Revista Colombiana de Estadística
Junio 2012, volumen 35, no. 1, pp. 1 a 14

Linearity Measures of the P-P Plot in the Two-Sample Problem

Aplicación de medidas de linealidad del gráfico P-P al problema de dos muestras

Francisco M. Ojeda[1,a], Rosalva L. Pulido[2,b], Adolfo J. Quiroz[2,3,c], Alfredo J. Ríos[1,d]

[1]Departamento de Matemáticas Puras y Aplicadas, Universidad Simón Bolívar, Caracas, Venezuela

[2]Departamento de Cómputo Científico y Estadística, Universidad Simón Bolívar, Caracas, Venezuela

[3]Departamento de Matemáticas, Universidad de Los Andes, Bogotá, Colombia

Abstract

We present a non-parametric statistic based on a linearity measure of the P-P plot for the two-sample problem by adapting a known statistic proposed for goodness of fit to a univariate parametric family. A Monte Carlo comparison is carried out to compare the method proposed with the classical Wilcoxon and Ansari-Bradley statistics and the Kolmogorov-Smirnov and Cramér-von Mises statistics the two-sample problem, showing that, for certain relevant alternatives, the proposed method offers advantages, in terms of power, over its classical counterparts. Theoretically, the consistency of the statistic proposed is studied and a Central Limit Theorem is established for its distribution.

***Key words*:** Nonparametric statistics, P-P plot, Two-sample problem.

Resumen

Se presenta un estadístico no-paramétrico para el problema de dos muestras, basado en una medida de linealidad del gráfico P-P. El estadístico propuesto es la adaptación de una idea bien conocida en la literatura en el contexto de bondad de ajuste a una familia paramétrica. Se lleva a cabo una comparación Monte Carlo con los métodos clásicos de Wilcoxon y Ansari-Bradley, Kolmogorov-Smirnov y Cramér-von Mises para el probelam de dos muestras. Dicha comparación demuestra que el método propuesto ofrece una

[a]Professor. E-mail: fojeda@usb.ve
[b]Professor. E-mail: rosalvaph@gmail.com
[c]Professor. E-mail: aj.quiroz1079@uniandes.edu.co
[d]Professor. E-mail: alfrios@usb.ve

potencia superior frente a ciertas alternativas relevantes. Desde el punto de vista teórico, se estudia la consistencia del método propuesto y se establece un Teorema del Límite Central para su distribución.

***Palabras clave*:** estadísticos no-paramétricos, gráfico P-P, problema de dos muestras.

1. Introduction

Probability plots, usually refered to as P-P plots, are, together with quantile-quantile plots, among the most commonly used tools for informal judgement of the fit of a data set to a hypothesized distribution or parametric family.

Gan & Koehler (1990) propose statistics that can be interpreted as measures of linearity of the P-P plot, for use in goodness of fit testing of univariate data sets to parametric families. They offer, as well, an interesting discussion on how the difference between a distribution and a hypothesized model will be reflected on the corresponding P-P plot. Their discussion is relevant to interpret the results in Section 3 below.

In order to describe the statistic that we will adapt to the two-sample problem, let $X_1, \ldots, X_m$ denote a univariate i.i.d. sample from a distribution that, we believe, might belong in the location-scale parametric family

$$F(\frac{x-\mu}{\sigma}), \qquad \mu \in I\!R, \sigma > 0 \tag{1}$$

for a fixed, continuous distribution F. Let $\hat{\mu}$ and $\hat{\sigma}$ be consistent estimators of μ and σ. Let $p_i = i/(n+1)$ and $Z_{(i)} = F((X_{(i)} - \hat{\mu})/\hat{\sigma})$, $i = 1, \ldots, m$. Let $\overline{Z}$ and $\overline{p}$ denote, respectively, the averages of the $Z_{(i)}$ and the p_i. Except for a squared power irrelevant in our case, one of the statistics proposed by Gan & Koehler (1990) is the following:

$$k(\widehat{X}) = \frac{\sum_{i=1}^{n}(Z_{(i)} - \overline{Z})(p_i - \overline{p})}{\left(\sum_{i=1}^{n}(Z_{(i)} - \overline{Z})^2 \sum_{i=1}^{n}(p_i - \overline{p})^2\right)^{1/2}} \tag{2}$$

Here, $\widehat{X}$ denotes the X sample. The p_i's used above, are the expected values, when we assume that the X_i has a fully specified distribution given by (1), of the transformed order statistics $F((X_{(i)} - \mu)/\sigma)$. Different possibilities for the plotting positions to be used in P-P plots (that is, for the choice of p_i's) are discussed in Kimball (1960). $k(\widehat{X})$ measures the linear correlation between the vectors $(Z_{(i)})_{i \leq n}$ and $(p_i)_{i \leq n}$, which should be high (close to 1) under the null hypothesis. In their paper, Gan & Koehler study some of the properties of $k(\widehat{X})$, obtain approximate (Monte Carlo) quantiles and, by simulation, perform a power comparison with other univariate goodness of fit procedures, including the Anderson-Darling statistic.

