Asymptotic Distribution of Two-Sample Empirical U-Quantiles for Dependent Data

Herold Dehling
(joint work with Roland Fried (TU Dortmund) and Martin Wendler)

RUHR-UNIVERSITÄT BOCHUM

Limit Theorems for Dependent Data and Applications
Conference in honor of Professor Magda PELIGRAD
**Motivating Example: Hodges-Lehmann Estimator**

**ONE SAMPLE CASE:** $X_1, \ldots, X_n$; define the Hodges-Lehmann estimator for the location

$$\text{median}\left\{ \frac{1}{2}(X_i + X_j) : 1 \leq i < j \leq n \right\}$$

**TWO SAMPLE CASE:** $X_1, \ldots, X_{n_1}$, $Y_1, \ldots, Y_{n_2}$; define the Hodges-Lehmann estimator for the difference in location

$$\text{median}\{(X_i - Y_j) : 1 \leq i \leq n_1, 1 \leq j \leq n_2\}.$$ 

We are interested in the asymptotic distribution of such estimators in the case of dependent data.
Definition (Halmos 1946, Hoeffding 1948, von Mises 1947)

Given a process \((X_i)_{i \geq 1}\) of iid random variables with marginal distribution \(F\) and a symmetric kernel \(h : \mathbb{R}^2 \to \mathbb{R}\), we define the bivariate \(U\)- and \(V\)-statistics statistics with kernel \(h\) by

\[
U_n(h) := \frac{1}{\binom{n}{2}} \sum_{1 \leq i < j \leq n} h(X_i, X_j),
\]

\[
V_n(h) := \frac{1}{n^2} \sum_{1 \leq i, j \leq n} h(X_i, X_j).
\]

▶ \(U\)- and \(V\)-statistics are generalized means of \(h(X_i, X_j)\), \(1 \leq i < j \leq n\) (resp. \(1 \leq i, j \leq n\))

▶ Analogously one can define \(m\)-variate \(U\)- and \(V\)-statistics
Examples

- $h(x, y) = \frac{1}{2}(x - y)^2$ leads to the sample variance
  
  $$U_n(h) = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$$

- $h(x, y) = \int (1_{(-\infty,s]}(x) - F_0(s))(1_{(-\infty,s]}(y) - F_0(s))w(s)dF_0(s)$;
  
  $$V_n(h) = \int (F_n(s) - F_0(s))^2w(s)dF_0(s);$$

Cramer-von Mises test statistic for testing the hypothesis $H : F = F_0$.

- $h(x, y) = \log(\|x - y\|)$ leads to the Takens’ estimator of the correlation dimension of the distribution $F$.
  (Floris Takens (12.11.1940–20.06.2010))
Hoeffding Decomposition

The tool for the analysis of $U$-statistics:

$$\theta := \mathbb{E} h(X_1, X_2)$$

$$h_1(x) := \mathbb{E} h(x, X) - \theta$$

$$h_2(x, y) := h(x, y) - h_1(x) - h_1(y) - \theta.$$ 

We obtain the decomposition of $h$ and of the $U$-statistic

$$h(x, y) = \theta + h_1(x) + h_1(y) + h_2(x, y)$$

$$U_n(h) = \theta + \frac{2}{n} \sum_{i=1}^{n} h_1(X_i) + U_n(h_2)$$

The functions $h_1$ and $h_2$ satisfy $\int h_1(x) dF(x) = 0$ and

$$\int h_2(x, y) dF(x) = 0 \quad \text{(degeneracy)}$$
The terms in the summands on the r.h.s. are uncorrelated (!) and thus
\[ \text{Var}\left(\frac{2}{n} \sum_{i=1}^{n} h_1(X_i)\right) = \frac{4}{n} \text{Var}(h_1(X_1)) \]
\[ \text{Var}(U_n(h_2)) = \frac{1}{\binom{n}{2}} \text{Var}(h_2(X_1, X_2)). \]

- Generally, the linear term \( \frac{2}{n} \sum_{i=1}^{n} h_1(X_i) \) is dominating. Limit theorems can be obtained by using classical limit theorems for partial sums and a control of the remainder term \( U_n(h_2) \).
- Non-classical limit theory in the \textit{degenerate case}, when \( \text{Var}(h_1(X)) = 0 \).
Non-degenerate U-Statistics Limit Theorems

(1) Law of Large Numbers (Hoeffding 1961, Berk 1966)

\[ U_n(h) \to \theta \quad \text{a.s.} \]

