The Dip Test of Unimodality

J. A. Hartigan; P. M. Hartigan

Annals of Statistics, Volume 13, Issue 1 (Mar., 1985), 70-84.

Stable URL:
http://links.jstor.org/sici?sici=0090-5364%28198503%2913%3A1%3C70%3ATDTOU%3E2.0.CO%3B2-A

Annals of Statistics

The Annals of Statistics
1985, Vol. 13, No. 1, 70–84

THE DIP TEST OF UNIMODALITY[1]

BY J. A. HARTIGAN AND P. M. HARTIGAN

Yale University and Veteran's Administration Hospital

The dip test measures multimodality in a sample by the maximum difference, over all sample points, between the empirical distribution function, and the unimodal distribution function that minimizes that maximum difference. The uniform distribution is the asymptotically least favorable unimodal distribution, and the distribution of the test statistic is determined asymptotically and empirically when sampling from the uniform.

1. Introduction. There are few statistical tests for discovering the presence of more than one mode in a distribution. One test, suggested by Wolfe (1970), uses the likelihood ratio for a two-component normal mixture against the normal null hypothesis. This test is defined in any number of dimensions. It is a formidable task to compute the statistic, and the usual maximum likelihood asymptotics do not apply (the likelihood ratio approaches infinity in probability in the null case as the number of observations increases). Worse, the test may be expected to be quite sensitive to the normality assumption, and may, for example, decide with high probability that a long-tailed unimodal distribution has more than one mode. Such a distribution may look more like a normal mixture than a normal.

A related test due to Engelman and Hartigan (1969) divides the sample into two subsets to maximize the likelihood ratio that the two subsets are sampled from normals with different means, against the null hypothesis that the means are equal. The test statistic is the maximum likelihood ratio over all divisions. The distribution is asymptotically normal (Hartigan, 1978), the statistic is easy to compute, but again the test will not work well when the bimodal alternative is not a normal mixture.

A test based on intervals between successive order statistics is suggested by J. B. Kruskal in Giacomelli, et al. (1971), but this test requires that the two modes be specified in advance. Another test based on intervals, using the idea that in a bimodal distribution, there should be a large interval accompanied by many small intervals on either side, is described in Hartigan (1977). In the many-dimensional case, a similar idea breaks the minimum spanning tree at a link such that there are large numbers of neighboring smaller links on each side of the break, and uses as test statistic the least of the two numbers of neighboring small links. Equivalently, take the maximum size of the smaller cluster among all pairs of disjoint single linkage clusters (Hartigan, 1981).

Why not compute the likelihood ratio test for unimodality versus bimodality

Received September 1983; revised June 1984.

[1] Research supported by the National Science Foundation Grant No. MCS-8102280.

AMS 1980 *subject classifications.* Primary 62G05; secondary 62F05.

Key words and phrases. Multimodality, isotonic regression, empirical distribution.

using monotone regression? The maximum likelihood density is infinite at the mode, and so it is necessary to constrain the estimate to be constant in the neighborhood of the mode (Wegman, 1970), perhaps by setting a bound on the density. Setting the width of the modal interval is analogous to the intractable problem of setting kernel widths in density estimation, which usually requires knowledge of higher derivatives of the density. See, for example, Silverman (1978). Setting the width is crucial in testing for bimodality, because contributions to the likelihood from observations near the mode and between the two modes in the bimodal fit, have the main effect—observations in the tail usually make the same contribution to the bimodal and unimodal likelihoods, and so have little effect, as one would desire.

Silverman (1981) suggests using the smallest window width, such that the resulting kernel density estimate is unimodal, as a test statistic for unimodality. The significance level of the test statistic is evaluated by empirically sampling from a rescaled version of the unimodal density estimate.

We propose the *dip statistic* as the maximum difference between the empirical distribution function, and the unimodal distribution function that minimizes that maximum difference. The statistic may be computed in order n operations, for n observations, and it is consistent for testing any unimodal against any multimodal distribution. We argue that the appropriate null distribution is uniform, by showing that the dip is asymptotically larger for the uniform than for any distribution in a wide class of unimodal distributions, those with exponentially decreasing tails. (We speculate that the result holds for the class of all unimodal distributions.) The asymptotic distribution of the dip is given in the uniform case, empirically derived distributions are specified for some finite sample sizes, and a few power computations are performed.

A modal interval is produced as an outcome of the dip calculation; it is not known how this competes with the various estimates of a mode considered by Chernoff (1964), Venter (1967), Wegman (1970), Sager (1979), Eddy (1980) and others. It does have the benefit of not requiring a kernel width.

2. The dip. A distribution function F is unimodal with mode m if F is convex in $(-\infty, m]$ and concave in $[m, \infty)$. The mode m is not necessarily unique. A unimodal F may have an atom only at a unique node m, and has a density, except possibly at a unique mode m, that increases in $(-\infty, m)$ and decreases in (m, ∞).

Define $\rho(F, G) = \sup_x |F(x) - G(x)|$ for any bounded functions F, G. Define $\rho(F, \mathscr{A}) = \inf_{G \in \mathscr{A}} \rho(F, G)$ for any class $\mathscr{A}$ of bounded functions. Let $\mathscr{U}$ be the class of unimodal distribution functions.

The *dip* of a distribution function F is defined by $D(F) = \rho(F, \mathscr{U})$. Note that $D(F_1) \leq D(F_2) + \rho(F_1, F_2)$ and $D(F) = 0$ for $F \in \mathscr{U}$, $D(F) > 0$ for $F \notin \mathscr{U}$; thus the dip measures departure from unimodality.

The *greatest convex minorant* (g.c.m.) of F in $(-\infty, a]$ is $\sup G(x)$ for $x \leq a$, where the sup is taken over all functions G that are convex in $(-\infty, a]$ and nowhere greater than F.

The *least concave majorant* (l.c.m.) of F in $[a, \infty)$ is $\inf L(x)$ for $x \geq a$, where the inf is taken over all functions L that are concave in $[a, \infty]$ and nowhere less than F.

