ON A KOLMOGOROV-SMIRNOV TYPE ALIGNED TEST

by

Pranab Kumar Sen

Department of Biostatistics
University of North Carolina at Chapel Hill

Institute of Statistics Mimeo Series No. 1442

June 1983

ON A KOLMOGOROV-SMIRNOV TYPE ALIGNED TEST *

PRANAB KUMAR SEN
University of North Carolina, Chapel Hill

For testing the hypothesis that two (symmetric) distributions differ only in locations, a Kolmogorov-Smirnov type test based on the aligned observations is considered and its properties studied.

1. Introduction. Let $X_1,\ldots,X_m$ be m independent and identically distributed random variables (i.i.d.r.v.) with a continuous distribution function (d.f.) F, defined on the real line $R = (-\infty,\infty)$. Also, let $Y_1,\ldots,Y_n$ be n i.i.d.r.v. with a d.f. G , defined on R. It is assumed that

(1.1) $\quad F(x) = F_o(x - \theta_1)$ and $G(x) = G_o(x - \theta_2)$, $x \in R$,

where $\underset{\sim}{\theta} = (\theta_1,\theta_2)$ is an unknown vector of the location parameters, and, we want to test for

(1.2) $\quad H_0: F_o = G_o$ against $H_1: F_o \neq G_o$,

treating $\underset{\sim}{\theta}$ as a nuisance parameter. If $\theta_1 = \theta_2$, then an omnibus goodness of fit test for (1.2) is based on the classical Kolmogorov-Smirnov statistics

(1.3) $\quad K_{mn}^{+} = \sup\{ F_m(x) - G_n(x) : x \in R \}$,

(1.4) $\quad K_{mn} = \sup\{ |F_m(x) - G_n(x)| : x \in R \}$,

where F_m and G_n are the two sample (empirical) d.f. This test is not suitable when θ_1 and θ_2 are not equal. One possibility to overcome this problem is to estimate $\underset{\sim}{\theta}$ by some convenient estimator $\hat{\underset{\sim}{\theta}}$ and to base the test on the aligned observations $X_i - \hat{\theta}_1$, $i=1,\ldots,m$ and $Y_j - \hat{\theta}_2$, $j=1,\ldots,n$. Unfortunately, the resulting test may not be distribution-free, even asymptotically. In this note, we consider

---

AMS 1980 Subject Classification : 62G10, 62G99.
Key Words & Phrases: Alignment; asymptotically distribution-free; empirical d.f.; empirical process; Kolmogorov-Smirnov statistics; tightness; weak convergence.

* Work partially supported by the National Heart, Lung and Blood Institute, Contract No. NIH-NHLBI-71-2243-L from the National Institutes of Health.

a variant form of the Kolmogorov-Smirnov statistics where the alignment procedure works out well when the underlying d.f.'s are symmetric. Along with the preliminary notions, this proposed test is considered in Section 2 and its properties are studied in the concluding section.

2. The proposed test. Let $\hat{\theta}_{1,m}$ and $\hat{\theta}_{2,n}$ be two arbitrary translation-invariant estimators of θ_1 and θ_2, respectively, such that

(2.1) $\quad m^{\frac{1}{2}} |\hat{\theta}_{1,m} - \theta_1| = O_p(1)$ and $n^{\frac{1}{2}} |\hat{\theta}_{2,n} - \theta_2| = O_p(1)$.

Consider then the aligned observations

(2.2) $\quad \hat{X}_{mi} = X_i - \hat{\theta}_{1,m}$, $i=1,\ldots,m$; $\hat{Y}_{nj} = Y_j - \hat{\theta}_{2,n}$, $j=1,\ldots,n$,

and let

(2.3) $\quad \hat{F}_m(x) = m^{-1}\Sigma_{i=1}^{m} I(\hat{X}_{mi} \leq x)$ and $\hat{G}_n(x) = n^{-1}\Sigma_{j=1}^{n} I(\hat{Y}_{nj} \leq x)$, $x \in R$

be the two empirical d.f.'s based on the aligned observations. Further, let

(2.4) $\quad \hat{F}_m^*(x) = \hat{F}_m(x) - \hat{F}_m(-x-)$ and $\hat{G}_n^*(x) = \hat{G}_n(x) - \hat{G}_n(-x-)$, $x \geq 0$.

