Notions of Limiting P Values Based on Data Depth and Bootstrap
Author(s): Regina Y. Liu and Kesar Singh
Source: *Journal of the American Statistical Association*, Vol. 92, No. 437 (Mar., 1997), pp. 266-277
Published by: American Statistical Association
Stable URL: http://www.jstor.org/stable/2291471
Accessed: 05/03/2014 21:40

Notions of Limiting P Values Based on Data Depth and Bootstrap

Regina Y. LIU and Kesar SINGH

In this article we propose some new notions of limiting P values for hypothesis testing. The limiting P value (LP) here not only provides the usual attractive interpretation of a P value as the strength in support of the null hypothesis coming from the observed data, but also has several advantages. First, it allows us to resample directly from the empirical distribution (in the bootstrap implementations), rather than from the estimated population distribution satisfying the null constraints. Second, it serves as a test statistic and as a P value simultaneously, and thus enables us to obtain test results directly without having to construct an explicit test statistic and then establish or approximate its sampling distribution. These are the two steps generally required in a standard testing procedure. Using bootstrap and the concept of data depth, we have provided LP's for a broad class of testing problems where the parameters of interest can be either finite or infinite dimensional. Some computer simulation results show the generality and the computational feasibility of our approach.

KEY WORDS: Bootstrap; Data depth; Hypothesis testing; Limiting P value.

1. INTRODUCTION

Consider a hypothesis testing problem with the null hypothesis H: $F \in \Omega_1$ versus K: $F \in \Omega_2$. The so-called P value is used to assess the strength of the observed evidence in support of H. The P value is also the attained level of significance of a given test, meaning that for any preset level of significance α, the given test will lead to the rejection of H if the obtained P value is less than or equal to α. To ensure that a P value is the attained level of a test, the classical P value, denoted by p_n^c, is defined as

$$p_n^c = \sup_{F \in \Omega_1} l_n(F), \qquad (1)$$

where $l_n(F)$ stands for the attained level of the test when F prevails. Here F is the population distribution from which the n random sample points are drawn. Except for some cases where the null space is specifically restricted, the exact computation of p_n^c is extremely difficult. To avoid such a difficulty, statisticians generally use large-sample approximations for the P value, without paying much attention to the uniformity of the approximation in the null space $F \in \Omega_1$.

In this article we broaden the definition of P values to a class of measures that assess the evidence supporting H contained in the observed data. To encompass the definition of the classical P value, we require this more general P value to satisfy a test that rejects H if P value $\leq \alpha$ will have its type I error probability $\leq \alpha$, at least asymptotically. Mathematically, this requirement may be viewed as one of the following two conditions: (a) type I error probability is $\leq \alpha$ for all large n and for any fixed $F \in \Omega_1$, or (b) the supremum of type I error probability over the null space $F \in \Omega_1$ is $\leq \alpha$, for all large n. Consequently, there are two definitions of this more general notion of P value proposed in Section 2. The first definition is termed the *limiting P value* (LP), which ensures condition (a); the second is termed the *limiting P value in the strong sense* [LP(S)], which ensures condition (b). Note that the classical P value is in fact a LP(S), and the commonly used P values based on large sample approximations are often only LP's. Usually, most of the notions of LP can be strengthened to LP(S) by further restricting the class of distributions belonging in H. The general notion of LP introduced here is to be used in hypothesis testing in exactly the same way as the classical P value, even though the two may not have the same interpretation. One advantage of using the concept of LP—or LP(S)—is that it allows us to combine the standard three-step test procedure—namely, constructing a test statistic, establishing its sampling distribution, and finally obtaining the P value—into a one-step procedure. In other words, LP or LP(S) itself serves as a test statistic as well as a P value.

Using LP or LP(S) as the new extended notion of P value, we are able to obtain sensible testing procedures for a wide range of hypothesis testing problems whose parameters of interest are allowed to be finite or infinite dimensional. By using bootstrap techniques and the notion of data depth, we present in Section 3 some methods for obtaining LP's for testing problems where the null space is a specified single value of the parameter. We extend this approach in Section 4 to deal with the situation in which the parameter space under H is a region rather than a singleton. Furthermore, we point out in Section 4 that the probability of the observed bootstrap values of the parameter falling in the null region provides a notion of LP, and it clearly indicates the support for H coming from the observed data. This bootstrap probability is termed the *empirical strength probability* (ESP) of H. This ESP is shown to be a LP if the null space is a smooth region. The proof here involves a new bootstrap limit result, which we refer to as the *bootstrap convolution lemma.* In Section 5 we extend all approaches discussed in Sections 3 and 4 further to cover the testing of infinite-dimensional parameters such as cdf's and other

Regina Y. Liu and Kesar Singh are Professors, Department of Statistics, Rutgers University, Piscataway, NJ 08855. The authors gratefully acknowledge the support from National Science Foundation grants DMS-90-04658, DMS-90-22126, and DMS-92-24839. The authors thank Kay Tatsuoka for his computing assistance and the referees, associate editor, and editor for their helpful comments.

© 1997 American Statistical Association
Journal of the American Statistical Association
March 1997, Vol. 92, No. 437, Theory and Methods

statistical curves. In Section 6 we give some computer simulations for various LP's discussed in Sections 3 and 4 and present a power comparison between our ESP method and the standard t test for testing the correlation coefficient in a bivariate normal distribution. The simulation results appear to be quite supportive of our proposals. We give some concluding remarks in Section 7, and provide all technical proofs in the Appendix.

There is an extensive literature on the classical P value. In particular, Bahadur (1971) gave a thorough treatise on the subject, and Lambert (1981) provided a concise account of the properties of both exact and approximate P values. Some bootstrap methods for obtaining approximate P values were given by Beran (1986), Loh (1985), and Romano (1988). Unlike our method, these earlier methods need to resample from an approximate distribution that obeys the constraints of the null hypothesis. In other words, they essentially bootstrap from a member of the null space under H that is "closest" to the empirical distribution, whereas we bootstrap directly from the empirical distribution. This alternative bootstrap P value was initiated by Singh and Berk (1994). Recently, Berger and Boos (1994) gave account and a definition of P values in the presence of nuisance parameter.

Finally, it is not difficult to see that all proposed LP's in this article tend to zero under their respective alternative hypotheses, since the test statistics in those situations are no longer properly centered. We shall avoid repeating this observation when each individual LP is discussed later.

2. DEFINITIONS

Throughout the article, we let $X_1, \ldots, X_n$, possibly multivariate, denote a random sample from a population with cdf F. Consider testing H: $F \in \Omega_1$ versus K: $F \in \Omega_2$. Let p_n be a statistic defined on $X_1, \ldots, X_n$.

Definition 2.1. (I) A sequence of statistics p_n is called a *limiting P value*, denoted by LP, if $p_n \in [0,1]$ and p_n satisfies the following:

(a) $\limsup_{n\to\infty} P_F(p_n \le t) \le t$, for all $F \in \Omega_1$ and for any $t \in [0,1]$.
(b) $p_n \to 0$ in probability for all $F \in \Omega_2$.

(II) A sequence p_n is called a *limiting P value in strong sense*, denoted by LP(S), if (a) is replaced by the following stronger requirement:

(a′) $\limsup_{n\to\infty} \sup_{F\in\Omega_1} P(p_n \le t) \le t$, for any $t \in [0,1]$.

Typically, to show that a given sequence p_n is a LP one needs to show that for any F, p_n converges weakly to a random variable, say Z_F, where Z_F is stochastically larger or equal to $U[0,1]$ for all $F \in \Omega_1$, and Z_F degenerates to 0 for all $F \in \Omega_2$. Here $U[0,1]$ stands for a uniform random variable with support $[0,1]$. If the foregoing weak convergence is made uniform over $F \in \Omega_1$, then p_n is a LP(S). Note that under Definition 2.1 the classical P value defined in (1) is a LP(S) (see, e.g., Bahadur 1971), provided that the underlying test is a consistent one; namely, under $F \in \Omega_2$, the power of the test tends to 1 if α is held fixed. Note that in reality, the P values presented by statisticians in most tests are derived from the limiting null distributions of test statistics, and they are only approximations of the true P values. As shown in Example 2.1, these approximate P values are usually LP's. In principle, a LP can become a LP(S) if the null space is reduced to some proper subset, as clearly demonstrated in Example 2.1.

Definition 2.1 can be motivated by considering type I and type II errors as follows: Consider the size α test ϕ_n derived from p_n; that is,

$$\phi_n = \begin{cases} 1 & \text{iff } p_n \le \alpha, \\ 0 & \text{otherwise,} \end{cases}$$

where $\phi_n = 1$ stands for the decision of rejecting H. If p_n is a LP sequence, then for any $F \in \Omega_1$

$$P_F(\text{committing type I error by using test } \phi_n) = P_F(p_n \le \alpha) \le \alpha$$

for all large n, and for any $F \in \Omega_2$,

$$P_F(\text{committing type II error by using test } \phi_n) = P_F(p_n > \alpha),$$

which clearly tends to 0. Note that in the classical setting, the probability of type I error of the test ϕ_n is actually defined as $\sup_{F\in\Omega_1} P_F(\phi_n = 1)$, which in this case is $\sup_{F\in\Omega_1} P_F(p_n \le \alpha)$. This in turn is $\le \alpha$ for all large n if the condition (a′) in Definition 2.1 for LP(S) is satisfied. In the following example, we list some commonly used "P values" that are actually only LP's.

