THE EXACT DISTRIBUTION OF A SMIRNOV STATISTIC

CENTER FOR NAVAL ANALYSES
1401 Wilson Boulevard
Arlington, Virginia 22209

Marine Corps Operations Analysis Group

By: E.A. Parent, James K. Tyson

February 1977

Approved for public release; distribution unlimited.

Prepared for:

OFFICE OF NAVAL RESEARCH
Department of the Navy
Arlington, Virginia 22217

DEPUTY CHIEF OF STAFF (RD&S)
Headquarters, Marine Corps
Washington, D.C. 20380
The exact distribution is given for the number of times the vertical steps of an empirical distribution cross the underlying theoretical distribution. This statistic was used for Smirnov tests to derive limiting distributions of statistics used in Kolmogorov-Smirnov tests, and may be used as a goodness-of-fit test in its own right.
MEMORANDUM FOR DISTRIBUTION LIST

Subj: Center for Naval Analyses Research Contribution 317

Encl: (1) CRC 317, "The Exact Distribution of a Smirnov Statistic," by E.A. Parent and James K. Tyson
February 1977

1. Enclosure (1) is forwarded as a matter of possible interest.

2. Research Contributions are distributed for their potential value in other studies or analyses. They do not necessarily represent the opinion of the Department of the Navy.

ROBERT J. CORN
Director
Marine Corps Operations
Analysis Group
Distribution List:

Commander, Operational Test and Evaluation Force
Deputy Commander, Operational Test and Evaluation Force, Pacific
Assistant Secretary of the Navy (Research and Development)
Office of Program Appraisal, Department of the Navy
Chief of Naval Research
DC/S for Research, Development & Studies, HqMC
National Defense University
Armed Forces Staff College
Annapolis Laboratory, David Taylor Naval Ship R&D Center
Carderock Laboratory, David Taylor Naval Ship R&D Center
New London Laboratory, Naval Undersea Center
AUTEC Detachment, Naval Undersea Center
San Diego Detachment, Naval Undersea Center
Newport Laboratory, Naval Undersea Center
Naval Surface Weapons Center
Naval Research Laboratory
Fleet Material Support Office
U.S. Naval Academy (Nimitz Library)
Naval Postgraduate School
Naval War College
Marine Corps Development and Education Command
CNO: Op-09BH and Op-96
Director, Defense Research and Engineering
Director, Planning and Evaluation (OSD) (2 copies)
Defense Documentation Center (12 copies)
Institute for Defense Analyses
The Rand Corporation
University of Rochester (2 copies)
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Introduction</td>
<td>1</td>
</tr>
<tr>
<td>2. The distribution of $C_n$</td>
<td>1</td>
</tr>
<tr>
<td>3. Related results</td>
<td>6</td>
</tr>
<tr>
<td>4. A goodness-of-fit test</td>
<td>11</td>
</tr>
<tr>
<td>References</td>
<td>12</td>
</tr>
</tbody>
</table>
1. INTRODUCTION

Smirnov [1] defined a statistic \( V_n \) as the number of times a continuous distribution \( F(x) \) crosses the vertical steps of its empirical distribution \( F_n(x) \). \( F_n(x) \) is based on \( n \) independent observations from a distribution \( F(x) \). Smirnov showed that

\[
P(V_n \leq t\sqrt{n}) \to 1 - e^{-t^2/2} \quad \text{for } t \geq 0 \quad \text{and} \quad P(V_n \leq t\sqrt{n}) \to 0 \quad \text{for } t < 0
\]
as \( n \to \infty \), independent of \( F(x) \). By generalizing this result to counting the number of crossings of \( F(x) + \lambda/\sqrt{n} \) with \( F_n(x) \), and \( F(x) = \lambda/\sqrt{n} \) with \( F_n(x) \), Smirnov was able to show that

\[
P(\sup F_n(x) - F(x) < t\sqrt{n}) \to 1 - e^{-2\lambda^2} \quad \text{and} \quad P(\sup | F_n(x) - F(x) | < t) \to e^{-2k\lambda^2} \quad \text{as } n \to \infty.
\]
The purpose of this paper is to derive the exact distribution of \( V_n \) and define a goodness-of-fit test based on \( V_n \). Actually we consider \( C_n \), which is the number of crossings of \( F(x) \) with the horizontal steps of \( F_n(x) \).

