UNIVERSIDAD CARLOS III DE MADRID working papers

Working Paper 09-19
Statistics and Econometrics Series 07
March 2009

Departamento de Estadística
Universidad Carlos III de Madrid
Calle Madrid, 126
28903 Getafe (Spain)
Fax (34) 91 624-98-49

ADJUSTED EMPIRICAL LIKELIHOOD ESTIMATION OF THE YOUDEN INDEX AND ASSOCIATED THRESHOLD FOR THE BIGAMMA MODEL

Emilio Letón* and Elisa-María Molanes-López*

Abstract

The Youden index is a widely used measure in the framework of medical diagnostic, where the effectiveness of a biomarker (screening marker or predictor) for classifying a disease status is studied. When the biomarker is continuous, it is important to determine the threshold or cut-off point to be used in practice for the discrimination between diseased and healthy populations. We introduce a new method based on adjusted empirical likelihood for quantiles aimed to estimate the Youden index and its associated threshold. We also include bootstrap based confidence intervals for both of them. In the simulation study, we compare this method with a recent approach based on the delta method under the bigamma scenario. Finally, a real example of prostatic cancer, well known in the literature, is analyzed to provide the reader with a better understanding of the new method.

Keywords: Confidence interval; Empirical likelihood; Optimal cut-off point; ROC curve; Youden index.

* Department of Statistics, Universidad Carlos III de Madrid, Avda. de la Universidad 30, 28911 Leganés (Madrid), Spain, e-mail: emilio.leton@uc3m.es.

Adjusted empirical likelihood estimation of the Youden index and associated threshold for the bigamma model

Emilio LETÓN [*] Elisa-María MOLANES -LÓPEZ[*]

Abstract

The Youden index is a widely used measure in the framework of medical diagnostic, where the effectiveness of a biomarker (screening marker or predictor) for classifying a disease status is studied. When the biomarker is continuous, it is important to determine the threshold or cut-off point to be used in practice for the discrimination between diseased and healthy populations. We introduce a new method based on adjusted empirical likelihood for quantiles aimed to estimate the Youden index and its associated threshold. We also include bootstrap based confidence intervals for both of them. In the simulation study, we compare this method with a recent approach based on the delta method under the bigamma scenario. Finally, a real example of prostatic cancer, well known in the literature, is analyzed to provide the reader with a better understanding of the new method.

Key Words: Confidence interval; Empirical likelihood; Optimal cut-off point; ROC curve; Youden index.

[*] Department of Statistics, Universidad Carlos III de Madrid, Avda. de la Universidad 30, 28911 Leganés (Madrid), Spain, e-mail: emilio.leton@uc3m.es.

1 Introduction

Diagnostic tests are often used for classifying diseased and healthy populations. They are based on biomarkers, and can be dicothomous, ordinal or continuous. From here on, we will focus on continuous biomarkers which are the most used in practice.

As Le (2006) points out, sometimes larger values of the biomarker, denoted by X, are associated with the diseased population (fasting blood glucose for diabetes, prostatic specific antigen (PSA) for prostate cancer, antibodies for infections, etc.), but other times is the opposite, with smaller values of X associated with the disease population (static admittance for ear infection, thyroid-specific hormone (TSH) for hyperthyroidism, etc.). We will asume along the paper, without loss of generality, the former case.

In this context, a person is classified as 'diseased' if the corresponding biomarker value is greater than a given threshold value, and is classified as 'healthy' otherwise. Denoting by c that threshold value, there are two important probabilities associated with it: the sensitivity, $q(c)$, and the specificity, $p(c)$. The sensitivity is defined as 'true positive subjects', i.e. correctly classified diseased individuals and the specificity as 'true negative subjects', i.e correctly classified healthy individuals. The pairs $(1-p(c), q(c))$, for all possible threshold values c, are usually drawn in a plot called the ROC curve. The name ROC stands for 'Receiver operating characteristic' because was originally created in the context of radar technology, in order to distinguish the signal and the noise (Erdreich and Lee, 1981). The first application of the ROC curves was for detecting arrivals of missiles.

The ROC curve describes graphically the performance of the biomarker under several cut-off points (see, for example, Pepe, 2003). The area under the ROC curve is denoted by AUC, and it is a summary measure of the accuracy of the diagnostic test. While a value of $AUC = 1$ represents a perfect test, a value of $AUC = 0.5$ represents a test that performs exactly the same as if we had used a fair coin (50-50 chance) as a diagnostic test.

A key point in this methodology is to find an optimal threshold, in order to maximize the effectiveness of the biomarker. In most instances, there is an inverse relationship between sensitivity and specificity, in the sense that moving the threshold increases one while decreasing the other. So a kind of balance between sensitivity and specificity is necessary.

There exist two main methods for identifying the optimal threshold: the northwest corner and the Youden index. These two methods can give different cut-off points as Perkins and Schisterman (2006) point out. From here on, we will concentrate on the latter defined by

Youden (1950) and recently studied by Fluss et al. (2005), Schisterman et al. (2005), Le (2006), and Schisterman and Perkins (2007), among others.

