de verosimilitud positiva es

$$H_0 : \omega^+ = 0 \quad \text{vs} \quad H_1 : \omega^+ \neq 0 \ ,$$

y para comparar las razones de verosimilitud negativas es

$$H_0 : \omega^- = 0 \quad \text{vs} \quad H_1 : \omega^- \neq 0 \ ,$$

siendo $\omega^+ = \log\left(LR_1(+)/LR_2(+)\right)$ y $\omega^- = \log\left(LR_1(-)/LR_2(-)\right)$. Los estimadores máximo verosímiles de ω^+ y ω^- son

$$\hat{\omega}^+ = \log\left\{\frac{(s_{10}+s_{11})(r_{01}+r_{11})}{(s_{01}+s_{11})(r_{10}+r_{11})}\right\} \tag{3.26}$$

y

$$\hat{\omega}^- = \log\left\{\frac{(s_{01}+s_{00})(r_{10}+r_{00})}{(s_{10}+s_{00})(r_{01}+r_{00})}\right\}. \tag{3.27}$$

Aplicando el método delta (Agresti, 2002), la varianza de ω^+ es

$$Var\left(\omega^{+}\right)=\left(\frac{\partial\omega^{+}}{\partial\boldsymbol{\theta}}\right)I_{\boldsymbol{\theta}}^{-1}\left(\frac{\partial\omega^{+}}{\partial\boldsymbol{\theta}}\right)^{T}+\left(\frac{\partial\omega^{+}}{\partial\boldsymbol{\eta}}\right)I_{\boldsymbol{\eta}}^{-1}\left(\frac{\partial\omega^{+}}{\partial\boldsymbol{\eta}}\right)^{T} \tag{3.28}$$

y el estadístico para el test de hipótesis de igualdad de las dos razones de verosimilitud positivas es

$$z_{\exp}=\frac{\hat{\omega}^{+}}{\sqrt{\hat{V}ar\left(\hat{\omega}^{+}\right)}}\xrightarrow[n\to\infty]{}N\left(0,1\right). \tag{3.29}$$

De forma similar, la varianza de ω^{-} es

$$Var\left(\omega^{-}\right)=\left(\frac{\partial\omega^{-}}{\partial\boldsymbol{\theta}}\right)I_{\boldsymbol{\theta}}^{-1}\left(\frac{\partial\omega^{-}}{\partial\boldsymbol{\theta}}\right)^{T}+\left(\frac{\partial\omega^{-}}{\partial\boldsymbol{\eta}}\right)I_{\boldsymbol{\eta}}^{-1}\left(\frac{\partial\omega^{-}}{\partial\boldsymbol{\eta}}\right)^{T} \tag{3.30}$$

y el estadístico para el test de hipótesis de igualdad de las dos razones de verosimilitud negativas es

$$z_{\exp}=\frac{\hat{\omega}^{-}}{\sqrt{\hat{V}ar\left(\hat{\omega}^{-}\right)}}\xrightarrow[n\to\infty]{}N\left(0,1\right). \tag{3.31}$$

Asimismo, un intervalo de confianza asintótico para ω es

$$\omega\in\hat{\omega}\pm z_{1-\alpha/2}\sqrt{\hat{V}ar\left(\hat{\omega}\right)}, \tag{3.32}$$

donde ω es ω^+ u ω^- y un intervalo para el cociente de las dos razones de verosimilitud positivas o negativas es,

$$\frac{LR_1}{LR_2} \in \exp\left\{\hat{\omega} \pm z_{1-\alpha/2}\sqrt{\hat{Var}\left(\hat{\omega}\right)}\right\}, \tag{3.33}$$

donde LR_i es $LR_i(+)$ o $LR_i(-)$ respectivamente.

Roldán Nofuentes y Luna del Castillo (2007) han realizado experimentos de simulación para estudiar el comportamiento asintótico de los contrastes de hipótesis individuales, obteniendo que, en términos generales, los dos test de hipótesis, $H_0 : \omega^+ = 0$ y $H_0 : \omega^- = 0$, tiene unos errores tipo I que suelen ser menores que el error nominal del 5%, y que con muestras de entre 200 y 500 individuos (dependiendo de la prevalencia de la enfermedad) ambas potencias son elevadas (superiores al 80%), aunque en algunas situaciones (cuando la prevalencia es superior al 50%) se requieren tamaños muestrales superiores a 500 para que la potencia del test $H_0 : \omega^- = 0$ sea elevada.

3.2.2. Comparación simultánea de las razones de verosimilitud

Las razones de verosimilitud positiva y negativa no son parámetros independientes, por lo que se pueden comparar de forma simultánea. En esta

situación, Roldán-Nofuentes y Luna del Castillo (2007) han estudiado el test de hipótesis

$$H_0:\left(\omega^- = 0\right)\cap\left(\omega^+ = 0\right) \quad \text{vs} \quad H_1:\left(\omega^- \neq 0\right)\cup\left(\omega^+ \neq 0\right).$$

Los datos de la Tabla 3.1 son la realización de una distribución multinomial por lo que aplicando el teorema central del límite multivariante se tiene que

$$\frac{1}{\sqrt{n}}\begin{pmatrix}\hat{\omega}^+ - \omega^+\\ \hat{\omega}^- - \omega^-\end{pmatrix}\xrightarrow[n\to\infty]{} N\left[\mu = \begin{pmatrix}0\\0\end{pmatrix}, \sum\nolimits_{\omega} = \begin{pmatrix}\sigma_{11} & \sigma_{12}\\ \sigma_{12} & \sigma_{22}\end{pmatrix}\right] \qquad (3.34)$$

donde los elementos de la matriz de varianzas-covarianzas se obtienen aplicando el método delta. Sea $\boldsymbol{\zeta} = \left(\omega^+, \omega^-\right)^T$, entonces su matriz de varianzas-covarianzas es

$$\sum\nolimits_{\zeta} = \left(\frac{\partial\boldsymbol{\zeta}}{\partial\boldsymbol{\theta}}\right) I_{\boldsymbol{\theta}}^{-1}\left(\frac{\partial\boldsymbol{\zeta}}{\partial\boldsymbol{\theta}}\right)^T + \left(\frac{\partial\boldsymbol{\zeta}}{\partial\boldsymbol{\eta}}\right) I_{\boldsymbol{\eta}}^{-1}\left(\frac{\partial\boldsymbol{\zeta}}{\partial\boldsymbol{\eta}}\right)^T \qquad (3.35)$$

y el estadístico para el test de hipótesis de comparación simultánea de las razones de verosimilitud positivas y negativas es

$$Q_{\exp}^2 = \hat{\boldsymbol{\zeta}}^T \hat{\Sigma}_{\hat{\zeta}}^{-1} \hat{\boldsymbol{\zeta}} \xrightarrow[n\to\infty]{} \chi_2^2, \qquad (3.36)$$

donde $\hat{\Sigma}_{\hat{\zeta}}$ se obtiene al sustituir en $\sum_{\zeta}$ cada parámetro por su correspondiente estimador máximo verosímil.

3.3. Comparación de los valores predictivos

La comparación de los valores predictivos de dos test diagnósticos binarios es un tópico de especial interés en el estudio de los métodos estadísticos para el diagnóstico y ha sido objeto también de diversos estudios en la literatura Estadística. Bennett (1972, 1985) ha estudiado la comparación de los valores predictivos positivos (negativos) de test diagnósticos binarios proponiendo un test basado en la distribución chi-cuadrado. Jamart (1993) ha discutido los resultados de Bennett, indicando que estos no son apropiados para resolver este problema de inferencia, ya que los métodos de Bennett depende del orden de los *TDBs*. Leisenring et al (2000) han estudiado la comparación de los *VPs* de dos *TDBs* mediante modelos de regresión marginal. Wang et al (2006) han estudiado el mismo problema mediante un modelo de mínimos cuadrados ponderados. Kosinski (2012) ha propuesto un estadístico score generalizado ponderado para resolver este mismo problema, demostrando que su método tiene un mejor

comportamiento en términos de error tipo I que los anteriores. Roldán Nofuentes et al (2012) han estudiado un test de hipótesis global para comparar simultáneamente los valores predictivos de dos (o más) test diagnósticos binarios, proponiéndose un método basado en la distribución chi-cuadrado y en comparaciones múltiples. A continuación se presentan el método de Kosinski (2012), por ser el método de comparación individual de los valores predictivos que presenta un mejor comportamiento asintótico, y el método de comparación simultánea de Roldán Nofuentes et al (2012).

3.3.1. Comparación individual de los VPs

Kosinski (2012) ha propuesto un estadístico score ponderado generalizado para resolver los test de hipótesis de comparación de los valores predictivos positivos y negativos respectivamente. El estadístico score ponderado generalizado para el test

$$H_0 : VPP_1 = VPP_2 \quad \text{vs} \quad H_1 : VPP_1 \neq VPP_2$$

es

$$T_{VPP}^{WGS} = \frac{\left(\hat{VPP}_1 - \hat{VPP}_2\right)^2}{\left\{\hat{VPP}_p\left(1-\hat{VPP}_p\right) - 2C_p^{VPP}\right\}\left(\frac{1}{n_{10}+n_{11}} + \frac{1}{n_{01}+n_{11}}\right)}, \qquad (3.37)$$

y el estadístico score ponderado generalizado para el test

$$H_0 : VPN_1 = VPN_2 \quad \text{vs} \quad H_1 : VPN_1 \neq VPN_2$$

es

$$T_{VPN}^{WGS} = \frac{\left(\hat{VPN}_1 - \hat{VPN}_2\right)^2}{\left\{\hat{VPN}_p\left(1-\hat{VPN}_p\right) - 2C_p^{VPN}\right\}\left(\frac{1}{n_{00}+n_{01}} + \frac{1}{n_{00}+n_{10}}\right)}, \qquad (3.38)$$

que se distribuyen asintóticamente según una chi-cuadrado con 1 grado de libertad cuando la hipótesis nula es cierta, siendo

$$\hat{VPP}_p = \frac{2s_{11}+s_{10}+s_{01}}{2n_{11}+n_{10}+n_{01}} \quad \text{y} \quad \hat{VPN}_p = \frac{2r_{00}+r_{01}+r_{10}}{2n_{00}+n_{01}+n_{10}} \qquad (3.39)$$

y

$$C_p^{VPP} = \frac{s_{11}\left(1-\hat{VPP}_p\right)^2 + r_{11}\hat{VPP}_p^2}{2n_{11}+n_{10}+n_{01}} \quad \text{y} \quad C_p^{VPN} = \frac{s_{00}\hat{VPN}_p^2 + r_{00}\left(1-\hat{VPN}_p\right)^2}{2n_{00}+n_{01}+n_{10}}. \qquad (3.40)$$

Kosinski también ha propuesto los siguientes intervalos de confianza para la diferencia de los valores predictivos,

$$VPP_1 - VPP_2 \in$$
$$\left(\hat{V}PP_1 - \hat{V}PP_2\right) \pm z_{1-\alpha/2}\sqrt{\left\{\hat{V}PP_p\left(1-\hat{V}PP_p\right) - 2C_p^{VPP}\right\}\left(\frac{1}{n_{10}+n_{11}} + \frac{1}{n_{01}+n_{11}}\right)} \quad (3.41)$$

y

$$VPN_1 - VPN_2 \in$$
$$\left(\hat{V}PN_1 - \hat{V}PN_2\right) \pm z_{1-\alpha/2}\sqrt{\left\{\hat{V}PN_p\left(1-\hat{V}PN_p\right) - 2C_p^{VPN}\right\}\left(\frac{1}{n_{00}+n_{01}} + \frac{1}{n_{00}+n_{10}}\right)}. \quad (3.42)$$

3.3.2. Comparación simultánea de los VPs de dos TDBs

Roldán-Nofuentes et al (2012) han estudiado la comparación simultánea de los valores predictivos de dos test diagnósticos binarios. En términos de las probabilidades de la Tabla 3.2, los valores predictivos positivos se escriben como

$$VVP_1 = \frac{p_{11}+p_{10}}{p_{11}+p_{10}+q_{11}+q_{10}} \quad (3.43)$$

y

$$VPP_2 = \frac{p_{11}+p_{01}}{p_{11}+p_{01}+q_{11}+q_{01}}, \quad (3.44)$$

y los valores predictivos negativos como

$$VPN_1 = \frac{q_{01} + q_{00}}{p_{01} + p_{00} + q_{01} + q_{00}} \tag{3.45}$$

y

$$VPN_2 = \frac{q_{10} + q_{00}}{p_{10} + p_{00} + q_{10} + q_{00}}. \tag{3.46}$$

Sean los vectores $\boldsymbol{\gamma} = \left(s_{11}, s_{10}, s_{01}, s_{00}, r_{11}, r_{10}, r_{01}, r_{00}\right)^T$ y $\boldsymbol{\pi} = \left(p_{11}, p_{10}, p_{01}, p_{00}, q_{11}, q_{10}, q_{01}, q_{00}\right)^T$. Como $\boldsymbol{\pi}$ es el vector de probabilidades de una distribución multinomial su matriz de varianzas-covarianzas es

$$\Sigma_{\hat{\boldsymbol{\pi}}} = \left\{\text{Diag}\left(\boldsymbol{\pi}\right) - \boldsymbol{\pi}\boldsymbol{\pi}^T\right\}\Big/ n \tag{3.47}$$

y su estimador es

$$\hat{\boldsymbol{\pi}} = \frac{\boldsymbol{\gamma}}{n}. \tag{3.48}$$

Finalmente los estimadores máximo verosímiles de los valores predictivos positivos de cada *TDB* son

$$\hat{VPP}_1 = \frac{s_{10} + s_{11}}{s_{10} + s_{11} + r_{10} + r_{11}} \quad \text{y} \quad \hat{VPP}_2 = \frac{r_{00} + r_{01}}{s_{00} + s_{01} + r_{00} + r_{01}} \tag{3.49}$$

y los estimadores de los valores predictivos negativos son

$$\hat{VPN}_1 = \frac{s_{01} + s_{11}}{s_{01} + s_{11} + r_{01} + r_{11}} \quad \text{y} \quad \hat{VPN}_2 = \frac{r_{00} + r_{10}}{s_{00} + s_{10} + r_{00} + r_{10}} \qquad (3.50)$$

Aplicando el método delta, las varianzas-covarianzas asintóticas estimadas de los estimadores de los valores predictivos son

$$\hat{Var}\left(\hat{VPP}_1\right) = \frac{\left(s_{10} + s_{11}\right)\left(r_{10} + r_{11}\right)}{\left(s_{10} + s_{11} + r_{10} + r_{11}\right)^2}, \; \hat{Var}\left(\hat{VPP}_2\right) = \frac{\left(s_{01} + s_{11}\right)\left(r_{01} + r_{11}\right)}{\left(s_{01} + s_{11} + r_{01} + r_{11}\right)^2},$$

$$\hat{Var}\left(\hat{VPN}_1\right) = \frac{\left(s_{00} + s_{01}\right)\left(r_{00} + r_{01}\right)}{\left(s_{00} + s_{01} + r_{00} + r_{01}\right)^2}, \; \hat{Var}\left(\hat{VPN}_2\right) = \frac{\left(s_{00} + s_{10}\right)\left(r_{00} + r_{10}\right)}{\left(s_{00} + s_{10} + r_{00} + r_{10}\right)^2},$$

$$\hat{Cov}\left(\hat{VPP}_1, \hat{VPP}_2\right) = \frac{s_{01}s_{10}r_{11} + s_{11}\left\{r_{01}\left(r_{10} + r_{11}\right) + r_{11}\left(s_{01} + s_{10} + s_{11} + r_{10} + r_{11}\right)\right\}}{\left(s_{01} + s_{11} + r_{01} + r_{11}\right)^2\left(s_{10} + s_{11} + r_{10} + r_{11}\right)^2},$$

$$\hat{Cov}\left(\hat{VPP}_1, \hat{VPN}_2\right) = -\frac{s_{00}\left(s_{10} + s_{11}\right)r_{10} + s_{10}r_{10}\left(s_{10} + s_{11} + r_{00} + r_{10}\right) + s_{10}\left(r_{00} + r_{10}\right)r_{11}}{\left(s_{00} + s_{10} + r_{00} + r_{10}\right)^2\left(s_{10} + s_{11} + r_{10} + r_{11}\right)^2},$$

$$\hat{Cov}\left(\hat{VPP}_2, \hat{VPN}_1\right) = -\frac{s_{00}\left(s_{01} + s_{11}\right)r_{01} + s_{01}r_{01}\left(s_{01} + s_{11} + r_{00} + r_{01}\right) + s_{01}\left(r_{00} + r_{01}\right)r_{11}}{\left(s_{00} + s_{01} + r_{00} + r_{01}\right)^2\left(s_{01} + s_{11} + r_{01} + r_{11}\right)^2},$$

$$\hat{Cov}\left(\hat{VPN}_1, \hat{VPN}_2\right) = \frac{s_{00}\left(r_{00} + r_{01}\right)r_{10} + r_{00}\left\{r_{00}^2 + s_{01}s_{10} + s_{00}\left(s_{01} + s_{10} + r_{00} + r_{01}\right)\right\}}{\left(s_{00} + s_{01} + r_{00} + r_{01}\right)^2\left(s_{00} + s_{10} + r_{00} + r_{10}\right)^2},$$

$$\hat{Cov}\left(\hat{VPP}_1, \hat{VPN}_1\right) = 0 \quad \text{y} \quad \hat{Cov}\left(\hat{VPP}_2, \hat{VPN}_2\right) = 0.$$

El test de hipótesis para la comparación simultánea de los valores predictivos es

$$H_0 : \left(VPP_1 = VPP_2 \cap VPN_1 = VPN_2\right) \quad \text{vs} \quad H_1 : \left(VPP_1 \neq VPP_2 \cup VPN_1 \neq VPN_2\right),$$

y el estadístico para este test de hipótesis viene dado por la expresión

$$Q^2_{\text{exp}} = \hat{\boldsymbol{\eta}}^T \boldsymbol{\varphi}^T \left(\boldsymbol{\varphi} \hat{\Sigma} \boldsymbol{\varphi}^T\right)^{-1} \boldsymbol{\varphi} \hat{\boldsymbol{\eta}} \xrightarrow[n\to\infty]{} \chi^2_2, \tag{3.51}$$

donde $\hat{\Sigma}$ es la matriz de varianzas-covarianzas asintóticas estimadas de $\hat{\boldsymbol{\eta}} = \left(\hat{VPP}_1, \hat{VPP}_2, \hat{VPN}_1, \hat{VPN}_2\right)$ y $\boldsymbol{\varphi}$ es la matriz de diseño cuya expresión es

$$\boldsymbol{\varphi} = \begin{pmatrix} 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 \end{pmatrix}.$$

El estadístico Q^2_{exp} se distribuye asintóticamente según una distribución chi-cuadrado central con dos grados de libertad bajo la hipótesis nula. Para poder aplicar este método es necesario que todos los valores predictivos se puedan estimar y que la matriz $\boldsymbol{\varphi} \hat{\Sigma} \boldsymbol{\varphi}^T$ sea no singular. Si el test de hipótesis es significativo al error α, la investigación de las causas de la significación se realiza comparando los valores predictivos positivos y los valores predictivos negativos de forma independiente (por ejemplo, utilizando el método de Wang et al o el de Kosinski) y posteriormente aplicando algún

método de comparaciones múltiples (por ejemplo, el método de Holm (1979) o el de Hochberg (1983)). Experimentos de simulación han demostrado que, en términos generales, es necesario un tamaño muestral entre 300 y 500 individuos para que la potencia del test de hipótesis sea alta (mayor del 80%). En cuanto al error tipo I, este tiene el comportamiento clásico de un test asintótico, fluctuando en torno al error nominal (5%) a partir de tamaños muestrales grandes.

3.3.3. Comparación simultánea de los VPs de más de dos TDBs

Roldán Nofuentes et al (2012) también han estudiado la comparación simultánea de los valores predictivos de *J TDBs* $\left(J \geq 3\right)$ cuando todos los test y el *GE* se aplican a una misma muestra de individuos. En esta situación, sea T_j la variable aleatoria que modeliza el resultado del j-ésimo *TDB* $\left(j = 1,...,J\right)$ y D la variable que modeliza el resultado del gold estándar. Cuando los *J* test binarios y el gold estándar se aplican a todos los individuos de una muestra aleatoria de tamaño *n*, $s_{i_1,...,i_J}$ es el número de individuos enfermos en los que $T_1 = i_1$, $T_2 = i_2,...,$ $T_J = i_J$, y $r_{i_1,...,i_J}$ es el número de individuos no enfermos en los que $T_1 = i_1$, $T_2 = i_2,...,$ $T_J = i_J$, con

$i_j = 0,1$ y $j = 1,...,J$. Sea $s = \sum_{i_1,...,i_J=0}^{1} s_{i_1,...,i_J}$ el número total de individuos enfermos y $r = \sum_{i_1,...,i_J=0}^{1} r_{i_1,...,i_J}$ el número total de individuos no enfermos, con $n = s + r$. En esta situación se definen las probabilidades

$$\begin{aligned} p_{i_1,...,i_J} &= P\left(D = 1, T_1 = i_1, T_2 = i_2, ..., T_J = i_J\right) \\ q_{i_1,...,i_J} &= P\left(D = 0, T_1 = i_1, T_2 = i_2, ..., T_J = i_J\right). \end{aligned} \tag{3.52}$$

Sea $\boldsymbol{\pi} = \left(p_{1,...,1}, ..., p_{0,...,0}, q_{1,...,1} ..., q_{0,...,0}\right)^T$ un vector de dimensión 2^{J+1} cuyas componentes son las probabilidades anteriores y sea $\boldsymbol{\eta} = \left(PPV_1, ..., PPV_J, NPV_1, ..., NPV_J\right)^T$ un vector de dimensión $2J$ cuyas componentes son los valores predictivos positivos y negativos de cada uno de los *J TDBs*. En términos de las probabilidades del vector $\boldsymbol{\pi}$, los valores predictivos del j-ésimo *TDB* vienen dados por las expresiones

$$VPP_j = \frac{\sum_{\substack{i_1,...,i_J=0 \\ i_j=1}}^{1} p_{i_1,...,i_J}}{\sum_{\substack{i_1,...,i_J=0 \\ i_j=1}}^{1} p_{i_1,...,i_J} + \sum_{\substack{i_1,...,i_J=0 \\ i_j=1}}^{1} q_{i_1,...,i_J}} \quad \text{y} \quad VPN_j = \frac{\sum_{\substack{i_1,...,i_J=0 \\ i_j=0}}^{1} q_{i_1,...,i_J}}{\sum_{\substack{i_1,...,i_J=0 \\ i_j=0}}^{1} p_{i_1,...,i_J} + \sum_{\substack{i_1,...,i_J=0 \\ i_j=0}}^{1} q_{i_1,...,i_J}}. \tag{3.53}$$

Como las probabilidades del vector $\boldsymbol{\pi}$ son las probabilidades de una distribución multinomial, sus estimadores máximo verosímiles son

$\hat{p}_{i_1,\ldots,i_J} = s_{i_1,\ldots,i_J}/n$ y $\hat{q}_{i_1,\ldots,i_J} = r_{i_1,\ldots,i_J}/n$, por lo que los estimadores máximo verosímiles de los valores predictivos del j-ésimo *TDB* son

$$\hat{VPP}_j = \frac{\sum_{\substack{i_1,\ldots,i_J=0 \\ i_j=1}}^{1} s_{i_1,\ldots,i_J}}{\sum_{\substack{i_1,\ldots,i_J=0 \\ i_j=1}}^{1} s_{i_1,\ldots,i_J} + \sum_{\substack{i_1,\ldots,i_J=0 \\ i_j=1}}^{1} r_{i_1,\ldots,i_J}} \quad \text{y} \quad \hat{VPN}_j = \frac{\sum_{\substack{i_1,\ldots,i_J=0 \\ i_j=0}}^{1} r_{i_1,\ldots,i_J}}{\sum_{\substack{i_1,\ldots,i_J=0 \\ i_j=0}}^{1} s_{i_1,\ldots,i_J} + \sum_{\substack{i_1,\ldots,i_J=0 \\ i_j=0}}^{1} r_{i_1,\ldots,i_J}}. \quad (3.54)$$

En cuanto a la matriz de varianzas-covarianzas asintóticas del vector $\hat{\boldsymbol{\eta}}$, aplicando el método delta se obtiene que

$$\Sigma_{\hat{\boldsymbol{\eta}}} = \left(\frac{\partial \boldsymbol{\eta}}{\partial \boldsymbol{\pi}}\right) \Sigma_{\hat{\boldsymbol{\pi}}} \left(\frac{\partial \boldsymbol{\eta}}{\partial \boldsymbol{\pi}}\right)^T, \quad (3.55)$$

donde $\Sigma_{\hat{\boldsymbol{\pi}}}$ es la matriz de varianzas-covarianzas de $\hat{\boldsymbol{\pi}}$ y su expresión es similar a la dada en la ecuación (3.47).

El test de hipótesis conjunto para comparar simultáneamente los valores predictivos positivos y negativos de los *J TDBs* es

$$H_0 : \left(PPV_1 = PPV_2 = \ldots = PPV_J\right) \cap \left(NPV_1 = NPV_2 = \ldots = NPV_J\right)$$
$$H_1 : \text{al menos una igualdad no es cierta.}$$

Este test de hipótesis es equivalente a contrastar

$$H_0 : \boldsymbol{\varphi}\boldsymbol{\eta} = \mathbf{0} \text{ vs } H_1 : \boldsymbol{\varphi}\boldsymbol{\eta} \neq \mathbf{0},$$

donde $\boldsymbol{\varphi}$ es una matriz de rango completo de dimensión $2(J-1)\times 2J$ cuyos elementos son constantes conocidas. Para $J=3$,

$$\boldsymbol{\varphi}=\begin{pmatrix}1&-1&0&0&0&0\\0&1&-1&0&0&0\\0&0&0&1&-1&0\\0&0&0&0&1&-1\end{pmatrix},$$

y para $J=4$

$$\boldsymbol{\varphi}=\begin{pmatrix}1&-1&0&0&0&0&0&0\\0&1&-1&0&0&0&0&0\\0&0&1&-1&0&0&0&0\\0&0&0&0&1&-1&0&0\\0&0&0&0&0&1&-1&0\\0&0&0&0&0&0&1&-1\end{pmatrix}.$$

Finalmente el estadístico para el test de hipótesis es

$$Q^2_{\text{exp}}=\hat{\boldsymbol{\eta}}^T\boldsymbol{\varphi}^T\left(\boldsymbol{\varphi}\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\theta}}}\boldsymbol{\varphi}^T\right)^{-1}\boldsymbol{\varphi}\hat{\boldsymbol{\eta}}\xrightarrow[n\to\infty]{}\chi^2_{2(J-1)} \tag{3.56}$$

es decir, el estadístico se distribuye asintóticamente según una distribución chi-cuadrado central con $2(J-1)$ grados de libertad cuando la hipótesis nula es cierta.

En esta situación, el método para comparar simultáneamente los valores predictivos de los *J TDBs* es: 1) resolver el test de hipótesis conjunto al error α (ecuación (3.56)); 2) si el test de hipótesis conjunto es significativo al error α, la investigación de las causas de la significación se debe realizar comparando los valores predictivos positivos (negativos) de cada pareja de *TDBs* mediante el método de Kosinski junto con un método de comparaciones múltiples (por ejemplo, método de Holm o de Hochberg) al error α.

3.4. Comparación de los coeficientes kappa ponderados

El coeficiente kappa ponderado es un parámetro válido para evaluar y comparar el rendimiento de dos o más *TDBs*. En esta sección se presentan el test de hipótesis para comparar las coeficientes kappa ponderados de dos y de más de dos *TDBs*.

