PROBABILITY
AND
MATHEMATICAL STATISTICS

Vol. 12, Fasc. 2 (1991), pp. 265–270

MINIMIZING L_1-DISTANCE BETWEEN DISTRIBUTION FUNCTIONS

BY

EUGENE F. SCHUSTER (EL PASO, TEXAS)

Abstract. The problem addressed is that of finding the closest distribution function G in a class of distributions $\mathscr{G}$ to a given theoretical or empirical distribution function F in the L_1-norm. Applications considered are those of estimating the center of symmetry θ in the one-sample problem and in estimating the shift θ in the two-sample problem by minimizing the L_1-distance between suitably chosen empirical distribution functions. In both cases, the minimizing $\hat{\theta}$ is shown to be Galton's estimator. The closest symmetric distribution function to the empirical in L_1-norm is identified as the average of the empirical distribution function and the empirical distribution function of the data reflected about Galton's estimator. The minimizing techniques employed can be used to give new proofs of the corresponding results for the L_2-norm where the minimizing $\hat{\theta}$ is the Hodges–Lehmann estimator.

1. Introduction and summary. In this paper we consider the problem of identifying the closest distribution function G in L_1-norm ϱ to a given distribution function (cdf) F when G ranges over a class of distribution functions $\mathscr{G}$. In Lemma 2.1 of Section 2, we interpret $\varrho(F, G)$ as an expectation $\mathrm{E}|X-Y|$, where X and Y are jointly distributed random variables having marginal cdf's F and G, respectively, and joint distribution identical to that of $(F^{-1}(U), G^{-1}(U))$, where U is uniform over the interval (0, 1). We use this representation in Section 2 to show that Galton's estimator (defined in Theorems 2.2 and 2.4) minimizes the L_1-distance between suitably chosen empirical distribution functions in estimating the center of symmetry in the one-sample problem and in estimating the shift in the two-sample problem. It then follows from [6] that the closest symmetric distribution to the empirical cdf is the average of the empirical cdf and the empirical cdf of the data reflected about Galton's estimator. In Section 3, we briefly show that the minimizing techniques employed in this paper can be used in the corresponding cases for the L_2-norm and for Hellinger distance.

In related work, Schuster [6] has explicitly identified the closest symmetric distribution to the empiric supnorm and L_2-norm. Bickel and Hodges [2] present history, properties, and asymptotic theory of Galton's estimator and test.

2. Galton's estimator minimizes L_1. Let F be a given (right continuous) theoretical or empirical cumulative distribution function and let F^{-1} be the corresponding quantile function defined on (0, 1) by $F^{-1}(u) = \inf\{x\colon F(x) \geqslant u\}$.

Let $\mathscr{G}$ be a class of distribution functions with finite means and, for $G \in \mathscr{G}$, let

$$\varrho(F, G) = \int_{-\infty}^{\infty} |F(x) - G(x)|\,dx$$

be the L_1-distance between F and G. Then:

LEMMA 2.1. $\varrho(F, G) = \mathrm{E}|X - Y|$ *for some pair of jointly distributed random variables* X *and* Y *having marginal cdf's* F *and* G, *respectively, and joint distribution identical with that of* $(F^{-1}(U), G^{-1}(U))$ *when* U *is uniform over the interval* $(0, 1)$.

Proof. As noted, for example, in [1]

$$\varrho(F, G) = \int_{-\infty}^{\infty} |F(x) - G(x)|\,dx = \int_{0}^{1} |F^{-1}(u) - G^{-1}(u)|\,du,$$

since both integrals represent the area between the graph of F and G. If U is uniform over the interval (0, 1), then the lemma follows from the well-known result that $X = F^{-1}(U)$ and $Y = G^{-1}(U)$ have distribution functions F and G, respectively. ∎

Next we will show that for some classes of distributions $\mathscr{G}$ we can use Lemma 2.1 to identify the cdf $G \in \mathscr{G}$ which minimizes $\varrho(F, G)$ over $G \in \mathscr{G}$. We first consider the problem of estimating the center of symmetry by minimizing the L_1-norm.

