*Biometrika* (1989), **76**, 3, *pp.* 585–92
*Printed in Great Britain*

# A family of nonparametric statistics for comparing diagnostic markers with paired or unpaired data

BY SAM WIEAND

*Department of Epidemiology and Statistics, Mayo Clinic, Rochester, Minnesota 55905, U.S.A.*

MITCHELL H. GAIL

*Epidemiologic Methods Section, National Cancer Institute, Bethesda, Maryland 20892, U.S.A.*

BARRY R. JAMES AND KANG L. JAMES

*Instituto de Matemática Pura e Aplicado, 22460 Rio de Janeiro, Brazil*

## SUMMARY

In this paper we study a broad class of nonparametric statistics for comparing two diagnostic markers. One can compare the sensitivities of these diagnostic markers over restricted ranges of specificity by selecting an appropriate statistic from this class. As special cases, one can compare the entire area under the receiver-operator curve (Hanley & McNeil, 1982), or one can compare the sensitivities at a fixed common specificity. Usually we would recommend a comparison based on an average of sensitivities over a restricted high level of specificities. Test procedures and confidence intervals are based on asymptotic normality. These procedures are applicable for paired data, in which both diagnostic markers are performed on each subject, and for unpaired data. The procedures may be used to compare two real functions of multiple diagnostic markers as well as to compare individual markers.

*Some key words*: Hanley–McNeil statistic; Quantile process; ROC curve; Sensitivity; Specificity.

## 1. INTRODUCTION

Consider a comparison of two diagnostic markers $\text{DM}_1$ and $\text{DM}_2$ that yield continuous measurements. If we use both diagnostic markers on the same $m$ controls and $n$ cases, we can represent the bivariate outcomes as $(X_{1j}, X_{2j})$ $(j=1,\ldots,m)$ and $(Y_{1k}, Y_{2k})$ $(k=1,\ldots,n)$ respectively. We denote the respective bivariate distributions by $F(x_1, x_2)$ and $G(y_1, y_2)$, and the marginals by $F_i(x_i)$ and $G_i(y_i)$, and we assume that the $m+n$ bivariate vectors are mutually independent. If a diagnostic test, $(\text{DM}, c)$, associated with a marker $M$ is called 'positive' if DM exceeds a cutoff value $c$, then the specificity of $(\text{DM}_i, c)$ is $F_i(c)$ and the corresponding sensitivity is $1-G_i(c)$. As $c$ takes on all possible values, $\text{DM}_i$ generates a locus of $\{F_i(c), 1-G_i(c)\}$ versus $c$ as in Fig. 1. Some investigators plot $\{1-F_i(c), 1-G_i(c)\}$ instead and use the term receiver operating characteristic curve (Swets & Pickett, 1982). Bamber (1975) noted that the area under this curve is equal to $\text{pr}(Y>X)$.

The advantage of a plot like Fig. 1 is that it permits one to compare the sensitivities of two markers for any particular specificity or range of specificities. Hanley & McNeil

(1982, 1983), McClish (1987) and DeLong, DeLong & Clarke-Pearson (1988) have advocated comparisons of markers based on the difference of areas under these curves on the ground that the marker with higher sensitivity will tend to have the greater area. However, many medical applications require that the diagnostic test have high specificity. Perhaps for this reason, Greenhouse & Mantel (1950) and Linnet (1987) considered the comparison of two sensitivities at a single fixed level of specificity, $p_0$, namely $\{1-G_1(\xi_{1p_0})\}-\{1-G_2(\xi_{2p_0})\}$, where $\xi_{ip}=F_i^{-1}(p)$. In this paper we present a class of statistics for such comparisons based on a weighted average of sensitivities. Denoting the sensitivity at specificity $p$ by $S_i(p)=1-G_i\{\xi_{ip}\}$, we express this comparison as

$$\Delta=\Delta_W=\int_0^1\{S_1(p)-S_2(p)\}\,dW(p), \tag{1·1}$$

where $W$ is a probability measure on the open unit interval. In particular the proposal of Greenhouse & Mantel (1950) corresponds to point mass at $p_0$ and that of Hanley & McNeil (1982, 1983) to $W(p)=p$, for $0<p<1$. We would often favour a weighting

$$W(p)=0 \quad (0<p<p_1), \quad W(p)=(p-p_1)/(p_2-p_1) \quad (p_1\leqslant p\leqslant p_2),$$
$$W(p)=1 \quad (p_2<p<1).$$

In many cases this third alternative is appealing because $p_1$ and $p_2$ can be confined to a range of usefully high specificities and the test statistics will often have greater power than a statistic based on a single point of comparison, $p_0$.

