Rev. Acad. Canar. Cienc., XII (Núms. 1-2), 147-159 (2000)

ROC CURVES DETERMINATION BY NONPARAMETRIC METHODS

Saavedra, P

Departamento de Matemáticas, Universidad de Las Palmas de Gran Canaria,
35017 Las Palmas de Gran Canaria,
email: saavedra@dma.ulpgc.es

Keywords and Phrases: Nonparametric estimation , optimal bandwidth, diabetes diagnosis.

ABSTRACT

A ROC curves estimation method is proposed, based on the nonparametric estimation of the distribution function. An optimal bandwidth expression based on the mean integrated squared error is estimated by means of a crossvalidation function. A simulation study is carried out and the methodology is applied to a set of patients data with diabetic diseases.

1. INTRODUCTION

Diseases, as a general rule, alter the standard values of several numeric variables. Thus, a CD4 lymphocytes depletion account may indicate a VIH infection and high basal glucose levels may suggest a diabetic illness. When a determined pathology diagnosis entails some risks to the patient or can be economically very costly, the variables presumably affected can be the basis of an alternative diagnosis. Thus, when certain pathology causes a decrease of the usual levels of certain variable, it can be possible to work on an alternative diagnosis trial, which consists on a patient classification as sick or healthy according to

whether the variable measurement is lower or not that certain cut-off value C. The medical practitioner will then choose that value of C whose sensibility and specificity is considered more acceptable. If the basic aim of a diagnosis trial consists on rejecting or not the disease in the patient, it would be essentially interesting that the test would have a high sensitivity even if that implies a high false positive coefficient too. The ROC curve gives the trial sensitivity as a function of the false positive coefficient. Each point of the curve will be associated to a cut-off value and therefore, it is enough to choose a point to fix the cut-off value, the false positive coefficient and the diagnosis trial sensibility.

The ROC curves estimation basically depends on the estimation of the probability distributions considered in the cases and controls populations. These distributions are frequently estimated supposing the data are normally distributed or, since the data present long queues to the right, considering that the log-transformations of the data follow normal distributions. However, it is very unusual to make this type of transformations to reach normality in the usual practice. More general transformations as those of Box-Cox can be used but generally it is very difficult that the same transformation will lead to normality for both populations. On the other hand, when the number of data is scarce, the test to determinate the goodness of the fit does not really clarify if the data are normally distributed.

An alternative methodology to estimate the ROC curve is based on the estimation of the probability distribution functions for the marker considered in the disease cases and controls groups by means of nonparametric methods. The density function estimation methodology for kernel estimates introduced by Rosenblatt (1956) is widely developed (Härdle, 1991; Cao *et al.*, 1994). Azzalini (1981) considered the kernel estimation of the distribution function integrating a kernel estimate of the function density. Recently, Bowman *et al.* (1998) discussed a procedure to estimate the smoothing parameter or bandwidth. In this paper we will consider kernel estimates of the distribution function and to obtain therefore

an estimate of the ROC curves. Two log-normal distributions with specific parameters that correspond to the characteristic considered for both groups will be simulated. Graphically we will compare the theoretical ROC curve with the obtained by means of the proposed methodology and with the curve obtained under data normality hypothesis.

2. ROC CURVES

Let's consider a population whose individuals can or not have certain illness for whose diagnosis we dispose of a numeric marker X. Let $F_1(x)$ and $F_2(x)$ be the probability distribution functions of that characteristic over the sick and healthy's populations respectively. Let's suppose that the disease produces a diminishing of the normal X values. The diagnosis criteria based on X will therefore consists on determining a cut-off value C such that a subject is diagnosed as sick when $X \leq C$ and as healthy otherwise. Then, the sensitivity and the positive false coefficients of the diagnosis trial are defined as $F_1(C)$ and $F_2(C)$ respectively. Therefore, the established ROC curve is then defined as the graph that results from plotting the sensibility versus the false positive coefficient. In those cases where the disease produces a rise of the characteristic considered, the individual is diagnosed as being ill when the corresponding X value will be over the cut-off value C. In this case, the sensibility of the diagnosis trial is given by $1-F_1(C)$ and the positive false coefficient by $1-F_2(C)$. The diagnostic power of the marker X can be measured as the area under the corresponding ROC curve. Obviously, areas close to one indicate a high diagnostic power, while values close to 0.5 or less show a poor diagnostic power.

