*The Annals of Statistics*
1989, Vol. 17, No. 1, 141–159

# BOOTSTRAP AND RANDOMIZATION TESTS OF SOME NONPARAMETRIC HYPOTHESES

BY JOSEPH P. ROMANO

*Stanford University*

In this paper, the asymptotic behavior of some nonparametric tests is studied in situations where both bootstrap tests and randomization tests are applicable. Under fairly general conditions, the tests are asymptotically equivalent in the sense that the resulting critical values and power functions are appropriately close. This implies, among other things, that the difference in the critical functions of the tests, evaluated at the observed data, tends to 0 in probability. Randomization tests may be preferable since an exact desired level of the test may be obtained for finite samples. Examples considered are: testing independence, testing for spherical symmetry, testing for exchangeability, testing for homogeneity, and testing for a change point.

**1. Introduction.** The main goal of this paper is to study the behavior of some nonparametric tests having a common structure. In particular, two methods to simulate a null distribution will be analyzed and compared. The bootstrap method, formulated by Efron (1979), has been shown to be a widely applicable method in testing problems; see Beran (1986) and Romano (1988). In this paper, the problem of testing a nonparametric hypothesis is considered in situations where randomization ideas apply. The idea of randomization dates back to Fisher (1935), and then Pitman (1937/38). Both bootstrap and randomization methods are the same in that rejection of a null hypothesis occurs when a common test statistic is large. However, the approaches differ in that critical values are determined by (usually) distinct resampling methods to estimate a null distribution.

The statistical problem considered here has the following form. Given a sample $X_1, \dots, X_n$ of $S$-valued random variables, we wish to test the null hypothesis $H_0$ that the unknown probability distribution $P$ on $S$ generating the data belongs to a certain class $\Omega_0$ against the alternative class $\Omega_1$. Here, if $\Omega$ represents the class of all probabilities on $S$, then $\Omega_1$ will typically be $\Omega - \Omega_0$. Moreover, $\Omega_0$ can be characterized as the set of probabilities $P$ satisfying $\tau P = P$ for some mapping $\tau$ from $\Omega$ to $\Omega_0$. Furthermore, if $\delta$ is a metric on the space of probabilities on $S, \Omega_0$ is specified by $\delta(P, \tau P) = 0$.

Let $\hat{P}_n$ be the empirical measure of $X_1, \dots, X_n$. Then, the proposed test rejects for large values of $T_n = T_n(X_1, \dots, X_n)$, where $T_n$ is of the form

$$T_n = n^{1/2}\delta(\hat{P}_n, \tau\hat{P}_n), \tag{1.1}$$

so that the test rejects when $\tau\hat{P}_n$ is sufficiently far from $\hat{P}_n$.

Received April 1987; revised June 1988.

*AMS* 1980 *subject classifications*. Primary 62G10; secondary 62E20, 62G30.

*Key words and phrases*. Bootstrap, nonparametric tests, randomization tests, testing independence, testing equality of distributions, testing for rotational invariance, Vapnik–Cervonenkis classes.

A typical choice for $\delta$, in the spirit of Kolmogorov–Smirnov test statistics, is

$$\delta_{\mathbf{V}}(P, Q) = \sup\{V \in \mathbf{V}: |P(V) - Q(V)|\} \tag{1.2}$$

for some collection of events $\mathbf{V}$. It has the advantages of being generally applicable (especially for complex data) and yields tests with good power properties; see Remark 2.2. Henceforth, we restrict attention to $T_n$ given by (1.1) and $\delta$ give by (1.2).

Next, we give a typical example of the testing problem just described.

Example 1 (Testing independence). Let $X_1, \dots, X_n$ be i.i.d. $S$-valued random variables and suppose $X_i = (X_{i,1}, \dots, X_{i,d})$ is made up of $d$ components. The problem is to test the joint independence of the components. To get started, suppose the $j$th component takes values in a space $S_j$ and $S$ is the product space $S = \times_{j=1}^{d} S_j$. If $P$ is a probability on $S$, let $P_j$ be the marginal probability on $S_j$ of the $j$th component. If $P$ is a probability on $S$ with marginals $P_j$, let $\tau P$ be the product probability $\times_{j=1}^{d} P_j$. Note that $\tau P = P$ if and only if $P$ is a product of its marginals. Then, the proposed test statistic is (1.1), with $\delta = \delta_{\mathbf{V}}$ given by (1.2), where $\mathbf{V} = \times_{j=1}^{d} \mathbf{V}_j$ and $\mathbf{V}_j$ is a collection of sets in $S_j$. Notice the generality of the problem and the flexibility of the choice of test statistic. Specifically, no continuity assumption on the underlying distribution is made. Also, the component spaces can, in fact, be quite general. Indeed, they can be different; some variables could be quantitative (continuous or discrete), while others might be qualitative or categorical. Also, the results allow for a choice in the collection of sets $\mathbf{V}_j$ defining $T_n$. For example, if $S_1$ is the plane, several reasonable choices for $\mathbf{V}_1$ exist: all lower left-hand quadrants, all half-spaces or all ellipses, for example. In summary, no assumptions will be made on the underlying probability law $P$ of the data. However, the class of sets $\mathbf{V}_j$ chosen must be a Vapnik–Cervonenkis class.

1.1. *Bootstrap test.* Let $J_n(P)$ be the law of $T_n(X_1, \dots, X_n)$ when $X_1, \dots, X_n$ are i.i.d. $P$. In order to obtain a critical value for a test based on $T_n$, $J_n(P)$ must be approximated for $P \in \Omega_0$; that is, $J_n(\tau P)$ must be approximated. The bootstrap procedure is to estimate $J_n(\tau P)$ by $J_n(\tau \hat{P}_n)$ and then use the corresponding critical value from this estimated sampling distribution. We formally define a bootstrap critical value as follows. Let $J_n(t, P) = P[T_n(X_1, \dots, X_n) \le t]$. For $\alpha \in (0, 1)$, let

$$b_n(\alpha, P) = \inf\{t: J_n(t, P) \ge 1 - \alpha\}.$$

Then, the nominal level $\alpha$ bootstrap test rejects when $T_n > b_n(\alpha, \tau \hat{P}_n)$. The random variable $b_n(\alpha, \tau \hat{P}_n)$ is called a bootstrap critical value.

In Romano (1988), such a bootstrap procedure is applied to several examples (testing goodness of fit, testing independence, testing for spherical symmetry, etc.), and it is established that

$$P_0\left[T_n > b_n\left(\alpha, \tau \hat{P}_n\right)\right] \to \alpha \quad \text{as } n \to \infty \tag{1.3}$$

for any $P_0$ in $\Omega_0$ and such tests are consistent against all alternatives.

In this paper, this general testing problem is specialized to cases where the distribution of the data $H_0$ is invariant under a transformation group, thus leading to a randomization procedure as a competitor to the bootstrap procedure.

1.2. *Randomization test.* The following is assumed. As the notation suggests [borrowed from Hoeffding (1952)] the objects considered are defined for an infinite sequence of positive integers $n$ in anticipation of some asymptotic results. The observation $\mathbf{x}_n$ takes values in a sample space $S^{(n)}$. Typically, $\mathbf{x}_n$ is a vector of $n$ i.i.d. $S$-valued random variables. Let $\mathbf{G}_n$ be a group of transformations of $S^{(n)}$ onto itself. For now, assume $\mathbf{G}_n$ is finite with $M_n$ elements. We assume the hypothesis implies that the distribution $P^{(n)}$ of $\mathbf{x}_n$ is invariant under $\mathbf{G}_n$; that is, for every $g$ in $\mathbf{G}_n$, $g\mathbf{x}_n$ and $\mathbf{x}_n$ have the same distribution. Let $T_n$ be any real-valued test statistic defined on $S^{(n)}$. For every $x$ in $S^{(n)}$, let

$$T_n^{(1)}(x) \le T_n^{(2)}(x) \le \cdots \le T_n^{(M_n)}(x)$$

be the ordered values of $T_n(gx)$ for all $g$ in $G_n$. Given a number $\alpha$ in $(0,1)$, let $k_n = k_n(\alpha)$ be defined by $k_n = M_n - [M_n\alpha]$, where $[t]$ denotes the largest integer less than or equal to $t$. Let $M_n^+(x)$ and $M_n^0(x)$ be the number of values $T_n^{(j)}(x)$, $j = 1, \dots, M_n$, which are greater than $T_n^{(k_n)}(x)$ and equal to $T_n^{(k_n)}$, respectively. Define

$$a_n(x) = \frac{M_n\alpha - M_n^+(x)}{M_n^0(x)}. \tag{1.4}$$

Let $\phi_n(x)$ be the test function equal to 1 if $T_n(x) > T_n^{(k_n)}(x)$, 0 if $T_n(x) < T_n^{(k_n)}(x)$ and equal to $a_n(x)$ if $T_n(x) = T_n^{(k_n)}(x)$. Define $r_n(\alpha, \mathbf{x}_n) = T_n^{(k_n)}(\mathbf{x}_n)$ to be a randomization critical value. Then, for any $P^{(n)}$ which is invariant under $\mathbf{G}_n$, $\mathbf{E}_{P^{(n)}}[\phi_n(\mathbf{x}_n)] = \alpha$. Such a test will be referred to as a randomization test.

