CONVERGENCE RATES FOR U-STATISTICS AND RELATED STATISTICS

by

William F. Grams and R. J. Serfling

FSU Statistics Report M222

September, 1971
The Florida State University
Department of Statistics
Tallahassee, Florida

Research of first author supported by the National Institute of General Medical Sciences under Training Grant 5-T01-GM00913-10.
ABSTRACT

Bounds are provided for the rates of convergence in the central limit theorem and the strong law of large numbers for U-statistics. The results are obtained by establishing suitable bounds upon the moments of the difference between a U-statistic and its projection. Analogous conclusions for the associated von Mises statistical functions are indicated. Statistics considered for exemplification are the sample variance and the Wilcoxon two-sample statistic.
1. **Introduction.** The data consists of c independent collections of
independent observations \( \{X_1^{(1)}, \ldots, X_{n_1}^{(1)}\}, \ldots, \{X_1^{(c)}, \ldots, X_{n_c}^{(c)}\} \) taken from
distributions \( F_1, \ldots, F_c \), respectively. Consider a parametric function
\( \theta = \theta(F_1, \ldots, F_c) \) for which there is an unbiased estimator. That is,

\[
(1.0) \quad \theta = \mathbb{E}_n(X_1^{(1)}, \ldots, X_{m_1}^{(1)}; \ldots; X_1^{(c)}, \ldots, X_{m_c}^{(c)})
\]

for some function \( h \) which will be assumed, without loss of generality, to
be symmetric within each of its c blocks of arguments. Corresponding to
the "kernel" \( h \), and assuming \( n_1 \geq m_1, \ldots, n_c \geq m_c \), the U-statistic for
estimation of \( \theta \) is obtained by averaging \( h \) symmetrically over the data:

\[
(1.1) \quad U = \frac{c}{n_1 \cdot \ldots \cdot n_c} \sum_{i=1}^{n_1} \ldots \sum_{i_m=1}^{n_m} h(X_{i_1}^{(1)}, \ldots, X_{i_m}^{(m)}; \ldots; X_{i_1}^{(c)}, \ldots, X_{i_m}^{(c)}).
\]

Here \( \{i_1, \ldots, i_m\} \) denotes a set of \( m \) distinct elements of the set
\( \{1, 2, \ldots, n_j\}, 1 \leq j \leq c \), and \( \sum_c \) denotes summation over all such
combinations.

The one-sample (c=1) U-statistics were introduced by Hoeffding [8]
and a central limit theorem (CLT) covering a wide class of such statistics
was proved. The treatment was generalized for c \( \geq 1 \) by Lehmann [10] and
Dwass [4]. In [9] Hoeffding proved the strong law of large numbers (SLLN)
for U-statistics (c=1). Later Berk [1] gave a different argument, exploiting
the reverse martingale character of a sequence of one-sample U-statistics.

It is the purpose of the present paper to exhibit rates of convergence
apropos to these limit theorems. Our method is to approximate \( U \) by its
projection,
(1.2) \( \hat{U} = \sum_{j=1}^{c} \sum_{i=1}^{n_j} E(U|X_1^{(j)}) - (N-1)\theta, \)

where \( N = n_1 + \ldots + n_c. \) (See Hájek [6] for exposition of the notion of projection of a statistic upon the basic observations.) Since the summands of \( \hat{U} \) are independent, it may be dealt with by standard theory.

We then infer conclusions about \( U \) by showing that \( U - \hat{U} \) is negligible.

This is accomplished by establishing, in §2, suitable bounds on the moments \( E|U-\hat{U}|^a. \) Application is made to the CLT in §3 and to the SLLN in §4. The main results of the paper are Theorems 2.1, 3.1 and 4.1.

It is shown in §5 that analogous results hold for the associated von Mises statistic, which is given by replacing \( F_1, \ldots, F_c \) by the respective sample df's in the formulation of \( \theta. \)

Many familiar statistics are of the U-type. See [8] and [10] for examples. We examine the sample variance and the Wilcoxon two-sample statistic in §6.

Finally, certain generalizations and further notions are discussed in §7.