It is necessary to extend the definition of the dip to bounded functions F that are constant on $[-\infty, 0]$ and on $[1, \infty]$. Define $\mathscr{V}$ to be the class of functions that are constant on $[-\infty, 0]$ and on $[1, \infty]$ and for some m, $0 \leq m \leq 1$, are convex on $[0, m]$ and concave on $[m, 1]$. Define $D(F) = \rho(F, \mathscr{V})$. We need to show that this definition is consistent with the previous one. Both definitions apply to distribution functions on $[0, 1]$.

THEOREM 1. *Let F be a distribution function with $F(0) = 0$, $F(1) = 1$. Then $\rho(F, \mathscr{U}) = \rho(F, \mathscr{V})$.*

PROOF. For $G \in \mathscr{U}$, define

$$H(x) = G(0)\{x < 0\} + G(x)\{0 \leq x \leq 1\} + G(1)\{x > 1\}.$$

(Here the set $\{x < 0\}$ is used as a $0 - 1$ function equal to 1 if $x < 0$ and to 0 if $x \geq 0$.)

Then $H \in \mathscr{V}$ and $\sup_x |F(x) - H(x)| = \sup_{0 \leq x \leq 1} |F(x) - G(x)| \leq \rho(F, G)$. Thus $\rho(F, \mathscr{V}) \leq \rho(F, \mathscr{U})$.

Conversely, suppose that $G \in \mathscr{V}$. If $G(0) \geq G(1)$, set

$$H(x) = (G(0) \wedge 1)\{0 \leq x < 1\} + \{x \geq 1\}.$$

Then $H \in \mathscr{U}$,

$$\begin{aligned} \rho(F, H) &\leq \max[|G(0) - F(0)|, |G(0) - F(1)|] \\ &\leq \max[|G(0) - F(0)|, |G(1) - F(1)|] \quad \text{since} \quad G(0) \geq G(1) \\ &\leq \rho(F, G). \end{aligned}$$

For $G(0) < G(1)$, define

$$H = G(0)\{G < G(0)\} + G\{G(0) \leq G(x) \leq G(1)\} + G(1)\{G > G(1)\}.$$

Then $H \in \mathscr{V}$, H is nondecreasing, and

$$\rho(F, H) = \sup_{G(0) \leq G(x) \leq G(1)} |F(x) - G(x)| \leq \rho(F, G).$$

Also $(H \vee 0) \wedge 1 \in \mathscr{V}$, is nondecreasing, and $\rho(F, (H \vee 0) \wedge 1) \leq \rho(F, H)$.

Thus for $G \in \mathscr{V}$, we can find H with $0 \leq H \leq 1$, H nondecreasing, $H \in \mathscr{V}$, $\rho(F, H) \leq \rho(F, G)$. Suppose H is convex on $[0, m]$ and concave on $[m, 1]$. For $a \geq 1$, define G_a to be the g.c.m. of $\{x \geq -a\}H$ on $(-\infty, m]$, and the l.c.m. of $\{x \leq a\} \vee H$ on $[m, \infty)$. Then $G_a \in \mathscr{U}$.

The function G_a is constant on $(-\infty, -a]$, consists of a linear segment on $[-a, x_1]$ for some x_1, $0 \leq x_1 \leq m$, and will be identical to H on $[x_1, m]$. As $a \to \infty$, the slope of the linear segment approaches zero, and since H is nondecreasing, $\sup_{0 \leq x \leq m} |H(x) - G_a(x)| \to 0$.

Similarly $\sup_{m \leq x \leq 1} |H(x) - G_a(x)| \to 0$ as $a \to \infty$,

$$\rho(F, G_a) = \sup_{0 \leq x \leq 1} |G_a(x) - F(x)| \to \sup_{0 \leq x \leq 1} |H(x) - F(x)| = \rho(F, H).$$

Thus for each $\varepsilon > 0$, each $G \in \mathscr{U}$, there exists $G_a \in \mathscr{U}$ with $\rho(F, G_a) \leq \rho(F, G) + \varepsilon$. It follows that $\rho(F, \mathscr{U}) \leq \rho(F, \mathscr{V})$. □

THEOREM 2. *Let F be a bounded function constant on $[-\infty, 0]$ and on $[1, \infty]$. Let I be the distribution function of the uniform on $(0, 1)$.*
Then $D(\alpha F + \beta I) = \alpha D(F)$ for $\alpha, \beta \geq 0$.

Thus if a distribution function F is mixed with a uniform, the dip of the resulting distribution is the dip of F multiplied by the mixing proportion of F.

PROOF:

$$D(\alpha F + \beta I) = \rho(\alpha F + \beta I, \mathscr{V}) = \inf_{G \in \mathscr{V}} \rho(\alpha F + \beta I, G)$$

$$= \inf_{G \in \mathscr{V}} \rho(\alpha F + \beta I, \alpha G + \beta I),$$

noting that $G \in \mathscr{V}$ if and only if $\alpha G + \beta I \in \mathscr{U}$, when $\alpha > 0$.
Thus $D(\alpha F + \beta I) = \inf_{G \in \mathscr{V}} \alpha \rho(F, G) = \alpha D(F)$.

3. Asymptotic behavior of the dip. If $X_1, X_2, \cdots, X_n$ is a sample from F, define the empirical distribution function F_n by

$$F_n(x) = (1/n) \sum \{X_i \leq x\}.$$

The Glivenko-Cantelli theorem states that $\rho(F_n, F) \to 0$ a.s., and so $D(F_n) \to D(F)$ a.s. A test based on the dip will thus asymptotically distinguish any unimodal F from any multimodal F.

In developing a test, it is necessary to choose a unimodal distribution as the null distribution, and we have chosen to use the uniform. This choice would be justified if $D(F_n)$ were stochastically larger for the uniform than for any other unimodal distribution—that is, if

$$P_I\{D(F_n) \geq x\} = \sup_{F \in \mathscr{U}} P_F\{D(F_n) \geq x\}.$$

Unfortunately this is not true for all x and n. For example, when $n = 4$, the dip is an increasing function of the statistic

$$T = 1 \vee [(X_{(3)} - X_{(2)})/\sup(X_{(2)} - X_{(1)}, X_{(4)} - X_{(3)})]$$

where $X_{(1)}, X_{(2)}, X_{(3)}, X_{(4)}$ are the order statistics.