(2) Central Limit Theorem (Hoeffding 1948)

\[ \sqrt{n}(U_n(h) - \theta) \to N(0, 4 \text{Var}(h_1(X))) \quad \text{in distribution,} \]

(3) Law of the Iterated Logarithm (Sen 1972)

\[ \limsup_{n \to \infty} \frac{\sqrt{n}}{\sqrt{2 \log \log n}} (U_n(h) - \theta) = 2\text{Var}(h_1(X)) \quad \text{a.s.} \]

The functional versions of these limit theorems also hold.
Degenerate U- Statistic Limit Theorems

Let \( h \in L_2([0, 1]^2) \) be degenerate and let \((X_i)_{i \geq 1}\) be independent \(U([0, 1])\)-distributed. Then

(1) Degenerate \(U\)-statistics CLT (Fillipova 1964)

\[
n(U_n(h) - \theta) \to \int \int h(x, y) dW_0(x) dW_0(y).
\]

where \((W_0(t))_{0 \leq t \leq 1}\) is standard Brownian bridge.

(2) Degenerate \(U\)-statistics LIL (D., Denker, Philipp 1984, D. 1989)

\[
\limsup_{n \to \infty} \frac{1}{n \log \log n} \sum_{1 \leq i < j \leq n} h(X_i, X_j) = \sup \left\{ \int \int f(x)f(y)h(x, y)\,dxdy : \int f^2(x)\,dx = 1 \right\} \quad a.s.
\]
Weakly Dependent Processes I

Definition (Absolutely regular process)

(i) Let \((\Omega, \mathcal{F}, P)\) be a probability space and let \(\mathcal{A}\) and \(\mathcal{B}\) be two sub-\(\sigma\)-fields of \(\mathcal{F}\). We then define

\[
\beta(\mathcal{A}, \mathcal{B}) := \sup \sum_{i=1}^{m} \sum_{j=1}^{n} |P(A_i \cap B_j) - P(A_i)P(B_j)|,
\]

supremum taken over all partitions of \(\Omega\) into sets \(A_1, \ldots, A_m \in \mathcal{A}\), all partitions of \(\Omega\) into sets \(B_1, \ldots, B_n \in \mathcal{B}\) and all \(m, n \geq 1\).

(ii) The process \((X_i)_{i \geq 1}\) is called absolutely regular, if for \(k \to \infty\)

\[
\beta(k) := \sup_n \beta(\mathcal{F}_1^n, \mathcal{F}_{n+k}^\infty) \to 0,
\]

where \(\mathcal{F}_k^l\) is the \(\sigma\)-field generated by \(X_k, \ldots, X_l\).
More generally, we consider functionals of absolutely regular processes, i.e. we assume that $(X_i)_{i \geq 1}$ has a representation

$$X_i = f((Z_{n+i})_{n \in \mathbb{Z}}),$$

where $(Z_n)_{n \in \mathbb{Z}}$ is an absolutely regular process and $f : \mathbb{R}^\mathbb{Z} \rightarrow \mathbb{R}$ satisfies some continuity property.

Large classes of processes can be expressed in this way, e.g.

- ARMA processes
- Many dynamical systems $X_n = T^n(X_0)$, e.g. if $T : [0, 1] \rightarrow [0, 1]$ is expanding (Hofbauer, Keller 1984).

For details and more examples, see Borovkova, Burton, D. (2001).
Theorem (Aaronson, Burton, D., Gilat, Hill, Weiss 1996)

If one of the following two conditions is satisfied,

(i) $h$ is $F \times F$ almost everywhere continuous and bounded
(ii) the process $(X_k)_{k \geq 1}$ is absolutely regular and $h$ is bounded,

the U-statistics ergodic theorem holds, i.e.

$$\frac{1}{\binom{n}{2}} \sum_{1 \leq i < j \leq n} h(X_i, X_j) \to \int \int h(x, y) dF(x) dF(y)$$

Aaronson et al. (1996) gave counterexamples in case the above conditions are not satisfied: the $U$-statistic ergodic theorem may fail for ergodic processes $(X_i)_{i \geq 1}$. 
Dependent U-Statistics CLT

Theorem

Under some technical conditions on \( h(x, y) \) and \((X_i)_{i \geq 1}\),

\[
\sqrt{n}(U_n(h) - \theta) \rightarrow N(0, 4\sigma^2),
\]

where

\[
\sigma^2 := \text{Var}(h_1(X_1)) + 2 \sum_{i=2}^{\infty} \text{Cov}(h_1(X_1), h_1(X_i))
\]

- Absolutely regular processes: Yoshihara (1976)

Results on degenerate kernels have been obtained by Babbel (1989), Kanagawa, Yoshihara (1998), Leucht, Neumann (2010).
Empirical U-Process CLT

Given a symmetric kernel $f(x, y)$, define the empirical $U$-distribution function

$$U_n(t) = \frac{1}{\binom{n}{2}} \sum_{1 \leq i < j \leq n} 1\{f(X_i, X_j) \leq t\}$$

and the empirical $U$-process $\sqrt{n}(U_n(t) - U(t))$, where $U(t) = P(f(X, Y) \leq t)$.

