A COMBINATORIAL CENTRAL LIMIT THEOREM.

by

Wassily Hoeffding

Special report to the Office of Naval Research of work at Chapel Hill under Project NR 642-031 for research in probability and statistics

Institute of Statistics
Mimeograph Series No. 46

July 1, 1951
A COMBINATORIAL CENTRAL LIMIT THEOREM.

By Wassily Hoeffding

Institute of Statistics, University of North Carolina.

1. Summary. Let \((Y_{n1}, \ldots, Y_{nn})\) be a random vector which takes on the \(n!\) permutations of \((1, \ldots, n)\) with equal probabilities. Let \(c_n(i,j)\), \(i,j = 1, \ldots, n\), be \(n^2\) real numbers. Sufficient conditions for the asymptotic normality of

\[ S_n = \sum_{i=1}^{n} c_n(i, Y_{ni}) \]

are given (theorem 3). For the special case \(c_n(i,j) = a_n(i)b_n(j)\) a stronger version of a theorem of Wald, Wolfowitz and Noether is obtained (theorem 4). A condition of Noether is simplified (theorem 1).

2. Introduction and statement of results. An example of what is here meant by a combinatorial central limit theorem is a solution of the following problem. For every positive integer \(n\) there are given \(2n\) real numbers \(a_n(i), b_n(i), i = 1, \ldots, n\). It is assumed that the \(a_n(i)\) are not all equal and the \(b_n(i)\) are not all equal. Let \((Y_{n1}, \ldots, Y_{nn})\) be a random vector which takes on the \(n!\) permutations of \((1, \ldots, n)\) with equal probabilities \(1/n!\). Under what conditions is

\[ S_n = \sum_{i=1}^{n} a_n(i)b_n(Y_{ni}) \]

asymptotically normally distributed as \(n \to \infty\)?

Throughout this paper a random variable \(S_n\) will be called asymptotically normal or asymptotically normally distributed if

1. Work done under the sponsorship of the Office of Naval Research.
\[
\lim_{n \to \infty} \Pr \left\{ S_n - ES_n \leq x \sqrt{\text{var } S_n} \right\} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-y^2/2} \, dy, \quad -\infty < x < \infty,
\]
where \( ES_n \) and \( \text{var } S_n \) are the mean and the variance of \( S_n \).

In the particular case \( a_n(i) = b_n(i) = i \) the asymptotic normality of \( S_n \) was proved by Hotelling and Pabst \( \sqrt{2} \). The first general result is due to Wald and Wolfowitz \( \sqrt{6} \) who showed that \( S_n \) is asymptotically normal if, as \( n \to \infty \),

\[
\frac{\frac{1}{n} \sum_{i=1}^{n} (a_n(i) - \bar{a}_n)^r}{\sqrt{\frac{1}{n} \sum_{i=1}^{n} (a_n(i) - \bar{a}_n)^2}^{r/2}} = o(1), \quad r = 3, 4, \ldots
\]

and

\[
\frac{\frac{1}{n} \sum_{i=1}^{n} (b_n(i) - \bar{b}_n)^r}{\sqrt{\frac{1}{n} \sum_{i=1}^{n} (b_n(i) - \bar{b}_n)^2}^{r/2}} = o(1), \quad r = 3, 4, \ldots
\]

where

\[
\bar{a}_n = \frac{1}{n} \sum_{i=1}^{n} a_n(i), \quad \bar{b}_n = \frac{1}{n} \sum_{i=1}^{n} b_n(i).
\]

Noether \( \sqrt{5} \) proved that condition (3) can be replaced by the weaker condition

\[
\lim_{n \to \infty} \frac{\sum_{i=1}^{n} (b_n(i) - \bar{b}_n)^r}{\left( \sum_{i=1}^{n} (b_n(i) - \bar{b}_n)^2 \right)^{r/2}} = 0, \quad r = 3, 4, \ldots
\]

This condition can be simplified as follows.

