Metodología para la elección de punto de corte óptimo para dicotomizar covariables continuas.

Methodology for the selection of optimal cutoff point to dichotomize continuous covariates.

Lisset Evelyn Fuentes Smith.

Resumen

La práctica de categorizar covariables continuas es común en investigaciones médicas y epidemiológicas por razones tanto clínicas como estadísticas. El siguiente trabajo se centra solo en la dicotomía de una covariable continua, por ser desde el punto de vista biológico, uno de los objetivos más deseados, sin embargo, pueden existir más de un punto de corte para el recorrido de una variable continua. Como metodología para la obtención de puntos de corte candidatos se proponen 1) Representación gráfica, 2) el cálculo de cuantiles, 3) al determinación de Ji-Cuadrado y el Odd Ratio máximo en tabla de contingencia 2x2 y 4) la regresión logística. Luego de determinar los puntos de corte candidatos para la selección de punto de corte óptimo se propone el cálculo de estadígrafos como la Sensibilidad, Especificidad e Índice de Youden, así como también el uso de la Curva ROC y el área bajo la curva (ABC) como índice de exactitud. La Sensibilidad, Especificidad y ABC son estimadores muestrales de parámetros poblacionales; por consiguiente, cada uno tiene asociado un error de estimación, que hace necesario reportar sus respectivos intervalos de confianza. Dado el uso generalizado y frecuente de la dicotomización de una covariable continua en las investigaciones de bioestadística y epidemiología, este trabajo ofrece una metodología práctica para la obtención de punto de corte óptimo.

Palabras clave: Punto de corte, categorización, dicotomización.

Abstract

Categorizing continuous covariables is a common practice in medical and epidemiological investigations due to clinical and statistical reasons. This work is only centered on the dichotomy of a continuous covariable, since it is one of the most wanted objectives from the biological point of view; however more than a single cut point may exist for the range of a continuous variable. In order to define the candidate cut points the following methodology is proposed: 1) graphical representation, 2) quantiles calculation, 3) determination of Xi-square and maximum odd ratio in the 2x2 contingency table and 4) logistic regression. After determining the candidate cut points, in order to select the optimum cut point it is proposed to use the calculation of statistics as sensitivity, specificity and Youden's index, as well as to use the ROC curve and the area below the curve (ABC) as an exactitude index. Sensitivity, specificity and ABC are sampling estimators of demographic parameters; therefore, each one of them has an associated estimation error that makes it necessary to report their respective reliability intervals. Since dichotomizing a continuous covariable is a generalized and frequent application in biostatistics and epidemiology investigations, this work proposes a practical methodology for obtaining the optimum cut point.

Keywords: Cut point, categorization, dichotomizing.

Introducción

La práctica de categorizar covariables continuas es común en investigaciones médicas y epidemiológicas por varias razones, tanto clínicas como estadísticas. Desde el punto de vista clínico, las covariables binarias ofrecen varias ventajas como: (1) ofrecer una clasificación simple de riesgo ("alto" y "bajo", "presencia" y "ausencia"), (2) modelar criterios de elegibilidad para los estudios prospectivos, (3) establecer criterios diagnósticos para la enfermedad, recomendar tratamiento diagnóstico, (4) estimar el pronóstico de la enfermedad y (5) la imposición de un umbral biológico.[1-11.]

Desde un punto de vista estadístico, las covariables binarias ofrecen una interpretación más simple a través de las medidas de asociación utilizadas en los modelos estadísticos correspondientes, tales como odds ratio y riesgo relativo, permiten evitar el supuesto de linealidad implícito en los modelos estadísticos comunes para covariables continuas y resumir los

Licenciada en Matemáticas. Investigador Agregado. Centro Nacional de Genética Médica. La Habana. Cuba. E-mail: evelynfuentes@infomed.sld.cu.

datos de manera más eficiente.[1-2,8-12]

Aunque resulta atractiva la utilización de un método sistemático para la elección de los puntos de corte, la categorización de una variable cuantitativa supone siempre una pérdida importante de información, y si además los puntos de corte se eligen en base a la información proporcionada por los propios datos del estudio puede dar lugar a que las conclusiones sean menos extrapolables a otras situaciones.[1-5]

El siguiente trabajo se centra solo en la dicotomía de una covariable continua, por ser, desde el punto de vista biológico, uno de los objetivos más deseados, sin embargo, pueden existir más de un punto de corte para el recorrido de una variable continua.