Theorem (Serfling 1984, Arcones, Yu 1994, Borovkova, Burton, D. 2001)

Let $(X_i)_{i \geq 1}$ be a functional of an absolutely regular process. Then under some technical conditions on $f(x, y)$ and $(X_i)_{i \geq 1}$,

$$(\sqrt{n}(U_n(t) - U(t)))_{t \geq 0} \xrightarrow{D} (W(t))_{t \geq 0},$$

where $(W(t))_{t \geq 0}$ is a mean-zero Gaussian process.
Example: The Hodges-Lehmann estimator of location

$$\text{median} \left\{ \frac{X_i + X_j}{2} : 1 \leq i < j \leq n \right\}$$

$$= \inf \left\{ t : \frac{2}{n(n-1)} \sum_{1 \leq i < j \leq n} \mathbb{1}\left\{ \frac{1}{2} (X_i + X_j) \leq t \right\} \geq \frac{1}{2} \right\}$$

is the 50% quantile of the empirical distribution $U_n(\cdot)$ of the pairwise means $\frac{1}{2}(X_i + X_j)$, $1 \leq i < j \leq n$. More general, we define the empirical (one sample) $U$-quantile

$$U^{-1}_n(p) := \inf \{ t : U_n(t) \geq p \}.$$

Herold Dehling
Theorem (Wendler, 2010)

Let \((X_i)_{i \geq 1}\) be a functional of an absolutely regular process. Then under some technical conditions, we have for any \(0 < p_1 < p_2 < 1\)

\[
\left( \sqrt{n} \left( U_n^{-1}(p) - U^{-1}(p) \right) \right)_{p \in (p_1, p_2)} \overset{D}{\rightarrow} \left( \frac{1}{U'(U^{-1}(p))} W(U^{-1}(p)) \right)_{p \in (p_1, p_2)}
\]

The functional LIL also holds.
Bahadur-Kiefer Representation

Basic tool in the treatment of the empirical $U$-quantiles is the Bahadur-Kiefer representation, i.e.

$$U_n^{-1}(p) - U^{-1}(p) = \frac{p - U_n(U^{-1}(p))}{U'(U^{-1}(p))} + R_n(p).$$

**Theorem (Wendler, 2010)**

*Under the same technical assumptions as in the previous theorem*

$$\sup_{p \in (p_1, p_2)} R_n(p) = o(n^{-\frac{23}{40}}) \quad \text{a.s.}$$
The two sample Hodges-Lehmann estimator

\[
\text{median}\{(X_i - Y_j) : 1 \leq i \leq n_1, 1 \leq j \leq n_2\}.
\]

is the 50\% quantile of the empirical distribution \(U_{n_1,n_2}(\cdot)\) of the differences \(X_i - Y_j, 1 \leq i \leq n_1, 1 \leq j \leq n_2,\)

\[
U_{n_1,n_2}(t) = \frac{1}{n_1 n_2} \#\{1 \leq i \leq n_1, 1 \leq j \leq n_2 : X_i - Y_j \leq t\}
\]

More generally, we define the two-sample empirical \(U\)-quantiles

\[
Q_{n_1,n_2}(p) = \inf\{t : U_{n_1,n_2}(t) \geq p\}, \ 0 \leq p \leq 1.
\]
Two Sample U-Process, U-Quantile Process

The empirical $U$-distribution function and $U$-quantiles,

\[
U_{n_1,n_2}(t) = \frac{1}{n_1 n_2} \# \{1 \leq i \leq n_1, 1 \leq j \leq n_2 : X_i - Y_j \leq t \}
\]

\[
Q_{n_1,n_2}(p) = \inf \{ t : U_{n_1,n_2}(t) \geq p \},
\]

are the natural estimator of the distribution function and the quantiles of $X - Y$, where $X, Y$ are independent,

\[
H(t) = P(X - Y \leq t)
\]

\[
Q(p) = \inf \{ t : H(t) \geq p \}.
\]

