The power of the one-sided and one-sample Kolmogorov-Smirnov test

by

SUNG-WOOK LEE

B.S., National Seoul University, 1956

A MASTER'S REPORT

submitted in partial fulfillment of the

requirements for the degree

MASTER OF SCIENCE

Department of Statistics

KANSAS STATE UNIVERSITY

Manhattan, Kansas

1966

Approved by:

W J Conover

Major Professor

TABLE OF CONTENTS

1. Introduction

Let the random variables $Y_i (i = 1, 2, . . . , n)$ be independent and have the same continuous distribution function $F(x)$. Let the ordered sample be represented by $X_1 \leq X_2 \leq . . . \leq X_n$. From the assumption that $F(x)$ is continuous and that the random variables are independent, it follows that the probability of any two X_i's being equal is zero.

We define the empirical distribution function $F_n(x)$ as:

$$F_n(x) = \begin{cases} 0 & \text{for } x < X_1 \\ \frac{i}{n} & \text{for } X_i \leq x < X_{i+1},\ i = 1, . . . , n-1 \\ 1 & \text{for } X_n \leq x . \end{cases}$$

It is known that the probability that the sequence $F_n(x)$ converges to $F(x)$ as $n \to \infty$, uniformly in x $(-\infty < x < +\infty)$, equals one (see Fisz, 1963, p. 391).

Let

$$D_n^+ = \sup_{-\infty<x<+\infty} [F(x) - F_n(x)] .$$

The distribution of D_n^+ was given by A. Wald and J. Wolfowitz (1939) and by Z. W. Birnbaum and F. H. Tingey (1951). The asymptotic expression for the distribution of D_n^+ was given by N . Smirnov (1939).

In the present report, the distribution of D_n^+ is studied in section 2. In section 3 the power of the test based on D_n^+ is discussed. Discussions in section 2 and section 3 are mainly based on Birnbaum and Tingey (1951) and Birnbaum (1953). In section 4 the greatest lower bound for the power of the test is obtained under a slight modification of Birnbaum's assumption given in section 3. In section 5 numerical tables are obtained in order to make a comparison of the power of the test with that of a parametric test under the assumption of normality.

2. The Distribution of D_n^+.

Let $F_n(x)$ be the empirical distribution function determined by a random sample (ordered) of size n from the continuous distribution function $F(x)$. It is known that the probability

$$P(D_n^+ \leq \varepsilon), \tag{1}$$

where ε is a constant, is independent of $F(x)$ (Wald and Wolfowitz, 1939). Hence we assume that $F(x)$ is the rectangular distribution in the interval $[0, 1]$, namely,

$$F(x) = \begin{cases} 0 & \text{for } x < 0 \\ x & \text{for } 0 \leq x < 1 \\ 1 & \text{for } 1 \leq x \end{cases} .$$

Figure 1 will show that (1) is equal to the probability that the ordered sample:

$$0 \leq X_1 \leq X_2 \leq \cdots \leq X_n \leq 1$$

satisfies the condition:

$$X_i \leq \min\left(\frac{i-1}{n} + \varepsilon,\ 1\right) \text{ for } i = 1, 2, \ldots, n .$$

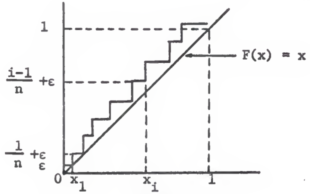

Figure 1.

We know that the probability element of $(X_1, X_2, \ldots, X_n)$ is $n!\,dx_1 dx_2 \ldots dx_n$ for the region $(x_1 < x_2 < \ldots < x_n)$ and $0 \cdot dx_1 dx_2 \ldots dx_n$ elsewhere. Therefore we conclude

$$(2) \qquad P(D_n^+ \leq \varepsilon)$$

$$= n! \int_0^{\varepsilon} \int_{x_1}^{\frac{1}{n}+\varepsilon} \cdots \int_{x_k}^{\frac{k}{n}+\varepsilon} \int_{x_{k+1}}^{1} \cdots \int_{x_{n-1}}^{1} dx_n \ldots dx_{k+2} dx_{k+1} \ldots dx_2 dx_1,$$

where k is the greatest integer j such that $\frac{j}{n} + \varepsilon < 1$.

The following theorem is from Birnbaum and Tingey (1951). This proof is an expanded version of their proof.

