Working Paper Series, N. 17, November 2009

Accurate likelihood inference on the area under the ROC curve for small samples

Giuliana Cortese
Department of Statistical Sciences
University of Padua
Italy

Laura Ventura
Department of Statistical Sciences
University of Padua
Italy

Abstract: The accuracy of a diagnostic test with continuous-scale results is of high importance in clinical medicine. Receiver operating characteristics (ROC) curves, and in particular the area under the curve (AUC), are widely used to examine the effectiveness of diagnostic markers. Classical likelihood-based inference about the AUC has been widely studied under various parametric assumptions, but it is well-known that it can be inaccurate when the sample size is small, in particular in the presence of unknown parameters. The aim of this paper is to propose and discuss modern higher-order likelihood based procedures to obtain accurate point estimators and confidence intervals for the AUC. The accuracy of the proposed methodology is illustrated by simulation studies. Moreover, two real data examples are used to illustrate the application of the proposed methods.

Keywords: Area under the ROC curve, diagnostic markers, higher-order likelihood inference, small sample size, confidence intervals,reliability.

Contents

Department of Statistical Sciences
Via Cesare Battisti, 241
35121 Padova
Italy

tel: +39 049 8274168
fax: +39 049 8274170
http://www.stat.unipd.it

Corresponding author:
Giuliana Cortese
tel: +39 049 827 4124
gcortese@stat.unipd.it
http://www.stat.unipd.it/~gcortese

Accurate likelihood inference on the area under the ROC curve for small samples

Giuliana Cortese
Department of Statistical Sciences
University of Padua
Italy

Laura Ventura
Department of Statistical Sciences
University of Padua
Italy

Abstract: The accuracy of a diagnostic test with continuous-scale results is of high importance in clinical medicine. Receiver operating characteristics (ROC) curves, and in particular the area under the curve (AUC), are widely used to examine the effectiveness of diagnostic markers. Classical likelihood-based inference about the AUC has been widely studied under various parametric assumptions, but it is well-known that it can be inaccurate when the sample size is small, in particular in the presence of unknown parameters. The aim of this paper is to propose and discuss modern higher-order likelihood based procedures to obtain accurate point estimators and confidence intervals for the AUC. The accuracy of the proposed methodology is illustrated by simulation studies. Moreover, two real data examples are used to illustrate the application of the proposed methods.

Keywords: Area under the ROC curve, diagnostic markers, higher-order likelihood inference, small sample size, confidence intervals,reliability.

1 Introduction

This paper deals with modern likelihood theory in order to better distinguish between healthy and diseased populations.

Receiver operating characteristic (ROC) curves are one of the main tools for medical decision-making [1] and they are mostly used to assess the effectiveness of continuous diagnostic markers in distinguishing between diseased and non-diseased individuals. A ROC curve can be obtained from the response values of a diagnostic test based on a continuous diagnostic marker, and thus it provides a global measure of the accuracy of the test [2].

A diagnostic test based on a continuous diagnostic marker provides usually a response about the clinical status of subjects, identifying them as diseased (test positive) or non-diseased (test negative) patients. In order to provide such a positive or negative answer, the diagnostic test requires that a certain cut-off point is chosen.

The sensitivity and specificity associated with a given cut-off are defined as the probabilities of the test of correctly classifying subjects as diseased and non-diseased, respectively. Sensitivity and specificity vary when different choices of cut-off points are made over the continuous scale of the diagnostic characteristic. The ROC curve is obtained by plotting sensitivity versus 1-specificity for all possible values of the cut-off point.

ROC curves can be obtained under the assumption that the measurements of the diagnostic marker on the diseased and non-diseased subjects are distributed as two random variables X_1 and X_2, respectively. The area under the ROC curve (AUC) is the most popular summary measure of diagnostic accuracy of a continuous-scale test, or equivalently, of the diagnostic effectiveness of a continuous diagnostic marker. Its advantage consists of providing a single index that summarizes the overall performance of a diagnostic test or continuous marker, other than an entire curve. Values of the AUC close to 1 indicate very high diagnostic accuracy, while very low accuracy corresponds to values close to 0.5. Bamber [3] showed that the AUC is equal to

$$A = P(X_1 < X_2)\,, \tag{1}$$

which can be interpreted as the probability that, in a randomly selected pair of diseased and non-diseased subjects, the diagnostic test value is higher for the diseased patient. In more general contexts, the AUC is also used as a measure of difference between distributions [4].

Quantity A appears in many statistical problems regardless the connection to diagnostic tests and markers. The area A was initially studied for electronic signal detection [5], and later on it has been used in a broad range of applied contexts such as radiology, reliability and inspection systems, earthquake resistance. In reliability, 1 is called the stress-strength model and measures the reliability of a component in an engineering system, that is the probability that the strength (X_2) of a component exceeds a certain applied stress (X_1), and thus the component is working without a failure. In medicine, a further example of application of A is given by treatment comparisons, where (1) measures the treatment effectiveness by defining X_1 and X_2 as the responses for a control group and a treatment group, respectively.

ROC curves and the AUC have been studied under both parametric and nonparametric assumptions. There is a substantial literature on statistical inference for A under various parametric assumptions for X_1 and X_2; see [6] and [7] for a comprehensive treatment of stress-strength models and ROC approaches. Parametric inference has been broadly handled by likelihood based procedures [8, 9], theory of unbiased estimation or under a Bayesian perspective [6]. Furthermore, some contributions addressing inference about A have been also provided in semiparametric settings [10]. In the nonparametric setting, the literature ranges from the pioneering works of Mann and Whitney [11] to more recent papers such as Qin and Zhou [12].

Recently, a special attention has been devoted to interval estimation of A [13] and the papers by Qin and Hotilovac [14] and by Obuchowski and Lieber [15] provide an exhaustive comparison of nonparametric intervals for the AUC. Regardless of the parametric and nonparametric assumptions, confidence interval estimation for the AUC is usually based on the normal approximation to the distribution of the estimators. Nevertheless, the existing asymptotic methods for the AUC do not have good coverage accuracy in all situations, e.g. for all values of the AUC and for both small and large sample sizes. In the nonparametric context, the recent work by Newcombe [16] developed asymptotic methods that have a good performance irrespective of sample size and the order of magnitude of the AUC, and Zhou [17] proposed asymptotic expansions in order to improve the estimation accuracy and have good finite-sample coverage.

The current paper deals with a similar problem in the parametric context and it addresses the problem of inaccurate parametric inference in case of small sample size. Classical likelihood based procedures for inference on A are available, but it is well-known that they can be inaccurate when the sample size is small, in particular in presence of many unknown parameters [**?**]. To overcome this drawback, in this paper we discuss and apply higher-order likelihood based procedures (see, e.g., [18] and [19], and references therein) to obtain accurate point estimators and confidence intervals for A. In particular, we focus on the modified directed likelihood, also called modified signed log-likelihood ratio, which is a higher-order pivotal quantity that can be easily computed in practice when the parameter of interest is A. The accuracy of the proposed methodology is illustrated by numerical studies. Two applications to real-life data with small sample sizes, about abdominal aorta Aneurysm measurements and ALCL lymphoma, are illustrated in order to describe the practical use of the proposed methods.

The paper is organized as follows. Section 2 gives a short review on interval and point estimation based on first-order likelihood procedures, in particular the Wald and signed log-likelihood ratio statistics. The higher-order technique used to obtain accurate confidence intervals and point estimators in the context of the AUC model is discussed in Section 3. In Section 4 the proposed method is presented for two examples (exponential and normal models) and simulations studies that compare classical and higher-order likelihood-based procedures are illustrated. Section 5 discusses two applications to real-life data. Finally, some final remarks are pointed out in Section 6.

2 First-order likelihood inference

It is well-known that the likelihood function plays a central role in both statistical theory and practice. In this section, we provide a brief overview of some basic well-known approximations for likelihood inference, called first–order asymptotics, with application to the AUC given in (1), which represents the parameter of interest.

