ROC curves as an aspect of classification

Jan Koláček

Department of Mathematics and Statistics
Faculty of Science
Masaryk University
Brno, Czech Republic
www.muni.cz

Contents

Measure of Diagnostic Accuracy

- $\mathcal{G}_0$ group of n_0 subjects without a condition.
- $\mathcal{G}_1$ group of n_1 subjects with a condition
- $D = 0,\ 1$ random variable denotes absence or presence of the condition
- $T = 1$ positive test result
- $T = 0$ negative test result

Test results: Confusion matrix

	Positive test, $T = 1$	Negative test, $T = 0$	Total
$\mathcal{G}_1$ $(D = 1)$	True positive (TP)	False negative (FN)	$TP + FN$
$\mathcal{G}_0$ $(D = 0)$	False positive (FP)	True negative (TN)	$FP + TN$
Total	$TP + FP$	$FN + TN$	$n = n_0 + n_1$

The *sensitivity* (Se) of the test is its ability to detect the condition when it is present.

$Se = P(T = 1|D = 1)$ is a probability P that the test result is positive $(T = 1)$, given that the condition is present $(D = 1)$.

The *specificity* (Sp) of a test is its ability to exclude the condition when it is absent.

$Sp = P(T = 0|D = 0)$ is a probability P that the test result is negative $(T = 0)$, given that the condition is absent $(D = 0)$.

$$Se = \frac{TP}{TP + FN}, \; Sp = \frac{TN}{FP + TN}$$

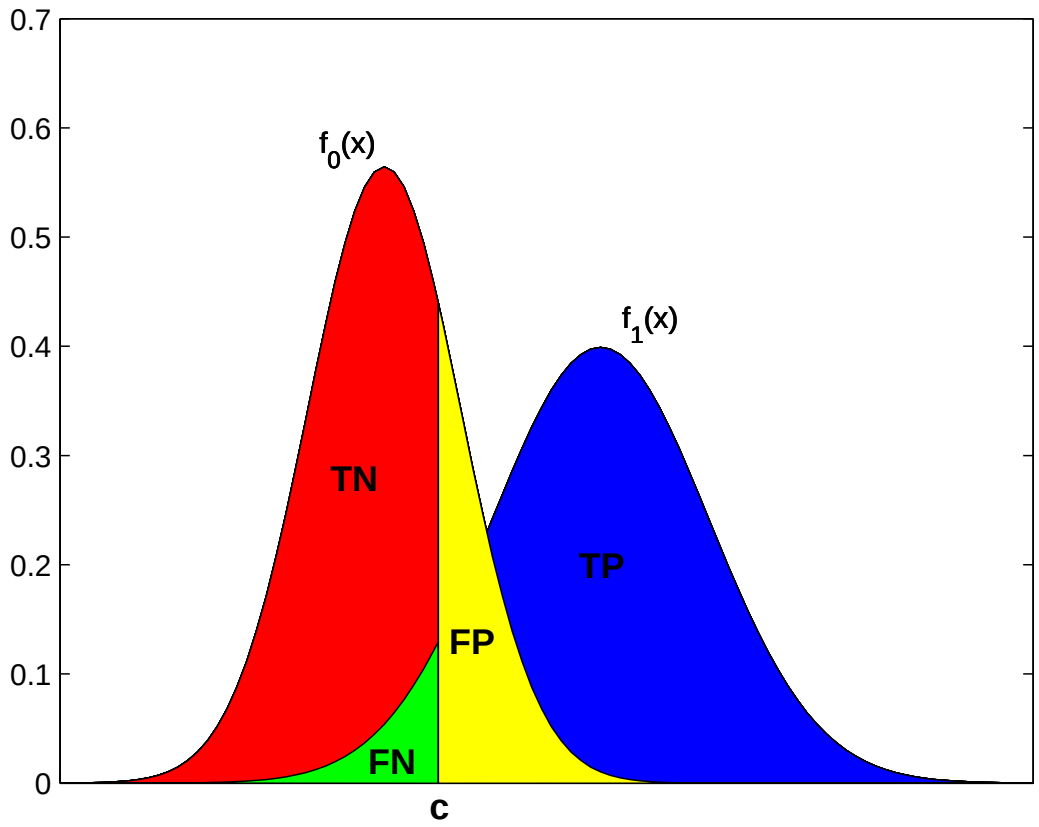
Accuracy: $Ac = \frac{TP + TN}{n}$

False positive rate: $FPR = 1 - Sp = \frac{FP}{FP + TN}$

Receiver Operating Characteristic (ROC) Curve

Construction of the test

- X – the diagnostic test variable (one-dimensional absolutely continuous random variable)
- c – given cutoff point, $c \in \mathbb{R}$
- The subject is classified as $\mathcal{G}_1$ if $X \geq c$ and $\mathcal{G}_0$ otherwise for given cutoff point c
- $$F_0(c) = P(X \leq c|\mathcal{G}_0) = \int_{-\infty}^{c} f_0(x)dx$$
  $$F_1(c) = P(X \leq c|\mathcal{G}_1) = \int_{-\infty}^{c} f_1(x)dx$$
  F_0 or F_1 are cumulative distribution functions (c.d.f.) of group $\mathcal{G}_0$ or $\mathcal{G}_1$ and f_0 and f_1 are corresponding probability density functions (p.d.f.).



TN – True Negative FN – False Negative

FP – False Positive TP – True Positive

$$f_0(x) = \frac{1}{\sqrt{\pi}} e^{-x^2}, \qquad f_1(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-2)^2}{2}}$$

- F_0 – the specificity (Sp) of the test
- $1 - F_1$ – the sensitivity (Se) of the test
- p – the probability of false classification of subject from $\mathcal{G}_0$
- q – the probability of true classification of subject from $\mathcal{G}_1$

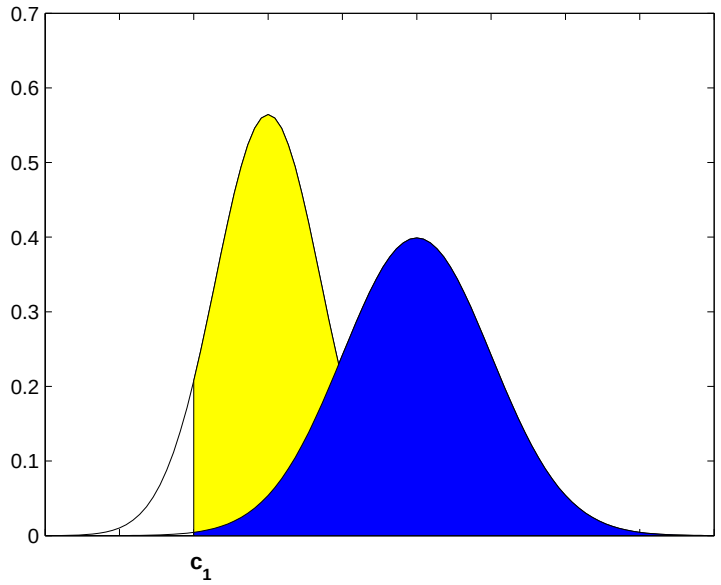
$$p = 1 - F_0(c) \Rightarrow c = F_0^{-1}(1-p),\ 0 \leq p \leq 1$$
$$q = 1 - F_1(c) = 1 - F_1(F_0^{-1}(1-p)),\ 0 \leq p \leq 1$$

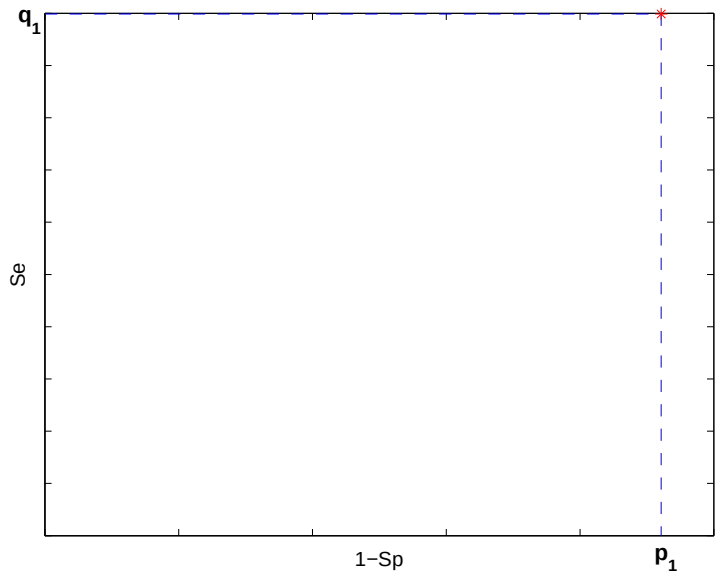
$$\Downarrow$$

$$ROC(p) = 1 - F_1(F_0^{-1}(1-p)),\ 0 \leq p \leq 1$$

ROC curve is displayed by plotting $\underbrace{1 - F_1(c)}_{Se}$ against $\underbrace{1 - F_0(c)}_{1-Sp}$

for a range of cutoff points $c \in \mathbb{R}$.



