A COMBINATORIAL CENTRAL LIMIT THEOREM[1]

By Wassily Hoeffding

Institute of Statistics, University of North Carolina

1. Summary. Let $(Y_{n1}, \cdots, Y_{nn})$ be a random vector which takes on the $n!$ permutations of $(1, \cdots, n)$ with equal probabilities. Let $c_n(i, j)$, $i, j = 1, \cdots, n$, 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 $a_n(i)$, $b_n(i)$, $i = 1, \cdots, n$. It is assumed that the $a_n(i)$ are not all equal and the $b_n(i)$ are not all equal. Let $(Y_{n1}, \cdots, Y_{nn})$ be a random vector which takes on the $n!$ permutations of $(1, \cdots, n)$ with equal probabilities $1/n!$. Under what conditions is

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

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

$$\lim_{n\to\infty} \Pr\{S_n - ES_n \leq x\sqrt{\operatorname{var} S_n}\} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{1}{2}y^2}\, dy, \qquad -\infty < x < \infty,$$

where ES_n and $\operatorname{var} S_n$ 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 [2]. The first general result is due to Wald and Wolfowitz [6], who showed that S_n is asymptotically normal if, as $n \to \infty$,

$$\frac{\frac{1}{n}\sum_{i=1}^{n} (a_n(i) - \bar{a}_n)^r}{\left[\frac{1}{n}\sum_{i=1}^{n} (a_n(i) - \bar{a}_n)^2\right]^{r/2}} = O(1), \qquad r = 3, 4, \cdots, \tag{2}$$

and

$$\frac{\frac{1}{n}\sum_{i=1}^{n} (b_n(i) - \bar{b}_n)^r}{\left[\frac{1}{n}\sum_{i=1}^{n} (b_n(i) - \bar{b}_n)^2\right]^{r/2}} = O(1), \qquad r = 3, 4, \cdots, \tag{3}$$

[1] Work done under the sponsorship of the Office of Naval Research.

where

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

Noether [5] proved that condition (3) can be replaced by the weaker condition

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

This condition can be simplified as follows.

THEOREM 1. *Condition* (4) *is equivalent to either of the following two conditions:*

$$\lim_{n\to\infty} \frac{\sum_{i=1}^{n} |\, b_n(i) - \bar{b}_n \,|^r}{\left[\sum_{i=1}^{n} (b_n(i) - \bar{b}_n)^2\right]^{r/2}} = 0 \qquad \textit{for some } r > 2; \tag{5}$$

$$\lim_{n\to\infty} \frac{\max_{1\le i\le n} (b_n(i) - \bar{b}_n)^2}{\sum_{i=1}^{n} (b_n(i) - \bar{b}_n)^2} = 0. \tag{6}$$

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

$$S_n = \sum_{i=1}^{n} c_n(i, Y_{ni}), \tag{7}$$

where $c_n(i, j)$, $i, j = 1, \cdots, n$, 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

$$d_n(i, j) = c_n(i, j) - \frac{1}{n}\sum_{g=1}^{n} c_n(g, j) - \frac{1}{n}\sum_{h=1}^{n} c_n(i, h) + \frac{1}{n^2}\sum_{g=1}^{n}\sum_{h=1}^{n} c_n(g, h). \tag{8}$$

THEOREM 2. *The mean and variance of*

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

are

$$ES_n = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} c_n(i,j), \tag{9}$$

$$\operatorname{var} S_n = \frac{1}{n-1}\sum_{i=1}^{n}\sum_{j=1}^{n} d_n^2(i,j). \tag{10}$$

Henceforth we assume that $d_n(i,j) \neq 0$ for some (i,j), so that $\operatorname{var} S_n > 0$. 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*

$$\lim_{n\to\infty} \frac{\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} d_n^r(i,j)}{\left[\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} d_n^2(i,j)\right]^{r/2}} = 0, \qquad r = 3, 4, \cdots. \tag{11}$$