Note that \( C_n = V_n - 1 \). The reason for choosing \( C_n \) is that there is a simple relationship between \( C_n \) and the transformed observations \( F(X_i) \), which allows the test to be applied without recourse to actually plotting \( F(x) \) and \( F_n(x) \) and counting the number of crossings. The procedure is described in section 4.

2. THE DISTRIBUTION OF \( C_n \)

We disregard the steps \( F_n(x) = 0 \) and \( F_n(x) = 1 \) and consider the \( n - 1 \) steps corresponding to \( F_n(x) = i/n \) for \( i = 1, 2, \ldots, n - 1 \) and seek \( P(C_n = k) \) for \( k = 0, 1, \ldots, n - 1 \). If there are exactly \( i \) observations less than \( F^{-1}(i/n) \), then the \( i^{th} \) step will cross \( F(x) \). Conversely, if the \( i^{th} \) step crosses, there
This scheme can be extended to the simultaneous occurrence of any $k$ events $E_i$, $k = 1, 2, \ldots, n - 1$. We define $S_j = \sum_{i=1}^{n-1} P_i$, $j = 1, 2, \ldots, n - 1$, \hspace{1cm} (2.4)

Using the notation $P(E_i)$, $P(E_i | E_j)$, $P(E_i E_j)$ etc., we have

\begin{align*}
P_i &= P(E_i) = \frac{i}{n}, \\
P_i | j = P_i | j = \frac{i}{j}, \\
P_i | j = P_i | j = \frac{i}{j}.
\end{align*}

Using this notation we get, for $1 \leq i < j < k < n - 1$,

\begin{align*}
P_{i,j,k} &= P(E_i E_j E_k) = P(E_i | E_j E_k) P(E_j E_k) \\
&= \left( \frac{i}{j} \right) \left( \frac{j}{k} \right) \left( \frac{k}{n} \right).
\end{align*}

For $1 \leq i < j < n - 1$, we compute the simultaneous occurrence of $E_i$ and $E_j$ by conditioning on $E_k$ as

\begin{align*}
P(E_i E_j) &= P(E_i | E_j) P(E_j) \\
&= \frac{i}{j} \left( \frac{j}{n} \right).
\end{align*}

Extending this notation to get

\begin{align*}
P(X \leq F(\frac{i}{n})) &= P(X \leq F(\frac{j}{n})) = \frac{i}{n},
\end{align*}

must have been exactly $i$ observations less than $F(\frac{i}{n})$. Since
where the summation is taken over all subscripts \(1 \leq i_1 < i_2 < \ldots < i_j < n\), and the indicator variables

\[
X_i = \begin{cases} 
1 & \text{if the } i^{th} \text{ step crosses} \\
0 & \text{otherwise} 
\end{cases}
\]

for \(i = 1, 2, \ldots, n - 1\). Since \(s_i = 1 + X_i(s - 1)\), we have

\[
\sum_{i=1}^{n-1} X_i = 1 + \sum X_i(s - 1) + \sum X_i X_j (s - 1)^2 + \ldots
\]

\[
+ \sum X_i X_2 \ldots X_{n-1}(s - 1)^{n-1},
\]

where the summations are taken over the ranges as in (2.4). Since \(C_n = \sum_{i=1}^{n-1} X_i\), we have the probability generating function of \(C_n\) as

\[
P(s) = E \left( \prod_{i=1}^{n-1} X_i \right) = 1 + \sum_{j=1}^{n-1} (s - 1)^j S_j
\]

(2.5)

because \(E(X_1 X_2 \ldots X_{i-1}) = n! P_{i_1 i_2 \ldots i_j}\).