In what follows, it will be considered that the distributions F_1 and F_2 are absolutely continuous and therefore, have a density function that it will be represented by $f_1(x)$ and $f_2(x)$ respectively. This supposes that the functions F_1 and F_2 are continuous and strictly increasing in their density functions supports.

Thus, the inverse function F_2^{-1} gets defined over $]0,1[$. Let´s also define $F_2^{-1}(0) = \sup\{x; F_2(x) = 0\}$ and $F_2^{-1}(1) = \inf\{x; F_2(x) = 1\}$.

If ϕ is the positive false coefficient and S the sensibility, it is easy to prove that the ROC curve corresponds to the graph of the function $S(\phi) = F_1 \circ F_2^{-1}(\phi)$, $\phi \in]0,1[$, which is obviously a continuous function.

3. KERNEL ESTIMATES OF THE DISTRIBUTION FUNCTION

Let's suppose that $F(x)$ is the probability distribution function of a random variable X and let $X_1,\ldots,X_n$ be a random sample of $F(x)$. The kernel estimate of the distribution function $F(x)$ is defined as:

$$\hat{F}_n(x,h) = \frac{1}{n}\sum_{i=1}^{n} W\left(\frac{x - X_i}{h}\right) \qquad (1)$$

where W is a distribution function and h the smoothing parameter or bandwidth. We call the function W the integrated kernel since its derivative, if there exists, is a kernel function in the ordinary sense. In this paper we consider integrated kernels W, such that the ordinary kernel $K(x) = W'(x)$ is lipschitzian, of compact support, continuous and with a finite second order moment.

In order to be able to define an optimal bandwidth for the distribution function of the kernel estimate (1), it should be first calculated the expressions for the bias and the variance of that estimate.

Theorem 1. Let´s suppose that $W(x)$ is derivable with $W'(x) = K(x)$, $K(x)$ verifying the conditions aforementioned and $F \in C^2$. Then:

i) $E\left[\hat{F}_n(x;h)\right] - F(x) = \dfrac{F''(x)\mu_2(K)}{2}h^2 + o(h^2), \quad h \to 0$

$$\text{ii) } \operatorname{var}\left(\hat{F}_n(x;h)\right) = \frac{F(x)(1-F(x))}{n} - \frac{F'(x)\mu_1\left(\left(W^2\right)_\bullet\right)}{n}h + o(h),\ h \to 0 \quad (2)$$

where $\mu_2(K) = \int x^2 K(x)dx$ and $\mu_1\left(\left(W^2\right)_\bullet\right) = \int y \cdot (\partial/\partial y) W^2(y)dy$

The proof is deferred to appendix.

If the function properties of the empirical distribution are compared with those of the estimate $\hat{F}_n(x;h)$, we may observe that the later implies a bias of order h^2, though its variance is lower.

We will then consider as optimal bandwidth h_0, that which minimizes the mean integrated squared error, given by:

$$\text{MISE}(h) = E\left[\int \left\{\hat{F}_n(x,h) - F(x)\right\}^2 dx\right] \quad (3)$$

which is equal to

$$\text{MISE}(h) = \int \operatorname{var}\left(\hat{F}_n(x;h)\right)dx + \int \left\{E\left[\hat{F}_n(x;h)\right] - F(x)\right\}^2 dx$$

According to theorem 1, the asymptotically optimal bandwidth is given by:

$$h_0 = \left\{\frac{\mu_1\left(W_\bullet^2\right)}{\mu_2(K)^2 \cdot \|F''\|_2^2 \cdot n}\right\}^{1/3} \quad (4)$$

where $\|F''\|_2^2 = \int F''(x)^2 dx$.

Since h_0 is unknown, we use a crossvalidation method due to Bowman *et al* for their estimation. Let the crossvalidation function be given by:

$$\mathrm{cv}(h) = \frac{1}{n}\sum_{i=1}^{n}\int\left\{I_{[0,\infty)}(x - X_i) - \hat{F}_{-i}(x,h)\right\}^2 dx \qquad (5)$$

where $\hat{F}_{-i}(x,h)$ denotes the kernel estimate evaluated at observation x, but constructed from the data with observation X_i omitted. The optimal smoothing parameter h_0 is then estimated by $\hat{h}_n$, such that $\mathrm{cv}\left(\hat{h}_n\right) = \min_h \mathrm{cv}(h)$.