In order to adapt the statistic just described to the two-sample problem, one can apply the empirical c.d.f. of one sample to the ordered statistics of the other,

and substitute the values obtained for the Z_i's in formula (2). How this can be done to obtain a fully non-parametric procedure for the univariate two-sample problem is discussed in Section 2, where we consider, as well, the consistency of the proposed statistic and establish a Central Limit Theorem for its distribution. In Section 3, a Monte Carlo study is presented that investigates the convergence of the finite sample quantiles of our statistic to their limiting values and compares, in terms of power, the proposed method with the classical Wilcoxon and Ansari-Bradley statistics for the two-sample problem.

2. Measures of Linearity for the Two-sample Problem

We will consider the non-parametric adaptation of the statistic of Gan & Koehler (1990), described above, to the univariate two-sample problem. In this setting we have two i.i.d. samples: $X_1, \ldots, X_m$, produced by the continuous distribution $F(x)$ and $Y_1, \ldots, Y_n$, coming from the continuos distribution $G(y)$. These samples will be denoted $\widehat{X}$ and $\widehat{Y}$, respectively. Our null hypothesis is $F = G$ and the general alternative of interest is $F \neq G$. Let $F_m(\cdot)$ denote the empirical cumulative distribution function (c.d.f.) of the X sample. By the classical Glivenko-Cantelli Theorem, as m grows, F_m becomes an approximation to F and, under our null hypothesis, to G. Therefore, if we apply F_m to the ordered statistics for the Y sample, $Y_{(1)}, \ldots, Y_{(n)}$, we will obtain, approximately (see below), the beta distributed variables whose expected values are the p_i of Gan and Koehler's statistics. Thus, the statistic that we will consider for the two-sample problem is

$$\eta(\widehat{X}, \widehat{Y}) = \frac{\sum_{i=1}^{n}(Z_{(i)} - \overline{Z})(p_i - \overline{p})}{\left(\sum_{i=1}^{n}(Z_{(i)} - \overline{Z})^2 \sum_{i=1}^{n}(p_i - \overline{p})^2\right)^{1/2}} \tag{3}$$

with $Z_{(i)} = F_m(Y_{(i)})$. Our first theoretical result is that $\eta(\cdot, \cdot)$, indeed, produces a non-parametric procedure for the two-sample problem.

Theorem 1. *Under the null hypothesis, the statistic $\eta(\widehat{X}, \widehat{Y})$, just defined, is distribution free (non-parametric), for the two-sample problem, over the class of i.i.d. samples from continuous distributions.*

Proof. The argument follows the idea of the proof of Theorem 11.4.3 in Randles & Wolfe (1979). Since the p_i are constants, $\eta(\widehat{X}, \widehat{Y})$ is a function only of the vector $(F_m(Y_1), F_m(Y_2), \ldots, F_m(Y_n))$ only. Thus, it is enough to show that the distribution of this vector does not depend on F under the null hypothesis. Now, for $i_1, i_2, \ldots, i_n$ in $\{0, 1, \ldots, m\}$, we have, by definition of F_m,

$$\begin{aligned}&\Pr(F_m(Y_1) = i_1/m, F_m(Y_2) = i_2/m, \ldots, F_m(Y_n) = i_n/m) = \\ &\Pr(X_{(i_1)} \leq Y_1 < X_{(i_1+1)}, X_{(i_2)} \leq Y_2 < X_{(i_2+1)}, \ldots X_{(i_n)} \leq Y_n < X_{(i_n+1)}),\end{aligned} \tag{4}$$

where, if $i_j = 0$, $X_{(0)}$ must be taken as $-\infty$ and, similarly, if $i_j = m$, $X_{(m+1)}$ must be understood as $+\infty$. Consider the variables $U_i = F(X_i)$, for $i \leq m$ and

$V_j = F(Y_j)$, for $j \leq n$. Under the null hypothesis, the U_i and V_j are i.i.d. Unif(0,1) and, since F is non-decreasing, the probability in (4) equals

$$\Pr(U_{(i_1)} \leq V_1 < U_{(i_1+1)}, U_{(i_2)} \leq V_2 < U_{(i_2+1)}, \ldots U_{(i_n)} \leq V_n < U_{(i_n+1)})$$

which depends only on i.i.d. uniform variables, finishing the proof. □

Theorem 11.4.4 in Randles & Wolfe (1979) identifies the distribution of $F_m(Y_{(i)})$ as the inverse hypergeometric distribution whose properties were studied in Guenther (1975). The study of these results in Randles & Wolfe (1979) is motivated by the consideration of the exceedance statistics of Mathisen (1943) for the two-sample problem.

Theorem 1 allows us to obtain generally valid approximate null quantiles to the distribution of $\eta(\widehat{X}, \widehat{Y})$, in the two-sample setting, by doing simulations in just one case: $F = G =$ the Unif(0,1) distribution.

We will now study the consistency of $\eta(\widehat{X}, \widehat{Y})$ (and a symmetrized version of it) as a statistic for the two sample problem. We begin by establishing a Strong Law of Large Numbers (SLLN) for $\eta(\widehat{X}, \widehat{Y})$.