Example 2.1. Consider testing the population mean μ in H: $\mu = \mu_0$ versus K: $\mu > \mu_0$, where $\mu_0 \in \mathbb{R}$. Though H here seems to be a simple null hypothesis, it is actually not. For instance, even if we only concentrate on the distributions with finite variances, we still have the following two *classes* of distributions:

$$H: \left\{ F \,\middle|\, \int x\,dF = \mu_0 \quad \text{and} \quad \int x^2\,dF < \infty \right\}$$

and

$$K: \left\{ F \,\middle|\, \int x\,dF > \mu_0 \quad \text{and} \quad \int x^2\,dF < \infty \right\}.$$

There are at least four, listed in (2)–(5), commonly "accepted" approximate P values. We briefly describe them and observe that they are all LP's. Some new notations are needed before we proceed. Let Φ denote the cdf of the standard normal distribution, and let $\bar{X}$ and S_n denote the sample mean and sample standard deviation of X_i's, respectively. Let $X_1^*, \ldots, X_n^*$ be a bootstrap sample obtained by resampling with replacement from X_i's, and let $P^*(\cdot)$ stand for the bootstrap probability distribution conditional on X_i's. The bootstrap counterparts of $\bar{X}$ and S_n are indicated by $\bar{X}^*$ and S_n^*.

Consider

$$p_n \equiv 1 - \Phi\left(\frac{\bar{X} - \mu_0}{S_n/\sqrt{n}}\right). \tag{2}$$

This P value stems from the central limit theorem of the studentized test statistic.

Consider

$$p_n \equiv P^*(\bar{X}^* \leq \mu_0). \tag{3}$$

This is referred to as a type-2 P value in Singh and Berk (1994).

Consider

$$p_n \equiv P^*(\bar{X}^* > 2\bar{X} - \mu_0), \tag{4}$$

or

$$p_n \equiv P^*\left(\frac{\bar{X}^* - \bar{X}}{S_n^*/\sqrt{n}} > \frac{\bar{X} - \mu_0}{S_n/\sqrt{n}}\right). \tag{5}$$

Both (4) and (5) are derived from various bootstrap approximations of relevant sampling distributions, (4) is from the so-called hybrid bootstrap method and (5) from the so-called bootstrap t. (See Hall 1988 for a detailed discussion of various bootstrap methods, and see Efron 1982 for the introduction and general description of bootstrap.)

It can be shown that each of the above four p_n's converges weakly to $U[0,1]$ for any fixed F in H, and it degenerates to zero in limit under any alternative hypothesis. Here $U[0,1]$ denotes a uniform distribution with support the unit interval (0, 1). Hence these p_n's are LP's. However, they are not LP(S) unless the null space is further restricted. For instance, if H satisfies additional moment conditions such that

$$H\colon \left\{F \,\middle|\, \int x\,dF = \mu_0, \sigma_F^2 \geq \delta_1 \quad\text{and}\quad \int x^{2+\delta_2}\,dF \right.$$
$$\left. \leq c \text{ for some positive } \delta_1 \text{ and } \delta_2 \text{ and for some } c < \infty \right\},$$

then all of the foregoing p_n sequences become LP(S)'s.

We remark here that the foregoing observations and discussion on Example 2.1 remain valid even if $\mu = \mu_0$ in H is replaced by the more general statement $\mu \leq \mu_0$. A general result pertaining to LP(S) for testing the same set of hypotheses is given as Theorem 4.2 in Section 4.

3. A LIMITING P VALUE BASED ON DATA DEPTH FOR H: $\theta_F = \theta_0$ VERSUS K: $\theta_F \neq \theta_0$

Let $X_1, \ldots, X_n$ be a random sample from F, a d-dimensional distribution, $d \geq 1$, and let θ_F be a finite-dimensional functional of F. Consider testing H: $\theta_F = \theta_0$ versus K: $\theta_F \neq \theta_0$, where θ_0 is fixed. Let $\theta_n \equiv \theta_n(X_1, \ldots, X_n)$ be an estimate of θ_F and let $\theta_n^* \equiv \theta_n(X_1^*, \ldots, X_n^*)$ be a bootstrap estimate of θ_F, where $X_1^*, \ldots, X_n^*$ is a bootstrap sample drawn with replacement from $X_1, \ldots, X_n$. We present in this section a LP based on the bootstrap and the notion of data depth. The motivation here is to look at the bootstrap distribution of θ_n^* as a plausible distribution to which θ belongs, conditional on the given data. This line of thinking is very similar to the Bayesian viewpoint of posterior distributions. In other words, we examine how plausible it is for θ_F to assume the value θ_0 if θ_F actually followed the bootstrap distribution of θ_n^*. To facilitate such an examination, we need to apply some notion of centrality or data depth to the bootstrap distribution to determine the fraction of possible θ_F values that are more outlying (or less plausible, so to speak) than θ_0. The LP discussed in this section is nothing but this fraction, and its precise definition appears after the following brief description of some well-known notions of data depth.

Given observations $W_1, \ldots, W_m$ from the distributions Ψ in $\mathbb{R}^k$, here are several ways to measure the depth (or the centrality) of a given point $w \in \mathbb{R}^k$ with respect to Ψ or with respect to the data cloud $\{W_1, \ldots, W_m\}$:

- *Mahalanobis depth* ($\mathrm{M_hD}$) (Mahalanobis 1936) at w with respect to Ψ is defined to be
$$\mathrm{M_hD}(\Psi; w) = [1 + (w - \boldsymbol{\mu}_\Psi)'\boldsymbol{\Sigma}_\Psi^{-1}(w - \boldsymbol{\mu}_\Psi)]^{-1},$$
where $\boldsymbol{\mu}_\Psi$ and $\boldsymbol{\Sigma}_\Psi$ are the mean and variance matrix of Ψ. The sample version of $\mathrm{M_hD}$ is obtained by replacing $\boldsymbol{\mu}_\Psi$ and $\boldsymbol{\Sigma}_\Psi$ by their sample counterparts.
- *Tukey's depth* (TD) (Tukey 1975) at a point w with respect to Ψ is defined to be
$$\mathrm{TD}(\Psi; w) = \inf_E\{P(E)\colon E \text{ is a closed half-space in } \mathbb{R}^d \text{ and } w \in E\}.$$
The sample version of $\mathrm{TD}(\Psi; w)$ is $\mathrm{TD}(\Psi_m; w)$. Here Ψ_m denotes the empirical distribution of the sample W_i's.
- *Simplicial depth* (SD) (Liu 1990) at w with respect to Ψ is defined to be
$$\mathrm{SD}(\Psi; w) = P_\Psi\{w \in S[W_1, \ldots, W_{k+1}]\},$$
where $S[W_1, \ldots, W_{k+1}]$ is the closed simple whose vertices $W_1, \ldots, W_{k+1}$ are $(k+1)$ random observations from Ψ. The sample version of $\mathrm{SD}(\Psi; w)$ is $\mathrm{SD}(\Psi_m; w)$.
- *Majority depth* ($\mathrm{M_jD}$) (proposed by Kesar Singh) of w with respect to Ψ is defined to be
$$\mathrm{M_jD}(\Psi; w) = P_\Psi\{w \text{ is in a major side determined by } (W_1, \ldots, W_k)\}.$$
Here a major side is the half-space with probability $\geq .5$, which is bounded by the hyperplane containing $(W_1, \ldots, W_k)$. The sample version of $\mathrm{M_jD}(\Psi; w)$ is $\mathrm{M_jD}(\Psi_m; w)$.

Some general properties of the foregoing four depths were provided by Liu and Singh (1993). Throughout the rest of the article, $D(\cdot;\cdot)$ is used to indicate any of the four depths unless stated otherwise. The value of $D(\Psi; w)$ may vary when a different notion of data depth is used, but for each notion of depth, the larger the value $D(\Psi; w)$, the deeper (or more central) the w with respect to Ψ. In what follows, we apply $D(\cdot;\cdot)$ to numerous bootstrap estimates of the parameter, θ_n^*'s, to determine the relative outlyingness of these estimates with respect to the hypothesized value θ_0. In this case, Ψ would be the sampling distribution of θ_n^*.

Let G_n and G_n^* denote the sampling distributions for θ_n and θ_n^*. For testing H: $\theta_F = \theta_0$ versus K: $\theta_F \neq \theta_0$, the following proposed p_n is a LP, as established in Theorem 3.1.

Definition 3.1. Let

$$p_n = P_{G_n^*}\{\theta_n^*\colon\ D(G_n^*;\theta_n^*) \le D(G_n^*;\theta_0)\}. \qquad (6)$$

In practice, we would calculate k values of θ_n^*, say $\theta_{n,1}^*, \ldots, \theta_{n,k}^*$. Based on the empirical distribution of these k values, say $G_{n,k}^*$, we then compare each $D(G_{n,k}^*;\theta_{n,\imath}^*)$ to $D(G_{n,k}^*;\theta_0)$ to obtain the fraction of $\theta_{n,\imath}^*$'s that have less depth than θ_0; namely, $k^{-1}\sum_{\imath=1}^{k} I\{D(G_{n,k}^*;\theta_{n,\imath}^*) < D(G_{n,k}^*;\theta_0)\}$. Here $I\{\cdot\}$ is an indicator function with $I(A) = 1$ if A occurs and $I(A) = 0$ otherwise. The θ here can be a location or scale functional. For example, if the $d \times d$ covariance matrix is the parameter of interest, then it is viewed as a $d(d+1)/2$ vector and denoted by θ.

In Theorem 3.1 we claim that p_n defined in (6) converges to $U[0,1]$ under H. Now we introduce some new notations to facilitate stating the theorem.

