The Annals of Statistics
1986, Vol. 14, No. 2, 431–443

CONFIDENCE SETS FOR A MULTIVARIATE DISTRIBUTION[1]

BY R. BERAN AND P. W. MILLAR

University of California, Berkeley

The confidence sets for a q-dimensional distribution studied in this paper have several attractive features: affine invariance, correct asymptotic level whatever the actual distribution may be, numerical feasibility, and a local asymptotic minimax optimality property. When dimension q equals one, the confidence sets reduce to the usual Kolmogorov–Smirnov confidence bands, except that critical values are determined by bootstrapping.

1. Introduction. Kolmogorov–Smirnov confidence bands for a one-dimensional cdf have two attractive features: affine invariance and distribution-free critical values over the class of all continuous cdf's. Neither property is retained by the analogous confidence sets for a multivariate distribution based on the q-dimensional Kolmogorov–Smirnov statistic ($q \geq 2$). Studied in this paper is an alternative multivariate version of the one-dimensional Kolmogorov–Smirnov confidence band which preserves affine invariance, has correct asymptotic level, and makes equally good sense whether the actual distribution of the data is discrete, possesses a Lebesgue density, or is singular with respect to Lebesgue measure.

1.1. Half-spaces and confidence sets. Let $|\cdot|$ and $\langle \cdot, \cdot \rangle$ denote, respectively, euclidean norm and inner product in R^q. Let $S_q = \{s \in R^q: |s| = 1\}$ be the unit sphere in R^q. For every $(s, t) \in S_q \times R$, let $A(s, t)$ be the half-space

$$A(s, t) = \{x \in R^q: \langle s, x \rangle \leq t\}. \tag{1.1}$$

Let $\mathscr{P}$ be the set of all probability measures defined on the Borel sets of R^q. The half-spaces $\mathscr{A} = \{A(s, t): (s, t) \in S_q \times R\}$ separate probabilities in the sense that, if $P, Q \in \mathscr{P}$ and $P(A) = Q(A)$ for every $A \in \mathscr{A}$, then P, Q agree on all Borel sets (Cramér and Wold, 1936). The class $\mathscr{A}$ is a Vapnik–Červonenkis class of index $q + 1$ (e.g., Dudley, 1978) and is invariant under affine transformation of R^q.

Consider the distance between $P, Q \in \mathscr{P}$ defined by

$$d(P, Q) = \sup\{|P(A(s, t)) - Q(A(s, t))|: (s, t) \in S_q \times R\}. \tag{1.2}$$

Introduced into statistics by Wolfowitz (1954), the half-space distance d has reemerged in recent discussions of projection pursuit (Diaconis and Freedman, 1984; Huber, 1985). Let $\hat{P}_n$ be the empirical distribution of $x_1, x_2, \ldots, x_n$, i.i.d.

Received June 1984; revised June 1985.
[1]Research supported by National Science Foundation Grant MCS84-03239.
AMS 1980 *subject classifications*. 62G05, 62H12.
Key words and phrases. Confidence set, multivariate distribution, affine invariance, local asymptotic minimax, bootstrap.

R^q-valued random vectors with unknown distribution $P \in \mathscr{P}$. The proposed confidence set for P has the form $\{Q \in \mathscr{P}\colon d(\hat{P}_n, Q) \le c\}$.

How is the critical value c to be chosen so that the confidence set has level $1 - \alpha$? Bounds for $P^n[d(\hat{P}_n, P) > c]$ obtained by Vapnik and Červonenkis (1971), Devroye (1982), Alexander (1984) appear far too conservative to yield accurate values of c. Numerical evidence for this assertion is presented in Section 3; see also Huber (1985). Direct asymptotic approximations appear no more useful, though for different reasons. Let L_∞ be the set of all bounded measurable functions on $S_q \times R$, metrized by the supremum norm $\|\cdot\|$. The σ-algebra in L_∞ is that generated by open balls. Let $W = \{W(s,t)\colon (s,t) \in S_q \times R\}$ be a Gaussian process with mean zero covariance function

$$E[W(s,t)W(s',t')] = P[A(s,t) \cap A(s',t')] - P[A(s,t)]P[A(s',t')],$$

and sample paths in L_∞. Then

$$\mathscr{L}\left[n^{1/2}d(\hat{P}_n, P)|P^n\right] \Rightarrow \mathscr{L}(\|W\|) \tag{1.3}$$

(cf. Dudley, 1978). In general, the cdf of this limit law depends upon the unknown distribution P and is not tractable.

1.2. Bootstrap confidence sets. A bootstrap construction (cf. Efron, 1979) for the critical value c avoids some of the difficulties. Let $t_n(\alpha, P)$ denote an upper α-point of $\mathscr{L}[n^{1/2}d(\hat{P}_n, P)|P^n]$. Define the confidence set

$$C_n(\alpha, \hat{P}_n) = \left\{Q \in \mathscr{P}\colon n^{1/2}d(\hat{P}_n, Q) \le t_n(\alpha, \hat{P}_n)\right\}. \tag{1.4}$$

(For a more precise description of $t_n(\alpha, \hat{P}_n)$, see Beran, 1984.) A *triangular array* version of weak convergence (1.3), derivable from a result in Le Cam (1983), implies that $\lim_{n\to\infty} P^n[C_n(\alpha, \hat{P}_n) \ni P] = 1 - \alpha$; that is, the asymptotic level of confidence set $C_n(\alpha, \hat{P}_n)$ is $1 - \alpha$. Theorem 2 in Section 2 gives a stronger version of this result.