A property of this approach follows by considering:

$$H(h) = \mathrm{cv}(h) - \frac{1}{n}\sum_{i=1}^{n}\left\{I_{[0,\infty)}(x - X_i) - F(x)\right\}^2 dx \qquad (6)$$

The new term does not involve h and so the crossvalidatory procedure is unaffected. It is straightforward to prove that:

$$E\left[H(h)\right] = E\left[\int\left\{\hat{F}_{n-1}(x,h) - F(x)\right\}^2\right] \qquad (7)$$

This equation suggests that $H(h)$ might be a good approximation to MISE(h). Bowman shows that under certain conditions, $\hat{h}_n/h_0 \to 1$ with probability 1 as $n \to \infty$, being h_0 the optimal bandwidth and $\hat{h}_n$ the one that minimizes $\mathrm{cv}(h)$.

4. ROC CURVE ESTIMATION

Let $X_{1,1}, X_{1,2}, \ldots, X_{1,n_1}$ and $X_{2,1}, X_{2,2}, \ldots, X_{2,n_2}$ be simple random samples of the $F_1(x)$ and $F_2(x)$ distributions respectively introduced in section 2. We consider for each distribution the estimate given in (1)

$$\hat{F}_{i,n_1}(x;h_i) = \frac{1}{n_i}\sum_{j=1}^{n_i} W\left(\frac{x - X_{ij}}{h_i}\right), \quad i = 1,2 \qquad (8)$$

where $W(x)=\int_{-\infty}^{x} K(t)dt$, and $K(t)$ is a kernel function that verifies the properties given in section 3. Taking into account the properties of the integrated kernel $W(x)$, it is obvious that the function $\hat{F}_{2,n_2}(x;h_2)$ is strictly increasing in $\left\{x; 0<\hat{F}_{2,n_2}(x;h_2)<1\right\}$. We also define $F_{2,n_2}^{-1}(0;h_2)=\sup\left\{x; F_{2,n_2}(x;h_2)=0\right\}$ and $F_{2,n_2}^{-1}(1)=\inf\left\{x; F_{2,n_2}(x;h_2)=1\right\}$ as in section 2. In this way, we estimate the ROC curves as:

$$\hat{S}(\phi;h_1,h_2)=\hat{F}_{1,n_1}\circ\hat{F}_{2,n_2}^{-1}(\phi);\ \phi\in[0,1] \tag{9}$$

For each considered cut-off value C, the estimated sensibility of the diagnosis trial is $\hat{F}_{1,n_1}(C;h_2)$ and the estimated false positive coefficient is given by $\hat{F}_{2,n_2}(C;h_2)$.

The consistency of the kernel estimates of the distribution functions implies the consistency of the ROC curves estimation given by (9), in the sense of that, for each cutt-off C, $\hat{F}_{1,n_1}(C,h)\to F_1(C)$ and $\hat{F}_{2,n_2}(C,h)\to F_2(C)$, when $\min\{n_1,n_2\}\to\infty$.

It is well known that estimates as (1) produce bias, whose asymptotic expressions are given by (2). Then, the estimate ROC curve can accumulate the biases corresponding to both of the estimated distribution functions. Such biases can be approximate estimating $f'(x)$ from the data as:

$$\hat{f}'(x)=\frac{1}{nl^2}\sum_{i=1}^{n}K'\left(\frac{x-X_i}{l}\right) \tag{10}$$

where $K(x)$ is a function kernel and l the corresponding bandwidth. According to (2), the expression for the estimated bias is $\hat{f}'(x)\cdot\mu_2(K)\cdot h^2/2$. Thus, the estimation of the distribution function corrected by bias is:

$$\tilde{F}(x;h) = \hat{F}(x;h) - \frac{\hat{f}'(x)\cdot\mu_2(K)\cdot h^2}{2} \qquad (11)$$

5. APPLICATIONS

A simulation study is carried out in this section; the theoretical ROC curve is compared with the one obtained by means of the methodology based on the nonparametric estimation of the distribution functions; it is also compared with the one obtained supposing normality of the variables within each group. The ROC curve calculated by using the nonparametric method proposed for the diabetes diagnosis from the basal glucose determination is given too. For each group, we have used the Epannnenikov kernel given by $K(t) = 3(1-t^2)\cdot I_{[-1,1]}(t)/4$.