**Theorem 1.** Condition (4) is equivalent to either of the following two conditions:
\[ (5) \quad \lim_{n \to \infty} \frac{\sum_{i=1}^{n} \left| b_n(i) - \overline{b}_n \right|^r}{\left( \sum_{i=1}^{n} (b_n(i) - \overline{b}_n)^2 \right)^{r/2}} = 0 \quad \text{for some } r > 2; \]

\[ (6) \quad \lim_{n \to \infty} \frac{\max_{1 \leq i \leq n} (b_n(i) - \overline{b}_n)^2}{\sum_{i=1}^{n} (b_n(i) - \overline{b}_n)^2} = 0. \]

Hence conditions (2) and (5) or (2) and (6) are sufficient for the asymptotic normality of (1).

The proof is given in section 3. For a more general condition and a stronger but simpler condition see theorem 4 below.

One extension of this problem was considered by Daniels [1] who studied the asymptotic distribution of

\[ \sum_{i=1}^{n} \sum_{j=1}^{n} a_n(i, j) b_n(Y_{ni}, Y_{nj}). \]

The present paper is concerned with an alternative extension. It considers the distribution of

\[ (7) \quad S_n = \sum_{i=1}^{n} c_n(i, Y_{ni}), \]

where \( c_n(i, j), i, j = 1, \ldots, n, \) are \( n^2 \) real numbers, defined for every positive integer \( n. \) In the particular case \( c_n(i, j) = a_n(i) b_n(j), \) (7) reduces to (1).

Let

\[ (8) \quad d_n(i, j) = c_n(i, j) - \frac{1}{n} \sum_{g=1}^{n} c_n(g, j) - \frac{1}{n} \sum_{h=1}^{n} c_n(i, h) + \frac{1}{n^2} \sum_{g=1}^{n} \sum_{h=1}^{n} c_n(g, h). \]
Theorem 2. The mean and variance of

\[ S_n = \sum_{i=1}^{n} c_n(i, Y_{ni}) \]

are

\[ \text{ES}_n = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} c_n(i,j), \]

(9)

\[ \text{var} S_n = \frac{1}{n-1} \sum_{i=1}^{n} \sum_{j=1}^{n} \sigma_n^2(i,j). \]

(10)

Henceforth we assume that \( \sigma_n^2(i,j) \neq 0 \) for some \( (i,j) \), so that \( \text{var} S_n > 0 \). In the special case \( c_n(i,j) = a_n(i)b_n(j) \) this corresponds to the assumption that the \( a_n(i) \) are not all equal and the \( b_n(j) \) are not all equal.

Theorem 3. The distribution of \( S_n = \sum_{i=1}^{n} c_n(i, Y_{ni}) \) is asymptotically normal if

\[ \lim_{n \to \infty} \frac{\frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} d_n^2(i,j) \cdot r/2}{\sqrt{\frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} d_n^2(i,j)}} = 0, \quad r = 3, 4, \ldots \]

(11)

Condition (11) is satisfied if

\[ \lim_{n \to \infty} \frac{\max_{1 \leq i,j \leq n} d_n^2(i,j)}{\left( \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} d_n^2(i,j) \right)^{1/2}} = 0. \]

(12)

Theorems 2 and 3 will be proved in sections 4 and 5.

For the special case \( c_n(i,j) = a_n(i)b_n(j) \), theorem 3 immediately gives
Theorem 4. The distribution of $S_n = \sum_{i=1}^{n} a_n(i)b_n(Y_{ni})$ is asymptotically normal if

$$\lim_{n \to \infty} \frac{r}{2} - 1 = \frac{\sum_{i=1}^{n} (a_n(i) - \bar{a}_n)^r}{\sum_{i=1}^{n} (a_n(i) - \bar{a}_n)^{2r/2}} \frac{\sum_{i=1}^{n} (b_n(i) - \bar{b}_n)^r}{\sum_{i=1}^{n} (b_n(i) - \bar{b}_n)^{2r/2}} = 0,$$

$$r = 3, 4, \ldots$$

Condition (13) is satisfied if

$$\lim_{n \to \infty} \frac{\max_{1 \leq i \leq n} (a_n(i) - \bar{a}_n)^2}{\sum_{i=1}^{n} (a_n(i) - \bar{a}_n)^2} \frac{\max_{1 \leq i \leq n} (b_n(i) - \bar{b}_n)^2}{\sum_{i=1}^{n} (b_n(i) - \bar{b}_n)^2} = 0.$$

It will be observed that the symmetrical condition (13) contains Noether's condition (2) and (4) as a special case.

