Journal of Statistical Software

November 2017, Volume 82, Issue 4. *doi: 10.18637/jss.v082.i04*

ThresholdROC: Optimum Threshold Estimation Tools for Continuous Diagnostic Tests in R

Sara Perez-Jaume
University of Barcelona

Konstantina Skaltsa
University of Barcelona

Natàlia Pallarès
IDIBELL

Josep L. Carrasco
University of Barcelona

Abstract

We introduce an R package that estimates decision thresholds in diagnostic settings with a continuous marker and two or three underlying states. The package implements parametric and non-parametric estimation methods based on minimizing an overall cost function, as well as confidence interval estimation approaches to account for the sampling variability of the cut-off. Further features of the package include sample size determination and estimation of diagnostic accuracy measures. We used randomly generated data and two real datasets to illustrate the capabilities and characteristics of the package.

Keywords: ROC curve, threshold estimation, cost function, diagnostic tests, R package, bootstrap.

1. Introduction

In the diagnostic area, it is of interest to predict the state of a subject (usually, “diseased” or “non-diseased”) using a continuous diagnostic test with a classifying threshold, that is, a value of the diagnostic marker that classifies subjects into two categories: positive and negative for the disease under study. However, the diagnostic problem can also include more than two classification states, such as “non-diseased”, “mild condition” or “severe condition”. The ability of the diagnostic marker to discriminate between states is usually evaluated with the area under the receiver operating characteristic (ROC) curve in the two-state setting (Metz 1978; Pepe 2003) and the volume under the surface (VUS) for the three-state setting (Nakas, Alonzo, and Yiannoutsos 2010).

To estimate the threshold that optimally discriminates between states, standard methods consist of choosing a threshold for a desired false positive/negative rate to be achieved or, more formally, by maximizing the Youden index, which is the sum, diminished by unity, of the two fractions showing the proportions correctly classified (Youden 1950; Nakas *et al.* 2010). Another method based on defining an overall cost function, which includes correct and incorrect classification rates and the relevant weights associated with each decision, thus allowing disease prevalence to be also considered, was proposed (Metz 1978; Pepe 2003) and further developed (Jund, Rabilloud, Wallon, and Ecochard 2005; Skaltsa, Jover, and Carrasco 2010; Skaltsa, Jover, Fuster, and Carrasco 2012). In this methodology, the estimation process is focused on minimizing the cost function. The methodology takes into consideration the following: (1) all classification rates should be taken into account; (2) each wrong or right decision can have a different impact on the final result and (3) disease prevalence can also play a role in threshold selection or estimation. Zweig and Campbell (1993) warned that although a diagnostic test can be highly accurate, "its cost or undesirability of false results may be so high that there is no threshold for which the trade-off between sensitivity and specificity is acceptable". Cantor, Sun, Tortolero-Luna, Richards-Kortum, and Follen (1999) recommended clinicians to weigh their decisions in different fields and provided reasonable values for different applications.

The cost-minimizing approach provides point estimates for the threshold(s) in a given setting. Confidence intervals can also be estimated to account for sampling variability, especially in very overlapping distributions where threshold estimation becomes cumbersome and an alternative management (e.g., further examination) may be required for those subjects with marker values close to the estimated threshold (Skaltsa *et al.* 2012). Further methodological issues related to sample size requirements have also been addressed for the classic two-state setting (Skaltsa *et al.* 2010).

The statistical software currently available for optimum threshold estimation mainly deals with accuracy. However, there are some programs addressing the costs involved in threshold estimation, providing either an expected value for each possible threshold, which should be maximized (e.g., **MedRoc**; StenStat 2017) or a cost function and its values for each threshold, which should be minimized (e.g., **Analyse-it**; Analyse-it Software, Ltd 2017). **ROCR** (Sing, Sander, Beerenwinkel, and Lengauer 2005) is a powerful R package (R Core Team 2017) for ROC visualization that provides tools to plot the cost function when costs for false positives and false negatives are defined. Another relevant R package in the field is **pROC** (Robin *et al.* 2011), which, among many other functions, provides confidence intervals of the sensitivity and the specificity of a given set of thresholds. However, these packages do not estimate the threshold confidence interval or address the sample size issue in this context.

Here, we present the R package **ThresholdROC** (Perez-Jaume, Pallarès, and Skaltsa 2017), which implements a wide range of techniques for threshold estimation and sample size computation. The package is available from the Comprehensive R Archive Network (CRAN) at `https://CRAN.R-project.org/package=ThresholdROC`. In this paper, we briefly present the methodology behind the **ThresholdROC** functions and refer to Skaltsa *et al.* (2010) and Skaltsa *et al.* (2012) for further details. We define the cost function and derive analytical threshold estimators under the binormality/trinormality assumption. We also develop analytical variance estimators and construct confidence intervals for the optimum diagnostic threshold. Moreover, we address the empirical method, which is an alternative approach when no distributional assumptions can be made. The optimal sample size ratio of diseased to non-diseased subjects may be of interest during study design and it can also be obtained using a function in the package under the assumption of binormality. Thus, **ThresholdROC** can perform a wide range of calculations when continuous measurements are involved and the true state of the subjects known. The rest of the article is structured as follows. We describe the methodology for estimation and inference in Section 2. Then, we illustrate how the package can be used to calculate optimum threshold estimates and their confidence intervals in two- and three-state settings using randomly generated data in Section 3 and on two real datasets in Section 4. Finally, a discussion and concluding remarks are given in Section 5.

2. Threshold estimation

In this section, we present the threshold estimation methods implemented by **ThresholdROC** for two- and three-state settings, as well as the methodology for estimating sample size in a binormal setting.

2.1. Optimum threshold estimation

First of all, we will focus on the two-state setting. Let S be a binary variable indicating the true disease status of the subjects (gold standard). We use D to denote the diseased population and $\bar{D}$ for the non-diseased individuals. Similarly, let Y be a second binary variable representing the result of the diagnostic test (1 for a positive test if $X > T$ and 0 for a negative result if $X < T$, where T is a threshold value and X a continuous biomarker). The sensitivity (Se) and specificity (Sp) of the test are defined as

$$\mathrm{Se} = \mathsf{P}\left(Y = 1|S = D\right) \quad \text{and} \quad \mathrm{Sp} = \mathsf{P}\left(Y = 0|S = \bar{D}\right).$$

The overall cost function should be minimized (Metz 1978; Pepe 2003). For the two-state setting, the expression for the cost function is

$$C = \mathrm{TP} \cdot C_{\mathrm{TP}} + \mathrm{FN} \cdot C_{\mathrm{FN}} + \mathrm{FP} \cdot C_{\mathrm{FP}} + \mathrm{TN} \cdot C_{\mathrm{TN}},$$

where TP (true positive) is the fraction of correctly-identified diseased subjects, FP (false positive) the fraction of individuals falsely identified as diseased, FN (false negative) the fraction of subjects falsely identified as non-diseased, and TN (true negative) the fraction of those correctly identified as non-diseased; C_{TP} and C_{TN} represent the costs of correct classifications, and C_{FP} and C_{FN} the costs of incorrect classifications.