## 2. AN ESTIMATE OF $\Delta$ AND ITS VARIANCE

The natural nonparametric estimate of $\Delta$ is

$$\hat{\Delta}=\int_0^1\{\hat{S}_1(p)-\hat{S}_2(p)\}\,dW(p), \tag{2·1}$$

where $\hat{S}_i(p)=1-\hat{G}_i(\hat{\xi}_{ip})$, $\hat{G}_i$ is the empirical distribution of $G_i$, and the sample quantile $\hat{\xi}_{ip}$ is the $[mp]$th order statistic among the $m$ values of $X_i$, where $[mp]$ is the smallest integer that equals or exceeds $mp$. Note that $\hat{\Delta}$ is a natural test statistic for testing the hypothesis $\Delta=0$ versus the alternative $\Delta>0$. The following theorem gives the asymptotic distribution of $\hat{\Delta}$ under general assumptions. Its proof is outlined in the Appendix.

THEOREM 2·1. *Define*

$$s_i(p)=S_i'(p)=-\frac{G_i'(\xi_{ip})}{F_i'(\xi_{ip})}, \quad p\wedge q=\min(p,q), \quad p\vee q=\max(p,q).$$

*Suppose that $W$ is a probability measure in $(0,1)$ and that there exists $\varepsilon>0$ such that $W$ has a bounded derivative in $(0,\varepsilon)$ and $(1-\varepsilon,1)$. Suppose further that $G_i(\xi_{ip})$, for $i=1,2$, have continuous derivatives in $(0,1)$ which are monotone in $(0,\varepsilon)$ and $(1-\varepsilon,1)$. Then as $N=n+m$ tends to $\infty$ with $m/N\to\lambda$, for $0<\lambda<1$, $N^{\frac{1}{2}}(\hat{\Delta}-\Delta)$ tends to a normal distribution with variance $\sigma^2=\sigma_{11}-2\sigma_{12}+\sigma_{22}$, where*

$$\sigma_{ii}=\int_0^1\int_0^1[S_i(p\vee q)\{1-S_i(p\wedge q)\}/(1-\lambda)$$
$$+s_i(p)s_i(q)(p\wedge q-pq)/\lambda]\,dW(p)\,dW(q), \tag{2·2}$$

$$\sigma_{12} = \int\int [G(\xi_{1p}, \xi_{2q}) - \{1 - S_1(p)\}\{1 - S_2(q)\}]\, dW(p)\, dW(q)/(1-\lambda)$$
$$+ \int\int \{F(\xi_{1p}, \xi_{2q}) - pq\} s_1(p) s_2(q)\, dW(p)\, dW(q)/\lambda. \quad (2{\cdot}3)$$

Theorem 2·1 may be used to derive the variance of the statistic used by Greenhouse & Mantel (1950) and to compare the difference in areas under the entire sensitivity curves. In particular, if $W$ has unit mass at $p_0$,

$$\sigma_{ii} = (1-\lambda)^{-1} S_i(p_0)\{1 - S_i(p_0)\} + \lambda^{-1} s_i^2(p_0)\, p_0(1-p_0),$$
$$\sigma_{12} = (1-\lambda)^{-1}[G(\xi_{1p_0}, \xi_{2p_0}) - \{1 - S_1(p_0)\}\{1 - S_2(p_0)\}]$$
$$+ \lambda^{-1}\{F(\xi_{1p_0}, \xi_{2p_0}) - p_0^2\} s_1(p_0) s_2(p_0),$$

(Greenhouse & Mantel, 1950). For $W$ uniform on $(0, 1)$, equations (2·2) and (2·3) are asymptotically equivalent to those of DeLong et al. (1988) as outlined in the Appendix.