Let L_n^* and L_n be the distribution of $a_n(\theta_n^* - \theta_n)$ and $a_n(\theta_n - \theta_0)$ for some positive sequence $a_n \to \infty$ as $n \to \infty$. We say that L_n *converges D regularly to the cdf* L, denoted by $L_n \to L, D$ *regularly,* if

(a) $L_n \xrightarrow{\mathcal{L}} L$, i.e. L_n converges weakly to L as $n \to \infty$, and

(b) $\lim_{n\to\infty} \sup_{x\in\mathbb{R}^d} |D(L_n;x) - D(L;x)| = 0$.

Remark 3.1. Note that if there exists some positive sequence a_n such that $L_n \xrightarrow{\mathcal{L}} L$, then this convergence is M$_\text{h}$D regular if $\theta_n \to \theta_0$ a.s. and $S_{\theta_n} \to \mathbf{\Sigma}_{\theta_0}$ a.s. as $n \to \infty$. It is TD-regular or SD-regular if L is absolutely continuous.

Theorem 3.1. Let $L_n \to L, D$-regular, and $L_n^* \to L, D$-regular a.s., where L is a continuous distribution symmetric around 0. Let L be the cdf of the random variable T. Assume that the distribution of $D(L;T)$ is continuous. Then p_n in (6) converges in distribution to $U[0,1]$.

If L here is not symmetric around 0, then Theorem 3.1 does not hold. However, a slight modification of our procedure can ensure the same desired property of p_n. The modification here is to flip θ_n^* around θ_n (see Remark 4.3).

Example 3.1. Consider the testing problem on the mean vector $\boldsymbol{\mu}_F$ in H: $\boldsymbol{\mu}_F = \mu_0$ versus K: $\boldsymbol{\mu}_F \neq \mu_0$, where μ_0 is given. Let D be M$_\text{h}$D. Then $p_n = P^*\{\bar{X}^*\colon (\bar{X}^* - \bar{X})'\mathbf{S}_n^{-1}(\bar{X}^* - \bar{X}) \ge (\bar{X} - \mu_0)'\mathbf{S}_n^{-1}(\bar{X} - \mu_0)\}$, where $\mathbf{S}_n$ is the sample covariance matrix. If the underlying distribution is multivariate normal, then the parametric bootstrap procedure yields

$$p_n = P\{\chi_d^2 \ge (\bar{X} - \mu_0)'\mathbf{S}_n^{-1}(\bar{X} - \mu_0)\},$$

where χ_d^2 stands for a chi-squared distribution with degree of freedom d. This is essentially the Z test stemming from the central limit theorem. Note that in this example if the bootstrap t is used to define p_n; that is, $p_n = P^*\{\bar{X}^*\colon (\bar{X}^* - \bar{X})'\mathbf{S}_n^{*-1}(\bar{X}^* - \bar{X}) \ge (\bar{X} - \mu_0)'\mathbf{S}_n^{-1}(\bar{X} - \mu_0)\}$, where $\mathbf{S}_n^*$ is the bootstrap sample covariance matrix, then this P value is the same as the one associated with the Hotelling T test.

4. LIMITING *P* VALUE AND LIMITING *P* VALUES IN A STRONG SENSE FOR TESTING *H*: $\theta \in \mathcal{R}$ WHERE $\mathcal{R}$ IS A REGION

For testing H: $\theta \in \mathcal{R}$ versus K: $\theta \notin \mathcal{R}$, where $\mathcal{R}$ is a region, we present two methods for obtaining LP's.

Method 4A.

$$p_n^{(A)} = P^*(\theta_n^* \in \mathcal{R}) \equiv \text{ ESP}(\mathcal{R}).$$

The $p_n^{(A)}$ defined here indicates the likelihood that the bootstrap estimate of θ—namely, θ_n^*—lies within the null space $\mathcal{R}$. In essence, it indicates the *empirical* strength of the support for H in light of the observed data. Hence $p_n^{(A)}$ is also denoted by ESP($\mathcal{R}$), indicating the *empirical strength probability* of null space $\mathcal{R}$. Some simulations results are provided in Section 6 to show the generality and the computational ease of ESP in (4.A). These two properties of ESP hold even in cases where parametric tests are no longer applicable, as seen in Section 6.2. In Theorem 4.1 we show that if $\mathcal{R}$ has a smooth boundary, then ESP($\mathcal{R}$) is a LP. It should be pointed out that ESP($\mathcal{R}$) is obviously not meaningful if the Lebesgue measure of $\mathcal{R}$ is 0.

Theorem 4.1. Let $\mathcal{R}$ be a closed-connected region. Let θ_0 be a boundary point of $\mathcal{R}$, where $\mathcal{R}$ admits an unique tangent plane. Assume that there exists a positive sequence $a_n \to \infty$ such that $a_n(\theta_n - \theta_0) \xrightarrow{\mathcal{L}} L$ and $a_n(\theta_n^* - \theta_0) \xrightarrow{\mathcal{L}} L$ a.s., where L is an absolutely continuous distribution symmetric around 0. Then $p_n^{(A)} \xrightarrow{\mathcal{L}} U[0,1]$ as $n \to \infty$, if $\theta = \theta_0$.

Method 4B.

$$p_n^{(B)} = 1 - \min_{0\le\gamma\le1}\{\gamma\colon\ \mathbb{B}(\gamma) \text{ is a data-depth–based confidence region for } \theta \text{ with coverage probability } \gamma \text{ and } \mathbb{B}(\gamma) \cap \mathcal{R} \neq \emptyset\}.$$

This definition of P value is motivated by the duality between the rejection region for a level-α two-sided test and a $(1-\alpha)$ confidence interval for a univariate parameter. In this case we expand the confidence region for θ by increasing its confidence level until the confidence region meets the null space $\mathcal{R}$, and then regard the level of this final confidence region as the strength of evidence against H. The actual implementation of this idea can be carried out, for example, as follows. Generate a large number, say N, of bootstrap estimates of θ—namely, $\theta_{n1}^*, \ldots, \theta_{nN}^*$; apply any data depth mentioned in Section 3 to obtain a center outward ranking of θ_{ni}^*'s, with the smallest rank associated with the deepest point. Let $r_{n,N}$ be the rank of the smallest ranked θ^*_{ni} that lies in $\mathcal{R}$. Then $p_n^{(B)}$ is approximately $[1-(r_{n,N}/N)]$.

Remark 4.1. If the null space $\mathcal{R}$ does not have a smooth boundary at θ_0 and is concave around θ_0, then ESP($\mathcal{R}$) defined in method 4A tends to be conservative at $\theta = \theta_0$, meaning that ESP($\mathcal{R}$) tends to be stochastically larger than $U[0,1]$ (cf. Fig. 3). Though ESP($\mathcal{R}$) in this case is still a LP, we emphasize that if such a boundary point is of vital importance, as some situations may require, then other approaches of LP, such as $p_n^{(B)}$, should be used instead.

Remark 4.2. If the null space $\mathcal{R}$ has finitely many nonsmooth boundary points, say $a_1, a_2, \ldots, a_k$, then we may define $\text{ESP}'(\mathcal{R}) = \max\{\text{ESP}(\mathcal{R}), p_n(a_1), \ldots, p_n(a_k)\}$, where ESP($\mathcal{R}$) is as defined in Method 4A and $p_n(t)$ denotes a LP value for testing the null hypothesis $\theta = t$ (cf. Def. 3.1). This modified ESP$'$($\mathcal{R}$) is then a LP. If $\mathcal{R}$ is a convex region with some nonsmooth boundary points, (e.g., a rectangle in $\mathbb{R}^2$), then ESP($\mathcal{R}$) has a negative bias at such points (cf. Fig. 4). As a result, when ESP($\mathcal{R}$) is found to be higher than the preset level α, we can conclude to keep H without obtaining ESP$'$($\mathcal{R}$). In other cases, only ESP$'$($\mathcal{R}$) should be used to render the final rejection of H (cf. Fig. 6).

We next present a simple result where ESP is actually a LP(S).

Theorem 4.2. Assume that the underlying distribution is of the form $F_\theta(\cdot) = G(\cdot - \theta)$, where $\theta \in \mathbb{R}^d$ and $G(\cdot)$ is a fixed but unspecified distribution. Let θ be a prespecified location functional such that $\theta(G) = 0$ (thus $\theta(F_\theta) = \theta$). Assume that $\theta_n = \theta(F_n)$ admits a limiting distribution that is symmetric around zero. Then in testing the hypotheses H: $\nu \in \mathcal{R}$ versus K: $\nu \notin \mathcal{R}$, where $\nu = \mathbf{L}\cdot\theta$ for some $d\times 1$ vector $\mathbf{L}$, the ESP($\mathcal{R}$) is an LP(S), if $\mathcal{R}$ is of the form of $(-\infty, a]$, $[a,b]$ or $[b,\infty)$ for any real numbers $a \le b$.

The parameter ν here is motivated by problems of testing multisample means. For instance, ν may be viewed as a contrast in population means. The functional θ goes beyond the scope of linear functions of means. It includes means, medians, modes, and even componentwise medians, centers, and various notions of multivariate medians.

Remark 4.3. If θ is on the boundary of the support of the underlying distribution, then the limiting distribution of $a_n(\theta_n - \theta)$ is not symmetric around 0. In this case clearly the results in Theorems 3.1, 4.1, and 4.2 would not hold. However, a simple modification can be made to correct this shortcoming. This modification is to flip θ^*_n around θ_n (i.e., let $\theta^{**}_n = 2\theta_n - \theta^*_n$), and then use θ^{**}_n instead of θ^*_n is our original approach. This method gives the same desired results in Theorems 3.1, 4.1, and 4.2 without requiring symmetry on the limiting distribution. Furthermore, if an estimate of the standard error of θ_n is available, then we can even take advantage of it by using the studentized version of θ^{**}_n—that is, $\theta^{**}_{n,st} = \theta_n - \mathcal{S}_n^{1/2}\mathcal{S}_n^{*-1/2}(\theta^*_n - \theta_n)$—instead of the θ^{**}_n. Here $\mathcal{S}_n(\mathcal{S}^*_n)$ is the estimated covariance matrix of $\theta_n(\theta^*_n)$. Note that the second-order asymptotic accuracy under H can be achieved in this way. This modification of using θ^{**}_n or $\theta^{**}_{n,st}$ can be applied even if θ^*_n already admits a symmetric limiting distribution under H. However, the definition of LP with θ^{**}_n does not seem as natural as the one with θ^*_n.