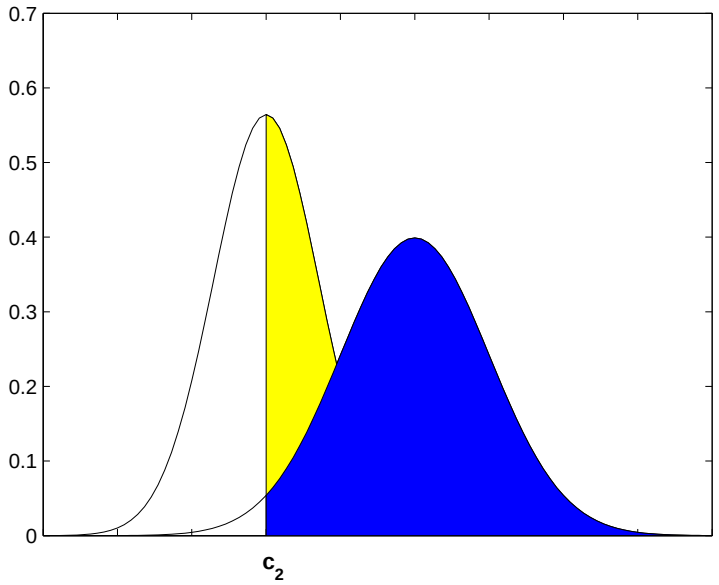


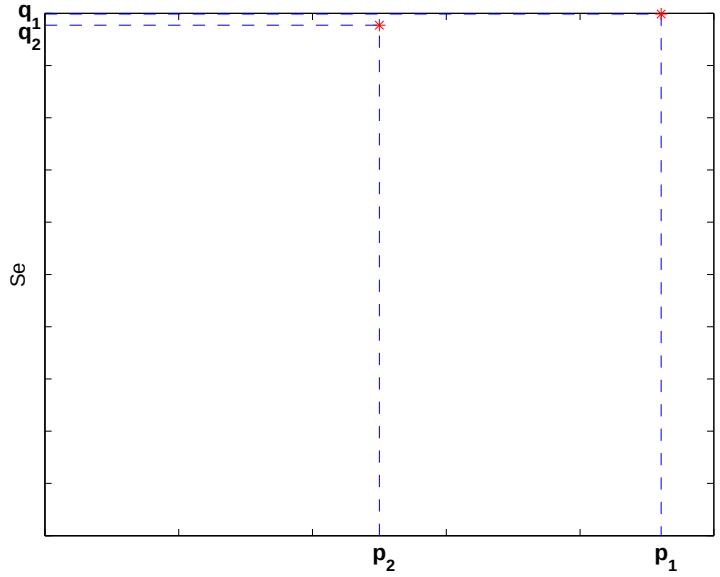
$$c_1 = -1$$

$$p_1 = 1 - F_0(c_1) = \int_{-1}^{\infty} f_0(x)dx = 0.9214$$

$$q_1 = 1 - F_1(c_1) = \int_{-1}^{\infty} f_1(x)dx = 0.9987$$

$$Sp = 0.0786, \qquad Se = 0.9987$$



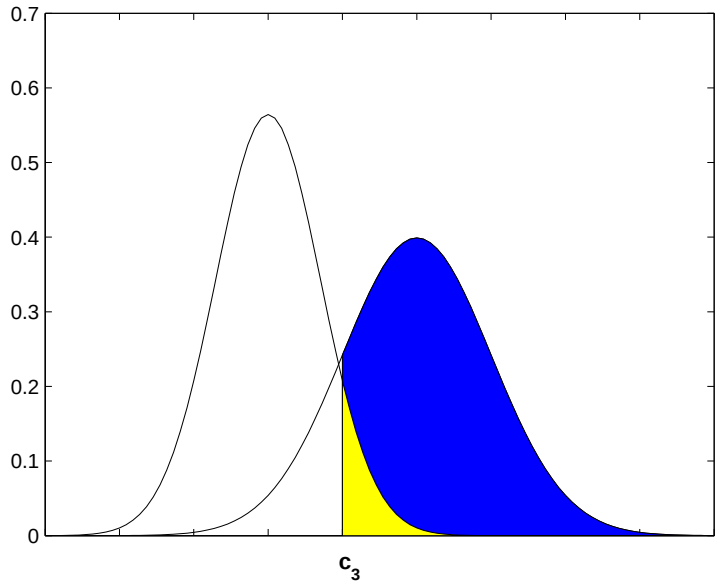


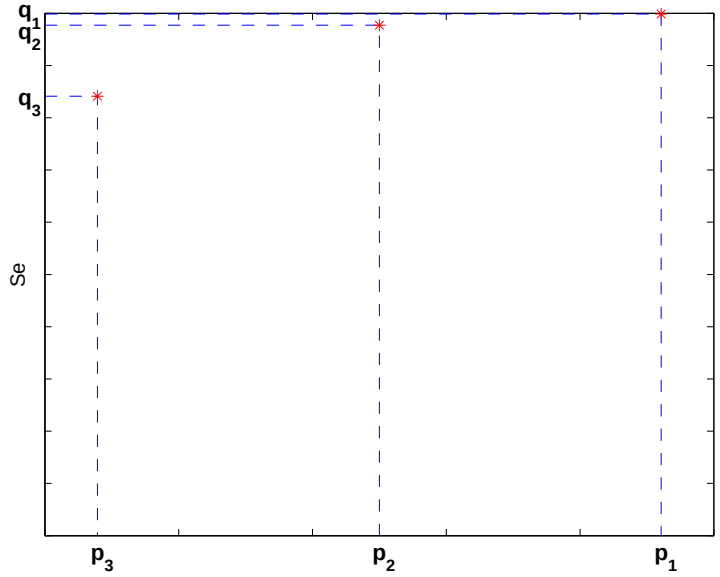
$$c_2 = 0$$

$$p_2 = 1 - F_0(c_2) = \int_0^\infty f_0(x)dx = 0.5$$

$$q_2 = 1 - F_1(c_2) = \int_0^\infty f_1(x)dx = 0.9772$$

$$Sp = 0.5, \qquad Se = 0.9772$$



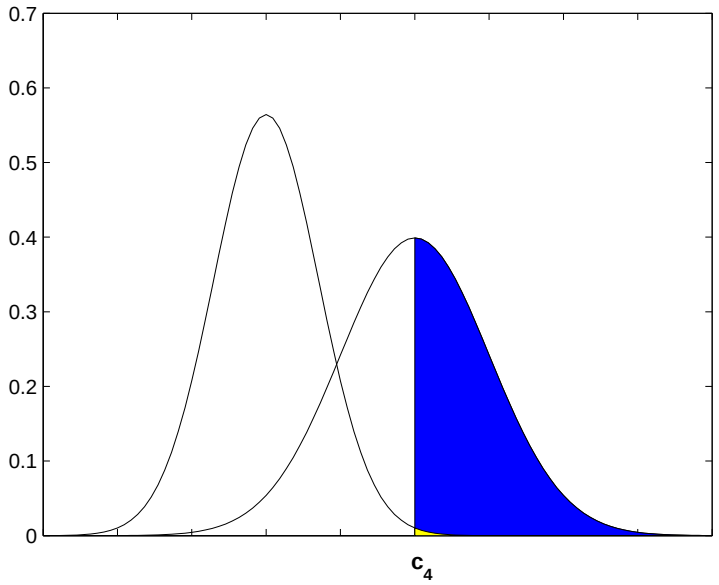


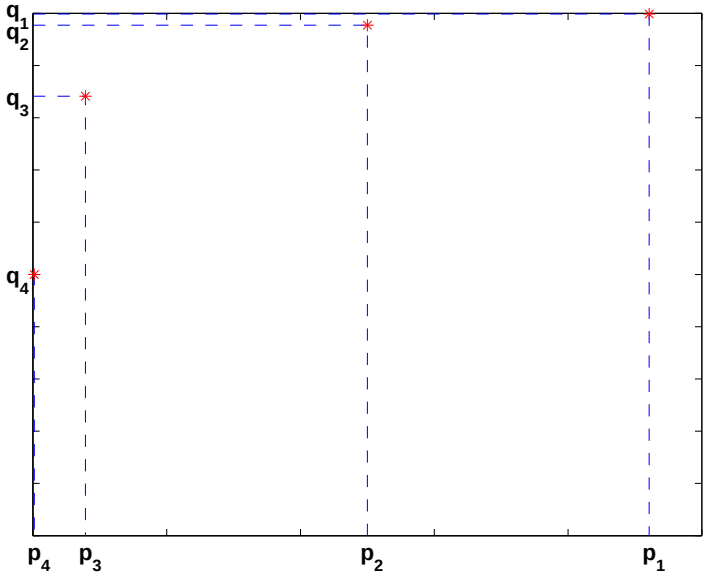
$$c_3 = 1$$

$$p_3 = 1 - F_0(c_3) = \int_1^\infty f_0(x)dx = 0.0786$$

$$q_3 = 1 - F_1(c_3) = \int_1^\infty f_1(x)dx = 0.8413$$

$$Sp = 0.9214, \qquad Se = 0.8413$$



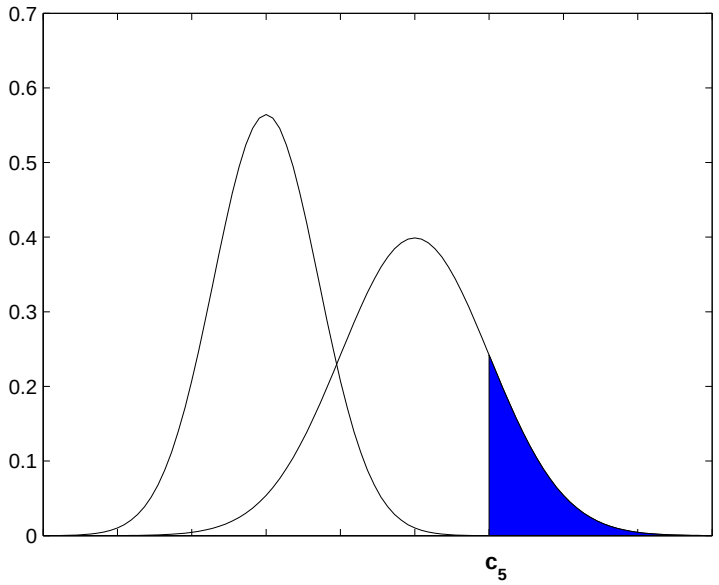


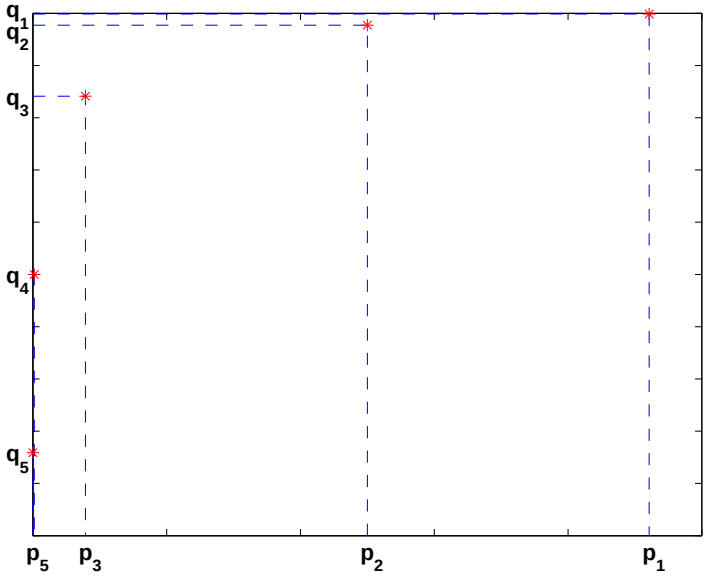
$$c_4 = 2$$

$$p_4 = 1 - F_0(c_4) = \int_2^\infty f_0(x)dx = 0.0023$$

$$q_4 = 1 - F_1(c_4) = \int_2^\infty f_1(x)dx = 0.5$$

$$Sp = 0.9977, \qquad Se = 0.5$$



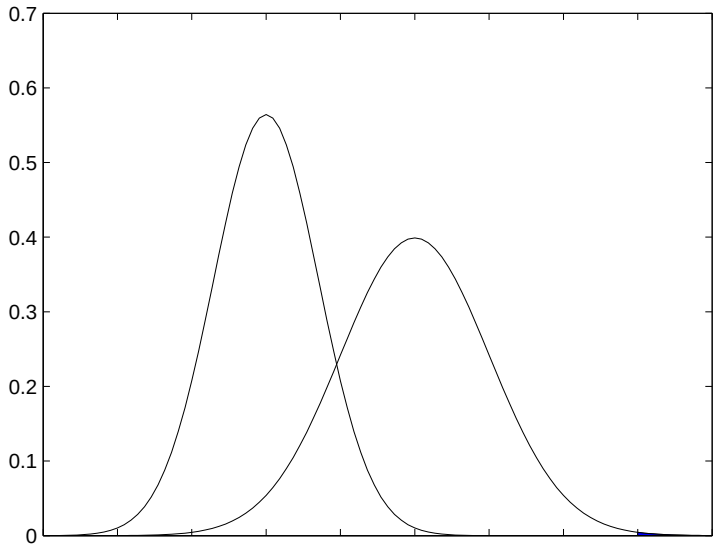


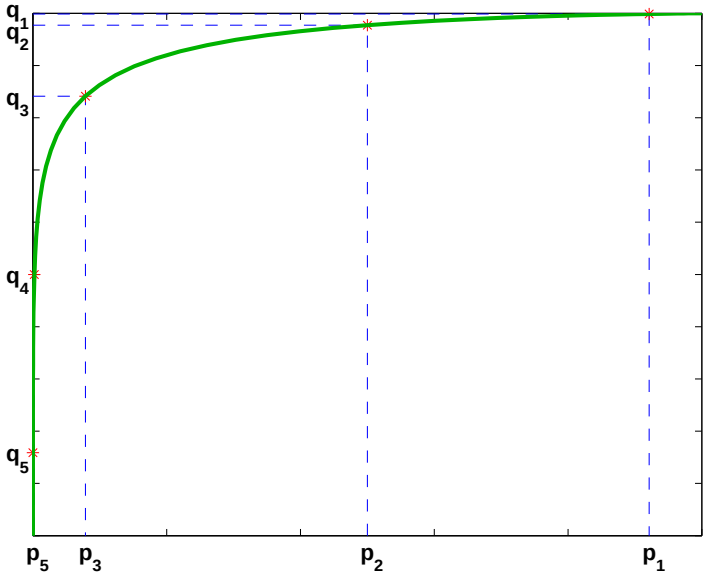
$$c_5 = 3$$

$$p_5 = 1 - F_0(c_5) = \int_3^\infty f_0(x)dx = 0.00001$$

$$q_5 = 1 - F_1(c_5) = \int_3^\infty f_1(x)dx = 0.1587$$

$$Sp = 0.99999, \qquad Se = 0.1587$$



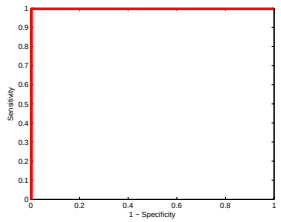


$$p_i = 1 - F_0(c_i) = \int_{c_i}^{\infty} f_0(x)dx, \qquad q_i = 1 - F_1(c_i) = \int_{c_i}^{\infty} f_1(x)dx$$

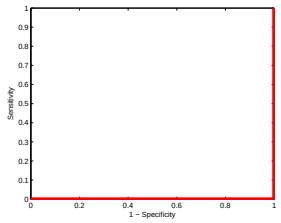
Point $(1, 1)$ - all subject are classified to be from $\mathcal{G}_1$

Point $(0, 0)$ - all subject are classified to be from $\mathcal{G}_0$

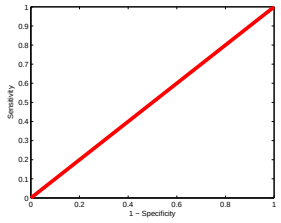
Extreme cases



A perfectly accurate test because sensitivity is 1.0 when 1-specificity is 0.0



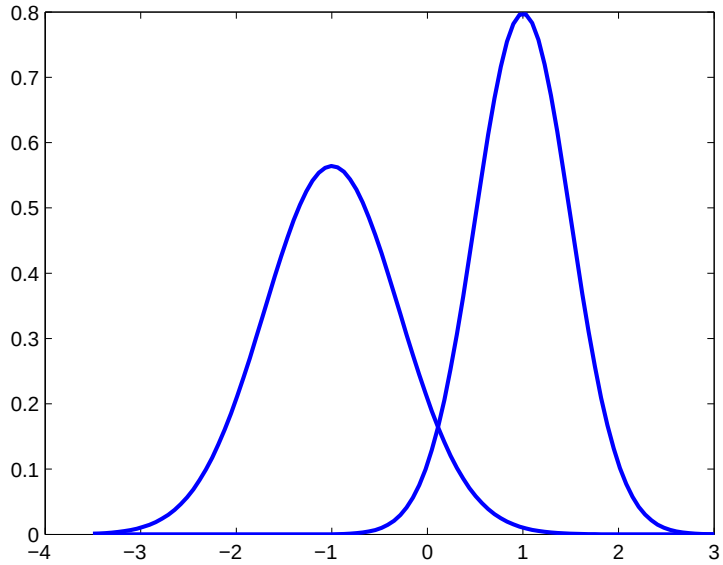
A perfectly inaccurate test, subjects with the condition are located incorrectly as negative and subjects without condition are located incorrectly as positive

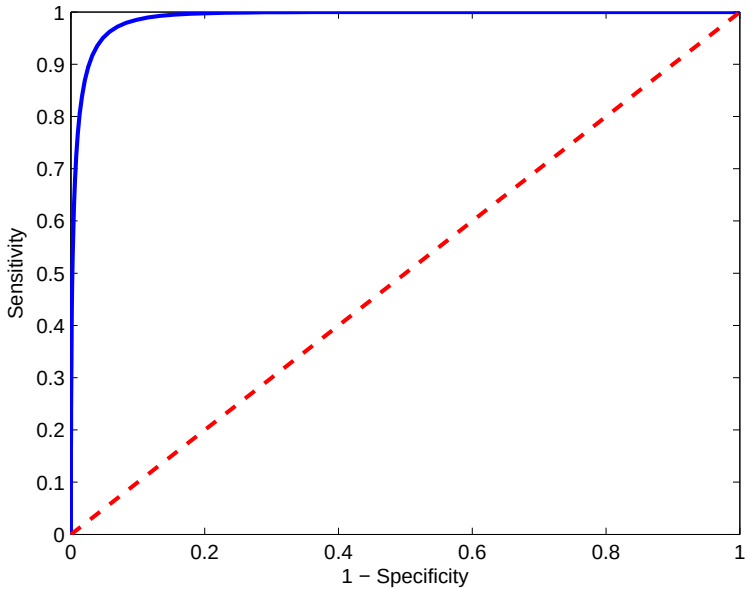


A diagonal – chance diagonal. The test is not usable for separation of the subjects.