<u>THEOREM</u>. For $0 < \varepsilon \leq 1$,

$$P(D_n^+ \leq \varepsilon) = 1 - \varepsilon \sum_{j=0}^{k} \binom{n}{j} \left(1 - \varepsilon - \frac{j}{n}\right)^{n-j} \left(\varepsilon + \frac{j}{n}\right)^{j-1} .$$

Before proving the theorem, let us give the following two formulae; namely, for any integer k, $1 \leq k \leq n$,

$$(3) \qquad \int_{x_{k-1}}^{1} \int_{x_k}^{1} \cdots \int_{x_{n-1}}^{1} dx_n \ldots dx_{k+1} dx_k = \frac{(1-x_{k-1})^{n-k+1}}{(n-k+1)!}$$

$$(4) \qquad \int_0^{\varepsilon} \int_{x_1}^{\frac{1}{n}+\varepsilon} \cdots \int_{x_k}^{\frac{k}{n}+\varepsilon} dx_{k+1} \ldots dx_2 dx_1 = \frac{\varepsilon}{(k+1)!} \left(\varepsilon + \frac{k+1}{n}\right)^{k} .$$

The formula (3) is easy to show by induction. As for (4), let us assume that it is valid for k, and put

$$x_2 = x_1 + y_1$$

$$x_3 = x_1 + y_2$$

$$\vdots$$

$$x_{k+2} = x_1 + y_{k+1}$$

and

$$\frac{1}{n} + \varepsilon - x_1 = \varepsilon' .$$

Then we have, for $k+1$,

$$f(\varepsilon,k+1,n) = \int_0^{\varepsilon} \int_{x_1}^{\frac{1}{n}+\varepsilon} \int_{x_2}^{\frac{2}{n}+\varepsilon} \cdots \int_{x_{k+1}}^{\frac{k+1}{n}+\varepsilon} dx_{k+2} \cdots dx_3 dx_2 dx_1$$

$$= \int_0^{\varepsilon} \left\{ \int_0^{\varepsilon'} \int_{y_1}^{\frac{1}{n}+\varepsilon'} \cdots \int_{y_k}^{\frac{k}{n}+\varepsilon'} dy_{k+1} \cdots dy_2 dy_1 \right\} dx_1 .$$

By the assumption of induction,

$$f(\varepsilon,k+1,n) = \int_0^{\varepsilon} \frac{\varepsilon'}{(k+1)!} \cdot (\varepsilon' + \frac{k+1}{n})^k \, dx_1$$

$$= \int_0^{\varepsilon} \frac{(\frac{1}{n} + \varepsilon - x_1')}{(k+1)!} \cdot (\frac{k+2}{n} + \varepsilon - x_1)^k \, dx_1 .$$

One can easily verify that the last integral yields (4) with k+1 instead of k, namely,

$$\frac{\varepsilon}{(k+2)!}\left(\varepsilon + \frac{k+2}{n}\right)^{k+1} .$$

Therefore (4) is valid for any integer k, $1 \leq k \leq n$.

Proof of the theorem: From (2) and (3) we have

$$P(D_n^+ \leq \varepsilon) = J(\varepsilon, k, n),$$

where $$J(\varepsilon, k, n) = n! \int_0^{\varepsilon} \int_{x_1}^{\frac{1}{n}+\varepsilon} \cdots \int_{x_k}^{\frac{k}{n}+\varepsilon} \frac{(1-x_{k+1})^{n-k-1}}{(n-k-1)!} \, dx_{k+1} \cdots dx_2 dx_1 .$$

Then,

$$J(\varepsilon, k, n) = n! \int_0^{\varepsilon} \int_{x_1}^{\frac{1}{n}+\varepsilon} \cdots \int_{x_{k-1}}^{\frac{k-1}{n}+\varepsilon} \left\{ \frac{(1-x_k)^{n-k}}{(n-k)!} - \frac{(1-\frac{k}{n}-\varepsilon)}{(n-k)!} \right\} dx_k \cdots dx_2 dx_1$$

$$= n! \int_0^{\varepsilon} \int_{x_1}^{\frac{1}{n}+\varepsilon} \cdots \int_{x_{k-1}}^{\frac{k-1}{n}+\varepsilon} \frac{(1-x_k)^{n-k}}{(n-k)!} \, dx_k \cdots dx_2 dx_1$$

$$- \frac{n!}{(n-k)!}\left(1-\frac{k}{n}-\varepsilon\right)^{n-k} \int_0^{\varepsilon} \int_{x_1}^{\frac{1}{n}+\varepsilon} \cdots \int_{x_{k-1}}^{\frac{k-1}{n}+\varepsilon} dx_k \cdots dx_2 dx_1 .$$

With this and from (4), we obtain

$$J = (\varepsilon,k,n) = J(\varepsilon,k-1,n) - \varepsilon\binom{n}{k}\left(1 - \frac{k}{n} - \varepsilon\right)^{n-k}\left(\varepsilon + \frac{k}{n}\right)^{k-1} .$$

Applying this procedure successively will give us

$$J(\varepsilon,k,n) = J(\varepsilon,0,n) - \varepsilon \sum_{j=1}^{k} \binom{n}{j}\left(1 - \frac{j}{n} - \varepsilon\right)^{n-j}\left(\varepsilon + \frac{j}{n}\right)^{j-1} .$$

Finally noting that

$$J(\varepsilon,0,n) = n! \int_0^{\varepsilon} \frac{(1-x_1)^{n-1}}{(n-1)!}\, dx_1 = 1 - (1-\varepsilon)^n ,$$

we have

$$P(D_n^+ \leq \varepsilon) = J(\varepsilon,k,n)$$

$$= 1 - \varepsilon \sum_{j=0}^{k} \binom{n}{j}\left(1 - \varepsilon - \frac{j}{n}\right)^{n-j}\left(\varepsilon + \frac{j}{n}\right)^{j-1} .$$

Thus the proof is complete.

