ALTERNATIVE INTERPRETATIONS
OF SOME KOLMOGOROV - SMIRNOV
TYPE STATISTICS.

By

Emil Spjøtvoll
SUMMARY

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 $Y_1, \ldots, Y_n$ 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 $X_1 < \cdots < X_m$ and $Y_1 < \cdots < Y_n$,

and let $R_1 < \cdots < R_m$ be the ordered ranks of the $X$'s in the combined sample, and $S_1 < \cdots < S_n$ 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 \( t_0 \) (not necessarily unique). If \( F_m(t_0) - G_n(t_0) < 0 \), the point \( t_0 \) 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

\[
(1.1) \quad F_m(t_0) - G_n(t_0) = \frac{1}{m} (S_i - i) - \frac{i-1}{n} = \frac{i-1}{m} - \frac{R_j - j}{n}
\]

If \( F_m(t_0) - G_n(t_0) > 0 \), the point \( t_0 \) is in some interval \([X(j), Y(i)]\) where \( X(j) \) and \( Y(i) \) are consecutive observations in the combined sample. We get

\[
(1.2) \quad F_m(t_0) - G_n(t_0) = \frac{1}{m} (S_i - i) - \frac{i-1}{n} = \frac{i}{m} - \frac{R_j - j}{n}
\]

Since the maximum must occur in some interval \([Y(i), X(j)]\) or \([X(j), Y(i)]\) it follows that

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

or in terms of the \( R \)'s,

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

Now consider \( D^+ \). If maximum occur at same point \( t_0 \) with

\[
F_m(t_0) - G_n(t_0) > 0, \text{ then equation (1.2) holds. It is also possible}
\]
that \( D^+ = 0 \). Then \( t_0 \) can be chosen equal to \( X_{(m)} \) and no \( Y_i \) is greater than \( X_{(m)} \). In that case
\[
F_m(t_0) - G_n(t_o) = 0 = \frac{m}{m} - \frac{R - m}{n}.
\]
It follows that
\[
(1.5) \quad D^+ = \frac{1}{m} \max_i \left[ \max \left( S_i - \frac{m+i}{n}, \frac{m}{n} \right), 0 \right]
\]
or
\[
(1.6) \quad D^+ = \frac{1}{n} \max_i \left( \frac{m+i}{m} S_i - R_i \right)
\]
Let \( E_{R_i} \) and \( ES_i \) be the expectations when \( F = G \) of \( R_i \) and \( S_i \), respectively. We shall prove that
\[
D = \frac{1}{m} \max_i \left[ \max \left( \frac{m}{n(n+1)} i - (S_i - ES_i), S_i - ES_i - \frac{m}{n(n+1)} i + \frac{m}{n} \right) \right]
\]
\[
(1.7) \quad = \frac{1}{n} \max_i \left[ \max \left( R_i - E_{R_i} - \frac{n}{m(m+1)} i + \frac{n}{m}, \frac{n}{m(m+1)} - (R_i - E_{R_i}) \right) \right]
\]
and
\[
D^+ = \frac{1}{m} \max_i \left[ \max \left( S_i - ES_i - \frac{m}{n(n+1)} i + \frac{m}{n} \right), 0 \right]
\]
\[
(1.8) \quad = \frac{1}{n} \max_i \left( \frac{n}{m(m+1)} i - (R_i - E_{R_i}) \right)
\]
To prove (1.7) and (1.8) we now find \( E_{R_i} \) and \( ES_i \). It is easily seen that when \( F = G \)
\[
(1.9) \quad P(S_i = x) = \frac{(x-1)(m+n-x)}{(i-1)(n-i)} \times i, \ldots , m+i.
\]
From (1.9) we obtain the identity
\[
(1.10) \quad \sum_{x=1}^{m+i} \frac{(x-1)(m+n-x)}{(i-1)(n-i)} = \frac{(m+n)}{(n)}.
\]
We find

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

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

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

where \( x' = x+1, i' = i+1 n' = n+1 \). Hence by (1.10).