Diagnostic tests with ROC curves above the chance diagonal have at least some ability do discriminate between subjects with and without condition.

ROC curve close to the perfectly accurate one





$$f_0(x) = \frac{1}{\sqrt{\pi}} e^{-(x+1)^2}, \qquad f_1(x) = \frac{1}{0.5\sqrt{2\pi}} e^{-\frac{(x-1)^2}{2 \cdot 0.5^2}}$$

ROC Measure

Area under curve

- The most common used global index of diagnostic accuracy is the *area under the ROC curve – AUC*.
- The area under the ROC curve is the probability that a pair of individuals known to be from different groups will be correctly classified

$$AUC = \int_0^1 ROC(p)dp.$$

- Values of AUC close to 1 indicate that the test has high diagnostic accuracy.
- Known also as *Gini coefficient*

$$Gini = 2\ AUC - 1$$

Nonparametric estimates of ROC curve

Suppose that independent samples $X_{01},\ldots,X_{0n_0}$ from $\mathcal{G}_0$ and $X_{11},\ldots,X_{1n_1}$ from $\mathcal{G}_1$.

Empirical ROC curve

F_0 and F_1 are replaced by their *empirical* cumulative distribution functions

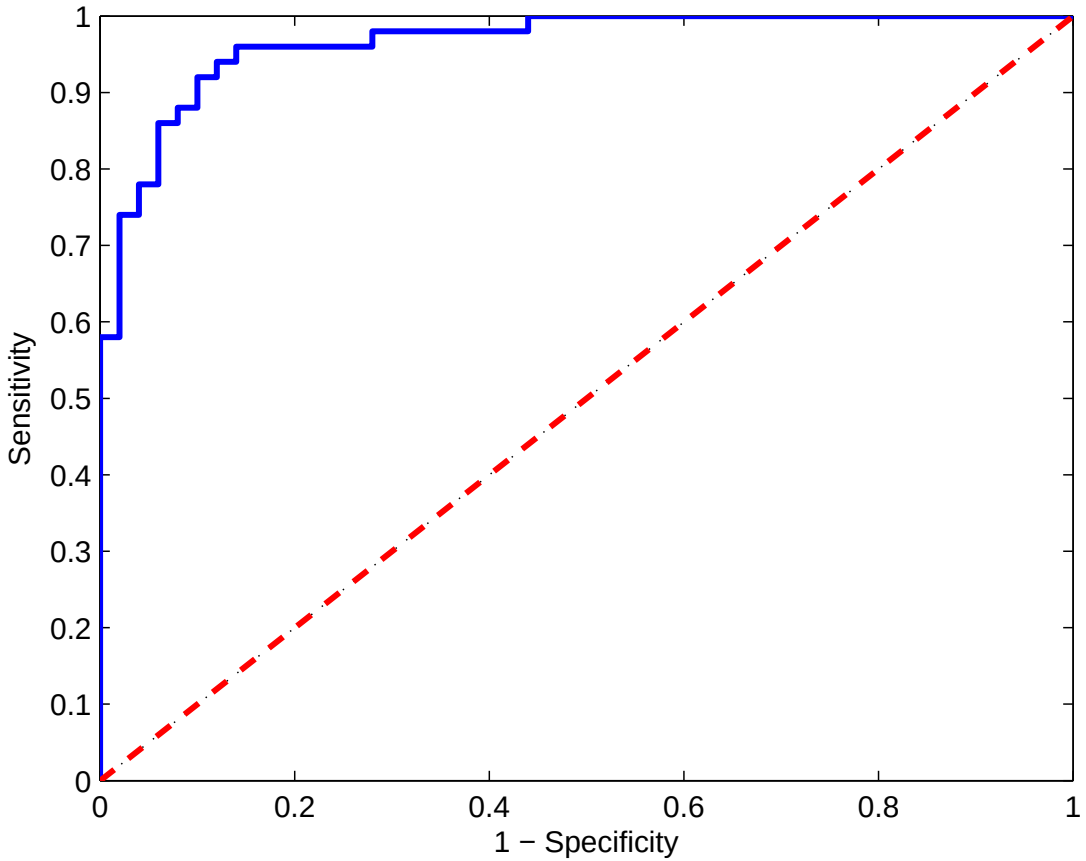
$$\hat{F}_0(x) = \frac{1}{n_0}\sum_{i=1}^{n_0} I(X_{0i} \le x), \quad \hat{F}_1(x) = \frac{1}{n_1}\sum_{i=1}^{n_1} I(X_{1i} \le x)$$

Hence the empirical estimator

$$\hat{R}(p) = 1 - \hat{F}_1(\hat{F}_0^{-1}(1-p)),\ 0 \le p \le 1$$

is the nonparametric estimator of $R(p)$.

Empirical ROC curve



Kernel estimate of ROC curve

Lloyd (1999) proposed a smooth estimator based on the technology of kernel smoothing.

The kernel estimators of F_0 and F_1 are

$$\widehat{F}_0(x) = \frac{1}{n_0}\sum_{i=1}^{n_0} W\left(\frac{x - X_{0i}}{h_0}\right), \quad \widehat{F}_1(x) = \frac{1}{n_1}\sum_{i=1}^{n_1} W\left(\frac{x - X_{1i}}{h_1}\right)$$

where

$$W(x) = \int_{-1}^{x} K(t)dt.$$

The corresponding estimator of $R(p)$ is

$$\widehat{R}(p) = 1 - \widehat{F}_1(\widehat{F}_0^{-1}(1-p)),\ 0 \le p \le 1.$$

Kernel estimate of CDF

Let $X_1,\ldots,X_n$ be independent real random variables each having the same cumulative distribution F. Assume $F \in C^{k_0}$, k_0 – a positive integer.

