The Annals of Statistics
1987, Vol. 15, No. 2, 865–874

IDENTIFYING THE CLOSEST SYMMETRIC DISTRIBUTION OR DENSITY FUNCTION

BY EUGENE F. SCHUSTER

The University of Texas at El Paso

The problem addressed is that of finding the "closest" symmetric distribution (density) to a given theoretical or empirical distribution (density) function. Measures of "closeness" considered include: weighted sup norm, weighted L_p norm and Hellinger distance. Explicit formulas are given for the closest symmetric distribution function to the empirical distribution function in both sup norm and integrated square error.

1. Introduction and motivation. The problem of finding the "closest" symmetric distribution or density is of interest in:

(1) quantifying the asymmetry of a distribution;
(2) estimating the distribution or density function of a symmetric distribution; and
(3) using (symmetric) bootstrap procedures when the underlying distribution is assumed symmetric.

Measures of "closeness" considered are: weighted sup norm, weighted L_p norm, weighted Cramér–von Mises and Hellinger distance. We became interested in finding the closest symmetric distribution to the empiric in related work in using the bootstrap in testing the nonparametric hypothesis of symmetry versus asymmetry, i.e., in testing:

$$H_0\colon F \text{ (unknown) is symmetric about some center } \theta \text{ (unknown)}$$

against

$$H_a\colon F \text{ is asymmetric}.$$

One natural way of testing this hypothesis is to estimate θ by θ_n and then use θ_n in a nonparametric test for symmetry about a known θ. For example, one might use an estimate of θ in the so-called Butler (1969) statistic [see also Orlov (1972), Smirnov (1947), Schuster and Narvarte (1973) and Koziol (1983)]

$$h_n(\theta) = \sup|F_n(x) - 1 + F_n((2\theta - x)-)|, \tag{1.1}$$

or the Cramér–von Mises statistic

$$h_n(\theta) = \int_{-\infty}^{\infty} \{F_n(x) - 1 + F_n((2\theta - x)-)\}^2\,dx, \tag{1.2}$$

where F_n is the usual empirical cumulative distribution function (cdf).

Received October 1985; revised September 1986.

AMS 1980 *subject classifications*. Primary 62E99; secondary 62G99.

Key words and phrases. Empirical distribution function, symmetrical distribution, symmetrical density, closest symmetrical distribution, symmetrical bootstrap, minimum distance.

The problem with this approach is that the resulting tests are not distribution-free. Hence, one cannot compute critical values and/or p-values, e.g., see Boos (1982) for a study of a test for asymmetry associated with $h_n(\theta)$ of (1.2) with θ estimated by the Hodges–Lehmann estimator of (2.3). One can sidestep this problem by using the bootstrap to estimate the critical and/or p-values. However, under the null hypothesis F is symmetric, so one should not bootstrap from F_n but from the "closest" symmetric distribution. Results of these simulation studies have been encouraging and are reported in Schuster and Barker (1987).

A second approach to testing the general hypothesis of symmetry might be to use a test statistic that measures asymmetry by the distance from F_n to the closest symmetric distribution. If the distance measure is sup norm or integrated square error, and θ_n is the natural estimator of θ obtained by minimizing (1.1) or (1.2), then Theorems 1 and 2 indicate that these two approaches are identical in the corresponding cases.

In Section 2, we identify the "closest" symmetric cdf. We give explicit formulas for the closest symmetric cdf to the empirical cdf in sup norm and in integrated square error. In Section 3, we identify the closest symmetric pdf to an arbitrary fixed pdf for weighted distance measures.

2. The "closest" symmetric cdf. Let $X_1, \dots, X_n$ be the order statistics of a random sample from a cdf F and let F_n be the corresponding empirical cdf. Suppose the distribution F from which we are sampling is symmetric with center of symmetry θ. Then natural estimates of θ are given by estimators θ_n minimizing (1.1) or (1.2).