Let us consider a random sample $y = (y_1, \ldots, y_n)$ of size n drawn from a random variable Y whose probability density function $p(y; \theta)$ depends on an unknown parameter $\theta \in \Theta \subseteq \mathrm{I\!R}^d$, $d > 1$. Let $\ell(\theta) = \ell(\theta; y) = \sum_{i=1}^{n} \log p(y_i; \theta)$ denote the log-likelihood function for θ, $\hat{\theta}$ the maximum likelihood estimator (MLE), and $j(\theta) = -\ell_{\theta\theta}(\theta) = -\partial^2 \ell(\theta)/(\partial\theta\partial\theta^{\mathsf{T}})$ the observed information. Under broad conditions, $\hat{\theta}$ may be found by solving the score equation $\ell_\theta(\hat{\theta}) = 0$ and its asymptotic variance is approximated using the inverse of the observed information matrix $j(\theta)$. When we distinguish between quantities of primary interest and others not of direct concern, the d-dimensional parameter θ can be expressed as $\theta = (\psi, \lambda)$, where $\psi = \psi(\theta)$ is the scalar parameter of interest and λ a $(d-1)$-dimensional nuisance parameter. This partitioning entails corresponding splits of the score vector $\ell_\theta(\psi, \lambda)$ into $\ell_\psi(\psi, \lambda)$ and $\ell_\lambda(\psi, \lambda)$, and of the observed information $j(\psi, \lambda)$ into the sub-matrices $j_{\psi\psi}(\psi, \lambda)$, $j_{\psi\lambda}(\psi, \lambda)$, $j_{\lambda\psi}(\psi, \lambda)$ and $j_{\lambda\lambda}(\psi, \lambda)$. In this setting, it is well-known that, by the invariance property, the MLE of ψ is $\hat{\psi} = \psi(\hat{\theta})$.

General likelihood inference for ψ is typically based on profile procedures, which require to eliminate the nuisance parameter λ by replacing it by the constrained MLE $\hat{\lambda}_\psi$ obtained by maximizing $\ell(\psi, \lambda)$ with respect to λ for fixed ψ. Then inference about ψ may be performed using the profile log-likelihood $\ell_p(\psi) = \ell(\psi, \hat{\lambda}_\psi)$. The corresponding observed information, $j_p(\psi) = -\partial^2 \ell_p(\psi)/\partial\psi^2$, can be expressed in terms of the full observed information through the identity

$$j_p(\psi) = j_{\psi\psi}(\hat{\theta}_\psi) - j_{\psi\lambda}(\hat{\theta}_\psi)\; j_{\lambda\lambda}^{-1}(\hat{\theta}_\psi)\; j_{\lambda\psi}(\hat{\theta}_\psi) \;, \tag{2}$$

where $\hat{\theta}_\psi = (\psi, \hat{\lambda}_\psi)$.

To a first order of approximation, inference on the scalar parameter of interest ψ may be based on the Wald statistic

$$w_p = w_p(\psi) = j_p(\hat{\psi})^{1/2}(\hat{\psi} - \psi) \;, \tag{3}$$

or on the signed log-likelihood ratio statistic (or directed likelihood)

$$r_p = r_p(\psi) = \text{sign}(\hat{\psi} - \psi)\left(2(\ell_p(\hat{\psi}) - \ell_p(\psi))\right)^{1/2} \;, \tag{4}$$

which have standard normal distributions up to the order $O(n^{-1/2})$. Hence, a $100(1-\alpha)\%$ approximate confidence interval for ψ based on the Wald statistic is in practice computed as

$$(\hat{\psi} - z_{1-\alpha/2}\; j_p(\hat{\psi})^{-1/2},\; \hat{\psi} + z_{1-\alpha/2}\; j_p(\hat{\psi})^{-1/2}) \;,$$

where z_α is the $\alpha-$quantile of the standard normal distribution. Alternatively, a $100(1-\alpha)\%$ confidence interval for ψ based on r_p is $\{\psi : |r_p(\psi)| \leq z_{1-\alpha/2}\}$, and typically a numerical solution is required. In practice, the Wald statistic based interval is often preferred because of the simplicity in calculations. However, it is well-known that in general Wald procedures have poor behaviour even for large samples and are less accurate than those based on the directed likelihood [18].

All the former results about standard likelihood procedures can be easily applied when the focus of scientific inquiry ψ is the parameter A of the area under the ROC curve. Let X_1 and X_2 be independent random variables with cumulative distribution functions $F_{X_1}(x; \theta_1)$ and $F_{X_2}(x; \theta_2)$, respectively, with $\theta_1 \in \Theta_1 \subseteq \mathbb{R}^{d_1}$ and $\theta_2 \in \Theta_2 \subseteq \mathbb{R}^{d_2}$, $d = d_1 + d_2$. The equality $A = P(X_1 < X_2)$ relates the AUC to the probability that the marker measurement X_2 on a diseased subject is stochastically larger than the marker measurement X_1 on a non-diseased subject. Therefore, the area A can be evaluated as a function of the entire parameter $\theta = (\theta_1, \theta_2)$, through the relation

$$A = A(\theta) = \int F_{X_1}(t; \theta_1)\; dF_{X_2}(t; \theta_2) \;. \tag{5}$$

Theoretical expressions for A are available under several distributional assumptions both for X_1 and X_2 [6]. For parametric inference about A based on the random sample $x_1 = (x_{11}, \ldots, x_{1n_1})$ of size n_1 from X_1 and on the random sample $x_2 = (x_{21}, \ldots, x_{2n_2})$ of size n_2 from X_2, the most popular inferential procedures are those based on the profile likelihood function, due to their flexibility and generality. In particular, if $\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2)$ is the MLE of $\theta = (\theta_1, \theta_2)$, then the MLE of A is given by $\hat{A} = A(\hat{\theta})$ due to the invariance property. Thus, if the statistical model is reparameterized so that $\psi = A(\theta) = A(\theta_1, \theta_2)$ is the scalar parameter of interest and $\lambda = \lambda(\theta) = \lambda(\theta_1, \theta_2)$ the $(d-1)$-dimensional nuisance parameter, first–order confidence intervals for A may be based on (3) or (4). We refer the reader to Kotz *et al.* [6] for several examples on first–order inference on A studied under different assumptions on the distributions of X_1 and X_2.

When the sample size is relatively small, the first–order approximations are often inaccurate and can give poor results, especially if the dimension of the nuisance parameter λ is high with respect to n or, for the ROC model, when A is close to one, that is A is nearly on the boundary of the parameter space ([20]). In these situations, it may be useful to resort to modern likelihood theory.

3 Higher-order likelihood asymptotics

The theory of higher–order asymptotic analysis provides more precise inferences than the standard theory (see, e.g., [21], [22], [18] and [19]). There are two aspects of the improvement on classical likelihood-based inference about the scalar parameter of interest ψ based on the directed likelihood r_p. A first adjustment reduces the effects due to the estimation of nuisance parameters, and a second adjustment improves approximations when the sample size is small. In this section, we discuss a modified version of the directed likelihood (4) which is more accurate in cases of small sample size, having standard normal distribution up to $O(n^{-3/2})$, compared with $O(n^{-1/2})$ for standard asymptotics. One intriguing feature of the higher-order methods discussed here is that relatively simple and simple likelihood quantities play a central role.

