EXACT NULL DISTRIBUTIONS AND ASYMPTOTIC EXPANSIONS
FOR RANK TEST STATISTICS

BY

WARREN F. ROGERS

TECHNICAL REPORT NO. 145
DECEMBER 26, 1971

SUPPORTED BY THE ARMY AND NAVY
UNDER CONTRACT N00014-67-A-0112-0052 (NR-042-002)
WITH THE OFFICE OF NAVAL RESEARCH

Reproduction in Whole or in Part is Permitted
for any purpose of the United States Government
Distribution of this Document is Unlimited

DEPARTMENT OF OPERATIONS RESEARCH
AND
DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
EXACT NULL DISTRIBUTIONS AND ASYMPTOTIC EXPANSIONS
FOR RANK TEST STATISTICS

by

WARREN F. ROGERS

TECHNICAL REPORT NO. 145
DECEMBER 26, 1971

Supported by the Army and Navy
with the Office of Naval Research

Gerald J. Lieberman, Project Director

Reproduction in whole or in Part is Permitted
for any purpose of the United States Government

Distribution of this document is unlimited

DEPARTMENT OF OPERATIONS RESEARCH
AND
DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
CONTENTS

<table>
<thead>
<tr>
<th>CHAPTER I - INTRODUCTION</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>CHAPTER II - DERIVATION OF GENERATING FUNCTIONS</td>
<td>4</td>
</tr>
<tr>
<td>Section 1</td>
<td>4</td>
</tr>
<tr>
<td>Section 2 - Linear Rank Statistics</td>
<td>5</td>
</tr>
<tr>
<td>Section 3 - Tests of Symmetry</td>
<td>9</td>
</tr>
<tr>
<td>Section 4 - Two Sample Tests of Location and Scale</td>
<td>11</td>
</tr>
<tr>
<td>Section 5 - Tests for Regression</td>
<td>14</td>
</tr>
<tr>
<td>Section 6 - Solution of Recurrence Relations</td>
<td>16</td>
</tr>
<tr>
<td>CHAPTER III - INVERTING ALGORITHMS</td>
<td>26</td>
</tr>
<tr>
<td>Section 1</td>
<td>26</td>
</tr>
<tr>
<td>Section 2 - Tests of Symmetry</td>
<td>27</td>
</tr>
<tr>
<td>Section 3 - Two Sample Tests</td>
<td>30</td>
</tr>
<tr>
<td>Section 4 - Tests of Regression</td>
<td>40</td>
</tr>
<tr>
<td>Section 5 - The Kolmogorov-Smirnov Test</td>
<td>44</td>
</tr>
<tr>
<td>CHAPTER IV - LOCAL LIMIT THEOREMS AND ASYMPTOTIC EXPANSIONS</td>
<td>46</td>
</tr>
<tr>
<td>Section 1</td>
<td>46</td>
</tr>
<tr>
<td>Section 2 - Preliminaries</td>
<td>47</td>
</tr>
<tr>
<td>Section 3 - Local Limit Theorems</td>
<td>73</td>
</tr>
<tr>
<td>Section 4 - Expansions for Distributions</td>
<td>81</td>
</tr>
<tr>
<td>Section 5 - Error Bounds</td>
<td>83</td>
</tr>
<tr>
<td>LIST OF REFERENCES</td>
<td>87</td>
</tr>
<tr>
<td>APPENDICES</td>
<td></td>
</tr>
<tr>
<td>1 - Computer Programs</td>
<td>88</td>
</tr>
<tr>
<td>2 - Sample Sizes Required to Achieve Specified Accuracy</td>
<td>103</td>
</tr>
</tbody>
</table>
CHAPTER I

Introduction

Tabulation of distributions of rank test statistics is generally accomplished either by direct enumeration of all possibilities or by reducing a recurrence relation. Both methods can be excessively time consuming when sample sizes are moderate. An alternative method is developed here which is applicable to a wide class of rank test statistics. The approach taken is to develop efficient inverting algorithms for the appropriate probability generating functions.

In Chapter II probability generating functions are derived for one and two sample scores statistics and the Kolmogorov-Smirnov test statistic. The method used is to demonstrate that the enumeration of rank vectors which give rise to a distinct value of a rank test statistic is generally equivalent to enumerating a restricted partition or composition of that value. The equivalence is established directly for linear rank statistics and, in two non-linear examples, by demonstrating the equivalence of enumerating recurrence relations. Enumerating generating functions are then derived by well known methods. In most cases the generating functions so derived are not in
closed form. For purposes of enumeration, however, the open
form is adequate. In 1956, Hodges and Fix (3) noted the simi-
lariry of the two sample Wilcoxon statistic to a partition
function which had been tabled by Euler. They did not further
exploit this discovery and their tabulation method consisted
essentially of reducing a recurrence.

Efficient inverting algorithms are developed in Chapter
III. Examples are given of their use in tabling the one sample
Normal Scores (Fraser), the two sample Wilcoxon, the median test,
the Spearman rank correlation, and the one sided Kolmogorov-
Smirnov test statistics. Computer programs are given in Appendix
I for tabling one and two sample scores statistics, general
scores regression test statistics of which the Spearman test is
a special case and the one and two sided Kolmogorov-Smirnov
tests.

In Chapter IV an approximation to the two sample Wilcoxon
distribution by several terms of an Edgeworth series is examined.
It is shown that the expansion to many terms provides a valid
asymptotic expansion for the density and that the one term ex-
pansion is valid for the distribution function. An explicit
bound for the error in approximating the distribution by the
normal and one correction term is given. An improved measure
of the rate of convergence of the distribution to the limit-
ing normal distribution is also derived.

Minimum sample sizes at which 2D, 3D, and 4D accuracy were achieved in approximating with the uncorrected Normal and the Normal with one and with three correction terms are given in Appendix II.
CHAPTER II

Derivation of Generating Functions

1. Generating functions for some rank statistics have been derived by means of recurrence by Kendall and others. Van Dantzig (10) rederived the generating functions of the Kendall rank correlation and Wilcoxon two sample statistics by an ingenious application of his theory of collective marks. Neither method, however, appears capable of extension to more general cases.

The methods developed here are applicable to a very wide class of rank statistics. The approach is to show that enumerating rank vectors which give rise to a particular value of a statistic is equivalent to enumerating certain restricted partitions of that value. There is a very extensive literature on the theory of restricted partitions, some of which is repeated here in detail for clarity of exposition. An excellent survey of the theory is given by Riordan (8).

We will consider samples of independent identically distributed variables $X_1, X_2, \ldots, X_N$ drawn from a continuous distribution. Let
\[ R_i \stackrel{\text{def}}{=} \text{The rank of } x_i \text{ in the ordered sample } x(1) < x(2) < \ldots < x(N) \]

\[ R^+_i \stackrel{\text{def}}{=} \text{The rank of } |x_i| \text{ in the ordered sample of absolute values } |x|(1) < |x|(2) < \ldots < |x|(N). \]

2. **Linear Rank Statistics.** Following Hajek (4), we designate as linear rank statistics those functions of a rank vector which are expressible as

\[
S = \sum_{i=1}^{N} a(i, R_i),
\]

where \( \{a(i, j)\} \) is an arbitrary \( N \times N \) matrix. By a partition of a real number \( S \) into \( N \) parts from a set \( \{b_i\} \), we mean a solution in integers, \( n_1, n_2, \ldots, n_N \), of the equation,

\[
b_1 n_1 + b_2 n_2 + \ldots + b_N n_N = S.
\]

It is clear, in this case, that enumerating all vectors of ranks which yield a value of \( S \) in (1), is equivalent to enumerating the partitions of \( S \) into \( N \) parts from the set \( \{a(i, j)\} \), subject to the constraint that only one element from each row and column is admissible in any partition.
The study of partitions is facilitated by the introduction of enumerating generating functions defined as,

\[ \phi(t) = \sum_{k} c_k t^k, \]

where \( c_k \) is the number of partitions of \( k \) into some specified parts and summation is over all realizable sums of elements from the specified part set. Such generating functions may frequently be derived by inspection. Thus the generating function for partitions with no restriction as to repetition of parts is,

\[ \phi(t) = (1 + t + t^2 + \ldots)(1 + t^2 + t^4 + \ldots) \ldots \\
= 1/(1 - t^b_1)(1 - t^b_2)(1 - t^b_3) \ldots. \]

This follows immediately from the observation that each term in the product represents the number of appearances in the partition of the indicated part. The generating function for partitions into parts from a set of \( N \) is,

\[ \phi_N(t) = 1/(1 - t^{b_1})(1 - t^{b_2}) \ldots (1 - t^{b_N}). \]

The generating function for partitions into distinct parts from a set of \( N \) is,
\[
\phi'_N(t) = (1 + t^{-1})(1 + t^2) \ldots (1 + t^{-N})
\]

By distinct parts we mean distinct in the indices of the parts \( b_i \). Some or all of the \( b_i \) could in fact be identical. There is a subtle distinction which must be made in this regard between the generating functions of (5) and (6). By repetition of parts in (4) and (5) we again mean repetition of indices. Thus, in determining those partitions of \( S \) whose enumeration is equivalent to the enumeration of the rank vectors which yield \( S \) in (1), it would clearly be inappropriate to treat the simultaneous appearance of two admissible scores, which happened to be identical, as the repeated (and thus inadmissible) appearance of one of them. That the distinction is real may be seen by considering the case \( N=2 \) and the identical parts \( a_1 \) and \( a_2 \). Considered as repetitions of the part \( a_1 \), (5) would yield

\[
\phi(t) = 1 + t a_1 + 2a_1 t^2
\]

whereas (6) would yield

\[
\phi'(t) = (1 + t^{-1})(1 + t^2) = (1 + t^{-1})^2 = 1 + 2t^{-1} + t^2
\]

The partition equivalents of sums of the form in (1) are usually further restricted in that the number of parts is
specified. Lacking additional conditions on the form of the summands, it is generally impossible to proceed further with explicit derivation of generating functions. Nonetheless, it will be shown in Chapter III that with even this limited specification, it is possible to tabulate a wide range of statistics efficiently.

A composition is a partition in which the order of parts is distinguished. Compositions with restrictions as to number of repetitions of parts, etc., can also be defined. The generating function for compositions of \( N \) parts from a set \( \{a_i\}_{i=1}^{M} \) with unlimited repetitions of the same part is

\[
\Phi(t) = (t^{a_1} + t^{a_2} + \ldots + t^{a_M})^N.
\]

Notice that even if \( a_1 = a_2 \) we count \( a_1 + a_2 \) distinct from \( a_2 + a_1 \). The \( k^{\text{th}} \) term in the product corresponds to the appearance of one of the parts in the \( k^{\text{th}} \) position in the composition.

The general theory of partitions and compositions typically addresses problems involving only positive parts. In the present context it will be occasionally necessary to consider non-positive parts. Negative parts present no difficulty in partitions
or compositions which are restricted as to number and range of parts. The treatment of zero as a part, however, requires special care, particularly when generating functions are employed. The enumerating generating function for compositions (7) presents no difficulty. Here $t^0$ may be interpreted as unity and its presence in the $k^{th}$ term of the product may be interpreted as the appearance of zero in the $k^{th}$ position in the composition. If, however, one or more of the summands (scores), $a(i, R_i)$, in (1) are zero, we are forced to consider partitions into exactly $N$ parts some of which may be zero. The alternative formulation would be to consider partitions into at most $N$ and at least $N-k$ parts, where $k$ is the number of zero scores. In practice, however, it is difficult to incorporate the latter specification in a generating function. Instead we treat zero as we would any other score with the understanding that $t^0$ must not be interpreted as unity.

3. **Tests of Symmetry.** The usual rank test statistics for symmetry are of the form

\[
S^+_N = \sum_{\chi_i > 0} a_N(R_i^+).
\]
Letting \( a_i = a_{iN}(R_i^+) \), \( i = 1, 2, \ldots, N \) we may identify the enumeration of those rank vectors yielding \( S_N^+ \) in (7) with the enumeration of the partitions of \( S_N^+ \) into distinct parts from the set \( \{a_i\}_{i=1}^{N} \). By distinct here we again refer to the indices \( i = 1, 2, \ldots, N \). Some or all of the \( a_i \) may be identical. The probability generating function for \( S_N^+ \),

\[
P_N(t) = \frac{1}{2^N} \prod_{k=1}^{N} (1 + t^{a_k})
\]

(9) follows from (6). Also, 

\[
E(S_N^+) = \frac{1}{2} \sum_{j=1}^{N} a_j'
\]

(10) so that substituting in (8) yields the characteristic function for \( S_N^+ - E(S_N^+) \),

\[
\eta_N(t) = \frac{1}{2^N} \exp \left( -\frac{it}{2} \sum_{j=1}^{N} a_j \right) \prod_{k=1}^{N} \left( 1 + \frac{ia_k t}{2} \right)
\]

(11) 

\[
= \prod_{k=1}^{N} e^{-\frac{ia_k t}{2} + \frac{ia_k t}{2}}
\]
Specializing to the case,

\[ a_n(R_i^+) = R_i^+ \]

yields the one sample Wilcoxon test and (11) simplifies to the well known form,

\[ \psi_N(t) = \prod_{k=1}^{N} \cos\left(\frac{kt}{2}\right). \]

4. Two Sample Tests. In this case the sample is drawn from two distributions \( X_1, \ldots, X_m \) from the first, \( X_{m+1}, \ldots, X_N \) from the second. Under the null hypothesis the distributions are identical. The statistic takes the form,

\[ S_{m,n} = \sum_{i=1}^{m} a_n(R_i), \quad m + n = N. \]

The appropriate partitions of \( S_{m,n} \) are restricted to have exactly \( m \) parts from the set \( a_1, \ldots, a_N \), where \( a_j = a_n(j) \). To obtain the enumerator we introduce Euler's two variable generating function,

\[ G(u,t) = (1 + ut^1)(1 + ut^2) \cdots (1 + ut^N) \]

\[ = \sum_{k} \phi(t)u^{k_{i,n}} \]
The enumerating generating function for the restricted partitions of $S_{m,n}$ is then the coefficient of $u^m$ in (14). We denote the probability generating function by

$$P_{m,N}(t) = \frac{1}{\binom{N}{m}} \prod_{k=1}^{N} \left( 1 + ut^k \right) \left| \frac{a_k}{u^m} \right.$$  

If we specialize to the case $a_N(R_i) = R_i$ the following recurrence is possible,

$$G(u,t) = (1 + ut^{N+1}) G(u,t) = (1 + ut) G(ut,t).$$

Equating coefficients of $u^m$ yields,

$$\phi(t) = \frac{t^m - t^{N+1}}{1 - t^m} \phi(t)_{m-1,N}.$$  

This, with the boundary condition $\phi(t) = 1$, yields the probability generating function,

$$P_{m,N}(t) = \frac{1}{\binom{N}{m}} \prod_{k=1}^{m} \left( 1 - t^{n+k} \right),$$

in agreement with Kendall and Van Dantzig.

The preceding technique is unfortunately restricted to this one example. It depends critically on the form of the scores and is applicable only when succeeding scores differ by a constant.
Any test with this structure is of course equivalent to the two sample Wilcoxon test described above.

For the general case we may proceed as follows. Taking partial derivatives in the first line of (14) yields

\[
\frac{\partial}{\partial u} G(u,t) = \sum_{k=1}^{N} t^{a_k} G(u,t) \frac{a_k}{(1+ut)^k}
\]

\[
= \sum_{k=1}^{N} t^{a_k} G(u,t) \left[ 1 - ut^{a_k} + ut^{2a_k} - ut^{3a_k} + \ldots \right]
\]

\[
= \sum_{j=0}^{\infty} \phi(t)u^j \sum_{k=1}^{N} t^{a_k} \sum_{i=0}^{j} (-1)^i i^{ia_k}
\]

The second line of (14) yields

\[
\frac{\partial}{\partial u} G(u,t) = \sum_{j=0}^{\infty} j \phi(t)u^{j-1}.
\]

Equating coefficients of \(u^m\) then yields the recurrence,

\[
\phi(t) \sum_{m,N}^{\infty} t^{a_k} - \phi(t) \sum_{m-1,N}^{\infty} t^{2a_k} + \ldots + (-1)^m \sum_{k=1}^{N} t^{(m+1)a_k} = (m+1) \phi(t).
\]
with initial conditions,

\[ \phi(t) = 1, \]

\[ \phi(t) = t_1 + t_2 + \ldots + t_N. \]

5. **Tests for Regression.** For testing against regression alternatives the general linear form is,

\[ S_N = \sum_{i=1}^{N} C_i a_N(R_i), \]

where the \( C_i \) are regression constants. The case \( C_i = 1, \ i = 1, 2, \ldots, m, \)
\( C_i = 0, \ i = m+1, \ldots, N \) is the two sample problem of the preceding section. For more general constants the problem is complicated by the fact that it is not generally possible to form a recurrence relation.

The sum in (21) is not entirely analogous to a composition but can be treated similarly. The part set consists of the \( N \times N \)-matrix, \( C_i a_N(R_i), \ i = 1, 2, \ldots, N \). The summands are restricted in that only one entry from any row or column is permitted. By analogy with (7) we write the generating function,
\[
\phi(t) = (t^{C_1 a_1} + t^{C_1 a_2} + \ldots + t^{C_1 a_N})(t^{C_2 a_1} + t^{C_2 a_2} + \ldots + t^{C_2 a_N})
\]
\[
\ldots (t^{C_N a_1} + t^{C_N a_2} + \ldots + t^{C_N a_N})
\]

thus incorporating the restriction that for each \(i\), \(C_i\) may appear only once in the sum. To apply the same restriction to the scores \(a_N(R_i) \equiv a_i\) we extend the Euler function of (14) to,

\[
G(t, d_1, \ldots, d_N) = \prod_{k=1}^{N} \left(d_1 t^{C_k a_1} + d_2 t^{C_k a_2} + \ldots + d_N t^{C_k a_N}\right).
\]

The required generator will then be the coefficient of \(d_1 \cdot d_2 \cdot \ldots \cdot d_N\) in (23). We denote the probability generating function by

\[
P_N(t) = \frac{1}{N!} \prod_{k=1}^{N} \left(\sum_{j=1}^{N} d_j t^{C_k a_j}\right)
\]

\[
d_1 \cdot d_2 \cdot \ldots \cdot d_N
\]

Even in the case,

\[
a_N(R_i) = R_i, C_i = i,
\]

which yields one form of the Spearman rank correlation statistic, it is not clear that any further simplification is possible.

The \(N^{th}\) partial derivative, \(\frac{\partial^N}{\partial d_1 \cdots \partial d_N} \) in (24) will yield the
desired coefficients but is operationally equivalent to enumerating all rank vectors. For tabulation purposes, however, the form in (24) is adequate. An inversion algorithm is given in Chapter III.