Condition (11) *is satisfied if*

$$\lim_{n\to\infty} \frac{\max_{1\le i,j\le n} d_n^2(i,j)}{\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} d_n^2(i,j)} = 0. \tag{12}$$

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*

$$\lim_{n\to\infty} n^{\frac{1}{2}r-1} \frac{\sum_{i=1}^{n} (a_n(i) - \bar{a}_n)^r}{\left[\sum_{i=1}^{n} (a_n(i) - \bar{a}_n)^2\right]^{r/2}} \frac{\sum_{i=1}^{n} (b_n(i) - \bar{b}_n)^r}{\left[\sum_{i=1}^{n} (b_n(i) - \bar{b}_n)^2\right]^{r/2}} = 0, \quad r = 3, 4, \cdots. \tag{13}$$

Condition (13) *is satisfied if*

$$\lim_{n\to\infty} n \frac{\max_{1\le i\le n} (a_n(i) - \bar{a}_n)^2}{\sum_{i=1}^{n} (a_n(i) - \bar{a}_n)^2} \frac{\max_{1\le i\le n} (b_n(i) - \bar{b}_n)^2}{\sum_{i=1}^{n} (b_n(i) - \bar{b}_n)^2} = 0. \tag{14}$$

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

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

THEOREM 5. *The random variable*

$$S_n' = \sum_{i=1}^{n} c_n(X_{ni}, Y_{ni}) \tag{15}$$

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 $(1, \cdots, n)$, 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 nonparametric tests (see, for example, [2], [6], [3]) and sampling from a finite population (which leads to the case $a_n(i) = 0$ for $i > m$; cf. also Madow [4]). More general sums of the form (7) or (15) are likewise of interest in nonparametric theory. Thus it follows from results of Lehmann and Stein [3] that a test of the hypothesis that $U_1, \cdots, U_n$ are independent and identically distributed, which is most powerful similar against the alternative that the joint frequency function is $f_1(u_1) \cdots f_n(u_n)$ is based on a statistic of the form (7) with

$$c_n(i, j) = \log f_i(u_j),$$

where the u_j are the observed sample values. If the n pairs $(U_1, V_1), \cdots, (U_n, V_n)$ are independent and identically distributed, a test of the hypothesis that U_i and V_i are independent which is most powerful similar against the alternative that their joint frequency function is $f(u, v)$ is based on a statistic of the form (15) with $c_n(i, j) = \log f(u_i, v_j)$, where $(u_1, v_1), \cdots, (u_n, v_n)$ are the observed values.

In these examples the numbers $c_n(i, j)$ 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

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

$$G_n = \max(|g_1|, \cdots, |g_n|).$$

Theorem 1 asserts the equivalence of the three relations

$$\lim_{n\to\infty} \sum_{i=1}^{n} g_i^r = 0, \qquad r = 3, 4, \cdots; \tag{16}$$

$$\lim_{n\to\infty} \sum_{i=1}^{n} |g_i|^r = 0 \qquad \text{for some } r > 2; \tag{17}$$

$$\lim_{n\to\infty} G_n = 0. \tag{18}$$

We have

$$\sum_{i=1}^{n} g_i^2 = 1,$$

and hence for $r > 2$

$$G_n^r \le \sum_{i=1}^{n} |g_i|^r \le G_n^{r-2} \sum_{i=1}^{n} g_i^2 = G_n^{r-2}.$$

The equivalence of (16), (17) and (18) follows immediately.

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 $i_1, \cdots, i_m$ are distinct, the expected value of a function $f(Y_{i_1}, \cdots, Y_{i_m})$ is equal to

$$\frac{1}{n(n-1)\cdots(n-m+1)} \sum_{j_1,\cdots,j_m}' f(j_1, \cdots, j_m),$$

where the sum Σ' is extended over all m-tuples $(j_1, \cdots, j_m)$ of distinct integers from 1 to n. Relation (9) follows immediately.