\[ ES_i = \binom{m+n}{n}^{-1} \binom{m+n+1}{n+1} = \frac{m+n+1}{n+1} i \]

By symmetry

\[ ER_i = \frac{m+n+1}{n+1} i \]

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

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) which seem 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

\[(2.1) \quad V^+ = \frac{1}{n} \max_i (2i - \frac{i}{n+1} - R_i) \]

and the statistic $D^+$ in the form

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

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

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

Then

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

since the difference of the lefthand side and righthand side of (2.3) must be $> 1$ while $|\frac{k-i}{n+1}| < 1$. 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

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

Consider

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

It is seen as above that the maximum value must take place for some $k_j$, $j = 1, \ldots, p$. 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 $nD^+ = 2i - R_1$. Then we have proved that
\begin{equation}
(2.5) \quad nV^+ = nD^+ - \frac{1}{n+1} I_1
\end{equation}
or
\begin{equation}
(2.6) \quad nD^+ = nV^+ + \frac{1}{n+1} I_1.
\end{equation}

Since $nD^+$ is an integer (see (2.2)), and $\frac{1}{n+1} I_1 < 1$ it follows that
\begin{equation}
(2.7) \quad nD^+ = \left\lfloor nV^+ \right\rfloor + 1
\end{equation}

where $\lfloor nV^+ \rfloor$ 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

\[
I_1 = n+1 - (n+1) \left( nV^+ - \lfloor nV^+ \rfloor \right).
\]

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 $I = R_1$. 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
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

$$D^+ = \frac{1}{n} \max_i \left[ \max_0 (S_i - 2i + 1), 0 \right].$$

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

$$D_0^+ = \frac{1}{n} \max_i (S_i - 2i + 1).$$

We have

$$D_0^+ = D^+ \text{ when } D^+ > 0.$$ 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 $D_0^+$ instead of $D^+$.

Analogous to (2.5) it is found that

$$nW^+ = nD_0^+ - 1 + \frac{1}{n+1} I_2$$

where $I_2$ is the maximum (not the minimum as in (2.5)) of the $i$'s such that $S_i - 2i + 1 = nD_0^+$. It is also found that $W^+$ is equivalent to the pair $(D_0^+, I_2)$. 

(2.8)
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

\[
V = \frac{1}{n} \max \mid R_i - 2i + \frac{i}{n+1} \mid
\]
and \( D \) in the form

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

Introduce

\[
D^- = \frac{1}{n} \max (R_i - 2i + 1)
\]
and

\[
V^- = \frac{1}{n} \max (R_i - 2i + \frac{i}{n+1})
\]

Then

\[
(2.9) \; D = \max \left[ D^-, D^+ \right]
\]
and it is found that

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

(2.10)

\[
nD^- = nV^- + 1 - \frac{1}{n+1} I_4
\]
where $I_3$ is the smallest $i$ such that $nD^+ = 2i - R_i$ and $I_4$ the largest
$i$ such that $nD^- = R_i - 2i$.

Combining (2.9) and (2.10)

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

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

$$nD = \left[nV\right] + 1.$$

In a similar way it is shown that

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

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 demon-
strated by the following example. Let $m = 2$ and $n = 12$. Then $12D^+ = \max$
$(7i - R_i)$ and $12V^+ = \max (5i - R_i)$. 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.
\[ (R_1, R_2) \quad I_1 \quad 12D^+ \quad I_2 \quad 12V^+ \]

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>6</th>
<th>1</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1,9)</td>
<td>2</td>
<td>7</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>(1,8)</td>
<td>2</td>
<td>7</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>(1,7)</td>
<td>2</td>
<td>7</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>(2,7)</td>
<td>2</td>
<td>7</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>(3,7)</td>
<td>2</td>
<td>7</td>
<td>3</td>
<td></td>
</tr>
</tbody>
</table>

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

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

where maximum is taken over all \( t \) such that

\[
(m+n)^{-1}(m F_m(t) + n F_n(t)) \geq a.
\]