Assume that a is an ancillary statistic, either exactly or at least to an approximate order of approximation, such that $\ell(\theta) = \ell(\theta; y) = \ell(\theta; \hat{\theta}, a)$. The modified directed likelihood for ψ (see, e.g.,[18], Chap. 7) is given by

$$r_p^* = r_p^*(\psi) = r_p + \frac{1}{r_p} \log \frac{q}{r_p} \ , \tag{6}$$

where

$$q = q(\psi) = \left| \ell_{;\hat{\psi}}(\hat{\theta}) - \ell_{;\hat{\psi}}(\hat{\theta}_\psi) - \ell_{\lambda;\hat{\psi}}(\hat{\theta}_\psi) \ell_{\lambda;\hat{\lambda}}(\hat{\theta}_\psi)^{-1} \left(\ell_{;\hat{\lambda}}(\hat{\theta}) - \ell_{;\hat{\lambda}}(\hat{\theta}_\psi) \right) \right| \frac{|\ell_{\lambda;\hat{\lambda}}(\hat{\theta}_\psi)|}{(|j(\hat{\theta})||j_{\lambda\lambda}(\hat{\theta}_\psi)|)^{1/2}} \, .$$

In the definition of q, the quantities appearing in the numerator are computed using the sample derivatives $\ell_{;\hat{\psi}} = \partial \ell(\theta)/\partial \hat{\psi}$, $\ell_{;\hat{\lambda}} = \partial \ell(\theta)/\partial \hat{\lambda}$, $\ell_{\lambda;\hat{\psi}} = \partial^2 \ell(\theta)/(\partial \lambda \partial \hat{\psi})$ and $\ell_{\lambda;\hat{\lambda}} = \partial^2 \ell(\theta)/(\partial \lambda \partial \hat{\lambda}^{\mathsf{T}})$. The modified directed likelihood r_p^* is a higher-order pivotal quantity with null standard normal distribution to order $O(n^{-3/2})$, conditionally on an appropriate ancillary a and hence also unconditionally at the same order. Moreover, it satisfies the requirement of parameterisation equivariance.

A confidence interval for ψ with approximate level $(1-\alpha)$ based on r_p^* is given by (ψ_1^*, ψ_2^*), with ψ_1^* and ψ_2^* solutions in ψ of the equations $r_p^*(\psi) = z_{1-\alpha/2}$ and $r_p^*(\psi) = z_{\alpha/2}$, respectively. Hence, a $100(1-\alpha)\%$ confidence interval for ψ based on r_p^* is $\{\psi : |r_p^*(\psi)| \leq z_{1-\alpha/2}\}$.

The modified directed likelihood r_p^* can also be used to derive a point estimator for ψ that improves the small sample properties of $\hat{\psi}$, respecting the requirement of parameterisation equivariance. As the MLE $\hat{\psi}$ can be seen as the solution of an estimating equation based on r_p, also the modified directed likelihood r_p^* can be used to define an estimating equation, following Pace and Salvan [23] and Giummolé and Ventura [24]. More precisely, the modified directed likelihood (6) gives rise to a simple estimating equation of the form

$$r_p^*(\psi) = 0 \ . \tag{7}$$

A numerical procedure is usually required in order to solve (7). The existence and uniqueness of the solution, denoted by $\hat{\psi}^*$, is asymptotically guaranteed, at least in a neighborhood of $\hat{\psi}$. The estimator $\hat{\psi}^*$ is a refinement of $\hat{\psi}$, with the estimating equation (7) giving implicitly a higher-order correction to the MLE. In view of the properties of r_p^*, the estimating equation (7) is mean unbiased as well as median unbiased at the third-order of accuracy. The median unbiasedness property also holds for the corresponding estimator $\hat{\psi}^*$, under the condition that the estimating equation is a monotone function of the parameter of interest. Moreover, since r_p^* is invariant under interest respecting reparameterisations, $\hat{\psi}^*$ is an equivariant estimator of ψ. Several numerical investigations [23, 24] show that the estimators based on r_p^* improve on the MLE.

The proposed higher–order procedures for inference about the AUC parameter A can be summarized into the following steps:

1. calculation of the AUC $A = A(\theta)$ as a function of θ, with $\theta = (\theta_1, \theta_2)$;
2. calculation of the likelihood $\ell(\psi, \lambda)$ with $\psi = A$ and λ nuisance parameter;
3. computation of $r_p^*(\psi)$ for a range of values around the MLE;
4. interpolation of the points $r_p^*(\psi)$ by a smoothing method;
5. invert the interpolating function and find the corresponding values in $z_{1-\alpha/2}$ and $z_{\alpha/2}$ to obtain a $100(1-\alpha)\%$ confidence interval (ψ_1^*, ψ_2^*) as solution to the equations $r_p^*(\psi) = z_{1-\alpha/2}$ and $r_p^*(\psi) = z_{\alpha/2}$, respectively, or
6. invert the interpolating function and find the corresponding values in 0 to obtain a point estimate $\hat{\psi}^*$ as solution to $r_p^*(\psi) = 0$.

It is important to note that the proposed procedures can be easily implemented in practice for many commonly used statistical models using modern statistical environments, such as `R` (`http://www.r-project.org/`). An illustration of how steps 4–6 can be implemented with `R` is given in Appendix B.

4 Examples and simulation studies

In this section the construction of confidence intervals and point estimators for A based on the modified directed likelihood r_p^* is illustrated for two examples. The first example is about the simple situation where both the measurements of the marker on non-diseased and diseased patients, X_1 and X_2, are exponentially distributed, whereas in the second example they are supposed to be independent gaussian variables. Both these statistical models are members of the exponential family and in this case the r_p^* statistic is simple to compute since $\ell(\theta; x_1, x_2) = \ell(\theta; \hat{\theta})$. This means that the MLE $\hat{\theta}$ is the sufficient statistic based on the sample and the likelihood can be written as a function of $\hat{\theta}$ only.

For both the examples numerical studies are considered to investigate the performance of r_p^* for the construction of both confidence intervals and point estimators for A.

The current section is supplemented by Appendix A, where the main formulas are reported, and Appendix B, where a package of functions written with the `R` software is illustrated. The `R` package is available online at the webpage `http://homes.stat.unipd.it/gcortese` and can be used for the analyses of the AUC and ROC curves under a parametric setting. Continuous diagnostic markers can be assumed to be exponentially or normally distributed with either equal or different variances. Note that in the following, for the sake of simplicity, the second example concerns the case where Gaussian models have equal variances. Analyses of more general examples with unequal variances can also be performed by using the `R` functions given in Appendix B.

4.1 Exponential distribution

Assume that X_1 and X_2 are independent and distributed as exponential random variables with parameters α and β, respectively, i.e. $X_1 \sim Exp(\alpha)$ and $X_2 \sim Exp(\beta)$. Let $x_1 = (x_{11}, \dots, x_{1n_1})$ be a random sample of size n_1 from X_1 and $x_2 = (x_{21}, \dots, x_{2n_2})$ a random sample of size n_2 from X_2. Moreover, it is assumed that the ratio of the sample sizes n_1/n_2 converges to some finite positive constant as n_1 and n_2 diverge. Under these assumptions, the probability A representing the AUC can be written as

$$A = A(\theta) = \frac{E(X_1)}{E(X_1) + E(X_2)} = \frac{\alpha}{\alpha + \beta}, \tag{8}$$

with $\theta = (\alpha, \beta)$.

For first–order and higher–order likelihood inference on A, it is convenient to reparameterize the log–likelihood function $\ell(\theta)$ so that $\theta = (\psi, \lambda)$, where $\psi = \alpha/(\alpha + \beta) = A$ is the scalar parameter of interest and $\lambda = \alpha + \beta$ is the scalar nuisance parameter. In this situation, standard likelihood based inference procedures for the parameter A are easy to perform [6]. For example, for the invariance property, the MLEs for ψ and λ are $\hat{\psi} = \hat{\alpha}/(\hat{\beta} + \hat{\alpha})$ and $\hat{\lambda} = \hat{\alpha} + \hat{\beta}$, respectively, with the MLEs of α and β given by $\hat{\alpha} = n_1/\sum x_{1i}$ and $\hat{\beta} = n_2/\sum x_{2i}$, respectively.

First–order inference about the parameter of interest ψ may be based on the Wald statistic

$$w_p = (\hat{\psi} - \psi)\sqrt{n_1 n_2 / \left(n\hat{\psi}^2 (1-\hat{\psi})^2 \right)} \ ,$$

or on the directed likelihood r_p given in (4) with

$$\ell_p(\psi) = n \log \hat{\lambda}_\psi + n_1 \log \psi + n_2 \log(1-\psi) \ , \qquad (9)$$

where $n = n_1 + n_2$ and $\hat{\lambda}_\psi = n\hat{\lambda} / \left(\frac{n_1 \psi}{\hat{\psi}} + \frac{n_2(1-\psi)}{(1-\hat{\psi})} \right)$.