It is a well-known argument why the test $\phi_n$ has exact level $\alpha$. In particular, let $\mathbf{G}_n x$ be the $\mathbf{G}_n$-orbit of $x$ in $S$; that is, $\mathbf{G}_n x$ is the set $\mathbf{G}_n x = \{gx | g \in \mathbf{G}_n\}$. Now, conditional on $\mathbf{x}_n \in \mathbf{G}_n x$, the test statistic is equally likely to be any of the values $T_n^{(j)}(x)$, $1 \le j \le M_n$. Hence, a conditional level $\alpha$ test has been constructed for each $x$, yielding a test with unconditional level $\alpha$ as well.

The randomization distribution of $T_n$ will be denoted by $J_n(P^{(n)}|\mathbf{G}_n\mathbf{x}_n)$. That is, $J_n(P^{(n)}|\mathbf{G}_n\mathbf{x}_n)$ is the conditional distribution of $T_n(\mathbf{y}_n)$ under $P^{(n)}$ given that $\mathbf{y}_n$ falls in $\mathbf{G}_n\mathbf{x}_n$. Thus, if $P^{(n)}$ is invariant under $\mathbf{G}_n$, $J_n(P^{(n)}|\mathbf{G}_n\mathbf{x}_n)$ is the random distribution assigning equal mass to each of the $M_n$ values $T_n(g_j\mathbf{x}_n)$. As a consequence, $J_n(P^{(n)}|\mathbf{G}_n\mathbf{x}_n)$ does not depend on $P^{(n)}$ if $P^{(n)}$ is invariant under $\mathbf{G}_n$.

The connection with the bootstrap set-up should be apparent. When $\mathbf{x}_n$ is a vector of $n$ i.i.d. variables with distribution $P$, then $P^{(n)} = P^n$ and $J_n(P^n|\mathbf{G}_n\mathbf{x}_n)$ is actually independent of $P$ for $P$ in $\Omega_0$. Similar to $J_n(t, P)$, define $J_n(t, P^{(n)}|\mathbf{G}_n\mathbf{x}_n)$ to be the conditional probability that $T_n(\mathbf{y}_n)$ is less than or equal to $t$ given that $\mathbf{y}_n$ falls in $\mathbf{G}_n\mathbf{x}_n$ and $\mathbf{y}_n$ has distribution $P^{(n)}$. Both bootstrap and randomization tests reject for large values of $T_n$. The difference is that critical values are determined by referring to distinct distributions $J_n(\tau\hat{P}_n)$ and $J_n(P_0^n|\mathbf{G}_n\mathbf{x}_n)$.

EXAMPLE 1 (Testing independence, continued). The sample space $S^{(n)}$ is $S^n$, where $S$ is as explained in Example 1. An element $\mathbf{x}_n = (x_1, \dots, x_n)$ in $S^n$ is therefore made up of $d$ components, so that $x_i = (x_{i,1}, \dots, x_{i,d})$ with $x_{ij}$ in $S_j$. Let $\pi_j = \pi_j^n$, $1 \le j \le n!$, be the $n!$ permutations of $\{1, \dots, n\}$. Given integers $i_1, \dots, i_d$, each between 1 and $n!$, let $g_{i_1,\dots,i_d} \in \mathbf{G}_n$ be defined to transform $\mathbf{x}_n$ into $\mathbf{y}_n$, where $\mathbf{y}_n$ has $k$th component $y_k$ (in $S$) given by $y_{k,j} = x_{\pi_{i_k}, j}$. Then, for any $g$ in $\mathbf{G}_n$ and any $P$ in $\Omega_0$, $gP^n = P^n$; that is, the distribution of $g\mathbf{x}_n$ is the same as $\mathbf{x}_n$ if $\mathbf{x}_n$ has distribution $P^n$. In words, under the hypothesis of independence, we can, for each $j$, permute the data values in $S_j$ with each other to form a new data set which has the same distribution as the original data set. For each such data set $\mathbf{x}_{n,b}^*$, $b = 1, \dots, M_n$, $T(\mathbf{x}_{n,b}^*)$ is computed and the empirical distribution of these values is the randomization distribution. Notice that each new data set $\mathbf{x}_{n,b}^*$ may be expressed as $\mathbf{x}_{n,b}^* = g\mathbf{x}_n$ for some $g$ in $\mathbf{G}_n$. So, one may choose $g$'s with or without replacement from $\mathbf{G}_n$. The case of choosing $B = M_n$ $g$'s without replacement from $\mathbf{G}_n$ corresponds to exact evaluation of the randomization distribution. Unfortunately, in the case of testing independence, the number of $g$'s one needs for an exact evaluation is $(n!)^{d-1}$, so this approach may not be practical. The results obtained here apply even when bootstrap and randomization distributions must be approximated by Monte Carlo; see Section 4.

The following point may help to understand the conceptual distinction between the bootstrap and randomization procedures. The bootstrap distribution may be viewed as an unconditional approximation to the null distribution of the test statistic while the randomization distribution may be viewed as a conditional distribution of the test statistic. In the notation previously defined, $J_n(\tau\hat{P}_n) = J_n(P_0^n|S)$, where $P_0$ is any member of $\Omega_0$. It if were the case that $J_n(P_0^n|\mathbf{G}_n\mathbf{x}_n)$ did depend on the actual $P$ in $\Omega_0$, an alternative or combined approach might be to approximate the conditional distribution $J_n(P^n|\mathbf{G}_n\mathbf{x}_n)$ by a bootstrap procedure, say $J_n(\tau\hat{P}_n^n|\mathbf{G}\mathbf{x}_n)$. In this way, the randomization distribution may be considered a conditional bootstrap distribution.

The results obtained in this paper may be summarized as follows. The bootstrap and randomization distributions are uniformly close in the following sense. If $\mathbf{x}_n$ has distribution $P_0^n$ with $P_0$ in $\Omega_0$, then

$$\sup_t |J_n(t, \tau\hat{P}_n) - J_n(t, P_0^n|\mathbf{G}_n\mathbf{x}_n)| \to 0 \quad \text{in probability.} \tag{1.5}$$

Moreover, each distribution, say $J_n(t, \tau\hat{P}_n)$ may be approximated by a strictly increasing continuous distribution, say $J(t, P_0)$ which is not random and depends only on $P_0$; that is, we also have

$$\sup_t |J_n(t, \tau\hat{P}_n) - J(t, P_0)| \to 0 \quad \text{in probability} \tag{1.6}$$

and

$$\sup_t |J_n(t, P_0) - J(t, P_0)| \to 0 \tag{1.7}$$

as well. Thus, the difference in corresponding critical values tends to 0 in probability:

$$d_n(\alpha, \tau\hat{P}_n) - r_n(\alpha, \mathbf{x}_n) \to 0 \quad \text{in probability.} \tag{1.8}$$

Moreover, analogous results hold for the power of the tests under alternatives because critical values are still determined under the null hypothesis. In particular, (1.8) still holds if $P$ is not in $\Omega_0$. Moreover, (1.5) is also true if $\mathbf{x}_n$ has distribution $P^n$. Actually, even more is true. The critical values for both procedures tend to the common finite value $d(\alpha, \tau P)$ in probability, where $d(\alpha, \tau P)$ is the upper $\alpha$-quantile of the liming distribution $J(t, \tau P)$. This implies that the difference in the critical functions of the tests, evaluated at the observed data, tends to 0 in probability. Also, the probability that the randomization test is randomized tends to 0. It also easily follows that both tests are consistent. In the same way, one can study the power functions of the tests against general alternatives $Q_n$ appropriately defined to yield a limiting power value less than 1. For instance, suppose $Q_n$ satisfies $\delta_{\mathbf{V}}(Q_n, P_0) = O(n^{-1/2})$ for some $P_0$ in $\Omega_0$. Then, if $T_n$ has a limiting continuous distribution under $Q_n$ and (1.8) holds, then the power of both tests tends to the same value. The result is that the power functions of both tests may be said to be asymptotically equivalent. Hence, the randomization test may be preferable since it has exact level $\alpha$ for finite samples.

In Section 2, these results are made clear and a general methodology for proving these claims is developed. Several examples are introduced in Section 3 for which the results apply. Details of the proofs are given in Section 5.

It should be pointed out that the program developed here is similar to that carried out by Hoeffding (1952). He obtained similar results for randomization tests based on test statistics arising from optimal parametric procedures. For example, he obtains results for the permutation test of whether a correlation is 0 based on the optimal test statistic in the case of Gaussian data. In contrast, the problems considered here are tackled from a purely nonparametric point of view. In addition, comparisons with the bootstrap are made. Moreover, the distributions of the test statistics considered in this paper cannot be approximated by a simple asymptotic distribution, such as a Gaussian or chi-squared approximation, which further shows the power of simulation techniques.

**2. Asymptotic results.** In this section, we outline the justification for the claims made in Section 1, followed by several examples in Section 3. Details of the proofs appear in Section 5.

