Approximation Theorems of Mathematical Statistics
Robert J. Serfling

CHAPTER 5

U-Statistics

From a purely mathematical standpoint, it is desirable and appropriate to view any given statistic as but a single member of some general class of statistics having certain important features in common. In such fashion, several interesting and useful collections of statistics have been formulated as generalizations of particular statistics that have arisen for consideration as special cases.

In this and the following four chapters, five such classes will be introduced. For each class, key features and propositions will be examined, with emphasis on results pertaining to consistency and asymptotic distribution theory. As a by-product, new ways of looking at some familiar statistics will be discovered.

The class of statistics to be considered in the present chapter was introduced in a fundamental paper by Hoeffding (1948). In part, the development rested upon a paper of Halmos (1946). The class arises as a generalization of the sample mean, that is, as a generalization of the notion of forming an *average*. Typically, although not without important exceptions, the members of the class are asymptotically normal statistics. They also have good consistency properties.

The so-called "*U-statistics*" are closely connected with a class of statistics introduced by von Mises (1947), which we shall examine in Chapter 6. Many statistics of interest fall within these two classes, and many other statistics may be approximated by a member of one of these classes.

The basic description of U-statistics is provided in Section 5.1. This includes relevant definitions, examples, connections with certain other statistics, martingale structure and other representations, and an optimality property of U-statistics among unbiased estimators. Section 5.2 deals with the moments, especially the variance, of U-statistics. An important tool in deriving the asymptotic theory of U-statistics, the "*projection*" of a U-statistic on the basic observations of the sample, is introduced in Section 5.3. Sections 5.4 and 5.5 treat, respectively, the almost sure behavior and asymptotic distribution theory of U-statistics. Section 5.6 provides some further probability bounds and limit theorems. Several complements are provided in Section 5.7, including a look at stochastic processes associated with a sequence of U-statistics, and an examination of the Wilcoxon one-sample statistic as a U-statistic in connection with the problem of confidence intervals for quantiles (recall **2.6.5**).

The method of "projection" introduced in Section 5.3 is of quite general scope and will be utilized again with other types of statistic in Chapters 8 and 9.

5.1 BASIC DESCRIPTION OF *U*-STATISTICS

Basic definitions and examples are given in **5.1.1**, and a class of closely related statistics is noted in **5.1.2**. These considerations apply to *one-sample U-statistics*. Generalization to *several samples* is given in **5.1.3**, and to *weighted versions* in **5.1.7**. An important *optimality* property of U-statistics in unbiased estimation is shown in **5.1.4**. The representation of a U-statistic as a *martingale* is provided in **5.1.5**, and as an *average of I.I.D. averages* in **5.1.6**.

Additional general discussion of U-statistics may be found in Fraser (1957), Section 4.2, and in Puri and Sen (1971), Section 3.3.

5.1.1 First Definitions and Examples

Let $X_1, X_2, \ldots$ be independent observations on a distribution F. (They may be vector-valued, but usually for simplicity we shall confine attention to the real-valued case.) Consider a "parametric function" $\theta = \theta(F)$ for which there is an unbiased estimator. That is, $\theta(F)$ may be represented as

$$\theta(F) = E_F\{h(X_1, \ldots, X_m)\} = \int \cdots \int h(x_1, \ldots, x_m) dF(x_1) \cdots dF(x_m),$$

for some function $h = h(x_1, \ldots, x_m)$, called a "kernel." Without loss of generality, we may assume that h is *symmetric*. For, if not, it may be replaced by the symmetric kernel

$$\frac{1}{m!} \sum_p h(x_{i_1}, \ldots, x_{i_m}),$$

where $\sum_p$ denotes summation over the $m!$ permutations $(i_1, \ldots, i_m)$ of $(1, \ldots, m)$.

For any kernel h, the corresponding *U-statistic* for estimation of θ on the basis of a sample $X_1, \ldots, X_n$ of size $n \geq m$ is obtained by averaging the kernel h symmetrically over the observations:

$$U_n = U(X_1, \ldots, X_n) = \frac{1}{\binom{n}{m}} \sum_c h(X_{i_1}, \ldots, X_{i_m}),$$

where $\sum_c$ denotes summation over the $\binom{n}{m}$ combinations of m distinct elements $\{i_1, \ldots, i_m\}$ from $\{1, \ldots, n\}$. Clearly, U_n is an *unbiased* estimate of θ.

Examples. (i) $\theta(F) = \text{mean of } F = \mu(F) = \int x\, dF(x)$. For the kernel $h(x) = x$, the corresponding U-statistic is

$$U(X_1, \ldots, X_n) = \frac{1}{n} \sum_{i=1}^{n} X_i = \bar{X},$$

the *sample mean.*

(ii) $\theta(F) = \mu^2(F) = [\int x\, dF(x)]^2$. For the kernel $h(x_1, x_2) = x_1 x_2$, the corresponding U-statistic is

$$U(X_1, \ldots, X_n) = \frac{2}{n(n-1)} \sum_{1 \le i < j \le n} X_i X_j.$$

(iii) $\theta(F) = \text{variance of } F = \sigma^2(F) = \int (x - \mu)^2\, dF(x)$. For the kernel

$$h(x_1, x_2) = \frac{x_1^2 + x_2^2 - 2x_1 x_2}{2} = \frac{1}{2}(x_1 - x_2)^2.$$

the corresponding U-statistic is

$$\begin{aligned} U(X_1, \ldots, X_n) &= \frac{2}{n(n-1)} \sum_{1 \le i < j \le n} h(X_i, X_j) \\ &= \frac{1}{n-1} \left(\sum_{i=1}^{n} X_i^2 - n\bar{X}^2 \right) \\ &= s^2, \end{aligned}$$

the *sample variance.*

(iv) $\theta(F) = F(t_0) = \int_{-\infty}^{t_0} dF(x) = P_F(X_1 \le t_0)$. For the kernel $h(x) = I(x \le t_0)$, the corresponding U-statistic is

$$U(X_1, \ldots, X_n) = \frac{1}{n} \sum_{i=1}^{n} I(X_i \le t_0) = F_n(t_0),$$

where F_n denotes the sample distribution function.

(v) $\theta(F) = \alpha_k(F) = \int x^k\, dF(x) = k\text{th moment of } F$. For the kernel $h(x) = x^k$, the corresponding U-statistic is

$$U(X_1, \ldots, X_n) = \frac{1}{n} \sum_{i=1}^{n} X_i^k = a_k,$$

the *sample* kth *moment.*

(vi) $\theta(F) = E_F|X_1 - X_2|$, a measure of concentration. For the kernel $h(x_1, x_2) = |x_1 - x_2|$, the corresponding U-statistic is

$$U(X_1, \ldots, X_n) = \frac{2}{n(n-1)} \sum_{1 \le i < j \le n} |X_i - X_j|,$$

the statistic known as "Gini's mean difference."

(vii) Fisher's k-statistics for estimation of cumulants are U-statistics (see Wilks (1962), p. 200).

(viii) $\theta(F) = E_F \gamma(X_1) = \int \gamma(x) dF(x)$; $U_n = n^{-1} \sum_1^n \gamma(X_i)$.

(ix) *The Wilcoxon one-sample statistic.* For estimation of $\theta(F) = P_F(X_1 + X_2 \le 0)$, a kernel is given by $h(x_1, x_2) = I(x_1 + x_2 \le 0)$ and the corresponding U-statistic is

$$U(X_1, \ldots, X_n) = \frac{2}{n(n-1)} \sum_{1 \le i < j \le n} I(X_i + X_j \le 0).$$

(x) $\theta(F) = \iint [F(x, y) - F(x, \infty)F(\infty, y)]^2 \, dF(x, y)$, a measure of dependence for a bivariate distribution F. Putting

$$\psi(z_1, z_2, z_3) = I(z_2 \le z_1) - I(z_3 \le z_1)$$

and

$$\begin{aligned} h((x_1, y_1), \ldots, (x_5, y_5)) = {} & \tfrac{1}{4}\psi(x_1, x_2, x_3)\psi(x_1, x_4, x_5) \\ & \times \psi(y_1, y_2, y_3)\psi(y_1, y_4, y_5), \end{aligned}$$

we have $E_F\{h\} = \theta(F)$, and the corresponding U-statistic is

$$U_n = \frac{5!}{n(n-1)(n-2)(n-3)(n-4)} \sum_c h((X_{i_1}, Y_{i_1}), \ldots, (X_{i_5}, Y_{i_5})). \quad ■$$

5.1.2 Some Closely Related Statistics: V-Statistics

Corresponding to a U-statistic

$$U_n = \frac{1}{\binom{n}{m}} \sum_c h(X_{i_1}, \ldots, X_{i_m})$$

for estimation of $\theta(F) = E_F\{h\}$, the *associated von Mises statistic* is

$$\begin{aligned} V_n &= \frac{1}{n^m} \sum_{i_1=1}^{n} \cdots \sum_{i_m=1}^{n} h(X_{i_1}, \ldots, X_{i_m}) \\ &= \theta(F_n), \end{aligned}$$

where F_n denotes the sample distribution function. Let us term this statistic, in connection with a kernel h, the associated *V-statistic*. The connection between U_n and V_n will be examined closely in 5.7.3 and pursued further in Chapter 6.

Certain other statistics, too, may be treated as approximately a U-statistic, the gap being bridged via Slutsky's Theorem and the like. Thus the domain of application of the asymptotic theory of U-statistics is considerably wider than the context of unbiased estimation.

5.1.3 Generalized U-Statistics

The extension to the case of several samples is straightforward. Consider k independent collections of independent observations $\{X_1^{(1)}, X_2^{(1)}, \ldots\}, \ldots, \{X_1^{(k)}, X_2^{(k)}, \ldots\}$ taken from distributions $F^{(1)}, \ldots, F^{(k)}$, respectively. Let $\theta = \theta(F^{(1)}, \ldots, F^{(k)})$ denote a parametric function for which there is an unbiased estimator. That is,

$$\theta = E\{h(X_1^{(1)}, \ldots, X_{m_1}^{(1)}; \ldots; X_1^{(k)}, \ldots, X_{m_k}^{(k)})\},$$

where h is assumed, without loss of generality, to be symmetric within each of its k blocks of arguments. Corresponding to the "kernel" h and assuming $n_1 \geq m_1, \ldots, n_k \geq m_k$, the U-statistic for estimation of θ is defined as

$$U_n = \frac{1}{\prod_{j=1}^{k} \binom{n_j}{m_j}} \sum_c h(X_{i_{11}}^{(1)}, \ldots, X_{i_{1m_1}}^{(1)}; \ldots; X_{i_{k1}}^{(k)}, \ldots, X_{i_{km_k}}^{(k)}).$$

Here $\{i_{j1}, \ldots, i_{jm_j}\}$ denotes a set of m_j distinct elements of the set $\{1, 2, \ldots, n_j\}$, $1 \leq j \leq k$, and $\sum_c$ denotes summation over all such combinations.

The extension of Hoeffding's treatment of one-sample U-statistics to the k-sample case is due to Lehmann (1951) and Dwass (1956). Many statistics of interest are of the k-sample U-statistic type.

Example. *The Wilcoxon 2-sample statistic.* Let $\{X_1, \ldots, X_{n_1}\}$ and $\{Y_1, \ldots, Y_{n_2}\}$ be independent observations from continuous distributions F and G, respectively. Then, for

$$\theta(F, G) = \int F\, dG = P(X \leq Y),$$

an unbiased estimator is

$$U = \frac{1}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} I(X_i \leq Y_j). \quad \blacksquare$$

5.1.4 An Optimality Property of U-Statistics

A U-statistic may be represented as the result of conditioning the kernel on the order statistic. That is, for a kernel $h(x_1, \ldots, x_m)$ and a sample $X_1, \ldots, X_n$, $n \geq m$, the corresponding U-statistic may be expressed as

$$U_n = E\{h(X_1, \ldots, X_m) | \mathbf{X}_{(n)}\},$$

where $\mathbf{X}_{(n)}$ denotes the order statistic $(X_{n1}, \ldots, X_{nn})$.

One implication of this representation is that any statistic $S = S(X_1, \ldots, X_n)$ for unbiased estimation of $\theta = \theta(F)$ may be "improved" by the corresponding U-statistic. That is, we have

Theorem. *Let* $S = S(X_1, \ldots, X_n)$ *be an unbiased estimator of* $\theta(F)$ *based on a sample* $X_1, \ldots, X_n$ *from the distribution* F. *Then the corresponding* U*-statistic is also unbiased and*

$$\mathrm{Var}_F\{U\} \leq \mathrm{Var}_F\{S\},$$

with equality if and only if $P_F(U = S) = 1$.