Theorem 2. *Suppose that F and G are continuous distributions on $\mathbb{R}$. Then, as m and n go to infinity, $\eta(\widehat{X}, \widehat{Y}) \to cor(F(Y), G(Y))$, almost sure (a.s.), where Y has distribution G and 'cor' stands for 'correlation'.*

Proof. We will only verify that $\frac{1}{n}\sum_{i=1}^{n}(Z_i - \overline{Z})(p_i - \overline{p})$ converges, a.s., as $n, m \to \infty$, to $\mathrm{Cov}(F(Y), G(Y))$. The quantities in the denominator of η are studied similarly. Let $G_n(\cdot)$ denote the empirical c.d.f. associated to the Y sample and let, also, $\overline{F}_m = (1/m)\sum F_m(Y_i)$ and $\overline{G}_n = (1/n)\sum G_n(Y_i)$. Observe that $p_i = (n/(n+1))G_n(Y_{(i)})$. It follows that

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^{n}(Z_i - \overline{Z})(p_i - \overline{p}) &= \frac{1}{n+1}\sum_{i=1}^{n}(F_m(Y_{(i)}) - \overline{F}_m)(G_n(Y_{(i)}) - \overline{G}_n) \\ &= \frac{1}{n}\sum_{i=1}^{n}(F(Y_i) - \mathbb{E}\,F(Y_1))(G(Y_i) - \mathbb{E}\,G(Y_1)) + r_{m,n} \end{aligned} \tag{5}$$

Repeated application of the Glivenko-Cantelli Theorem and the SLLN shows that $r_{m,n} \to 0$, a.s., as $m, n \to \infty$, finishing the proof. □

According to Theorem 2, when the null hypothesis: $F = G$ holds, $\eta(\widehat{X}, \widehat{Y})$ will converge to 1. In order to have consistency of the corresponding statistic for the two-sample problem, we would like to have the reciprocal of this statement to hold: If $F \neq G$ then the limit of $\eta(\widehat{X}, \widehat{Y})$ is strictly less than one. Unfortunately, this is not the case, as the following example shows.

Example 1. Let F and G be the Unif(0,2) distribution and the Unif(0,1) distribution, respectively. Then, $\mathrm{cor}(F(Y), G(Y)) = 1$ and, therefore, $\eta(\widehat{X}, \widehat{Y})$ applied to samples from F and G will converge to 1.

The counter-example just given suggests the consideration of a 'symmetrized' version of η in order to attain consistency of the statistic against the general alternative $F \neq G$. For this purpose, one could define

$$\eta_{\text{symm}} = \frac{1}{2}(\eta(\widehat{X}, \widehat{Y}) + \eta(\widehat{Y}, \widehat{X})) \tag{6}$$

For η_{symm}, we have the following result.

Theorem 3. *Let the X and Y samples be obtained from the continuous distributions F and G with densities f and g, respectively, such that the sets $\mathcal{S}_f = \{x : f(x) > 0\}$ and $\mathcal{S}_g = \{x : g(x) > 0\}$ are open. Then, η_{symm} converges to 1, a.s., as $n, m \to \infty$ if, and only if, $F = G$.*

Proof. In view of Theorem 2, we only need to show that, if $F \neq G$, then either $\text{corr}(F(Y), G(Y)) \neq 1$ or $\text{corr}(F(X), G(X)) \neq 1$, where the variables X and Y have distributions F and G, respectively. Let λ denote Lebesgue measure in $\mathbb{R}$. Suppose first that $\lambda(\mathcal{S}_g \setminus \mathcal{S}_f) > 0$. Then, there is an interval $J \subset \mathbb{R}$ such that $g(x) > 0$ for $x \in J$, while $f(x) \equiv 0$ on J. Suppose $\text{corr}(F(Y), G(Y)) = 1$. Then, there are constants a and b, with $a \neq 0$ such that, with probability 1, $G(Y) = a\,F(Y) + b$. By the continuity of the distributions and the fact that g is positive on J, it follows that

$$G(y) = a\,F(y) + b, \text{ for all } y \in J \tag{7}$$

Taking derivatives on both sides, we have, for all $y \in J$,

$$0 < g(y) = a\,f(y) = 0$$

a contradiction. The case $\lambda(\mathcal{S}_f \setminus \mathcal{S}_g) > 0$ is treated similarly.

It only remains to consider the case when $\lambda(\mathcal{S}_f \,\Delta\, \mathcal{S}_g) = 0$, where Δ denotes "symmetric difference" of sets. In this case we will show that $\text{corr}(F(Y), G(Y)) = 1$ implies $F = G$. Suppose that $\text{corr}(F(Y), G(Y)) = 1$. For J any open interval contained in $\mathcal{S}_g$, we have, by the argument of the previous case, $g(x) = a\,f(x)$ in J. Since $\mathcal{S}_g$ is open, it follows that $a\,f$ and g coincide on $\mathcal{S}_g$. Since $\lambda(\mathcal{S}_f \,\Delta\, \mathcal{S}_g) = 0$ and f and g are probability densities, a must be 1 and $F = G$, as desired. □

The result in Theorem 3 establishes the consistency of η_{symm} against general alternatives, and is, therefore, satisfactory from the theoretical viewpoint. According to the results given so far in this section, η would fail to be consistent only in the case when one of the supports of the distributions considered is strictly contained in the other and, in the smaller support, the densities f and g are proportional, which is a very uncommon situation in statistical practice. Therefore, we feel that, in practice, both the statistics η and η_{symm} can be employed with similar expectations for their performances. The results from the power analysis in Section 3 support this belief, since the power numbers for both statistics considered tend to be similar, with a slight superiority of η_{symm} in some instances.