Following the approach of Skaltsa *et al.* (2010), the cost-minimizing threshold, T, is the value such that

$$\frac{f_D\left(T\right)}{f_{\bar{D}}\left(T\right)} = \frac{1-\rho}{\rho} \cdot \frac{C_{\mathrm{TN}} - C_{\mathrm{FP}}}{C_{\mathrm{TP}} - C_{\mathrm{FN}}} = R, \tag{1}$$

where $f_D\left(T\right)$ and $f_{\bar{D}}\left(T\right)$ stand for the diseased and non-diseased probability density functions, respectively, and ρ is the prevalence of the disease. It should be noted that R is the product of the cost ratio and the non-diseased odds. Thus, the optimum threshold is such that the ratio of densities equals R.

Assuming normal distributions for each population and under the hypothesis of either equal or unequal variances, two analytical formulas for the optimum threshold can be obtained for the two-state setting (Skaltsa *et al.* 2010). In a distribution-free setting, the optimum threshold can also be estimated on the basis of the empirical costs. In this case, each sample value is used as a threshold and the overall cost calculated. Thus, the optimum threshold is that with the lowest cost. Parametric estimation can also be applied when the distributions for both diseased and non-diseased populations are known.

To generalize the approach used for the two-state setting we will follow Skaltsa *et al.* (2012). Let k be the number of possible states, X a continuous marker and T_l the thresholds between the k states, with $l = 1, \ldots, k-1$. If n is the sample size, ρ_i the prevalence of the ith state, C_{ij} the cost of classifying an individual of class i as class j and F_i the distribution function of the population in the ith class, then the cost function is defined as

$$C = n \sum_{i=1}^{k} \sum_{j=1}^{k} \rho_i C_{ij} \left(F_i \left(T_j \right) - F_i \left(T_{j-1} \right) \right),$$

where $F_i\left(T_0\right) = 0$ and $F_i(T_k) = 1$.

This cost function, which depends on $k-1$ variables, has to be minimized to find the optimum thresholds. For $k = 3$ states, Skaltsa *et al.* (2012) proposed to find the roots of the cost function's first derivatives, that is, to solve the following equations:

$$\begin{cases} \frac{\partial C}{\partial T_1} = n \sum_{i=1}^{3} \rho_i f_i \left(T_1\right) \left(C_{i1} - C_{i2}\right) = 0, \\ \\ \frac{\partial C}{\partial T_2} = n \sum_{i=1}^{3} \rho_i f_i \left(T_2\right) \left(C_{i2} - C_{i3}\right) = 0, \end{cases} \tag{2}$$

where f_i represents the probability density function of measurements on individuals of state i.

Only parametric estimation is considered for the three-state setting. However, solutions can also be obtained in this setting through numerical methods based on non-linear optimization involving bisection and secant methods with inverse quadratic interpolation (Brent 1973).

Now we focus on the variance estimation, which is needed to calculate confidence intervals for the optimum thresholds. In the two-state setting, the sample variance corresponding to the estimated threshold can be approximated by the delta method when assuming binormality using (Jund *et al.* 2005; Skaltsa *et al.* 2010):

$$\mathsf{VAR}\left(T\right) = d \Sigma d^\top,$$

where d is the vector of derivatives of T with respect to $\theta = (\mu_D, \mu_{\bar{D}}, \sigma_D, \sigma_{\bar{D}})$, where $\mu_D, \mu_{\bar{D}}, \sigma_D$ and $\sigma_{\bar{D}}$ stand for the means and standard deviations of the diseased and non-diseased populations, respectively, and Σ is the variance-covariance matrix of θ.

In the three-state setting, variance can be estimated using parametric methods based on non-linear equations (Skaltsa *et al.* 2012; Mak 1993) with the expression:

$$\mathsf{VAR}\left(T_i\right) = \frac{V_{ii}}{A_{ii}^2},$$

where $V_{ii} = \mathsf{VAR}\left(\frac{\partial C}{\partial T_i}\right)$ and A_{ii} are the diagonal elements of the matrix $A = \left(\mathsf{E}\left(\frac{\partial^2 C}{\partial T^2}\Big|_{\hat{T}}\right)\right)^\top$, $\hat{T}$ being a root of $\frac{\partial C}{\partial T}$.

Bootstrapping can also be applied to calculate confidence intervals in both two- and three-state settings (Efron and Tibshirani 1998). There are two ways in which bootstrapping can be used. In the first approach, the standard error of the threshold is estimated using bootstrapping, with the corresponding confidence interval being obtained on the assumption that the threshold estimators follow a normal distribution. In the second approach, the bootstrap percentile confidence interval is calculated.

2.2. Sample size estimation in a binormal setting

To determine the optimal sample size in a two-state setting, we provide a function that computes the optimal sample size ratio, $\varepsilon = n_D/n_{\bar{D}}$, where n_D and $n_{\bar{D}}$ stand for the number of diseased and non-diseased subjects, respectively, needed to achieve a desired width $2L$ and confidence level α in a binormal setting (i.e., when assuming normal distributions for both diseased and non-diseased populations). When we assume equal variances for both groups, the sample sizes for diseased and non-diseased samples are calculated using (Skaltsa *et al.* 2010):

$$n_{\bar{D}} \geq \left(\frac{z_{\alpha/2}}{L}\right)^2 \left(\frac{a\sigma_D^2}{\varepsilon} + b\sigma_{\bar{D}}^2 + \frac{2c\sigma_D^4}{\varepsilon} + 2d\sigma_{\bar{D}}^4\right),$$

$$n_D \geq \left(\frac{z_{\alpha/2}}{L}\right)^2 \left(a\sigma_D^2 + b\varepsilon\sigma_{\bar{D}}^2 + 2c\sigma_D^4 + 2d\varepsilon\sigma_{\bar{D}}^4\right),$$

where

$$a = \left(\frac{\partial T}{\partial \mu_D}\right)^2, \qquad b = \left(\frac{\partial T}{\partial \mu_{\bar{D}}}\right)^2, \qquad c = \left(\frac{\partial T}{\partial \sigma_D}\right)^2, \qquad d = \left(\frac{\partial T}{\partial \sigma_{\bar{D}}}\right)^2,$$

$z_{\alpha/2}$ is the $\alpha/2$th quantile of a standard normal distribution and $\mu_D, \mu_{\bar{D}}, \sigma_D$ and $\sigma_{\bar{D}}$ represent the means and standard deviations of the diseased and non-diseased populations, respectively. Please see Skaltsa *et al.* (2010) for details on the formula used when the variances for the diseased and non-diseased populations are different.

3. The R package ThresholdROC

The **ThresholdROC** package aims to provide functions that implement the methods introduced in the previous section. It contains algorithms that calculate population-based thresholds, point and confidence interval estimates for sample data, sample size estimates and diagnostic accuracy measures. The package also generates plots to provide a better analysis and understanding of the data and results obtained.

3.1. Two-state setting

The **ThresholdROC** functions for the two-state setting were developed on the assumption that the non-diseased distribution takes lower values, although this is just convention. These functions are: `thresTH2()`, which provides an algorithm to calculate the threshold when the distributions of non-diseased and diseased populations are known; `thres2()`, which offers a variety of options to compute point estimates and confidence intervals when sample measurements are available; `secondDer2()`, which allows the assessment of the validity of the threshold calculated; and `SS()` which calculates sample size. **ThresholdROC** also includes functions providing plots related to the threshold estimation.