For higher–order inference, the modified directed likelihood r_p^* can be computed. Under the exponential model, computation of r_p^* can follow the formula given in (6) which requires computation of the adjustment term q. Straightforward calculations lead to

$$q = \left(n_1(1-\hat{\psi}) - n_2\hat{\psi} + n\frac{n_2\hat{\psi}^2(1-\psi) - n_1\psi(1-\hat{\psi})^2}{n_1\psi(1-\hat{\psi}) + n_2\hat{\psi}(1-\psi)} \right) \sqrt{\frac{n}{n_1 n_2}} \ . \qquad (10)$$

The statistical accuracy of the modified directed likelihood r_p^* under the exponential model is illustrated through a simulation study, based on 5000 Monte Carlo trials. The performance of r_p^* is compared with the classical procedures, i.e. the directed likelihood r_p and the Wald statistic w_p. The numerical study was carried out by fixing the parameter α and determining β so that $A = \psi = \alpha/(\alpha+\beta) = 0.5$, for different combinations of sample sizes (n_1, n_2). The simulation study was repeated for $\psi = 0.8$ and $\psi = 0.95$. Table 1 reports empirical coverages for the equitailed confidence intervals for A with nominal levels 90% and 95%. These intervals were obtained, for w_p and r_p, on the basis of the normal approximation to their distribution as mentioned in Section 2, and for r_p^*, by using the step procedure described in Section 3.

For the example of independently exponentially distributed X_1 and X_2, results in Table 1 show that confidence intervals based on r_p^* and r_p have a considerable improvement in the two-sided coverage accuracy as compared to confidence intervals based on the Wald statistic w_p. Moreover, in all cases there is evidence of a strong asymmetry in the confidence intervals for ψ based on the Wald statistic, due to different non-coverage probabilities for the left and right tails, in contrast to the equitailed results based on r_p^*. Furthermore, the results in Table 1 tell that, in this example, confidence intervals derived from r_p and r_p^* have mean coverage very close to the nominal value, but r_p^* is more accurate than r_p in particular when the sample sizes n_1 and n_2 are small and in the coverage probabilities for the left and right tails.

4.2 Gaussian distribution

Let us assume that X_1 and X_2 are independent Gaussian random variables, with equal variances, that is $X_1 \sim N(\mu_1, \sigma^2)$ and $X_2 \sim N(\mu_2, \sigma^2)$. This is a typical setting commonly used in the literature on ROC curves, two samples comparisons and stress-strength models.

		$\psi = 0.5$ $(\alpha = \beta = 1)$		$\psi = 0.8$ $(\alpha = 1,\ \beta = 0.25)$		$\psi = 0.95$ $(\alpha = 1,\ \beta = 0.05)$	
(n_1, n_2)	statistic	90%	95%	90%	95%	90%	95%
(3,3)	w_p	0.792	0.842	0.807	0.857	0.839	0.865
		(0.099,0.109)	(0.077,0.081)	(0.150,0.043)	(0.116,0.026)	(0.156,0.005)	(0.133,0.003)
	r_p	0.885	0.937	0.886	0.936	0.881	0.939
		(0.063,0.053)	(0.031,0.032)	(0.056,0.057)	(0.031,0.033)	(0.056,0.062)	(0.030,0.031)
	r_p^*	0.901	0.951	0.901	0.947	0.897	0.949
		(0.054,0.045)	(0.022,0.027)	(0.050,0.049)	(0.025,0.028)	(0.049,0.054)	(0.025,0.026)
(5,5)	w_p	0.829	0.874	0.848	0.893	0.864	0.896
		(0.082,0.089)	(0.064,0.062)	(0.116,0.035)	(0.089,0.018)	(0.132,0.003)	(0.104,0.001)
	r_p	0.890	0.937	0.890	0.945	0.890	0.942
		(0.056,0.054)	(0.031,0.032)	(0.058,0.052)	(0.030,0.025)	(0.056,0.053)	(0.030,0.028)
	r_p^*	0.898	0.944	0.899	0.952	0.900	0.948
		(0.052,0.050)	(0.027,0.029)	(0.054,0.047)	(0.027,0.021)	(0.051,0.049)	(0.026,0.025)
(10,10)	w_p	0.864	0.917	0.863	0.916	0.879	0.919
		(0.064,0.071)	(0.042,0.040)	(0.105,0.031)	(0.070,0.013)	(0.113,0.008)	(0.081,0.001)
	r_p	0.894	0.946	0.893	0.945	0.893	0.947
		(0.056,0.050)	(0.026,0.027)	(0.053,0.054)	(0.026,0.029)	(0.052,0.054)	(0.026,0.027)
	r^*	0.899	0.950	0.898	0.949	0.901	0.951
		(0.053,0.048)	(0.025,0.025)	(0.051,0.051)	(0.024,0.027)	(0.050,0.049)	(0.024,0.025)
(30,30)	w_p	0.882	0.933	0.893	0.935	0.898	0.938
		(0.059,0.059)	(0.033,0.034)	(0.068,0.039)	(0.049,0.016)	(0.083,0.020)	(0.056,0.006)
	r_p	0.891	0.947	0.892	0.946	0.896	0.948
		(0.054,0.055)	(0.026,0.027)	(0.063,0.045)	(0.029,0.024)	(0.051,0.052)	(0.028,0.024)
	r_p^*	0.892	0.949	0.894	0.949	0.898	0.950
		(0.053,0.055)	(0.025,0.027)	(0.062,0.044)	(0.029,0.023)	(0.050,0.051)	(0.027,0.024)

Table 1: Two-sided empirical coverage of equitailed confidence intervals with 90% and 95% nominal levels for A, under the exponential assumption. The values in brackets are the non-coverage probabilities on the left and right tail, expressing the lower and upper errors, respectively.

In this situation, the entire parameter θ is given by $\theta = (\mu_1, \mu_2, \sigma^2)$ and the AUC can be written as [6]

$$A = A(\theta) = \Phi\left(\frac{\mu_2 - \mu_1}{\sqrt{2\sigma^2}}\right) , \tag{11}$$

where $\Phi(\cdot)$ is the standard normal cumulative distribution function. Let $x_1 = (x_{11}, \ldots, x_{1n_1})$ be a random sample of size n_1 from X_1 and $x_2 = (x_{21}, \ldots, x_{2n_2})$ a random sample of size n_2 from X_2. By the invariance property, the MLE of A is

$$\hat{A} = A(\hat{\theta}) = \Phi\left(\frac{\hat{\mu}_2 - \hat{\mu}_1}{\sqrt{2\hat{\sigma}^2}}\right) , \tag{12}$$

where $\hat{\mu}_1 = \sum x_{1i}/n_1$, $\hat{\mu}_2 = \sum x_{2i}/n_2$ and $\hat{\sigma}^2 = (1/n)(\sum(x_{1i} - \hat{\mu}_1)^2 + \sum(x_{2i} - \hat{\mu}_2)^2)$ are the MLEs of μ_1, μ_2 and σ^2, respectively.

Standard first–order inference on A can be based on the standard normal approximation to w_p and r_p, with log–likelihood function $\ell(\theta)$ reparametrized to $\theta = (\psi, \lambda)$.