Los métodos estadísticos para la determinación del punto de corte caen en dos grandes categorías: orientado a datos y orientado a resultados. Los métodos orientados a datos se basan en el cálculo de los cuantiles, mientras que los métodos orientados a resultados proporcionan un valor de punto de corte en correspondencia con la relación más significativa con los resultados.[13]

Metodología para la obtención de puntos de corte candidatos

1. Representación gráfica

Una vez que se ha determinado que una covariable cuantitativa se asocia de forma significativa con la variable resultado o con la presencia del suceso, se debe examinar de forma gráfica dicha relación. Si el resultado es una variable cualitativa, por ejemplo dicotómica, la gráfica será en general poco informativa salvo que haya una separación bastante definida.

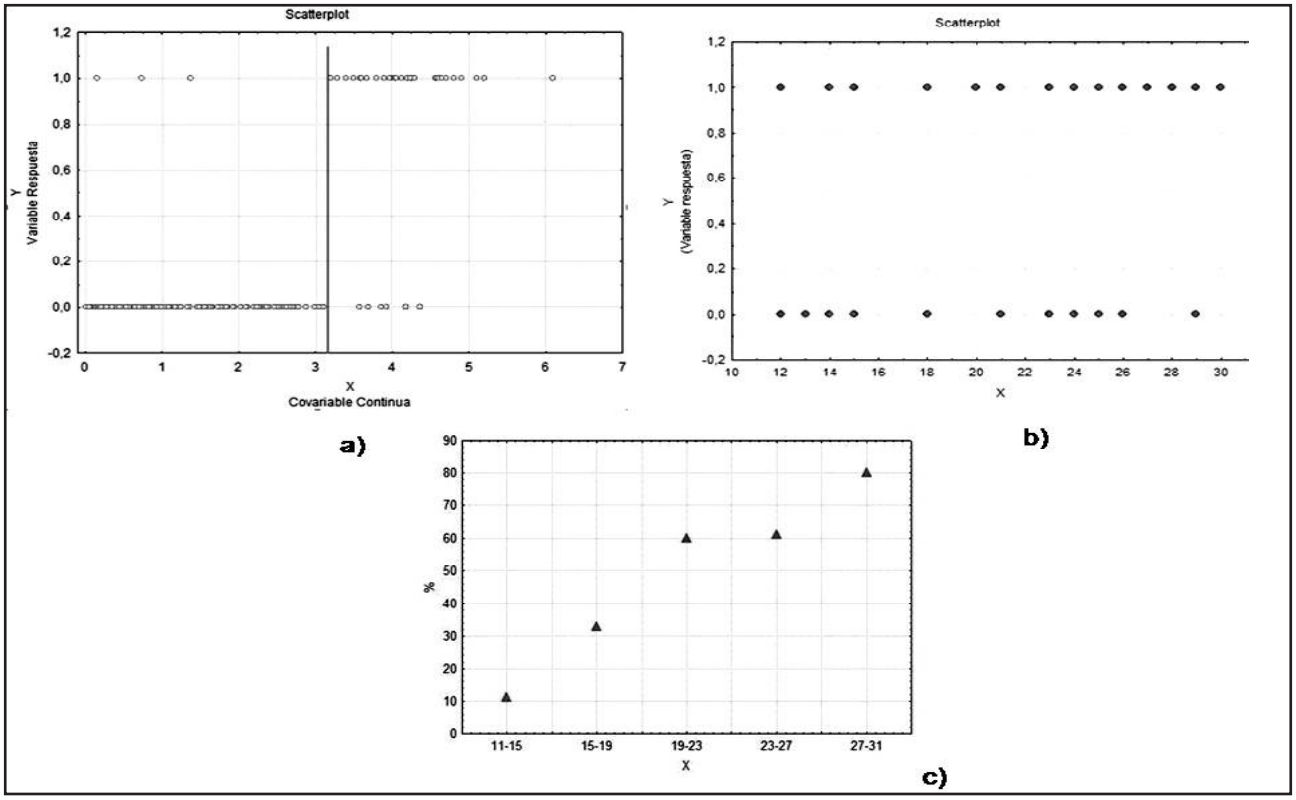
En la figura 1a se ve un ejemplo de este tipo de gráfica que sugiere un punto de corte en torno al valor 3,2.