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 \( X(i) \) or \( Y(j) \). We have

\[
m F_m(X(i)) + n G_n(X(i)) = R_i
\]

and

\[
m F_m(Y(j)) + n G_n(Y(j)) = S_j
\]
and
\[ F_m(X(i)) = \frac{i}{m} - \frac{R_i - i}{n} \]
(4.4)
\[ F_m(Y(j)) = \frac{S_j}{m} - \frac{j}{n} \]

Hence (4.1) can be written
\[ R^+ = (\frac{mn}{m+n})^{\frac{1}{2}} \max \left( \max \left( \frac{\frac{i}{m} - \frac{R_i - i}{n}}{R_i} \right), \frac{\frac{S_j}{m} - \frac{j}{n}}{S_j} \right) \]

By (4.3) the condition (4.2) is
\[ R_i > (m+n)a \quad \text{and} \quad S_j > (m+n)a \]

The maximum must take place at same point \( X(i) \) with the exception of the case where the smallest \( S_j \), say \( S_0 \), greater than \( (m+n)a \) is smaller than the smallest \( R_i \) greater than \( (m+n)a \), and the maximum take place at \( S_0 \). In that case \( S_0 = (m+n)a \) (if \( (m+n)a \) is an integer), otherwise \( S_0 = \left\lfloor \frac{(m+n)a}{2} + 1 \right\rfloor \), and
\[ R^+ = (\frac{mn}{m+n})^{\frac{1}{2}} \frac{(m+n)(\frac{I}{m} - (m+n)a - I)}{(m+n)a} \]

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

It follows that
\[ R^+ = (\frac{mn}{m+n})^{\frac{1}{2}} \left[ \max \left( \frac{\frac{(m+n)}{m} + \frac{i}{n} - \frac{R_i}{n}}{R_i}, \frac{I}{m} + \frac{I}{n} - \frac{m+n}{n} a \right) \right] \]
If we neglect the possibility (4.6) we get the statistic
\[
\left( \frac{m(m+n)}{n} \right)^{\frac{1}{2}} \max_{R_i > (m+n)a} \left( \frac{m+n}{m} \frac{i - R_i}{R_i} \right)
\]
\[
= \left( \frac{m(m+n)}{n} \right)^{\frac{1}{2}} \max_{R_i > (m+n)a} \left( \frac{m}{m(m+1)} \frac{i - (R_i - ER_i)}{R_i} \right).
\]

This should be closely related to
\[
\max_{R_i > (m+n)a} \frac{ER_i - R_i}{R_i},
\]
which in turn suggest the use of the statistic
\[
\max_{R_i > (m+n)a} \frac{ER_i - R_i}{ER_i}.
\]

5. THE CRAMÉR – VON MISES TEST

The test statistic is
\[
M = \frac{mn}{m+n} \int_{-\infty}^{\infty} \left( F_m(t) - G_n(t) \right)^2 d\left( \frac{F_m(t) + G_n(t)}{m+n} \right).
\]

By (4.4) this can be written
\[
M = \frac{m}{n(m+n)^2} \sum_{i=1}^{m} \left( R_i - \frac{m+n}{m} i \right)^2 + \frac{n^2}{m} \sum_{j=1}^{n} \left( S_j - \frac{m+n}{n} j \right)^2.
\]

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

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

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

\[
Q = \frac{n}{m} \sum_{i=1}^{m} (R_i - i \frac{m+n}{m})^2
\]

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

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

(5.2) \[
M = \frac{1}{m(m+n)} \sum_{i=1}^{m} (R_i - \frac{m+n}{m} + \frac{1}{2} \frac{n}{m})^2 + \frac{2m+n}{12m(m+n)}.
\]

This can also be written

\[
M = \frac{1}{m(m+n)} \sum_{i=1}^{m} (R_i - \frac{m+n}{m+1} - \frac{1}{m(m+1)} i + \frac{1}{2} \frac{n}{m})^2 + \frac{2m+n}{12m(m+n)}.
\]

Hence the Cramér–von Mises statistic \( M \) is closely related to the astatistic

\[
\sum_{i=1}^{m} (R_i - \frac{m+n}{m})^2.
\]
REFERENCES