The kernel estimate of a cumulative distribution function F

$$\widehat{F}_{h,K}(x) = \frac{1}{n}\sum_{i=1}^{n} W\left(\frac{x - X_i}{h}\right), \qquad W(x) = \int_{-1}^{x} K(t)dt \qquad (1)$$

- K – a *kernel*, a nonnegative symmetric function, supported on $[-1, 1]$, integrated to unity
  Epanechnik kernel: $K(x) = \frac{3}{4}(1 - x^2)I_{[-1,1]}$
  quartic kernel: $K(x) = \frac{15}{16}(1 - x^2)^2 I_{[-1,1]}$
- h – a smoothing parameter – *banwidth*
  $(h = h(n) > 0, \lim_{n\to\infty} h = 0, \lim_{n\to\infty} nh = \infty)$

Bandwidth selection

Mean Integrated Square Error

$$\mathrm{MISE}(\widehat{F}_{h,K}) = \mathrm{E} \int (\widehat{F}_{h,K}(x) - F(x))^2 dx$$

The leading term of MISE (Bowman, A., Hall, P., Prvan, T. (1998))

$$\overline{\mathrm{MISE}}(\widehat{F}_{h,K}) = \underbrace{\frac{1}{n} \int F(x)(1 - F(x))dx - q_1 \frac{h}{n}}_{\overline{var}(\widehat{F}_{h,K})} + \underbrace{q_2 h^4}_{\overline{bias}^2(\widehat{F}_{h,K})},$$

$$q_1 = \int_{-1}^{1} W(x)(1 - W(x))dx > 0, \qquad q_2 = \frac{\beta_2^2}{4} \int (F^{(2)}(x))^2 dx.$$

Optimal bandwidth minimizing $\overline{\mathrm{MISE}}$

$$h_{opt}^F = n^{-1/3} \left(\frac{q_1}{4q_2} \right)^{1/3}$$

Methods for estimation of the optimal bandwidth:

- Terrell and Scott (1985), Terrell (1990) – maximal smoothing principle
- Sarda (1993) – a cross-validation method
- Altman and Léger (1995), Zhou and Harezlak (2002) – a method of the reference (Gaussian) density
- Lloyd and Yong (1999) – a more complex selection of bandwidth, procedure based on two-stage plug-in method
- Peng and Zhou (2004) – a method is based on local linear smoothing
- Horová and Zelinka (2007) – an iterative method

Boundary Effects

Assumptions:

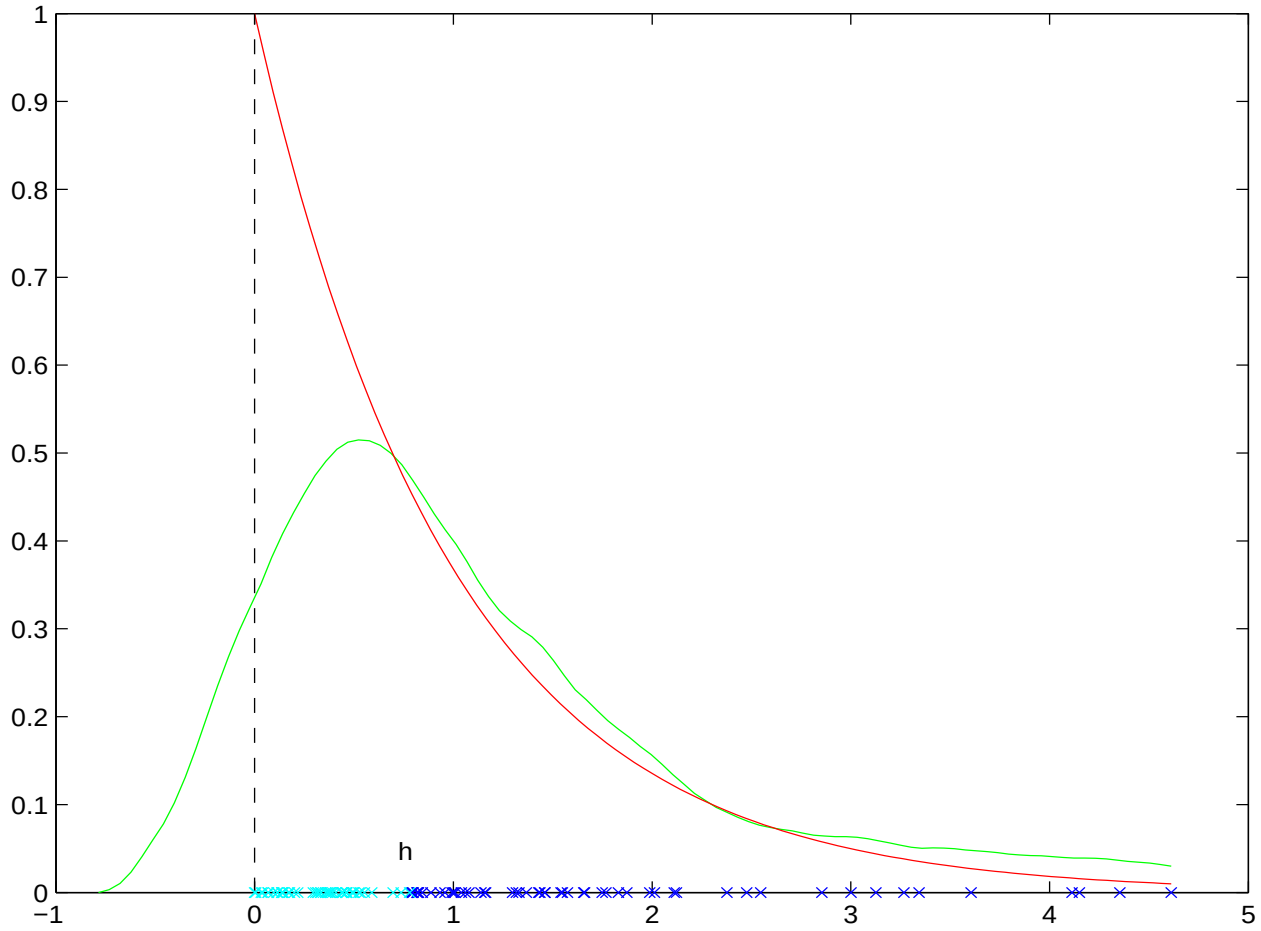
- $X_i,\ i = 1, \ldots, n$ are nonnegative
- the distribution function F has a support $[0, \infty)$
- $f(0) \neq 0$

Boundary effects arise by estimates in points "near" the left boundary, it is for $x \in [0, h]$.

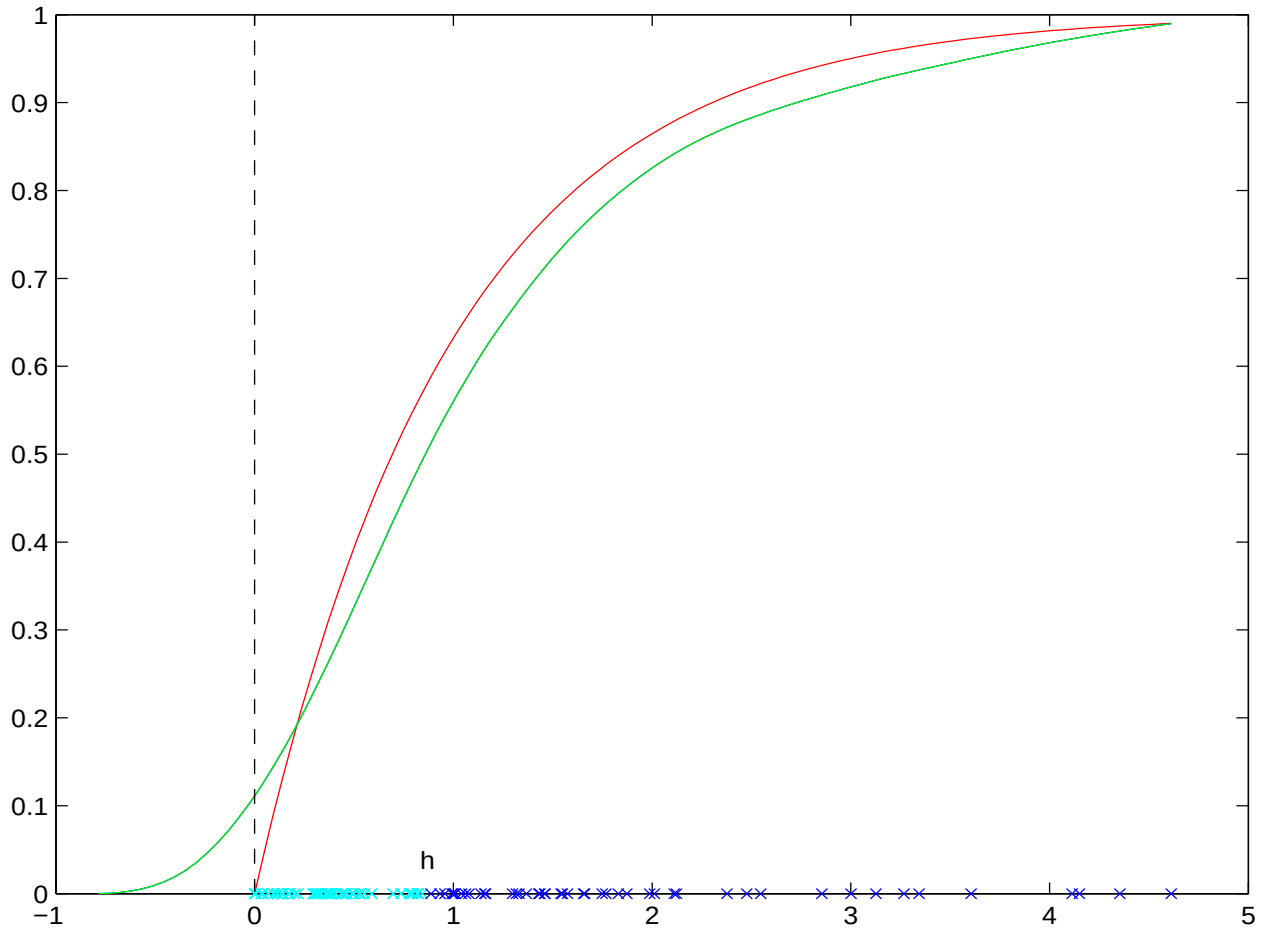
In next, we will write

$$x = ch,\ 0 \leq c \leq 1.$$

$X \sim Exp(1)$ – the kernel estimate of f $(n = 100,\ h^f_{opt} = 0.786)$



$X \sim Exp(1)$ – the kernel estimate of F $(n = 100,\ h^F_{opt} = 0.8479)$



The *Bias* of $\widehat{F}_{h,K}(x)$ in $x = ch$,

- "near" the left boundary $(0 \leq c < 1)$:

$$\mathrm{E}(\widehat{F}_{h,K}(x)) - F(x) = hf(0) \int_{-1}^{-c} W(t)dt + h^2 f^{(1)}(0) \left\{ \frac{c^2}{2} + c \int_{-1}^{-c} W(t)dt - \int_{-1}^{c} tW(t)dt \right\} + o(h^2)$$