In order to maximize the effectiveness of the biomarker, the Youden index, J, is defined as follows,

$$J = \max\{J(c); c \in \Re\}, \text{ where } J(c) = q(c) + p(c) - 1.$$

The notation X_0 and X_1 will be used to refer to the values of the biomarker on the healthy and diseased populations, respectively. Denoting by F_0 and F_1 their corresponding cumulative distribution functions (cdf's), and by $\bar{F}_0$ and $\bar{F}_1$ their complementary ones, it follows that

$$\begin{aligned} J(c) &= \Pr(X_1 > c) + \Pr(X_0 < c) - 1 \qquad (1) \\ &= \bar{F}_1(c) + F_0(c) - 1 \\ &= \bar{F}_1(c) - \bar{F}_0(c) = F_0(c) - F_1(c). \end{aligned}$$

It is easy to see that J is the maximal vertical distance from the ROC curve to the diagonal or chance line. Another way of visualizing J is that $J(c)$ in (1) can also be determined computing the difference of the area under f_1 and f_0 to the right of the cut-off point c (see, for instance, Schisterman et al., 2005), with f_0 and f_1 the density functions associated to F_0 and F_1, respectively. Using the latter reasoning, the maximum difference, J, is achieved at the corresponding c value where f_0 and f_1 intersect. Additionally, the Youden index can be seen as the Kolmogorov-Smirnov distance between X_0 and X_1 (Pepe, 2003).

The paper is organized as follows. In Section 2, we revise the delta method for computing the optimal J and c, and their confidence intervals applied to the bigamma model. In Section 3, we propose a new method, not requiring parametric assumptions. Through a simulation study we check its performance in Section 4, where it is shown to be competitive with the delta method. Finally, in Section 5, the new approach is illustrated through a well known real example.

2 Delta method

The delta method, based on Taylor's theorem, is useful to approximate the moments of transformed random variables, using the moments of the original non-transformed ones. The main application of the delta method is for computing the variance of transformed random variables, from which approximate confidence intervals can be obtained (see, for instance, Miller (1981) and Collett (2003)).

In the univariate case, let X be a random variable with mean μ_X and variance $\mathbb{V}\text{ar}[X]$ and consider the transformation $h(X)$. Applying the delta method, it follows that

$$\mathbb{V}\text{ar}[h(X)] \simeq \mathbb{V}\text{ar}\left[h(\mu_x) + h'(\mu_x)(X - \mu_x)\right] = [h'(\mu_x)]^2 \mathbb{V}\text{ar}[X].$$

On the other hand, in the bivariate case, let X and Y be two random variables, with mean vector (μ_x, μ_y) and covariance matrix given by the expression below,

$$\begin{pmatrix} \mathbb{V}\text{ar}[X] & \mathbb{C}\text{ov}(X,Y) \\ \mathbb{C}\text{ov}(X,Y) & \mathbb{V}\text{ar}[X] \end{pmatrix}$$

and consider the transformation $h(X,Y)$. Once again, based on the delta method, the variance of $h(X,Y)$ can be approximated by

$$\begin{aligned} \mathbb{V}\text{ar}[h(X,Y)] &\simeq \mathbb{V}\text{ar}\left[h(\mu_x,\mu_y) + \frac{\partial h}{\partial x}(\mu_x,\mu_y)(X-\mu_x) + \frac{\partial h}{\partial y}(\mu_x,\mu_y)(Y-\mu_y)\right] \\ &= \left[\frac{\partial h}{\partial x}(\mu_x,\mu_y)\right]^2 \mathbb{V}\text{ar}[X] + \left[\frac{\partial h}{\partial y}(\mu_x,\mu_y)\right]^2 \mathbb{V}\text{ar}[Y] \\ &+ 2\left[\frac{\partial h}{\partial x}(\mu_x,\mu_y)\right]\left[\frac{\partial h}{\partial x}(\mu_x,\mu_y)\right]\mathbb{C}\text{ov}(X,Y). \end{aligned}$$

In the context of ROC curves, a 4-dimensional version of the delta method can be applied to the bigamma model (Schisterman and Perkins, 2007) for approximating the variance of the Youden index and the associated threshold. With these approximated variances and the corresponding point estimates, $\hat{J}$ and $\hat{c}$, the following asymptotic $(1-\alpha)100\%$ confidence intervals for J and c are obtained:

$$CI_{(1-\alpha)100\%}(J) = \hat{J} \mp z_{1-\alpha/2}\sqrt{\widehat{\mathbb{V}\text{ar}}[\hat{J}]}, \quad \text{and} \quad CI_{(1-\alpha)100\%}(c) = \hat{c} \mp z_{1-\alpha/2}\sqrt{\widehat{\mathbb{V}\text{ar}}[\hat{c}]},$$

with $z_{1-\alpha/2}$ referring to the $(1-\alpha/2)$-quantile of the standard gaussian distribution.