Our proposed tests are based on the statistics

(2.5) $\quad \hat{K}_{mn}^{*+} = \sup\{ \hat{F}_m^*(x) - \hat{G}_n^*(x) : x \geq 0\}$,

(2.6) $\quad \hat{K}_{mn}^{*} = \sup\{ |\hat{F}_m^*(x) - \hat{G}_n^*(x)| : x \geq 0\}$.

We intend to show that the tests based on the statistics in (2.5)-(2.6) are asymptotically distribution-free and have the same properties as the ones based on (1.3)-(1.4), when F_o is symmetric about 0.

3. Properties of the test. Because of the translation-invariance of the estimators $\hat{\theta}_{1,m}$ and $\hat{\theta}_{2,n}$, without any loss of generality, we may take $\theta_1 = \theta_2 = 0$ and note that the residuals in (2.2) are translation invariant too. Further, we assume that there exists a positive λ_o, such that on letting $N = m+n$,

(3.1) $\quad 0 < \lambda_o \leq \lambda_N = m/N \leq 1 - \lambda_o < 1$, for all N .

Also, we assume that the d.f. F_o (and G_o) are symmetric and have uniformly continuous probability density functions (p.d.f.) f_o (and g_o) almost everywhere (a.e.). Finally, replacing $\hat{\theta}_{1,m}$ and $\hat{\theta}_{2,n}$ by θ_1 and θ_2, respectively, in (2.2), we denote the corresponding empirical d.f.s in (2.4) by F_m^* and G_n^*, respectively.

Then, we have the following

Theorem 1. Under (2.1), (3.1) and the assumed regularity conditions on F_o and G_o,

$$(3.2) \qquad \sup\{ m^{\frac{1}{2}} |\hat{F}_m^*(x) - F_m^*(x)| : x \geq 0 \} = o_p(1) ,$$

$$(3.2) \qquad \sup\{ n^{\frac{1}{2}} |\hat{G}_n^*(x) - G_n^*(x)| : x \geq 0 \} = o_p(1), \text{ as } N \to \infty.$$

Proof. We shall only prove (3.2) , and (3.3) follows precisely on the same line. Let $t_m = m^{\frac{1}{2}}(\hat{\theta}_{1,m} - \hat{\theta}_1)$. Then, by (2.2) and (2.3), for every $x \geq 0$,

$$(3.4) \qquad \hat{F}_m^*(x) = F_m(x + m^{-\frac{1}{2}} t_m) - F_m(-x + m^{-\frac{1}{2}} t_m -)$$
$$= F_m^*(x) + \{F_m(x+m^{-\frac{1}{2}}t_m) - F_m(x) - F_m(-x+m^{-\frac{1}{2}}t_m -) + F_m(-x-)\} ,$$

so that

$$(3.5) \qquad m^{\frac{1}{2}}\{\hat{F}_m^*(x) - F_m^*(x)\} = m^{\frac{1}{2}}\{ F_o(x+m^{-\frac{1}{2}}t_m) - F_o(x) - F_o(-x+m^{-\frac{1}{2}}t_m) + F_o(-x)\}$$
$$+ m^{\frac{1}{2}}\{F_m(x+m^{-\frac{1}{2}}t_m) - F_o(x+m^{-\frac{1}{2}}t_m) - F_m(x) + F_o(x)\}$$
$$- m^{\frac{1}{2}}\{F_m(-x+m^{-\frac{1}{2}}t_m) - F_o(-x+m^{-\frac{1}{2}}t_m) - F_m(-x-) + F_o(-x)\} .$$

Now, the first term on the right hand side of (3.5) is equal to $t_m\{f_o(x+am^{-\frac{1}{2}}t_m) - f_o(-x+bm^{-\frac{1}{2}}t_m)\}$, where $0 < a,b < 1$, and using the symmetry and uniform continuity of the pdf f_o, we conclude that under (2.1) (insuring the boundedness of t_m, in probability), this converges to 0, in probability, as $m \to \infty$, uniformly in $x \geq 0$. On the other hand, if we make use of the weak convergence of the empirical process $m^{\frac{1}{2}}\{F_m - F_o\}$ to a Brownian bridge , then, by the tightness part of this weak convergence, (2.1) and the uniform continuity of f_o , each of the other two terms on the right hand side of (3.5) converges to 0, in probability, as $m \to \infty$; the conclusion remains true under the sup-norm metric when we make use of the modulus of continuity for the empirical process $m^{\frac{1}{2}}(F_m - F_o)$. Hence, the proof of (3.2) is complete.

