SCHOOL OF OPERATIONS RESEARCH
AND INDUSTRIAL ENGINEERING
COLLEGE OF ENGINEERING
CORNELL UNIVERSITY
ITHACA, NY 14853-3801

TECHNICAL REPORT NO. 1024

August 1992

NONPARAMETRIC METHODS FOR EVALUATING DIAGNOSTIC TESTS

by

Fushing Hsieh[1]
and Bruce Turnbull[2]

[1]National Tsing Hua University, Institute of Statistics, Hsinchu, Taiwan 30043. Research supported in part by a fellowship from the U.S. Army Research Office through the Mathematical Sciences Institute at Cornell University.

[2]Research supported in part by Grant GM 28364 from the U.S. National Institutes of Health.

NONPARAMETRIC METHODS FOR EVALUATING DIAGNOSTIC TESTS

Fushing Hsieh and Bruce W. Turnbull

National Tsing Hua University and Cornell University

Abstract: We consider the performance of a diagnostic test based on continuous measurements in its ability to distinguish between healthy and diseased individuals. For a performance criterion we use Youden's (1950) index which is essentially the sum of the sensitivity and specificity. Based on available training set data, two types of nonparametric estimators for the optimal cutoff level and for the index are proposed. The first type is constructed from empirical distribution functions, the other from kernel smoothed density estimates. We compare their asymptotic properties, including rates of convergence. Finite sample properties are investigated by means of a small simulation study. Finally, the methods are applied to results of a glucose tolerance test for diabetes in a sample of 578 individuals from the NHANES-II study.

Key words and phrases: Classification, consistency, convergence rates, diagnostic markers, discrimination, empirical distribution function, empirical processes, kernel density estimate, sensitivity, specificity, Youden index.

1. Introduction

A diagnostic test giving a measurement on a continuous scale is used to classify patients into either the " healthy" or "diseased" categories. Typically, a cutoff point, c, is selected, and patients with test results greater than this are classified as "diseased" , otherwise as "healthy". The test score of a healthy patient is represented as a real random variable X with distribution function F and density f. Similarly a diseased patient's score will be denoted by Y with distribution function G, density g. Typically the supports of X and Y will overlap, but we will assume that:

(A1) there exists a value θ such that g(θ)=f(θ),

$g(t) < f(t)$ for $t < \theta$, and $g(t) > f(t)$ for $t > \theta$.

This is satisfied if, for example, the likelihood ratio is monotone. The assumption implies that X is stochastically smaller than Y, i.e. $F(t) \geq G(t)$ for all t.

The sensitivity of the test is defined as SE(c)=1-G(c), which is the probability of correctly clasifying a diseased individual when cutoff point c is used. Similarly we define the test's specificity SP(c)=F(c) as the probability of correctly classifying a healthy patient. Clearly these are the complements of the familiar Type I and Type II errors. A simple measure of the merit of a diagnostic test is the sum SP(c)+SE(c), which under assumption (A1) is maximized by choosing $c = \theta$. We have

$$\begin{aligned}
\max_c[SE(c) + SP(c)] &= SE(\theta) + SP(\theta) \\
&= 1 + F(\theta) - G(\theta) \\
&= 1 + \max_c[F(c) - G(c)]. \qquad (1)
\end{aligned}$$

Youden (1950) proposed $\eta = F(\theta) - G(\theta)$= $\max_c[F(c) - G(c)]$ as an index of performance of the diagnostic test and he listed a number of its desirable features. This index or measure assumes false positives and false negatives are equally undesirable. Gail and

Green (1976) discussed a generalization whereby the index was a weighted sum of sensitivity and specificity. For simplicity we will consider only Youden's original unweighted index, although our results can easily be extended. In any case, the relative cost of a false positive to a false negative is often difficult to ascertain. Brownie, Habicht and Cogill (1986) have used Youden's index for rating indicators of nutritional status (e.g. skin fold thickness, arm circumference, weight/height etc.) for a population of rural Bangladeshi children. The value θ is also clearly of interest as the value that yields the maximum in (1). In certain circumstances, θ also approximates the optimal choice of cutoff value for estimating the prevalence of the disease in a population (cf. Habicht and Brownie (1982), Brownie and Habicht (1984)).

When the distributions F and G are unknown, we wish to estimate the value of Youden's index η and the optimal cutoff value θ. We suppose that a training data set $X_1, X_2, ..., X_m$ of readings from the healthy population is available as is a set $Y_1, Y_2, ..., Y_n$, from the diseased population. Our approach will be nonparametric and in the next section we consider estimators of η and θ , based on empirical distribution functions F_m and G_n for F and G, respectively. There we will state and prove a theorem about the convergence in distribution of these estimators ($\hat{\eta}$, $\hat{\theta}$, say) with rates $n^{-\frac{1}{2}}$ and $n^{-\frac{1}{3}}$, respectively. In Section 3, we discuss alternative "smoothed" estimators, $\tilde{\theta}, \tilde{\eta}$ say, based on kernel density estimates of f and g and demonstrate their consistency and convergence properties. The rate of convergence of $\tilde{\theta}$ is shown to be the same as that of the density estimate and depends on the smoothness of the underlying densities, f and g. The estimator $\tilde{\eta}$ is shown to be $\sqrt{n}$ mean square consistent and has considerably lower mean square error than the empirical estimator $\hat{\eta}$. Details of all the proofs of lemma and theorems are given in Appendix I. In Section 4 simulation results for comparing estimators discussed in Section 2 and 3 are reported. Also we apply our methods to a glucose tolerance test for the diagnosis of diabetes based on data from the Second National Health and Nutrition Examination Survey (NHANES-II, 1976-1980).

There have been other approaches to the problem of assessment of a merit of a diagnostic test. Altham (1973) used a weighted sum of differences $\sum u_j[F(\xi_j) - G(\xi_j)]$ for given rating levels ξ_j and weights u_j, $1 \leq j \leq r$ for what she terms a measure of "signal discriminability". Greenhouse and Mantel (1950) proposed that a test be acceptable if there existed a cutoff point c such that $SE(c) > \alpha$ and $SP(c) > \beta$ for some prespecified fractions α and β . They went on to describe a hypothesis testing approach for determining whether a diagnostic test was acceptable under this criterion given an available training data set. Schäfer (1989) described a procedure where the cutoff value is chosen to be a specified sample quantile from the X sample or, alternatively, an upper confidence limit for $F^{-1}(p)$, for specified p. He illustrated his method with an application to a marker for bone marrow metastases in patients with small cell lung cancer. Miller and Siegmund (1982) estimated the cutoff point θ by choosing that value θ that maximized the Pearson chi-square statistic based on the 2×2 table formed when the healthy and diseased individuals in the training data set are classified as having test values either above or below θ. Halpern (1982) presented simulation results comparing this maximum chi-square-based statistic, one based on the maximum square of a standardized log cross-product ratio, and the statistic proposed by Gail and Green (1976). Yet another approach involves measures based on the receiver operating characteristic (ROC) curve, given by $1 - G(F^{-1}(1-t))$. For recent papers, see Swets (1988), Wieand et al. (1989), Goddard and Hinberg (1990).

Statistical evaluation of diagnostic tests has been important in many fields, including medicine, nutrition, epidemiology, psychology, electrical engineering and polygraph testing. We shall not attempt to give a review of the large amount of literature on the subject; much of it relates to binary or discrete responses rather than ones on a continuous scale which is our concern. The reader is referred to the book by Swets and Pickett (1982), also the more recent paper by Gastwirth (1987) with accompanying discussion.

2. An empirical estimate of η and θ

A natural estimate of η is obtained by replacing cdf's F and G in the definition by their empirical distribution functions, F_m and G_n, *i.e.*

$$\hat{\eta} = \max_x (F_m(x) - G_n(x)) \tag{2}$$

Analogously we can use the location of the maximum of (2) as an estimate of θ. Since this may not be unique, we define the empirical estimator, $\hat{\theta}$, by

$$\hat{\theta} = \mathrm{median}\{x_0 \mid F_m(x_0) - G_n(x_0) = \max_x (F_m(x) - G_n(x))\}. \tag{3}$$

(Alternatively, in the definition (3), we could use the maximum or minimum value instead of the median.) These estimators $\hat{\eta}$ and $\hat{\theta}$ are nonparametric generalized maximum likelihood estimators in the sense of Kiefer and Wolfowitz (1956).

