STICHTING
MATHEMATISCH CENTRUM
2e BOERHAAVESTRAAT 49
AMSTERDAM

AFDELING MATHEMATISCHE STATISTIEK

Report S 333

Asymptotic Efficiency of Certain Distance Criteria

by

F.C. Andrews[1)] and K. Jogdeo

(Preliminary Report)

1) This research was supported by the National Science Foundation under grant NSF - GP - 1643.

September 1964

Printed at the Mathematical Centre at Amsterdam,49,2nd Boerhaavestraat.
The Netherlands.

The Mathematical Centre, founded the 11th of February 1946, is a non - profit institution aiming at the promotion of pure mathematics and its applications, and is sponsored by the Netherlands Government through the Netherlands Organization for Pure Scientific Research (Z.W.O.) and the Central National Council for Applied Scientific Research in the Netherlands (T.N.O.), by the Municipality of Amsterdam and by several industries.

1. Introduction.

During the last dozen years several papers have been published each giving some information about the asymptotic behaviour of a two sample test based upon the integral of the squared difference of two empirical distribution functions. This preliminary report adds to the literature by providing mathematics necessary for the calculation of asymptotic relative efficiencies of the above test with respect to certain other tests both of parametric and nonparametric form. Computations of these asymptotic relative efficiencies are proposed and will be carried out by the authors. An observation pertaining to the powers of the two sided two sample test of Lehmann (1953) in comparison with that of the two sided Wilcoxon test is given.

Although the methods used here are similar to those used by Anderson and Darling (1952), Rosenblatt (1952) and Blum, Kiefer and Rosenblatt (1961) for purposes of exposition we find it best to give them again in some detail.

2. Asymptotic Distributions.

Suppose F and G are two continuous probability distribution functions for which corresponding independent random samples of size n are available. Let $X_1', \cdots, X_n'$ be the independent random variables of the random sample from F and $Y_1', \cdots, Y_n'$ corresponding from G. Consider the test of H_0 : F = G that rejects H_0 whenever

$$2.1 \qquad D_n = \int_{-\infty}^{\infty} \left[F_n(x) - G_n(x) \right]^2 d \frac{F_n(x)+G_n(x)}{2}$$

is found too large. Here F_n and G_n are the sample cumulative distribution functions of the respective samples i.e.

$$2.2 \qquad n F_n(x) = \text{number of } X_i' \leq x.$$

An equivalent test which differs from D_n by a linear term is based on

2.3 $$D_n' = \sum_{i=1}^{n} \left[(r_i - i)^2 + (s_i - i)^2 \right]$$

where r_i and s_i are the ranks of i^{th} largest X_i' and the i^{th} largest Y_i' among the pooled sample of 2n observations.

Rosenblatt (1952) has shown that under the null hypothesis $nD_n'/2$ has the same limiting distribution as the von Mises' statistic $n\omega_n^2$ whose limit distribution was derived and tabulated by Anderson and Darling (1952). In order to set the stage for computations of asymptotic relative efficiency of (2.1) we consider a sequence of alternative hypotheses $\{H_n\}$ where $H_n : G^{(n)}(x) = F(x+a/n^{\frac{1}{2}})$ so that under H_n the pair $(F, G^{(n)})$ is the alternative to the pair (F,F) under H_0. Since F is continuous we find it convenient to effect the probability integral transformation

2.4 $$X_i = F(X_i'), \qquad Y_i = F(Y_i'),$$

with the result that the X's are independent and uniformly distributed on (0,1) and the Y's are independent and have the distribution function $G\,F^{-1}$ on (0,1). The statistic D_n can then be written as

2.5 $$D_n = \int_0^1 \left[F_n(F^{-1}(y)) - G_n^{(n)}(F^{-1}(y)) \right]^2 d\, \frac{F_n(F^{-1}(y)) + G_n^{(n)}(F^{-1}(y))}{2}$$

where $F_n F^{-1}$ and $G_n^{(n)} F^{-1}$ are the respective sample cumulative distributions of the transformed samples.