Let $F = \alpha I + (1 - \alpha)\delta_1$ where $\delta_1(x) = \{x \geq 1\}$.

Let E_1, E_2, E_3 denote independent exponentials; then T is distributed as $[E_1/(E_2 + \frac{1}{2}E_3)] \vee 1$ in sampling from the uniform.

In sampling from F, T is distributed

as	$[E_1/(E_2 + \frac{1}{2}E_3)] \vee 1$	given that	$X_{(3)} < 1$;
as	$[E_1/E_2] \vee 1$	given that	$X_{(3)} = 1, X_{(2)} < 1$;
as	1	given that	$X_{(2)} = 1$.

The three conditioning events occur with respective probabilities $\alpha^4 + 4\alpha^3(1 - \alpha)$, $6\alpha^2(1 - \alpha)^2$, $4\alpha(1 - \alpha)^3 + (1 - \alpha)^4$.

Now $[E_1/E_2] \vee 1$ is stochastically larger than $[E_1/(E_2 + \frac{1}{2}E_3)] \vee 1$.

And it will always be possible to choose α close to 1 so that the conditioning event $X_{(2)} = 1$ has arbitrarily small probability compared to the event $X_{(3)} = 1$, $X_{(2)} < 1$. (The event $X_{(3)} < 1$ is of no importance since the conditional distribution of T is the same as for the uniform.)

Thus for each x, α may be chosen so that $P_F[T \geq x] > P_I[T \geq x]$; the uniform does not give the stochastically largest distribution of the dip.

However, we conjecture that, asymptotically, the distribution of the dip is stochastically largest for the uniform; we have been able to prove that $\sqrt{n}D(F_n)$ is asymptotically positive for the uniform and asymptotically zero for distributions whose densities decrease exponentially away from the mode.

THEOREM 3. *Let F_n be the empirical distribution function for a sample from the uniform on* (0, 1), *and let B be the Brownian bridge process with* $\text{cov}[B(s), B(t)] = s(1 - t)$, $0 \leq s \leq t \leq 1$. *Assume that B is zero outside* (0, 1).

Then $\sqrt{n}D(F_n) \to D(B)$ in distribution as $n \to \infty$.

PROOF. From the Skorohod embedding of Donsker's theorem (Breiman, 1968, page 296), it is possible to construct a probability space in which $\tilde{F}_n$ has the same distribution as F_n for each n, and in which B is a Brownian bridge, and

$$\sup_{0 \leq x \leq 1} | \sqrt{n}(\tilde{F}_n(x) - x) - B(x) | \to 0 \quad \text{in probability.}$$

Since $D(F_1) - D(F_2) \leq \rho(F_1, F_2)$,

$$| D(\sqrt{n}\tilde{F}_n) - D(\sqrt{n}I + B) | \leq \sup_x | \sqrt{n}(\tilde{F}_n(x) - x) - B(x) | \to 0,$$

$$| D(\sqrt{n}\tilde{F}_n) - D(B) | \to 0 \quad \text{in probability,} \quad \text{from Theorem 2.}$$

Thus $\sqrt{n}D(F_n) \to D(B)$ in distribution.

THEOREM 4. *Let F_n be the empirical distribution for a sample of size n from I. Then*

$$\sup_{0 \leq x \leq y \leq 1} \{ \sqrt{n} \, [F_n(y) - F_n(x) - (y - x)] / [(\sqrt{y - x} + 1/\sqrt{n})(\log n)^2] \} \to 0$$

in probability as $n \to \infty$.

PROOF. This result differs from Theorem 1.3 of Shorack and Wellner (1982) only in the standardization factor $\sqrt{y - x} + (1/\sqrt{n})$ where the $(1/\sqrt{n})$ is introduced to obtain a result uniform over all intervals (they exclude very small intervals), and in the $(\log n)^2$ which is a little larger than their factor.

From Bennett (1962), if Z is binomial with mean np and variance $np(1 - p)$, $P(|Z - np| > t) \leq 2 \exp(-t^2/[2np(1 - p) + \frac{2}{3}t])$.

Set $Z = n(F_n(y) - F_n(x))$, $p = y - x$, $t = \delta\sqrt{n}\,(\sqrt{y - x} + (1/\sqrt{n}))(\log n)^2$.

Let $Z^n_{x,y} = \sqrt{n} \, [F_n(y) - F_n(x) - (y - x)] / [(\sqrt{y - x} + (1/\sqrt{n}))(\log n)^2]$.

Then $P[|Z^n_{x,y}| > \delta] = P[|Z - np| > t] \leq 2 \exp[-\frac{1}{4}\delta^2(\log n)^2]$ for $\delta \leq 3$, $n \geq 3$, considering the two cases $t \leq 3np$ and $t > 3np$ separately.

Now consider A_n, the set of intervals in which the endpoints are of form kn^{-3}, $0 \leq k \leq n^3$, k an integer. There are $\frac{1}{2}n^6$ such intervals.

As x changes over the interval $[kn^{-3}, (k+1)n^{-3}]$, $F_n(x) - x$ changes by at most $(1/n)$ if the interval contains at most one point. Let B_n be the event that all intervals of the form $[kn^{-3}, (k+1)n^{-3}]$ contain at most one point. Then when B_n holds, $Z^n_{x,y}$ changes by at most

$$\sqrt{n}\ (2/n)/[(1/\sqrt{n})(\log n)^2] = 2/(\log n)^2$$

as x and y change over intervals of form $[kn^{-3}, (k+1)n^{-3}]$.