The problem of estimating θ is similar to that of estimating the mode of a density function. Chernoff (1964) provided an estimator of mode of a density with an $O_p(n^{-\frac{1}{3}})$ rate of convergence, whose distribution was expressed by means of a functional of Brownian motion with quadratic drift. More general development on this cube root asymptotics via functional limit theorems for empirical processes indexed by class of functions can be found in Kim and Pollard (1990).

A heuristic argument given below, which is similar to that of Chernoff (1964) and Kim and Pollard (1990), will lead us to the Theorem 2.1 which is the principal result of this section.

We will assume that θ is unique in the following sense;

(A1$'$) For any $\delta > 0$, there exists ε (> 0) , such that

$$\sup_{|x-\theta|>\delta} [F(x) - G(x)] < F(\theta) - G(\theta) - \varepsilon.$$

Note that (A1$'$) is slightly weaker than (A1). We shall be concerned with the asymptotic properties of our estimators, $\hat{\eta}$ and $\hat{\theta}$. We will assume that the sample sizes are increasing such that $\frac{m}{n} \to \lambda^2(> 0)$, say.

<u>Lemma 2.1</u> Suppose that sequences $\{F_m^*\}$ and $\{G_n^*\}$ are strongly uniform consistent estimators of F and G ; *i.e.*

$$\sup_x \mid F_m^*(x) - F(x) \mid \stackrel{a.s}{\to} 0$$
$$\sup_x \mid G_n^*(x) - G(x) \mid \stackrel{a.s}{\to} 0$$

as $n \to \infty$. Define $\hat{\theta}^*$ and $\hat{\eta}^*$ analogously to $\hat{\theta}$ and $\hat{\eta}$, with F_m^* and G_n^* replacing F_m and G_n, respectively in the definitions (2) and (3). Then, under the condition (A1$'$), we have $\hat{\theta}^*$ and $\hat{\eta}^*$ converge almost surely to θ and η respectively.

The proof of this lemma is given in the Appendix I. Lemma 2.1 together with the Glivenko-Cantelli theorem, which guarantees the strongly uniform convergence of empirical distributions F_m and G_n, show that $\hat{\theta}$ and $\hat{\eta}$, as defined in (2) and (3), are strongly consistent.

Now we define a functional H by $H(H_1,\ H_2,\ x,\ \theta) = (H_1(x) - H_2(x)) - (H_1(\theta) - H_2(\theta))$ for any two functions H_1 and H_2 . Let $\mathcal{C}^{(k)}(C)$ denote the class of functions with a continuous k-th derivative on interval C, $C \subset \Re$. From the strong approximation of empirical processes (Csörgö and Révész, 1981 Theorem 4.41, p.133), we have that, almost surely:

$$H(F_m, G_n, x, \theta) - H(F, G, x, \theta) = \qquad (4)$$
$$\frac{1}{\sqrt{m}}\,[B_1(F(x)) - B_1(F(\theta))] - \frac{1}{\sqrt{n}}[B_2(G(x)) - B_2(G(\theta))] + O(n^{-1} \log\, n)$$

Here $\{B_1\}$ and $\{B_2\}$ are two independent Brownian bridge processes on [0,1]. Further, we assume F and G satisfy (A2) and (A3) below:

(A2) F and G are in $\mathcal{C}^{(2)}(a_0, b_0)$, for some a_0, b_0, with $\theta \in (a_0, b_0)$. F and G have connected intervals as their supports with intersection containing (a_0, b_0).

(A3) $\mid f'(\theta) - g'(\theta) \mid = a,\ a > 0$.

From (A2), if x is close to θ, we see that (4) is approximately distributed as,

$$n^{-\frac{1}{2}}[\lambda^{-2} f(\theta) + g(\theta)]^{\frac{1}{2}} Z((x-\theta)) \tag{5}$$

where $Z(\cdot)$ is a two-sided standard Brownian motion, i.e. Brownian motion on $(-\infty, \infty)$ with $Z(0) = 0$ (Chernoff 1964, page 35). Also the assumptions imply

$$H(F, G, x, \theta) \approx \frac{1}{2}(f'(\theta) - g'(\theta))(x-\theta)^2 \tag{6}$$

From (4),(5) and (6), we have

$$\begin{aligned}
&\max_x (H(F_m, G_n, x, \theta)) \\
&= \max_x (H(F_m, G_n, x, \theta) - H(F, G, x, \theta) + H(F, G, x, \theta))
\end{aligned}$$

converges in distribution to

$$\begin{aligned}
&\max_x \{\frac{1}{\sqrt{n}}[\lambda^{-2} f(\theta) + g(\theta)]^{\frac{1}{2}} Z(x-\theta) - \frac{a}{2}(x-\theta)^2\} \\
&= C \cdot n^{-\frac{2}{3}} \max_z (Z(z) - z^2) \qquad (7)
\end{aligned}$$

where $z = (x-\theta)/\gamma z$ with $\gamma = (\frac{4K}{na^2})^{\frac{1}{3}}$, $K = (\lambda^{-2} f(\theta) + g(\theta))$, $C = \frac{1}{2} \cdot (\frac{4K}{a^2})^{\frac{2}{3}}$ and a is as defined in (A3). As above, $Z(z)$ is defined as a two-sided standard Brownian motion process. Hence we have that

$$\begin{aligned}
\sqrt{n}(\hat{\eta} - \eta) \quad &= \quad \sqrt{n}[F_m(\theta) - G_m(\theta) - (F(\theta) - G(\theta))] \\
&\qquad + \max_x \{\sqrt{n} \cdot H(F_m, G_n, x, \theta)\}
\end{aligned}$$

converges in distribution to

$$\lambda^{-1}B_1(F(\theta)) - B_2(G(\theta)) + O_p(n^{-\frac{1}{6}}) \tag{8}$$

where B_1 and B_2 are two independent Brownian bridges.

The above results are summarized in the Theorem 2.1 below. A rigorous proof may be obtained by a slightly modification of the proof of the main theorem in Kim and Pollard (1990).

Theorem 2.1 : Let F and G satisfy (A1′), (A2) and (A3).

Then we have:

1. $\sqrt{n}(\hat{\eta} - \eta)$ converges in distribution to $\lambda^{-1}B_1(F(\theta)) - B_2(G(\theta)) + O_p(n^{-\frac{1}{6}})$
2. $\hat{\theta}$ converges to θ almost surely and $(\frac{a^2}{4K})^{\frac{1}{3}} n^{\frac{1}{3}}(\hat{\theta} - \theta)$ converges in distribution to the distribution of the random variable which maximizes process $(Z(z) - z^2)$; $z \in \Re$.

Remark 1 From (7), it is clear that

$$\text{Bias}(\hat{\eta}) \doteq C \cdot n^{-\frac{2}{3}} \cdot E\{\max_z(Z(z) - z^2)\}$$

is always positive. Hsieh(1991) considered nonsmoothed bootstrap estimates of η which can reduce the bias, but the bootstrap bias-correction introduces extra variation and the simulation results given there indicate that bootstrapping does not lower the mean square error (MSE).

Remark 2 The MSE of $\hat{\eta}$ can be obtained by squaring (8) and taking the expectation.

$$nE(\hat{\eta} - \eta)^2 = \lambda^{-2}F(\theta)(1 - F(\theta)) + G(\theta)(1 - G(\theta)) + O(n^{-\frac{1}{3}}). \tag{9}$$

Theorem 2.1 shows that $\hat{\eta}$ is first order efficient in estimating η, doing as well asymptotically as if the true θ were known. However, in Section 3, we show that, under stricter smoothness conditions on F and G, another estimator of η can be constructed which yields a lower mean square error. Theorem 2.1 shows that $\hat{\theta}$ converges to θ at rate $n^{-\frac{1}{3}}$. Also in Section 3 we show that a better rate of convergence can be obtained if a smoother condition than (A2) is assumed. However under (A2), it is shown in Hsieh and Turnbull (1992) that $n^{-\frac{1}{3}}$ is the best rate in the sense of being locally asymptotic minimax.