Since the focus of our paper is on the bigamma model, we introduce it in the following. The bigamma model is given by two gamma distributed random variables, $X_0 \sim \gamma(\alpha_0, \beta_0)$ and $X_1 \sim \gamma(\alpha_1, \beta_1)$, with density functions defined by

$$f_i(\alpha_i, \beta_i, x) = \frac{e^{-x/\beta_i} x^{\alpha_i - 1}}{\beta_i^{\alpha_i}\Gamma(\alpha_i)}, \text{ with } x > 0,$$

where $\alpha_i > 0$ and $\beta_i > 0$ are the shape and scale parameters, for $i = 0, 1$, and Γ denotes the gamma function

$$\Gamma(p) = \int_0^\infty e^{-x} x^{p-1} \mathrm{d}x, \text{ with } p > 0.$$

Now, for the bigamma model, $J(c)$ in (1) can be rewritten as follows

$$J(c) = \int_{c(\alpha_0,\alpha_1,\beta_0,\beta_1)}^\infty \frac{e^{-x/\beta_1} x^{\alpha_1 - 1}}{\beta_1^{\alpha_1} \Gamma(\alpha_1)} \mathrm{d}x - \int_{c(\alpha_0,\alpha_1,\beta_0,\beta_1)}^\infty \frac{e^{-x/\beta_0} x^{\alpha_0 - 1}}{\beta_0^{\alpha_0} \Gamma(\alpha_0)} \mathrm{d}x.$$

For the sake of easy reading, we will denote from here on $c(\alpha_0, \alpha_1, \beta_0, \beta_1)$ by c.

Based on the delta method and assuming independence between X_0 and X_1, the variance of $\hat{J}$ can be approximated as follows

$$\begin{aligned} \mathbb{V}\mathrm{ar}[\hat{J}] &\simeq \left(\frac{\partial J}{\partial \alpha_1}\right)^2 \mathbb{V}\mathrm{ar}[\hat{\alpha_1}] + \left(\frac{\partial J}{\partial \beta_1}\right)^2 \mathbb{V}\mathrm{ar}[\hat{\beta_1}] \\ &+ \left(\frac{\partial J}{\partial \alpha_0}\right)^2 \mathbb{V}\mathrm{ar}[\hat{\alpha_0}] + \left(\frac{\partial J}{\partial \beta_0}\right)^2 \mathbb{V}\mathrm{ar}[\hat{\beta_0}] \\ &+ 2\left(\frac{\partial J}{\partial \alpha_1}\right)\left(\frac{\partial J}{\partial \beta_1}\right) \mathbb{C}\mathrm{ov}(\hat{\alpha_1}, \hat{\beta_1}) \\ &+ 2\left(\frac{\partial J}{\partial \alpha_0}\right)\left(\frac{\partial J}{\partial \beta_0}\right) \mathbb{C}\mathrm{ov}(\hat{\alpha_0}, \hat{\beta_0}). \end{aligned} \tag{2}$$

Analogously, an expression for $\mathbb{V}\mathrm{ar}[\hat{c}]$ can be obtained using the delta method. The only difference is that the partial derivatives of J appearing in (2) are now replaced by the partial derivatives of c.

The variances and covariances appearing in the right-hand side of (2) can be found in Schisterman and Perkins (2007). Below, the expressions for the rest of terms in the right-hand side of (2) have been rewritten in a more compact way than in Schisterman and Perkins (2007), correcting some minor missprints. Moreover, closed forms for the computation of the partial derivatives of c, appearing in those expressions, have been obtained, inspired in the envelope of curves (see, for instance, Hairer and Wanner, 1996). This approach for computing the partial derivatives of c differs from that used in Schisterman and Perkins (2007), who approximate them numerically. We detail below the four partial derivatives of J:

$$\frac{\partial J}{\partial \alpha_1}(c) = \bar{F}_1(c)\left[-\frac{\delta_1}{\Gamma(\alpha_1)} - \log(\beta_1)\right] + \int_c^\infty f_1(x)\log(x)\mathrm{d}x + \frac{\partial c}{\partial \alpha_1}(f_0(c) - f_1(c)),$$

$$
\begin{aligned}
\frac{\partial J}{\partial \beta_1}(c) &= \frac{\alpha_1}{\beta_1}\left[\bar{F}_1(\alpha_1+1,\beta_1,c)-\bar{F}_1(c)\right]+\frac{\partial c}{\partial \beta_1}(f_0(c)-f_1(c)),\\
\frac{\partial J}{\partial \alpha_0}(c) &= \bar{F}_0(c)\left[\log(\beta_0)+\frac{\delta_0}{\Gamma(\alpha_0)}\right]-\int_c^\infty f_0(x)\log(x)\mathrm{d}x+\frac{\partial c}{\partial \alpha_0}(f_0(c)-f_1(c)),\\
\frac{\partial J}{\partial \beta_0}(c) &= \frac{\alpha_0}{\beta_0}[\bar{F}_0(c)-\bar{F}_0(\alpha_0+1,\beta_0,c)]+\frac{\partial c}{\partial \beta_0}(f_0(c)-f_1(c)),
\end{aligned}
$$