Sin embargo esta gráfica suele ser en general difícil de interpretar (por ejemplo como muestra la figura 1b al distribuirse todos los puntos en dos valores del eje Y).

Una gráfica mucho más ilustrativa se obtiene dividiendo la variable X en intervalos iguales y calculando la proporción de sucesos para cada uno de esos intervalos, representando entonces dicha proporción frente al valor correspondiente al centro de cada intervalo (Figura 1c).

El ejemplo concreto de la gráfica 1c, nos sugiere un valor de corte para X en el entorno del valor 21.

Figura 1. La figura 1a) sugiere un punto de corte en torno al valor 3,2, la figura 1b) no permite discernir un punto de corte, la figura 1c) Representa la proporción de sucesos por intervalos, sugiere punto de corte para la covariable X en el intervalo de 19 a 23.



Por otro lado el inconveniente de este tipo de representación radica en que puede ser sensible a la amplitud del intervalo empleado, por lo que es una buena idea considerar diferentes amplitudes, y sobre todo verificar la frecuencia, número de datos, en cada intervalo, ya que si éste es muy pequeño la información será poco precisa.[11]

2. Cálculo de cuantiles

Se debe verificar que los datos siguen una distribución normal, sin embargo lo más frecuente es que los resultados de pruebas analíticas presenten distribuciones asimétricas, concentradas en el lado izquierdo y con una cola larga al lado derecho, por lo que será preciso realizar una transformación de los datos.

a. Transformación de los datos

Las transformaciones $\sqrt{X}$, Ln(x) y 1/x comprimen los valores altos de los datos y expanden los bajos, por su parte si la concentración de datos está, en el lado de la derecha y la cola en la izquierda, se puede utilizar la transformación x^2, que comprime la escala para valores pequeños y la expande para valores altos. Cuando los datos son proporciones o porcentajes de una distribución binomial, las diferencias con una distribución normal son más acusadas para valores pequeños o grandes de las proporciones, utilizándose entonces transformaciones basadas en $\text{arcsen}\sqrt{X}$.[14]

En todos los casos para los cálculos estadísticos basados en la teoría gaussiana, se utilizarán los valores transformados, pero después para la presentación de los resultados se efectuará la transformación inversa para presentarlos en su escala de medida natural.

2.2. Cálculos de Media, Percentiles y Cuartiles

Media

Una vez comprobada la normalidad se suele estimar como punto de corte el valor $\bar{x}$-2DS en una muestra de controles.[11]

$\bar{x}$: media y DS: desviación estándar.

Percentil 2,5 y 97,5

Un procedimiento muy empleado para fijar intervalos de referencia de pruebas analíticas, a partir de una muestra representativa de la población, se basa en seleccionar los valores de dos percentiles centrados en torno a la mediana de la distribución, concretamente los percentiles 2,5 y 97,5, que definen un intervalo de referencia del 95 %, tomando como puntos de corte los valores extremos del intervalo respectivamente.

Si el objetivo del estudio es determinar punto de corte para guiar la toma de decisiones, el cálculo de los percentiles de la distribución a partir de los valores de la muestra da lugar a estimaciones sesgadas si el tamaño de la muestra no es suficientemente grande, y en general sus valores pueden variar en gran medida de una muestra a otra, por lo que se prefiere realizar su cálculo a partir de un modelo de distribución de probabilidad.[11,15]

Tercer Cuartil

De modo que, se propone, una vez calculada la media y la desviación estándar, estimar el tercer cuartil consultando en una tabla de probabilidad de la distribución normal el valor para el cual $Pr(x \leq z)=0,75$, que corresponde a 0,674, por lo que se estima dicho cuartil como $\bar{x}+0,674*DS$.[11]

3. Ji-Cuadrado y el Odd Ratio en tabla de contingencia 2x2

Frente a la elección como punto de corte de un percentil, o a partir de la inspección visual de las gráficas, existe una alternativa sistemática que nos puede ayudar en la decisión. Consiste en determinar, para todos los valores de la variable X que se desea categorizar, el valor que mejor separa a los pacientes de acuerdo a la prueba de asociación del Ji-Cuadrado.