Note that we can write

\[
n_{i_1 i_2 \ldots i_j} = \frac{n!}{i_1 ! i_2 ! \ldots i_j !} \frac{(n-i)^{n-i}}{(n-i)!},
\]

and, in general,

\[
n_{i_1 i_2 \ldots i_j} = \frac{n!}{i_1 ! i_2 ! \ldots i_j !} \frac{(i_2-i_1)^{i_2-i_1}}{(i_2-i_1)!} \ldots \frac{(n-i_j)^{n-i_j}}{(n-i_j)!}.
\]

Now

\[
S_j = \sum_{i_1 i_2 \ldots i_j} = \sum_{i_1 i_2 \ldots i_j} \frac{(j+1)}{n} \sum_{i_1 = 1}^{j+1} \frac{\lambda_i^{i_1}}{\lambda_i^{i_1}!}.
\]

(2.6)

where this summation is taken over \(\sum_{i=1}^{i+1} \lambda_i = n\) and \(\lambda_i \geq 1\).
We calculate the unconstrained sum

\[
\sum_{\lambda_1 \geq 1} x^{\lambda_1} \cdots x^{\lambda_j} \left( \prod_{i=1}^{j+1} \frac{x^{\lambda_i}}{\lambda_i!} \right)
\]  

(2.7)

and use the coefficient of \( x^n \) to get \( S_{jn} \).

Using the Residue Theorem, we have

\[
\frac{\lambda^\lambda}{\lambda!} = \frac{1}{2\pi i} \oint \frac{e^{\lambda z}}{z^{\lambda+1}} \, dz,
\]

(2.8)

and

\[
\sum_{\lambda=1}^{\infty} \frac{x^{\lambda}}{\lambda!} = \frac{1}{2\pi i} \oint \frac{\sum_{\lambda=1}^{\infty} z^{\lambda}}{z} \left( \frac{x}{z} \right)^\lambda \, dz
\]

\[
= \frac{1}{2\pi i} \oint \frac{1}{z} \frac{xe^z}{(z-xe^z)} \, dz
\]

\[
= \frac{z(x)}{1-z(x)},
\]

(2.9)

where \( z(x) \) is the solution to \( z = xe^z \) such that \( \frac{z(x)}{x} \to 1 \) as \( x \to 0 \).

The coefficient of \( x^n \) is obtained from

\[
\frac{1}{2\pi i} \oint \frac{x^{n+1}}{x^n+1} \left( \frac{z(x)}{1-z(x)} \right)^{j+1} \, dx;
\]

(2.10)

and changing the variable of integration from \( z \) to \( x \), we have

\[
S_j = \frac{n!}{n^n} \frac{1}{2\pi i} \oint \frac{e^{nz}}{z^n} \left( \frac{z}{1-z} \right)^j \, dz = \frac{n!}{n^n(j-1)!} \sum_{i=0}^{n-j-1} \frac{n-i}{i!} \frac{(n-2-i)!}{(n-j-1-i)!}.
\]

(2.11)

Notice that \( S_{jn} = 0 \) for \( j \geq n \) in (2.4) and (2.11). Thus from (2.5)
we have

\[ P(s) = 1 + \sum_{j=1}^{\infty} (s-1)^j s_j \]

\[ = 1 + \frac{n!}{n^n} \frac{1}{2\pi i} \oint \frac{e^{nz}}{z^n} \sum_{j=1}^{\infty} (s-1)^j \left(\frac{z}{1-z}\right)^j \, dz \]

\[ = 1 + \frac{n!}{n^n} \frac{(s-1)}{2\pi i} \oint \frac{e^{nz}}{z^{n-1}} \frac{1}{1-sz} \, dz. \quad (2.12) \]

We have

\[ C(s) = \frac{P(s)}{1-s} = \sum_{k=0}^{n-1} \sum_{i=n-k}^{n-1} \frac{P(C_n<k)s^k}{k!} \]

\[ = \frac{1}{1-s} \frac{n!}{n^n} \frac{1}{2\pi i} \oint \frac{e^{nz}}{z^{n-1}} \frac{1}{1-sz} \, dz, \]

which is the generating function of the cumulative probabilities.