2. The moments of \( U-\hat{U}. \) From (1.1) and (1.2) we readily obtain

\[
(2.1) \quad \hat{U} - \theta = \sum_{j=1}^{c} \sum_{i=1}^{n_j} \frac{m_j}{n_j} h_j^*(x_1^{(j)}),
\]

where

\[
(2.2) \quad h_j^*(x) = E[h(X_1^{(1)}, \ldots, \frac{X_1^{(j)}}{m_1}, \ldots, \frac{X_c^{(j)}}{m_c})|X_1^{(j)} = x] - \theta.
\]
It follows that \( \hat{U} - \theta \) may be expressed in the form (1.1) with the role of kernel played by

\[
(2.3) \quad g(x_{i1}^{(1)}, \ldots, x_{ic}^{(1)} ; \ldots; x_{i1}^{(c)}, \ldots, x_{ic}^{(c)}) = \sum_{j=1}^{c} \sum_{k=1}^{m} h^*(x_{ij}^{(j)}).
\]

Define now a further "kernel" \( H \) by

\[
(2.4) \quad H = h - g - \theta,
\]

and we have

\[
(2.5) \quad U - \hat{U} = \prod_{j=1}^{n} \left( \frac{1}{m_j} \right)^{-1} E H(x_{i1}^{(1)}, \ldots, \ldots, x_{ic}^{(c)})\Bigg|_{i1 \leq i \leq c}.
\]

Thus \( U - \hat{U} \) is of the form (1.1) with kernel \( H \). Note that \( E[H] = 0 \) and \( E[H|x_{1j}^{(j)}] = 0 \).

Let \( n = \min_{1 \leq i \leq c} n_i \).

**THEOREM 2.1.** Let \( r \) be a positive integer. If \( E[h^{2r}] = \infty \) (implied by \( E[h^{2r}] < \infty \)), then

\[
(2.6) \quad E(U - \hat{U})^{2r} = o(n^{-2r}), \quad n \to \infty.
\]

**PROOF.** By (2.5), the quantity in (2.6) may be written

\[
(2.7) \quad \prod_{j=1}^{c} \left( \frac{1}{m_j} \right)^{-2r} E \sum_{a=1}^{2r} E[H(x_{i1}^{(1)}, \ldots, \ldots, x_{ic}^{(c)})],
\]

where the indices are as in (1.1), with the additional suffix "a" identifying
the factor within the product, and $\Sigma$ denotes summation over all $\prod_{j=1}^{c} n_j^{2r}$ of the indicated terms. Clearly, the hypothesis of the theorem implies that $E(U - \hat{U})^{2r} < \infty$.

Let $M = m_1 + \ldots + m_c$ and consider a typical term in the sum in (2.7). If all $M$ indices occurring in one of the factors inside the expectation occur only in that factor, then the independence of that factor from the other factors implies that the product of factors has expectation zero. If $(M-1)$ of the indices in one of the factors occur only in that factor, then again the product of factors has expectation zero. For the conditional expectation, given all variables but the $(M-1)$ designated, is zero. Hence a term in (2.7) may have non-zero expectation only if each factor in the product contains at least two indices which appear in other factors of the product.

For the $a$-th factor and the $j$-th sample, let $q_{a}^{(j)}$ be the number of indices not repeated in other factors, and let $p_{a}^{(j)} = m_j - q_{a}^{(j)}$ be the number of indices repeated elsewhere. Among the repeated indices within the $j$-th sample, let $q_{o}^{(j)}$ be the number of distinct elements. Then clearly

$$2q_{o}^{(j)} \leq \sum_{a=1}^{2r} p_{a}^{(j)}.$$  

The number of ways of selecting these $\sum_{a=0}^{2r} q_{a}^{(j)}$ indices for the $j$-th sample is, making use of (2.8), of order

$$o(n_{j}^{\sum_{a=0}^{2r} q_{a}^{(j)}}) = o(n_{j}^{2m_{j} - \frac{1}{2} \sum_{a=1}^{2r} p_{a}^{(j)}}) = o(n_{j}^{2rm_{j} - \frac{1}{2} \sum_{a=1}^{2r} p_{a}^{(j)}}).$$
Here the implicit constants depend upon $r$ and $m_1, \ldots, m_c$, but not upon $n_1, \ldots, n_c$. This is true also of the number of ways of selecting the values $q_1^{(j)}, \ldots, q_{2r}^{(j)}$. Therefore the overall number of ways of selecting the indices for the $j$-th sample is of order given by the right-most term in (2.9). It follows that the number of terms in the sum in (2.7) for which the expectation is possibly non-zero is of order

$$O\left( \prod_{j=1}^{c} \prod_{n_1}^{2r} \prod_{a=1}^{n_j} p_a^{(j)} \right) = O\left( \prod_{j=1}^{c} n_j^{n-2r} \right),$$

(2.10)

since $\prod_{j=1}^{c} p_a^{(j)} > 2$. Thus (2.6) follows.

The case $r=1$ of Theorem 2.1 was proved by Hoeffding [8] and suffices for applications such as the CLT and SLLN. For information on the rates of these convergences, however, the generalization for $r > 1$ is relevant.