Note that

$$P[\sup_{x,y\in A_n} |Z^n_{x,y}| > \delta] \le n^6 \sup_{x,y\in A_n} P[|Z^n_{x,y}| > \delta]$$

$$\le 2n^6 \exp[-\tfrac{1}{4}\delta^2(\log n)^2] \to 0 \quad \text{as} \quad n \to \infty.$$

$$P[B_n^c] \le n^3 P\{[kn^{-3}, (k+1)n^{-3}] \text{ contains more than one point}\}$$

$$\le n^5 P\{[kn^{-3}, (k+1)n^{-3}] \text{ contains the first two observations}\}$$

$$\le 1/n \to 0 \quad \text{as} \quad n \to \infty,$$

$$\{\sup_{0\le x\le y\le 1} |Z^n_{x,y}| > \delta\} \le \{\sup_{x,y\in A_n} |Z^n_{x,y}| > \delta - 2/(\log n)^2\} + B_n^c$$

$$\to 0 \quad \text{in probability as required.} \ \square$$

THEOREM 5. *Let F be unimodal with nonzero kth derivative at the mode m, for some $k \ge 2$, and let F have exponentially decreasing density, that is,*

$$\inf_{0<F'(x)<F'(m)-\varepsilon} |(d/dx)\log F'(x)| > 0 \quad \textit{for each} \quad \varepsilon > 0.$$

Then $\sqrt{n}D(F_n) \to 0$ in probability.

PROOF. Let G_n be the unimodal distribution equal to the g.c.m. of F_n on $(-\infty, m]$ and the l.c.m. of F_n on $[m, \infty)$. We show that G_n is linear in segments of length $o_p(1)$, such that $G_n = F_n$ at the endpoints of the segments. It then follows from Theorem 4 that, in probability, $\sqrt{n}\rho(G_n, F_n) \to 0$, so $\sqrt{n}D(F_n) \to 0$.

A similar result is proved by Kiefer and Wolfowitz (1976) for concave distribution functions, but their conditions do not allow infinite tails or $F''(m) = 0$. The condition on the tails is undoubtedly too strong, but at least the normal distribution is covered; it is sufficient to have, for some $k \ge 2$, (C)

$$\sup_{m\notin(x_1,x_2),-\infty\le x_1\le x_2\le\infty} [F(x_2) - F(x_1)]^k / |F(x_2) + F(x_1) - 2F(\bar{x})| < \infty$$

where $\bar{x} = \frac{1}{2}(x_1 + x_2)$.

Let $m = 0$ without loss of generality. We will first show that condition (C) is implied by the conditions given in the theorem.

(i) For $x_1 < x_2 < -\varepsilon$, $F'(x) > 0$,

$$(d/dx)\log F'(x) > B > 0.$$

$$F'(x_2)/F'(x_1) \ge \exp[B(x_2 - x_1)]$$

$$F(x_2) - F(x_1) \ge [F(\bar{x}) - F(x_1)]\exp(B\delta) \quad \text{where} \quad \delta = (x_2 - x_1)/2.$$

For $B\delta \geq 1$,

$$[F(x_2) - F(x_1)]^k/[F(x_2) + F(x_1) - 2F(\bar{x})]$$
$$\leq [F(x_2) - F(\bar{x})]^k 2^k/[F(x_2) - F(\bar{x})][1 - e^{-1}] \leq 2^k/(1 - e^{-1}).$$

For $B\delta < 1$,

$$[F(x_2) - F(x_1)]^k/[F(x_2) + F(x_1) - 2F(\bar{x})]$$
$$\leq [F(x_2) - F(\bar{x}_1)]^k 2^k/[F(x_2) - F(\bar{x})]B\delta$$
$$\leq 2^k \delta^{k-2}[F'(0)]^{k-1}/B \leq 2^k[F'(0)]^{k-1}/B^{k-1}.$$

Similar bounds will apply for $0 < \varepsilon < x_1 < x_2$; note that B will depend on ε.

(ii) For $-\varepsilon < x_1 < x_2 < 0$,

$$F(x_2) + F(x_1) - 2F(\bar{x}) = \tfrac{1}{3}(x_2 - x_1)^2[F''(y_1) + F''(y_2)]$$

where $x_1 \leq y_1 \leq \bar{x} \leq y_2 \leq x_2$.

If $F^{(k)}(0)$ is the first nonzero derivative at the mode (except for $F'(0)$),

$$F''(y_1) = y_1^{k-2}H(y_1)/(k-2)! \quad \text{where} \quad H(y) \to F^{(k)}(0) \quad \text{as} \quad y \to 0.$$

Thus $[F(x_2) + F(x_1) - 2F(\bar{x})] \geq \tfrac{1}{3}(x_2 - x_1)^2 |\bar{x}|^{k-2} A(x_1, x_2)/(k-2)!$ where $A(x_1, x_2) \to -F^{(k)}(0)$ as $x_1 \to 0$.

$$[F(x_2) - F(x_1)]^k/[F(x_2) + F(x_1) - 2F(\bar{x})] \leq 8F'(0)^k 2^{k-2}/A(x_1, x_2)$$
$$\leq 2^{k+2}F'(0)^k/|F^{(k)}(0)|$$

for ε sufficiently close to zero.

(iii) It is necessary to consider also the case when x_1 is far from zero and x_2 is close to zero. Suppose $x_1 \leq -\varepsilon \leq -\eta \leq x_2 \leq 0$, where $F(0) - F(-\eta) \leq F(-\eta) - F(-\varepsilon)$. Then

$$[F(x_2) - F(x_1)]^k/[F(x_2) + F(x_1) - 2F(\bar{x})]$$
$$\leq 2^k[F(-\eta) - F(x_1)]^k/[F(-\eta) + F(x_1) - 2F(x_1 - \eta/2)]$$
$$\leq 2^k C_\eta \quad \text{say,}$$

using the established bound for $x_1 \leq x_2 \leq -\eta$.

Choose ε to establish (ii), then η to establish (iii); (i) holds for that η. The cases where x_1, x_2 are positive are handled similarly. This establishes condition C, which we will now use to prove the theorem.

Let G_n be the greatest convex minorant of F_n in $[-\infty, 0]$, and the least concave majorant of F_n in $[0, \infty]$. Then G_n is a unimodal distribution; we will show that $\sqrt{n}\rho(F_n, G_n) \to 0$ in probability, which implies that $\sqrt{n}D(F_n) \to 0$ in probability.