- interior points $(c \geq 1)$:

$$\mathrm{E}(\widehat{F}_{h,K}(x)) - F(x) = \frac{h^2}{2} f^{(1)}(0) \int_{-1}^{1} tW(t)dt + o(h^2)$$

Possible solutions

- *boundary kernels* – estimators could be negative, some remedies have been proposed
- *pseudo-data* – generating some extra data nearby the boundary and then combining them with the original data
- *data transformation*
  (a) a transformation is selected from a parametric family,
  (b) a kernel estimator is applied to transformed data,
  (c) estimated values are converted by an inverse formula
- *reflection method* – reflecting the data and applying the classical kernel estimator

$$\widehat{F}_{h,K}(x) = \frac{1}{n}\sum_{i=1}^{n}\left\{W\left(\frac{x - X_i}{h}\right) - W\left(-\frac{x + X_i}{h}\right)\right\} \qquad (2)$$

Proposed estimator

Generalized reflection method

- the *density* case – Zhang et al. (1999), Karunamuni and Alberts (2005)
- the *CDF* case

$$\widetilde{F}_{h,K}(x) = \frac{1}{n}\sum_{i=1}^{n}\left\{W\left(\frac{x-\widehat{g}_c(X_i)}{h}\right) - W\left(-\frac{x+\widehat{g}_c(X_i)}{h}\right)\right\}$$

where

$$\widehat{g}_c(y) = \widehat{A}_c^2 y^3 + \frac{1}{2}\widehat{A}_c y^2 + y,$$

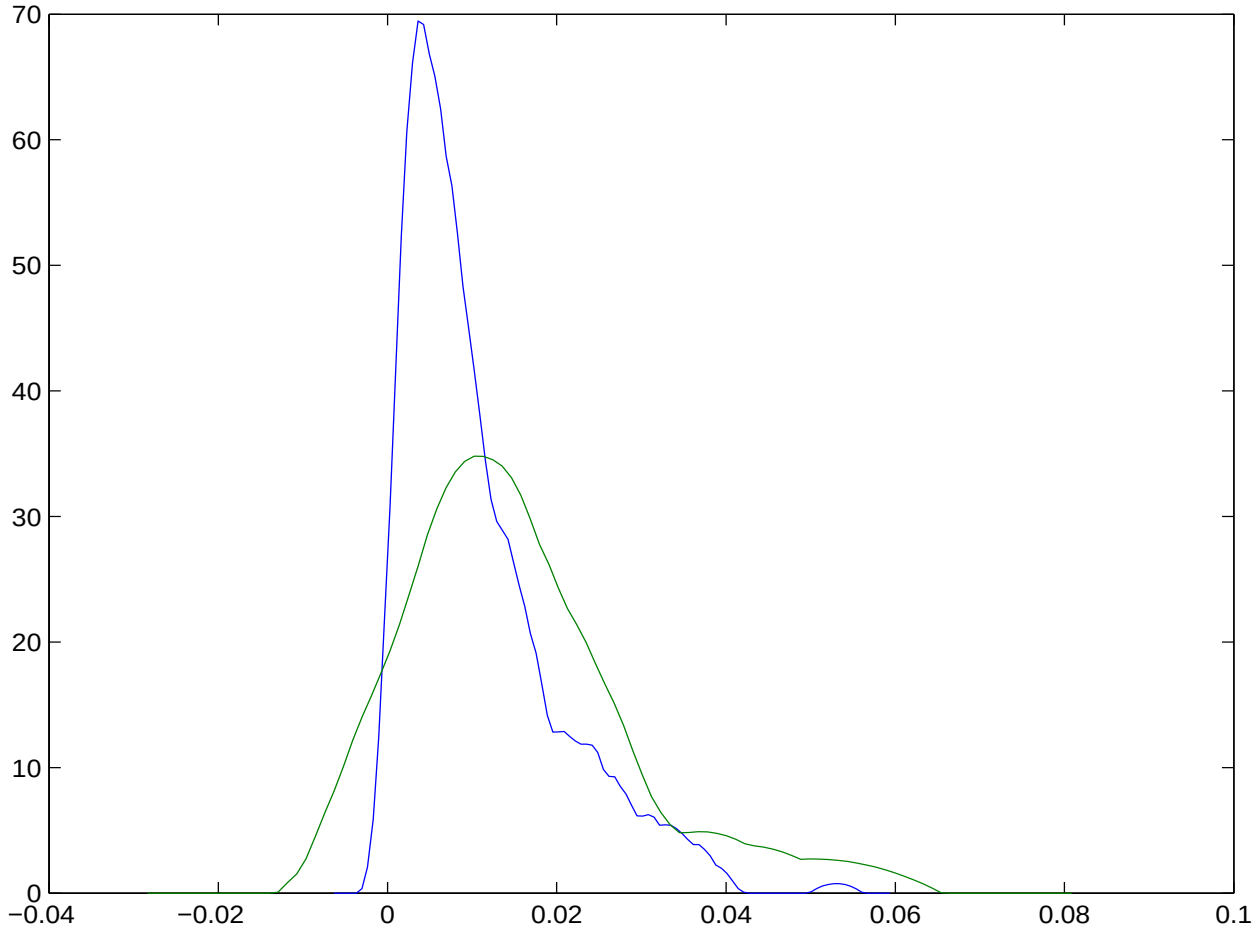
for more see Koláček and Karunamuni (2009).

Real data

Consumer loans data

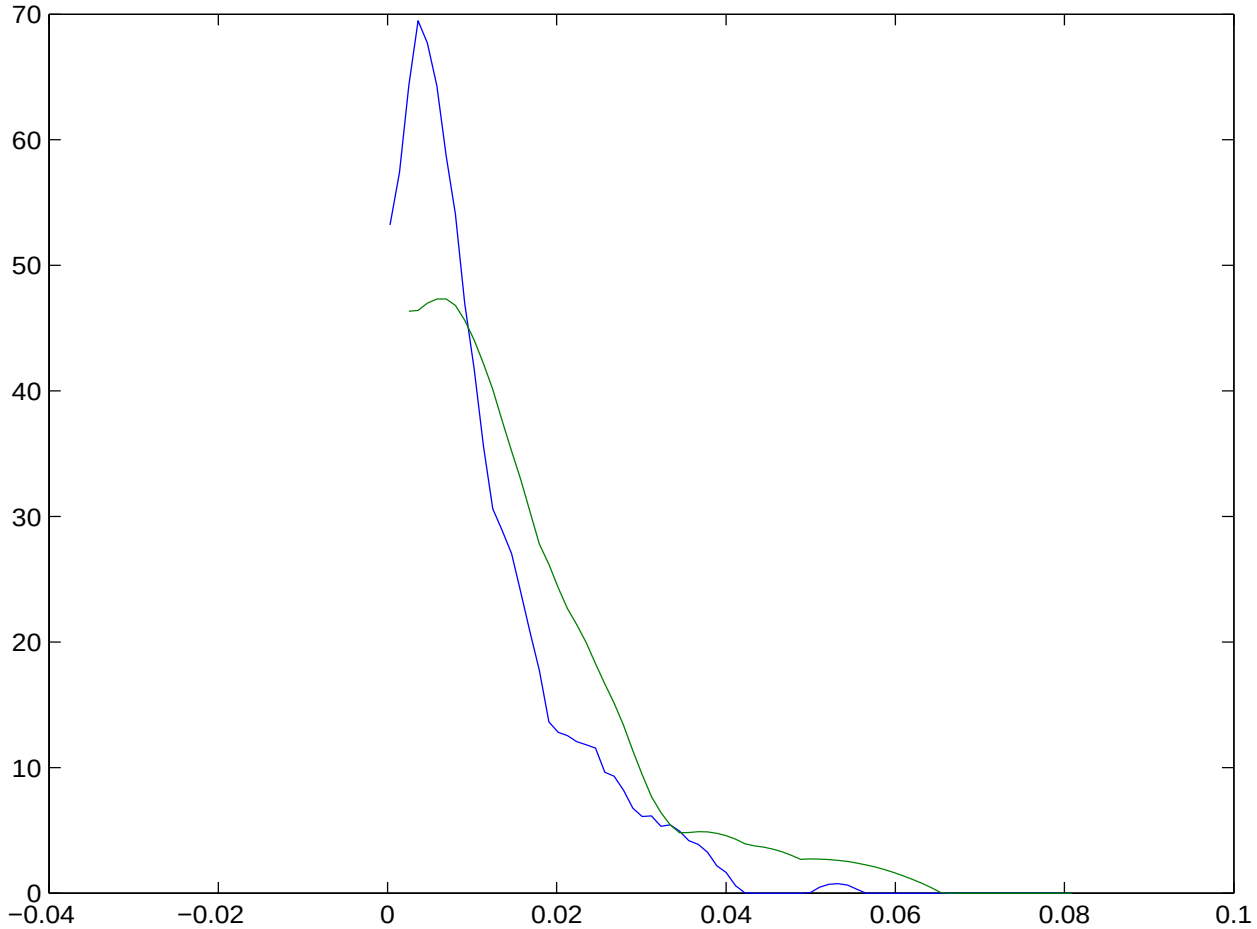
- The use of some (not specified) scoring function for predicting the likelihood of repayment of a client.
- We are interested in determining which clients are able to repay their loans.
- A test set: 332 clients – 309 have repaid their loans (group $\mathcal{G}_0$) and 23 had problems with payments or did not pay (group $\mathcal{G}_1$).
- We use the ROC curve to assess the discrimination power of given scoring function.

The estimate of $f_0(x)$ ($\hat{h}_{opt}^{f_0} = 0.0032$) and $f_1(x)$ ($\hat{h}_{opt}^{f_1} = 0.0153$) with boundary effects