The parameter of interest is $\psi = A$ given in (11) and the nuisance parameter is set to be $\lambda = (\lambda_1, \lambda_2) = (\mu_1/\sqrt{2\sigma^2}, \sqrt{2\sigma^2})$. Note that other choices for λ are possible and they would lead to the same results. Computation of the Wald statistic in (3) requires that the profile observed information is obtained from the identity in (2) (an explicit expression is given in Appendix A). The directed likelihood r_p is given by (4) with

$$\ell_p(\psi) = -n\left(\log\tilde{\lambda}_2 + \frac{\hat{\lambda}_2^2}{2\tilde{\lambda}_2^2}\right) - \frac{1}{\tilde{\lambda}_2^2}\left[n_2\left(D(\hat{\theta}_\psi) - D(\hat{\theta})\right)^2 + n_1\left(\tilde{\lambda}_1\tilde{\lambda}_2 - \hat{\lambda}_1\hat{\lambda}_2\right)^2\right]\ ,(13)$$

where $D(\theta) = (\Phi^{-1}(\psi)\lambda_2 + \lambda_1\lambda_2)$ and $\hat{\lambda}_\psi = (\tilde{\lambda}_1, \tilde{\lambda}_2)$ is obtained by numerical procedures.

For higher–order inference, computation of the modified directed likelihood r_p^* is straightforward, although more laborious expressions are obtained for computing the correction term q (see Appendix A). For this reason, in order to help the reader in applying the r_p^* for the AUC, direct implementation of the formulas by using the R software is provided in Appendix B.

As in the previous example (see Subsection 4.1), the accuracy of r_p^* is illustrated through a simulation study, based on 5000 Monte Carlo trials. Table 2 reports empirical coverages for the equitailed confidence intervals for A for different combinations of sample sizes (n_1, n_2) and different values of ψ. Results in Table 2 show that r_p^* is more accurate than r_p when the sample sizes n_1 and n_2 are small, in terms of both central coverage probability and the symmetry of error rates. It is also to note that confidence intervals based on r_p^* and r_p have a considerable improvement in the two-sided coverage accuracy as compared to confidence intervals based on the Wald statistic w_p.

We used also simulation studies to evaluate the properties of the r_p^*-based estimator of A, $\hat{\psi}^*$, in comparison with the MLE $\hat{\psi}$. The two estimators are compared in terms of median bias and results are shown in Table 3. Estimated standard errors of median bias are given in parentheses. It can be noted that the estimator $\hat{\psi}^*$ is preferable to the MLE in terms of the considered criteria, since it is less median-biased than the MLE, in particular for high values of ψ and small sample sizes. In particular, from the table we observe that the estimator $\hat{\psi}^*$ performs better for values of ψ equal to or higher than 0.8, especially when the sample sizes are lower or equal to 20. The choice of the median-bias as a comparison criteria is due to the fact that the median unbiasedness property holds for the r_p^*-based estimator, which is more robust under model misspecifications. For the previous example about exponential model assumptions, simulation studies for point estimates were not reported in the paper, since conclusions were very similar to those given for the Gaussian model, and $\hat{\psi}^*$ and $\hat{\psi}$ differ only slightly in the median biases.

5 Data examples

Two real-life data examples are illustrated in the following. Both the datasets consist of samples with small sizes. For the first example an exponential model is assumed,

		$\psi = 0.5$		$\psi = 0.8$		$\psi = 0.95$	
		$(\mu_1 = \mu_2 = 5, \sigma = 1)$		$(\mu_1 = 5, \mu_2 = 6.55, \sigma = 1.3)$		$(\mu_1 = 5, \mu_2 = 8.5, \sigma = 1.5)$	
(n_1, n_2)	statistic	90%	95%	90%	95%	90%	95%
(5,5)	w_p	0.781	0.833	0.735	0.790	0.663	0.685
		(0.101,0.118)	(0.082,0.085)	(0.030,0.235)	(0.016,0.194)	(0.003,0.334)	(0.001,0.314)
	r_p	0.843	0.911	0.841	0.911	0.835	0.899
		(0.071,0.085)	(0.045,0.044)	(0.041,0.118)	(0.025,0.064)	(0.031,0.134)	(0.018,0.083)
	r_p^*	0.898	0.945	0.898	0.950	0.897	0.950
		(0.047,0.056)	(0.028,0.027)	(0.043,0.059)	(0.024,0.027)	(0.050,0.053)	(0.025,0.026)
(10,10)	w_p	0.845	0.895	0.819	0.862	0.764	0.802
		(0.076,0.079)	(0.056,0.049)	(0.026,0.155)	(0.013,0.125)	(0.004,0.232)	(0.001,0.197)
	r_p	0.877	0.933	0.880	0.930	0.870	0.932
		(0.061,0.061)	(0.035,0.032)	(0.043,0.076)	(0.024,0.047)	(0.040,0.090)	(0.015,0.053)
	r_p^*	0.900	0.951	0.900	0.944	0.888	0.951
		(0.051,0.048)	(0.027,0.023)	(0.053,0.047)	(0.027,0.029)	(0.062,0.050)	(0.024,0.026)
(20,20)	w_p	0.863	0.922	0.859	0.909	0.812	0.922
		(0.065,0.071)	(0.040,0.038)	(0.028,0.113)	(0.011,0.080)	(0.009,0.180)	(0.040,0.038)
	r_p	0.882	0.940	0.890	0.945	0.886	0.940
		(0.056,0.062)	(0.030,0.030)	(0.042,0.068)	(0.020,0.035)	(0.034,0.080)	(0.030,0.030)
	r_p^*	0.895	0.950	0.903	0.951	0.901	0.948
		(0.050,0.056)	(0.027,0.025)	(0.048,0.048)	(0.024,0.025)	(0.049,0.050)	(0.027,0.025)
(30,30)	w_p	0.886	0.930	0.867	0.926	0.844	0.892
		(0.063,0.051)	(0.035,0.035)	(0.025,0.108)	(0.013,0.061)	(0.011,0.145)	(0.003,0.105)
	r_p	0.893	0.941	0.892	0.944	0.886	0.945
		(0.059,0.047)	(0.029,0.030)	(0.038,0.070)	(0.022,0.034)	(0.042,0.071)	(0.020,0.034)
	r_p^*	0.900	0.947	0.899	0.951	0.89	0.950
		(0.055,0.044)	(0.025,0.027)	(0.044,0.057)	(0.024,0.024)	(0.055,0.055)	(0.025,0.026)

Table 2: Two-sided empirical coverage of equitailed confidence intervals with 90% and 95% nominal levels for A, under the normal assumption. The values in brackets are the non-coverage probabilities on the left and right tail, expressing the lower and upper errors, respectively.

(n_1, n_2)	estimator	$\psi = 0.5$		$\psi = 0.8$		$\psi = 0.95$	
		BI		BI		BI	
(5,5)	$\hat{\psi}$	0.0003	(0.194)	0.035	(0.138)	0.022	(0.042)
	$\hat{\psi}^*$	-0.0003	(0.186)	0.003	(0.134)	0.001	(0.046)
(10,10)	$\hat{\psi}$	-0.005	(0.134)	0.015	(0.100)	0.012	(0.042)
	$\hat{\psi}^*$	-0.005	(0.131)	-0.001	(0.098)	0.002	(0.046)
(20,20)	$\hat{\psi}$	0.007	(0.096)	0.013	(0.066)	0.005	(0.029)
	$\hat{\psi}^*$	0.006	(0.094)	0.005	(0.065)	0.0001	(0.031)
(30,30)	$\hat{\psi}$	0.005	(0.072)	0.006	(0.055)	0.003	(0.024)
	$\hat{\psi}^*$	0.005	(0.071)	0.001	(0.055)	0.0001	(0.025)

Table 3: Empirical median biases (BI) and estimated standard errors (in parentheses) of the r_p^*-based estimator, $\hat{\psi}^*$, and the MLE $\hat{\psi}$ for the AUC parameter A, under the Gaussian model.

while the second example is studied under the assumption of normally distributed variables.

