Confidence regions for optimal sensitivity and specificity of a diagnostic test

Regioni di confidenza per sensibilità e specificità ottimali di un test diagnostico

Gianfranco Adimari, Duc-Khanh To and Monica Chiogna

Abstract We propose new methods that provide approximate joint confidence regions for the optimal sensitivity and specificity of a diagnostic test, fixed by the Youden index criterion. Such methods are semiparametric and overcome limitations of alternative approaches available in the literature. Our proposal is based on empirical likelihood pivots and covers two situations: binormal model and binormal model after the use of Box-Cox transformations. In the last case, we show how to use two different transformations, for the healthy and the diseased subjects.

Key words: Empirical likelihood, Box-Cox transformation, Youden index, ROC analysis, bootstrap calibration

1 Introduction and background

The accuracy of a diagnostic test can be assessed by its receiver operating characteristic (ROC) curve. The test result can be dichotomized at a specified cutpoint. Given the cutpoint, the sensitivity is the probability of a true positive, i.e., the probability that the test correctly identifies a diseased subject. The specificity is the probability of a true negative, i.e., the probability that the test correctly identifies a non-diseased subject. When one varies the cutpoint throughout the entire real line, the resulting pairs (1 − specificity, sensitivity) form the ROC curve (see Pepe[1], as general reference). Let X denote the result of a continuous diagnostic test for a non-diseaded subject and Y the result of the test for a diseaded subject. We as-

Gianfranco Adimari, Duc-Khanh To
Department of Statistical Sciences, University of Padova, Via C. Battisti, 241; I-35121 Padova, Italy. e-mail: gianfranco.adimari@unipd.it; e-mail: duckhanh.to@unipd.it

Monica Chiogna
Department of Statistical Sciences "Paolo Fortunati", University of Bologna, Via Belle Arti, 41; 40126 Bologna, Italy. e-mail: monica.chiogna2@unibo.it

sume, without loss of generality, that high test values indicate high likelihood of disease. Then, for a given cutpoint τ, the sensitivity and the specificity of the test are $\theta(\tau) = \Pr(Y > \tau) = 1 - F_Y(\tau)$ and $\eta(\tau) = \Pr(X \leq \tau) = F_X(\tau)$, respectively, where $F_X(\cdot)$ denotes the cumulative distribution function of X and $F_Y(\cdot)$ the cumulative distribution function of Y. In practice, for the purpose of making diagnosis, that is, classifying a subject as either diseased or healthy, a diagnostic threshold is required. In order to select an "optimal" diagnostic cutpoint, there exists a variety of approaches. Among them, the one based on the Youden index [2], J, is certainly the most popular. The Youden index is the maximum of the sum of sensitivity and specificity minus one, i.e., $J = \max_\tau J(\tau)$, with $J(\tau) = \theta(\tau) + \eta(\tau) - 1$. The corresponding optimal cutpoint, τ^+ say, is the value that maximizes $J(\tau)$, i.e., the cutpoint that has the largest value in the associated sum of sensitivity and specificity. Clearly, sensitivity and specificity at the optimal cutpoint τ^+ are relevant measures for the diagnostic ability of the test.

Actually, the ROC curve of a new diagnostic test is unknown, being unknown (at least partially) the distributions of the test results, for both diseased and non-diseased populations. Hence, the statistical evaluation of the discriminatory ability of the test is obtained by making inference about its ROC curve and other quantities of interest, such as optimal thresholds and associated sensitivities and specificities. Generally, inference is based on data from some suitable sample of patients for which the disease status can be exactly assessed by means of a so-called gold standard test (GS). When an optimal threshold is estimated from data of both diseased and healthy samples, the corresponding estimated sensitivity and specificity are correlated, and joint inference is necessary to take into account such a correlation. Methods to built joint confidence regions for sensitivity and specificity at the optimal cutpoint based on the Youden index are proposed by Bantis et al.[3] and Yin and Tian[4]. Both works deal, in particular, with parametric (or semi-parametric) methods for the binormal model. When the hypothesis of normality is not adequate, in both articles it is suggested to resort, when possible, to a single Box-Cox transformation. The main limits of such methodologies are: (i) the related confidence regions have (or derive from regions that have) elliptical shape, because they are based on the asymptotic normality of appropriate estimators or pivots; (ii) when the normality of the biomarker does not meet, application of Box-Cox transformations is considered, but limited to a single transformation for both populations.