If we let

$$D_n^- = \sup_{-\infty<x<+\infty} [F_n(x) - F(x)] ,$$

then it can be shown, by the symmetry of D_n^+ and D_n^-, that

$$P(D_n^- \leq \varepsilon) = P(D_n^+ \leq \varepsilon) .$$

By making use of the theorem, we can compute, for given values of α

and n, value of ε for which

(5) $$P(D_n^+ \leq \varepsilon) = 1 - \alpha,$$

where $0 < \alpha < 1$.

Values of ε for several values of n and α are given in Table 1, taken from Birnbaum and Tingey (1951).

Table 1. The values of ε for (5).

α/n	.10	.05	.01	.001
5	.4470	.5094	.6271	.7480
8	.3583	.4096	.5065	.6130
10	.3226	.3687	.4566	.5550
20	.2316	.2647	.3285	.4018
40	.1655	.1891	.2350	.2877
50	.1484	.1696	.2107	.2581
	(.1517)	(.1731)	(.2146)	(.2628)

In practice, when n is greater than 50, one can use an approximation based on the asymptotic distribution function of D_n^+ due to Smirnov (1939), namely,

(6) $$P(D_n^+ \leq \varepsilon) \doteq 1 - e^{-2n\varepsilon^2} .$$

Numbers in parentheses of Table 1 are due to (6) for n = 50, and one can see that these are fairly accurate when n = 50.

3. The Power of the Test based on D_n^+.

Let $F(x)$, the distribution function of the random variable X, be continuous. We want to test the null hypothesis

$$H_o: \; F(x) \equiv H(x)$$

against the alternative hypothesis

$$H_1: \; F(x) \equiv G(x) \; .$$

We use D_n^+ for the test statistic. For a test of size α (significance level) we draw a sample of size n from the population considered, and compute from the sample the empirical distribution function $F_n(x)$. We will reject H_o at α level if the inequality

$$D_n^+ > \varepsilon$$

is satisfied, where ε is the value determined in such a way that, provided H_o is true, we have

$$P(D_n^+ \leq \varepsilon \mid H(x)) = 1 - \alpha \; .$$

The value of ε is given in Table 1. For $n > 50$, we use the asymptotic distribution function of D_n^+ given by (6).

The power of this test is given by

$$Q = 1 - P(D_n^+ \leq \varepsilon \mid G(x)) \; .$$

One can easily verify that the inequality

$$D_n^+ \leq \varepsilon$$

is satisfied if and only if

$$H(X_i) \leq \frac{i-1}{n} + \varepsilon \quad \text{for } i = 1, 2, \ldots, n$$

is true (refer to Fig. 1). Hence

$$P(D_n^+ \leq \varepsilon \mid G(x))$$

$$= P(H(X_i) \leq \frac{i-1}{n} + \varepsilon,\ i = 1, 2, \ldots, n \mid G(x))$$

(7) $$= P(X_i \leq H^{-1}(\frac{i-1}{n} + \varepsilon),\ i = 1,2,\ldots,n \mid G(x))$$

$$= P(G(X_i) \leq G[H^{-1}(\frac{i-1}{n} + \varepsilon)],\ i = 1,2,\ldots,n),$$

where H^{-1} is the inverse function of H.

We recall that since G(x) is continuous, the new random variable $Z = G(X)$ has the uniform distribution in the interval [0, 1]. Hence the $Z_i = G(x_i)$ are independent order statistics drawn from a population with the uniform distribution in the interval [0, 1]. So we obtain

(8) $$1 - Q = P(Z_i \leq G[H^{-1}(\frac{i-1}{n} + \varepsilon)],\ i = 1, \ldots, n \mid \mathcal{U}(Z)),$$

where $\mathcal{U}(Z)$ is the uniform distribution function in the interval [0, 1].

By the fact that the probability element of $(Z_1, Z_2, \ldots, Z_n)$ is

$$n!\, dz_1 \cdot dz_2 \ldots dz_n \quad \text{for } z_1 < z_2 < \ldots < z_n$$

and

$$0 \cdot dz_1 dz_2 \ldots dz_n \qquad \text{elsewhere,}$$

we conclude

$$Q = \text{Power} = 1 - n!\int_0^{R(\varepsilon)} \int_{Z_1}^{R(\frac{1}{n}+\varepsilon)} \cdots \int_{Z_{n-1}}^{R(\frac{n-1}{n}+\varepsilon)} dZ_n \cdots dZ_2 dZ_1,$$

where

$$(9) \qquad R(v) = \begin{cases} \lim\limits_{0<v\to 0} G[H^{-1}(v)] & \text{for } v \leq 0 \\ G[H^{-1}(v)] & \text{for } 0 < v < 1 \\ \lim\limits_{1>v\to 1} G[H^{-1}(v)] & \text{for } v \geq 1 \end{cases}$$