3. Smoothed estimators of η and θ

The estimators of η and θ, $\tilde{\eta}$ and $\tilde{\theta}$ say, considered here are obtained by substituting kernel smoothed estimates in their definitions (2), (3). Their properties are compared with those of the estimators in Section 2; in particular, we show that $\tilde{\eta}$ has an asymptotic mean square error which is smaller than that of $\hat{\eta}$.

3.1 Estimation of θ

We will define kernel density estimates f_m and g_n of f and g, respectively. We will show that the estimator, $\tilde{\theta}$, defined as a solution of $f_m(x) - g_n(x) = 0$, converges to θ at a certain rate.

First suppose $\gamma > 2$, let α be the largest integer less than γ and set $\beta = \gamma - \alpha$. Define $\mathcal{F}(\gamma, \gamma_1)$ to be the class of distribution functions $Q(x)$, of Hölder continuity of order γ. That is they satisfy the following conditions:

$$(i) \quad \text{There exist } (a_0, b_0), \text{ such that } Q(x) \in \mathcal{C}^{(\alpha)}(a_0, b_0) \text{ with } \theta \in (a_0, b_0).$$

$$(ii) \quad \sup \mid x_1 - x_2 \mid^{-\beta} \mid Q^{(\alpha)}(x_1) - Q^{(\alpha)}(x_2) \mid < \gamma_1, \text{over } x_1, x_2 \in (a_0, b_0)$$

From here on, we will assume that

(A2$'$) F and G are in $\mathcal{F}(\gamma, \gamma_1)$, for some γ_1 and $\gamma(> 2)$.

In order to construct smooth density estimators of f and g, we will need to introduce the kernel function $k(\cdot)$. This function can be taken to satisfy the following conditions.

(B1) $k(\cdot)$ is bounded and has a bounded continuous first derivative of bounded variation. Also for some $\delta(>0), \mid k(\cdot)) \mid^{2+\delta}$ is integrable, and $\int k(z)dz = 1, I(k) = \int k^2(z)dz < \infty$ and $H(r,k) = \int \mid z \mid^{\gamma-1} \mid k(z) \mid dz < \infty$. And for any $\delta > 0$

$$\frac{1}{h_n^j} \int_{\{z:|z|>\delta/h_n\}} \mid k^{(j)} \mid dz \to 0 \quad \text{for } j = 0, 1 \text{ as } h_n \to 0.$$

(B2) $k(\ )$ is an α th-order kernel. That is

$$\int z^j k(z)dz = 0, \quad j = 1, 2, \ldots \alpha - 1.$$
$$\text{and} \quad \int z^\alpha k(z)dz \neq 0.$$

Kernel density estimates, $f_m(x)$ and $g_n(x)$, of $f(x)$ and $g(x)$ are given by:

$$\begin{aligned} f_m(x) &= \frac{1}{m} \sum_1^m \frac{1}{h_m} k\left(\frac{x - x_i}{h_m}\right) \\ g_n(x) &= \frac{1}{n} \sum_1^n \frac{1}{h_n} k\left(\frac{x - y_i}{h_n}\right) \end{aligned}$$

where bandwidths $h_m = c \cdot m^{-\frac{1}{2\gamma-1}}$ and $h_n = c \cdot n^{-\frac{1}{2\gamma-1}}$ for an appropriate constant c.

For convenience, we now assume θ is the unique solution of the equation

$$f(x) = g(x)$$

on $(a_0\ b_0)$ and maximizes $F(x) - G(x)$. Under above convention, the condition (A1′) is equivalent to the following assumption (A1″).

(A1″) For $\delta > 0$, sufficiently small, there exists an $\varepsilon > 0$ such that

$$\inf \mid f(x) - g(x) \mid > \ \varepsilon, \text{for} \mid x - \theta \mid > \ \delta \text{ and } x \in (a_0, b_0)$$

We define $\tilde{\theta}$ as follows:

$$\tilde{\theta} = \text{median}\{x \mid x \in (a_0, b_0), \text{and } f_m(x_0) = g_n(x_0)\}. \tag{10}$$

We now have the following theorem.

Theorem 3.1 Let F and G satisfy (A1$''$) and (A2$'$), and kernel $k(\cdot)$ satisfy (B1). Then $\tilde{\theta}$ converges to θ almost surely. Further if (A3) is assumed, the equation $f_m(x) = g_n(x)$ has a unique solution almost surely.

The proof of this strong consistency of $\tilde{\theta}$ is given in Appendix I. Recall that, for our asymptotic theory, $\frac{m}{n} \to \lambda^2$. The next theorem shows that the rate of convergence of $\tilde{\theta}$ is $n^{-\frac{\gamma-1}{2\gamma-1}}$.

Theorem 3.2 Assume that the underlying distribution functions F and G satisfy conditions (A1$''$), (A2$'$) and (A3), kernel function $k(\cdot)$ satisfies conditions (B1) and (B2). Then

$$(nh_n)^{\frac{1}{2}}(\tilde{\theta} - \theta) \longrightarrow Z + c^* \quad \text{(in distribution)}$$

as $n \to \infty$, where Z is normally distributed with mean 0 and variance σ^2 given by

$$\sigma^2 = [(\lambda^{\frac{4\gamma}{2\gamma-1}})f(\theta) + g(\theta)]I(k)/(f'(\theta) - g'(\theta))^2$$

And

$$c^* = (\lambda^{\frac{2(2\gamma-2)}{2\gamma-1}})[C(\gamma, f, \theta) - C(\gamma, g, \theta)]c^{\frac{2\gamma-1}{2}} H(\gamma, k)/[g'(\theta)) - f'(\theta)]$$

Here $C(\gamma, f, \theta)$ is defined by

$$\begin{aligned} C(\gamma, f, \theta)h_m^{\beta} \int \mid z \mid^{r-1} \mid k(z) \mid dz \cdot (1 + o(1)) \\ = \frac{(-1)^{\alpha-1}}{(\alpha-1)!} \int z^{(\alpha-1)}(f^{(\alpha-1)}(\theta - h_m z) - f^{\alpha-1}(\theta))k(z)dz. \end{aligned}$$

and similarly for $C(\gamma, g, \theta)$.

From Theorem 3.2, we have that the rate of convergence $n^{-\frac{\gamma-1}{2\gamma-1}}$ of $\tilde{\theta}$ is the same as the optimal rate for estimation of the density function under the same smoothness conditions (see e.g. Farrell 1972). (It is shown in Hsieh and Turnbull(1992) that this rate is indeed optimal in a sense of being locally asymptotical minimax for estimating θ as well.)