6. Solution of Recurrence Relations. Partitions and compositions may generally be enumerated by means of recurrence relations. Let \( A = \{a_i\}_{i=1}^N \) be a sequence of real numbers and \( A_N = \{a_i\}_{i=1}^N \). Let \( \pi(z, m, A_N) \) be the number of partitions of \( z \) into \( m \) distinct parts, (distinct in their indices), from the set \( A_N \). Then considering separately those partitions which contain \( a_N \) and those which do not yields

\[
\pi(z, m, A_N) = \pi(z-a_N, m-1, A_{N-1}) + \pi(z, m, A_{N-1}).
\]

Initial and boundary conditions are determined by the form of the parts. For example, let

\[
A_N = 1, 2, \ldots, N,
\]

then

\[
\pi(0, 0, A_N) = 1
\]

\[
\pi(z, m, A_N) = 0, \quad z < \frac{1}{2}m(m+1)
\]

\[
\pi(z, 0, A_N) = 0, \quad z > 0
\]

\[
\pi(z, m, A_N) = 0, \quad N < m.
\]
For the number of partitions of \( z \) into at most \( m \) parts with possible repetition of parts, the recurrence is again formed by distinguishing those partitions which contain \( a_N \) to yield,

\[
\pi(z, m, A_N) = \pi(z-a_N, m-1, A_m) + \pi(z, m, A_{N-1}).
\]

Alternatively if \( \pi(z, m, A_N) \) is the number of compositions of \( z \) into \( m \) parts from \( A_N \) and we permit repetition of parts we may consider separately those compositions whose first part is \( a_1 \) or \( a_2 \) or ... or \( a_N \) to yield

\[
\pi(z, m, A_N) = \sum_{i=1}^{N} \pi(z-a_i, m-1, A_n).
\]

Parts may be identical and or non-positive. Additional restrictions as to the order of parts or the form of partial sums of parts are incorporated as boundary conditions.

Recurrence relations also enter naturally in the study of statistics based on ranks. In K-sample tests for location the probability-generating recurrence is formed by considering separately which of the K-samples contributed the largest (or smallest) observation. In the case of paired observations, one may consider separately the possible values of the second coordinate rank of the pair with the largest first coordinate.
Tests of the Kolmogorov-Smirnov type can usually be related to a random walk and the recurrence can be derived by considering separately whether a specified step was up or down.

Kendall's method is to solve recursively the functional equation induced by the recurrence on the generating function for the statistic in question. I have had some success with the following approach: Given the probability-generating recurrence, it is often possible to "guess" a corresponding restricted partition or composition and to verify that it is enumerated by the same recurrence. In general, a statistic whose enumeration is equivalent to enumerating a partition or composition must have a representation as a linear form which could be treated as in the preceding cases. In practice, however, I have found it easier to infer the form of the equivalent partitions or compositions from the recurrence, particularly in the case of Kolmogorov-Smirnov statistics and in K-sample problems which are not treated here.

When the equivalence is established it is usually possible to derive the corresponding enumerating generating function at least in an open form such as (24). It will be shown in Chapter III that it is generally more efficient to invert a generating function, even if in open form, than it is to reduce a recurrence.
EXAMPLE 1. The Mann-Whitney form, $U$, of the Wilcoxon two sample statistic has

\begin{equation}
(25) \quad P(U = K) = \frac{1}{\binom{N}{m}} \pi(K, m, n),
\end{equation}

where

\begin{equation}
(26) \quad \pi(K, m, n) = \pi(K-m, m, n-1) + \pi(K, m-1, n) \\
\pi(K, 0, n) = \pi(K, m, 0) = 0 \quad K > 0 \\
\pi(K, m, n) = 0 \quad K < 0 \\
\pi(0, m, n) = 1
\end{equation}

Now $U$ is clearly a linear rank statistic and could be treated by the methods of section 2.4. However, the recurrence relation (31) is the enumerating relation for partitions of the argument $K$ into at most $n$ parts from the set 1, 2, ..., $m$. To see that this is so, we form the partition recurrence by considering separately those partitions which contain $m$ and those which do not. The first term on the right of (26) and the first boundary condition imply that there are at most $n$ parts. The probability generating function may be derived by an induction analogous to that of (17) to yield
(27) \[ p_{m,n}'(t) = \frac{1}{\binom{n}{m}} \left( 1 + \sum_{k=1}^{n} \sum_{j=1}^{k} \frac{k!}{(1-t)^{m+j-1}} \right) \]

\[ = t^{-(m+1)} \binom{m+1}{2} p_{m,n}(t) \]

where \( p_{m,n}(t) \) is defined in (17).

Other forms of recurrence are discussed in the following examples.

EXAMPLE 2. Hajek (4) shows that the statistic \( T_N \) related to the Kendall rank correlation statistic \( \tau \) by,

\[ \tau = \frac{4T_N}{N(N-1)} - 1, \]

satisfies the following relations.

(28) \[ P(T_N = k) = \frac{\pi_N(k)}{N!}, k = 0, 1, \ldots, \frac{1}{2}N(N-1) \]

with

(29) \[ \pi_N(k) = \pi_{N-1}(k) + \pi_{N-1}(k-1) + \ldots + \pi_{N-1}(k-N+1) \]

and

\[ \pi_{N}(0) = 1, \]

\[ \pi_{N}(k) = 0, k > 0 \]

\[ \pi_{N}(k) = 0, k < 0. \]
Each term on the right of (29) has the index $N$ reduced by 1. This suggests that the recurrence may be equivalent to the enumerating relation for a composition since in forming such a relation one distinguishes compositions by their first part. A composition which satisfied the recurrence (29) could have as its first part any one of $0,1,2,\ldots,N-1$ and as its $j^{th}$ part any one of $0,1,2,\ldots,N-j$. The initial condition restricts the number of parts to be $N-1$ (counting zeroes as parts). Summarizing then, we see that (29) is the enumerating recurrence relation for the compositions of $K$ into $N-1$ parts from the set $0,1,2,\ldots,N-1$, subject to the constraint that the $j^{th}$ part must be one of $0,1,2,\ldots,N-j$.

The probability generating function follows immediately.

\begin{equation}
(30) \quad P_n(t) = \frac{1}{N!} \left(1 + t + t^2 + \ldots + t^{N-1}\right) \left(1 + t + \ldots + t^{N-2}\right) \ldots \left(1 + t\right)
\end{equation}

\begin{equation}
= \frac{1}{N!} \prod_{j=1}^{N} \frac{1-t^j}{1-t}
\end{equation}

in agreement with Kendall's result.

EXAMPLE 3. Hajek (4) shows that the one sided Kolmogorov-Smirnov statistic is a multiple of $\max_{1 \leq k \leq m+n} (\frac{km}{m+n} - C_{d1} - \ldots - C_{dk})$. 

21
where $C_{d_k}$ = 1 whenever the $k^{th}$ order statistic $x^{(k)}$ belongs to
the first sample and $C_{d_k}$ = 0 whenever $x^{(k)}$ belongs to the second
sample and that

$$\Pr\left[ \max_{1 \leq k \leq m+n} \left( \frac{km}{m+n} - C_{d_k} \right) \leq \frac{h}{m+n} \right] \overset{\text{def}}{=} \Pr\left[ \frac{k}{m, n} \leq \frac{h}{m+n} \right]$$

$$= \binom{m+n}{m}^{-1} \sum_{i=h-mn}^{h} \pi^+_m(i)$$

where $h$ is an integer, $0 \leq h \leq mn$ and $\pi^+_m(i)$ satisfies the
following relations,

$$\pi^+_j(i) = \pi^+_j(i+n) + \pi^+_j(i+n-m) + \pi^+_j(i+n-2m) + \ldots + \pi^+_j(i+n-nm),$$

The initial and boundary conditions as stated by Hajek are:

$$\pi^+_0(h) = 1, \pi^+_0(i) = 0, i \neq h.$$ 

For $j \geq 1$, $\pi^+_j(i) \neq 0$ only for integers $i$ such that $i-h+nj$
is divisible by $m$ and

$$h - jn \leq i \leq 2h - n.$$ 

Again, as in Example 2, the reduction of the index $j$ by 1
in each term on the right of (32) suggests that the recurrence
may be equivalent to the enumerating relation for a composition. A composition which satisfied (32) would be formed from parts drawn from the set \(-n, m-n, 2m-n, \ldots, mn-n\). The initial and boundary conditions are satisfied by compositions of \(i-h\) into \(m\) parts, \(a_1, a_2, \ldots, a_m\), such that

\[
h - nj \leq h + \sum_{k=1}^{j} a_k \leq 2h - n.
\]

The condition that \(i-h+nj\) be divisible by \(m\) is automatically satisfied provided \(i-h+nm\) is divisible by \(m\).

Thus \(\pi^+_m(i)\) is equivalent to the number of compositions of \(i-h\) into \(m\) parts from the set \(-n, m-n, 2m-n, \ldots, mn-n\) such that each of the partial sums of parts satisfies,

\[
-nj \leq \sum_{i=1}^{j} a_i \leq h-n, \quad j = 1, \ldots, m.
\]

The enumerating generating function would then be

\[
(33) \quad \Phi_{h,m,n}(t) = \sum_{i=1}^{m} t^{i\pi^+_m(i)}
\]

\[
= t^h \left[ \prod_{j=1}^{m} (t^{-n} + t^{m-n} + t^{2m-n} + \ldots + t^{mn-n}) \right]_{-nj}^{h-n}
\]

23
The notation \( (-n_j)_{j-th}^{h-n} \) means that when the \( j \)-th term is multiplied into the "product" of the first \( j-1 \) terms only those terms with exponent between \(-n_j\) and \(h-n\) are retained. A closed form is obviously impossible, but for purposes of enumeration this is of no consequence as will be shown in Chapter III.

The cumulative probability generating function follows directly.

\[
P_{m,n}(t) = \sum_{n=0}^{mn} p\left(k_{m,n} \leq \frac{h}{m+n}\right) t^h
\]

\[
= \binom{m+n}{n}^{-1} \sum_{h=0}^{mn} t^h \sum_{i=h-mn}^{n^+} \binom{m}{i}^+(i)
\]

\[
= \binom{m+n}{n}^{-1} \sum_{h=0}^{mn} t^h \sum_{k=0}^{n^+} \binom{m}{h-mn+km}
\]

The final equality follows from the divisibility condition.

If, as before, the coefficient of \( u^h \) in \( G(u) \) is denoted by \( G(u)_{h} \), the cumulative probability generating function may be written

\[
P_{m,n}(t) = \binom{m+n}{m}^{-1} \sum_{h=0}^{mn} \left[ \phi_{h,m,n}(u)/(1-u^m) \right]_{u=h} \cdot t^h
\]
The two sided form follows analogously with slight modification of the lower limit of summation in (14) and the bounds on the exponents in the product in (44). The generating function then becomes

\[(36) \quad p^\prime_{m,n}(t) = \binom{m+n}{m}^{-1} \sum_{h=0}^{\min(m,n)} \left[ \phi^\prime_{h,m,n}(u)/(1-u^m) \right]_{u^h} \cdot t^h\]

where \( \ell = \max(\min(m,n), \frac{1}{2}\max(m,n)) \) and

\[(37) \quad \phi^\prime_{h,m,n}(t) = t^h \prod_{j=1}^{m} \left( t^{-n} + t^{m-n} + \ldots + t^{mn-n} \right)_{u^{-h}} \]
CHAPTER III

Inverting Algorithms

1. The generating functions developed in Chapter II have a rather simple structure in common. They are generally in the form of products of simple polynomials with various restrictions on the way the products are taken. To determine the coefficients this property is exploited using the exponents in the component polynomials as place holders. Thus to determine the coefficients in

$$(1 + t + t^2 + \ldots t^N) \cdot (1 + t + t^2 + \ldots t^M)$$

one may write

\[
\begin{array}{cccccccc}
0 & 1 & 2 & \ldots & N & N+1 & \ldots & N+M \\
1 & 1 & 1 & \ldots & 1 & 0 & 0 & \ldots & 0 \\
1 & 1 & 1 & \ldots & 1 & 0 & 0 & \ldots & 0 \\
1 & 1 & 1 & \ldots & 1 & 0 & 0 & \ldots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\
1 & 1 & 1 & \ldots & 1 & \ldots & \ldots & \ldots & \ldots \\
1 & 2 & 3 & \ldots & N & N-1 & \ldots & 1
\end{array}
\]

Each of the $M+1$ lines is indented by 1 to represent multiplication by the next highest power of $t$.
2. **Tests of Symmetry.** The probability generating function, $p_N$, for one sample signed scores tests of symmetry is given by

$$(1) \quad 2^N p_N(t) = \prod_{k=1}^{N} (1 + t^{a_k})$$

where the scores $\{a_k\}_{k=1}^{N}$ are some arbitrary real numbers. In the simple (Wilcoxon) case $a_k = k$ we proceed as follows to obtain the coefficients of powers of $t$ on the right side of (1):

\[
\begin{array}{cccccc}
0 & 1 & 2 & 3 & 4 & 5 & \ldots & \frac{1}{2}N(N+1) \\
(l+t) & 1 & 1 \\
(l+t^2) & + & 1 & 1 \\
(l+t^3) & + & 1 & 1 & 1 \\
(l+t^4) & + & 1 & 1 & 1 & 1 \\
& & 1 & 1 & 1 & 2 & 1 & 1 & 1 \\
& & \vdots & \vdots & \vdots & \vdots & \vdots \\
(l+t^K) & \text{indent preceding line $K$ places and add.} \\
\end{array}
\]

At each stage in the process the coefficient of $t^K$ is represented by the appropriate entry in the $K^{th}$ column of the array.
To generalize to arbitrary scores it is merely necessary to represent the scores as $K$ digit integers where $K$ is the number of significant figures in the scores tabulation. Of course, if the scores are computed with very high precision, machine computation may be complicated by computer memory limitations. Unfortunately it is not possible in current computers to index an array without declaring an array size at least as large as the greatest index. This difficulty may be overcome by printing the first $a_K$ values at the $K^{th}$ stage of the computation and re-using that storage in subsequent calculations.

EXAMPLE 1. The Fraser (normal scores) test uses as scores the expected values of absolute normal order statistics. The tabulation here for $N=5$ is carried out with the scores computed by Klotz (7) but rounded off at the first decimal. Klotz computed to 5 figures. Scores are entered in the top row without parentheses, zero is not a score.
Thus

\[ P[S_N^+ = K] = \frac{1}{2^5}, \quad K = 0, 0.2, 0.4, 0.6, 0.7, 0.9, 1.0, 1.1, 1.2, 1.3, 1.4, 1.7, 1.8, 1.9, 2.0, 2.1, 2.2, 2.5, 2.6, 2.7, 2.8, 2.9, 3.0, 3.2, 3.3, 3.5, 3.7, 3.9 \]

\[ = \frac{1}{2^4}, \quad K = 1.6, 2.3 \]

The efficiency of the procedure in the general case is evident. The only alternative is to compute all \( 2^N \) possible values and order the results. In the one sample Wilcoxon case one has the alternative of reducing the recurrence given Hajek (4),
\( \pi_N(K) = \pi_{N-1}(K) + \pi_{N-1}(K-N), \)

\( \pi_N(0) = 1 \)

\( \pi_0(K) = 0 \quad K > 0 \)

\( \pi_N(K) = 0 \quad K < 0. \)

Computationally of course one starts with the initial and boundary conditions and builds the array,

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

so that in this case reducing the recurrence and inverting the generating function are equivalent. In the case of two or more samples, however, considerable economies are realized in that fewer unwanted tables need be computed.

3. **Two Sample Tests.** To invert the general form given in (2.14),

\[
\Phi_m(t) = \prod_{k=1}^{N} (1 + ut^k)^{a_k} \bigg|_{u_m}^N
\]
a two dimensional array \( \{ b_{i,j} \}_{i=1}^{m} \) is developed as follows. We first consider the case where all scores are positive and distinct. The scores are represented, as before, by integers without parentheses.

\[
\begin{array}{ccccccc}
0 & 1 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 1 & 1 \\
2 & 0 & 0 & 0 & 0 & 0 & 1 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
m & & & & & & \\
\end{array}
\]

First \( b_{0,0} \) and \( b_{1,a_1} \) are set to 1, all other entries to zero. Then all entries are shifted to the right \( a_2 \) units, down one unit and the resulting matrix added to the original. This procedure is repeated \( N \) times, at the \( k^{th} \) stage moving all entries to the right by \( a_k \), down 1 and adding to the \( k-1 \) stage matrix. Entries which would fall in rows below the \( m^{th} \) row are ignored.

At the conclusion the \( m^{th} \) row will contain the appropriate coefficients. This follows because the \( k^{th} \) movement to the right or left generates the new coefficients introduced by multiplying the first \( k-1 \) terms in the product by \( l+ut \) while the vertical movement registers the increase by 1 of the number of summands in each exponent (i.e., the power of \( u \)). Note that if the operation were carried
out for $N=m+n$ rows and the rows added, the resulting vector would contain the coefficients for the one sample case. This graphically illustrates the well known relation between the one and two sample statistics.

$$p\left[S^+_N = k\right] = \sum_{j=0}^{N} \binom{N}{j} 2^N p\left[S^+_{j,N-j} = k\right].$$

Zero and negative scores may be treated in either of two ways. A large enough constant may be added to every score so that a new and equivalent set of positive scores is generated. Alternatively if $a_k$ is negative (or zero) then at the $k$th stage all entries may be shifted $a_k$ units to the left rather than to the right and then down 1 unit as before. The effect when $a_k$ is zero is to register the increase by 1 in the exponent of $u$ while leaving the exponents of $t$ unchanged. Repeated scores require no adjustment in the procedure. Example 4 illustrates the application of the procedure to a case with repeated zero scores.

EXAMPLE 2. With $a_k=k$, $n=5$, $m=3$ to find $p(S_{m,n} = 11)$. Note that in this and all other examples, when the value at one specific argument is desired, the procedure may be truncated at that argument.