Let $x - \delta$, $x + \delta$ denote a maximal interval in $[-\infty, 0]$ where G_n is linear. Then $F_n(x - \delta) = G_n(x - \delta)$, $F_n(x + \delta) = G_n(x + \delta)$. Since G_n is the greatest convex

minorant,

$$F_n(y) \geq G_n(y) \quad \text{for} \quad x - \delta \leq y \leq x + \delta.$$

Thus $F_n(x) - F_n(x - \delta) \geq F_n(x + \delta) - F_n(x)$.

Let $H(x, \delta, n) = ([F(x + \delta) - F(x - \delta)]^{1/2}/n^{1/2} + 1/n)(\log n)^2$. From Theorem 4,

$$F_n(x) - F_n(x - \delta) = F(x) - F(x - \delta) + \Delta(H(x, \delta, n)).$$

(The notation $A(x, \delta, n) = \Delta(B(x, \delta, n))$ means $\sup_{x,\delta}(|A(x, \delta, n)|/B(x, \delta, n)) \to 0$ in probability as $n \to \infty$.)

Thus

$$F(x) - F(x - \delta) \geq F(x + \delta) - F(x) + \Delta(H(x, \delta, n))$$

$$0 \leq F(x - \delta) + F(x - \delta) - 2F(x) \leq \Delta(H(x, \delta, n)).$$

For some $k \geq 2$,

$$[F(x + \delta) - F(x - \delta)]^k = \Delta[H(x, \delta, n)] \quad \text{from condition (C).}$$

$$[F(x + \delta) - F(x - \delta)]^{2k-1} = \Delta[(\log n)^4/n]$$

$$H(x, \delta, n) = \Delta(n^{-1/2}).$$

Now consider $\sup_{x-\delta \leq y \leq x+\delta} |F_n(y) - G_n(y)|$.

$$\begin{aligned} F_n(y) - G_n(y) &= F_n(y) - F_n(x - \delta) - (y - x + \delta)[F_n(x + \delta) - F_n(x - \delta)]/2\delta \\ &= F(y) - F(x - \delta) - (y - x + \delta)[F(x + \delta) - F(x - \delta)]/2\delta \\ &\quad + \Delta(H(x, \delta, n)). \end{aligned}$$

Since F is concave in $[-\infty, 0]$, the first expression is negative, and

$$F_n(y) - G_n(y) = \Delta(H(x, \delta, n)) = \Delta[n^{-1/2}].$$

This holds uniformly over all x, δ and y, so

$$\sup_{x \leq 0} |G_n(x) - F_n(x)| = \Delta(n^{-1/2})$$

Similarly $\sup_{x \geq 0} |G_n(x) - F_n(x)| = \Delta(n^{-1/2})$, so $\sqrt{n}\rho(G_n, F_n) \to 0$ as required.

4. Computing the dip.

THEOREM 6. *Let F be an arbitrary distribution function. Then $D(F) = d$ only if there exists a nondecreasing function G such that, for some $x_L \leq x_U$,*

(i) *G is the greatest convex minorant of $F + d$ in $(-\infty, X_L)$*

(ii) *G has constant maximum slope in (x_L, x_U)*

(iii) *G is the least concave majorant of $F - d$ in $[x_U, \infty)$*

(iv) $d = \sup_{x \notin (x_L, x_U)} |F(x) - G(x)| \geq \sup_{x \in (x_L, x_U)} |F(x) - G(x)|$.

PROOF. We need some preliminary facts.

(A) The maximum value of $D(F)$ is ¼, achieved when F has two atoms of size ½. To see that $D(F) \leq ¼$, consider a unimodal G that is long-tailed and symmetric about the median of F with an atom of size ½ at the median.

(B) If a unimodal G has mode m, for $A > 0$, $F(A) - F(-A) > 2\rho(F, G)$, then

$$|m| \leq 2A/[F(A) - F(-A) - 2\rho(F, G)].$$

(The worst case occurs when G is linear between a negative m and A, $G(m) = 0$, $G(-A) = F(-A) + \rho(F, G)$, $G(A) = F(A) - \rho(F, G)$.)

(C) Let $\mathcal{G}$ be the class of functions G that are nondecreasing, convex in $(-\infty, m]$ and concave in $[m, \infty)$ for some mode m. Then there exists G in $\mathcal{G}$ such that $\rho(F, G) = D(F)$. (It is true that there exists G in $\mathcal{U}$ such that $\rho(F, G) = D(F)$ but we do not need the slight extra generality.)

Consider a sequence G_i of unimodal distribution functions with $\rho(F, G_i) \to D(F)$. Take $\rho(F, G_i) \leq ½$ so $|m_i| \leq 2A/(F(A) - F(-A) - ¼)$. The m_i thus have a limit point m, say, and by a standard diagonalization argument we can specify a function $G(r)$ on the rationals and m such that G is nondecreasing, convex on $(-\infty, m)$ and concave on (m, ∞), and $G_{n_i}(r) \to G(r)$ on some subsequence of G_i. The function $G(x) = \sup_{r \leq x} G(r)$ has these properties on the real line. (Note that G is continuous on the rationals except possibly at m.)

Thus $\rho(F, G) = \lim \rho(F, G_{n_i}) = D(F)$ as required.

(D) Note that $F(x) - D(F) \leq G(x) \leq F(x) + D(F)$ all x. Let G_0 be the greatest convex minorant of $G(m) \wedge [F + D(F)]$ in $(-\infty, m]$ and the least concave majorant of $G(m) \vee [F - D(F)]$ in $[m, \infty)$. Then $G_0 \in \mathcal{G}$, $\rho(F, G_0) = D(F)$, since G_0 is bounded by $F - D(F)$, $F + D(F)$ by its definition.

Let $x_L = \sup\{x \mid G_0(x) = D(F) + F(x)\}$, $x_U = \inf\{x \mid G_0(x) = F(x) - D(F)\}$. G_0 will be linear in the interval $[x_L, m]$ and in the interval $[m, x_U]$.

We may assume that $m = x_L$ or $m = x_U$ because we can repeat the construction with mode at x_L if G_0 has higher slope in $[x_L, m]$ and with mode at x_U otherwise. (If $x_L = m = x_U$ then F is unimodal and the conditions (i)–(iv) are trivially satisfied. Otherwise G_0 is continuous.)