Definition (1.2) leads to an exact algorithm for computing $d(P, Q)$ when both P and Q are supported on a finite set of cardinality m; the number of mathematical operations required is of order $2^q\binom{m}{q}$. When dimension q is small, this algorithm can be used to compute Monte Carlo approximations to $t_n(\alpha, \hat{P}_n)$ and to determine whether a given distribution Q lies in confidence set $C_n(\alpha, \hat{P}_n)$. In the latter application, Q is first replaced by a discrete approximation.

When dimension q is larger, confidence sets for P based upon a stochastic approximation to d become attractive for computational reasons. Let $s_1, s_2, \ldots, s_{k_n}$ be i.i.d. random unit vectors, uniformly distributed on S_q and independent of the $\{x_i;\ 1 \le i \le n\}$. For $P, Q \in \mathscr{P}$, define

$$d_n(P, Q) = \max_{1 \le k \le k_n} \sup_{t \in R} \{|P(A(s_k, t)) - Q(A(s_k, t))|\}. \tag{1.5}$$

Random selection of the $\{s_k;\ 1 \le k \le k_n\}$ has some advantages over systematic selection. The condition $\lim_{n\to\infty} k_n = \infty$ ensures that $\lim_{n\to\infty} d_n(P, Q) = d(P, Q)$ w.p. 1, the rate of convergence in probability of $d_n(P, Q)$ to $d(P, Q)$

being exponential in k_n. More precisely, let μ denote the uniform distribution on S_q and let $Y(s) = \sup_t |P[A(s,t)] - Q[A(s,t)]|$. Then, for every positive ε,

$$\mu\{d_n(P,Q) > (1-\varepsilon)d(P,Q)\} = 1 - [1 - b(\varepsilon)]^{k_n}, \tag{1.6}$$

where $b(\varepsilon) = \mu[Y(s_1) > (1-\varepsilon)d(P,Q)]$.

Let $\mathbf{s}_n = (s_1, s_2, \ldots, s_{k_n})$ and let $u_n(\alpha, P, \mathbf{s}_n)$ denote an upper α-point of the conditional distribution of $n^{1/2}d_n(\hat{P}_n, P)$ under P^n, given $\mathbf{s}_n$. Consider the modified bootstrap confidence set

$$\tilde{C}_n(\alpha, \hat{P}_n) = \{Q \in P: n^{1/2}d_n(\hat{P}_n, Q) \le u_n(\alpha, \hat{P}_n, \mathbf{s}_n)\}. \tag{1.7}$$

The distribution of $\tilde{C}_n(\alpha, \hat{P}_n)$ depends upon the joint distribution of the observations $\{x_i;\ 1 \le i \le n\}$ and of the search sample $\mathbf{s}_n$. If $\lim_{n\to\infty} k_n = \infty$, the asymptotic level of $\tilde{C}_n(\alpha, \hat{P}_n)$ is $1 - \alpha$ (Theorem 3 in Section 2). The number of mathematical operations required to compute $d_n(P,Q)$ depends linearly upon the product $k_n q$.

1.3. Confidence sets and risk. In general, the problem of confidence set construction has a natural decision theoretic formulation, which views it as a set-valued estimation problem, subject to the level constraint (see Beran and Millar, 1985). For confidence set $C_n(\alpha, \hat{P}_n)$, the decision space treated in this formulation is the collection of all balls $C(z,r) = \{Q \in \mathscr{P}: d(Q,z) \le r\}$ with center $z \in \mathscr{P}$ and radius r. If Z_n, R_n are estimates of center and radius based upon the n observations, then the loss function for the confidence set $C(Z_n, R_n)$ is taken to be

$$l_n(Z_n, R_n; P) = n^{1/2} \sup_{Q \in C(Z_n, R_n)} d(Q,P) \tag{1.8}$$

or a monotone function thereof. Evidently, this loss penalizes for excessive size or miscentering of $C(Z_n, R_n)$. The risk of $C(Z_n, R_n)$ is then

$$\rho_n(Z_n, R_n; P) = \int l_n(Z_n, R_n; P)\, dP^n. \tag{1.9}$$

Among all confidence sets of the form just described whose asymptotic level is at least $1 - \alpha$, the confidence set $C_n(\alpha, \hat{P}_n)$, defined in (1.4), is locally asymptotically minimax (Theorems 1 and 2 in Section 2). Moreover, the risk (1.9) of $C_n(\alpha, \hat{P}_n)$ can itself be estimated from the data.

2. Asymptotic properties of confidence sets. This section formulates and proves the three theorems described in the introduction. Notation introduced there is retained.

2.1. Local asymptotic minimax bound for confidence sets. Any probability $P \in \mathscr{P}$ can be regarded as an element of L_∞ by identifying P with the function which maps $(s,t) \in S_q \times R$ into $P(A(s,t))$. With this identification, $d(P,Q) = \|P - Q\|$ for every $P, Q \in \mathscr{P}$, where $\|\cdot\|$ is supremum norm on $S_q \times R$.

Let $\mathscr{A}(n, \alpha)$ denote the collection of all confidence sets $C(Z_n, R_n)$, Z_n and R_n depending on $x_1, x_2, \ldots, x_n$, which satisfy the criterion of asymptotic level $1 - \alpha$:

$$\liminf_{n\to\infty} P^n\big[C(Z_n, R_n) \ni P\big] \geq 1 - \alpha \quad \text{for every } P \in \mathscr{P}. \tag{2.1}$$

Define a norm $|\cdot|_\delta$ by $|P - Q|_\delta = \sup\{|P(A \cap B) - Q(A \cap B)|\colon A, B \in \mathscr{V}\}$ for $P, Q \in \mathscr{P}$. Let $\mathscr{F}(n, c, P)$ be the set of all probabilities $Q \in \mathscr{P}$ such that $|Q - P|_\delta \leq ca_n$, where $0 < c < \infty$ and a_n is a preselected sequence of real numbers subject to the constraints $a_n \geq n^{-1/2}$, $a_n \downarrow 0$.