We propose new methods that provide approximate joint confidence regions for the sensitivity and specificity of a test, corresponding to the optimal cutpoint fixed by the Youden index criterion. They are based on empirical likelihood pivots, so give rise to likelihood-type regions that have no predetermined constraints on the shape and are automatically range respecting. The proposal covers binormal model and binormal model after the use of Box-Cox transformations. Importantly, in the second case, we show how to use two different transformations, for the healthy and the diseased subjects.

Background Let $x_1, \ldots, x_m$ be a random sample from X, i.e., the test results from m non-diseased patients, and $y_1, \ldots, y_n$ a random sample from Y, i.e., the test results from n diseased patients. The true disease status of each patient is assumed to

be ascertained by a GS test. Let $\hat{F}_X$ denote the empirical distribution function based on $x_1,\ldots,x_m$ and $\hat{F}_Y$ the empirical distribution function based on $y_1,\ldots,y_n$. Hence, for a fixed value t, $\hat{F}_X(t) = (1/m)\sum_{i=1}^{m} I(x_i \le t)$, where $I(\cdot)$ denotes the indicator function. Recall that we consider a continuous diagnostic test and that, for a fixed threshold τ, $\theta = \theta(\tau) = 1 - F_Y(\tau)$ and $\eta = \eta(\tau) = F_X(\tau)$. Adimari and Chiogna[5] define the empirical likelihood [6] statistic

$$\ell(\theta,\eta,\tau) = 2m\left\{\hat{F}_X(\tau)\log\frac{\hat{F}_X(\tau)}{\eta} + \left[1-\hat{F}_X(\tau)\right]\log\frac{1-\hat{F}_X(\tau)}{1-\eta}\right\} + 2n\left\{\hat{F}_Y(\tau)\log\frac{\hat{F}_Y(\tau)}{1-\theta} + \left[1-\hat{F}_Y(\tau)\right]\log\frac{1-\hat{F}_Y(\tau)}{\theta}\right\},$$

for $\theta \in (0,1)$, $\eta \in (0,1)$, $\tau \in \mathscr{T}$, where $\mathscr{T} = \left[\max\{x_{(1)},y_{(1)}\}, \min\{x_{(m)},y_{(n)}\}\right)$, and $x_{(i)}$, $i = 1,\ldots,m$, and $y_{(j)}$, $j = 1,\ldots,n$, denote the order statistics from the samples. When $\tau \notin \mathscr{T}$, then $\ell(\theta,\eta,\tau) = +\infty$. The Authors prove that, under very weak conditions, for each triplet τ_0, $\theta_0 = 1 - F_Y(\tau_0)$ and $\eta_0 = F_X(\tau_0)$ of true parameter values, when $\min\{m,n\} \to +\infty$ and $\lim m/n$ is finite and not zero, $\ell(\theta_0,\eta_0,\tau_0) \xrightarrow{d} \chi^2_2$, where χ^2_2 indicates the chi-squared distribution with 2 df. Then, such result can be used to construct non-parametric confidence regions for a pair of parameters, for example (θ_0,η_0), when the third remaining parameter is fixed: the set $\mathscr{R}_\alpha = \{(\theta,\eta) : \ell(\theta,\eta,\tau_0) \le c_\alpha\}$, where $\alpha \in (0,1)$ and c_α is such that $\Pr(\chi^2_2 > c_\alpha) = \alpha$, is a confidence region with nominal coverage probability $1-\alpha$ for the pair (sensitivity, specificity), at a fixed cutt-off τ_0. Regions obtained by $\ell(\theta,\eta,\tau_0)$ retain all good features of empirical likelihood confidence regions, whose shape and orientation is determined only by the data, and are range respecting.

2 The proposed method

Let τ^+ be the true optimal cutpoint based on the Youden index approach, and let θ^+ and η^+ be the corresponding values of sensitivity and specificity. If τ^+ was known, the pivot $\ell(\theta^+,\eta^+,\tau^+)$ could be used directly to build approximate confidence regions for (θ^+,η^+). In practice, τ^+ is unknown and must be estimated from the available data. Let $\hat{\tau}^+$ be a suitable estimator for τ^+. By resorting to the plug-in method, we can consider the quantity $\ell(\theta^+,\eta^+,\hat{\tau}^+)$, where an estimate replaces an unknown value. Unfortunately, such a replacement is not untroublesome. Indeed, the standard χ^2 approximation is no longer applicable in such a case. It is well known in the literature that when estimated entities enter in an empirical likelihood pivot, this has an asymptotic distribution which is a linear combination of independent χ^2_1 random variables (see, for example, Hjort et al.[9]), whose coefficients are (unknown) eigenvalues of a suitable 2×2 matrix. Since accurate estimation of such coefficients can be somewhat challenging, in the following we resort to the bootstrap calibration and apply it to the distribution of $\ell_*(\theta^+,\eta^+) = \ell(\theta^+,\eta^+,\hat{\tau}^+)$. This allows us to obtain estimates of the quantiles of the distribution of $\ell_*(\theta^+,\eta^+)$, which we will use to define the desired confidence regions.