However the interval (x_L, x_U) is not necessarily of maximum slope. Suppose that $m = x_L$; it is then possible that G_0 has maximum slope at points less than x_L. In this case consider the minimum m such that G_0 has mode m, $G_0(m) = D(F) + F(m)$; G_0 is the greatest convex minorant of $F + D(F)$ in $(-\infty, m]$, G_0 is the least concave majorant of $F - D(F)$ in $[m, \infty)$. (A compactness argument similar to the one showing $\rho(G, F) = D(F)$ establishes the existence of G_0.)

Let m_1 be the greatest $x < m$ such that $G_0(x) = F(x) + D(F)$. If $m_1 < m$ then G_0 is linear in $[m_1, m]$. If G_0 has greater slope in $[m_1, m]$ than in $[m, x_U]$, then G_0 is concave in $[m_1, \infty]$ and so the requirement that m is minimal is contradicted; thus if $m_1 < m$, conditions (i)–(iv) of the theorem are satisfied.

If $m_1 = m$, but G_0 has greater slope less than m than in $[m, x_U]$, take x_0 close enough to m so that the line segment $[x_0, G_0(x_0)]$ to $[x_U, G_0(x_U)]$ lies between

$F - D(F)$ and $F + D(F)$; this is possible because the interior of the segment $[m, G_0(m)]$ to $[x_U, G_0(x_U)]$ lies in the region $\{(x, y) \mid F(x) - D(F) < y < F(x) + D(F)\}$. Now define G^* to be equal to G_0 except for $x_0 \le x \le x_U$, and define $G^*(x)$ to be linear in (x_0, x_U). Then G^* contradicts the minimality of m.

Thus in both cases G_0 has greater slope in $[m, x_U]$, and conditions (i)–(iv) are satisfied. □

For a purely discrete distribution function such as the empirical distribution function, the theorem suggests the following algorithm:

Let $x_1, x_2, \cdots, x_n$ be the atoms of F.

The only possible endpoints of the modal interval (x_L, x_U) are the atoms. Consider the $n(n-1)/2$ possible modal intervals, and compute for each $[x_i, x_j]$ the greatest convex minorant of F in $[-\infty, x_i]$ and the least concave majorant of F in $[x_j, \infty]$. Let d_{ij} be the maximum distance of F to these computed curves. Then $2D(F)$ is the minimum value of d_{ij} over all modal intervals (x_i, x_j) such that the line segment $[x_i, F(x_i) + \frac{1}{2}d_{ij}]$ to $[x_j, F(x_j) - \frac{1}{2}d_{ij}]$ lies in $\{(x, y) \mid F(x) - \frac{1}{2}d_{ij} \le y \le F(x) + \frac{1}{2}d_{ij}\}$.

The minorant and majorant computations may be made once and for all in order n. At first sight, it looks as if $n(n-1)/2$ modal intervals must be examined, but many possibilities may be excluded—for x_i the lower endpoint, only those x_j where the least concave majorant of F in $[x_i, \infty)$ touches F need be considered.

An order n algorithm exists. Consider a taut string stretched between $[x_1, d]$ and $[x_n, F(x_n) - d]$ where $x_1 \le x_i \le x_n$ for $1 \le i \le n$. Assume that the curves $\{x, F(x) + d\}$ and $\{x, F(x) - d\}$ are solid. As d decreases, the string bends to form a convex minorant from x_1 to x_L and a concave majorant from x_U to x_n where x_L increases with d decreasing, and x_U decreases with d decreasing. Then $D(F)$ is the value of d such that any further decrease forces the string out of its unimodal shape. It is necessary to consider at most n changes in d, and order 1 calculations for each change, so the computation is order n.

The following algorithm implements the stretched string notion:

(i) Begin with $x_L = x_1$, $x_U = x_n$, $D = 0$.

(ii) Compute the g.c.m. G and l.c.m. L for F in $[x_L, x_U]$; suppose the points of contact with F are respectively $g_1, g_2, \cdots, g_k$ and $l_1, l_2, \cdots, l_m$.

(iii) Suppose $d = \sup \mid G(g_i) - L(g_i) \mid > \sup \mid G(l_i) - L(l_i) \mid$ and that the sup occurs at $l_j \le g_i \le l_{j+1}$. Define $x_L^0 = g_i$, $x_U^0 = l_{j+1}$.

(iv) Suppose $d = \sup \mid G(l_i) - L(l_i) \mid \ge \sup \mid G(g_i) - L(g_i) \mid$ and that the sup occurs at $g_i \le l_j \le g_{i+1}$. Define $x_L^0 = g_i$, $x_U^0 = l_j$.

(v) If $d \le D$, stop and set $D(F) = D$.

(vi) If $d > D$, set

$$D = \sup\{D, \sup_{x_L \le x \le x_L^0} \mid G(x) - F(x) \mid, \sup_{x_U^0 \le x \le x_U} \mid L(x) - F(x) \mid\}.$$

(vii) Set $x_U = x_U^0$, $x_L = x_L^0$ and return to (ii).

5. Percentage points and power of the dip. In Table 1 appear the percentage points

.01 .05 .10 .50 .90 .95 .99 .995 .999

of the DIP, for sample sizes n = 4–10, 15, 20, 30, 50, 100, 200, based on 9999 repetitions from the uniform. From Theorem 3, $\sqrt{n}D(F_n)$ converges in distribution to the dip computed for a Brownian bridge, and the table shows that $\sqrt{n}D(F_n)$ has very nearly the same percentage points for $n = 100$ as $n = 200$. It is suggested that interpolation be based on $\sqrt{n}$DIP.