Let

$$T_n = \sum_{i=1}^{n} d(i, Y_i), \tag{19}$$

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

$$T_n = S_n - ES_n. \tag{20}$$

Also

$$\sum_{i=1}^{n} d(i, j) = 0 \text{ for all } j, \qquad \sum_{j=1}^{n} d(i, j) = 0 \text{ for all } i. \tag{21}$$

Hence

$$Ed(i, Y_i) = 0,$$

$$Ed^2(i, Y_i) = \frac{1}{n} \sum_{j=1}^{n} d^2(i, j),$$

and if $i \neq j$,

$$\begin{aligned} Ed(i, Y_i)d(j, Y_j) &= \frac{1}{n(n-1)} \sum_{g,h}' d(i, g)d(j, h) \\ &= \frac{-1}{n(n-1)} \sum_{g=1}^{n} d(i, g)d(j, g). \end{aligned}$$

Therefore

$$\begin{aligned} \operatorname{var} S_n = \operatorname{var} T_n &= \sum_{i=1}^{n} Ed^2(i, Y_i) + \sum_{i,j}' Ed(i, Y_i)d(j, Y_j) \\ &= \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} d^2(i, j) - \frac{1}{n(n-1)} \sum_{g=1}^{n} \sum_{i,j}' d(i, g)d(j, g) \\ &= \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} d^2(i, j) + \frac{1}{n(n-1)} \sum_{g=1}^{n} \sum_{i=1}^{n} d^2(i, g), \end{aligned}$$

which gives relation (10).

5. Proof of Theorem 3. Let

$$M_{r,n} = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} d^r(i, j), \tag{22}$$

$$\bar{M}_{r,n} = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} | d(i,j) |^r, \tag{23}$$

$$D_n = \max_{1 \le i,j \le n} | d(i,j) |. \tag{24}$$

Then var $S_n = n/(n-1)\, M_{2,n}$. Since, by hypothesis, var $S_n > 0$, we may and shall assume that

$$M_{2,n} = 1. \tag{25}$$

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

$$\lim_{n \to \infty} M_{r,n} = 0, \qquad r = 3, 4, \cdots, \tag{26}$$

and

$$\lim_{n \to \infty} D_n = 0. \tag{27}$$

That (27) implies (26) is seen from the inequalities

$$| M_{r,n} | \le \bar{M}_{r,n} \le D_n^{r-2} M_{2,n} = D_n^{r-2} \qquad \text{for } r > 2.$$

Since

$$\bar{M}_{2k+1,n}^2 \le M_{2k,n} M_{2k+2,n}, \qquad k = 1, 2, \cdots,$$

condition (26) implies

$$\lim_{n \to \infty} \bar{M}_{r,n} = 0, \qquad r = 3, 4, \cdots. \tag{28}$$

As var $S_n \to 1$, it is now sufficient to demonstrate that under conditions (25) and (28), $T_n = S_n - ES_n$ has a normal limiting distribution with mean 0 and variance 1. This will be proved by showing that

$$\lim_{n \to \infty} ET_n^r = \begin{cases} 1 \cdot 3 \cdots (r-1) & \text{if } r \text{ is even,} \\ 0 & \text{if } r \text{ is odd.} \end{cases} \tag{29}$$

The rth moment of T_n,

$$ET_n^r = E \sum_{i_1=1}^{n} \cdots \sum_{i_r=1}^{n} d(i_1, Y_{i_1}) \cdots d(i_r, Y_{i_r}), \tag{30}$$

can be written as a sum of terms of the form

$$I(r, e_1, \cdots, e_m) = \sum_{i_1, \cdots, i_m}{}' E d^{e_1}(i_1, Y_{i_1}) \cdots d^{e_m}(i_m, Y_{i_m}), \tag{31}$$

where $e_i \ge 1$, $e_1 + \cdots + e_m = r$. The number of terms (31) is independent of n. It will be shown that

$$\lim_{n \to \infty} I(r, e_1, \cdots, e_m) = 0 \quad \text{unless } r = 2m, \quad e_1 = \cdots = e_m = 2, \tag{32}$$

$$\lim_{n \to \infty} I(r, 2, \cdots, 2) = 1 \qquad \text{if } r \text{ even,} \tag{33}$$

and that the number of terms $I(r, 2, \cdots, 2)$ in (30) with r even equals $1 \cdot 3 \cdots (r-1)$. Then (29) holds, and the theorem will be proved.