Let $X_n = (X_{n1}, \ldots, X_{nn})$ be independent of and have the same distribution as $Y_n = (Y_{n1}, \ldots, Y_{nn})$.

Theorem 5. The random variable

$$S'_n = \sum_{i=1}^{n} c_n(X_{ni}, Y_{ni})$$

has the same distribution as $S_n$ in (7).

In fact, the conditional distribution of $S'_n$ given that $X_n = p$, a fixed permutation of $(1, \ldots, n)$, is independent of $p$ because the distribution of $Y_n$ is invariant under permutations of its components.

The distribution of sums of the form (1) has attracted the attention of statisticians in connection with non-parametric tests (see, for example, $\mathbb{F}_2$, $\mathbb{F}_6$, $\mathbb{F}_3$) and sampling from a finite population (which leads to the case $a_n(i) = 0$ for $i > m$; cf. also Madow $\mathbb{F}_4$). More general sums of the form (7) or (15) are likewise of interest in non-parametric theory. Thus it follows from results of Lehmann and Stein $\mathbb{F}_3$ that a test of the hypothesis that
U₁, ..., Uₙ are independent and identically distributed, which is most
powerful
similar against the alternative that the joint frequency function is f₁(u₁)...
fₙ(uₙ) is based on a statistic of the form (7) with

\[ c_n(i,j) = \log f_i(u_j), \]

where the uⱼ are the observed sample values. If the n pairs (U₁, V₁), ..., 
(Uₙ, Vₙ) are independent and identically distributed, a test of the hypothesis
against that Uᵢ and Vᵢ are independent which is most powerful similar/the alternative
that their joint frequency function is f(u,v) is based on a statistic of the
form (15) with cᵢ(𝑖, 𝑗) = \log fᵢ(𝑖, 𝑣ᵢ), where (𝑢₁, 𝑣₁), ..., (𝑢ₙ, 𝑣ₙ) are the
observed values.

In these examples the numbers cᵢ(𝑖, 𝑗) are random variables. An
application of some of the present results to such cases will be considered
by the author in a forthcoming paper.

3. Proof of theorem 1. Let

\[ \xi_i = \frac{b_i - \bar{b}_n}{\left( \sum_{i=1}^{n} (b_i - \bar{b}_n)^2 \right)^{1/2}}, \]

\[ G_n = \max\left( \left| \xi_1 \right|, \ldots, \left| \xi_n \right| \right). \]

Theorem 1 asserts the equivalence of the three relations

\[ \lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} \xi_i^r = 0, \quad r = 3, 4, \ldots; \]

\[ \lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} \left| \xi_i \right|^r = 0 \quad \text{for some } r > 2; \]

\[ \lim_{n \to \infty} G_n = 0. \]
We have

(19) \[ \sum_{i=1}^{n} g_i^2 = 1, \]

and hence

(20) \[ G_n \leq 1. \]

If \( 0 \leq r < s < t \), Hölder's inequality gives

(21) \[ \sum_{i=1}^{n} |g_i|^s = \sum_{i=1}^{n} |g_i|^{t-r} \frac{r(t-s)}{t-r} \frac{t(s-r)}{t-r} \leq \left( \sum_{i=1}^{n} |g_i|^r \right)^{t-r} \left( \sum_{i=1}^{n} |g_i|^t \right)^{s-r} \]

If we let \( r = 2k, t = 2k + 2, k = 1, 2, \ldots \), it follows that relation (16) is equivalent to

(22) \[ \lim_{n \to \infty} \sum_{i=1}^{n} |g_i|^r = 0, \quad 2 < r < \infty \]

From (21) with \( r = 2 \) and (19) we have

(23) \[ \sum_{i=1}^{n} |g_i|^s \leq \left( \sum_{i=1}^{n} |g_i|^{t} \right)^{s-2} \frac{s-2}{t-2} \text{ if } 2 < s < t. \]