First Term: $1 + ut$

\[
\begin{array}{cccccccccccc}
0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 \\
0 & & & & & & & & & & & \\
1 & & & & & & & & & & & \\
2 & & & & & & & & & & & \\
3 & & & & & & & & & & & \\
\end{array}
\]
\[ 1 + ut^2 \]

\[
\begin{array}{cccccccccccc}
0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 \\
0 & 1 \\
1 & 1 & 1 \\
2 & 1 \\
3 & \\
\end{array}
\]

\[ 1 + ut^3 \]

\[
\begin{array}{cccccccccccc}
0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 \\
0 & 1 \\
1 & 1 & 1 & 1 \\
2 & 1 & 1 & 1 \\
3 & 1 \\
\end{array}
\]

\[ 1 + ut^4 \]

\[
\begin{array}{cccccccccccc}
0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 \\
0 & 1 \\
1 & 1 & 1 & 1 & 1 \\
2 & 1 & 1 & 2 & 1 & 1 \\
3 & 1 & 1 & 1 & 1 \\
\end{array}
\]

\[ 1 + ut^5 \]

\[
\begin{array}{cccccccccccc}
0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 \\
0 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 \\
2 & 1 & 1 & 2 & 2 & 2 & 1 & 1 \\
3 & 1 & 1 & 2 & 2 & 2 & 1 \\
\end{array}
\]
$1 + ut^6$

$1 + ut^7$

$1 + ut^8$

\[
P(t) = \frac{1}{\binom{8}{3}} \left[ t^6 + t^7 + 2t^8 + 3t^9 + 4t^{10} + 5t^{11} \ldots \right]
\]

\[
P \left[ S_{3,5} = 11 \right] = 0.08929
\]
It is interesting to compare this procedure with another which is suggested by the closed form for the special case $a_k = k$

$$
\phi (t) = \prod_{m,n}^{\infty} \frac{(1-t^{m+k})}{(1-t^k)}
$$

(4)

This procedure is similar to one developed by Kemperman (5) for this special case.

Representing the first term in the product by a line of $m+1$ 1's followed by $mn-m-1$ zeroes, the second term is incorporated as follows:

```
1 1 1 . . . 1 0 0 . . . . . . 0
1 1 1 . . . 1 0 0 . . . . . . 0
1 1 1 . . . 1 0 0 . . . . . . 0
 . . . . . . . . . . . . . . . . . . .
 . . . . . . . . . . . . . . . . . . .
 . . . . . . . . . . . . . . . . . . .
1 1 2 2 3 3 . . . . . . . . . . .
```

Indenting by two and adding corresponds to division of the first term by $1-t^2$. The row thus derived is then indented by $m+2$ to correspond to multiplication by $t^{m+2}$ and subtracted. Succeeding terms are incorporated similarly.
EXAMPLE 3. \( n = 5, \ m = 3 \) to find \( P(S_3, 5) \)

First Term: \( \frac{1-t^4}{1-t} = 111100 \)

Second Term: \( \frac{1}{(1-t^2)} = 11111 \)
\( \frac{1}{11} = 11 \)
\( \frac{1}{112222} = 112222 \)
\( \frac{1}{112222} = 112221 \)

Third Term: \( \frac{1}{(1-t^3)} = 112221 \)
\( \frac{1}{112} = 112333 \)
\( \frac{1}{112333} = - - - \)

Fourth Term: \( \frac{1}{(1-t^4)} = 112333 \)
\( \frac{1}{11} = 112344 \)
\( \frac{1}{112344} = - - - \)

Fifth Term: \( \frac{1}{(1-t^5)} = 112344 \)
\( \frac{1}{112345} = - - - \)

\[ \varphi(t) = \frac{1}{\binom{8}{3}} \left[ 1 + t^1 + 2t^2 + 3t^3 + 4t^4 + 5t^5 + \ldots \right] \]

Either of these procedures is more efficient than reducing the recurrence (2.26). Only \( m-1 \) unwanted tables are produced here while reduction of the recurrence involves generation of \( m \times n \) tables.
It is perhaps surprising that the second procedure which follows from a closed form solution is less efficient than the first. More transposition and addition operations are required. As in the one sample case scores which are carried to many significant figures may require more sophisticated programming to avoid exceeding computer memory limitations. A suitable computer program for tabling the general two sample scores statistic is given in the appendix.

EXAMPLE 4. Hajek (4) gives as one form of the median test statistic

\[ S = \sum_{i=1}^{m} \frac{1}{2} \left[ \text{sign} \left( R_i - \frac{1}{2}(m+n+1) \right) + 1 \right]. \]

A closed form does exist for \( P(S=k) \). This example is given to illustrate the procedure when zero scores and repeated scores are present.

Let \( m = 3, n = 4 \). Then

\[ a_k = 0, \quad k = 1, 2, 3 \]
\[ = \frac{1}{2}, \quad k = 4 \]
\[ = 1, \quad k = 5, 6, 7. \]

The probability generating function is then given by

\[ P_{m,n}(t) = \binom{m+n}{m}^{-1} (1+ut)^3 (1+ut^{1/2}) (1+ut)^3 \]
We form the arrays as before

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>1/2</th>
<th>1</th>
<th>3/2</th>
<th>2</th>
<th>5/2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\[ (1+ut^0) \]

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>1/2</th>
<th>1</th>
<th>3/2</th>
<th>2</th>
<th>5/2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\[ (1+ut^0) \]

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>1/2</th>
<th>1</th>
<th>3/2</th>
<th>2</th>
<th>5/2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\[ (1+ut^0) \]

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>1/2</th>
<th>1</th>
<th>3/2</th>
<th>2</th>
<th>5/2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td>3</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
\[(1+ut)\]

\[
\begin{array}{ccccccc}
0 & 1/2 & 1 & 3/2 & 2 & 5/2 & 3 \\
0 & 1 & & & & & \\
1 & 3 & 1 & 1 & & & \\
2 & 3 & 3 & 3 & 1 & & \\
3 & 1 & 3 & 3 & 3 & & \\
\end{array}
\]

\[(1+ut)\]

\[
\begin{array}{ccccccc}
0 & 1/2 & 1 & 3/2 & 2 & 5/2 & 3 \\
0 & 1 & & & & & \\
1 & 3 & 1 & 2 & & & \\
2 & 3 & 3 & 6 & 2 & 1 & \\
3 & 1 & 3 & 6 & 6 & 3 & 1 \\
\end{array}
\]

\[(1+ut)\]

\[
\begin{array}{ccccccc}
0 & 1/2 & 1 & 3/2 & 2 & 5/2 & 3 \\
0 & 1 & & & & & \\
1 & 3 & 1 & 3 & & & \\
2 & 3 & 3 & 9 & 3 & 3 & \\
3 & 1 & 3 & 9 & 9 & 9 & 3 & 1 \\
\end{array}
\]

Then \( P_3(t) = \binom{7}{3}^{-1} (1 + 3t^{1/2} + 9t + 9t^{3/2} + 9t^2 + 3t^{5/2} + t^3) \)

The result is readily confirmed by Hajek's formula,

\[
P(S = K) = \binom{m+n}{m}^{-1} \left( \begin{bmatrix} \frac{1}{2}(m+n) \\ k \end{bmatrix} \right) \left( \begin{bmatrix} \frac{1}{2}(m+n) \\ m-k \end{bmatrix} \right)
\]
4. **Tests of Regression.** To invert the general form given in (2.24)

\[ P_N(t) = \frac{1}{N} \prod_{k=1}^{N} \left( \sum_{j=1}^{d_j t} c_k a_j \right) \]

we form a sequence of \( N \) two dimensional arrays \( B^k = \{ b_{i,j}^k \} \) as follows.

**\( B^1 \):** Let \( b_{0,j}^1 = c_1 a_j \) \( j = 1, 2, \ldots, N \)

\( b_{j,j}^1 = -1 \) \( j = 1, 2, \ldots, N \)

\( b_{i,j}^1 = 0 \) otherwise.

**\( B^2 \):**

\( b_{0,j}^2 = b_{0,j}^1 + c_2 a_1 \)

\( b_{1,j}^2 = b_{1,j}^1 - 1 \)

\( b_{i,j}^2 = b_{i,j}^1 \) \( i = 2, \ldots, N \)

\( b_{0,j+N}^2 = b_{0,j}^1 + c_2 a_2 \)

\( b_{2,j+N}^2 = b_{2,j}^1 - 1 \)

\( b_{i,j+N}^2 = b_{i,j}^1 \) \( i > 0, i \neq 2 \)

\[ \ldots \]

\[ \ldots \]
\[ b_{o,j+kN}^2 = b_{o,j}^1 + c_2a_{k+1} \]
\[ b_{k+1,j+kN}^2 = b_{k+1,j}^1 - 1 \]
\[ b_{i,j+kN}^2 = b_{i,j}^1 \quad i > 0, \ i \neq k+1 \]
\[ \ldots \]
\[ \ldots \]
\[ \ldots \]
\[ b_{o,j+N(N-1)}^2 = b_{o,j}^1 + c_2a_N \]
\[ b_{N,j+N(N-1)}^2 = b_{N,j}^1 - 1 \]
\[ b_{i,j+N(N-1)}^2 = b_{o,j}^1 \quad i = 1, 2, \ldots, N-1 \]
\[ j = 1, 2, \ldots, N \]

\( B^r: \)
\[ b_{o,j+kN}^r = b_{o,j}^{r-1} + c_r a_{k+1} \]
\[ b_{k+1,j+kN}^r = b_{k+1,j}^{r-1} - 1 \]
\[ b_{i,j+kN}^r = b_{i,j}^{r-1} \quad i > 0, \ i \neq k+1 \]
\[ j = 1, 2, \ldots, N \]
\[ k = 0, 1, \ldots, N^r-N \]

We now form a vector \( C = \{c_j\} \) as follows:

Let \( c_j = 0 \quad j = 1, 2, \ldots, N^N \)

Then \( c_{b_{o,j}} = c_{b_{o,j}} + 1 \) If \( b_{i,j} = -1 \)
\[ i = 1, 2, \ldots, N \]
\[ j = 1, 2, \ldots, N^N \]
The vector $c$ will now contain the desired coefficients.

Now the procedure described above would require very large computer memory capacity. Some memory can be conserved, however, by a simple device. In forming the entry $b_{o,j+kN}^r$ we first add 1 to $b_{k+1,j+kN}^r$. Let

$$b_{k+1,j+kN}^r + 1 = \delta$$

and set

$$b_{o,(j+kN)\delta}^r = b_{o,j}^{r-1} + c_{r,k+1}^r$$

$$b_{k+1,(j+kN)\delta}^r = b_{k+1,j-1}^r$$

The effect is to load unwanted terms in the zeroth column of the $B$ array.

EXAMPLE 4: $c_i = i$, $a_N(R_i) = R_i$,

$N = 3$

$$B^1: \begin{array}{cccc}
0 & 1 & 2 & 3 \\
1 & & & \\
2 & & -1 & \\
3 & & & -1 \\
\end{array}$$

42
\[ B^2 : \]
\[
\begin{array}{ccccccc}
0 & 1 & 2 & 3 & 4 & 5 & 6 \\
0 17 & 4 & 5 & 5 & 7 & 7 & 8 \\
1 & -1 & -1 & -1 & -1 & -1 & -1 \\
2 & -1 & -1 & -1 & -1 & -1 & -1 \\
3 & -1 & -1 & -1 & -1 & -1 & -1 \\
\end{array}
\]

\[ B^3 : \]
\[
\begin{array}{ccccccc}
0 & 1 & 2 & 3 & 4 & 5 & 6 \\
0 140 & 10 & 11 & 11 & 13 & 13 & 14 \\
1 & -8 & -1 & -1 & -1 & -1 & -1 \\
2 & -8 & -1 & -1 & -1 & -1 & -1 \\
3 & -8 & -1 & -1 & -1 & -1 & -1 \\
\end{array}
\]

\[ C(10) = 1 \]
\[ C(11) = 2 \]
\[ C(13) = 2 \]
\[ C(14) = 1 \]

\[ p_3(t) = \frac{1}{6} (t^{10} + 2t^{11} + 2t^{13} + t^{14}) \]

For machine computation using this procedure the computer memory must have the capacity to store \((2N+1) \times N!\) words. This effectively limits the procedure to moderate sample sizes. It does not appear that this storage requirement may be reduced for general scores and regression coefficients. In expanding the product in \((5)\), \(N!\) admissible exponents are formed and with general scores and regression
coefficients they may all be distinct. For the Spearman case, 
\( C_i = i \), \( a_N(R_i) = i \), a slightly modified procedure yields some 
进一步 economies.

5. **The Kolmogorov-Smirnov Test.** To invert the probability gen-
erating function, the auxiliary function,

\[
\hat{h}(t) = t^h \left[ \prod_{i=1}^{m} (t^{-n} + t^{m-n} + \cdots + t^{mn-n}) \right] h-n
\]

is first inverted and the coefficients of \( t^K, K \geq h \) are added. 
Recall from Chapter II that the notation used here signifies that, 
in taking the product, exponents which exceed \( h-n \) or fall 
below \( -nj \) are discarded at each stage. We proceed in a manner 
 analogous to the one sample case using exponents as place 
 holders and registering a multiplication by shifting a vector 
of coefficients to the right, if the multiplier has a positive 
exponent, to the left otherwise. The method is best illustrated 
by example.

**EXAMPLE 5.** \( m = 3, n = 4 \) to find

\[
P \left[ \max_{1 \leq k \leq 12} \left( \frac{3k}{7} - c_{d_1} - \cdots - c_{d_k} \right) \leq \frac{8}{7} \right] \overset{\text{def}}{=} P \left[ K_{3,4} \leq \frac{8}{7} \right],
\]

\[
\hat{h}(t) = t^8 \prod_{j=1}^{3} \left[ (t^{-4} + t^{-1} + t^2 + t^5 + t^8) \right]^{-4j}
\]

\[
8, 3, 4
\]
\[ (t^{-4} + t^{-1} + t^{2} + t^{5} + t^{8})_{4}^{4} : \]
\[ (t^{-4} + t^{-1} + t^{2} + t^{5} + t^{8})_{8}^{4} : \]
\[ (t^{-4} + t^{-1} + t^{2} + t^{5} + t^{8})_{12}^{4} : \]
\[ .t^{8} : \]

Adding coefficients of \( t^{K} \), \( K \leq 8 \) yields

\[ P\left( K_{3,4} \leq \frac{8}{7} \right) = \frac{31}{\binom{7}{3}} = 0.885 \]

Vertical lines denote the restriction that the exponent lie between \( -n_j \) and \( h-n \), entries which would fall to the right or left of the vertical lines are discarded at each stage in the multiplication.
CHAPTER IV
Local Limit Theorems and Asymptotic Expansions

1. In this chapter local limit theorems and asymptotic expansions for the distribution of the two sample Wilcoxon statistic are developed. The use of a formal Edgeworth expansion in this application was first suggested by Hodges and Fix (3) who noted that the uncorrected normal approximation was subject to considerable error, particularly at high significance levels. Fellingham and Stoker (2) derived the formal expansion for the case of the one sample Wilcoxon statistic. Stoker (9) proved, by an application of the methods of Chernoff and Savage, that the error in the normal approximation was of the order of $\max(n^{-1/2}, m^{-1/2})$.

It is shown here that the formal expansion to one correction term is justified in the two sample case. It is also shown that the error in approximating by the one term expansion may be bounded by an explicit decreasing function of sample sizes. As a corollary, it is shown that the error in the uncorrected normal approximation is in fact of the order $\max(n^{-1}, m^{-1})$.

The methods of proof apply without significant change to the one sample case which is not treated here. The proofs essentially
parallel those of Feller (1) for the case of independent sums.

2. Preliminaries. Kendall (6) gives the characteristic function of the distribution of the two sample Wilcoxon statistic centered at the origin, \( U_{m,n} - 1/2mn \), as,

\[
(2.1) \quad f_{m,n}(t) = \left( \frac{m+n}{m} \right)^{-1} \prod_{k=1}^{n} \frac{\sin \left( \frac{m+k}{2} \cdot t \right)}{\sin \left( \frac{kt}{2} \right)}.
\]

Let \( g_{m,k}(t) = \frac{k}{m+k} \cdot \sin \left( \frac{m+k}{2} \cdot t \right)/\sin \left( \frac{kt}{2} \right) \) and observe that whenever \( m+k \) is divisible by \( k \), \( g_{m,k}(t) \) is a characteristic function. It will be shown that, even when this condition fails to hold, \( g_{m,k}(t) \) behaves like a c.f. when \( |t| \) is small. Also, it will be shown that \( f_{m,n}(t) \) tends to zero rapidly in intervals bounded away from its maxima at \( 2j\pi, j=0, 1, 2, \ldots \). These qualities are exploited in proving the limit theorems.

For \( |t| < 2\pi/(m+k) \), \( \log g_{m,k}(t) \) may be expanded in a Taylor series;

\[
(2.2) \quad \log g_{m,k}(t) = \sum_{j=1}^{\infty} \frac{(-1)^{j+1} B_{2j}}{2j(2j)!} \left[ (m+k)^{2j} - k^{2j} \right] \cdot (it)^{2j}
\]

where \( B_k \) is the \( k^{th} \) Bernoulli number. Now
\[ (2.3) \quad \frac{2(2j)!}{(2\pi)^{2j}} \left( \frac{2^{2j-1}}{2^{2j-1} - 1} \right) > (-1)^{j+1} B_{2j} > \frac{2(2j)!}{(2\pi)^{2j}} \]

and

\[ \frac{1}{j} \cdot \frac{2^{2j}}{2^{2j}} > \frac{1}{j+1} \cdot \frac{2^{2j+1}}{2^{2j+1} - 1} \]

so that for \( t < 2\pi/(m+n) \) and \( k \leq n \leq m \)

\[ (2.4) \quad \frac{(-1)^{j+1} B_{2j}}{2j(2j)!} \left( (m+k)^{2j} - k^{2j} \right) t^{2j} \]

\[ > \frac{1}{j} \frac{(m+k)^{2j} - k^{2j}}{(m+k)^{2j+2} - k^{2j+2}} (m+n)^{2j+2} t^{2j+2} \]

\[ = \frac{1}{j(2\pi)^{2j+2}} \frac{(m+k)^{2j} - k^{2j}}{(m+k)^{2j+2} - k^{2j+2}} (m+n)^{2j+2} \frac{1 - \frac{k^{2j}}{(m+k)^{2j}}}{(m+k)^{2j+2} - k^{2j+2}} t^{2j+2} \]

\[ > \frac{1}{j(2\pi)^{2j+2}} \frac{(m+k)^{2j+2} - k^{2j+2}}{(m+k)^{2j+2} - k^{2j+2}} t^{2j+2} \]

\[ > \frac{2}{(2j+2)(2\pi)^{2j+2}} \left( \frac{2^{2j+1}}{2^{2j+1} - 1} \right) (m+k)^{2j+2} - k^{2j+2} t^{2j+2} \]

\[ > \frac{(-1)^{j+2} B_{2j+2}}{(2j+2)(2j+2)!} \left( (m+k)^{2j+2} - k^{2j+2} \right) t^{2j+2}. \]
Thus it follows that the error in approximating \( \log q_{m,n}(t) \) by a partial sum in (2) is no greater than the magnitude of the first term omitted.