Let $\mathscr{Q}_0$ be the distribution of the Gaussian process W described in Section 1, viewed as a random element of L_∞. Let ρ_n be the risk of confidence set $C(Z_n, R_n)$ as defined in (1.9).

THEOREM 1. *Fix $\alpha \in (0,1)$ and $P \in \mathscr{P}$. Then*

(2.2)

$$\lim_{c\to\infty} \liminf_{n\to\infty} \inf_{C(Z_n, R_n)\in\mathscr{A}(n,\alpha)} \sup_{Q\in\mathscr{F}(n,c,P)} \rho_n(Z_n, R_n; Q) \geq \int_{L_\infty} (\|z\| + r)\mathscr{Q}_0(dz),$$

where the constant r is determined by

$$\mathscr{Q}_0\big[\|z\| \leq r\big] = 1 - \alpha. \tag{2.3}$$

This theorem rests, in part, upon the abstract Hájek–Le Cam lower bound for minimax risk of a decision procedure. Theorem 1 remains valid for various restrictions of $\mathscr{P}$: for instance, if $\mathscr{P}$ is replaced by the set of all probabilities supported on the unit sphere in R^q; or if $\mathscr{P}$ is replaced by the set of all probabilities supported on a finite subset of R^q. Section 3 discusses examples of these two situations.

PROOF. Theorem 1 will be deduced from (4.5) of Beran and Millar (1985), cited hereafter as BM. We also draw on results in Millar (1983)—these to be referenced by notations of the form X.2 (Chapter X, Section 2). Let f be the density of P with respect to some σ finite measure ν. Let H be the Hilbert space of real functions h on R^q such that $\int hf\,d\nu = 0$, $\int h^2 f\,d\nu < \infty$, support $h \subset$ support f. Let H_0 be the subset of H consisting of all h such that $f(1 + n^{-1/2}h)$ is a probability density for all sufficiently large n. Then H_0 is dense in H. Define τ, a map from H to L_∞, by $(\tau h)(A) = \int_A hf\,d\nu$, $A \in \mathscr{V}$. Let B be the closure of τH in the supremum norm $\|\cdot\|$, so $B \subset L_\infty$. Calculations as in V.2 show that (τ, H, B) is an abstract Wiener space; its canonical Gaussian measure on L_∞ is, in fact, $\mathscr{Q}_0$.

Let $\{\mathscr{Q}_h,\ h \in H\}$ be the Gaussian shift experiment (V.3) associated with (τ, H, B). If P_h^n is the n-fold product measure of $P_{n^{-1/2}h}(dx) = f(x)[1 + n^{-1/2}h(x)]\nu(dx)$ then $\{P_h^n,\ h \in H_0\}$ converges to $\{\mathscr{Q}_h,\ h \in H_0\}$ (see VI.1). Define ξ of Section 4 of BM by $\xi(P_h^n) = P_{n^{-1/2}h} \in L_\infty$. If ξ' is the linear operator on L_∞ given by $\xi' x = x$ then clearly $\xi(P_h^n) = \xi(P_0^n) + \xi'(\tau n^{-1/2}h)$, so (4.2) of BM holds. Since the set of measures over which the supremum in (2.2) above is

taken contains the measures $\{P_{n\ 1/2h},\ |h| \le c\}$, the theorem follows from the locally asymptotically minimax lower bound, Theorem 4.5, in BM. □

2.2. Confidence set $C_n(\alpha, \hat{P}_n)$ is LAM. This subsection demonstrates, in particular, that the bootstrap confidence set $C_n(\alpha, \hat{P}_n)$ defined in (1.4) has asymptotic level $1 - \alpha$, uniformly over $|\cdot|_\delta$ compacts in $\mathscr{P}$. Moreover, $C_n(\alpha, \hat{P}_n)$ is locally asymptotically minimax (LAM) in the sense that its maximum risk over $\mathscr{F}(n, c, P)$ attains the lower bound for minimax risk given in Theorem 1.

THEOREM 2. *Fix $\alpha \in (0,1)$ and $P \in \mathscr{P}$. If $\lim_{n\to\infty} |P_n - P|_\delta = 0$, then*

$$\lim_{n\to\infty} P_n^n\left[C_n(\alpha, \hat{P}_n) \ni P_n\right] = 1 - \alpha \tag{2.4}$$

and

$$t_n(\alpha, \hat{P}_n) \to r \quad \text{in } P_n^n\text{-probability}, \tag{2.5}$$

where r is defined by (2.3).

Moreover, for every positive c

$$\lim_{n\to\infty} \sup_{Q \in \mathscr{F}(n, c, P)} \rho_n\left(\hat{P}_n, n^{-1/2} t_n(\alpha, \hat{P}_n); Q\right) = \int (\|z\| + r)\mathscr{Q}_0(dz). \tag{2.6}$$

PROOF. Suppose $\{P_n \in \mathscr{P};\ n \ge 1\}$ satisfies the hypothesis of the theorem. From Proposition 1 of Section 4,

$$\mathscr{L}\left[n^{1/2}\|\hat{P}_n - P_n\| | P_n^n\right] \Rightarrow \mathscr{L}(\|W\|). \tag{2.7}$$

Since the limit law has a continuous, strictly monotone cdf (Proposition 2 of Section 4), it follows that

$$\lim_{n\to\infty} t_n(\alpha, P_n) = r. \tag{2.8}$$

The Vapnik and Červonenkis (1971) inequality implies, in particular, that $|\hat{P}_n - P|_\delta \to 0$ in P_n^n-probability. This convergence and (2.8) yield (2.5). In turn, (2.5) and (2.7) imply (2.4) (cf. Beran, 1984).