PROOF. The "kernel" associated with S is

$$\frac{1}{n!} \sum_p S(x_{i_1}, \ldots, x_{i_n}),$$

which in this case $(m = n)$ is the U-statistic associated with itself. That is, the U-statistic associated with S may be expressed as

$$U = E\{S | \mathbf{X}_{(n)}\}.$$

Therefore,

$$E_F\{U^2\} = E_F\{E^2\{S | \mathbf{X}_{(n)}\}\} \leq E_F\{E\{S^2 | \mathbf{X}_{(n)}\}\} = E_F\{S^2\},$$

with equality if and only if $E\{S | \mathbf{X}_{(n)}\}$ is degenerate and equals S with P_F-probability 1. Since $E_F\{U\} = E_F\{S\}$, the proof is complete. ∎

Since the order statistic $\mathbf{X}_{(n)}$ is sufficient (in the usual technical sense) for any family $\mathcal{F}$ of distributions containing F, the U-statistic is the result of conditioning on a sufficient statistic. Thus the preceding result is simply a special case of the Rao–Blackwell theorem (see Rao (1973), §5a.2). In the case that $\mathcal{F}$ is rich enough that $\mathbf{X}_{(n)}$ is *complete* sufficient (e.g., if $\mathcal{F}$ contains all absolutely continuous F), then U_n is the minimum variance *unbiased* estimator of θ.

5.1.5 Martingale Structure of U-Statistics

Some important properties of U-statistics (see **5.2.1**, **5.3.3**, **5.3.4**, Section 5.4) flow from their martingale structure and a related representation.

Definitions. Consider a probability space $(\Omega, \mathscr{A}, P)$, a sequence of random variables $\{Y_n\}$, and a sequence of σ-fields $\{\mathscr{F}_n\}$ contained in $\mathscr{A}$, such that Y_n is $\mathscr{F}_n$-measurable and $E|Y_n| < \infty$. Then the sequence $\{Y_n, \mathscr{F}_n\}$ is called a *forward martingale* if

(a) $\mathscr{F}_1 \subset \mathscr{F}_2 \subset \ldots,$

(b) $E\{Y_{n+1} | \mathscr{F}_n\} = Y_n$ *wp*1, all n,

and a *reverse martingale* if

(a′) $\mathscr{F}_1 \supset \mathscr{F}_2 \supset \ldots,$

(b′) $E\{Y_n | \mathscr{F}_{n+1}\} = Y_{n+1}$ *wp*1, all n. ■

The following lemmas, due to Hoeffding (1961) and Berk (1966), respectively, provide both forward and reverse martingale characterizations for U-statistics. For the first lemma, some preliminary notation is needed. Consider a symmetric kernel $h(x_1, \ldots, x_m)$ satisfying $E_F|h(X_1, \ldots, X_m)| < \infty$. We define the associated functions

$$h_c(x_1, \ldots, x_c) = E_F\{h(x_1, \ldots, x_c, X_{c+1}, \ldots, X_m)\}$$

for each $c = 1, \ldots, m - 1$ and put $h_m \equiv h$. Since

$$\int_A h_c(x_1, \ldots, x_c) dF(x_1) \cdots dF(x_c) = \int_{A \times R^{m-c}} h(x_1, \ldots, x_m) dF(x_1) \cdots dF(x_m)$$

for every Borel set A in R^c, h_c is (a version of) the conditional expectation of $h(X_1, \ldots, X_m)$ given $X_1, \ldots, X_c$:

$$h_c(x_1, \ldots, x_c) = E_F\{h(X_1, \ldots, X_m) | X_1 = x_1, \ldots, X_c = x_c\}.$$

Further, note that for $1 \le c \le m - 1$

$$h_c(x_1, \ldots, x_c) = E_F\{h_{c+1}(x_1, \ldots, x_c, X_{c+1})\}.$$

It is convenient to center at expectations, by defining

$$\theta(F) = E_F\{h(X_1, \ldots, X_m)\},$$
$$\tilde{h} = h - \theta(F),$$

and

$$\tilde{h}_c = h_c - \theta(F), \qquad 1 \le c \le m.$$

We now define

$$g_1(x_1) = \tilde{h}_1(x_1),$$
$$g_2(x_1, x_2) = \tilde{h}_2(x_1, x_2) - g_1(x_1) - g_1(x_2),$$
$$g_3(x_1, x_2, x_3) = \tilde{h}_3(x_1, x_2, x_3) - \sum_{i=1}^{3} g_1(x_i) - \sum_{1 \le i < j \le 3} g_2(x_i, x_j),$$
$$\ldots,$$
$$(*) \quad g_m(x_1, \ldots, x_m) = \tilde{h}(x_1, \ldots, x_m) - \sum_{i=1}^{m} g_1(x_i) - \sum_{1 \le i_1 < i_2 \le m} g_2(x_{i_1}, x_{i_2})$$
$$- \cdots - \sum_{1 \le i_1 < \cdots < i_{m-1} \le m} g_{m-1}(x_{i_1}, \ldots, x_{i_{m-1}}).$$

Clearly, the g_c's are symmetric in their arguments. Also, it is readily seen (check) that

$$E_F\{g_1(X_1)\} = 0,$$
$$E_F\{g_2(x_1, X_2)\} = 0,$$
$$\ldots,$$
$$E_F\{g_m(x_1, \ldots, x_{m-1}, X_m)\} = 0.$$

Now consider a sample $X_1, \ldots, X_n (n \ge m)$ and note that the U-statistic U_n corresponding to the kernel h satisfies

$$U_n - \theta(F) = \binom{n}{m}^{-1} S_n,$$

where

$$(1) \qquad S_n = \sum_{1 \le i_1 < \cdots < i_m \le n} \tilde{h}(X_{i_1}, \ldots, X_{i_m}).$$

Finally, for $1 \le c \le m$, put

$$S_{cn} = \sum_{1 \le i_1 < \cdots < i_c \le n} g_c(X_{i_1}, \ldots, X_{i_c}).$$

Hoeffding's lemma, which we now state, asserts a martingale property for the sequence $\{S_{cn}\}_{n \ge c}$ for each $c = 1, \ldots, m$, and gives a representation for U_n in terms of $S_{1n}, \ldots, S_{mn}$.

Lemma A (Hoeffding). *Let* $h = h(x_1, \ldots, x_m)$ *be a symmetric kernel for* $\theta = \theta(F)$, *with* $E_F|h| < \infty$. *Then*

$$(2) \qquad U_n - \theta = \sum_{c=1}^{m} \binom{m}{c} \binom{n}{c}^{-1} S_{cn}.$$

Further, for each $c = 1, \ldots, m$,

(3) $$E_F\{S_{cn} | X_1, \ldots, X_k\} = S_{ck}, \qquad c \le k \le n.$$

Thus, with $\mathscr{F}_k = \sigma\{X_1, \ldots, X_k\}$, *the sequence* $\{S_{cn}, \mathscr{F}_n\}_{n \ge c}$ *is a forward martingale.*

PROOF. The definition of g_m in (*) expresses $\tilde{h}$ in terms of $g_1, \ldots, g_m$. Substitution in (1) then yields

$$S_n = S_{mn} + \sum_{c=1}^{m-1} \sum_{1 \le i_1 < \cdots < i_m \le n} \sum_{1 \le j_1 < \cdots < j_c \le m} g_c(X_{i_{j_1}}, \ldots, X_{i_{j_c}}).$$

On the right-hand side, the term for $c = 1$ may be written

$$\sum_{1 \le i_1 < \cdots < i_m \le n} \sum_{j=1}^{m} g(X_{i_j}).$$

In this sum, each $g(X_i)$, $1 \le i \le n$, is represented the same number of times. Since the sum contains $\binom{n}{m} \cdot m$ terms, each $g(X_i)$ appears $n^{-1}\binom{n}{m}m$ times. That is, the sum $S_{1n} = \sum_1^n g(X_i)$ appears $\binom{n}{1}^{-1}\binom{n}{m}\binom{m}{1}$ times. In this fashion we obtain

$$S_n = \sum_{c=1}^{m} \binom{n}{c}^{-1} \binom{n}{m} \binom{m}{c} S_{cn},$$

which yields (2). To see the martingale property (3), observe that

$$E_F\{g_c(X_{i_1}, \ldots, X_{i_c}) | X_1, \ldots, X_k\} = 0$$

if one of $i_1, \ldots, i_c$ is not contained in $\{1, \ldots, k\}$. For example, if $i_1 \notin \{1, \ldots, k\}$, then

$$\begin{aligned}
&E_F\{g_c(X_{i_1}, \ldots, X_{i_c}) | X_1, \ldots, X_k\} \\
&\quad = E_F\{E_F[g_c(X_{i_1}, \ldots, X_{i_c}) | X_1, \ldots, X_k, X_{i_2}, \ldots, X_{i_c}] | X_1, \ldots, X_k\} \\
&\quad = E_F\{E_F[g_c(X_{i_1}, \ldots, X_{i_c}) | X_{i_2}, \ldots, X_{i_c}] | X_1, \ldots, X_k\} \\
&\quad = E_F\{0 | X_1, \ldots, X_k\} = 0.
\end{aligned}$$

Thus

$$E_F\{S_{cn} | X_1, \ldots, X_k\} = \sum_{1 \le i_1 < \cdots < i_c \le k} g_c(X_{i_1}, \ldots, X_{i_c}) = S_{ck}. \quad \blacksquare$$

Example A. For the case $m = 1$ and $h(x) = x$, Lemma A states simply that

$$U_n - \theta = \frac{1}{n} \sum_{i=1}^{n} (X_i - \theta)$$

and that $\{\sum_1^n (X_i - \theta), \sigma(X_1, \ldots, X_n)\}$ is a forward martingale. $\blacksquare$

The other martingale representation for U_n is much simpler:

Lemma B (Berk). *Let* $h = h(x_1, \ldots, x_m)$ *be a symmetric kernel for* $\theta = \theta(F)$, *with* $E_F|h| < \infty$. *Then, with* $\mathscr{F}_n = \sigma\{X_{(n)}, X_{n+1}, X_{n+2}, \ldots\}$, *the sequence* $\{U_n, \mathscr{F}_n\}_{n \geq m}$ *is a reverse martingale.*

PROOF. (exercise) Apply the representation

$$U_n = E\{h(X_1, \ldots, X_m)|X_{(n)}\}$$

considered in 5.1.4. ∎

Example B (*continuation*). For the case $m = 1$ and $h(x) = x$, Lemma B asserts that $\bar{X}$ is a reverse martingale. ∎

5.1.6 Representation of a *U*-Statistic as an Average of (Dependent) Averages of I.I.D. Random Variables

Consider a symmetric kernel $h(x_1, \ldots, x_m)$ and a sample $X_1, \ldots, X_n$ of size $n \geq m$. Define $k = [n/m]$, the greatest integer $\leq n/m$, and define

$$W(x_1, \ldots, x_n) = \frac{h(x_1, \ldots, x_m) + h(x_{m+1}, \ldots, x_{2m}) + \cdots + h(x_{km-m+1}, \ldots, x_{km})}{k}.$$

Letting $\sum_p$ denote summation over all $n!$ permutations $(i_1, \ldots, i_n)$ of $(1, \ldots, n)$ and $\sum_c$ denote summation over all $\binom{n}{m}$ combinations $\{i_1, \ldots, i_m\}$ from $\{1, \ldots, n\}$, we have

$$k \sum_p W(x_{i_1}, \ldots, x_{i_n}) = km!(n-m)! \sum_c h(x_{i_1}, \ldots, x_{i_m}),$$

and thus

$$\sum_p W(X_{i_1}, \ldots, X_{i_n}) = m!(n-m)!\binom{n}{m} U_n,$$

or

$$U_n = \frac{1}{n!} \sum_p W(X_{i_1}, \ldots, X_{i_n}).$$

This expresses U_n as an average of $n!$ terms, each of which is itself an average of k I.I.D. random variables. This type of representation was introduced and utilized by Hoeffding (1963). We shall apply it in Section 5.6.