3.2 Estimation of η

To estimate η, we will need first to construct kernel smoothed estimates, $\tilde{F_m}$ and $\tilde{G_n}$ say, of the distribution functions F and G. Because we are now estimating distribution functions rather than densities as in Section 3.1, we will use a kernel function $\tilde{k}(\cdot)$ of order $\alpha+1$, rather than α as above. (This can be seen from the Taylor expansion of the bias in (11) below.) Define kernel distribution $\tilde{K} = \int \tilde{k}$. Now we construct kernel smoothed estimates of F and G with bandwidths $h_m = c \cdot m^{-\frac{1}{2\gamma-1}}$ and $h_n = c \cdot n^{-\frac{1}{2\gamma-1}}$,

$$\tilde{F}_m(t) = \frac{1}{m}\sum_{i=1}^{m} \tilde{K}(\frac{t-x_i}{h_m})$$

and

$$\tilde{G}_n(t) = \frac{1}{n}\sum_{j=1}^{n} \tilde{K}(\frac{t-y_j}{h_n})$$

Then, we have the expectations

$$\begin{aligned} E(\tilde{F}_m(t)) &= F(t) + (-1)^{\alpha}\frac{h_m^{\alpha}}{\alpha!}\int z^{\alpha}[F^{(\alpha)}(t-h_m z) - F^{(\alpha)}(t)]\tilde{k}(z)dz \qquad (11) \\ &= F(t) + C_1(\gamma, F, t)h_m^{\gamma}(1+o(1)), \text{say}, \end{aligned}$$

and similarly,

$$E(\tilde{G}_n(t)) = G(t) + C_1(\gamma, G, t)h_n^{\gamma}(1+o(1)).$$

Variances are given by

$$var(\tilde{F}_m(t)) = \frac{1}{m}F(t)(1-F(t)) - \frac{h_m}{m}f(t)\cdot d_0(1+o(1)) \qquad (12)$$

and

$$var(\tilde{G}_n(t)) = \frac{1}{n}G(t)(1-G(t)) - \frac{h_n}{n}g(t)\cdot d_0(1+o(1)) \tag{13}$$

where

$$d_0 = 2\int z\tilde{k}(z)\tilde{K}(z)dz. \tag{14}$$

From the above expressions, we will choose the kernel $\tilde{K}$ such that d_0 defined above is positive in order that the variances in (12) and (13) are reduced. This we list as Assumption (B3).

(B3) $\tilde{K}$ is chosen so that d_0 in (14) is positive.

From (11) and (12) and by choosing suitable bandwidth constants in constructing the smoothed distribution estimators, we have that the MSE of $\tilde{F}_m(t)$ is

$$E(\tilde{F}_m(t) - F(t))^2 = \frac{1}{m}(F(t)\cdot(1-F(t))) - d^*\cdot\frac{h_m}{m}(1+o(1)) \tag{15}$$

where d^* is positive. That is that the smoothed distribution function, $\tilde{F}_m(t)$, has a MSE smaller than that of $F_m(t)$ by an amount of order $m^{-\frac{2\gamma}{2\gamma-1}}$. (In fact, this rate of improvement upon $F_m(t)$ can be shown to be the optimal one by using the argument found in Hsieh and Levit (1991).)

We can now define the smoothed estimator, $\tilde{\eta}$, as follows:

$$\tilde{\eta} = \tilde{F}_m(\tilde{\theta}) - \tilde{G}_n(\tilde{\theta}) \tag{16}$$

where $\tilde{\theta}$ is defined in (10). We might expect that $\tilde{\eta}$ will improve upon $\hat{\eta}$ by a term that is of the same magnitude as the improvement in MSE of $\tilde{F}_m(t)$ and $\tilde{G}_n(t)$ over $F_m(t)$ and $G_n(t)$. The following theorem, proved in the Appendix, says just this.

Theorem 3.3: We impose the same conditions on F, G and kernel $k(\cdot)$ as assumed in Theorem 3.2. Let $\tilde{k}$ be a kernel function of order $\alpha+1$, uniformly continuous and of bounded variation. Also we assume $\tilde{K}$ is bounded and satisfies (B3). Then, choosing a bandwidth of order $n^{-\frac{1}{2\gamma-1}}$ with appropriate bandwidth constants for kernels k and $\tilde{k}$, the MSE expansion of $\tilde{\eta}$ is ;

$$nE(\tilde{\eta}-\eta)^2 = \lambda^{-2}F(\theta)(1-F(\theta)) + G(\theta)(1-G(\theta)) - d_0^* \cdot h_n(1+o(1))$$

where d_0^* is a positive constant.

Comparing this expression to (9) we see that the improvement in MSE by using $\tilde{\eta}$ over $\hat{\eta}$ can be substantial. Using the same methods mentioned above (Hsieh and Levit 1991), it can be proved that this rate is optimal under the assumed conditions on F and G. It is also clear that a "good" kernel $\tilde{k}$ will be the one that gives a large value of d_0.

4. Simulations

Here we report the results of a small simulation study comparing the MSE's of various estimators of η and θ to see how they perform with finite samples. Simulated training sets of $m = 200$ X-values and $n = 200$ Y-values were generated where X is distributed as $\mathcal{N}(0,1)$ and Y as $\mathcal{N}(2\theta,1)$. Four values of θ were chosen, namely $\theta = 0.5, 1.0, 1.5$ and2.0. Table 1 shows the mean values (with mean square errors in parentheses) for five different estimators of η based on 1000 simulations. The first estimator $\hat{\eta}_1 = \hat{\eta} = \max(F_m(x) - G_n(x))$ is that based on the empirical cdf's. The second is $\hat{\eta}_2 = F_m(\overline{\theta}) - G_n(\overline{\theta})$, where $\overline{\theta} = \frac{1}{2}(\overline{X}_n + \overline{Y}_n)$. This estimator is a natural one to use if f and g are symmetric and differ only by a translation, as is the case simulated here. The next two estimators are of the form $\tilde{\eta} = \tilde{F}_m(\tilde{\theta}) - \tilde{G}_n(\tilde{\theta})$. In both cases the argument $\tilde{\theta}$ is defined as in (10) with bandwidth $h = 1.06n^{-\frac{1}{5}}$ and Gaussian kernels k for f_m and g_n. For the estimates of functions $\tilde{F}_m$, $\tilde{G}_n$, a Gaussian kernel $\tilde{k}$ was also used. However, for $\hat{\eta}_3$ we use bandwidth $h = 1.06n^{-\frac{1}{5}}$, while for $\hat{\eta}_4$, the bandwidth is $h = 1.06n^{-\frac{1}{3}}$. Here of course $n = 200$. The constant 1.06 was chosen following the suggestion by Silverman (1986, page 45). The final estimator, $\hat{\eta}_5$, is defined as $\max(\tilde{F}_m(x) - \tilde{G}_n(x))$

using a Gaussian kernel $\tilde{k}$ for $\tilde{F}_m$ and $\tilde{G}_n$ with bandwidth $h = 1.06n^{-\frac{1}{3}}$. This selection of estimators, kernels and bandwidths, though limited, enables us to see the potential benefits in using the smoothed estimates.

[Table 1 about here.]

The results shown in Table 1 indicate, for the situations investigated, that the non-smoothed estimator $\hat{\eta}_1$ fares poorly in terms of both bias and mean square error. The estimator $\hat{\eta}_2$ is not based on a smoothed estimates of F and G but does use a very accurate estimate of θ in this particular situation where F and G are symmetric and differ only by a translation. The estimator has low bias here, but the mean square errors are higher than the next three estimators which are all based on smoothed estimates of F and G. These last three estimators perform similarly, with low bias and mean square error.

Table 2 shows results from the same simulation study for three estimators of the crossing point θ. The first estimator is $\hat{\theta} = \arg\ \max(F_m(x) - G_n(x))$ as given in Section 2. The second estimator is $\arg\max(\tilde{F}_m(x) - \tilde{G}_n(x))$ using the same Gaussian kernel with bandwidth $h = 1.06n^{-\frac{1}{3}}$. The last estimator is $\tilde{\theta}$ as defined in Theorem 3.1 as the solution to $f_m(x) = g_n(x)$. Again the non-smoothed estimator $\hat{\theta}$ fares poorest both in terms of bias and mean square error. Both smoothed estimators show low bias, but $\tilde{\theta}$ has the lowest mean square error for all the cases considered.

[Table 2 about here.]

Hsieh (1991) also carried out simulations to compare a smoothed bootstrap approach (De Angelis and Young 1992) to obtain bias corrected estimates of η and θ. Although successful in reducing bias, the mean square errors were not significantly reduced and so the extra computation needed did not seem worthwhile when compared to the performance of the smoothed estimators used in Tables 1 and 2.