5. SOME LIMITING *P* VALUES FOR TESTING INFINITE-DIMENSIONAL PARAMETERS

For testing about cdf's such as H: $F = F_0$ versus K: $F \ne F_0$, where F_0 is a given continuous cdf, we first describe the following notions of depth for curves, and then construct a LP for the test by using the same idea of depth based ranking discussed in Section 3. Let $X_1, \ldots, X_n$ be a random sample from the underlying distribution F. Let F_n be the empirical cdf for the sample and F^*_n its bootstrap counterpart. Define

$$d_b(F, F_0) = \begin{cases} \left\{\int |F(x) - F_0(x)|^b \, dW(x)\right\}^{b^{-1}} \\ \quad \text{for } 1 \le b < \infty, \\ \sup_x |F(x) - F_0(x)| \equiv (\|F - F_0\|_\infty), \\ \quad \text{for } b = \infty \end{cases}$$

and

$$D_b(F, F_0) = \{1 + d_b(F, F_0)\}^{-1}. \tag{7}$$

Here $W(\cdot)$ is a given cdf, which can be taken as F_0. The definition of $D_b(F, F_0)$ clearly gives a measure of "depth" of F with respect to the "center" F_0, meaning that the larger $D_b(F, F_0)$, the "closer" F to F_0 within the class of all cdf's. Throughout the rest of the article, the index b in d_b and D_b shall be omitted when it does not call for special attention.

Now, for testing H: $F = F_0$ versus K: $F \ne F_0$, we present the following construction of a LP. A rigorous justification of this LP can be established using arguments similar to those used in the proof of Theorem 3.1, and we omit the details.

Definition 5.1. Let

$$p_n \equiv p_n(F_0) = P^*(F^*_n\colon\ D(F^*_n, F_n) \le D(F_0, F_n)).$$

Because $\sqrt{n}d_b(F_n, F_0)$ and $\sqrt{n}d_b(F^*_n, F_n)$ generally have the same continuous limiting distribution, the p_n defined here is a LP for the cases when $b = 1, 2$, or ∞. As a matter of fact, these cases correspond to the well-known Mallows, Cramer–von Mises, and Kolmogorov–Smirnov goodness-of-fit tests, where the exact P values are obtainable. We point out that the scope of the LP construction in this testing concept goes far beyond the empirical process. For example, it should cover the product limit estimate in the survival analysis. Moreover, similar concepts of LP can be defined for testing curves such as density functions and regression functions.

For testing H: $F \in \mathcal{F}_0$ versus K: $F \notin \mathcal{F}_0$, where $\mathcal{F}_0$ is a class of specified cdf's, the following proposed definition is a LP for the test.

Definition 5.2. Let

$$p_n \equiv p_n(\mathcal{F}_0) = \sup_{F \in \mathcal{F}_0} p_n(F),$$

where $p_n(F)$ is given in Definition 5.1.

Note that

$$p_n(\mathcal{F}_0) = P^*(F_n^*\colon\ D(F_n^*, F_n) \le D(F_0, F_n), \\ \text{where } d(F, F_n) \text{ is minimized by } F_0 \in \mathcal{F}_0).$$

This expression clearly shows the association between $p_n(\mathcal{F}_0)$ and the so-called minimum distance estimation. It would be interesting in its own right to further investigate the properties of this proposed P value along the line of minimum distance estimation. Often $\mathcal{F}_0$ in the foregoing composite null hypothesis H is taken to be a parametric family of distributions; for example, the class of all normal distributions with a finite mean and a bounded variance. Note that Definition 5.2 is especially useful if $\mathcal{F}_0$ consists of only finitely or countably many members, as in the case of discrete parameter problems. In practical situations, one may also be interested in the case when $\mathcal{F}_0$ is a class of distributions that in some sense are not substantially different from a specified distribution, say F_0. For instance, let $\mathcal{F}_0 = \mathcal{R}_\varepsilon \equiv \{G\colon d(G, F_0) \le \varepsilon\}$ for some preset value $\varepsilon \ge 0$. The ε value can be determined by previous knowledge in constructing confidence bounds for F_0 or it can be assumed to be the maximal allowed deviation from the preset standard distribution F_0. For testing H: $F \in \mathcal{R}_\varepsilon$ versus K: $F \notin \mathcal{R}_\varepsilon$, where $\mathcal{R}_\varepsilon = \{G\colon d(G, F_0) \le \varepsilon$, for the given F_0 and $\varepsilon\}$, we adapt the same idea of ESP given in Section 4 to provide an ESP in the next definition. Clearly, this definition readily extends to similar testing problems concerning curves other than cdf's.

Definition 5.3. $p_n(\mathcal{R}_\varepsilon) \equiv \text{ESP}(\mathcal{R}_\varepsilon) = P^*\{F_n^*\colon d(F_n^*, F_0) \le \varepsilon\}$. In Theorem 5.1 we prove that $\text{ESP}(\mathcal{R}_\varepsilon)$ is a LP.

Theorem 5.1. Let $\mathcal{R}_\varepsilon = \{G\colon d_0(G, F_0) \le \varepsilon\}$ for some F_0 and $\varepsilon > 0$. Assume that $d_b(F, F_0) = \varepsilon$, where F in the underlying population distribution. Then as $n \to \infty$,

$$\text{ESP}(\mathcal{R}_\varepsilon) \xrightarrow{\mathcal{L}} U[0,1]$$

holds for $1 < b < \infty$ and for $b = \infty$ under the additional assumptions that $\|F - F_0\|_\infty = \varepsilon$ is attained uniquely at a point, say x_0, and both F and F_0 have bounded second derivatives in a neighborhood of x_0. The foregoing result for ESP also holds for $b = 1$ if $F(\cdot) = F_0(\cdot)$ occurs only at finitely many points on the interior of the union of the supports of $F(\cdot)$ and $F_0(\cdot)$ and if the weight function $W(\cdot)$ is continuous.

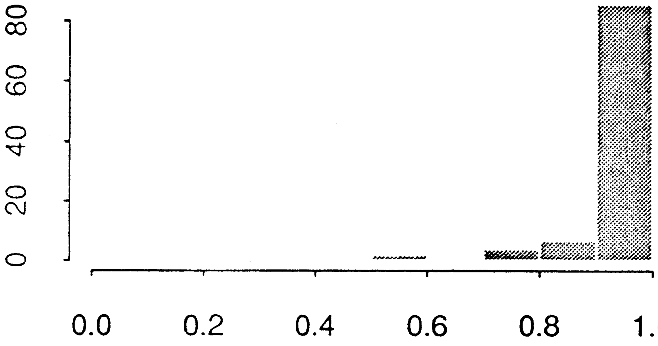

Figure 1. Histogram of ESP Values When the True Mean is an Interior Point of the Null Space $\mathcal{R}$ = Rectangle With Corners $\{(-1, -1), (1, -1), (1, 1), (-1, 1)\}$.

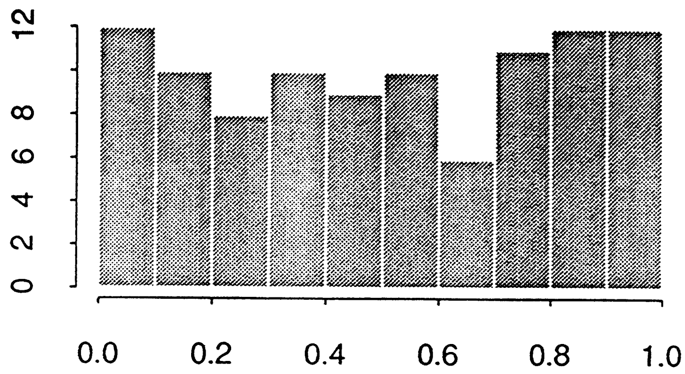

Figure 2. Histogram of ESP Values When the True Mean is on the Smooth Boundary of the Null Space $\mathcal{R}$ = Rectangle With Corners $\{(-1, -4), (0, -4), (0, 4), (-1, 4)\}$.

Remark 5.1. The definition of $d_b(F, F_0)$ in (7) measures the deviation of F from F_0. Different b values lead to different properties of their associated LP values, and in turn to different power functions of their resulting tests. For example, the special cases $b = \infty$ and $b = 2$ correspond to the Kolmogorov–Smirnov and Cramer–von Mises tests with quite different power functions. Even though in this article we restrict ourselves to LP's based on such normed metric distances, we think other measures of the difference between two cdf's can be incorporated in the proposed LP's. For example, the so-called *quality index* $Q(F, F_0)$ of Liu and Singh (1993) quantifies the difference between F and F_0 in a probabilistic geometry approach. It may be interesting to derive a LP from it.