For each real number $t \in [0,1]$ consider the random variable

2.6 $$V_n(t) = \left(\frac{n}{2}\right)^{\frac{1}{2}} \left[F_n(F^{-1}(t)) - G_n^{(n)}(F^{-1}(t)) \right].$$

The family $\{V_n(t)\}$, $t \in [0,1]$ defines a stochastic process with mean function

2.7 $$E\, V_n(t) = \left(\frac{n}{2}\right)^{\frac{1}{2}} \left[F(F^{-1}(t)) - G_n^{(n)}(F^{-1}(t)) \right].$$

and

$$2.8 \qquad \sigma(V_n(s), V_n(t)) = \tfrac{1}{2}\{\min[s,t] - st + \min\left[G_n^{(n)}(F^{-1}(t)), G_n^{(n)}(F^{-1}(s))\right] - G_n^{(n)}(F^{-1}(t))\, G_n^{(n)}(F^{-1}(s))\}.$$

If we assume that F is differentiable we find that

$$2.9 \qquad \lim_{n \to \infty} E\, V_n(t) = -\frac{a}{\sqrt{2}} F'[F^{-1}(t)],$$

and

$$2.10 \qquad k(s,t) = \lim_{n \to \infty} \sigma(V_n(s), V_n(t)) = \min(s,t) - st; \quad 0 \leq s, t \leq 1.$$

considering the above limiting covariance function as the kernel of an integral operator we can derive the corresponding eigen values and normalized eigen functions from the integral equation

$$2.11 \qquad f(s) = \lambda \int_0^1 [\min(s,t) - st]\, f(t)dt.$$

They are $\lambda_k = 1/\pi^2 k^2$ and $f_k(t) = (1/\pi k) \sin \pi kt$. Now $|k(s,t)| \leq 1$ and so by theorem 2 (S.G.Mikhlin, 1960, p.140) the system of orthogonal functions $\{f_k\}$ is complete and any square integrable function on (0,1) has a generalized Fourier expansion in terms of $\{f_k\}$ that converges pointwise to the original function. If we further assume that

$$2.12 \qquad \int_{-\infty}^{\infty} \left[F'(x)\right]^3 dx < \infty$$

then we see that $F'[F^{-1}(t)]$ is square integrable on (0,1) and hence possesses a convergent generalized Fourier expansion namely

$$2.13 \qquad F'\left[F^{-1}(t)\right] = \sum_{k=1}^{\infty} c_k f_k(t), \qquad t \in [0,1].$$

Following the construction given by Doob (1949) and Anderson and Darling (1952) we define an infinite sequence of independent random variables $X_1, X_2, \cdots$ where X_k has the $N(\mu_k, 1)$ distribution with

$$\mu_k = -\frac{a\pi k}{\sqrt{2}} c_k. \tag{2.14}$$

Defining the stochastic process

$$v(t) = \sum_{k=1}^{\infty} \frac{X_k f_k(t)}{\pi k}, \tag{2.15}$$

one readily computes the mean function to be

$$E\, v(t) = \sum_{k=1}^{\infty} \frac{\mu_k f_k(t)}{\pi k} = -\frac{a}{\sqrt{2}} F'\left[F^{-1}(t)\right] \tag{2.16}$$

and the covariance function

$$\sigma(v(s), v(t)) = \min(s,t) - st. \tag{2.17}$$

Applying a theorem of Donsker (1952) we can assert that

$$\lim_{n \to \infty} P\left[\left\{\int_0^1 (v_n(t))^2 dt\right\} \leq x\right] = P\left[\left\{\int_0^1 (v(t))^2 dt\right\} \leq x\right]. \tag{2.18}$$

The random variable $\int_0^1 (v(t))^2 dt$ can be expressed as

$$\int_0^1 \left[\sum_{k=1}^{\infty} \frac{X_k f_k(t)}{\pi k}\right]^2 dt = \sum_{k=1}^{\infty} \frac{X_k^2}{k^2 \pi^2} \tag{2.19}$$

which is a weighted sum of non-central chi-square variables each with one degree of freedom.