5.1 ALCL lymphoma

The dataset about ALCL lymphoma is part of a retrospective study on the anaplastic large cell lymphoma carried out by the Clinic of Pediatric Hematology Oncology, University of Padova, Italy. The anaplastic large cell lymphoma is a rare cancer disease which affects both children and adults. The aim of the study was to assess the role of the Hsp70 protein in association with the ALCL lymphoma. Diseased patients seem to have higher Hsp70 levels than healthy subjects. It is known that Hsp70 can induce the development of pathological states such as oncogenesis ([25]). Moreover, excessive Hsp70 protein levels in diseased patients seem to limit the efficacy of the chemotherapy treatment. Thus, Hsp70 protein levels can be studied as a biomarker for detecting early ALCL lymphoma and therefore, its effectiveness in diagnosing the disease was evaluated by the AUC approach. The interest was also to interpret the AUC as the probability that the Hsp70 protein level is higher in ALCL cancer patients than in healthy individuals.

The data consist of a small sample: 10 patients with ALCL lymphoma in the group of 'cases' and 4 healthy subjects in the group of 'controls'. Hsp70 protein level was recorded on a continuous scale for each individual. Two independent exponential random variables, $X_1 \sim \exp(\alpha)$ and $X_2 \sim \exp(\alpha)$, were assumed for the protein level in cancer patients and in non-diseased subjects, respectively. Results from a Kolmogorov-Smirnov nonparametric test supported the choice of an exponential model assumption for these data ($p = 0.865$ and $p = 0.846$), although this conclusion may be instable due to the considered small sample sizes.

The two protein level samples result to have both different means (equal to 0.23 and 1.44 in the controls and cases, respectively) and variances, as observed in Figure 1 (a). Therefore, under the exponential model, the MLE for the exponential parameters, $\hat{\alpha} = 4.25$ and $\hat{\beta} = 0.70$, are substantially different in the two samples, suggesting thus a high value of the AUC. Confidence intervals (CI) for the AUC based on the Wald, r_p and r_p^* statistics are reported in Table 4 together with the MLE and the r_p^*-based point estimate for the AUC. These values have been obtained by applying the theory described in Subsection 4.1, and thus by inverting the interpolating functions r_p and r_p^* shown in Figure 1 (b). Horizontal and vertical lines in the plot identify the interpolation points on the curves and the corresponding ψ values for the CIs and point estimates, as explained in the step procedure in Section 3. Table 4 reports that the estimated probability that a cancer patient has higher Hsp70 protein level than a healthy patient is about 0.85. This value may also suggest a sufficiently high effectiveness of the protein level in early detecting ALCL patients. Results about CIs in Table 4 do not differ substantially, as point estimates neither. Nevertheless, it is possible to note that the upper bound of the Wald CI (0.89) is lower than the r_p- and r_p^*-based CIs (0.95). Moreover r_p^*-based CI seems to be more protective in estimating the accuracy of the protein level biomarker, since its lower bound (0.60) is further below the lower bound of the r_p-based CI (0.62) (see Table 4).

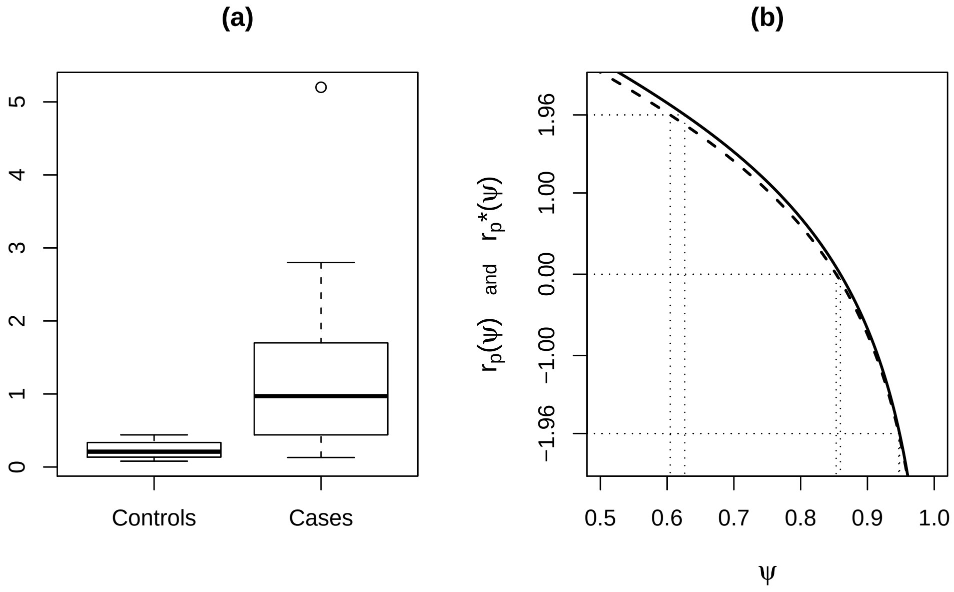

Figure 1: Panel (a): Boxplot of Hsp70 protein levels in cases and controls subjects. Panel (b): Plot of r_p (thick solid line) and r_p^* (thick dashed line) statistics for a range of values of the parameter ψ. The upper and lower horizontal lines are drawn to show the difference in confidence intervals based on the two statistics, while the central horizontal line identifies point estimates of the AUC.

	ALCL lymphoma	**Abdominal aortic aneurysm**
Confidence intervals		
Wald	(0.719, 0.999)	(0.897, 0.994)
r_p	(0.626, 0.948)	(0.794, 0.992)
r_p^*	(0.605, 0.947)	(0.764, 0.989)
Point estimates		
MLE	0.859	0.950
r_p^*	0.853	0.933

Table 4: Point estimates and 95% confidence intervals for the AUC in the real data examples on ALCL lymphoma and abdominal aortic aneurysm.

5.2 Abdominal aortic aneurysm

The abdominal aortic aneurysm is a localized blood-filled dilation of the abdominal aorta. Accurate measurements of the diameter of the aneurysm are essential for screening and in assessing the seriousness of the disease. Surgical intervention is planned when the aneurysm diameter exceeds a certain threshold, often fixed at 5 cm, since it is known that the risk of aneurysm rupture increases as the size becomes larger, causing death. For decision making about interventions, it is thus important that the available measurements instruments are very accurate and provide the actual diameter values.

The aneurysm study considered two groups of patients who have been classified with low (L) and high (H) rupture risk, that is with small and large aneurysm diameter, by using a gold standard measurement instrument (computed tomography). The dataset consists of measurements of the diameter aneurysm on the two groups of patients obtained by a newer instrument based on ultrasounds (US). The aim of the study was to evaluate the diagnostic accuracy of this latter instrument in discriminating between patients with low and high rupture risk.

Two samples of US measurements with small sizes $n_1 = n_2 = 10$ were obtained from the L and H groups. It was assumed that US measurements were distributed in the two groups as normal variables with different means and equal variances. This last hypothesis was supported by the boxplots in Figure 2 (a) showing a similar variability for the two samples, and verified by the F-test ($p = 0.641$).

In this example, confidence intervals (CI) and point estimates for the AUC based on the Wald, r_p and r_p^* statistics were found by applying the theory in Subsection 4.2, and computed from the step procedure described in Section 3. Estimates are represented graphically in Figure 2 (b), where the interpolating functions r_p and r_p^* are inverted analogously to the previous example about ALCL data.

The MLEs of the parameters of the Gaussian distributions were $\hat{\mu}_1 = 4$, $\hat{\mu}_2 = 6$ and $\hat{\sigma}^2 = 0.78$. The MLE of A ($\hat{\psi} = 0.95$) was higher than the r_p^*-based estimate ($\hat{\psi}^* = 0.93$). The Wald CI was also found to differ from the r_p- and r_p^*-based CIs substantially, especially in the lower bounds. The r_p- and r_p^*-based CIs were found to be similar, although the r_p^*-based CI is slightly shifted to the left. In summary, in this example the use of the r_p^* statistic seems to yield more protective results about the accuracy of US measurements with respects to the other classical procedures.