where, for $i=0,1$, $\delta_i=\frac{\partial}{\partial\alpha_i}\Gamma(\alpha_i)$, $\frac{\delta_i}{\Gamma(\alpha_i)}=\frac{\partial}{\partial\alpha_i}\log(\Gamma(\alpha_i))$ is known as the digamma function, $\bar{F}_i(\alpha_i+1,\beta_i,c)$ denotes the complementary cdf of a gamma distributed random variable, with parameters α_i+1 and β_i, and the four partial derivatives of c are given by the closed forms below

$$
\begin{aligned}
\frac{\partial c}{\partial \alpha_0} &= \frac{\frac{-\partial^2 J}{\partial\alpha_0\partial c}}{\frac{\partial^2 J}{\partial c^2}} = \frac{-\frac{\partial}{\partial\alpha_0}f_0(x)}{\frac{\partial}{\partial c}(f_0(c)-f_1(c))}, \qquad & \frac{\partial c}{\partial \beta_0} &= \frac{\frac{-\partial^2 J}{\partial\beta_0\partial c}}{\frac{\partial^2 J}{\partial c^2}} = \frac{-\frac{\partial}{\partial\beta_0}f_0(x)}{\frac{\partial}{\partial c}(f_0(c)-f_1(c))},\\
\frac{\partial c}{\partial \alpha_1} &= \frac{\frac{-\partial^2 J}{\partial\alpha_1\partial c}}{\frac{\partial^2 J}{\partial c^2}} = \frac{\frac{\partial}{\partial\alpha_1}f_1(x)}{\frac{\partial}{\partial c}(f_0(c)-f_1(c))}, \qquad & \frac{\partial c}{\partial \beta_1} &= \frac{\frac{-\partial^2 J}{\partial\beta_1\partial c}}{\frac{\partial^2 J}{\partial c^2}} = \frac{\frac{\partial}{\partial\beta_1}f_1(x)}{\frac{\partial}{\partial c}(f_0(c)-f_1(c))}.
\end{aligned}
$$

Note that, based on the fact that $\frac{\partial J}{\partial c}=0$ when evaluated at the optimal threshold, it is satisfied that

$$
\frac{\partial^2 J}{\partial\alpha_0\partial c}+\frac{\partial^2 J}{\partial^2 c}\frac{\partial c}{\partial\alpha_0}=0,
$$

from where the closed form of $\frac{\partial c}{\partial\alpha_0}$, previously detailed, is straightforwardly obtained. Similarly, the other partial derivatives of c are obtained.

3 New method

Likelihood-based methods can deal with incomplete data, pool information from different sources and, when there exists extra information from the outside, they can include it as constraints, restricting the domain of the likelihood function, or as prior distributions multiplying the likelihood function. On the other hand, parametric assumptions can yield wrong estimates when the model is misspecified. Nonparametric estimates, however, avoid this misspecification inherent to parametric model-based estimates. The combination of these two methodologies has the advantage of using likelihood methods without the restriction of having to assume that the data follow a known parametric model of distributions. The combination of likelihood and nonparametric methods has been termed in the literature as empirical likelihood. It was first proposed by Thomas and Grunkemeier (1975) to obtain better confidence intervals for the Kaplan-Meier estimator (see Kaplan and Meier, 1958). Later on, Owen (1990, 2001) and other authors have shown the potential of this approach, which nowadays is still an active area of research (see, for instance, Cao and Van Keilegom, 2006, Chen et al., 2009, Hjort et al., 2009, and Molanes-López et al., 2009). One of the main advantages of empirical likelihood based confidence intervals is that they respect the range of the parameter space, are invariant under transformations and their shape is data-driven.

We present a new approach, based on EL and bootstrapping, for estimating the optimal threshold and the associated Youden index, and their corresponding confidence intervals. However, since in the medical field is more relevant to know which is the cut-off of the biomarker to classify the individuals, our main focus is on correctly estimate the optimal threshold. As a byproduct of our method, we get as well an estimate of J. Besides, the new method has the additional advantages of easy implementation and not requiring any particular parametric assumption.