The purpose of next theorem is to show that an appropriate standarization of the statistic η has a limiting Gaussian distribution, as m and n tend to infinite.

This will allow the user to employ the Normal approximation for large enough sample sizes. Of course, for smaller sample sizes the user can always employ Monte Carlo quantiles for η, which are fairly easy to generate according to Theorem 1. Some of these quantiles appear in the tables presented in Section 3.

Theorem 4. *Suppose that the X and Y samples, of size m and n, respectively, are obtained from the continuous distribution F (=G). Let $N = m + n$ and suppose that $N \to \infty$ in such a way that $m/N \to \alpha$, with $0 < \alpha < 1$ (the "standard" conditions in the two-sample setting). Let $\xi_{1,0} = 0.00138\overline{8}$ and $\xi_{0,1} = 0.00\overline{5}/36$, where the bar over a digit means that this digit is to be repeated indefinitely. Let*

$$D = D(\widehat{X}, \widehat{Y}) = \frac{1}{n}\left(\sum_{i=1}^{n}(Z_{(i)} - \overline{Z})^2 \sum_{i=1}^{n}(p_i - \overline{p})^2\right)^{1/2}$$

$D(\widehat{X}, \widehat{Y})$ is the denominator of $\eta(\widehat{X}, \widehat{Y})$ after division by n. Then, as $N \to \infty$, the distribution of

$$W = W(\widehat{X}, \widehat{Y}) = \sqrt{N}\left(\eta(\widehat{X}, \widehat{Y}) - \frac{1}{12D}\right) \tag{8}$$

converges to a Gaussian distribution with mean 0 and variance

$$\sigma_W^2 = 144 \times \left(\frac{\xi_{1,0}}{\alpha} + \frac{9\,\xi_{0,1}}{1-\alpha}\right) \tag{9}$$

Proof. Let $C = C(\widehat{X}, \widehat{Y}) = \frac{1}{n}\sum_{i=1}^{n}(Z_{(i)} - \overline{Z})(p_i - \overline{p})$. C is the numerator of $\eta(\widehat{X}, \widehat{Y})$ after division by n. The idea of the proof is to show that, essentially, C is a two sample V-statistic of degrees (1,3), and then to use the classical Central Limit Theorem for V-statistics which, in the present case, gives the same limit distribution of the corresponding U-statistic. Then the result will follow by observing that D satisfies a Law of Large Numbers. □

Using, as in Theorem 2, that $p_i = G_n(Y_{(i)})$, we can show that, with probability one (ignoring ties between sample points, which have probability zero)

$$C = \frac{1}{m\,n^2(n+1)} \sum_{j,i,k,r} \mathbb{1}_{\{X_j < Y_i, Y_k < Y_i\}} - \mathbb{1}_{\{X_j < Y_i, Y_k < Y_r\}} \tag{10}$$

where, j goes from 1 to m, while i, k and r range from 1 to n. Thus, except for an irrelevant multiplying factor of $n/(n+1)$, C is the V-statistic associated to the kernel

$$h^*(X; Y_1, Y_2, Y_3) = \mathbb{1}_{\{X < Y_1, Y_2 < Y_1\}} - \mathbb{1}_{\{X < Y_1, Y_2 < Y_3\}} \tag{11}$$

The symmetric version of this kernel is

$$h(X; Y_1, Y_2, Y_3) = \frac{1}{6}\sum_{\tau} \mathbb{1}_{\{X < Y_{\tau(1)}, Y_{\tau(2)} < Y_{\tau(1)}\}} - \mathbb{1}_{\{X < Y_{\tau(1)}, Y_{\tau(2)} < Y_{\tau(3)}\}} \tag{12}$$

where τ runs over the permutations of $\{1, 2, 3\}$. It is easy to see that, under the null hypothesis, the expected value of $h(X; Y_1, Y_2, Y_3)$ is $\gamma = 1/12$. By the two-sample version of the Lemma in Section 5.7.3 of Serfling (1980), it follows that the limiting distribution of C, after standardization, is the same as that for the corresponding U-statistic, for which the sum in (10) runs only over distinct indices i, j and k. Then, according to Theorem 3.4.13 in Randles & Wolfe (1979), $\sqrt{N}(C - \gamma)$ converges, in distribution, to a zero mean Normal distribution, with variance

$$\sigma_C^2 = \frac{\xi_{1,0}}{\alpha} + \frac{9\,\xi_{0,1}}{1-\alpha}$$

where

$$\begin{aligned} \xi_{1,0} = &\ \mathrm{Cov}(h(X; Y_1, Y_2, Y_3), h(X; Y_1', Y_2', Y_3')) \quad \text{while} \\ \xi_{0,1} = &\ \mathrm{Cov}(h(X; Y_1, Y_2, Y_3), h(X'; Y_1, Y_2', Y_3')) \end{aligned}$$

for i.i.d. $X, Y_1, Y_2, Y_3, X', Y_1'$, Y_2' and Y_3' with distribution F. These covariances depend on the probabilities of certain sets of inequalities between the variables involved. Since the vector of ranks of the variables involved has the uniform distribution on the set $\mathcal{S}_7$ of permutations of seven elements, the required probabilities can be computed by inspection on $\mathcal{S}_7$ (with the help of an *ad hoc* computer program), to obtain the numbers given in the statement of the Theorem.