3.4.1. Comparación de dos coeficientes kappa ponderados

Bajo un diseño apareado Bloch (1997) ha estudiado la comparación de los coeficientes kappa ponderados de dos *TDBs*. Adaptando el método

propuesto por Bloch a la notación utilizada en las Tablas 3.1 y 3.2, el coeficiente kappa ponderado del test 1 es

$$\kappa_1(c) = \frac{q\sum_{j=0}^{1} p_{1j} + p\sum_{j=0}^{1} q_{0j} - pq}{cp\left(1-\sum_{j=0}^{1} p_{1j} - \sum_{j=0}^{1} q_{1j}\right) + (1-c)q\left(\sum_{j=0}^{1} p_{1j} + \sum_{j=0}^{1} q_{1j}\right)} \quad (3.57)$$

y el coeficiente kappa ponderado del test 2 es

$$\kappa_2(c) = \frac{q\sum_{i=0}^{1} p_{i1} + p\sum_{i=0}^{1} q_{i0} - pq}{cp\left(1-\sum_{i=0}^{1} p_{i1} - \sum_{i=0}^{1} q_{i1}\right) + (1-c)q\left(\sum_{i=0}^{1} p_{i1} + \sum_{i=0}^{1} q_{i1}\right)}, \quad (3.58)$$

siendo c el índice de ponderación. Sustituyendo en las dos ecuaciones anteriores cada parámetro por su estimador máximo verosímil se obtienen los estimadores máximo verosímiles de cada coeficiente kappa ponderado, siendo sus expresiones

$$\hat{\kappa}_1(c) = \frac{r\sum_{j=0}^{1} s_{1j} + s\sum_{j=0}^{1} r_{0j} - sr}{cs\left(n-\sum_{j=0}^{1} s_{1j} - \sum_{j=0}^{1} r_{1j}\right) + (1-c)r\left(\sum_{j=0}^{1} s_{1j} + \sum_{j=0}^{1} r_{1j}\right)} \quad (3.59)$$

y

$$\hat{\kappa}_2(c)=\frac{r\sum_{i=0}^{1}s_{i1}+s\sum_{i=0}^{1}r_{i0}-sr}{cs\left(n-\sum_{i=0}^{1}s_{i1}-\sum_{i=0}^{1}r_{i1}\right)+(1-c)r\left(\sum_{i=0}^{1}s_{i1}+\sum_{i=0}^{1}s_{i1}\right)}. \qquad (3.60)$$

Aplicando el método delta, las varianzas-covarianzas asintóticas de los estimadores de los coeficientes kappa ponderados son

$$\hat{Var}\left(\hat{\kappa}_1(c)\right)=\frac{1}{n\left\{\left(c\hat{p}\frac{n_{00}+n_{01}}{n}+(1-c)\hat{q}\frac{n_{11}+n_{10}}{n}\right)\right\}^2}\times$$
$$\left\{\left(\hat{q}\left(1-\hat{\kappa}_1(c)\right)\frac{n_{00}+n_{01}}{n}\right)^2\frac{s_{11}+s_{10}}{n}+\left(\hat{p}\left(1-\hat{\kappa}_1(c)\right)\frac{n_{00}+n_{01}}{n}+(1-c)\hat{\kappa}_1(c)\right)^2\frac{r_{11}+r_{10}}{n}+\right. \qquad (3.61)$$
$$\left.\left(\frac{n_{11}+n_{10}}{n}\hat{q}\left(1-\hat{\kappa}_1(c)\right)+c\hat{\kappa}_1(c)\right)^2\frac{s_{01}+s_{00}}{n}+\left(\frac{n_{11}+n_{10}}{n}\hat{p}\left(1-\hat{\kappa}_1(c)\right)\right)^2\frac{r_{01}+r_{00}}{n}\right\},$$

$$\hat{Var}\left(\hat{\kappa}_2(c)\right)=\frac{1}{n\left\{\left(c\hat{p}\frac{n_{00}+n_{10}}{n}+(1-c)\hat{q}\frac{n_{11}+n_{01}}{n}\right)\right\}^2}\times$$
$$\left\{\left(\hat{q}\left(1-\hat{\kappa}_2(c)\right)\frac{n_{00}+n_{10}}{n}\right)^2\frac{s_{11}+s_{01}}{n}+\left(\hat{p}\left(1-\hat{\kappa}_2(c)\right)\frac{n_{00}+n_{10}}{n}+(1-c)\hat{\kappa}_2(c)\right)^2\frac{r_{11}+r_{01}}{n}+\right. \qquad (3.62)$$
$$\left.\left(\frac{n_{11}+n_{01}}{n}\hat{q}\left(1-\hat{\kappa}_2(c)\right)+c\hat{\kappa}_2(c)\right)^2\frac{s_{10}+s_{00}}{n}+\left(\frac{n_{11}+n_{01}}{n}\hat{p}\left(1-\hat{\kappa}_2(c)\right)\right)^2\frac{r_{10}+r_{00}}{n}\right\}$$

y

$$\hat{C}ov\left(\hat{\kappa}_1(c),\hat{\kappa}_2(c)\right)=$$

$$\frac{n}{\left\{c\hat{p}\left(n_{00}+n_{01}\right)+\left(1-c\right)\hat{q}\left(n_{11}+n_{10}\right)\right\}\left\{c\hat{p}\left(n_{00}+n_{10}\right)+\left(1-c\right)\hat{q}\left(n_{11}+n_{01}\right)\right\}}\times$$

$$\left\{\left(1-\hat{\kappa}_1(c)\right)\left(1-\hat{\kappa}_2(c)\right)\left[\left(\frac{\left(r_{00}-r_{11}\right)\left(n_{11}+n_{10}\right)}{n^2}+\frac{r_{11}}{n}\right)\hat{p}^2+\right.\right.$$

$$\left(\frac{\left(s_{11}-s_{00}\right)\left(n_{11}+n_{10}\right)}{n^2}+\frac{s_{00}}{n}\right)\hat{q}^2-\frac{\left(s_{11}+s_{10}\right)\left(n_{00}+n_{01}\right)\left(n_{11}+n_{01}\right)}{n^3}\hat{q}^2-$$

$$\frac{\left(r_{11}+r_{10}\right)\left(n_{00}+n_{01}\right)\left(n_{11}+n_{01}\right)}{n^3}\hat{p}^2-\frac{\left(s_{01}+s_{00}\right)\left(n_{00}+n_{10}\right)\left(n_{11}+n_{10}\right)}{n^3}\hat{q}^2-$$

$$\left.\frac{\left(r_{01}+r_{00}\right)\left(n_{00}+n_{10}\right)\left(n_{11}+n_{10}\right)}{n^3}\hat{p}^2\right]+$$

$$\left(1-\hat{\kappa}_1(c)\right)\hat{\kappa}_2(c)\left[\left(1-c\right)\hat{p}\frac{r_{11}}{n}-\left(1-c\right)\hat{p}\frac{\left(r_{11}+r_{01}\right)\left(n_{11}+n_{10}\right)}{n^2}+\right.$$

$$\left.c\hat{q}\frac{s_{00}}{n}-c\hat{q}\frac{\left(s_{10}+s_{00}\right)\left(n_{00}+n_{01}\right)}{n^2}\right]+$$

$$\hat{\kappa}_1(c)\left(1-\hat{\kappa}_2(c)\right)\left[\left(1-c\right)\hat{p}\frac{r_{11}}{n}-\left(1-c\right)\hat{p}\frac{\left(r_{11}+r_{10}\right)\left(n_{11}+n_{01}\right)}{n^2}+\right.$$

$$\left.\left.c\hat{q}\frac{s_{00}}{n}-c\hat{q}\frac{\left(s_{01}+s_{00}\right)\left(n_{00}+n_{10}\right)}{n^2}\right]+\hat{\kappa}_1(c)\hat{\kappa}(c)_2\left[c^2\frac{s_{00}}{n}+\left(1-c\right)^2\frac{r_{11}}{n}\right]\right\},$$

(3.63)

siendo $n_{ij}=s_{ij}+r_{ij}$, $\hat{p}=s/n$ y $\hat{q}=1-\hat{p}=r/n$. Una vez obtenidos todos los estimadores, el estadístico para contrastar

$$H_0:\kappa_1(c)=\kappa_2(c) \quad \text{vs} \quad H_1:\kappa_1(c)\neq\kappa_2(c)$$

es

$$z_{\exp} = \frac{\left|\hat{\kappa}_1(c) - \hat{\kappa}_2(c)\right|}{\sqrt{\hat{V}ar\left(\hat{\kappa}_1(c)\right) + \hat{V}ar\left(\hat{\kappa}_2(c)\right) - 2\hat{C}ov\left(\hat{\kappa}_1(c), \hat{\kappa}_2(c)\right)}} \to N(0,1), \quad (3.64)$$

y el intervalo de confianza a nivel $100(1-\alpha)\%$ para la diferencia de los dos coeficientes kappa ponderados es

$$\hat{\kappa}_1(c) - \hat{\kappa}_2(c) \pm z_{1-\alpha/2}\sqrt{\hat{V}ar\left(\hat{\kappa}_1(c)\right) + \hat{V}ar\left(\hat{\kappa}_2(c)\right) - 2\hat{C}ov\left(\hat{\kappa}_1(c), \hat{\kappa}_2(c)\right)}. \quad (3.65)$$

3.4.2. Comparación de más de dos coeficientes kappa ponderados

Roldán Nofuentes y Luna del Castillo (2010) han estudiado la comparación los coeficientes kappa ponderados de *J TDBs* $(J \geq 3)$ cuando estos y el *GE* se aplican a todos los individuos de una muestra aleatoria de tamaño *n*. En esta situación es similar a la presentada en la Sección 3.4.3. Generalizando el método de Bloch (1997), el estimador máximo verosímil del coeficiente kappa ponderado del j-ésimo *TDB* es

$$\hat{\kappa}_j(c) = \frac{r\left(\sum_{\substack{i_1,\ldots,i_J=0 \\ i_j=1}}^{1} s_{i_1,\ldots,i_J}\right) + s\left(\sum_{\substack{i_1,\ldots,i_J=0 \\ i_j=0}}^{1} r_{i_1,\ldots,i_J}\right) - sr}{s\left(s - \sum_{\substack{i_1,\ldots,i_J=0 \\ i_j=1}}^{1} s_{i_1,\ldots,i_J} + \sum_{\substack{i_1,\ldots,i_J=0 \\ i_j=0}}^{1} r_{i_1,\ldots,i_J}\right)c + r\left(r + \sum_{\substack{i_1,\ldots,i_J=0 \\ i_j=1}}^{1} s_{i_1,\ldots,i_J} - \sum_{\substack{i_1,\ldots,i_J=0 \\ i_j=0}}^{1} r_{i_1,\ldots,i_J}\right)(1-c)}, \quad (3.66)$$

siendo $s_{i_1,\ldots,i_J}$ el número de individuos enfermos en los que $T_1 = i_1$, $T_2 = i_2,\ldots,$ $T_J = i_J$, con $i_j = 0,1$ y $j = 1,\ldots,J$; $r_{i_1,\ldots,i_J}$ el número de individuos no enfermos en los que $T_1 = i_1$, $T_2 = i_2,\ldots,$ $T_J = i_J$ con $i_j = 0,1$ y $j = 1,\ldots,J$; $s = \sum_{i_1,\ldots,i_J=0}^{1} s_{i_1,\ldots,i_J}$ el número total de individuos enfermos; $r = \sum_{i_1,\ldots,i_J=0}^{1} r_{i_1,\ldots,i_J}$ el número total de individuos no enfermos y $n = s + r$. Para un mismo valor del índice de ponderación *c*, el test de hipótesis para contrastar la igualdad de los *J* coeficientes kappa ponderados es

$$H_0 : \kappa_1(c) = \kappa_2(c) = \ldots = \kappa_J(c) \quad \text{vs} \quad H_1 : \text{al menos una igualdad no es cierta.}$$

Este test de hipótesis es equivalente a contrastar

$$H_0 : \boldsymbol{\varphi\kappa} = \mathbf{0} \quad \text{vs} \quad H_1 : \boldsymbol{\varphi\kappa} \neq \mathbf{0},$$

donde $\boldsymbol{\varphi}$ es una matriz $(J-1)\times J$ de rango completo cuyos elementos son constantes conocidas. Para tres *TDBs* se tiene que

$$\boldsymbol{\varphi} = \begin{pmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \end{pmatrix}.$$

El estadístico para el test de hipótesis de igualdad de los *J* coeficientes kappa ponderados es

$$Q^2 = \hat{\boldsymbol{\kappa}}^T \boldsymbol{\varphi}^T \left(\boldsymbol{\varphi} \hat{\Sigma}_{\hat{\boldsymbol{\kappa}}} \boldsymbol{\varphi}^T \right)^{-1} \boldsymbol{\varphi} \hat{\boldsymbol{\kappa}} \xrightarrow[n \to \infty]{} \chi^2_{J-1}, \tag{3.67}$$

siendo $\hat{\boldsymbol{\kappa}} = \left(\hat{\kappa}_1(c), \hat{\kappa}_2(c), ..., \hat{\kappa}_J(c) \right)^T$ y $\hat{\Sigma}_{\hat{\boldsymbol{\kappa}}}$ la matriz de varianzas-covarianzas asintóticas estimadas de $\hat{\boldsymbol{\kappa}}$. Esta matriz de varianzas-covarianzas se puede estimar mediante el método delta o bien mediante bootstrap. Roldán Nofuentes y Luna del Castillo (2010) han comprobado mediante experimentos de simulación que para tres test diagnósticos binarios, el test de hipótesis conjunto tiene un buen rendimiento en términos de error tipo I y potencia para muestras de al menos 500 individuos (el error tipo I fluctúa en torno al error nominal y la potencia es superior al 80%). Cuando el test de hipótesis es significativo al error α, la investigación de las causas de la significación se realiza resolviendo las comparaciones por parejas de *TDBs* mediante el método de Bloch (1997) al error $2\alpha / \left(J(J-1) \right)$, es decir aplicando el método de Bonferroni (u otro método de comparaciones múltiples).

3.5. Comparación de los coeficientes kappa promedios

El coeficiente kappa promedio es un parámetro que también permite comparar el rendimiento de dos o más *TDBs*. En esta sección se estudian, en primer lugar, el test de hipótesis para comparar los coeficientes kappa promedio de dos *TDBs* bajo un diseño apareado, y en segundo lugar se generaliza el método al caso de múltiples *TDBs*.

3.5.1. Comparación de dos coeficientes kappa promedios

Considérense dos *TDBs*, test 1 y test 2, que se comparan con respecto a un mismo *GE*, dando lugar a la Tabla 3.1 y cuyo modelo teórico se muestra en la Tabla 3.2. En primer lugar se estudia la comparación de los dos coeficientes kappa promedios cuando el clínico considera que $L' > L$ (y por tanto $0 < c < 0.5$). El contraste de hipótesis para comparar los dos coeficientes kappa promedios es

$$H_0 : \kappa_{11} = \kappa_{21} \quad \text{vs} \quad H_1 : \kappa_{11} \neq \kappa_{21},$$

donde κ_{11} y κ_{21} son los coeficientes kappa promedios para cada uno de los *TDBs* cuando $L' > L$. En términos de las probabilidades de la Tabla 3.2, κ_{11} se escribe como

$$\kappa_{11} = \frac{2\kappa_1(0)\kappa_1(1)}{\kappa_1(0)-\kappa_1(1)} \ln\left\{\frac{\kappa_1(0)+\kappa_1(1)}{2\kappa_1(1)}\right\} =$$

$$2\left(\frac{\sum_{j=0}^{1}\left(p_{0j}+q_{0j}\right)}{\frac{1}{p}\sum_{j=0}^{1}p_{1j}-\sum_{j=0}^{1}\left(p_{1j}+q_{1j}\right)}-\frac{\sum_{j=0}^{1}\left(p_{1j}+q_{1j}\right)}{\frac{1}{q}\sum_{j=0}^{1}q_{0j}-\sum_{j=0}^{1}\left(p_{0j}+q_{0j}\right)}\right)^{-1} \times \quad (3.68)$$

$$\ln\left\{\frac{1}{2}\left(\frac{\left(\sum_{j=0}^{1}\left(p_{0j}+q_{0j}\right)\right)\left(\frac{1}{q}\sum_{j=0}^{1}q_{0j}-\sum_{j=0}^{1}\left(p_{0j}+q_{0j}\right)\right)}{\left(\sum_{j=0}^{1}\left(p_{1j}+q_{1j}\right)\right)\left(\frac{1}{p}\sum_{j=0}^{1}p_{1j}-\sum_{j=0}^{1}\left(p_{1j}+q_{1j}\right)\right)}+1\right)\right\},$$

cuando $p \neq Q_1$, y

$$\kappa_{11} = Y_1 = \frac{1}{p}\sum_{j=0}^{1}p_{1j} + \frac{1}{q}\sum_{j=0}^{1}q_{0j} - 1 \quad (3.69)$$

cuando $p = Q_1$. Con respecto a κ_{21}, este se escribe como

$$\kappa_{21} = \frac{2\kappa_2(0)\kappa_2(1)}{\kappa_2(0)-\kappa_2(1)} \ln\left\{\frac{\kappa_2(0)+\kappa_2(1)}{2\kappa_2(1)}\right\} =$$

$$2\left(\frac{\sum_{i=0}^{1}(p_{i0}+q_{i0})}{\frac{1}{p}\sum_{i=0}^{1}p_{i1}-\sum_{i=0}^{1}(p_{i1}+q_{i1})} - \frac{\sum_{i=0}^{1}(p_{i1}+q_{i1})}{\frac{1}{q}\sum_{i=0}^{1}q_{i0}-\sum_{i=0}^{1}(p_{i0}+q_{i0})}\right)^{-1} \times \qquad (3.70)$$

$$\ln\left\{\frac{1}{2}\left(\frac{\left(\sum_{i=0}^{1}(p_{i0}+q_{i0})\right)\left(\frac{1}{q}\sum_{i=0}^{1}q_{i0}-\sum_{i=0}^{1}(p_{i0}+q_{i0})\right)}{\left(\sum_{i=0}^{1}(p_{i1}+q_{i1})\right)\left(\frac{1}{p}\sum_{i=0}^{1}p_{i1}-\sum_{i=0}^{1}(p_{i1}+q_{i1})\right)}+1\right)\right\},$$

cuando $p \neq Q_2$, y como

$$\kappa_{21} = Y_2 = \frac{1}{p}\sum_{i=0}^{1}p_{i1} + \frac{1}{q}\sum_{i=0}^{1}q_{i0} - 1 \ , \qquad (3.71)$$

cuando $p = Q_2$. Sustituyendo en las ecuaciones anteriores cada parámetro por la expresión de su correspondiente estimador, se obtiene que

$$\hat{\kappa}_{11} = \frac{2\{(s_{10}+s_{11})(r_{00}+r_{01})-(s_{00}+s_{01})(r_{10}+r_{11})\}}{n\left(\sum_{j=0}^{1}(s_{0j}-r_{1j})\right)} \times \ln\left\{\frac{1}{2}\left(\frac{s\sum_{j=0}^{1}(s_{0j}+r_{0j})}{r\sum_{j=0}^{1}(s_{1j}+r_{1j})}+1\right)\right\} \qquad (3.72)$$

cuando $\hat{p} \neq \hat{Q}_1$, es decir cuando $s_{01}+s_{00} \neq r_{10}+r_{11}$, y

$$\hat{\kappa}_{11} = \frac{r(s_{10}+s_{11})+s(r_{00}+r_{01})-sr}{sr} \qquad (3.73)$$

cuando $\hat{p} = \hat{Q}_1$, es decir cuando $s_{01} + s_{00} = r_{10} + r_{11}$. Con respecto al estimador de κ_{21}, sus expresiones son

$$\hat{\kappa}_{21} = \frac{2\left\{\left(s_{01}+s_{11}\right)\left(r_{00}+r_{10}\right)-\left(s_{00}+s_{10}\right)\left(r_{01}+r_{11}\right)\right\}}{n\left(\sum_{i=0}^{1}\left(s_{i0}-r_{i1}\right)\right)} \times \ln\left\{\frac{1}{2}\left(\frac{s\sum_{i=0}^{1}\left(s_{i0}+r_{i0}\right)}{r\sum_{i=0}^{1}\left(s_{i1}+r_{i1}\right)}+1\right)\right\} \tag{3.74}$$

cuando $\hat{p} \neq \hat{Q}_2$ $\left(s_{00}+s_{10} \neq r_{01}+r_{11}\right)$, y

$$\hat{\kappa}_{21} = \frac{r\left(s_{01}+s_{11}\right)+s\left(r_{00}+r_{10}\right)-sr}{sr} \tag{3.75}$$

cuando $\hat{p} = \hat{Q}_2$ $\left(s_{00}+s_{10} = r_{01}+r_{11}\right)$. Aplicando el método delta, la estimación de la matriz de varianzas-covarianzas asintóticas es

$$\sum\nolimits_{\hat{\boldsymbol{\kappa}}_1} = \left(\frac{\partial \boldsymbol{\kappa}_1}{\partial \boldsymbol{\pi}}\right) \sum\nolimits_{\hat{\boldsymbol{\pi}}} \left(\frac{\partial \boldsymbol{\kappa}_1}{\partial \boldsymbol{\pi}}\right)^T \tag{3.76}$$

con $\boldsymbol{\kappa}_1 = \left(\kappa_{11}, \kappa_{21}\right)^T$, $\boldsymbol{\pi} = \left(p_{11}, p_{10}, p_{01}, p_{00}, q_{11}, q_{10}, q_{01}, q_{00}\right)^T$ y $\sum_{\hat{\boldsymbol{\pi}}}$ la matriz de varianzas-covarianzas definida como

$$\sum\nolimits_{\hat{\boldsymbol{\pi}}} = \frac{\mathrm{Diag}\left(\boldsymbol{\pi}\right) - \boldsymbol{\pi}\boldsymbol{\pi}^T}{n}. \tag{3.77}$$

Sustituyendo en la expresión (3.76) cada parámetro por su estimador, se obtienen las expresiones de las varianzas-covarianzas asintóticas estimadas

de $\boldsymbol{\kappa}_1$. Estas expresiones no se presentan por ser muy extensas y complicadas (su cálculo se realiza con el programa en R elaborado para resolver este test de hipótesis). Finalmente, el estadístico para contrastar la igualdad de los coeficientes kappa promedios cuando $L' > L$ $(0 < c < 0.5)$ es

$$z_{\exp} = \frac{\left|\hat{\kappa}_{11} - \hat{\kappa}_{21}\right|}{\sqrt{\hat{V}ar\left(\hat{\kappa}_{11}\right) + \hat{V}ar\left(\hat{\kappa}_{21}\right) - 2\hat{C}ov\left(\hat{\kappa}_{11}, \hat{\kappa}_{21}\right)}} \xrightarrow[n\to\infty]{} N(0,1). \quad (3.78)$$

Asimismo, un intervalo de confianza asintótico para la diferencia de los dos coeficientes kappa promedios es

$$\kappa_{11} - \kappa_{21} \in \hat{\kappa}_{11} - \hat{\kappa}_{21} \pm z_{1-\alpha/2}\sqrt{\hat{V}ar\left(\hat{\kappa}_{11}\right) + \hat{V}ar\left(\hat{\kappa}_{21}\right) - 2\hat{C}ov\left(\hat{\kappa}_{11}, \hat{\kappa}_{21}\right)}. \quad (3.79)$$

Si el clínico considera que $L > L'$, y por tanto que $0.5 < c < 1$, el contraste de hipótesis para comparar los dos coeficientes kappa promedios es

$$H_0 : \kappa_{12} = \kappa_{22} \quad \text{vs} \quad H_1 : \kappa_{12} \neq \kappa_{22},$$

donde κ_{i2} es el coeficiente kappa promedio del i-ésimo *TDB* cuando $L > L'$. En términos de las probabilidades de la Tabla 3.2, κ_{12} se escribe como

$$\kappa_{12} = \frac{2\kappa_1(0)\kappa_1(1)}{\kappa_1(0)-\kappa_1(1)} \ln\left\{\frac{2\kappa_1(0)}{\kappa_1(0)+\kappa_1(1)}\right\} =$$

$$2\left(\frac{\sum_{j=0}^{1}\left(p_{0j}+q_{0j}\right)}{\frac{1}{p}\sum_{j=0}^{1}p_{1j}-\sum_{j=0}^{1}\left(p_{1j}+q_{1j}\right)}-\frac{\sum_{j=0}^{1}\left(p_{1j}+q_{1j}\right)}{\frac{1}{q}\sum_{j=0}^{1}q_{0j}-\sum_{j=0}^{1}\left(p_{0j}+q_{0j}\right)}\right)^{-1}\times \quad (3.80)$$

$$\ln\left\{2\left[1+\frac{\left(\frac{1}{p}\sum_{j=0}^{1}p_{1j}-\sum_{j=0}^{1}\left(p_{1j}+q_{1j}\right)\right)\left(\sum_{j=0}^{1}\left(p_{1j}+q_{1j}\right)\right)}{\left(\sum_{j=0}^{1}\left(p_{0j}+q_{0j}\right)\right)\left(\frac{1}{q}\sum_{j=0}^{1}q_{0j}-\sum_{j=0}^{1}\left(p_{0j}+q_{0j}\right)\right)}\right]^{-1}\right\}$$

cuando $p \neq Q_1$, y

$$\kappa_{12} = Y_1 = \frac{1}{p}\sum_{j=0}^{1}p_{1j}+\frac{1}{q}\sum_{j=0}^{1}q_{0j}-1 \quad (3.81)$$

cuando $p = Q_1$. Con respecto a κ_{22}, este se escribe como

$$\kappa_{22} = \frac{2\kappa_2(0)\kappa_2(1)}{\kappa_2(0)-\kappa_2(1)} \ln\left\{\frac{2\kappa_2(0)}{\kappa_2(0)+\kappa_2(1)}\right\} =$$

$$2\left(\frac{\sum_{i=0}^{1}(p_{i0}+q_{i0})}{\frac{1}{p}\sum_{i=0}^{1}p_{i1}-\sum_{i=0}^{1}(p_{i1}+q_{i1})} - \frac{\sum_{i=0}^{1}(p_{i1}+q_{i1})}{\frac{1}{q}\sum_{i=0}^{1}q_{i0}-\sum_{i=0}^{1}(p_{i0}+q_{i0})}\right)^{-1} \times \qquad (3.82)$$

$$\ln\left\{2\left(1+\frac{\left(\frac{1}{p}\sum_{i=0}^{1}p_{i1}-\sum_{i=0}^{1}(p_{i1}+q_{i1})\right)\left(\sum_{i=0}^{1}(p_{i1}+q_{i1})\right)}{\left(\sum_{i=0}^{1}(p_{i0}+q_{i0})\right)\left(\frac{1}{q}\sum_{i=0}^{1}q_{i0}-\sum_{i=0}^{1}(p_{i0}+q_{i0})\right)}\right)^{-1}\right\},$$

cuando $p \neq Q_2$, y como

$$\kappa_{22} = Y_2 = \frac{1}{p}\sum_{j=0}^{1}p_{1j} + \frac{1}{q}\sum_{j=0}^{1}q_{0j} - 1 \qquad (3.83)$$

cuando $p = Q_2$. Sustituyendo en las ecuaciones anteriores cada parámetro por la expresión de su correspondiente estimador, se obtiene que

$$\hat{\kappa}_{12} = \frac{2\left\{(s_{10}+s_{11})(r_{00}+r_{01})-(s_{00}+s_{01})(r_{10}+r_{11})\right\}}{n\left(\sum_{j=0}^{1}(s_{0j}-r_{1j})\right)} \times \ln\left\{2\frac{s\sum_{j=0}^{1}(s_{0j}+r_{0j})}{s\sum_{j=0}^{1}(s_{0j}+r_{0j})+r\sum_{j=0}^{1}(s_{1j}+r_{1j})}\right\} \qquad (3.84)$$

cuando $\hat{p} \neq \hat{Q}_1$, es decir cuando $s_{01}+s_{00} \neq r_{10}+r_{11}$, y

$$\hat{\kappa}_{12} = \frac{r(s_{10}+s_{11})+s(r_{00}+r_{01})-sr}{sr} \qquad (3.85)$$

cuando $\hat{p}=\hat{Q}_1$ $\left(s_{01}+s_{00}=r_{10}+r_{11}\right)$. Con respecto al estimador de κ_{22}, sus expresiones son

$$\hat{\kappa}_{22}=\frac{2\left\{\left(s_{01}+s_{11}\right)\left(r_{00}+r_{10}\right)-\left(s_{00}+s_{10}\right)\left(r_{01}+r_{11}\right)\right\}}{n\left(\sum_{i=0}^{1}\left(s_{i0}-r_{i1}\right)\right)}\times\ln\left\{2\frac{s\sum_{i=0}^{1}\left(s_{i0}+r_{i0}\right)}{s\sum_{i=0}^{1}\left(s_{i0}+r_{i0}\right)+r\sum_{i=0}^{1}\left(s_{i1}+r_{i1}\right)}\right\} \tag{3.86}$$

cuando $\hat{p}\neq\hat{Q}_2$ $\left(s_{00}+s_{10}\neq r_{01}+r_{11}\right)$, y

$$\hat{\kappa}_{22}=\frac{r\left(s_{01}+s_{11}\right)+s\left(r_{00}+r_{10}\right)-sr}{sr} \tag{3.87}$$

cuando $\hat{p}=\hat{Q}_2$ $\left(s_{00}+s_{10}=r_{01}+r_{11}\right)$.