Population-based threshold

The `thresTH2()` function solves the one-variable equation (1) of the density ratio, as detailed in Section 2, given the population probability distribution functions, parameters and the cost and prevalence values. The `thresTH2()` algorithm uses the `uniroot()` function of the **stats** package, which searches a pre-specified interval for a root of a given function. The probability distribution assumed for the populations (arguments `dist1` and `dist2`, which indicate the probability distribution assumed for the non-diseased and diseased populations, respectively, and can be chosen from any two-parameter continuous distribution implemented in R) has to be specified, as well as their parameters (four parameters in the function that the user should use to specify the first and the second parameter of both distributions), the disease prevalence (`rho`) and the classification costs (`costs` argument). Default values are specified for further options available in the function. It should be noted that the classification costs must be provided in an object of class '`matrix`' as follows:

$$\begin{pmatrix} C_{\text{TP}} & C_{\text{TN}} \\ C_{\text{FP}} & C_{\text{FN}} \end{pmatrix}.$$

If we consider an example in which the non-diseased population follows a standard normal distribution with a mean of 0 and a standard deviation of 1 and the diseased population follows a lognormal distribution with a mean of 1 and a standard deviation of 0.5 on the log scale, with a disease prevalence of 0.3, the following code can be applied to calculate the optimum threshold based on the distributions of the two populations:

```
R> thresTH2(dist1 = "norm", dist2 = "lnorm", par1.1 = 0, par1.2 = 1,
+    par2.1 = 1, par2.2 = 0.5, rho = 0.3)

Threshold: 1.235043

Parameters used
  Disease prevalence: 0.3
  Costs (Ctp, Cfp, Ctn, Cfn): 0 1 0 2.333333
  R: 1
```

We should remark that we used the default cost matrix here, that is, a combination of costs that leads to $R = 1$, which is equivalent to using the Youden index method to obtain the optimum threshold (Skaltsa *et al.* 2010). As we can see in the output provided by `thresTH2()`, the optimum threshold for the example was 1.24. Disease prevalence, costs and R values used in the `thresTH2()` computations are also reported.

Point estimation and confidence intervals

To analyze sample measurements from both non-diseased and diseased populations, **ThresholdROC** contains the point estimation function `thres2()`, which uses as main arguments a vector containing the non-diseased sample values (`k1`), a second vector containing the diseased sample measurements (`k2`), the disease prevalence (`rho`) and the cost matrix (`costs`), which must be passed to `thres2()` as explained before for function `thresTH2()`. Through the `method` argument the user can choose the estimation method. The options currently available include:

- `method = "equal"`: Assumes binormality and equal variances for non-diseased and diseased populations. This is the default value.
- `method = "unequal"`: Assumes binormality with unequal variances.
- `method = "empirical"`: Excludes any distributional assumptions. In this case, each sample value is used as a threshold and the overall cost calculated. The optimum threshold is then chosen as the value that leads to the lowest cost.
- `method = "parametric"`: This estimation method is based on the probability distributions assumed for the two populations. These distributions must be specified through the arguments `dist1` (non-diseased population) and `dist2` (diseased population). As mentioned before, any two-parameter distribution implemented in R can be chosen. Their parameters are estimated from the samples in `k1` and `k2` using the `fitdistr()` function in the **MASS** package (Venables and Ripley 2002) and the threshold estimation is then provided by `thresTH2()`.

The user can also choose the method for calculating the confidence interval corresponding to the threshold estimate using the argument `ci.method`. The choices currently available are:

- `ci.method = "delta"`: Delta method is used to estimate the threshold standard error assuming an underlying binormal model. Thus, this option can only be used when `method` is `"equal"` or `"unequal"`. This is the default value.
- `ci.method = "boot"`: The confidence interval is computed by bootstrapping according to the `method` selected. When `method = "parametric"`, parametric bootstrapping is used (Efron and Tibshirani 1998). Otherwise, non-parametric bootstrapping is applied. The parameter `B`, whose default value is 1000, allows the user to change the number of bootstrap replications to be used.

For further details on these methods please see Skaltsa *et al.* (2010).

Package **ThresholdROC** also includes a function that evaluates the second derivative of the cost function at the estimated threshold (`secondDer2()`), enabling the assessment of whether the estimated threshold leads to a minimum in the cost function. A value close to zero would imply that the minimum of the cost function is found in a plateau of the cost function and it would be advisable to revise the cost assignments.

To illustrate how `thres2()` works, we will use two random samples of size 100 from two different normal distributions. Data from the non-diseased sample are stored in the vector `k1`, whereas those from the diseased population are stored in `k2`.

```
R> set.seed(1234)
R> n <- 100
R> k1 <- rnorm(n, 0, 1)
R> k2 <- rnorm(n, 2, 1)
```

If we assume the disease prevalence to be 0.2 and a binormal setting with equal variances, the optimum threshold and its corresponding confidence interval can be calculated as follows:

```
R> (thr2 <- thres2(k1, k2, 0.2))
```

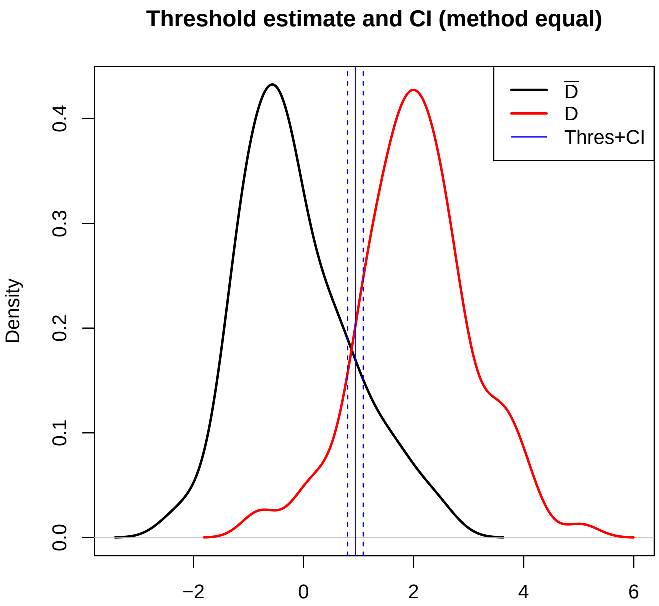

Figure 1: Estimates of the probability density functions for non-diseased and diseased populations, respectively. Also the threshold estimate and its 95% confidence interval are depicted.

```
Estimate:
  Threshold:  0.9422407

Confidence interval (delta method):
  Lower Limit: 0.8011015
  Upper Limit: 1.08338

Parameters used:
  Disease prevalence: 0.2
  Costs (Ctp, Cfp, Ctn, Cfn): 0 1 0 4
  R: 1
  Method: equal
  Significance Level:  0.05
```