Let $X_1, \ldots, X_n$ be independent identically distributed (iid) r.v.'s with cdf F. Suppose the distribution F is symmetric with center θ. In this case, $2\theta - X_1, \ldots, 2\theta - X_n$ are also iid with cdf F. Let us consider estimating θ by $\hat{\theta}$, where $\hat{\theta}$ minimizes the L_1-distance between the empirical cdf's F_n and $\bar{F}_n(\cdot; a)$ based on $X_1, \ldots, X_n$ and $2a - X_1, \ldots, 2a - X_n$, respectively. Our next theorem indicates that the minimizing $\hat{\theta}$ is Galton's estimator.

Here and in Section 3, let $X_{(1)}, \ldots, X_{(n)}$ be the order statistics corresponding to $X_1, \ldots, X_n$. Then:

THEOREM 2.2. *Galton's estimator* $\hat{\theta} = \mathrm{median}\{(X_{(i)} + X_{(n+1-i)})/2\colon\ 1 \leqslant i \leqslant n\}$ *minimizes*

$$h_n(a) = \varrho(F_n, \bar{F}_n(\cdot; a)) = \int_{-\infty}^{\infty} |F_n(x) - \bar{F}_n(x; a)|\,dx \qquad \textit{over all } a.$$

Proof. Using Lemma 2.1, we see that $\varrho(F_n, \bar{F}_n(\cdot; a)) = \mathrm{E}|X - Y|$ with X and Y jointly distributed as $(F_n^{-1}(U), \bar{F}_n^{-1}(U; a))$ when F_n^{-1}, $\bar{F}_n^{-1}(\cdot; a)$ are the quantile functions corresponding to F_n and $\bar{F}_n(\cdot; a)$, respectively, and U is uniform over (0, 1). Now for $(i-1)/n < U \leqslant i/n$, $i = 1, 2, \ldots, n$, we see that $(F_n^{-1}(U), \bar{F}_n^{-1}(U; a)) = (X_{(i)}, 2a - X_{(n+1-i)})$. Thus (X, Y) is jointly discrete with distribution the same as that of an empirical cdf over the n pairs $(X_{(i)}, 2a - X_{(n+1-i)})$. It then follows that

$$\varrho(F_n, \bar{F}_n(\cdot; a)) = \mathrm{E}|X - Y| = \frac{1}{n} \sum_{i=1}^{n} |X_{(i)} - (2a - X_{(n+1-i)})| = \frac{2}{n} \sum_{i=1}^{n} |Y_{(i)} - a|,$$

where $Y_{(i)} = \{X_{(i)}+X_{(n+1-i)}\}/2$. Since the median minimizes the sum of the absolute deviations of the data from any constant a, the proof is complete. ∎

In our next theorem, we note that the closest symmetric cdf to the empirical cdf in L_1-norm is the average of the empirical cdf and the empirical cdf of the data reflected about Galton's estimator $\hat{\theta}$.

THEOREM 2.3. *$\varrho(F_n, G)$ is minimized over all symmetric distributions G by $G_n(\cdot; \theta)$, where*

$$G_n(x; \hat{\theta}) = \{F_n(x)+1-F_n((2\hat{\theta}-x)-)\}/2$$

for all x, and $\hat{\theta}$ is Galton's estimator.

Proof. The theorem follows directly from Theorem 2.2 above and Theorem 3 of [6]. ∎

Remark 1. Let I_a be the distribution function of the constant random variable which always assumes the value a and let $\mathscr{G}$ be the class of all such single point distributions. Then one can use Lemma 2.1 and the equality $I_a^{-1} = a$ to see that the usual median $\hat{\theta}$ minimizes $h_n(a) = \int_{-\infty}^{\infty} |F_n(x)-I_a(x)|\,dx$ over all a, and hence $I_{\hat{\theta}}$ is the closest cdf to F_n in the class $\mathscr{G}$.