Consistent estimates of $\sigma^2$ are obtained by substituting empirical distributions and quantiles into (2·2) and (2·3), and by replacing $s_i(p)$ by $\hat{s}_i(p) = m\{\hat{G}(\hat{\xi}_{ip}^-) - \hat{G}(\hat{\xi}_{ip})\}$ when $W$ is continuous. Here $\hat{\xi}_{ip}^-$ is the $([mp]-1)$th order statistic of the sample of values $X_i$. For $W$ discrete at the point $p$, we let

$$\hat{s}_i(p) = \{\hat{S}_i(p+h) - \hat{S}_i(p-h)\}/(2h), \quad (2{\cdot}4)$$

where $h$ tends to zero as $N \to \infty$.

## 3. A PARAMETRIC TEST FOR A DIFFERENCE IN AREAS UNDER THE SENSITIVITY CURVES

If $X$ is $N(\mu_x, \sigma_x^2)$, $Y$ is $N(\mu_y, \sigma_y^2)$ and $X$ and $Y$ are independent, then the area under the sensitivity curve, $\mathrm{pr}\,(Y > X)$, is $\Phi(\delta)$, where

$$\delta = (\mu_y - \mu_x)(\sigma_x^2 + \sigma_y^2)^{-\frac{1}{2}},$$

which can be estimated by

$$\hat{\delta} = (\bar{y} - \bar{x})(\hat{\sigma}_x^2 + \hat{\sigma}_y^2)^{-\frac{1}{2}},$$

where $\bar{x} = \Sigma X_j/m$, $\hat{\sigma}_x^2 = (\Sigma X_j^2 - m\bar{x}^2)/(m-1)$ and so on. Reiser & Guttman (1986), discuss this estimator in some detail and provide earlier references.

Under normality and with $W(p) = p$, for $0 < p < 1$, the hypothesis $\Delta_W = 0$ corresponds to the hypothesis $\delta_1 - \delta_2 = 0$. A natural statistic for testing this hypothesis is $T = \hat{\delta}_1 - \hat{\delta}_2$, whose asymptotic variance can be obtained using the delta method (Rao, 1973, p. 388). Defining

$$\sigma_i^2 = \sigma_{x_i}^2 + \sigma_{y_i}^2, \quad C_x = \mathrm{cov}\,(X_{1j}, X_{2j}), \quad C_y = \mathrm{cov}\,(Y_{1k}, Y_{2k}),$$

we find that $N^{\frac{1}{2}}T$ is asymptotically normal with

$$\mathrm{var}\,(N^{\frac{1}{2}}T) = \sigma_T^2 = \sigma_{11} - 2\sigma_{12} + \sigma_{22},$$

where

$$N^{-1}\sigma_{ii} = \sigma_i^{-2}(m^{-1}\sigma_{x_i}^2 + n^{-1}\sigma_{y_i}^2) + \tfrac{1}{2}\delta_i^2\sigma_i^{-4}\{(m-1)^{-1}\sigma_{x_i}^4 + (n-1)^{-1}\sigma_{y_i}^4\},$$
$$N^{-1}\sigma_{12} = (\sigma_1\sigma_2)^{-1}(m^{-1}C_x + n^{-1}C_y) + \tfrac{1}{2}\delta_1\delta_2(\sigma_1\sigma_2)^{-2}\{(m-1)^{-1}C_x^2 + (n-1)^{-1}C_y^2\}. \quad (3{\cdot}1)$$

A consistent estimator of $\sigma_T^2$ is obtained by replacing the parameters in (3·1) by sample estimates such as $\hat{\delta}_i$ and $\hat{\sigma}_x^2$.