We have not completed very extensive power computations. The following special case is illuminating however. Let F_0 be uniform on (0, 1) and let F_1 be a mixture, in the proportions 3:2:3, of a uniform on (0, ¼), a uniform on (¼, ¾), and a uniform on (¾, 1). Thus F_1 has a density f_1

$$f_1(x) = \tfrac{3}{2}\{0 \le x \le \tfrac{1}{4}\} + \tfrac{1}{2}\{\tfrac{1}{4} < x < \tfrac{3}{4}\} + \tfrac{3}{2}\{\tfrac{3}{4} \le x \le 1\}.$$

Consider the three tests for bimodality:

(i) the *dip*

(ii) the *depth*:

$$\sup[\inf[F_n(x_5, x_6),\ F_n(x_1, x_2] - F_n(x_3, x_4)]$$

over all points $x_1 \le x_2 \le x_3 \le x_4 \le x_5 \le x_6$ such that $x_4 - x_3 \ge x_2 - x_1$, $x_6 - x_5$; where $F_n(x, y) = F_n(y) - F_n(x)$. The depth is similar in aim to

TABLE 1

Percentage points of the dip in uniform samples

(1) *Dip is the maximum distance between the empirical distribution and the best fitting unimodal distribution.*

(2) *Based on 9999 dips. Maximum standard error is* .001.

sample size	probability of dip less than tabled value is								
	.01	.05	.10	.50	.90	.95	.99	.995	.999
4	.1250	.1250	.1250	.1250	.1863	.2056	.2325	.2387	.2458
5	.1000	.1000	.1000	.1217	.1773	.1872	.1966	.1981[3]	.1996[3]
6	.0833	.0833	.0833	.1224	.1586	.1645	.1904	.2034	.2224
7	.0714	.0714	.0822	.1181	.1445	.1597	.1832	.1900	.2035
8	.0625	.0745	.0828	.1109	.1428	.1552	.1744	.1801	.1978
9	.0618	.0735	.0807	.1041	.1362	.1458	.1623	.1693	.1851
10	.0610	.0718	.0780	.0979	.1302	.1394	.1623[3]	.1699	.1828
15	.0544	.0606	.0641	.0836	.1097	.1179	.1365	.1424	.1538
20	.0474	.0529	.0569	.0735	.0970	.1047	.1209	.1262	.1382
30	.0395	.0442	.0473	.0617	.0815	.0884	.1012	.1061	.1177
50	.0312	.0352	.0378	.0489	.0645	.0702	.0804	.0842	.0926
100	.0228	.0256	.0274	.0355	.0471	.0510	.0586	.0619	.0687
200	.0165	.0185	.0197	.0255	.0341	.0370	.0429	.0449	.0496

(3) *Repeated computations.*

(4) *Interpolate on* $\sqrt{n}$ *dip.*

the dip; it identifies three intervals of equal length such that the middle interval has low empirical probability relative to both the outside intervals. It has similar asymptotic properties to the dip, but is $O(n^2)$ in computation.

(iii) *likelihood ratio*:

$$\sup_{f \in U_{1C}} \sum_{i=1}^{n} \log f(x_i) / \sup_{f \in U_{2C}} \sum_{i=1}^{n} \log f(x_i)$$

where U_{1C}, is the class of unimodal densities, and U_{2C} is the class of bimodal densities, constrained to have maximum density less than C.

The constraint on the density is necessary because $\sum \log f(x_i)$ has maximum value ∞ over all unimodal densities. Wegman (1970) considers instead constraining the density to be constant in an interval about the mode—the two constraints are equivalent with appropriate choice of C and interval length.

The constrained unimodal density estimate has value C in some interval $[x_L, x_U]$, and is proportional to the density of the g.c.m. in $(-\infty, X_L)$ and to the density of the l.c.m. in (x_U, ∞). It is thus quite similar in form to the DIP estimate. There are two important differences.

(i) The measure of distance for the likelihood ratio test is $\int \log f \cdot dF_n$ and for the dip is $\sup |F - F_n|$, so that the dip is relatively insensitive to large changes in f that cause small changes in F.

(ii) The dip automatically determines $[x_L, x_U]$, the modal interval, but the likelihood ratio statistic requires some specification of C. We have used $C = 1/ng_{\sqrt{n}}$ where $g_{\sqrt{n}}$ is the $\sqrt{n}$ smallest interval between the order statistics; this may well result in too large a C (allowing exaggerated contributions near the mode) when there is some discreteness or rounding in the data. The dip is insensitive to such rounding.

It will be seen from Table 2 that the dip appears slightly superior to the depth, and markedly superior to the likelihood ratio. The particular alternative distribution, a mixture of uniforms, is very suitable for the type of density estimation used in the depth, and the depth should be expected to do relatively worse in other applications. The poor performance of the likelihood ratio test is probably due to poor choice of C; but the real defect of this method is the difficulty in choosing C.

We have used as reference distribution the "least-favorable" unimodal distribution, the uniform. There may be evidence in the data that, if the true

TABLE 2

Power of dip, depth, and likelihood ratio.
In sampling from F_1, the probability that the statistic exceeds the 95% point computed in sampling from F_0(based on 1000 repetitions).