La matriz de varianzas-covarianzas asintótica se estima de forma similar al acaso anterior, esto es

$$\sum_{\hat{\boldsymbol{\kappa}}_2}=\left(\frac{\partial\boldsymbol{\kappa}_2}{\partial\boldsymbol{\pi}}\right)\sum_{\hat{\boldsymbol{\pi}}}\left(\frac{\partial\boldsymbol{\kappa}_2}{\partial\boldsymbol{\pi}}\right)^T, \tag{3.88}$$

con $\boldsymbol{\kappa}_2=\left(\kappa_{12},\kappa_{22}\right)^T$ y $\sum_{\hat{\boldsymbol{\pi}}}$ la matriz de varianzas-covarianzas definida en la ecuación (3.77). Sustituyendo en la ecuación (3.88) cada parámetro por su estimador y realizando las operaciones algebraicas se obtienen las expresiones de los estimadores de las varianzas-covarianzas, que al igual que el caso anterior, estas son unas expresiones extensas y complicadas y

por tanto no se muestran. Finalmente, el estadístico para contrastar los coeficientes kappa promedios cuando $L > L'$ $(0.5 < c < 1)$ es

$$z_{\exp} = \frac{\left|\hat{\kappa}_{12} - \hat{\kappa}_{22}\right|}{\sqrt{\hat{Var}\left(\hat{\kappa}_{12}\right) + \hat{Var}\left(\hat{\kappa}_{22}\right) - 2\hat{Cov}\left(\hat{\kappa}_{12}, \hat{\kappa}_{22}\right)}} \xrightarrow[n\to\infty]{} N(0,1), \quad (3.89)$$

y un intervalo de confianza asintótico para la diferencia de los dos coeficientes kappa promedios es

$$\kappa_{12} - \kappa_{22} \in \hat{\kappa}_{12} - \hat{\kappa}_{22} \pm z_{1-\alpha/2}\sqrt{\hat{Var}\left(\hat{\kappa}_{12}\right) + \hat{Var}\left(\hat{\kappa}_{22}\right) - 2\hat{Cov}\left(\hat{\kappa}_{12}, \hat{\kappa}_{22}\right)}. \quad (3.90)$$

La comparación de los coeficientes kappa promedios también se puede realizar utilizando alguna transformación, como por ejemplo la transformación del logaritmo neperiano, que es una transformación muy común. El problema se resuelve de forma similar al caso anterior, quedando los estimadores puntuales afectados por dicha transformación, siendo la matriz de varianzas-covarianzas

$$\sum_{\ln(\hat{\boldsymbol{\kappa}}_i)} = \left(\frac{\partial \ln(\boldsymbol{\kappa}_i)}{\partial \boldsymbol{\pi}}\right) \sum_{\hat{\boldsymbol{\pi}}} \left(\frac{\partial \ln(\boldsymbol{\kappa}_i)}{\partial \boldsymbol{\pi}}\right)^T, \quad (3.91)$$

y su estimación se obtiene sustituyendo en esta ecuación cada parámetro por su estimador, y donde $\ln(\boldsymbol{\kappa}_i) = \left(\ln(\kappa_{1i}), \ln(\kappa_{2i})\right)^T$ es $\ln(\boldsymbol{\kappa}_1)$ cuando $L' > L$ y $\ln(\boldsymbol{\kappa}_2)$ cuando $L > L'$. El correspondiente test de hipótesis es

$$H_0 : \ln(\kappa_{1i}) = \ln(\kappa_{2i}) \quad \text{vs} \quad H_1 : \ln(\kappa_{1i}) \neq \ln(\kappa_{2i}),$$

y el estadístico de contraste es

$$z_{\exp} = \frac{\left|\ln(\hat{\kappa}_{1i}) - \ln(\hat{\kappa}_{2i})\right|}{\sqrt{\hat{V}ar\left[\ln(\hat{\kappa}_{1i})\right] + \hat{V}ar\left[\ln(\hat{\kappa}_{2i})\right] - 2\hat{C}ov\left[\ln(\hat{\kappa}_{1i}), \ln(\hat{\kappa}_{2i})\right]}} \xrightarrow[n\to\infty]{} N(0,1) \ . \quad (3.92)$$

Finalmente, un intervalo de confianza para el cociente de los dos coeficientes kappa promedios es

$$\frac{\kappa_{1i}}{\kappa_{2i}} \in \exp\left\{\ln\left(\frac{\hat{\kappa}_{1i}}{\hat{\kappa}_{2i}}\right) \pm z_{1-\alpha/2}\sqrt{\hat{V}ar\left[\ln(\hat{\kappa}_{1i})\right] + \hat{V}ar\left[\ln(\hat{\kappa}_{2i})\right] - 2\hat{C}ov\left[\ln(\hat{\kappa}_{1i}), \ln(\hat{\kappa}_{2i})\right]}\right\}. \quad (3.93)$$

3.5.2. Comparación de múltiples coeficientes kappa promedios

Considérense la misma situación dada en las Secciones 3.3.3 y 3.4.2, es decir, se aplican J *TDBs* y el *GE* a todos los *n* individuos de una muestra

aleatoria. Cuando $L' > L \quad (0 < c < 0.5)$, la expresión del coeficiente kappa ponderado para el j-ésimo *TDB* es

$$\kappa_{j1} = \begin{cases} \dfrac{2\kappa_j(0)\kappa_j(1)}{\kappa_j(0)-\kappa_j(1)} \ln\left\{\dfrac{\kappa_j(0)+\kappa_j(1)}{2\kappa_j(1)}\right\}, & p \neq Q_j \\ Y_j, & p = Q_j \end{cases} \tag{3.94}$$

y cuando $L > L' \quad (0.5 < c < 1)$, su expresión es

$$\kappa_{j2} = \begin{cases} \dfrac{2\kappa_j(0)\kappa_j(1)}{\kappa_j(0)-\kappa_j(1)} log\left[\dfrac{2\kappa_j(0)}{\kappa_j(0)+\kappa_j(1)}\right], & p \neq Q_j \\ Y_j, & p = Q_j, \end{cases} \tag{3.95}$$

con $\kappa_j(0) = \dfrac{Sp_j - (1-Q_j)}{Q_j}$, $\kappa_j(1) = \dfrac{Se_j - Q_j}{1-Q_j}$ y

$Q_j = pSe_j + (1-p)(1-Sp_j)$. Sea

$$p = P(D=1) = \sum_{i_1,\ldots,i_J=0}^{1} p_{i_1,\ldots,i_J} \tag{3.96}$$

la prevalencia de la enfermedad y

$$q = 1 - p = P(D=0) = \sum_{i_1,\ldots,i_J=0}^{1} q_{i_1,\ldots,i_J} \tag{3.97}$$

la probabilidad de que un individuo no tenga la enfermedad. La sensibilidad y la especificidad del j-ésimo *TDB* se escriben como

$$Se_j = P\left(T_j = 1 \middle| D = 1\right) = \frac{\sum_{i_1,\ldots,i_J=0}^{1} p_{i_1,\ldots,i_J}}{\sum_{i_1,\ldots,i_J=0}^{1} p_{i_1,\ldots,i_J}} \quad (3.98)$$

y

$$Sp_j = P\left(T_j = 0 \middle| D = 0\right) = \frac{\sum_{i_1,\ldots,i_J=0}^{1} q_{i_1,\ldots,i_J}}{\sum_{i_1,\ldots,i_J=0}^{1} q_{i_1,\ldots,i_J}} \quad (3.99)$$

respectivamente. Sustituyendo en las expresiones (3.94) y (3.95) cada parámetro por su expresión se obtiene que

$$\kappa_{j1} = \begin{cases} a \times \ln\left[\frac{1}{2}\left(b_1 + 1\right)\right], & p \neq Q_j \\ Y_j, & p = Q_j \end{cases} \quad (3.100)$$

y

$$\kappa_{j2} = \begin{cases} a \times log\left[\frac{2}{\left(1 + b_2\right)}\right], & p \neq Q_j \\ Y_j, & p = Q_j, \end{cases} \quad (3.101)$$

donde

$$a = \frac{2}{\dfrac{\left(p - \sum\limits_{\substack{i_1,\ldots,i_J=0 \\ i_j=1}}^{1} p_{i_1,\ldots,i_J} + \sum\limits_{\substack{i_1,\ldots,i_J=0 \\ i_j=0}}^{1} q_{i_1,\ldots,i_J}\right)}{\left(-q + \dfrac{\sum\limits_{\substack{i_1,\ldots,i_J=0 \\ i_j=1}}^{1} p_{i_1,\ldots,i_J}}{\sum\limits_{i_1,\ldots,i_J=0}^{1} p_{i_1,\ldots,i_J}} - \sum\limits_{\substack{i_1,\ldots,i_J=0 \\ i_j=1}}^{1} p_{i_1,\ldots,i_J} + \sum\limits_{\substack{i_1,\ldots,i_J=0 \\ i_j=0}}^{1} q_{i_1,\ldots,i_J}\right)} - \dfrac{\left(q + \sum\limits_{\substack{i_1,\ldots,i_J=0 \\ i_j=1}}^{1} p_{i_1,\ldots,i_J} - \sum\limits_{\substack{i_1,\ldots,i_J=0 \\ i_j=0}}^{1} q_{i_1,\ldots,i_J}\right)}{\left(-p + \dfrac{\sum\limits_{\substack{i_1,\ldots,i_J=0 \\ i_j=0}}^{1} q_{i_1,\ldots,i_J}}{\sum\limits_{i_1,\ldots,i_J=0}^{1} q_{i_1,\ldots,i_J}} + \sum\limits_{\substack{i_1,\ldots,i_J=0 \\ i_j=1}}^{1} p_{i_1,\ldots,i_J} - \sum\limits_{\substack{i_1,\ldots,i_J=0 \\ i_j=0}}^{1} q_{i_1,\ldots,i_J}\right)}, \tag{3.102}$$

$$b_1 = \left(\frac{\left(p - \sum\limits_{\substack{i_1,\ldots,i_J=0 \\ i_j=1}}^{1} p_{i_1,\ldots,i_J} + \sum\limits_{\substack{i_1,\ldots,i_J=0 \\ i_j=0}}^{1} q_{i_1,\ldots,i_J}\right)\left(-p + \dfrac{\sum\limits_{\substack{i_1,\ldots,i_J=0 \\ i_j=0}}^{1} q_{i_1,\ldots,i_J}}{\sum\limits_{i_1,\ldots,i_J=0}^{1} q_{i_1,\ldots,i_J}} + \sum\limits_{\substack{i_1,\ldots,i_J=0 \\ i_j=1}}^{1} p_{i_1,\ldots,i_J} - \sum\limits_{\substack{i_1,\ldots,i_J=0 \\ i_j=0}}^{1} q_{i_1,\ldots,i_J}\right)}{\left(q + \sum\limits_{\substack{i_1,\ldots,i_J=0 \\ i_j=1}}^{1} p_{i_1,\ldots,i_J} - \sum\limits_{\substack{i_1,\ldots,i_J=0 \\ i_j=0}}^{1} q_{i_1,\ldots,i_J}\right)\left(-q + \dfrac{\sum\limits_{\substack{i_1,\ldots,i_J=0 \\ i_j=1}}^{1} p_{i_1,\ldots,i_J}}{\sum\limits_{i_1,\ldots,i_J=0}^{1} p_{i_1,\ldots,i_J}} - \sum\limits_{\substack{i_1,\ldots,i_J=0 \\ i_j=1}}^{1} p_{i_1,\ldots,i_J} + \sum\limits_{\substack{i_1,\ldots,i_J=0 \\ i_j=0}}^{1} q_{i_1,\ldots,i_J}\right)}\right), \tag{3.103}$$

y

$$b_2 = \frac{\left(q + \sum_{\substack{i_1,\ldots,i_J=0 \\ i_j=1}}^{1} p_{i_1,\ldots,i_J} - \sum_{\substack{i_1,\ldots,i_J=0 \\ i_j=0}}^{1} q_{i_1,\ldots,i_J} \right)\left(-q + \frac{\sum_{\substack{i_1,\ldots,i_J=0 \\ i_j=1}}^{1} p_{i_1,\ldots,i_J}}{\sum_{i_1,\ldots,i_J=0}^{1} p_{i_1,\ldots,i_J}} - \sum_{\substack{i_1,\ldots,i_J=0 \\ i_j=1}}^{1} p_{i_1,\ldots,i_J} + \sum_{\substack{i_1,\ldots,i_J=0 \\ i_j=0}}^{1} q_{i_1,\ldots,i_J} \right)}{\left(p - \sum_{\substack{i_1,\ldots,i_J=0 \\ i_j=1}}^{1} p_{i_1,\ldots,i_J} + \sum_{\substack{i_1,\ldots,i_J=0 \\ i_j=0}}^{1} q_{i_1,\ldots,i_J} \right)\left(-p + \frac{\sum_{\substack{i_1,\ldots,i_J=0 \\ i_j=0}}^{1} q_{i_1,\ldots,i_J}}{\sum_{i_1,\ldots,i_J=0}^{1} q_{i_1,\ldots,i_J}} + \sum_{\substack{i_1,\ldots,i_J=0 \\ i_j=1}}^{1} p_{i_1,\ldots,i_J} - \sum_{\substack{i_1,\ldots,i_J=0 \\ i_j=0}}^{1} q_{i_1,\ldots,i_J} \right)}$$

. (3.104)

Como los estimadores máximo verosímiles las probabilidades $p_{i_1,\ldots,i_J}$ y $q_{i_1,\ldots,i_J}$ son

$$\hat{p}_{i_1,\ldots,i_J} = \frac{s_{i_1,\ldots,i_J}}{n} \quad \text{y} \quad \hat{q}_{i_1,\ldots,i_J} = \frac{r_{i_1,\ldots,i_J}}{n}, \tag{3.105}$$

con $i_1,\ldots,i_J = 0,1$, el estimador de cada coeficiente kappa promedio se obtiene sustituyendo en las expresiones (3.100) y (3.101) cada parámetro $p_{i_1,\ldots,i_J}$ y $q_{i_1,\ldots,i_J}$ por su correspondiente estimador.

Sea $\boldsymbol{\kappa}_i = \left(\kappa_{1i}, \kappa_{2i}, \ldots, \kappa_{Ji}\right)^T$ el vector de coeficientes kappa promedios y $\hat{\boldsymbol{\kappa}}_i = \left(\hat{\kappa}_{1i}, \hat{\kappa}_{2i}, \ldots, \hat{\kappa}_{Ji}\right)^T$ su estimador, siendo $i = 1$ cuando $L' > L$ e $i = 2$ cuando $L > L'$. La matriz de varianzas-covarianzas asintóticas del vector $\hat{\boldsymbol{\kappa}}_i$ se puede estimar aplicando el método delta. De esta forma,

$$\sum_{\hat{\boldsymbol{\kappa}}_i} = \left(\frac{\partial \boldsymbol{\kappa}_i}{\partial \boldsymbol{\pi}} \right) \sum_{\hat{\boldsymbol{\pi}}} \left(\frac{\partial \boldsymbol{\kappa}_i}{\partial \boldsymbol{\pi}} \right)^T \tag{3.106}$$

Realizando las operaciones algebraicas y sustituyendo en esta expresión cada parámetro por su estimador, se obtiene la matriz de varianzas-covarianzas asintóticas estimadas $\hat{\sum}_{\hat{\boldsymbol{\kappa}}_i}$. El test de hipótesis para la contrastar la igualdad de los coeficientes kappa promedios es

$$H_0 : \kappa_{1i} = \kappa_{2i} = ... = \kappa_{Ji} \quad \text{vs} \quad H_1 : \text{al menos una igualdad no es cierta},$$

siendo $i = 1$ cuando $L' > L$ e $i = 2$ cuando $L > L'$. Este test de hipótesis es equivalente a contrastar

$$H_0 : \boldsymbol{\varphi} \boldsymbol{\kappa}_i = \mathbf{0} \quad \text{vs} \quad H_1 : \boldsymbol{\varphi} \boldsymbol{\kappa}_i \neq \mathbf{0},$$

con $i = 1,2$, y siendo $\boldsymbol{\kappa}_i = \left(\kappa_{1i}, \kappa_{2i}, ..., \kappa_{Ji} \right)^T$ y $\boldsymbol{\varphi}$ una matriz de rango completo cuya dimensión es $\left(J - 1 \right) \times J$. Por ejemplo, para tres *TDBs* la matriz $\boldsymbol{\varphi}$ es

$$\boldsymbol{\varphi} = \begin{pmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \end{pmatrix}.$$

Aplicando el teorema central del límite multivariante se verifica que

$$\sqrt{n} \left(\hat{\boldsymbol{\kappa}}_i - \boldsymbol{\kappa}_i \right) \xrightarrow[n \to \infty]{} N_{J-1} \left(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\kappa}_i} \right), \tag{3.107}$$

por lo que el estadístico

$$Q^2_{\text{exp}} = \hat{\boldsymbol{\kappa}}_i^{\,T} \boldsymbol{\varphi}^T \left(\boldsymbol{\varphi} \hat{\Sigma}_{\hat{\boldsymbol{\kappa}}_i} \boldsymbol{\varphi}^T \right)^{-1} \boldsymbol{\varphi} \hat{\boldsymbol{\kappa}}_i \tag{3.108}$$

se distribuye según una distribución T^2 de Hotelling de dimensión $J-1$ y n grados de libertad, donde $J-1$ es la dimensión del vector $\boldsymbol{\varphi}\hat{\boldsymbol{\kappa}}_i$. Para n grande, el estadístico Q^2_{exp} se distribuye según una distribución chi-cuadrado central con $J-1$ grados de libertad cuando la hipótesis nula es cierta, es decir,

$$Q^2_{\text{exp}} = \hat{\boldsymbol{\kappa}}_i^{\,T} \boldsymbol{\varphi}^T \left(\boldsymbol{\varphi} \hat{\Sigma}_{\hat{\boldsymbol{\kappa}}_i} \boldsymbol{\varphi}^T \right)^{-1} \boldsymbol{\varphi} \hat{\boldsymbol{\kappa}}_i \xrightarrow[n\to\infty]{} \chi^2_{J-1}. \tag{3.109}$$

El procedimiento final sería muy similar al utilizado por Roldán Nofuentes y Luna del Castillo (2010) para comparar simultáneamente los valores predictivos de múltiples *TDBs*: 1) resolver el test global (ecuación (3.109)) al error α; 2) si el test global es no significativo a ese error, entonces no se rechaza la homogeneidad de los J coeficientes kappa promedios, y si el test es significativo entonces la investigación de las causas de la significación se realiza comparando las parejas de coeficientes kappa promedios utilizando los resultados de la Sección anterior y

penalizando el nivel de significación mediante algún método de comparaciones múltiples (por ejemplo, Holm, Hocheberg y Bonferroni).

Finalmente, al igual que para dos *TDBs*, la comparación de múltiples coeficientes kappa promedios se puede realizar utilizando la transformación logarítmica, siendo el procedimiento similar al utilizado para el caso sin transformación.

3.5.3. Experimentos de simulación

En la práctica clínica lo más frecuente es la comparación de dos *TDBs*, por lo que se han realizado experimentos de simulación para estudiar el comportamiento asintótico de los test de hipótesis de comparación de los coeficientes kappa promedios de dos *TDBs*. Por tanto se han estudiado los errores de tipo I y las potencias de los test $H_0 : \kappa_{1k} = \kappa_{2k}$ y $H_0 : \ln(\kappa_{1k}) = \ln(\kappa_{2k})$, para $k = 1,2$. Para ello se han generado 5000 muestras aleatorias de distribuciones multinomiales con tamaños 100, 200, 300, 400, 500, 1000 y 2000. Las probabilidades de las distribuciones multinomiales se han calculado utilizando el modelo de dependencia condicional de Vacek (1985), esto es,

$$\begin{aligned} p_{ij} &= P\left(T_1 = i, T_2 = j \middle| D = 1\right) = P\left(T_1 = i \middle| D = 1\right) \times P\left(T_2 = j \middle| D = 1\right) + \delta_{ij}\varepsilon_1 \\ q_{ij} &= P\left(T_1 = i, T_2 = j \middle| D = 0\right) = P\left(T_1 = i \middle| D = 0\right) \times P\left(T_2 = j \middle| D = 0\right) + \delta_{ij}\varepsilon_0, \end{aligned} \tag{3.110}$$

donde $\delta_{ij} = 1$ si $i = j$ y $\delta_{ij} = -1$ si $i \neq j$, ε_1 es la covarianza entre los dos *TDBs* cuando $D = 1$ y ε_0 es la covarianza entre los dos *TDBs* cuando $D = 0$. Vacek (1985) ha demostrado que

$$\varepsilon_1 \leq \begin{cases} Se_1\left(1 - Se_2\right) & \text{si } Se_2 > Se_1 \\ Se_2\left(1 - Se_1\right) & \text{si } Se_1 > Se_2 \end{cases} \tag{3.111}$$

y que

$$\varepsilon_0 \leq \begin{cases} Sp_1\left(1 - Sp_2\right) & \text{si } Sp_2 > Sp_1 \\ Sp_2\left(1 - Sp_1\right) & \text{si } Sp_1 > Sp_2. \end{cases} \tag{3.112}$$

Si $\varepsilon_1 = \varepsilon_0 = 0$ entonces los dos *TDBs* son condicionalmente independientes dado el estado de enfermedad. En la práctica la suposición de independencia condicional es poco realista por lo que suele ocurrir que $\varepsilon_1 > 0$ y/o $\varepsilon_0 > 0$.

Los experimentos de simulación se han diseñado a partir de las ecuaciones de los coeficientes kappa promedios de los dos *TDBs*, esto es

$$\kappa_{i1} = \frac{2\kappa_i(0)\kappa_i(1)}{\kappa_i(0) - \kappa_i(1)} \ln\left\{\frac{\kappa_i(0) + \kappa_i(1)}{2\kappa_i(1)}\right\} \tag{3.113}$$

y

$$\kappa_{i2} = \frac{2\kappa_i(0)\kappa_i(1)}{\kappa_i(0)-\kappa_i(1)} \ln\left\{\frac{2\kappa_i(0)}{\kappa_i(0)+\kappa_i(1)}\right\} \tag{3.114}$$

Como prevalencia de la enfermedad se han considerado los valores 10%, 30% y 50% y para los coeficientes kappa promedios se han considerado los valores 0.2, 0.4, 0.6 y 0.8. Una vez fijados los valores de la prevalencia y del coeficiente kappa promedio, utilizando el método de Newton-Raphson se ha resuelto el sistema formado por las ecuaciones (3.113) y (3.114) para así obtener los valores de $\kappa_i(0)$ y $\kappa_i(1)$. Aquí se han considerado aquellos valores cuyas soluciones se han encontrado entre 0 y 1. Por último, para obtener los valores de la sensibilidad y especificidad de cada *TDB* (Se_i y Sp_i) se ha resuelto el sistema formado por las ecuaciones

$$\kappa_i(0) = \frac{Sp_i - (1-Q_i)}{Q_i}, \tag{3.115}$$

y

$$\kappa_i(1) = \frac{Se_i - Q_i}{1-Q_i}. \tag{3.116}$$

Una vez obtenidos los valores de Se_i y Sp_i, a partir de las ecuaciones (3.111) y (3.112)se han calculado los valores máximos de las covarianzas ε_1

y ε_0. Finalmente, las probabilidades de las distribuciones multinomiales se han calculado a partir de las expresiones (3.110). Asimismo, las muestras se han generado de tal forma que en todas ellas los índices de Youden estimados han sido mayores que 0 y se han podido estimar todos los parámetros y sus varianzas-covarianzas. Para todo el estudio se ha tomado como error nominal $\alpha = 5\%$. A continuación se analizan los resultados obtenidos.

3.5.3.1. Errores tipo I

En las Tablas 3.3 a 3.6 se muestran los resultados para los errores tipo I de los test de hipótesis $H_0 : \kappa_{11} = \kappa_{21}$ y $H_0 : \ln(\kappa_{11}) = \ln(\kappa_{21})$ (es decir, cuando se comparan los coeficientes kappa promedio considerando que $L' > L$ para ambos *TDBs*), y en las Tablas 3.7 a 3.10 se muestran los resultados para los errores tipo I de los test de hipótesis $H_0 : \kappa_{12} = \kappa_{22}$ y $H_0 : \ln(\kappa_{12}) = \ln(\kappa_{22})$ (es decir, cuando se comparan los coeficientes kappa promedio considerando que $L > L'$ para ambos *TDBs*). En estas tablas se indican los valores de las sensibilidades, especificidades, prevalencia y covarianzas con las que se han generado las muestras multinomiales.

Cuando $L' > L$ (Tablas 3.3 a 3.6), la prevalencia de la enfermedad y las covarianzas entre los dos *TDBs* tienen un importante efecto sobre el error tipo I del test $H_0 : \kappa_{11} = \kappa_{21}$. El aumento de la prevalencia implica un aumento del error tipo I, sobre todo en muestras de tamaño 100 y 200, aunque sin desbordar al error nominal (situación que se ha considerado cuando el error tipo I es mayor que el 6.5%). El incremento de los valores de las covarianzas implica una disminución del error tipo I, sobre todo para $n \leq 500$. En términos generales, cuando los valores de las covarianzas son altas, el test de hipótesis $H_0 : \kappa_{11} = \kappa_{21}$ es conservador (su error tipo I es menor que el nominal) para un tamaño muestral $n \leq 500$ (dependiendo de la prevalencia de la enfermedad). La prevalencia y las covarianzas prácticamente no tienen ningún efecto sobre el error tipo I cuando las muestras son muy grandes $(n = 1000 - 2000)$. Por tanto, en términos generales, el error tipo I del test $H_0 : \kappa_{11} = \kappa_{21}$ es menor que el error nominal y a partir de un cierto tamaño muestral fluctúa en torno al error nominal sin desbordarlo. En cuanto al error tipo I del test $H_0 : \ln(\kappa_{11}) = \ln(\kappa_{21})$, su comportamiento es, en términos generales, muy similar al del test

$H_0 : \kappa_{11} = \kappa_{21}$, si bien para muestras de tamaño 100 y 200 su error tipo I es algo menor que el del test de hipótesis sin la transformación.

Tabla 3.3. Errores tipo I cuando $\kappa_{11} = 0.2$ y $\kappa_{21} = 0.2$.