We have for $n \to \infty$

$$I(r, e_1, \cdots, e_m) \sim n^{-m} \sum_{i_1, \cdots, i_m}' \sum_{j_1, \cdots, j_m}' d^{e_1}(i_1, j_1) \cdots d^{e_m}(i_m, j_m). \tag{34}$$

The right-hand side can be written as a sum of terms which, apart from the sign, are of the form

$$\begin{aligned} &n^{-m} J(r, p, q, e_1, \cdots, e_m) \\ &\quad = n^{-m} \sum_{i_1=1}^{n} \cdots \sum_{i_p=1}^{n} \sum_{j_1=1}^{n} \cdots \sum_{j_q=1}^{n} d^{e_1}(i_{c_1}, j_{d_1}) \cdots d^{e_m}(i_{c_m}, j_{d_m}), \end{aligned} \tag{35}$$

where

$$\begin{aligned} &1 \le p \le m, \qquad 1 \le q \le m, \\ &1 \le c_g \le p, \qquad 1 \le d_h \le q, \qquad (g, h = 1, \cdots, m), \end{aligned}$$

and for every integer u, $1 \le u \le p (1 \le u \le q)$ at least one $c_g (d_h)$ is equal to u. The number of terms (35) is independent of n.

The sum J in (35) can be written as a product of $s \ge 1$ sums of a similar form,

$$J(r, p, q, e_1, \cdots, e_m) = \prod_{k=1}^{s} J(r_k, p_k, q_k, e_{k1}, \cdots, e_{km_k}), \tag{36}$$

where

$$(e_{k1}, \cdots, e_{km_k}), \qquad k = 1, \cdots, s,$$

are s disjoint subsets of $(e_1, \cdots, e_m)$,

$$\begin{aligned} &e_{k1} + \cdots + e_{km_k} = r_k, \qquad r_1 + \cdots + r_s = r, \\ &p_1 + \cdots + p_s = p, \qquad q_1 + \cdots + q_s = q, \\ &\qquad m_1 + \cdots + m_s = m. \end{aligned} \tag{37}$$

We observe that

$$1 \le p_k \le m_k, \qquad 1 \le q_k \le m_k, \qquad m_k \le r_k. \tag{38}$$

It will be assumed that s is the greatest possible number of factors into which $J(r, p, q, e_1, \cdots, e_m)$ can be decomposed in the form (36). If $s = 1$, the number of equalities between the subscripts c or between the subscripts d in (35) 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 $p + q \le m + 1$. If

$$(c_g, d_g) = (c_h, d_h) \qquad \text{for some } (g, h),\ g \ne h, \tag{39}$$

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

$$p_k + q_k \le m_k + 1, \qquad k = 1, \cdots, s, \tag{40}$$

and hence

$$p + q \leq m + s, \tag{41}$$

with strict inequality in the case (39).