Se confeccionará una tabla de contingencia 2x2 para cada valor de la covariable continua X y se calculará el estadígrafo Ji-Cuadrado y el Odd Ratio (OR) para cada tabla.

	X≤B	**X>B**	
Y=0	n_{11}	n_{12}	OR= (n11 x n12)/(n21 x n22)
Y=1	n_{21}	n_{22}	

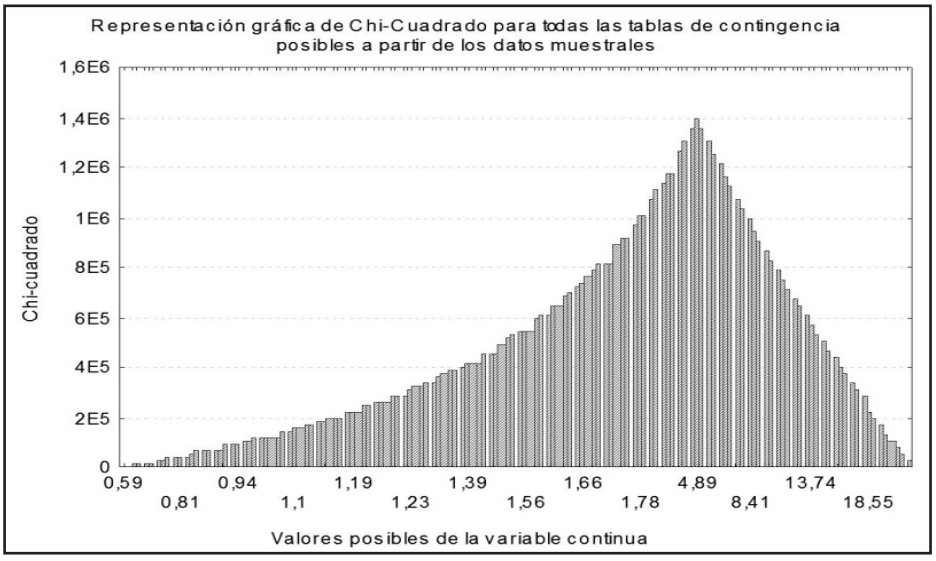
X es la covariable continua dicotomizada, B es un valor cualquiera de la covariable X, Y es la variable respuesta de tipo binaria (0=ausencia del rasgo o enfermedad, 1=presencia del rasgo o enfermedad). Se toma como punto de corte candidato el valor B para el cual siendo Ji-Cuadrado máximo, maximice el OR.

Para evaluar los posibles puntos de corte se recomienda no considerar los valores más extremos de la variable a ambos lados, excluyendo entonces entre el 5 % o el 10 % de ellos en cada extremo. Asimismo debido al aumento de la probabilidad de error de tipo I, que se produce al efectuar comparaciones múltiples, es también aconsejable utilizar alguna fórmula de ajuste para el valor de probabilidad mínimo obtenido.

Altman et al. proponen una fórmula de corrección muy sencilla para el caso de que se excluya el 5 % de los valores más extremos de X a ambos lados (percentiles 5 y 95): $p = -3,13\ p\_min(1 + 1,65\ \ln(p\_min))$, y otra

para cuando se excluye el 10 % (percentiles 10 y 90): $p = -1{,}63\ p\_min(1 + 2{,}35\ \ln(p\_min))$ donde p_min es el valor de probabilidad mínimo obtenido y p es el valor corregido.[11,15-17] (Figura 2)

Figura 2. Representación gráfica de todos los valores de Ji-cuadrado obtenidos al confeccionar todas las tablas de contingencia posibles. Se asume como punto de corte candidato el valor para el cual la gráfica alcanza su máximo (4,89).