By (20), for \( r < s \)

(24) \[ \sum_{i=1}^{n} |g_i|^s \leq G_n^s \sum_{i=1}^{n} |g_i|^r \leq \sum_{i=1}^{n} |g_i|^r. \]

Inequalities (23) and (24) show that (17) implies (22). Since obviously (22) implies (17), these relations are equivalent.
The equivalence of (17) and (18) is seen from the inequalities

\[ G_n^r \leq \sum_{i=1}^{n} |s_i|^r \leq C_n^{r-2} \sum_{i=1}^{n} s_i^2 = C_n^{r-2}, \quad r > 2. \]

This completes the proof.

4. Proof of theorem 2. The subscript \( n \) in \( Y_{ni}, c_{n}(i,j) \) etc. will henceforth be omitted. We note that if the subscripts \( i_1, \ldots, i_m \) are distinct, the expected value of a function \( f(Y_{i_1}, \ldots, Y_{i_m}) \) is equal to

\[ \frac{1}{n(n-1)\ldots(n-m+1)} \sum' f(j_1, \ldots, j_m), \]

where the sum \( \Sigma' \) is extended over all \( m \)-tuples \( (j_1, \ldots, j_m) \) of distinct integers from 1 to \( n \).

Relation (9) follows immediately.

Let

\[ T_n = \sum_{i=1}^{n} d(i, Y_i), \]

where \( d(i, j) = d_n(i, j) \) is defined by (8). Using (9), we get

\[ T_n = S_n - ES_n. \]

Also

\[ \sum_{i=1}^{n} d(i, j) = 0 \quad \text{for all } j, \quad \sum_{j=1}^{n} d(i, j) = 0 \quad \text{for all } i. \]

Hence

\[ Ed(i, Y_i) = 0, \]

\[ Ed^2(i, Y_i) = \frac{1}{n} \sum_{j=1}^{n} d^2(i, j), \]
and if \(i \neq j\),

\[
\text{Ed}(i,Y_i)\text{d}(j,Y_j) = \frac{1}{n(n-1)} \sum_{g,h} \text{Ed}(i,g)\text{d}(j,h)
\]

\[
= \frac{-1}{n(n-1)} \sum_{g=1}^{n} \text{d}(i,g)\text{d}(j,g).
\]

Hence

\[
\text{var } S_n = \text{var } T_n = \sum_{i=1}^{n} \text{Ed}^2(i,Y_i) + \sum_{i,j} \text{Ed}(i,Y_i)\text{d}(j,Y_j)
\]

\[
= \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} \text{d}^2(i,j) - \frac{1}{n(n-1)} \sum_{g=1}^{n} \sum_{i,j} \text{d}(i,g)\text{d}(j,g)
\]

\[
= \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} \text{d}^2(i,j) + \frac{1}{n(n-1)} \sum_{g=1}^{n} \sum_{i=1}^{n} \text{d}^2(i,g),
\]

which gives relation (10).

5. Proof of theorem 3. Let

\[
M_r,n = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} \text{d}^r(i,j),
\]

(28)

\[
\overline{M}_r,n = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} |\text{d}(i,j)|^r,
\]

(29)

\[
D_n = \max_{1 \leq i,j \leq n} |\text{d}(i,j)|.
\]

(30)

Then \(\text{var } S_n = \frac{n}{n-1} M_2,n\). Since, by hypothesis, \(\text{var } S_n > 0\), we may

and shall assume that

\[
M_2,n = 1.
\]

(31)

Conditions (11) and (12) can now be written as

\[
\lim_{n \to \infty} M_r,n = 0, \quad r = 3, 4, \ldots
\]

(32)
and

\[ \lim_{n \to \infty} D_n = 0. \] (33)

That (33) implies (32) is seen from the inequalities

\[ |M_{r,n}| \leq \bar{M}_{r,n} \leq D_n^{r-2} M_{2,n} = D_n^{r-2} \quad \text{for } r > 2. \]