Finally we remark that

$$\int_{-\infty}^{\infty} n\left[F_n(x) - G_n^{(n)}(x)\right]^2 d\, \frac{F_n(x) + G_n^{(n)}(x)}{2} - \int_{-\infty}^{\infty} n\left[F_n(x) - G_n^{(n)}(x)\right] d\, \frac{F(x) + G^{(n)}(x)}{2} \longrightarrow 0 \tag{2.20}$$

in probability as $n \longrightarrow \infty$.

This can be seen by a slight extension of a lemma of Kiefer (1959, section 2). For this extension we need a boundedness condition viz.

$$2.21 \qquad \left| \frac{F(x)-G^{(n)}(x)}{1/n^{\frac{1}{2}}} \right| < K$$

uniformly in n, for some positive constant K. Assuming this, the foregoing results are summarized in the following.

Theorem 2.1. If the distribution function F possesses a density function F' which is uniformly continuous and bounded and such that $\int_{-\infty}^{\infty} (F'(x))^3 dx < \infty$ then under the sequence of hypotheses $\{F(x), F(x+a/n^{\frac{1}{2}})\}$ (distribution functions of X' and Y' respectively), the statistic $nD_n/2$ has as limit distribution the distribution of $\sum_{k=1}^{\infty} X_k^2/\pi^2 k^2$ where $X_1, X_2, \ldots, X_k, \ldots$ are independent normally distributed random variables with

$$2.22 \qquad E(X_k) = \frac{a\pi k}{\sqrt{2}} \int_0^1 F'\left[F^{-1}(t)\right] \sin(\pi k t)\,dt, \text{ and } \sigma^2(X_k) = 1.$$

Since the characteristic function of the noncentral chi-square distribution with one degree of freedom and noncentrality parameter λ^2 is

$$2.23 \qquad e^{\frac{it\lambda^2}{1-2it}} (1-2it)^{-\frac{1}{2}}$$

we have the

Corollary 2.1. Under the hypotheses of the theorem 2.1, the characteristic function of the limiting distribution of the random variable $nD_n/2$ is

$$2.24 \qquad \varphi(t) = e^{\frac{a^2}{2}\sum_{k=1}^{\infty}\frac{itc_k^2}{1-\frac{2it}{\pi^2k^2}}} \prod_{k=1}^{\infty}\left(1-\frac{2it}{\pi^2k^2}\right)^{-\frac{1}{2}}.$$

The possibility of inverting the above characteristic function in general seems to be rather remote. It might, however be possible in a few isolated cases. For instance suppose that

$$2.25 \qquad F\left[F^{-1}(t)\right] = \frac{\pi}{2}\sin\pi t; \quad t \in [0,1].$$

We then compute the generalized Fourier coefficients as $c_1 = 1$, $c_k = 0$, for $k > 1$. It is quite possible that the characteristic function in this case can be inverted using methods similar to those of Anderson and Darling (1952) since the characteristic function reduces to

$$2.26 \qquad \varphi(t) = e^{\frac{a^2}{2}\frac{it}{1-\frac{2it}{\pi^2}}} \prod_{k=1}^{\infty}\left(1-\frac{2it}{\pi^2k^2}\right)^{-\frac{1}{2}}.$$

In cases where an inversion is possible the authors propose to calculate sufficient number of percentage points of the limit distribution to enable exact asymptotic efficiency computations relative to two sample tests for which the limiting distribution is known. In other cases lower bounds to the asymptotic relative efficiencies can be obtained by using percentage points of the first few terms of the series $\sum_{k=1}^{\infty} X_k^2/\pi^2k^2$. In those instances where the first few terms do not yield specific conclusions we plan to carry out Monte Carlo estimates of the desired percentage points.

The above theory and the methodology can be extended to cover similar criteria used for other situations. We cite two examples here. First is the k-sample problem which was considered by Kiefer (1959).

The sequence of alternatives would be $F_i(x) = F(x+a_i/n^{\frac{1}{2}})$ and the limiting distribution of the criteria under such a sequence of alternatives would be weighted sum of noncentral chi-square random variables with k degrees of freedom.