On the other hand, under the null hypothesis, using that $F(Y_i)$ has the U(0,1) distribution, and following the procedure in the proof of Theorem 2, one can check that both $(1/n)\sum_{i=1}^n (Z_{(i)} - \overline{Z})^2$ and $(1/n)\sum_{i=1}^n (p_i - \overline{p})^2$ converge, a.s. to $1/12$. It follows that $D(\widehat{X}, \widehat{Y})$ converges, in probability, to $1/12$. Then, Theorem 2.4 follows from an application of Slutsky's Theorem.

For small values of m and n, the distribution of W in (8) displays a negative skewness, that makes inadequate the use of the Gaussian approximation given by Theorem 4. Figure 1 displays the histogram of a sample of 10,000 values of W obtained from simulated X and Y samples of size 500 ($m = n = 500$) from the Unif(0,1) distribution. We see that for these sample sizes, the distribution of W, displayed in Figure 1, is near the bell shape of the Gaussian family. For this combination of m and n, the asymptotic variance of W, given by (9), is $\sigma_W^2 = 0.8$. Figure 2 shows the P-P plot obtained by applying the N(0,0.8) cummulative distribution function to the order statistics of the W sample and plotting these against the plotting positions, p_i. The closeness to a 45° straight line suggests that the Gaussian approximation is valid for this combination of m and n. We conclude that, when the smaller of m and n is, at least, 500, the Gaussian approximation given by Theorem 4 can be used for the distribution of $\eta(\widehat{X}, \widehat{Y})$, rejecting the null hypothesis when W falls below a prescribed quantile, say 5%, of the N(0,σ_W^2) distribution.

3. Monte Carlo Evaluation of $\eta(\widehat{X}, \widehat{Y})$

All the simulations described here were programmed using the R Statistical Language (see R Development Core Team 2011) on a laptop computer. Tables 1

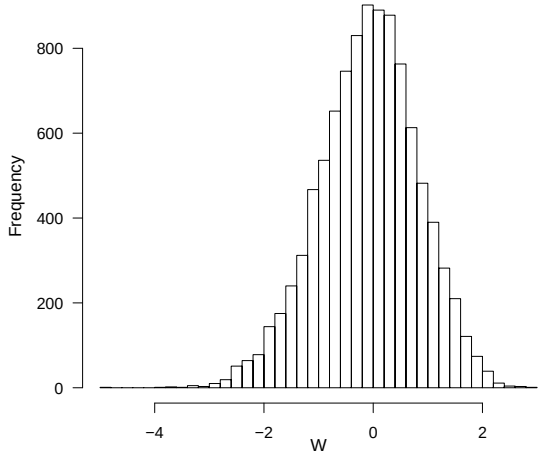

FIGURE 1: Histogram of W for $m = n = 500$.

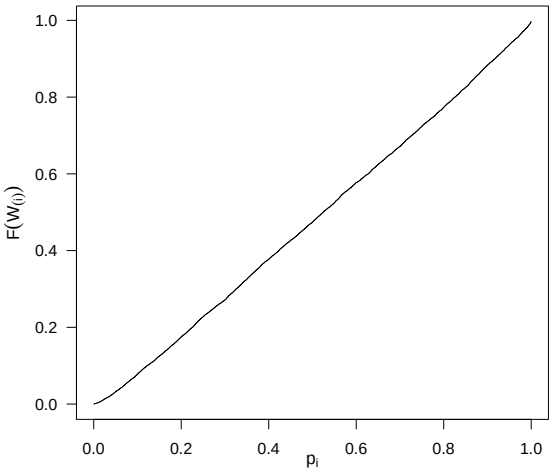

FIGURE 2: P-P plot of W for $m = n = 500$.

and 2 display Monte Carlo null quantiles for the statistics η and η_{symm}, obtained from 10,000 independent pairs of samples for each choice of m and n, using, without loss of generality, data with the Unif(0,1) distribution. Table 2 contains entries for sample size pairs of the form $m \leq n$ only, since, by the symmetry of the statistic, the quantiles are the same when the roles of m and n are interchanged. We see in these tables the convergence towards 1 of all quantiles, as m and n grow, as predicted by Theorem 3. We see, as well, that the quantiles are very similar for both statistics.