6. SIMULATION RESULTS AND POWER ASPECTS OF EMPIRICAL STRENGTH PROBABILITY

The simulation work can be divided into two parts. The first part (Figs. 1–4) is designed to show that ESP is easily adaptable to various unusual forms of the null space. The second part involves testing the correlation coefficient, which provides an opportunity for comparing ESP and the

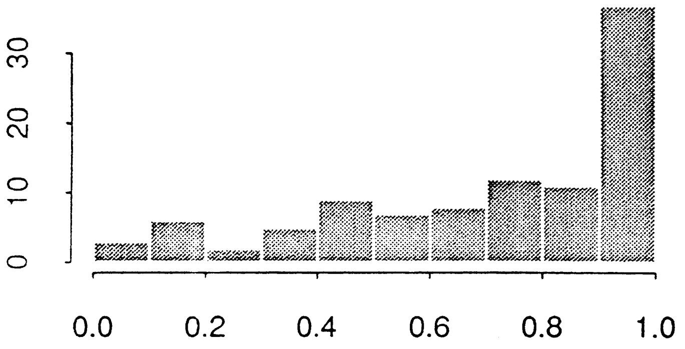

Figure 3. Histogram of ESP Values When the True Mean is a Boundary Point Around Which the Null Space is Concave.

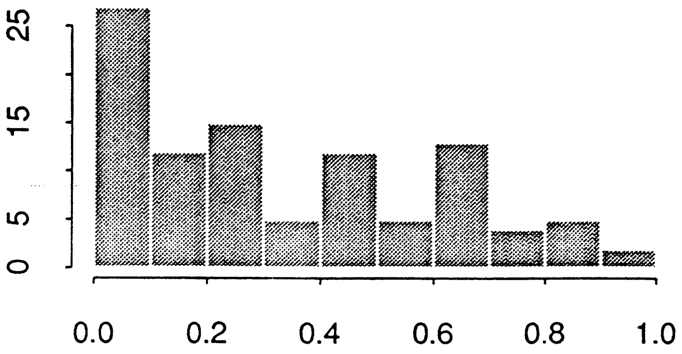

Figure 4. Histogram of ESP Values When the True Mean is a Nonsmooth Boundary Point of a Convex Null Space in Case 4, Section 6.1.

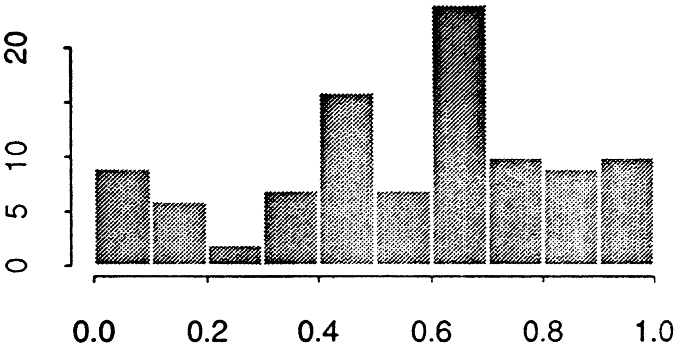

Figure 6. Histogram of ESP′ Values Corresponding to Figures 4 and 5.

standard t test under the assumption of bivariate normality. In the absence of normality, we use three examples (Figs. 11–13) to show that ESP is still applicable while the t test fails completely. We would also like to point out that the great simplicity of the ESP procedure, counting the number of the bootstrapped estimates falling inside the null space, remains unaffected even if the null space and the test parameter assume more complicated forms. All simulations are carried out using S-language on a SUN workstation.

6.1 Testing the Bivariate Normal Mean Vector

All results in this section are simulated from a bivariate normal random variable (X_1, X_2), with the mean vector (0, 0) and the covariance matrix $(\begin{smallmatrix}1 & .8\\ .8 & 4\end{smallmatrix})$. The parameter of interest is the mean vector of (X_1, X_2), which is (0, 0).

Figures 1–4 show the histograms of the simulated ESP values under different null spaces $\mathcal{R}$'s under H: $\mu \in \mathcal{R}$ (cf. Method 4A). A sample of size 30 is taken from (X_1, X_2), and for each sample 500 bootstrap samples of size 30 are drawn to compute 500 bootstrap estimates of the mean. Our ESP in Method 4A is nothing but the fraction of the 500 bootstrap estimates that fall in $\mathcal{R}$. For each given $\mathcal{R}$, this procedure is repeated 100 times to obtain 100 values of ESP, to form a histogram. The three $\mathcal{R}$'s in Figures 1, 2, and 4 are rectangles with the following corners: 1) $\{(-1,-1),(1,-1),(1,1),(-1,1)\}$; 2) $\{(-1,-4),(0,-4),(0,4),(-1,4)\}$; and 4) $\{(0,0),(0,-4),$ $(-1,-4),(-1,0)\}$. The $\mathcal{R}$ in Figure 3 is the complement of the quadrant $\{(x_1,x_2)\colon x_1 > 0, x_2 > 0\}$. Figure 1 shows that ESP tends to assume values close to 1 when the true parameter is in the interior of $\mathcal{R}$. The histogram in Figure 2 is close to $U[0,1]$, because in this case the true parameter is on the smooth boundary of $\mathcal{R}$. Figure 3 is somewhat skewed to the left and gives conservative values of ESP, because the true parameter is a boundary point around which the region is concave. As for Figure 4, we see that ESP tends to be close to 0, as the true parameter is a nonsmooth boundary point of a convex region. As suggested in Remark 4.2, this situation can be corrected by using ESP′, the modified ESP. The histogram in Figure 6 reflects such a correction.

To compute ESP′ for Figure 6, we first follow Definition 3.1 and apply the simplicial depth to obtain 100 simulated p_n values as if we were testing H: $\mu = (0,0)$. Figure 5 is a histogram of these 100 simulated p_n values. Here each p_n value is the fraction of the 500 bootstrap means which have less simplicial depth than the point (0, 0). The histogram appears to be close to $U[0,1]$ as expected. Calculation of the simplicial depth here uses the FORTRAN algorithm developed by Rousseeuw and Ruts (1992). This algorithm is highly efficient since it requires only $O(n \log n)$ steps, instead of $O(n^4)$ steps as required by the direct computation based on solving systems of linear equations. The p_n values obtained in Figure 5 are compared to their corresponding ESP's in Figure 4 to determine the maximum values. These are the ESP′ values presented in Figure 6. It is clear that compared to Figure 4, the histogram of Figure 6 is closer to $U[0,1]$.

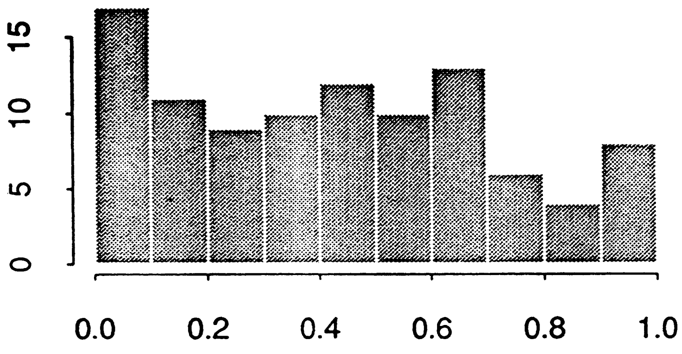

Figure 5. Histogram of Simplicial Depth–Based p_n Values Corresponding to Figure 4.

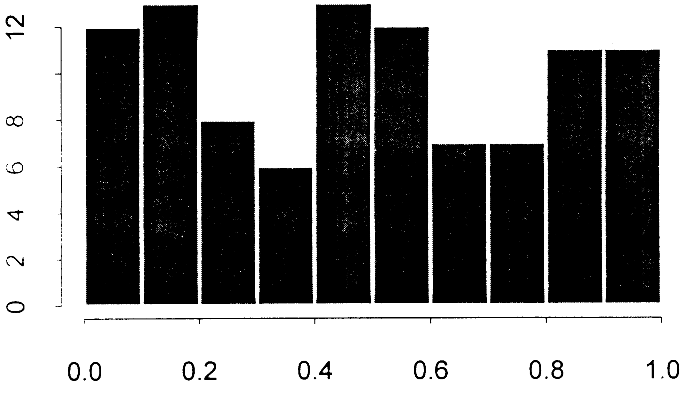

Figure 7. Histogram of ESP Values for Testing H: $\rho \leq 0$ Versus K: $\rho = \rho_a$, where $\rho_a > 0$.

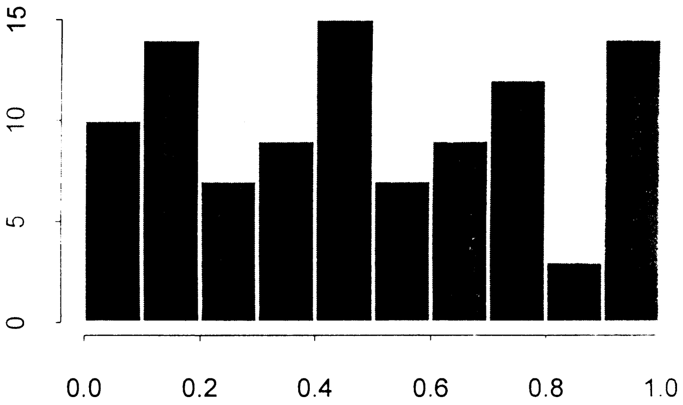

Figure 8. Histogram of P Values Based on the t Test for Testing H: $\rho \le 0$ Versus K: $\rho = \rho_a$, Where $\rho_a > 0$.