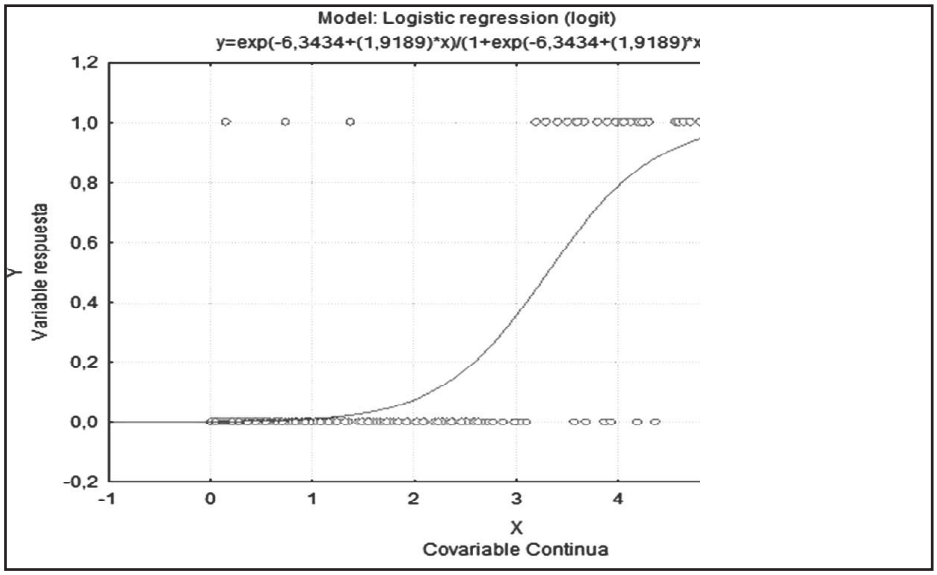
4. Regresión Logística Binaria

La regresión logística binaria de forma general toma como valor de punto de corte para la clasificación el valor 0,5.

La clasificación para Y tiene como punto de corte $Y = \frac{e^{(\beta_0+\beta_1 X)}}{1+e^{(\beta_0+\beta_1 X)}} = 0{,}5$ lo que permite obtener como punto de corte candidato para X el valor: $X = \frac{-\beta_0}{\beta_1}$.

Verificando que se cumpla para el modelo: que el coeficiente de determinación R^2 sea un valor cercano a 1 y la prueba de bondad de ajuste de Hosmer y Lemeshow.[18-20] (Figura 3)

Figura 3. Modelo de Regresión Logística, punto de corte candidato: 3,3.



Metodología para selección de punto de corte óptimo

Supongamos la existencia de k posibles puntos de corte, determinados por los métodos anteriores, para la selección del punto de corte óptimo se sugiere verificar el comportamiento de algunos estadígrafos como el número de falsos positivos (FP), el número de falsos negativos (FN), la Sensibilidad (S), la Especificidad (E) y el Índice de Youden que a continuación se definen:

Sea C el valor de punto de corte candidato obtenido por alguno de los métodos anteriores se calcularán los estadígrafos de la siguiente manera.

	X≤C	X>C	
Controles (Y=0)	VN	FP	TS =VN+FP
Casos (Y=1)	FN	VP	TE= VP+FN

Glosario:

FP: *Falsos positivos*: Individuos clasificados como paciente, sin la enfermedad.

FN: *Falsos negativos*: individuos clasificados como controles, con la enfermedad.

VP: *Verdaderos positivos*: Individuos clasificados como pacientes con la enfermedad.

VN: *Verdaderos Negativos*: Individuos clasificados como controles sin la enfermedad.

TE: *Total de Enfermos*: Falsos Negativos + Verdaderos Positivos.

TS: *Total de Sanos*: Falsos Positivos + Verdaderos Negativos.

Sensibilidad: Es la probabilidad de diagnosticar un positivo cuando el individuo en realidad está enfermo. S= (VP/TE)* 100.

Especificidad: Es la probabilidad de diagnosticar un negativo cuando el individuo en realidad está sano. E= (VN/TS)*100.