5.1.7 Weighted *U*-Statistics

Consider now an arbitrary kernel $h(x_1, \ldots, x_m)$, not necessarily symmetric, to be applied as usual to observations $X_1, \ldots, X_n$ taken m at a time. Suppose also that each term $h(X_{i_1}, \ldots, X_{i_m})$ becomes weighted by a factor $w(i_1, \ldots, i_m)$

depending only on the indices $i_1, \ldots, i_m$. In this case the U-statistic sum takes the more general form

$$T_n = \sum_c w(i_1, \ldots, i_m) h(X_{i_1}, \ldots, X_{i_m}).$$

In the case that h is symmetric and the weights $w(i_1, \ldots, i_m)$ take only 0 or 1 as values, a statistic of this form represents an "incomplete" or "reduced" U-statistic sum, designed to be computationally simpler than the usual sum. This is based on the notion that, on account of the dependence among the $\binom{n}{m}$ terms of the complete sum, it should be possible to use less terms without losing much information. Such statistics have been investigated by Blom (1976) and Brown and Kildea (1978).

Certain "permutation statistics" arising in nonparametric inference are asymptotically equivalent to statistics of the above form, with weights not necessarily 0- and 1-valued. For these and other applications, the statistics of form T_n with h symmetric and $m = 2$ have been studied by Shapiro and Hubert (1979).

Finally, certain "weighted rank statistics" for simple linear regression take the form T_n. Following Sievers (1978), consider the simple linear regression model

$$y_i = \alpha + \beta x_i + e_i, \qquad 1 \le i \le n,$$

where α and β are unknown parameters, $x_1, \ldots, x_n$ are known regression scores, and $e_1, \ldots, e_n$ are I.I.D. with distribution F. Sievers considers inferences for β based on the random variables

$$T_\beta = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} a_{ij} \phi(Y_i - \alpha - \beta x_i, Y_j - \alpha - \beta x_j),$$

where $\phi(u, v) = I(u \le v)$, the weights $a_{ij} \ge 0$ are arbitrary, and it is assumed that $x_1 \le \cdots \le x_n$ with at least one strict inequality. For example, a test of $H_0: \beta = \beta_0$ against $H_1: \beta > \beta_0$ may be based on the statistic T_{β_0}. Under the null hypothesis, the distribution of T_{β_0} is the same as that of T_0 when $\beta = 0$. That is, it is the same as

$$\sum_{i=1}^{n} \sum_{j=i+1}^{n} a_{ij} \phi(e_i, e_j),$$

which is of the form T_n above. The a_{ij}'s here are selected to achieve high asymptotic efficiency. Recommended weights are $a_{ij} = x_j - x_i$.

5.2 THE VARIANCE AND OTHER MOMENTS OF A U-STATISTIC

Exact formulas for the variance of a U-statistic are derived in **5.2.1**. The higher moments are difficult to deal with exactly, but useful bounds are obtained in **5.2.2**.

5.2.1 The Variance of a U-Statistic

Consider a symmetric kernel $h(x_1, \ldots, x_m)$ satisfying

$$E_F\{h^2(X_1, \ldots, X_m)\} < \infty.$$

We shall again make use of the functions h_c and $\tilde{h}_c$ introduced in **5.1.5**. Recall that $h_m = h$ and, for $1 \le c \le m - 1$,

$$h_c(x_1, \ldots, x_c) = E_F\{h(x_1, \ldots, x_c, X_{c+1}, \ldots, X_m)\},$$

that $\tilde{h} = h - \theta$, $\tilde{h}_c = h_c - \theta (1 \le c \le m)$, where

$$\theta = \theta(F) = E_F\{h(X_1, \ldots, X_m)\},$$

and that, for $1 \le c \le m - 1$,

$$h_c(x_1, \ldots, x_c) = E_F\{h_{c+1}(x_1, \ldots, x_c, X_{c+1})\}.$$

Note that

$$E_F\tilde{h}_c(X_1, \ldots, X_c) = 0, \qquad 1 \le c \le m.$$

Define $\zeta_0 = 0$ and, for $1 \le c \le m$,

$$\zeta_c = \mathrm{Var}_F\{h_c(X_1, \ldots, X_c)\} = E_F\{\tilde{h}_c^2(X_1, \ldots, X_c)\}.$$

We have (Problem 5.P.3(i))

$$0 = \zeta_0 \le \zeta_1 \le \cdots \le \zeta_m = \mathrm{Var}_F\{h\} < \infty.$$

Before proceeding further, let us exemplify these definitions. Note from the following example that the functions h_c and $\tilde{h}_c$ depend on F for $c \le m - 1$. The role of these functions is technical.

Example A. $\theta(F) = \sigma^2(F)$. Writing $\mu = \mu(F)$, $\sigma^2 = \sigma^2(F)$ and $\mu_4 = \mu_4(F)$, we have

$$\begin{aligned}
h(x_1, x_2) &= \tfrac{1}{2}(x_1^2 + x_2^2 - 2x_1x_2) = \tfrac{1}{2}(x_1 - x_2)^2, \\
\tilde{h}(x_1, x_2) &= h(x_1, x_2) - \sigma^2, \\
h_1(x) &= \tfrac{1}{2}(x^2 + \sigma^2 + \mu^2 - 2x\mu), \\
\tilde{h}_1(x) &= \tfrac{1}{2}(x^2 - \sigma^2 + \mu^2 - 2x\mu) = \tfrac{1}{2}[(x - \mu)^2 - \sigma^2], \\
E\{h^2\} &= \tfrac{1}{4}E\{[(X_1 - \mu) - (X_2 - \mu)]^4\} \\
&= \frac{1}{4}\sum_{j=0}^{4}\binom{4}{j}(-1)^{4-j}E\{(X_1 - \mu)^j\}E\{(X_2 - \mu)^{4-j}\} \\
&= \tfrac{1}{4}(2\mu_4 + 6\sigma^4), \\
\zeta_2 &= E\{h^2\} - \sigma^4 = \tfrac{1}{2}(\mu_4 + \sigma^4), \\
\zeta_1 &= E\{\tilde{h}_1^2\} = \tfrac{1}{4}\,\mathrm{Var}_F\{(X_1 - \mu)^2\} = \tfrac{1}{4}(\mu_4 - \sigma^4). \quad \blacksquare
\end{aligned}$$

Next let us consider two sets $\{a_1, \ldots, a_m\}$ and $\{b_1, \ldots, b_m\}$ of m distinct integers from $\{1, \ldots, n\}$ and let c be the number of integers common to the two sets. It follows (Problem 5.P.4) by symmetry of $\tilde{h}$ and by independence of $\{X_1, \ldots, X_n\}$ that

$$E_F\{\tilde{h}(X_{a_1}, \ldots, X_{a_m})\tilde{h}(X_{b_1}, \ldots, X_{b_m})\} = \zeta_c.$$

Note also that the number of distinct choices for two such sets having exactly c elements in common is $\binom{n}{m}\binom{m}{c}\binom{n-m}{m-c}$.

With these preliminaries completed, we may now obtain the variance of a U-statistic. Writing

$$U_n - \theta = \binom{n}{m}^{-1} \sum_c \tilde{h}(X_{i_1}, \ldots, X_{i_m}),$$

we have

$$\begin{aligned} \mathrm{Var}_F\{U_n\} &= E_F\{(U_n - \theta)^2\} \\ &= \binom{n}{m}^{-2} \sum_c \sum_c E_F\{\tilde{h}(X_{a_1}, \ldots, X_{a_m})\tilde{h}(X_{b_1}, \ldots, X_{b_m})\} \\ &= \binom{n}{m}^{-2} \sum_{c=0}^{n} \binom{n}{m}\binom{m}{c}\binom{n-m}{m-c}\zeta_c. \end{aligned}$$

This result and other useful relations from Hoeffding (1948) may be stated as follows.

Lemma A. *The variance of* U_n *is given by*

$$(*) \qquad \mathrm{Var}_F\{U_n\} = \binom{n}{m}^{-1} \sum_{c=1}^{m} \binom{m}{c}\binom{n-m}{m-c}\zeta_c$$

and satisfies

(i) $\dfrac{m^2}{n}\zeta_1 \le \mathrm{Var}_F\{U_n\} \le \dfrac{m}{n}\zeta_m;$

(ii) $(n+1)\mathrm{Var}_F\{U_{n+1}\} \le n\,\mathrm{Var}_F\{U_n\};$

(iii) $\mathrm{Var}_F\{U_n\} = \dfrac{m^2\zeta_1}{n} + O(n^{-2}), \qquad n \to \infty.$

Note that (*) is a fixed sample size formula. Derive (i), (ii), and (iii) from (*) as an exercise.

Example B (*Continuation*).

$$\begin{aligned}\mathrm{Var}_F\{s^2\} &= \binom{n}{2}^{-1}[2(n-2)\zeta_1 + \zeta_2] \\ &= \frac{4\zeta_1}{n} + \frac{2\zeta_2}{n(n-1)} - \frac{4\zeta_1}{n(n-1)} \\ &= \frac{\mu_4 - \sigma^4}{n} + \frac{2\sigma^4}{n(n-1)} \\ &= \frac{\mu_4 - \sigma^4}{n} + O(n^{-2}). \quad \blacksquare\end{aligned}$$

The extension of (*) to the case of a *generalized* U-statistic is straightforward (Problem 5.P.6).

An alternative formula for $\mathrm{Var}_F\{U_n\}$ is obtained by using, instead of h_c and $\tilde{h}_c$, the functions g_c introduced in **5.1.5** and the representation given by Lemma 5.1.5A.

Consider a set $\{i_1, \ldots, i_c\}$ of c distinct integers from $\{1, \ldots, n\}$ and a set $\{j_1, \ldots, j_d\}$ of d distinct integers from $\{1, \ldots, n\}$, where $1 \le c, d \le m$. It is evident from the proof of Lemma 5.1.5A that if one of $\{i_1, \ldots, i_c\}$ is not contained in $\{j_1, \ldots, j_d\}$, then

$$E_F\{g_c(X_{i_1}, \ldots, X_{i_c}) | X_{j_1}, \ldots, X_{j_d}\} = 0.$$

From this it follows that $E_F\{g_c(X_{i_1}, \ldots, X_{i_c})g_d(X_{j_1}, \ldots, X_{j_d})\} = 0$ *unless* $c = d$ and $\{i_1, \ldots, i_c\} = \{j_1, \ldots, j_d\}$. Therefore, for the functions

$$S_{cn} = \sum_{1 \le i_1 < \cdots < i_c \le n} g_c(X_{i_1}, \ldots, X_{i_c}),$$

we have

$$E\{S_c S_d\} = \begin{cases} \binom{n}{c} E\{g_c^2\}, & c = d, \\ 0, & c \ne d. \end{cases}$$

Hence

Lemma B. *The variance of* U_n *is given by*

$$(**) \qquad \mathrm{Var}_F\{U_n\} = \sum_{c=1}^{m} \binom{m}{c}^{-2} \binom{n}{c}^{-1} E_F\{g_c^2\}.$$

The result (iii) of Lemma A follows slightly more readily from (**) than from (*).

Example C (Continuation). We have

$$g_1(x) = \tilde{h}_1(x) = \tfrac{1}{2}[(x - \mu)^2 - \sigma^2],$$
$$g_2(x_1, x_2) = \tilde{h}_2(x_1, x_2) - \tilde{h}_1(x_1) - \tilde{h}_1(x_2) = \mu^2 + x_1\mu + x_2\mu - x_1x_2,$$
$$E\{g_1^2\} = \zeta_1 = \tfrac{1}{4}(\mu_4 - \sigma^4), \text{ as before,}$$
$$E\{g_2^2\} = \sigma^4,$$

and thus

$$\mathrm{Var}_F\{s^2\} = \frac{4}{n}E\{g_1^2\} + \frac{2}{n(n-1)}E\{g_2^2\}$$
$$= \frac{\mu_4 - \sigma^4}{n} + \frac{2\sigma^4}{n(n-1)}, \text{ as before. } \blacksquare$$

The rate of convergence of $\mathrm{Var}\{U_n\}$ to 0 depends upon the least c for which ζ_c is nonvanishing. From either Lemma A or Lemma B, we obtain

Corollary. *Let* $c \geq 1$ *and suppose that* $\zeta_0 = \cdots = \zeta_{c-1} = 0 < \zeta_c$. *Then*

$$E(U_n - \theta)^2 = O(n^{-c}), \qquad n \to \infty.$$

Note that the condition $\zeta_d = 0$, $d < c$, is equivalent to the condition $E\{\tilde{h}_d^2\} = 0$, $d < c$, and also to the condition $E\{g_d^2\} = 0$, $d < c$.

5.2.2 Other Moments of U-Statistics

Exact generalizations of Lemmas 5.2.1 A, B for moments of order other than 2 are difficult to work out and complicated even to state. However, for most purposes, suitable bounds suffice. Fortunately, these are rather easily obtained.