Now let

\[
\psi_{m,n}(t) = \log f_{m,n}(t) + \frac{mn(m+n+1)t^2}{24}
\]

\[
= \sum_{k=1}^{n} \log q_{m,k}(t) + \sum_{k=1}^{n} \frac{(m+k)^2 - k^2}{24} \cdot t^2
\]

and let \( \psi_{m,n}^r(t) \) be the partial sum approximation to \( \psi_{m,n}(t) \) up to and including terms of degree \( 2r \). Then for \( t < 2\pi/(m+n) \)

\[
(2.5) \quad |\psi_{m,n}(t)| \leq \frac{(-1)^3P_4}{4(4)!} \sum_{k=1}^{n} \left( (m+k)^4 - k^4 \right) t^4
\]

\[
< \frac{1}{2} \cdot \frac{8}{7} \cdot \frac{1}{(2\pi)^4} \sum_{k=1}^{n} \left( (m+k)^4 - k^4 \right) t^4
\]

\[
< \frac{8}{14} \cdot \frac{t^2}{(2\pi)^2} \sum_{k=1}^{n} \frac{(m+k)^4 - k^4}{(m+n)^2}
\]

\[
= \frac{8}{14} \cdot \frac{t^2}{(2\pi)^2} \sum_{k=1}^{n} \frac{(m+k)^2 - k^2}{((m+k)^2 - k^2)(m+n)^2} \left( (m+k)^4 - k^4 \right)
\]
\[
< \frac{8}{14} \cdot \frac{t^2}{(2\pi)^2} \sum_{k=1}^{n} \frac{(m+k)^2 - k^2}{1 - \left(\frac{k}{(m+k)^2}\right)} \frac{(m+k)^2}{(m+n)^2} \\
< \frac{8}{14} \cdot \frac{t^2}{(2\pi)^2} \frac{1}{1 - \left(\frac{1}{2}\right)^2} \sum_{k=1}^{n} \left((m+k)^2 - k^2\right) \\
= \frac{16}{21} \cdot \frac{t^2}{(2\pi)^2} \cdot (mn(m+n+1)) \\
< \frac{mn(m+n+1)}{48} t^2 .
\]

\( \psi_{m,n}^{r}(t) \) is a \( 2r \) degree polynomial with alternating signs and for \( t < 2\pi/(m+n) \) the terms are of decreasing magnitude so that the same reasoning applies and

(2.6) \[ | \psi_{m,n}^{r}(t) | < \frac{mn(m+n+1)}{48} t^2 . \]

Also

(2.7) \[ | \psi_{m,n}(t) - \psi_{m,n}^{r}(t) | \leq \frac{(-1)^{r+1}B_{2r+2}}{(2r+2)(2r+2)!} \sum_{k=1}^{n} \left((m+k)^{2r+2} - k^{2r+2}\right) t^{2r+2} \\
< \frac{1}{(r+1)(2\pi)^{2r+2}} \left(\frac{2^{2r+1}}{2^{2r+1} - 1}\right) \sum_{k=1}^{n} \left((m+k)^{2r+2} - k^{2r+2}\right) t^{2r+2} \]
\[
< \frac{t^{2r}}{(r+1)(2n)^{2r}} \left( \frac{2^{2r+1}}{2^{2r+1} - 1} \right) \sum_{k=1}^{n} \frac{(\frac{2r}{m+k} - \frac{k}{m+k}) \frac{2r}{(m+k)^{2r}}}{(m+k)^{2r} - k^{2r}} \frac{2r+2}{(m+k)^{2r+2} - k^{2r+2}} (m+n)^2
\]

\[
< \frac{t^{2r}}{(r+1)(2n)^{2r}} \left( \frac{2^{2r+1}}{2^{2r+1} - 1} \right) \sum_{k=1}^{n} \left( \frac{\frac{2r}{m+k} - \frac{k}{m+k}}{1 - \frac{k^{2r}}{(m+k)^{2r}}} \right) \frac{2r}{(m+k)^{2r}} \frac{2r}{(m+k)^{2r} - k^{2r}} (m+n)^2
\]

Now \( \frac{1}{r+1} \left( \frac{2^{2r+1}}{2^{2r+1} - 1} \right) < \frac{1}{r} \) so that denoting the \( 2r \)th cumulant of the distribution as \( \lambda_{m,n}^r \)

\[
\left| \psi_{m,n}(t) - \psi_{m,n}^r(t) \right| < \frac{\lambda_{m,n}^r}{(2r)!} t^{2r}.
\]

We now examine the behavior of \( f_{m,n}(t) \) in the interval

\[
\frac{2\pi}{m+n} \leq t \leq 2\pi - \frac{2\pi}{m+n}.
\]

We first treat the case where for each \( k=1, 2, \ldots, n, m+k \) is divisible by \( k \), in which case

\[
\frac{k}{m+k} \sin \left( \frac{m+k}{2} \cdot t \right) / \sin \left( \frac{kt}{2} \right)
\]

is a characteristic function. We then show that the results extend to the general case.
We will need the following lemmata the first of which is an extension of a well known theorem of Weyl.

Lemma 1. Let \( t \) be such that

\[
2\pi/(m+n) \leq t \leq 2\pi - 2\pi/(m+n), \quad n \leq m.
\]

Then for at least \( 2 \cdot \left\lfloor \frac{n}{5} \right\rfloor \) of the integers \( k=1, 2, \ldots, n, \)

\[
n\pi/(m+n) \leq k \cdot t \mod 2\pi \leq 2\pi - n\pi/(m+n).
\]

Remark. Weyl's theorem guarantees that, asymptotically as \( n \to \infty \), at least \( \left\lfloor \frac{n}{2} \right\rfloor \) of the terms, \( k \cdot t \mod 2\pi \), will have the desired property if \( t \) is an irrational multiplier of \( 2\pi \).

This would be adequate for asymptotic error bounds but since the interest here extends to producing explicit bounds for finite sample sizes the more precise statement is required. The conclusion is intuitive but the proof is long. The reader is advised to skip to page 65 on first reading.

Proof of Lemma. Let \( n \leq m \) and

\[
2\pi/m+n \leq t \leq 3\pi/2n.
\]

Then

\[
\frac{n}{2} \cdot t \geq \frac{n\pi}{m+n}
\]

and

\[
nt \leq 3\pi/2 \leq 2\pi - n\pi/m+n.
\]
Now let \( t = \frac{2i\pi}{j} \) for some rational \( i/j \) in lowest terms such that \( 1/m+n \leq i/j < 1/2 \) and observe that for \( l=0, 1, 2, \ldots, i-2 \)

\[
\left[ \frac{(l+1)j}{i} \right] \cdot \frac{2i\pi}{j} \mod 2\pi \leq 2\pi \left( \frac{j-1}{j} \right)
\]

and

\[
\left( \left[ \frac{lj}{i} \right] + 1 \right) \cdot \frac{2i\pi}{j} \mod 2\pi > 0.
\]

Thus for

\[
k = \left( \left[ \frac{lj}{i} \right] + 1 \right), \left( \left[ \frac{lj}{i} \right] + 2 \right), \ldots, \left[ \frac{(l+1)j}{i} \right]
\]

\[
k \cdot t \mod 2\pi = \frac{k \cdot 2i\pi}{j} - 2l\pi \approx \pi/2
\]

implies

\[
k > j/4i + lj/i
\]

and

\[
k \cdot t \mod 2\pi < 3\pi/2
\]

implies

\[
k < 3j/4i + lj/i.
\]

It follows then that for at least \( \left[ (l + 3/4)j/i \right] - \left[ (l + 1/4)j/i \right] \)
of \( k = \left( \left[ \frac{lj}{i} \right] + 1 \right), \ldots, \left[ \frac{(l+1)j}{i} \right], l=0, 1, \ldots, i-1 \)

\[
\pi/2 \leq k \cdot t \mod 2\pi \leq 3\pi/2.
\]
Let 
\[ N_n = \text{the number of integers, } k = 1, 2, ..., n \text{ for which} \]
\[ \pi/2 \leq \frac{k \cdot 2i\pi}{j} \mod 2\pi \leq \frac{3\pi}{2}, \]

\[ N_L = \sum_{L=0}^{L-1} \left( \left\lfloor \left( L + \frac{3}{4} \right) j/i \right\rfloor - \left\lfloor \left( L + \frac{1}{4} \right) j/i \right\rfloor \right), \]

\[ j/i = \alpha + \frac{s}{i} \text{ for some integers } \alpha \text{ and } 1 \leq s < i, \]

\[ j/4i = \beta + r/4i, \text{ } 1 \leq r < 4i \text{ and} \]

\[ I(x < y) = 1 \text{ if } x < y \]

\[ = 0 \text{ otherwise.} \]

Then

\[ [Lj/i] = \sum_{L=0}^{L-1} \left( \left\lfloor \left( L+1 \right) j/i \right\rfloor - \left\lfloor Lj/i \right\rfloor \right) \]

\[ = \sum_{L=0}^{L-1} \left( \left\lfloor j/i \right\rfloor + I \left( \left( i-L \cdot s \mod i \right) \leq s \right) \right) \]

and for \( L > 0 \)

\[ \left\lfloor \left( L + \frac{3}{4} \right) j/i \right\rfloor - \left\lfloor \left( L + \frac{1}{4} \right) j/i \right\rfloor = \left\lfloor 3j/4i \right\rfloor - \left\lfloor j/4i \right\rfloor \]

\[ - I \left( \left( i-L \cdot s \mod i \right) > \left\lfloor 3j/4 \right\rfloor - \left\lfloor 3j/4i \right\rfloor \cdot i \right) \]

\[ + I \left( \left( i-L \cdot s \mod i \right) > \left\lfloor j/4 \right\rfloor - \left\lfloor j/4i \right\rfloor \cdot i \right). \]

Along the sequence \( n = [Lj/i], L=1, 2, ..., N_n = N_L \), but if \( j/4i > 1 \), then as \( n \) increases from \( n = [Lj/i] \) to \( n = [Lj/i] + [j/4i] \),

\[ \frac{k \cdot 2i\pi}{j} \mod 2\pi < \frac{\pi}{2} \]
for \( k = \lceil \frac{L_j}{i} \rceil + 1, \ldots, n \). Thus if \( L \) is the largest integer such that \( \lceil \frac{L_j}{i} \rceil \leq n \) it will suffice to show that

\[
N_L \geq 2 \left[ \frac{\lceil \frac{L_j}{i} \rceil + \lceil \frac{j}{4i} \rceil}{5} \right].
\]

Now

\[
\lceil \frac{3j}{4i} \rceil - \lceil \frac{j}{4i} \rceil = 2\theta \text{ if } r < \frac{4i}{3}
\]

\[
= 2\theta + 1 \text{ if } \frac{4i}{3} < r < \frac{8i}{3}
\]

\[
= 2\theta + 2 \text{ if } \frac{8i}{3} < r < 4i.
\]

Note that only strict inequality applies for the bounds on \( r \) since \( i/j \) is in lowest terms.

\[
\lceil \frac{j}{i} \rceil = 4\theta \text{ if } r < i,
\]

\[
= 4\theta + 1 \text{ if } i < r < 2i,
\]

\[
= 4\theta + 2 \text{ if } 2i < r < 3i,
\]

\[
= 4\theta + 3 \text{ if } 3i < r < 4i.
\]

\[
\lceil \frac{3j}{4} \rceil - \lceil \frac{3j}{4i} \rceil \cdot i = \lceil \frac{3r}{4} \rceil - \lceil \frac{3r}{4i} \rceil \cdot i
\]

\[
\lceil \frac{j}{4} \rceil - \lceil \frac{j}{4i} \rceil \cdot i = \lceil \frac{r}{4} \rceil
\]

and

\[
\lceil \frac{3r}{4} \rceil - \lceil \frac{3r}{4i} \rceil \cdot i = \lceil \frac{3r}{4} \rceil \text{ if } r < 4i/3
\]

\[
= \lceil \frac{3r}{4} \rceil - i \text{ if } 4i/3 < r < 8i/3
\]

\[
= \lceil \frac{3r}{4} \rceil - 2i \text{ if } 8i/3 < r < 4i.
\]
Thus for \( r < \frac{4i}{3} \),

\[
[j/4] - \lfloor j/4i \rfloor \cdot i - [3j/4] + [3j/4i] \cdot i = \lfloor r/4 \rfloor - [3r/4] \leq 0.
\]

Since \( \lfloor j/i \rfloor \geq 2 \) we may assume \( \beta > 0 \) for \( r < 2i \).

CASE I. \( r < i \)

\[
\frac{N_L}{\left\lfloor \frac{Li}{i} \right\rfloor + \left\lfloor \frac{i}{4i} \right\rfloor} \geq \frac{2L\beta}{4L\beta + \left( \sum_{\ell=0}^{L-1} I((i - \ell \cdot s \mod i) \leq s) \right) + \beta}
\]

\[
\geq \frac{2L\beta}{4L\beta + L - 1 + \beta} \geq \frac{2}{5}
\]

CASE II. \( i < r < 4i/3 \)

As in Case I, whenever

\[
I \left( (i - \ell \cdot s \mod i) > [3j/4] - [3j/4i] \cdot i \right) = 1
\]

then

\[
I \left( (i - \ell \cdot s \mod i) > [j/4] - [j/4i] \cdot i \right) = 1
\]

Also

\[
\frac{\ell}{i} = \alpha + \frac{s}{i} = 4\beta + \frac{r}{i}, \quad 1 \leq s < i
\]

and so \( s = r - i < \frac{i}{3} \).
Now if \( l \) is such that \( i - l \cdot s \mod i \leq s < \frac{i}{3} \) then

\[
i - (l - 1) \cdot s \mod i = (i - l \cdot s \mod i + s) \mod i = i - l \cdot s \mod i + s
\]

and thus

\[
\frac{r}{4} < \frac{i}{3} < i - (l - 1) \cdot s \mod i < \frac{2i}{3} < \frac{3i}{4} \leq \left[ \frac{3r}{4} \right].
\]

Finally \( \frac{x-1}{y} < \frac{x}{y+1} \) whenever \( x < y+1 \) so we may approximate

\[
\frac{N_L}{[L/j/i] + [j/4i]} > \frac{2L\beta}{4L\beta + L + \beta} \geq \frac{2}{5}
\]

if both \( L > 1 \) and \( \beta > 1 \). If \( L = 1 \) or \( \beta = 1 \) or both, then

\[
2 \cdot \left[ \frac{4L\beta + L + \beta}{5} \right] \leq 2L\beta \quad \text{Q.E.D.}
\]

CASE III. \( 4i/3 < r < 2i \)

Then \( s = r - i > i/3 \). Let \( l \) be such that

\[
i - l \cdot s \mod i \leq \left[ j/4 \right] - \left[ j/4i \right] \cdot i = \left[ r/4 \right] < i/2.
\]

Then

\[
i - (l + 1) \cdot s \mod i = (i - l \cdot s \mod i - s) \mod i < i/2 \quad \text{only if} \quad s < i/2.
\]
Now if \( i/3 < s < i/2 \)

\[
i - s > i/2, \\
0 < i - 2s < i/2, \\
i - 3s \mod i > i - s > i/2
\]

and in general if

\[
i - \ell \cdot s \mod i < i/2 \\
i - (\ell+1) \cdot s \mod i < i/6 \\
i - (\ell+2) \cdot s \mod i > i - s > i/2.
\]

Also \( i/3 < s < i/2 \) implies that if \( \ell > 0 \) is such that

\[
(i - \ell \cdot s \mod i) \leq s < i/2
\]

then

\[
i - (\ell+1) \cdot s \mod i = i - \ell \cdot s \mod i + i - s \\
> i - \ell \cdot s \mod i + i/2 \\
> s.
\]

It follows that if \( i/3 < s < i/2, \) then for at least \( L/3 \) of \( \ell=1, 2, \ldots, L-1 \)

\[
I((i - \ell \cdot s \mod i) > \lfloor j/4 \rfloor - \lfloor j/4i \rfloor \cdot i) = 1
\]

and for at most \( L/2 \) of \( \ell=1, 2, \ldots, L-1, \)

\[
I((i - \ell \cdot s \mod i) \leq s) = 1
\]
Thus for \( i/3 < s < i/2 \),

\[
\frac{N_L}{[Lj/i] + [j/4i]} \geq \frac{2L\beta + L/3}{4L\beta + 3L/2 + \beta} > \frac{2}{5}
\]

if \( L > 1 \) and \( \beta > 1 \).

If either \( L = 1 \) or \( \beta = 1 \) or both

\[
2 \cdot \left[ \frac{4L\beta + 3L/2 + \beta}{5} \right] \leq 2L\beta + L/3 .
\]

If \( i/2 < s \),

\[
\frac{N_L}{[Lj/i] + [j/4i]} \geq \frac{2L\beta + L/2}{4L\beta + 2L - 1 + \beta} > \frac{2}{5} .
\]

Before proceeding to the remaining cases it will be necessary to examine the expressions \( i - \ell \cdot s \mod i \). The ordered \((i - 1)\) - tuple

\[
(i - \ell \cdot s \mod i)_{\ell=1}^{i-1}
\]

is the permutation

\[
(i-s, i-2s, \ldots, i-[i/s] \cdot s, 2i-[i/s] \cdot s-s, \\
2i-[i/s] \cdot s-2s, \ldots, 2i-[2i/s] \cdot s, \\
3i-[2i/s] \cdot s-s, \ldots, ki-[ki/s] \cdot s, \\
(k+1)i-[ki/s] \cdot s-s, \ldots)
\]

of the integers 1, 2, ..., \( i-1 \). Obviously when \( s=i-1 \) it is the ordered set \((1, 2, \ldots, i-1)\). We will refer to the numbers
where $k = 1, 2, \ldots, s$.

as a cycle of $(i - \ell \cdot s \mod i)$. Note that the last term in
every cycle but the $s$'th is greater than zero and the first term
of every cycle is derived by adding $i - s$ to the last term of
the preceding cycle and thus for $k > 1$

$$ki - \left[ \frac{(k-1)i}{s} \right] \cdot s - s > i - s.$$  

Obviously $i - \ell \cdot s \mod i \leq s$ for at most one term in any cycle.

