ACTA UNIVERSITATIS LODZIENSIS
FOLIA OECONOMICA 206, 2007

*Agnieszka Rossa**

ON SOME INFERENTIAL PROCEDURES FOR RECEIVER OPERATING CHARACTERISTIC CURVES

Abstract. In the paper two significance tests for receiver operating characteristic curves (*ROC*) are proposed. Both tests use an asymptotic χ^2 distribution of the test statistics.

Key words: *ROC* curve, goodness-of-fit tests.

1. NOTATION

Suppose a diagnostic test is used to detect the presence of a disease. Let X be a random variable representing the test result. Denote by π_0, π_1 a disease group and a control group, respectively. We will assume that an individual comes from π_0, if X exceeds a fixed threshold X, say, and from π_1, otherwise. Let

$$C = X \mid \pi_0, \quad Z = X \mid \pi_1,$$

variables C and Z represent the diagnostic variable X in the respective populations π_0 and π_1. Let F and G be cumulative distribution functions (CDF) of C and Z, respectively.

2. THE *ROC* CURVE

The *ROC* curve (receiver operating characteristic curve, see: Green, Swets 1966; Lloyd 1998) is a plot of $p_0 = 1 - F(x)$ against $p_1 = 1 - G(x)$ as x varies over the support of X. In biomedical context p_0 is termed "sensitivity", and $1 - p_1$ is termed "specificity".

* Ph.D., Associate Professor, Department of Statistical Methods, University of Łódź.

In statistical terms, the *ROC* curve displays the trade-off between "power" and "size" of the test with a rejection region $\{X > x\}$ as x is varied. The power $P(X > x \mid \pi_0)$ is the probability of a true positive diagnosis, and the size $P(X > x \mid \pi_1)$ is the probability of false positive diagnosis. If X is continuous, then *ROC* depends on F, G *via* the formula

$$ROC(v) = 1 - F(G^{-1}(1 - v)), \quad v \in [0, 1] \tag{1}$$

Indeed, let us denote $v = 1 - G(x)$ then $G(x) = 1 - v$ and $x(v) = G^{-1}(1 - v)$. Thus, for $v \in [0, 1]$ we receive $ROC(v) = 1 - F(x(v))$ what leads directly to (1).

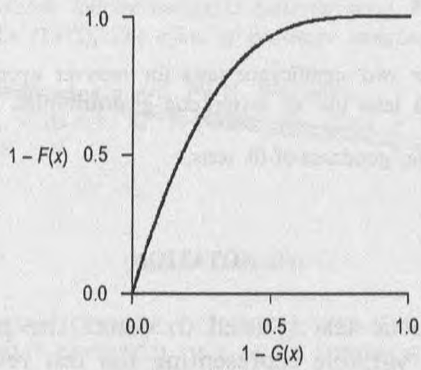

Fig. 1. An example of a *ROC* curve

Source: own elaboration.

Estimation of $ROC(v)$ is usually based on replacing F and G by their empirical counterparts F_m and G_n defined as follows

$$F_m(x) = \frac{1}{m} \sum_{i=1}^{m} \mathbf{1}(C_i \leqslant x) \tag{2}$$

$$G_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}(Z_i \leqslant x) \tag{3}$$

where $\mathbf{1}(A)$ denotes a characteristic function of an event A, and

$$C_1, C_2, \ldots, C_m, \quad Z_1, Z_2, \ldots, Z_n \tag{4}$$

are two independent random samples drawn from populations π_0 and π_1, respectively.

The *ROC* curve summarizes the separation between the distributions F and G in two populations π_0 and π_1. The higher is the *ROC* curve, the greater is the prediction accuracy of X. If the *ROC* curve of a variable

X lies on the diagonal of the unit space then there are no difference in distributions of X in the populations π_0 and π_1. This concept constitutes a background for two χ^2 goodness-of-fit tests discussed in details in the next section.

3. THE GOODNESS-OF-FIT TESTS FOR *ROC* CURVES

We will consider the problem of testing two null hypotheses. The first one states, that the *ROC* curve lies on the diagonal

$$H_0: \forall_{\nu \in [0,1]} \; ROC(\nu) = \nu \tag{5}$$

The second null hypothesis assumes that *ROC* functions for two diagnostic variables X_A and X_B, say, are equal

$$H_0^{'}: \forall_{\nu \in [0,1]} \; ROC_A(\nu) = ROC_B(\nu) \tag{6}$$

The alternatives for both cases (5) and (6) take the general form $H_1 : \sim H_0$ and $H_1^{'} : \sim H_0^{'}$, respectively.