We will investigate the asymptotic distributions of

\[
\sqrt{n_1 + n_2}(U_{n_1,n_2}(t) - H(t))
\]

\[
\sqrt{n_1 + n_2}(Q_{n_1,n_2}(p) - Q(p)).
\]
In the standard two sample problem,

\[ X_1, \ldots, X_{n_1} \sim F \]
\[ Y_1, \ldots, Y_{n_2} \sim G \]

all observations are independent. We study two situations

1. Given are two stationary ergodic processes \((X_i)_{i \geq 1}\) and \((Y_j)_{j \geq 1}\), independent of each other.

2. Given is one stationary ergodic process \((X_i)_{i \geq 1}\) and

\[ Y_j = X_{n_1+j}, \quad 1 \leq j \leq n_2. \]

The asymptotic distributions of our statistics are the same in both cases, at least for weakly dependent observations.
The two sample empirical $U$-distribution function,

\[ U_{n_1,n_2}(t) = \frac{1}{n_1 n_2} \# \{ 1 \leq i \leq n_1, 1 \leq j \leq n_2 : X_i - Y_j \leq t \} \]

\[ = \frac{1}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} 1\{X_i - Y_j \leq t\}, \]

is a special case of a two sample $U$-statistic, defined as

\[ U_{n_1,n_2} = \frac{1}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} h(X_i, Y_j). \]

We will begin our investigations by studying the asymptotic distribution of $U_{n_1,n_2}$ as $n_1, n_2 \to \infty$. 
As in the case of independent observations, the analysis of the asymptotic behavior of $U$-statistics uses the Hoeffding decomposition. We introduce the following quantities,

$$
\begin{align*}
\theta &= Eh(X, Y) \\
\theta_1(x) &= Eh(x, Y) - \theta \\
\theta_2(y) &= Eh(X, y) - \theta \\
g(x, y) &= h(x, y) - \theta_1(x) - \theta_2(y) - \theta, \\
\end{align*}
$$

and observe that

$$
h(x, y) = \theta + \theta_1(x) + \theta_2(y) + g(x, y).
$$
The decomposition of the kernel $h(x, y)$ leads to the Hoeffding decomposition of the $U$-statistic,

$$U_{n_1,n_2} = \theta + \frac{1}{n_1} \sum_{i=1}^{n_1} h_1(X_i) + \frac{1}{n_2} \sum_{j=1}^{n_2} h_2(Y_j) + \frac{1}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} g(X_i, Y_j).$$

The functions $h_1(x)$, $h_2(y)$ have the property

$$Eh_1(X) = Eh_2(Y) = 0,$$

i.e. $\sum_{i=1}^{n_1} h_1(X_i)$ and $\sum_{i=1}^{n_2} h_2(Y_i)$ are sums of mean zero random variables. Moreover,

$$Eg(X, y) = Eg(x, Y) = 0 \quad (\text{degenerate})$$
Theorem (D., Fried (2010))

Let \( (X_i)_{i \geq 1} \) and \( (Y_i)_{i \geq 1} \) be functionals of absolutely regular processes satisfying \( \sum_{k=1}^{\infty} k \beta(k) < \infty \) and assume that \( E|h(X, Y)|^{2+\epsilon} < \infty \), for some \( \epsilon > 0 \). Then, as \( n_1, n_2 \to \infty \) so that \( \frac{n_1}{n_1 + n_2} \to \lambda \in (0, 1) \), we have

\[
\sqrt{n_1 + n_2}(U_{n_1,n_2} - \theta) \to N(0, \sigma^2),
\]

where

\[
\sigma^2 = \frac{1}{\lambda} \left( \text{Var}(h_1(X)) + 2 \sum_{i=2}^{\infty} \text{Cov}(h_1(X_1), h_1(X_i)) \right)
+ \frac{1}{1 - \lambda} \left( \text{Var}(h_2(Y)) + 2 \sum_{i=2}^{\infty} \text{Cov}(h_2(Y_1), h_2(Y_i)) \right)
\]
Lemma (D., Fried 2010)

Let \((X_i)_{i \geq 1}\) and \((Y_i)_{i \geq 1}\) be functionals of absolutely regular processes with mixing coefficients satisfying \(\sum_{k=1}^{\infty} k \beta(k) < \infty\). Then

\[
E \left( \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} g(X_i, Y_j) \right)^2 \leq C n_1 n_2
\]

where \(C\) is some constant, not depending on \(n_1\) and \(n_2\).