## 4. SIMULATIONS

To examine the small-sample behaviour of the hypothesis testing procedure in § 2, we performed simulations in which the $F(x_1, x_2)$ and $G(y_1, y_2)$ were respectively bivariate normal distributions with means $E(X_1) = E(X_2) = 0$, $E(Y_1) = \mu_1$, $E(Y_2) = \mu_2$, and with common variances 1 and covariances $\rho$. Thus $S_i(p) = 1 - \Phi\{\Phi^{-1}(p) - \mu_i\}$, and alternatives $\mu_1 > \mu_2$ may be expressed equivalently as $S_1(p) > S_2(p)$ for all $p \neq 0, 1$. The power of a one-sided test based on the parametric statistic, $T$, is compared with the powers of four procedures based on $\hat{\Delta}$ for four different weight functions in Table 1. One-sided rejection regions were determined using $\hat{\Delta} > 1{\cdot}645\hat{\sigma}$ and $T > 1{\cdot}645\hat{\sigma}_T$. For the point mass at 0·9 we chose $h = \frac{1}{30}$ in (2·4). Note that even for $n = m = 30$, the size of each procedure is near the nominal 0·05 level, and unreported simulations confirm this result for $n = m = 15$ as well.

The powers of the parametric procedure and the procedure based on the uniform (0, 1) weighting are quite comparable. The test based on a point mass at $p_0 = 0{\cdot}9$ has the lowest power, and procedures with uniform weight on (0·8, 0·9) and (0·7, 0·95) have intermediate power. In each case we find that the asymptotic theory yields good predictions of power for these small-sample experiments. Of course, if the distributions are not normal, the parametric procedure can be highly misleading, whereas the other procedures are asymptotically distribution-free.

Table 1. *Theoretical and simulated power with $n = m = 30$ and $\mu_2 = 1{\cdot}282$ for five one-sided tests based on standard deviates*

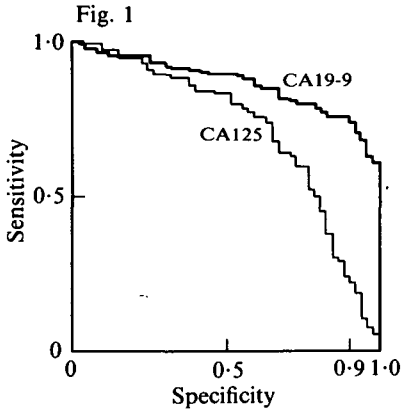
| | | Weighting distribution | | | | |
|---|---|---|---|---|---|---|
| $\rho$ | $\mu_1$ | Point mass at 0·9 | Uniform on (0·80, 0·90) | Uniform on (0·70, 0·95) | Uniform on (0, 1) | Parametric test, $T$ |
| 0 | 1·282 | 0·050, 0·059 | 0·050, 0·052 | 0·050, 0·051 | 0·050, 0·056 | 0·050, 0·047 |
| | 1·956 | 0·34, 0·30 | 0·36, 0·40 | 0·41, 0·44 | 0·45, 0·49 | 0·48, 0·48 |
| | 2·564 | 0·75, 0·65 | 0·72, 0·76 | 0·79, 0·81 | 0·81, 0·87 | 0·88, 0·89 |
| 0·5 | 1·282 | 0·050, 0·052 | 0·050, 0·048 | 0·050, 0·045 | 0·050, 0·048 | 0·050, 0·047 |
| | 1·956 | 0·42, 0·33 | 0·44, 0·47 | 0·53, 0·56 | 0·61, 0·67 | 0·68, 0·68 |
| | 2·564 | 0·82, 0·71 | 0·79, 0·85 | 0·87, 0·92 | 0·89, 0·97 | 0·98, 0·98 |

Left-hand entry, power of procedure based on asymptotic theory; right-hand entry, proportion of rejections in 1000 independent simulations.

Values $\mu_i = 1{\cdot}282$, 1·956 and 2·564 chosen to correspond to $S_i(0{\cdot}9) = 0{\cdot}5$, 0·75 and 0·90 respectively.