Before going on more details, we first require to introduce the concept of relative distribution (see Handcock and Morris, 1999, for more details), which is very related to the concept of ROC curve. Specifically, the relative distribution of X_1 with respect to (w.r.t.) X_0 is defined as the cdf of the random variable $Z = F_0(X_1)$, i.e.

$$R_{01}(t) = \Pr(Z \leq t) = \Pr(F_0(X_1) \leq t) = F_1\left(F_0^{-1}(t)\right).$$

To provide the reader with some insight on the interpretation of $R_{01}(t)$ for a fixed $t \in (0,1)$, let denote by $s = R_{01}(t)$, with $s \in (0,1)$. Then, $F_1(c) = s$ for some c in $\Re$ such as $F_0(c) = t$, i.e. c is the s-th quantile of X_1 and the t-th quantile of X_0. On the other hand, it is easy to see that $R_{01}(t)$ is a reparametrization of the ROC curve,

$$R_{01}(t) = 1 - ROC(1-t),$$

where $ROC(t) = \bar{F}_1\left(\bar{F}_0^{-1}(t)\right)$, for $t \in (0,1)$, denotes the ROC curve, i.e. the cdf of the random variable $1 - Z$, known in the literature as 'the placement value' (see Cai, 2004 and Pepe and Cai, 2004). It is interesting to note here that when a fair coin is used as a diagnostic test to classify the individuals, Z and $1 - Z$ both follow a uniform distribution in $(0,1)$ and consequently J=0 and there is not optimal c to distinguish between both populations. Moreover, since $E[Z] = \Pr(X_0 \leq X_1) = AUC$ (Bamber, 1975), when a fair coin is used as a diagnostic test, it follows that $AUC = 0.5$, the minimum attainable value for AUC.

Consider $\{X_{0k}\}_{k=1}^{n_0}$ and $\{X_{1k}\}_{k=1}^{n_1}$, two independent samples taken from both populations,

X_0 and X_1, with sample sizes n_0 and n_1, respectively. Based on these observations, we detail below the 4 steps of our method to obtain estimates of J and c.

Step 1. We obtain $\hat{R}_{01}(t)$, a kernel-type estimate of the relative distribution of X_1 w.r.t. X_0,

$$\hat{R}_{01}(t) = \frac{1}{n_1} \sum_{k=1}^{n_1} G\left(\frac{t - F_{0n_0}(X_{1k})}{h_1} \right), \tag{3}$$

and then we find the value t, let say t_0, that maximizes the distance between $\hat{R}_{01}(t)$ and t. In equation (3) above, $G(x) = \int_{-\infty}^{x} K(y)\mathrm{d}y$, K denotes a kernel function, h_1 is the smoothing parameter, also known as bandwidth, and F_{0n_0} refers to the empirical cdf of X_0.

Note that, since F_0 is assumed unknown, it is required to estimate it through F_{0n_0}. Consequently, $\hat{R}_{01}(t)$ in (3) can be seen as a traditional kernel-type cdf estimate of Z, based on the pseudosample $\{F_{0n_0}(X_{1k})\}_{k=1}^{n_1}$ rather than on the unobserved sample $\{F_0(X_{1k})\}_{k=1}^{n_1}$, straightforwardly drawn from Z. Note that $F_{0n_0}(X_{1k})$ above gives the rank of X_{1k} in the healthy sample $\{X_{01}, \ldots, X_{0n_0}\}$.

Analogously, interchanging the roles of X_0 and X_1, we obtain $\hat{R}_{10}(t)$, a kernel-type estimate of the relative distribution of X_0 w.r.t. X_1, and find the value t, let say t_1, that maximizes the distance between $\hat{R}_{10}$ and t.

Step 2. With the two values previously computed, t_0 and t_1, we then apply the adjusted EL method for quantiles proposed by Zhou and Jing (2003), and estimate the t_0-th quantile of the healthy population, $c_0 = F_0^{-1}(t_0)$, and the t_1-th quantile of the diseased population, $c_1 = F_1^{-1}(t_1)$.

Specifically, for $i = 0, 1$, we find the value $\hat{c}_i$, that minimizes the adjusted log-empirical likelihood ratio given by the expression below,

$$\hat{\ell}(c_i) \quad = \quad 2n_i \left(\hat{F}_i(c_i) \log \frac{\hat{F}_i(c_i)}{t_i} + (1 - \hat{F}_i(c_i)) \log \frac{1 - \hat{F}_i(c_i)}{1 - t_i} \right),$$

where $\hat{F}_i$ denotes a kernel-type estimate of F_i,

$$\hat{F}_i(x) = \frac{1}{n_i} \sum_{k=1}^{n_i} G\left(\frac{x - X_{ik}}{g_i} \right), \tag{4}$$

with g_i the smoothing parameter.

It is interesting to note here that, although most of the empirical likelihood approaches lead to log-likelihood functions implicitly defined by a nonlinear equation, this is not the case for the approach of Zhou and Jing (2003), where a closed form is available for the log-likelihood function.

Step 3. With the two estimates previously computed in Step 2, $\hat{c}_1$ and $\hat{c}_2$, we then propose $\hat{c} = \frac{n_0}{n}\hat{c}_0 + \frac{n_1}{n}\hat{c}_1$ as an estimate of the optimal threshold c, where $n = n_0 + n_1$. Finally, as a byproduct, an estimate of the Youden index is given by $\hat{J} = \hat{F}_0(\hat{c}) - \hat{F}_1(\hat{c})$, where $\hat{F}_i$ has been previously introduced in equation (4).

Step 4. In order to obtain confidence intervals for c and J, we resample independently from both populations and repeat the three steps given above a large number of times, let say B, using the bootstrap resamples. These bootstrap resamples are drawn from smoothed versions of the corresponding empirical cdf's.