5. Application to NHANES data

In this section, we apply the methods discussed in Sections 2 and 3 to a training data set from the NHANES-II survey involving glucose tolerance measurements for the diagnosis of diabetes. For each individual, the data consist of three responses, namely fasting glucose level L_0, one-hour glucose level L_1 and two-hour glucose level L_2. These glucose levels of an individual are measured in the following fashion; the fasting glucose level is taken after this individual has been fasting for 12 hours. A 75-gram dose of oral glucose is then administered. The one- and two-hour glucose measurements are then taken after the corresponding intervals. For sample sizes we have $n = 96$ individuals in the diabetic group excluding 6 individuals with missing responses; for the healthy group we have $m = 482$, chosen from the first five hundred and excluding 18 individuals with missing responses. The data are listed in Appendix II. Usually, linear combinations of marker values offer improved performance (Su and Liu 1993). A fourth diagnostic response variable L_3 can be constructed from a linear combination of the three glucose levels as given by,

$$L_3 = 0.5(L_0 + L_2) + L_1.$$

The weights are chosen such that this linear combination is the area under the polygon connecting the three glucose levels by line segments. The nonsmoothed estimators $\hat{\eta}$, $\hat{\theta}$ and smoothed estimators $\tilde{\eta}$, $\tilde{\theta}$ for this data set are displayed in Table 3. For the smoothed estimators in this table $\tilde{F}_m$ and $\tilde{G}_n$ were constructed using a Gaussian kernel with bandwidths, $\hat{\sigma}_x \cdot m^{-\frac{1}{3}}$ and $\hat{\sigma}_y \cdot n^{-\frac{1}{3}}$, respectively, where $\hat{\sigma}_x$ and $\hat{\sigma}_y$ are sample standard deviations. Here $\tilde{\theta}$ is the solution of equation $g_n(x) = f_m(x)$, also constructed with a Gaussian kernel, but with bandwidth $\hat{\sigma}_x \cdot m^{-\frac{1}{5}}$ and $\hat{\sigma}_y \cdot n^{-\frac{1}{5}}$ respectively.

[Table 3 about here.]

From Table 3 it can be seen that the diagnostic variable L_3 has the highest Youden index value η. It is interesting to note the following recommendation for classification and diagnosis of diabetes from the National Diabetes Data Group (1979, page 1040).

"8. The diagnosis of diabetes in non-pregnant adults be restricted to (a) those with the classic symptoms of diabetes and unequivocal hyperglycemia; (b) those with fasting vemous plasma glucose (PG) concentrations greater than or equal to 140 $mg/d\ell$ on more than one occasion; and (c) those who, if fasting plasma glucose is less than 140 $mg/d\ell$ exhibit sustained elevated venous PG values during the oral glucose tolerance test greater than or equal 200 $mg/d\ell$, both at 2-hours after ingestion of the glucose dose and also at some other time point between time 0 and 2-hr."

The table shows that the smoothed estimator of θ recovers the above recommendations on fasting and one-hour glucose levels. However, both the non-smoothed and smoothed method give much lower optimal cut off values fo 2-hour glucose level than 200 $mg/d\ell$ as recommended.

Appendix I

Proof of Lemma 2.1:

From condition (A1′), for any small $\delta(>0)$, choose an $\varepsilon(>0)$ accordingly. Let $\varepsilon' = \frac{\varepsilon}{5}$. From the strong consistency of F_m^* and G_n^*, there is a pair (m_0, n_0) such that for all $m > m_0$ and $n > n_0$:

$$F(x) - G(x) - 2\varepsilon' < F_m^*(x) - G_n^*(x) < F(x) - G(x) + 2\varepsilon' \qquad \text{for all } x$$

Hence

$$\begin{aligned} \sup_{|x-\theta|>\delta} [F_m^*(x) - G_n^*(x)] &\le 2\varepsilon' + \sup_{|x-\theta|>\delta} (F(x) - G(x)) \\ &< F(\theta) - G(\theta) - 3\varepsilon' \\ &< F_m^*(\theta) - G_n^*(\theta) \end{aligned}$$

Therefore

$$\sup_{|x-\theta|<\delta} (F_m^*(x) - G_n^*(x)) = \sup_x (F_m^*(x) - G_n^*(x))$$

Thus $\hat{\theta}^* \longrightarrow \theta$ a.s. Similarly,

$$\sup_x (F_m^*(x) - G_n^*(x)) = F_m^*(\hat{\theta}^*) - G_n^*(\hat{\theta}^*) \overset{a.s}{\to} F(\theta) - G(\theta) = \sup_x (F(x) - G(x)).$$

and thus $\hat{\eta}^* \longrightarrow \eta$ a.s. which completes the proof of the Lemma.

Before going on to the proof of Theorem 3.1, we need to state the following lemma.

Lemma A1 Let f be a density with distribution function $F \in \mathcal{F}(\gamma, \gamma_1)$; for some $\gamma > 2$. Let the kernel k have a bounded and continuous integrable j-th derivative of bounded variation. Set

$$\hat{f}_{n,h}(t) = \frac{1}{h} \int k(\frac{t-x}{h}) dF_n \text{ and } f_{n,h}(t) = \frac{1}{h} \int k(\frac{t-x}{h}) dF$$

We take h_n to be a fixed bandwidth sequence such that $nh_n^{2j+1}/logn \to \infty$, then, for $j \leq \alpha$, the largest integer less than γ:

$$\sup_t \mid \hat{f}_{n,h_n}^{(j)}(t) - f_{n,h_n}^{(j)}(t) \mid \to 0 \quad a.s.$$

This lemma follows directly from Theorem 37 of Pollard (1984, page 34). See also Romano (1988,Corollary 5.1)

<u>Proof of Theorem 3.1</u>

From the condition (B1) on kernel $k(\cdot)$ and smoothness conditions on F and G, we have

$$\sup_{x \in (a_0,b_0)} \mid E(f_m(x) - g_n(x)) - (f(x) - g(x)) \mid \to 0$$

as $n \to \infty$. By Lemma A1 above,

$$\sup_{x \in (a_0,b_0)} \mid f_m(x) - g_n(x) - (f(x) - g(x)) \mid \to 0 \quad a.s.$$

Using the argument similar to that in the proof of Lemma 1 with (A1$''$), we have that $\tilde{\theta}$ converges to θ almost surely.

When (A3) is assumed, from Lemma A1 and the uniform continuity of $f'(x)$ and $g'(x)$, for $x \in (a_0, b_0)$, it follows that $\tilde{\theta}$ will be the unique solution of equation $f_m(x) = g_n(x)$ almost surely. This completes the proof.

<u>Proof of Theorem 3.2</u>

From Theorem 3.1, using a Taylor expansion, we have

$$(nh_n)^{\frac{1}{2}}(\tilde{\theta} - \theta) = (nh_n)^{\frac{1}{2}}(f_m(\theta) - g_n(\theta))/(g_n'(\theta^*) - f_m'(\theta^*))$$

where θ^* lies between θ and $\tilde{\theta}$. To prove the theorem, it is sufficient to show that

$$(i) \quad (nh_n^{\frac{1}{2}}(f_m(\theta) - g_n(\theta))/(g'(\theta) - f'(\theta)) \to Z + C^* \text{ (in dist.)}$$

$as\ n \to \infty$, where Z is normally distributed and C^* is a constant, and

$$(ii) \quad g_n'(\theta^*) - f_m'(\theta^*) \to g'(\theta) - f'(\theta) \quad a.s.$$

For (i), by simple calculations, we have

$$\begin{aligned} E(f_m(\theta)) &= f(\theta) + C(\gamma, f, \theta) h_m^{\gamma-1}(1+o(1)), \text{ and} \\ E(g_n(\theta)) &= g(\theta) + C(\gamma, g, \theta) h_n^{\gamma-1}(1+o(1)). \end{aligned}$$

Also,

$$\begin{aligned} var(f_m(\theta)) &= \frac{1}{m} var(\frac{1}{h_m} k(\frac{\theta - X_1}{h_m})) \\ &= \frac{1}{m h_m} \cdot f(\theta) \int k^2(z) dz \cdot (1+o(1)), \end{aligned}$$

and similarly,

$$var(g_n(\theta)) = \frac{1}{n h_n} g(\theta) \int k^2(z) dz \cdot (1+o(1)).$$

It is easy to check the Liapounov condition, since $| k(x) |^{2+\delta_o}$ is integrable and so (i) follows by the central limit theorem. The constant C^* depends on $C(\gamma, f, \theta), C(\gamma, g, \theta), \lambda$ and $(g'(\theta) - f'(\theta))$.