Binormal model In some situations it is reasonable to assume that the diagnostic test has normal distribution, in both populations of healthy and diseased subjects. Let $X \sim N(\mu_x, \sigma_x^2)$ and $Y \sim N(\mu_y, \sigma_y^2)$. To formalize the assumption that the values of the test are positively associated with the disease status, we impose hereafter that $\mu_y > \mu_x$. For the binormal model, the optimal threshold provided by the criterion based on the Youden index can be obtained analytically and results to be

$$\tau^+ = \frac{\mu_x((\sigma_y/\sigma_x)^2-1) - (\mu_y-\mu_x) + (\sigma_y/\sigma_x)\sqrt{(\mu_y-\mu_x) + ((\sigma_y/\sigma_x)^2-1)\sigma_x^2 \log((\sigma_y/\sigma_x)^2)}}{(\sigma_y/\sigma_x)^2-1}.$$

When variances are equal, i.e. $\sigma_x^2 = \sigma_y^2$, then $\tau^+ = (\mu_x+\mu_y)/2$. Then, by the plug-in principle, a consistent estimator of τ^+ can be obtained by substituting in the above expressions the empirical counterparts $\bar{X} = (1/m)\sum_{i=1}^m X_i$, $\bar{Y} = (1/n)\sum_{i=1}^n Y_i$, $S_x = \sqrt{(1/(m-1))\sum_{i=1}^m (X_i-\bar{X})^2}$ and $S_y = \sqrt{(1/(n-1))\sum_{i=1}^n (Y_i-\bar{Y})^2}$ of μ_x, μ_y, σ_x and σ_y, respectively. Observe that the resulting $\hat{\tau}^+$ estimator is, essentially, the maximum likelihood estimator of τ^+.

Let $\Phi(\cdot)$ denote the cumulative distribution function of the standard normal. Given the observed data $x_1, \ldots, x_m$ and $y_1, \ldots, y_n$, we compute the estimate $\hat{\tau}^+$ and the corresponding optimal estimated sensitivity $\hat{\theta}^+ = 1 - \Phi((\hat{\tau}^+ - \bar{y})/S_y)$ and specificity $\hat{\eta}^+ = \Phi((\hat{\tau}^+ - \bar{x})/S_x)$. To conveniently calibrate $\ell_*(\theta^+, \eta^+)$ we resort to a simple bootstrap procedure:

1. use parametric bootstrap to get B bootstrap samples $\{x\}_b$ and $\{y\}_b$, for $b = 1, \ldots, B$, of sizes m and n, respectively;
2. add to each bootstrap sample the extremes (min and max) of the corresponding original sample, so as to obtain "enlarged" bootstrap samples of size $m+2$ and $n+2$, respectively;
3. compute $(\bar{x}_b, \bar{y}_b, S_{xb}, S_{yb})$, $\hat{\tau}_b^+$ and $\ell(\hat{\theta}^+, \hat{\eta}^+, \hat{\tau}_b^+)$ from the b-th pair of (enlarged) bootstrap samples, where $\hat{\theta}^+$ and $\hat{\eta}^+$ are the estimates obtained from the original data and the empirical distributions are taken from bootstrap samples;
4. get the estimate $\hat{c}_\alpha$ as the sample quantile of order $1-\alpha$ from the values $\ell(\hat{\theta}^+, \hat{\eta}^+, \hat{\tau}_b^+)$, $b = 1, \ldots, B$.

Then, the set $\mathscr{R}_\alpha = \{(\theta^+, \eta^+) : \ell_*(\theta^+, \eta^+) \leq \hat{c}_\alpha\}$, is a confidence region, with nominal coverage probability $1-\alpha$, for the optimal pair (sensitivity, specificity), corresponding to the Youden index criterion. In the above presented bootstrap procedure, at step 2, we "enlarged" bootstrap samples in order to reduce effects of the so-called convex hull problem, i.e., to reduce the probability that $\ell(\hat{\theta}^+, \hat{\eta}^+, \hat{\tau}_b^+)$ may be not finite. Moreover, we process only (enlarged) bootstrap samples whose averages respect the fixed order ($\mu_y > \mu_x$).