Finally, the confidence intervals for the Youden index J and the optimal threshold c are given by the percentile method.

4 Simulation study

A study of interval width and coverage probability is done through a simulation study based on the bigamma model with several parameters, similar to those considered by other authors. Within the parametric assumptions in Schisterman and Perkins (2007), the bigamma model can deal with asymmetric situations what make it more realistic and flexible than the binormal model in real applications. The specific values, under the bigamma assumption, for the shape and scale parameters of the healthy population were fixed to $\alpha_0 = 1.5$ and $\beta_0 = 1$. However, the parameters of the diseased population were accordingly selected to yield different values of J, as collected in Table 1.

	Youden index J			
Shape parameter α_1 of X_1	$J = 0.4$	$J = 0.6$	$J = 0.8$	$J = 0.9$
$\alpha_1 = 1.5$	2.4828	4.3565	9.7847	19.8020
$\alpha_1 = 2.0$	1.6622	2.7650	5.6517	10.3842

Table 1: *Bigamma model: scale parameter of diseased population, β_1, with $\alpha_0 = 1.5$ and $\beta_0 = 1$.*

The simulations were carried out in MATLAB. For every scenario specified in Table 1, 300 trials were considered. For each trial, a sample of $n_0 = 50$ i.i.d. observations, $\{X_{01}, \ldots, X_{0n_0}\}$, and a sample of $n_1 = 50$ i.i.d. observations, $\{X_{11}, \ldots, X_{1n_1}\}$, were independently drawn from X_0 and X_1, respectively. The uniform kernel, K, was considered to estimate the relative distributions involved in the first step of our algorithm, R_{01} and R_{10}, and the cdf's, F_i, for $i = 0, 1$, required to estimate J in Step 3. For these kernel type estimates we considered the following bandwidths, $h_i = n_i^{-1/3}$ and $g_i = 2n_i^{-1/3}$, for $i = 0, 1$, that are of optimal order to estimate cdf's in two-sample and one-sample problems. However, given the regularity conditions required by the adjusted empirical likelihood method for quantiles of Zhou and Jing (2003), we used in (4) bandwidths given by $n_i^{-1/2}$, for $i = 0, 1$, and a kernel different from the commonly used in nonparametric density estimation,

$$K(x) = \left\{ \frac{21 - 9\sqrt{21}}{8} x^2 + \frac{-3 + 3\sqrt{21}}{8} \right\} 1_{\{|x| \le 1\}}.$$

From each pair of samples, we generated $B = 299$ bootstrap resamples to obtain 95%-confidence intervals for the optimal cut-off point c and the Youden index J. Although in classical resampling methodology the resamples are drawn from the empirical cdf's, we have used instead kernel type cdf estimates of F_i with gaussian kernel and bandwidths given by $0.2 \max\{\text{iqr}(\{X_{ik}\}_{i=1}^{n_i}), \text{std}(\{X_{ik}\}_{i=1}^{n_i})\} n_i^{-1/5}$, for $i = 0, 1$, with iqr and std referring to, respectively, the sample interquartile range and sample standard deviation.

We would like to remark here that sometimes, in Step 2 of our algorithm, it may be required to deal with upper and lower quantiles, more extreme than those considered by Zhou and Jing (2003). For instance, for the bigamma model with parameters $\alpha_0 = 1.5$ $\beta_0 = 1$ $\alpha_1 = 2$ and $\beta_1 = 10.38$, which corresponds to a setting of $J = 0.90$, it is necessary to estimate the 0.97-quantile of the healthy population, which can be very challenging, specially if the sample size is small.

We collected in Tables 2 and 3 the results from this simualion study. To get a more visual understanding of them we also included Figure 1. For the sake of simplicity, we will refer to the new method by ELM (Empirical Likelihood Method). From the results collected in Tables 2 and 3, we observe that in general the ELM for CI's of both parameters of interest tend to present overcoverage, while those based on the delta method present undercoverage. In terms of width average, the ELM for the CI of c behaves better than the other when the biomarker X separates both populations reasonably well (see, for instance, the results in Table 2 for $J = 0.8, 0.9$).

Bigamma model: $CI_{95\%}(c)$ with $(n_0, n_1) = (50, 50)$					
		ELM		Delta method	
α_1	J	coverage(%)	width	coverage(%)	width
1.5	0.4	96.33	1.3734	90.33	1.0598
1.5	0.6	96.00	1.4164	93.00	0.9643
1.5	0.8	99.00	1.3263	93.33	1.3414
1.5	0.9	95.33	1.9039	94.33	1.9033
2.0	0.4	96.33	1.2632	96.33	4.1990
2.0	0.6	98.33	1.2676	96.33	1.1748
2.0	0.8	94.67	1.3755	94.67	1.4047
2.0	0.9	93.33	1.5828	93.33	1.9389

Table 2: *Coverage probabilities and width averages of* $CI_{95\%}(c)$.