Evidentemente los valores de sensibilidad y especificidad son estimaciones realizadas mediante un experimento diseñado al efecto, por lo que es necesario calcular algún indicador de su grado de incertidumbre, como puede ser un intervalo de confianza del 95%.[15-16,21]

Luego podemos construir un intervalo de 95 % de confianza para la sensibilidad y la especificidad como sigue:

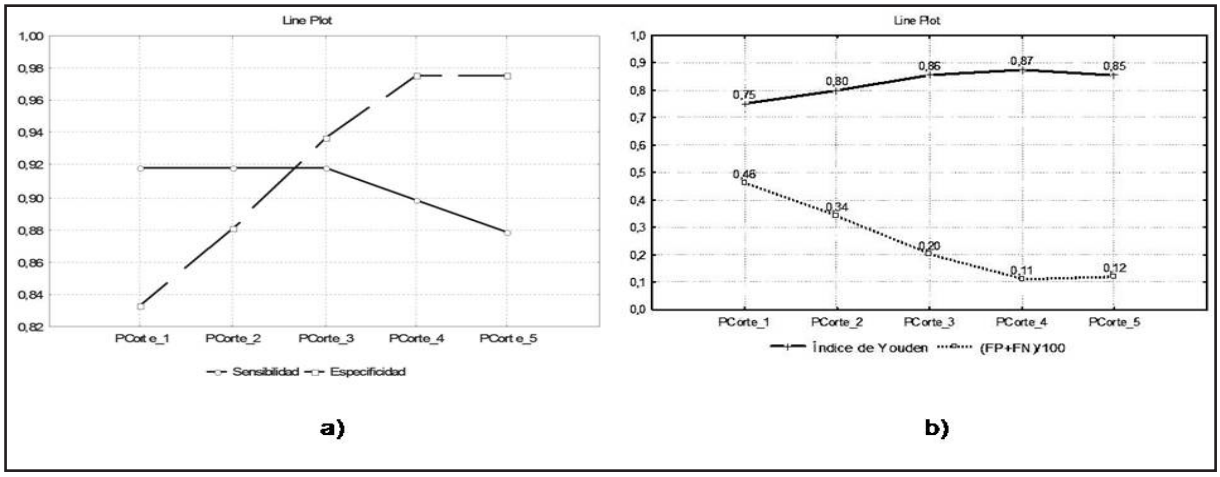
$$\hat{S}-1{,}96\times\sqrt{((\hat{S}(1-\hat{S}))/n)}\leq S\leq\hat{S}+1{,}96\times\sqrt{((\hat{S}(1-\hat{S}))/n)}$$
$$\hat{E}-1{,}96\times\sqrt{((\hat{E}(1-\hat{E}))/n)}\leq E\leq\hat{E}+1{,}96\times\sqrt{((\hat{E}(1-\hat{E}))/n)}$$

Se puede observar que si la sensibilidad es máxima, S=100 %, eso significa que todos los enfermos han sido bien diagnosticados (VP=TE), es decir, no hay falsos negativos (FN=0). Se deduce que este caso se requiere en enfermedades de consecuencias graves si no se les detecta a tiempo.

De la misma manera se puede inferir que una especificidad máxima E=100 % implica que no existen falsos positivos (FP=0), este caso se desea cuando es muy grave informarle al paciente que está enfermo si en realidad está sano. Sería para el caso de enfermedades que podrían ocasionar serios daños en lo psicológico, moral o económico.

De forma general se observa (Figura 4a) que al establecer un punto de corte para una variable continua, si la sensibilidad aumenta disminuye la

Figura 4. Determinación de punto de corte óptimo. a) Representación de la Sensibilidad y Especificidad para 5 puntos de corte de una covariable continua, b) Representación gráfica de punto de corte óptimo, mediante el índice de Youden y los valores FP+FN.



especificidad y viceversa.

Luego, el problema real consiste en decidir cuál de los dos estadígrafos debe ser maximizado. En consecuencia, surge el Índice de Youden para los casos donde ambos errores FP y FN son igualmente graves.

Índice de Youden: Se establece una solución donde se optimice el valor de ambos estadígrafos: sensibilidad y especificidad IY= ((S+E)/100)-1

De modo que se elige como punto de corte óptimo aquel que minimizando la suma de los errores (FP+FN) maximice el Índice de Youden.[15,16]

Presentación de la curva COR y el área bajo la curva (ABC) como índice de exactitud.