Birnbaum (1953) found the greatest lower bound for the power of this test under the assumption that

$$(10) \qquad \sup_{-\infty<x<+\infty} [H(x) - G(x)] = \delta > 0$$

and

$$(11) \qquad H(x_0) - G(x_0) = \delta .$$

He established that

$$(12) \qquad \text{Power} \geq \sum_{i=0}^{j} \binom{n}{i} U_0^{\,i} (1 - U_0)^{n-i} ,$$

where $U_0 = G(x_0)$, $j = [n(v_0 - \varepsilon)]$ and $v_0 = H(x_0)$.

He also gave the least upper bound for the power, namely,

$$\text{Power} \le 1 \quad \text{for } \varepsilon < \delta$$

and

$$\text{(13)} \qquad \text{power} \le (\varepsilon-\delta) \sum_{i=0}^{k} \binom{n}{i} \left(1-\varepsilon+\delta - \frac{i}{n}\right)^{n-i} \left(\varepsilon-\delta+ \frac{i}{n}\right)^{i-1} \quad \text{for } \varepsilon \ge \delta,$$

where $k = [n(1-\varepsilon+\delta)]$.

In fact, the right-hand side of (12) is the power when $G(x)$ is such that

$$G(x) = G^*(x) = \begin{cases} H(x_0) - \delta & \text{for } x \le x_0 \\ 1 & \text{for } x > x_0 \end{cases}$$

and the right-hand side of (13) is the power when $G(x)$ is such that

$$G(x) = G^{**}(x) = \max\,[H(x) - \delta,\ 0]\,.$$

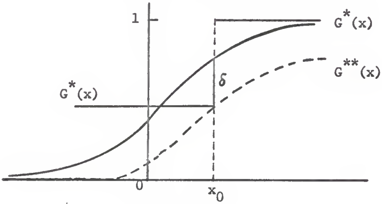

Figure 2 .

In practice, there exists neither $G^*(x)$ nor $G^{**}(x)$ as a distribution function. However we can construct a $G(x)$ arbitrarily close to $G^*(x)$ or $G^{**}(x)$.

4. Lower Bound for the Power under Certain Assumptions.

In this section we make a slight modification of the assumptions of (10) and (11), and we will find the greatest lower bound for the test. Occasionally it is plausible to make this modified assumption, and in such cases the lower bound given by (12) may sometimes be sharpened.

Let us assume that

(14) $$H(x) \geq G(x) \qquad \text{for all } x$$

and

(15) $$H(x_0) - G(x_0) = d \quad .$$

Under this assumption we can find the greatest lower bound for the power of the forementioned test. To see this let

$$\delta(x) = H(x) - G(x) \quad .$$

Then it can be seen that $d = \delta(x_0)$ and

$$G[H^{-1}(\frac{i-1}{n} + \varepsilon)] = \frac{i-1}{n} + \varepsilon - \delta[H^{-1}(\frac{i-1}{n} + \varepsilon)], \text{ (see Fig. 3)} \quad .$$

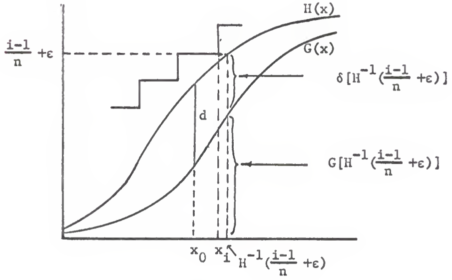

Figure 3 .

From (7) and (8) we have

$$(16) \qquad P(D_n^+ \leq \varepsilon \mid G(x))$$

$$= P(Z_i \leq \frac{i-1}{n} + \varepsilon - \delta[H^{-1}(\frac{i-1}{n} + \varepsilon)],\ i = 1, \ldots, n \mid \bigcup(Z)) .$$

The probability (16) will give its maximum value when $G(x)$ is such that the values of $\delta[H^{-1}(\frac{i-1}{n} + \varepsilon)]$ are small as possible for all i under consideration. This would occur when $\delta(x)$ is very close to $\delta_1(x)$ defined by

$$\delta_1(x) = \begin{cases} H(x) - G(x_0) & \text{for } H^{-1}(G(x_0)) < x \leq x_0 \\ 0 & \text{elsewhere.} \end{cases}$$

In fact, for any $G(x)$ under the given assumption we have

$$P(D_n^+ \leq \varepsilon \mid G(x))$$

$$\leq P(Z_i \leq \frac{i-1}{n} + \varepsilon - c_i,\ i = 1,2,\ldots,n \mid \bigcup(Z)),$$

where

$$(17) \qquad c_i = \begin{cases} 0 & \text{for } i \leq k \\ \frac{i-1}{n} + \varepsilon - G(x_0) & \text{for } k < i \leq \ell \\ 0 & \text{for } \ell < i \leq n, \end{cases}$$

$k = [n(G(x_0) - \varepsilon) + 1]$, and $\ell = [n(H(x_0) - \varepsilon) + 1]$, (see Fig. 4) .