It is also interesting to point out the results collected in the third row of Table 2. While the ELM for estimating c presents the higher observed overcoverage, 99%, it shows a width average shorter than the delta method (a width average of 1.3263 versus 1.3414). On the other hand, an isolated case has been observed for the delta method in the fifth row of Table 2, where an atypical trial had a negative effect on the width average.

However, when estimating J, the width of the ELM for the CI is larger than the width of the delta method. This can be explained due to the fact that our approach is focused on correctly estimating c, and once $\hat{c}$ is computed, $\hat{J}$ is obtained as a byproduct. As it was already observed in the literature, even though J and c are strongly related, a good method for estimating one of them is not necessarily good for the other (see Fluss et al., 2005).

From the previous discussion on the results, we conclude that the new confidence intervals have good performance in terms of nominal coverage and width, being competitive with the delta method, recently used by Schisterman and Perkins (2007) under parametric assumptions. The delta method is dependent on distributional assumptions, and violations of them can yield substantial bias in estimation. Therefore, we suggest using the new method when the underlying distributions, F_0 and F_1, are unknown, although it is more time-consuming than the delta method.

In view of the promising results of the new methodology we plan to extend this simulation study to other models, different sample sizes, including balanced and non-balanced designs,

Bigamma model: $CI_{95\%}(J)$ with $(n_0, n_1) = (50, 50)$					
		ELM		Delta method	
α_1	J	coverage(%)	width	coverage(%)	width
1.5	0.4	95.67	0.3301	95.67	0.2757
1.5	0.6	97.67	0.2860	94.67	0.2432
1.5	0.8	96.67	0.2123	94.33	0.1790
1.5	0.9	96.33	0.1746	93.67	0.1220
2.0	0.4	97.00	0.3286	92.33	0.2829
2.0	0.6	96.67	0.2892	93.00	0.2485
2.0	0.8	94.67	0.2106	93.00	0.1810
2.0	0.9	92.67	0.1675	92.00	0.1219

Table 3: *Coverage probabilities and width averages of* $CI_{95\%}(J)$.

and a larger number of trials to get more reliable estimates of coverage probabilities and width averages. We also plan to incorporate more sophisticated data-driven bandwidth selectors and scenarios where different costs are assumed for the two types of errors involved (false positives and false negatives).

5 Example

A real example analyzed in Le (2006) is used to illustrate the application of the new approach. There are 53 patients with prostate cancer: 20 out of them with nodal involvement and 33 without. The biomarker used in this example is the level of acid phosphatase in blood serum (×100).

It is easy to check that these data do not follow any of the parametric models (binormal or bigamma) studied in Schisterman and Perkins (2007) via the delta method. Consequently, a straightforward application of the delta method would not be possible. First, the appropriate parametric model should be find, which not always may be possible, and then all the formulation required by the delta method should be rewritten.

After analyzing this example using our method, which does not require any parametric assumption, we obtain a point estimate of $\hat{c} = 60.67$ for the optimal threshold and the following confidence interval $CI_{95\%}(c) = (51.40, 67.50)$. The point estimate of c differs from that

