Statistical Research Report
Institute of Mathematics
University of Oslo

No. 2

June 1968

ALTERNATIVE INTERPRETATIONS OF SOME KOLMOGOROV - SMIRNOV TYPE STATISTICS.

By

Emil Spjøtvoll

S U M M A R Y

It is shown how the Kolmogorov - Smirnov two-sample test can be expressed as a function of the difference of ordered ranks and their expectations. In the case of equal sample sizes this lead to the consideration of a statistic which in some sense is finer than the Kolmogorov - Smirnov statistic.

Similar interpretations are also given of Rényi's and Cramér - von Mises test.

1. THE KOLMOGOROV - SMIRNOV TWO SAMPLE STATISTICS.

Let $X_1, \ldots, X_m$ and $X_1, \ldots, Y_n$ be samples from continuous distribution functions F and G, respectively. Let F_m and G_n be the empirical distribution functions formed from the samples, that is, $m F_m(t)$ is the number which do not exceed t, with $n G_n(t)$ defined analogously.

The Kolmogorov - Smirnov two-sample tests for the hypotheses

H_1: $F = G$ against $F > G$

H_2: $F = G$ against $F \neq G$

are based **upon the statistics**

$$D^+ = \max_t (F_m(t) - G_n(t))$$

and

$$D = \max_t |F_m(t) - G_n(t)|,$$

respectively.

Let the ordered observations be $X_{(1)} < \ldots < X_{(m)}$ and $Y_{(1)} < \ldots < Y_{(n)}$, and let $R_1 < \ldots < R_m$ be the ordered ranks of the X's in the combined sample, and $S_1 < \ldots < S_n$ the ordered ranks of the Y's. Here R_i and S_j are the ranks of $X_{(i)}$ and $Y_{(j)}$ respectively, in the combined sample. We shall now express D^+ and D in terms of the ranks.

First consider D. Suppose maximum occur at a point t_o (not necessarily unique). If $F_m(t_o) - G_n(t_o) < 0$, the point t_o must be contained in some interval $[Y_{(i)}, X_{(j)} >$ where $Y_{(i)}$ and $X_{(j)}$ are consecutive observations in the combined ordered sample. Since the rank of $Y_{(i)}$ is S_i and of $X_{(j)}$ is R_j we find

$$(1.1) \qquad F_m(t_o) - G_n(t_o) = \frac{1}{m}(S_i - i) - \frac{i}{n} = \frac{j-1}{m} - \frac{R_j - j}{n}$$

If $F_m(t_o) - G_n(t_o) > 0$, the point t_o is in some interval $[X_{(j)}, Y_{(i)} >$ where $X_{(j)}$ and $Y_{(i)}$ are consecutive observations in the combined sample. We get

$$(1.2) \qquad F_m(t_o) - G_n(t_o) = \frac{1}{m}(S_i - i) - \frac{i-1}{n} = \frac{j}{m} - \frac{R_j - j}{n} .$$

Since the maximum must occur in some interval $[Y_{(i)}, X_{(j)} >$ or $[X_{(j)}, Y_{(i)} >$ it follows that

$$D = \max_i \left[\max \left(\frac{i}{n} - \frac{1}{m}(S_i - i), \frac{1}{m}(S_i - i) - \frac{i-1}{n} \right) \right]$$

$$(1.3) \qquad = \frac{1}{m} \max_i \left[\max \left(\frac{m+n}{n} i - S_i, S_i - \frac{m+n}{n} i + \frac{m}{n} \right) \right]$$

or in terms of the R's ,

$$(1.4) \qquad D = \frac{1}{n} \max_i \left[\max \left(R_i - \frac{m+n}{m} i + \frac{n}{m}, \frac{m+n}{m} i - R_i \right) \right]$$