$Se_1 = 0.7773$ $Sp_1 = 0.7308$ $Se_2 = 0.7773$ $Sp_2 = 0.7308$ $p = 10\%$ $\varepsilon_1 \le 0.1731$ $\varepsilon_0 \le 0.1967$						
	$\varepsilon_1 = 0$ $\varepsilon_0 = 0$		$\varepsilon_1 = 0.08$ $\varepsilon_0 = 0.09$		$\varepsilon_1 = 0.16$ $\varepsilon_0 = 0.19$	
n	Método 1	Método 2	Método 1	Método 2	Método 1	Método 2
100	0.022	0.009	0.012	0.008	0	0
200	0.044	0.026	0.031	0.022	0.001	0
300	0.047	0.040	0.035	0.029	0.004	0.004
400	0.045	0.040	0.050	0.042	0.004	0.004
500	0.050	0.048	0.044	0.042	0.010	0.008
1000	0.048	0.046	0.047	0.046	0.020	0.020
2000	0.055	0.056	0.056	0.055	0.044	0.043
$Se_1 = 0.6901$ $Sp_1 = 0.5904$ $Se_2 = 0.6901$ $Sp_2 = 0.5904$ $p = 30\%$ $\varepsilon_1 \le 0.2138$ $\varepsilon_0 \le 0.2418$						
	$\varepsilon_1 = 0$ $\varepsilon_0 = 0$		$\varepsilon_1 = 0.10$ $\varepsilon_0 = 0.11$		$\varepsilon_1 = 0.20$ $\varepsilon_0 = 0.22$	
n	Método 1	Método 2	Método 1	Método 2	Método 1	Método 2
100	0.043	0.008	0.043	0.012	0.001	0
200	0.047	0.025	0.051	0.032	0.013	0.007
300	0.042	0.030	0.059	0.045	0.023	0.016
400	0.052	0.039	0.046	0.037	0.036	0.029
500	0.053	0.045	0.050	0.042	0.038	0.031
1000	0.049	0.049	0.051	0.047	0.052	0.048
2000	0.048	0.048	0.050	0.048	0.054	0.053
$Se_1 = 0.9374$ $Sp_1 = 0.3194$ $Se_2 = 0.9374$ $Sp_2 = 0.3194$ $p = 50\%$ $\varepsilon_1 \le 0.0586$ $\varepsilon_0 \le 0.2173$						
	$\varepsilon_1 = 0$ $\varepsilon_0 = 0$		$\varepsilon_1 = 0.02$ $\varepsilon_0 = 0.10$		$\varepsilon_1 = 0.04$ $\varepsilon_0 = 0.20$	
n	Método 1	Método 2	Método 1	Método 2	Método 1	Método 2
100	0.049	0.019	0.043	0.024	0.004	0
200	0.051	0.040	0.044	0.038	0.015	0.007
300	0.049	0.042	0.059	0.052	0.031	0.024
400	0.049	0.046	0.058	0.051	0.040	0.035
500	0.048	0.044	0.045	0.041	0.052	0.045
1000	0.052	0.050	0.051	0.051	0.050	0.049
2000	0.043	0.042	0.053	0.053	0.048	0.047

Método 1: $H_0 : \kappa_{11} = \kappa_{21}$ vs $H_1 : \kappa_{11} \neq \kappa_{21}$.

Método 2: $H_0 : \ln(\kappa_{11}) = \ln(\kappa_{21})$ vs $H_1 : \ln(\kappa_{11}) \neq \ln(\kappa_{21})$.

Tabla 3.4. Errores tipo I cuando $\kappa_{11} = 0.4$ y $\kappa_{21} = 0.4$.

$Se_1 = 0.8209$ $Sp_1 = 0.8670$ $Se_2 = 0.8209$ $Sp_2 = 0.8670$ $p = 10\%$ $\varepsilon_1 \le 0.1470$ $\varepsilon_0 \le 0.1153$						
	$\varepsilon_1 = 0$ $\varepsilon_0 = 0$		$\varepsilon_1 = 0.06$ $\varepsilon_0 = 0.05$		$\varepsilon_1 = 0.12$ $\varepsilon_0 = 0.10$	
n	Método 1	Método 2	Método 1	Método 2	Método 1	Método 2
100	0.038	0.028	0.022	0.013	0	0
200	0.048	0.043	0.039	0.033	0.005	0.004
300	0.046	0.043	0.034	0.032	0.014	0.012
400	0.047	0.043	0.043	0.042	0.027	0.025
500	0.049	0.047	0.048	0.046	0.028	0.027
1000	0.047	0.047	0.047	0.046	0.045	0.044
2000	0.046	0.046	0.046	0.046	0.056	0.056
$Se_1 = 0.8864$ $Sp_1 = 0.6746$ $Se_2 = 0.8864$ $Sp_2 = 0.6746$ $p = 30\%$ $\varepsilon_1 \le 0.1007$ $\varepsilon_0 \le 0.2195$						
	$\varepsilon_1 = 0$ $\varepsilon_0 = 0$		$\varepsilon_1 = 0.04$ $\varepsilon_0 = 0.10$		$\varepsilon_1 = 0.08$ $\varepsilon_0 = 0.20$	
n	Método 1	Método 2	Método 1	Método 2	Método 1	Método 2
100	0.058	0.049	0.045	0.035	0.001	0
200	0.050	0.046	0.049	0.045	0.022	0.018
300	0.047	0.046	0.052	0.050	0.033	0.031
400	0.052	0.051	0.048	0.047	0.037	0.037
500	0.048	0.047	0.040	0.040	0.045	0.044
1000	0.049	0.048	0.050	0.049	0.058	0.058
2000	0.046	0.046	0.048	0.048	0.054	0.054
$Se_1 = 0.9315$ $Sp_1 = 0.5421$ $Se_2 = 0.9315$ $Sp_2 = 0.5421$ $p = 50\%$ $\varepsilon_1 \le 0.0638$ $\varepsilon_0 \le 0.2482$						
	$\varepsilon_1 = 0$ $\varepsilon_0 = 0$		$\varepsilon_1 = 0.02$ $\varepsilon_0 = 0.10$		$\varepsilon_1 = 0.04$ $\varepsilon_0 = 0.22$	
n	Método 1	Método 2	Método 1	Método 2	Método 1	Método 2
100	0.059	0.056	0.042	0.033	0.007	0.003
200	0.057	0.050	0.053	0.047	0.028	0.025
300	0.048	0.045	0.057	0.055	0.038	0.037
400	0.044	0.042	0.051	0.051	0.048	0.043
500	0.048	0.047	0.051	0.050	0.055	0.053
1000	0.050	0.049	0.052	0.051	0.042	0.042
2000	0.056	0.056	0.050	0.050	0.051	0.051

Método 1: $H_0 : \kappa_{11} = \kappa_{21}$ vs $H_1 : \kappa_{11} \neq \kappa_{21}$.

Método 2: $H_0 : \ln(\kappa_{11}) = \ln(\kappa_{21})$ vs $H_1 : \ln(\kappa_{11}) \neq \ln(\kappa_{21})$.

Tabla 3.5. Errores tipo I cuando $\kappa_{11} = 0.6$ y $\kappa_{21} = 0.6$.

$Se_1 = 0.7929$ $Sp_1 = 0.9421$ $Se_2 = 0.7929$ $Sp_2 = 0.9421$ $p = 10\%$ $\varepsilon_1 \le 0.1642$ $\varepsilon_0 \le 0.0546$						
	$\varepsilon_1 = 0$ $\varepsilon_0 = 0$		$\varepsilon_1 = 0.07$ $\varepsilon_0 = 0.02$		$\varepsilon_1 = 0.14$ $\varepsilon_0 = 0.04$	
n	Método 1	Método 2	Método 1	Método 2	Método 1	Método 2
100	0.028	0.018	0.013	0.007	0.001	0
200	0.038	0.031	0.034	0.028	0.010	0.008
300	0.051	0.046	0.046	0.041	0.025	0.023
400	0.050	0.045	0.041	0.036	0.031	0.030
500	0.057	0.053	0.054	0.051	0.038	0.036
1000	0.053	0.052	0.049	0.049	0.054	0.053
2000	0.056	0.055	0.045	0.044	0.051	0.051
$Se_1 = 0.8495$ $Sp_1 = 0.8375$ $Se_2 = 0.8495$ $Sp_2 = 0.8375$ $p = 30\%$ $\varepsilon_1 \le 0.1279$ $\varepsilon_0 \le 0.1361$						
	$\varepsilon_1 = 0$ $\varepsilon_0 = 0$		$\varepsilon_1 = 0.05$ $\varepsilon_0 = 0.06$		$\varepsilon_1 = 0.10$ $\varepsilon_0 = 0.12$	
n	Método 1	Método 2	Método 1	Método 2	Método 1	Método 2
100	0.046	0.035	0.052	0.042	0.003	0.001
200	0.049	0.046	0.054	0.048	0.019	0.014
300	0.045	0.042	0.053	0.048	0.031	0.030
400	0.048	0.046	0.046	0.045	0.044	0.041
500	0.053	0.052	0.061	0.060	0.050	0.048
1000	0.056	0.056	0.051	0.051	0.046	0.046
2000	0.051	0.051	0.054	0.054	0.043	0.043
$Se_1 = 0.6816$ $Sp_1 = 0.8624$ $Se_2 = 0.6816$ $Sp_2 = 0.8624$ $p = 50\%$ $\varepsilon_1 \le 0.2170$ $\varepsilon_0 \le 0.1187$						
	$\varepsilon_1 = 0$ $\varepsilon_0 = 0$		$\varepsilon_1 = 0.10$ $\varepsilon_0 = 0.05$		$\varepsilon_1 = 0.20$ $\varepsilon_0 = 0.10$	
n	Método 1	Método 2	Método 1	Método 2	Método 1	Método 2
100	0.048	0.034	0.041	0.029	0.004	0.003
200	0.049	0.044	0.046	0.041	0.011	0.008
300	0.050	0.046	0.047	0.044	0.026	0.022
400	0.057	0.055	0.043	0.039	0.033	0.032
500	0.062	0.059	0.047	0.046	0.047	0.046
1000	0.055	0.055	0.044	0.043	0.049	0.047
2000	0.052	0.052	0.049	0.049	0.047	0.047

Método 1: $H_0 : \kappa_{11} = \kappa_{21}$ vs $H_1 : \kappa_{11} \neq \kappa_{21}$.

Método 2: $H_0 : \ln(\kappa_{11}) = \ln(\kappa_{21})$ vs $H_1 : \ln(\kappa_{11}) \neq \ln(\kappa_{21})$.

Tabla 3.6. Errores tipo I cuando $\kappa_{11} = 0.8$ y $\kappa_{21} = 0.8$.

$Se_1 = 0.9637$ $Sp_1 = 0.9701$ $Se_2 = 0.9637$ $Sp_2 = 0.9701$ $p = 10\%$ $\varepsilon_1 \le 0.0350$ $\varepsilon_0 \le 0.0290$						
	$\varepsilon_1 = 0$ $\varepsilon_0 = 0$		$\varepsilon_1 = 0.01$ $\varepsilon_0 = 0.01$		$\varepsilon_1 = 0.02$ $\varepsilon_0 = 0.02$	
n	Método 1	Método 2	Método 1	Método 2	Método 1	Método 2
100	0.005	0.003	0.003	0.002	0	0
200	0.034	0.024	0.023	0.021	0.007	0.007
300	0.043	0.040	0.042	0.040	0.013	0.011
400	0.055	0.053	0.041	0.039	0.028	0.026
500	0.048	0.047	0.038	0.035	0.032	0.032
1000	0.049	0.048	0.052	0.051	0.045	0.044
2000	0.040	0.040	0.052	0.052	0.051	0.051
$Se_1 = 0.7425$ $Sp_1 = 0.9654$ $Se_2 = 0.7425$ $Sp_2 = 0.9654$ $p = 30\%$ $\varepsilon_1 \le 0.1912$ $\varepsilon_0 \le 0.0334$						
	$\varepsilon_1 = 0$ $\varepsilon_0 = 0$		$\varepsilon_1 = 0.09$ $\varepsilon_0 = 0.01$		$\varepsilon_1 = 0.18$ $\varepsilon_0 = 0.02$	
n	Método 1	Método 2	Método 1	Método 2	Método 1	Método 2
100	0.025	0.020	0.017	0.011	0.002	0.001
200	0.046	0.038	0.039	0.034	0.005	0.004
300	0.043	0.041	0.053	0.048	0.026	0.023
400	0.049	0.044	0.054	0.050	0.032	0.029
500	0.058	0.057	0.048	0.048	0.046	0.045
1000	0.048	0.047	0.036	0.034	0.055	0.054
2000	0.042	0.042	0.058	0.058	0.049	0.048
$Se_1 = 0.8063$ $Sp_1 = 0.9392$ $Se_2 = 0.8063$ $Sp_2 = 0.9392$ $p = 50\%$ $\varepsilon_1 \le 0.1562$ $\varepsilon_0 \le 0.0571$						
	$\varepsilon_1 = 0$ $\varepsilon_0 = 0$		$\varepsilon_1 = 0.07$ $\varepsilon_0 = 0.02$		$\varepsilon_1 = 0.14$ $\varepsilon_0 = 0.04$	
n	Método 1	Método 2	Método 1	Método 2	Método 1	Método 2
100	0.033	0.025	0.023	0.019	0.002	0.001
200	0.048	0.045	0.043	0.039	0.011	0.008
300	0.045	0.044	0.036	0.034	0.027	0.024
400	0.053	0.049	0.049	0.047	0.040	0.037
500	0.056	0.055	0.056	0.055	0.037	0.036
1000	0.048	0.048	0.053	0.052	0.043	0.043
2000	0.045	0.045	0.056	0.055	0.051	0.050

Método 1: $H_0 : \kappa_{11} = \kappa_{21}$ vs $H_1 : \kappa_{11} \neq \kappa_{21}$.

Método 2: $H_0 : \ln(\kappa_{11}) = \ln(\kappa_{21})$ vs $H_1 : \ln(\kappa_{11}) \neq \ln(\kappa_{21})$.

Cuando $L > L'$ (Tablas 3.7 a 3.10), la prevalencia y las covarianzas también tienen un efecto importante (y similar a la situación anterior) sobre el error tipo I del test $H_0 : \kappa_{12} = \kappa_{22}$. Al igual que en la situación anterior, el aumento de la prevalencia implica un aumento del error tipo I, sobre todo en muestras de tamaño 100 y 200, aunque sin desbordar al error nominal. El incremento de las covarianzas implica una disminución del error tipo I, sobre todo para $n \leq 500$. Por tanto, en términos generales, cuando los valores de las covarianzas son altos, para un tamaño muestral $n \leq 500$ (dependiendo de la prevalencia de la enfermedad) el test de hipótesis $H_0 : \kappa_{12} = \kappa_{22}$ es conservador. La prevalencia y las covarianzas prácticamente no tienen ningún efecto sobre el error tipo I cuando las muestras son muy grandes $(n = 1000 - 2000)$. Por tanto, en términos generales, el error tipo I del test $H_0 : \kappa_{12} = \kappa_{22}$ tiene un comportamiento muy similar al del test de hipótesis de comparación de los dos coeficientes kappa promedios cuando $L' > L$ $(H_0 : \kappa_{11} = \kappa_{21})$; es decir, es un test conservador y a partir de un determinado tamaño muestral su error tipo I fluctúa en torno al error nominal sin desbordarlo. En cuanto al error tipo I del test $H_0 : \ln(\kappa_{12}) = \ln(\kappa_{22})$, su comportamiento es, en términos generales, muy

similar al del test $H_0 : \kappa_{12} = \kappa_{22}$, si bien para muestras de tamaño 100-200 su error tipo I es, al igual que para el caso $L' > L$, algo menor que el del test de hipótesis sin la transformación.

Tabla 3.7. Errores tipo I cuando $\kappa_{12} = 0.2$ y $\kappa_{22} = 0.2$.

$Se_1 = 0.2091$ $Sp_1 = 0.9715$ $Se_2 = 0.2091$ $Sp_2 = 0.9715$ $p = 10\%$ $\varepsilon_1 \le 0.1654$ $\varepsilon_0 \le 0.0277$						
	$\varepsilon_1 = 0$ $\varepsilon_0 = 0$		$\varepsilon_1 = 0.07$ $\varepsilon_0 = 0.01$		$\varepsilon_1 = 0.14$ $\varepsilon_0 = 0.02$	
n	Método 1	Método 2	Método 1	Método 2	Método 1	Método 2
100	0.012	0	0.003	0	0	0
200	0.041	0.010	0.018	0.004	0.001	0.000
300	0.052	0.021	0.023	0.010	0.001	0.001
400	0.054	0.029	0.051	0.030	0.006	0.002
500	0.051	0.032	0.052	0.033	0.008	0.004
1000	0.057	0.045	0.046	0.034	0.020	0.016
2000	0.057	0.049	0.055	0.049	0.046	0.045
$Se_1 = 0.3019$ $Sp_1 = 0.9030$ $Se_2 = 0.3019$ $Sp_2 = 0.9030$ $p = 30\%$ $\varepsilon_1 \le 0.2108$ $\varepsilon_0 \le 0.0876$						
	$\varepsilon_1 = 0$ $\varepsilon_0 = 0$		$\varepsilon_1 = 0.10$ $\varepsilon_0 = 0.11$		$\varepsilon_1 = 0.20$ $\varepsilon_0 = 0.22$	
n	Método 1	Método 2	Método 1	Método 2	Método 1	Método 2
100	0.052	0.003	0.038	0.008	0	0
200	0.048	0.022	0.046	0.025	0.001	0.001
300	0.049	0.029	0.055	0.040	0.003	0.002
400	0.046	0.033	0.049	0.037	0.006	0.003
500	0.053	0.044	0.053	0.045	0.012	0.008
1000	0.050	0.046	0.050	0.046	0.042	0.040
2000	0.046	0.044	0.045	0.044	0.047	0.045
$Se_1 = 0.4237$ $Sp_1 = 0.8131$ $Se_2 = 0.4237$ $Sp_2 = 0.8131$ $p = 50\%$ $\varepsilon_1 \le 0.2442$ $\varepsilon_0 \le 0.1520$						
	$\varepsilon_1 = 0$ $\varepsilon_0 = 0$		$\varepsilon_1 = 0.02$ $\varepsilon_0 = 0.10$		$\varepsilon_1 = 0.04$ $\varepsilon_0 = 0.20$	
n	Método 1	Método 2	Método 1	Método 2	Método 1	Método 2
100	0.047	0.006	0.045	0.013	0.003	0
200	0.048	0.021	0.049	0.027	0.020	0.006
300	0.056	0.037	0.042	0.030	0.030	0.021
400	0.055	0.044	0.052	0.043	0.045	0.034
500	0.058	0.051	0.042	0.037	0.040	0.034
1000	0.046	0.043	0.055	0.052	0.041	0.039
2000	0.046	0.044	0.048	0.048	0.058	0.057

Método 1: $H_0 : \kappa_{12} = \kappa_{22}$ vs $H_1 : \kappa_{12} \neq \kappa_{22}$.

Método 2: $H_0 : \ln(\kappa_{12}) = \ln(\kappa_{22})$ vs $H_1 : \ln(\kappa_{12}) \neq \ln(\kappa_{22})$.

Tabla 3.8. Errores tipo I cuando $\kappa_{12} = 0.4$ y $\kappa_{22} = 0.4$.

$Se_1 = 0.7773$ $Sp_1 = 0.7308$ $Se_2 = 0.7773$ $Sp_2 = 0.7308$ $p = 10\%$ $\varepsilon_1 \le 0.1731$ $\varepsilon_0 \le 0.1967$						
	$\varepsilon_1 = 0$ $\varepsilon_0 = 0$		$\varepsilon_1 = 0.08$ $\varepsilon_0 = 0.09$		$\varepsilon_1 = 0.16$ $\varepsilon_0 = 0.18$	
n	Método 1	Método 2	Método 1	Método 2	Método 1	Método 2
100	0.022	0.002	0.006	0	0	0
200	0.043	0.026	0.022	0.008	0	0
300	0.055	0.040	0.030	0.018	0.001	0.000
400	0.049	0.037	0.047	0.038	0.002	0.001
500	0.039	0.032	0.047	0.042	0.002	0.002
1000	0.049	0.047	0.053	0.050	0.014	0.011
2000	0.056	0.054	0.051	0.050	0.030	0.030
$Se_1 = 0.8837$ $Sp_1 = 0.4575$ $Se_2 = 0.8837$ $Sp_2 = 0.4575$ $p = 30\%$ $\varepsilon_1 \le 0.1028$ $\varepsilon_0 \le 0.2482$						
	$\varepsilon_1 = 0$ $\varepsilon_0 = 0$		$\varepsilon_1 = 0.04$ $\varepsilon_0 = 0.11$		$\varepsilon_1 = 0.08$ $\varepsilon_0 = 0.22$	
n	Método 1	Método 2	Método 1	Método 2	Método 1	Método 2
100	0.038	0.006	0.029	0.007	0.004	0
200	0.046	0.026	0.052	0.035	0.010	0.003
300	0.052	0.039	0.057	0.048	0.027	0.020
400	0.047	0.040	0.056	0.052	0.038	0.031
500	0.048	0.042	0.055	0.052	0.040	0.034
1000	0.050	0.046	0.049	0.047	0.039	0.036
2000	0.044	0.043	0.045	0.044	0.048	0.047
$Se_1 = 0.8112$ $Sp_1 = 0.5293$ $Se_2 = 0.8112$ $Sp_2 = 0.5293$ $p = 50\%$ $\varepsilon_1 \le 0.1532$ $\varepsilon_0 \le 0.2491$						
	$\varepsilon_1 = 0$ $\varepsilon_0 = 0$		$\varepsilon_1 = 0.07$ $\varepsilon_0 = 0.11$		$\varepsilon_1 = 0.14$ $\varepsilon_0 = 0.22$	
n	Método 1	Método 2	Método 1	Método 2	Método 1	Método 2
100	0.052	0.019	0.041	0.016	0.001	0
200	0.058	0.041	0.054	0.040	0.013	0.006
300	0.059	0.052	0.051	0.041	0.034	0.028
400	0.049	0.043	0.046	0.042	0.038	0.033
500	0.055	0.047	0.048	0.045	0.045	0.041
1000	0.054	0.051	0.058	0.054	0.061	0.059
2000	0.054	0.052	0.051	0.051	0.048	0.047

Método 1: $H_0 : \kappa_{12} = \kappa_{22}$ vs $H_1 : \kappa_{12} \neq \kappa_{22}$.

Método 2: $H_0 : \ln(\kappa_{12}) = \ln(\kappa_{22})$ vs $H_1 : \ln(\kappa_{12}) \neq \ln(\kappa_{22})$.

Tabla 3.9. Errores tipo I cuando $\kappa_{12} = 0.6$ y $\kappa_{22} = 0.6$.

$Se_1 = 0.8209$ $Sp_1 = 0.8670$ $Se_2 = 0.8209$ $Sp_2 = 0.8670$ $p = 10\%$ $\varepsilon_1 \le 0.1470$ $\varepsilon_0 \le 0.1153$						
	$\varepsilon_1 = 0$ $\varepsilon_0 = 0$		$\varepsilon_1 = 0.06$ $\varepsilon_0 = 0.05$		$\varepsilon_1 = 0.12$ $\varepsilon_0 = 0.10$	
n	Método 1	Método 2	Método 1	Método 2	Método 1	Método 2
100	0.017	0.005	0.005	0.002	0.001	0
200	0.039	0.031	0.018	0.012	0.001	0
300	0.056	0.046	0.039	0.030	0.004	0.001
400	0.056	0.049	0.045	0.039	0.011	0.007
500	0.058	0.056	0.044	0.040	0.009	0.007
1000	0.052	0.049	0.050	0.048	0.033	0.028
2000	0.051	0.049	0.056	0.055	0.045	0.044
$Se_1 = 0.8864$ $Sp_1 = 0.6746$ $Se_2 = 0.8864$ $Sp_2 = 0.6746$ $p = 30\%$ $\varepsilon_1 \le 0.1007$ $\varepsilon_0 \le 0.2195$						
	$\varepsilon_1 = 0$ $\varepsilon_0 = 0$		$\varepsilon_1 = 0.04$ $\varepsilon_0 = 0.10$		$\varepsilon_1 = 0.08$ $\varepsilon_0 = 0.20$	
n	Método 1	Método 2	Método 1	Método 2	Método 1	Método 2
100	0.050	0.034	0.024	0.014	0.001	0
200	0.053	0.048	0.044	0.036	0.009	0.006
300	0.044	0.040	0.058	0.051	0.017	0.015
400	0.053	0.050	0.052	0.048	0.030	0.028
500	0.052	0.050	0.054	0.050	0.033	0.032
1000	0.054	0.054	0.049	0.047	0.051	0.051
2000	0.055	0.054	0.063	0.062	0.056	0.055
$Se_1 = 0.7458$ $Sp_1 = 0.8991$ $Se_2 = 0.7458$ $Sp_2 = 0.8991$ $p = 50\%$ $\varepsilon_1 \le 0.1896$ $\varepsilon_0 \le 0.0908$						
	$\varepsilon_1 = 0$ $\varepsilon_0 = 0$		$\varepsilon_1 = 0.08$ $\varepsilon_0 = 0.04$		$\varepsilon_1 = 0.16$ $\varepsilon_0 = 0.08$	
n	Método 1	Método 2	Método 1	Método 2	Método 1	Método 2
100	0.055	0.047	0.043	0.033	0.005	0.004
200	0.048	0.044	0.058	0.056	0.026	0.022
300	0.039	0.037	0.057	0.056	0.034	0.033
400	0.053	0.051	0.058	0.057	0.047	0.045
500	0.049	0.048	0.046	0.046	0.052	0.049
1000	0.058	0.058	0.048	0.048	0.051	0.050
2000	0.056	0.055	0.049	0.049	0.052	0.052

Método 1: $H_0 : \kappa_{12} = \kappa_{22}$ vs $H_1 : \kappa_{12} \neq \kappa_{22}$.

Método 2: $H_0 : \ln(\kappa_{12}) = \ln(\kappa_{22})$ vs $H_1 : \ln(\kappa_{12}) \neq \ln(\kappa_{22})$.

Tabla 3.10. Errores tipo I cuando $\kappa_{12} = 0.8$ y $\kappa_{22} = 0.8$.

$Se_1 = 0.9569$ $Sp_1 = 0.9224$ $Se_2 = 0.9569$ $Sp_2 = 0.9224$ $p = 10\%$ $\varepsilon_1 \le 0.0413$ $\varepsilon_0 \le 0.0716$						
	$\varepsilon_1 = 0$ $\varepsilon_0 = 0$		$\varepsilon_1 = 0.02$ $\varepsilon_0 = 0.03$		$\varepsilon_1 = 0.04$ $\varepsilon_0 = 0.06$	
n	Método 1	Método 2	Método 1	Método 2	Método 1	Método 2
100	0.002	0.001	0	0	0	0
200	0.011	0.009	0.004	0.003	0	0
300	0.018	0.015	0.012	0.011	0	0
400	0.031	0.028	0.020	0.017	0.001	0.001
500	0.042	0.038	0.021	0.018	0	0
1000	0.054	0.051	0.036	0.035	0.003	0.003
2000	0.043	0.043	0.043	0.043	0.013	0.012
$Se_1 = 0.9102$ $Sp_1 = 0.8773$ $Se_2 = 0.9102$ $Sp_2 = 0.8773$ $p = 30\%$ $\varepsilon_1 \le 0.0818$ $\varepsilon_0 \le 0.1077$						
	$\varepsilon_1 = 0$ $\varepsilon_0 = 0$		$\varepsilon_1 = 0.03$ $\varepsilon_0 = 0.04$		$\varepsilon_1 = 0.06$ $\varepsilon_0 = 0.08$	
n	Método 1	Método 2	Método 1	Método 2	Método 1	Método 2
100	0.030	0.025	0.017	0.011	0.002	0.001
200	0.059	0.053	0.033	0.029	0.008	0.008
300	0.051	0.048	0.052	0.049	0.021	0.019
400	0.048	0.045	0.046	0.044	0.035	0.034
500	0.056	0.054	0.047	0.046	0.038	0.037
1000	0.053	0.052	0.047	0.047	0.049	0.048
2000	0.050	0.049	0.059	0.058	0.048	0.047
$Se_1 = 0.9392$ $Sp_1 = 0.8063$ $Se_2 = 0.9392$ $Sp_2 = 0.8063$ $p = 50\%$ $\varepsilon_1 \le 0.0571$ $\varepsilon_0 \le 0.1562$						
	$\varepsilon_1 = 0$ $\varepsilon_0 = 0$		$\varepsilon_1 = 0.02$ $\varepsilon_0 = 0.07$		$\varepsilon_1 = 0.04$ $\varepsilon_0 = 0.14$	
n	Método 1	Método 2	Método 1	Método 2	Método 1	Método 2
100	0.040	0.030	0.022	0.019	0.003	0.001
200	0.050	0.046	0.042	0.040	0.012	0.010
300	0.051	0.049	0.050	0.049	0.022	0.021
400	0.053	0.050	0.058	0.057	0.036	0.033
500	0.049	0.048	0.051	0.048	0.043	0.040
1000	0.053	0.052	0.054	0.053	0.054	0.053
2000	0.054	0.053	0.055	0.054	0.049	0.049

Método 1: $H_0 : \kappa_{12} = \kappa_{22}$ vs $H_1 : \kappa_{12} \neq \kappa_{22}$.