By (2.7) and the exponential bound of Alexander (1984), $n^{1/2}E_{P_n^n}\|\hat{P}_n - P_n\|$ converges to $E\|W\|$. Since $\{P_n\}$ could be chosen to be an arbitrary sequence in $\mathscr{F}(n, c, P)$, a straightforward argument yields (2.6), as follows:

$$\begin{aligned}
&\lim_{n\to\infty} \sup_{Q \in \mathscr{F}(n, c, P)} \rho_n\left(\hat{P}_n, n^{-1/2} t_n(\alpha, \hat{P}_n); Q\right) \\
&\quad = \lim_{n\to\infty} \rho_n\left(\hat{P}_n, n^{-1/2} t_n(\alpha, \hat{P}_n); P_n\right) \\
&\quad = \lim_{n\to\infty} \left\{n^{1/2}E_{P_n^n}\|\hat{P}_n - P_n\| + E_{P_n^n} t_n(\alpha, \hat{P}_n)\right\} \\
&\quad = E\|W\| + r.
\end{aligned} \tag{2.9}$$

□

2.4. Confidence set $\tilde{C}_n(\alpha, \hat{P}_n)$ has asymptotic level $1 - \alpha$. This subsection establishes that the computationally simpler confidence set $\tilde{C}_n(\alpha, \hat{P}_n)$ defined in (1.7) has asymptotic level $1 - \alpha$ and that its critical value $u_n(\alpha, \hat{P}_n, \mathbf{s}_n)$ converges

in probability to r. Both convergences are uniform over $|\cdot|_\delta$ compacts in $\mathscr{P}$. Let μ denote the uniform distribution on S_q and let $P^n \times \mu^{k_n}$ designate the product measure generated by P^n and μ^{k_n}.

THEOREM 3. *Fix* $\alpha \in (0, 1)$ *and* $P \in \mathscr{P}$. *If* $\lim_n |P_n - P|_\delta = 0$ *and* $\lim_{n\to\infty} k_n = \infty$, *then*

$$\lim_{n\to\infty} \left(P_n^n \times \mu^{k_n}\right)\left[\tilde{C}_n(\alpha, \hat{P}_n) \ni P_n\right] = 1 - \alpha \tag{2.10}$$

and

$$u_n(\alpha, \hat{P}_n, \mathbf{s}_n) \to r \quad \textit{in } \left(P_n^n \times \mu^{k_n}\right)\textit{-probability}, \tag{2.11}$$

where r *is defined by* (2.3).

PROOF. Suppose $\{P_n \in \mathscr{P};\ n \geq 1\}$ and $\{k_n;\ n \geq 1\}$ satisfy the hypothesis of the theorem. Let $W_n(s, t) = n^{1/2}\{\hat{P}_n[A(s, t)] - P_n[A(s, t)]\}$. Since $W_n \Rightarrow W$ as random elements of L_∞ (Proposition 1 of Section 4), there exist versions of $\{W_n\}$, W such that $\lim_{n\to\infty}\|W_n - W\| = 0$ for every realization (Wichura, 1970).

Fix the realization of $\{W_n\}$ and W. Let $Z_n(s) = \sup_t|W_n(s, t)|$ and $Z(s) = \sup_t|W(s, t)|$. From above, $\limsup_{n\to\infty}\{|Z_n(s) - Z(s)|:\ s \in S_q\} = 0$. From this and the evident convergence w.p. $1(\mu)$ of $\max\{Z(s_k): 1 \leq k \leq k_n\}$ to $\operatorname{ess\,sup}_\mu Z(s)$, it follows that

$$\lim_{n\to\infty} \max_{1\leq k\leq k_n} Z_n(s_k) = \operatorname*{ess\,sup}_\mu Z(s) \quad \text{w.p. } 1(\mu). \tag{2.12}$$

Let $\mathscr{L}[\max_{1\leq k\leq k_n}\sup_t|W_n(s_k, t)||\mathbf{s}_n, P_n^n]$ denote the conditional distribution of the first argument, given $\mathbf{s}_n$. In view of (2.12),

$$\begin{aligned}&\mathscr{L}\left[\max_{1\leq k\leq k_n}\sup_t|W_n(s_k, t)||\mathbf{s}_n, P_n^n\right]\\ &\quad\Rightarrow \mathscr{L}\left[\operatorname*{ess\,sup}_\mu \sup_t|W(s, t)|\right] \quad \text{w.p. } 1(\mu)\end{aligned} \tag{2.13}$$

for the original versions of $\{W_n\}$ and W.