Now consider D^+. If maximum occur at same point t_o with $F_m(t_o) - G_n(t_o) > 0$, then equation (1.2) holds. It is also possible that $D^+ = 0$. Then t_o can be chosen equal to $X_{(m)}$ and no Y_i is greater than $X_{(m)}$. In that case

$$F_m(t_o) - G_n(t_o) = 0 = \frac{m}{m} - \frac{R_m - m}{n} \quad .$$

It follows that

$$(1.5) \qquad D^+ = \frac{1}{m} \max \left[\max_i \left(S_i - \frac{m+n}{n} i + \frac{m}{n} \right), 0 \right]$$

or

$$(1.6) \qquad D^+ = \frac{1}{n} \max_i \left(\frac{m+n}{m} i - R_i \right)$$

Let ER_i and ES_i be the expectations when $F = G$ of R_i and S_i respectively. We shall prove that

$$(1.7) \qquad \begin{aligned} D &= \frac{1}{m} \max_i \left[\max \left(\frac{m}{n(n+1)} i - (S_i - ES_i), S_i - ES_i - \frac{m}{n(n+1)} i + \frac{m}{n} \right) \right] \\ &= \frac{1}{n} \max_i \left[\max \left(R_i - ER_i - \frac{n}{m(m+1)} i + \frac{n}{m}, \quad \frac{n}{m(m+1)} - (R_i - ER_i) \right) \right] \end{aligned}$$

and

$$(1.8) \qquad \begin{aligned} D^+ &= \frac{1}{m} \max \left[\max_i \left(S_i - ES_i - \frac{m}{n(n+1)} i + \frac{m}{n} \right), 0 \right] \\ &= \frac{1}{n} \max \left(\frac{n}{m(m+1)} i - (R_i - ER_i) \right) \end{aligned}$$

To prove (1.7) and (1.8) we now find ER_i and ES_i. It is easily seen that when $F = G$

$$(1.9) \qquad P(S_i = x) = \frac{\binom{x-1}{i-1}\binom{m+n-x}{n-i}}{\binom{m+n}{n}} \quad x = i, \ldots, m+i \; .$$

From (1.9) we obtain the identity

$$(1.1o) \qquad \sum_{x=i}^{m+i} \binom{x-1}{i-1}\binom{m+n-x}{n-i} = \binom{m+n}{n} ,$$

We find

$$ES_i = \sum_{x=i}^{m+i} x \binom{x-1}{i-1}\binom{m+n-x}{n-i}\binom{m+n}{n}^{-1}$$

$$= \binom{m+n}{n}^{-1} i \sum_{x=i}^{m+i} \binom{x}{i}\binom{m+n-x}{n-i}$$

$$= \binom{m+n}{n}^{-1} i \sum_{x'=i'}^{m+i'} \binom{x'-1}{i'-1}\binom{m+n'-x'}{n'-i'}$$

where $x' = x+1$, $i' = i+1$ $n' = n+1$. Hence by (1.1o.)

$$ES_i = \binom{m+n}{n}^{-1} \binom{m+n+1}{n+1} = \frac{m+n+1}{n+1} i .$$

By symmetry

$$ER_i = \frac{m+n+1}{m+1} i .$$

The identities (1.7) and (1.8) now follows.

In (1.7) and (1.8) the Kolmogorov-Smirnov statistics are given in terms of ranks. The form indicates that they are closely related to the statistics

$$V = \frac{1}{n} \max_i |R_i - ER_i|$$

(1.11)

$$W = \frac{1}{m} \max_i |S_i - ES_i|$$

and

$$V^+ = \frac{1}{n} \max_i (ER_i - R_i)$$

(1.12) $$W^+ = \frac{1}{m} \max_i (S_i - ES_i) .$$

If we were interested in rank test for the hypothesis H_1 and H_2 the test statistics (1.11) and (1.12) would appear to have a more intuitive appeal than the statistics (1.7) and (1.8) whichseem somewhat artificial. The three sets of statistics will be composed in Sections 2 and 3.