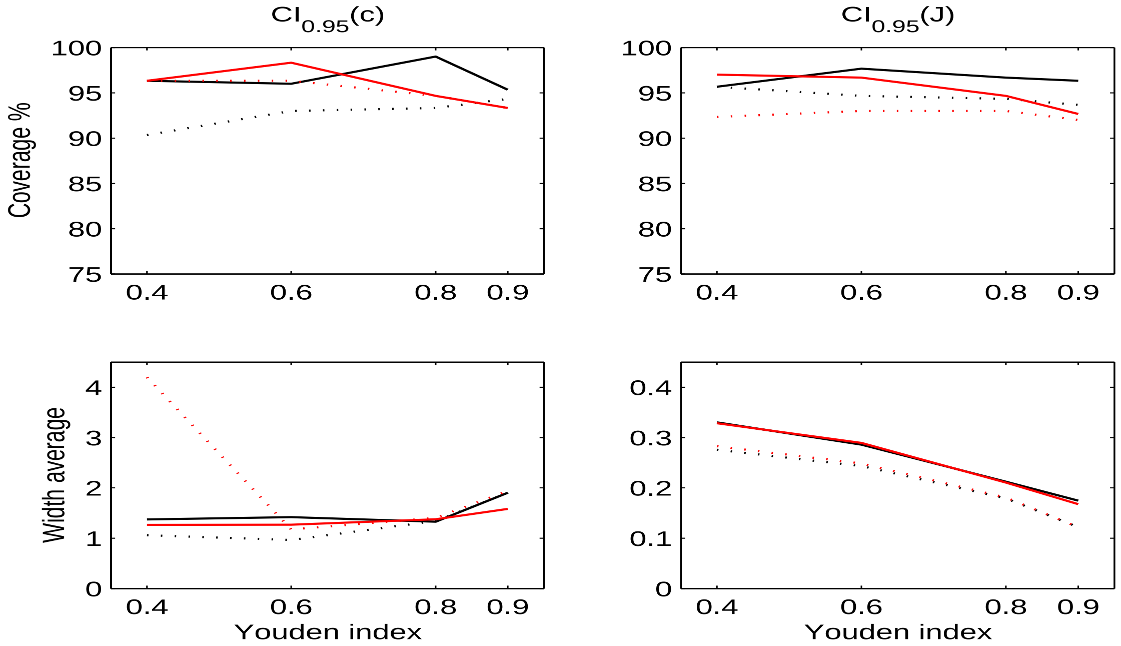

Figure 1: *Coverages and width averages of $CI_{95\%}(c)$ and $CI_{95\%}(J)$ using ELM (solid line) and the delta method (dotted line) with $\alpha_1 = 1.5$ (black) and $\alpha_1 = 2$ (red).*

obtained in Le (2006), $\hat{c} = 75.00$, who proposed to model the ROC function by proportional hazards model, also called the Lehmann's alternatives. Notice that this Lehmann-based estimate is outside our confidence interval. This suggests that the assumption of Lehmann's alternatives may be not tenable for this data set.

References

[1] Bamber, D.C. (1975). The area above the ordinal dominance graph and the area below the receiver operating characteristic curve graph, *Journal of Mathematical Psychology*, **12**, 387–415.

[2] Cai, T. (2004). Semi-parametric ROC regression analysis with placement values, *Biostatistics*, **5**, 45–60.

[3] Cao, R. & Van Keilegom, I. (2006). Empirical likelihood tests for two-sample problems via nonparametric density estimation, *Scandinavian Journal of Statistics*, **34**, 61–77.

[4] Chen, J., Peng, L. & Zhao, Y. (2009). Empirical likelihood based confidence intervals for copulas, *Journal of Multivariate Analysis*, **100**, 137–151.

[5] Collett, D. (2003). *Modelling survival data in medical research.* Florida: Chapman and Hall.

[6] Erdreich, L.S. & Lee, E.T. (1981). Use of relative operating characteristic analysis in epidemiology, *American Journal of Epidemiology*, **114**, 649–662.

[7] Fluss, R., Faraggi, D. & Reiser, B. (2005). Estimation of the Youden index and its associated cutoff point, *Biometrical Journal*, **47**, 458–472.

[8] Hairer, E. & Wanner, G. (1996). *Analysis by its history.* New York: Springer.

[9] Handcock, M.S. & Morris, M. (1999). *Relative distribution methods in social sciences.* New York: Springer.

[10] Hjort, N.L., McKeague, I.W. & Van Keilegom, I. (2009). Extending the scope of empirical likelihood, *Annals of Statistics* (in press).

[11] Kaplan, E.L & Meier, P. (1958). Non-parametric estimation from incomplete observations, *Journal of the American Statistical Association,* **53**, 457–481.

[12] Le, C.T. (2006). A solution for the most basic optimization problem associated with an ROC curve, *Statistical Methods in Medical Research*, **15**, 571–584.

[13] Miller, R.G.Jr. (1981). *Survival Analysis.* New York: John Wiley & Sons.

[14] Molanes-López, E.M., Van Keilegom, I. & Veraverbeke, N. (2009). Empirical likelihood for non-smooth criterion functions, *Scandinavian Journal of Statistics* (in press).

[15] Owen, A.B. (1990). Empirical likelihood ratio confidence regions, *The Annals of Statistics*, **18**, 90–120.

[16] Owen, A.B. (2001). *Empirical likelihood.* New York, Chapman & Hall.

[17] Pepe, M.S. (2003) *The statistical evaluation of medical tests for classification and prediction.* New York: Oxford University Press.

[18] Pepe, M.S. & Cai, T. (2004). The analysis of placement values for evaluating discriminatory measures, *Biometrics*, **60**, 528–535.

[19] Perkins, N.J. & Schisterman, E.F (2006). The inconsistency of “optimal” cutpoints obtained using two criteria based on the receiver operating characteristic curve, *American Journal of Epidemiology*, **163**, 670–675.

[20] Schisterman, E.F. & Perkins, N.J. (2007). Confidence intervals for the Youden index and corresponding optimal cut-point, *Communications in Statistics – Simulation & Computation*, **36**, 549–563.

[21] Schisterman, E.F., Perkins, N.J., Aiyi, L. & Bondell, H. (2005). Optimal cut-point and its corresponding Youden Index to discriminate individuals using pooled blood samples, *Epidemiology*, **16**, 73–81.

[22] Thomas, D.R. & Grunkemeier, G.L. (1975). Confidence interval estimation of survival probabilities for censored data, *Journal of the American Statistical Association,* **70**, 865–871.

[23] Youden, W.J. (1950). Index for rating diagnostic tests, *Cancer*, 3, 32–35.

[24] Zhou, W. & Jing, B.Y. (2003). Adjusted empirical likelihood method for quantiles, *Annals of the Institute of Statistical Mathematics*, **55**, 689–703.