Since \( C^{(k)}(0) = P(C_n<k)k! \), we have

\[ P(C_n<k) = 1 - \frac{n!}{n^n} \frac{1}{2\pi i} \oint \frac{e^{nz}}{z^{n-1-k}} \, dz \]

\[ = 1 - \frac{n!}{(n-2-k)!n^{2+k}}, \quad (2.13) \]

and

\[ P(C_n=k) = \frac{(1+k)n!}{n^{2+k}(n-1-k)!}. \quad (2.14) \]

Using Stirling's formula and \( \ln(1-x) \approx -x - x^2/2 \) for small \( x \), we have, for \( k = t\sqrt{n} \),
\[ P(C_{n<k}) \approx 1 - \left( \frac{n}{n-(k+2)} \right)^{1/2} \left( \frac{n}{n-(k+2)} \right)^{n-(k+2)} \exp(-(k+2)) \]

\[ = 1 - \exp \left( -\frac{1}{2} \ln \left( 1 - \frac{k+2}{n} \right) - (n-(k+2)) \ln \left( 1 - \frac{k+2}{n} \right)^{(k+2)} \right) \]

\[ \approx 1 - \exp \left( \frac{1}{2n} (k+1)(k+2) \right) + 1 - \exp \left( \frac{1}{2} t^2 \right), \]

which is Smirnov's result.

3. RELATED RESULTS

The characteristic function of \( C_n \) may be obtained from (2.12) by substituting \( s = e^{it} \) and noting that \( S_{jn} = 0 \) for \( j > n \). We get

\[ \Phi_n(t) = \sum_{j=0}^{n-1} (e^{it} - 1)^j S_u, \quad (3.1) \]

and calculate the central moments as

\[ E(C_n^r) = \frac{\Phi_n^{(r)}(0)}{1^r} = \sum_{j=0}^{n-1} \left( \sum_{k=j}^{j} \binom{j}{k} (-1)^{j-k} k^r \right) S_{jn}, \quad (3.2) \]

where we can use the known result

\[ \sum_{k=0}^{j} \binom{j}{k} (-1)^{j-k} k^r = \begin{cases} 0 & \text{if } r < j, \\ j & \text{if } r = j. \end{cases} \quad (3.3) \]
In general, we get the recursion relation

\[ S_k = \sum_{j=k}^{n-1} n^j p_j S_{k-1} \quad k = 1, 2, 3, \ldots, n-1, \quad (3.4) \]

where \( S_0 = 1 \) for all \( n \).

The characteristic function of \( C_n \) may be calculated from

\[ \phi_n(t) = E(e^{itC_n}) = \sum_{m=0}^{n-1} \sum_{k=0}^{n-1-m} \binom{m+k}{m}(-1)^k i^m e^{itm} S_{m+k}; \quad (3.5) \]

and by a change of variable \( u = m+k \), and a change in the order of summation we get

\[ \phi_n(t) = \sum_{u=0}^{n-1} (e^{it}-1)^u S_u. \quad (3.6) \]

Using (3.4), we get the form

\[ \phi_n(t) = 1 + (e^{it}-1) \sum_{j=1}^{n-1} n^j p_j \phi_j(t). \quad (3.7) \]