Let $\{t_k;\ k \geq 1\}$ be a fixed countable dense subset of R. With probability $1(\mu)$, the random set $\{s_k;\ k \geq 1\}$ is a dense subset of S^q. For every n,

$$\begin{aligned}\operatorname*{ess\,sup}_\mu \sup_t|W_n(s, t)| &= \sup_{k\geq 1}\sup_t|W_n(s_k, t)|\\ &= \sup_{k\geq 1}|W_n(s_k, t_k)| \quad \text{w.p. } 1(\mu)\\ &= \|W_n\|.\end{aligned} \tag{2.14}$$

The final equality holds because $\|W_n\|$ equals the supremum of $W_n(s, t)$ over any countable dense subset of $S_q \times R$. Letting n tend to infinity in (2.14) proves

$$\mathscr{L}\left[\operatorname*{ess\,sup}_\mu \sup_t|W(s, t)|\right] = \mathscr{L}[\|W\|]. \tag{2.15}$$

Combining (2.13) with (2.15) yields the convergences:

$$u_n(\alpha, P_n, \mathbf{s}_n) \to r \text{ w.p. } 1(\mu)$$

$$\text{and } \mathscr{L}\left[\max_{1 \le k \le k_n} \sup_t |W_n(s_k, t)| \,|P_n^n \times \mu^{k_n}\right] \Rightarrow \mathscr{L}[\|W\|].$$

Theorem 3 follows. □

2.4. Extensions. The asymptotic theory of Theorems 1 to 3 can be extended in several directions.

Estimation of risk. The risk $\rho_n(\hat{P}_n, n^{-1/2}t_n(\alpha, \hat{P}_n); P)$ of confidence set $C_n(\alpha, \hat{P}_n)$ has the bootstrap estimate $\rho_n(\hat{P}_n, n^{-1/2}t_n(\alpha, \hat{P}_n); \hat{P}_n)$. Convergence in probability of this estimate to the actual risk can be proved using the triangular array reasoning of Theorem 2. A simpler risk estimate, which relies on the asymptotic constancy (2.5) of $t_n(\alpha, \hat{P}_n)$ and on formula (2.9), is the mean of the bootstrap distribution for $n^{1/2}d(\hat{P}_n, P)$ plus $t_n(\alpha, \hat{P}_n)$. This in turn can be approximated by the mean of the bootstrap distribution for $n^{1/2}d_n(\hat{P}_n, P)$ plus $u_n(\alpha, \hat{P}_n, \mathbf{s}_n)$; see the proof of Theorem 3. In principle, risk estimates provide a means for directly comparing confidence set $C_n(\alpha, \hat{P}_n)$ with other confidence sets of the same asymptotic level.

Other roots for confidence sets. Alternative confidence sets for P can be obtained from the weighted metric

$$\sup\left\{\frac{|\hat{P}_n[A(s,t)] - P[A(s,t)]|}{[P[A(s,t)](1 - P[A(s,t)])]^{1/2}} : (s,t) \in S_q \times R\right\} \tag{2.16}$$

of Anderson–Darling type. The supremum norm in (2.16) or in the definition of $d(\hat{P}_n, P)$ may be replaced by other norms, such as the $L_p(m)$ norm on $S_q \times R$, m being a finite measure. The asymptotic theory for confidence sets $C_n(\alpha, \hat{P}_n)$ and $\tilde{C}_n(\alpha, \hat{P}_n)$ can be extended to the analogous confidence sets based on these alternate roots. $L_p(m)$ norms with σ finite m can also be treated, at the price of reducing the class of possible distributions.

Confidence sets for the difference of two distributions. Let P, Q be probabilities in P. Suppose $x_1, x_2, \ldots, x_{n_1}$ are i.i.d. (P) and $y_1, y_2, \ldots, y_{n_2}$ are i.i.d. (Q), the two samples being independent. Let $\hat{P}_n$ and $\hat{Q}_n$ be the empirical distributions of the $\{x_i: 1 \le i \le n_1\}$ and $\{y_j: 1 \le j \le n_2\}$, respectively, where $n = n_1 + n_2$ and $n_1/n \to \lambda$ as $n \to \infty$, $0 < \lambda < 1$. Let $t_n(\alpha, P, Q)$ denote an upper α-point of $\mathscr{L}[n^{1/2}\|(\hat{P}_n - \hat{Q}_n) - (P - Q)\| | P^{n_1} \times Q^{n_2}]$. Define the confidence set

$$C_n(\alpha, \hat{P}_n, \hat{Q}_n) = \{P - Q: \|(P - Q) - (\hat{P}_n - \hat{Q}_n)\| \le n^{-1/2}t_n(\alpha, \hat{P}_n, \hat{Q}_n);\ P, Q \in \mathscr{P}\}. \tag{2.17}$$

The development of this paper, including correct asymptotic level, LAM

optimality, and computationally feasible variants can be carried through for this confidence set.

3. Numerical study. Further insight into the applicability and performance of confidence sets $C_n(\alpha, \hat{P}_n)$ and $\tilde{C}_n(\alpha, \hat{P}_n)$ is gained from three numerical case studies.

3.1. First case study: five dimensional data. Mardia, Kent, and Bibby (1979) reported test scores for $n = 88$ college students, each of whom took two closed book and three open book tests. Marginally, the scores for each test appear to be normally distributed (perhaps the consequence of a grading curve). Is it reasonable to approximate the joint distribution of the five test scores by a normal distribution?

The Monte Carlo approximation to the conditional bootstrap distribution for $d_n(\hat{P}_n, P)$, given $\mathbf{s}_n$, recorded in Table 1, was calculated from 200 bootstrap samples, with d_n defined by $k_n = 200$ randomly generated unit vectors in R^5. Two points are noteworthy: (a) The bootstrap distribution of $d_n(\hat{P}_n, P)$ in this example is necessarily supported on $\{j/88\colon 0 \le j \le 88\}$. The Monte Carlo approximation in Table 1 is supported on $\{j/88\colon 8 \le j \le 22\}$, the cdf having sizable jumps. Convergence of the conditional bootstrap distribution to its continuous limit (Theorem 3) may not be swift. (b) The standard inequalities for $P^n[d(\hat{P}_n, P) > c]$ are extremely conservative here. Both the Vapnik–Červonenkis (1971) and Devroye (1982) inequalities yield the trivial conclusion $P^n[d(\hat{P}_n, P) > 0.250] \le 1$, in contrast to Table 1. Alexander's (1984, Theorem 2.11) inequality is not even applicable for $c < 8(88)^{-1/2} = 0.853$.