Lemma A. *Let* r *be a real number* ≥ 2. *Suppose that* $E_F|h|^r < \infty$. *Then*

$$(*) \qquad E|U_n - \theta|^r = O(n^{-(1/2)r}), \qquad n \to \infty.$$

PROOF. We utilize the representation of U_n as an average of averages of I.I.D.'s (5.1.6),

$$U_n - \theta = (n!)^{-1} \sum_p \tilde{W}(X_{i_1}, \ldots, X_{i_n}),$$

where $\tilde{W}(X_{i_1}, \ldots, X_{i_n}) = W(X_{i_1}, \ldots, X_{i_n}) - \theta$ is an average of $k = [n/m]$ I.I.D. terms of the form $\tilde{h}(X_{i_1}, \ldots, X_{i_m})$. By Minkowski's inequality,

$$E|U_n - \theta|^r \leq E|\tilde{W}(X_1, \ldots, X_n)|^r.$$

By Lemma 2.2.2B, $E|\tilde{W}(X_1, \ldots, X_n)|^r = O(k^{-(1/2)r})$, $k \to \infty$. $\blacksquare$

Lemma B. Let $c \geq 1$ *and suppose that* $\zeta_0 = \cdots = \zeta_{c-1} = 0 < \zeta_c$. *Let* r *be an integer* ≥ 2 *and suppose that* $E_F|h|^r < \infty$. *Then*

$$(**) \qquad E(U_n - \theta)^r = O(n^{-[(1/2)(rc+1)]}), \qquad n \to \infty,$$

where $[\cdot]$ *denotes integer part.*

PROOF. Write

$$(1) \qquad E(U_n - \theta)^r = \binom{n}{m}^{-r} \sum E\left\{\prod_{j=1}^{r} \tilde{h}(X_{i_{j1}}, \ldots, X_{i_{jm}})\right\},$$

where "j" identifies the factor within the product, and $\sum$ denotes summation over all $\binom{n}{m}^r$ of the indicated terms. Consider a typical term. For the jth factor, let p_j denote the number of indices repeated in other factors. If $p_j \leq c - 1$, then (justify)

$$E\{\tilde{h}(X_{i_{j1}}, \ldots, X_{i_{jm}}) | \text{the } p_j \text{ repeated } X_{i_{jk}}\text{'s}\} = 0.$$

Thus a term in (1) can have nonzero expectation only if each factor in the product contains at least c indices which appear in other factors in the product. Denote by q the number of distinct elements among the repeated indices in the r factors of a given product. Then (justify)

$$(2) \qquad 2q \leq \sum_{j=1}^{r} p_j.$$

For fixed values of $q, p_1, \ldots, p_r$, the number of ways to select the indices in the r factors of a product is of order

$$(3) \qquad O(n^{q+(m-p_1)+\cdots+(m-p_r)}),$$

where the implicit constants depend upon r and m, but not upon n. Now, by (2), $q \leq [\frac{1}{2}\sum_{j=1}^r p_j]$. Thus

$$q + \sum_{j=1}^{r}(m - p_j) \leq rm + \left[\frac{1}{2}\sum_{j=1}^{m} p_j\right] - \sum_{j=1}^{m} p_j = rm - \left[\frac{1}{2}\left(\sum_{j=1}^{r} p_j + 1\right)\right],$$

since (verify), for any integer x, $x - [\frac{1}{2}x] = [\frac{1}{2}(x + 1)]$. Confining attention to the case that $p_1 \geq c, \ldots, p_r \geq c$, we have $\sum_{j=1}^r p_j \geq rc$, so that

$$(4) \qquad q + \sum_{j=1}^{r}(m - p_j) \leq rm - [\tfrac{1}{2}(rc + 1)].$$

The number of ways to select the values $q, p_1, \ldots, p_r$ depends on r and m, but not upon n. Thus, by (3) and (4), it follows that the number of terms in the sum in (1) for which the expectation is possibly nonzero is of order

$$O(n^{rm-[(1/2)(rc+1)]}), \qquad n \to \infty.$$

Since $\binom{n}{m}^{-1} = O(n^{-m})$, the relation (*) is proved. ∎

Remarks. (i) Lemma A generalizes to rth order the relation $E(U_n - \theta)^2 = O(n^{-1})$ expressed in Lemma 5.2.1A.

(ii) Lemma B generalizes to rth order the relation $E(U_n - \theta)^2 = O(n^{-c})$, given $\zeta_{c-1} = 0$, expressed in Corollary 5.2.1.

(iii) In the proof of Theorem 2.3.3, it was seen that

$$E(\bar{X} - \mu)^3 = \mu_3 n^{-2} = O(n^{-2}).$$

This corresponds to (**) in the case $m = 1, c = 1, r = 3$ of Lemma B.

(iv) For a *generalized* U-statistic based on k samples, (**) holds with n given by $n = \min\{n_1, \ldots, n_k\}$. The extension of the preceding proof is straightforward (Problem 5.P.8).

(v) An application of Lemma B in the case $c \geq 2$ arises in connection with the approximation of a U-statistic by its *projection*, as discussed in **5.3.2** below. (Indeed, the proof of Lemma B is based on the method used by Grams and Serfling (1973) to prove Theorem 5.3.2.) ■

5.3 THE PROJECTION OF A U-STATISTIC ON THE BASIC OBSERVATIONS

An appealing feature of a U-statistic is its simple structure as a sum of identically distributed random variables. However, except in the case of a kernel of dimension $m = 1$, the summands are *not* all independent, so that a direct application of the abundant theory for sums of *independent* random variables is not possible. On the other hand, by the special device of "projection," a U-statistic may be approximated within a sufficient degree of accuracy by a sum of I.I.D. random variables. In this way, classical limit theory for sums does carry over to U-statistics and yields the relevant asymptotic distribution theory and almost sure behavior.

Throughout we consider as usual a U-statistic U_n based on a symmetric kernel $h = h(x_1, \ldots, x_m)$ and a sample $X_1, \ldots, X_n$ $(n \geq m)$ from a distribution F, with $\theta = E_F\{h(X_1, \ldots, X_m)\}$.

In **5.3.1** we define and evaluate the projection $\hat{U}_n$ of a U-statistic U_n. In **5.3.2** the moments of $U_n - \hat{U}_n$ are characterized, thus providing the basis for negligibility of $U_n - \hat{U}_n$ in appropriate senses. As an application, a representation for U_n as a mean of I.I.D.'s plus a negligible random variable is obtained in **5.3.3**. Further applications are made in Sections 5.4 and 5.5.

In the case $\zeta_1 = 0$, the projection $\hat{U}_n$ serves no purpose. Thus, in **5.3.4**, we consider an extended notion of projection for the general case $\zeta_0 = \cdots = \zeta_{c-1} = 0 < \zeta_c$.

In Chapter 9 we shall further treat the concept of projection, considering it in general for an arbitrary statistic S_n in place of the U-statistic U_n.

5.3.1 The Projection of U_n

Assume $E_F|h| < \infty$. The *projection* of the U-statistic U_n is defined as

$$(1) \qquad \hat{U}_n = \sum_{i=1}^{n} E_F\{U_n | X_i\} - (n-1)\theta.$$

Note that it is exactly a sum of I.I.D. random variables. In terms of the function $\tilde{h}_1$ considered in Section 5.2. we have

$$(2) \qquad \hat{U}_n - \theta = \frac{m}{n} \sum_{i=1}^{n} \tilde{h}_1(X_i).$$

Verify (Problem 5.P.9) this in the wider context of a *generalized* U-statistic. From (2) it is evident that $\hat{U}_n$ is of no interest in the case $\zeta_1 = 0$. However, in this case we pass to a certain analogue (5.3.4).

5.3.2 The Moments of $U_n - \hat{U}_n$

Here we treat the difference $U_n - \hat{U}_n$. It is useful that $U_n - \hat{U}_n$ may itself be expressed as a U-statistic, namely (Problem 5.P.10).

$$U_n - \hat{U}_n = \frac{1}{\binom{n}{m}} \sum_c H(X_{i_1}, \ldots, X_{i_m}),$$

based on the symmetric kernel

$$H(x_1, \ldots, x_m) = h(x_1, \ldots, x_m) - \tilde{h}_1(x_1) - \cdots - \tilde{h}_1(x_m) - \theta.$$

Note that $E_F\{H\} = E_F\{H | X_1\} = 0$. That is, in an obvious notation, $\zeta_1^{(H)} = 0$. An application of Lemma 5.2.2B with $c = 2$ thus yields

Theorem. *Let* ν *be an even integer. If* $E_F H^\nu < \infty$ *(implied by* $E_F h^\nu < \infty$*), then*

$$(*) \qquad E_F(U_n - \hat{U}_n)^\nu = O(n^{-\nu}), \qquad n \to \infty.$$

For $\nu = 2$, relation (*) was established by Hoeffding (1948) and applied to obtain the CLT for U-statistics, as will be seen in Section 5.5. It also yields the LIL for U-statistics (Section 5.4). Indeed, as seen below in 5.3.3, it leads to an almost sure representation of U_n as a mean of I.I.D.'s. However, for information on the *rates* of convergence in such results as the CLT and SLLN for U-statistics, the case $\nu > 2$ in (*) is apropos. This extension was obtained by Grams and Serfling (1973). Sections 5.4 and 5.5 exhibit some relevant rates of convergence.

5.3.3 Almost Sure Representation of a U-Statistic as a Mean of I.I.D.'s

Theorem. *Let* ν *be an even integer. Suppose that* $E_F h^\nu < \infty$. *Put*

$$U_n = \hat{U}_n + R_n.$$

Then, for any $\delta > 1/\nu$, *with probability* 1

$$R_n = o(n^{-1}(\log n)^\delta), \qquad n \to \infty,$$

PROOF. Let $\delta > 1/\nu$. Put $\lambda_n = n(\log n)^{-\delta}$. It suffices to show that, for any $\varepsilon > 0$, *wp*1 $\lambda_n |R_n| < \varepsilon$ for all n sufficiently large, that is,

(1) $$P(\lambda_n |R_n| > \varepsilon \text{ for infinitely many } n) = 0.$$

Let $\varepsilon > 0$ be given. By the Borel–Cantelli lemma, and since λ_n is nondecreasing for large n, it suffices for (1) to show that

(2) $$\sum_{k=0}^{\infty} P\left(\lambda_{2^{k+1}} \max_{2^k \le n \le 2^{k+1}} |R_n| > \varepsilon\right) < \infty.$$

Since $R_n = U_n - \hat{U}_n$ is itself a U-statistic as noted in **5.3.2** and hence a reverse martingale as noted in Lemma 5.1.5B, we may apply a standard result (Loeve (1978), Section 32) to write

$$P\left(\sup_{j \ge n} |U_j - \hat{U}_j| > t\right) \le t^{-\nu} E|U_n - \hat{U}_n|^\nu.$$

Thus, by Theorem 5.3.2, the kth term in (2) is bounded by (check)

$$\varepsilon^{-\nu}\lambda_{2^{k+1}} E_F |U_{2^k} - \hat{U}_{2^k}|^\nu = O((k+1)^{-\delta\nu}).$$

Since $\delta\nu > 1$, the series in (2) is convergent. ■

The foregoing result is given and utilized by Geertsema (1970).

5.3.4 The "Projection" of U_n for the General Case $\zeta_0 = \cdots = \zeta_{c-1} = 0 < \zeta_c$

(It is assumed that $E_F h^2 < \infty$.) Since $\zeta_d = 0$ for $d < c$, the variance formula for U-statistics (Lemma 5.2.1A) yields

$$\mathrm{Var}_F\{U_n\} = \frac{c!\binom{m}{c}^2 \zeta_c}{n^c} + O(n^{-c-1}), \qquad n \to \infty,$$

and thus

(1) $$\mathrm{Var}_F\{n^{(1/2)c}(U_n - \theta)\} \to c!\binom{m}{c}^2 \zeta_c, \qquad n \to \infty.$$

This suggests that in this case the random variable $n^{(1/2)c}(U_n - \theta)$ converges in distribution to a nondegenerate law.