Método 2: $H_0 : \ln(\kappa_{12}) = \ln(\kappa_{22})$ vs $H_1 : \ln(\kappa_{12}) \neq \ln(\kappa_{22})$.

3.6.3.2. Potencias

En las Tablas 3.11 a 3.15 se muestran los resultados para las potencias de los test de hipótesis $H_0 : \kappa_{11} = \kappa_{21}$ y $H_0 : \ln(\kappa_{11}) = \ln(\kappa_{21})$ (es decir, cuando se comparan los coeficientes kappa promedio considerando que $L' > L$ para ambos *TDBs*), y en las Tablas 3.16 a 3.20 se muestran los resultados para los potencias de los test de hipótesis $H_0 : \kappa_{12} = \kappa_{22}$ y $H_0 : \ln(\kappa_{12}) = \ln(\kappa_{22})$ (es decir, cuando se comparan los coeficientes kappa promedio considerando que $L > L'$ para ambos *TDBs*). En estas tablas, al igual que para los errores tipo I, se indican los valores de las sensibilidades, especificidades, prevalencia y covarianzas con las que se han generado las muestras multinomiales.

Cuando $L' > L$ (Tablas 3.11 a 3.15), la prevalencia de la enfermedad tiene un importante efecto en las potencias de los test $H_0 : \kappa_{11} = \kappa_{21}$ y $H_0 : \ln(\kappa_{11}) = \ln(\kappa_{21})$. En términos generales, cuando la prevalencia es pequeña $(p = 10\%)$, para un mismo tamaño muestral y covarianzas, las potencias son menores que cuando la prevalencia es mayor $(p \geq 30\%)$. En cuanto a las covarianzas entre los dos *TDBs*, su efecto sobre las potencias es, en términos generales, de menor importancia que el de la prevalencia, si

bien para muestras de tamaño entre 100 y 300 individuos, el aumento de los valores de las covarianzas implica un aumento de las potencias. En términos generales, cuando la diferencia entre los dos coeficientes kappa promedios es pequeña, por ejemplo para $|\kappa_{11} - \kappa_{21}| = 0.2$, se obtienen las siguientes conclusiones:

a) Si la prevalencia es pequeña $(p = 10\%)$ y los valores de los coeficientes kappa son bajos $(\kappa_{i1} \leq 0.4)$, con un tamaño muestral $n \geq 200$ se obtienen potencias muy elevadas (superiores al 80% o 90% dependiendo de las covarianzas); cuando alguno de los dos coeficientes kappa promedios toma un valor mayor, entonces con un tamaño muestral $n \geq 400 - 500$ se obtienen potencias superiores al 80%-90% (dependiendo de las covarianzas).

b) Si la prevalencia es elevada $(p \geq 30\%)$, con un tamaño muestral $n \geq 200$ las potencias de ambos test de hipótesis son muy elevadas (superiores al 80% o 90% dependiendo de las covarianzas).

Cuando la diferencia entre los dos coeficientes kappa promedios es grande, por ejemplo para $|\kappa_{11} - \kappa_{12}| = 0.4$, se obtienen las siguientes conclusiones:

a) Si la prevalencia es pequeña $(p = 10\%)$, con un tamaño muestral $n \geq 200$ se obtienen potencias superiores al 90%.

b) Si la prevalencia es elevada $(p \geq 30\%)$, con un tamaño muestral $n \geq 100$ las potencias de ambos test de hipótesis son superiores al 90%.

Finalmente, en términos generales, el test $H_0 : \kappa_{11} = \kappa_{21}$ es más potente que el test $H_0 : \ln(\kappa_{11}) = \ln(\kappa_{21})$, sobre todo cuando $n \leq 200$, debido a que su error tipo I es algo mayor (sin desbordar al error nominal).

Tabla 3.11. Potencias cuando $\kappa_{11} = 0.4$ y $\kappa_{21} = 0.2$.

$Se_1 = 0.8209$ $Sp_1 = 0.8670$ $Se_2 = 0.7773$ $Sp_2 = 0.7308$ $p = 10\%$ $\varepsilon_1 \le 0.1392$ $\varepsilon_0 \le 0.0972$						
	$\varepsilon_1 = 0$ $\varepsilon_0 = 0$		$\varepsilon_1 = 0.06$ $\varepsilon_0 = 0.04$		$\varepsilon_1 = 0.12$ $\varepsilon_0 = 0.08$	
n	Método 1	Método 2	Método 1	Método 2	Método 1	Método 2
100	0.498	0.452	0.613	0.593	0.767	0.755
200	0.831	0.837	0.927	0.935	0.995	0.996
300	0.937	0.941	0.987	0.988	1	1
400	0.986	0.987	1	1	1	1
500	0.990	0.991	1	1	1	1
1000	1	1	1	1	1	1
2000	1	1	1	1	1	1
$Se_1 = 0.7413$ $Sp_1 = 0.7441$ $Se_2 = 0.6901$ $Sp_2 = 0.5904$ $p = 30\%$ $\varepsilon_1 \le 0.1785$ $\varepsilon_0 \le 0.1511$						
	$\varepsilon_1 = 0$ $\varepsilon_0 = 0$		$\varepsilon_1 = 0.08$ $\varepsilon_0 = 0.07$		$\varepsilon_1 = 0.16$ $\varepsilon_0 = 0.14$	
n	Método 1	Método 2	Método 1	Método 2	Método 1	Método 2
100	0.535	0.401	0.679	0.561	0.908	0.819
200	0.804	0.779	0.929	0.919	0.999	0.998
300	0.918	0.909	0.988	0.987	1	1
400	0.968	0.964	0.994	0.994	1	1
500	0.991	0.990	1	1	1	1
1000	1	1	1	1	1	1
2000	1	1	1	1	1	1
$Se_1 = 0.8315$ $Sp_1 = 0.6111$ $Se_2 = 0.8131$ $Sp_2 = 0.4237$ $p = 50\%$ $\varepsilon_1 \le 0.1370$ $\varepsilon_0 \le 0.1648$						
	$\varepsilon_1 = 0$ $\varepsilon_0 = 0$		$\varepsilon_1 = 0.06$ $\varepsilon_0 = 0.07$		$\varepsilon_1 = 0.12$ $\varepsilon_0 = 0.14$	
n	Método 1	Método 2	Método 1	Método 2	Método 1	Método 2
100	0.550	0.391	0.672	0.549	0.852	0.743
200	0.811	0.785	0.923	0.908	0.997	0.994
300	0.919	0.915	0.980	0.979	1	1
400	0.962	0.959	0.993	0.993	1	1
500	0.987	0.986	0.999	0.999	1	1
1000	1	1	1	1	1	1
2000	1	1	1	1	1	1

Método 1: $H_0 : \kappa_{11} = \kappa_{21}$ vs $H_1 : \kappa_{11} \neq \kappa_{21}$.

Método 2: $H_0 : \ln(\kappa_{11}) = \ln(\kappa_{21})$ vs $H_1 : \ln(\kappa_{11}) \neq \ln(\kappa_{21})$.

Tabla 3.12. Potencias cuando $\kappa_{11} = 0.6$ y $\kappa_{21} = 0.4$.

$Se_1 = 0.7929$ $Sp_1 = 0.9421$ $Se_2 = 0.6318$ $Sp_2 = 0.9056$ $p = 10\%$ $\varepsilon_1 \le 0.1308$ $\varepsilon_0 \le 0.0525$						
	$\varepsilon_1 = 0$ $\varepsilon_0 = 0$		$\varepsilon_1 = 0.06$ $\varepsilon_0 = 0.02$		$\varepsilon_1 = 0.12$ $\varepsilon_0 = 0.04$	
n	Método 1	Método 2	Método 1	Método 2	Método 1	Método 2
100	0.217	0.146	0.218	0.153	0.204	0.139
200	0.510	0.468	0.584	0.545	0.709	0.677
300	0.667	0.647	0.797	0.786	0.915	0.907
400	0.788	0.775	0.887	0.878	0.976	0.973
500	0.861	0.849	0.949	0.946	0.993	0.993
1000	0.986	0.985	0.999	0.998	1	1
2000	1	1	1	1	1	1
$Se_1 = 0.8495$ $Sp_1 = 0.8375$ $Se_2 = 0.8864$ $Sp_2 = 0.6746$ $p = 30\%$ $\varepsilon_1 \le 0.0965$ $\varepsilon_0 \le 0.1096$						
	$\varepsilon_1 = 0$ $\varepsilon_0 = 0$		$\varepsilon_1 = 0.04$ $\varepsilon_0 = 0.04$		$\varepsilon_1 = 0.08$ $\varepsilon_0 = 0.08$	
n	Método 1	Método 2	Método 1	Método 2	Método 1	Método 2
100	0.570	0.554	0.692	0.677	0.820	0.812
200	0.845	0.844	0.917	0.916	0.985	0.986
300	0.939	0.939	0.982	0.983	0.998	0.998
400	0.984	0.984	0.997	0.997	1	1
500	0.992	0.992	0.998	0.998	1	1
1000	1	1	1	1	1	1
2000	1	1	1	1	1	1
$Se_1 = 0.6816$ $Sp_1 = 0.8624$ $Se_2 = 0.9315$ $Sp_2 = 0.5421$ $p = 50\%$ $\varepsilon_1 \le 0.0467$ $\varepsilon_0 \le 0.0746$						
	$\varepsilon_1 = 0$ $\varepsilon_0 = 0$		$\varepsilon_1 = 0.02$ $\varepsilon_0 = 0.03$		$\varepsilon_1 = 0.04$ $\varepsilon_0 = 0.06$	
n	Método 1	Método 2	Método 1	Método 2	Método 1	Método 2
100	0.531	0.513	0.579	0.569	0.598	0.582
200	0.784	0.780	0.823	0.824	0.896	0.895
300	0.896	0.896	0.945	0.946	0.972	0.973
400	0.960	0.960	0.974	0.975	0.996	0.996
500	0.985	0.986	0.993	0.993	0.999	0.999
1000	1	1	1	1	1	1
2000	1	1	1	1	1	1

Método 1: $H_0 : \kappa_{11} = \kappa_{21}$ vs $H_1 : \kappa_{11} \neq \kappa_{21}$.

Método 2: $H_0 : \ln(\kappa_{11}) = \ln(\kappa_{21})$ vs $H_1 : \ln(\kappa_{11}) \neq \ln(\kappa_{21})$.

Tabla 3.13. Potencias cuando $\kappa_{11} = 0.6$ y $\kappa_{21} = 0.2$.

$Se_1 = 0.9569$ $Sp_1 = 0.9224$ $Se_2 = 0.5114$ $Sp_2 = 0.8326$ $p = 10\%$ $\varepsilon_1 \le 0.0220$ $\varepsilon_0 \le 0.0645$						
	$\varepsilon_1 = 0$ $\varepsilon_0 = 0$		$\varepsilon_1 = 0.009$ $\varepsilon_0 = 0.025$		$\varepsilon_1 = 0.018$ $\varepsilon_0 = 0.05$	
n	Método 1	Método 2	Método 1	Método 2	Método 1	Método 2
100	0.763	0.606	0.812	0.656	0.870	0.731
200	0.992	0.982	0.998	0.994	1	0.994
300	1	1	1	1	1	0.998
400	1	1	1	1	1	1
500	1	1	1	1	1	1
1000	1	1	1	1	1	1
2000	1	1	1	1	1	1
$Se_1 = 0.9707$ $Sp_1 = 0.7969$ $Se_2 = 0.6901$ $Sp_2 = 0.5904$ $p = 30\%$ $\varepsilon_1 \le 0.0202$ $\varepsilon_0 \le 0.1199$						
	$\varepsilon_1 = 0$ $\varepsilon_0 = 0$		$\varepsilon_1 = 0.09$ $\varepsilon_0 = 0.01$		$\varepsilon_1 = 0.18$ $\varepsilon_0 = 0.02$	
n	Método 1	Método 2	Método 1	Método 2	Método 1	Método 2
100	0.981	0.938	0.997	0.967	0.998	0.977
200	1	0.999	1	1	1	0.998
300	1	1	1	1	1	1
400	1	1	1	1	1	1
500	1	1	1	1	1	1
1000	1	1	1	1	1	1
2000	1	1	1	1	1	1
$Se_1 = 0.8991$ $Sp_1 = 0.7458$ $Se_2 = 0.8131$ $Sp_2 = 0.4237$ $p = 50\%$ $\varepsilon_1 \le 0.0820$ $\varepsilon_0 \le 0.1076$						
	$\varepsilon_1 = 0$ $\varepsilon_0 = 0$		$\varepsilon_1 = 0.06$ $\varepsilon_0 = 0.01$		$\varepsilon_1 = 0.12$ $\varepsilon_0 = 0.02$	
n	Método 1	Método 2	Método 1	Método 2	Método 1	Método 2
100	0.954	0.907	0.973	0.949	0.982	0.961
200	0.998	0.998	0.998	0.998	0.999	0.999
300	1	1	1	1	1	1
400	1	1	1	1	1	1
500	1	1	1	1	1	1
1000	1	1	1	1	1	1
2000	1	1	1	1	1	1

Método 1: $H_0 : \kappa_{11} = \kappa_{21}$ vs $H_1 : \kappa_{11} \neq \kappa_{21}$.

Método 2: $H_0 : \ln(\kappa_{11}) = \ln(\kappa_{21})$ vs $H_1 : \ln(\kappa_{11}) \neq \ln(\kappa_{21})$.

Tabla 3.14. Potencias cuando $\kappa_{11} = 0.8$ y $\kappa_{21} = 0.4$.

$Se_1 = 0.9637$ $Sp_1 = 0.9701$ $Se_2 = 0.8209$ $Sp_2 = 0.8670$ $p = 10\%$ $\varepsilon_1 \le 0.0298$ $\varepsilon_0 \le 0.0260$						
	$\varepsilon_1 = 0$ $\varepsilon_0 = 0$		$\varepsilon_1 = 0.01$ $\varepsilon_0 = 0.01$		$\varepsilon_1 = 0.02$ $\varepsilon_0 = 0.02$	
n	Método 1	Método 2	Método 1	Método 2	Método 1	Método 2
100	0.601	0.570	0.612	0.588	0.671	0.654
200	0.951	0.949	0.937	0.937	0.936	0.936
300	0.989	0.989	0.984	0.984	0.985	0.985
400	0.994	0.994	0.995	0.995	0.995	0.995
500	1	1	0.998	0.998	0.999	0.999
1000	1	1	1	1	1	1
2000	1	1	1	1	1	1
$Se_1 = 0.7425$ $Sp_1 = 0.9654$ $Se_2 = 0.7413$ $Sp_2 = 0.7441$ $p = 30\%$ $\varepsilon_1 \le 0.1909$ $\varepsilon_0 \le 0.0258$						
	$\varepsilon_1 = 0$ $\varepsilon_0 = 0$		$\varepsilon_1 = 0.09$ $\varepsilon_0 = 0.01$		$\varepsilon_1 = 0.18$ $\varepsilon_0 = 0.02$	
n	Método 1	Método 2	Método 1	Método 2	Método 1	Método 2
100	0.901	0.891	0.931	0.926	0.958	0.957
200	0.995	0.995	0.993	0.993	0.994	0.994
300	1	1	0.999	0.999	1	1
400	1	1	0.999	0.999	1	1
500	1	1	1	1	1	1
1000	1	1	1	1	1	1
2000	1	1	1	1	1	1
$Se_1 = 0.8063$ $Sp_1 = 0.9392$ $Se_2 = 0.8315$ $Sp_2 = 0.6111$ $p = 50\%$ $\varepsilon_1 \le 0.1359$ $\varepsilon_0 \le 0.0371$						
	$\varepsilon_1 = 0$ $\varepsilon_0 = 0$		$\varepsilon_1 = 0.06$ $\varepsilon_0 = 0.01$		$\varepsilon_1 = 0.12$ $\varepsilon_0 = 0.02$	
n	Método 1	Método 2	Método 1	Método 2	Método 1	Método 2
100	0.945	0.940	0.949	0.945	0.969	0.968
200	0.998	0.998	0.999	0.999	0.997	0.997
300	1	1	1	1	1	1
400	1	1	1	1	1	1
500	1	1	1	1	1	1
1000	1	1	1	1	1	1
2000	1	1	1	1	1	1

Método 1: $H_0 : \kappa_{11} = \kappa_{21}$ vs $H_1 : \kappa_{11} \neq \kappa_{21}$.

Método 2: $H_0 : \ln(\kappa_{11}) = \ln(\kappa_{21})$ vs $H_1 : \ln(\kappa_{11}) \neq \ln(\kappa_{21})$.

Tabla 3.15. Potencias cuando $\kappa_{11} = 0.8$ y $\kappa_{21} = 0.6$.

$Se_1 = 0.6849$ $Sp_1 = 0.9890$ $Se_2 = 0.9569$ $Sp_2 = 0.9224$ $p = 10\%$ $\varepsilon_1 \le 0.0295$ $\varepsilon_0 \le 0.0101$						
	$\varepsilon_1 = 0$ $\varepsilon_0 = 0$		$\varepsilon_1 = 0.01$ $\varepsilon_0 = 0.004$		$\varepsilon_1 = 0.02$ $\varepsilon_0 = 0.008$	
n	Método 1	Método 2	Método 1	Método 2	Método 1	Método 2
100	0.128	0.106	0.114	0.093	0.130	0.105
200	0.474	0.459	0.499	0.483	0.476	0.466
300	0.686	0.678	0.728	0.723	0.739	0.735
400	0.808	0.805	0.839	0.837	0.874	0.873
500	0.866	0.865	0.924	0.924	0.926	0.926
1000	0.994	0.994	0.997	0.997	0.999	0.999
2000	1	1	1	1	1	1
$Se_1 = 0.9732$ $Sp_1 = 0.9146$ $Se_2 = 0.9707$ $Sp_2 = 0.7969$ $p = 30\%$ $\varepsilon_1 \le 0.0261$ $\varepsilon_0 \le 0.0681$						
	$\varepsilon_1 = 0$ $\varepsilon_0 = 0$		$\varepsilon_1 = 0.01$ $\varepsilon_0 = 0.03$		$\varepsilon_1 = 0.02$ $\varepsilon_0 = 0.06$	
n	Método 1	Método 2	Método 1	Método 2	Método 1	Método 2
100	0.600	0.580	0.697	0.672	0.778	0.762
200	0.875	0.870	0.959	0.956	0.990	0.990
300	0.965	0.963	0.994	0.994	1	1
400	0.990	0.990	1	1	1	1
500	0.998	0.998	1	1	1	1
1000	1	1	1	1	1	1
2000	1	1	1	1	1	1
$Se_1 = 0.9817$ $Sp_1 = 0.8644$ $Se_2 = 0.8991$ $Sp_2 = 0.7458$ $p = 50\%$ $\varepsilon_1 \le 0.0164$ $\varepsilon_0 \le 0.1012$						
	$\varepsilon_1 = 0$ $\varepsilon_0 = 0$		$\varepsilon_1 = 0.005$ $\varepsilon_0 = 0.04$		$\varepsilon_1 = 0.01$ $\varepsilon_0 = 0.08$	
n	Método 1	Método 2	Método 1	Método 2	Método 1	Método 2
100	0.560	0.536	0.654	0.631	0.782	0.742
200	0.862	0.853	0.936	0.932	0.987	0.986
300	0.959	0.957	0.990	0.989	1	1
400	0.988	0.987	0.998	0.998	1	1
500	0.998	0.998	1	1	1	1
1000	1	1	1	1	1	1
2000	1	1	1	1	1	1

Método 1: $H_0 : \kappa_{11} = \kappa_{21}$ vs $H_1 : \kappa_{11} \neq \kappa_{21}$.

Método 2: $H_0 : \ln(\kappa_{11}) = \ln(\kappa_{21})$ vs $H_1 : \ln(\kappa_{11}) \neq \ln(\kappa_{21})$.

Cuando $L > L'$ (Tablas 3.16 a 3.20), las potencias de los test de hipótesis $H_0 : \kappa_{12} = \kappa_{22}$ y $H_0 : \ln(\kappa_{12}) = \ln(\kappa_{22})$ tienen un comportamiento similar al del caso anterior $(L' > L)$. La prevalencia de la enfermedad y las covarianzas tienen un efecto muy similar, y las conclusiones sobre las potencias son en algunos casos también muy similares. En términos generales se obtienen las siguientes conclusiones. Cuando la diferencia entre los dos coeficientes kappa promedios es pequeña $(|\kappa_{12} - \kappa_{22}| = 0.2)$:

a) Si la prevalencia es pequeña $(p = 10\%)$ y los valores de los coeficientes kappa son bajos $(\kappa_{i2} \leq 0.4)$, se necesita un tamaño muestral muy alto $(n > 500$ o $n \geq 1000)$ para que las potencias de ambos test de hipótesis sean muy elevadas (superiores al 90%, dependiendo de las covarianzas); cuando alguno de los dos coeficientes kappa promedios toma valores mayores, entonces con un tamaño muestral $n \geq 400 - 500$ se obtienen potencias superiores al 80%-90% (dependiendo de las covarianzas).

b) Si la prevalencia es elevada $\left(p \geq 30\%\right)$, con un tamaño muestral $n \geq 200$ o 300 (dependiendo de las covarianzas) las potencias de ambos test de hipótesis son superiores al 80% o 90%.

Cuando la diferencia entre los dos coeficientes kappa promedios es grande $\left(\left|\kappa_{12} - \kappa_{22}\right| = 0.4\right)$ se obtienen las siguientes conclusiones:

a) Si la prevalencia es pequeña $\left(p = 10\%\right)$, con un tamaño muestral $n \geq 200$ se obtienen potencias superiores al 95%.

b) Si la prevalencia es elevada $\left(p \geq 30\%\right)$, con un tamaño muestral $n \geq 100$ las potencias de ambos test de hipótesis son superiores al 90%.

Finalmente, y al igual que en el caso anterior, el test $H_0 : \kappa_{12} = \kappa_{22}$ es más potente que el test $H_0 : \ln\left(\kappa_{12}\right) = \ln\left(\kappa_{22}\right)$, sobre todo cuando $n \leq 200$, debido a que su error tipo I es también levemente mayor (sin desbordar al error nominal).

Tabla 3.16. Potencias cuando $\kappa_{12} = 0.4$ y $\kappa_{22} = 0.2$.

$Se_1 = 0.7773$ $Sp_1 = 0.7308$ $Se_2 = 0.2091$ $Sp_2 = 0.9715$ $p = 10\%$ $\varepsilon_1 \le 0.0466$ $\varepsilon_0 \le 0.0208$						
	$\varepsilon_1 = 0$ $\varepsilon_0 = 0$		$\varepsilon_1 = 0.02$ $\varepsilon_0 = 0.009$		$\varepsilon_1 = 0.04$ $\varepsilon_0 = 0.018$	
n	Método 1	Método 2	Método 1	Método 2	Método 1	Método 2
100	0.090	0.013	0.075	0.012	0.076	0.013
200	0.322	0.152	0.328	0.159	0.314	0.132
300	0.501	0.358	0.527	0.368	0.505	0.362
400	0.661	0.555	0.678	0.569	0.642	0.526
500	0.763	0.700	0.771	0.695	0.773	0.694
1000	0.955	0.942	0.972	0.962	0.971	0.965
2000	0.999	0.999	0.999	0.999	1	1
$Se_1 = 0.7021$ $Sp_1 = 0.6817$ $Se_2 = 0.3019$ $Sp_2 = 0.9030$ $p = 30\%$ $\varepsilon_1 \le 0.0900$ $\varepsilon_0 \le 0.0661$						
	$\varepsilon_1 = 0$ $\varepsilon_0 = 0$		$\varepsilon_1 = 0.04$ $\varepsilon_0 = 0.03$		$\varepsilon_1 = 0.08$ $\varepsilon_0 = 0.06$	
n	Método 1	Método 2	Método 1	Método 2	Método 1	Método 2
100	0.410	0.238	0.416	0.252	0.433	0.278
200	0.630	0.587	0.733	0.693	0.784	0.749
300	0.790	0.773	0.862	0.851	0.931	0.927
400	0.878	0.876	0.938	0.936	0.978	0.977
500	0.941	0.940	0.970	0.969	0.991	0.991
1000	0.998	0.998	1	1	1	1
2000	1	1	1	1	1	1
$Se_1 = 0.8112$ $Sp_1 = 0.5293$ $Se_2 = 0.4237$ $Sp_2 = 0.8131$ $p = 50\%$ $\varepsilon_1 \le 0.0800$ $\varepsilon_0 \le 0.0989$						
	$\varepsilon_1 = 0$ $\varepsilon_0 = 0$		$\varepsilon_1 = 0.03$ $\varepsilon_0 = 0.04$		$\varepsilon_1 = 0.06$ $\varepsilon_0 = 0.08$	
n	Método 1	Método 2	Método 1	Método 2	Método 1	Método 2
100	0.474	0.371	0.489	0.403	0.545	0.461
200	0.692	0.682	0.760	0.755	0.835	0.832
300	0.842	0.842	0.886	0.890	0.939	0.946
400	0.925	0.927	0.951	0.953	0.983	0.984
500	0.968	0.969	0.979	0.982	0.992	0.992
1000	1	1	1	1	1	1
2000	1	1	1	1	1	1

Método 1: $H_0 : \kappa_{12} = \kappa_{22}$ vs $H_1 : \kappa_{12} \neq \kappa_{22}$.

Método 2: $H_0 : \ln(\kappa_{12}) = \ln(\kappa_{22})$ vs $H_1 : \ln(\kappa_{12}) \neq \ln(\kappa_{22})$.

Tabla 3.17. Potencias cuando $\kappa_{12} = 0.6$ y $\kappa_{22} = 0.4$.

$Se_1 = 0.8209$ $Sp_1 = 0.8670$ $Se_2 = 0.7773$ $Sp_2 = 0.7308$ $p = 10\%$ $\varepsilon_1 \le 0.1392$ $\varepsilon_0 \le 0.0972$						
	$\varepsilon_1 = 0$ $\varepsilon_0 = 0$		$\varepsilon_1 = 0.06$ $\varepsilon_0 = 0.05$		$\varepsilon_1 = 0.12$ $\varepsilon_0 = 0.10$	
n	Método 1	Método 2	Método 1	Método 2	Método 1	Método 2
100	0.283	0.145	0.288	0.167	0.283	0.163
200	0.560	0.519	0.681	0.656	0.907	0.891
300	0.695	0.680	0.859	0.852	0.988	0.987
400	0.825	0.817	0.933	0.931	1	1
500	0.881	0.878	0.970	0.969	1	1
1000	0.990	0.991	0.999	0.999	1	1
2000	1	1	1	1	1	1
$Se_1 = 0.8864$ $Sp_1 = 0.6746$ $Se_2 = 0.7021$ $Sp_2 = 0.6817$ $p = 30\%$ $\varepsilon_1 \le 0.0797$ $\varepsilon_0 \le 0.2147$						
	$\varepsilon_1 = 0$ $\varepsilon_0 = 0$		$\varepsilon_1 = 0.03$ $\varepsilon_0 = 0.10$		$\varepsilon_1 = 0.06$ $\varepsilon_0 = 0.20$	
n	Método 1	Método 2	Método 1	Método 2	Método 1	Método 2
100	0.398	0.295	0.463	0.347	0.556	0.414
200	0.682	0.646	0.788	0.757	0.930	0.913
300	0.822	0.803	0.915	0.900	0.987	0.986
400	0.902	0.894	0.967	0.965	0.998	0.997
500	0.954	0.949	0.991	0.989	1	1
1000	0.999	0.999	1	1	1	1
2000	1	1	1	1	1	1
$Se_1 = 0.8624$ $Sp_1 = 0.6816$ $Se_2 = 0.8112$ $Sp_2 = 0.5293$ $p = 50\%$ $\varepsilon_1 \le 0.1116$ $\varepsilon_0 \le 0.1686$						
	$\varepsilon_1 = 0$ $\varepsilon_0 = 0$		$\varepsilon_1 = 0.05$ $\varepsilon_0 = 0.07$		$\varepsilon_1 = 0.10$ $\varepsilon_0 = 0.14$	
n	Método 1	Método 2	Método 1	Método 2	Método 1	Método 2
100	0.434	0.347	0.567	0.473	0.731	0.623
200	0.680	0.650	0.840	0.820	0.987	0.984
300	0.825	0.813	0.948	0.945	0.999	0.999
400	0.901	0.899	0.981	0.980	1	1
500	0.956	0.952	0.996	0.995	1	1
1000	1	0.999	1	1	1	1
2000	1	1	1	1	1	1

Método 1: $H_0 : \kappa_{12} = \kappa_{22}$ vs $H_1 : \kappa_{12} \neq \kappa_{22}$.