Schuster and Narvarte (1973) have shown that (1.1) is minimized over all θ by any

$$\theta_n \in [m(L), M(L)], \tag{2.1}$$

where for $k = 0, 1, \dots, n-1$ ($[x]$ is the greatest integer $\le x$),

$$m(k) = \max\{(X_i + X_j)/2 \colon 1 \le i \le [(n-k+1)/2],\ j = n-k+1-i\},$$

$$M(k) = \min\{(X_i + X_j)/2 \colon k+1 \le i \le [(n+k+1)/2],\ j = n+k+1-i\},$$

and

$$L = \min\{k \colon m(k) \le M(k)\}.$$

The natural unbiased estimator of θ in this case is the SN (Schuster and Narvarte) estimator,

$$\theta_n = [m(L) + M(L)]/2. \tag{2.2}$$

As noted by Boos (1982) and Knusel (1969), one can use the argument in the two-sample location problem given in Fine (1966) to see that the Cramér–von Mises statistic (1.2) is minimized over all θ by the HL (Hodges–Lehmann) estimator

$$\theta_n = \text{median}\{(X_i + X_j)/2, 1 \le i, j \le n\}. \tag{2.3}$$

If θ_n is a good estimator of θ and F is symmetric about θ, then a natural symmetric nonparametric estimator G_n of F is obtained from the empirical cdf F_n by reflecting the data about θ_n, i.e.,

$$G_n(x) = G_n(x; \theta_n) = \{F_n(x) + 1 - F_n((2\theta_n - x) -)\}/2 \tag{2.4}$$

is the empirical cdf of the $2n$ data points $X_1, \dots, X_n, 2\theta_n - X_1, \dots, 2\theta_n - X_n$ [see also Schuster (1975) and Hinkley (1976)]. The following two theorems indicate that the "closest" symmetric distribution to the empiric F_n in sup norm (integrated square error) is the estimator $G_n(\cdot\,; \theta_n)$, where θ_n is the SN estimator (HL estimator). These are the only two cases where we can solve explicitly for the closest symmetric distribution function to the empiric.

Let $\mathbf{G}$ be the class of all symmetric distribution functions. Then:

THEOREM 1. *Let* $h_n(\theta)$ *be defined by* (1.1). *Then*

$$\begin{aligned} h_n(\theta_n)/2 &= \sup_x |F_n(x) - G_n(x; \theta_n)| \\ &= \min\Big\{\sup_x |F_n(x) - G(x)| : G \in \mathbf{G}\Big\}, \end{aligned}$$

where θ_n *is the SN estimator of* (2.2).

THEOREM 2. *Let* $h_n(\theta)$ *be defined by* (1.2). *Then*

$$\begin{aligned} h_n(\theta_n)/2 &= \int_{-\infty}^{\infty} [F_n(x) - G_n(x; \theta_n)]^2\, dx \\ &= \min\Big\{\int_{-\infty}^{\infty} [F_n(x) - G(x)]^2\, dx : G \in \mathbf{G}\Big\}, \end{aligned}$$

where θ_n *is the HL estimator of* (2.3).

PROOF OF THEOREM 1. Let

$$\begin{aligned} G_n(x; \theta) &= \{F_n(x) + 1 - F_n((2\theta - x) -)\}/2 \\ &= \{F_n(x) + F_n^{\theta}(x)\}/2, \end{aligned}$$

and let G be a cdf that is symmetric about 0. Take $G_\theta(x) = G(x - \theta)$. Then the cdf G_θ has center of symmetry θ and, for any symmetric G and θ, we see that

$$\begin{aligned} |F_n(x) - G_n(x, \theta)| &= |\{F_n(x) - F_n^{\theta}(x)\}/2| \\ &\le \{|F_n(x) - G_\theta(x)| + |G_\theta(x) - F_n^{\theta}(x)|\}/2. \end{aligned} \tag{2.5}$$

Since F_n and G_θ are right continuous and G_θ is symmetric about θ, it follows that

$$\begin{aligned}\sup_x |F_n^\theta(x) - G_\theta(x)| &= \sup_x |1 - F_n((2\theta - x)-) - G_\theta(x)| \\ &= \sup_x |F_n((2\theta - x)-) - G_\theta((2\theta - x)-)| \\ &= \sup_x |F_n(x-) - G_\theta(x-)| \\ &= \sup_x |F_n(x) - G_\theta(x)|.\end{aligned} \tag{2.6}$$

Using (2.5) and (2.6), we see that for any G and θ

$$\begin{aligned}&\sup_x |F_n(x) - G_n(x, \theta)| \\ &\quad\leq \left\{\sup_x |F_n(x) - G_\theta(x)| + \sup_x |F_n^\theta(x) - G_\theta(x)|\right\}\Big/2 \\ &\quad= \sup_x |F_n(x) - G_\theta(x)|.\end{aligned} \tag{2.7}$$