In order to evaluate the performance of η and η_{symm} as test statistics for the null hypothesis of equality of distributions, we will consider their power against different alternatives, in comparison to the classical non-parametric tests of Wilcoxon and

TABLE 1: Monte Carlo null quantiles for $\eta(\widehat{X}, \widehat{Y})$.

m	n	1%	2.5%	5%	10%
25	25	0.8956	0.9137	0.9290	0.9436
25	50	0.9203	0.9371	0.9469	0.9576
50	25	0.9235	0.9365	0.9472	0.9578
25	100	0.9363	0.9466	0.9555	0.9646
100	25	0.9360	0.9479	0.9569	0.9656
50	50	0.9471	0.9572	0.9644	0.9715
50	100	0.9624	0.9682	0.9740	0.9788
100	50	0.9598	0.9680	0.9735	0.9786
100	100	0.9744	0.9787	0.9822	0.9858

TABLE 2: Monte Carlo null quantiles for η_{symm}.

m	n	1%	2.5%	5%	10%
25	25	0.8969	0.9171	0.9313	0.9441
25	50	0.9248	0.9374	0.9482	0.9584
25	100	0.9348	0.9474	0.9565	0.9652
50	50	0.9483	0.9581	0.9649	0.9720
50	100	0.9602	0.9682	0.9738	0.9791
100	100	0.9743	0.9790	0.9823	0.9857

Ansari-Bradley, described, for instance, in Hollander & Wolfe (1999). Wilcoxon's test is specifically aimed at detecting differences in location while the statistic of Ansari-Bradley is designed to discover differences in scale. We will also include in our comparison two of the classical tests based on the empirical distribution function (EDF), namely, the two-sample versions of the Kolmogorov-Smirnov and Cramér-von Mises statistics, which are consistent against arbitrary differences in the distribution functions of the samples. These EDF statistics are described in Darling (1957). We will use the particular implementation of the Cramér-von Mises statistic studied by Anderson (1962). As alternatives, we include first the classical scenarios of difference in mean and difference in scale, between Gaussian populations. More precisely, in our first alternative, denoted Δ-mean in the tables below, the sample $\widehat{X}$ has a $N(0,1)$ distribution and $\widehat{Y}$ has the $N(0.4,1)$ distribution, while for our second alternative, denoted Δ-scale in the tables, $\widehat{X}$ has the $N(0,1)$ distribution and $\widehat{Y}$ has a normal distribution with mean zero and variance $\sigma_Y^2 = 3$. Our remaining alternatives seek to explore the advantages of η and η_{symm} when the X and Y distributions have the same mean and variance, but differ in their shape. The Weibull distribution, as described in Johnson, Kotz & Balakrishnan (1995), Chapter 21, with shape parameter $a = 1.45$ and scale parameter $b = 2.23$, has mean and variance both nearly 2.0, and exhibits right skewness. For our third alternative, denoted Gaussian vs. right-skewed, the sample $\widehat{X}$ has the $N(2,2)$ distribution, while $\widehat{Y}$ has the Weibull distribution with parameters (1.45,2.23). In order to produce a distribution with mean and variance equal 2

and left skewness, we take $X = 4 - Z$, where Z has the Gamma distribution with shape parameter $a = 2$ and scale $s = 1$. In our fourth scenario, denoted left-skewed vs. Gaussian, the sample $\widehat{X}$ comes from the distribution just described, while $\widehat{Y}$ has the $N(2, 2)$ distribution. Finally, we consider the situation of right skewness vs. left skewness, in which $\widehat{X}$ comes from the Weibull(1.45,2.23) distribution, while $\widehat{Y}$ is distributed as $4 - Z$, with $Z \sim$ Gamma(2,1).

Tables 3 to 7 display, as percentages, the power against the alternatives just described, of the six statistics compared, namely, Wilcoxon (W), Ansari-Bradley (AB), Kolmogorov-Smirnov (KS), Cramér-von Mises (CvM), η, and η_{symm}, at level 10%. The power is computed based on 1,000 independent pair of samples for each m and n combination with the given alternative distributions, using as reference the 10% quantiles given in Tables 1 and 2 for η and η_{symm}.

Table 3: Monte Carlo power against Δ-mean at level 10%.

m	n	W	AB	KS	CvM	η	η_{symm}
25	25	38.5	8.5	32.4	36.1	22.8	23.5
25	50	47.9	10.0	43.7	45.0	29.5	27.0
50	25	49.3	10.6	42.9	44.1	24.3	28.1
50	50	63.9	10.1	58.3	61.5	36.2	39.8
50	100	73.4	9.8	65.3	70.0	43.2	44.6
100	50	72.2	9.4	63.0	69.7	44.2	43.8
100	100	87.3	10.2	80.7	85.3	55.5	56.1

Table 4: Monte Carlo power against Δ-scale at level 10%.

m	n	W	AB	KS	CvM	η	η_{symm}
25	25	10.9	66.6	25.4	24.0	13.9	22.3
25	50	7.9	77.2	33.2	28.9	13.9	22.1
50	25	14.7	76.1	39.7	32.0	21.8	29.2
50	50	6.3	88.0	47.5	50.0	27.4	35.6
50	100	8.1	96.2	56.4	62.9	36.1	34.9
100	50	13.1	95.1	56.7	64.8	42.7	45.6
100	100	11.5	99.2	77.6	85.5	61.7	56.1

In Table 3 we see, as expected, that for the shift in mean scenario, the Wilcoxon test has the best performance, followed by the KS and CvM statistics. In this case the performances of η and η_{symm} are similar and inferior to that of the EDF statistics, while the Ansari-Bradley statistic has practically no power beyond the test level against the location alternative. The situation depicted in Table 4 (shift in scale) is similar, but now the Ansari-Bradley statistic is the one displaying the best power by far, followed by KS, CvM, η_{symm}, and η, in that order, while the Wilcoxon test shows basically no power against this alternative, as should be expected.