2. THE CASE $m = n$.

We shall compare the statistic V^+ (1.12) which now becomes

(2.1) $$V^+ = \frac{1}{n} \max_i \left(2i - \frac{i}{n+1} - R_i\right)$$

and the statistic D^+ in the form

(2.2) $$D^+ = \frac{1}{n} \max_i (2i - R_i) .$$

Suppose that $2i - R_i$ has a unique maximum for $i = k$, such that

(2.3) $$2i - R_i < 2k - R_k \quad \text{when} \quad i \neq k.$$

Then

(2.4) $$2i - R_i < 2k - R_k - \frac{k-i}{n+1} \quad \text{when } i \neq k$$

since the difference of the lefthand side and righthand side of (2.3) must be ≥ 1 while $\left|\frac{k-i}{n+1}\right| < 1$. But (2.4) is equivalent to

$$2i - \frac{i}{n+1} - R_i < 2k - \frac{k}{n+1} - R_k \quad \text{when} \quad i \neq k .$$

Hence $2i - \frac{i}{n+1} - R_i$ has a unique maximum for $i = k$. It follows that

$$V^+ = D^+ - \frac{k}{n(n+1)} .$$

Suppose now that the maximum of $2i - R_i$ is not unique, and let

$$nD^+ = 2k_j - R_{k_j} \qquad j = 1,\dots,p .$$

Consider

$$2i - \frac{i}{n+1} - R_i .$$

It is seen as above that the maximum value must take place for some k_j , $j=1,\dots,p$. Since $2k_j - R_{k_j}$ is constant, the maximum is attained when k_j is smallest. Hence

$$nV^+ = nD^+ - \frac{1}{n+1} \min_j k_j .$$

Let I_1 be the smallest i such that $nD^+ = 2i - R_i$. Then we have proved that

(2.5) $$nV^+ = nD^+ - \frac{1}{n+1} I_1$$

or

(2.6) $$nD^+ = nV^+ + \frac{1}{n+1} I_1 .$$

Since nD^+ is an integer (see (2.2)), and $|\frac{1}{n+1} I_1| < |$ it follows that

(2.7) $$nD^+ = [nV^+] + 1$$

where $[nV^+]$ is the largest integer less or equal to nV^+. Equation (2.7) gives D^+ as a function of V^+. V^+ is a "finer" statistic than D^+, since V^+ may have several values for each value of D^+. In fact V^+ is equivalent to the pair of statistics (D^+, I_1). V^+ is given as a function of D^+ and I_1 in equation (2.5). Conversely if V^+ is given, D^+ is found from (2.7) Combining (2.5) and (2.7) we then find

$$I_1 = n+1 - (n+1)(nV^+ - [nV^+]) .$$

The distribution of V^+ when $F = G$ may be found from a result proved by Vincze (1957). In theorem 1 of his paper is given the distribution of D^+ and I, where $I = R_{I_1}$. We have $I_1 = \frac{1}{2}(I+nD^+)$.

Hence

$$P(nD^+ = k, I = r) = P(nD^+ = k, I_1 = s)$$
$$= P(nV^+ = k - \frac{s}{n+1})$$

where $s = \frac{1}{2}(r+k)$. From Vincze's result we find that the above is equal to

$$\frac{1}{(2s-1)\,(2n-2s+2)} \quad \frac{\binom{2s}{s}\binom{2n-2s}{n-s}}{\binom{2n}{n}} \qquad \begin{array}{l} k = 0 \\ s = 1,\dots, n. \end{array}$$

(2.8)

$$\frac{k(k+1)}{(2s-k)\,(2n-2s+k+1)} \quad \frac{\binom{2s-k}{s}\binom{2n-2s+k+1}{n-s}}{\binom{2n}{n}} \qquad \begin{array}{l} k = 1,\dots, n \\ s = k,\dots, n\ . \end{array}$$

Since the statistic V^+ is finer than D^+ we can in some cases by using V^+ avoid randomization when trying to find a test for a given level of significance. The probabilities $P(nD^+ \geq k) = P(nV^+ \geq k - 1)$ is given in statistical tables (for some values of k and n) . Hence it is only necessary to compute the probabilities (2.8) for a given value of k, if we, for a given level of significance, want to find a constant c such that we reject the hypothesis H_1 when $V^+ \geq c$.