Sample Size	dip	depth	likelihood ratio
50	.795	.749	.540
100	.973	.961	.905

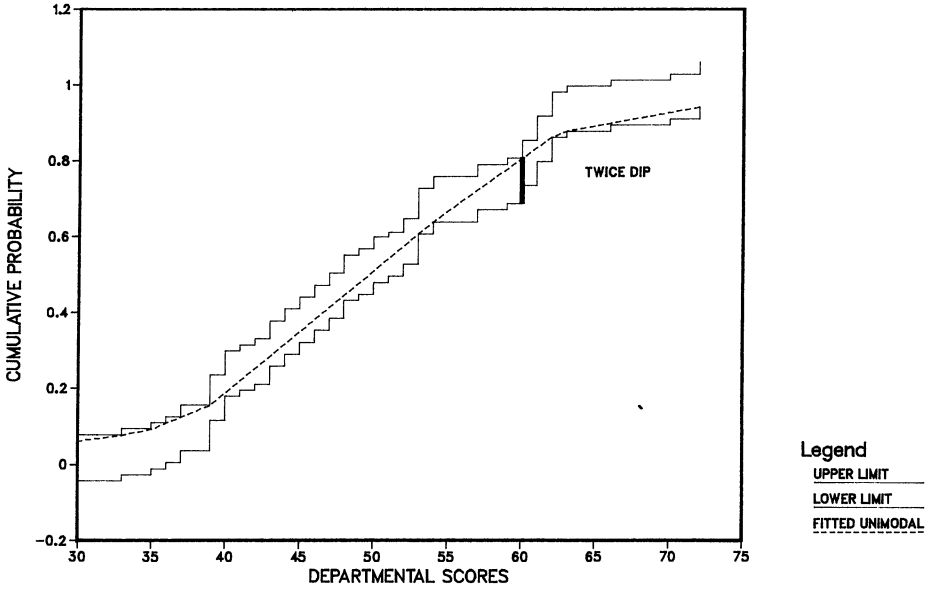

FIG. 1. *Faculty quality in statistics department.*

distribution is unimodal, it is far from uniform. Following Silverman (1981), it would be possible to evaluate the significance of a computed dip, against the null distribution of dips obtained by sampling from the best fitting unimodal distribution as specified in Theorem 6. This procedure should have better power than the present test for discovering two relatively close modes with pronounced tails—the new procedure effectively conditions on the observations not in the tails.

6. An example. From Scully (1982), the quality of faculty in 63 statistics departments was assessed on a range of 30 to 72, distributed in the counts

$$10010112054113222231212520020134410010001 02.$$

There seem to be modes about 40 and about 60. The dip illustrated in Figure 1 is .059 which has a tail probability about 10% from Table 1. ($\sqrt{63}D(F_{63})$ is distributed approximately as $\sqrt{100}D(F_{100})$ which has 90% point .474. Thus $D(F_{63})$ has 90% point .060.)

7. The multivariate case. Empirical distribution functions do not generalize nicely to more than 1 dimension; but we can linearize the higher dimensional problem in various ways. The most promising linearization is through the minimum spanning tree, which unfortunately requires some measure of distance on the space. The minimum spanning tree is the graph of minimum total length connecting all sample points.

(i) Use the greatest dip over all linear combinations of the original variables.

(ii) Let a particular data point x_0 be a trial value of the mode. All data points have a partial order in which $x < y$ if y lies between x and x_0 on the tree. Define the empirical probability on the tree, relative to x_0, by $F_n(x) =$ [# of points $\leq x$]$/n$. A probability distribution F is unimodal with respect to x_0 if F is supported by the minimum spanning tree and has increasing density according to the partial order. The dip for x_0 will be the maximum distance between F_n and the closest unimodal distribution; then select x_0 to minimize the dip for x_0.

(iii) A simpler technique begins again with a trial value of the mode, say x_0; the closest point to x_0, say x_1, is found; then the closest point to either x_0 or x_1, say x_2; the closest point to x_0, x_1 or x_2; and so on. These points are easily determined from the minimum spanning tree and vice-versa.

Let the successive closest distances be $d_1, d_2, \cdots, d_{n-1}$ and define $y_i = \sum_{j=1}^{i} d_j$. Let F_n be the empirical distribution on the y_i and define $D(x_0) = \rho(F_n, F)$ where F is the least concave minorant of F_n. (We expect the d_i to be roughly increasing if x_0 is the unique mode.) The test statistic would be $\inf_{x_0} D(x_0)$.

REFERENCES

BENNETT, G. (1962). Probability inequalities for the sum of independent random variables. *J. Amer. Statist. Assoc.* **57** 33–45.

BREIMAN, L. (1968). *Probability.* Addison, Reading.

CHERNOFF, H. (1964). Estimation of the mode. *Ann. Instit. Statist. Math.* **16** 31–41.

EDDY, W. F. (1980). Optimum kernel estimators of the mode. *Ann. Statist.* **8** 870–882.

ENGELMAN, L. and HARTIGAN, J. A. (1969). Percentage points of a test for clusters. *J. Amer. Statist. Assoc.* **64** 1647–1648.

GIACOMELLI, F., WIENER, J., KRUSKAL, J. B., POMERAN, J. W., and LOUD, A. V. (1971). Subpopulations of blood lymphocytes as demonstrated by quantitative cytochemistry. *J. Histochemistry Cytochemistry* **19** 426–433.

HARTIGAN, J. A. (1977). Distribution problems in clustering. *Classification and Clustering*, ed. J. V. Ryzin. Academic, New York.

HARTIGAN, J. A. (1978). Asymptotic distributions for clustering criteria. *Ann. Statist.* **6** 117–131.

HARTIGAN, J. A. (1981). Consistency of single linkage for high density clusters. *J. Amer. Statist. Assoc.* **76** 388–394.

KIEFER, J. and WOLFOWITZ, J. (1976). Asymptotically minimax estimation of concave and convex distribution functions. *Z. Wahrsch. verw. Gebiete* **34** 73–85.

SAGER, T. (1979). An iterative procedure for estimating a multivariate mode and isopleth. *J. Amer. Statist. Assoc.* **74** 329–339.

SCULLY, M. G. (1982). Evaluation of 596 programs in mathematics and physical sciences. *Chronicle Higher Educ. (29 Sept, 1982)* **25** *5* 8–10.

SHORACK, G. R. and WELLNER, JON A. (1982). Limit theorems and inequalities for the uniform empirical process indexed by intervals. *Ann. Probab.* **10** 639–652.

SILVERMAN, B. W. (1978). Choosing the window width when estimating a density. *Biometrika* **65** 1–11.

SILVERMAN, B. W. (1981). Using kernel density estimates to investigate multimodality. *J. Roy. Statist. Soc. B* **43** 97–99.

VENTER, J. H. (1967). On estimation of the mode. *Ann. Math. Statist.* **38** 1446–1455.

WEGMAN, E. J. (1970). Maximum likelihood estimation of a unimodal density, II. *Ann. Math. Statist.* **6** 2169–2174.

WOLFE, J. H. (1970). Pattern clustering by multivariate mixture analysis. *Multivariate Behavioural Res.* **5** 329–350.

STATISTICS DEPARTMENT
YALE UNIVERSITY
NEW HAVEN, CONNECTICUT 06520

COOPERATIVE STUDIES UNIT
VETERAN'S ADMINISTRATION HOSPITAL
WEST HAVEN, CONNECTICUT