From (3.6), the mean and variance of \( C_n \) are

\[ E(C_n) = S_1 \quad (3.8) \]

\[ Var(C_n) = S_1(1-S_1) + 2S_2, \]

and moments may be calculated using (3.5) to get

\[ E(C_n^r) = \frac{\phi_r(0)}{ir} = \sum_{u=0}^{n-1} \left( \sum_{m=0}^{u} \binom{u}{m}(-1)^{u-m} r^m \right) S_u. \quad (3.9) \]
A table of the exact distribution of $C_n$ up to $n = 30$ is given in table 1. In some goodness-of-fit tests, convergence to the asymptotic distribution is rapid and the asymptotic distribution may be used for small sample sizes. In the case of the Cramer-Von Mises statistic, Marshall [2] has shown that the asymptotic distribution may be used for sample sizes of three or four. In the case of $C_n$, convergence is slow and it is not until $n > 100$ that the exact and asymptotic distributions become reasonably close. In fact, calculating

$$P(C_n + 1 \leq \sqrt{n} t) = P(C_n \leq \sqrt{n} t - 1) = P(C_n \leq k)$$

for $k = 0, 1, 2, \ldots, n - 1$; and comparing these values with $1 - e^{-t^2/2}$, where $t = (k + 1)^2/n$ for $k = 0, 1, \ldots, n - 1$, we find the maximum difference between the exact and asymptotic cumulative distributions decreases as follows:

<table>
<thead>
<tr>
<th>n</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>10</th>
<th>20</th>
<th>30</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maximum difference</td>
<td>0.3679</td>
<td>0.2912</td>
<td>0.2315</td>
<td>0.2146</td>
<td>0.1469</td>
<td>0.1013</td>
<td>0.0822</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>n</th>
<th>40</th>
<th>50</th>
<th>60</th>
<th>100</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maximum difference</td>
<td>0.0708</td>
<td>0.0632</td>
<td>0.0575</td>
<td>0.0443</td>
</tr>
<tr>
<td>Sample size</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>-------------</td>
<td>----</td>
<td>----</td>
<td>----</td>
<td>----</td>
</tr>
<tr>
<td>2</td>
<td>5000</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>3333</td>
<td>7777</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>2500</td>
<td>6250</td>
<td>9062</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>2000</td>
<td>5200</td>
<td>8080</td>
<td>9616</td>
</tr>
<tr>
<td>6</td>
<td>1666</td>
<td>4444</td>
<td>7222</td>
<td>9074</td>
</tr>
<tr>
<td>7</td>
<td>1429</td>
<td>3877</td>
<td>6501</td>
<td>8500</td>
</tr>
<tr>
<td>8</td>
<td>1250</td>
<td>3437</td>
<td>5828</td>
<td>7949</td>
</tr>
<tr>
<td>9</td>
<td>1111</td>
<td>3086</td>
<td>5590</td>
<td>7459</td>
</tr>
<tr>
<td>10</td>
<td>1000</td>
<td>2800</td>
<td>4960</td>
<td>6976</td>
</tr>
<tr>
<td>11</td>
<td>0909</td>
<td>2562</td>
<td>4591</td>
<td>6538</td>
</tr>
<tr>
<td>12</td>
<td>0833</td>
<td>2361</td>
<td>4271</td>
<td>6181</td>
</tr>
<tr>
<td>13</td>
<td>0769</td>
<td>2189</td>
<td>3992</td>
<td>5840</td>
</tr>
<tr>
<td>14</td>
<td>0714</td>
<td>2041</td>
<td>3746</td>
<td>5533</td>
</tr>
<tr>
<td>15</td>
<td>0666</td>
<td>1911</td>
<td>3529</td>
<td>5255</td>
</tr>
<tr>
<td>16</td>
<td>0625</td>
<td>1797</td>
<td>3355</td>
<td>5001</td>
</tr>
<tr>
<td>17</td>
<td>0588</td>
<td>1695</td>
<td>3161</td>
<td>4770</td>
</tr>
<tr>
<td>18</td>
<td>0555</td>
<td>1605</td>
<td>3064</td>
<td>4559</td>
</tr>
<tr>
<td>19</td>
<td>0526</td>
<td>1523</td>
<td>2862</td>
<td>4365</td>
</tr>
<tr>
<td>20</td>
<td>0500</td>
<td>1450</td>
<td>2732</td>
<td>4186</td>
</tr>
</tbody>
</table>
### TABLE 1 (CONT'D)