To summarize the problem, we now focus on the i.i.d. case. Slight extensions will be presented in Examples 4 and 5. Given a sample $X_1, \dots, X_n$ of i.i.d. $S$-valued random variables, we wish to test the null hypothesis that $\tau P = P$ for some specified $\tau$. The test statistic $T_n$ is given by (1.1) with $\delta$ given by (1.2). Furthermore, we will assume the choice of sets $\mathbf{V}$ in the definition of $\delta$ is a countable Vapnik–Cervonenkis (V.C.) class of subsets of $S$. The restriction to countable classes of sets is to avoid measurability problems and may be weakened, but there do not appear to be any applications which warrant the need to do so. Tacit is the assumption that we choose $\mathbf{V}$ large enough so that $\delta_{\mathbf{V}}$ is indeed a metric. In general, the results may apply to testing $\delta_{\mathbf{V}}(P, \tau P) = 0$.

To study the asymptotic behavior of the bootstrap and randomization tests, it is first helpful to review the bootstrap. All details may be found in Romano (1988). In order for the bootstrap to succeed, the distribution of the test statistic $J_n(P)$ must be smooth as $P$ varies. But, smoothness in $J_n(P)$ can be traced to smoothness of the mapping $\tau$. The following smoothness condition on $\tau$ holds for the problems considered in this paper. It is assumed that $\tau$ is differentiable in the following sense: If $P \in \Omega_0$, then there exists a mapping $f(\cdot, \cdot, P)$ on $S \times \mathbf{V}$ so that

$$\tau Q(V) = \tau P(V) + \int_S f(x, V, P)\, d(Q - P) + o(|Q - P|_{\mathbf{V}}) \tag{2.1}$$

as $|Q - P|_{\mathbf{V}} \to 0$ and $|P - P_0|_{\mathbf{V}} \to 0$ for some $P_0$ in $\Omega_0$. Here, $|\cdot|_{\mathbf{V}}$ is the supremum norm in $L_\infty(\mathbf{V})$, the metric space of real-valued bounded functions on $\mathbf{V}$. In order to analyze $J_n(P_0)$, consider the process $S_n(\cdot)$, given by

$$S_n(V) = n^{1/2}\left[\hat{P}_n(V) - \tau\hat{P}_n(V)\right]. \tag{2.2}$$

Regard $S_n$ as a random variable on $L_\infty(\mathbf{V})$. Then the test statistic $T_n$ is just the norm of the process $S_n$. Letting

$$\psi(x, V, P) = 1(x \in V) - f(x, V, P), \tag{2.3}$$

show that the test statistic behaves approximately as $|Z_n(\cdot)|_{\mathbf{V}}$, where

$$Z_n(V) = \int_S \psi(x, V, P_0)\, d\left(\hat{P}_n - P_0\right). \tag{2.4}$$

Of course, $Z_n$ is just the empirical process indexed by the class of functions $\mathbf{F}_{\mathbf{V}}(P_0) = \{\psi(\cdot, V, P_0), V \in \mathbf{V}\}$. Because of the linear structure here [and assumptions on the functions $\psi(\cdot, V, P_0)$], $Z_n$ is approximately a mean 0 Gaussian process $Z$ indexed by $\mathbf{V}$ with covariance function

$$\operatorname{Cov}\left[Z_n(V), Z_n(W)\right] = \int_S \psi(x, V, P_0)\psi(x, W, P_0)\, dP_0(x). \tag{2.5}$$

If this approximation is valid as $P_0$ varies as well, then the bootstrap will succeed. In fact, uniformity in $P$ can often be expressed in the following way. Define the metric (or possibly a pseudometric if $\mathbf{V}$ is not large enough) $d_{\mathbf{V}}(P, Q)$ between probabilities $P$ and $Q$ to be the supremum of $|P(V) - Q(V)|$ over sets $V$ in $\mathbf{V}$ and $\mathbf{V} \cap \mathbf{V}$. Then, the following condition typically holds.

Condition A. Fix $P_0$ in $\Omega_0$. If $P_n$ is in $\Omega_0$ with $d_{\mathbf{V}}(P_n, P_0) \to 0$, then $J_n(\cdot, P_n)$ converges weakly to a continuous strictly increasing distribution $J(\cdot, P_0)$.

The approach to verifying Condition A then consists in analyzing the process $S_n$ defined in (2.2). So, let $L_n(P)$ be the law of $S_n$ [as a r.v. on $L_\infty(\mathbf{V})$] based on $n$ observations from $P$. From the smoothness of $\tau$, Condition A is implied by

CONDITION B. Fix $P_0$ in $\Omega_0$. If $P_n$ is in $\Omega_0$ with $d_{\mathbf{V}}(P_n, P_0) \to 0$, then

$$\rho(L_n(P_n), L(P_0)) \to 0 \quad \text{as } n \to \infty,$$

where $L(P_0)$ is the distribution of a mean 0 Gaussian process $Z$ with covariance given by (2.5) and $\rho$ is any metric metrizing weak convergence of probabilities on $L_\infty(\mathbf{V})$. Moreover, assume $Z$ has its paths in a separable subset of $L_\infty(\mathbf{V})$.

Condition B has been verified for several examples in Romano (1988). Of course, Condition A implies that the bootstrap is valid in the sense (1.3). Furthermore, if $b(\alpha, P)$ denotes the upper $\alpha$-quantile of $J(\cdot, P)$, then the bootstrap critical value $b_n(\alpha, \tau\hat{P}_n) \to b(\alpha, P_0)$ in probability if $P_0$ (assumed to be in $\Omega_0$) is the true law. To study the consistency of the bootstrap test, assume

CONDITION C. The map $\tau$ is continuous in the following sense. For any sequence $P_n$ and any $P$, if $d_{\mathbf{V}}(P_n, P) \to 0$, then $d_{\mathbf{V}}(\tau P_n, \tau P) \to 0$.

Condition C is evidently weak and is easy to check. Conditions A and C imply that if $P$ is the true distribution (whether or not in $\Omega_0$), then the bootstrap critical value tends to $b(\alpha, \tau P)$ in probability, but under an alternative the test statistic $T_n \to \infty$ in probability. In summary, we have

PROPOSITION 2.1. *Condition* B *implies Condition* A, *which implies* (1.3). *If Condition* C *holds as well, then for any $P$ satisfying $\delta_{\mathbf{V}}(P, \tau P) > 0$, we have*

$$P\left[T_n > b_n\left(\alpha, \tau\hat{P}_n\right)\right] \to 1 \quad \text{as } n \to \infty. \tag{2.6}$$

We now proceed to analyzing the randomization test. Since, in fact, Condition B holds in all the examples we will consider, the methodology used will depend on already having verified Condition B. In particular, if we fix $P_0$ in $\Omega_0$, we already know the unconditional distribution $J_n(\cdot, P_0)$ has a continuous strictly increasing weak limit $J(\cdot, P_0)$ so that both (1.6) and (1.7) are true. Hence, to verify (1.5) it suffices to show

$$\sup_t |J_n(t, P_0^n | \mathbf{G}_n \mathbf{x}_n) - J(t, P_0)| \to 0 \quad \text{in probability}. \tag{2.7}$$

Roughly speaking, this means we must show that the conditional distribution of $T_n$ given the $\sigma$-field generated by the partition of $\mathbf{G}_n$-orbits converges weakly to the same limit as the unconditional distribution of $T_n$. The following elementary condition, due to Hoeffding (1952) implies (2.7) is true.

CONDITION D. Let $G_n$ and $G_n'$ be random transformations which are uniformly distributed over $\mathbf{G}_n$ and independent of the observation $\mathbf{x}_n$. Here, $\mathbf{x}_n$ has distribution $P^n$ which need not be invariant under $\mathbf{G}_n$. Then, $T_n(G_n \mathbf{x}_n)$ and

$T_n(G_n'\mathbf{x}_n)$ are asymptotically independent, each with a continuous increasing limiting cdf $J(\cdot, Q)$.

It follows from Hoeffding (1952), Theorem 3.2, and the assumptions on $J(\cdot, Q)$ that Condition D implies

$$\sup_t |J_n(t, P_0^{(n)} | \mathbf{G}_n \mathbf{x}_n) - J(t, Q)| \to 0 \quad \text{in probability,} \tag{2.8}$$

where $P_0^{(n)}$ is any distribution invariant under $\mathbf{G}_n$. Actually, Hoeffding proves a pointwise (for fixed $t$) result, but a stronger statement is possible when (as is the case here) the limit distribution $J(\cdot, Q)$ is known to be continuous. Hence, in the case $P^{(n)} = P_0^n$ with $P_0$ in $\Omega_0$, $Q$ necessarily equals $P_0$ and (2.7) holds. Moreover, in the case $P^{(n)} = P^n$ but $P$ is not in $\Omega_0$, Condition D will be verified to imply (2.8) with $Q = \tau P$, yielding consistency results about the test.