Taking the infimum over all G and θ of both sides of (2.7), we have

$$\begin{aligned}\inf_\theta \sup_x |F_n(x) - G_n(x;\theta)| &\leq \inf_{G,\theta} \sup_x |F_n(x) - G_\theta(x)| \\ &= \inf\left\{\sup_x |F_n(x) - H(x)|: H \in \mathbf{G}\right\},\end{aligned} \tag{2.8}$$

where $\mathbf{G}$ is the class of all symmetric distributions.

Schuster and Narvarte (1973) have shown that

$$\begin{aligned}h_n(\theta) &= \sup_x |F_n(x) - F_n^\theta(x)| \\ &= 2\sup_x |F_n(x) - G_n(x;\theta)|\end{aligned}$$

is minimized over all θ at any θ_n satisfying (2.1) and hence at θ_n of (2.2) and the validity of Theorem 1 follows. □

Proof of Theorem 2. In the following, all integrals are assumed to have limits of integration from $-\infty$ to ∞. Using the same notation as in the proof of Theorem 1, we can add and subtract $G_\theta(x)$ and then use Minkowski's inequality to see that

$$\begin{aligned}h_n^{1/2}(\theta) &= \left\{\int [F_n(x) - G_n(x;\theta)]^2\, dx\right\}^{1/2} \\ &= \left\{\int [(F_n(x) - F_n^\theta(x))/2]^2\, dx\right\}^{1/2} \\ &\leq \left\{\left[\int (F_n(x) - G_\theta(x))^2\, dx\right]^{1/2} \right. \\ &\qquad \left. + \left[\int (F_n^\theta(x) - G_\theta(x))^2\, dx\right]^{1/2}\right\}\Big/2.\end{aligned} \tag{2.9}$$

Now

$$\begin{aligned}\int[F_n^\theta(x) - G_\theta(x)]^2\,dx &= \int[1 - F_n((2\theta - x) -) - G_\theta(x)]^2\,dx \\ &= \int[F_n((2\theta - x) -) - G_\theta((2\theta - x) -)]^2\,dx \\ &= \int[F_n(x -) - G_\theta(x -)]^2\,dx \\ &= \int[F_n(x) - G_\theta(x)]^2\,dx.\end{aligned} \tag{2.10}$$

But then (2.9) and (2.10) imply that for any G symmetric about 0 and any θ

$$h_n^{1/2}(\theta) \le \left\{\int[F_n(x) - G_\theta(x)]^2\,dx\right\}^{1/2}. \tag{2.11}$$

Taking the infimum with respect to G and θ over both sides of (2.11), we can proceed as in the proof of Theorem 1 using the observations in Knusel (1969) or Boos (1982) to arrive at the conclusion in Theorem 2. □

We now examine the proofs of Theorems 1 and 2 to find the key conditions to generalize these theorems to the problem of identifying the closest cdf to a given cdf in a weighted distance measure. In this direction, let $\mathbf{F}$ be the class of all cumulative distribution functions (cdf's), $\mathbf{G}$ be the subset of $\mathbf{F}$ consisting of all symmetric cdf's and let w be a nonnegative weight function, which is a function of two real arguments. For fixed a, we will use w_a to denote the weight function $w(\cdot\,; a)$ and G_a to denote a symmetric cdf with center of symmetry a. For $F \in \mathbf{F}$, F_a denotes the cdf defined by $F_a(x) = 1 - F((2a - x) -)$ for all x.