The threshold estimate and its confidence interval (and the method used to compute it between brackets) are provided in the output as are the disease prevalence, costs, the R term, estimation method and significance level. Moreover, we can apply the `plot()` method to the '`thres2`' object returned. This method produces a plot that allows visual examination of the problem: estimates of the probability density functions for both samples, as well as vertical lines representing the threshold and its confidence interval. The `plot()` method calls the `density()` function of the **stats** package to compute the density curves, and its default options can be modified with further arguments in the `plot()` function (Figure 1).

```
R> plot(thr2, col = c(1, 2, 4), lwd = c(2, 2, 1), leg.pos = "topright")
```

Now we can check whether the threshold estimate is a minimum of the cost function:

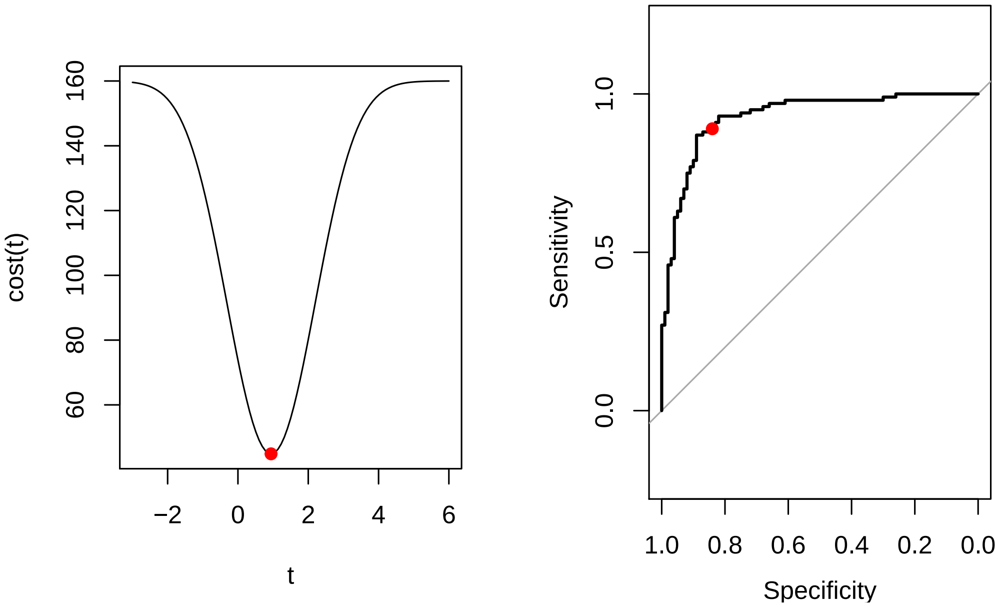

Figure 2: Cost function and ROC curve for the two-state example data.

```
R> round(secondDer2(thr2), 2)
```

```
[1] 74.24
```

The value of the second derivative at the threshold estimate is positive and quite far from zero. Therefore, we can conclude that the threshold estimate is a reliable optimum. The validity of the estimate can be confirmed by plotting the cost function using the `plotCostROC()` function. Notice that we additionally obtain the corresponding ROC curve (Figure 2).

```
R> par(mfrow = c(1, 2))
R> plotCostROC(thr2)
```

Sample size

Package **ThresholdROC** contains the `SS()` function to estimate the optimum sample size ratio (diseased to non-diseased) and the sample size required to achieve a specified confidence interval width and confidence level, assuming a binormal model with either equal or unequal variances. To demonstrate how `SS()` works, we will use the following example in which the non-diseased population follows a normal distribution with a mean of 0 and a standard deviation of 1 and the diseased population follows a normal distribution with a mean of 2 and the same standard deviation, with the disease prevalence being 0.3. Default costs are used. The following code calculates the sample size needed to achieve a desired confidence interval width of 0.5 and a 95% confidence level (default option):

```
R> par1.1 <- 0
R> par1.2 <- 1
R> par2.1 <- 2
```

```
R> par2.2 <- 1
R> rho <- 0.3
R> width <- 0.5
R> SS(par1.1, par1.2, par2.1, par2.2, rho, width, var.equal = TRUE)

Optimum SS Ratio:  1

Sample size for
  Diseased:  30.73167
  Non-diseased:  30.73167

Parameters used
  Significance Level:  0.05
  CI width:  0.5
  Disease prevalence: 0.3
  Costs (Ctp, Cfp, Ctn, Cfn): 0 1 0 2.333333
  R: 1
```

The output shows that the optimum ratio is 1:1, i.e., an equal number of diseased and non-diseased subjects are needed. The minimum sample size required to achieve the desired width of the confidence interval is 31 diseased and 31 non-diseased subjects.

Consider now that the standard deviation of the diseased population is set at 3:

```
R> par2.2 <- 3
R> SS(par1.1, par1.2, par2.1, par2.2, rho, width, var.equal = FALSE)

Optimum SS Ratio:  0.4099684

Sample size for
  Diseased:  44.12022
  Non-diseased:  107.6186

Parameters used
  Significance Level:  0.05
  CI width:  0.5
  Disease prevalence: 0.3
  Costs (Ctp, Cfp, Ctn, Cfn): 0 1 0 2.333333
  R: 1
```

The optimum ratio is now around 0.41; thus, 41 diseased individuals are needed for every 100 non-diseased subjects. Hence, the optimum sample size is 153 subjects, 45 diseased and 108 non-diseased individuals.

3.2. Three-state setting

In a three-state setting, it is assumed that the first population takes lower values and the third population shows the highest values. However, if the populations are labeled in a different way when using the package, they are automatically reordered. The functions related to this setting are: `thresTH3()`, which computes the optimum thresholds based on the distributions assumed for the three states; `thres3()`, which calculates threshold estimates and their confidence intervals when sample measurements for each population are available; and `secondDer3()`, which computes the second derivative of the cost function to validate the estimates. Functions providing plots related to the thresholds and their confidence intervals are also included in package **ThresholdROC**.

Population-based threshold

Similar to `thresTH2()`, `thresTH3()` estimates the theoretical optimum thresholds for specific distribution parameters, decision costs and prevalences in a three-state setting. The equations to be solved to find the optimum thresholds in this setting are given in (2). As before, this is done using the function `uniroot()`. The arguments in this function are similar to those in `thresTH2()`, although here `rho` must be a 3-dimensional vector of prevalences (indicating the prevalence of each underlying state) and `costs` should be a 3×3 `matrix` object as follows:

$$\begin{pmatrix} C_{11} & C_{12} & C_{13} \\ C_{21} & C_{22} & C_{23} \\ C_{31} & C_{32} & C_{33} \end{pmatrix},$$

where C_{ij} is the cost of classifying an individual of class i as class j, for $i, j = 1, 2, 3$. The arguments `dist1`, `dist2` and `dist3` are used to specify the distribution assumed for each population.