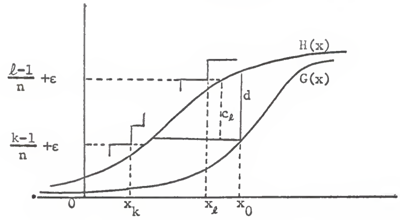

Figure 4.

Hence we have

$$P(D_n^+ \le \varepsilon \mid G(x))$$

$$(18) \quad \le n! \int_0^{\varepsilon}\int_{z_1}^{\frac{1}{n}+\varepsilon}\cdots\int_{z_{k-1}}^{\frac{k-1}{n}+\varepsilon}\int_{z_k}^{b}\cdots\int_{z_{\ell-1}}^{b}\int_{z_\ell}^{\frac{\ell}{n}+\varepsilon}\cdots\int_{z_{m-1}}^{\frac{m-1}{n}+\varepsilon}\int_{z_m}^{1}\cdots\int_{z_{n-1}}^{1}$$

$$dz_n \ldots dz_{m+1}dz_m \ldots dz_{\ell+1}dz_\ell \ldots dz_{k+1}dz_k \ldots dz_2 dz_1,$$

where $b = G(x_0)$ and $m = [n(1-\varepsilon)+1]$.

Since the power of the test is the complementary probability of $P(D_n^+ \le \varepsilon \mid G(x))$, we obtain:

<u>THEOREM</u>. Under the assumption of $H(x) \ge G(x)$, for all x, with $H(x_0) - G(x_0) = d$, the greatest lower bound for the power of the D_n^+ test is

$$(19) \qquad 1 - P_n,$$

where P_n is given in the right-hand side of (18), and also by (20).

One can see that, when d in (15) equals δ in (10), the greatest lower bound given by (19) is greater than or equal to that of Birnbaum given by (12), because $G^*(x)$ in Fig. 2 is greater than or equal to the corresponding curve, Min $(G^*(x), H(x))$, as indicated in Fig. 4. This fact is not contrary to our intuition because the assumption of $G(x)$ given by (14) is more restrictive than the assumption of $G(x)$ given by Birnbaum.

The integration of P_n is tedious, but straightforward, and one may verify the following results, step by step.

$$P_1 = \int_{z_m}^{1} \cdots \int_{z_{n-1}}^{1} dz_n \cdots dz_{m+1} = \frac{(1-z_m)^{n-m}}{(n-m)!}$$

$$P_2 = \int_{z_\ell}^{\frac{\ell}{n}+\varepsilon} \cdots \int_{z_{m-1}}^{\frac{m-1}{n}+\varepsilon} P_1 dz_m \cdots dz_{\ell+1}$$

$$= \frac{(1-z_\ell)^{n-\ell}}{(n-\ell)!} - \frac{(\frac{\ell}{n}+\varepsilon-z_\ell)}{(n-\ell)!} \sum_{j=\ell}^{m-1} [\binom{n-\ell}{n-j}(1-\frac{j}{n}-\varepsilon)^{n-j} (\frac{j}{n}+\varepsilon-z_\ell)^{j-\ell-1}].$$

Now let P_{21} and P_{22} be the first and second term of the right-hand side of the last expression respectively, then we get

$$P_{31} = \int_{z_k}^{b} \cdots \int_{z_{\ell-1}}^{b} P_{21} dz_\ell \cdots dz_{k+1} = \frac{(1-z_k)^{n-k}}{(n-k)!}$$

$$- \frac{1}{(n-k)!} \sum_{j=0}^{\ell-k-1} \binom{n-k}{j} (1-b)^{n-k-j} (b-z_k)^j,$$

and

$$P_{32} = \int_{z_k}^{b} \cdots \int_{z_{\ell-1}}^{b} P_{22} dz_\ell \cdots dz_{k+1}$$

$$= -\frac{1}{(n-k)!} \binom{n-k}{n-\ell} (1 - \frac{\ell}{n} - \varepsilon)^{n-\ell} (b-z_k)^{\ell-k}$$

$$+ \sum_{i=1}^{\ell-k} \frac{(b-z_k)^{\ell-k-i}}{(n-k)!} \binom{n-k}{n-\ell+i} \sum_{j=\ell+1}^{m-1} [\binom{n-\ell+i}{n-j} (1 - \frac{j}{n} - \varepsilon)^{n-j}$$

$$(\frac{j}{n} + \varepsilon - b)^{j-\ell-1+i} (\varepsilon - b + \frac{\ell-i}{n})]$$

$$- \frac{1}{(n-k)!} \sum_{j=\ell+1}^{m-1} \binom{n-k}{n-j} (1 - \frac{j}{n} - \varepsilon)^{n-j} (\frac{j}{n} + \varepsilon - z_k)^{j-k-1} (\varepsilon - z_k + \frac{k}{n})$$