For fixed cdf $F \in \mathbf{F}$, weight function w and distance measure ρ, we define an associated minimization problem as follows:

$$\begin{gathered}\text{minimize}\\ h(a) = h(a; F, \rho, w) = \rho(F, F_a; w_a)\\ \text{over all } a.\end{gathered} \tag{2.12}$$

We say that θ is a solution to the associated minimization problem if

$$h(\theta) = \min_a h(a). \tag{2.13}$$

Definition. We say that a symmetric cdf G_θ is the closest symmetric cdf to a cdf F in distance measure ρ using weight function w if

$$\rho(F, G_\theta; w_\theta) = \min\{\rho(F, G_a; w_a) \colon G_a \in \mathbf{G}\}. \tag{2.14}$$

We assume that the distance measure ρ depends on the weight function w in such a manner that for fixed a and any $F, H, G \in \mathbf{F}$, $G_a \in \mathbf{G}$ with center of symmetry a, the following hold:

$$(2.15)\qquad \begin{aligned} &\text{(i) } 0 \le \rho(F, H; w_a) \le \rho(F, G; w_a) + \rho(G, H; w_a),\\ &\text{(ii) } \rho(\alpha F, \alpha H; w_a) = \alpha\rho(H, F; w_a), \quad \text{for any } \alpha > 0,\\ &\text{(iii) } \rho(F, H; w_a) = \Phi(|F - H|; w_a), \quad \text{for some function } \Phi,\\ &\text{(iv) } \rho(F, G_a; w_a) = \rho(F_a, G_a; w_a),\\ &\text{(v) the associated minimization problem has a solution.} \end{aligned}$$

In addition, most interesting weighted distance measures would also satisfy:

$$\begin{aligned} &\text{(vi) } \rho(F, F; w_a) = 0,\\ &\text{(vii) } \rho(F, F_a; w_a) = 0 \text{ implies } F \in \mathbf{G}. \end{aligned}$$

We have found (v) the most difficult condition to verify.

Examples of weighted distance measures satisfying these conditions when F is continuous or $F = F_n$ (the empirical cdf) are $\rho(F, H; w_a) =$

(2.16) (i) $\sup_x|[F(x) - H(x)]\, w(x; a)|$ (weighted sup norm) with $w(x; a) = [F_a^*(x)(1 - F_a^*(x))]^{-1}I(x; a)$, where F_a^* is defined in (2.17) and $I(\cdot; a)$ is the indicator function of $\{x\colon 0 < F_a^*(x) < 1\}$

(ii) $\{\int_{-\infty}^{\infty} |F(x) - H(x)|^p d[w(x; a)]\}^{1/p}$ (weighted L_p distance, $p \ge 1$), with $w(\cdot; a) = F_a^*$.

Weight functions w used with (2.16) (i) [(ii)] would normally require that $w(\cdot; a)$ be symmetric about the line $x = a$ [a point (a, C), some fixed C].

Our next theorem identifies the closest symmetric cdf in a weighted distance measure to a given theoretical or empirical cdf F. The proof of this theorem closely parallels the proofs of Theorems 1 and 2 and is omitted.

THEOREM 3. *Suppose ρ with weight function w satisfies conditions* (i)–(v) *of* (2.15) *for a fixed cdf F. Let θ be a solution to the associated minimization problem of* (2.12). *Then*

$$\rho(F, F_\theta^*; w_\theta) = \min\{\rho(F, G_a; w_a)\colon G_a \in \mathbf{G}\},$$

where

$$(2.17)\qquad F_\theta^*(x) = \{F(x) + 1 - F((2\theta - x)-)\}/2, \quad \textit{all } x.$$

REMARK 1. $\rho(F, F_\theta^*; w_\theta)$ would be a natural measure of the asymmetry of F.

REMARK 2. Symmetry of a cdf corresponds to symmetry of a function about a point (θ, C), where θ is the center of symmetry and $C = \frac{1}{2}$. Properties of a cdf are not critical in Theorems 1–3 and these theorems can easily be modified to identify the closest symmetric [about some point (θ, C)] function to an arbitrary right- (or left-) continuous function R. For example, if C is known and finite θ minimizes

$$h(a) = \sup_x |R(x) + R((2a - x) -) - C|$$

over all a, then

$$R_\theta^*(x) = \{R(x) + C - R((2\theta - x) -)\}/2$$

is a closest function in sup norm to R in the class of right-continuous functions symmetric about a point (α, C), some α.

REMARK 3. Theorems 1 and 2 give explicit formulas for the closest symmetric cdf to the empiric in sup norm and integrated square error, respectively. As indicated in Theorem 3, the main practical problem with finding the closest symmetric cdf to a given cdf F is the problem of finding the center θ minimizing (2.12). In general, the minimizing θ must be found by numerical methods. However, there are cases where one can solve explicitly for the center. For example, if $F(x) = 1 - \exp(-x/\beta)$, $x \geq 0$, is the exponential cdf, then one can show that the center of the closest symmetric cdf in sup norm to F is

$$\theta = -\beta \ln(2^{1/2} - 1).$$

In Section 3, we establish the corresponding version of Theorem 3 for probability density functions.