When comparing D^+ and W^+ we use the form

$$D^+ = \frac{1}{n} \max \left[\max_i (Si - 2i + 1), 0 \right] .$$

Then 0 terms here will introduce some technical difficulty. We therefore introduce the statistic

$$D_o^+ = \frac{1}{n} \max (Si - 2i + 1) \quad .$$

We have

$$D_o^+ = D^+ \quad \text{when} \quad D^+ > 0 \quad .$$

Since we reject the hypothesis H_1 for large values of D^+ and since $P(D^+ > 0) = 1 - \frac{1}{n+1}$ (see e.g. Hodges (1957) p.473) under the hypothesis, it follows that when testing H_1 we can use D_o^+ instead of D^+ .

Analogous to (2.5) it is found that

$$nW^+ = nD_o^+ - 1 + \frac{1}{n+1} I_2$$

where I_2 is the maximum (not the minimum as in (2.5)) of the i's such that $S_i - 2i + 1 = nD_o^+$. It is also found that W^+ is equivalent to the pair $(D_o^+, I_2$

The statistics V^+ and W^+ are not equivalent as shown by the following example. Let $n = 4$, and consider two cases. In both cases $R_1 = 1$, $R_3 = 4$, $R_4 = 7$, $S_2 = 5$, $S_3 = 6$, $S_4 = 8$. In the first case $R_2 = 2$, $S_1 = 3$ while in the second case $R_2 = 3$, $S_1 = 2$. In both cases $W^+ = \frac{7}{20}$, while V^+ is equal to $\frac{8}{20}$ and $\frac{7}{20}$ respectively

By reasons of symmetry we have that

$P(nD^+ = k, I_2 = n-s+1) = P(nW^+ = k + \frac{s}{n+1})$ is equal to $P(nD^+ = k, I_1 = s)$ which is given by (2.8).

Finally compare the statistic

$$V = \frac{1}{n} \max \left| R_i - 2_i + \frac{i}{n+1} \right|$$

and D in the form

$$D = \frac{1}{n} \max_i \left[\max \left(R_i - 2i + 1,\ 2i - R_i\right) \right]$$

$$= \frac{1}{n} \max \left[\max_i \left(R_i - 2i + 1\right),\ \max_i \left(2i - R_i\right) \right].$$

Introduce

$$D_o^- = \frac{1}{n} \max_i \left(R_i - 2i + 1\right)$$

and

$$V^- = \frac{1}{n} \max_i \left(R_i - 2i + \frac{i}{n+1}\right).$$

Then

$$D = \max \left[D_o^-, D^+ \right] \tag{2.9}$$

and it is found that

$$nD^+ = nV^+ + \frac{1}{n+1} I_3$$

(2.10)

$$nD_o^- = nV^- + 1 - \frac{1}{n+1} I_4$$

where I_3 is the smallest i such that $nD^+ = 2i - R_i$ and I_4 the largest i such that $nD_o^- = R_i - 2i + 1$.

Combining (2.9) and (2.10)

$$D = \max \left[V^- + \frac{1}{n}\left(1 - \frac{1}{n+1} I_4\right),\ V^+ + \frac{1}{n(n+1)} I_3 \right].$$

Since $V^- \leq V$ and $V^+ \leq V$ with at least one equality, we get

$$\frac{1}{n(n+1)} \min (n+1-I_4, I_3) \leq D - V \leq \frac{1}{n(n+1)} \max (n+1-I_4, I_3).$$

From the above it is seen that

$$nD = [nV] + 1 .$$

In a similar way it is shown that

$$\frac{1}{n(n+1)} \min (I_5, n+1-I_6) \leq D - W \leq \frac{1}{n(n+1)} \max (I_5, n+1-I_6)$$

where I_5 is the smallest i such that $2i - S_i = nD$, and I_6 the largest i such that $S_i - 2i + 1 = nD$.