**CUMULATIVE PROBABILITIES**

|    | 0   | 1   | 2   | 3   | 4   | 5   | 6   | 7   | 8   | 9   | 10  | 11  | 12  | 13  | 14  | 15  | 16  | 17  | 18  | 19b |
|----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| 21 | 0.476 | 1383 | 2614 | 4021 | 5445 | 6746 | 7831 | 8657 | 9233 | 9598 | 9809 | 9918 | 9969 | 9989 | 9997 | 9999 |
| 22 | 0.435 | 1222 | 2506 | 3868 | 5262 | 6554 | 7650 | 8505 | 9116 | 9518 | 9759 | 9890 | 9955 | 9984 | 9995 | 9999 |
| 23 | 0.435 | 1267 | 2406 | 3726 | 5090 | 6371 | 7476 | 8354 | 8998 | 9434 | 9704 | 9859 | 9939 | 9976 | 9992 | 9997 | 9999 |
| 24 | 0.416 | 1215 | 2313 | 3594 | 4929 | 6197 | 7306 | 8204 | 8877 | 9345 | 9645 | 9823 | 9919 | 9966 | 9987 | 9996 | 9999 |
| 25 | 0.400 | 1168 | 2229 | 3471 | 4777 | 6031 | 7142 | 8057 | 8756 | 9254 | 9582 | 9783 | 9890 | 9954 | 9982 | 9993 | 9998 | 9999 |
| 26 | 0.385 | 1124 | 2148 | 3356 | 4634 | 5872 | 6984 | 7912 | 8635 | 9160 | 9515 | 9739 | 9869 | 9940 | 9975 | 9990 | 9997 | 9999 |
| 27 | 0.370 | 1084 | 2074 | 3249 | 4499 | 5721 | 6831 | 7770 | 8513 | 9064 | 9445 | 9692 | 9840 | 9923 | 9966 | 9986 | 9995 | 9998 | 9999 |
| 28 | 0.357 | 1046 | 2005 | 3147 | 4371 | 5577 | 6683 | 7631 | 8392 | 8966 | 9572 | 9641 | 9808 | 9904 | 9955 | 9981 | 9992 | 9997 | 9999 |
| 29 | 0.345 | 1011 | 1941 | 3052 | 4250 | 5440 | 6541 | 7495 | 8272 | 8868 | 9297 | 9588 | 9772 | 9882 | 9943 | 9975 | 9989 | 9996 | 9999 |
| 30 | 0.333 | 0977 | 1880 | 2965 | 4136 | 5308 | 6403 | 7362 | 8154 | 8769 | 9220 | 9532 | 9735 | 9859 | 9929 | 9967 | 9986 | 9994 | 9998 |

---

*Decimal point omitted.*

*bValues larger than .9999 omitted.*
4. A GOODNESS-OF-FIT TEST

We use $C_n$ as our test statistic and note that if the hypothesized distribution $F(x)$ is correct we expect $C_n$ to be large. Thus, acceptance regions coincide with large values of $C_n$. We divide the interval $(0,1)$ into $n$ equal cells of length $1/n$. The $i^{th}$ step of the empirical distribution function crosses $F(x)$ if and only if there are exactly $i$ observations less than $F^{-1}(i/n)$. This occurs if and only if there are exactly $i$ transformed observations $F(x)$ in the first $i$ cells of length $1/n$. Thus, if we denote the cell frequencies as $f_i$, $C_n$ will be equal to the number of times

$$f_1 + f_2 + \ldots + f_i = i \text{ for } i = 1, 2, \ldots, n-1. \quad (4.1)$$

To apply the test we need simply transform the sample $X_1, X_2, \ldots, X_n$ to $F(X_1), F(X_2), \ldots, F(X_n)$ and count according to (4.1). Critical points and significance levels may be calculated from (2.13).
REFERENCES