From Table 1, the confidence set

$$\tilde{C}_n(0.930, \hat{P}_n) = \{Q \in \mathscr{P}\colon d_n(Q, \hat{P}_n) \le 0.193\}, \tag{3.1}$$

with $n = 88$, has approximate level 0.93. Let $\hat{N}_n$ denote the normal distribution on R^5 whose mean and covariance matrix are given by the sample mean and sample covariance matrix of the 88 test score vectors. For the d_n of the previous paragraph, $d_n(\hat{P}_n, \hat{N}_n) = 0.129$. The multivariate normal model for the test score

TABLE 1

Monte Carlo approximation to the bootstrap distribution of $d_n(\hat{P}_n, P)$, given $\mathbf{s}_n$, for the five-dimensional test score data. The number of random directions used is 200; the number of bootstrap samples drawn is 200.

x	**0.091**	**0.102**	**0.114**	**0.125**	**0.136**	**0.148**	**0.159**	**0.170**
Bootstrap cdf at x	0.010	0.025	0.080	0.185	0.345	0.525	0.685	0.805
x	**0.182**	**0.193**	**0.205**	**0.216**	**0.227**	**0.239**	**0.250**	
Bootstrap cdf at x	0.885	0.930	0.970	0.980	0.985	0.995	1.000	

vectors appears satisfactory, in the sense that the trustworthy set estimate $\tilde{C}_n(0.930, \hat{P}_n)$ for P contains at least one normal distribution, namely $\hat{N}_n$. (The reasoning here is not that of a classical goodness-of-fit test.)

Computations for this example, in FORTRAN 77, using Kolmogorov–Smirnov subroutines and pseudorandom numbers provided by IMSL, took approximately one hour on a VAX 11/750.

3.2. *Second case study: directional data.* Steinmetz (1962) reported $n = 20$ cross-bed measurements of azimuth and dip from sandstone bodies in the Eocene Cathedral Bluffs member of the Wasatch formation in Wyoming. These directional measurements can be represented as unit vectors in R^3 or as points on the surface of a unit sphere. Is it reasonable to regard the data as a sample from a Fisher distribution? A Schmidt-net plot (Mardia 1972, page 218) indicates that the observations form a sausage-like cluster on the surface of the unit sphere, a configuration suspiciously inconsistent with the axial symmetry of the Fisher density.

Let $\mathscr{P}_s$ denote the set of all distributions which are supported on S_3, the unit sphere in R^3. Suppose the 20 observations form a random sample from an unknown distribution $P \in \mathscr{P}_s$. The asymptotic theory of Section 2 remains valid if $\mathscr{P}$ is replaced by $\mathscr{P}_s$. Intersection of S_3 with all half-spaces of R^3 yields the collection of spherical caps on S_3. The Monte Carlo approximation to the bootstrap distribution for $d_n(\hat{P}_n, P)$, given $\mathbf{s}_n$, recorded in Table 2, was calculated from 200 bootstrap samples, with d_n defined by $k_n = 200$ randomly generated unit vectors in R^3. In contrast, the Vapnik–Červonenkis and Devroye inequalities yield the trivial bound $P^n[d(\hat{P}_n, P) > 0.45] \leq 1$, while Alexander's inequality is not applicable at values in the support $\{j/20\colon 3 \leq j \leq 9\}$ of the distribution in Table 2.

The confidence set

$$\tilde{C}_n(0.985, \hat{P}_n) = \{Q \in \mathscr{P}_s\colon d_n(Q, \hat{P}_n) \leq 0.35\}, \tag{3.2}$$

with $n = 20$, has approximate level 0.985. If $\hat{F}_n$ denotes the Fisher distribution fitted to the sample by maximum likelihood, then $d_n(\hat{P}_n, \hat{F}_n) \simeq 0.43$. This distance was approximated by first replacing $\hat{F}_n$ with the empirical distribution of a Monte Carlo sample of size 1000 drawn from $\hat{F}_n$. While $\hat{F}_n$ does not lie in $\tilde{C}_n(0.985, \hat{P}_n)$, there might be some other Fisher distribution which does. Settling

TABLE 2

Monte Carlo approximation to the bootstrap distribution of $d_n(\hat{P}_n, P)$, given $\mathbf{s}_n$, for the directional data. The number of random directions used is 200; the number of bootstrap samples drawn is 200.

x	0.15	0.20	0.25	0.30	0.35	0.40	0.45
Bootstrap cdf at x	0.100	0.335	0.685	0.890	0.985	0.995	1.000

this question conclusively—by computing the minimum value of $d_n(\hat{P}_n, F)$ as F ranges over all Fisher distributions—appears very time-consuming.

3.3. Simulation study: categorical data. A sample $z = (z_1, z_2, \ldots, z_q)$ from the multinomial $(n; p_1, p_2, \ldots, p_q)$ distribution can be thought to arise from n independent multivariate Bernoulli trials as follows. At each trial, one of q possible outcomes occurs, the probability of outcome j being p_j. The random variable z_j is the number of times outcome j occurs in the n trials.