Since (cf. inequality (21))

\[ \bar{M}_{2k+1,n}^2 \leq M_{2k,n} M_{2k+2,n}, \quad k = 1, 2, \ldots, \]

condition (32) implies

\[ \lim_{n \to \infty} \bar{M}_{r,n} = 0, \quad r = 3, 4, \ldots \] (34)

As var \( S_n \to 1 \), it is now sufficient to demonstrate that under conditions (31) and (34), \( T_n = S_n - E S_n \) has a normal limiting distribution with mean 0 and variance 1. This will be proved by showing that

\[ \lim_{n \to \infty} E T_n^r = \begin{cases} 
1 \cdot 3 \cdots (r-1) & \text{if } r \text{ is even} \\
0 & \text{if } r \text{ is odd. } 
\end{cases} \] (35)

The \( r \)-th moment of \( T_n \),

\[ E T_n^r = E \sum_{i_1=1}^{n} \ldots \sum_{i_r=1}^{n} d(i_{1,Y_1}) \ldots d(i_{r,Y_r}), \] (36)

can be written as a sum of terms of the form

\[ I(r, e_1, \ldots, e_m) = \sum\limits_{i_1, \ldots, i_m} E d_{i_1,Y_{i_1}}^1 \ldots d_{i_m,Y_{i_m}}^m, \] (37)

where \( e_1 \geq 1, e_1 + \ldots + e_m = r \). The number of terms (37) is independent of \( n \).

It will be shown that
\[
(38) \quad \lim_{n \to \infty} I(r, e_1, \ldots, e_m) = 0 \quad \text{unless } r \text{ even, } m = r/2, e_1 = \ldots = e_m = 2,
\]
and that the number of terms \(I(r, 2, \ldots, 2)\) in (36) with \(r\) even equals \(1 \cdot 3 \cdots (r-1)\). Then (35) holds, and the theorem will be proved.

We have for \(n \to \infty\)
\[
(40) \quad I(r, e_1, \ldots, e_m) \sim n^{-m} \sum_{i_1} \sum_{j_1} \cdots \sum_{i_m} \sum_{j_m} d_1^{e_1(i_1, j_1)} \cdots d_m^{e_m(i_m, j_m)}.
\]

The right-hand side can be written as a sum of terms which, apart from the sign, are of the form
\[
(41) \quad n^{-m} J(r, p, q, e_1, \ldots, e_m) = n^{-m} \sum_{i_1} \sum_{j_1} \cdots \sum_{i_m} \sum_{j_m} d_1^{e_1(i_1, j_1)} \cdots d_m^{e_m(i_m, j_m)}
\]
where
\[
1 \leq p \leq m, \quad 1 \leq q \leq m, \quad 1 \leq c \leq p, \quad 1 \leq d_h \leq q, \quad (g, h = 1, \ldots, m),
\]
and for every integer \(u, 1 \leq u \leq p (1 \leq u \leq q)\) at least one \(c_g(d_h)\) is equal to \(u\). The number of terms (41) is independent of \(n\).

The sum \(J\) in (41) can be written as a product of \(s \geq 1\) sums of a similar form,
\[
(42) \quad J(r, p, q, e_1, \ldots, e_m) = \prod_{k=1}^s J(r_k, p_k, q_k, e_{k1}, \ldots, e_{km_k}),
\]
where
\[
(e_{k1}, \ldots, e_{km_k}), \quad k = 1, \ldots, s,
\]
are \(s\) disjoint subsets of \((e_1, \ldots, e_m)\),
\[
\begin{align*}
\{ & e_{kl} + \ldots + e_{km_k} = r_k, \\
& r_1 + \ldots + r_s = r, \\
& p_1 + \ldots + p_s = p, \quad q_1 + \ldots + q_s = q, \\
& m_1 + \ldots + m_s = m.
\end{align*}
\]

We observe that

\[(44) \quad 1 \leq p_k \leq m_k, \quad 1 \leq q_k \leq m_k, \quad m_k \leq r_k.\]

It will be assumed that \(s\) is the greatest possible number of factors into which \(J(r,p,q,e_1,\ldots,e_m)\) can be decomposed in the form (42). If \(s = 1\), the number of equalities between the subscripts \(c\) or between the subscripts \(d\) in (44) must be at least \(m-1\). The total number of subscripts \(c,d\) being \(2m\), there are at most \(m + 1\) distinct subscripts, so that \(p + q \leq m + 1\). If

\[(45) \quad (c_{g},d_{g}) = (c_{h},d_{h}) \quad \text{for some } (g,h), \ g \neq h,
\]

we have strict inequality. For an arbitrary \(s\) we have in a similar way

\[(46) \quad p_k + q_k \leq m_k + 1, \quad k = 1, \ldots, s,
\]

and hence

\[(47) \quad p + q \leq m + s,
\]

with strict inequality in the case (45).