For (ii), note that Lemma A1 implies the strong consistency of f'_m and g'_n, i.e.

$$\sup_{x \in (a_0, b_0)} | f'_m(x) - g'_n(x) - (f'(x) - g'(x)) | \to 0$$

as $n \to \infty$. So that, for θ^* between θ and $\tilde{\theta}$, we have

$$g'_n(\theta^*) - f'_n(\theta^*) \to g'(\theta) - f'(\theta) \quad a.s.$$

This completes the proof of Theorem 3.2.

<u>Proof of Theorem 3.3</u>

From the definition of $\tilde{\eta}$,

$$\begin{aligned} \tilde{\eta} - \eta &= [(\tilde{F}_m(\tilde{\theta}) - G_n(\tilde{\theta})) - (F(\theta) - G(\theta))] \\ &= [(\tilde{F}_m(\theta) - F(\theta)) - (\tilde{G}_n(\theta) - G(\theta))] \\ &\quad + [\tilde{F}_m(\tilde{\theta}) - \tilde{F}_m(\theta) - (\tilde{G}_m(\tilde{\theta}) - \tilde{G}_m(\theta))] \\ &= U + V, \text{ say.} \end{aligned}$$

To prove the theorem, by using (15) we need only to show that

$$E(| V |) = O(n^{-4(\gamma-1)/(2\gamma-1)}).$$

Let A_n be the event

$$A_n = \{| f'_m(x) - g'_n(x) |> \frac{a}{2}, \ \forall x \in (\theta - \epsilon_0, \theta + \epsilon_0) \text{ for some } \epsilon_0\}$$

and define $\| \tilde{K} \| = \sup_x | \tilde{K}(x) |$. Then, using $f(\theta) = g(\theta)$, we have

$$\begin{aligned}
| V | &= | \tilde{F}_m(\tilde{\theta}) - \tilde{G}_n(\tilde{\theta}) - (\tilde{F}_m(\theta) - \tilde{G}_n(\theta)) | \\
&\leq 4 \| \tilde{K} \| 1_{A_n^c} + | (\tilde{f}_m(\tilde{\theta}^*) - \tilde{g}_n(\tilde{\theta}^*))(\tilde{\theta} - \theta) | \cdot 1_{A_n} \\
&= 4 \cdot \| \tilde{K} \| \cdot 1_{A_n^c} + 1_{A_n} \cdot | (\tilde{\theta} - \theta) \cdot \{[\tilde{f}_m(\tilde{\theta}^*) - f(\tilde{\theta}^*)] \\
&\quad + [f(\tilde{\theta}^*) - f(\theta)] - [\tilde{g}_n(\tilde{\theta}^*) - g(\tilde{\theta}^*)] - [g(\tilde{\theta}^*) - g(\theta)]\} | \\
&\leq 4 \| \tilde{K} \| \cdot 1_{A_n^c} + 1_{A_n} | \{| \tilde{f}_m(\tilde{\theta}^*) - f(\tilde{\theta}^*) | \\
&\quad + | \tilde{g}_n(\tilde{\theta}^*) - g(\tilde{\theta}^*) |\}(\tilde{\theta} - \theta) + (\| f' \| + \| g' \|)(\tilde{\theta} - \theta) |
\end{aligned}$$

where $\tilde{\theta}^*$ is between $\tilde{\theta}$ and θ, and $\| f' \|$ and $\| g' \|$ are defined respectively as

$$\sup_{x \epsilon (\theta_0 - \epsilon_0, \theta + \epsilon_0)} | f'(x) | \text{ and } \sup_{x \epsilon (\theta - \epsilon_0, \ \theta + \epsilon_0)} | g'(x) | \ .$$

By applying the maximum inequality, (Pollard (1984, page 31)), we have

$$(a) \quad Prob(A_n^c) \leq exp(-nh_n^3 \cdot d^*)$$

for some constant $d^*(> 0)$. In the statement of Theorem 3.3 we have assumed $\tilde{K}$ is bounded. Therefore, $4 \| \tilde{K} \| E(1_{A_n^c})$ can be smaller than any polynomial in n^{-1} for n sufficiently large.

From the definition of $\tilde{\theta}$, we have

$$\begin{aligned}
(b) \ E((\tilde{\theta} - \theta)^2 1_{A_n}) &= E\left(\frac{(f_m(\theta) - g_n(\theta))^2}{(f'_m(\theta^*) - g'_n(\theta^*))^2} \cdot 1_{An}\right) \\
&\leq \frac{8}{a} \cdot E\{| f_m(\theta) - f(\theta) |^2 + | g_n(\theta) - g(\theta) |^2\} \\
&= O(n^{-\frac{4(\gamma-1)}{2\gamma-1}})
\end{aligned}$$

Again by another maximum inequality result of Pollard (1990, page 37), we have

$$
\begin{aligned}
(c)\ \ E \mid \tilde{f}(\tilde{\theta}^*) - f(\tilde{\theta}^*) \mid^2 &\leq E(\sup_{x \in N(\varepsilon_0)} \mid \tilde{f}_m(x) - f(x) \mid^2) \\
&\leq 2 \cdot \{E \mid \sup_{x \in N(\varepsilon_0)} \mid \tilde{f}_m(x) - E\tilde{f}_m(x) \mid^2) \\
&\quad + \sup_{x \in N(\varepsilon_0)} (E(\tilde{f}_m(x) - f(x))^2\} \\
&= O(n^{-\frac{4(\gamma-1)}{2\gamma-1}})
\end{aligned}
$$

where $N(\varepsilon_0) = (\theta - \varepsilon_0, \theta + \varepsilon_0)$. Similarly

$$
E \mid g_n(\tilde{\theta}^*) - g(\tilde{\theta}^*) \mid^2 \ = \ O(n^{-\frac{4(\gamma-1)}{2\gamma-1}})
$$

Combining (a),(b) and (c), and the Cauchy-Schwarz inequality, the proof of Theorem 3.3 follows.

Appendix II NHANES-II Data used in Section 5.