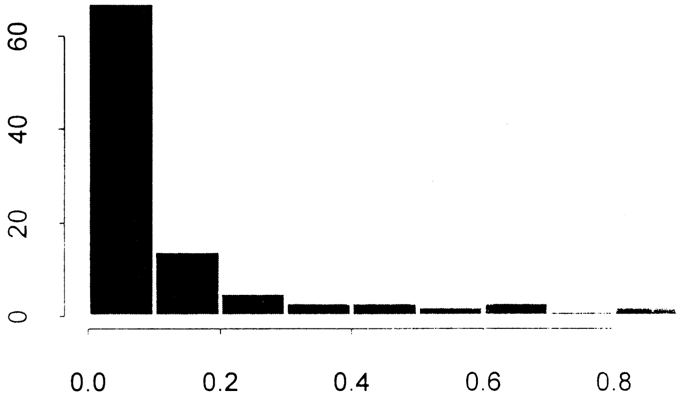

Figure 10. Histogram of P Values Based on the t Test for Testing H: $\rho \le 0$ Versus K: $\rho = .4$.

6.2 Testing the Correlation Coefficient

This section contains seven simulated histograms of P values for testing the correlation coefficient, ρ. The first four comprise a power comparison of our ESP in Method 4A with the existing t test under normality assumption. Samples are simulated from a bivariate normal random variable (X_1, X_2) with mean vector (0, 0) and covariance matrix $\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$. For testing H: $\rho \le 0$ versus K: $\rho = \rho_a$, where $\rho_a > 0$, the standard t test is based on the statistic $r/\sqrt{(1-r^2)/(n-2)}$, where r is the sample correlation coefficient and n the sample size, taken to be 20 here. In fact, the specification of the mean and the covariance matrix is unnecessary, because it would not affect the test procedure or its result. To observe the behavior of the ESP P value under H, we draw a sample of 20 from the standard bivariate normal distribution. From the given sample, we generate 500 bootstrap samples each of size 20 to obtain 500 bootstrap sample correlation coefficients, r^*'s. An ESP P value, (i.e., $P^*(r^* \le 0)$), is then simply the fraction of the r^*'s that are ≤ 0. This procedure is repeated 100 times to obtain 100 such P values. The histogram of these 100 ESP P values in Figure 7 appears to be uniformly distributed, as we expected. Figure 8 contains the histogram of 100 P values obtained by the t test that also seems to be close to uniform. Concerning the power of the two tests, Figures 9 and 10 give two histograms of 100 P values under K: $\rho = .4$, for ESP and the t test. The histograms both are heavily skewed to the right and of similar shapes, showing that the two methods are *almost equally powerful* under K: $\rho = .4$. Table 1 lists the estimated power when $\rho_a = .2, .4, .6$, and .8, given $\alpha = .05$. The results clearly show that the ESP method is quite comparable to the t test, even though ESP is completely nonparametric and does not utilize the normality assumption.

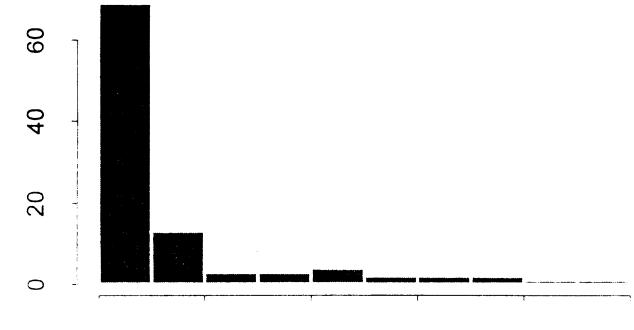

Figure 9. Histogram of ESP Values for Testing H: $\rho \le 0$ Versus K: $\rho = .4$.

We proceed to carry out the same ESP procedure for three more bivariate distributions, where no other methods are readily available for testing H: $\rho = 0$. The three distributions are as follows:

a. The component variables are two uncorrelated unvariate exponential variables each with mean 1.
b. The exponential variables in distribution a are replaced by double exponential variables each with mean 0 and variance 2.
c. The first component variable is a standard normal, and the second component variable is the square of the first one.

Figures 11, 12, and 13 are the histograms of the corresponding resulting 100 ESP P values, and they all appear more or less uniform, which once again confirms the validity of the ESP approach. Note that the t test described in the previous paragraph is not applicable here, because the assumption of bivariate normality is violated. As a matter of fact, if we blindly apply the t test to the case of distribution c, to obtain 100 P values, then the histogram of these P values appears to be V-shaped rather than uniform.

7. CONCLUDING REMARKS

We have shown how bootstrap and data depth can combine to provide nonparametric methods to determine P values in testing hypotheses. In particular, we have introduced the new concept *empirical strength probability* (ESP) as a P value. This approach can handle very general classes of

Table 1. Estimated Power Under K: $\rho = \rho_a$

ρ_a	.2	.4	.6	.8
t test	.27	.56	.89	1.00
ESP	.31	.56	.90	1.00

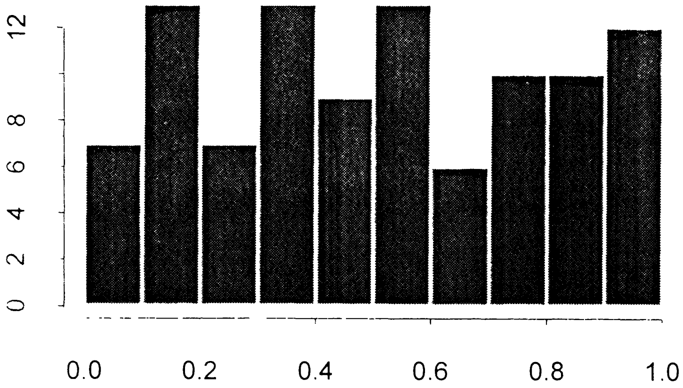

Figure 11. Histogram of ESP Values for Testing H: ρ = 0 Versus K: ρ ≠ 0 in Case a, Section 6.2.

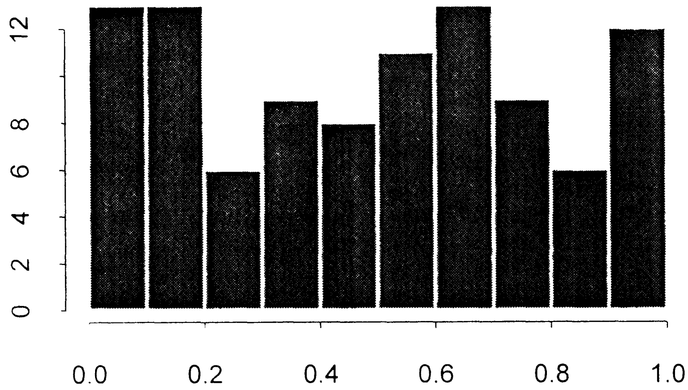

Figure 13. Histogram of ESP Values for Testing H: ρ = 0 Versus K: ρ ≠ 0 in Case c, Section 6.2.

hypotheses, including the ones with any smooth bounded region as a null space. Throughout, the bootstrap samples are drawn directly from the empirical distribution. This is simpler than bootstrapping from distributions obeying the null hypothesis, as currently required (see, e.g., Beran 1986, Loh 1985, and Romano 1988).

If a parametric model (say multivariate normal) is assumed and the likelihood ratio test can be obtained, then this parametric test in principle should be superior to ours or to all other nonparametric methods. However, our methods here are nonparametric and *free from* analytical work, which is often complicated and arduous, as seen in the simulations in Section 6.2. Are our tests comparable to the parametric ones in any way? The answer here is an emphatic "yes"! This sums up the key merit of our test: If the assumed model fails, then the parametric tests may become worthless, while ours still stay optimal under some minor regularity conditions. This conclusion is also supported by the simulation results in Section 6.2. A detailed power/efficiency–related study in this regard shall be presented elsewhere, but a brief heuristic account is given as follows. Considering Hodges–Lehmann efficacy, Singh and Berk (1994) established some rigorous results on the efficacy for the so-called type 2 P values, which are special cases of our ESP. Following similar arguments, our tests retain the Hodges–Lehmann efficacy–type optimality if the estimator θ_n happens to be the maximum likelihood estimator (MLE) of θ under an assumed parametric model from which the sample was drawn. A technically easier task is to look at Pitman efficiency where a local alternative is clearly defined. With respect to Pitman efficiency, all of our tests, including the ones based on data depth, turn out to be locally equivalent to parametric tests that use the true distribution of θ_n, the MLE of θ, except for some multidimensional regions where defining Pitman efficiency is a nontrivial task in itself. Looking back at the simulation study of testing H: $\rho \le 0$ versus K: $\rho = \rho_a$ in Section 6.2, our ESP is equivalent to the t test in terms of Hodges–Lehmann efficacy and Pitman efficiency.

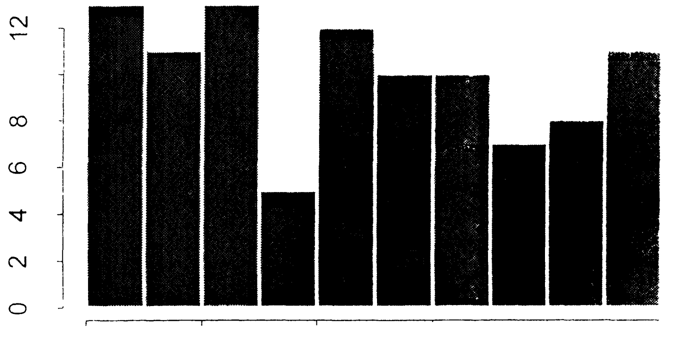

Figure 12. Histogram of ESP Values for Testing H: ρ = 0 Versus K: ρ ≠ 0 in Case b, Section 6.2.