3. Rate of convergence in the CLT. The variance of the projection $\hat{U}$ is found from (2.1) to be

$$\sigma^2(\hat{U}) = \sum_{i=1}^{m_1} \frac{\xi_i}{n_i},$$

(3.1)

where $\xi_i = \text{Var}[h_{x_i}^{(1)}(x_{i})]$. Asymptotic normality theorems [8],[4],[10] for U-statistics state that

$$P\left[ \frac{(U-\theta) - \phi(t)}{\sigma(\hat{U})} \leq t \right] \rightarrow \phi(t), \quad n \rightarrow \infty,$$

(3.2)
where $\phi(t) = \int_{-\infty}^{t} (2\pi)^{-\frac{1}{2}} e^{-u^2/2} du$, $n = \min\{n_1, \ldots, n_c\}$ as previously, and it is assumed that $E h^2 < \infty$ and $n \sigma^2(\hat{U}) > B > 0$ as $n \to \infty$.

The rate of convergence in (3.2) is seen in the theorem below to satisfy a bound which improves with the order of the moments that may be assumed on $h$ (or $H$). If moments of all orders may be assumed, the bound may be brought "close" to the order $O(n^{-\frac{1}{2}})$, which in view of the Berry-Esseen theorem [3] is the best possible without specific assumptions on the underlying distributions $F_1, \ldots, F_c$ of $X^{(1)}_1, \ldots, X^{(c)}_1$. Thus, e.g., regarding the two-sample Wilcoxon statistic, which has a bounded kernel, we are able to corroborate remarks of Stoker [13] on improvement of the order $O(n^{-2/5})$ which he obtained.

**THEOREM 3.1.** Assume that $h$ satisfies

(3.3) \[ E|h_i^*(X^{(i)}_1)|^3 < \infty, \quad 1 \leq i \leq c \quad (\text{implied by } E|h|^3 < \infty) \]

and

(3.4) \[ n \sum_{i=1}^{c} m_i^2 r_i/n_i > B > 0, \quad n \to \infty. \]

16. further, $E h^{2r} < \infty$ (implied by $E h^{2r} < \infty$) for a positive integer $r$, then

(3.5) \[ \sup_{t} |P[(U-\theta)/\sigma(\hat{U}) \leq t] - \phi(t)| = O(n^{-r/(2r+1)}), \quad n \to \infty. \]

**PROOF.** We will apply a standard device (see, e.g., [2]). Namely, if

(3.6) \[ \sup_{t} |P[(U-\theta)/\sigma(\hat{U}) \leq t] - \phi(t)| = O(a_n), \quad n \to \infty, \]

for a sequence of constants $\{a_n\}$, then
(3.7) \( \sup_t \left| P[(U-\theta)/\sigma(\hat{U}) \leq t] - \Phi(t) \right| = O(a_n) + P[|U-\hat{U}|/\sigma(\hat{U}) > a_n], \ n \to \infty. \)

The proof is elementary.

Now, by the classical Berry-Esseen theorem, as stated in Loève [11], p. 288, it follows directly from (2.1) that

\[
(3.8) \sup_t \left| P[(\hat{U}-\theta)/\sigma(\hat{U}) \leq t] - \Phi(t) \right| \leq C \sum_{j=1}^{n} \sum_{i=1}^{m_j} \left| 3 \frac{\sum_{i=1}^{n_j} \sum_{j=1}^{m_i} E[h^*(X_i^j)|^3]}{\sigma^3(\hat{U})} \right|
\]

where \( C \) is a universal constant. It is checked easily that the RHS of (3.8) is \( O(n^{-\frac{1}{2}}) \) as \( n \to \infty \) subject to (3.4).

Consequently, for any sequence of constants \( a_n \) satisfying \( n^{-\frac{1}{2}} = O(a_n) \) as \( n \to \infty \), we have (3.6) and thus (3.7). We now utilize Theorem 2.1 in selecting the best sequence \( \{a_n\} \). By (3.1), (3.4) and Markov's inequality, we have

\[
(3.9) \ P[|U-\hat{U}|/\sigma(\hat{U}) > a_n] \leq n^{-r}(a_n)^{-2r} B^{-r} E(U-\hat{U})^{2r}.
\]

By Theorem 2.1, the RHS of (3.9) is \( O(n^{-r}a_n^{-2r}) \). Setting \( a_n = O(n^{-r}a_n^{-2r}), \) we obtain \( a_n = O(n^{-r/(2r+1)}). \]

**COROLLARY 3.1.** Assume that \( h \) has finite moments of all orders and that (3.4) holds. Then, for every \( \varepsilon > 0, \)

\[
(3.10) \sup_t \left| P[(U-\theta)/\sigma(\hat{U}) \leq t] - \Phi(t) \right| = O(n^{-\frac{1}{2}+\varepsilon}), \ n \to \infty.
\]
4. **Rate of convergence in the SLLN.** In this section we consider one-sample U-statistics (c=1). The following lemma will be required.