To obtain the validity of Condition D, consider the process

$$S_n(\mathbf{x}_n, V) = n^{1/2}\left[\hat{P}_n(\mathbf{x}_n, V) - \tau\hat{P}_n(\mathbf{x}_n, V)\right], \tag{2.9}$$

where $\hat{P}_n(\mathbf{x}_n, V)$ is the empirical measure of an observation $\mathbf{x}_n$ from $S^{(n)}$ evaluated at the set $V$. Regard $(S_n(G_n\mathbf{x}_n, \cdot), S_n(G_n'\mathbf{x}_n, \cdot))$ as a random variable on the product space of $L_\infty(\mathbf{V})$ with itself. The joint distribution of $(T_n(G_n\mathbf{x}_n), T_n(G_n'\mathbf{x}_n))$ is just the joint distribution of

$$\left(|S_n(G_n\mathbf{x}_n, \cdot)|_{\mathbf{V}}, |S_n(G_n'\mathbf{x}_n, \cdot)|_{\mathbf{V}}\right).$$

Hence, Condition D is trivially implied by the following.

CONDITION E. Suppose $\mathbf{x}_n$ has distribution $P^{(n)}$. Then, $S_n(G_n\mathbf{x}_n, \cdot)$ and $S_n(G_n'\mathbf{x}_n, \cdot)$ are asymptotically independent each with law $L(Q)$.

In order to verify Condition E in the case $P^{(n)}$ is invariant under $\mathbf{G}_n$, one may replace $G_n'$ by the identity transformation.

Now consider the case $P^{(n)} = P_0^n$ and suppose $P_0$ is in $\Omega_0$. From what we already know, $S_n(\mathbf{x}_n, \cdot)$ and $S_n(G_n\mathbf{x}_n, \cdot)$ each have weak Gaussian limits $L(P_0)$, so when considered jointly on the product space, the random variable

$$R_n = R_n(\mathbf{x}_n, G_n) = \left(S_n(\mathbf{x}_n, \cdot), S_n(G_n\mathbf{x}_n, \cdot)\right)$$

is uniformly tight. Therefore, all we need do is analyze the finite dimensional distributions of the process $R_n$. Using the differentiability (2.1) will help establish the joint asymptotic Gaussianity of the limiting finite dimensional distributions of $R_n$, and a covariance calculation should determine the required independence. In summary:

PROPOSITION 2.2. *Condition* E *implies Condition* D. *If Condition* D *holds when* $P^{(n)} = P^n$ *for some* $P$ *not in* $\Omega_0$ *with* $Q = \tau P$, *then the randomization critical value* $r_n(\alpha, \mathbf{x}_n) \to J(\alpha, \tau P)$ *in probability. Hence, if Condition* C *holds as well, the randomization test is consistent.*

We are now in a position to more fully compare the two procedures in the i.i.d. case. Assume Conditions A–C hold and D holds when $P^{(n)} = P^n$. For any fixed $P$ (in $\Omega_0$ or not), the difference between the randomization and bootstrap distributions tends to 0 uniformly in $t$; that is, Propositions 2.1 and 2.2 yield

$$\sup_t |J_n(t, \tau\hat{P}_n) - J_n(t, P_0^n|\mathbf{G}_n\mathbf{x}_n)| \to 0 \quad \text{in } P^n\text{-probability}, \tag{2.10}$$

where $P_0$ is any member of $\Omega_0$. It follows (using 2.8) that the difference in critical values tends to 0 in $P^n$-probability. This implies that the probability that the tests differ (in whether or not to reject the null hypothesis) tends to 0. Furthermore, both tests are consistent against all alternatives. However, such a result is practically useless in comparing the power of the two tests. In order to obtain a more useful result, consider a sequence of alternatives $Q_n$ to $\Omega_0$ and study the asymptotic power (if it exists) against such a sequence. In order to get an interesting limit for the asymptotic power, it should be clear that one needs to get close to $\Omega_0$ and, in fact, $Q_n$ should satisfy $\delta_{\mathbf{V}}(Q_n, \tau Q_n) = O(n^{-1/2})$. The next proposition gives conditions when the bootstrap and randomization distributions are uniformly close under general sequences of alternatives $Q_n$ and when the limiting power of each test against $Q_n$ is the same. Note, the proposition includes the possibility that the function $H(\cdot)$ could be identically 0.

PROPOSITION 2.3. *Assume Conditions* A–C *hold. Let* $Q_n$ *be a sequence of alternatives to* $\Omega_0$ *satisfying* $d_{\mathbf{V}}(Q_n, Q_\infty) \to 0$. *Assume Condition* E *holds with* $P^{(n)} = Q_n^n$ *and* $Q = \tau Q_\infty$. *Then*

$$\sup_t |J_n(t, \tau\hat{P}_n) - J_n(t, P_0^n|\mathbf{G}_n\mathbf{x}_n)| \to 0 \quad \textit{in } Q_n^n\textit{-probability}, \tag{2.11}$$

*where* $P_0$ *is any member of* $\Omega_0$. *Also, the corresponding critical values satisfy*

$$b_n(\alpha, \tau\hat{P}_n) - r_n(\alpha, \mathbf{x}_n) \to 0 \quad \textit{in } Q_n^n\textit{-probability}. \tag{2.12}$$

*Furthermore, suppose there exists a continuous function* $H(\cdot)$ *so that*

$$Q_n^n(T_n \le t) \to H(t).$$

*Then, the asymptotic power of both tests against the sequence of alternatives* $Q_n$ *is* $1 - H(b(\alpha, \tau Q_\infty))$.

REMARK 2.1. In order to compute the limiting distribution of $T_n$ under $Q_n^n$ (or show its existence), consider the process $S_n$ defined in (2.2). Suppose $Q_n$ satisfies an approximation like $Q_n \approx P_0 + \Delta n^{-1/2}$, where $\Delta$ is some element in $L_\infty(\mathbf{V})$. The differentiability condition (2.1) shows $S_n$ is approximately a Gaussian process with covariance (2.5), but this time with mean $\Delta$. Thus, the limiting distribution of $T_n$ under $Q_n^n$ is the distribution of the supremum of such a process.

REMARK 2.2. Why use the tests considered in this paper anyway? The power properties of Kolmogorov–Smirnov distance type tests have been well studied in several problems; see Blum, Kiefer and Rosenblatt (1961) and Kiefer (1959), for example. It is clear that these properties hold quite generally in problems of the type considered in this paper. To be more specific, consider a sequence of alternatives $Q_n$ to $\Omega_0$ satisfying an approximation like $Q_n \approx P_0 + \Delta\varepsilon_n$. Then, the argument in Remark 2.1 shows the power of the bootstrap and randomization tests against $Q_n \to 1$ if $(\varepsilon_n/n^{1/2}) \to \infty$. Moreover, in the case $\varepsilon_n = n^{1/2}$, the power can be made arbitrarily close to 1 if $|\Delta|$ is large. To see why, choose $V_0$ so $\Delta(V_0) \neq 0$ and let $\Delta_0 = \Delta(V_0)$. Then, the power of the bootstrap test, for example, is bounded below by

$$Q_n^n\big(n^{1/2}|\hat{P}_n(V_0) - \tau\hat{P}_n(V_0)| \geq b_n\big(\alpha, \tau\hat{P}_n\big)\big).$$

The distribution of the random variable on the left side of the inequality tends to the distribution of the absolute value of a real-valued Gaussian random variable with mean $\Delta_0$ (and variance which only depends on $P_0$ and $V_0$ and not the original choice of $\Delta$). Also, the right side of the inequality tends to $b(\alpha, P_0)$ in probability. Hence, the limiting power is bounded below by

$$P\big(|Z + \Delta_0| \geq \sigma b(\alpha, P_0)\big),$$

for some $\sigma = \sigma(V_0, P_0)$. Now, increase $\Delta_0$. In summary, the limiting power of supremum tests against alternatives converging to $\Omega_0$ at the $n^{1/2}$ rate is not degenerate. In typical smooth problems, this is the best obtainable rate.

**3. Examples.** In this section, the methodology of Section 2 is applied to some examples.

EXAMPLE 1 (Testing independence, continued).