5.1. Simulation study. Let's suppose that a marker X follows a probability distribution log-normal such that $\log(X)$ is $N(2,1)$ in the sick group and $N(3,1/2)$ in the healthy group. A random sample of size 60 has been simulated for each group. We have estimated the ROC curves by using the proposed method based on the nonparametric estimation of the distribution function. Figure 1 shows a simultaneous representation of the theoretical ROC curve, the one obtained by means of the method based on the nonparameric estimation of the distribution functions and finally, the one obtained under the normality assumptions. Figure 2 shows the theoretical ROC curve jointly with the nonparametric estimation with and without correction by bias.

5.2. Diabetes diagnosis. The diabetes diagnosis requires to make several trials, e.g. the preparation of a glucose metabolic curve. However, the determination of basal glucose can be used as a trial to discard the disease. Therefore, we have made a ROC curve (figure 3) based on the measurement of the basal glucose in 67 patients having a confirmed diagnosis of diabetes type 2 with 73 controls carried out at the Hospital Insular of Gran Canaria. The area under ROC curve is 0.7636.

Since the main purpose of the diagnosis trial is to discard the disease, the trial must have a high sensibility, in spite of the fact that in this case the false positive coefficient has to be high too. In order to obtain a 80% sensibility a very high false positive coefficient is required (41%). That produces a cut-off value C=97,64 gram/dl and supposes that for a basal glucose lower than this value, the illness can be reasonably discarded. However, higher values would require to make complementary trials. Figure 3 provides the different sensibilities as a function of the chosen cut-off value C. Figure 4 shows the ROC curve corresponding to Glucose Tolerance Test (GTT). Fifty-one diabetics patients and the same amount of controls received the load of glucose, and then, its concentration in blood was measured. Afterwards, the area under the ROC curve is 0.9549, which it implies a higher diagnostic capacity. Both ROC curves have been estimated by using the bias correction given by $\tilde{F}(x;h)$.

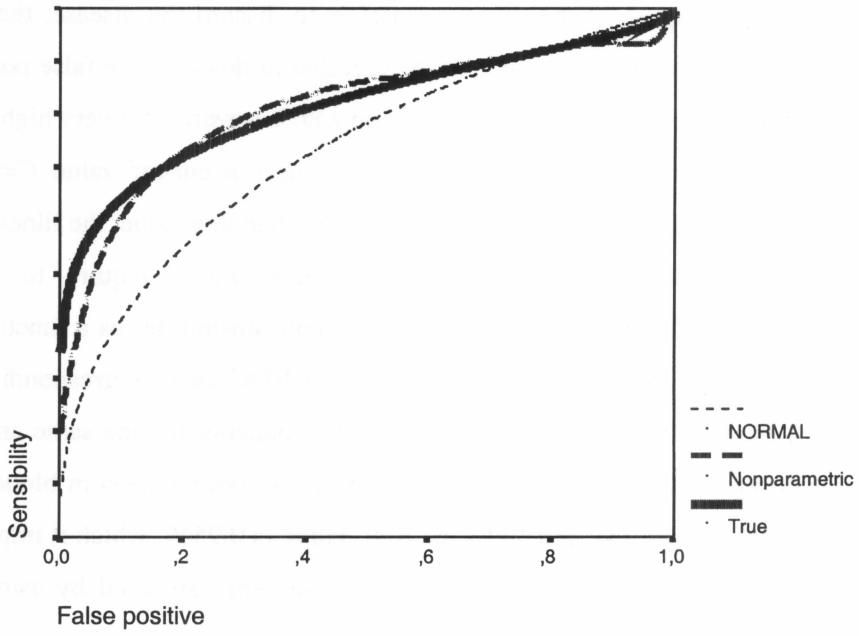

Figure 1. Simulation study. Estimated roc curve by the nonparametric method and under the hypothesis of normality.