Now, generalizing 5.3.1, let us define the "projection" of U_n to be $\hat{U}_n$ given by

$$\hat{U}_n - \theta = \sum_{1 \le i_1 < \cdots < i_c \le n} E_F\{U_n | X_{i_1}, \ldots, X_{i_c}\} - \binom{n}{c}\theta.$$

Verify (Problem 5.P.11) that

$$\hat{U}_n - \theta = \frac{m(m-1)\cdots(m-c+1)}{n(n-1)\cdots(n-c+1)} \sum_{1 \le i_1 < \cdots < i_c \le n} \tilde{h}_c(X_{i_1}, \ldots, X_{i_c}). \tag{2}$$

Again (as in 5.3.2), $U_n - \hat{U}_n$ is itself a U-statistic, based on the kernel

$$H(x_1, \ldots, x_m) = h(x_1, \ldots, x_m) - \sum_{1 \le i_1 < \cdots < i_c \le m} \tilde{h}_c(x_{i_1}, \ldots, x_{i_c}) - \theta,$$

with $E_F\{H\} = E_F\{H|X_1\} = \cdots = E_F\{H|X_1, \ldots, X_c\} = 0$, and thus $\zeta_c^{(H)} = 0$. Hence the variance formula for U-statistics yields

$$E(U_n - \hat{U}_n)^2 = O(n^{-(c+1)}), \tag{3}$$

so that $E\{n^{(1/2)c}(U_n - \hat{U}_n)^2\} = O(n^{-1})$ and thus

$$n^{(1/2)c}(U_n - \hat{U}_n) \xrightarrow{p} 0.$$

Hence the limit law of $n^{(1/2)c}(U_n - \theta)$ may be found by obtaining that of $n^{(1/2)c}(\hat{U}_n - \theta)$. For the cases $c = 1$ and $c = 2$, this approach is carried out in Section 5.5.

Note that, more generally, for any even integer v, if $E_F H^v < \infty$ (implied by $E_F h^v < \infty$), then

$$E|U_n - \hat{U}_n|^v = O(n^{-(1/2)v(c+1)}), \qquad n \to \infty, \tag{4}$$

The foregoing results may be extended easily to generalized U-statistics (Problem 5.P.12).

In the case under consideration, that is, $\zeta_{c-1} = 0 < \zeta_c$, the "projection" $\hat{U}_n - \theta$ corresponds to a term in the martingale representation of U_n given by Lemma 5.1.5A. Check (Problem 5.P.13) that $S_{0n} = \cdots = S_{c-1,n} = 0$ and

$$\hat{U}_n - \theta = \binom{m}{c}\binom{n}{c}^{-1} S_{cn}.$$

5.4 ALMOST SURE BEHAVIOR OF U-STATISTICS

The classical SLLN (Theorem 1.8B) generalizes to U-statistics:

Theorem A. *If* $E_F|h| < \infty$, *then* $U_n \xrightarrow{wp1} \theta$.

This result was first established by Hoeffding (1961), using the forward martingale structure of U-statistics given by Lemma 5.1.5A. A more direct proof, noted by Berk (1966), utilizes the *reverse* martingale representation of Lemma 5.1.5B. Since the classical SLLN has been generalized to reverse martingale sequences (see Doob (1953) or Loève (1978)), Theorem A is immediate.

For *generalized* k-sample U-statistics, Sen (1977) obtains strong convergence of U under the condition $E_F\{|h|(\log^+|h|^{k-1}\} < \infty$.

Under a slightly stronger moment assumption, namely $E_F h^2 < \infty$, Theorem A can be proved very simply. For, in this case, we have

$$E_F(U_n - \hat{U}_n)^2 = O(n^{-2})$$

as established in **5.3.2**. Thus $\sum_{n=1}^{\infty} E_F(U_n - \hat{U}_n)^2 < \infty$, so that by Theorem 1.3.5 $U_n - \hat{U}_n \xrightarrow{wp1} 0$. Now, as an application of the classical SLLN,

$$\hat{U}_n - \theta = \frac{m}{n}\sum_{i=1}^{n} \tilde{h}_1(X_i) \xrightarrow{wp1} mE_F\{\tilde{h}_1(X_1)\} = 0.$$

Thus $U_n \xrightarrow{wp1} \theta$. This argument extends to *generalized* U-statistics (Problem 5.P.14).

As an alternate proof, also restricted to the second order moment assumption, Theorem 5.3.3 may be applied for the part $U_n - \hat{U}_n \xrightarrow{wp1} 0$.

In connection with the strong convergence of U-statistics, the following rate of convergence is established by Grams and Serfling (1973). The argument uses Theorem 5.3.2 and the reverse martingale property to reduce to $\hat{U}_n$.

Theorem B. *Let* ν *be an even integer. If* $E_F h^\nu < \infty$, *then for any* $\varepsilon > 0$,

$$P\left(\sup_{k \geq n} |U_n - \theta| > \varepsilon\right) = O(n^{1-\nu}), \qquad n \to \infty.$$

The classical LIL may also be extended to U-statistics. As an exercise (Problem 5.P.15), prove

Theorem C. *Let* $h = h(x_1, \ldots, x_m)$ *be a kernel for* $\theta = \theta(F)$, *with* $E_F h^2 < \infty$ *and* $\zeta_1 > 0$. *Then*

$$\overline{\lim_{n \to \infty}} \frac{n^{1/2}(U_n - \theta)}{(2m^2\zeta_1 \log\log n)^{1/2}} = 1 \text{ wp1}.$$

5.5 ASYMPTOTIC DISTRIBUTION THEORY OF U-STATISTICS

Consider a kernel $h = h(x_1, \ldots, x_m)$ for unbiased estimation of $\theta = \theta(F) = E_F\{h\}$, with $E_F h^2 < \infty$. Let $0 = \zeta_0 \le \zeta_1 \le \cdots \le \zeta_m = \mathrm{Var}_F\{h\}$ be as defined in 5.2.1. As discussed in 5.3.4, in the case $\zeta_{c-1} = 0 < \zeta_c$, the random variable

$$n^{(1/2)c}(U_n - \theta)$$

has variance tending to a positive constant and its asymptotic distribution may be obtained by replacing U_n by its projection $\hat{U}_n$. In the present section we examine the limit distributions for the cases $c = 1$ and $c = 2$, which cover the great majority of applications. For $c = 1$, treated in 5.5.1, the random variable $n^{1/2}(U_n - \theta)$ converges in distribution to a normal law. Corresponding rates of convergence are presented. For $c = 2$, treated in 5.5.2, the random variable $n(U_n - \theta)$ converges in distribution to a weighted sum of (possibly infinitely many) independent χ_1^2 random variables.

5.5.1 The Case $\zeta_1 > 0$

The following result was established by Hoeffding (1948). The proof is left as an exercise (Problem 5.P.16).

Theorem A. *If* $\mathrm{E_F h^2} < \infty$ *and* $\zeta_1 > 0$, *then* $\mathrm{n^{1/2}(U_n - \theta)} \xrightarrow{d} \mathrm{N(0, m^2\zeta_1)}$, *that is,*

$$\mathrm{U_n} \quad \textit{is} \quad \mathrm{AN}\left(\theta, \frac{\mathrm{m^2}\zeta_1}{\mathrm{n}}\right).$$

Example A. *The sample variance.* $\theta(F) = \sigma^2(F)$. As seen in 5.1.1 and 5.2.1, $h(x_1, x_2) = \frac{1}{2}(x_1^2 + x_2^2 - 2x_1x_2)$, $\zeta_1 = (\mu_4 - \sigma^4)/4$, and

$$U_n = s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

Assuming that F is such that $\sigma^4 < \mu_4 < \infty$, so that $E_F h^2 < \infty$ and $\zeta_1 > 0$, we obtain from Theorem A that

$$s^2 \quad \text{is} \quad AN\left(\sigma^2, \frac{\mu_4 - \sigma^4}{n}\right).$$

Compare Section 2.2, where the same conclusion was established for $m_2 = (n-1)s^2/n$.

In particular, suppose that F is binomial $(1, p)$. Then $\bar{X} = \hat{p}$, say, and (check) $s^2 = n\hat{p}(1 - \hat{p})/(n-1)$. Check that $\mu_4 - \sigma^4 > 0$ if and only if $p \ne \frac{1}{2}$. Thus Theorem A is applicable for $p \ne \frac{1}{2}$. (The case $p = \frac{1}{2}$ will be covered by Theorem 5.5.2.) ∎

By routine arguments (Problem 5.P.18) it may be shown that a *vector* of several U-statistics based on the same sample is asymptotically multivariate normal. The appropriate limit covariance matrix may be found by the same method used in **5.2.1** for the computation of variances to terms of order $O(n^{-1})$.

It is also straightforward (Problem 5.P.19) to extend Theorem A to *generalized* U-statistics. In an obvious notation, for a k-sample U-statistic, we have

$$U \quad \text{is} \quad AN\left(\theta, \sum_{j=1}^{k} \frac{m_j^2 \zeta_{1j}}{n_j}\right),$$

provided that $n \sum m_j^2 \zeta_{1j}/n_j \geq B > 0$ as $n = \min\{n_1, \ldots, n_k\} \to \infty$.

Example B. *The Wilcoxon 2-sample statistic* (continuation of Example 5.1.3). Here $\theta = P(X \leq Y)$ and $h(x, y) = I(x \leq y)$. Check that $\zeta_{11} = P(X \leq Y_1, X \leq Y_2) - \theta^2$, $\zeta_{12} = P(X \leq Y, X_2 \leq Y) - \theta^2$. Under the null hypothesis that $\mathscr{L}(X) = \mathscr{L}(Y)$, we have $\theta = \frac{1}{2}$ and $\zeta_{11} = P(Y_3 \leq Y_1, Y_3 \leq Y_2) - \frac{1}{4} = \frac{1}{3} - \frac{1}{4} = \frac{1}{12} = \zeta_{12}$. In this case

$$U \quad \text{is} \quad AN\left(\frac{1}{2}, \frac{1}{12}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right). \quad \blacksquare$$

The rate of convergence in the asymptotic normality of U-statistics has been investigated by Grams and Serfling (1973), Bickel (1974), Chan and Wierman (1977) and Callaert and Janssen (1978), the latter obtaining the sharpest result, as follows.

Theorem B. *If* $\nu = E|h|^3 < \infty$ *and* $\zeta_1 > 0$, *then*

$$\sup_{-\infty < t < \infty} \left| P\left(\frac{n^{1/2}(U_n - \theta)}{m\zeta_1^{1/2}} \leq t\right) - \Phi(t)\right| \leq C\nu(m^2\zeta_1)^{-3/2} n^{-1/2},$$

where C *is an absolute constant.*

5.5.2 The Case $\zeta_1 = 0 < \zeta_2$

For the function $\tilde{h}_2(x_1, x_2)$ associated with the kernel $h = h(x_1, \ldots, x_m)$ ($m \geq 2$), we define an operator A on the function space $L_2(R, F)$ by

$$Ag(x) = \int_{-\infty}^{\infty} \tilde{h}_2(x, y) g(y) dF(y), \qquad x \in R, g \in L_2.$$

That is, A takes a function g into a new function Ag. In connection with any such operator A, we define the associated eigenvalues $\lambda_1, \lambda_2, \ldots$ to be the real numbers λ (not necessarily distinct) corresponding to the distinct solutions $g_1, g_2, \ldots$ of the equation

$$Ag - \lambda g = 0.$$

We shall establish

Theorem. *If* $E_F h^2 < \infty$ *and* $\zeta_1 = 0 < \zeta_2$, *then*

$$n(U_n - \theta) \xrightarrow{d} \frac{m(m-1)}{2} Y,$$

where Y *is a random variable of the form*

$$Y = \sum_{j=1}^{\infty} \lambda_j (\chi_{1j}^2 - 1),$$

where $\chi_{11}^2, \chi_{12}^2, \ldots$ *are independent* χ_1^2 *variates, that is,* Y *has characteristic function*

$$E_F\{e^{itY}\} = \prod_{j=1}^{\infty} (1 - 2it\lambda_j)^{-1/2} e^{-it\lambda_j}.$$

Before developing the proof, let us illustrate the application of the theorem.