Diseased group			Healthy group																	
0-hr	1-hr	2-hr	0-hr	1-hr	2-hr	0-hr	1-hr	2-hr	0-hr	1-hr	2-hr	0-hr	1-hr	2-hr	0-hr	1-hr	2-hr
102	240	270	111	221		97	205	189	102	222	196	85	130	113	144	264	298
108	184	168	92	137	66	113	204	173	85	142	106	100	230	204	97	142	151
200	375	438	105	220	139	90	149	118	102	169	103	86	113	100	89	173	93
147	247	261	88	171	90	92	123	69	103	145	86	90	191	46	105	204	163
272	407	461	102	212	134	102	138	117	78	140	66	92	83	46	104	210	127
88	273	52	119	224	189	93			86	85	77	142	358	200	93	161	111
103	228	242	66	144	91	90	91	84	96	163	93	87	130	95	83	190	90
133	173	99	98	191	110	100	188	112	90	150	103	155	320	153	93	142	101
122	258	278	87	161	115	103	123	79	81	161	128	87	135	137	88	136	125
145	287	358	133	288	242	127	266	295	85	186	110	105	163	77	88	100	109
124	136	116	99	144	85	90	131	70	82	141	77	77	83	101	93	188	131
223	389	415	93	119	56	105			88	125	108	78	116	87	91	83	48
103	178	175	95	242	115	73	117	79	90	103	100	77	96	83	91	156	112
269	458	472	69	127	133	83	115	83	82	158	126	103	195	94	100	141	101
100	194	130	84	122	79	118	224	193	93	157	98	86	118	58	122	294	146
97	217	200	112	222	224	95	216	189	84	145	170	95	149	164	97	89	102
89	160	104	75	93		93	218	151	136	324	240	78	87	85	96	87	119
118	198	195	108	230	135	89	99	90	95	149	111	94	162	142	78	57	59
163	296	293	85	147	74	111	168	99	100	208	124	80	96	103	97	199	119
151	319	296	109	212	157	81	152	127	98	249	227	97	173	108	122	187	123
115	292	190	68	70	53	88	150	63	82	142	120	91	82	75	95	129	85
100	141	107	96	178	91	84	92	120	92	120	80	149	255	217	100	148	132
111	229	180	95	147	100	101	96	118	113	197	113	90	237	59	91	128	112
98	119	97	94	153	135	83	104	107	82	202	156	93	199	90	85	149	116
85	136	103	87	138	128	83	93	91	85	149	83	115	144	161	88	195	103
181	328	318	82	112	109	90	124	94	108	128	119	92	138	124	189	313	393
135	279	276	97	165	126	84	122	83	84	127	111	122	184	118	84	107	79
155	324	382	82	83	69	84	67	53	89	162	85	102	95	113	103	157	146
400	581	703	83	109	84	81			92	166	132	89	201	201	83	128	77
93	188	144	104	104	118	99	117	94	95	143	90	93	169	58	91	174	99
95	214	131	89	142	97	87	145	128	102	155	119	108	128	91	132	231	177
112			95	180	138	55	152	149	87	117	89	95	243	159	91	71	118
107	247	228	85	73	99	94	67	108	109	217	155	94	111	95	101	183	101
196	354	291	110	189	183	105	260	138	104	222	175	92	137	85	89	175	145
178	378	356	90	138	125	96	178	152	90	119	91	92	135	132	87	130	107
250	409	456	100	206	95	80	122	96	129	305	196	88	109	114	97	151	103
98	188	143	80	84	70	101	151	133	72	74	78	95	137	94	100	158	109
196	365	379	100	198	119	81	112	128	92	124	80	89	113	96	105	161	110
117	198	212	100	154	115	92	184	202	79	105	80	95	67	74	96	114	39
83	164	165	92			90	75	64	95	184	127	90	122	97	97	163	168
105	189	151	89	122	108	92	160	133	81	129	108	165	307	304	93	132	101
231	332	382	84	136	111	88	155	154	84	120	103	89	75	75	94	158	81
89	239	153	104	188	81	97	238		87	180	116	97	103	84	104	136	113
140	247	240	86	151	136	126	267	257	80	166	102	94	139	150	99	185	99
110	215	208	90	194	109	92	134	89	90	146	81	90	125	91	113	240	186
73	187	154	91	153	36	103	218	56	108	202	87	82	160	80	95	160	115
180	396	282	90	154	117	95	139	107	94	143	96	101	135	120	81	126	92
85	155	106	99	182	162	89	197	112	89	148	105	94	223	111	83	80	95
158	253	202	88	90	85	89	129	85	81	93	110	84	147	89	87	148	130
226	334		133	290	326	107	348	278	79	209	85	91	199	172	85	102	74
138	266	312	95	192	111	98	152	121	79	180	168	97	141	116	87	147	

Diseased group			Healthy group														
0-hr	1-hr	2-hr	0-hr	1-hr	2-hr	0-hr	1-hr	2-hr	0-hr	1-hr	2-hr	0-hr	1-hr	2-hr	0-hr	1-hr	2-hr
89	230		100	114	101	90	159	85	94	171	73	90	188	140	92	124	95
167	310	373	90	207	175	86	94	74	80	79	51	104	172	122	94	172	110
121	226	187	84	121	97	98	203	98	102	197	54	117	124	116	91	198	82
90		161	93	134	105	83	111	100	82	135	80	99	148	118	82	198	116
120	170	93	79	103	95	86	138	112	102	187	132	95	158	131	110	230	189
172	329		90	181	121	77	94	112	93	179	140	99	120	81	74	81	66
91	101	96	75	129		79	110	119	81	105	90	125	216	165	78	112	99
122	217	137	84	120	99	89	159	109	102	110	99	96	202	84	97	165	89
87	211	133	90	115	78	98	190	186	76	92	75	81	115	81	88	96	65
155	330	259	93	133	108	92	96	98	90	174	155	74	83		99	189	166
93	193	178	99	232		75	151	130	91	136	100	89	168	94	87	97	100
155	313	293	102	167	102	98	206	177	84	126	50	90	224	194	83	146	111
239	436	405	104	214	120	85	118	96	93	167		93	180	156	85	141	91
198	312	349	85	100	51	103	81	111	91	108	90	96	184	146	98	143	115
161	226	169	102	225	207	95	169	97	81	142	142	87	103	106	78	148	115
106	271	155	91	130	104	92	94	102	89	134	95	108	228	138	93	97	94
170	304	201	83	178	120	86	116	111	95		147	99	190	137	88	120	115
87	133	104	104	177	130	74	140	140	91	122	126	103	247	178	86	77	118
98	189	213	96	206	149	79	186	152	93	112	142	94	150	131	84	79	103
400	617	603	92	111	89	101	148	119	97	165	101	100	172	148	81	141	110
136	285	368	97	182	131	85	90	113	87	90	116	87	145	100	79	110	112
121	249	148	104	151	128	77	109	96	82	122	116	110			81	185	122
127	273	265	95	109	59	81	93	98	95	152	116	106	190	133	161	306	215
108	239	165	79		81	78	97	106	90	132	91	122	313	275	92	118	105
163	305	278	108	164	151	87	156	103	97	193	190	118	256	240	94	125	102
94	165	105	116	241	167	97	160	102	122	267	221	86	99	67	94	185	59
88	255	281	86	158	117	90	104	101	90	199	73	120	257	175	89	111	108
109	208	223	76	168	160	63	67	64	144	293	301	95	187	94	85	167	101
101	212	149	79	102	101	90	95	101	91	150	122	84	162	147	99	144	94
138	334	319	99	172	140	92	213	156	87	96	112	91	216	136	88	96	94
87	149	112	87	146	130	88	85	63	97	148	123	82	100	73	108	218	213
164	317	314	109	231	128	89	85	122	99	182	137	86	142	106	78	58	53
89	176	52	94	133	86	79	117	85	83	147	98	89	152	87	110	201	125
117	248	244	86	78	113	90	198	163	94	205	98	105	222	129	105	134	87
85	203	103	80	112	86	91	143	103	84	113	89	97	185	78	88	136	73
105	217	82	104	212	171	81	146	99	139	264	273	110	252	151	89	99	88
203	307	344	83	113	80	89	202	166	92	148	46	91	115	93	82	129	95
211	345	315	89	209	75	88	163	139	95	191	154	123	207	185	82	95	82
125		245	92	129	111	81	174	160	88	142	89	119	265	259	80	77	63
100	232	225	87	88	105	77	116	99	83	172	130	98	179	124	100	133	92
163	289	329	94	118	71	68	153		100	217	130	81	106	86	111	261	224
67	80	83	89			96	129	96	92	187	160	100	146	141	93	170	57
104	243	175	79	106	87	92	83	74	80	77	77	83	93	93	92	121	123
86	169	128	95	173	124	76	136	122	89	195	121	87	127	57	79	81	93
236	347	402	88	214	151	91	197	96	112	220	189	88	170	108	98	146	122
83	96	117	77	129	73	97	80	79	82	120	128	109	244	232	103	171	64
111	279	107	87	110	102	101	94	95	80	115	91	96	177	180	94	307	134
262	400	462	88	71	71	97	237	181	80	187	89	96	145	139	94	171	90
103	215	141	82	53	126	85	131	108	88	165	135	93	194		104	222	71
110	223	106															
120	246	176															

Acknowledgements The first author was supported in part by a fellowship from the U.S. Army Research Office through the Mathematical Sciences Institute at Cornell University. The second author was supported in part by Grant GM 28364 from the U.S. National Institutes of Health.

References:

Altham, P.M.E. (1973). A non-parametric measure of signal discriminability. *Brit. J. of Math. and Statist. Psychology* **26**, 1-12.

Brownie, C. and Habicht, J.-P. (1984). Selecting a screening cutoff point or diagnostic criterion for comparing prevalences of disease. *Biometrics* **40** 675-684.