Normal models after Box-Cox transformations Suppose now that distributions of X and Y cannot reasonably be considered normal. Moreover, assume that $X > 0$ and $Y > 0$, with probability 1. In such a situation, Box-Cox transformations can help to achieve normality and are frequently used in ROC studies [3, 7]. However, to the best of our knowledge, papers available in the literature discuss, in this regard, a single transformation for the test results, both for healthy and diseased subjects. Although this approach has the advantage of leaving the ROC curve un-

changed (due to its invariance property with respect to increasing monotonic transformations), it can be inappropriate in cases where the two distributions of X and Y are very different from each other. In the following, we show how a more general approach involving two different transformations may be used.

Let

$$X^{(\lambda_1)} = \begin{cases} \frac{X^{\lambda_1}-1}{\lambda_1} & \lambda_1 \neq 0 \\ \log(X) & \lambda_1 = 0 \end{cases} \qquad Y^{(\lambda_2)} = \begin{cases} \frac{Y^{\lambda_2}-1}{\lambda_2} & \lambda_2 \neq 0 \\ \log(Y) & \lambda_2 = 0, \end{cases}$$

the transformed test results. The parameters defining the transformations are denoted by λ_1 and λ_2. Our proposal starts from the following observation. Since the optimal threshold τ^+ meets the relation

$$\begin{aligned} \tau^+ &= \arg\max_{\tau} [F_X(\tau) - F_Y(\tau)] = \arg\max_{\tau} [\Pr(X \leq \tau) - \Pr(Y \leq \tau)] \\ &= \arg\max_{\tau} \left[\Pr\left(X^{(\lambda_1)} \leq \tau^{(\lambda_1)}\right) - \Pr\left(Y^{(\lambda_2)} \leq \tau^{(\lambda_2)}\right) \right], \end{aligned}$$

where $\tau^{(\lambda_1)}$ and $\tau^{(\lambda_2)}$ are the transformed values of a same generic threshold τ, the double transformation leaves the optimal values of sensitivity and specificity unchanged. Moreover, $\theta^+ = 1 - \Pr\left(Y^{(\lambda_2)} \leq \tau^{+(\lambda_2)}\right)$ and $\eta^+ = \Pr\left(X^{(\lambda_1)} \leq \tau^{+(\lambda_1)}\right)$. Clearly, in the transformed scales, such optimal values match with two different threshold values.

$$\begin{aligned} \ell(\theta^+, \eta^+, \hat{\tau}^+) = {} & 2m \left\{ \hat{F}_X^{(1)}(\hat{\tau}_1^+) \log \frac{\hat{F}_X^{(1)}(\hat{\tau}_1^+)}{\eta^+} + \left[1 - \hat{F}_X^{(1)}(\hat{\tau}_1^+)\right] \log \frac{1 - \hat{F}_X^{(1)}(\hat{\tau}_1^+)}{1 - \eta^+} \right\} \\ & + 2n \left\{ \hat{F}_Y^{(2)}(\hat{\tau}_2^+) \log \frac{\hat{F}_Y^{(2)}(\hat{\tau}_2^+)}{1 - \theta^+} + \left[1 - \hat{F}_Y^{(2)}(\hat{\tau}_2^+)\right] \log \frac{1 - \hat{F}_Y^{(2)}(\hat{\tau}_2^+)}{\theta^+} \right\}, \end{aligned} \tag{1}$$

which is infinite when $\hat{\tau}_1^+$ is out of the range of the sample from $X^{(\hat{\lambda}_1)}$, or $\hat{\tau}_2^+$ is out of the range of the sample from $Y^{(\hat{\lambda}_2)}$. Observe that now the pivot in (1) depends on three estimated nuisance parameters: $\hat{\lambda}_1$, $\hat{\lambda}_2$ and $\hat{\tau}^+$.