CASE IV. $2i < r < 8i/3$

$$\left[ \frac{3r}{4} \right] - \left[ \frac{3r}{4i} \right] \cdot i = \left[ \frac{3r}{4} \right] - i \geq \left[ \frac{r}{4} \right] + \left[ \frac{r}{2} \right] - i$$

$$> \left[ \frac{r}{4} \right].$$

Thus, as in cases I and II,

$$I(i - \ell \cdot s \mod i) > \left[ \frac{i}{4} \right] - \left[ \frac{i}{4i} \right] \cdot i = 1$$

whenever

$$I(i - \ell \cdot s \mod i) > \left[ \frac{3i}{4} \right] - \left[ \frac{3i}{4i} \right] \cdot i = 1.$$  

Then if $\beta > 0$

$$\frac{N_L}{\left[ \frac{L}{i} \right] + \left[ \frac{i}{4i} \right]} > \frac{2L\beta + L}{4L\beta + 3L - 1 + \beta} > \frac{2}{5}$$

If $\beta = 0$ then let $\ell$ be such that

$$i - \ell \cdot s \mod i \leq s < 2i/3.$$
Then either

\[ i - (l+1) \cdot s \mod i > 2i/3 \]

or

\[ i - (l+1) \cdot s \mod i = i - l \cdot s \mod i + i - s < \frac{2i}{3} \]

and

\[ i - l \cdot s \mod i < \frac{i}{3}. \]

Then \[ i - (l-1) \cdot s \mod i = i - l \cdot s \mod i + s > s \] and thus for at most \[ L/2 \] of \[ l = 0, 1, 2, \ldots, L-1, \] \[ \mathbb{I}(i - l \cdot s \mod i) \leq s = 1. \]

Then

\[ \frac{N_L}{L_i \left[ \frac{L_i}{i} \right]} > \frac{L}{2L + \frac{L}{2}} = \frac{2}{5}. \]

CASE V: \( 8i/3 < r < 3i \)

If \( \beta > 0 \)

\[ \frac{N_L}{\left[ \frac{L_i}{i} \right] + \left[ \frac{i}{4i} \right]} > \frac{2L\beta + L + 1}{4L\beta + 3L - 1 + \beta} > \frac{2}{5}. \]

If \( \beta = 0 \) then consider those cycles of \( (i - l \cdot s \mod i) \) whose least term is such that

\[ i - l \cdot s \mod i \leq s. \]

Now if there are \( k < L \) successive terms of this type and \( i - l_1 \cdot s \mod i \) is the first then

\[ i - l_1 \cdot s \mod i \leq i - s \]

61
and if \( k \geq 6 \) then

\[
6i - 6s \leq s
\]

and

\[
s \geq \frac{6i}{7}, \quad i - s \leq \left\lfloor \frac{i}{7} \right\rfloor,
\]

so that

\[
\left\lfloor \frac{3r}{4} \right\rfloor - 2i = \frac{3s+2i}{4} - i \geq \frac{8i}{7} - i = \left\lfloor \frac{i}{7} \right\rfloor \geq i - s \geq i - \ell_1 \cdot s \mod i.
\]

If

\[
k \geq 6p, \quad p > 1
\]

then

\[
s \geq \frac{6pi}{6p + 1}
\]

and

\[
\left\lfloor \frac{3r}{4} \right\rfloor - 2i \geq p(i - s) \geq i - \ell_1 \cdot s \mod i + (p - 1)i - (p - 1) \cdot s.
\]

Thus

\[
\frac{N_L}{[L i]} \geq \frac{L + 1}{3L - [\frac{L}{5}]}
\]

and

\[
2 \cdot \left[ 3L - [\frac{L}{5}] \right] \leq L + 1.
\]

CASE VI. \( 3i < r < 4i \)

\[
s = r - 3i
\]
\[
\begin{align*}
\left[ \frac{3i}{4} \right] - \left[ \frac{3i}{4i} \right] \cdot i &= \left[ \frac{3i}{4} \right] - 2i \\
&= \left[ \frac{3s + 9i}{4} \right] - 2i \\
&= \left[ \frac{3s + i}{4} \right] > s.
\end{align*}
\]

Thus whenever

\[ I((i - \ell \cdot s \mod i) \leq s) = 1 \]

then

\[ I((i - \ell \cdot s \mod i) > \left[ \frac{3i}{4} \right] - \left[ \frac{3i}{4i} \right] \cdot i) = 0 \]

and if \( \beta > 0 \)

\[
\frac{N_L}{\left[ \frac{i}{4i} \right]} > \frac{2L \beta + L + 1}{4L \beta + 3L + \beta} > \frac{2}{5}.
\]

If \( \beta = 0 \) we consider those terms for which

\[
\left[ \frac{x}{4} \right] \geq i - \ell \cdot s \mod i > \left[ \frac{3i}{4} \right] - \left[ \frac{3i}{4i} \right] \cdot i \geq \frac{i}{4}.
\]

If the first term in any cycle of \( (i - \ell \cdot s \mod i) \) does not exceed \( \frac{x}{4} \) then, indexing that term by \( \ell_1 \), \( i - s \leq i - \ell_1 \cdot s \mod i \leq \frac{x}{4} = \frac{3i + s}{4} \)

and \( s > \frac{i}{5} \).

Also

\[
\left[ \frac{x}{4} \right] - \left[ \frac{3x}{4} \right] + 2i < \frac{2i - 2s + 3}{4} < \frac{2i}{5} + 1
\]
and it follows that there are at most two terms $l_1 > l_2$ such that

$$s < \left[ \frac{3s + i}{4} \right] < i - l \cdot s \mod i \leq \left[ \frac{3i + s}{4} \right].$$

Also if in any cycle there are $3k$, $k \geq 1$, terms such that

$$\left[ \frac{3s + i}{4} \right] < i - l \cdot s \mod i \leq \left[ \frac{3i + s}{4} \right]$$

then $s \leq \frac{i}{6k-1}$ and

$$i - ks \geq \frac{5ki - i}{6k - 1} > \frac{9ki - i}{2(6k - 1)} \geq \frac{3i + s}{4}.$$ 

Thus in any cycle of length $k$, there are at least $\left[ \frac{k-1}{3} \right] + 1$ terms for which

$$I \left( (i - l \cdot s \mod i) > \left[ \frac{i}{4} \right] \right) - I \left( (i - l \cdot s \mod i) > \left[ \frac{3i}{4} \right] - \left[ \frac{3i}{4i} \cdot i \right] \right) = 0$$

and it follows that

$$\frac{N}{L} \leq \frac{L + \left[ \frac{L}{3} \right]}{3L}$$

and

$$2 \cdot \left[ \frac{3L}{5} \right] < L + \left[ \frac{L}{3} \right].$$
Finally the entire argument is symmetric for \( t = \frac{2i\pi}{j} > \pi \) so that the conclusion follows for all rational \( t \). Extension to irrationals is straightforward.

Note. The bound is "tight." To see this let \( t = 2\pi/5 \) and \( m = n \). Then for \( n = 5i + 1, i = 1, 2, \ldots \), \( N_n = 2\left[\frac{n}{5}\right] \).

Lemma 2. For any positive real numbers \( \alpha, \beta, \) and \( t \) such that \( \alpha < \beta < 2\alpha \) and \( 0 \leq t \leq \pi/\alpha \),

\[
\frac{1}{\beta} \sin \left( \frac{\beta}{2} \cdot t \right)
\]

is non-increasing in \( \beta \).

Proof.

\[
\frac{d}{dt} \left( \frac{1}{\beta} \tan \left( \frac{\beta}{2} \cdot t \right) \right) = \frac{1}{2} \sec^2 \left( \frac{\beta}{2} \cdot t \right)
\]

\[
\geq \frac{1}{2} \text{ if } 0 \leq t \leq \frac{2\pi}{\beta}
\]

whence for \( 0 \leq t \leq \pi/\alpha \),

\[
\frac{t}{2} \leq \frac{1}{\beta} \tan \left( \frac{\beta}{2} \cdot t \right)
\]

and

\[
\frac{t}{2\beta} \cos \left( \frac{\beta}{2} \cdot t \right) - \frac{1}{\beta^2} \sin \left( \frac{\beta}{2} \cdot t \right) \leq 0.
\]

The left hand side of the final line is the derivative of \( \frac{1}{\beta} \sin \left( \frac{\beta}{2} \cdot t \right) \) with respect to \( \beta \) and the conclusion follows.
Let $m+k$ be divisible by $k$ for $k=1, 2, \ldots, n$ and let

$$g_k(t) = \left| \frac{k}{m+k} \sin \left( \frac{m+k}{2} \cdot t \right) / \sin \left( \frac{kt}{2} \right) \right| \geq \frac{2J\pi}{k} - \frac{\pi}{m+k} < t < \frac{2J\pi}{k} + \frac{\pi}{m+k},$$

$$J = 0, 1, 2, \ldots,$$

$$= \left| \frac{k}{m+k} \csc \left( \frac{m+k}{2} \right) \right| \text{ otherwise}$$

Then

$$\left| \frac{k}{m+k} \sin \left( \frac{m+k}{2} \cdot t \right) / \sin \left( \frac{kt}{2} \right) \right| \leq g_k(t).$$

Let $h_k(t) = g_k(t/k)$ and then

$$\left| f_{m,n}(t) \right| \leq \prod_{k=1}^{n} h_k(k+t).$$

Now $\frac{k}{m+k} \sin \left( \frac{m+k}{2k} \cdot t \right)$ and $\sin(t/2)$ have identical slope at the origin while for $0 \leq t \leq \frac{k\pi}{m+k}$, $\cos \left( \frac{m+k}{2k} \cdot t \right) \leq \cos(t/2)$.

It follows then that $h_k(t)$ is monotone decreasing in the interval $0 \leq t \leq \pi$. It is periodic with period $2\pi$ and symmetric about $\pi$.

By definition

$$h_k(t) \leq \frac{k}{m+k} \csc \left( \frac{t}{2} \right) \leq \frac{n}{m+n} \csc \left( \frac{t}{2} \right).$$

Also if $\frac{2\pi}{m+n} \leq t \leq 2\pi - \frac{2\pi}{m+n}$ and $n \leq m$, it follows from Lemma 1 that for at least $2 \left[ \frac{n}{5} \right]$ of $k=1, 2, \ldots, n$

$$\frac{n\pi}{m+n} \leq k \cdot t \mod 2\pi \leq 2\pi - \frac{n\pi}{m+n}.$$

Thus setting to unity those terms for which
\[ 2\pi - \frac{n\pi}{m+n} < k \cdot t \mod 2\pi < \frac{n\pi}{m+n}, \]

\[ |f_{m,n}(t)| \leq \left( \frac{n}{m+n} \csc \left( \frac{n\pi}{2(m+n)} \right) \right)^{2\left[ \frac{n}{5} \right]} \]

Now introducing the general inequality

\[ \sin(\pi\chi) > \pi \cdot \chi (1 - \chi), \text{ valid for } 0 < \chi < 1, \]

it follows that

\[ |f_{m,n}(t)| < \left( \frac{n}{m+n} \cdot \frac{4(m+n)^2}{\pi \cdot (2n(m+n) - n^2)} \right)^{2\left[ \frac{n}{5} \right]} \]

\[ = \left( \frac{4}{\pi \cdot 2m+n} \right)^{2\left[ \frac{n}{5} \right]} \]

\[ \leq \left( \frac{8}{3\pi} \right)^{2\left[ \frac{n}{5} \right]} = o(n^{-k}) \text{ for any } k. \]

The argument is valid with slight modification if we require only that there is some permutation \( \{k_i\}_{i=1}^{n} \) of the integers \( k = 1, 2, \ldots, n \) such that for each \( k \), \( m+k \) is divisible by \( k_i \) for some \( i \).

Before proceeding to the general case we will require the following lemmata.
Lemma 3. For every $k$ and $J$, $k = 1, \ldots, n$ $0 \leq J \leq k$ let

$$A = \text{greatest common divisor (}\ J, k\ )$$

and

$$K = k - \text{Remainder}\left(\frac{m \cdot A}{k}\right),$$

then

1. $1 < K < k$

2. $\frac{(m+k) \cdot J}{k}$ is an integer

and

3. If $(J,k)$ and $(J',k')$ lead to the same value of $K$, then

$$\left|\frac{J}{k} - \frac{J'}{k'}\right| \geq \frac{1}{m+K}.$$

Proof.

1. $\frac{k}{A}$ is an integer $\leq k$ so the remainder of $m$ after division by $k/A$ lies between $0$ and $k-1$.

2. $k = k' \cdot A$ and $J = J' \cdot A$.

Write $m = k' \cdot \mu + r$, then $K = k-r$ and $m+K = k' \cdot (\mu+1)$ so

$$\frac{(m+k) \cdot J}{k} = \frac{k' \cdot (\mu+1) \cdot J' \cdot A}{k' \cdot A} = (\mu+1) \cdot J'. $$

3. Let $\frac{J}{k} = \frac{I}{m+K}$ and $\frac{J'}{k'} = \frac{I'}{m+K}$. If $\frac{J}{k} \neq \frac{J'}{k'}$, then $I \neq I'$, so

$$\frac{J}{k} - \frac{J'}{k'} = \frac{|I-I'|}{m+K} \geq \frac{1}{m+K}.$$
If on the other hand \( \frac{J}{k} = \frac{J'}{k'} \) and, say, \( k' < k \) then \( Jk' = kJ' \). Hence \( k' \) is a factor of \( k \cdot J' \). Since \( J' < k' \) it follows that \( k' \) is a factor of \( k \). Say that \( k = B \cdot k' \).

Therefore \( J = BJ' \) and so \( J \) and \( k \) have a common divisor bigger than \( I \). In fact if \( A' \) is the g.c.d. of \( (J', k') \), then \( A = BA' \) is the g.c.d. of \( (J, k) \).

Let \( K' = k' - \text{Remainder} \left( \frac{m \cdot A'}{k} \right) \)

\[
= k - \text{Remainder} \left( \frac{m \cdot A'B}{k'B} \right)
\]

\[
= k - \text{Remainder} \left( \frac{m \cdot A'}{k'} \right).
\]

We assumed that \( K = K' \) so we must have \( k = k' \).

Contradiction.

Lemma 4.

Let us define

\[
g_k(t) = \begin{cases} 
\sin \left( \frac{m+K}{2} t \right), & |t - \frac{2J\pi}{k}| < \frac{\pi}{m+K} \\
\frac{1}{\sin \left( \frac{k}{2} t \right)}, & \frac{\pi}{m+K} < |t - \frac{2J\pi}{k}| < \frac{\pi}{k} \\
& J = 0, 1, \ldots, k
\end{cases}
\]

then

\[
(*) \quad \prod_{k=1}^{n} |g_k(t)| \geq \prod_{k=1}^{n} \sin \left( \frac{m+K}{2} t \right) / \prod_{k=1}^{n} \sin \left( \frac{k}{2} t \right).
\]
Proof. It is sufficient to show that the term \( \sin \left( \frac{m+k}{2} t \right) \) occurs at most once on the left side of (*). Suppose not, then we must have

\[
\left| t - \frac{2J\pi}{k} \right| \leq \frac{\pi}{m+K}
\]

and

\[
\left| t - \frac{2J'\pi}{k'} \right| \leq \frac{\pi}{m+k'},
\]

and \( K = K' \).

But then

\[
\left| \frac{2J\pi}{k} - \frac{2J'\pi}{k'} \right| \leq \frac{2\pi}{m+K},
\]

which is impossible.

Lemma 5. \( |g_k(t)| \) is (1) symmetric about \( t = \frac{2J\pi}{2} \) for \( |t - \frac{2J\pi}{k}| < \frac{\pi}{k} \) and (2) decreasing in \( t \) for \( \frac{2J\pi}{k} \leq t < \frac{(2J+1)\pi}{k} \).

Proof. Because of periodicity it suffices to prove that the function

\[
h(t) = \begin{cases} 
\frac{\sin(Bt)}{\sin(At)}, & 0 \leq t \leq \frac{\pi}{2B} \\
\frac{1}{\sin(At)}, & \frac{\pi}{2B} \leq t \leq \frac{\pi}{2A}
\end{cases}
\]

is decreasing in \( t \) provided \( A < B \).
Since \( \frac{d}{dt} \log h(t) = \frac{1}{t} \left\{ \frac{Bt}{\tan(Bt)} - \frac{At}{\tan(At)} \right\} \), the result follows at once from the fact that \( \tan(x)/x \) is increasing for \( 0 \leq x \leq \frac{\pi}{2} \).

Lemma 6.

\[
| g_k \left( \frac{2J \pi}{k} + \frac{n \pi}{k(m+n)} \right) | = | g_k \left( \frac{n \pi}{k(m+n)} \right) | \leq \frac{8}{3\pi} \cdot \frac{m+k}{k}
\]

Proof. By periodicity, we can assume \( J = 0 \). We use the inequality \( \sin(n \pi x) \geq n \pi (1-x) \), \( 0 \leq x \leq 1 \).

\[
| g_k \left( \frac{n \pi}{k(m+n)} \right) | \leq \frac{1}{\sin \left( \frac{n \pi}{2(m+n)} \right)} \leq \frac{4(m+n)^2}{n \cdot (2m+n) \cdot \pi}
\]

\[
= \frac{m+n}{n} \cdot \frac{4(m+n)}{\pi(2m+n)}
\]

\[
\leq \frac{m+k}{k} \cdot \frac{8}{3\pi}.
\]

Lemma 7.

\[
| g_k \left( \frac{2J \pi}{k} \right) | = \frac{m+k}{k} \leq \frac{m+k}{k}
\]

Proof. L'Hospital.

Theorem. If \( \frac{2m}{m+n} \leq t \leq 2\pi - \frac{2m}{m+n} \) then

\[
| f_{m,n}(t) | \leq \left( \frac{8}{3\pi} \right)^2 \left[ \frac{n}{5} \right]
\]

71
Proof. Lemma 1, page 52, can be paraphrased as follows: for fixed \( t \), satisfying the above inequality, if \( J \) is the nearest integer to \( \frac{tk}{2\pi} \), then

\[
|t - \frac{2J\pi}{k}| \leq \frac{mn}{m+n}
\]

for at most \( n - 2 \cdot \left[ \frac{n}{5} \right] \) values of \( k, 1 \leq k \leq n \).

Let \( S \) be this set of \( k \)-values,

\[
\prod_{k=1}^{n} |g_k(t)| \leq \prod_{k \in S} |g_k\left(\frac{2J\pi}{k}\right)| \prod_{k \notin S} |g_k\left(\frac{2J\pi}{k} + \frac{mn}{k(m+n)}\right)|
\]

\[
\leq \prod_{k=1}^{\left[ \frac{n}{5} \right]} \frac{m+k}{k} \cdot \left(\frac{8}{3\pi}\right)^{2\left[ \frac{n}{5} \right]}
\]

We summarize all of the foregoing in the following lemma.