PROPOSITION 3.1. *Assume* $\mathbf{V}_j$ (*in the definition of the test statistic* $T_n$) *is a countable V.C. class in* $S_j$. *Conditions* A–E *all hold when* $P^{(n)} = P^n$ *for any fixed* $P$ (*whether* $P$ *is in* $\Omega_0$ *or not*). *Thus, both bootstrap and randomization tests are consistent and asymptotically equivalent in the sense* (2.10). *Moreover, suppose* $Q_n$ *is any sequence of alternatives to* $\Omega_0$ *satisfying* $d_{\mathbf{V}}(Q_n, Q_\infty) \to 0$ *for some* $Q_\infty$. *Then, Condition* E *holds for* $P^{(n)} = Q_n^n$ *and* $Q = \tau Q_\infty$, *so that* (2.11) *and* (2.12) *follow.*

EXAMPLE 2 (Testing for rotational invariance). The problem is to test whether the underlying probability distribution on $S = \mathbf{R}^p$ belongs to the class $\Omega_0$ of rotationally invariant or spherically symmetric distributions. Let $S_r \subset S$ be the sphere of radius $r$ and center 0. If $X = (X_1, \dots, X_p)$ has probability distribution $P$ on $\mathbf{R}^p$, then $P$ is completely specified by the marginal distribution $P_R$ of $R = (\sum_{i=1}^p X_i^2)^{1/2}$ and the conditional distribution $P_{X|R}$ of $X$ given $R$. Of course, if $P$ is spherically symmetric, then $P_{X|R=r}$ is always uniformly distributed on $S_r$. Let $\tau P$ be the distribution $Q$ in $\Omega_0$ such that $P_R = Q_R$. Then, the proposed test statistic becomes

$$T_n = n^{1/2} \sup_{V \in \mathbf{V}} |\hat{P}_n(V) - \tau\hat{P}_n(V)|. \tag{3.1}$$

To see that a randomization test is applicable, we must identify the appropriate class of transformations $\mathbf{G}_n$ on $S^p$. If $\theta$ is a point on $S_1$ and $x$ is a point on $S$, let $g_\theta x$ be the new point $\theta|x|_p$, where $|\cdot|_p$ is the usual Euclidean norm on $\mathbf{R}^p$. Then, $\mathbf{G}_n = \{(g_{\theta_q}, \dots, g_{\theta_n}): \theta_i \in S_1\}$, so an element $(g_{\theta_1}, \dots, g_{\theta_n})$ of $\mathbf{G}_n$ transforms a point $\mathbf{x}_n = (X_1, \dots, X_n)$ in $S^n$ to the point $(g_{\theta_1}X_1, \dots, g_{\theta_n}X_n)$.

Notice that $\mathbf{G}_n$ is not finite except in the case $p = 1$. Thus, the description of the test given in Section 1 is not quite accurate. In this case, $J_n(P^n|\mathbf{G}_n\mathbf{x}_n)$ is the distribution of $T_n(G_n\mathbf{y}_n)$, where $\mathbf{y}_n$ is any element of $\mathbf{G}_n\mathbf{x}_n$ and $G_n$ is uniformly distributed over $\mathbf{G}_n$. The methodology described in Section 2 for comparing the bootstrap and randomization tests is then applicable. The only technicality involved is showing that Hoeffding's Condition D implies (2.7), but an easy generalization of his proof shows the argument carries over as long as $J_n(P^{(n)}|\mathbf{G}_n\mathbf{x}_n)$ is the distribution of $T_n(\mathbf{x}_n)$ when $\mathbf{x}_n$ is uniform over $\mathbf{G}_n\mathbf{x}_n$. In general, one needs to be able to put a uniform probability distribution on $\mathbf{G}_n$. In the example here, it is clear how to do this.

To describe the class of sets $\mathbf{V}$ allowed in (3.1), embed the sample space $S$ into $S_1 \times R$, where $R$ denotes the nonnegative real numbers. A point $x$ in $S$ is identified with the point $(s_1, s_2)$ in $S_1 \times R$ if it is at distance $s_2$ from the origin and $x/|x|_p = s_1$. In the case $x$ is the origin, identify it with $(0, 0)$.

PROPOSITION 3.2. *Assume the collection of sets* $\mathbf{V}$ *in* (3.1) *is of the form* $\mathbf{V} = \mathbf{V}_1 \times \mathbf{V}_2$, *where* $\mathbf{V}_1$ *is a (countable) V.C. class in* $S_1$ *and* $\mathbf{V}_2$ *is a (countable) V.C. class in* $R$. *Conditions* A–E *all hold when* $P^{(n)} = P^n$ *for any fixed* $P$ *(whether* $P$ *is in* $\Omega_0$ *or not). Thus, both bootstrap and randomization tests are consistent and asymptotically equivalent in the sense* (2.10). *Moreover, suppose* $Q_n$ *is any sequence of alternatives to* $\Omega_0$ *satisfying* $d_{\mathbf{V}}(Q_n, Q_\infty) \to 0$ *for some* $Q_\infty$. *Then, Condition* E *holds for* $P^{(n)} = Q_n^n$ *and* $Q = \tau Q_\infty$, *so that* (2.11) *and* (2.12) *follow.*

A slight modification of the previous set-up is needed in some situations. Suppose $\Omega_0$ is now specified as the set of probabilities $P$ satisfying $\tau_j(P) = P$ for $1 \le j \le k$, where $\tau_j$ is a mapping from $\Omega_0$. Then, $P$ lies in $\Omega_0$ if and only if $\max_{1 \le j \le k} \delta(P, \tau_j(P)) = 0$ and the proposed test rejects for values of $\max_{1 \le j \le k} \delta(\hat{P}_n, \tau_j(\hat{P}_n))$.

EXAMPLE 3 (Testing whether a probability law is exchangeable). Let $X_1, \dots, X_n$ be i.i.d. $S$-valued random variables, where $X_i = (X_{i,1}, \dots, X_{i,d})$ is made up of $d$ components each living in a space $E$. Let $P$ be the probability law generating the data. The problem is to test whether $P$ is exchangeable. That is, if $D = \{1, \dots, d\}$ and $\pi_j: D \to D$, $1 \le j \le d!$, are the $d!$ permutations of $D$, then the problem is to test whether the law of $(X_{i,1}, \dots, X_{i,d})$ is the same as the law of $(X_{i,\pi_j(1)}, \dots, X_{i,\pi_j(d)})$ for every $j$. Given any probability $P$ on $S$, let $\tau_j P$ denote the law of $(X_{\pi_j(1)}, \dots, X_{\pi_j(d)})$ if $(X_1, \dots, X_d)$ has law $P$. Then the proposed test statistic takes the form

$$T_n = n^{1/2} \max_j \sup_{V \in \mathbf{V}} |\hat{P}_n(V) - \tau_j \hat{P}_n(V)|. \tag{3.2}$$

Note that, when $d = 2$, $\tau P = P$ if and only if $P \in \Omega_0$ and the test statistic is of the form (1.1). In this case, $P$ is exchangeable if $P(V) = P(V^t)$ for all (measurable) $V$, where $V^t$ is the set $\{(y, x): x \in E, y \in E, (x, y) \in V\}$.

The appropriate group of transformations $\mathbf{G}_n$ for this problem may be described as follows. A transformation $g$ in $\mathbf{G}_n$ takes an element $\mathbf{x}_n = (X_1, \dots, X_n)$ to an element $y_n = (Y_1, \dots, Y_n)$ if $Y_i$ is some permutation of $X_i$. That is, $Y_n = (X_{i, \pi(i)}, \dots, X_{i, \pi(d)})$ for some permutation $\pi$ of $D$. The permutation $\pi$ transforming $X_i$ may also depend on $i$ so that $g$ may be identified by a vector $\pi = (\pi^{(1)}, \dots, \pi^{(n)})$, where $\pi^{(k)}$ is some permutation of $D$. Then, $\mathbf{G}_n$ is the collection of all such $g$, so $\mathbf{G}_n$ has $(d!)^n$ elements.

To avoid introducing new notation, we restrict attention to the case $d = 2$. Suffice it to say that bootstrap and randomization tests are asymptotically equivalent procedures in the sense described here even if $d > 2$. The asymptotics for the bootstrap for $d > 2$ are given in Romano (1988).

PROPOSITION 3.3. *Let* $\mathbf{V}$ *be a countable V.C. class in the definition of the test statistic* $T_n$ *given by* (3.2). *Assume* $\mathbf{V}$ *contains elements* $V$ *with* $V \neq V^t$. *Then, Conditions* A–E *all hold when* $P^{(n)} = P^n$ *for any fixed* $P$ (*whether* $P$ *is in* $\Omega_0$ *or not*). *Thus, both bootstrap and randomization tests are consistent and asymptotically equivalent in the sense* (2.10). *Moreover, suppose* $Q_n$ *is any sequence of alternatives to* $\Omega_0$ *satisfying* $d_{\mathbf{V}}(Q_n, Q_\infty) \to 0$ *for some* $Q_\infty$. *Then, Condition* E *holds for* $P^{(n)} = Q_n^n$ *and* $Q = \tau Q_\infty$, *so that* (2.11) *and* (2.12) *follow.*

EXAMPLE 4 ($K$-sample test of homogeneity). The structure of the previous tests can easily be adapted to $k$ independent samples from possibly different populations. For $i = 1, \dots, k$, let $X_{ij}$, $1 \le j \le n_i$, be a sample of $S$-valued random variables with probability distribution $P_i$. The problem is to test the homogeneity hypothesis $H_0$: $P_1 = \cdots = P_k$. Let $\hat{P}_n$ be the empirical measure of all $n = \sum_{i=1}^k n_i$ observations combined and let $\hat{P}_{n_i, i}$ be the empirical measure of the $i$th sample. Then, one possible test statistic for testing $H_0$ takes the form

$$T_n = n^{1/2} \max_{1 \le i \le k} \left[ c_{n,i} \sup_{V \in \mathbf{V}} |\hat{P}_n(V) - \hat{P}_{n_i, i}(V)| \right], \tag{3.3}$$

where the $c_{n,i}$ are constants depending on the sample sizes $n_i$ and, again, $\mathbf{V}$ is an appropriate class of sets. One possible choice for $c_{n,i}$ is $(n_i/n)^{1/2}$.