La curva COR (en inglés *ROC*, del acrónimo de *Receiver Operating Characteristic*) es una representación gráfica de la sensibilidad frente a (1 – especificidad) para un sistema clasificador binario según se varía el umbral de discriminación, recorre todo el rango de valores posibles de los puntos de corte, obteniendo una serie de pares de Sensibilidad-Especificidad que definen la prueba diagnóstica.

Se entiende como Área Bajo la Curva (ABC) a la probabilidad de clasificar correctamente un par de individuos sano y enfermo (positivo o negativo) seleccionados al azar.[22-25]

Los valores del ABC oscilan entre 0,5 (representado el azar) y el máximo es 1. Se suele aceptar como valor aceptable de discriminación cuando supera el valor 0,7.

El ABC es un estimador muestral de un parámetro poblacional, por ello se debe presentar junto a su IC 95 %. Si este intervalo incluyese el valor 0,50, no sería posible afirmar que el ABC es diferente a la no-discriminación (en el ejemplo que mostramos, el ABC de la variable glicemia tiene IC 95 % 0,47-0,57, lo que la hace una variable no discriminatoria para el análisis en cuestión). (Tabla 1 y Figura 5)

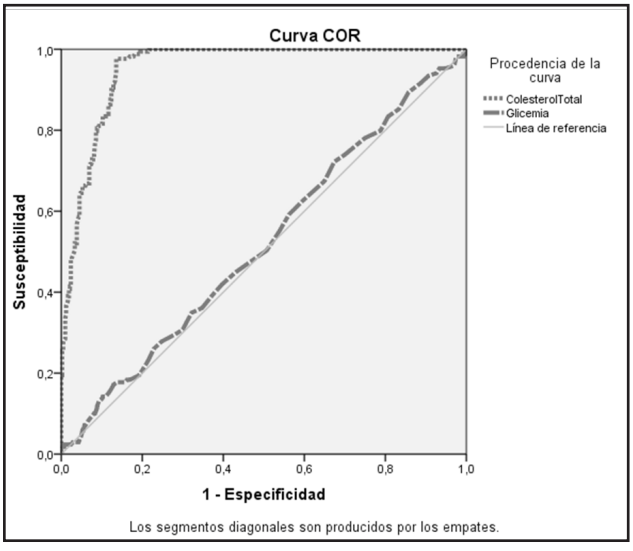
Tabla1. Área bajo la curva.

Variables resultado de contraste	Área	Error típ.[a]	Sig. asintótica[b]	Intervalo de confianza asintótico al 95 %	
				Límite inferior	Límite superior
Colesterol	0,953	0,008	0,000	0,938	0,968
Glicemia	0,520	0,026	0,444	0,469	0,572

a. Bajo el supuesto no paramétrico

b. Hipótesis nula: área verdadera=0,5

Figura 5. Gráfico de curva COR. Cada punto de la curva COR corresponde a un posible punto de corte del test diagnóstico, y nos informa su respectiva sensibilidad (eje Y) y 1-especificidad (eje X). Ambos ejes del gráfico incluyen valores entre 0 y 1 (0 % a 100 %). La línea trazada desde el punto 0,0 al punto 1,1 recibe el nombre de diagonal de referencia, o línea de no-discriminación.



El ABC puede ser usada además para determinar cuál de dos variables continuas tiene mejor capacidad discriminatoria para el diagnóstico que se desea realizar. Hanley & McNeil o DeLong describieron unos métodos que permiten comparar estadísticamente ambas áreas bajo la curva COR.[22-25]

Dado el uso generalizado y frecuente de la dicotomización de una covariable continua en las investigaciones de bioestadística y epidemiología, este trabajo ofrece una práctica metodología para la obtención de punto de corte óptimo.