To give an example of how this function works, we will consider the following three populations: a standard normal distribution; a lognormal distribution with a mean of 1 and a standard deviation of 0.5 on the log scale; and a lognormal distribution with a mean of 2 and a standard deviation of 0.5 on the log scale. The prevalence of each state is assumed to be 1/3 and the default costs, which lead to the same results as the Youden's method for the three-state setting (Skaltsa *et al.* 2012), will be used.

```
R> thresTH3(dist1 = "norm", dist2 = "lnorm", dist3 = "lnorm",
+    par1.1 = 0, par1.2 = 1, par2.1 = 1, par2.2 = 0.5,
+    par3.1 = 2, par3.2 = 0.5, rho = rep(1/3, 3))

Threshold 1:  1.235043
Threshold 2:  4.481689

Parameters used
  Prevalences: 0.3333333 0.3333333 0.3333333
  Costs
    C11,C12,C13: 0 1 1
    C21,C22,C23: 1 0 1
    C31,C32,C33: 1 1 0
```

As we can see from the results, the threshold estimates are 1.24 and 4.48. The object returned by the function is of class '`thresTH3`', which, in addition to the threshold estimates, also contains information on the parameters used.

Point estimation and confidence intervals

The `thres3()` function calculates threshold estimates and their confidence intervals when sample measurements for each population are available. Similar to `thres2()`, `thres3()` receives as main arguments the samples of the three distributions (namely, `k1`, `k2` and `k3`), the vector of prevalences and the cost matrix (with the same structure as in `thresTH3()`). It is also necessary to specify the distributions assumed for the three populations through the arguments `dist1`, `dist2` and `dist3`. The `start` parameter gives starting values for the thresholds if all the distributions are `"norm"` (that is, when assuming trinormality). If this is not the case, this argument is not required.

The argument `ci.method` can be used for calculating the confidence intervals corresponding to the threshold point estimates. The options available are:

- `ci.method = "param"`: This provides parametric confidence intervals based on methods on non-linear equations. This option can only be used when all the populations are assumed to follow normal distributions.

- `ci.method = "boot"`: This provides confidence intervals computed by bootstrapping. The confidence intervals for each threshold can be calculated in two ways. The first one involves bootstrapping to compute the standard error and then the confidence limits are calculated using a standard normal distribution. The second option is the common percentile approach based on calculating the empirical percentiles of the bootstrap threshold estimates. Parametric bootstrapping is used if any one of the population is not assumed to follow a normal distribution. The number of bootstrap resamples can be chosen through argument `B`, whose default value is 1000.

The object returned by function `thres3()` is of class '`thres3`' and contains the results about the threshold estimates, their confidence intervals and further information.

In this setting, as in the two-state setting, package **ThresholdROC** also contains the function `secondDer3()`, which calculates the second partial derivatives of the cost function to assess if the threshold estimates lead to a minimum in the cost function (when the derivatives are positive) or if such a minimum does not exist (when the derivatives are close to zero).

To illustrate the usage of function `thres3()`, we will use three random samples of size 100: a lognormal distribution with a mean of 0 and a standard deviation of 1 on the log scale, and two normal distributions with means of 3 and 5, respectively, and both with a standard deviation of 1. Prevalences are assumed to be $\frac{1}{3}$ and default costs are used.

```
R> set.seed(1234)
R> n <- 100
R> k1 <- rlnorm(n)
R> k2 <- rnorm(n, 3, 1)
R> k3 <- rnorm(n, 5, 1)
R> rho <- c(1/3, 1/3, 1/3)
R> (thr3 <- thres3(k1, k2, k3, rho, dist1 = "lnorm", dist2 = "norm",
+    dist3 = "norm", ci.method = "boot"))

Estimate:
  Threshold 1:  1.750509
  Threshold 2:  4.102581

Confidence intervals (parametric bootstrap):
  CI based on normal distribution for Threshold 1:  1.61558  -  1.885438
  CI based on percentiles for Threshold 1:  1.626109  -  1.893765
  CI based on normal distribution for Threshold 2:  3.964265  -  4.240896
  CI based on percentiles for Threshold 2:  3.973393  -  4.241973
  Bootstrap resamples:  1000

Parameters used:
  Prevalences: 0.3333333 0.3333333 0.3333333
  Costs
    C11,C12,C13: 0 1 1
    C21,C22,C23: 1 0 1
    C31,C32,C33: 1 1 0
  Confidence Level:  0.05
  Distribution assumed for the first sample: lnorm(-0.16, 1)
  Distribution assumed for the second sample: norm(3.04, 1.03)
  Distribution assumed for the third sample: norm(5.15, 0.96)
```

The threshold estimates and their confidence intervals are provided in the output. As bootstrapping was used, two confidence intervals for each threshold are generated, one based on the normal distribution and the other on percentiles. The output of this function also displays information about the other parameters used. Applying the method `plot()` to the object returned by this function, we obtain a graph showing the estimations of the three probability density functions and vertical lines representing the threshold estimates and their confidence intervals (Figure 3).

```
R> plot(thr3, col = 1:4, lwd = c(2, 2, 2, 1), leg.pos = "topright")
```

Through `secondDer3()`, we can assess the validity of the estimate:

```
R> round(secondDer3(thr3), 2)

Value for thres1 Value for thres2
            0.13             0.16
```

The values obtained are positive but quite close to zero. We can also plot the contribution of each threshold to the cost function (Figure 4).

```
R> par(mfrow = c(1, 2))
R> plotCostROC(thr3)
```

As we can see in Figure 4, both thresholds lead to a minimum in the cost function. Furthermore, the cost functions do not show any plateau, indicating that these minimums can be considered reliable optima.

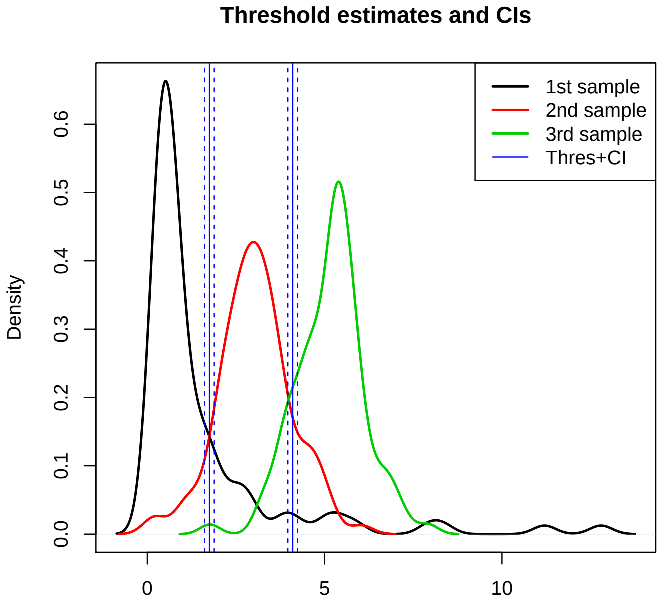

Figure 3: Estimates of the probability density functions of the three populations. Also the threshold estimates and their confidence intervals are depicted.