TABLE 5: Monte Carlo power for Gaussian vs. right-skewed at level 10%.

m	n	W	AB	KS	CvM	η	η_{symm}
25	25	9.4	10.3	16.0	14.3	23.5	22.3
25	50	10.9	10.9	18.5	14.8	28.8	29.6
50	25	9.9	12.9	18.8	14.6	25.9	27.6
50	50	11.9	10.8	19.6	19.1	35.3	35.3
50	100	11.8	10.5	24.5	22.5	41.5	42.8
100	50	13.3	13.7	23.0	22.1	41.8	43.9
100	100	14.3	14.1	27.6	24.4	55.8	53.2

TABLE 6: Monte Carlo power for left-skewed vs. Gaussian at level 10%.

m	n	W	AB	KS	CvM	η	η_{symm}
25	25	12.9	13.2	18.2	17.5	23.9	27.4
25	50	15.3	13.5	22.9	18.7	28.1	33.2
50	25	11.5	15.9	20.7	15.8	30.6	33.7
50	50	16.6	16.0	25.1	23.2	39.5	42.0
50	100	18.2	15.8	28.4	25.7	46.7	53.8
100	50	14.9	18.9	30.2	27.5	52.9	53.9
100	100	19.6	18.9	36.4	35.4	66.7	65.4

TABLE 7: Monte Carlo power for right-skewed vs. left-skewed at level 10%.

m	n	W	AB	KS	CvM	η	η_{symm}
25	25	17.7	14.7	31.5	28.7	53.7	54.1
25	50	22.4	15.4	43.3	38.3	69.1	70.5
50	25	20.5	15.2	43.9	38.1	65.4	70.9
50	50	25.9	15.0	50.4	48.4	84.5	85.2
50	100	30.7	15.8	60.2	60.8	92.6	93.0
100	50	27.8	17.7	60.3	61.7	93.2	92.0
100	100	38.2	15.4	80.5	83.2	98.7	98.8

In Tables 5, 6 and 7, in which the distributions considered have the same mean and variance, with differences in their skewness, the results change significantly respect to the previous tables. In these scenarios, the best power clearly corresponds to η_{symm} and η, which for some of the sample sizes nearly double the power of the KS and CvM statistics, which come next in power after η_{symm} and η. In order to understand why the proposed statistics enjoy such good power in the "difference in skewness" scenarios, the reader is advised to see Section 2 in Gan & Koehler (1990), where through several examples (and figures) it is shown the marked departure from linearity that differences in skewness can produce on a P-P plot.

From the power results above, we conclude that η and η_{symm} can be considered a useful non-parametric statistic for the null hypothesis of equality of distributions,

and its application can be recommended specially when differences in shape between F and G are suspected, instead of differences in mean or scale. The power of the two statistics studied here tends to be similar, with η_{symm} being slightly superior in some cases.

We finish this section with the application of our statistic to a real data set. For this purpose, we consider the well known drilling data of Penner & Watts (1991), that has been used as illustrative example of a two-sample data set in Hand, Daly, Lunn, McConway & Ostrowski (1994) and Dekking, Kraaikamp, Lopuhaa & Meester (2005). In these data, the times (in hundreths of a minute) for drilling 5 feet holes in rock were measured under two different conditions: *wet drilling*, in which cuttings are flushed with water, and *dry drilling*, in which cuttings are flushed with compressed air. Each drilling time to be used in our analysis is actually the average of three measures performed at the same depth with the same method, except when some of the three values might be missing, in which case the reported value is the average of the available measurements at the given depth. The sample sizes for these data are $m = n = 80$. Figure 3 shows the P-P plot for the drilling data. In this case, in order to compare the empirical cummulative distribution for the two data sets, the plot consists of the pairs $(F_m(z), G_n(z))$, where z varies over the combined data set and F_m and G_n are, respectively, the EDFs for the dry drilling and wet drilling data. In this figure a strong departure from linearity is evident. This is due to the fact that most of the smallest drilling times correspond to dry drilling, while a majority of the largest drilling times reported correspond to wet drilling, making the plot very flat in the left half and steep in the right half.