The second example is that of the criteria considered by Blum, Kiefer and Rosenblatt (1961) for testing independence. Restricting to the case of bivariate distributions the test statistic takes the form

$$2.27 \qquad B_n = \iint \left[S_n(x,y) - S_n(x)S_n(y)\right]^2 dS_n(x,y)$$

where $S_n(x,y)$ is the bivariate sample cumulative distribution function and $S_n(x)$ and $S_n(y)$ corresponding to the marginals, all computed from the sample of $(X_1,Y_1),\cdots,(X_n,Y_n)$. Blum, Kiefer and Rosenblatt (1961) point out that in the case of the sequence of alternatives $F^{(n)}$, where $F^{(n)}$ are the bivariate distribution functions and such that

$$2.28 \qquad n^{\frac{1}{2}}\left[F^{(n)}(x,y) - F_1^{(n)}(x)\, F_2^{(n)}(y)\right] \longrightarrow q(x,y)$$

(finite and continuous) as $n \longrightarrow \infty$, the limiting distribution of B_n is again a weighted sum of noncentral chi-square random variables. The above mentioned authors discuss general properties of the power of the test; however, specific cases are not available. Here we consider two types of alternatives to independence.

$$2.29 \qquad \begin{aligned} X &= (1 - \theta n^{-\frac{1}{4}})\, Z_1 + \theta n^{-\frac{1}{4}}\, Z_3 \\ Y &= (1 - \theta n^{-\frac{1}{4}})\, Z_3 + \theta n^{-\frac{1}{4}}\, Z_3. \end{aligned}$$

where Z_1, Z_2, Z_3 are independent random variables. Such alternatives were considered by Bhuchongkul (1964) and are useful when there is a suspicion that the independence is disrupted by some contaminating random variable. With these alternatives it can be verified that the joint distribution of (X,Y) is

2.30 $$F_n(x,y) = \int_{-\infty}^{\infty} F_1\left(\frac{x - \theta n^{-\frac{1}{4}} z}{1 - \theta n^{-\frac{1}{4}}}\right) F_2\left(\frac{y - \theta n^{-\frac{1}{4}} z}{1 - \theta n^{-\frac{1}{4}}}\right) d F_3(z)$$

where F_1, F_2, F_3 are the distribution functions of Z_1, Z_2 and Z_3 respectively. It can be seen that

2.31 $$n^{\frac{1}{2}} \left[F^{(n)}(x,y) - F_1(x)F_2(y)\right] \longrightarrow -\theta^2 F_1'(x)F_2'(y).$$

It will be possible to compare the test based on (2.27) and the bivariate analogue of normal scores test considered by Bhuchongkul (1964). Another type of alternatives for which a class of nonparametric tests was considered by Jogdeo (1962) is the following.

2.32 $$Y = \alpha + \beta n^{-\frac{1}{2}} X + Z,$$

where X and Z are independent random variables. With this sequence of regression alternatives

2.33 $$n^{\frac{1}{2}} \left[F^n(x,y) - F_1(x)F_2(y)\right] \longrightarrow -\beta F_2'(y) \int_{-\infty}^{x} z dF_1(z),$$

and again this leads to the possibility of comparison with various nonparametric tests.

3. Remarks on Lehmann's Tests.

Lehmann (1951) considers a two sample test based on $\binom{m}{2}\binom{n}{2}$ quadruples chosen from the samples $(X_1, \cdots, X_m)$, $(Y_1, \cdots; Y_n)$. Wegner (1956) has shown that this test is equivalent to the two sample test based on D_n of section 2. Hence the discussion regarding the asymptotic efficiency of the test based on D_n directly applies to Lehmann's test.