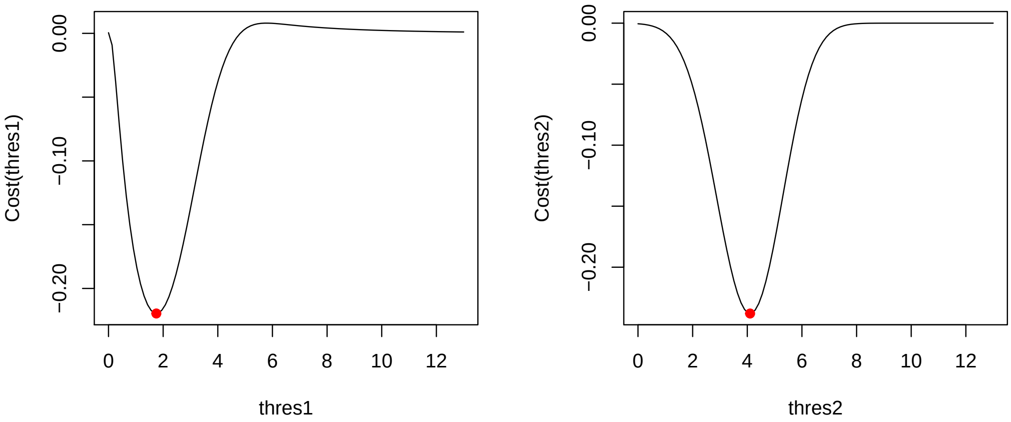

Figure 4: Cost function with respect to both thresholds in the three-state example.

4. Case examples

In order to illustrate the techniques described in the previous sections and the use of the respective R functions, we applied package **ThresholdROC** to two real datasets, one for each diagnostic setting. The datasets were both analyzed in Skaltsa *et al.* (2010, 2012), and we present them here for illustration purposes. Both datasets are available in the **ThresholdROC** package.

4.1. Two-state setting

For the two-state setting, we used a dataset from Kapaki, Paraskevas, Zalonis, and Zournas (2003), which contains measurements of tau protein levels in the cerebrospinal fluid of 49 control subjects and 49 patients with Alzheimer's disease (AD). The authors reported that the cut-off of 317 led to an optimal combination of sensitivity (0.88) and specificity (0.96), producing $R = 1$. We calculated an alternative threshold for this dataset using the functions in **ThresholdROC**. This approach accounts for specific characteristics of the problem by choosing a reasonable combination of costs based on clinical criteria.

We set the costs corresponding to correct classification at zero (that is, $C_{\mathrm{TP}} = C_{\mathrm{TN}} = 0$) given that there are no consequences when correct decisions are made. The costs corresponding to false classifications were set at 1 (i.e., $C_{\mathrm{FP}} = C_{\mathrm{FN}} = 1$), placing the same weight on false positives and false negatives. The value for AD prevalence was set at 0.2 based on the literature (Tsolaki, Fountoulakis, Pavlopoulos, Chatzi, and Kazis 1999; Ferri *et al.* 2005).

To determine if the measurements from both the control and diseased groups could be assumed to follow a normal distribution in deciding the method to use for threshold estimation, we applied Shapiro-Wilk's test to the measurements, obtaining $p < 0.01$ in both cases. Since the data failed the normality tests, the empirical method was used. Thus, using `thres2()`, the threshold that minimizes the cost function for these cost and prevalence values was estimated to be 384.45, corresponding to a sensitivity of 0.76 and a specificity of 1. Using a confidence level of 95% and applying the bootstrap methodology with 1000 resamples, the confidence intervals of the threshold were $(304.40, 464.51)$ for the bootstrap method based on the normal distribution and $(295.37, 444.03)$ for the percentile technique. It should be noted that this approach leads to $R = 4$. Figure 5, obtained through the `plot` method for '`thres2`' objects, shows the estimates for the probability density function of each group, as well as the estimated threshold and its confidence intervals determined by bootstrapping. We also plotted the empirical cost function for this application, and the empirical ROC curve (Figure 6, obtained with `plotCostROC()`). Using the empirical cost function graph, we can conclude that the

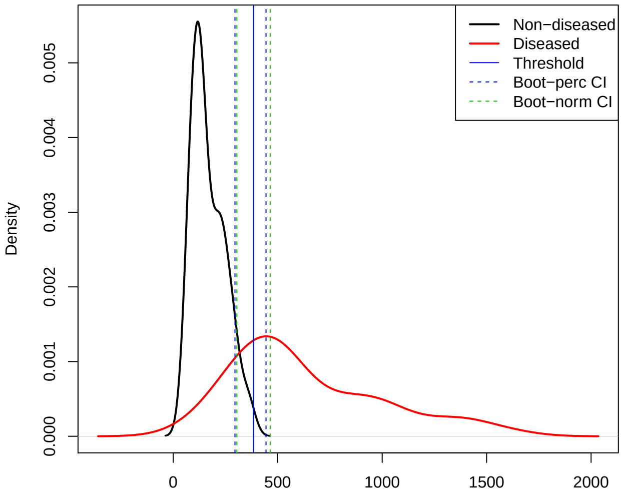

Figure 5: Alzheimer's disease data: Estimates of the probability density functions for tau protein measurements in non-diseased and diseased groups. Also the threshold estimate and its confidence intervals calculated by bootstrapping are depicted.

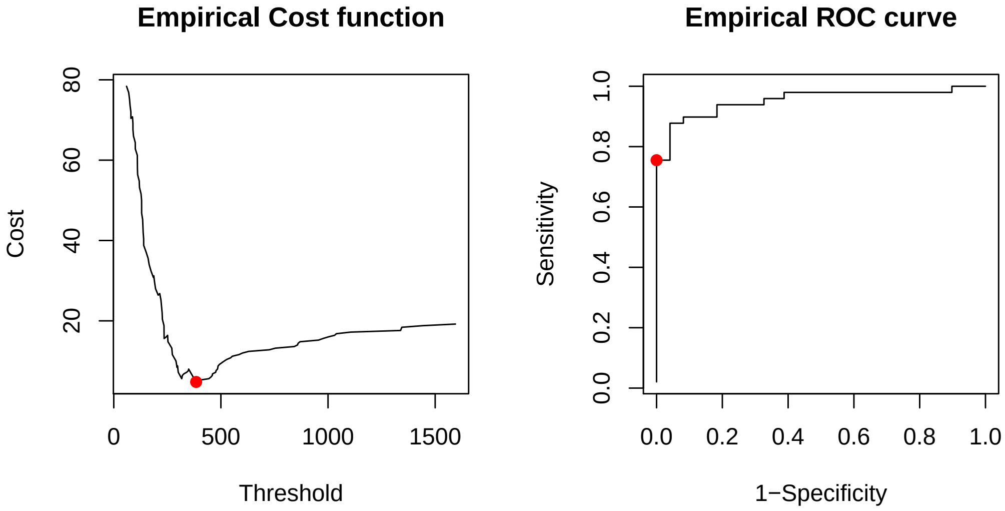

Figure 6: Alzheimer's disease data: Empirical cost function and ROC curve.

threshold estimate (shown as a red dot on the plot) leads to a minimum in the cost function. However, the function is noticeably flat around this point, implying that any value around the estimate, of 384.45, can be considered a plausible threshold. Regarding the empirical ROC curve, we must point out that our threshold estimate did not lead to an optimal combination of specificity and sensitivity because the choice of costs did not lead to the same results as those obtained with Youden's method.