It would be useful to know the effects of different choices of data depth in our approach. Clearly, the properties of depths themselves vary greatly, because some of them are metric or norm dependent (e.g., Mahalanobis depth) and others are not (e.g., Tukey, simplicial, and majority depth). A "normed" depth is moment dependent, and hence more sensitive to outliers, and thus tends to be less robust. In addition, "normed" depths (such as the Mahalanobis depth) may not reflect well the geometric shape of the underlying distributions—in particular, the asymmetry if any exists. A full comparison of the various depths should include aspects of inference such as efficiency, robustness, and computational feasibility. Some earlier results on some individual depths were given by, for example Arcones, Chen, and Gińe (1994), Donoho and Gasko (1993), Liu (1990), Liu and Singh (1993), and Rousseeuw and Ruts (1992). We plan to return to these issues in a more systematic manner in a later project.

APPENDIX

Proof of Theorem 3.1

Recall that $L_n^*(\cdot)$ and $L_n(\cdot)$ are the cdf's of $a_n(\theta_n^* - \theta_n)$ and $a_n(\theta_n - \theta_0)$. Let $T_n^* = a_n(\theta_n^* - \theta_n), T_n = a_n(\theta_n - \theta_0)$, and $T_n' = -T_n$. Because $D(\cdot;\cdot)$ is affine invariant, the definition $p_n = P_{G_n^*}\{\theta_n^*\colon D(G_n^*;\theta_n^*) \le D(G_n^*;\theta_0)\}$ can also be expressed as

$$p_n = P_{L_n^*}\{T_n^*\colon D(L_n^*;T_n^*) \le D(L_n^*;T_n')\}.$$

Define

$$p_n' = P_L\{T\colon D(L;T) \le D(L;T_n')\}.$$

Using Slutsky's theorem, the claim of our theorem will follow if we show both $p_n' \xrightarrow{\mathcal{L}} U[0,1]$ and $p_n - p_n' \to 0$ in probability as $n \to \infty$.

To show that $p_n' \xrightarrow{\mathcal{L}} U[0,1]$, we first let $H(\cdot)$ denote the cdf of the random variable $D(L;T)$; that is, $H(s) = P\{D(L;T) \leq s\}$. Because $H(\cdot)$ is assumed to be continuous, $H(D(L;T))$ follows $U[0,1]$. The remaining task now is to show that $D(L;T_n') \xrightarrow{\mathcal{L}} H$, as $p_n' = H(D(L;T_n'))$. Note that $T_n' \xrightarrow{\mathcal{L}} L$ under the assumption that L is symmetric around zero. Consequently, under the continuous transformation $D(L;\cdot), D(L;T_n')$ converges in law to the distribution of $D(L;T)$, which is H.

We now turn to the proof of $p_n - p_n' \to 0$ in probability as $n \to \infty$. For any $\varepsilon > 0$, it is clear that for all large n, p_n lies between the following two quantities:

$$P_{L_n^*}\{T_n^*\colon\ D(L;T_n^*) \leq D(L;T_n') + \varepsilon\}$$

and

$$P_{L_n^*}\{T_n^*\colon\ D(L;T_n^*) \leq D(L;T_n') - \varepsilon\},$$

almost surely. Because $D(L;T_n^*) \xrightarrow{\mathcal{L}} D(L;T)$ and because the cdf of $D(L;T)$ is continuous, by Polya's theorem, p_n then lies almost surely between

$$P_L\{T\colon\ D(L;T) \leq D(L;T_n') + \varepsilon\} + o(1)$$

and

$$P_L\{T\colon\ D(L;T) \leq D(L;T_n') - \varepsilon\} + o(1)$$

for all large n. Here $o(1)$ indicates a random variable that tends to zero a.s. Because ε is arbitrary, $p_n - p_n'$ converges to 0 in probability.

The following lemma is needed in the proof of Theorem 4.1.

Lemma (Bootstrap Convolution Lemma). Assume that there exist a positive sequence a_n and a cdf L such that $a_n(\theta_n - \theta_0) \xrightarrow{\mathcal{L}} L$ and $a_n(\theta_n^* - \theta_n) \xrightarrow{\mathcal{L}} L$ a.s. Then $a_n(\theta_n^* - \theta_0) \xrightarrow{\mathcal{L}} L * L$ unconditionally. Here * indicates the convolution of two distributions.

Proof. Rewrite $a_n(\theta_n^* - \theta_0)$ as $\{a_n(\theta_n^* - \theta_n) + a_n(\theta_n - \theta_0)\}$. Denote the unconditional characteristic function of $a_n(\theta_n^* - \theta_0)$ by $\phi_n(t)$. Then, taking the conditional expectation, we have

$$\phi_n(t) = E[\exp\{e^{ita_n(\theta_n - \theta_0)}[E_L e^{\imath tY} + o(1)]\}],$$

where $Y \sim L$. Here $t \in \mathbb{R}$ and $i = \sqrt{-1}$. Using the bounded convergence theorem, we now obtain that $\phi_n(t) \to [E_L e^{\imath tY}]^2$, and thus the desired result follows.

Proof of Theorem 4.1

Without any loss of generality, we give the proof in $\mathbb{R}^2$. For any fixed $\varepsilon > 0$, there exists a $\delta > 0$, such that $C(\theta_0,\varepsilon) \cap S(\theta_0,\delta)$ contains $B \cap S(\theta_0,\delta)$, where B denotes the boundary of $\mathcal{R}$, $S(\theta_0,\delta)$ denotes the disc with center θ_0 and radius δ, and $C(\theta_0,\varepsilon)$ denotes the two cone regions within which all lines passing through θ_0 will form an angle no larger than ε on either side of the tangent line of θ_0. Note that the tangent line at θ_0 divides the parameter space into two closed half-spaces with the tangent line as their common boundary. Let TL^+ be the half-space containing the perpendicular at θ_0 pointing toward the interior of the region $\mathcal{R}$. For example, TL^+ contains $\mathcal{R}$ when $\mathcal{R}$ is convex. We show later that

$$p_n^{(A)} = P^*(\theta_n^* \in TL^*) + o_p(1). \tag{A.1}$$

Now this last step essentially reduces the original multivariate test to a test of a fixed linear combination of the coordinates of θ, which is univariate in nature. The claim of the theorem therefore follows from theorem 2.1 of Singh and Berk (1994).

To show that (A.1) holds, we first note that the bootstrap convolution lemma yields. If $\theta = \theta_0$, then, for a fixed $\delta > 0$,

$$E\{P^*(\theta_n^* \in \mathcal{R}) - P^*(\theta_n^* \in [\mathcal{R} \cap S(\theta_0,\delta)])\} = o(1).$$

Because $P(|\theta_n^* - \theta_0| > \delta) \to 0$ and

$$\begin{aligned}&|P^*(\theta_n^* \in [\mathcal{R} \cap S(\theta_0,\delta)]) - P^*(\theta_n^* \in [TL^+ \cap S(\theta_0,\delta)])|\\ &\qquad \leq\ P^*(\theta_n^* \in [C(\theta_0,\varepsilon) \cap S(\theta_0,\delta)]),\end{aligned} \tag{A.2}$$

it suffices to show that the right side of (A.2) is of order $O(\varepsilon)$. This is done by using the bootstrap convolution lemma and by recalling that L is absolutely continuous.

Proof of Theorem 4.2

Because ν can be viewed as a location shift parameter for the distribution of random variable $\mathbf{L} \cdot X$, where X follows the distribution $F(\cdot - \theta)$, it suffices to prove the result for the univariate location shift distributions. Thus we assume without loss of generality ν is the location shift parameter of the univariate distribution $G(\cdot - \nu)$.

Consider the case $\mathcal{R} = (-\infty, a]$. If H holds, then $\nu \in (-\infty, a]$, and for any t such that $0 \leq t \leq 1$, we have

$$\begin{aligned}P_\nu\{\text{ESP}(-\infty,a] \leq t\} &= P_a\{\text{ESP}(-\infty, 2a - \nu] \leq t\}\\ &\leq P_a\{\text{ESP}(-\infty,a] \leq t\}.\end{aligned}$$

The last expression converges to $U[0,1]$ following Theorem 4.1.

A similar argument holds for $\mathcal{R} = [b,\infty)$.

In case $\mathcal{R} = [a,b]$, we give only the proof for $\nu \in [a,(a+b)/2]$, because similar arguments hold for the remaining region. Following the shift invariance property of ν, we have

$$\begin{aligned}P_\nu\{\text{ESP}[a,b] \leq t\} &= P_a\{\text{ESP}[2a - \nu, a + b - \nu] \leq t\}\\ &\leq P_a\{\text{ESP}[a, a + b - \nu] \leq t\}\\ &= P_a\{\text{ESP}[a,b] - r_n) \leq t\},\end{aligned}$$

where $r_n =$ ESP$(a + b - \nu, b]$. Our result is obtained by taking $\limsup_{n\to\infty}$ on both sides of the foregoing inequality and by making the following observations: Because ESP$\{a + b - \nu, b] \leq$ ESP$((a+b)/2, b], P_a(r_n \geq \varepsilon) \leq P_a(ESP((a+b)/2, b] \geq \varepsilon)$ for any $\varepsilon > 0$. The last probability tends to 0, because the distance between a and $((a+b)/2, b]$ is at least $(b-a)/2$.