By Hölder's inequality, from (44),

\[
\begin{align*}
\left| J(r,p,q,e_1,\ldots,e_m) \right| & \leq \prod_{g=1}^{m} \left( \sum_{i_1} \ldots \sum_{j_g} d(i_c,j_d) \right)^{r_{c,g} / r} \\
& = \prod_{g=1}^{m} \left( n^{p+q-1} r_{c,g} \right) \frac{e^{g / r}}{n} = n^{p+q-1} \frac{e^{g / r}}{r_{c,g}.n}. 
\end{align*}
\]
Similarly,

\[ |J(r, p, q, e_1, \ldots, e_m)| \leq n^{p+q-3} M_{r_1, n} \cdots M_{r_s, n}. \]

Hence, by (42),

\[ n^{-m} |J(r, p, q, e_1, \ldots, e_m)| \leq n^{p+q-s-m} M_{r_1, n} \cdots M_{r_s, n}. \]

If, for some \( k \), \( r_k = 1 \), then, by (44) and (43), \( p_k = q_k = m_k = e_{kl} = 1 \), and hence \( J = 0 \) by (27). Thus we may assume \( r_k \geq 2, k = 1, \ldots, s \). Then, by (34), 
\[ M_{r_1, n} \cdots M_{r_s, n} \rightarrow 0 \] unless \( r_1 = \ldots = r_s = 2 \). It now follows from (48) and (47) that

\[ \lim_{n \to \infty} n^{-m} J(r, p, q, e_1, \ldots, e_m) = 0 \]

except perhaps when \( r_1 = \ldots = r_s = 2 \).

If \( r_1 = \ldots = r_s = 2 \), we have

\[ n^{-m} J(r, p, q, e_1, \ldots, e_m) = 0 \quad (n^{p+q-s-m}). \]

By (44), \( r_k = 2 \) implies \( m_k = 1 \) or 2.

If \( m_k = 2 \), then \( e_{kl} = e_{k2} = 1 \) and \( p_k + q_k \leq 3 \) by (46). If \( p_k + q_k = 3 \), the corresponding \( J \)-factor is of the form

\[ \Sigma \Sigma \Sigma d(i, j)d(i, k) \quad \text{or} \quad \Sigma \Sigma \Sigma d(i, k)d(j, k), \]

both of which vanish by (27). If \( m_k = 2 \) and \( p_k + q_k = 2 \), we have case (45) and hence, by the remark following (47), \( p + q - s - m < 0 \). By (50), this implies (49).

Thus the only case where (49) need not hold is \( r_k = 2, m_k = 1 \) for \( k = 1, \ldots, s \). Then \( p_k = q_k = 1, e_{kl} = 2 \), hence

\[ r = 2s = 2m, \quad p = q = r/2 \]

\[ e_1 = \ldots = e_m = 2. \]
This proves relation (38), and (39) follows from

\[ I(r, 2, \ldots, 2) \sim n^{-\frac{r}{2}} J(r, \frac{r}{2}, \frac{r}{2}, 2, \ldots, 2) \]
\[ = n^{-\frac{r}{2}} \sqrt{J(2, 1, 1, 2)} \]
\[ = \pi^{r/2} n^{-1}. \]

It remains to determine the number of terms \( I(r, 2, \ldots, 2) \) in (36) when \( r \) is even. This is the number of ways the subscripts \( i_1, \ldots, i_r \) can be tied in \( r/2 \) groups of two, which is \((r-1)(r-3)\ldots3\cdot1\). The proof is complete.

REFERENCES