Next we consider the problem of estimating the shift θ in the two-sample problem. In this direction, let F_n and G_m be the empirical cdf's based on two independent samples, say $X_1, \ldots, X_n$ iid as F and $Y_1, \ldots, Y_m$ iid as G, where $G(x) = F(x-\theta)$ for all x. Let $X_{(1)}, \ldots, X_{(n)}$ and $Y_{(1)}, \ldots, Y_{(m)}$ be the corresponding order statistics. Now, $G(x+\theta) = F(x)$, and so both $F_n(x)$ and $G_m(x+\theta)$ estimate $F(x)$. Thus one can estimate θ by that value of a which minimizes the L_1-distance between the empirical cdf's F_n and $G_m(\cdot+a)$. Our next theorem indicates that the minimizing value of a, say $\hat{\theta}$, is Galton's estimator of shift when $n = m$.

THEOREM 2.4. *Galton's estimator $\hat{\theta} = \mathrm{med}\{Y_{(i)}-X_{(i)}\colon 1 \leqslant i \leqslant n\}$ minimizes*

$$h_n(a) = \varrho(F_n, G_n(\cdot+a)) = \int_{-\infty}^{\infty} |F_n(x)-G_n(x+a)|\,dx \qquad \textit{over all } a.$$

Proof. Let us use the notation $G_n(\cdot; a)$ for the cdf defined by $G_n(x; a) = G_n(x+a)$ and let $G_n^{-1}(\cdot; a)$ be the quantile function corresponding to $G_n(\cdot; a)$. Then, using Lemma 2.1, we see that

$$\varrho(F_n, G_n(\cdot; a)) = \int_{-\infty}^{\infty} |F_n(x)-G_n(x+a)|\,dx = \mathrm{E}|X-Y|,$$

where (X, Y) is distributed as $(F_n^{-1}(U), G_n^{-1}(U; a))$ when U is uniformly distributed over $(0, 1)$. Thus, X and Y are jointly discrete with joint distribution given by the empirical cdf over the n pairs of order statistics $(X_{(i)}, Y_{(i)}-a)$. But then

$$\mathrm{E}|X-Y| = \frac{1}{n}\sum_{i=1}^{n} |X_{(i)}-(Y_{(i)}-a)| = \frac{1}{n}\sum_{i=1}^{n} |Y_{(i)}-X_{(i)}-a|$$

is minimized at $\mathrm{median}\{Y_{(i)}-X_{(i)}\colon 1 \leqslant i \leqslant n\}$ and the proof is complete. ∎

Suppose then that $n \neq m$. Let us define $X'_{(1)}, \ldots, X'_{(nm)}$ and $Y'_{(1)}, \ldots, Y'_{(nm)}$ as the order statistics of the nm random variables consisting of m replications of each of $X_{(1)}, \ldots, X_{(n)}$ and n replications of each of $Y_{(1)}, \ldots, Y_{(m)}$, respectively. Then:

COROLLARY 2.5. *The Galton type estimator* $\hat{\theta} = \operatorname{med}\{Y'_{(i)} - X'_{(i)}:\ 1 \leqslant i \leqslant nm\}$ *minimizes*

$$h_n(a) = \int_{-\infty}^{\infty} |F_n(x) - G_m(x+a)|\,dx \quad \textit{over all } a.$$

Proof. Using the notation as in the proof of Theorem 2.4, we take U to be uniform over (0, 1). Then for $(i-1)/nm < U \leqslant i/nm$, $i = 1, 2, \ldots, nm$, we see that $\big(F_n^{-1}(U), G_m^{-1}(U; a)\big) = (X'_{(i)}, Y'_{(i)} - a)$. Proceeding as in the proof of Theorem 2.4, the validity of the corollary easily follows. ∎

3. Minimizing L_2-norm and Hellinger distance. In this section we will sketch the proofs of the theorems of Section 2 for the L_2-norm. In this case it is known (see [3]–[5]) that the Hodge–Lehmann estimator minimizes the L_2-distance between empirical cdf's and replaces Galton's estimator in restating Theorems 2.2–2.4 for the L_2-norm. Simple proofs of these theorems follow from the following analog of Lemma 2.1 for the L_2-norm.