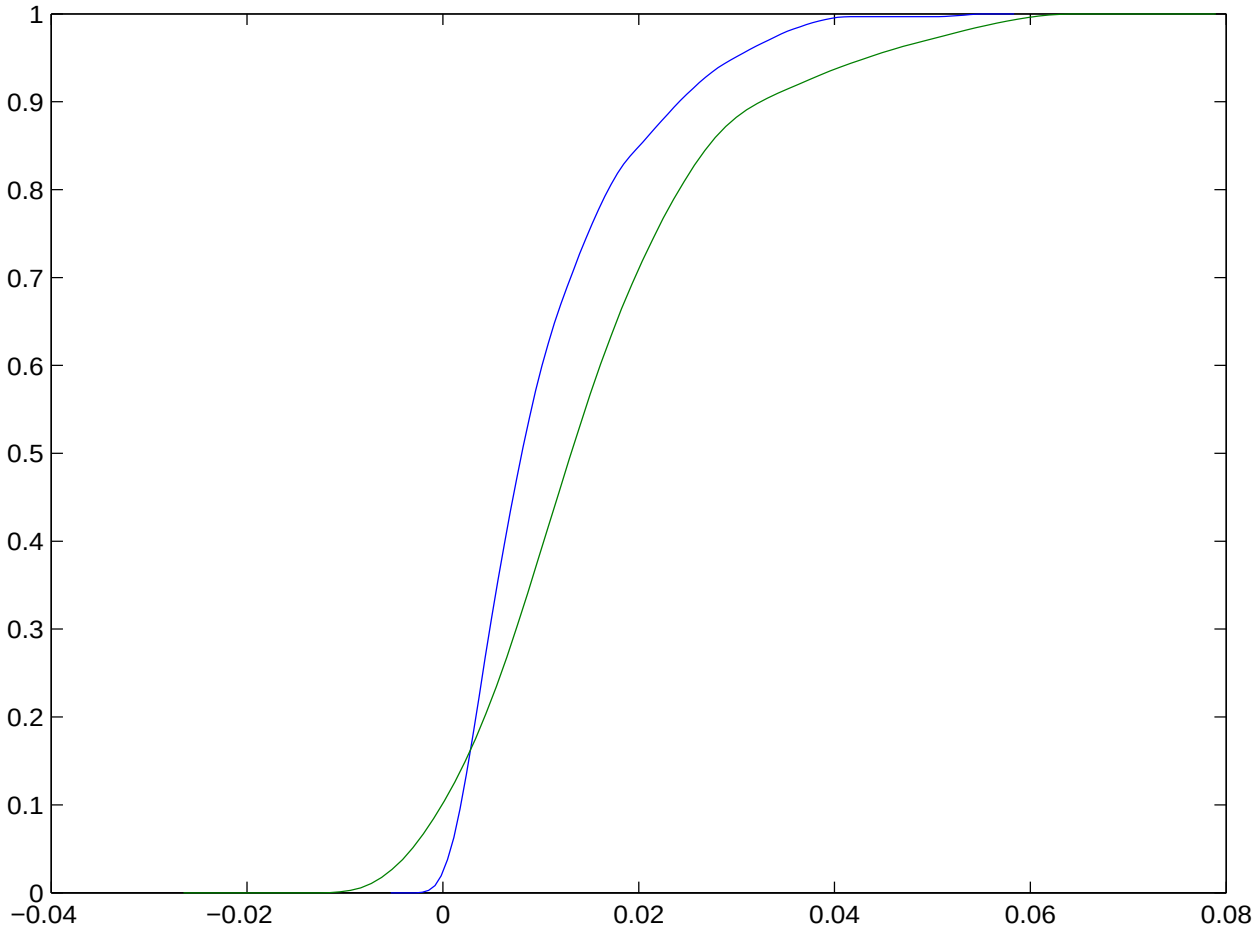
– $\widehat{f_0}$, – $\widehat{f_1}$

The estimate of $f_0(x)$ ($\hat{h}_{opt}^{f_0} = 0.0032$) and $f_1(x)$ ($\hat{h}_{opt}^{f_1} = 0.0153$) with NO boundary effects



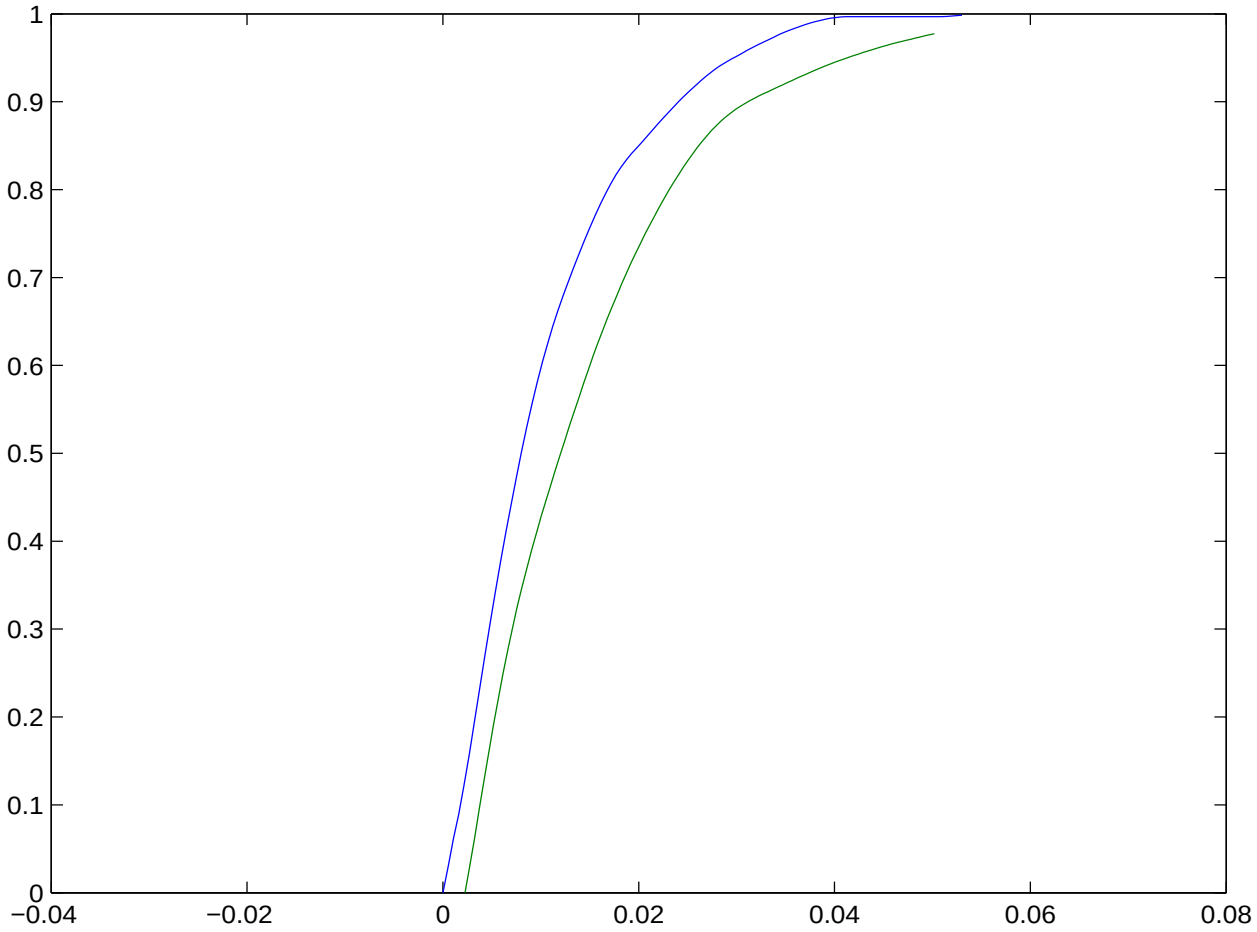
– $\widehat{f_0}$, – $\widehat{f_1}$

The estimate of $F_0(x)$ ($\hat{h}_{opt}^{F_0} = 0.0068$) and $F_1(x)$ ($\hat{h}_{opt}^{F_1} = 0.0286$) with boundary effects



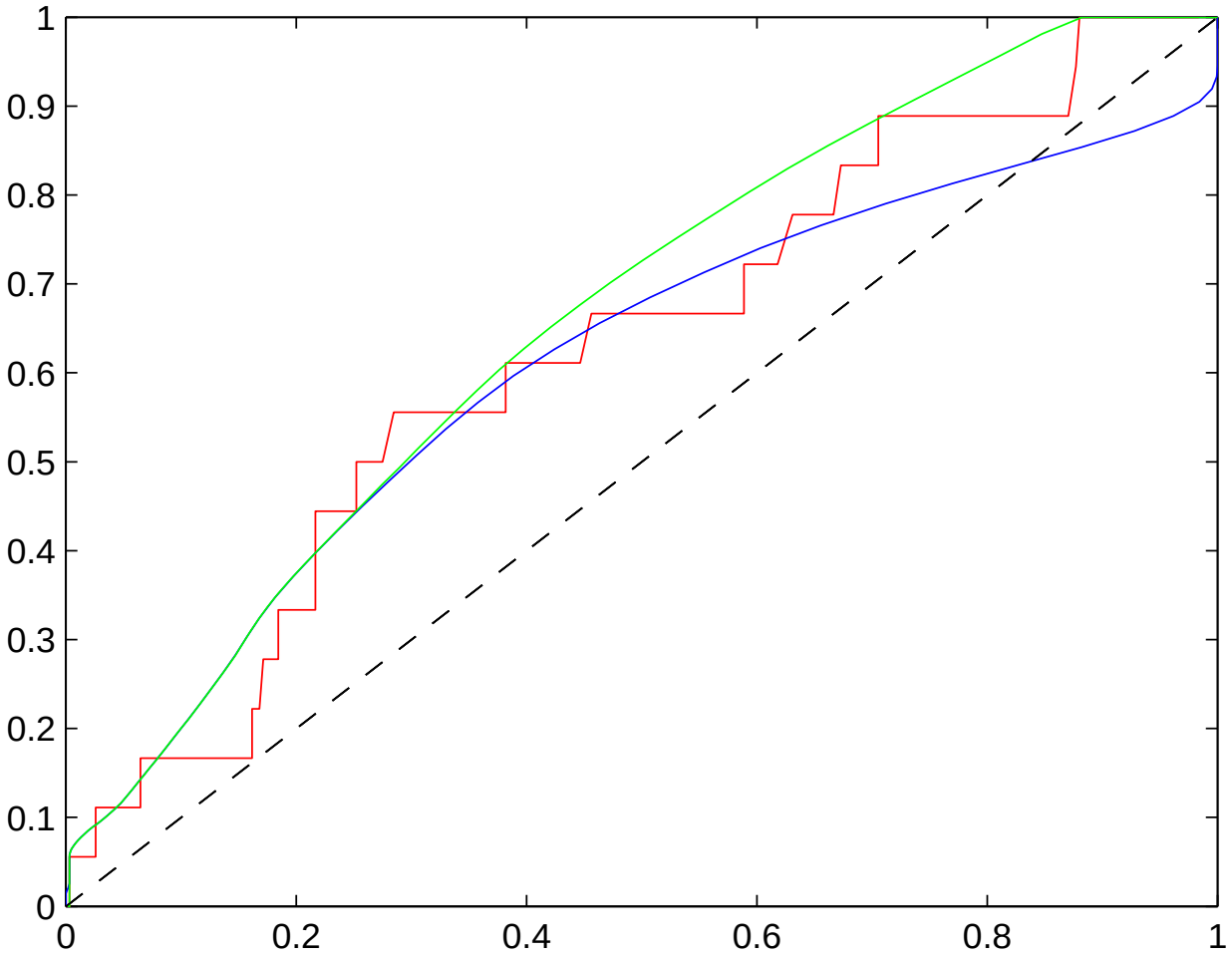
– $\widehat{F}_0$, – $\widehat{F}_1$

The estimate of $F_0(x)$ ($\hat{h}_{opt}^{F_0} = 0.0068$) and $F_1(x)$ ($\hat{h}_{opt}^{F_1} = 0.0286$) with NO boundary effects



– $\widehat{F}_0$, – $\widehat{F}_1$

The estimate of ROC



– empirical, – boundary effects, – proposed

Some statistics or measures

	K–S	Gini	AUC
Empirical ROC	0.2708	0.2175	0.6088
Estimate with boundary effects	0.2088	0.2117	0.6059
Estimate with NO boundary effects	0.2303	0.3223	0.6612

References

[1] Azzalini, A.: *A note on the estimation of a distribution function and quantiles by a kernel method*. Biometrika, 68, No 1, pp. 326–328, 1981.