By virtue of Theorem 1, we conclude that under the hypothesis of Theorem 1,

$$(3.6) \qquad \sup\{ N^{\frac{1}{2}} | \hat{F}_m^*(x) - \hat{G}_n^*(x) - F_m^*(x) + G_n^*(x)| : x \geq 0 \} \xrightarrow{p} 0, \text{ as } N \to \infty.$$

Note that F_m^* is the empirical d.f. of the $X_i - \theta_1$ whose true d.f. is $F_o^*(x) = F_o(x) - F_o(-x)$, $x \geq 0$. Similarly, G_n^* is the empirical d.f. corresponding to the true d.f. $G_o^*(x) = G_o(x) - G_o(-x)$, $x \geq 0$. Hence, if we define

(3.7) $K_{mn}^{*+} = \sup\{ F_m^*(x)-G_n^*(x) : x \geq 0\}$ and $K_{mn}^{*} = \sup\{|F_m^*(x)-G_n^*(x)| : x \geq 0\}$,

then, these are respectively the one and two-sided Kolmogorov-Smirnov statistics for the two samples with observations $|X_i - \theta_1|$, i=1,...,m and $|Y_j-\theta_2|$, j=1,...,n. From (2.5), (2.6), (3.6) and (3.7), we conclude that under the hypothesis of Theorem 1, as $N \to \infty$,

(3.8) $N^{\frac{1}{2}}| \hat{K}_{mn}^{*+} - K_{mn}^{*+} | \xrightarrow{p} 0$ and $N^{\frac{1}{2}}| \hat{K}_{mn}^{*} - K_{mn}^{*}| \xrightarrow{p} 0.$

This exhibits the asymptotic equivalence of the proposed tests and the ones based on the statistics in (3.7). In particular, under $F_o = G_o$, F_o^* and G_o^* are also the same, and hence, we have for every $d \geq 0$,

(3.9) $\lim_{N\to\infty} P_{H_0}\{ (mn/N)^{\frac{1}{2}} K_{mn}^{*+} \geq d \} = \exp(-2d^2)$,

(3.10) $\lim_{N\to\infty} P_{H_0}\{ (mn/N)^{\frac{1}{2}} K_{mn}^{*} \geq d \} = 2\Sigma_{r=1}^{\infty}(-1)^{r-1}\exp(-2r^2d^2)$,

see for example, Hájek and Šidák (1967, p.190). Thus, under H_0: $F_o=G_o$ and the regularity conditions of Theorem 1, for $(mn/N)^{\frac{1}{2}}\hat{K}_{mn}^{*+}$ and $(mn/N)^{\frac{1}{2}}\hat{K}_{mn}^{*}$, the limiting distributions in (3.9) and (3.10) apply. Also, for contiguous alternatives, the power properties of K_{mn}^{*+} and K_{mn}^{*} are shared by the proposed tests.

In this context, we may use the sample medians for $\hat{\theta}_{1,m}$ and $\hat{\theta}_{2,n}$ and these satisfy (2.1) whenever $f_o(0)$ and $g_o(0)$ are positive. Other robust estimates of location may also be used, and, this may avoid the assumption of a finite second moment which is usually needed with the sample mean. It is also possible to improve (3.8) by using the Bahadur-Kiefer representation for the sample quantiles in (3.5) and the Komlos-Major-Tusnady strong approximation for the reduced empirical processes. However, for our specific purpose, these are not really needed.

REFERENCE

HÁJEK, J. and ŠIDÁK, Z. (1967). Theory of Rank Tests. Academic Press, New York.