3. THE CASE $m \neq n$.

In the case $m \neq n$ there seems to be no simple functional relation between the variables $D^+(D)$ and $V^+(V)$ or $W^+(W)$. This is demonstrated by the following example. Let $m = 2$ and $n = 12$. Then $12D^+ = \max_i (7i - R_i)$ and $12V^+ = \max_i (5i - R_i)$. Let J_1 be the set of i's such that $7i - R_i = 12D^+$, and let J_2 be the set of i's such that $5i - R_i = 12V^+$. Consider the following table.

(R_1, R_2)	I_1	$12D^+$	I_2	$12V+$
(1,9)	1	6	1	4
(1,8)	{1,2}	6	1	4
(1,7)	2	7	1	4
(2,7)	2	7	{1,2}	3
(3,7)	2	7	2	3

It is seen from the above table that we in general have no relationship of the form given in Section 3. I_1 does not determine I_2, and the value of V^+ does not determine D^+ uniquely. Neither does one of the pair (I_1, D^+) and (I_2, V^+) determine any of the other two variables.

4. THE RÉNYI STATISTIC.

We shall consider the statistic

$$(4.1) \qquad R_a^+ = \left(\frac{mn}{m+n}\right)^{\frac{1}{2}} \max \frac{(m+n)\left[F_m(t) - G_n(t)\right]}{m F_m(t) + n G_n(t)}$$

where maximum is taken over all t such that

$$(4.2) \qquad (m+n)^{-1}(m F_m(t) + n F_n(t)) \geq a \quad .$$

The above statistic was introduced by Rényi (1953). The hypothesis H_1 is rejected for large values of R_a^+. A similar statistic for the hypothesis H_2 can be constructed by taking absolute values of the weighted differences after max in (4.1). The maximum in (4.1) must occur at some point $X_{(i)}$ og $Y_{(j)}$. We have

$$(4.3) \qquad \begin{aligned} m F_m(X_{(i)}) + n G_n(X_{(i)}) &= R_i \\ m F_m(Y_{(j)}) + n G_n(Y_{(j)}) &= S_j \quad . \end{aligned}$$

and

$$F_m(X_{(i)}) - G_n(X_{(i)}) = \frac{i}{m} - \frac{R_i - i}{n}$$

(4.4)

$$F_m(Y_{(j)}) - G_n(Y_{(j)}) = \frac{S_j - j}{m} - \frac{j}{n} \quad .$$

Hence (4.1) can be written

$$(4.5) \qquad R_a^{+} = \left(\frac{mn}{m+n}\right)^{\frac{1}{2}} \max \left[\max \frac{(m+n)\left(\frac{i}{m} - \frac{R_i - i}{n}\right)}{R_i} \; , \; \max \frac{(m+n)\left(\frac{S_j - j}{m} - \frac{j}{n}\right)}{S_j} \right.$$

By (4.3) the condition (4.2) is

$$R_i \geq (m+n)a \quad \text{and} \quad S_j \geq (m+n)a \quad .$$

The maximum must take place at same point $X_{(i)}$ with the exception of the of the case where the smallest S_j, say S_o, greater than $(m+n)a$ is smaller than the smallest R_i greater than $(m+n)a$, and the maximum take place at S_o. In that case $S_o = (m+n)a$ (if $(m+n)a$ is an integer , otherwise $S_o = [(m+n)a] + 1$), and

$$(4.6) \; R_a^{+} = \left(\frac{mn}{m+n}\right)^{\frac{1}{2}} \frac{(m+n)\left(\frac{I}{m} - \frac{(m+n)a - I}{n}\right)}{(m+n)a}$$

where I is the index of the largest $R_i < (m+n)a$.