[2] Bowman, A., Hall, P., Prvan, T.: *Bandwidth selection for the smoothing of distribution functions*. Biometrika, 85, No 4, pp. 799–808, 1998.

[3] Horová I., Zelinka J.: *Contribution to the bandwidth choice for kernel density estimates* Computational Statistics, 22, No. 1, pp. 31–47, 2007

[4] Lloyd, C.J., Zhou Yong: *Kernel estimators of the ROC curve are better than empirical*. Statistics and Prob. Letters 44, pp. 221–228, 1999.

[5] Karunamuni, R.J., Alberts T.: *On boundary correction in kernel density estimation*. Statistical Methodology 2, pp. 191–212, 2005.

[6] Koláček, J., Karunamuni, R.J. *On boundary correction in kernel estimation of ROC curves*. Austrian Journal of Statistics, 2009, Vol. 38, No. 1, pp. 17–32.

[7] Terrell, G. R.: *The maximal smoothing principle in density estimation*. Journal of the American Statistical Association. Vol. 85, No. 410, pp. 440-447, 1990.

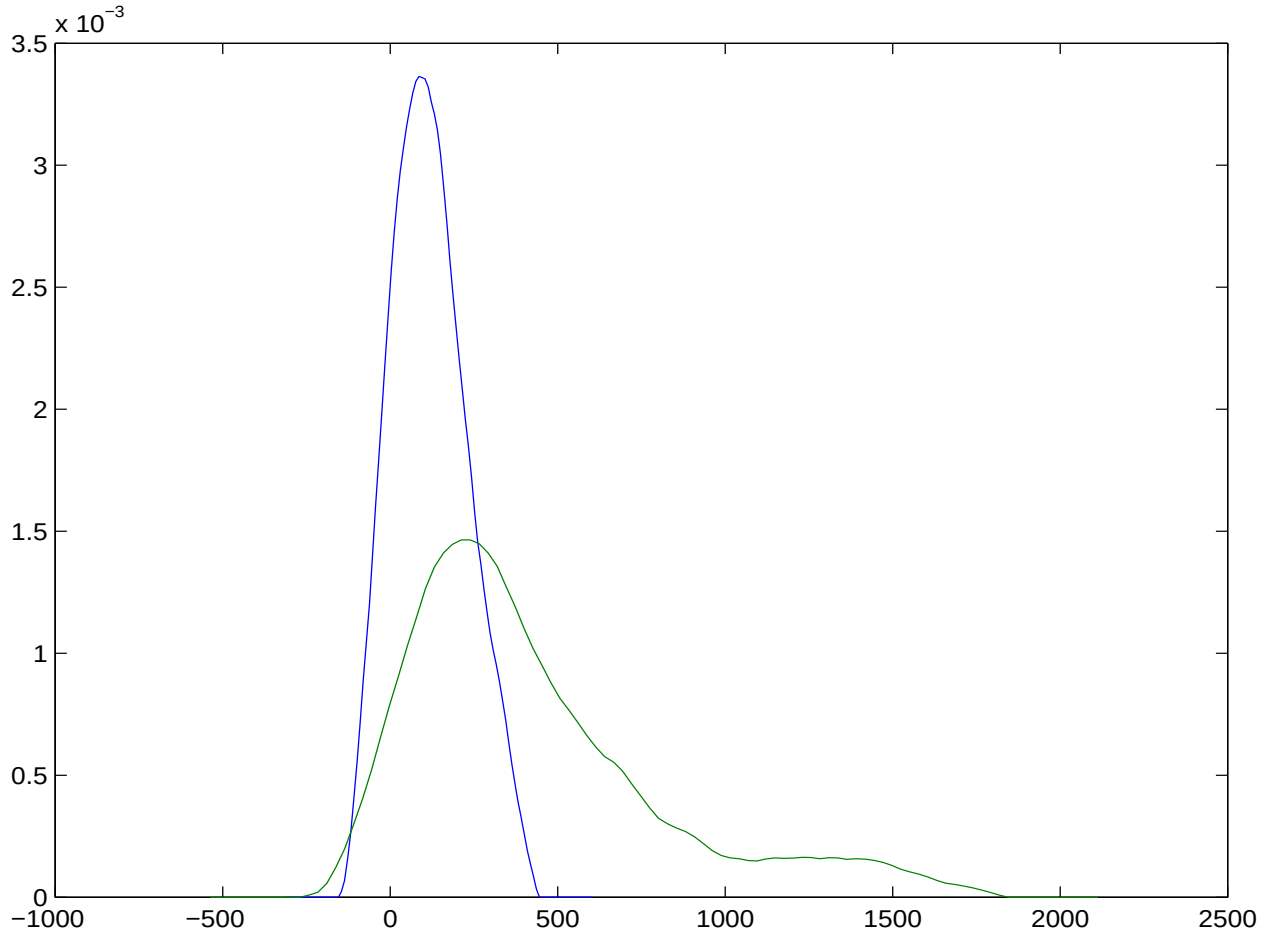
[8] Wand, I.P. and Jones, M.C.: *Kernel smoothing*. Chapman & Hall, London, 1995.

[9] Zhang, S., Karunamuni, R.J., Jones, M.C.: *An improved estimator of the density function at the boundary*. Journal of the Amer. Stat. Assoc., 448, pp. 1231–1241, 1999.

Head trauma data

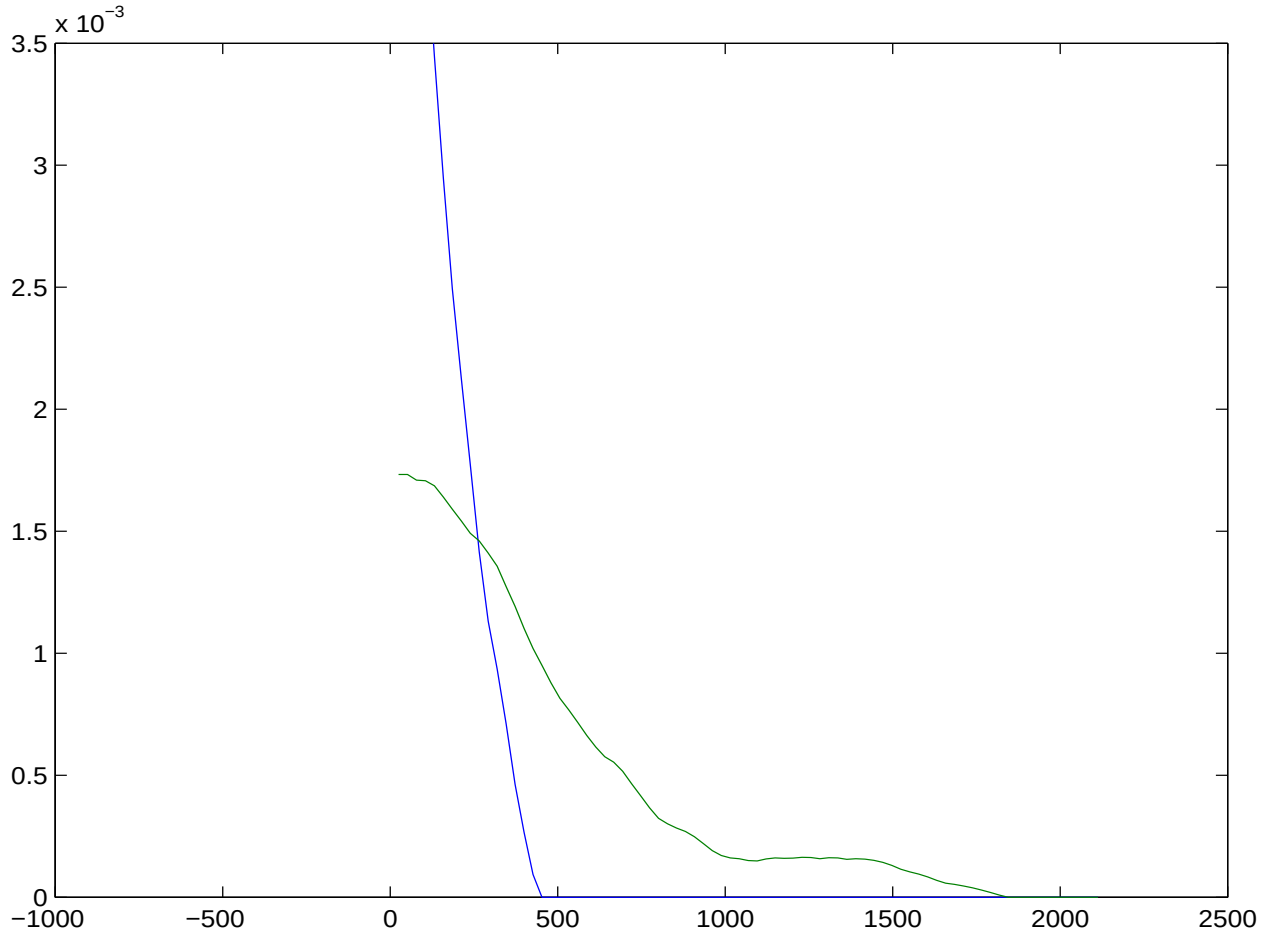
- The use of cerebrospinal fluid CK–BB (creative kinase – BB) isoenzyme measured within 24 hours of injury for predicting the outcome of severe head trauma.
- We are interested in determining which patients have a poor outcome after suffering a severe head trauma.
- 60 patients: 19 had moderate to full recovery and 41 eventually had poor or not recovery.
- We use the ROC curve to assess the discrimination between patients with and without a poor outcome.
- We want to know if the CK–BB isoenzyme is a good predictor of the outcome.

The estimate of $f_0(x)$ ($\hat{h}^{f_0}_{opt} = 145.7135$) and $f_1(x)$ ($\hat{h}^{f_1}_{opt} = 253.6472$) with boundary effects.