Método 2: $H_0 : \ln(\kappa_{12}) = \ln(\kappa_{22})$ vs $H_1 : \ln(\kappa_{12}) \neq \ln(\kappa_{22})$.

Tabla 3.18. Potencias cuando $\kappa_{12} = 0.6$ y $\kappa_{22} = 0.2$.

$Se_1 = 0.8209$ $Sp_1 = 0.8670$ $Se_2 = 0.2091$ $Sp_2 = 0.9715$ $p = 10\%$ $\varepsilon_1 \le 0.0374$ $\varepsilon_0 \le 0.0247$						
	$\varepsilon_1 = 0$ $\varepsilon_0 = 0$		$\varepsilon_1 = 0.015$ $\varepsilon_0 = 0.01$		$\varepsilon_1 = 0.03$ $\varepsilon_0 = 0.02$	
n	Método 1	Método 2	Método 1	Método 2	Método 1	Método 2
100	0.376	0.135	0.359	0.146	0.411	0.164
200	0.805	0.683	0.814	0.693	0.838	0.720
300	0.945	0.914	0.965	0.928	0.874	0.947
400	1	0.978	0.993	0.972	0.996	0.990
500	1	1	1	1	1	1
1000	1	1	1	1	1	1
2000	1	1	1	1	1	1
$Se_1 = 0.8864$ $Sp_1 = 0.6746$ $Se_2 = 0.3019$ $Sp_2 = 0.9030$ $p = 30\%$ $\varepsilon_1 \le 0.0342$ $\varepsilon_0 \le 0.0654$						
	$\varepsilon_1 = 0$ $\varepsilon_0 = 0$		$\varepsilon_1 = 0.015$ $\varepsilon_0 = 0.03$		$\varepsilon_1 = 0.03$ $\varepsilon_0 = 0.06$	
n	Método 1	Método 2	Método 1	Método 2	Método 1	Método 2
100	0.855	0.761	0.901	0.787	0.913	0.817
200	0.992	0.990	0.998	0.994	0.999	0.995
300	1	1	1	1	1	1
400	1	1	1	1	1	1
500	1	1	1	1	1	1
1000	1	1	1	1	1	1
2000	1	1	1	1	1	1
$Se_1 = 0.9315$ $Sp_1 = 0.5421$ $Se_2 = 0.4237$ $Sp_2 = 0.8131$ $p = 50\%$ $\varepsilon_1 \le 0.0290$ $\varepsilon_0 \le 0.1013$						
	$\varepsilon_1 = 0$ $\varepsilon_0 = 0$		$\varepsilon_1 = 0.01$ $\varepsilon_0 = 0.04$		$\varepsilon_1 = 0.02$ $\varepsilon_0 = 0.08$	
n	Método 1	Método 2	Método 1	Método 2	Método 1	Método 2
100	0.916	0.874	0.945	0.910	0.969	0.927
200	1	0.998	0.997	0.997	0.999	0.997
300	1	1	1	1	1	1
400	1	1	1	1	1	1
500	1	1	1	1	1	1
1000	1	1	1	1	1	1
2000	1	1	1	1	1	1

Método 1: $H_0 : \kappa_{12} = \kappa_{22}$ vs $H_1 : \kappa_{12} \neq \kappa_{22}$.

Método 2: $H_0 : \ln(\kappa_{12}) = \ln(\kappa_{22})$ vs $H_1 : \ln(\kappa_{12}) \neq \ln(\kappa_{22})$.

Tabla 3.19. Potencias cuando $\kappa_{12} = 0.8$ y $\kappa_{22} = 0.4$.

$Se_1 = 0.9569$ $Sp_1 = 0.9224$ $Se_2 = 0.7773$ $Sp_2 = 0.7308$ $p = 10\%$ $\varepsilon_1 \le 0.0335$ $\varepsilon_0 \le 0.0567$						
	$\varepsilon_1 = 0$ $\varepsilon_0 = 0$		$\varepsilon_1 = 0.015$ $\varepsilon_0 = 0.02$		$\varepsilon_1 = 0.03$ $\varepsilon_0 = 0.04$	
n	Método 1	Método 2	Método 1	Método 2	Método 1	Método 2
100	0.776	0.654	0.803	0.677	0.851	0.729
200	0.991	0.990	0.997	0.997	1	1
300	1	1	1	1	1	1
400	1	1	1	1	1	1
500	1	1	1	1	1	1
1000	1	1	1	1	1	1
2000	1	1	1	1	1	1
$Se_1 = 0.9102$ $Sp_1 = 0.8773$ $Se_2 = 0.7021$ $Sp_2 = 0.6817$ $p = 30\%$ $\varepsilon_1 \le 0.0631$ $\varepsilon_0 \le 0.0837$						
	$\varepsilon_1 = 0$ $\varepsilon_0 = 0$		$\varepsilon_1 = 0.025$ $\varepsilon_0 = 0.035$		$\varepsilon_1 = 0.05$ $\varepsilon_0 = 0.07$	
n	Método 1	Método 2	Método 1	Método 2	Método 1	Método 2
100	0.900	0.883	0.931	0.920	0.961	0.955
200	0.993	0.993	0.995	0.995	0.995	0.995
300	1	1	1	1	1	1
400	1	1	1	1	1	1
500	1	1	1	1	1	1
1000	1	1	1	1	1	1
2000	1	1	1	1	1	1
$Se_1 = 0.9392$ $Sp_1 = 0.8063$ $Se_2 = 0.8112$ $Sp_2 = 0.5293$ $p = 50\%$ $\varepsilon_1 \le 0.0493$ $\varepsilon_0 \le 0.1025$						
	$\varepsilon_1 = 0$ $\varepsilon_0 = 0$		$\varepsilon_1 = 0.02$ $\varepsilon_0 = 0.04$		$\varepsilon_1 = 0.04$ $\varepsilon_0 = 0.08$	
n	Método 1	Método 2	Método 1	Método 2	Método 1	Método 2
100	0.939	0.925	0.966	0.958	0.971	0.966
200	0.998	0.998	0.998	0.998	0.996	0.996
300	1	1	1	1	1	1
400	1	1	1	1	1	1
500	1	1	1	1	1	1
1000	1	1	1	1	1	1
2000	1	1	1	1	1	1

Método 1: $H_0 : \kappa_{12} = \kappa_{22}$ vs $H_1 : \kappa_{12} \neq \kappa_{22}$.

Método 2: $H_0 : \ln(\kappa_{12}) = \ln(\kappa_{22})$ vs $H_1 : \ln(\kappa_{12}) \neq \ln(\kappa_{22})$.

Tabla 3.20. Potencias cuando $\kappa_{12} = 0.8$ y $\kappa_{22} = 0.6$.

$Se_1 = 0.9569$ $Sp_1 = 0.9224$ $Se_2 = 0.8209$ $Sp_2 = 0.8670$ $p = 10\%$ $\varepsilon_1 \le 0.0354$ $\varepsilon_0 \le 0.0672$						
	$\varepsilon_1 = 0$ $\varepsilon_0 = 0$		$\varepsilon_1 = 0.015$ $\varepsilon_0 = 0.03$		$\varepsilon_1 = 0.03$ $\varepsilon_0 = 0.06$	
n	Método 1	Método 2	Método 1	Método 2	Método 1	Método 2
100	0.126	0.069	0.122	0.065	0.088	0.044
200	0.535	0.500	0.591	0.551	0.646	0.584
300	0.776	0.755	0.852	0.837	0.931	0.923
400	0.905	0.896	0.951	0.947	0.987	0.987
500	0.956	0.953	0.985	0.985	0.998	0.998
1000	1	0.999	1	1	1	1
2000	1	1	1	1	1	1
$Se_1 = 0.9707$ $Sp_1 = 0.7969$ $Se_2 = 0.7923$ $Sp_2 = 0.7940$ $p = 30\%$ $\varepsilon_1 \le 0.0232$ $\varepsilon_0 \le 0.1613$						
	$\varepsilon_1 = 0$ $\varepsilon_0 = 0$		$\varepsilon_1 = 0.01$ $\varepsilon_0 = 0.07$		$\varepsilon_1 = 0.02$ $\varepsilon_0 = 0.14$	
n	Método 1	Método 2	Método 1	Método 2	Método 1	Método 2
100	0.414	0.352	0.447	0.384	0.510	0.435
200	0.843	0.813	0.880	0.863	0.954	0.949
300	0.944	0.937	0.980	0.978	0.998	0.998
400	0.984	0.983	0.996	0.995	1	1
500	0.997	0.996	1	1	1	1
1000	1	1	1	1	1	1
2000	1	1	1	1	1	1
$Se_1 = 0.9814$ $Sp_1 = 0.6996$ $Se_2 = 0.8624$ $Sp_2 = 0.6816$ $p = 50\%$ $\varepsilon_1 \le 0.0160$ $\varepsilon_0 \le 0.2047$						
	$\varepsilon_1 = 0$ $\varepsilon_0 = 0$		$\varepsilon_1 = 0.007$ $\varepsilon_0 = 0.09$		$\varepsilon_1 = 0.014$ $\varepsilon_0 = 0.18$	
n	Método 1	Método 2	Método 1	Método 2	Método 1	Método 2
100	0.467	0.414	0.555	0.508	0.629	0.561
200	0.856	0.837	0.918	0.909	0.982	0.980
300	0.955	0.950	0.991	0.990	0.998	0.998
400	0.993	0.993	0.996	0.996	1	1
500	0.996	0.996	0.999	0.999	1	1
1000	1	1	1	1	1	1
2000	1	1	1	1	1	1

Método 1: $H_0 : \kappa_{12} = \kappa_{22}$ vs $H_1 : \kappa_{12} \neq \kappa_{22}$.

Método 2: $H_0 : \ln(\kappa_{12}) = \ln(\kappa_{22})$ vs $H_1 : \ln(\kappa_{12}) \neq \ln(\kappa_{22})$.

3.5.3.3. Conclusiones

Los resultados de los experimentos de simulación realizados han demostrado que los test de hipótesis de comparación de los coeficientes kappa promedios, tanto si $L' > L$ como si $L > L'$, tienen un comportamiento asintótico que valida su aplicación en la práctica. Los errores tipo I de ambos test de hipótesis (sin transformación y con la transformación logarítmica) no desbordan al error nominal, y ambos test tienen unas potencias elevadas con un tamaño muestral no excesivamente grande.

Comparando los comportamientos asintóticos del test de hipótesis sin transformación y del test con la transformación logarítmica, ambos test tienen un comportamiento muy similar (sobre todo para tamaños muestrales grandes, $n \geq 300$ o 400) tanto en términos de error tipo I como de potencia. Para muestras de tamaño no muy grande (100 o 200 individuos), el error tipo I del test con la transformación logarítmica es levemente menor que el error tipo I del test sin transformación (el cual no desborda al error nominal) y la potencia de este último es más alta que la del primero. Por consiguiente, si bien no hay una gran diferencia entre los comportamientos de ambos test de hipótesis, por su facilidad de interpretación es preferible utilizar el test sin transformación.

3.5.4. El programa "cakctbt"

El programa "cakctbt" (comparison of average kappa coefficients of two binary diagnostic test) es un programa escrito en R que resuelve los test de hipótesis que contrasta la igualdad de los coeficientes kappa promedios de dos *TDBs*, es decir

$$H_0 : \kappa_{11} = \kappa_{21} \quad \text{vs} \quad H_0 : \kappa_{11} \neq \kappa_{21}$$

y

$$H_0 : \kappa_{12} = \kappa_{22} \quad \text{vs} \quad H_0 : \kappa_{12} \neq \kappa_{22}$$

Este programa se ejecuta con el comando

$$\text{cakctbt}\left(s_{11}, s_{10}, s_{01}, s_{00}, r_{11}, r_{10}, r_{01}, r_{00}\right)$$

cuando el error α es igual al 5%, y con el comando

$$\text{cakctbt}\left(s_{11}, s_{10}, s_{01}, s_{00}, r_{11}, r_{10}, r_{01}, r_{00}, \alpha\right)$$

cuando el error α es diferente al 5%. El programa proporciona las estimaciones de cada coeficiente kappa promedio y su respectivo error estándar, el valor del estadístico y el P-valor de cada test de hipótesis. También proporciona los intervalos de confianza para la diferencia de los dos coeficientes kappa promedios en cada situación ($L' > L$ y $L > L'$). Los

resultados obtenidos al ejecutar el programa se guardan en un fichero denominado "Results_cakctbt.txt" en la misma carpeta desde donde se ejecuta el programa. El código de este programa se muestra en el Apéndice II.

3.6. Ejemplo

Los resultados de las Secciones anteriores se han aplicado al estudio de Weiner et al (1979) sobre el diagnóstico de la enfermedad coronaria, que es un ejemplo clásico cuando se comparan parámetros de dos *TDBs* bajo un diseño apareado. En la Tabla 3.21 se muestran los resultados al aplicar dos *TDBs*, una prueba de esfuerzo cardiaco y la historia clínica de enfermedad coronaria del individuo, y el *GE* (arteriografía coronaria) a una muestra de 871 individuos, y donde la variable T_1 modeliza el resultado de la prueba de esfuerzo (test 1), T_2 modeliza el resultado de la historia clínica de enfermedad coronaria del individuo (test 2) y la variable D modeliza el resultado de la angiografía coronaria.

Tabla 3.21. Datos del estudio de Weiner et al (1979).

	$T_1 = 1$		$T_1 = 0$		
	$T_2 = 1$	$T_2 = 0$	$T_2 = 1$	$T_2 = 0$	Total
$D = 1$	473	29	81	25	608
$D = 0$	22	46	44	151	263
Total	495	75	125	176	871

En la Tablas 3.22 a 3.28 se muestran las estimaciones de los parámetros, los resultados de los test de hipótesis $(\alpha = 5\%)$ y los intervalos de confianza al 95%. A partir de estos resultados se obtienen las siguientes conclusiones:

a) Comparación de las sensibilidades y especificidades (Tabla 3.22). Se rechaza el test de hipótesis global de igualdad de ambas sensibilidades y especificidades. La sensibilidad del test 2 (historia clínica de enfermedad coronaria) es significativamente mayor que la del test 1 (prueba de esfuerzo), un valor entre el 5.21% y el 11.84% mayor con una confianza del 95%; y no se rechaza la igualdad entre las dos especificidades. La historia clínica de enfermedad coronaria es un test más útil para descartar la enfermedad coronaria que la prueba de esfuerzo, no rechazándose que ambos test diagnósticos

son igualmente útiles para confirmar dicha enfermedad. Por tanto, al no rechazarse la igualdad de las especificidades y al ser la sensibilidad de la historia clínica de enfermedad coronaria mayor que la de la prueba de esfuerzo, la historia clínica de enfermedad coronaria es una prueba diagnóstica mejor que la prueba de esfuerzo para el diagnóstico de la enfermedad coronaria.

Tabla 3.22. Comparación de las sensibilidades y de las especificidades.

$\hat{Se}_1 = 82.57\%$	$\hat{Var}\left(\hat{Se}_1\right) = 0.000237$	$\hat{Sp}_1 = 74.14\%$	$\hat{Var}\left(\hat{Sp}_1\right) = 0.000729$
$\hat{Se}_2 = 91.12\%$	$\hat{Var}\left(\hat{Se}_2\right) = 0.000133$	$\hat{Sp}_2 = 74.90\%$	$\hat{Var}\left(\hat{Sp}_2\right) = 0.000714$
$H_0 : Se_1 = Se_2 \cap Sp_1 = Sp_2$ vs $H_1 : Se_1 \neq Se_2 \cup Sp_1 \neq Sp_2$			
$Q^2_{\text{exp}} = 24.63$ P-valor $= 4.49 \times 10^{-6}$			
$H_0 : Se_1 = Se_2$ vs $H_1 : Se_1 \neq Se_2$			
$z_{\text{exp}} = 4.91$ P-valor $= 9.09 \times 10^{-7}$			
$Se_1 - Se_2 \in (-11.84\% \ ; \ -5.21\%)$ al 95% de confianza			
$H_0 : Sp_1 = Sp_2$ vs $H_1 : Sp_1 \neq Sp_2$			
$z_{\text{exp}} = 0.16$ P-valor $= 0.87$			
$Sp_1 - Sp_2 \in (-6.30\% \ ; \ 7.81\%)$ al 95% de confianza			

b) Comparación de las razones de verosimilitud (Tabla 3.23). Se rechaza el test de hipótesis global de igualdad de las dos razones de verosimilitud positivas y de las dos razones de verosimilitud negativas. Resolviendo los test individuales y aplicando algún

método de comparaciones múltiples (por ejemplo, Holm u Hochberg), no se rechaza la igualdad de las dos razones de verosimilitud positivas y se rechaza la igualdad de las dos razones de verosimilitud negativas. La razón de verosimilitud negativa de la prueba de esfuerzo es significativamente mayor que la del test 2, un valor entre 1.487 y 2.644 veces mayor con una confianza del 95%. Por tanto, un resultado negativo de la prueba de esfuerzo es más indicativo de la ausencia de la enfermedad coronaria que un resultado negativo de la historia clínica de la enfermedad coronaria.

Tabla 3.23. Comparación de las razones de verosimilitud.

$\hat{LR}_1(+) = 3.193$	$\hat{Var}\left(\hat{LR}_1(+)\right) = 0.11473$	$\widehat{LR}_1(-) = 0.235$	$\hat{Var}\left(\hat{LR}_1(-)\right) = 0.000504$
$\hat{LR}_2(+) = 3.631$	$\hat{Var}\left(\hat{LR}_2(+)\right) = 0.15174$	$\hat{LR}_2(-) = 0.119$	$\hat{Var}\left(\hat{LR}_2(-)\right) = 0.000255$
$H_0 : \left(\omega^- = 0\right) \cap \left(\omega^+ = 0\right)$ vs $H_1 : \left(\omega^- \neq 0\right) \cup \left(\omega^+ \neq 0\right)$			
$Q^2_{\text{exp}} = 23.44$ P-valor $= 8.14 \times 10^{-6}$			
$H_0 : \omega^+ = 0$ vs $H_1 : \omega^+ \neq 0$			
$z_{\text{exp}} = 0.90$ P-valor $= 0.37$			
$\frac{LR_1(+)}{LR_2(+)} \in (0.665 \; ; \; 1.164)$ al 95% de confianza			
$H_0 : \omega^- = 0$ vs $H_1 : \omega^- \neq 0$			
$z_{\text{exp}} = 4.66$ P-valor $= 3.12 \times 10^{-6}$			
$\frac{LR_1(-)}{LR_2(-)} \in (1.487 \; ; \; 2.644)$ al 95% de confianza			

c) Comparación de los valores predictivos (Tabla 3.24). Se rechaza el test de hipótesis global de igualdad de los dos valores predictivos positivos y de los dos valores predictivos negativos. Resolviendo los test individuales aplicando el método de Kosinski y aplicando un método de comparaciones múltiples (Holm u Hochberg), no se rechaza la igualdad de los dos valores predictivos positivos y se rechaza la igualdad de los dos valores predictivos negativos. El valor predictivo negativo de la historia clínica de enfermedad coronaria es significativamente mayor que el de la prueba de esfuerzo, un valor entre el 8.04% y el 19.36% con una confianza del 95%. Por tanto, cuando ambos test se aplican a la población de donde se ha extraído la muestra, la historia clínica de enfermedad coronaria es más útil para descartar la enfermedad que la prueba de esfuerzo, no rechazándose que ambos test diagnósticos son igualmente útiles para confirmar la enfermedad coronaria (pues no se rechaza la igualdad de los dos valores predictivos positivos).

Tabla 3.24. Comparación de los valores predictivos.

$\hat{V}PP_1 = 88.07\%$ $\hat{V}ar\left(\hat{V}PP_1\right) = 0.000184$ $\hat{V}PN_1 = 64.78\%$ $\hat{V}ar\left(\hat{V}PN_1\right) = 0.000758$
$\hat{V}PP_2 = 89.35\%$ $\hat{V}ar\left(\hat{V}PP_2\right) = 0.000153$ $\hat{V}PN_2 = 78.49\%$ $\hat{V}ar\left(\hat{V}PN_2\right) = 0.000673$
$H_0 : \left(VPP_1 = VPP_2 \cap VPN_1 = VPN_2\right)$ vs $H_1 : \left(VPP_1 \neq VPP_2 \cup VPN_1 \neq VPN_2\right)$ $Q^2_{\text{exp}} = 25.94$ P-valor $= 2.32 \times 10^{-6}$
$H_0 : VPP_1 = VPP_2$ vs $H_1 : VPP_1 \neq VPP_2$ $T^{WGS}_{VPP} = 0.81$ P-valor $= 0.37$ $VPP_1 - VPP_2 \in (-5.21\% \ ; \ 2.64\%)$ al 95% de confianza
$H_0 : VPN_1 = VPN_2$ vs $H_1 : VPN_1 \neq VPN_2$ $T^{WGS}_{VPP} = 22.50$ P-valor $= 2.1 \times 10^{-6}$ $VPN_1 - VPN_2 \in (-19.36\% \ ; \ -8.04\%)$ al 95% de confianza

d) Comparación de los coeficientes kappa ponderados. En las Tablas 3.25, 3.26 y 3.27 se muestran los resultados obtenidos al comparar los coeficientes kappa ponderados para diferentes valores del índice de ponderación $(c = 0.1,\ 0.2,...,\ 0.9)$. Cuando $L' > L$, y por tanto $0 < c < 0.5$ (Tabla 3.25), si el índice de ponderación es muy bajo $(c = 0.1)$ no se rechaza la igualdad de los dos coeficientes kappa ponderados, para $c = 0.2$ no se rechaza la igualdad de ambos coeficientes kappa ponderados pero hay fuertes indicios de significación y para $0.3 \leq c < 0.5$ se rechaza la igualdad de los dos

coeficientes kappa ponderados, obteniéndose que el coeficiente kappa ponderado de la historia clínica de enfermedad coronaria es mayor que el de la prueba de esfuerzo. En la Tabla 3.26 se muestran los resultados en el caso de $L = L'$ $(c = 0.5)$, también se rechaza la igualdad de los dos coeficientes kappa ponderados y la conclusión es la misma. Para $L > L'$ $(0.5 < c < 1)$, Tabla 3.27, se rechaza siempre la igualdad de los dos coeficientes kappa ponderados, obteniéndose que el coeficiente kappa ponderado de la historia clínica de enfermedad coronaria es significativamente mayor que el de la prueba de esfuerzo. Por tanto, si el clínico tiene una mayor preocupación por los falsos negativos que por los falsos positivos $(L > L')$, entonces el acuerdo más allá del azar entre la historia clínica de enfermedad coronaria y la angiografía (que es siempre "bueno" en términos de los valores de los estimadores puntuales) es significativamente mayor que el acuerdo más allá del azar entre la prueba de esfuerzo y la angiografía (que es siempre "moderado"). Si el clínico tiene una mayor preocupación por los falsos positivos que por los falsos negativos $(L' > L)$, entonces el acuerdo más allá del azar entre la historia clínica de enfermedad coronaria y la

angiografía es significativamente mayor que el acuerdo más allá del azar entre la prueba de esfuerzo y la angiografía dependiendo del valor del índice de ponderación.

Tabla 3.25. Comparación de los coeficientes kappa ponderados cuando $0 < c < 0.5$.

c = 0.1
$\hat{\kappa}_1(c) = 0.592 \quad \hat{Var}(\hat{\kappa}_1(c)) = 0.001191 \quad \hat{\kappa}_2(c) = 0.652 \quad \hat{Var}(\hat{\kappa}_2(c)) = 0.001038$
$H_0 : \kappa_1(c) = \kappa_2(c)$ vs $H_1 : \kappa_1(c) \neq \kappa_2(c) \quad z_{\exp} = 1.35$ P-valor $= 0.18$ $\kappa_1(c) - \kappa_2(c) \in (-0.1469 \; ; \; 0.0273)$ al 95% de confianza
c = 0.2
$\hat{\kappa}_1(c) = 0.579 \quad \hat{Var}(\hat{\kappa}_1(c)) = 0.001055 \quad \hat{\kappa}_2(c) = 0.656 \quad \hat{Var}(\hat{\kappa}_2(c)) = 0.000093$
$H_0 : \kappa_1(c) = \kappa_2(c)$ vs $H_1 : \kappa_1(c) \neq \kappa_2(c) \quad z_{\exp} = 1.82$ P-valor $= 0.07$ $\kappa_1(c) - \kappa_2(c) \in (-0.1587 \; ; \; 0.0057)$ al 95% de confianza
c = 0.3
$\hat{\kappa}_1(c) = 0.567 \quad \hat{Var}(\hat{\kappa}_1(c)) = 0.000970 \quad \hat{\kappa}_2(c) = 0.660 \quad \hat{Var}(\hat{\kappa}_2(c)) = 0.000850$
$H_0 : \kappa_1(c) = \kappa_2(c)$ vs $H_1 : \kappa_1(c) \neq \kappa_2(c) \quad z_{\exp} = 2.32$ P-valor $= 0.02$ $\kappa_1(c) - \kappa_2(c) \in (-0.1711 \; ; \; -0.0145)$ al 95% de confianza
c = 0.4
$\hat{\kappa}_1(c) = 0.556 \quad \hat{Var}(\hat{\kappa}_1(c)) = 0.000926 \quad \hat{\kappa}_2(c) = 0.664 \quad \hat{Var}(\hat{\kappa}_2(c)) = 0.000795$
$H_0 : \kappa_1(c) = \kappa_2(c)$ vs $H_1 : \kappa_1(c) \neq \kappa_2(c) \quad z_{\exp} = 2.83$ P-valor $= 0.005$ $\kappa_1(c) - \kappa_2(c) \in (-0.1840 \; ; \; -0.0333)$ al 95% de confianza

Tabla 3.26. Comparación de los coeficientes kappa ponderados cuando $c = 0.5$.