Let X_1, X_2 be iid as X where X has cdf F and let Y_1, Y_2 be independent of X_1, X_2 and iid as Y where Y has cdf G. Suppose X and Y have finite means. Take

$$\varrho(F, G) = \int_{-\infty}^{\infty} (F(x) - G(x))^2\,dx.$$

Then:

THEOREM 3.1. $\varrho(F, G) = \mathrm{E}|X_1 - Y_1| - \{\mathrm{E}|X_1 - X_2| + \mathrm{E}|Y_1 - Y_2|\}/2$. *Furthermore,* $\varrho(F, G) = \mathrm{E}|X_1 + X_2 - 2a| - \mathrm{E}|X_1 - X_2|$ *when* $Y \sim$ *(is distributed as)* $2a - X$ *(G reflects F about the point* $(a, 1/2)$*) and* $\varrho(F, G) = \mathrm{E}|X_1 - X_2 - a| - \mathrm{E}|X_1 - X_2|$ *when* $Y \sim X - a$ *(G shifts F by an amount a).*

Proof. For simplicity we will write $\int_a^b f(x)\,dx$ as $\int_a^b f$ in the following. Noting that $H = (F^2 + G^2)/2$ and $K = FG$ are both cdf's with $H \geqslant K$ and $K^{-1} \geqslant H^{-1}$, we can proceed as in the proof of Lemma 2.1 to see that

$$\int_{-\infty}^{\infty} (F-G)^2 = 2\int_{-\infty}^{\infty}\left(\frac{F^2+G^2}{2} - FG\right) = 2\int_{-\infty}^{\infty}\left|\frac{F^2+G^2}{2} - FG\right|$$

$$= 2\int_{-\infty}^{\infty} |H-K| = 2\int_0^1 |H^{-1} - K^{-1}| = 2\int_0^1 (K^{-1} - H^{-1}) = 2\{\mathrm{E}(Z) - \mathrm{E}(W)\},$$

where Z has cdf K and W has cdf H. Since $K = FG$, it follows that K is the cdf of $Z = \max\{X_1, Y_1\}$. Noting that $\max\{a, b\} = \{a + b + |a-b|\}/2$, we see that $2\mathrm{E}(Z) = \mathrm{E}(X_1) + \mathrm{E}(Y_1) + \mathrm{E}|X_1 - Y_1|$. Similarly, since W has cdf $H = (F^2 + G^2)/2$, it follows that

$$2\mathrm{E}(W) = \mathrm{E}\max\{X_1, X_2\} + \mathrm{E}\max\{Y_1, Y_2\}$$
$$= \mathrm{E}(X_1) + \mathrm{E}(Y_1) + (\mathrm{E}|X_1 - X_2| + \mathrm{E}|Y_1 - Y_2|)/2.$$

Thus

$$\int_{-\infty}^{\infty} (F-G)^2 = 2\{\mathrm{E}(Z) - \mathrm{E}(W)\} = \mathrm{E}|X_1 - Y_1| - \{\mathrm{E}|X_1 - X_2| + \mathrm{E}|Y_1 - Y_2|\}/2,$$

and the proof of the theorem can easily be completed by considering the two cases $Y \sim 2a - X$ and $Y \sim X - a$ separately. ∎

Remark 2. Note that, in the latter two cases of Theorem 3.1, the value of a which minimizes $h(a) = \varrho(F, G)$ is always a median.