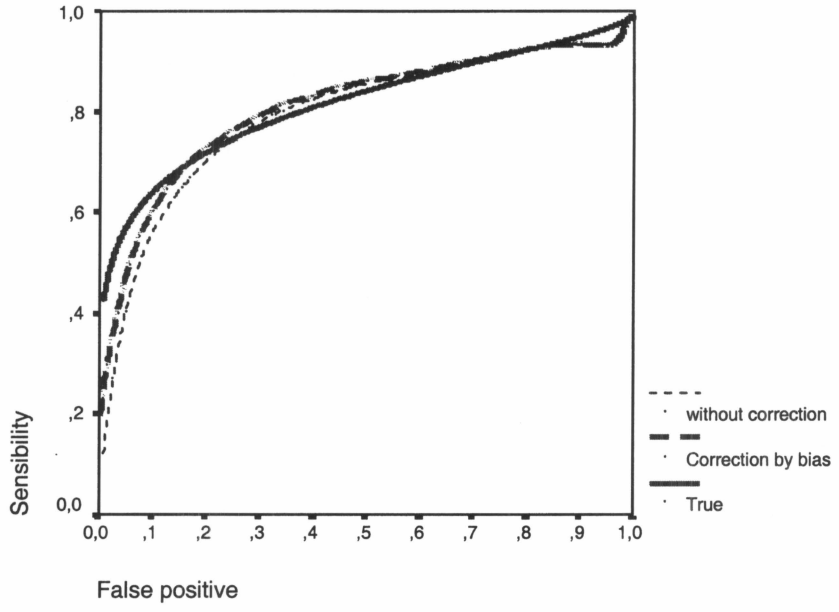

Figure 2. Simulation study. Theoretical roc curve jointly with the nonparametric estimation with and without correction by bias.