## 5. EXAMPLES

Sera from $m = 51$ 'control' patients with pancreatitis and $n = 90$ 'cases' with pancreatic cancer were studied at the Mayo Clinic with a cancer antigen (CA125) (Bast et al., 1983) and with a carbohydrate antigen (CA19-9) (Del Villano et al., 1983). The plots in Fig. 1 of sensitivity against specificity suggest that CA19-9 has higher sensitivity, especially in the region of medical interest with specificity above 0·80. Using a uniform weighting on (0·8, 1·00), we obtained a standardized deviate 4·49, compared to 4·00 for a fixed point mass at $p_0 = 0{\cdot}9$ and to 2·74 for a uniform weight on (0, 1). Thus not only did the uniform (0·8, 1·00) procedure appear to have greater power than a comparison of areas under the entire sensitivity curve, but, more importantly, this procedure demonstrated the superiority of CA19-9 in the specificity range of medical interest.



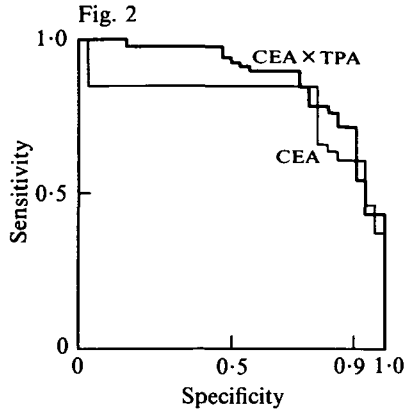




Fig. 1. Sensitivities of two monoclonal antibodies as markers for pancreatic cancer. Observed sensitivity versus specificity of CA125 and CA19-9 used as markers for discriminating between pancreatitis patients, i.e. controls, and pancreatic cancer patients, cases.

Fig. 2. Sensitivities of two antigens as markers for gastric cancer. Observed sensitivity versus specificity of CEA and CEA × TPA used as markers for discriminating between gastritis patients, i.e. controls, and gastric cancer patients, cases.

As another example, two assays, carcinoembryonic antigen, that is CEA, and tissue polypeptide antigen, TPA, were measured on $m = 31$ 'control' patients with benign bowel diseases and $n = 46$ 'cases' with gastrointestinal cancer. The product CEA × TPA appears to discriminate cases from controls somewhat better than CEA alone, shown in Fig. 2, but not in the range of specificity greater than 0·8. The uniform weighting on (0, 1) produces a standardized deviate 2·41, significantly in favour of CEA × TPA, whereas the weighting with point mass at $p_0 = 0{\cdot}9$ yielded a standardized deviate −0·36, slightly in favour of CEA alone and the uniform weighting on (0·8, 1·00) yielded a standardized deviate 0·81. We conclude that the product of CEA × TPA may not be superior to CEA alone in the specificity range of interest. Oehr et al. (1982) also compared CEA × TPA to CEA alone as a diagnostic marker for gastrointestinal cancers, using blood donors as 'control' patients. Their data indicated that CEA × TPA was more sensitive than CEA alone in regions of high specificity, although they did not attempt to assess statistical variability.

Although calculations of $\hat{\Delta}$ based on restricted ranges of specificity are more useful in determining clinical utility than more global assessments, restricting the specificity range also increases the variability of the estimate $\hat{\Delta}$. In the previous example, the variance of $\hat{\Delta}$ for the uniform weighting on (0·80, 1·00) was 0·0064, compared with 0·0011 for the uniform weighting on (0, 1). This suggests that one may wish to use the broadest specificity range consistent with intended applications.

## 6. Discussion

The family of statistics introduced in § 2 is quite general and allows us to restrict the sensitivity comparisons as dictated by practical requirements of specificity. If $W$ corresponds to a point mass, the asymptotic variance of $\hat{\Delta}$ is that obtained by Greenhouse & Mantel (1950). The uniform (0, 1) weighting corresponds to the difference of two correlated two-sample Wilcoxon statistics (Hanley & McNeil, 1983). The quantities $\sigma_{ii}$ and $\sigma_{12}$ in (2·2) and (2·3) are the appropriate asymptotic variances and covariance for these two statistics, as is verified in the Appendix.