Let F_n be the empirical cdf based on $X_1, \ldots, X_n$ iid as F, where F has center of symmetry θ, and let $\bar{F}_n(\cdot; a)$ be the empirical cdf based on $2a - X_1, \ldots, 2a - X_n$. An application of the second case of Theorem 3.1 gives the L_2-analog of Theorem 2.2:

COROLLARY 3.2. *The Hodges–Lehmann estimator* $\hat{\theta} = \mathrm{med}\{(X_{(i)} + X_{(j)})/2 : 1 \leqslant i, j \leqslant n\}$ *minimizes*

$$h_n(a) = \int_{-\infty}^{\infty} (F_n(x) - \bar{F}_n(x; a))^2 \, dx \quad \textit{over all } a.$$

In a similar fashion one can obtain:

COROLLARY 3.3. *The Hellinger distance* $h_n(a) = \int_{-\infty}^{\infty} (F_n^{1/2}(x) - \bar{F}_n^{1/2}(x; a))^2 \, dx$ *is minimized over all a at* $\hat{\theta} = \mathrm{med}\{F_n^{1/2} * F_n^{1/2}\}/2$, *where* $*$ *denotes convolution.*

Schuster [6] uses the result in Corollary 3.2 to show that the average of the empirical and the empirical reflected about the Hodges–Lehmann estimator is the closest symmetric cdf to the empirical in L_2-norm. A similar result would hold for the distance measure and estimator of Corollary 3.3.

Applying case three of Theorem 3.1, it is easily seen that the L_2-analog of Theorem 2.4 holds with the Hodges–Lehmann estimator

$$\hat{\theta} = \mathrm{med}\{Y_{(i)} - X_{(j)} : 1 \leqslant i \leqslant m,\ 1 \leqslant j \leqslant n\}$$

minimizing $h_n(a) = \int_{-\infty}^{\infty} (F_n(x) - G_m(x; a))^2 \, dx$ over all a.

In both Sections 2 and 3, one can use $h_n(\hat{\theta})$ to measure asymmetry or test for the nonparametric hypothesis of symmetry about an unknown center. However, the statistic $h_n(\hat{\theta})$ is not distribution free (see [3] and [6]). Hence, one cannot compute universal critical values or p-values for these tests. Schuster and Barker [7] sidestep this problem with a symmetric bootstrap procedure which uses bootstrap samples from the closest symmetric distribution to the empirical of the data to estimate these values.

Boos [3] attributes the first proof of Corollary 3.2 to Knüsel [5] and uses Corollary 3.2 in testing the nonparametric null hypothesis of symmetry about an unknown center θ. Fine [4] had previously proved the two-sample version of the corollary.

REFERENCES

[1] P. J. Bickel and D. A. Freedman, *Some asymptotic theory for the bootstrap*, Ann. Math. Statist. 9 (1981), pp. 73–79.

[2] P. J. Bickel and J. L. Hodges, Jr., *The asymptotic theory of Galton's test and a related simple estimate of location*, ibidem 38 (1967), pp. 73–79.

[3] D. D. Boos, *A test for asymmetry associated with the Hodges–Lehmann estimator*, J. Statist. Assoc. 77 (1982), pp. 647–649.

[4] T. Fine, *On the Hodges and Lehmann estimator for shift*, Ann. Math. Statist. 37 (1966), pp. 1814–1818.

[5] L. F. Knüsel, *Über minimum-distance Schätzungen*, Ph. D. Thesis, Swiss Federal Institute of Technology, Zurich, 1969.

[6] E. F. Schuster, *Identifying the closest symmetric distribution or density function*, Ann. Statist. 15 (1987), pp. 865–874.

[7] – and R. C. Barker, *Using the bootstrap in testing symmetry versus asymmetry*, Commun. Statist.-Simula. 16 (1) (1987), pp. 69–84.

Department of Mathematical Sciences
The University of Texas at El Paso
El Paso, Texas 79968, USA

Received on 6.3.1990