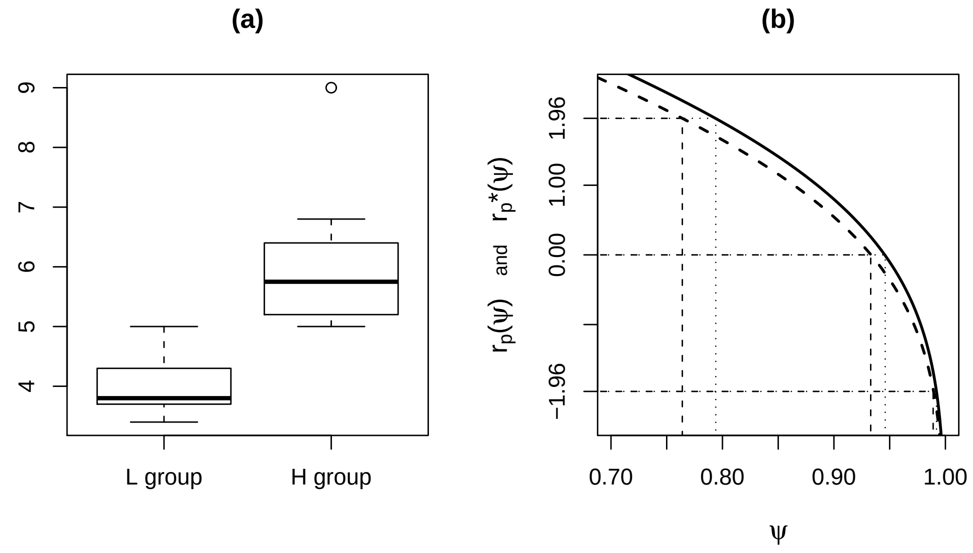

Figure 2: Panel (a): Boxplot of the sample distributions of L and H groups. Panel (b): Plot of r_p (thick solid line) and r_p^* (thick dashed line) for a range of values of the parameter ψ. Horizontal lines are drawn to identify confidence intervals and point estimates of the AUC based on the two statistics.

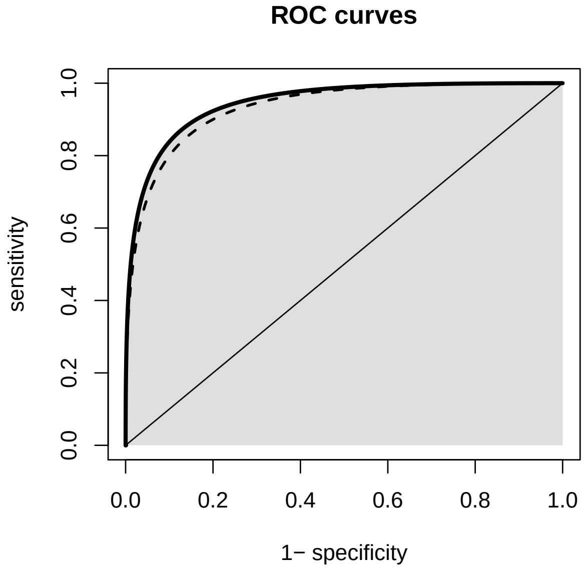

Figure 3: Area under the estimated ROC curves corresponding to the MLEs (solid line) and to the r_p^*-based estimates (dashed line).

Given the MLEs for the parameters of the two Gaussian distribution of X_1 and X_2, it is possible to draw the corresponding estimated ROC curve. This is possible since specificity and sensitivity are evaluated as $P(X_1 < t)$ and $P(X_2 > t)$, respectively, and thus they can be estimated as $F_{X_1}(t; \hat{\mu}_1, \hat{\sigma}^2)$ and $(1 - F_{X_2}(t; \hat{\mu}_2, \hat{\sigma}^2))$, respectively, for each cut-off point t. The resulting curve, represented in Figure 3 with a solid line, suggests values of sensitivity nearly equal to 1 in correspondence of desired high values of specificity, although presence of high variability is to be accounted. In our example, this fact means that the US instrument is highly accurate and has very low error rates.

The ROC curve can also be estimated by means of the r_p^* statistic. Estimates of the original parameters (μ_1, μ_2, σ^2) were obtained from the constrained estimate $\hat{\lambda}_{\hat{\psi}^*}$, given in (13), computed for $\psi = \hat{\psi}^*$. The resulting estimates $\hat{\theta}^* = (\hat{\mu}_1^*, \hat{\mu}_2^*, (\hat{\sigma}^2)^*)$ were then used to evaluate sensitivity and specificity via the cumulative distribution functions as done before with the MLEs. The resulting ROC curve is shown in Figure 3 with a dashed line. By comparing the r_p^*-based ROC curve with the ROC curve estimated from the MLEs, a non-negligible discrepancy is noted. From the upper curve based on MLEs a slight overestimation of sensitivity values is observed for fixed specificities, when compared with the more accurate ROC estimate based on r_p^*.

6 Discussion

The proposal discussed in this paper is centered on general distributional assumptions on both X_1 and X_2. Two examples have been presented in the context of exponential and gaussian model assumptions, and we pointed out, in case of small sample sizes, the improved accuracy of confidence intervals for the AUC when they are based on the modified directed likelihood r_p^*. The simple statistic r_p, as compared with the classical Wald statistic, yields also to more accurate inferential results both for small and large sample sizes. Our conclusions are in agreement with the simulation results given by Jiang and Wong [26].

The method we propose in this paper can be extended to more complex models. In particular, expression (6) requires determination of the sample space derivatives $\ell_{;\hat{\theta}}$, which may be difficult since it is necessary to write explicitly the ancillary statistic a of the model. [GIVE SOME EXAMPLES]. In cases where a is not explicitly available (see examples discussed in [24]), there exist alternative versions of the modified directed likelihood which to some extent share several properties of (6). In particular, an approximation to r_p^* may be derived by replacing the various sample space derivatives with suitable approximations based on covariances of the score function and the log–likelihood and by their derivatives [27, 18].

The problem presented in the current paper might be extended to include linear regression models by assuming that the mean of X_1 and X_2 depend on some covariates [28, 29]. When the interest is only on the AUC, this situation would lead to accounting for more nuisance parameters, and application of higher-order likelihood procedures, which adjust for that, might significantly improve inference in terms of precision.

Higher–order procedures have been presented for complete data. However, it would be of interest to extend our proposal to truncated or censored data, in order to investigate the gain given by these inferential procedures in presence of such incomplete information.

A final point may concern the extension of the problem to the partial area under the ROC curve, when only a restricted range of specificity values are of relevant interest.

Appendix A

In the example about the Gaussian distribution described in Subsection 4.2, in order to compute the Wald and the r_p^* statistics, the profile observed information is obtained from (2). Its expression is

$$j_p(\hat{\psi}) = \frac{\hat{A}}{(n\,\hat{\lambda}_2)^2 \, + \, n_1\, n_2\, (\hat{\lambda}_1 \hat{\lambda}_2 \, - \, D(\hat{\theta}))^2}\,, \tag{14}$$

where $\hat{A} = 2nn_1n_2\hat{\lambda}_2^2\hat{d}^2$, with $\hat{d} = \partial\Phi^{-1}(\hat{\psi})/\partial\hat{\psi} = 1/\phi(\Phi^{-1}(\hat{\psi}))$ and $\phi(\cdot)$ standard normal probability density function.

In the same example, for computing the adjustment term q of the r_p^* statistic, given in (6), we have

$$\frac{|\ell_{\lambda;\hat{\lambda}}(\hat{\theta}_\psi)|}{(|j(\hat{\theta})||j_{\lambda\lambda}(\hat{\theta}_\psi)|)^{1/2}} = \frac{2\hat{\lambda}_2^2\left(2n_1n_2\hat{B}^2 + (n\hat{\lambda}_2)^2 + n_1n_2\hat{B}(\tilde{\lambda}_1\tilde{\lambda}_2 - D(\hat{\theta}_\psi))\right)}{\tilde{\lambda}_2^2\left(\hat{A}\,(12n_1n_2\hat{B}^2 + 8\hat{C}^2 - 2n^2(\tilde{\lambda}_2^2 - 3\hat{\lambda}_2^2) - 8n\hat{C})\right)^{1/2}},$$

with $\hat{B} = \hat{\lambda}_1\hat{\lambda}_2 - D(\hat{\theta})$, $\hat{C} = n_1\hat{\lambda}_1\hat{\lambda}_2 + n_2D(\hat{\theta})$ and $\hat{F} = n_1\tilde{\lambda}_1\tilde{\lambda}_2\hat{\lambda}_1\hat{\lambda}_2 + n_2D(\hat{\theta})D(\hat{\theta}_\psi)$. The remaining terms in q are directly computed in the R software, as shown in Appendix B.