4.2. Three-state setting

For the three-state setting, we used a dataset from a study on chemotherapy response in breast cancer patients (Duch *et al.* 2009). Positron emission tomography (PET) was performed just before the beginning of chemotherapy and after the second cycle to decide whether treatment should be discontinued or modified. Uptake in PET studies was quantified by the difference in the standardized uptake value (SUV), which was our continuous diagnostic measurement. After surgery, response to chemotherapy was evaluated using the pathology results from the surgical specimen which was taken as gold standard by assigning one of the following three states: stable disease, partial remission and complete remission. The aim of that study was to estimate the optimum thresholds of the SUV variable between the three states of the response variable. Out of 50 subjects, 12 remained stable, 29 presented partial response and 9 showed complete remission. Radiologists were asked to assign plausible cost values (see Table 1; prevalence values are also shown) on a scale ranging from 0 to 5. Null costs were assigned for correct classifications. The cost of classifying a patient with a stable tumor who had responded partially was $C_{12} = 2$, whereas the inverse error was penalized with a cost of $C_{21} = 1$, given that the latter situation is less serious than the first one. Classifying an individual with a complete response as having a partial response and the inverse situation were both assigned a cost of $C_{23} = C_{32} = 1$. Classifying a patient with a stable tumor as having shown a complete response was considered to be the most serious error because of its medical implications and was penalized with a cost of $C_{13} = 5$. The inverse error was also considered important and its cost set at $C_{31} = 4$.

	Costs			
Prevalences	Correct classification	Incorrect classification		
$\rho_1 = 0.24$	$C_{11} = 0$	$C_{12} = 2$	$C_{13} = 5$	$C_{23} = 1$
$\rho_2 = 0.58$	$C_{22} = 0$	$C_{21} = 1$	$C_{31} = 4$	$C_{32} = 1$
$\rho_3 = 0.18$	$C_{33} = 0$			

Table 1: The prevalence and cost values for the chemotherapy response dataset. 1 denotes stable subjects, 2 the partial responders and 3 the complete responders.

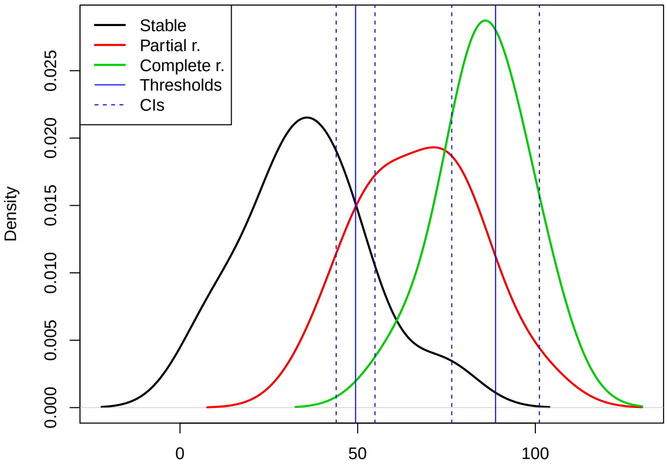

Figure 7: Chemotherapy response data: SUV difference densities for patients who remained stable, partially responded or completely responded to treatment. Also the threshold estimates and their confidence intervals are depicted.

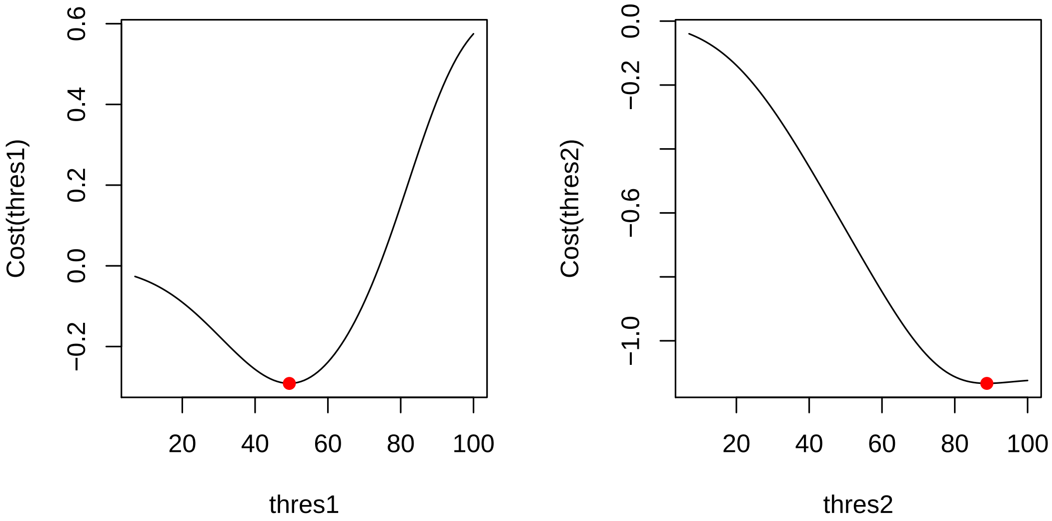

Figure 8: Chemotherapy response data: Cost function with respect to both thresholds.

We applied Shapiro-Wilk's test to the measurements from each population to assess the normality of the data, obtaining $p = 0.82$ for the group with stable tumor, $p = 0.49$ for the partial responders, and $p = 0.24$ for the complete responders. Thus, we could assume a situation of trinormality. Under this assumption, using `thres3()`, the threshold estimates and their 95% confidence intervals were 49.41 $(43.95, 54.87)$ and 88.80 $(76.47, 101.13)$. Confidence intervals were estimated based on the parametric method. A representation of the results is shown in Figure 7 (obtained with the `plot` method for '`thres3`' objects). Evaluating the second derivatives of the cost function in the threshold estimates through `secondDer3()` we obtained positive values (0.044 and 0.017), confirming that the estimates lead to a minimum in the cost function. The cost function corresponding to both estimates was plotted with `plotCostROC()` (Figure 8) graphically confirming that they lead to a minimum.

Using `VUS()` from **DiagTest3Grp** package (Luo and Xiong 2012), we calculated the volume under surface (VUS) for the biomarker SUV to be 0.72 (95% confidence interval, [0.57, 0.88]), thus, underlining the highly discriminatory capacity of the SUV.

5. Conclusions

The **ThresholdROC** package, which is publicly available from CRAN at `https://CRAN.R-project.org/package=ThresholdROC`, contains a set of functions intended to provide direct calculations of the optimum thresholds for continuous diagnostic tests using the methods described briefly in this article and more extensively in Skaltsa *et al.* (2010, 2012). Here, we illustrate the capabilities of package **ThresholdROC** in estimating optimum thresholds based on minimizing an overall cost function in a two- and three-state settings. Package **ThresholdROC** can also be used to calculate population-based thresholds, point estimates and confidence intervals for both two- and three-state settings. Moreover, it provides graphical tools related to the threshold estimates, allowing a deeper understanding of both the data and the results obtained. Package **ThresholdROC** also contains a function that estimates optimal sample sizes.

In addition to estimating optimum thresholds and sample sizes, package **ThresholdROC** also includes the function `diagnostic()`, which calculates common measures of the accuracy of diagnostic tests involving 2×2 contingency tables of classification results (usually, test outcome versus status tables). Specifically, it calculates the following statistical measures: sensitivity, specificity, positive and negative predictive value, positive and negative likelihood ratio, odds ratio, Youden's index, accuracy, error rate and appropriate confidence intervals for each index (Zhou, Obuchowski, and McClish 2002). This can be useful in a two-state setting when assessing the validity of a dichotomic test based on categorizing a continuous marker using a threshold estimate.