Example A. *The sample variance* (continuation of Examples 5.2.1A and 5.5.1A). We have $\tilde{h}_2(x, y) = \frac{1}{2}(x - y)^2 - \sigma^2$, $\zeta_1 = (\mu_4 - \sigma^4)/4$, and $\zeta_2 = \frac{1}{2}(\mu_4 + \sigma^4)$. Take now the case $\zeta_1 = 0$, that is, $\mu_4 = \sigma^4$. Then $\zeta_2 = \sigma^4 > 0$ and the preceding theorem may be applied. We seek values λ such that the equation

$$\int [\tfrac{1}{2}(x - y)^2 - \sigma^2] g(y) dF(y) = \lambda g(x)$$

has solutions g in $L_2(R, F)$. It is readily seen (justify) that any such g must be quadratic in form: $g(x) = ax^2 + bx + c$. Substituting this form of g in the equation and equating coefficients of x^0, x^1 and x^2, we obtain the system of equations

$$\frac{1}{2}\int y^2 g(y) dF(y) - \sigma^2 \int g(y) dF(y) = \lambda c, \qquad -\int y g(y) dF(y) = \lambda b,$$

$$\frac{1}{2}\int g(y) dF(y) = \lambda a.$$

Solutions (a, b, c, λ) depend upon F. In particular, suppose that F is binomial $(1, p)$, with $p = \frac{1}{2}$. Then (check) $\sigma^2 = \frac{1}{4}$, $\mu_4 = \sigma^4$, $\int y^k \, dF(y) = \frac{1}{2}$ for all k. Then (check) the system of equations becomes equivalently

$$a + b + 2c = 4a\lambda, \qquad a + b + c = -2b\lambda, \qquad a + b + c = (4c + 2a)\lambda.$$

It is then easily found (check) that $a = 0$, $b = -2c$, and $\lambda = -\frac{1}{4}$, in which case $g(x) = c(2x - 1)$, c arbitrary. The theorem thus yields, for this F,

$$n(s^2 - \tfrac{1}{4}) \xrightarrow{d} -\tfrac{1}{4}(\chi_1^2 - 1). \quad \blacksquare$$

Remark. Do s^2 and m_2 $(=(n-1)s^2/n)$ always have the same asymptotic distribution? Intuitively this would seem plausible, and indeed typically it is true. However, for F binomial $(1, \frac{1}{2})$, we have (Problem 5.P.22)

$$n(m_2 - \tfrac{1}{4}) \xrightarrow{d} -\tfrac{1}{4}\chi_1^2,$$

which differs from the above result for s^2. $\blacksquare$

Example B. $\theta(F) = \mu^2(F)$. We have $h(x_1, x_2) = x_1 x_2$ and

$$U_n = \frac{1}{\binom{n}{2}} \sum_{i<j} X_i X_j.$$

Check that $\zeta_1 = \mu^2\sigma^2$ and $\zeta_2 = \sigma^4 - 2\mu^2\sigma^2$. Assume that $0 < \sigma^2 < \infty$. Then we have the case $\zeta_1 > 0$ if $\mu \neq 0$ and the case $\zeta_1 = 0 < \zeta_2$ if $\mu = 0$. Thus

(i) If $\mu \neq 0$, Theorem 5.5.1A yields

$$U_n \quad \text{is} \quad AN\left(\mu^2, \frac{4\mu^2\sigma^2}{n}\right);$$

(ii) If $\mu = 0$, the above theorem yields (check)

$$\frac{nU_n}{\sigma^2} \xrightarrow{d} \chi_1^2 - 1. \quad \blacksquare$$

Example C. (*continuation of Example* 5.1.1(ix)). Here find that $\zeta_1 > 0$ for any value of θ, $0 < \theta < 1$. Thus Theorem 5.5.1A covers all situations, and the present theorem has no role. $\blacksquare$

PROOF OF THE THEOREM. On the basis of the discussion in **5.3.4**, our objective is to show that the random variable

$$n(\hat{U}_n - \theta) = \frac{m(m-1)}{n-1} \sum_{1 \le i < j \le n} \tilde{h}_2(X_i, X_j)$$

converges in distribution to

$$\frac{m(m-1)}{2} Y,$$

where

$$Y = \sum_{j=1}^{\infty} \lambda_j (W_j^2 - 1),$$

with $W_1^2, W_2^2, \ldots$ being independent χ_1^2 random variables. Putting

$$T_n = \frac{1}{n} \sum_{i \neq j} \tilde{h}_2(X_i, X_j),$$

we have

$$n(\hat{U}_n - \theta) = \frac{m(m-1)}{2} \frac{n}{n-1} T_n.$$

Thus our objective is to show that

$$(*) \qquad T_n \xrightarrow{d} Y.$$

We shall carry this out by the method of characteristic functions, that is, by showing that

$$(**) \qquad E_F\{e^{ixT_n}\} \to E\{e^{ixY}\}, \qquad n \to \infty, \text{ each } x.$$

A special representation for $\tilde{h}_2(x, y)$ will be used. Let $\{\phi_j(\cdot)\}$ denote orthonormal eigenfunctions corresponding to the eigenvalues $\{\lambda_j\}$ defined in connection with $\tilde{h}_2$. (See Dunford and Schwartz (1963), pp. 905, 1009, 1083, 1087). Thus

$$E_F\{\phi_j(X)\phi_k(X)\} = \begin{cases} 1, & j = k \\ 0, & j \neq k, \end{cases}$$

and $\tilde{h}_2(x, y)$ may be expressed as the mean square limit of $\sum_{k=1}^{K} \lambda_k \phi_k(x)\phi_k(y)$ as $K \to \infty$. That is,

$$(1) \qquad \lim_{K \to \infty} E_F\left\{\left[\tilde{h}_2(X_1, X_2) - \sum_{k=1}^{K} \lambda_k \phi_k(X_1)\phi_k(X_2)\right]^2\right\} = 0,$$

and we write

$$(2) \qquad \tilde{h}_2(x, y) = \sum_{k=1}^{\infty} \lambda_k \phi_k(x)\phi_k(y).$$

Then (Problem 5.P.24(a)), in the same sense,

$$(3) \qquad \tilde{h}_1(x) = \sum_{k=1}^{\infty} \lambda_k \phi_k(x) E_F\{\phi_k(X)\}.$$

Therefore, since $\zeta_1 = 0$,

$$E_F\{\phi_k(X)\} = 0, \qquad \text{all } k.$$

Furthermore (Problem 5.P.24(b)),

$$(4) \quad E_F\left\{\left[\tilde{h}_2(X_1, X_2) - \sum_{k=1}^{K} \lambda_k \phi_k(X_1)\phi_k(X_2)\right]^2\right\} = E_F\{\tilde{h}_2^2(X_1, X_2)\} - \sum_{k=1}^{K} \lambda_k^2,$$

whence (by (1))

$$\sum_{k=1}^{\infty} \lambda_k^2 = E_F\{\tilde{h}_2^2(X_1, X_2)\} < \infty.$$

In terms of the representation (2), T_n may be expressed as

$$T_n = \frac{1}{n} \sum_{i \neq j} \sum_{k=1}^{\infty} \lambda_k \phi_k(X_i)\phi_k(X_j).$$

Now put

$$T_{nK} = \frac{1}{n} \sum_{i \neq j} \sum_{k=1}^{K} \lambda_k \phi_k(X_i)\phi_k(X_j).$$

Using the inequality $|e^{iz} - 1| \leq |z|$, we have

$$\begin{aligned} |E\{e^{ixT_n}\} - E\{e^{ixT_{nK}}\}| &\leq E|e^{ixT_n} - e^{ixT_{nK}}| \\ &\leq |x|E|T_n - T_{nK}| \\ &\leq |x|[E(T_n - T_{nK})^2]^{1/2}. \end{aligned}$$

Next it is shown that

(5) $$E(T_n - T_{nK})^2 \leq 2 \sum_{k=K+1}^{\infty} \lambda_k^2.$$

Observe that $T_n - T_{nK}$ is basically of the form of a U-statistic, that is,

$$T_n - T_{nK} = \frac{2}{n}\binom{n}{2} U_{nK},$$

where

$$U_{nK} = \binom{n}{2}^{-1} \sum_{i<j} g_K(X_i, X_j),$$

with

$$g_K(x, y) = \tilde{h}_2(x, y) - \sum_{k=1}^{K} \lambda_k \phi_k(x)\phi_k(y).$$

Justify (Problem 5.P.24(c)) that

(6a) $$E_F\{g_K(X_1, X_2)\} = 0$$

(6b) $$E_F\{g_K^2(X_1, X_2)\} = \sum_{k=K+1}^{\infty} \lambda_k^2,$$

(6c) $$E_F\{g_K(x, X)\} \equiv 0.$$

Hence $E\{U_{nk}\} = 0$ and, by Lemma 5.2.1A,

$$E\{U_{nk}^2\} = \binom{n}{2}^{-1} \sum_{k=K+1}^{\infty} \lambda_k^2.$$

Thus

$$E(T_n - T_{nK})^2 = (n-1)^2 \binom{n}{2}^{-1} \sum_{k=K+1}^{\infty} \lambda_k^2 \leq 2 \sum_{k=K+1}^{\infty} \lambda_k^2,$$

yielding (5).

Now fix x and let $\varepsilon > 0$ be given. Choose and fix K large enough that

$$|x| \left(2 \sum_{k=K+1}^{\infty} \lambda_k^2\right)^{1/2} < \varepsilon.$$

Then we have established that

$$|E\{e^{ixT_n}\} - E\{e^{ixT_{nK}}\}| < \varepsilon, \qquad \text{all } n. \tag{7}$$

Next let us show that

$$T_{nK} \xrightarrow{d} Y_K = \sum_{k=1}^{K} \lambda_k (W_k^2 - 1). \tag{8}$$

We may write

$$T_{nK} = \sum_{k=1}^{K} \lambda_k (W_{kn}^2 - Z_{kn}),$$

where

$$W_{kn} = n^{-1/2} \sum_{i=1}^{n} \phi_k(X_i)$$

and

$$Z_{kn} = n^{-1} \sum_{i=1}^{n} \phi_k^2(X_i).$$

From the foregoing considerations, it is seen that

$$E\{W_{kn}\} = 0, \qquad \text{all } k \text{ and } n,$$

and

$$\operatorname{Cov}\{W_{jn}, W_{kn}\} = \begin{cases} 1, & j = k, \\ 0, & j \neq k, \end{cases} \qquad \text{all } j, k \text{ and } n.$$

Therefore, by the Lindeberg–Lévy CLT,

$$(W_{1n}, \ldots, W_{Kn}) \xrightarrow{d} N(0, \mathbf{I}_{K \times K}).$$

Also, since $E_F\{\phi_k^2(X)\} = 1$, the classical SLLN gives

$$(Z_{1n}, \ldots, Z_{Kn}) \xrightarrow{wp1} (1, \ldots, 1).$$

Consequently (8) holds and thus

(9) $$|E\{e^{ixT_{nK}}\} - E\{e^{ixY_K}\}| < \varepsilon, \qquad \text{all } n \text{ sufficiently large.}$$

Finally, we show that

(10) $$|E\{e^{ixY_K}\} - E\{e^{ixY}\}| < \varepsilon[E(W_1^2 - 1)^2]^{1/2}, \qquad \text{all } n.$$

To accomplish this, let the random variables $W_1^2, W_2^2, \ldots$ be defined on a common probability space and represent Y as the limit in mean square of Y_K as $K \to \infty$. Then

$$\begin{aligned} |E\{e^{ixY_K}\} - E\{e^{ixY}\}| &\le |x|[E(Y - Y_K)^2]^{1/2} \\ &\le |x|[E(W_1^2 - 1)^2]^{1/2}\left[\sum_{k=K+1}^{\infty} \lambda_k^2\right]^{1/2}, \end{aligned}$$

yielding (10). Combining (7), (9) and (10), we have, for any x and any $\varepsilon > 0$, and for all n sufficiently large,

$$|E\{e^{ixT_n}\} - E\{e^{ixY}\}| \le \varepsilon\{2 + [E(W_1^2 - 1)^2]^{1/2}\},$$

proving (**). ■

This theorem has also been proved, independently, by Gregory (1977).

5.6 PROBABILITY INEQUALITIES AND DEVIATION PROBABILITIES FOR U-STATISTICS

Here we augment the convergence results of Sections 5.4 and 5.5 with exact exponential-rate bounds for $P(U_n - \theta \ge t)$ and with asymptotic estimates of moderate deviation probabilities

$$P\left(\frac{n^{1/2}(U_n - \theta)}{(m^2\zeta_1)^{1/2}} \ge c(\log n)^{1/2}\right).$$

5.6.1 Probability Inequalities for U-Statistics

For any random variable Y possessing a moment generating function $E\{e^{sY}\}$ for $0 < s < s_0$, one may obtain a probability inequality by writing

$$P(Y - E\{Y\} \ge t) = P(s[Y - E\{Y\} - t] \ge 0) \le e^{-st}E\{e^{s[Y - E\{Y\}]}\}$$

and minimizing with respect to $s \in (0, s_0]$. In applying this technique, we make use of the following lemmas. The first lemma will involve the function

$$f(x, y) = \frac{x}{x+y} e^{-y} + \frac{y}{x+y} e^{x}, \qquad x > 0, y > 0.$$

Lemma A. *Let* $E\{Y\} = \mu$ *and* $\mathrm{Var}\{Y\} = \nu$.