Appendix B

We present the `R` code for the AUC and ROC curves analyses under a parametric setting. Continuous markers can be assumed as exponential variables or Gausssian variables with either equal or different variances. The `R` package `AROC` can be downloaded at `http://homes.stat.unipd.it/gcortese`.

The two data sets from the healthy and diseased populations, respectively, are called `xdata` and `ydata`. MLEs for ϕ and λ, the nuisance parameter, can be obtained from the following code

```
MLEs(xdata,ydata,distr),
```

where `distr` can be set equal to either `"exp"`, `"norm_EV"` or `"norm_DV"`, according as the distributions assumed for the continuous markers are exponential or Gaussian with equal or unequal variances, respectively. The loglikelihood can be computed for a given set of values of ψ and λ by means of the function

```
loglik(xdata,ydata,lambda,psi,distr),
```

where `lambda` and `psi` are, respectively, the nuisance parameter λ and the parameter of interest ψ meaning the area under the ROC curve. For the case of Gaussian distributions with different variances the following simpler reparameterisation has been used:

$$\psi = A(\theta)) = \Phi\left(\frac{\mu_2 - \mu_1}{\sqrt{\sigma_1^2 + \sigma_2^2}}\right), \qquad \lambda = (\lambda_1, \lambda_2, \lambda_3) = (\mu_1, \sigma_1^2, \sigma_2^2),$$

with σ_1^2, σ_2^2 variances of the Gaussian variables X_1 and X_2, respectively.

Point estimates and confidence intervals for the AUC are obtained by using the R function `aroc` as follows:

```
aroc(xdata,ydata,distr,method,level),
```

where the argument `method`, set to be equal to either "Wald", "RP" or "RPstar", allows to choose confidence intervals based on the Wald, r_p or r_p^* statistic, respectively (eq. (3) and (4) and Section 3). When the methods "Wald" or "RP" are selected, point estimate for ψ corresponds to the MLE, while the method "RPstar" yields the r_p^*-based point estimate for ψ as presented in Section 3. The confidence level $(1-\alpha)$ can be decided by setting `level` $= \alpha$.

The Wald, r_p and r_p^* statistics can also be computed for a given value of ψ by applying, respectively, the R functions

```
   w=wald(xdat,ydat,psi,distr),
r=rp(xdat,ydat,psi,distr),
r_star=rpstar(xdat,ydat,psi,distr).
```

The steps 4–6, given at the end of Section 3, for application of the higher-order procedures based on the signed log-likelihood ratio statistic r_p^*, can be implemented by means of the following `R` commands:

```
   smoother = smooth.spline(V.r_star,r_star.range)
psi1 = predict(smoother,z1)$y
psi2 = predict(smoother,z2)$y
hatpsi = predict(smoother,0)$y
```

where `V.r_star` is the vector of values of r_p^* calculated by applying the R function `rpstar` ripetutivaly on an appropriate range `r_star.range` of values of ψ, `psi1` and `psi2` are the limits ψ_1^* and ψ_2^* of the confidence interval corresponding to the percentiles `z1=` $z_{1-\alpha/2}$ and `z2=` $z_{\alpha/2}$, and `hatpsi` is the point estimate $\hat{\psi}^*$.

References

[1] Zweig MH, Campbell G. Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clinical Chemistry* 1993; **39**:561–577.

[2] Zhou XH, Obuchowski NA, McClish DK. *Statistical Methods in Diagnostic Medicine.* Wiley & Sons: New York, , 2002.

[3] Bamber D. The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *J. Mathematical Psycology* 1975; **12**: 387–415.

[4] Wolfe DA, Hogg RV. On constructing statistics and reporting data. *The American Statistician* 1971; **25**:27–30.

[5] Hanley JA. Receiver operating characteristic (ROC) methodology: the state of the art. *Critical Reviews in Diagnostic Imaging* 1989; **29**: 307–335.

[6] Kotz S, Lumelskii Y, Pensky M. *The Stress-Strength Model and its Generalizations. Theory and Applications.* World Scientific: Singapore, 2003.

[7] Pepe MS. *The statistical evaluation of medical tests for classification and prediction.* Oxford University Press: Oxford, 2003.

[8] Tong H. On the estimation of $\Pr\{Y < X\}$ for exponential families. *IEEE Transactions on Reliability* 1977; **26**:54–56.

[9] Metz CE, Herman BA, Shen J-H. Maximum-likelihood estimation of ROC curves from continuously-distributed data. *Statistics in Medicine* 1998; **17**: 1033–1053.

[10] Adimari G, Chiogna M. Partially parametric interval estimation of $\Pr(Y > X)$. *Computational Statistics and Data Analysis* 2006; **51**: 1875–1891.

[11] Mann HB, Whitney DR. On a test whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics* 1947; **18**: 50–60.

[12] Qin GS, Zhou XH. Empirical likelihood inference for the area under the ROC curve. *Biometrics* 2006; **62**: 613–622.

[13] Reiser B, Faraggi D. Confidence intervals for the generalized ROC criterion. *Biometrics* 1997; **53**: 644-652.

[14] Qin G, Hotilovac L. Comparison of non-parametric confidence intervals for the area under the ROC curve of a continuous-scale diagnostic test. *Statistical Methods in Medical Research* 2008; **17**: 207–221.

[15] Obuchoski NA, Lieber ML. Confidence intervals for the receiver operating characteristic area in studies with small samples. *Academic Radiology* 1998; **5**: 561–71.

[16] Newcombe RG. Confidence intervals for an effect size measure based on the Mann-Whitney statistic. Part 2: Asymptotic methods and evaluation. *Statistics in Medicine* 2006; **25**: 559–573.

[17] Zhou W. Statistical inference for $P(X < Y)$. *Statistics in Medicine* 2008; **27**:257–279.

[18] Severini TA. *Likelihood Methods in Statistics.* Oxford University Press: New York, 2000.

[19] Brazzale AR, Davison AC, Reid N. *Applied Asymptotics. Case-Studies in Small Sample Statistics.* Cambridge University Press: Cambridge, 2007.

[20] Newcombe RG. Two-sided confidence intervals for the single proportion: comparison of seven methods. *Statistics in Medicine* 1998; **17**: 857–872.

[21] Barndorff-Nielsen OE, Cox DR. *Inference and Asymptotics.* Chapman and Hall: London, 1994.

[22] Pace L, Salvan A. *Principles of Statistical Inference.* World Scientific: Singapore, 1997.

[23] Pace L, Salvan A. Point estimation based on confidence intervals. *Journal of Statistical Computation and Simulation* 1999; **64**: 1–21.

[24] Giummolé F, Ventura L. Practical point estimation from higher-order pivots. *Journal of Statistical Computation and Simulation* 2002; **72**: 419–430.

[25] Mayer MP,Bukau B. Hsp70 chaperones: cellular functions and molecular mechanism. *Cellular and Molecular Life Sciences* 2005; **62**: 670–84.

[26] Jiang L, Wong ACM. A note on inference for $P(X < Y)$ for right truncated exponentially distributed data. *Statistical Papers* 2008; **49**: 637–651.

[27] Severini TA. An empirical adjustment to the likelihood ratio statistic. *Biometrika* 1999; **86**:235–247.

[28] Guttman I, Johnson RA, Bhattacharyya GK, Reiser B. Confidence limits for stress-strength models with explanatory variables. *Technometrics* 1988; **30**: 161–168.

[29] Schisterman EF, Faraggi D, Reiser B. Adjusting the generalized ROC curve for covariates. *Statistics in Medicine* 2004; **23**: 3319–3331.