Further, Lehmann (1953) considers problem of testing $H_{0_2} : F_0 = G_0$, against the alternative hypothesis $H_1 : F_0 = F$, $G_0 = qF + pF^2$ or $F_0 = qF + pF^2$, $G_0 = F$ where F_0 and G_0 are the distribution functions

of X and Y respectively and p + q = 1. He shows that the locally most powerful rank test for the above problem is (for equal sample sizes) to reject when

$$3.1 \qquad L_n' = \left[\sum_{i=1}^{n} s_i - \frac{n(2n+1)}{2} \right]^2 + \left[\sum_{i=1}^{n} r_i - \frac{n(2n+1)}{2} \right]^2 + \sum_{i=1}^{n} (s_i - i)^2 + \sum_{i=1}^{n} (r_i - i)^2$$

is too large. (Here r_i and s_i are the ranks of i^{th} largest X and i^{th} largest Y in the pooled sample of $(X_1, \cdots, X_n, Y_1, \cdots, Y_n)$). An equivalent test statistic which is linearly related to (3.1) can be written as

$$3.2 \qquad L_n = \left[\sum_{i=1}^{n} s_i - \frac{n(n+1)}{2} \right]^2 + \left[\sum_{i=1}^{n} r_i - \frac{n(2n+1)}{2} \right]^2 + n^3 \left[\int (F_n - G_n)^2 \, d \, \frac{F_n + G_n}{2} \right].$$

We will make two remarks regarding the above test statistic, First, although the test statistic is not equivalent to two sided Wilcoxon test, under proper normalization (division by n^3) under the sequence of hypotheses $\{G^{(n)}\}$ the term with the integral sign goes to zero and as far as the asymptotic relative efficiency is concerned the two sided Wilcoxon test and the L_n-test are equivalent. The second remark, which follows, is of more interest because of its applications. As seen above, under the hypothesis, the limiting distribution of $n D_n$ is of the same form as that of a von Mises statistic and hence it can be seen that

$$3.3 \qquad n^{\frac{1}{2}} D_n \longrightarrow 0 \quad \text{in probability, as } n \longrightarrow \infty.$$

However under a fixed alternative $F \neq G$ the statistic D_n is of the form of a Hoeffding's nonstationary U-statistic and $n^{\frac{1}{2}} D_n$ has a limiting normal distribution. Similarly for large n

3.4 $$D_n \lessapprox 0 \text{ for } F = G \text{ and } D_n \lessapprox \int (F-G)^2 dF > 0 \text{ for } F \neq G$$

and an addition of D_n to a test statistic may result in considerable increase in power with a neglible change in the probability of the error of the first kind. Especially in situation where the alternative is not very close to the null hypothesis and n is not small (for very large n any consistent test will have power close to unity and the addition of D_n is not meaningful). The authors plan to investigate applications of this property.

REFERENCES

[1] ANDERSON, T.W., and DARLING, D.A. (1952). Asymptotic Theory of Certain "Goodness of Fit" Criteria Based on Stochastic Processes. Ann.Math.Statist. 23 193-212.

[2] BHUCHONGKUL, S. (1964). A Class of Nonparametric Tests for Independence in Bivariate Populations. Ann.Math.Statist. 35 138-149.

[3] BLUM, J.R., KIEFER, J. and ROSENBLATT, M. (1961). Distribution Free Tests of Independence Based on the Sample Distribution Function. Ann.Math.Statist. 32 485-498.

[4] DOOB, J.L. (1949). Heuristic Approach to the Kolmogorov-Smirnov Theorems. Ann.Math.Statist. 20 393-403.

[5] JOGDEO, K. (1962). Nonparametric Methods for Regression. Unpublished Ph.D. Thesis, University of California, Berkeley.

[6] KIEFER, J. (1959). K-sample Analogues of the Kolmogorov-Smirnov and Cramér-v.Mises Tests. Ann.Math.Statist. 30 420-447.

[7] LEHMANN, E.L. (1951). Consistency and Unbiasedness of Certain Nonparametric Tests. Ann.Math.Statist. 22 165-179.

[8] LEHMANN, E.L. (1953). The Power of Rank Tests. Ann.Math.Statist. 24 23-44.

[9] ROSENBLATT, M. (1952). Limit Theorems Associated with Variants of the von Mises Statistics. Ann.Math.Statist. 23 617-624.

[10] WEGNER, L.H. (1956) Properties of Some Two-Sample Tests Based on a Particular Measure of Discrepancy. Ann.Math.Statist. 27 1006-1016.