Suppose possible outcome j is represented as the vector e_j in R^q whose jth component is 1 and whose other components are all 0. Then, the outcome of the ith multivariate Bernoulli trial is a random vector x_i whose distribution P is supported on the subset $\Gamma = \{e_j: 1 \le j \le q\}$ of R^q and is given by $P[\{e_j\}] = p_j$, $1 \le j \le q$. Moreover, $z = \Sigma_{i=1}^n x_i$; and the empirical distribution $\hat{P}_n$ of the $\{x_i: 1 \le i \le n\}$ is supported on Γ, with $\hat{P}_n[\{e_j\}] = \hat{p}_{jn} = z_j/n$, which is the relative frequency of outcome j.

In this situation, the half-space distance $d(\hat{P}_n, P)$ is equal to the variation norm distance between $\hat{P}_n$ and P. Consequently, $d(\hat{P}_n, P) = 2^{-1}\Sigma_{j=1}^q |\hat{p}_{j,n} - p_j|$. This algebraic simplification provides an opportunity to compare the actual level of confidence set $C_n(\alpha, \hat{P}_n)$, defined in (1.4), with its asymptotic level $1 - \alpha$. Table 3 reports the results of a Monte Carlo study for five-dimensional multinomial random vectors with $n = 20, 40, 80$. For each vector of outcome probabilities $\{p_j: 1 \le j \le 5\}$ considered, 1000 multinomial $(n; p_1, p_2, \ldots, p_5)$ vectors $\{(z)_m: 1 \le m \le 1000\}$ were generated. For each such sample vector $(z)_m$, the estimated outcome probabilities $\{(\hat{p}_{jn})_m: 1 \le j \le 5\}$ were calculated and 200 vectors were drawn from the multinomial $(n; (\hat{p}_{1n})_m, (\hat{p}_{2n})_m, \ldots, (\hat{p}_{5n})_m)$ distribution, in order to build up a bootstrap distribution for the L_1-distance $\Sigma_{j=1}^5 |(\hat{p}_{jn})_m - p_j|$ and so determine the associated confidence ball for the $\{p_j: 1 \le j \le 5\}$.

As might be expected, the agreement in Table 3 between nominal and actual levels of the confidence set $C_n(\alpha, \hat{P}_n)$ is best when the probabilities $\{p_j; 1 \le j \le 5\}$ are all equal. Even when one or more of the $\{p_j\}$ is very small, the convergence of actual level to nominal level as sample size n increases is evident. Equally noteworthy is the apparent tendency of the nominal level to exceed the actual

TABLE 3

Observed levels in 1000 *Monte Carlo trials of the bootstrap confidence set* $C_n(\alpha, \hat{P}_n)$. *The data is multinomial and of dimension* 5. *The number of bootstrap samples used to construct each confidence set is* 200.

Outcome probabilities $\{p_j\}$					$n = 20$			$n = 40$			$n = 80$		
					Nominal levels			Nominal levels			Nominal levels		
					0.90	0.95	0.99	0.90	0.95	0.99	0.90	0.95	0.99
0.2	0.2	0.2	0.2	0.2	0.899	0.935	0.985	0.898	0.498	0.987	0.897	0.947	0.983
0.1	0.1	0.2	0.3	0.3	0.871	0.937	0.986	0.881	0.932	0.981	0.893	0.954	0.987
0.05	0.1	0.15	0.3	0.4	0.879	0.925	0.977	0.903	0.947	0.983	0.872	0.932	0.979
0.01	0.04	0.1	0.2	0.65	0.829	0.901	0.957	0.870	0.932	0.980	0.881	0.929	0.978

level of the confidence regions (34 cases out of 36 in Table 3). It remains to be seen if bootstrap critical values can be refined to reduce this effect.

4. Empirical process results. This section collects facts about the empirical process W_n and its limit distribution which are needed for the proofs in Section 2.

4.1. Convergence of the empirical process W_n. Let $x_{1n}, x_{2n}, \ldots, x_{nn}$ be i.i.d. random vectors in R^q, each having distribution $P_n \in \mathscr{P}$. Let $\hat{P}_n$ be the empirical measure of the $\{x_{in}\colon\ 1 \le i \le n\}$ and let $W_n(s,t) = n^{1/2}\{\hat{P}_n[A(s,t)] - P_n[A(s,t)]\}$. Recall the norm $|\cdot|_\delta$ on L_∞ defined after display (2.1).

PROPOSITION 1. *Suppose* $\lim_{n\to\infty}|P_n - P|_\delta = 0$ *for some distribution* $P \in \mathscr{P}$. *The empirical processes* $\{W_n;\ n \ge 1\}$ *converge weakly, as random elements of* L_∞, *to a Gaussian process* W *on* $S_q \times R$ *having mean zero and covariance function*

$$E[W(s,t)W(s',t')] = P[A(s,t) \cap A(s',t')] - P[A(s,t)]P[A(s',t')] \tag{4.1}$$

for $(s,t),(s',t')$ *in* $S_q \times R$.

This triangular array weak convergence result may be deduced from Le Cam (1983). Le Cam's Lemma 4, together with his analysis of $M(F, \mathscr{D})$ at the bottom of p. 317, implies the equicontinuity property: For every $\varepsilon > 0$ and $\eta > 0$ there exists $\gamma > 0$ such that

$$\limsup_{n\to\infty} P_n^n\left[\sup_{G(n,\gamma)} |W_n(s,t) - W_n(s',t')| > \eta\right] < \varepsilon, \tag{4.2}$$

where

$$G(n,\gamma) = \{(s,t),(s',t') \in S_q \times R\colon\ P_n[A(s,t)\,\Delta\,A(s',t')] < \gamma\}. \tag{4.3}$$

(An elementary argument involving the definition of Le Cam's $\mathscr{D}_\alpha$ converts the supremum in Le Cam's lemma to ours.) The convergence $\lim_{n\to\infty}|P_n - P|_\delta = 0$ permits replacement of $G(n,\gamma)$ in (4.2) by

$$G(0,\gamma) = \{(s,t),(s',t') \in S_q \times R\colon\ P[A(s,t)\,\Delta\,A(s',t')] < \gamma\}. \tag{4.4}$$

This yields the classical criterion for tightness of the processes $\{W_n\colon\ n \ge 1\}$ in L_∞. The proposition follows immediately.