$\hat{\kappa}_1(c) = 0.545 \quad \hat{Var}(\hat{\kappa}_1(c)) = 0.000145 \quad \hat{\kappa}_2(c) = 0.669 \quad \hat{Var}(\hat{\kappa}_2(c)) = 0.000770$
$H_0 : \kappa_1(c) = \kappa_2(c)$ vs $H_1 : \kappa_1(c) \neq \kappa_2(c) \quad z_{\exp} = 3.31$ P-valor $< 10^{-3}$
$\kappa_1(c) - \kappa_2(c) \in (-0.1975\ ;\ -0.0507)$ al 95% de confianza

Tabla 3.27. Comparación de los coeficientes kappa ponderados cuando $0.5 < c < 1$.

c = 0.6
$\hat{\kappa}_1(c) = 0.534 \quad \hat{Var}(\hat{\kappa}_1(c)) = 0.000929 \quad \hat{\kappa}_2(c) = 0.673 \quad \hat{Var}(\hat{\kappa}_2(c)) = 0.000777$
$H_0 : \kappa_1(c) = \kappa_2(c)$ vs $H_1 : \kappa_1(c) \neq \kappa_2(c) \quad z_{\exp} = 3.77$ P-valor $< 10^{-3}$
$\kappa_1(c) - \kappa_2(c) \in (-0.2116\ ;\ -0.0668)$ al 95% de confianza
c = 0.7
$\hat{\kappa}_1(c) = 0.524 \quad \hat{Var}(\hat{\kappa}_1(c)) = 0.000962 \quad \hat{\kappa}_2(c) = 0.678 \quad \hat{Var}(\hat{\kappa}_2(c)) = 0.000819$
$H_0 : \kappa_1(c) = \kappa_2(c)$ vs $H_1 : \kappa_1(c) \neq \kappa_2(c) \quad z_{\exp} = 4.17$ P-valor $< 10^{-4}$
$\kappa_1(c) - \kappa_2(c) \in (-0.2264\ ;\ -0.0815)$ al 95% de confianza
c = 0.8
$\hat{\kappa}_1(c) = 0.514 \quad \hat{Var}(\hat{\kappa}_1(c)) = 0.001011 \quad \hat{\kappa}_2(c) = 0.682 \quad \hat{Var}(\hat{\kappa}_2(c)) = 0.000899$
$H_0 : \kappa_1(c) = \kappa_2(c)$ vs $H_1 : \kappa_1(c) \neq \kappa_2(c) \quad z_{\exp} = 4.49$ P-valor $< 10^{-5}$
$\kappa_1(c) - \kappa_2(c) \in (-0.2418\ ;\ -0.0949)$ al 95% de confianza
c = 0.9
$\hat{\kappa}_1(c) = 0.505 \quad \hat{Var}(\hat{\kappa}_1(c)) = 0.001071 \quad \hat{\kappa}_2(c) = 0.687 \quad \hat{Var}(\hat{\kappa}_2(c)) = 0.001019$
$H_0 : \kappa_1(c) = \kappa_2(c)$ vs $H_1 : \kappa_1(c) \neq \kappa_2(c) \quad z_{\exp} = 4.74$ P-valor $< 10^{-5}$
$\kappa_1(c) - \kappa_2(c) \in (-0.2578\ ;\ -0.1071)$ al 95% de confianza

e) Comparación de los coeficientes kappa promedios (Tabla 3.28). Si el clínico tiene una mayor preocupación por los falsos positivos que por los falsos negativos $(L' > L)$, entonces se rechaza la igualdad de los dos coeficientes kappa promedio, obteniéndose que el coeficiente kappa promedio de la historia clínica de la enfermedad coronaria (que toma un valor "bueno" en términos de la estimación puntual) es significativamente mayor que el de la prueba de esfuerzo (que toma un valor "moderado" en términos de la estimación puntual). Por tanto, el acuerdo promedio más allá del azar entre la historia clínica de enfermedad coronaria y la angiografía es, con una confianza del 95%, un valor entre 0.0041 y 0.1644 mayor que el acuerdo promedio más allá del azar entre la prueba de esfuerzo y la angiografía.

Si el clínico tiene una mayor preocupación por los falsos negativos que por los falsos positivos $(L > L')$, entonces se rechaza la igualdad de los dos coeficientes kappa promedio, obteniéndose que el coeficiente kappa promedio de la historia clínica de la enfermedad coronaria (que toma un valor "bueno" en términos de la estimación puntual) es significativamente mayor que el de la prueba de esfuerzo (que toma un valor "moderado" en términos de la estimación

puntual). Por tanto, el acuerdo promedio más allá del azar entre la historia clínica de enfermedad coronaria y la angiografía es, con una confianza del 95%, un valor entre 0.0881 y 0.2336 mayor que el acuerdo promedio más allá del azar entre la prueba de esfuerzo y la angiografía.

Las conclusiones obtenidas en las comparaciones de los coeficientes kappa promedios son concordantes con las obtenidas al comparar las sensibilidades, especificidades y los valores predictivos. Los resultados obtenidos permiten concluir que la historia clínica de enfermedad coronaria es un método mejor que la prueba de esfuerzo para el diagnóstico de la enfermedad coronaria.

Tabla 3.28. Comparación de los coeficientes kappa promedios.

$\boldsymbol{L' > L}$			
$\hat{\kappa}_{11} = 0.574$	$\hat{Var}(\hat{\kappa}_{11}) = 0.031820$	$\hat{\kappa}_{21} = 0.658$	$\hat{Var}(\hat{\kappa}_{11}) = 0.029746$
$H_0 : \kappa_{11} = \kappa_{21}$ vs $H_1 : \kappa_{11} \neq \kappa_{21}$			
$z_{\exp} = 2.06$ P-valor $= 0.039$			
$\kappa_{21} - \kappa_{11} \in (0.0041\ ;\ 0.1644)$ al 95% de confianza			
$\boldsymbol{L > L'}$			
$\hat{\kappa}_{12} = 0.519$	$\hat{Var}(\hat{\kappa}_{12}) = 0.031303$	$\hat{\kappa}_{22} = 0.680$	$\hat{Var}(\hat{\kappa}_{22}) = 0.029260$
$H_0 : \kappa_{12} = \kappa_{22}$ vs $H_1 : \kappa_{12} \neq \kappa_{22}$			
$z_{\exp} = 4.33$ P-valor $= 1.46 \times 10^{-5}$			
$\kappa_{22} - \kappa_{12} \in (0.0881\ ;\ 0.2336)$ al 95% de confianza			

Conclusiones

Los parámetros fundamentales para evaluar y comparar el rendimiento de test diagnósticos binarios son la sensibilidad y la especificidad, que dependen únicamente de la capacidad del propio test para distinguir entre individuos con la enfermedad e individuos sin la enfermedad. El resto de parámetros clásicos (razones de verosimilitud y valores predictivos) dependen a su vez de la sensibilidad y especificidad del test (los valores predictivos también dependen de la prevalencia de la enfermedad). Cuando se consideran las pérdidas de una clasificación errónea con el test diagnóstico, el parámetro adecuado para evaluar el rendimiento del test diagnóstico es el coeficiente kappa ponderado. Este parámetro depende de la sensibilidad y especificidad del test, de la prevalencia de la enfermedad y de la importancia relativa entre los falsos positivos y los falsos negativos (índice de ponderación c). El problema que presenta el uso del coeficiente kappa ponderado es la asignación de valores al índice de ponderación. El

asignar valores 0 o 1 es una cuestión muy extrema, pues asume que una de las pérdidas $(L \text{ o } L')$ es 0, lo que no es realista. Asimismo, el clínico no siempre tiene un conocimiento del problema que le permita determinar cuánto es mayor una pérdida que la otra y por tanto asignar un valor al índice de ponderación. Una solución a este problema es definir un coeficiente kappa que no depende del índice de ponderación: el coeficiente kappa promedio. Este nuevo parámetro es un promedio de coeficientes kappa ponderados en dos situaciones distintas, una cuando $L' > L$ $(0 < c < 0.5)$ y otra cuando $L > L'$ $(0.5 < c < 1)$. El clínico debe determinar cada situación; así, si el test diagnóstico se utiliza como un test definitivo previo a un tratamiento de riesgo (por ejemplo, un test confirmatorio antes de una operación quirúrgica) entonces el clínico tiene una mayor preocupación por los falsos positivos y $L' > L$, y si el test se utiliza como un test de screening (por ejemplo, la mamografía para el diagnóstico del cáncer de mama en mujeres mayores de 55 años) entonces el clínico tiene una mayor preocupación por los falsos negativos y $L > L'$. El coeficiente kappa promedio (en cada una de las dos situaciones anteriores) depende únicamente de la sensibilidad y especificidad del test diagnóstico y de la prevalencia de la enfermedad (mismos parámetros de los que dependen los

valores predictivos), y se define como una medida del acuerdo promedio más allá del azar entre el test diagnóstico y el gold estándar. Esta Tesis Doctoral se ha centrado en el estudio de este nuevo parámetro.

En el Capítulo 1 se han definido los parámetros de un test diagnóstico binario y se han presentado sus propiedades, y se ha definido y caracterizado el nuevo parámetro objeto principal de esta Tesis: el coeficiente kappa promedio. Entre las propiedades de este nuevo parámetro cabe destacar que a partir de su estimación se puede calcular el valor del índice de ponderación cuando $L' > L$ o $L > L'$ y por tanto determinar cuánto es mayor una pérdida que la otra. Las propiedades del coeficiente kappa promedio validan a esta medida como un parámetro válido para evaluar y comparar el rendimiento de test diagnósticos binarios.

En el Capítulo 2 se han estudiado las estimaciones e intervalos de confianza de los parámetros definidos en el Capítulo 1 bajo un muestreo transversal. La aportación realizada en este Capítulo es la estimación del coeficiente kappa promedio. Se ha estudiado su estimación puntual y se han propuesto tres intervalos de confianza para este parámetro: un intervalo tipo Wald, un intervalo logit y un intervalo mediante bootstrap. Se han realizado experimentos de simulación para estudiar la cobertura asintótica de estos

intervalos, obteniéndose que en términos generales el intervalo de confianza tipo Wald es preferible para una muestra de tamaño 100 a 500, y que para muestras de mayor tamaño se puede utilizar cualquiera de los tres intervalos de confianza, si bien el intervalo bootstrap requiere un mayor esfuerzo computacional que los otros dos intervalos.

Finalmente, en el Capítulo 3 se ha estudiado la comparación de los parámetros definidos en el Capítulo 1 de dos test diagnósticos binarios (y en algunos casos de más de dos test) bajo un diseño apareado. La aportación en este Capítulo ha sido la comparación de dos coeficientes kappa promedios. Se han propuesto dos test de hipótesis para resolver este problema, uno sin transformar los coeficientes kappa promedios y otro utilizando la transformación del logaritmo neperiano, en ambos casos cuando $L' > L$ y cuando $L > L'$. Los estadísticos de contraste de estos test de hipótesis se basan en la aproximación a la distribución normal, y a partir de ellos se pueden obtener unos intervalos de confianza para la diferencia de los dos coeficientes kappa promedios y para el cociente de ambos. Se han realizado experimentos de simulación Monte Carlo para estudiar el error tipo I y la potencia de estos test de hipótesis al error $\alpha = 5\%$. De los resultados de estos experimentos se ha obtenido que ambos test de hipótesis tienen un

comportamiento asintótico muy similar (en ambas situaciones, $L' > L$ y $L > L'$), sus errores tipo I no desbordan al error nominal y sus potencias son, en términos generales, elevadas sin necesidad de un tamaño muestral demasiado grande. Como ambos test de hipótesis tienen un comportamiento asintótico muy similar, por una interpretación más sencilla, es preferible utilizar el test de hipótesis sin transformación que con la transformación logarítmica. Estos test de hipótesis se han extendido a la situación en la que se comparan más de dos test diagnósticos binarios. En esta situación se resuelve un test de hipótesis global (sin transformación o con la transformación logarítmica) basado en la aproximación a la distribución chi-cuadrado. Si el test de hipótesis global es significativo al error α, la investigación de las causas de la significación se realiza comparando los coeficientes kappa promedios dos a dos y aplicando algún método de comparaciones múltiples.

Las dos aportaciones realizadas en esta Tesis Doctoral se han aplicado a ejemplos reales de la Medicina, utilizando para ello dos programas escritos en *R*.

Apéndice I: Programa “akcbdt”

```
akcbdt <- function (s1, s0, r1, r0, conflevel = 0.95, B = 2000)
{
  root <- function (x)
    {
    f <- function (y) { pnorm (y) - (1 - x / B) }
    uniroot(f, c(0,1), lower = -10000, upper = 10000, maxiter = 10000)$root
    }
  set.seed(1231)
  list1 <- c()
  list2 <- c()
  list3 <- c()
  x1 <- 0
```

```
x2 <- 0
x3 <- 0
if (conflevel >= 1 | conflevel <= 0)
 {
 stop("Confidence level must be a value between 0 and 1. Introduces a new value \n")
 }
if (B <= 0)
 {
 stop("The number of samples with replacement is not correct. Introduce a new value \n")
 }
if (abs(B - trunc (B)) > 0)
 {
 stop("The number of samples with replacement is not correct. Introduce a new value \n")
 }
if (s1 < 0 | s0 < 0 | r1 < 0 | r0 < 0)
 {
 stop("All observed frequencies must be positive integer values. Introduces new values
 \n")
```

```
 }
if (abs(s1 - trunc (s1)) > 0 | abs(s0 - trunc (s0)) > 0 | abs(r1 - trunc (r1)) > 0 | abs(r0 -
trunc(r0)) > 0)
 {
 stop("All observed frequencies must be positive integer values. Introduces new values
 \n")
 }
if (s0 == 0 && r1 == 0)
 {
 stop("Observed frecuencies s0 and r1 can not be zero. Introduces new values \n")
 }
z  = qnorm (1 - (1-conflevel) / 2, 0, 1)
n1 <- s1 + r1
n0 <- s0 + r0
n <- n1 + n0
#Estimated sensitivity, specificity and prevalence
Se <- s1 / (s1 + s0)
Sp <- r0 / (r1 + r0)
p <- (s1 + s0) / n
```

```
Y <- Se + Sp - 1

if (Y <= 0)

  {

  stop("Estimated Youden index must be greater than zero. Introduces new values \n")

  }

VarSe <- Se * (1 - Se) / (s1 + s0)

VarSp <- Sp * (1 - Sp) / (r1 + r0)

Varp <- p * (1 - p) / n

#Yu et al CI for sensitivity

LSe <- 0.5 + (((s1 + s0) + z^4 / 53) * (Se - 0.5)) / ((s1 + s0) + z^2) - (z / ((s1 + s0) + z^2))
* sqrt((s1 + s0) * Se * (1 - Se) + z^2 / 4)

USe <- 0.5 + (((s1 + s0) + z^4 / 53) * (Se - 0.5)) / ((s1 + s0) + z^2) + (z / ((s1 + s0) +
z^2)) * sqrt((s1 + s0) * Se * (1 - Se) + z^2 / 4)

#Yu et al CI for specificity

LSp <- 0.5 + (((r1 + r0) + z^4 / 53) * (Sp - 0.5)) / ((r1 + r0) + z^2) - (z / ((r1 + r0) + z^2))
* sqrt((r1 + r0) * Sp * (1 - Sp) + z^2 / 4)

USp <- 0.5 + (((r1 + r0) + z^4 / 53) * (Sp - 0.5)) / ((r1 + r0) + z^2) + (z / ((r1 + r0) + z^2))
* sqrt((r1 + r0) * Sp * (1 - Sp) + z^2 / 4)

Q <- p * Se + (1 - p) * (1 - Sp)
```

```
#Estimated k(0)

kappa0 <- (Sp - (1 - Q)) / Q

Varkappa0 <- ((1 - Sp)^2 * Y^2 * Varp + p^2 * ((1 - Sp)^2 * VarSe + Se^2 * VarSp)) /
Q^4

#Estimated k(1)

kappa1 <- (Se - Q) / (1 - Q)

Varkappa1 <- ((1 - Se)^2 * Y^2 * Varp + (1 - p)^2 * (Sp^2 * VarSe + (1 - Se)^2 *
VarSp)) / (1 - Q)^4

if (abs(s0 - r1) > 0)

 {

 #Estimated average kappa coefficient k1

 k1 <- (2 * (s1 * r0 - s0 * r1) / (n0 * (s1 + s0) - n1 * (r1 + r0))) * log((n1 * (r1 + r0) + n0
 * (s1 + s0)) / (2 * n1 * (r1 + r0)))

 Vark1 <- (1 /((kappa0 + kappa1)^2 * (kappa0 - kappa1)^2)) * (((2 * kappa0^2 * kappa1
 - kappa1 * (kappa0 + kappa1) * k1) / kappa0)^2 * ((1 - Sp)^2 * Y^2 * Varp + p^2 * ((1
 - Sp)^2 * VarSe + Se^2 * VarSp)) / Q^4 + (kappa0 * ((kappa0 + kappa1) * k1 - 2 *
 kappa0 * kappa1) / kappa1)^2 * ((1 - Se)^2 * Y^2 * Varp + (1 - p)^2 * (Sp^2 * VarSe +
  (1 - Se)^2 * VarSp)) / (1 - Q)^4 + 2 * ((2 * kappa0^2 * kappa1 - kappa1 * (kappa0 +
 kappa1) * k1) / kappa0) * (kappa0 * ((kappa0 + kappa1) * k1 - 2 * kappa0 * kappa1) /
```

```
kappa1) * ((p * (1-p) * ((1-Se) * Se * VarSp + (1-Sp) * Sp * VarSe) - (1-Se) * (1-Sp) *
Y^2 * Varp) / (Q^2 * (1-Q)^2)))

L1 <- k1 - z * sqrt(Vark1)

if (L1 < 0) {L1 <- 0}

L2 <- k1 + z * sqrt(Vark1)

if (L2 > 1) {L2 <- 1}

Varlogitk1 <- (1 / ((kappa0 + kappa1) * (kappa0 - kappa1) * k1 * (1 - k1))^2) *(((2 *
kappa0^2 * kappa1 * (1 - k1) - kappa1 * (kappa0 * (1 - 2 * kappa0) + kappa1) * k1) /
kappa0)^2 * ((1 - Sp)^2 * Y^2 * Varp + p^2 * ((1 - Sp)^2 * VarSe + Se^2 *
VarSp))/Q^4 + ((kappa0 * ((kappa0 + kappa1) * k1 - 2 * kappa0 * kappa1)) / kappa1)^2
* ((1 - Se)^2 * Y^2 * Varp + (1 - p)^2 * (Sp^2 * VarSe + (1 - Se)^2 * VarSp)) / (1 -
Q)^4 +2 * ((2 * kappa0^2 * kappa1 * (1 - k1) - kappa1 * (kappa0 * (1 - 2 * kappa0) +
kappa1) * k1) / kappa0) *((kappa0 * ((kappa0 + kappa1) * k1 - 2 * kappa0 * kappa1)) /
kappa1) * ((p * (1-p) * ((1-Se) * Se * VarSp + (1-Sp) * Sp * VarSe) - (1-Se) * (1-Sp) *
Y^2 * Varp) / (Q^2 * (1-Q)^2)))

#Estimated average kappa coefficient k2

k2 <- (2 * (s1 * r0 - s0 * r1) / (n0 * (s1 + s0) - n1 * (r1 + r0))) * log((2 * n0 * (s1 + s0)) /
(n1 * (r1 + r0) + n0 * (s1 + s0)))

Vark2 <- (1 /((kappa0 + kappa1)^2 * (kappa0 - kappa1)^2)) * ((kappa1 * (2 * kappa0 *
kappa1 - (kappa0 + kappa1) * k2)/ kappa0)^2 * ((1 - Sp)^2 * Y^2 * Varp + p^2 * ((1 -
Sp)^2 * VarSe + Se^2 * VarSp)) / Q^4 + ((kappa0 * (kappa0 + kappa1) * k2 - 2 *
```

```
kappa0 * kappa1^2) / kappa1)^2 * ((1 - Se)^2 * Y^2 * Varp + (1 - p)^2 * (Sp^2 * VarSe
+ (1 - Se)^2 * VarSp)) / (1 - Q)^4 + 2 * (kappa1 * (2 * kappa0 * kappa1 - (kappa0 +
kappa1) * k2)/ kappa0) * ((kappa0 * (kappa0 + kappa1) * k2 - 2 * kappa0 * kappa1^2) /
kappa1) * ((p * (1-p) * ((1-Se) * Se * VarSp + (1-Sp) * Sp * VarSe) - (1-Se) * (1-Sp) *
Y^2 * Varp) / (Q^2 * (1-Q)^2)))

L3 <- k2 - z * sqrt(Vark2)

if (L3 < 0) {L3 <- 0}

L4 <- k2 + z * sqrt(Vark2)

if (L4 > 1) {L4 <- 1}

Varlogitk2 <- (1 / ((kappa0 + kappa1) * (kappa0 - kappa1) * k2 * (1 - k2))^2) *
((kappa1 * (2 * kappa0 * kappa1 - (kappa0 + kappa1) * k2) / kappa0)^2 * ((1 - Sp)^2 *
Y^2 * Varp + p^2 * ((1 - Sp)^2 * VarSe + Se^2 * VarSp))/Q^4 + (kappa0 * ((kappa0 +
kappa1) * k2 - 2 * kappa1^2) / kappa1)^2 * ((1 - Se)^2 * Y^2 * Varp + (1 - p)^2 *
(Sp^2 * VarSe + (1 - Se)^2 * VarSp)) / (1 - Q)^4 + 2 * (kappa1 * (2 * kappa0 *
kappa1 - (kappa0 + kappa1) * k2) / kappa0) * (kappa0 * ((kappa0 + kappa1) * k2 - 2 *
kappa1^2) / kappa1) * ((p * (1-p) * ((1-Se) * Se * VarSp + (1-Sp) * Sp * VarSe) - (1-
Se) * (1-Sp) * Y^2 * Varp) / (Q^2 * (1-Q)^2)))

#Bootstrap

data <- c(s1, s0, r1, r0)

data1 <- matrix(1,1,data[1])
```

```
data2 <- matrix(2,1,data[2])

data3 <- matrix(3,1,data[3])

data4 <- matrix(4,1,data[4])

datatot <- c(data1,data2,data3,data4)

samplesB <- B

while (samplesB >= 1)

 {

 sampleboot <- sample(datatot,length(datatot),replace = TRUE)

 data.sample <- tabulate(sampleboot)

 if (length(data.sample) < length(data)) next

 d1 <- data.sample[1]

 d0 <- data.sample[2]

 e1 <- data.sample[3]

 e0 <- data.sample[4]

 m1 <- d1 + e1

 m0 <- d0 + e0

 Sen <- d1 / (d1 + d0)

 Spe <- e0 / (e1 + e0)
```

```
Yo <- Sen + Spe - 1

if ((d0 == 0 && e1 == 0) | Yo <= 0)  next

if ((d1 > 0 && d0 > 0 && e1 > 0 && e0 > 0 && abs(d0 - e1) > 0 && Yo > 0) | (d0 ==
0 && (d1 * e1 * e0) > 0 && Yo > 0) | (e1 == 0 && (d1 * d0 * e0) > 0 && Yo > 0))

  {

  k1boot <- (2 * (d1 * e0 - d0 * e1) / (m0 * (d1 + d0) - m1 * (e1 + e0))) * log((m1 * (e1
 + e0) + m0 * (d1 + d0)) / (2 * m1 * (e1 + e0)))

  k2boot <- (2 * (d1 * e0 - d0 * e1) / (m0 * (d1 + d0) - m1 * (e1 + e0))) * log((2 * m0 *
 (d1 + d0)) / (m1 * (e1 + e0) + m0 * (d1 + d0)))

  list1[samplesB] <- k1boot

  if (k1boot < k1) { x1 <- x1 + 1 }

  list2[samplesB] <- k2boot

  if (k2boot < k2) { x2 <- x2 + 1 }

  }

if (d1 > 0 && e0 > 0 && (d0 * e1) > 0 && (d0 - e1) == 0 && Yo > 0)

  {

  kboot <- Yo

  list1[samplesB] <- kboot

  if (kboot < k1) { x1 <- x1 + 1 }
```

```
    list2[samplesB] <- kboot
    if (kboot < k2) { x2 <- x2 + 1 }
    }
  samplesB <- samplesB - 1
  }
u1 <- root (x1)
t1 <- pnorm (2 * u1 - z)
a1 <- quantile (list1, t1, names = FALSE)
t2 <- pnorm (2 * u1 + z)
a2 <- quantile (list1, t2, names = FALSE)
u2 <- root (x2)
v1 <- pnorm (2 * u2 - z)
b1 <- quantile (list2, v1, names = FALSE)
v2 <- pnorm (2 * u2 + z)
b2 <- quantile (list2, v2, names = FALSE)
#Result
sink("Results.txt", split=TRUE)
cat("\n")
```

```
cat("        R E S U L T S \n")

cat("        ---------------\n")

cat("\n")

cat("AVERAGE KAPPA COEFFICIENT FOR  L' > L (0 < c < 0.5) \n")

cat("\n")

cat("Estimated average kappa coefficient is ",k1," and its standard error is ", sqrt(Vark1), "\n")

cat("\n")

cat(100 * conflevel,"% Wald confidence interval: (", L1," ; ", L2,") \n")

cat("\n")

cat(100 * conflevel,"% Logit confidence interval: (", exp(log(k1 / (1 - k1)) - z * sqrt(Varlogitk1)) / (1 + exp(log(k1 / (1 - k1)) - z * sqrt(Varlogitk1))),"  ; ", exp(log(k1 / (1 - k1)) + z * sqrt(Varlogitk1)) / (1 + exp(log(k1 / (1 - k1)) + z * sqrt(Varlogitk1))),") \n")

cat("\n")

cat(100 * conflevel,"% Bias corrected confidence interval: (", a1," ; ", a2,") \n")

cat("\n")

cat("AVERAGE KAPPA COEFFICIENT FOR  L > L' (0.5 < c < 1) \n")

cat("\n")
```

```
cat("Estimated average kappa coefficient is ",k2," and its standard error is ", sqrt(Vark2), "\n")

cat("\n")

cat(100 * conflevel,"% Wald confidence interval: (", L3," ; ", L4,") \n")

cat("\n")

cat(100 * conflevel,"% Logit confidence interval: (", exp(log(k2 / (1 - k2)) - z * sqrt(Varlogitk2)) / (1 + exp(log(k2 / (1 - k2)) - z * sqrt(Varlogitk2))),"  ; ", exp(log(k2 / (1 - k2)) + z * sqrt(Varlogitk2)) / (1 + exp(log(k2 / (1 - k2)) + z * sqrt(Varlogitk2))),") \n")

cat("\n")

cat(100 * conflevel,"% Bias corrected confidence interval: (", b1," ; ", b2,") \n")

cat("\n")

cat("SENSITIVITY \n")

cat("\n")

cat("Estimated sensitivity is ",Se," and its standard error is", sqrt(VarSe), "\n")

cat("\n")

cat(100 * conflevel,"% modified score confidence interval (Yu et al, 2014): (", LSe," ; ", USe,") \n")

cat("\n")
```

```
cat("SPECIFICITY \n")

cat("\n")

cat("Estimated specificity is ",Sp," and its standard error is", sqrt(VarSp), "\n")

cat("\n")

cat(100 * conflevel,"% modified score confidence interval (Yu et al, 2014): (", LSp," ; ",
USp,") \n")

cat("\n")

cat("DISEASE PREVALENCE \n")

cat("\n")

cat("Estimated disease prevalence is ",p," and its standard error is", sqrt(Varp), "\n")

cat("\n")

cat("CHANCE-CORRECTED SENSITIVITY AND SPECIFICITY \n")

cat("\n")

cat("Estimated chance-corrected sensitivity is ",kappa1," and its standard error is",
sqrt(Varkappa1), "\n")

cat("\n")

cat("Estimated chance-corrected specificity is ",kappa0," and its standard error is",
sqrt(Varkappa0), "\n")

sink()
```

```
  }
else
  {
  #Estimated average kappa coefficient. As s0 = r1, then k1 = k2 = k
  k <- Y
  Vark <- VarSe + VarSp
  L5 <- k - z * sqrt(Vark)
  if (L5 < 0) {L5 <- 0}
  L6 <- k + z * sqrt(Vark)
  if (L6 > 1) {L6 <- 1}
  Varlogitk <- (1 / (Y^2 * (1 - Y)^2)) * (VarSe + VarSp)
  #Bootstrap
  data <- c(s1, s0, r1, r0)
  data1 <- matrix(1,1,data[1])
  data2 <- matrix(2,1,data[2])
  data3 <- matrix(3,1,data[3])
  data4 <- matrix(4,1,data[4])
  datatot <- c(data1,data2,data3,data4)
```

```
samplesB <- B
while (samplesB >= 1)
 {
 sampleboot <- sample(datatot,length(datatot),replace = TRUE)
 data.sample <- tabulate(sampleboot)
 if (length(data.sample) < length(data)) next
 d1 <- data.sample[1]
 d0 <- data.sample[2]
 e1 <- data.sample[3]
 e0 <- data.sample[4]
 m1 <- d1 + e1
 m0 <- d0 + e0
 Sen <- d1 / (d1 + d0)
 Spe <- e0 / (e1 + e0)
 Yo <- Sen + Spe - 1
 if ((d0 == 0 && e1 == 0) | Yo <= 0)  next
 if ((d1 > 0 && d0 > 0 && e1 > 0 && e0 > 0 && abs(d0 - e1) > 0 && Yo > 0) | (d0 ==
0 && (d1 * e1 * e0) > 0 && Yo > 0) | (e1 == 0 && (d1 * d0 * e0) > 0 && Yo > 0))
```

```
    {
    kboot <- (2 * (d1 * e0 - d0 * e1) / (m0 * (d1 + d0) - m1 * (e1 + e0))) * log((m1 * (e1+
   e0) + m0 * (d1 + d0)) / (2 * m1 * (e1 + e0)))
    list3[samplesB] <- kboot
    if (kboot < k) { x3 <- x3 + 1 }
    }
  if (d1 > 0 && d0 > 0 && e1 > 0 && e0 > 0 && abs(d0 - e1) == 0 && Yo > 0)
    {
    kboot <- Yo
    list3[samplesB] <- kboot
    if (kboot < k) { x3 <- x3 + 1 }
    }
  samplesB <- samplesB - 1
  }
u3 <- root (x3)
w1 <- pnorm (2 * u3 - z)
c1 <- quantile (list3, w1, names = FALSE)
w2 <- pnorm (2 * u3 + z)
```

```
c2 <- quantile (list3, w2, names = FALSE)

#Result

sink("Results.txt", split=TRUE)

cat("\n")

cat("R E S U L T S \n")

cat("---------------\n")

cat("\n")

cat("AVERAGE KAPPA COEFFICIENT \n")

cat("\n")

cat("As s0 = r1, estimated average kappa coefficient for  L' > L (0 < c < 0.5) is equal to
estimated average kappa coefficient for  L > L'  (0.5 < c < 1) \n")

cat("\n")

cat("Estimated average kappa coefficient is ",k," and its standard error is", sqrt(Vark),
"\n")

cat("\n")

cat(100 * conflevel,"% Wald confidence interval: (", L5," ; ", L6,") \n")

cat("\n")
```

```
cat(100 * conflevel,"% Logit confidence interval: (", exp(log(k / (1 - k)) - z * sqrt(Varlogitk)) / (1 + exp(log(k / (1 - k)) - z * sqrt(Varlogitk))),"  ; ", exp(log(k / (1 - k)) + z * sqrt(Varlogitk)) / (1 + exp(log(k / (1 - k)) + z * sqrt(Varlogitk))),") \n")

cat("\n")

cat(100 * conflevel,"% Bias corrected confidence interval: (", c1," ; ", c2,") \n")

cat("\n")

cat("SENSITIVITY \n")

cat("\n")

cat("Estimated sensitivity is ",Se," and its standard error is", sqrt(VarSe), "\n")

cat("\n")

cat(100 * conflevel,"% modified score confidence interval (Yu et al, 2014): (", LSe," ; ", USe,") \n")

cat("\n")

cat("SPECIFICITY \n")

cat("\n")

cat("Estimated specificity is ",Sp," and its standard error is", sqrt(VarSp), "\n")

cat("\n")

cat(100 * conflevel,"% modified score confidence interval (Yu et al, 2014): (", LSp," ; ", USp,") \n")
```

```
    cat("\n")
    cat("DISEASE PREVALENCE \n")
    cat("\n")
    cat("Estimated disease prevalence is ",p," and its standard error is", sqrt(Varp), "\n")
    cat("\n")
    cat("CHANCE-CORRECTED SENSITIVITY AND SPECIFICITY \n")
    cat("\n")
    cat("Estimated chance-corrected sensitivity is ",kappa1," and its standard error is",
    sqrt(Varkappa1), "\n")
    cat("\n")
    cat("Estimated chance-corrected specificity is ",kappa0," and its standard error is",
    sqrt(Varkappa0), "\n")
    cat("\n")
    sink()
    }
}
```