Brownie, C., Habicht, J.-P. and Cogill, B. (1986). Comparing indicators of health or nutritional status. *Am. J. of Epidemiology* **124**, 1031-1044.

Chernoff H. (1964). Estimation of the mode. *Ann. Inst. Statist. Math.* **16**, 31-41.

Csörgö, M. and Révész, P. (1981) *Strong Approximations in Probability and Statistics.* Academic Press, New York.

De Angelis D and Young G.A (1992). Smoothing the bootstrap. *International Statistical Review* **60**, 45-56.

Farrell, R.H. (1972) On the best obtainable asymptotic rates of convergence in estimation of a density function at a point. *Ann. Math. Statist.* **43**, 170-180.

Gail, M. H. and Green, S.B. (1976). A generalization of the one-sided two sample Kolmogorov-Smirnov Statistic for evaluation diagnostic tests. *Biometrics* **32**, 561-570.

Gastwirth, J.L. (1987). The statistical precision of medical screening procedures (with discussion). *Statistical Science* **2**, 213-238.

Goddard, M.J. and Hinberg, I. (1990). Receiver operator characteristic (ROC) curves and non-normal data : an empirical study. *Statistics in Medicine* **9**, 325-337.

Greenhouse, S.W. and Mantel, N. (1950). The evaluation of diagnostic tests. *Biometrics* **6**, 399-412.

Habicht, J-P. and Brownie C. (1982) Reply to letter by Bairagi on best cutoff point for nutritional monitoring. *Am. J. of Clinical Nutrition* **35**, 369-371.

Halpern, J. (1982). Maximally selected chi square statistics for small samples. *Biometrics* **38**, 1017-1023.

Hsieh, F.S. (1991) Performance of diagnostic tests in a nonparametric setting. Ph.D. Thesis, Cornell University.

Hsieh F.S. and Levit B. (1991). On the optimal rates of improvement of the sample median. Technical Report 684, Department of Mathematics, University of Utrecht. The Netherlands.

Hsieh F.S and Turnbull W.B (1992). Locally asymptotic minimax rate of crossing point estimation. Technical Report, Department of Operations Research, Cornell University.

Kiefer J. and Wolfowitz J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Ann. Math. Statist.* **27**, 642-669.

Kim J. and Pollard D. (1990). Cube root asymptotics. *Ann. Statist.* **18**, 191-219.

Mantel, N. (1951). Evaluation of a class of diagnostic tests. *Biometrics* **7**, 240-246.

Miller, R. and Siegmund, D. (1982). Maximally selected chi square statistics. *Biometrics* **38**, 1011-1016.

National Diabetes Data Group (1979). Classification and diagnosis of diabetes mellitus and other categories of glucose intolerance. *Diabetes* **28**, 1039-1057.

Pollard D. (1984). *Convergence of Stochastic Processes.* Springer-Verlag, New York.

Pollard D. (1990). *Empirical Processes: Theory and Applications.* NSF-CBMS Regional Conference series in Probability and Statistics Vol. 2., Institute of Mathematical Statistics, Hayward, Calif.

Romano, J. (1988). On weak convergence and optimality of kernel density estimates of the mode. *Ann. Statist.* **16**, 629-647.

Schäfer, H. (1989). Constructing a cutoff point for a quantitative diagnostic test. *Statistics in Medicine* **8**, 1381-1391.

Silverman, B.W. (1978). Weak and strong consistency of the kernel estimate of a density and its derivatives. *Ann. Statist.* **6**, 177-184.

Silverman, B.W. (1986). *Density Estimation for Statistics and Data Analysis.* Chapman and Hall, New York.

Su, J.Q. and Liu, J.S. (1993). Linear combinations of multiple diagnostic markers. *J. Amer. Statist. Assoc.* To appear. Research Report R-430, Department of Statistics, Harvard University.

Swets J.A. and Pickett R.M. (1982). *Evaluation of Diagnostic Systems: Methods from Signal Detection Theory.* Academic Press, New York.

Swets, J.A. (1988). Measuring the accuracy of diagnostic systems. *Science* **240**, 1285-1293.

Youden, W.J. (1950). Index for rating diagnostic tests. *Cancer* **3**, 32-35.

Wieand, S., Gail, M.H., James, B.R. and James, K.L. (1989). A family of nonparametric statistics for comparing diagnostic markers with paired or unpaired data. *Biometrika* **76**, 585-592.

Institute of Statistics, National Tsing Hua University, Hsinchu, 30043, Taiwan.

School of Operations Research and Industrial Engineering, 227 E&TC Building, Cornell University, Ithaca, NY 14853-3801, USA.

Table 1: Simulation results for Youden's index η

θ	$\theta = 0.5$	$\theta = 1.0$	$\theta = 1.5$	$\theta = 2.0$
$\eta = \max(F(x) - G(x))$	0.38292	0.68269	0.86639	0.95450
$\hat{\eta}_1 = \max(F_m(x) - G_n(x))$	0.41066 (2.544×10^{-3})	0.70199 $(1.475 \times 10^{-3}$	0.88001 $(0.706 \times 10^{-3}$	0.96265 (0.233×10^{-3})
$\hat{\eta}_2 = F_m(\overline{\theta}) - G_n(\overline{\theta})$	0.38240 (2.119×10^{-3})	0.68249 (1.300×10^{-3})	0.86636 (0.643×10^{-3})	0.95478 (0.231×10^{-3})
$\hat{\eta}_3 = {}_5\tilde{F}_m(\tilde{\theta}) - {}_5\tilde{G}_n(\tilde{\theta})$	0.38671 (1.889×10^{-3})	0.68357 (1.196×10^{-3})	0.86740 (0.576×10^{-3})	0.95540 (0.194×10^{-3})
$\hat{\eta}_4 = {}_3\tilde{F}_m(\tilde{\theta}) - {}_3\tilde{G}_n(\tilde{\theta})$	0.38128 (1.728×10^{-3})	0.67688 (1.135×10^{-3})	0.86183 (0.537×10^{-3})	0.95221 (0.183×10^{-3})
$\hat{\eta}_5 = \max({}_3\tilde{F}_m(x) - {}_3\tilde{G}_n(x))$	0.38277 (1.707×10^{-3})	0.67776 (1.115×10^{-3})	0.86247 (0.525×10^{-3})	0.95274 (0.178×10^{-3})

Note:

1. Normal kernel is used with bandwidth constant 1.06.
2. $\tilde{\theta}$ is defined in (10) with $h = 1.06n^{-\frac{1}{5}}$ and $\overline{\theta} = \frac{1}{2}(\overline{X}_n + \overline{Y}_n)$.
3. ${}_k\tilde{F}_m$ and ${}_k\tilde{G}_n$ are smoothed distibution functions with bandwidth of order $n^{-\frac{1}{k}}$.
4. The number in parentheses is the MSE.

Table 2: Simulation results for the crossing point θ

Estimator	$\theta = 0.5$	$\theta = 1.0$	$\theta = 1.5$	$\theta = 2.0$
$\hat{\theta}$	0.4876 (5.188×10^{-2})	0.9777 (3.06×10^{-2})	1.4791 (2.736×10^{-2})	1.9251 (3.825×10^{-2})
location of maximum of $({}_3F_m(x) -{}_3 G_n(x))$	0.5002 $(3.230 \times^{-2})$	1.0003 (1.41×10^{-2})	1.5094 (1.352×10^{-2})	1.9974 (1.778×10^{-2})
$\tilde{\theta}$	0.5005 (1.781×10^{-2})	1.0024 (0.670×10^{-2})	1.5053 (0.673×10^{-2})	1.9998 (0.896×10^{-2})

Table 3: Comparison of diagnostic tests for diabetes.

Tests	Fasting	1-hour	2-hour	L_3
$\hat{\eta}$	0.4174	0.5469	0.5300	0.5925
$\hat{\theta}$	(160.0)	(187.0)	(141.0)	(306.5)
$\tilde{\eta}$	0.4203	0.5298	0.5184	0.5634
$\tilde{\theta}$	(142.2)	(198.5)	(145.7)	(311.5)