Special cases of these tests were proposed by Smirnov (1939) and Kiefer (1959) in the case $S = \mathbf{R}, \mathbf{V} = \{(-\infty, t]: t \in \mathbf{R}\}$ and the assumption that the underlying probability distribution is continuous. Bickel (1969) considers the two-sample problem ($k = 2$) in the case $S = \mathbf{R}^p$ and $\mathbf{V} = \{(-\infty, t): t \in \mathbf{R}^p\}$ and shows that the randomization test is consistent against (fixed) alternatives. Here, the sample space $S$ is arbitrary and $\mathbf{V}$ is assumed to be any (countable) V.C. class. No assumptions at all (such as continuity assumptions) are made on the underlying population.

For simplicity, assume the sample size $n_i$ is a function of $n$, where $n_i = n_i(n)$ is the integer part of $\lambda_i n$ if $i < k$ and $n_k = n - \sum_{i<k} n_i$. Let $J_n(P_1, \dots, P_k)$ be the sampling distribution of $T_n$ when the $k$ samples have distributions $P_1, \dots, P_k$ and let $J_n(t, P_1, \dots, P_k)$ be the corresponding distribution function. The bootstrap null distribution is then $J_n(\hat{P}_n, \dots, \hat{P}_n)$.

To see that a randomization test is applicable, we must identify $\mathbf{G}_n$. Think of the observed data $\mathbf{x}_n$ as a vector of length $n$, ordered so that the first $n_1$ data values of $\mathbf{x}_n$ are thought of as coming from the first population, the next $n_2$ data values from the second and so on. Let $\pi$ be a permutation of the integers 1 to $n$ given $\mathbf{x}_n = (X_1, \dots, X_n)$ is in $S^n$ and let $g\mathbf{x}_n = (X_{\pi(1)}, \dots, X_{\pi(n)})$. Then, $\mathbf{G}_n$ is the collection of all such $g$. Under the null hypothesis, $g\mathbf{x}_n$ and $\mathbf{x}_n$ have the same distribution. As before, let $J_n(P^{(n)}|\mathbf{G}_n\mathbf{x}_n)$ denote the conditional distribution of the test statistic $T_n(\mathbf{y}_n)$ given that $\mathbf{y}_n$ falls in $\mathbf{G}_n\mathbf{x}_n$ and $\mathbf{y}_n$ has distribution $P^{(n)}$.

PROPOSITION 3.4. *Let* $\mathbf{V}$ *be a (countable) V.C. class in* (3.3). *Assume the sample sizes* $n_i$ *satisfy* $n_i/n \to \lambda_i$ *as* $n \to \infty$ *for some* $\lambda_i$ *in* $(0,1)$ *and* $c_{n,i} \to c_i$ *for some* $c_i > 0$. *Let* $Q^{(n)} = \times_{j=1}^{k} Q_{n,j}^{n_j}$ *be a sequence of possible distributions of the data, so that* $Q_{n,i}$ *represents the distribution of the ith sample. Assume* $d_{\mathbf{V}}(Q_{n,i}, Q_{\infty,i}) \to 0$ *as* $n \to \infty$, *for some probability* $Q_{\infty,i}$ *on S. Then,*

$$\sup_t |J_n(t, \hat{P}_n, \dots, \hat{P}_n) - J_n(t, P^{(n)}|\mathbf{G}_n\mathbf{x}_n)| \to 0 \quad \text{in } Q^{(n)}\text{-probability},$$

*where* $P^{(n)} = P^n$ *is any distribution made up of n i.i.d. components. Moreover, the corresponding critical values of the bootstrap and randomization tests tend to a common finite limit in probability. Also, bootstrap and randomization tests are consistent against any alternative of the form* $Q^{(n)} = \times_{j=1}^{k} Q_i^{n_i}$, *if the* $Q_i$ *are not all equal.*

REMARK 3.1. Similar remarks to those presented after Proposition 2.3 are applicable in this context as well. In particular, let the hypothesis on $Q^{(n)}$ in Proposition 3.4 be satisfied. If the distribution of the test statistic under $Q^{(n)}$ tends weakly to a continuous distribution, then the power of the randomization and bootstrap tests tends to a common value.

EXAMPLE 5 (Testing for a change point). Let $\mathbf{x}_n = (X_1, \dots, X_n)$ be a sample of $n$ independent random variables taking values in a sample space $S$. The null hypothesis asserts that the $X_i$'s have a common (unknown) distribution $P$. The alternative hypothesis asserts that, for some $J$, $\mathbf{x}_J = (X_1, \dots, X_J)$ are i.i.d. with a distribution $P_1$ and $\mathbf{y}_J = (X_{J+1}, \dots, X_n)$ are i.i.d. with a different distribution $P_2$. Let $\hat{P}_j$ be the empirical of the first $j$ observations and let $\hat{Q}_j$ be the empirical of all observations but the first $j$. In the spirit of the test statistics considered in this paper, a natural test statistic might be

$$T_n = \max_{1 \le j \le n} c_{n,j} \delta_{\mathbf{V}}(\hat{P}_j, \hat{Q}_j),$$

where $\delta_{\mathbf{V}}$ is the metric (1.2) and $c_{n,j}$ is some sequence of norming constants. As in the other examples, both bootstrap and randomization methods apply. In particular, the bootstrap method consists in resampling (conditional on the data) $n$ i.i.d. points from the empirical distribution of the data. The randomization method, on the other hand, consists in generating samples of size $n$ by sampling the data without replacement, or equivalently, transforming $\mathbf{x}_n$ into $\mathbf{y}_n = (X_{\pi(1)}, \dots, X_{\pi(n)})$ for some permutation $\pi$ of $\{1, 2, \dots, n\}$. The methods applied in this paper can be extended to analyze and compare the two techniques. The reader may consult James, James and Siegmund (1987) for references on the change-point problem.

**4. Stochastic approximations.** The attractiveness of the testing procedures described in this paper is marred by the burdening amount of computation involved. Monte Carlo approximations to bootstrap null distributions are described in Romano (1988). Here, we focus on the problem of implementation of the randomization procedure. Also, see Vadiveloo (1983) in this connection. Two main difficulties are apparent.

(i) The exact value of the observed test statistic $T_n(\mathbf{x}_n)$ may be difficult to obtain because the supremum of $\hat{P}_n - \tau\hat{P}_n$ over the class of sets $\mathbf{V}$ may be hard to compute. Instead, one may have to resort to computing the supremum discrepancy between $\hat{P}_n$ and $\tau\hat{P}_n$ over a finite number of search sets. One possibility is to choose these search sets $V_1, \dots, V_s$ i.i.d. (and independent of $\mathbf{x}_n$) according to a probability on $\mathbf{V}$.

(ii) The randomization null distribution, which assigns equal mass to the values $T_n(g\mathbf{x}_n)$, $g \in \mathbf{G}_n$, may be hard to compute because the number of elements in $\mathbf{G}_n$ is too large. Here, one can sample $g_1, \dots, g_r$ from $\mathbf{G}_n$ with or without replacement and approximate the randomization distribution by the distribution assigning equal mass to the values $T_n(g_i\mathbf{x}_n)$, $1 \le i \le r$.

Of course, if problem (i) is present, then $T_n(g_i\mathbf{x}_n)$ will have to be approximated as well. One possibility is to use the same search sets $V_1, \dots, V_s$ for all $i$. An alternative possibility is to allow the search sets to depend on $i$; that is, for each $i$ choose $s$ independent random sets from $\mathbf{V}$.

The important fact is the following. The above stochastic approximations do not affect the level of the test, so that the resulting test has exact desired level $\alpha$. This follows because the random vector $(T_n(\mathbf{x}_n), T_n(g_1\mathbf{x}_n), \dots, T_n(g_r\mathbf{x}_n))$ is still exchangeable, even if the $g_i$'s are chosen at random with or without replacement from $T_n$ or if $T_n$ is computed by replacing the supremum over $\mathbf{V}$ by a maximum over random search sets. In the case the same $V_j$'s are used to approximate $T_n(g_i\mathbf{x}_n)$ for each $i$, an easy way to see that the sets $V_j$ chosen do not change this exact finite sampling result is by conditioning on these chosen sets and then regard the resulting test statistic as a new test statistic in its own right of the form described in Section 1.

For the remainder of this section, we focus on the asymptotic behavior of the randomization test employing such stochastic approximations. A general methodology to handle such computational difficulties is described in Beran and Millar (1987), where it is suggested to approximate a supremum over a collection of sets by a maximum over randomly chosen sets in the context of constructing a confidence set for a measure. In general, their arguments presented in this case extend to cover the approximations described in (i) and (ii) above, where in (ii) the $g_i$'s are chosen with replacement from $\mathbf{G}_n$. In general, consistency results and the equivalency of stochastic bootstrap tests and stochastic randomization tests continue to hold, assuming they hold without the use of stochastic approximations. In order to handle the more natural case (in the context here) where the $g_i$'s are chosen without replacement from $\mathbf{G}_n$, the following result applies. We give a more general result than what is needed here, and it may be considered a Glivenko–Cantelli theorem for sampling without replacement in the general setting of Vapnik and Cervonenkis. The proof is analogous to the proof of Theorem 1, page 828, in Shorack and Wellner (1986); see Romano (1987) for details.