3. The "closest" symmetric pdf. Let $\mathbf{f}$ be the class of all pdf's, $\mathbf{g}$ be the subset of $\mathbf{f}$ consisting of all symmetric cdf's and let the nonnegative weight function w be a function of two real arguments. For fixed a, we use w_a to denote the weight function $w(\cdot; a)$, $g_a \in \mathbf{g}$ to denote a symmetric pdf with center of symmetry a, and for $f \in \mathbf{f}$, f_a denotes the pdf defined by

$$f_a(x) = f(2a - x), \tag{3.1}$$

for all x.

For fixed pdf $f \in \mathbf{f}$, weight function w and distance measure ρ, we replace $\mathbf{F}$ by $\mathbf{f}$, F by f, F_a by f_a, $\mathbf{G}$ by $\mathbf{g}$, G_θ by g_θ and G_a by g_a in (2.12)–(2.14) to define the closest symmetric pdf to the pdf f in distance measure ρ using weight function w. Here, we assume that the distance measure ρ depends on the weight function w in such a manner that (i)–(v) of (2.15) hold for the fixed pdf f when the cdf's are replaced by the corresponding pdf's. Our next theorem identifies the closest symmetric pdf to a given theoretical or empirical based pdf f. The proof of this theorem closely parallels the proofs in Section 2 and is omitted.

THEOREM 4. *Let θ be a solution to the associated minimization problem of* (2.12) *stated in terms of the pdf f, i.e., θ minimizes $h(a) = \rho(f, f_a; w_a)$ over all a. Then*

$$\rho(f, f_\theta^*; w_\theta) = \min\{\rho(f, g_a; w_a); g_a \in \mathbf{g}\},$$

where $f_\theta^*(x) = \{f(x) + f(2\theta - x)\}/2$.

REMARK 4. Symmetry of a pdf corresponds to symmetry about a line, say $x = \theta$, where θ is the center of symmetry. Theorem 4 can easily be modified to identify the closest symmetric (about a line) function to a given arbitrary function. For example, if the function r is given and finite θ minimizes

$$h(a) = \sup_x |r(x) - r(2a - x)|$$

over all a, then

$$r_\theta^*(x) = [r(x) + r(2\theta - x)]/2$$

is a closest symmetric (about a line) function in sup norm to r.

The Hellinger distance between two densities f and g is [see Beran (1977) and (1978)]

$$\rho(f, g) = \left\{\int_{-\infty}^{\infty} [f^{1/2}(x) - g^{1/2}(x)]^2 \, dx\right\}^{1/2}. \tag{3.2}$$

However, this distance measure does not satisfy condition (iii) of (2.15) in the class of densities and so Theorem 4 does not apply. As mentioned in Remark 4, the closest symmetric function to $f^{1/2}$ in integrated square error distance is the function $v_\theta^{1/2}$ defined by

$$v_\theta^{1/2}(x) = \{f^{1/2}(x) + f^{1/2}(2\theta - x)\}/2,$$

where θ minimizes $h(a) = \rho(f, f_a)$ over all a. Thus v_θ is the closest symmetric function to f in Hellinger distance. However, v_θ is not, in general, a pdf. Let $s_\theta = v_\theta / c^2$, where the constant $c = c(\theta)$ is chosen to make s_θ a pdf. Then

THEOREM 5. *s_θ is a closest symmetric pdf to f in Hellinger distance [ρ defined in* (3.2)].