Referencias bibliográficas

1. Altman DG. Categorising continuous variables. Br J Cancer. 1991;64:975.
2. Altman DG, Lausen B, Sauerbrei W, Schumacher M. Dangers of using "optimal"cutpoints in the evaluation of prognostic factors. Journal of the National Cancer Institute. 1994;86:829-835.
3. Altman, D. G. Categorizing continuous variables. In Armitage, P. and Colton, T. (eds). Encyclopedia of Biostatistics. Chichester: John Wiley;1998. pp.563 - 567.
4. Abdolell M, LeBlanc M, Stephens D, et al. Binary partitioning for continuous longitudinal data: categorizing a prognostic variable. Statist Med. 2002; 21:3395-3409.
5. Brent A. Williams, Jayawant N. Mandrekar, Sumithra J. Mandrekar, Stephen S. Cha, Alfred F. Furth. Finding Optimal Cutpoints for Continuous Covariates with Binary and Time-to-Event Outcomes. Technical Report Series #79. June 2006.
6. Contal, C., O'Quigley, J. An application of changepoint methods in studying the effect of age on survival in breast cancer. Computational Statistics and Data Analysis. 1999;30:253 - 270.
7. Faraggi D, Simon R. A simulation study of cross-validation for selecting an optimal cutpoint in univariate survival analysis. Statist Med. 1996;15:2203-2213.
8. Mazumdar M, Glassman JR. Categorizing a prognostic variable: review of methods, code for easy implementation and applications to decision-making about cancer treatments. Statistics in Medicine. 2000;19:113-132.
9. Mazumdar M, Smith A, Bacik J. Methods for categorizing a prognostic variable in a multivariable setting. Statistics in Medicine. 2003;22:559-571.
10. Baneshi MR, Talei AR. Dichotomisation of Continuous Data: Review of Methods, Advantages, and Disadvantages. Iran J Cancer Prev. 2011;4(1):26-32.
11. Luis M. Molinero. Elección de los puntos de corte para convertir una variable cuantitativa en cualitativa. URL disponible en: www.seh−lelha.org/stat1.htm.
12. Magder LS, Fix AD. Optimal choice of a cut point for a quantitative diagnostic test performed for research purposes. J Clin Epidemiol. 2003;56:956-962.
13. Mandrekar JN , Mandrekar SJ , Cha SS . Cutpoint determination methods in survival analysis using SAS®. Proceedings of the 28th SAS Users Group International Conference (SUGI). 2003:261-28.
14. Luis M. Molinero. ¿Y si los datos no siguen una distribución normal?...Bondad de ajuste a una normal. Transformaciones. Pruebas no paramétricas. URL disponible en: www.seh−lelha.org/stat1.htm.
15. Azzimonti Renzo JC. Bioestadística para Bioquímicos. Tema 4. Estadígrafos. URL disponible en: http://es.scribd.com/doc/2904463/Bioestadistica-Aplicada-a-Bioquimica-y-Farmacia.
16. Bioestadística amigable. Miguel A. Martínez-González, Almudena Sánchez-Villegas. Ediciones Díaz de Santos. ISBN: 84-7978-791-0. Depósito legal: M.40.343-2006.
17. Miller R, Siegmund D. Maximally selected chi square statistics. Biometrics. 1982;38:1011-1016.
18. Cumsille F, Bangdiwala SI, Sen PK, et al. Effect of dichotomizing a continuous variable on the model structure in multiple linear regression models. Commun Statist – Theory Meth. 2000;29:643-654.
19. Liquet B, Commenges D. Correction of the p-value after multiple coding of an explanatory variable in logistic regression. Statistics in Medicine. 2001;20:2815-2826.
20. Harrell, F.E. Jr. Regression modelling strategies with applications to linear models, logistic regression, and survival analysis. New York: Springer-Verlag; 2001.
21. Hollander N, Sauerbrei W, Schumacher M. Confidence intervals for the effect of a prognostic factor after selection of an `optimal' cutpoint. Statistics in Medicine. 2004;23:170-713.
22. Hanley J A, McNeil B J. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. Radiology. 1983; 148:839-43.
23. DeLong E R, DeLong D M, Clarke-Pearson D L. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics. 1988;44:837-45.
24. Jaime Cerda y Lorena Cifuentes. Uso de curvas ROC en investigación clínica. Aspectos teórico-prácticos. Rev Chil Infect. 2012;29(2):138-141.
25. Emma Domínguez Alonso y Roberto González Suárez. Análisis de las curvas receiver-operating characteristic: un método útil para evaluar procederes diagnósticos. Rev Cubana Endocrinol. 2002;13(2):173-80.