(i) *If* $P(Y \le b) = 1$, *then*

$$E\{e^{s(Y-\mu)}\} \le f(s(b-\mu), s\nu/(b-\mu)), \; s > 0.$$

(ii) *If* $P(a \le Y \le b) = 1$, *then*

$$E\{e^{s(Y-\mu)}\} \le e^{(1/8)s^2(b-a)^2}, \; s > 0.$$

PROOF. (i) is proved in Bennett (1962), p. 42. Now, in the proof of Theorem 2 of Hoeffding (1963), it is shown that

$$qe^{-zp} + pe^{zq} \le e^{(1/8)z^2},$$

for $0 < p < 1$, $q = 1 - p$. Putting $p = y/(x + y)$ and $z = (x + y)$, we have

$$f(x, y) \le e^{(1/8)(x+y)^2},$$

so that (i) yields

$$E\{e^{s(Y-\mu)}\} \le e^{(1/8)s^2[(b-\mu)+\nu/(b-\mu)]^2}.$$

Now, as pointed out by Hoeffding (1963), $\nu = E(Y - \mu)^2 = E(Y - \mu)(Y - a) \le (b - \mu)E(Y - a) = (b - \mu)(\mu - a)$. Thus (ii) follows. ■

The next lemma may be proved as an exercise (Problem 5.P.25).

Lemma B. *If* $E\{e^{sY}\} < \infty$ *for* $0 < s < s_0$, *and* $E\{Y\} = \mu$, *then*

$$E\{e^{s(Y-\mu)}\} = 1 + O(s^2), \qquad s \to 0.$$

In passing to U-statistics, we shall utilize the following relation between the moment generating function of a U-statistic and that of its kernel.

Lemma C. *Let* $h = h(x_1, \ldots, x_m)$ *satisfy*

$$\psi_h(s) = E_F\{e^{sh(X_1,\ldots,X_m)}\} < \infty, \qquad 0 < s \le s_0.$$

Then

$$E_F\{e^{sU_n}\} \le \psi_h^k\left(\frac{s}{k}\right), \qquad 0 < s \le s_0 k,$$

where $k = [n/m]$.

PROOF. By 5.1.6, $U_n = (n!)^{-1} \sum_p W(X_{i_1}, \ldots, X_{i_n})$, where each $W(\cdot)$ is an average of $k = [n/m]$ I.I.D. random variables. Since the exponential function is convex, it follows by Jensen's inequality that

$$e^{sU_n} = e^{s(n!)^{-1} \Sigma_p W(\cdot)} \leq (n!)^{-1} \sum_p e^{sW(X_{i_1}, \ldots, X_{i_n})}.$$

Complete the proof as an exercise (Problem 5.P.26). ■

We now give three probability inequalities for U-statistics. The first two, due to Hoeffding (1963), require h to be bounded and give very useful explicit exponential-type bounds. The third, due to Berk (1970), requires less on h but asserts only an implicit exponential-type bound.

Theorem A. *Let* $h = h(x_1, \ldots, x_m)$ *be a kernel for* $\theta = \theta(F)$, *with* $a \leq h(x_1, \ldots, x_m) \leq b$. *Put* $\theta = E\{h(X_1, \ldots, X_m)\}$ *and* $\sigma^2 = \mathrm{Var}\{h(X_1, \ldots, X_m)\}$, *Then, for* $t > 0$ *and* $n \geq m$,

(1) $$P(U_n - \theta \geq t) \leq e^{-2[n/m]t^2/(b-a)^2}$$

and

(2) $$P(U_n - \theta \geq t) \leq e^{-[n/m]t^2/2[\sigma^2 + (1/3)(b-\theta)t]}.$$

PROOF. Write, by Lemmas A and C, with $k = [n/m]$ and $s > 0$,

$$P(U_n - \theta \geq t) \leq E_F\{e^{s(U_n - \theta - t)}\} \leq e^{-st}\left[e^{-(s/k)\theta}\psi_h\left(\frac{s}{k}\right)\right]^k$$
$$\leq e^{-st + (1/8)s^2(b-a)^2/k}.$$

Now minimize with respect to s and obtain (1). A similar argument leads to

(2′) $$P(U_n - \theta \geq t) \leq \exp\left[\frac{-kt\left\{\left[1 + \frac{\sigma^2}{(b-\theta)t}\right]\log\left[1 + \frac{(b-\theta)t}{\sigma^2}\right] - 1\right\}}{(b-\theta)}\right].$$

It is shown in Bennett (1962) that the right-hand side of (2′) is less than or equal to that of (2). ■

(Compare Lemmas 2.3.2 and 2.5.4A.)

Theorem B. *Let* $h = h(x_1, \ldots, x_m)$ *be a kernel for* $\theta = \theta(F)$, *with*

$$E_F\{e^{sh(X_1, \ldots, X_m)}\} < \infty, \qquad 0 < s \leq s_0.$$

Then, for every $\varepsilon > 0$, *there exist* $C_\varepsilon > 0$ *and* $\rho_\varepsilon < 1$ *such that*

$$P(U_n - \theta \geq \varepsilon) \leq C_\varepsilon \rho_\varepsilon^n, \qquad \textit{all } n \geq m.$$

PROOF. For $0 < t \le s_0 k$, where $k = [n/m]$, we have by Lemma C that

$$P(U_n - \theta \ge \varepsilon) \le e^{-t\varepsilon}\left[e^{-(t/k)\theta}\psi_h\left(\frac{t}{k}\right)\right]^k$$
$$= [e^{-s\varepsilon}e^{-s\theta}\psi_h(s)]^k,$$

where $s = t/k$. By Lemma B, $e^{-s\theta}\psi_h(s) = 1 + O(s^2)$, $s \to 0$, so that

$$e^{-s\varepsilon}e^{-s\theta}\psi_h(s) = 1 - \varepsilon s + O(s^2), \qquad s \to 0,$$
$$< 1 \text{ for } s = s_\varepsilon \text{ sufficiently small.}$$

Complete the proof as an exercise. ∎

Note that Theorems A and B are applicable for n small as well as for n large.

5.6.2 "Moderate Deviation" Probability Estimates for *U*-Statistics

A "moderate deviation" probability for a U-statistic is given by

$$q_n(c) = P\left(\frac{n^{1/2}(U_n - \theta)}{(m^2\zeta_1)^{1/2}} > c(\log n)^{1/2}\right),$$

where $c > 0$ and it is assumed that the relevant kernel h has finite second moment and $\zeta_1 > 0$. Such probabilities are of interest in connection with certain asymptotic relative efficiency computations, as will be seen in Chapter 10. Now the CLT for U-statistics tells us that $q_n(c) \to 0$. Indeed, Chebyshev's inequality yields a bound,

$$q_n(c) \le \frac{1}{c^2 \log n} = O((\log n)^{-1}).$$

However, this result and its analogues, $O((\log n)^{-(1/2)\nu})$, under ν-th order moment assumptions on h are quite weak. For, in fact, if h is bounded, then (Problem 5.P.29) Theorem 5.6.1A implies that for any $\delta > 0$

(1) $$q_n(c) = O(n^{-[(1-\delta)2m\zeta_1/(b-a)^2]c^2}),$$

where $a \le h \le b$. Note also that if merely $E_F|h|^3 < \infty$ is assumed, then for c sufficiently small (namely, $c < 1$), the Berry-Esséen theorem for U-statistics (Theorem 5.5.1B) yields an *estimate*:

(*) $$q_n(c) \sim 1 - \Phi(c(\log n)^{1/2}) \sim \frac{1}{(2\pi c^2 \log n)^{1/2}} n^{-(1/2)c^2}.$$

However, under the stronger assumption $E_F|h|^\nu < \infty$ for some $\nu > 3$, this approach does not yield greater latitude on the range of c. A more intricate analysis is needed. To this effect, the following result has been established by Funk (1970), generalizing a pioneering theorem of Rubin and Sethuraman (1965a) for the case U_n a sample mean.

Theorem. *If* $E_F|h|^\nu < \infty$, *where* $\nu > 2$, *then* (*) *holds for* $c^2 < \nu - 2$.

5.7 COMPLEMENTS

In **5.7.1** we discuss stochastic processes associated with a sequence of U-statistics and generalize the CLT for U-statistics. In **5.7.2** we examine the Wilcoxon one-sample statistic and prove assertions made in **2.6.5** for a particular confidence interval procedure. Extension of U-statistic results to the related V-statistics is treated in **5.7.3**. Finally, miscellaneous further complements and extensions are noted in **5.7.4**.

5.7.1 Stochastic Processes Associated with a Sequence of U-Statistics

Let $h = h(x_1, \ldots, x_m)$ be a kernel for $\theta = \theta(F)$, with $E_F(h^2) < \infty$ and $\zeta_1 > 0$. For the corresponding sequence of U-statistics, $\{U_n\}_{n \geq m}$, we consider two associated sequences of stochastic processes on the unit interval [0, 1].

In one of these sequences of stochastic processes, the nth random function is based on $U_m, \ldots, U_n$ and summarizes the *past* history of $\{U_i\}_{i \leq n}$. In the other sequence of processes, the nth random function is based on $U_n, U_{n+1}, \ldots$ and summarizes the *future* history of $\{U_i\}_{i \geq n}$. Each sequence of processes converges in distribution to the *Wiener* process on [0, 1], which we denote by $W(\cdot)$ (recall **1.11.4**).

The process pertaining to the *future* was introduced and studied by Loynes (1970). The nth random function, $\{Z_n(t), 0 \leq t \leq 1\}$, is defined by

$$Z_n(0) = 0;$$

$$Z_n(t_{nk}) = \frac{U_k - \theta}{(\mathrm{Var}\{U_n\})^{1/2}}, \ k \geq n, \qquad \text{where} \quad t_{nk} = \frac{\mathrm{Var}\{U_k\}}{\mathrm{Var}\{U_n\}};$$

$$Z_n(t) = Z_n(t_{nk}), \qquad t_{n,k+1} < t < t_{nk}.$$

For each n, the "times" $t_{nn}, t_{n,n+1}, \ldots$ form a sequence tending to 0 and $Z_n(\cdot)$ is a step function continuous from the left. We have

Theorem A (Loynes). $Z_n(\cdot) \xrightarrow{d} W(\cdot)$ *in* D[0, 1].

This result generalizes Theorem 5.5.1A (asymptotic normality of U_n) and provides additional information such as

Corollary. *For* $x > 0$,

$$\text{(1)} \qquad \lim_{n\to\infty} P\left(\sup_{k \geq n}(U_k - \theta) \geq x(\mathrm{Var}\{U_n\})^{1/2}\right) = P\left(\sup_{0 \leq t \leq 1} W(t) \geq x\right) = 2[1 - \Phi(x)]$$

and

(2)
$$\lim_{n\to\infty} P\left(\inf_{k\geq n}(U_k - \theta) \leq -x(\mathrm{Var}\{U_n\})^{1/2}\right)$$
$$= P\left(\inf_{0\leq t\leq 1} W(t) \leq -x\right) = 2[1 - \Phi(x)].$$

As an exercise, show that the *strong convergence* of U_n to θ follows from this corollary, under the assumption $E_F\{h^2\} < \infty$.

The process pertaining to the *past* has been dealt with by Miller and Sen (1972). Here the nth random function, $\{Y_n(t), 0 \leq t \leq 1\}$, is defined by

$$Y_n(t) = 0, \qquad 0 \leq t \leq \frac{m-1}{n};$$

$$Y_n\left(\frac{k}{n}\right) = \frac{k(U_k - \theta)}{(m^2\zeta_1)^{1/2}n^{1/2}}, \qquad k = m, m+1, \ldots, n;$$

$Y_n(t)$ defined elsewhere, $0 \leq t \leq 1$, by linear interpolation.

Theorem B (Miller and Sen). $Y_n(\cdot) \xrightarrow{d} W(\cdot)$ in $C[0, 1]$.

This result likewise generalizes Theorem 5.5.1A and provides additional information such as

(3)
$$\lim_{n\to\infty} P\left(\sup_{m\leq k\leq n} k(U_k - \theta) \geq x(m^2\zeta_1)^{1/2}n^{1/2}\right) = 2[1 - \Phi(x)], \qquad x > 0.$$

Comparison of (1) and (3) illustrates how Theorems A and B complement each other in the type of additional information provided beyond Theorem 5.5.1A.