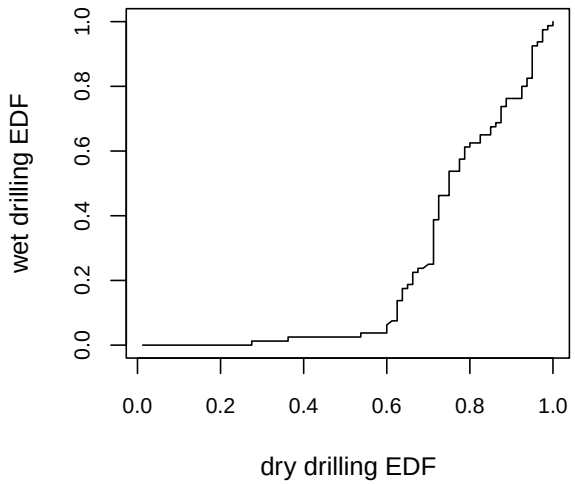

FIGURE 3: P-P Plot for dry drilling vs. wet drilling data.

In order to apply the statistic η to the drilling data, we compute first Monte Carlo null quantiles for η in the case $m = n = 80$, using, as done for Table 1,

10,000 pairs of samples of size 80 from the Unif(0,1) distribution. These quantiles turn out to be the following

1%	2.5%	5%	10%
0.9664	0.9728	0.9777	0.9821

The value of $\eta(\widehat{X}, \widehat{Y})$, taking the dry drilling data as $\widehat{X}$, is 0.9508, which is significant against the null hypothesis of equality of distributions, at the 1% level. Furthermore, comparing the actual value of $\eta(\widehat{X}, \widehat{Y})$ for the drilling data with the 10,000 values calculated for the Monte Carlo null quantile estimation, we obtain an approximate p-value for this data set of 0.0013. Thus, the evidence against equality of distribution is strong in this case.

Statistics based on ideas similar to those leading to $\eta(\widehat{X}, \widehat{Y})$ have been considered in the multivariate case by Liu, Parelius & Singh (1999), who consider statistics based on the Depth-Depth plot. Although generalization of $\eta(\widehat{X}, \widehat{Y})$ to the multivariate case is possible, we do not pursue this line of work, since in the generalization, the full non-parametric character of the statistic is lost and the computation of reference quantiles becomes computationally expensive, thus losing the ease of computation that the statistic enjoys in the univariate case.

4. Conclusions

A modified non-parametric version of the statistic proposed by Gan & Koehler (1990) for the goodness of fit of a univariate parametric family was presented based on a linearity measure of the P-P plot for the two-sample problem. A Monte Carlo comparison was carried out to compare the proposed method with the classical ones of Wilcoxon and Ansari-Bradley for the two-sample problem and the two-sample versions of the Kolmogorov-Smirnov and Cramer-von Mises statistics, showing that, for certain relevant alternatives, the method proposed offers advantages, in terms of power, over its classical counterparts. Theoretically, the consistency of the statistic proposed was studied and a Central Limit Theorem was established for its distribution.

[Recibido: febrero de 2010 — Aceptado: octubre de 2011]

References

Anderson, T. W. (1962), 'On the distribution of the two sample Cramér- von Mises criterion', *Annals of Mathematical Statistics* **33**(3), 1148–1159.

Darling, D. A. (1957), 'The Kolmogorov-Smirnov, Cramér-von Mises tests', *Annals of Mathematical Statistics* **28**(4), 823–838.

Dekking, F. M., Kraaikamp, C., Lopuhaa, H. P. & Meester, L. E. (2005), *A Modern Introduction to Probability and Statistics*, Springer-Verlag, London.

Gan, F. F. & Koehler, K. J. (1990), 'Goodness-of-fit tests based on P-P probability plots', *Technometrics* **32**(3), 289–303.

Guenther, W. C. (1975), 'The inverse hypergeometric - a useful model', *Statistica Neerlandica* **29**, 129–144.

Hand, D. J., Daly, F., Lunn, A. D., McConway, K. J. & Ostrowski, E. (1994), *A Handbook of Small Data Sets*, Chapman & Hall, Boca Raton, Florida.

Hollander, M. & Wolfe, D. A. (1999), *Nonparametric Statistical Methods*, 2 edn, John Wiley & Sons, New York.

Johnson, N. L., Kotz, S. & Balakrishnan, N. (1995), *Continuous Univariate Distributions*, 2 edn, John Wiley & Sons, New York.

Kimball, B. F. (1960), 'On the choice of plotting positions on probability paper', *Journal of the American Statistical Association* **55**, 546–560.

Liu, R. Y., Parelius, J. M. & Singh, K. (1999), 'Multivariate analysis by data depth: Descriptive statistics, graphics and inference', *The Annals of Statistics* **27**(3), 783–858.

Mathisen, H. C. (1943), 'A method for testing the hypothesis that two samples are from the same population', *The Annals of Mathematical Statistics* **14**, 188–194.

Penner, R. & Watts, D. G. (1991), 'Mining information', *The Annals of Statistics* **45**(1), 4–9.

R Development Core Team (2011), 'R: A language and environment for statistical computing'. Vienna, Austria.
*http://www.R-project.org/

Randles, R. H. & Wolfe, D. A. (1979), *Introduction to the Theory of Nonparametric Statistics*, Krieger Publishing, Malabar, Florida.

Serfling, R. J. (1980), *Approximation Theorems of Mathematical Statistics*, John Wiley and Sons, New York.