**Lemma 4.1 (Katz-Baum [7]).** Let \( \xi_1, \xi_2, \ldots \) be i.i.d. random variables. If \( r > 1 \), the following are equivalent:

\[
(4.1) \quad P[|\xi_1| > n] = o(n^{-r}) \quad \text{and} \quad E\xi_1 = \mu;
\]

\[
(4.2) \quad P[n^{-1} \sum_{i=1}^{n} \xi_i - \mu > \varepsilon] = o(n^{1-r}), \quad \text{for each} \quad \varepsilon > 0;
\]

\[
(4.3) \quad P[\sup_{k \geq n} |k^{-1} \sum_{i=1}^{k} \xi_i - \mu| > \varepsilon] = o(n^{1-r}), \quad \text{for each} \quad \varepsilon > 0.
\]

A corollary of this lemma is that if \( E|\xi_1|^r < \infty \), then (4.3) holds. This corollary is generalized to U-statistics in the following result.

**Theorem 4.1.** Let \( \{U_n\} \) be the sequence of U-statistics generated by a kernel \( h \) applied to a sequence of observations \( \{X_i\} \). Assume \( Eh^{2r} < \infty \) for a positive integer \( r \). Then, for any \( \varepsilon > 0 \),

\[
(4.4) \quad P[\sup_{k \geq n} |U_k - \theta| > \varepsilon] = o(n^{1-2r}), \quad n \to \infty.
\]

**Proof.** Let \( \varepsilon > 0 \) be given. Then

\[
(4.5) \quad P[\sup_{k \geq n} |U_k - \theta| > \varepsilon] \leq P[\sup_{k \geq n} |U_k - \hat{U}_k| > \frac{\varepsilon}{2}] + P[\sup_{k \geq n} |\hat{U}_k - \theta| > \frac{\varepsilon}{2}].
\]

Since \( \hat{U}_k - \theta = k^{-1} \sum_{j=1}^{k} h(X_j) \), the right-most term in (4.5) is \( o(n^{1-2r}) \) by Lemma 4.1. Now the sequence \( \{U_n - \hat{U}_n\} \) is a reverse martingale (see Geertsema [5] for discussion), so that by Loève [11], p. 391,
(4.6) \[ P[\sup_{k>n}|U_k - \hat{U}_k| > c] \leq C^{-2r}E(U_n - \hat{U}_n)^{2r} \]

for any constant C. Therefore, by Theorem 2.1, the first term of the RHS of (4.5) is \( O(n^{-2r}) \). Thus (4.4) follows. 

5. Analogous results for related von Mises statistics. We shall deal with the one-sample case. Let \( \{X_i\} \) be an i.i.d. sequence with df F. Let \( \theta = \theta(F) = \text{E}h(X_1, \ldots, X_m) \). For a sample of size n, the U-statistic for estimation of \( \theta \) is, recalling (1.1),

\[ U_n = \binom{n}{m}^{-1} \sum_{i_1, \ldots, i_m} h(X_{i_1}, \ldots, X_{i_m}) \]

where \( \sum_{c} \) denotes summation over all combinations \( \{i_1, \ldots, i_m\} \) from \( \{1, \ldots, n\} \). The associated von Mises statistic (see Hoeffding [8] for discussion) is

\[ V_n = n^{-m} \prod_{i_1=1}^{n} \prod_{i_m=1}^{n} h(X_{i_1}, \ldots, X_{i_m}) \]

The following result parallels Theorem 2.1 and shows that properties of \( V_n \) may be inferred from those of \( U_n \), just as in §3 and §4 properties of \( U_n \) were inferred from those of \( \hat{U}_n \).