Apéndice II: Programa “cakctbt”

```
cakctbt <- function(s11, s10, s01, s00, r11, r10, r01, r00, alpha = 0.05)

{

  if (s11 < 0 | s10 < 0 | s01 < 0 | s00 < 0 | r11 < 0 | r10 < 0 | r01 < 0 | r00 < 0)

    {

    cat("\n")

    stop("Any observed frequency can be negative. Introduces new values \n")

    cat("\n")

    }

  if (abs(s00 - trunc (s00)) > 0 | abs(s01 - trunc (s01)) > 0 | abs(s10 - trunc (s10)) > 0 |

  abs(s11 - trunc (s11)) > 0 | abs(r00 - trunc (r00)) > 0 | abs(r01 - trunc (r01)) > 0 | abs(r10

  - trunc (r10)) > 0 | abs(r11 - trunc (r11)) > 0)

    {

    cat("\n")

    stop("Observed frequencies can not have decimals. Introduces new values \n")
```

```
  cat("\n")
  }
if (alpha >= 1 | alpha <= 0)
  {
  cat("\n")
  stop("Alpha should take a value between 0 and 1. Introduces a new value \n")
  cat("\n")
  }
if ((s11 + s10 + s01 + s00) == 0 | (r11 + r10 + r01 + r00) == 0)
  {
  cat("\n")
   stop("Accuracy of a Binary Test can not be estimated. There are many observed
  frequencies equal to zero. Introduces new values \n")
  cat("\n")
  }
Se1 <- (s11 + s10) / (s11 + s10 + s01 + s00)
Se2 <- (s11 + s01) / (s11 + s10 + s01 + s00)
Sp1 <- (r01 + r00) / (r11 + r10 + r01 + r00)
Sp2 <- (r10 + r00) / (r11 + r10 + r01 + r00)
```

```
Y1 <- Se1 + Sp1 - 1

Y2 <- Se2 + Sp2 - 1

if (Y1 <= 0)

  {

  cat("\n")

  cat("Estimated Youden index of Binary Test 1 is ",Y1, "\n")

  cat("\n")

  }

if (Y2 <= 0)

  {

  cat("\n")

  cat("Estimated Youden index of Binary Test 2 is ",Y2, "\n")

  cat("\n")

  }

if (Y1 <= 0 | Y2 <= 0)

  {

  cat("\n")

  stop("Estimated Youden index of a Binary Test must be greater than zero. Introduces
new values \n")
```

```
 cat("\n")

 }

zalpha  = qnorm(1 - alpha/2,0,1)

n <- s00 + s01 + s10 + s11 + r00 + r01 + r10 + r11

k1test1 <- expression((2 *(-1 + p00 + p01 + p10 + p11) * (-(p10 + p11) * (q00 + q01) +
(p00 + p01) * (q10 + q11)) * log(-(2 * p00 * p00* (q10 + q11) + 2 * p01 * p01 * (q10 +
q11) - (-1 + p10 + p11) * (2 * (p10 + p11)* (q00 + q01) + q10 + q11) + p01 * ((1 - 2 *
p10 - 2 * p11) * q00 + q01 - 2 * (q10 + q11) + 2 * (p10 + p11) * (- q01 + q10 + q11)) +
p00 * ((1 - 2 * p10 - 2 * p11) * q00 + q01 + 2 * (-1 + 2 * p01) * (q10 + q11) + 2 * (p10
+ p11) * (-q01 + q10 + q11)))/(2 * (-1 + p00 + p01 + p10 + p11) * ((p10 + p11) * (q00 +
q01) - (-1 + p00 + p01) * (q10 + q11))))) / ((-1 + p10 + p11) * (q10 + q11) + p00 * (q00
+ q01 + 2 * (q10 + q11)) + p01 * (q00 + q01 + 2 * (q10 + q11))))

dk1test1p00 <- deriv(k1test1, "p00")

dk1test1p01 <- deriv(k1test1, "p01")

dk1test1p10 <- deriv(k1test1, "p10")

dk1test1p11 <- deriv(k1test1, "p11")

dk1test1q00 <- deriv(k1test1, "q00")

dk1test1q01 <- deriv(k1test1, "q01")

dk1test1q10 <- deriv(k1test1, "q10")

dk1test1q11 <- deriv(k1test1, "q11")
```

```
k1test2 <- expression((2 * (-1 + p00 + p01 + p10 + p11) * (-(p01 + p11) * (q00 + q10) +
(p00 + p10) * (q01 + q11)) * log((-1 + 2 * p00 + 2 * p10 + (p00 + p10)/(-1 + p00 + p01 +
p10 + p11) + ((p00 + p01 + p10 + p11) * (q00 + q10))/(-(p01 + p11) * (q00 + q10) + (-1
+ p00 + p10) * (q01 + q11)))/(2 * (-1 + p00 + p10))))/((-1 + p01 + p11) * (q01 + q11) +
p00 * (q00 + 2 * q01 + q10 + 2 * q11) + p10 * (q00 + 2 * q01 + q10 + 2 * q11)))

dk1test2p00 <- deriv(k1test2, "p00")

dk1test2p01 <- deriv(k1test2, "p01")

dk1test2p10 <- deriv(k1test2, "p10")

dk1test2p11 <- deriv(k1test2, "p11")

dk1test2q00 <- deriv(k1test2, "q00")

dk1test2q01 <- deriv(k1test2, "q01")

dk1test2q10 <- deriv(k1test2, "q10")

dk1test2q11 <- deriv(k1test2, "q11")

k2test1 <- expression((2 *(-1 + p00 + p01 + p10 + p11) * (-(p10 + p11) * (q00 + q01) +
(p00 + p01) * (q10 + q11)) * log((2 * (p00 + p01 + p10 + p11) * (-(-1 + p10 + p11) * q00
+ q01 - (p10 + p11) * q01 + (p00 + p01) * (q10 + q11)))/(2 * p00 * p00 * (q10 + q11) + 2
* p01 * p01 * (q10 + q11) - (-1 + p10 + p11) * (2 * (p10 + p11) * (q00 + q01) + q10 +
q11) + p01 * ((1 - 2 * p10 - 2 * p11) * q00 + q01 - 2 * (q10 + q11) + 2 * (p10 + p11) * (-
q01 + q10 + q11)) + p00 * ((1 - 2 * p10 - 2 * p11) * q00 + q01 + 2 * (-1 + 2 * p01) *
(q10 + q11) + 2 * (p10 + p11) * (-q01 + q10 + q11)))))/((-1 + p10 + p11) * (q10 + q11) +
p00 * (q00 + q01 + 2 * (q10 + q11)) + p01 * (q00 + q01 + 2 * (q10 + q11))))
```

```
dk2test1p00 <- deriv(k2test1, "p00")

dk2test1p01 <- deriv(k2test1, "p01")

dk2test1p10 <- deriv(k2test1, "p10")

dk2test1p11 <- deriv(k2test1, "p11")

dk2test1q00 <- deriv(k2test1, "q00")

dk2test1q01 <- deriv(k2test1, "q01")

dk2test1q10 <- deriv(k2test1, "q10")

dk2test1q11 <- deriv(k2test1, "q11")

k2test2 <- expression((2 * (-1 + p00 + p01 + p10 + p11) * (-(p01 + p11) * (q00 + q10) +
(p00 + p10) * (q01 + q11)) * log((2 * (p00 + p01 + p10 + p11) * (-(-1 + p01 + p11) * q00
+ q10 - (p01 + p11) * q10 + (p00 + p10) * (q01 + q11)))/ (p10 * q00 + 2 * p11 * q00 - 2
* p10 * p11 * q00 - 2 * p11 * p11 * q00 + q01 - 2 * p10 * q01 + 2 * p10* p10 * q01 -
p11 * q01 + 2 * p10 * p11 * q01 + p10 * q10 + 2 * p11 * q10 - 2 * p10 * p11 * q10 - 2 *
p11 * p11 * q10 - 2 * p01 * p01 * (q00 + q10) + (1 - p11 + 2 * p10 * (-1 + p10 + p11)) *
q11 + 2 * p00 * p00 * (q01 + q11) - p01 * (2 * (-1 + p10 + 2 * p11) * q00 + q01 - 2 * p10
* q01 + 2 * (-1 + p10 + 2 * p11) * q10 + q11 - 2 * p10 * q11) + p00 * ((1 - 2 * p01 - 2 *
p11) * q00 + 2 * (-1 + p01 + 2 * p10 + p11) * q01 + q10 - 2 * p01 * q10 - 2 * p11 * q10
+ 2 * (-1 + p01 + 2 * p10 + p11) * q11))))/((-1 + p01 + p11) * (q01 + q11) + p00 * (q00
+ 2 * q01 + q10 + 2 * q11) + p10 * (q00 + 2 * q01 + q10 + 2 * q11)))

dk2test2p00 <- deriv(k2test2, "p00")

dk2test2p01 <- deriv(k2test2, "p01")
```

```
dk2test2p10 <- deriv(k2test2, "p10")

dk2test2p11 <- deriv(k2test2, "p11")

dk2test2q00 <- deriv(k2test2, "q00")

dk2test2q01 <- deriv(k2test2, "q01")

dk2test2q10 <- deriv(k2test2, "q10")

dk2test2q11 <- deriv(k2test2, "q11")

p00 <- s00 / n

p01 <- s01 / n

p10 <- s10 / n

p11 <- s11 / n

q00 <- r00 / n

q01 <- r01 / n

q10 <- r10 / n

q11 <- r11 / n

a1 <- attr(eval(dk1test1p00), "gradient")[1]

a2 <- attr(eval(dk1test1p01), "gradient")[1]

a3 <- attr(eval(dk1test1p10), "gradient")[1]

a4 <- attr(eval(dk1test1p11), "gradient")[1]

a5 <- attr(eval(dk1test1q00), "gradient")[1]
```

```
a6 <- attr(eval(dk1test1q01), "gradient")[1]

a7 <- attr(eval(dk1test1q10), "gradient")[1]

a8 <- attr(eval(dk1test1q11), "gradient")[1]

b1 <- attr(eval(dk1test2p00), "gradient")[1]

b2 <- attr(eval(dk1test2p01), "gradient")[1]

b3 <- attr(eval(dk1test2p10), "gradient")[1]

b4 <- attr(eval(dk1test2p11), "gradient")[1]

b5 <- attr(eval(dk1test2q00), "gradient")[1]

b6 <- attr(eval(dk1test2q01), "gradient")[1]

b7 <- attr(eval(dk1test2q10), "gradient")[1]

b8 <- attr(eval(dk1test2q11), "gradient")[1]

mat1 <- matrix(0,2,8)

mat1[1,1] <- a1

mat1[1,2] <- a2

mat1[1,3] <- a3

mat1[1,4] <- a4

mat1[1,5] <- a5

mat1[1,6] <- a6

mat1[1,7] <- a7
```

```
mat1[1,8] <- a8

mat1[2,1] <- b1

mat1[2,2] <- b2

mat1[2,3] <- b3

mat1[2,4] <- b4

mat1[2,5] <- b5

mat1[2,6] <- b6

mat1[2,7] <- b7

mat1[2,8] <- b8

c1 <- attr(eval(dk2test1p00), "gradient")[1]

c2 <- attr(eval(dk2test1p01), "gradient")[1]

c3 <- attr(eval(dk2test1p10), "gradient")[1]

c4 <- attr(eval(dk2test1p11), "gradient")[1]

c5 <- attr(eval(dk2test1q00), "gradient")[1]

c6 <- attr(eval(dk2test1q01), "gradient")[1]

c7 <- attr(eval(dk2test1q10), "gradient")[1]

c8 <- attr(eval(dk2test1q11), "gradient")[1]

d1 <- attr(eval(dk2test2p00), "gradient")[1]

d2 <- attr(eval(dk2test2p01), "gradient")[1]
```

```
d3 <- attr(eval(dk2test2p10), "gradient")[1]

d4 <- attr(eval(dk2test2p11), "gradient")[1]

d5 <- attr(eval(dk2test2q00), "gradient")[1]

d6 <- attr(eval(dk2test2q01), "gradient")[1]

d7 <- attr(eval(dk2test2q10), "gradient")[1]

d8 <- attr(eval(dk2test2q11), "gradient")[1]

mat2 <- matrix(0,2,8)

mat2[1,1] <- c1

mat2[1,2] <- c2

mat2[1,3] <- c3

mat2[1,4] <- c4

mat2[1,5] <- c5

mat2[1,6] <- c6

mat2[1,7] <- c7

mat2[1,8] <- c8

mat2[2,1] <- d1

mat2[2,2] <- d2

mat2[2,3] <- d3

mat2[2,4] <- d4
```

```
mat2[2,5] <- d5

mat2[2,6] <- d6

mat2[2,7] <- d7

mat2[2,8] <- d8

k11 <- eval(k1test1)

k21 <- eval(k1test2)

k12 <- eval(k2test1)

k22 <- eval(k2test2)

vec <- vector("numeric", 8)

vec[1] <- p00

vec[2] <- p01

vec[3] <- p10

vec[4] <- p11

vec[5] <- q00

vec[6] <- q01

vec[7] <- q10

vec[8] <- q11

mat3 <- matrix(0,8,8)

mat3[1,1] <- p00
```

```
mat3[2,2] <- p01

mat3[3,3] <- p10

mat3[4,4] <- p11

mat3[5,5] <- q00

mat3[6,6] <- q01

mat3[7,7] <- q10

mat3[8,8] <- q11

sigma1 <- matrix(0,8,8)

sigma1 <- (1 / n) * (mat3 - vec %*% t(vec))

sigma2 <- matrix(0,2,2)

sigma2 <- mat1 %*% sigma1 %*% t(mat1)

sigma3 <- matrix(0,2,2)

sigma3 <- mat2 %*% sigma1 %*% t(mat2)

z1 <- abs(k11 - k21) / sqrt(sigma2[1,1] + sigma2[2,2] - 2 * sigma2[1,2])

pvalue1 <- 2*(1 - pnorm(z1,0,1))

z2 <- abs(k12 - k22) / sqrt(sigma3[1,1] + sigma3[2,2] - 2 * sigma3[1,2])

pvalue2 <- 2*(1 - pnorm(z2,0,1))

#Result

sink("Results_cakctbt.txt", split=TRUE)
```

```
cat("\n")

cat(" R E S U L T S \n")

cat("---------------\n")

cat("\n")

cat("COMPARISON OF AVERAGE KAPPA COEFFICIENTS FOR  L' > L (0 < c < 0.5) \n")

cat("\n")

cat("Estimated average kappa coefficient of Binary Test 1 is ",k11," and its standard error is", sqrt(sigma2[1,1]), "\n")

cat("\n")

cat("Estimated average kappa coefficient of Binary Test 2 is ",k21," and its standard error is", sqrt(sigma2[2,2]), "\n")

cat("\n")

cat("Statistics for hypothesis test H0: (K11 is equal to K21) vs H1: (K11 is not equal to K21) is ", z1, " and the P-value is ", pvalue1,"\n")

cat("\n")

if (pvalue1 > alpha)

  {

   cat("We do not reject ( to an error alpha = ",alpha,") the null hypothesis H0: (K11 is equal to K21) \n")
```

```
  cat("\n")
  cat("Therefore, we do not reject that the two average kappa coefficients are equal \n")
  cat("\n")
  }
else
  {
  cat("We reject ( to an error alpha = ",alpha,") the null hypothesis H0: (K11 is equal to
  K21) \n")
  cat("\n")
  }
if (k11 > k21)
  {
  cat(100 * (1 - alpha),"% confidence interval for K11 - K21 is (",(k11 - k21) - zalpha *
  sqrt(sigma2[1,1] + sigma2[2,2] - 2 * sigma2[1,2]), " ; ",(k11 - k21) + zalpha *
  sqrt(sigma2[1,1] + sigma2[2,2] - 2 * sigma2[1,2]),") \n")
  cat("\n")
  }
else
  {
```

```
 cat(100 * (1 - alpha),"% confidence interval for K21 - K11 is (",(k21 - k11) - zalpha *
 sqrt(sigma2[1,1] + sigma2[2,2] - 2 * sigma2[1,2]), " ; ",(k21 - k11) + zalpha *
 sqrt(sigma2[1,1] + sigma2[2,2] - 2 * sigma2[1,2]),") \n")
 cat("\n")
 }
cat("\n")
cat("\n")
cat("COMPARISON OF AVERAGE KAPPA COEFFICIENTS FOR  L > L'  (0.5 < c <
1) \n")
cat("\n")
cat("Estimated average kappa coefficient of Binary Test 1 is ",k12," and its standard error
is", sqrt(sigma3[1,1]), "\n")
cat("\n")
cat("Estimated average kappa coefficient of Binary Test 2 is ",k22," and its standard error
is", sqrt(sigma3[2,2]), "\n")
cat("\n")
cat("Statistics for hypothesis test H0: (K12 is equal to K22) vs H1: (K12 is not equal to
K22) is ", z2, " and the P-value is ", pvalue2,"\n")
cat("\n")
if (pvalue2 > alpha)
```

```
  {
  cat("We do not reject ( to an error alpha = ",alpha,") the null hypothesis H0: (K12 is
  equal to K22) \n")
  cat("\n")
  cat("Therefore, we do not reject that the two average kappa coefficients are equal \n")
  cat("\n")
  }
else
  {
  cat("We reject ( to an error alpha =",alpha,") the null hypothesis H0: (K12 is equal to
  K22) \n")
  cat("\n")
  }
if (k12 > k22)
  {
  cat(100 * (1 - alpha),"% confidence interval for K12 - K22 is (",(k12 - k22) - zalpha *
  sqrt(sigma3[1,1] + sigma3[2,2] - 2 * sigma3[1,2]), " ; ",(k12 - k22) + zalpha *
  sqrt(sigma3[1,1] + sigma3[2,2] - 2 * sigma3[1,2]),") \n")
  cat("\n")
  }
```

```
  else
    {
    cat(100 * (1 - alpha),"% confidence interval for K22 - K12 is (",(k22 - k12) - zalpha *
    sqrt(sigma3[1,1] + sigma3[2,2] - 2 * sigma3[1,2]), " ; ",(k22 - k12) + zalpha *
    sqrt(sigma3[1,1] + sigma3[2,2] - 2 * sigma3[1,2]),") \n")
    cat("\n")
    }
  sink()
}
```

Bibliografía

Agresti, A., (2002). *Categorical Data Analysis*. John Wiley and Sons, New York.

Agresti, A., Min, Y., (2005). Sample improved confidence intervals for comparing matched proportions. *Statistics in Medicine,* 24:729 – 740.

Bennett, B.M., (1972). On comparison of sensitivity, specificity and predictive value of a number of diagnostic procedures. *Biometrics,* 28:793 – 800.

Bennett, B.M., (1983). Further results on indices of diagnostic screening. *Biometrical Journal*, 24:59 – 62.

Bloch D.A., (1997). Comparing two diagnostic test against the same "gold standard" in the same sample. *Biometrics*, 53:73-85.

Bonferroni, C.E., 1936. Teoria statistica delle classi e calcolo delle probabilità. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze*, 8:3-62.

Cicchetti D.V., (2001). Methodological Commentary The Precision of Reliability and Validity Estimates Re-Visited: Distinguishing Between Clinical and Statistical Significance of Sample Size Requirements. *Journal of Clinical and Experimental Neuropsychology*, 23(5):695-700.

Cicchetti D.V., (2001). Methodological Commentary The Precision of Reliability and Validity Estimates Re-Visited: Distinguishing Between Clinical and Statistical Significance of Sample Size Requirements. *Journal of Clinical and Experimental Neuropsychology*, 23(5):695-700.

Clopper C.P., Pearson E.S., (1934). The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4):404-413.

Dolgun N.A., Gozukara H., Karaagaoglu E. (2012). Comparing diagnostic test: test of hypothesis for likelihood ratios. *Journal of Statistical Computation and Simulation*, 82:369-381.

Efron B., Tibshirani R., (1993). *An introduction to the Bootstrap*. Chapman & Hall. London.

Hochberg, Y., 1988. A sharper Bonferroni procedure for multiple test of significance. *Biometrika*, 75:800–802.

Holm, S., 1979. A simple sequential rejective multiple testing procedure. *Scandinavian Journal of Statistics*, 6:65–70.

Jamart, J., (1993). Letter to the editor: on test for equality of predictive values for *t* diagnostic procedures. *Statistics in Medicine,* 12:185 – 186.

Kraemer H.C., (1992). *Evaluating medical test. Objective and quantitative guidelines*. Sage publications, Newbury Park.

Kraemer H.C., Periyakoil V.S., Noda A., (2002). Kappa coefficients in medical research. *Statistics in Medicine*, 21:2109-2129.

Kosinski A.S., (2012). A weighted generalized score statistic for comparison of predictive values of diagnostic test. *Statistics in Medicine*, 32(6):964-77.

Lachenbruch P.A., Lynch C.J., (1998). Assessing screening test: Extensions of McNemar's test. *Statistics in Medicine*, 17(19):2207-2217.

Landis J.R., Koch G.G., (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33:159-174.

Leisenring W., Pepe M.S., (1998). Regression modelling of diagnostic likelihood ratios for the evaluation of medical diagnostic test. *Biometrics*, 54:444–452.

Leisenring W., Alonzo T., Pepe M.S., (2000). Comparisons of predictive values of binary medical diagnostic test for paired designs. *Biometrics*, 56:344–351.

Martín Andrés, A., Álvarez Hernández M., (2014). Two-tailed asymptotic inferences for a proportion. *Journal of Applied Statistics*, 41(7):1516-1529.

Martín Andrés, A., Álvarez Hernández M., (2015). Comment on 'An improved score interval with a modified midpoint for a binomial proportion'. *Journal of Statistical Computation and Simulation*. In press. DOI: 10.1080/00949655.2015.1015128.

McNeil B., Adelstein J., (1976). Determining the value of diagnostic and screening test. *Journal of Nuclear Medicine*, 17(6):439-448.

Roldán Nofuentes J.A., Luna del Castillo J.D., (2007) Comparing of the likelihood ratios of two binary diagnostic test in paired designs. *Statistics in Medicine*, 26:4179–4201.

Roldán Nofuentes J.A., Luna del Castillo J.D., 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.

Roldán Nofuentes, J.A., Luna del Castillo, J.D., (2010). Comparison of weighted kappa coefficients of multiple binary diagnostic tests done on the same subjects. Statistics in Medicine, 29:2149-2165.

Roldán Nofuentes J.A., Luna del Castillo J.D., Montero Alonso M.A., (2012). Global hypothesis test to simultaneously compare the predictive values of two binary diagnostic test. *Computational statistics & data analysis, Special issue "Computational Statistics for Clinical Research"*, 56(5):1161-1173.

Serfling R.J., (1980). *Approximation Theorems of Mathematical Statistics*. Wiley, New York.

Sox H.C., Blatt M.A., Higgins M.C., Marton K.I., (1989). *Medical decision making*. Butterworths-Heinemann, Boston.

Vacek, P.M., (1985). The effect of conditional dependence on the evaluation of diagnostic test. *Biometrics*, 41:959–968.

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.

Wang, W., Davis, C.S., Soong, S., (2006). Comparison of predictive values of two diagnostic test from the same sample of subjects using weighted least squares. *Statistics in Medicine*, 25:2215–2229.

Wilson E.B., (1927). Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158):209-212.

Yee J, Akerkar GA, Hung RK, Steinauer-Gebauer AM, Wall SD, McQuaid KR, (2001). Colorectal neoplasia: performance characteristics of CT colonography for detection in 300 patients. *Radiology*, 219:685-692.

Youden W.J., (1950). Index for rating diagnostic test. *Cancer*, 3:32-35.

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(5):1022–1038.

Zhou, X.H., (1998). Comparing accuracies of two screening test in a two-phase study for dementia. *Journal of Royal Statistical Society Series C Applied Statistics,* 47: 135-147.

Zhou X.H., Obuchowski N.A., McClish D.K., (2002). *Statistical methods in diagnostic medicine*. John Wiley and Sons, New York.