PROOF. For fixed a and f, let f_a be as in (3.1) and $c = c(a, f) > 0$ be that constant that makes $h_a = [(f + f_a^{1/2})/2c]^2$ a pdf. Let g be a pdf that is symmetric about zero and let g_a be defined by $g_a(x) = g(x - a)$, i.e., g_a is symmetric about a. Then

$$\rho^2(f, g_a) = 2 - 2\int f^{1/2} g_a^{1/2} \tag{3.3}$$

and

$$\rho^2(f_a, g_a) = 2 - 2\int f_a^{1/2} g_a^{1/2}. \tag{3.4}$$

Noting the equality of the left sides of (3.3) and (3.4), we see that

$$2\rho^2(f, g_a) = 4 - 2\int g_a^{1/2}(f^{1/2} + f_a^{1/2}) = 4 - 4c(a)\int g_a^{1/2} h_a^{1/2}.$$

Thus, for fixed a and f, the symmetric pdf g_a which minimizes $\rho(f, g_a)$ is the symmetric pdf g_a which maximizes $\int g_a^{1/2}h_a^{1/2}$. But since

$$0 \le \rho^2(g_a, h_a) = 2 - 2\int g_a^{1/2}h_a^{1/2},$$

we see that

$$0 \le \int g_a^{1/2}h_a^{1/2} \le 1.$$

Hence the maximum of $\int g_a^{1/2}h_a^{1/2}$ is attained at $g_a = h_a$, i.e., for fixed a

$$\min\{\rho(f, g_a)\colon g \in \mathbf{g}_0\} = [2 - 2c(a)]^{1/2}, \tag{3.5}$$

where $\mathbf{g}_0$ is the class of densities which are symmetric about 0.

Now

$$\begin{aligned} 4c^2(a) &= \int (f^{1/2} + f_a^{1/2})^2 \\ &= 2 + 2\int f^{1/2}f_a^{1/2} \\ &= 4 - \left(2 - 2\int f^{1/2}f_a^{1/2}\right) \\ &= 4 - \rho^2(f, f_a). \end{aligned}$$

Thus the value of a which minimizes the right side of (3.5) is the value of a which minimizes $\rho(f, f_a)$. Since

$$\begin{aligned} \min\{\rho(f,g)\colon g \in \mathbf{g}\} &= \min\{\rho(f, g_a)\colon g \in \mathbf{g}_0, -\infty < a < \infty\} \\ &= \min_a \min\{\rho(f, g_a)\colon g \in \mathbf{g}_0\}, \end{aligned}$$

the validity of the theorem follows. □

REMARK 5. There are no cases where we can solve explicitly for the center θ_n of the closest symmetric density to an empirical based density estimator. Beran (1978) gives an algorithm for finding θ_n for Hellinger distance and studies properties of θ_n.

REFERENCES

BERAN, R. (1977). Minimum Hellinger distance estimates for parametric models. *Ann. Statist.* **5** 445–463.

BERAN, R. (1978). An efficient and robust adaptive estimator of location. *Ann. Statist.* **6** 292–313.

BOOS, D. D. (1982). A test for asymmetry associated with the Hodges–Lehmann estimator. *J. Amer. Statist. Assoc.* **77** 647–649.

BUTLER, C. C. (1969). A test for symmetry using the sample distribution function. *Ann. Math. Statist.* **40** 2209–2210.

FINE, T. (1966). On the Hodges and Lehmann shift estimator in the two sample problem. *Ann. Math. Statist.* **37** 1814–1818.

HINKLEY, D. (1976). On estimating a symmetric distribution. *Biometrika* **63** 680–681.

KNUSEL, L. F. (1969). Über Minimum-Distanz-Schätzungen. Ph.D. thesis, Swiss Federal Institute of Technology, Zürich, KOZIOL, J. A. (1983). Tests for symmetry about an unknown value based on the empirical distribution function. *Comm. Statist. A—Theory Methods* **12** 2823–2844.

ORLOV, A. I. (1972). On testing the symmetry of distributions. *Theory Probab. Appl.* **17** 357–361.

SCHUSTER, E. F. (1975). Estimating the distribution function of a symmetric distribution. *Biometrika* **62** 631–635.

SCHUSTER, E. F. and BARKER, R. (1987). Using the bootstrap in testing symmetry vs. asymmetry. *Comm. Statist. B—Simulation Comput.* To appear.

SCHUSTER, E. F. and NARVARTE, J. A. (1973). A new nonparametric estimator of the center of a symmetric distribution. *Ann. Statist.* **1** 1096–1104.

SMIRNOV, N. V. (1947). On criteria for the symmetry of distribution laws of random variables. *Dokl. Akad. Nauk SSSR* **56** 13–16. (In Russian.)

DEPARTMENT OF MATHEMATICAL SCIENCES
UNIVERSITY OF TEXAS
EL PASO, TEXAS 79968-0514