Again letting $P_{31}^{(k)}$ and $P_{32}^{(k)}$ be the kth term of the right-hand sides of P_{31} and P_{32} respectively, we can express P_n as:

$$P_n = n! \int_0^{\varepsilon} \int_{z_1}^{\frac{1}{n}+\varepsilon} \cdots \int_{z_{k-1}}^{\frac{k-1}{n}+\varepsilon} (P_{31}^{(1)} + P_{31}^{(2)} + P_{32}^{(1)} + P_{32}^{(2)} + P_{32}^{(3)})$$

$$dz_k \cdots dz_2 dz_1 .$$

By performing integrations term by term, we obtain the final result

$$(20)\quad P_n = 1 - \varepsilon \sum_{j=0}^{k-1} \binom{n}{j} \left(\varepsilon + \frac{j}{n}\right)^{j-1} \left(1 - \frac{j}{n} - \varepsilon\right)^{n-j}$$

$$- \varepsilon \sum_{j=\ell+1}^{m-1} \binom{n}{j} \left(1 - \frac{j}{n} - \varepsilon\right)^{n-j} \left(\frac{j}{n} + \varepsilon\right)^{j-1}$$

$$+ \varepsilon \sum_{i=1}^{k} \binom{n}{k-1} \binom{n-k+i}{n-\ell} \left(1 - \frac{\ell}{n} - \varepsilon\right)^{n-\ell} \left(b - \varepsilon - \frac{k-i}{n}\right)^{\ell-k+i} \left(\varepsilon + \frac{k-i}{n}\right)^{k-1-i}$$

$$- \binom{n}{\ell} \left(1 - \frac{\ell}{n} - \varepsilon\right)^{n-\ell} b^{\ell}$$

$$- \sum_{j=0}^{\ell-k-1} \binom{n}{j+k} (1-b)^{n-k-j}\, b^{j+k}$$

$$+ \varepsilon \sum_{i=0}^{k-1} \sum_{j=0}^{\ell-k-1} \binom{n}{i} \binom{n-i}{k+j-1} (1-b)^{n-k-j} \left(\varepsilon + \frac{i}{n}\right)^{i-1} \left(b - \frac{i}{n} - \varepsilon\right)^{j+k-i}$$

$$- \varepsilon \sum_{r=1}^{k} \sum_{i=1}^{\ell-k} \sum_{j=\ell+1}^{m-1} \left[\binom{n}{k-r} \binom{n-k+r}{n-\ell+i} \left(b - \varepsilon - \frac{k-r}{n}\right)^{\ell-k-i+r} \left(\varepsilon + \frac{k-r}{n}\right)^{k-1-r}\right.$$

$$\left. \cdot K(i, j)\right]$$

$$+ \sum_{i=1}^{\ell-k} \sum_{j=\ell+1}^{m+1} \left[\binom{n}{n-\ell+i} b^{\ell-i} \cdot K(i,j)\right],$$

where $K(i,j) = \binom{n-\ell+1}{n-j}(1 - \frac{i}{n} - \varepsilon)^{n-j} (\frac{j}{n} + \varepsilon - b)(\varepsilon - b + \frac{\ell-i}{n})$.

If $k = \ell$, $c_i \equiv 0$ in (17), and we have a simple form for P_n, resulting in

$$P_n = 1 - \varepsilon \sum_{j=0}^{m-1} \binom{n}{j} (1 - \varepsilon - \frac{j}{n})^{n-j} (\varepsilon + \frac{j}{n})^{j-1} \quad ,$$

in agreement with the theorem in section 2.

5. Some Numerical Results for the Power.

For the purpose of making a comparison between the power of the test based on D_n^+ and that of a parametric test, let us consider the following hypotheses:

$$H_0: \ H(x) = N_{\mu,\sigma}(x)$$

$$H_1: \ G(x) = N_{\mu_1,\sigma}(x) \quad ,$$

where $\mu_1 = \mu + K$, $K > 0$, and $N_{\mu,\sigma}(x)$ denotes the normal distribution function with mean μ, and standard deviation σ.

We draw a sample of size n from the population considered. Let X_1, $X_2, \ldots, X_n$ be the ordered sample and $F_n(x)$ be the empirical distribution function determined by the sample. For a test of size α, choose a corresponding value of ε from Table 1.