In order to obtain confidence regions for the pair (θ^+, η^+), we propose the following algorithm. Given the observed data $x_1, \ldots, x_m$ and $y_1, \ldots, y_n$, we firstly compute the estimates $\hat{\lambda}_1$ and $\hat{\lambda}_2$. Then, using transformed data $x'_1, \ldots, x'_m$ and $y'_1, \ldots, y'_n$, say, we obtain sample means and standard deviations $\bar{x}'$, $\bar{y}'$, S'_x, S'_y, and maximize, with respect to τ, $\Phi\left(\frac{\tau^{(\hat{\lambda}_1)} - \bar{x}'}{S'_x}\right) - \Phi\left(\frac{\tau^{(\hat{\lambda}_2)} - \bar{y}'}{S'_y}\right)$ to get $\hat{\tau}^+$. Next, we can get the estimates $\hat{\theta}^+ = 1 - \Phi\left(\frac{\hat{\tau}_2^+ - \bar{y}'}{S'_y}\right)$ and $\hat{\eta}^+ = \Phi\left(\frac{\hat{\tau}_1^+ - \bar{x}'}{S'_x}\right)$. Finally, we use bootstrap calibration:

1. get B parametric bootstrap samples $\{x'\}_b$ and $\{y'\}_b$, for $b = 1, \ldots, B$, of sizes m and n, respectively;
2. add to each bootstrap sample the extremes (min and max) of the corresponding original (transformed) sample, so as to obtain "enlarged" bootstrap samples;
3. from the b-th pair of (enlarged) bootstrap samples, compute $(\bar{x}'_b, \bar{y}'_b, S'_{xb}, S'_{yb})$, $\hat{\tau}_b^+$, $\hat{\tau}_{1b}^+$, $\hat{\tau}_{2b}^+$ and $\ell(\hat{\theta}^+, \hat{\eta}^+, \hat{\tau}_b^+)$, using (1) where $\hat{\theta}^+$ and $\hat{\eta}^+$ are the estimate obtained from the original (transformed) data;

4. get the estimate $\hat{c}_\alpha$ as the sample quantile of order $1-\alpha$ from the values $\ell(\hat{\theta}^+, \hat{\eta}^+, \hat{\tau}_b^+)$, $b = 1, \ldots, B$.

The set $\mathscr{R}_\alpha = \{(\theta^+, \eta^+) : \ell_*(\theta^+, \eta^+) \leq \hat{c}_\alpha\}$, is the searched confidence region with nominal coverage $1-\alpha$. In the bootstrap procedure, we process only (enlarged) bootstrap samples whose averages respect the order between the averages of the original (transformed) samples. As an example, Table 1 gives some results of a large simulation study conducted to evaluate the finite sample behaviour of our confidence regions.

Table 1 Simulation results for Log Normal distributions: empirical coverages (over 10000 Monte Carlo replications) for some values of m, n, θ^+ and η^+. $B = 1000$ and $\alpha = 0.1, 0.05, 0.01$.

θ^+	η^+	m	n	0.90	0.95	0.99
0.696	0.861	50	50	0.869	0.929	0.983
		50	75	0.866	0.927	0.984
		75	50	0.862	0.932	0.983
		75	75	0.865	0.932	0.984
		100	100	0.866	0.927	0.983
0.790	0.895	50	50	0.874	0.936	0.986
		50	75	0.877	0.936	0.988
		75	50	0.870	0.935	0.986
		75	75	0.877	0.934	0.986
		100	100	0.867	0.935	0.987
0.888	0.940	50	50	0.903	0.944	0.974
		50	75	0.900	0.946	0.977
		75	50	0.887	0.948	0.991
		75	75	0.891	0.950	0.991
		100	100	0.886	0.947	0.989

References

1. Pepe MS. *The statistical evaluation of medical tests for classification and prediction*. Oxford University Press. 2003.
2. Youden WJ. Index for rating diagnostic tests. *Cancer* 1950; 3(1): 32–35.
3. Bantis LE, Nakas CT, Reiser B. Construction of confidence regions in the ROC space after the estimation of the optimal Youden index-based cut-off point. *Biometrics* 2014; 70(1): 212–223.
4. Yin J, Tian L. Joint inference about sensitivity and specificity at the optimal cut-off point associated with Youden index. *Comput Stat Data An* 2014; 77: 1–13.
5. Adimari G, Chiogna M. Simple nonparametric confidence regions for the evaluation of continuous-scale diagnostic tests. *Int J Biostat* 2010; 6(1).
6. Owen AB. *Empirical likelihood*. Chapman and Hall/CRC . 2001.
7. Fluss R, Faraggi D, Reiser B. Estimation of the Youden Index and its associated cutoff point. *Biometrical J* 2005; 47(4): 458–472.
8. Box GE, Cox DR. An analysis of transformations. *J R Stat Soc B* 1964; 26(2): 211–243.
9. Hjort NL, McKeague IW, Van Keilegom I. Extending the scope of empirical likelihood. *Ann Stat* 2009; 37(3): 1079–1111.