PROPOSITION 4.1. *Let $P_n$ be a sequence of probabilities on a space $S$. Suppose $P_n$ represents the distribution of a finite population $\mathbf{X}_n$ of $N_n$ elements, some of which may be equal. Let $\mathbf{x}_n = (X_{n,1}, \dots, X_{n,r_n})$ be a sample of size $r_n$ chosen at random without replacement from $\mathbf{X}_n$. Let $\hat{P}_n$ be the empirical measure corresponding to these $r_n$ values; that is,*

$$\hat{P}_n(V) = r_n^{-1} \sum_{i=1}^{r_n} 1(X_{n,i} \in V).$$

*Let $\mathbf{V}$ be a (countable) Vapnik–Cervonenkis class of (measurable) subsets of $S$. Then, if $r_n \to \infty$,*

$$\sup_{V \in \mathbf{V}} |\hat{P}_n(V) - P_n(V)| \to 0 \quad \textit{in } P_n^{r_n}\textit{-probability}.$$

REMARK 4.1. The case $P_n = P$, independent of $n$, is trivial. Note, however, that no assumptions are made on the sequence $P_n$.

COROLLARY 4.1. *Under the conditions of Proposition 4.1, suppose $S = \mathbf{R}$. Let $F_n$ be the cdf corresponding to the population $\mathbf{X}_n$ and let $\hat{F}_n$ be the empirical cdf based on $r_n$ observations chosen without replacement from $F_n$. Then*

$$\sup_t |\hat{F}_n(t) - F_n(t)| \to 0 \quad \textit{in probability}.$$

Corollary 4.1 can be applied to stochastic null distributions in the following way. As before, if $P_0^{(n)}$ is invariant under $\mathbf{G}_n$, then the randomization null distribution $J_n(P_0^{(n)}|\mathbf{G}_n\mathbf{x}_n)$ is the distribution corresponding to the population $\mathbf{X}_n$ of $M_n$ elements $T_n(g\mathbf{x}_n)$, $g \in \mathbf{G}_n$. A stochastic approximation to $J_n(x, P_0^{(n)}|\mathbf{G}_n\mathbf{x}_n)$, is $\hat{J}_n(x, P_0^{(n)}|\mathbf{G}_n\mathbf{x}_n)$ given by the empirical distribution of $r_n$ values, $T_n(g_j\mathbf{x}_n)$, $1 \le j \le r_n$, where the $g_j$ are sampled without replacement from $G_n$. Having established (2.8) that the randomization null distribution can be uniformly approximated by a continuous distribution $J(t, Q)$, it follows (by applying Corollary 4.1 conditional on $\mathbf{x}_n$) that the stochastic approximation $\hat{J}_n$

has the same property,

$$\sup_t |\hat{J}_n(t, P_0^{(n)}|\mathbf{G}_n\mathbf{x}_n) - J(t, Q)| \to 0 \quad \text{in probability}.$$

Then, for example, a critical value $\hat{r}_n$ based on $\hat{J}_n$ tends to $r(\alpha, Q)$ in probability, and consistency of the stochastic randomization test then follows as before.

## 5. Proofs.

PROOF OF PROPOSITION 3.1. Conditions A–C are verified in Romano (1988). Let $P_n$ be a sequence in $\Omega_0$ satisfying $d_{\mathbf{V}}(P_n, P_0) \to 0$ for some $P_0$ in $\Omega_0$. Then, from Romano (1988), $J_n(P_n)$ converges weakly to a continuous, strictly increasing limit law $J(P_0)$ and $J(P_0)$ is the distribution of the norm of a certain Gaussian process $L(P_0)$. We now verify Condition E when $P^{(n)} = Q_n^n$ and $Q = \tau Q_\infty$, where $Q_n$ is a sequence satisfying $d_{\mathbf{V}}(Q_n, Q_\infty) \to 0$. Let $S_n(\mathbf{x}_n, V)$ be given by (2.9). Here, $\mathbf{x}_n = (X_{n,1}, \dots, X_{n,n})$ is a vector of $n$ i.i.d. variables with distribution $P_n$ and $X_{n,i}$ is made up of $d$ independent components $(X_{n,i,1}, \dots, X_{n,i,d})$. Let $G_n$ and $G_n'$ be independent of $\mathbf{x}_n$ and each other, each uniformly distributed over $\mathbf{G}_n$. We must show $S_n(G_n\mathbf{x}_n, \cdot)$ $S_n(G_n'\mathbf{x}_n, \cdot)$ are asymptotically independent with law $L(Q)$. From Section 2, it suffices to examine the finite dimensional distributions of the process $(S_n(G_n\mathbf{x}_n, \cdot), S_n(G_n'\mathbf{x}_n, \cdot))$. Using the differentiability of $\tau$,

$$S_n(\mathbf{x}_n, V) = n^{-1/2} \sum_{i=1}^{n} \left\{ \prod_{j=1}^{d} \left[ Z_{n,i,j} - \mathbf{E}(Z_{n,i,j}) \right] \right\} + o_{P_n^n}(1),$$

where $Z_{n,i,j} = 1(X_{n,i,j} \in A_j)$ and $V = \times_{j=1}^{d} A_j$. Thus, Lemma 5.1 is applicable to yield that $(S_n(G_n\mathbf{x}_n, V), S_n(G_n'\mathbf{x}_n, W))$ converges weakly to a bivariate Gaussian distribution with correlation 0. An argument similar to the proof of Lemma 5.1 shows a finite linear combination of elements $S_n(G_n\mathbf{x}_n, V_i)$ is asymptotically independent of a linear combination of elements $S_n(G_n'\mathbf{x}_n, W_i)$. The result follows. □

LEMMA 5.1. *For $1 \le i \le n$, let $(Y_{n,i,1}, \dots, Y_{n,i,d})$ be $n$ i.i.d. random vectors made up of $d$ independent components. Moreover, suppose $Y_{n,i,j}$ has mean 0 and variance $\sigma_{n,j}^2$ and is bounded in absolute value by 1. Suppose the law of $Y_{n,i,j}$ converges weakly (as $n \to \infty$) to a distribution with variance $\sigma_j^2$ so that $\sigma_{n,j}^2 \to \sigma_j^2$. For $1 \le j \le d$, let $G_{n,j}$ and $G_{n,j}'$ be independent random permutations of $\{1, 2, \dots, n\}$. Define*

$$T_n = n^{-1/2} \sum_{i=1}^{n} \left[ \prod_{j=1}^{d} Y_{n, G_{nj}'(i), j} \right]$$

*and*

$$W_n = n^{-1/2} \sum_{i=1}^{n} \left[ \prod_{j=1}^{d} Y_{n, G_{nj}'(i), j} \right].$$

*Then the law of* $(T_n, W_n)$ *converges weakly to the law of a Gaussian random variable* $(T, W)$, *where* $T$ *and* $W$ *are i.i.d. with mean* 0 *and variance* $\sigma^2 = \prod_{j=1}^{d} \sigma_j^2$.

PROOF OF LEMMA 5.1. The proof follows by conditioning on all variables except the $Y_{n,i,1}$, $i = 1, \dots, n$. Using the Cramér–Wold device, the conditions for the Lindeberg central limit theorem [or Theorem A10 of Hettmansperger (1984)] hold with probability tending to 1. A precise argument is available in Romano (1987). □

PROOF OF PROPOSITION 3.2. Conditions A–C are verified in Romano (1988). Let $P_n$ be a sequence in $\Omega_0$ satisfying $d_{\mathbf{V}}(P_n, P_0) \to 0$ for some $P_0$ in $\Omega_0$. Then $J_n(P_n)$ converges weakly to a continuous strictly increasing limit law $J(P_0)$ and $J(P_0)$ is the distribution of the norm of a certain Gaussian process $L(P_0)$. We now verify Condition E when $P^{(n)} = P_n^n$ and $Q = P_0$. Let $S_n(\mathbf{x}_n, V)$ be given by (2.9). Here, $\mathbf{x}_n = (X_{n,1}, \dots, X_{n,n})$ is a vector of $n$ i.i.d. variables with distribution $P_n$. Let $G_n$ be independent of $\mathbf{x}_n$ and uniformly distributed over $\mathbf{G}_n$. We must show $S_n(\mathbf{x}_n, \cdot)$ and $S_n(G_n\mathbf{x}_n, \cdot)$ are asymptotically independent with law $L(Q)$. As in Example 1, it suffices to examine the finite dimensional distributions of these processes. Now, if $V = V_1 \times V_2$ and $W = W_1 \times W_2$, then the covariance between $S_n(\mathbf{x}_n, V)$ and $S_n(G_n\mathbf{x}_n, W)$ is