4.2. The maximum of a Gaussian process. The proofs of Theorems 2 and 3 required the fact that the random variable $\|W\|$ has a strictly increasing continuous cdf. This fact follows at once from the more general Proposition 2, which is useful in the analysis of other bootstrap procedures as well.

PROPOSITION 2. *Let (τ, H, B) be an abstract Wiener space with P_0 the canonical normal distribution on B. If $|\cdot|$ is the norm of B, then under P_0 the random variable $z \to |z|$, defined on B, has a density and a strictly increasing cdf on $[0, \infty)$.*

PROOF. The result will be deduced from a theorem of Tsirel'son (1975).

Let B^* be the dual of B. By the Hahn–Banach theorem, $|z| = \sup m(z)$, where the supremum is computed over $m \in B^*$, $|m| = 1$. Since B is separable, this supremum can be computed over a countable set of m. Since each random variable $z \to m(z)$ is Gaussian under P_0, we therefore may analyze $|z|$ as the maximum of a countable collection of mean zero Gaussian random variables.

According to Tsirel'son (Theorem 1), $|z|$ has a continuous distribution except possibly at the point $a_0 = \inf\{a\colon P_0\{|z| \le a\} > 0\}$. Let us show first that the distribution of $|z|$ has no atom at $a = 0$. Since $|z| = 0$ iff $z = 0$, it suffices to show that P_0 has no atom at $0 \in B$. Let $\{P_h,\ h \in H\}$ be the Gaussian shift family for (τ, H, B), so that $P_h(A) = P_0(A - \tau h)$. If P_0 had an atom at $0 \in B$, then because of the mutual absolute continuity of the P_h, P_0 would have an atom at each point $\tau h \in B$; this is an uncountable collection of atoms, which is impossible. Thus P_0 has no atom at 0 (and by the same argument, none at any other point).

On the other hand, every ball in B centred at 0 must have positive P_0 probability; this is immediate from a straightforward extension of Anderson's lemma (Anderson, 1955) to Gaussian measures on Banach space. By the previous paragraph, the cdf of $|z|$ must be continuous everywhere. Since every ball about 0 has positive P_0 probability, the mutual absolute continuity of the $\{P_h\}$ shows that every ball (center arbitrary) has positive probability. This implies that the cdf of $|z|$ is strictly increasing. The existence of the density of $|z|$ follows from another part of Tsirel'son's Theorem 1. This completes the proof. □

REMARK. Facts cited on page 854 of Tsirel'son's paper assert that the density of $|z|$ is strictly positive at every point of $(0, \infty)$.

REFERENCES

ALEXANDER, K. (1984). Probability inequalities for empirical processes and a law of the interated logarithm. *Ann. Probab.* **12** 1041–1067.

ANDERSON, T. W. (1955). The integral of a symmetric unimodal function. *Proc. Amer. Math. Soc.* **6** 1970–1976.

BERAN, R. (1984). Bootstrap methods in statistics. *Jber. Deutsch. Math.-Verein.* **86** 14–30.

BERAN, R. J. and MILLAR, P. W. (1985). Asymptotic theory of confidence sets. In *Proc. Berkeley Conf. in Honor of J. Neyman and J. Kiefer* **2** (L. Le Cam and R. Olshen, eds.) 865–887. Wadsworth, Monterey, Calif.

CRAMÉR, H. and WOLD, H. (1936). Some theorems on distribution functions. *J. London Math. Soc.* **11** 290–294.

DEVROYE, L. (1982). Bounds for the uniform deviation of empirical measures. *J. Multivariate Anal.* **12** 72–79.

DIACONIS, P. and FREEDMAN, D. (1984). Asymptotics of graphical projection pursuit. *Ann. Statist.* **12** 793–815.

DUDLEY, R. M. (1978). Central limit theorem for empirical measures. *Ann. Probab.* **6** 899–929.

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

HUBER, P. J. (1985). Projection pursuit. *Ann. Statist.* **13** 435–475.

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

MARDIA, K. V. (1972). *Statistics of Directional Data*. Academic, New York.

MARDIA, K. V., KENT, J. T. and BIBBY, J. M. (1979). *Multivariate Analysis*. Academic, New York.

MILLAR, P. W. (1983). The minimax principle in asymptotic statistical theory. *Lecture Notes in Mathematics* **976** 75–265. Springer, Berlin.

STEINMETZ, R. (1962). Analysis of vectorial data. *Jour. Sed. Petr.* **32** 801–812.

TSIREL'SON, V. S. (1975). The density of the distribution of the maximum of a Gaussian process. *Theory Probab. Appl.* **20** 847–856.

VAPNIK, V. N. and ČERVONENKIS, A. YU. (1971). On the uniform convergence of relative frequencies of events to their probabilities. *Theory Probab. Appl.* **16** 264–280.

WICHURA, M. J. (1970). On the construction of almost uniformly convergent random variables with given weakly convergent image laws. *Ann. Math. Statist.* **41** 284–291.

WOLFOWITZ, J. (1954). Generalization of the theorem of Glivenko–Cantelli. *Ann. Math. Statist.* **25** 131–138.

DEPARTMENT OF STATISTICS
UNIVERSITY OF CALIFORNIA
BERKELEY, CALIFORNIA 94720