Although the hypotheses of Theorem 2·1 exclude ties, our results may still be used if only a few ties result from rounding. In particular, the statistic proposed by Greenhouse & Mantel (1950) is unaffected unless ties occur at the mass point $p_0$, and a statistic based on a $W$ with support only on $(p_1, p_2)$ is unaffected if ties fall outside this interval. The special case of $W$ uniform on $(0, 1)$ can be treated by using modified Wilcoxon scores 1 if $Y > X$, $\frac{1}{2}$ if $Y = X$ and 0 if $Y < X$, as described in the Appendix. For other weighting functions and distributions of ties, one may choose to break ties at random or to compute the average standardized deviate $\hat{\Delta}/\hat{\sigma}$ over all possible ways that ties could be broken.

As illustrated by the second example, we can compare any two real functions of multiple measurements using the methods in this paper. For example two linear discriminants based on multiple assays could be compared.

If $X_{1j}$ and $X_{2j}$ are independent and if $Y_{1k}$ and $Y_{2k}$ are independent, as when independent samples of cases and controls are used to evaluate the two diagnostic tests, then Theorem 2·1 may be used with $\sigma_{12} = 0$. In this circumstance we must be aware of biases that result from noncomparability of the two case groups and two control groups.

## ACKNOWLEDGEMENT

The authors would like to thank Dr Bertil Björkland, who generously provided the data for the second example, and Doug Midthune and Steven Cha for valuable assistance in programming, Jim Hanley and Jim Goin for helpful discussions, and a referee for numerous constructive suggestions. This work was supported in part by the Mayo Comprehensive Cancer Center Grant, National Cancer Institute.

## APPENDIX

### *Outline of the proof of Theorem* 2·1 *and results for the paired Wilcoxon test*

*Proof of Theorem* 2·1. The proof of Theorem 2·1 uses techniques for dealing with $R$- and $L$-statistics, based on empirical and quantile representations, similar to those of Shorack & Wellner (1986, Ch. 19). A detailed proof is available from the authors, but only a heuristic outline is presented here.

The key step in the proof is to note that

$$\begin{aligned}\int [\{\hat{S}_1(p) - \hat{S}_2(p)\} - \{S_1(p) - S_2(p)\}]\, dW(p) \\ = \int [\{\hat{G}_2(\xi_{2p}) - G_2(\xi_{2p})\} - \{\hat{G}_1(\xi_{1p}) - G_1(\xi_{1p})\}]\, dW(p) \\ + \int [\{G_2(\hat{\xi}_{2p}) - G_2(\xi_{2p})\} - \{G_1(\hat{\xi}_{1p}) - G_1(\xi_{1p})\}]\, dW(p) \\ - \int [\{\hat{G}_1(\hat{\xi}_{1p}) - G_1(\hat{\xi}_{1p})\} - \{\hat{G}_1(\xi_{1p}) - G_1(\xi_{1p})\}]\, dW(p) \\ + \int [\{\hat{G}_2(\hat{\xi}_{2p}) - G_2(\hat{\xi}_{2p})\} - \{\hat{G}_2(\xi_{2p}) - G_2(\xi_{2p})\}]\, dW(p).\end{aligned}$$

The third and fourth terms can be shown to be of lower order in probability than the first two terms. To see this, note that the approximation theorem for uniform empirical processes of Komlós, Major & Tusnády (1975, p. 113) implies that the processes $\alpha_{in}$ defined by $\alpha_{in} = n^{\frac{1}{2}}[\hat{G}_i\{G_i^{-1}(p)\} - p]$, for $0 < p < 1$, can be approximated by standard Brownian bridges $b_{in}$ in such

a way that

$$\mathrm{pr}\left(\sup_{0<p<1}|\alpha_{in}(p)-B_{in}(p)|\geqslant an^{-\frac{1}{2}}\log n\right)\leqslant bn^{-c},$$

for all $n$, some positive constants $a$, $b$ and $c$, and $i=1,2$. This, together with the fact that $\sup_p|B_{in}\{G_i(\hat{\xi}_{ip})\}-B_{in}\{G_i(\xi_{ip})\}|$ converges to 0 in probability, for $i=1,2$, implies that the third and fourth terms are $o(N^{-\frac{1}{2}})$ in probability.