From (8), we have

$$(21) \quad 1 - Q = P(Z_i \leq N_{\mu+K,\sigma}[N_{\mu,\sigma}^{\;-1} (\frac{i-1}{n} + \varepsilon)], \ i = 1,\ldots,n \ | \bigcup(z) \ ).$$

Let

$$N_{\mu,\sigma}^{\;-1}(y) = x_0 \ .$$

Then we can write

$$y = \int_{-\infty}^{x_0} n(t;\mu,\sigma^2)\ dt$$

$$= \int_{-\infty}^{\frac{x_0-\mu}{\sigma}} n(t;\ 0,\ 1)\ dt,$$

where $n(t,\ \mu,\ \sigma^2)$ is the normal density function with mean μ, variance σ^2. Hence we have

$$N_{0,1}^{-1}(y) = \frac{x_0-\mu}{\sigma}$$

$$= \frac{1}{\sigma} N_{\mu,\sigma}^{-1}(y) - \frac{\mu}{\sigma},$$

which is written as

$$N_{\mu,\sigma}^{-1}(y) = \sigma N_{0,1}^{-1}(y) + \mu. \tag{22}$$

From (22) and with the fact that $N_{\mu,\sigma}(x) = N_{0,1}(\frac{x-\mu}{\sigma})$,

we have

$$N_{\mu+K,\sigma}[N_{\mu,\sigma}^{-1}(y)] = N_{\mu+K,\sigma}[\sigma N_{0,1}^{-1}(y) + \mu] \tag{23}$$

$$= N_{0,1}[N_{0,1}^{-1}(y) - \frac{K}{\sigma}] .$$

From (21) and (23), we can write

$$1 - Q = P(Z_i \leq N_{0,1}[N_{0,1}^{-1}(\frac{i-1}{n} + \varepsilon) - \frac{K}{\sigma}],\ i = 1,\ldots,n \mid \mathcal{U}(z)\ ),$$

where $N_{0,1}^{-1}\left(\frac{i-1}{n} + \varepsilon\right)$ is similarly defined as (9).

Therefore we have the power function

$$(24) \qquad Q = 1 - n! \int_0^{N_1} \int_{z_1}^{N_2} \cdots \int_{z_{n-1}}^{N_n} dz_n \ldots dz_2 dz_1 ,$$

where $N_i = N_{0,1}[N_{0,1}^{-1}\left(\frac{i-1}{n} + \varepsilon\right) - \frac{K}{\sigma}]$, $i = 1, 2, \ldots, n$.

We notice that under the given hypotheses, the power function is independent of actual values of μ and σ, but it depends on the value of $\frac{K}{\sigma}$, or $\frac{\mu_1 - \mu}{\sigma}$.

To illustrate the use of (24), let

$$n = 5, \quad \alpha = .10 \text{ and } \frac{K}{\sigma} = 1 .$$

From Table 1 $\varepsilon = .4470$, and then we obtain

$$N_1 = N_{0,1}[N_{0,1}^{-1}(.4470) - 1] = N_{0,1}[-1.133] = .1285$$

$$N_2 = N_{0,1}[N_{0,1}^{-1}(.6470) - 1] = N_{0,1}[-.623] = .2667$$

$$N_3 = N_{0,1}[N_{0,1}^{-1}(.8470) - 1] = N_{0,1}[.024] = .5096$$

$$N_4 = N_{0,1}[N_{0,1}^{-1}(1.0470) - 1]\ N_{0,1}[+\infty] = 1$$

$$N_5 = N_{0,1}[N_{0,1}^{-1}(1.2470) - 1] = N_{0,1}[+\infty] = 1.$$

By replacing these values for N_i's in (24), and after a little calculation, we have the power

$$Q = .7495.$$

Table 2 is the result of several such calculations, and gives the values of power for $\alpha = .10$, .05 and .01 when $n = 5$, with $\frac{K}{\sigma} = .5$, 1.0, 1.5 and 2.0 .

Table 2. Values of the power using D_n^+.

K/σ	α = .10	α = .05	α = .01
.5	31.65%	25.09%	8.56%
1.0	74.95	61.57	33.16
1.5	94.93	89.50	68.19
2.0	99.50	97.89	90.53

The most powerful parametric test under the same hypotheses would be the following.

By noting that

$$\alpha = P(\bar{X}_n > c \mid N(x;\, \mu,\, \sigma^2)) \tag{25}$$

$$= P(Z > \frac{c-\mu}{\sigma/\sqrt{n}} \mid N(z;\, 0,\, 1)) \;,$$

we have, for the power,

$$Q = P(\bar{X}_n > c \mid N(x;\, \mu + K,\, \sigma^2)) \tag{26}$$

$$= P(z > \frac{c-\mu-K}{\sigma/\sqrt{n}} \mid N(z;\, 0,\, 1)) \quad .$$

If we let α = .10, n = 5 and $\frac{\alpha}{K}$ = 1, then from (25) and (26) we have

$$\frac{c-\mu}{\sigma/\sqrt{n}} = 1.282$$

$$\frac{c-\mu-K}{\sigma/\sqrt{n}} = 1.282 - \sqrt{5} = -\,.954,$$

and therefore

$$Q = .8299 .$$

The following Table 3 is the values of the power of the parametric test with the same sizes of α, n and $\frac{K}{\sigma}$ as are in Table 2.