**THEOREM 5.1.** Assume \( E|h(X_{i_1}, \ldots, X_{i_m})|^r \leq A < \infty \) for all \( 1 \leq i_1, \ldots, i_m \leq m \) and \( r \) a positive integer. Then

\[ E|U_n - V_n|^r = O(n^{-r}) \]
PROOF. Let \( n_{(m)} = n(n-1) \cdots (n-m+1) \). Clearly,

\[
(5.4) \quad n^m(U_n - V_n) = [n^m - n_{(m)}] U_n - \sum_{\# h(x_{i_1}, \ldots, x_{i_m})}
\]

where \( \sum_{\#} \) denotes summation over choices \( \{i_1, \ldots, i_m\} \) from \( \{1, 2, \ldots, n\} \)
where at least one equality \( i_a = i_b, \ a \neq b \), holds. As the number of
terms in \( \sum_{\#} \) is \( n^m - n_{(m)} \), we have by Minkowski's inequality that

\[
(5.5) \quad E|\sum_{\#} h(x_{i_1}, \ldots, x_{i_m})|^{r} \leq A \cdot [n^m - n_{(m)}]^{r}.
\]

Likewise, \( E|U_n|^{r} \leq A \). Thus

\[
(5.6) \quad n^{rm} E|U_n - V_n|^{r} \leq 2^{r} A [n^m - n_{(m)}]^{r}.
\]

But \( n^m - n_{(m)} = o(n^{m-1}) \), which yields (5.3). \( \square \)

With the use of Theorem 5.1 where earlier Theorem 2.1 was needed,
Theorem 3.1 may be extended to apply to \( V_n \) in place of \( U_n \). Theorem 4.1
can also be extended, but the order of magnitude in (4.4) is reduced
to \( O(n^{1+\delta-2r}) \).

6. Examples. Typical examples of U-statistics are given by the
sample variance, Fisher's k-statistics, Gini's mean difference, Kendall's
\( \tau \), the grade correlation coefficient, and Wilcoxon's one- and two-sample
statistics. (See [8] [10]). Let us briefly examine two of these
examples.
(i) **Sample variance.** Let \( X_1, \ldots, X_n \) be independent observations from a distribution \( F \). Assume \( E(X_1) = 0 \) and \( \theta(F) = \sigma^2 = E(X_1^2) > 0 \).

The sample variance is given by

\[
(6.1) \quad U = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2 = \left( \frac{n}{2}\right)^{-1} \sum_{i<j} h(X_i, X_j),
\]

where \( \bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \) and \( h(x_1, x_2) = \frac{1}{2} (x_1 - x_2)^2 \). Thus \( h^*_1(x) = \frac{1}{2} (x^2 - \sigma^2) \), \( H(x_1, x_2) = x_1 x_2 \), and \( \sigma^2(U) = n^{-1} E(X_1^2 - \sigma^2)^2 \). If \( E(X_1^{2r}) < \infty \) for an integer \( r \geq 3 \), then clearly the hypothesis of Theorem 3.1 is satisfied and the rate of convergence in (3.2) is \( o(n^{-r/(2r+1)}) \).

(ii) **Wilcoxon two-sample statistic.** Let \( \{X_1^{(1)}, \ldots, X_{n_1}^{(1)}\} \) and \( \{X_1^{(2)}, \ldots, X_{n_2}^{(2)}\} \) be independent observations from continuous distributions \( F_1 \) and \( F_2 \). Then, for \( \theta(F_1, F_2) = \int F_2 dF_1 \), an unbiased estimator is the Wilcoxon two-sample statistic, which may be written

\[
(6.2) \quad U = (n_1 n_2)^{-1} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} I_{[0,\infty)}(X_i^{(1)} - X_j^{(2)}),
\]

where \( I_A(\cdot) \) is the indicator of the set \( A \). In this case the kernel is bounded and (3.4) is satisfied. Hence, by Corollary 3.1, the rate of convergence in (3.2) is, for any \( \epsilon > 0 \), of order \( o(n^{-\frac{1}{2} + \epsilon}) \).

7. **Concluding remarks.** (i) Since the \( X_i^{(j)} \)'s enter into the definition of \( U \) only through the kernel \( h \), our results apply also to vector observations.

(ii) In similar fashion as we have dealt with the CLT and SLLN, rates of convergence could also be established for the law of the iterated logarithm for U-statistics, which has been given in [12].
(iii) Although our Theorem 4.1 was restricted to the 1-sample case, a generalization to the c-sample case proceeds as follows. Consider a sequence of c-vectors \( \{(n_1, \ldots, n_c)\} \) in which all components increase strictly from one vector to the next. For the associated sequence of c-sample U-statistics, the sequence of differences \( U - \hat{U} \) again forms a reverse martingale, as established by the smoothing properties of conditional expectations (see [11], p. 350, #4). Hence, in view of (2.1), and by application of Lemma 4.1 c times, the method of proof of Theorem 4.1 extends to the general case.

Acknowledgement. The authors thank Professor J. Sethuraman for stimulating conversations in the genesis of this work.
REFERENCES