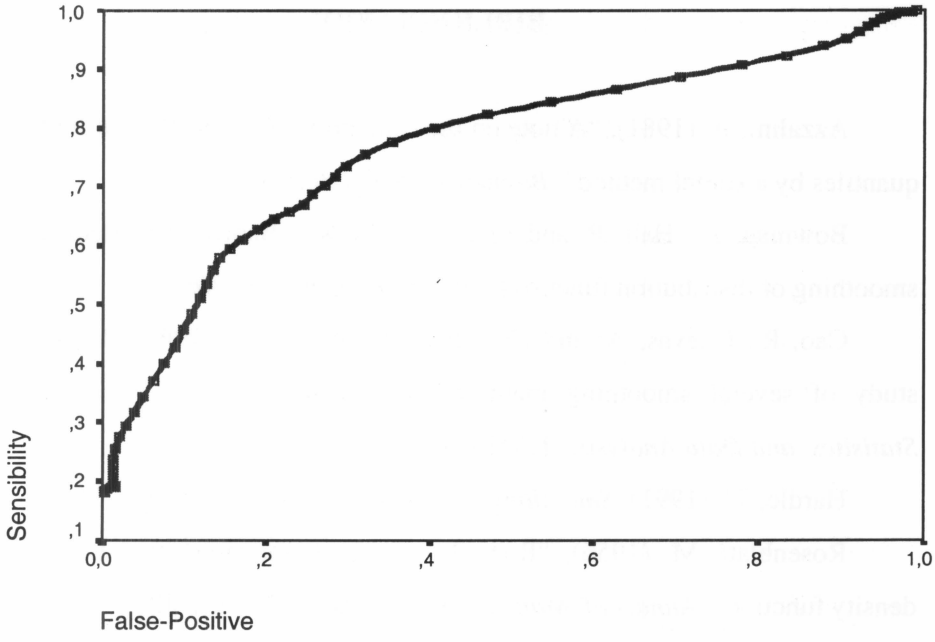

Figure 3. Roc curve of basal glucose

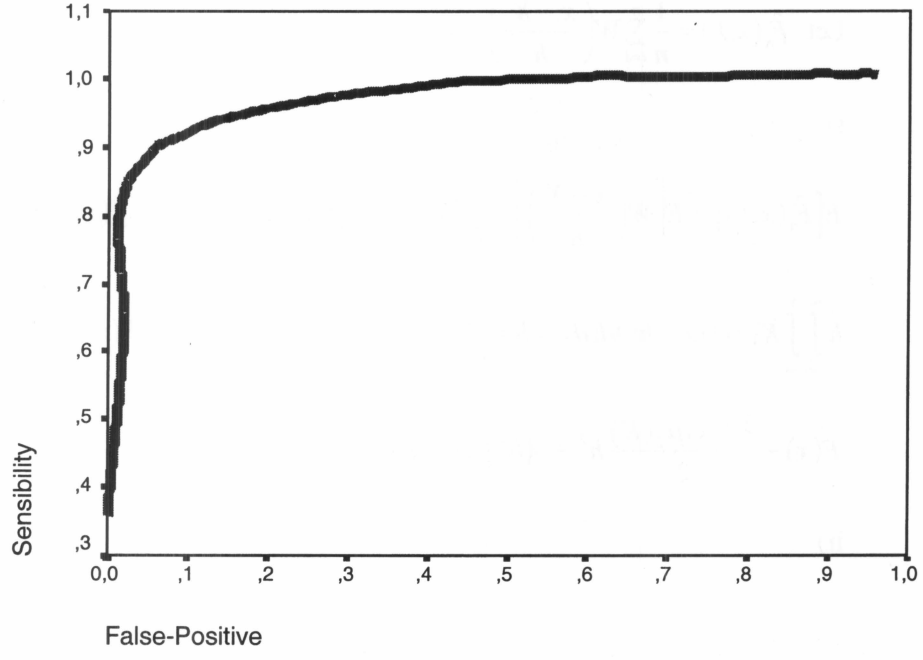

Figure 4. Roc curve of GTT

BIBLIOGRAPHY

Azzalini, A. (1981), "A note on the estimation of a distribution function and quantiles by a kernel method". *Biometrika,* **68**, *326-328.*

Bowman, A., Hall, P. and Prvan, T. (1998), "Bandwidth selection for the smoothing of distribution functions". *Biometrika,* **85**, *799-808.*

Cao, R., Cuevas, A. and González Manteiga, W. (1994), "A comparative study of several smoothing methods in density estimation". *Computational. Statistics. and Data Analysis.,* **1**, 153-176.

Härdle, W. (1991), *Smoothing Techniques.* Springer-Verlag.

Rosenblatt, M. (1956), "Remarks on some nonparametric estimates of a density function". *Annals of Mathematical Statistics*, **27**, 832-837.

APPENDIX

Proof of theorem 1.

Let $\hat{F}_n(x,h)=\frac{1}{n}\sum_{i=1}^{n} W\left(\frac{x-X_i}{h}\right)$. Then,

i)

$$E\left[\hat{F}_n(x;h)\right]=E\left[W\left(\frac{x-X}{h}\right)\right]=\int_{-\infty}^{\infty} W\left(\frac{x-u}{h}\right)f(u)du=h\int_{-\infty}^{\infty} W(y)f(x-hy)dy=$$

$$h\int_{-\infty}^{\infty}\int_{-\infty}^{y} K(z)f(x-hy)dzdy=h\int_{-\infty}^{\infty} K(z)\int_{z}^{\infty} f(x-hy)dydx=\int_{-\infty}^{\infty} K(z)F(x-hz)dz=$$

$$F(x)+\frac{F''(x)\mu_2(K)}{2}h^2+o\left(h^2\right)\text{, for } h\to 0.$$

ii)

$$E\left[W^2\left(\frac{x-X}{h}\right)\right]=\int_{-\infty}^{\infty}W^2\left(\frac{x-u}{h}\right)f(u)du=h\int_{-\infty}^{\infty}W^2(y)f(x-hy)dy=$$

$$\int_{-\infty}^{\infty}F(x-hy)\frac{\partial}{\partial y}\left(W^2\right)(y)dy=F(x)-h\cdot f(x)\int_{-\infty}^{\infty}y\cdot\frac{\partial}{\partial y}\left(W^2\right)(y)dy+o(h),\qquad \text{for}$$

$h\to 0$.

In this way,

$$\operatorname{var}\left(W\left(\frac{x-X}{h}\right)\right)=E\left[W^2\left(\frac{x-X}{h}\right)\right]-\left\{E\left[W\left(\frac{x-X}{h}\right)\right]\right\}^2=$$

$$F(x)-F^2(x)-h\cdot f(x)\int_{-\infty}^{\infty}y\cdot\frac{\partial}{\partial y}\left(W^2\right)(y)dy+o(h)$$

Finally, having in mind that $\operatorname{var}\left(\hat{F}_n(x;h)\right)=\frac{1}{n}\operatorname{var}\left(W\left(\frac{x-X}{h}\right)\right)$, *the result follows.*