It follows that

$$R_a^{+} = \left(\frac{mn}{m+n}\right)^{\frac{1}{2}} \left[\max_{R_i \geq (m+n)a} \frac{(m+n)\left(\frac{i}{m} + \frac{i}{n} - \frac{R_i}{n}\right)}{R_i} \; , \; \frac{\frac{I}{m} + \frac{I}{n} - \frac{m+n}{n} a}{a} \right]$$

If we neglect the possibility (4.6) we get the statistic

$$\left(\frac{m(m+n)}{n}\right)^{\frac{1}{2}} \max_{R_i \geq (m+n)a} \frac{\left(\frac{m+n}{m} i - R_i\right)}{R_i}$$

$$= \left(\frac{m(m+n)}{n}\right)^{\frac{1}{2}} \max_{R_i \geq (m+n)a} \frac{\left(\frac{m}{m(m+1)} i - (R_i - ER_i)\right)}{R_i} .$$

This should be closely related to

$$\max_{R_i \geq (m+n)a} \frac{ER_i - R_i}{R_i} ,$$

which in turn suggest the use of the statistic

$$\max_{R_i \geq (m+n)a} \frac{ER_i - R_i}{ER_i} .$$

5. THE CRAMÉR - VON MISES TEST.

The test statistic is

$$M = \frac{mn}{m+n} \int_{-\infty}^{\infty} (F_m(t) - G_n(t))^2 \, d\frac{(F_m(t) + G_n(t))}{m+n} .$$

By (4.4) this can be written

$$(5.1) \qquad M = \frac{m}{n(m+n)^2} \left(\sum_{i=1}^{m} \left(R_i - \frac{m+n}{m} i\right)^2 + \frac{n^2}{m^2} \sum_{j=1}^{m} \left(S_j = \frac{m+n}{n} j\right)^2 \right) .$$

Introduce

$$Q = \sum_{j=1}^{n} \left(S_j - \frac{m+n}{n} j\right)^2$$

Since the ranks S_j is uniquely determined when the ranks R_i are given, we can express Q in terms of the R_i. We find

$$Q = \sum_{k=1}^{m} \sum_{j=R_k+1}^{R_{k+1}-1} \left(j-(j-k)\frac{m+n}{n}\right)^2$$

$$+ \sum_{j=0}^{R_1-1} \left(j-j\frac{m+n}{n}\right)^2 + \sum_{j=R_m+1}^{m+n} \left(j-(j-m)\frac{m+n}{n}\right)^2 ,$$

where the last sum is 0 if $R_m = m + n$. After some long and tedious computations we find

$$Q = \frac{m}{n} \sum_{i=1}^{m} \left(R_i - i\frac{m+n}{m}\right)^2$$

$$+ \frac{m+n}{n} \sum_{i=1}^{m} \left(R_i - \frac{1}{6}\frac{m+n}{n}(3m^2 - 3mn - 2m - n)\right)^2$$

Combining this with (5.1) it is found after some more computations

$$(5.2) \qquad M = \frac{1}{m(m+n)} \sum_{i=1}^{m} \left(R_i - i\frac{m+n}{m} + \frac{1}{2}\frac{n}{m}\right)^2 + \frac{2m+n}{12m(m+n)} .$$

This can also be written

$$M = \frac{1}{m(m+n)} \sum_{i=1}^{m} \left(R_i - ER_i - \frac{1}{m(m+1)} i + \frac{1}{2}\frac{n}{m}\right)^2 + \frac{2m+n}{12m(m+n)} .$$

Hence the Cramér - von Mises statistic M is closely related to the astatistic

$$\sum_{i=1}^{m} (R_i - ER_i)^2 .$$

R E F E R E N C E S

H O D G E S , J. L. jr. (1957). The significance probability of the Smirnov two-sample test. Arkiv för Matematik, 3, 469-486.

R É N Y I, A. (1953). On the theory of order statistics. Acta. math. Acad. sci. hung. 4, 191-231.

V I N C Z E , I. (1957). Einige zweidimensionale Verteilungs- und Grenzverteilungssätze in der Theorie der geordneten Stichproben I. Magyar Tud. Akad. Mat. Kutató. Int. Közleményei 2, 183-209.

ESp/elr
6/1968.