See the Loynes paper for treatment of other random variables besides U-statistics. See the Miller and Sen paper for discussion of the use of Theorem B in the *sequential* analysis of U-statistics.

5.7.2 The Wilcoxon One-Sample Statistic as a *U*-Statistic

For testing the hypothesis that the median of a continuous symmetric distribution F is 0, that is, $\xi_{1/2} = 0$, the Wilcoxon one-sample test may be based on the statistic

$$\sum_{1\leq i<j\leq n} I(X_i + X_j > 0).$$

Equivalently, one may perform the test by estimating $G(0)$, where G is the distribution function $G(t) = P(\frac{1}{2}(X_1 + X_2) \leq t)$, with the null hypothesis to be rejected if the estimate differs sufficiently from the value $\frac{1}{2}$. In this way one may treat the related statistic

$$U_n = \frac{1}{\binom{n}{2}} \sum_{1 \le i < j \le n} I(X_i + X_j \le 0)$$

as an estimate of $G(0)$. This, of course, is a U-statistic (recall Example 5.1.1(ix)), so that we have the convenience of asymptotic normality (recall Example 5.5.2C-check as exercise).

In **2.6.5** we considered a related confidence interval procedure for $\xi_{1/2}$. In particular, we considered a procedure of Geertsema (1970), giving an interval

$$I_{Wn} = (W_{na_n}, W_{nb_n})$$

formed by a pair of the ordered values $W_{n1} \le \cdots \le W_{nN_n}$ of the $N_n = \binom{n}{2}$ averages $\frac{1}{2}(X_i + X_j)$, $1 \le i < j \le n$. We now show how the properties stated for I_{Wn} in **2.6.5** follow from a treatment of the U-statistic character of the random variable

$$G_n(x) = \frac{1}{\binom{n}{2}} \sum_{1 \le i < j \le n} I[\tfrac{1}{2}(X_i + X_j) \le x].$$

Note that G_n, considered as a function of x, represents a "sample distribution function" for the averages $\frac{1}{2}(X_i + X_j)$, $1 \le i < j \le n$. From our theory of U-statistics, we see that $G_n(x)$ is asymptotically normal. In particular, $G_n(\xi_{1/2})$ is asymptotically normal. The connection with the W_{nj}'s is as follows. Recall the Bahadur representation (**2.5.2**) relating order statistics X_{nk_n} to the sample distribution function F_n. Geertsema proves the analogue of this result for W_{nkn} and G_n. The argument is similar to that of Theorem 2.5.2, with the use of Theorem 5.6.1A in place of Lemma 2.5.4A.

Theorem. *Let* F *satisfy the conditions stated in* **2.6.5**. *Let*

$$\frac{k_n}{\binom{n}{2}} = \frac{1}{2} + o\left(\frac{\log n}{n^{1/2}}\right), \qquad n \to \infty.$$

Then

$$W_{nk_n} = \xi_{1/2} + \frac{\binom{n}{2}^{-1} k_n - G_n(\xi_{1/2})}{g(\xi_{1/2})} + R_n$$

where with probability 1

$$R_n = O(n^{-3/4} \log n), \qquad n \to \infty.$$

It is thus seen, via this theorem, that properties of the interval I_{W_n} may be derived from the theory of U-statistics.

5.7.3 Implications for V-Statistics

In connection with a kernel $h = h(x_1, \ldots, x_m)$, let us consider again the V-statistic introduced in 5.1.2. Under appropriate moment conditions, the U-statistic and V-statistic associated with h are closely related in behavior, as the following result shows.

Lemma. *Let* r *be a positive integer. Suppose that*

$$E_F|h(X_{i_1}, \ldots, X_{i_m})|^r < \infty, \qquad \textit{all } i \le i_1, \ldots, i_m \le m.$$

Then

$$E|U_n - V_n|^r = O(n^{-r}).$$

PROOF. Check that

$$n^m(U_n - V_n) = (n^m - n_{(m)})(U_n - W_n),$$

where $n_{(m)} = n(n-1)\cdots(n-m+1)$ and W_n is the average of all terms $h(X_{i_1}, \ldots, X_{i_m})$ with at least one equality $i_a = i_b$, $a \ne b$. Next check that

$$n^m - n_{(m)} = O(n^{m-1}).$$

Finally, apply Minkowski's inequality. ■

Application of the lemma in the case $r = 2$ yields

$$n^{1/2}(U_n - V_n) \xrightarrow{p} 0,$$

in which case $n^{1/2}(U_n - \theta)$ and $n^{1/2}(V_n - \theta)$ have the same limit distribution, a useful relationship in the case $\zeta_1 > 0$. In fact, this latter result can actually be obtained under slightly weaker moment conditions on the kernel (see Bönner and Kirschner (1977).)

5.7.4 Further Complements and Extensions

(i) *Distribution-free estimation of the variance of a U-statistic* is considered by Sen (1960).

(ii) Consideration of U-statistics when the distribution of $X_1, X_2, \ldots$ are *not necessarily identical* may be found in Sen (1967).

(iii) *Sequential confidence intervals based on U-statistics* are treated by Sproule (1969a, b).

(iv) *Jackknifing* of estimates which are functions of U-statistics, in order to reduce bias and to achieve other properties, is treated by Arvesen (1969).

(v) Further results on *probabilities of deviations* (recall **5.6.2**) of U-statistics are obtained, via some further results on *stochastic processes* associated with U-statistics (recall **5.7.1**), by Sen (1974).

(vi) Consideration of U-statistics for *dependent* observations $X_1, X_2, \ldots$ arises in various contexts. For the case of *m-dependence*, see Sen (1963), (1965). For the case of *sampling without replacement* from a finite population, see Nandi and Sen (1963). For a treatment of the Wilcoxon 2-sample statistic in the case of samples from a *weakly dependent stationary process*, see Serfling (1968).

(vii) A somewhat different treatment of the case $\zeta_1 = 0 < \zeta_2$ has been given by Rosén (1969). He obtains asymptotic *normality* for U_n when the observations $X_1, \ldots, X_n$ are assumed to have a common distribution $F^{(n)}$ which behaves in a specified fashion as $n \to \infty$. In this treatment $F^{(n)}$ is constrained *not* to remain fixed as $n \to \infty$.

(viii) A general treatment of *symmetric* statistics exploiting an orthogonal expansion technique has been carried out by Rubin and Vitale (1980). For example, U-statistics and V-statistics are types of symmetric statistics. Rubin and Vitale provide a unified approach to the asymptotic distribution theory of such statistics, obtaining as limit random variable a weighted sum of products of Hermite polynomials evaluated at $N(0, 1)$ variates.

5.P PROBLEMS

Section 5.1

1. Check the relations $E_F\{g_1(X_1)\} = 0$, $E_F\{g_2(x_1, X_2)\} = 0, \ldots$ in **5.1.5**.

2. Prove Lemma 5.1.5B.

Section 5.2

3. (i) Show that $\zeta_0 \le \zeta_1 \le \cdots \le \zeta_m$.

(ii) Show that $\zeta_1 \le \frac{1}{2}\zeta_2$. (Hint: Consider the function g_2 of **5.1.5**.)

4. Let $\{a_1, \ldots, a_m\}$ and $\{b_1, \ldots, b_m\}$ be two sets of m distinct integers from $\{1, \ldots, n\}$ with exactly c integers in common. Show that

$$E_F\{\tilde{h}(X_{a_1}, \ldots, X_{a_m})\tilde{h}(X_{b_1}, \ldots, X_{b_m})\} = \zeta_c.$$

5. In Lemma 5.2.1A, derive (iii) from (*).

6. Extend Lemma 5.2.1A(*) to the case of a generalized U-statistic.

7. Complete the details of proof for Lemma 5.2.2B.

8. Extend Lemma 5.2.2B to generalized U-statistics.

Section 5.3

9. The projection of a generalized U-statistic is defined as

$$\hat{U} = \sum_{j=1}^{k} \sum_{i=1}^{n_j} E_F\{U \mid X_i^{(j)}\} - (N-1)\theta,$$

where $N = n_1 + \cdots + n_k$. Define

$$\tilde{h}_{1j}(x) = E_F\{h(X_1^{(1)}, \ldots, X_{m_1}^{(1)}; \ldots; X_1^{(k)}, \ldots, X_{m_k}^{(k)}) \mid X_1^{(j)} = x\} - \theta.$$

Show that

$$\hat{U} - \theta = \sum_{j=1}^{k} \sum_{i=1}^{n_j} \frac{m_j}{n_j} \tilde{h}_{1j}(X_i^{(j)}).$$

10. (continuation) Show that $U_n - \hat{U}_n$ is a U-statistic based on a kernel H satisfying $E_F\{H\} = E_F\{H \mid X_i^{(j)}\} = 0$.

11. Verify relation (2) in 5.3.4..

12. Extend (2), (3) and (4) of 5.3.4 to generalized U-statistics.

13. Let g_c and S_{cn} be as defined in 5.1.5. Define a kernel G_c of order m by

$$G_c(x_1, \ldots, x_m) = \sum_{1 \le i_1 < \cdots < i_c \le m} g_c(x_{i_1}, \ldots, x_{i_c})$$

and let U_{cn} be the U-statistic corresponding to G_c. Show that

$$U_{cn} = \binom{m}{c}\binom{n}{c}^{-1} S_{cn}$$

and thus

$$U_n - \theta = \sum_{c=1}^{m} U_{cn}.$$

Now suppose that $\zeta_{c-1} = 0 < \zeta_c$. Show that $\hat{U}_n$ defined in 5.3.4 satisfies

$$\hat{U}_n - \theta = U_{cn}.$$

Section 5.4

14. For $E_F h^2 < \infty$, show strong convergence of generalized U-statistics.

15. Prove Theorem 5.4C, the LIL for U-statistics. (Hint: apply Theorem 5.3.3.)

Section 5.5

16. Prove Theorem 5.5.1A, the CLT for U-statistics.

17. Complete the details for Example 5.5.1A.

18. Extend Theorem 5.5.1A to a vector of several U-statistics defined on the same sample.

19. Extend Theorem 5.5.1A to generalized U-statistics (continuation of Problems 5.P.9, 10, 12).

20. Check the details of Example 5.5.1B.

21. Check the details of Example 5.5.2A.

22. (continuation) Show, for F binomial $(1, \frac{1}{2})$, that

$$n(m_2 - \tfrac{1}{4}) \xrightarrow{d} -\tfrac{1}{4}\chi_1^2.$$

(Hint: One approach is simply to apply the result obtained in Example 5.5.2A. Another approach is to write $m_2 = \hat{p} - \hat{p}^2$ and apply the methods of Chapter 3.)

23. Check the details of Example 5.5.2B.

24. Complete the details of proof of Theorem 5.5.2.

(a) Prove (3). (Hint: write $\tilde{h}_1(x) = E_F\{\tilde{h}_2(x, X_2)\}$ and use Jensen's inequality to show that

$$\lim_{K\to\infty} E_F\left\{\left[\tilde{h}_1(X_1) - \sum_{k=1}^{K} \lambda_k \phi_k(X_1) E_F\{\phi_k(X_2)\}\right]^2\right\}$$
$$\leq \lim_{K\to\infty} E_F\left\{\left[\tilde{h}_2(X_1, X_2) - \sum_{k=1}^{K} \lambda_k \phi_k(X_1)\phi_k(X_2)\right]^2\right\} = 0.$$

(b) Prove (4).

(c) Prove (6).

Section 5.6

25. Prove Lemma 5.6.1B. (Hint: Without loss assume $E\{Y\} = 0$. Show that $e^{sY} = 1 + sY + \frac{1}{2}s^2 Z$, where $0 < Z < Y^2 e^{s_0 Y}$.)

26. Complete the proof of Lemma 5.6.1C.

27. Complete the proof of Theorem 5.6.1A.

28. Complete the proof of Theorem 5.6.1B.

29. In **5.6.2**, show that (1) follows from Theorem 5.6.1A and that (*) for $c \leq 1$ follows from Theorem 5.5.1B.

Section 5.7

30. Derive the strong convergence of U-statistics, under the assumption $E_F\{h^2\} < \infty$, from Corollary 5.7.1.

31. Check the claim of Example 5.5.2C.

32. Apply Theorem 5.7.2 to obtain properties of the confidence interval I_{Wn}.

33. Complete the details of proof of Lemma 5.7.3.