Table 3. Values of the power using the most powerful test.

$\frac{K}{\sigma}$	$\alpha = .10$	$\alpha = .05$	$\alpha = .01$
.5	43.48%	29.91%	11.35%
1.0	82.99	72.27	46.41
1.5	98.09	95.58	84.80
2.0	99.93	99.77	98.40

By comparing Table 2 with Table 3, we see that the one-sided D_n^+ test turns out to be, under the hypotheses of normal distributions with equal variances, less powerful than the one-sided classical test. A similar result was obtained by Van Der Wearden (1953), when H(x) is normal with mean 0, variance 1, and G(x) is normal with mean $\mu > 0$, variance 1, for n = 2,3,5 and for $\alpha = .01$.

However it should be noted that the comparison is not quite fair. The Kolmogorov-Smirnov test may be used when the actual functional form of the distribution is not known, whereas the classical parametric test is used when the functional form is known and only a parameter is unknown. As van der Wearden noted, if, for instance, the true distribution is normal with mean 0 and variance much smaller than 1, Kolmogorov-Smirnov test may enable us to reject the hypothesis that variance equals 1, whereas the classical test used in this section is quite useless for this purpose.

When the functional forms of H(x) and G(x) are known one can construct a more powerful test of two simple hypotheses than that based on D_n^+. If the hypotheses are composite it may not be the case. The usefulness of the test based on D_n^+ is that, with a small loss of power, we have our test for all continuous distributions. The test is distribution free.

ACKNOWLEDGEMENT

The writer wishes to express his sincere appreciation to his major professor, Dr. W. J. Conover, for suggestion of this topic and for his advice and assistance during the preparation of this report.

REFERENCES

Birnbaum, Z. W. and Tingey, F. H. (1951). One sided confidence contours for probability distribution function. Ann. Math. Statist. **22** 592-596.

Birnbaum, Z. W. (1953). On the power of a one-sided test of fit for continuous probability functions. Ann. Math. Statist. **24** 484-489.

Birnbaum, Z. W. (1953). Distribution-free tests of fit for continuous distribution functions. Ann. Math. Statist. **24** 1-8.

Chapman, D. G. (1958). A comparative study of several one-sided goodness-of-fit tests. Ann. Math. Statist. **29** 655-674.

Darling, D. A. (1957). The Kolmogorov-Smirnov, Cramér-von mises tests. Ann. Math. Statist. **28** 823-824.

Fisz, M. (1963). Probability Theory and Mathematical Statistics (3rd ed.). Wiley, New York.

Massey, F. J. (1950). A note on the power of a nonparametric test. Ann. Math. Statist. **21** 440-443.

Massey, F. J. (1951). The Kolmogorov-Smirnov test for goodness of fit. J. Amer. Statist. **46** 68-78.

Smirnov, N. (1939). Sur les écarts de la courbe de distribution empirique. Rec. Math. **6** 3-26.

van der Wearden, B. L. (1953). Testing a distribution function. Nederl. Akad. Wetensch. proc. Ser. **A56** (Indagationes Math. 15) 201-207.

Wald, A. and Wolfowitz, J. (1939). Confidence limits for continuous distribution functions. Ann. Math. Statist. **10** 105-118.

The power of the one-sided and one-sample Kolmogorov-Smirnov test

by

Sung-Wook Lee

B.S., National Seoul University, 1956

AN ABSTRACT OF A MASTER'S REPORT

submitted in partial fulfillment of the requirements for the degree

MASTER OF SCIENCE

Department of Statistics
Kansas State University
Manhattan, Kansas

1966

ABSTRACT

The power of the one-sided and one-sample Kolmogorov-Smirnov test is studied in this report.

Let $F_n(x)$ be the empirical distribution function determined by an ordered sample of size n drawn from a population with a continuous distribution function $F(x)$ which is unknown. Then, for an alternative $G(x)$, the power is defined as

$$Pr[D_n^+ \leq \varepsilon\ (\alpha,n) \mid G(x)],$$

where $D_n^+ = \sup_{-\infty<x<+\infty} [F(x) - F_n(x)]$, and $\varepsilon(\alpha,n)$ is some constant which depends on α (level of significance) and n .

The main difficulty of studies on the power for the test (in general, for all non-parametric tests) is how to select the alternative hypothesis from among all possible alternative hypotheses.

Birnbaum (1953) gave the greatest lower bound and the least upper bound for the test under the assumption that

$$\sup_{-\infty<x<\infty} [F(x) - G(x)] = \delta$$

and

$$F(x_o) - G(x_o) = \delta\ .$$

Under a slight modified assumption of the above, the greatest lower bound for the test is found.

The power for the test is compared with the power for a parametric test under the assumption of normal distributions with equal variances for α = .10, .05, and .01 when n = 5. The result of the comparison is that the Kolmogorov-Smirnov test is less powerful than the parametric test considered. Needless to say, a non-parametric test is a tool which may be used when the functional form of the hypothesis tested is not known.