Proof of Theorem 5.1

Let $1 < b < \infty$. Following the arguments in theorem 2.3 of Singh and Berk (1994), it suffices to show that $\sqrt{n}\{d_b(F_n,F_0) - \varepsilon\}$ and $\sqrt{n}\{d_b(F_n^*,F_0) - d_b(F_n,F_0)\}$ both converge to the same normal distribution, with the latter converging almost surely. To achieve this, we express both expressions in the form of normalized sample means. Now consider the first case, in which we note that

$$\begin{aligned}d_b(F_n,F_0) &= \int |F_n - F_0|^b\, dW\\ &= \int |F_n - F + F - F_0|^b\, dW.\end{aligned}$$

We then evaluate this integral separately in three ranges: (a) $C_1 = \{x\colon |F(x) - F_0(x)| \leq n^{-1/2}\log n\}$, (b) $C_2 = \{x\colon F(x) - F_0(x) > n^{-1/2}\log n\}$, and (c) $C_3 = \{x\colon F_0(x) - F(x) > n^{-1/2}\log n\}$. The facts that $\sup_x |F_n(x) - F_0(x)| = O_p(n^{-1/2})$ and that $b > 1$ imply that the integral over the range C_1 has the order

$O_p(n^{-(1/2+\delta)})$ for some $\delta > 0$. In the range $C_2, (F - F_0)$ dominates $(F_n - F)$, because the latter is of $O_p(n^{-1/2}(\log n)^{1/2})$ a.s. Thus for all large n,

$$\int_{C_2} |F_n - F + F - F_0|^b \, dW = \int_{C_2} (F_n - F + F - F_0)^b \, dW$$

a.s. Next, we apply Taylor's expansion to the integrand on the right side, which leads to

$$(F_n - F + F - F_0)^b = (F - F_0)^b + (F_n - F)b(F - F_0)^{b-1} + \frac{1}{2}\,(F_n - F)^2 b(b-1)(*)^{b-2},$$

where (*) lies between $(F - F_0)$ and $(F_n - F_0)$. Arguing separately in the cases of $b \geq 2$ and $1 < b < 2$, we can show that the third term in the foregoing Taylor expansion is of $O(n^{-(1/2+\delta_1)})$ uniformly for some $\delta_1 > 0$, and hence is negligible for our purpose. Consequently,

$$\int_{C_2} |F_n - F|^b \, dW = \int_{C_2} |F - F_0|^b \, dW + \int_{C_2} b(F_n - F)(F - F_0)^{b-1} \, dW + o_p(n^{-1/2}).$$

Similar arguments hold for the range C_3 and lead to the following result:

$$\int_{C_3} |F - F|^b \, dW = \int_{C_3} |F - F_0|^b \, dW + \int_{C_3} b(F - F_n)(F_0 - F)^{b-1} \, dW + o_p(n^{-1/2}).$$

Now, combining observations that we have made for the three ranges, we obtain that

$$\int |F_n - F|^b \, dW = \int |F - F_0|^b \, dW + \int g \, dW + o_p(n^{-1/2}),$$

where

$$g(x) = \begin{cases} b(F_n(x) - F(x))(F(x) - F_0(x))^{b-1} & \text{if } F(x) > F_0(x), \\ b(F(x) - F_n(x))(F_0(x) - F(x))^{b-1} & \text{if } F(x) < F_0(x). \end{cases}$$

The claimed result now follows if we can show that $Eg(X_i) = 0$. This is obtained by interchanging the orders of the integrals involved.

Now consider the case $b = \infty$. Assume that $F(x) - F_0(x) = \varepsilon$ is attained uniquely at $x = x_0$. (The same argument holds if $F - F_0 = -\varepsilon$ at $x = x_0$.) Because $(F - F_0)$ attains its maximum at $x_0, F'(x_0) = F_0'(x_0)$. Recall that $\text{ESP}(\mathcal{R}_\varepsilon) = P^*(\|F_n^* - F_0\|_\infty \leq \varepsilon)$. We first concentrate on $\|F_n - F_0\|_\infty$. Because $\|F_n - F\|_\infty = O(n^{-1/2}(\log n)^{1/2})$ a.s., the supremum of $|F_n - F_0|$ occurs within the interval $I_n = (x_0 - n^{-1/4} \log n, x_0 + n^{-1/4} \log n)$ a.s. for all large n. In this particular interval, the differences $(F_n - F_0)$ and $(F - F_0)$ are both positive. Note that $\sup_{x \in I_n}(F_n(x) - F_0(x)) = \sup_{x \in I_n}(F_n(x) - F(x) + F(x) - F_0(x))$. A standard argument in asymptotics (see, e.g., Bahadur 1967) can show that

$$\sup_{x \in I_n} |F_n(x) - F(x) - F_n(x_0) - F(x_0)| = o(n^{-1/2}) \quad \text{a.s.}$$

Hence

$$\begin{aligned} &\sup_{x \in I_n} (F_n(x) - F_0(x)) \\ &\quad = F_n(x_0) - F(x_0) + \sup_{x \in I_n} (F(x) - F_0(x)) + o(n^{-1/2}) \quad \text{a.s.} \\ &\quad = F_n(x_0) - F(x_0) + F(x_0) - F_0(x_0) + o(n^{-1/2}) \quad \text{a.s.} \end{aligned}$$

Because $F(x_0) - F_0(x_0) = \varepsilon$, we now have

$$\sqrt{n}(\|F_n - F_0\|_\infty - \varepsilon) \xrightarrow{\mathcal{L}} \mathcal{N}(0, c_0(1 - c_0)),$$

where $c_0 = F_0(x_0) + \varepsilon$. Similar arguments will prove that $\sqrt{n}(\|F_n^* - F_0\|_\infty - \|F_n - F_0\|_\infty)$ converges to the same normal distribution. Thus the claim follows using the arguments provided by Singh and Berk (1994).

Consider the case $b = 1$. As in the previous two cases, we know the result of the theorem will follow if we show

$$\int |F_n - F_0| \, dW = \int |F - F_0| \, dW + \int \text{sign}(F - F_0)(F_n - F) \, dW + o_p(n^{-1/2}) \qquad \text{(A.3)}$$

and

$$\int |F_n^* - F_0| \, dW = \int |F_n - F_0| \, dW + \int \text{sign}(F - F_0)(F_n^* - F_n) \, dW + o_p(n^{-1/2}).$$

The proofs of the two statements are similar, so we focus only on the first one. Without loss of generality, we may assume that $F = F_0$ occurs only at one point in the interior of the union of the supports of F and F_0, say x_0. We carry out the integral in two separate regions: (I) = $\{x : |F(x) - F_0(x)| > \|F_n - F\|_\infty$ and (II) = $\{x : |F - F_0| \leq \|F_n - F\|_\infty$. In (I), the sign of $(F_n - F_0)$ is the same as that of $(F - F_0)$. Thus the statement (A.3) on the region (I) holds without any remainder. Next we turn to (II). Because x_0 is the only point where $F = F_0$, the set (II) is of the form $[x_0 - a_n, x_0 + b_n] \cup (-\infty, c_n] \cup [d_n, \infty)$, for $n \geq n_0$. Here a_n and b_n both tend to 0, and c_n and d_n tend to the left and the right endpoints of the union of the supports of F and F_0. Because $W(\cdot)$ is a cdf, and $\|F_n - F\|_\infty = O_p(n^{-1/2})$, it follows that

$$\int_{(\text{II})} |F_n - F_0| \, dW(t) + \int_{(\text{II})} |F - F_0| \, dW(t) + \int_{(\text{II})} |F_n - F| \, dW(t) = O_p(n^{-1/2}).$$

This establishes (A.3).

[Received June 1993. Revised April 1996.]

REFERENCES

Arcones, M., Chen, Z., and Giñe, E. (1994), "Estimators Related to U-Processes With Applications to Multivariate Medians: Asymptotic Normality," *The Annals of Statistics*, 92, 1460–1477.

Bahadur, R. R. (1967), "Rates of Convergence of Estimates and Test Statistics," *Annals of Mathematical Statistics*, 38, 303–324.

——— (1971), *Some Limit Theorems in Statistics*, Philadelphia: SIAM.

Beran, R. (1986), "Simulated Power Functions," *The Annals of Statistics*, 14, 151–173.

Berger, R., and Boos, D. (1994), "P-Values Maximized Over a Confidence Set for the Nuisance Parameter," *Journal of the American Statistical Association*, 89, 1012–1016.

Donoho, D., and Gasko, M. (1992), "Breakdown Properties of Location Estimators Based on Halfspace Depth and Projected Outlyingness," *The Annals of Statistics*, 90, 1803–1827.

Efron, B. (1982), *The Jackknife, the Bootstrap and Other Resampling Plan*, Philadelphia: SIAM.

Hall, P. (1988), "Theoretical Comparison of Bootstrap Confidence Intervals," *The Annals of Statistics*, 16, 927–953.

Lambert, D. (1981), "Inference Functions for Testing," *Journal of the American Statistical Association*, 76, 649–657.

Liu, R. Y. (1990), "On a Notion of Data Depth Based on Random Simplices," *The Annals of Statistics*, 18, 405–414.

Liu, R. Y., and Singh, K. (1993), "A Quality Index Based on Data Depth and Multivariate Rank Tests," *Journal of the American Statistical Association*, 88, 252–260.

Loh, W. (1985), "A New Method for Testing Separate Families of Hypotheses," *Journal of the American Statistical Association*, 80, 362–368.

Mahalanobis, P. C. (1936), "On the Generalized Distance in Statistics," *Proceedings of the National Academy of India*, 12, 49–55.

Romano, J. (1988), "A Bootstrap Revival of Some Nonparametric Distance Tests," *Journal of the American Statistical Association*, 83, 698–708.

Rousseeuw, P. J., and Leroy, A. H. (1987), *Robust Regression and Outlier Detection*, New York: Wiley.

Rousseeuw, P. J., and Ruts, I. (1992), "Bivariate Simplicial Depth," technical report, University of Antwerp, Dept. of Math & Computer Science.

Singh, K., and Berk, R. H. (1994), "A Concept of Type-2 *P*-Value," *Statistica Sinica*, 4, 493–504.

Tukey, J. W. (1975), "Mathematics and Picturing Data," *Proceedings of the 1974 International Congress of Mathematicians, Vancouver*, 2, 523–531.