In order to test the hypotheses (5) and (6) we will focus on a random variable $G(C)$. It easy to proof that a CDF of a variable defined as

$$W = 1 - G(C)$$

is equivalent to the *ROC* function (1). Indeed, we have for $\nu \in [0,1]$

$$P(W < \nu) = P(1 - G(C) < \nu) = P(G(C) > 1 - \nu) = P(C > G^{-1}(1 - \nu)) =$$
$$= 1 - P(C \leqslant G^{-1}(1 - v)) = 1 - F(G^{-1}(1 - v)) = ROC(v).$$

Unfortunately, it is usually impossible to observe $G(C)$ without any parametric assumptions concerning the function G. In our considerations we will replace the unknown function G with its empirical counterpart G_n defined in (3). Thus, we will consider a random variable $G_n(C)$ instead of $G(C)$.

It can be seen, that $G_n(C)$ takes values from the finite set

$$\left\{0, \frac{1}{n}, \ldots, \frac{n-1}{n}, 1\right\}.$$

To find its probability distribution function let us denote by R, F the CDF's of $G(C)$ and C, and by r and f the density functions of $G(C)$ and C, respectively. Notice, that for $x \in [0, 1]$ the following equalities hold

$$R(x) = F(G^{-1}(x)), \quad r(x) = f(G^{-1}(x))[G^{-1}(x)]' \tag{7}$$

Thus, for a fixed $i \in \{0, 1, ..., n\}$ we have

$$P\left(G_n(C) = \frac{i}{n}\right) = \binom{n}{i} \int_{-\infty}^{\infty} G^i(x)[1 - G(x)]^{n-i} f(x) \mathrm{d}x.$$

Denoting by $y = G(x)$ we obtain

$$x = G^{-1}(y), \quad \mathrm{d}x = [G^{-1}(y)]' \mathrm{d}y,$$

and

$$P\left(G_n(C) = \frac{i}{n}\right) = \binom{n}{i} \int_0^1 y^i (1 - y)^{n-i} f(G^{-1}(y))[G^{-1}(y)]' \mathrm{d}y$$

This result and (7) lead to the probability distribution function of $G_n(C)$ of the form

$$P\left(G_n(C) = \frac{i}{n}\right) = \binom{n}{i} \int_0^1 y^i (1 - y)^{n-i} r(y) \mathrm{d}y, \quad i \in \{0, 1, ..., n\} \tag{8}$$

3.1. Testing the null hypothesis H_0

If the hypothesis (5) is true then $r(y) = 1$ for $y \in [0, 1]$ and (8) reduces to

$$P\left(G_n(C) = \frac{i}{n} \mid H_0\right) = \binom{n}{i} \int_0^1 y^i (1 - y)^{n-i} \mathrm{d}y, \quad i \in \{0, 1, ..., n\} \tag{9}$$

Let $B(\alpha, \beta)$ be the beta function with parameters α, β, i.e.

$$B(\alpha, \beta) = \int_0^1 y^{\alpha-1}(1 - y)^{\beta-1} \mathrm{d}y \tag{10}$$

Denoting by $\Gamma(\cdot)$ the gamma function, the following properties are well-known

$$B(\alpha, \beta) = \Gamma(\alpha)\Gamma(\beta)/\Gamma(\alpha + \beta), \quad \Gamma(n+1) = n! \quad n \in N \tag{11}$$

From (9)–(11) we receive

$$P\left(G_n(C) = \frac{i}{n} \mid H_0\right) = \binom{n}{i}\int_0^1 y^i(1-y)^{n-i}dy = \binom{n}{i}B(i+1, n-i+1) =$$
$$= \binom{n}{i}\frac{\Gamma(i+1)\Gamma(n-i+1)}{\Gamma(n+2)} = \binom{n}{i}\frac{i!(n-i)!}{(n+1)!} = \frac{1}{n+1} \tag{12}$$

Consider the random sample

$$G_n(C_1), G_n(C_2), \ldots, G_n(C_m) \tag{13}$$

derived by a simple transformation of sequences (4). Basing the use of the sample (13), we can verify the null hypothesis (5) by means of the goodness-of-fit test statistic

$$Z_1 = \sum_{i=0}^{n} \frac{(m_i - mp)^2}{mp} \tag{14}$$

where m is a size of the sample (13), $p = 1/(n+1)$ represents the theoretical probability (12) that $G_n(C) = i/n$, and m_i stands for the empirical number of observations in (13) equal to i/n. If $m \to \infty$, then Z_1 under H_0 has an asymptotic χ^2 distribution with n degrees of freedom.