Lemma 8. Let \( f_{m,n}(t) \) be the characteristic function of the two sample Wilcoxon distribution. Let \( \chi^r_{m,n} \) be its \( 2r \)th cumulant and \( \sigma^2_{m,n} = mn(m+n+1)/12 \), its variance. Further let

\[
\psi_{m,n}(t) = \log f_{m,n}(t) + \frac{1}{2} \sigma^2_{m,n} t^2
\]

and let \( \varphi^r_{m,n}(t) \) be the Taylor series approximation to \( \psi_{m,n}(t) \) up to
and including the term of order $2r$. Then, for $0 \leq t < 2\pi/(m+n)$

(a) $\left| \psi_{m,n}(t) \right| < \frac{1}{4} \sigma_{m,n}^2 t^2$

(b) $\left| \psi^r_{m,n}(t) \right| < \frac{1}{4} \sigma_{m,n}^2 t^2$

(c) $\left| \psi_{m,n}(t) - \psi^r_{m,n}(t) \right| < \frac{\lambda_{m,n}^r t^{2r}}{(2r)!}$

and for $2\pi/(m+n) \leq t \leq 2\pi - 2\pi/(m+n)$,

(d) $f_{m,n}(t) < \left( \frac{8}{3\pi} \right)^2 \left[ \frac{n}{5} \right]$

3. **Local Limit Theorems.** We will assume in all that follows that there exists some $k \geq 1$ such that with $n \leq m$, $\frac{m}{n^k} \to 0$. With this understanding we will suppress the dependence on $m$ in the notation. Let

$$P_n(k) = P[u_{m,n} - 1/2 > m \cdot n]$$

$$Z = [u_{m,n} - 1/2 > m \cdot n]/\sigma_n$$

$$z = k/\sigma_n,$$ where $k$ is an integer if $m \cdot n$

is even and is an integer + 1/2 if $m \cdot n$ is odd.

Let

$$P_n(z) = P[Z \leq z]$$

$$\varphi(z) = (2\pi)^{-\frac{1}{2}} \exp(-z^2/2)$$

$$\Phi(z) = \int_{-\infty}^{z} \varphi(x) dx .$$

73
From the expansion (2.2) we have

\[
\frac{\lambda_n}{\sigma_n^{2r}} = \frac{B_{2r}}{2r!} \cdot \frac{(12)^r}{(mn(m+n+1))^{2r}} \sum_{k=1}^{n} \left( (n+k)^{2r} - k^{2r} \right) = O\left( \frac{1}{n^{r-1}} \right)
\]

Theorem 1. \[ |\sigma_n P_n(k) - \varphi(z) | \to 0. \]

Proof. From the formula for Fourier coefficients

\( (3.1) \quad 2\pi \sigma_n P_n(k) = \int_{-\pi}^{\pi} f_n(t) e^{-ikt} dt \)

\( (3.2) \quad = \int_{-\pi}^{\pi} e^{-ikt} \sigma_n z f_n(t) dt \)

Let

\[ \chi = t\sigma_n \]

then

\( (3.3) \quad 2\pi \sigma_n P_n(k) = \int_{-\pi \sigma_n}^{\pi \sigma_n} e^{-ixz} f_n\left( \frac{x}{\sigma_n} \right) d\chi. \)

Now

\[
\frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-izx - \frac{x^2}{2}} dx.
\]
Let
\[ R_n = 2\pi \left[ \sigma_n \mathcal{P}_n(k) - \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \right]. \]

and with \( \xi_n = \frac{2\pi}{m+n} \) write

\[ R_n = I_1 + I_2 + I_3 + I_4 \]

where,

\[ I_1 = \int_{-A}^{A} e^{-ixz} \left[ f_n \left( \frac{x}{\sigma_n} \right) - e^{-\frac{x^2}{2}} \right] dx. \]

\[ I_2 = \int_{\xi_n \sigma_n}^{\sigma_n} e^{-izx} f_n \left( \frac{x}{\sigma_n} \right) dx \]

\[ I_3 = \int_{\xi_n \sigma_n}^{\pi \sigma_n} e^{-izx} f_n \left( \frac{x}{\sigma_n} \right) dx \]

\[ I_4 = -\int_{|x| > A} e^{-izx - \frac{x^2}{2}} dx \]

Note that

\[ \xi_n \sigma_n = \frac{2\pi}{(m+n)} \cdot \left( \frac{mn(m+n+1)}{12} \right)^{1/2} \to \infty \text{ as } n \to \infty. \]
Now, by lemma 8 (a), for $|t| < 2\pi/(m+n)$

$$|f_{m,n}(t)| = \exp\left(\frac{\sigma_n^2 \cdot t^2}{2}\right) < \exp\left(-\frac{\sigma_n^2 \cdot t^2}{4}\right)$$

so that

$$|I_2| \leq 2 \int_A \epsilon \sigma_n |f_n\left(\frac{x}{\sigma_n}\right)| \, dx \leq 2 \int_A \epsilon \sigma_n e^{-\frac{x^2}{4}} \, dx$$

$$< 2 \int_A e^{-\frac{x^2}{4}} \, dx$$

Thus choosing $A$ sufficiently large, both $I_2$ and $I_4$ may be made arbitrarily small.

By lemma 8 (d)

$$|I_3| < 2\pi \sigma_n \sup_{\epsilon_n \leq \frac{\sigma_n}{\sigma_n}} f_n\left(\frac{t}{\sigma_n}\right) \leq 2\pi \left(\frac{mn(m+n+1)}{12}\right)^{1/2} \left(\frac{8}{3\pi}\right)^2 \frac{n}{5} \to 0$$

Therefore $R_n = I_1 + I_2 + I_3 + I_4 \to 0$ uniformly in $k$.

Q.E.D.

We now proceed to estimate the remainder when $\sigma_n P_n(k)$ is approximated by $\varphi(z)$. Theorem 2 is implied by Theorem 3. However, because of the notational complexity of 3 we follow Feller's scheme of proving them separately. Let $H_k(z) \overset{\text{def}}{=} \text{the } k^{\text{th}}$
Hermite Polynomial, \((\frac{d}{dx})^k e^{\frac{-x^2}{2}} = (-1)^k H_k(x) e^{\frac{-x^2}{2}}\)

Theorem 2.

\[| \sigma_n^P n(k) - \varphi(z) - \frac{\lambda_n^2}{4!} \frac{H_4(z)}{\sigma_n^4} \varphi(z) | = o\left(\frac{1}{n}\right)\]

Proof. From (3.1) we have

\[\sigma_n^P n(k) = \frac{1}{2\pi} \int_{-\infty}^{\infty} f_n\left(\frac{t}{\sigma_n}\right) e^{-i\pi z} dt\]

Also

\[\varphi(z) + \frac{\lambda_n^2}{4! \sigma_n^4} H_4(z) \varphi(z) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \left[ e^{-\frac{t^2}{2}} + \frac{\lambda_n^2}{4! \sigma_n^4} e^{-\frac{t^2}{2}} \right] e^{-i\pi z} dt\]

Again letting \(\xi_n = 2\pi/(m+n)\),

\[\int_{|t| \geq \xi_n \sigma_n} e^{-\frac{t^2}{2}} \left[ 1 + \left| \frac{\lambda_n^2}{4! \sigma_n^4} (it)^4 \right| \right] dt\]

tends to zero faster than any power of \(1/n\).

\[\int_{\xi_n \sigma_n \leq |t| \leq \pi \sigma_n} \left| f_n\left(\frac{t}{\sigma_n}\right) \right| dt = o\left(\frac{1}{n^k}\right) \text{ for any } k\]

so that

\[| \sigma_n^P n(k) - \varphi(z) - \frac{\lambda_n^2}{4! \sigma_n^4} H_4(z) \varphi(z) | <\]
\[
\int \left| \frac{f_n(t)}{\sigma_n} - e^{-\frac{t^2}{2}} - \frac{\lambda_n^2}{4!\sigma_n^4} t^4 e^{-\frac{t^2}{2}} \right| \, dt + o(n^{-1})
\]

Now introducing the general inequality

\[
(3.6) \quad |e^{a} - 1 - \beta| \leq (|a - \beta| + \frac{1}{2} \beta^2) e^\nu
\]

valid for \( \nu \leq \max(|a|, |\beta|) \), we have by lemma 8 (b)

\[
(3.7) \quad \int_{|t| < \varepsilon_n \sigma_n} -\frac{t^2}{2} \left| e_{n}(t/\sigma_n) - 1 - \frac{\lambda_n^2 t^4}{4!\sigma_n^4} \right| \, dt \leq
\]

\[
\int_{|t| < \varepsilon_n \sigma_n} -\frac{t^2}{2} \left( \left| \psi_n(t/\sigma_n) - \frac{\lambda_n t^4}{4!\sigma_n^4} \right| + \frac{1}{2} \left( \frac{\lambda_n t^4}{4!\sigma_n^4} \right)^2 \right) e^{-\frac{t^2}{4}} \, dt
\]

\[
< \int_{|t| < \varepsilon_n \sigma_n} -\frac{t^2}{4} \left( \frac{\lambda_n^2}{4!\sigma_n^4} \left( \frac{\lambda_n}{4!\sigma_n^4} \right)^2 t^8 \right) = o\left( \frac{1}{n} \right)
\]

Q.E.D.

Before proceeding to higher order expansions we introduce the following notation.

As before let \( \psi_n(t) = \log f_n(t) + \frac{1}{2} \sigma_n^2 t^2 \)

and

\[
\psi_n^r(t) = \sum_{j=2}^{r} \frac{\lambda_n}{(2j)!} (it)^{2j}
\]

78
Let \( P(it) = \sum_{j=1}^{[\frac{r}{2}]} \frac{1}{j!} \left[ \psi_n \left( \frac{t}{\sigma_n} \right) \right]^j \). The definition of \( P(it) \) differs from that of Feller because odd order cumulants vanish due to the symmetry of the distribution. Now \( P(it) \) is a polynomial of order \( r \left[ \frac{r}{2} \right] \) in \( \frac{1}{\sigma_n} \) whose coefficients are polynomials in \( t, \lambda_n^2, \lambda_n^3, \ldots \). We may rearrange \( P(it) \) in powers of \( \frac{1}{\sigma_n} \).

Denote the coefficient of \( \sigma_n^{-k} \) by \( q_k(it) \) and let \( Q_k(z) \) be the polynomial such that \( Q_k(z) \cdot \phi(z) \) has Fourier transform \( e^{-\frac{t^2}{2}} q_k(it) \).

Thus

\[
Q_4(z) = \frac{\lambda_n^2}{4!} H_4(z) \\
Q_6(z) = \frac{\lambda_n^3}{6!} H_6(z) \\
Q_8(z) = \left[ \frac{\lambda_n^4}{8!} + \frac{\lambda_n^2}{4!} \left( \frac{2}{4!} \right)^2 \right] H_8(z)
\]

where \( H_k(z) \) is the \( k^{th} \) Hermite polynomial.

**Theorem 3.** \[ | \sigma_n P_n(k) - \phi(z) - \sum_{j=2}^{r} \sigma_n^{-2j} Q_{2j}(z) \phi(z) | = O \left( \frac{n}{2} - 1 \right) \]

**Proof.** As in Theorem 2 we write

\[
\sigma_n P_n(k) = \frac{1}{2\pi} \int_{-\infty}^{\infty} f_n \left( \frac{t}{\sigma_n} \right) e^{-itz} dt
\]
(3.8) \( \varphi(z) + \sum_{k=2}^{r \left[ \frac{n}{2} \right]} \frac{Q_{2k}(z)}{\sigma_{2k}^n} \varphi(z) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-itz} \left[ (e^{- \frac{t^2}{2}} + p(it)e^{- \frac{t^2}{2}}) \right] dt \)

so that

\[ | \sigma_n p_n(k) - \varphi(z) - \sum_{k=2}^{r \left[ \frac{n}{2} \right]} \frac{Q_{2k}(z)}{\sigma_{2k}^n} \varphi(z) | \leq \]

\[ + \int_{|t| < \frac{\pi \sigma_n}{\epsilon_n \sigma_n}} |f_n(t/\sigma_n) - (1 + p(it))e^{- \frac{t^2}{2}} | dt + \]

\[ \int_{\epsilon_n \sigma_n}^{\pi \sigma_n} |f_n(t/\sigma_n) | dt + \int_{\epsilon_n \sigma_n}^{\pi \sigma_n} e^{- \frac{t^2}{2}} | 1 + p(it) | dt. \]

With \( \epsilon_n = 2\pi/(m+n) \), the last two terms tend to zero at the required rate. To estimate the first term we use the general inequality

(3.9) \[ | e^\alpha - 1 - \sum_{k=1}^{r-1} \frac{1}{k!} \alpha^k | \leq e^\nu \left( \| \alpha - \beta \| + \frac{1}{(r-1)!} \| \beta \| ^r \right) \]

valid for \( \nu > \max(\| \alpha \|, \| \beta \|) \). Now for \( |t| < \epsilon_n \)

\[ | \psi_n(t) - \psi_n^{r_2}(t) | < \frac{r t^{2r}}{\lambda_n^{(2r)!}}, \]

and

\[ | \psi_n^{r_2}(t) | < \frac{t^2 \sigma_n^2}{4}, \quad | \psi_n(t) | < \frac{t^2 \sigma_n^2}{4}. \]
Therefore

\begin{equation}
(3.10) \quad \int_{|t| < \frac{\epsilon}{\sigma_n}} |f_n(t/\sigma_n) - (1 + p(it))e^{-\frac{t^2}{2}}| \, dt
\end{equation}

\[ = \int_{|t| < \frac{\epsilon}{\sigma_n}} e^{-\frac{t^2}{2}} \left| e_n(t/\sigma_n) - 1 - \sum_{j=1}^{[\frac{r}{2}]} \frac{1}{j! \left[ \frac{r}{2} \right]!} \left[ \psi_n(t/\sigma_n) \right]^j \right| \]

\[ < \int_{|t| < \frac{\epsilon}{\sigma_n}} e^{-\frac{t^2}{2}} \left( \frac{r t^{2r}}{(2r)! \sigma_n^{2r}} + \frac{1}{\left[ \frac{r}{2} \right]!} \left| \psi_n(t/\sigma_n) \right|^r \right) e^{\frac{t^2}{4}} \]

\[ \left| \psi_n \left( \frac{t}{\sigma_n} \right) \right| < \left( \frac{2^4 t^4}{4! \sigma_n^4} \right) \quad \text{for} \quad |t| < \frac{\epsilon}{\sigma_n} \]

and the additional terms on the left of (3.8) are all of order less than \( n^{-r} \) so that the conclusion follows.

4. **Expansions for Distributions.** We will need the following lemma due to Feller (1).

**Lemma 2.** Let \( F \) be a probability distribution with vanishing expectation and characteristic function \( f \). Suppose that \( F-G \) vanishes at \( \pm \infty \) and that \( G \) has a derivative \( g \) such that \( |g| \leq m \). Finally suppose that \( g \) has a continuously differentiable Fourier transform \( \gamma \) such that \( \gamma(0) = 1 \) and \( \gamma'(0) = 0 \). Then
\[ (4.1) \quad \left| F(x) - G(x) \right| \leq \frac{1}{\pi} \int_{-T}^{T} \left| \frac{f(t) - \gamma(t)}{t} \right| \, dt + \frac{24m}{\pi T} \]

Theorem 4.

\[ \left| F_n(z) - \Phi(z) - \frac{\lambda_n}{4!} \frac{H_3(z)}{\sigma_n^4} \varphi(z) \right| = o\left( \frac{1}{n} \right) \]

Proof. In Lemma 2 put;

\[ G(x) = \Phi(x) + \frac{\lambda_n}{4!} \frac{H_3(x)}{\sigma_n^4} \varphi(x) \]

then

\[ |G'(x)| = |g(x)| = |\varphi(x) + \frac{\lambda_n}{4!} \frac{H_4(x)}{\sigma_n^4} \varphi(x)| < m \]

\[ \gamma(t) = e^{-\frac{\lambda_n}{4!} \frac{(it)^4}{\sigma_n^4}} \left( 1 + \frac{\lambda_n}{4!} \frac{(it)^4}{\sigma_n^4} \right) \]

Let \( T = \pi \sigma_n \) so that

\[ \frac{24m}{\pi T} \leq \frac{24}{\pi \sigma_n^2} \left( \frac{1}{\sqrt{2\pi}} + \frac{\lambda_n}{4!} \frac{\sup H_4(x) e^{-\frac{x^2}{2}}}{\sqrt{2\pi}} \right) \]

Now \( H_4(x) \leq K \sqrt{24} e^{-\frac{x^2}{2}} \) for all \( x \) where \( K < 2 \) is a constant.

\[ (4.2) \quad \frac{24m}{\pi T} < \frac{24}{\pi \sqrt{2\pi} \sigma_n} \left( 1 + \frac{\lambda_n^2}{2\sigma_n^4} \right) = o\left( \frac{1}{n^{3/2}} \right) \]

82
Thus

\[(4.3) \quad |F_n(\chi) - G(\chi)| \leq \int_{-\sigma_n}^{\sigma_n \pi} \left| \frac{f_n(t/\sigma_n) - \gamma(t)}{t} \right| \, dt + o\left(\frac{1}{n^{3/2}}\right)\]

As in theorem 2, the contribution of \(|t| \geq \varepsilon_n \sigma_n\) tends to zero as \(\frac{1}{n^k}\) for all \(k\). Also for \(|t| < \varepsilon_n \sigma_n\) we may estimate the numerator in the integrand by

\[e^{-\frac{t^2}{2}} \left(\frac{24}{4! \sigma_n^4} + \left(\frac{\lambda_n}{4 \sigma_n^4}\right)^2 \cdot t^8 \right)\]

and thus

\[|F_n(\chi) - G(\chi)| = o\left(\frac{1}{n}\right)\]

Corollary 1.

\[|F_n(\chi) - \phi(\chi)| = o\left(\frac{1}{n}\right)\]

Proof: The proof follows immediately from

\[\frac{\lambda_n^2}{4! \sigma_n^4} H_3(\chi) \quad \phi(\chi) = \frac{\lambda_n^2}{4 \sigma_n^4} \cdot \frac{1}{\sqrt{2\pi}} \cdot \frac{2\sqrt{6}}{24} = o\left(\frac{1}{n}\right)\]