By Hölder's inequality, from (35),

$$| J(r, p, q, e_1, \cdots, e_m) | \leq \prod_{g=1}^{m} \Big(\sum_{i_1} \cdots \sum_{i_p} \sum_{j_1} \cdots \sum_{q} | d(i_{c_g}, j_{d_g}) |^r\Big)^{e_g/r}$$

$$= \prod_{g=1}^{m} (n^{p+q-1} \bar{M}_{r,n})^{e_g/r} = n^{p+q-1} \bar{M}_{r,n}.$$

Similarly,

$$| J(r_k, p_k, q_k, e_{k1}, \cdots, e_{km_k}) | \leq n^{p_k+q_k-1} \bar{M}_{r_k,n}.$$

Hence, by (36),

$$n^{-m} | J(r, p, q, e_1, \cdots, e_m) | \leq n^{p+q-s-m} \bar{M}_{r_1,n} \cdots \bar{M}_{r_s,n}. \tag{42}$$

If, for some k, $r_k = 1$, then, by (38) and (37), $p_k = q_k = m_k = e_{k1} = 1$, and hence $J = 0$ by (21). Thus we may assume $r_k \geq 2$, $k = 1, \cdots, s$. Then, by (28), $\bar{M}_{r_1 n} \cdots \bar{M}_{r_s n} \to 0$ unless $r_1 = \cdots = r_s = 2$. It now follows from (42) and (41) that

$$\lim_{n\to\infty} n^{-m} J(r, p, q, e_1, \cdots, e_m) = 0 \tag{43}$$

except perhaps when $r_1 = \cdots = r_s = 2$.

If $r_1 = \cdots = r_s = 2$, we have

$$n^{-m} J(r, p, q, e_1, \cdots, e_m) = O(n^{p+q-s-m}). \tag{44}$$

By (38), $r_k = 2$ implies $m_k = 1$ or 2. If $m_k = 2$, then $e_{k1} = e_{k2} = 1$ and $p_k + q_k \leq 3$ by (40). If $p_k + q_k = 3$, the corresponding J-factor is of the form

$$\sum_i \sum_j \sum_k d(i, j) d(i, k) \qquad \text{or} \qquad \sum_i \sum_j \sum_k d(i, k) d(j, k),$$

both of which vanish by (21). If $m_k = 2$ and $p_k + q_k = 2$, we have case (39) and hence, by the remark following (41), $p + q - s - m < 0$. By (44), this implies (43).

Thus the only case where (43) need not hold is $r_k = 2$, $m_k = 1$ for $k = 1, \cdots, s$. Then $p_k = q_k = 1$, $e_{k1} = 2$, hence

$$r = 2s = 2m, \qquad p = q = r/2$$

$$e_1 = \cdots = e_m = 2.$$

This proves relation (32), and (33) follows from

$$I(r, 2, \cdots, 2) \sim n^{-r/2} J\Big(r, \frac{r}{2}, \frac{r}{2}, 2, \cdots, 2\Big)$$

$$= n^{-r/2} [J(2, 1, 1, 2)]^{r/2}$$

$$= M_{2,n}^{r/2} = 1.$$

It remains to determine the number of terms $I(r, 2, \cdots, 2)$ in (30) when r is even. This is the number of ways the subscripts $i_1, \cdots, i_r$ can be tied in $r/2$ groups of two, which is $(r-1)(r-3)\cdots 3\cdot 1$. The proof is complete.

REFERENCES

[1] H. E. DANIELS, "The relation between measures of correlation in the universe of sample permutations," *Biometrika*, Vol. 33 (1944), pp. 129–135.

[2] H. HOTELLING AND M. PABST, "Rank correlation and tests of significance involving no assumption of normality," *Annals of Math. Stat.*, Vol. 7 (1936), pp. 29–43.

[3] E. L. LEHMANN AND C. STEIN, "On the theory of some nonparametric hypotheses," *Annals of Math. Stat.*, Vol. 20 (1949), pp. 28–45.

[4] W. G. MADOW, "On the limiting distributions of estimates based on samples from finite universes," *Annals of Math. Stat.*, Vol. 19 (1948), pp. 535–545.

[5] G. E. NOETHER, "On a theorem by Wald and Wolfowitz," *Annals of Math. Stat.*, Vol. 20 (1949), pp. 455–458.

[6] A. WALD AND J. WOLFOWITZ, "Statistical tests based on permutations of the observations," *Annals of Math. Stat.*, Vol. 15 (1944), pp. 358–372.