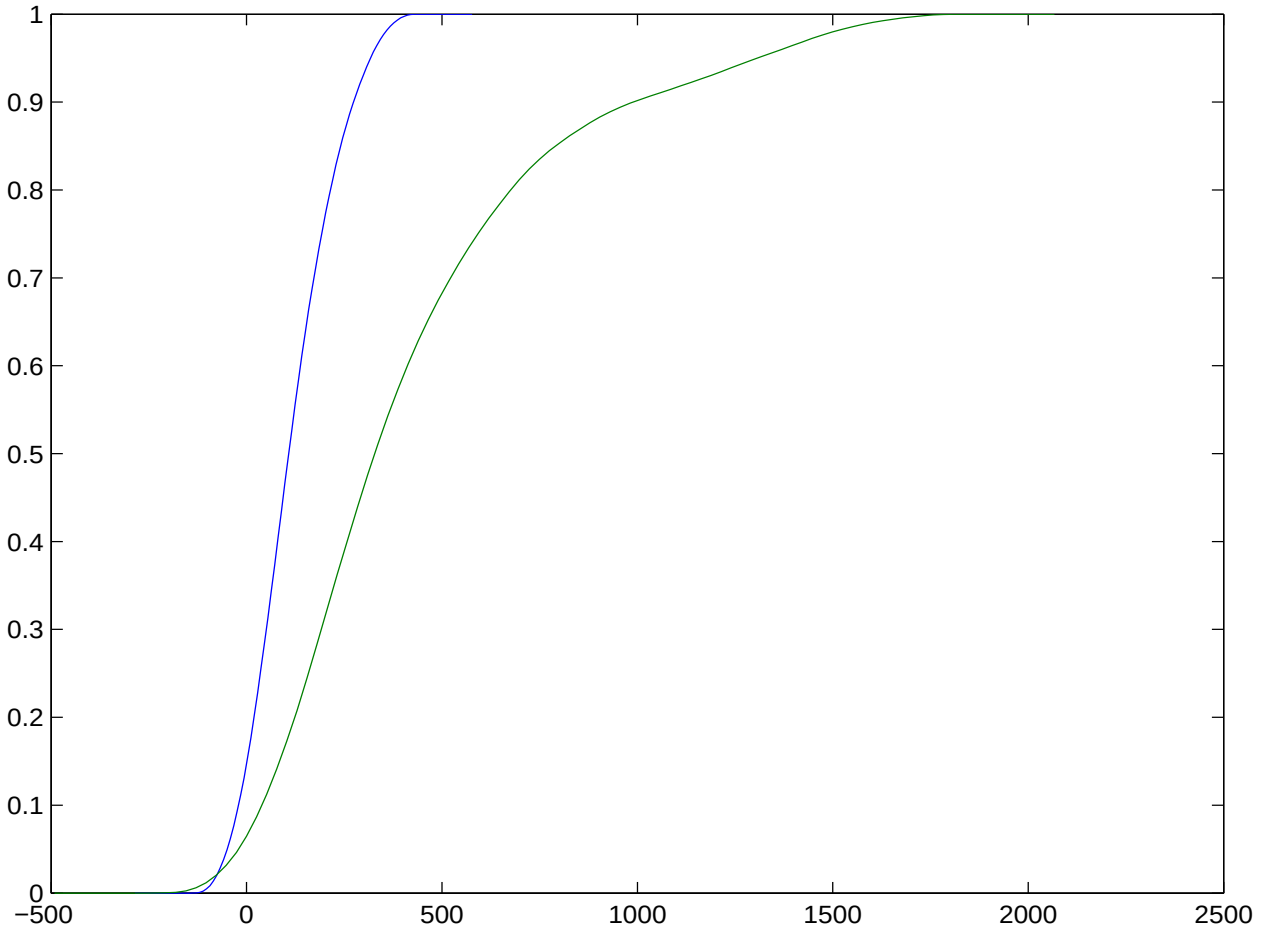
– $\widehat{f_0}$, – $\widehat{f_1}$

The estimate of $f_0(x)$ ($\hat{h}^{f_0}_{opt} = 145.7135$) and $f_1(x)$ ($\hat{h}^{f_1}_{opt} = 253.6472$) with NO boundary effects.



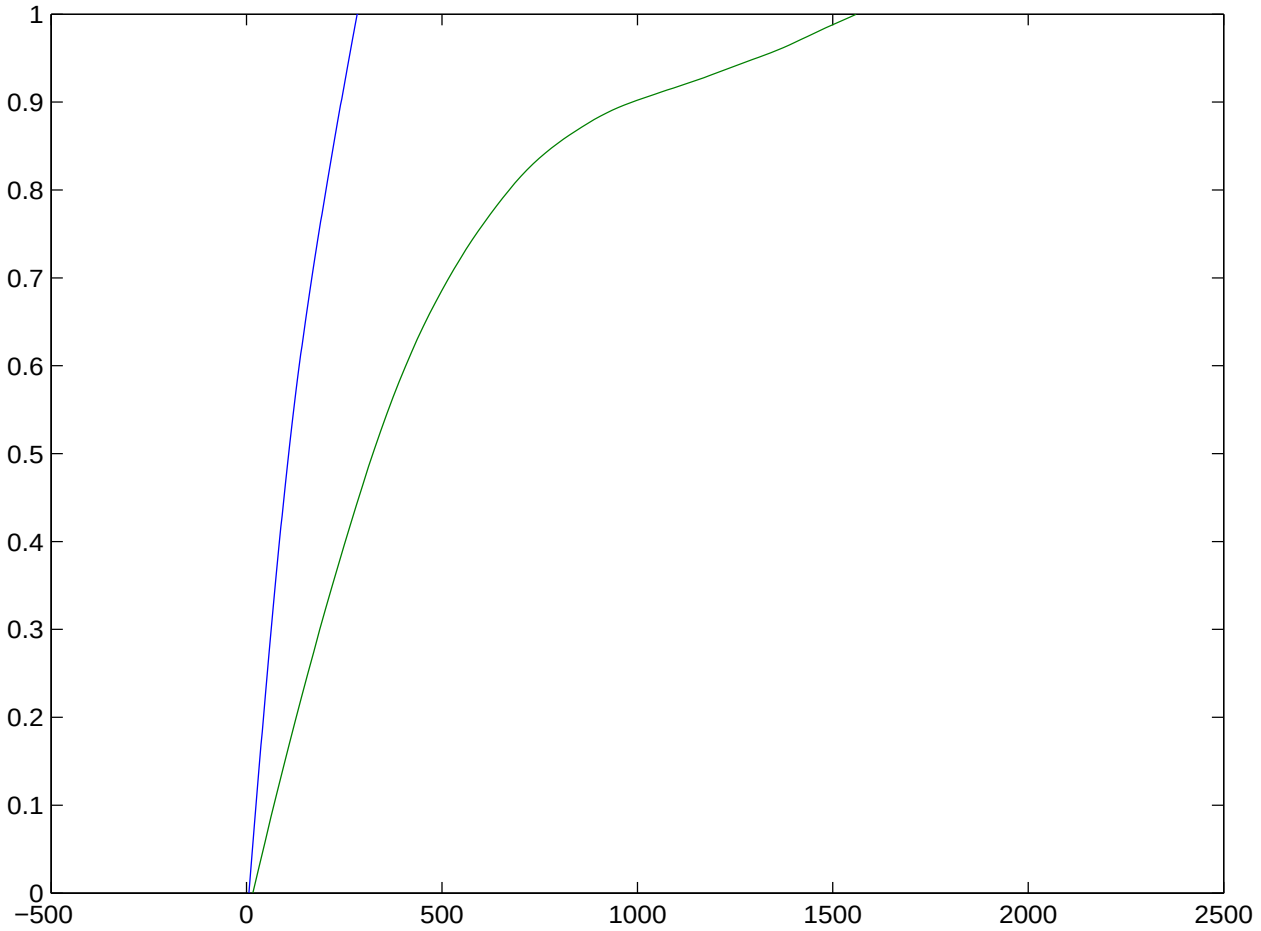
– $\widehat{f_0}$, – $\widehat{f_1}$

The estimate of $F_0(x)$ ($\hat{h}_{opt}^{F_0} = 158.6975$) and $F_1(x)$ ($\hat{h}_{opt}^{F_1} = 276.5697$) with boundary effects.



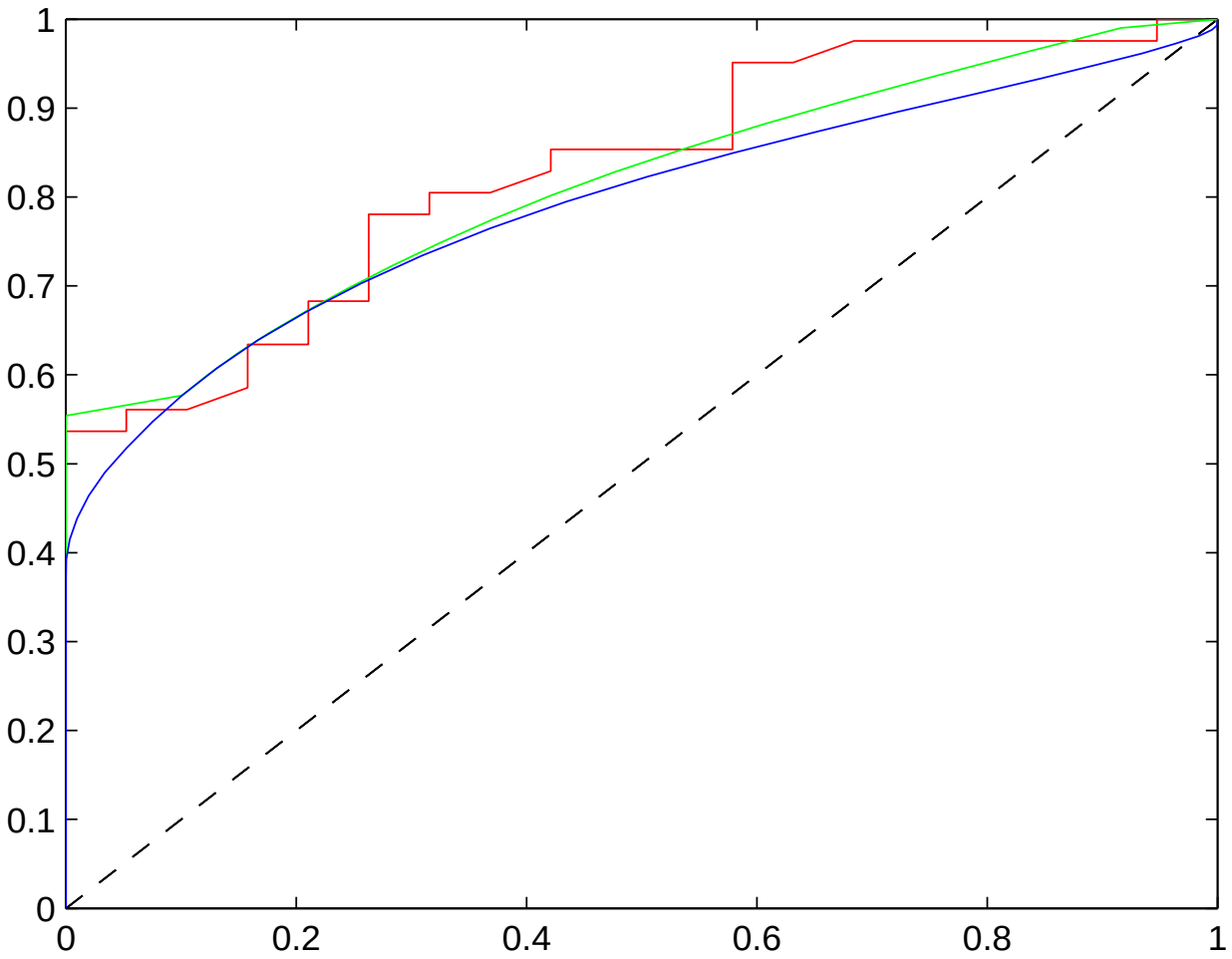
– $\widehat{F}_0$, – $\widehat{F}_1$

The estimate of $F_0(x)$ ($\hat{h}_{opt}^{F_0} = 158.6975$) and $F_1(x)$ ($\hat{h}_{opt}^{F_1} = 276.5697$) with NO boundary effects.



– $\widehat{F}_0$, – $\widehat{F}_1$

The estimate of ROC



– empirical, – boundary effects, – proposed

Some statistics or measures

	K–S	Gini	AUC
Empirical ROC	0.5366	0.6573	0.8286
Estimate with boundary effects	0.4761	0.5802	0.7901
Estimate with NO boundary effects	0.5541	0.6239	0.8119