5. Error Bounds. The calculations used in theorems 3 and 4 may also be used to derive explicit bounds on approximation error. Thus from (3.4), (3.5), (3.7), and (4.3) with \(\varepsilon_n = 2\pi/(m+n)\)
\[ (5.1) \quad | F_n(z) - \phi(z) - \frac{\lambda^2 \cdot H_3(z)}{4 \cdot \sigma_n^4} \phi(z) | < \]

\[ 2\pi \left( \frac{8}{3\pi} \right) \log (m+n) + \int_{|t| \geq \varepsilon \sigma_n} e^{-\frac{2}{4}} \left( \frac{\frac{1}{2} \left( \frac{\lambda t^2}{4 \cdot \sigma_n^4} \right)^2}{\lambda t^3} \right) dt + \frac{\lambda^2 t^3}{4 \cdot \sigma_n^4} \]

\[ \int_{|t| \leq \varepsilon \sigma_n} e^{\frac{2}{4}} \left( \frac{\frac{1}{2} \left( \frac{\lambda t^2}{4 \cdot \sigma_n^4} \right)^2}{\lambda t^3} \right) dt + \frac{24m}{\pi^2 \sigma_n} \]

where \( m \) is an upper bound for

\[ | \phi(z) + \frac{\lambda^2}{4 \cdot \sigma_n^4} \cdot H_4(z) \cdot \phi(z) | \]

and thus we may choose,

\[ m = \frac{1}{\sqrt{2\pi}} \left( 1 + \frac{2\lambda}{\sigma_n \sqrt{4\pi}} \right) < \frac{1}{\sqrt{2\pi}} \left( 1 + \frac{2.25 \sum_{k=1}^{n} (m+k)^4 - k^4)}{(mn(m+n+1))^2} \right) \]

Thus,

\[ (5.3) \quad \frac{24m}{\pi \sigma_n^2} < \frac{3.5}{(mn(m+n+1))^{\frac{1}{2}}} \left( 1 + \frac{2.25 \sum_{k=1}^{n} (m+k)^4 - k^4)}{(mn(m+n+1))^2} \right) \]

84
\[
\int e^{-\frac{t^2}{2}} \left( \left| \frac{1}{t} \right| + \left| \frac{\lambda_n t^3}{4! \sigma_n^4} \right| \right) dt < \\
\text{if } |t| \geq \epsilon_n \sigma_n
\]

\[
\frac{12 \sqrt{2 \pi (m+n)}}{2\sqrt{n} (mn(m+n+1))^{\frac{1}{2}}} \left( 1 - \Phi \left( \frac{2\pi (mn(m+n+1))^{\frac{1}{2}}}{12 \epsilon_n (m+n)} \right) \right) + \\
\frac{1}{5} \left( \sum_{k=1}^{n} \left( \frac{(m+k)^4 - k^4}{mn(m+n+1)} \right) \right) \cdot \left( \frac{2\pi^2 (mn(m+n+1))}{12(m+n)^2} \right)
\]

where

\[
\Gamma(a, x) = \int_{x}^{\infty} e^{-t} t^{a-1} dt
\]

and

\[
\int e^{-\frac{t^2}{4}} \left( \left| \frac{\lambda_n t^3}{4! \sigma_n^4} \right| + \left| \frac{1}{2} \left( \frac{\lambda_n}{4! \sigma_n^4} \right) t^7 \right| \right) dt < \\
\text{if } |t| \leq \epsilon_n \sigma_n
\]

\[
\frac{4}{5} \left( \sum_{k=1}^{n} \frac{(m+k)^4 - k^4}{mn(m+n+1)} \right) \left( 1 + \frac{6}{5} \sum_{k=1}^{n} \frac{(m+k)^4 - k^4}{mn(m+n+1)} \right).
\]
The first term on the right of (5.1) is asymptotically of lower order than the remaining correction terms. For moderate sample sizes, however, it provides trivial bounds. In computing numerical bounds the product of the approximating functions given in lemma 4 is easily computed numerically with any desired precision. The resulting bound then disappears very rapidly.

The term on the right of (5.6) is the dominating term in the bound and thus to assure uniform 2D accuracy with this approximation requires \( m, n > 100 \).
REFERENCES


APPENDIX 1

COMPUTER PROGRAMS

All programs are written in FORTRAN IV and have been run on a CDC 3400 computer. Running times are given for the maximum sample sizes given. Maximum sample sizes were dictated by the 25,000 word memory capacity of this computer. Comment statements in each program describe how sample sizes may be enlarged if computer memory permits.

Program 1     One Sample Scores

Program 1 computes the distribution of the general one sample scores statistic for sample sizes \( N \leq 50 \) when scores are computed to two significant figures. For the special case of the one sample Wilcoxon statistic, \( N \) may be increased to 99. Scores are entered as two digit non-negative integers. Scores computed with greater precision may be used with reduced sample size.

**Input**                          **Format**
Sample Size                        I2
Scores                             I5

**Output**
Program prints the argument, mass and cumulative distribution functions.
Programs 2A and 2B Two Sample Scores    Pages 92, 94

Programs 2A and 2B compute the distribution of the general two sample scores statistic. 2A is more efficient but requires more memory space than 2B. In the sample programs given, 2A is restricted to scores computed to two significant figures, 2B is unrestricted as to scores. For the special case of the two sample Wilcoxon, Program 2A should be used and sample sizes may be increased to \( mn \leq 1000 \).

**Input**

<table>
<thead>
<tr>
<th>Sample Sizes</th>
<th>Format</th>
</tr>
</thead>
<tbody>
<tr>
<td>I5</td>
<td></td>
</tr>
</tbody>
</table>

**Output**

Argument, mass and cumulative distribution functions.

Program 3 Spearman Rank Correlation Statistic Page 96

Tables the distribution of the statistic

\[
S_N = \sum iR_i
\]

**Input**

<table>
<thead>
<tr>
<th>Sample Size</th>
<th>Format</th>
</tr>
</thead>
<tbody>
<tr>
<td>I5</td>
<td></td>
</tr>
</tbody>
</table>

**Output**

Argument, mass and cumulative distribution function.
Program 4  General Rank Tests of Regression

Tables the distribution of the statistic

\[ S_N = \sum c_i a_n(R_i^i). \]

This program requires the capacity to store

\((N+3)\cdot N!\) words in memory. Regression constants and

scores are entered as integers.

**Input**

<table>
<thead>
<tr>
<th>Sample Size</th>
<th>I2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scores</td>
<td>I5</td>
</tr>
<tr>
<td>Regression Constants</td>
<td>I5</td>
</tr>
</tbody>
</table>

**Output**

Argument, mass and cumulative distribution functions.

Program 5  One and Two Sided Kolmogorov-Smirnov Test Statistics

Computes the distribution of the \((mn(m+n))^{\frac{1}{2}}\) -

multiple of the one and two sided Kolmogorov-

Smirnov test statistics.

**Input**

<table>
<thead>
<tr>
<th>Options: 1 if one sided</th>
<th>I1</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 if two sided</td>
<td>I1</td>
</tr>
<tr>
<td>Sample Sizes</td>
<td>2I5</td>
</tr>
</tbody>
</table>

**Output**

Argument and cumulative distribution function.
PROGRAM ONE
C PROGRAM TO TABLE ONE SAMPLE SCORES STATISTIC N LESS THAN OR EQUAL TO 50
C SCORES COMPUTED TO TWO SIGNIFICANT DIGITS
DIMENSION A(5000), B(5000), KK(50)
C IF SCORES ARE COMPUTED TO MORE THAN TWO SIGNIFICANT DIGITS
C A AND B ARRAYS SHOULD BE ENLARGED TO ACCOMMODATE THE SUM OF
C N SCORES EXPRESSED AS INTEGERS
READ 5, N
5 FORMAT (12)
DO 10 I=1, N
9 FORMAT(15)
10 CONTINUE
C SCORES ARE LOADED IN THE KK ARRAY AS INTEGERS
JJ=1
DO 15 I=1, N
JJ=JJ*KK(I)
15 CONTINUE
DO 20 I=1, JJ
A(I)=0,
B(I)=0,
20 CONTINUE
JJ=KK(I)*1
A(I)=1,
A(JJ)=1,
DO 25 I=2, N
K=KK(I)
DO 30 J=1, JJ
L=K+J
B(L)=A(J)
30 CONTINUE
JJ=JJ*KK(I)
DO 35 J=1, JJ
A(J)=A(J)+B(J)
B(J)=0,
35 CONTINUE
25 CONTINUE
F=0
DO 40 I=1, JJ
P=A(I)
DO 999 J=1, N
P=P/2
999 CONTINUE
F=F+P
J=1
PRINT 45, J, P, F
45 FORMAT(5X, I3, 5X, F0.4, 5X, F6.4)
40 CONTINUE
C CUMULATIVE DISTRIBUTION F,
END

Running Time  1 min 10 sec  N = 50
PROGRAM TWO

C PROGRAM TO TABLE TWO SAMPLE SCORES STATISTIC FOR SAMPLE SIZES
C M+N LE 10, SCORES COMPUTED TO TWO SIGNIFICANT DIGITS
DIMENSION A(11, 1000), B(2, 1000), KK(20), NB(11, 30)

C ARRAY DIMENSIONS
C A HAS M+1 ROWS AND THE SUM OF THE M LARGEST SCORES EXPRESSED AS INTEGERS
C KK HAS M+N, NB HAS M+1 ROWS AND M+N+1 COLUMNS
READ 10, M, N
10 FORMAT (215)
   LL = M+N
   KKP = 0
   DO 20 I = 1, LL
      READ 22, KK(I)
20 CONTINUE
   FORMAT (15)
   KKP = KKP + KK(I)
   CONTINUE

C SCORES ENTERED AS POSITIVE INTEGERS
   JJ = 1
   MP = M+1
   DO 30 I = 1, LL
      JJ = JJ + KK(I)
30 CONTINUE
   JJ = KK(I)

C CONTINUE
   JJ = KK(I)
   JJ = JJ + KK(I)
   A(1+JJ) = 1
   A(2+JJ) = 1
   MM = M-1
   DO 60 I = 2, MM
      K = KK(I)
      IP = I+1
      DO 70 J = 2, IP
         DO 80 J2 = 1, JJ
            L = K + J2
            B(2+L) = B(1+J2)
            B(1+J2) = A(J+J2)
300 CONTINUE
   DO 90 J2 = 1, L
      A(J+J2) = A(J+J2) + B(2+J2)
90 CONTINUE
   B(1+1) = 1
   JJ = JJ + KK(I)
   CONTINUE
   JJ = JJ + KK(I)

60 CONTINUE
   DO 100 I = MP, LL
   CONTINUE
   DO 110 J = 2, MP
      DO 120 J2 = 1, JJ
         L = K + J2
100 CONTINUE
   DO 110 J = 1, JJ
110 CONTINUE
   DO 120 J2 = 1, JJ
120 CONTINUE

92
B(2*L)=B(1,J2)
B(1,J2)=A(J*,J2)
120 CONTINUE
DO 130 J2=1,L
A(J*,J2)=A(J*,J2)*B(2,J2)
B(2,J2)=0
130 CONTINUE
110 CONTINUE
JJ=JJ+KK(I)
DO 115 J=1,JJ
B(1,J,J)=A(1,J,J)
115 CONTINUE
100 CONTINUE
NP=LL+1
DO 300 I=1,NP
NB(I,I)=1
300 CONTINUE
DO 400 I=2,NP
DO 500 J=1,NP
NB(I,J)=NB(I,(J-1))*NB((I-1),(J-1))
500 CONTINUE
400 CONTINUE
DO 600 I=2,NP
DO 700 J=2,NP
NB(I,J)=NB(I,J-1)+NB((I-1),J-1)
700 CONTINUE
600 CONTINUE
JJ=NB(MP,NP)
F=0
DO 200 I=1,KP
P=A(MP,I)/JJ
F=F+P
J=I-1
PRINT 210,J,P,F
210 FORMAT(15*5X,F6.4,5X,F6.4)
200 CONTINUE
END

Running Time 17 sec
PROGRAM TWO B

THIS PROGRAM TABLES THE GENERAL TWO SAMPLE SCORES STATISTIC FOR SAMPLE
SIZES M AND N LESS THAN OR EQUAL TO 100 , SCORES COMPUTED TO ANY DESIRED

PRECISION

READ 10,M,N
10 FORMAT (215)

LL=M+N
DO 20 I=1,LL
READ 15,NA(1,I)
15 FORMAT (F5+0)
20 CONTINUE

SCORES ENTERED IN NA ARRAY AS POSITIVE INTEGERS

MP=M+1
NP=N+1
II(2)=0
DO 30 I=2,M
II(I)=I-1
II(1)=0
DO 40 J=2,N
JM=J-1
IND=II(JM)
INDP=II(J)
DO 50 K=1,IND
INDPP=INDP*K
NA(J,INDPP)=NA(JM,K)+NA(1,I)
50 CONTINUE
JJ(J)=INDPP
40 CONTINUE
DO 60 J=2,N
II(J)=JJ(J)
60 CONTINUE
DO 70 I=MP,LL
II(I)=I-1
DO 80 J=2,M
JM=J-1
IND=II(JM)
INDP=II(J)
DO 90 K=1,IND
INDPP=INDP*K
NA(J,INDPP)=NA(JM,K)+NA(1,I)
90 CONTINUE
JJ(J)=INDPP
80 CONTINUE
DO 100 J=2,M
II(J)=JJ(J)
100 CONTINUE
70 CONTINUE
DO 110 I=1,NNDPP
NB(I)=NA(M*I)
110 CONTINUE
CALL ALGSORT(NH,INDPP)

C ALGSORT IS A SUBROUTINE WHICH REARRANGES THE NB ARRAY IN INCREASING ORDER
C
INDM=INDPP-1
K=1
NA(1,I)=NH(I)
NA(2,I)=1
DO 120 I=2,INDM
NA(2,I)=0
120 CONTINUE
DO 220 I=1,INDM
IF(NH(I)*EH,NH(I+1))130,140
130 NA(2,K)=NA(2,K)+1
GO TO 220
140 K=K+1
NA(I+1,K)=NH(I+1)
NA(2+K)=NA(2,K)+1
PRINT 100 U,K,NA(2,K)
100 FORMAT(5X,I5,5X,I5)
220 CONTINUE
DO 150 I=1,LL
ND(1,I)=I
150 CONTINUE
DO 160 J=2,M
DO 170 J=1,LL
ND(I,J)=0
170 CONTINUE
160 CONTINUE
DO 180 I=2*M
DO 190 J=2,LL
ND(I,J)=ND(I-J+1)+ND(I-J+1)
190 CONTINUE
180 CONTINUE
JD=ND(M,LL)
PRINT 111 I,J
1111 FORMAT(5X,I5)
F=0
DO 210 I=1,K
PRINT 100 U+1,NA(2,I)
P=NA(2+I)/JD
F=F*F
NARG=NA(I+1)
PRINT 200,NARG,P,F
200 FORMAT(5X,I5,5X,F7.4,5X,F7.4)
210 CONTINUE
END

Running Time 19 sec
PROGRAM Three
DIMENSION NA(11*1200), NR(11*1200)

C C C
C PROGRAM TO TABLE THE SPEARMAN RANK CORRELATION STATISTIC....
C N LESS THAN OR EQUAL TO 10
C ARRAY DIMENSIONS
C NA+NR REQUIRES N+1 ROWS AND ((N/5)+1)*N*NR(N+1)/2 COLUMNS
C
P=N,
N=6,
P=0,
LL=NP*(N+1)/2,
LP=(LL+2*N+1))/3,
IND=N,
NP=N+1,
LIMU=LL,
LIML=LIMU-LP+LL,
DO 20 I=1,LIMU
DO 30 J=1,NP
NA(J+1)=0
NH(J+1)=0
30 CONTINUE
20 CONTINUE
DO 40 I=1,N
NA(I+1)=1
40 CONTINUE
DO 50 J=2,NP
JM=J-1
NA(J,JM)=1
50 CONTINUE
DO 60 I=2,N
DO 70 J=1,N
INC=I-J
JP=J+1
DO 80 K=1,INC
NS=K+INC
IH=NH(J+NS)+NA(J+K)
IN=IH/5
NS=NS+LIMU*IN
NR=10*NP(J+NS)
NB(NJ+NS)=NH(J+NS)+NA(J+K)
NHA=NA(J+K)
NH8=0
DO 90 L=1,NRA
LM=LM+1
NRH=NRH*(10**LM)
90 CONTINUE
NRC=NRH*NK
DO 100 L=2,NP
NAP=NA(L+K)*NR
NB(L+NS)=NH(L+NS)*NAP
100 CONTINUE
NB(JP+NS)=NH(JP+NS)*NRC
80 CONTINUE
70 CONTINUE
IND=NS
DO 110 J=1,NP
DO 120 K=1,IND
PHTNT 700*I, J*K*NA(J,K)*NB(J,K)

700 FORMAT(9*I=9,2X,I2,9*J=9,2X,I2,9*K=9,2X,I2,5X,I10,5X,I10)
   NA(J,K)=NH(J,K)
   NB(J,K)=0

120 CONTINUE
110 CONTINUE
60 CONTINUE
   IS=1
   LIML=LIML
   LPI=LPI
   INDEX=((IND-1)/LIMU)+1
   DO 180 I=1,INDEX
      L=M-1
      LIML=LIML*LIMU*LIMU
      LPI=LPI*LIMU*LIMU
      DO 130 J=1,L
      IND=NA(I,J)-1
      NA(I,J)=0
      DO 140 K=1,IND
      DO 150 J=1,IND
      DO 160 K=2,NP
      IX=NA(K+1)/10
      IY=IX*10
      IZ=NA(K+1)-IY
      IS=IS+IZ
      NA(K+1)=I

150 CONTINUE
   NA(1,I)=NA(1,I)+IS
   IS=1
140 CONTINUE
   DO 170 J=2,NP
      IS=NA(J+1)+IS
170 CONTINUE
   NA(1,I)=NA(1,I)+IS
   IS=1
130 CONTINUE
160 CONTINUE
   DO 190 J=LIML+LPI
   DO 200 I=1,INDEX
      IM=I-1
      K=J*LIMU+1
      P=P+NA(I,K)
200 CONTINUE
   DO 160 JJ=2,N
      P=P/JJ
160 CONTINUE
   F=F+P
   PRINT 500*J,P,F
500 FORMAT(5X*13,5X*F10.4,5X*F10.6)
   P=0
190 CONTINUE
END

Running Time 25 sec
PROGRAM Four

PROGRAM TO TABLE GENERAL SCORES TESTS FOR REGRESSION
SAMPLE SIZE N LESS THAN OR EQUAL TO 6
DIMENSION A(7,800), B(7,800), K(800), KS(800)
ARRAY DIMENSIONS
A,B HAS N+1 ROWS AND N FACTORIAL COLUMNS
K IS N FACTORIAL
KS IS N
READ 20,N
20 FORMAT(5)
DO 30 I=1,N
   READ 35, KS(I), K(I)
35 FORMAT(2I5)
30 CONTINUE