The proof uses generalized correlation inequalities, i.e. bounds on

\[
Ef(\xi_1, \xi_2) - Ef(\xi'_1, \xi'_2)
\]

where \(\xi'_1, \xi'_2\) are independent with the same marginal distributions as \(\xi_1, \xi_2\).
Recall the definition of the empirical $U$-distribution function and $U$-quantiles:

$$U_{n_1,n_2}(t) = \frac{1}{n_1 n_2} \#\{1 \leq i \leq n_1, 1 \leq j \leq n_2 : X_i - Y_j \leq t\}$$

$$= \frac{1}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} 1\{X_i - Y_j \leq t\}$$

$$Q_{n_1,n_2}(p) = \inf\{t : U_{n_1,n_2}(t) \geq p\},$$

together with

$$H(t) = P(X - Y \leq t)$$

$$Q(p) = \inf\{t : H(t) \geq p\}.$$
Two Sample Empirical U-Process CLT

Theorem (D., Fried 2010)

Let \((X_i)_{i \geq 1}\) and \((Y_i)_{i \geq 1}\) be functionals of absolutely regular processes satisfying \(\sum_{k=1}^{\infty} k \beta(k) < \infty\). Let \(n_1, n_2 \to \infty\) so that \(\frac{n_1}{n_1 + n_2} \to \lambda \in (0, 1)\). Then, for any \(t \in \mathbb{R}\),

\[
\sqrt{n_1 + n_2} (U_{n_1,n_2}(t) - H(t)) \to N \left( 0, \frac{\sigma_1^2(t)}{\lambda} + \frac{\sigma_2^2(t)}{1 - \lambda} \right)
\]

in distribution, where

\[
\sigma_1^2(t) = \text{Var}(G(X_1 - t)) + 2 \sum_{k=2}^{\infty} \text{Cov}(G(X_1 - t), G(X_k - t))
\]

\[
\sigma_2^2(t) = \text{Var}(F(Y_1 + t)) + 2 \sum_{k=2}^{\infty} \text{Cov}(F(Y_1 + t), F(Y_k + t))
\]
The asymptotic distribution of the empirical $U$-quantiles can be derived from that of the empirical $U$-process with the help of the Bahadur-Kiefer representation

$$Q_{n_1, n_2}(p) = Q(p) + \frac{p - U_{n_1, n_2}(Q(p))}{H'(Q(p))} + R_{n_1, n_2},$$

where $R_{n_1, n_2}$ is a "small" remainder term.

**Theorem (D., Fried 2010)**

Let $(X_i)_{i \geq 1}$ and $(Y_i)_{i \geq 1}$ be functionals of absolutely regular processes with mixing coefficients $\beta(k)$ satisfying $\sum_{k=1}^{\infty} k \beta(k) < \infty$. Then for any $0 < p < 1$ we have

$$Q_{n_1, n_2}(p) = Q(p) + \frac{p - U_{n_1, n_2}(Q(p))}{H'(Q(p))} + R_{n_1, n_2}$$

where $R_{n_1, n_2} = o_P\left(\frac{1}{\sqrt{n_1 + n_2}}\right)$. 
Theorem (D., Fried 2010)

Let \((X_i)_{i \geq 1}\) and \((Y_i)_{i \geq 1}\) be stationary, absolutely regular processes satisfying \(\sum_{k=1}^{\infty} k \beta(k) < \infty\). Let \(n_1, n_2 \to \infty\) so that \(\frac{n_1}{n_1 + n_2} \to \lambda \in (0, 1)\). Then

\[
\sqrt{n_1 + n_2} (Q_{n_1, n_2}(p) - Q(p)) \quad \to \quad N \left( 0, \frac{1}{(H'(Q(p))))^2 \left( \frac{\sigma_1^2(Q(p))}{\lambda} + \frac{\sigma_2^2(Q(p))}{1 - \lambda} \right) \right),
\]

where \(\sigma_1^2(Q(p))\) and \(\sigma_2^2(Q(p))\) are defined as above.

2. Study of the process

$$\sum_{i=1}^{[\lambda n]} \sum_{j=[\lambda n]+1}^{n} 1\{X_i - X_j \leq t\}, \quad 0 \leq \lambda \leq 1,$$

as well as the associated $U$-quantile process.

3. Application to robust change-point tests with dependent data.
Selected Articles

**Herold Dehling and Roland Fried:** Robust estimation for two sample problems with dependent data. *Work in progress*

**Martin Wendler:** Bahadur representation for $U$-quantiles of dependent data. *Preprint*

**Herold Dehling and Aneas Rooch:** Two sample $U$-statistics for long-range dependent data. *Work in progress*