3.2. Testing the null hypothesis H_0'

Let X_A, X_B be two diagnostic variables and π_0, π_1 be two populations of individuals. Let us denote

$$C_A = X_A|\pi_0, \quad C_B = X_B|\pi_0, \quad Z_A = X_A|\pi_1, \quad Z_B = X_B|\pi_1,$$

and consider m independent copies of C_A, k independent copies of C_B, and n independent copies of both Z_A and Z_B. In other words, let us consider four independent random samples drawn from π_0, π_1

$$C_{A1}, C_{A2}, \ldots, C_{Am}, \quad C_{B1}, C_{B2}, \ldots, C_{Bk} \tag{15}$$

$$Z_{A1}, Z_{A2}, \ldots, Z_{An}, \quad Z_{B1}, Z_{B2}, \ldots, Z_{Bn} \tag{16}$$

Two sequences in (15) can be treated as independent random samples of two diagnostic variables X_A, X_B drawn from π_0 and two sequences in (16) – as independent random samples of X_A, X_B drawn from π_1. Simple transformations of (15)–(16) lead to the following two independent, random sequences

$$G_{An}(C_{A1}), \quad G_{An}(C_{A2}), ..., G_{An}(C_{Am}) \tag{17}$$

$$G_{Bn}(C_{B1}), \quad G_{Bn}(C_{B2}), ..., G_{Bn}(C_{Bk}) \tag{18}$$

where G_{An}, G_{Bn} represent empirical distribution functions derived according to the formula (3) from $Z_{A1}, Z_{A2}, ..., Z_{An}$ and $Z_{B1}, Z_{B2}, ..., Z_{Bn}$ respectively.

If the null hypothesis (6) is true, then the variables $G_{An}(C_{Ai})$, $i = 1, 2, ..., m$ and $G_{Bn}(C_{Aj})$, $j = 1, 2, ..., k$ in (17)–(18) are identically distributed with probability distribution functions expressed by (8). Further we will consider samples (17)–(18) grouped into the Tab. 1.

Table 1

i/n	$G_{An}(C_A)$	$G_{Bn}(C_B)$	Σ
0	N_{A0}	N_{B0}	$B_{\cdot 0}$
$1/n$	N_{A1}	N_{B1}	$N_{\cdot 1}$
...	...	...	...
$(n-1)/n$	N_{An-1}	N_{Bn-1}	$N_{\cdot n-1}$
1	N_{An}	N_{Bn}	$N_{\cdot n}$
Σ	$N_{A\cdot} = m$	$N_{B\cdot} = k$	$N = m + k$

Now, we can verify the null hypothesis (6) by means of the test statistic of the form

$$Z_2 = N \sum_{i=0}^{n} \frac{\left(N_{Ai} - \frac{N_{A\cdot} N_{\cdot i}}{N}\right)^2}{N_{A\cdot} N_{\cdot i}} + N \sum_{i=0}^{n} \frac{\left(N_{Bi} - \frac{N_{B\cdot} N_{\cdot i}}{N}\right)^2}{N_{B\cdot} N_{\cdot i}} \tag{19}$$

If $N \to \infty$, then Z_2 under H_0' has an asymptotic χ^2 distribution with n degrees of freedom.

REFERENCES

Green D. M., Swets J. A. (1966), *Signal Detection Theory and Psychophysics*, Wiley, New York.

Lloyd C. J. (1998), *Using Smoothed Receiver Operating Characteristic Curves to Summarize and Compare Diagnostic Systems*, JASA, 93, 1356–1364.

Agnieszka Rossa

O PEWNYCH TESTACH DLA KRZYWYCH OPERACYJNO-CHARAKTERYSTYCZNYCH

W pracy zaproponowano dwa testy istotności dla krzywych operacyjno-charakterystycznych (*ROC*), oparte na asymptotycznym rozkładzie statystyk testowych χ^2.