Acknowledgments

The authors would like to thank Dr. D. Fuster for the chemotherapy response dataset used to illustrate the **ThresholdROC** functions for a three-state setting. We also thank Toffa Evans from Language Services (University of Barcelona), who improved the English text. We are grateful to the reviewers for their valuable comments, which have led to substantial improvements in this article.

References

Analyse-it Software, Ltd (2017). ***Analyse-it** 4.80: Your New Go-To Statistics Package.* URL https://www.analyse-it.com/.

Brent R (1973). *Algorithms for Minimization without Derivatives.* Prentice-Hall.

Cantor SB, Sun CC, Tortolero-Luna G, Richards-Kortum R, Follen M (1999). "A Comparison of C/B Ratios from Studies Using Receiver Operating Characteristics Curve Analysis." *Journal of Clinical Epidemiology*, **52**(9), 885–892. doi:10.1016/s0895-4356(99)00075-x.

Duch J, Fuster D, Munoz M, Fernandez PL, Paredes P, Fontanillas M, Guzman F, Rubi S, Lomena FJ, Pons F (2009). "^{18}F-FDG PET/CT for Early Prediction of Response to Neoadjuvant Chemotherapy in Breast Cancer." *European Journal of Nuclear Molecular Imaging*, **36**(10), 1551–1557. doi:10.1007/s00259-009-1116-y.

Efron B, Tibshirani RJ (1998). *An Introduction to the Bootstrap.* Chapman and Hall/CRC.

Ferri CP, Prince M, Brayne C, Brodaty H, Fratiglioni L, Ganguli M, Hall K, Hasegawa K, Hendrie H, Huang Y, Jorm A, Mathers C, Menezes PR, Rimmer E, Scazufca M (2005). "Global Prevalence of Dementia: A Delphi Consensus Study." *Lancet*, **366**(9503), 2112–2117. doi:10.1016/s0140-6736(05)67889-0.

Jund J, Rabilloud M, Wallon M, Ecochard R (2005). "Methods to Estimate the Optimal Threshold for Normally or Log-Normally Distributed Biological Tests." *Medical Decision Making*, **25**(4), 406–415. doi:10.1177/0272989x05276855.

Kapaki E, Paraskevas GP, Zalonis I, Zournas C (2003). "CSF Tau Protein and β-Amyloid (1-42) in Alzheimer's Disease Diagnosis: Discrimination from Normal Ageing and the Other Dementias in the Greek Population." *European Journal of Neurology*, **10**(2), 119–128. doi:10.1046/j.1468-1331.2003.00562.x.

Luo J, Xiong C (2012). "**DiagTest3Grp**: An R Package for Analyzing Diagnostic Tests with Three Ordinal Groups." *Journal of Statistical Software*, **51**(3), 1–24. doi:10.18637/jss.v051.i03.

Mak TK (1993). "Solving Non-Linear Estimation Equations." *Journal of the Royal Statistical Society B*, **55**(4), 945–955.

Metz C (1978). "Basic Principles of ROC Analysis." *Seminars in Nuclear Medicine*, **8**(4), 283–298. doi:10.1016/s0001-2998(78)80014-2.

Nakas CT, Alonzo A, Yiannoutsos CT (2010). "Accuracy and Cut-Off Point Selection in Three-Class Classification Problems Using a Generalization of the Youden Index." *Statistics in Medicine*, **29**(28), 2946–2955. doi:10.1002/sim.4044.

Pepe MS (2003). *The Statistical Evaluation of Medical Tests for Classification and Prediction.* Oxford University Press.

Perez-Jaume S, Pallarès N, Skaltsa K (2017). ***ThresholdROC**: Threshold Estimation.* R package version 2.6, URL https://CRAN.R-project.org/package=ThresholdROC.

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez JC, Müller M (2011). "**pROC**: An Open-Source Package for R and S+ to Analyze and Compare ROC Curves." *BMC Bioinformatics*, **12**(77). doi:10.1186/1471-2105-12-77.

Sing T, Sander O, Beerenwinkel N, Lengauer T (2005). "**ROCR**: Visualizing Classifier Performance in R." *Bioinformatics*, **21**(20), 3940–3941. doi:10.1093/bioinformatics/bti623.

Skaltsa K, Jover L, Carrasco JL (2010). "Estimation of the Diagnostic Threshold Accounting for Decision Costs and Sampling Uncertainty." *Biometrical Journal*, **52**(5), 676–697. doi:10.1002/bimj.200900294.

Skaltsa K, Jover L, Fuster D, Carrasco JL (2012). "Optimum Threshold Estimation Based on Cost Function in a Multistate Diagnostic Setting." *Statistics in Medicine*, **31**(11–12), 1098–1109. doi:10.1002/sim.4369.

StenStat (2017). ***MedRoc** 2.0: Software for ROC Analysis of Biomedical Data*. URL https://stenstat.com/MedRoc/MedRoc.htm.

Tsolaki M, Fountoulakis C, Pavlopoulos I, Chatzi E, Kazis A (1999). "Prevalence and Incidence of Alzheimer's Disease and Other Dementing Disorders in Pylea, Greece." *American Journal of Alzheimer's Disease & Other Dementias*, **14**(3), 138–148. doi:10.1177/153331759901400308.

Venables WN, Ripley BD (2002). *Modern Applied Statistics with S*. 4th edition. Springer-Verlag, New York. doi:10.1007/978-0-387-21706-2.

Youden WJ (1950). "Index for Rating Diagnostic Tests." *Cancer*, **3**(1), 32–35. doi:10.1002/1097-0142(1950)3:1<32::aid-cncr2820030106>3.0.co;2-3.

Zhou XH, Obuchowski NA, McClish DK (2002). *Statistical Methods in Diagnostic Medicine*. John Wiley & Sons. doi:10.1002/9780470317082.

Zweig MH, Campbell G (1993). "Receiver Operating Characteristics (ROC) Plots: A Fundamental Tool in Clinical Medicine." *Clinical Chemistry*, **39**(4), 561–577.

Affiliation:

Sara Perez-Jaume, Konstantina Skaltsa, Josep L. Carrasco
Biostatistics. Department of Basic Clinical Practice
School of Medicine
University of Barcelona

Casanova 143, 08036 Barcelona, Spain
E-mail: jlcarrasco@ub.edu
and
Sara Perez-Jaume
Developmental Tumor Biology Laboratory
Hospital-Fundació Sant Joan de Déu
Passeig Sant Joan de Déu 2
08950 Esplugues de Llobregat, Barcelona, Spain
E-mail: sperezj@fsjd.org

Natàlia Pallarès
Statistics Advisory Service
Institute of Biomedical Research of Bellvitge (IDIBELL)
Gran Via de l'Hospitalet 199
08908 Hospitalet de Llobregat, Barcelona, Spain
E-mail: npallares@idibell.cat

Journal of Statistical Software http://www.jstatsoft.org/
published by the Foundation for Open Access Statistics http://www.foastat.org/
November 2017, Volume 82, Issue 4 *Submitted:* 2015-11-24
doi:10.18637/jss.v082.i04 *Accepted:* 2016-12-16