$$\begin{gathered}\operatorname{Cov}\big[1(X_{n,1} \in V) - \mu(V_1)1(X_{n,1} \in S_1 \times V_2),\\ 1(G_{n,1}X_{n,1} \in W) - \mu(W_1)1(G_{n,1}X_{n,1} \in S_1 \times W_2)\big],\end{gathered}$$

where $G_{n,1}$ is an independent uniform element from $\mathbf{G}_1$ and $\mu$ is the uniform measure on $S_1$. The independence of the events $\{X_{n,1} \in V_1 \times R\}$ and $\{G_{n,1}X_{n,1} \in W_1 \times R\}$ and the equivalency of the events $\{G_{n,1}X_{n,1} \in S_1 \times W_2\}$ and $\{X_{n,1} \in S_1 \times W_2\}$ shows this covariance is 0. It remains to show that any finite linear combination of elements $S_n(\mathbf{x}_n, V_i)$ or $S_n(\mathbf{x}_n, W_i)$ is asymptotically Gaussian. But, $G_n\mathbf{x}_n = \mathbf{y}_n$ is a vector $(Y_{n,1}, \dots, Y_{n,n})$ of i.i.d. variables with $X_{n,i}$ independent of $Y_{n,i}$ if $i$ is different from $j$. Hence, the Lindeberg C.L.T. is directly applicable, yielding the result. In the case $Q_n$ is not in $\Omega_0$ with $d_{\mathbf{V}}(Q_n, Q_\infty) \to 0$, we need to verify Condition E when $P^{(n)} = Q_n^n$ and $Q = \tau Q_\infty$. The above method extends to this case. Alternatively, let $P_n = \tau Q_n$ and $P_0 = Q$ so that $d_{\mathbf{V}}(P_n, P_0) \to 0$. Observe that the behavior of the critical value from the randomization test (under $Q_n^n$) as obtained from $J_n(\cdot, P_0^n|\mathbf{G}_n\mathbf{x}_n)$ only depends on the distribution of $\mathbf{G}_n\mathbf{x}_n$. But $\mathbf{G}_n\mathbf{x}_n$ has the same distribution whether $\mathbf{x}_n$ has distribution $Q_n$ or $P_n$. Thus, the above analysis is applicable, and Condition E is verified. □

PROOF OF PROPOSITION 3.3. The proof is completely analogous to the proofs of Propositions 3.1 and 3.2. Indeed, the differentiability condition holds with no error term at all and the pertinent covariance calculation is identically 0. See Romano (1987) for details. □

PROOF OF PROPOSITION 3.4. Let $P_n$ be any sequence of probabilities on $S$ satisfying $d_{\mathbf{V}}(P_n, P_0) \to 0$ as $n \to \infty$. Define a process $S_n(\mathbf{x}_n, \cdot): \mathbf{V} \to \mathbf{R}^k$ by $S_n(\mathbf{x}_n, V)$ to be the vector with $i$th component

$$S_{n,i}(\mathbf{x}_n, V) = n^{1/2} c_{n,i}\left[\hat{P}_n(V) - \hat{P}_{n_i,i}(V)\right],$$

where $\mathbf{x}_n$ is an observation in $S^n$, $\hat{P}_n$ is the empirical of all $n$ observations and $\hat{P}_{n,i}$ is the empirical of those $n_i$ observations corresponding to the $i$th sample. Regard $S_n$ as a random variable on $L_\infty^k(\mathbf{V})$, the metric space of bounded $\mathbf{R}^k$-valued functions on $\mathbf{V}$ with metric $\rho$ given by

$$\rho(S, S^*) = \max_{1 \le j \le k} \sup_{V \in \mathbf{V}} |S_j(V) - S_j^*(V)|.$$

Let $L_n(P^{(n)})$ be the distribution of $S_n(\mathbf{x}_n, \cdot)$ when $\mathbf{x}_n$ has distribution $P^{(n)}$. From Romano (1988), $L_n(P^n)$ converges in $L_\infty^k(\mathbf{V})$ to a mean 0 Gaussian limit law $L(P_0)$ [whose paths lie in a separable subset of $L_\infty^k(\mathbf{V})$]. By the continuous mapping theorem, $J_n(P_n, \dots, P_n)$ converges weakly to a limit law $J(P_0, \dots, P_0)$ if $d_{\mathbf{V}}(P_n, P_0) \to 0$. From Romano (1988), this limit law is continuous and strictly increasing.

Let $\hat{P}_n$ be the empirical measure of $n$ observations from $P_0$. It follows by the generalized Glivenko–Cantelli theorem that $d_{\mathbf{V}}(\hat{P}_n, P_0) \to 0$ a.s. as $n \to \infty$. Hence, if $b_n(\alpha, P_1, \dots, P_k)$ represents an upper $\alpha$ quantile of $J_n(P_1, \dots, P_k)$, then $b_n(\alpha, \hat{P}_n, \dots, \hat{P}_n)$ tends to a finite limit $b(\alpha, P_0)$ in probability. Consistency of the test now follows easily, for suppose the actual distributions of the $k$ populations are $P_1, \dots, P_k$. Clearly, the observed test statistic $T_n \to \infty$ in probability. On the other hand, the bootstrap critical value tends to a finite limit $b(\alpha, P_0)$ in probability, where $P_0 = \sum_i \lambda_i P_i$.

To see that the randomization null distribution behaves in the same way as the bootstrap distribution, we must verify Condition D. Now, assume the conditions on the distribution $Q^{(n)}$ of $\mathbf{x}_n$ as given in the statement of the proposition. Let $G_n$ and $G_n'$ be independent of $\mathbf{x}_n$ and i.i.d. uniform random transformations from $\mathbf{G}_n$. By analogy with Condition E, we must show the processes $S_n(G_n\mathbf{x}_n, \cdot)$ and $S_n(G_n'\mathbf{x}_n, \cdot)$ are asymptotically independent. As before, it suffices to examine the finite dimensional distributions of these processes. But, the linearity of these processes easily implies a limiting Gaussian distribution for linear combinations of elements $S_n(G_n\mathbf{x}_n, V_i)$ and $S_n(G_n'\mathbf{x}_n, W_i)$. Moreover, a covariance calculation (easily obtained by conditioning on $G_n$ and $G_n'$) shows the covariance between $S_n(G_n\mathbf{x}_n, V)$ and $S_n(G_n'\mathbf{x}_n, W)$ to be of order $O(n^{-1})$, and the result follows. □

## REFERENCES


BERAN, R. (1986). Simulated power functions. *Ann. Statist.* **14** 151–173.

BERAN, R. and MILLAR, P. W. (1987). Stochastic estimation and testing. *Ann. Statist.* **15** 1131–1154.

BICKEL, P. J. (1969). A distribution free version of the Smirnov two sample test in the $p$-variate case. *Ann. Math. Statist.* **40** 1–23.

BLUM, J. R., KIEFER, J. and ROSENBLATT, M. (1961). Distribution free tests of independence based on the sample distribution function. *Ann. Math. Statist.* **32** 481–498.

EFRON, B. (1979). Bootstrap methods: Another look at the jackknife. *Ann. Statist.* **7** 1–26.

FISHER, R. A. (1935). *The Design of Experiments*. Oliver and Boyd, Edinburgh.

HETTMANSPERGER, T. (1984). *Statistical Inference Based on Ranks*. Wiley, New York.

HOEFFDING, W. (1952). The large-sample power of tests based on permutations of observations. *Ann. Math. Statist.* **23** 169–192.

JAMES, B., JAMES, K. L. and SIEGMUND, D. (1987). Tests for a change-point. *Biometrika* **74** 71–83.

KIEFER, J. (1959). *K*-sample analogues of the Kolmogorov–Smirnov and Cramér–v. Mises tests. *Ann. Math. Statist.* **30** 420–447.

LE CAM, L. (1983). A remark on empirical measures. In *A Festschrift for Erich L. Lehmann* (P. J. Bickel, K. A. Doksum and J. L. Hodges, Jr., eds.) 305–327. Wadsworth, Belmont, Calif.

PITMAN, E. J. G. (1937 / 38). Significance tests which may be applied to samples from any population. *J. Roy. Statist. Soc. Suppl.* 4 119–130, 225–232; *Biometrika* **29** 322–335.

POLLARD, D. (1984). *Convergence of Stochastic Processes*. Springer, New York.

ROMANO, J. (1987). Bootstrap and randomization tests of some nonparametric hypotheses. Technical Report No. 270, Dept. Statistics, Stanford Univ.

ROMANO, J. (1988). A bootstrap revival of some nonparametric distance tests. *J. Amer. Statist. Assoc.* **83** 698–708.

SHORACK, G. and WELLNER, J. (1986). *Empirical Processes with Applications to Statistics*. Wiley, New York.

SMIRNOV, N. (1939). On the estimation of the discrepancy between empirical curves of distribution for two independent samples. *Bull. Math. Univ. Moscow.* **2** 3–16.

VADIVELOO, J. (1983). On the theory of modified randomization tests for nonparametric hypotheses. *Comm. Statist. A—Theory Methods* **12** 1581–1596.

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA 94305