The first two terms are independent, since the first depends only on the $Y$'s and the second on the $X$'s. The first term is the difference of two correlated sample means; its variance can be calculated via Fubini's theorem and is equal to $n^{-1}\sigma_1^2$, where

$$\sigma_1^2=\int\int[S_1(p\vee q)\{1-S_1(p\wedge q)\}+S_2(p\vee q)\{1-S_2(p\wedge q)\}]\,dW(p)\,dW(q)$$
$$-2\int\int G(\xi_{1p},\xi_{2q})\,dW(p)\,dW(q)$$
$$+2\int\int\{1-S_1(p)\}\{1-S_2(q)\}\,dW(p)\,dW(q).$$

The above integrals are taken over the unit square.

The second term is the difference between two correlated $L$-statistics. Its asymptotic variance can be obtained heuristically by observing that, by Taylor's theorem, the term is essentially equal to

$$\int[s_2(p)\{F_2(\hat{\xi}_{2p})-p\}-s_1(p)\{F_1(\hat{\xi}_{1p})-p\}]\,dW(p).$$

A result of Kiefer (1970) allows us to replace the quantile process $F_i(\hat{\xi}_{ip})-p$ by the empirical process $p-\hat{F}_i(\xi_{ip})$. This procedure yields another difference of two correlated sample means, whose variance can again be calculated via Fubini's theorem and equals $m^{-1}\sigma_2^2$, where

$$\sigma_2^2=\int\int\{s_1(p)s_1(q)+s_2(p)s_2(q)\}(p\wedge q-pq)\,dW(p)\,dW(q)$$
$$-2\int\int F(\xi_{1p},\xi_{2q})s_1(p)s_2(q)\,dW(p)\,dW(q)$$
$$+2\int\int pqs_1(p)s_2(q)\,dW(p)\,dW(q).$$

This variance can also be obtained following Shorack & Wellner's (1986, especially p. 660–4) derivation of the asymptotic distribution of $L$-statistics, at least for the case of differentiable $W$, by following their steps for the corresponding difference between $L$-statistics.

The variance given in Theorem 2·1 can be obtained from the above variances by rearranging the terms.

Pyke & Shorack (1968) prove a similar result in the one-dimensional case, and it may be that their result could be adapted for the present proof.

*Results for the paired Wilcoxon test.* When $W$ is the uniform (0, 1) weight function,

$$mn\hat{\Delta}=mn\int\{\hat{S}_2(p)-\hat{S}_1(p)\}\,dp=\Sigma\,\Sigma\{\psi(Y_{2k},X_{2j})-\psi(Y_{1k},X_{1j})\},$$

where $\psi(Y,X)=1$ if $Y>X$ and 0 otherwise. The variance of the latter form has been obtained by Wieand, Gail & Hanley (1983) and, independently, by DeLong et al. (1988), using results

from the theory of $U$-statistics. We use the notation of the latter authors, who define $\theta_r = E\{\psi(Y_r, X_r)\}$,

$$\xi_{10}^{rs} = E\{\psi(Y_{rk}, X_{rj})\psi(Y_{sk'}, X_{sj})\} - \theta_r\theta_s \quad (k' \neq k),$$
$$\xi_{01}^{rs} = E\{\psi(Y_{rk}, X_{rj})\psi(Y_{sk}, X_{sj'})\} - \theta_r\theta_s \quad (j' \neq j).$$

It follows from their equations (2) and (4) that the asymptotic variance of $N^{\frac{1}{2}}\hat{\Delta}$ is

$$\sigma_L^2 = \lambda^{-1}(\xi_{10}^{11} + \xi_{10}^{22} - 2\xi_{10}^{12}) + (1-\lambda)^{-1}(\xi_{01}^{11} + \xi_{01}^{22} - 2\xi_{01}^{12}).$$

To show the equivalence of our $\sigma^2$ with $\sigma_L^2$ when $W$ is the uniform weight function, we substitute $x_i = F_i^{-1}(p)$ and $y_i = F_i^{-1}(q)$ in our (2·2) and after some manipulation obtain the required forms for $\sigma_{ii}$ and $\sigma_{12}$.