SCORES ARE LOADED IN KS ARRAY, REGRESSION COEFFICIENTS IN K ARRAY
BOTH ARE EXPRESSED AS POSITIVE INTEGERS

LP=2
NP=N+1
N2=NP
DO 40 I=1,N
   DO 50 J=1,N
      A(I,J)=0,
50 CONTINUE
40 CONTINUE
DO 60 I=1,N
   IM=I+1
   A(I,1)=KS(IM)*K(I)
   A(I,1)=1,
60 CONTINUE
DO 70 J=1,N
   JM=J+1
   JS=K(J)*N2
   LX=A(J,K)+1
   L=-LP=LX
   B(I,J,L)=A(I,K)*SK
   B(J,J,L)=A(J,K)+1
   DO 100 K1=2,NM
      B(K1,L)=A(K1,K)
100 CONTINUE
DO 110 K1=JP,NP
   B(K1,L)=A(K1,K)
110 CONTINUE
LP=L=LX
90 CONTINUE
80 CONTINUE
N2=LP+1
LP=2
DO 120 J=2,N2
   DO 130 K=1,NP
      A(K,J)=B(K,J)
B(K,J)=0,
130 CONTINUE
120 CONTINUE
70 CONTINUE
   DO 140 J=1,800
   KC(J)=0
140 CONTINUE
   JM=A(1,2)
   DO 150 I=2,N2
   J=A(I,1)
   KC(J)=KC(J)+1
   IF(J,GT,JM)190,150
190 JM=J
150 CONTINUE
   F=0,
   DO 160 JJ=1,JM
   PP=KC(JJ)
   DO 170 I=2,N
   P=P/I
170 CONTINUE
   F=F+P
   JARG=JJ
   PRINT 180,JARG,P,F
180 FORMAT(5X,I3,5X,F6.4,5X,F6.4)
160 CONTINUE
END

Running Time 15 Sec
PROGRAM Five
TYPE REAL NA, NB, NU, NC, NE, NT, JBIN
DIMENSION NA(1000), NB(21, 1000), NC(21)
DIMENSION NE(10, 20)
C PROGRAM TO TABLE THE ONE AND TWO SIDED KOLMOGOROV SMIRNOV STATISTIC
C SAMPLE SIZES N, N LESS THAN OR EQUAL TO 20
C ARRAY DIMENSIONS
C NA IS 3*M*N+2*M+2
C NB IS N+1, 3*M*N+2*N+2
C NC IS N+1
C NE IS N+1, M+N
C ISIDE=0 ONE SIDED
C ISIDE=1 TWO SIDED
READ 10, M, N, ISIDE
10 FORMAT(215, 4X, I1)
NP=N+1
LL=M+N
MP=M+1
MN=M+1
MN=M+N
F=0;
MNP=MN+1
LU=LL+MN+2*N+1
X=MIN(F(M, N),
Y=5*MAXOF(M, N)
Z=MAX1F(X, Y)+5
IZ=Z
DO 20 I=1, LU
NA(I)=0
DO 30 J=1, NP
NB(J, I)=0
30 CONTINUE
20 CONTINUE
DO 40 I=1, NP
IH=I+1
IHM=IH*M
NC(I)=IMM-N
40 CONTINUE
IF(ISIDE)200, 200, 201
200 MNL=MNP
GO TO 202
201 MNL=MNP+IZ
202 DO 50 IH=MNL, LH
IHM=IH-N+1
DO 60 I=1, NP
IS=NC(I)+MN
NA(IS)=1
60 CONTINUE
DO 70 I=1, IHM, LU
NA(I)=0
70 CONTINUE
IF(ISIDE)203, 203, 204
204 IHM=2*MNP-IH
IHBM=IHBM+1
DO 205 I=1, IHBM
NA(I)=0
205 CONTINUE
203 DO 80 HM=1,NM
204 DO 90 I=1,NI
205 IS=N(I)
206 DO 100 J=NP,IM
207 NB(I,(J-IS))=NA(J)
100 CONTINUE
210 DO 110 J=IM,LU
211 NB(I,J)=0
110 CONTINUE
IF(ISIDE)90,90,207
207 DO 208 J=1,IM
208 NB(I,J)=0
208 CONTINUE
90 CONTINUE
DO 120 I=1,LU
NA(I)=0
120 CONTINUE
NT=0
DO 130 I=1,IM
DO 140 J=1,NI
NT=NT+NB(J,I)
140 CONTINUE
NA(I)=NT
NT=0
130 CONTINUE
IHU=IH+MNP
DO 150 I=1,IM
NB(I, I+IHU)=NA(I)
NA(I)=0
150 CONTINUE
DO 160 I=1,LU
NA(I)=0
160 CONTINUE
NT=0
DO 170 I=1,IL
NT=NT+NB(1,1)
NB(1,1)=0
170 CONTINUE
DO 600 I=1,LL
NE(I,1)=1
600 CONTINUE
DO 610 I=2,LM
DO 620 J=1,LL
NE(I,J)=NE(I,(J-1))+NE((I-1),(J+1))
620 CONTINUE
610 CONTINUE
DO 640 I=2,LM
DO 650 J=1,LL
NE(I,J)=NE(M,LL)
650 CONTINUE
640 CONTINUE
JBIN=NE(M,LL)
NARG=IH*MNP
P=NT/UBIN
F=F*P
NT=0
PRINT 500,NARG,P
500 FORMAT(5X,I5,5X,F6.4)
DO 180 I=1,LU
NB(1,I)=0
180 CONTINUE
50 CONTINUE
END

Running Time 1 min 10 sec
APPENDIX II

SAMPLE SIZES REQUIRED TO ACHIEVE SPECIFIED ACCURACY

Direct comparisons were made between the exact distribution and approximations by the uncorrected Normal, the Normal with one correction term, and the Hodges and Fix approximation which continues the expansion to three terms of the Edgeworth series. The minimum sample sizes at which the desired accuracies were attained were derived from these comparisons. These sample sizes are presented for a range of percentiles of the distribution as considerable variability was observed over the range. It should be noted that the validity of the three term expansion of Hodges and Fix remains to be proved. The observed errors in all cases decreased monotonically in \( m \) and \( n \) over the ranges considered.

<table>
<thead>
<tr>
<th>Range</th>
<th>2D</th>
<th>3D</th>
<th>4D</th>
</tr>
</thead>
<tbody>
<tr>
<td>.00 - .05</td>
<td>11</td>
<td>2</td>
<td>11</td>
</tr>
<tr>
<td>.05 - .1</td>
<td>11</td>
<td>3</td>
<td>45</td>
</tr>
<tr>
<td>.1 - .15</td>
<td>11</td>
<td>6</td>
<td>65</td>
</tr>
<tr>
<td>.15 - .2</td>
<td>11</td>
<td>7</td>
<td>85</td>
</tr>
<tr>
<td>.2 - .25</td>
<td>11</td>
<td>8</td>
<td>85</td>
</tr>
<tr>
<td>.25 - .3</td>
<td>11</td>
<td>7</td>
<td>85</td>
</tr>
<tr>
<td>.3 - .35</td>
<td>11</td>
<td>7</td>
<td>65</td>
</tr>
<tr>
<td>.35 - .4</td>
<td>11</td>
<td>4</td>
<td>65</td>
</tr>
<tr>
<td>.4 - .45</td>
<td>11</td>
<td>3</td>
<td>45</td>
</tr>
<tr>
<td>.45 - .5</td>
<td>11</td>
<td>2</td>
<td>19</td>
</tr>
</tbody>
</table>

In Excess of 300
### ONE TERM EXPANSION

<table>
<thead>
<tr>
<th>Range</th>
<th>2D</th>
<th></th>
<th>3D</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>m</td>
<td>n</td>
<td>m</td>
</tr>
<tr>
<td>.00 - .05</td>
<td>11</td>
<td>2</td>
<td>11</td>
</tr>
<tr>
<td>.05 - .1</td>
<td>11</td>
<td>2</td>
<td>11</td>
</tr>
<tr>
<td>.1 - .15</td>
<td>11</td>
<td>3</td>
<td>11</td>
</tr>
<tr>
<td>.15 - .2</td>
<td>11</td>
<td>2</td>
<td>11</td>
</tr>
<tr>
<td>.2 - .25</td>
<td>11</td>
<td>2</td>
<td>11</td>
</tr>
<tr>
<td>.25 - .3</td>
<td>11</td>
<td>2</td>
<td>11</td>
</tr>
<tr>
<td>.3 - .35</td>
<td>11</td>
<td>2</td>
<td>11</td>
</tr>
<tr>
<td>.35 - .4</td>
<td>11</td>
<td>3</td>
<td>11</td>
</tr>
<tr>
<td>.4 - .45</td>
<td>11</td>
<td>3</td>
<td>11</td>
</tr>
<tr>
<td>.45 - .5</td>
<td>11</td>
<td>2</td>
<td>11</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Range</th>
<th>4D</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>m</td>
<td>n</td>
<td>m</td>
<td>n</td>
</tr>
<tr>
<td>.00 - .05</td>
<td>16</td>
<td>16</td>
<td>18</td>
<td>14</td>
</tr>
<tr>
<td></td>
<td>25</td>
<td>12</td>
<td>37</td>
<td>11</td>
</tr>
<tr>
<td>.05 - .1</td>
<td>18</td>
<td>18</td>
<td>19</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td>22</td>
<td>15</td>
<td>26</td>
<td>14</td>
</tr>
<tr>
<td>.1 - .15</td>
<td>19</td>
<td>19</td>
<td>20</td>
<td>18</td>
</tr>
<tr>
<td></td>
<td>24</td>
<td>16</td>
<td>28</td>
<td>15</td>
</tr>
<tr>
<td>.15 - .2</td>
<td>19</td>
<td>19</td>
<td>21</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td>27</td>
<td>15</td>
<td>34</td>
<td>14</td>
</tr>
<tr>
<td>.2 - .25</td>
<td>18</td>
<td>18</td>
<td>19</td>
<td>16</td>
</tr>
<tr>
<td></td>
<td>30</td>
<td>13</td>
<td></td>
<td></td>
</tr>
<tr>
<td>.25 - .3</td>
<td>15</td>
<td>15</td>
<td>17</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td>28</td>
<td>11</td>
<td></td>
<td></td>
</tr>
<tr>
<td>.3 - .35</td>
<td>11</td>
<td>9</td>
<td>13</td>
<td>7</td>
</tr>
<tr>
<td>.35 - .4</td>
<td>11</td>
<td>7</td>
<td>13</td>
<td>6</td>
</tr>
<tr>
<td>.4 - .45</td>
<td>11</td>
<td>7</td>
<td>12</td>
<td>6</td>
</tr>
</tbody>
</table>

104
### THREE TERM EXPANSION

<table>
<thead>
<tr>
<th>Range</th>
<th>2D</th>
<th></th>
<th>3D</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>m</td>
<td>n</td>
<td>m</td>
<td>n</td>
</tr>
<tr>
<td>.00 - .35</td>
<td>11</td>
<td>2</td>
<td>11</td>
<td>4</td>
</tr>
<tr>
<td>.35 - .45</td>
<td>11</td>
<td>3</td>
<td>11</td>
<td>4</td>
</tr>
<tr>
<td>.45 - .5</td>
<td>11</td>
<td>2</td>
<td>11</td>
<td>4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Range</th>
<th>4D</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>m</td>
<td>n</td>
<td>m</td>
<td>n</td>
</tr>
<tr>
<td>.00 - .05</td>
<td>11</td>
<td>8</td>
<td>13</td>
<td>7</td>
</tr>
<tr>
<td>.05 - .1</td>
<td>11</td>
<td>8</td>
<td>13</td>
<td>7</td>
</tr>
<tr>
<td>.1 - .15</td>
<td>11</td>
<td>9</td>
<td>13</td>
<td>7</td>
</tr>
<tr>
<td>.15 - .2</td>
<td>11</td>
<td>9</td>
<td>13</td>
<td>7</td>
</tr>
<tr>
<td>.2 - .25</td>
<td>11</td>
<td>9</td>
<td>13</td>
<td>7</td>
</tr>
<tr>
<td>.25 - .3</td>
<td>11</td>
<td>9</td>
<td>13</td>
<td>7</td>
</tr>
<tr>
<td>.3 - .35</td>
<td>11</td>
<td>9</td>
<td>13</td>
<td>7</td>
</tr>
<tr>
<td>.35 - .4</td>
<td>11</td>
<td>7</td>
<td>14</td>
<td>5</td>
</tr>
<tr>
<td>.4 - .45</td>
<td>11</td>
<td>7</td>
<td>16</td>
<td>6</td>
</tr>
<tr>
<td>.45 - .5</td>
<td>11</td>
<td>7</td>
<td>15</td>
<td>5</td>
</tr>
</tbody>
</table>
**UNCLASSIFIED**

**Security Classification**

**DOCUMENT CONTROL DATA - R&D**

<table>
<thead>
<tr>
<th>1. ORIGINATING ACTIVITY (Corporate author)</th>
<th>5. AUTHOR(S) (Last name, first name, initial)</th>
</tr>
</thead>
<tbody>
<tr>
<td>DEPT. OF STATISTICS &amp; OPERATIONS RESEARCH STANFORD UNIVERSITY</td>
<td>ROGERS, W.F.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>2. REPORT SECURITY CLASSIFICATION</th>
<th>9b. ORIGINATOR'S REPORT NUMBER(S)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td># 145</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>3. REPORT TITLE</th>
<th>9a. TOTAL NO. OF PAGES</th>
</tr>
</thead>
<tbody>
<tr>
<td>EXACT NULL DISTRIBUTIONS AND ASYMPTOTIC EXPANSIONS FOR RANK TEST STATISTICS</td>
<td>105</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>6. REPORT DATE</th>
<th>7a. NO. OF REPS</th>
</tr>
</thead>
<tbody>
<tr>
<td>December 26, 1971</td>
<td>10</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>8a. CONTRACT OR GRANT NO.</th>
</tr>
</thead>
<tbody>
<tr>
<td>N00014-67-A-0112-0052</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>8b. PROJECT NO.</th>
<th>9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NR-042-002</td>
<td>#35 NSF GP-30711X</td>
</tr>
</tbody>
</table>

10. A V A I L A B I L I T Y / L I M I T A T I O N N O T I C E S

DISTRIBUTION OF THIS DOCUMENT IS UNLIMITED.

11. SUPPLEMENTARY NOTES

12. SPONSORING MILITARY ACTIVITY

Statistics & Probability Program
Office of Naval Research Code 436
Arlington, VA.

13. ABSTRACT.

Efficient algorithms are developed for inverting the probability generating functions of the distributions of rank test statistics. A method is given for deriving probability generating functions in a form suitable for inversion. Cases treated include one and two sample linear rank statistics and the Kolmogorov-Smirnov test statistics. A proof is given that one term of the Edgeworth series provides a valid asymptotic expansion for the Wilcoxon two sample distribution. Explicit bounds are given for the error in approximating the distribution by the one term expansion. It is shown that the error in the uncorrected normal approximation is of the order \( \max(n^{-1}, m^{-1}) \).
Rank Test
Kolmogorov-Smirnov Test
Two sample linear rank statistics
Wilcoxon

INSTRUCTIONS

1. ORIGINATING ACTIVITY: Enter the name and address of the contractor, subcontractor, grantee, Department of Defense activity, or other organization (corporate entity) issuing the report.

2a. REPORT SECURITY CLASSIFICATION: Enter the overall security classification of the report. Indicate whether "Restricted Data" is included. Marking is to be in accordance with appropriate security regulations.

2b. GROUP: Automatic downgrading is specified in DoD Directive 5200.10 and Armed Forces Industrial Manual. Enter the group number. Also, when applicable, show that optional markings have been used for Group 3 and Group 4 as authorized.

3. REPORT TITLE: Enter the complete report title in all capital letters. Titles in all cases should be unclassified. If a meaningful title cannot be selected without classification, show title classification in all capitals in parenthesis immediately following the title.

4. DESCRIPTIVE NOTES: If appropriate, enter the type of report, e.g., interim, progress, summary, annual, or final. Give the inclusive dates when a specific reporting period is covered.

5. AUTHOR(S): Enter the name(s) of author(s) as shown on or in the report. Enter last name, first name, middle initial.

6. REPORT DATE: Enter the date of the report as day, month, year, or month, year. If more than one date appears on the report, use date of publication.

7a. TOTAL NUMBER OF PAGES: The total page count should follow normal pagination procedures, i.e., enter the number of pages containing information.

7b. NUMBER OF REFERENCES: Enter the total number of references cited in the report.

8a. CONTRACT OR GRANT NUMBER: If appropriate, enter the applicable number of the contract or grant under which the report was written.

8b, 8c, & 8d. PROJECT NUMBER: Enter the appropriate military department identification, such as project number, subproject number, system numbers, task number, etc.

9a. ORIGINATOR'S REPORT NUMBER(S): Enter the official report number by which the document will be identified and controlled by the originating activity. This number must be unique to this report.

9b. OTHER REPORT NUMBERS: If the report has been assigned any other report numbers (either by the originator or by the sponsor), also enter this number(s).

9c. AVAILABILITY/LIMITATION NOTICES: Enter any limitations on further dissemination of the report, other than those imposed by security classification, using standard statements such as:

(1) "Qualified requesters may obtain copies of this report from DDC."

(2) "Foreign announcement and dissemination of this report by DDC is not authorized."

(3) "U.S. Government agencies may obtain copies of this report directly from DDC. Other qualified DDC users shall request through __________________________."

(4) "U.S. military agencies may obtain copies of this report directly from DDC. Other qualified users shall request through __________________________."

(5) "All distribution of this report is controlled. Qualified DDC users shall request through __________________________."

If the report has been furnished to the Office of Technical Services, Department of Commerce, for sale to the public, indicate this fact and enter the price, if known.

11. SUPPLEMENTARY NOTES: Use for additional explanatory notes.

12. SPONSORING MILITARY ACTIVITY: Enter the name of the departmental project office or laboratory sponsoring (paying for) the research and development. Include address.

13. ABSTRACT: Enter an abstract giving a brief and factual summary of the document indicative of the report, even though it may also appear elsewhere in the body of the technical report. If additional space is required, a continuation sheet shall be attached.

It is highly desirable that the abstract of classified reports be unclassified. Each paragraph of the abstract shall end with an indication of the military security classification of the information in the paragraph, represented as (TS), (S), (C), or (U).

There is no limitation on the length of the abstract. However, the suggested length is from 150 to 225 words.

14. KEY WORDS: Key words are technically meaningful terms or short phrases that characterize a report and may be used as index entries for cataloging the report. Key words must be selected so that no security classification is required. Identifiers, such as equipment model designation, trade name, military project code name, geographic location, may be used as key words but will be followed by an indication of technical context. The assignment of links, rules, and weights is optional.