Each of the terms in the Wilcoxon variance is a probability which can be estimated by a mean of indicator functions. Ties, or even grouped data, can be handled by defining $\psi(Y, X) = 0{\cdot}5$ if $Y = X$ and adjusting the variance accordingly. The variance estimates, which are obtained using a method of Sen (1960), are presented by DeLong et al. (1988).

## REFERENCES


BAMBER, D. (1975). The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *J. Math. Psychol.* **12**, 387–415.

BAST, R. C., KLUG, T. L., ST. JOHN, E., JENISON, E., NILOFF, J. M., LAZARUS, H., BERKOWITZ, R. S., LEAVITT, T., GRIFFITHS, C. T., PARKER, L., ZURAWSKI, V. R. & KNAPP, R. C. (1983). Radioimmunoassay using a monoclonal antibody to monitor the course of epithelial ovarian cancer. *New England J. Med.* **309**, 883–7.

DELONG, E. R., DELONG, D. M. & CLARKE-PEARSON, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* **44**, 837–45.

DEL VILLANO, B. C., BRENNAN, S., BROCK, P., BUCHER, C., LIU, V., MCCLURE, M., RAKE, M., SPACE, B. & ZURAWSKI, V. R. (1983). Radioimmunometric assay for a monoclonal antibody-defined tumor marker, CA19-9. *Clin. Chemistry* **29**, 549–52.

GREENHOUSE, S. W. & MANTEL, N. (1950). The evaluation of diagnostic tests. *Biometrics* **6**, 399–412.

HANLEY, J. A. & MCNEIL, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* **143**, 29–36.

HANLEY, J. A. & MCNEIL, B. J. (1983). A method of comparing the areas under receiver operating characteristics curves derived from the same cases. *Radiology* **148**, 839–43.

KIEFER, J. (1970). Deviations between the sample quantile process and the sample df. In *Nonparametric Techniques in Statistical Inference*, Ed. M. L. Puri, pp. 299–319. Cambridge University Press.

KOMLÓS, J., MAJOR, P. & TUSNÁDY, G. (1975). An approximation of partial sums of independent RV's, and the sample DF.I. *Z. Wahr. verw. Geb.* **32**, 111–31.

LINNET, K. (1987). Comparison of quantitative diagnostic tests: type 1 error, power, and sample size. *Statist. Med.* **6**, 147–58.

MCCLISH, D. K. (1987). Comparing the areas under more than two independent ROC curves. *Medical Decision Making* **7**, 149–55.

OEHR, P., FISCHER, L., KERSJES, W., KUNATH, U., BIERSACK, H. J., SIPEER, U. & WINKLER, C. (1982). Verteilung, Sensitivität und Spezifität von TPA, CEA und CEA×TPA Marker-Produktwerten bei Patienten mit gastrointestinalem Carcinom. *Tumor Diagnostik Therapie* **3**, 195–8.

PYKE, R. & SHORACK, G. (1968). Weak convergence of a two-sample empirical process and a new approach to Chernoff–Savage theorems. *Ann. Math. Statist.* **39**, 755–71.

RAO, C. R. (1973). *Linear Statistical Inference and its Applications*, 2nd ed. New York: Wiley.

REISER, B. & GUTTMAN, I. (1986). Statistical inference for pr $(Y < X)$: the normal case. *Technometrics* **28**, 253–7.

SEN, P. K. (1960). On some convergence properties of $U$-statistics. *Calcutta Statist. Assoc. Bull.* **10**, 1–18.

SHORACK, G. R. & WELLNER, J. A. (1986). *Empirical Processes with Applications to Statistics.* New York: Wiley.

SWETS, J. A. & PICKETT, R. M. (1982). *Evaluation of Diagnostic Systems. Methods from Signal Detection Theory.* New York: Academic.

WIEAND, H. S., GAIL, M. M. & HANLEY, J. A. (1983). A nonparametric procedure for comparing diagnostic tests with paired or unpaired data. *I.M.S. Bull.* **12**, 213–4.


[*Received June* 1988. *Revised November* 1988]