Variations of U-statistics with applications to random strings and random permutations

Svante Janson

28th Nordic Congress of Mathematicians, Helsinki
18 August, 2022

U-statistics

A (standard) *U-statistic* is a sum

$$U_n = U_n(f) = \sum_{i_1 < \cdots < i_m} f\left(X_{i_1}, \ldots, X_{i_m}\right)$$

where $X_1, \ldots, X_n$ is an i.i.d. sequence of random variables, and f is a measurable function of $m \geq 1$ variables.

X_i may take values in any measurable space. For example, X_i may be real-valued or vectors.

Traditionally (Hoeffding, 1948), f is supposed to be symmetric (equivalently, the sum is taken over all distinct $i_1, \ldots, i_m$). This is the case in the original statistical applications (e.g., Kendall's τ) but in my applications, I usually need the asymmetric version above.

Variations will come later.

Remark

The asymmetric case can be reduced to the symmetric as follows: Let $Y_1, \ldots, Y_n$ be uniform random variables on $[0,1]$, independent of (X_i) and each other, and define $Z_i := (X_i, Y_i)$. Let

$$F(Z_1, ..., Z_m) := \sum_{\pi \in \mathfrak{S}_m} f(X_{\pi(1)}, \ldots, X_{\pi(m)}) \mathbf{1}\{Y_{\pi(1)} < \cdots < Y_{\pi(m)}\}$$

Then $U_n(F)$ is a symmetric U-statistic, and

$$U_n(F) \stackrel{\mathrm{d}}{=} U_n(f).$$

Remark

The asymmetric case can be reduced to the symmetric as follows: Let $Y_1, \ldots, Y_n$ be uniform random variables on $[0,1]$, independent of (X_i) and each other, and define $Z_i := (X_i, Y_i)$. Let

$$F(Z_1, ..., Z_m) := \sum_{\pi \in \mathfrak{S}_m} f(X_{\pi(1)}, \ldots, X_{\pi(m)}) \mathbf{1}\{Y_{\pi(1)} < \cdots < Y_{\pi(m)}\}$$

Then $U_n(F)$ is a symmetric U-statistic, and

$$U_n(F) \stackrel{\mathrm{d}}{=} U_n(f).$$

This does not work in the extensions below.

Theorem (Hoeffding, 1948)

Let $\mathbb{E}\,|f(X_1,\ldots,X_m)|^2 < \infty$. *Then*

$$\frac{U_n - \binom{n}{m}\mu}{n^{m-1/2}} \xrightarrow{\mathrm{d}} N(0,\sigma^2),$$

where

$$\mu = \mathbb{E}\, f(X_1,\ldots,X_m)$$

and

$$\sigma^2 \geq 0.$$

(Explicit formula, but omitted today.)

Theorem (Hoeffding, 1948)

Let $\mathbb{E}\,|f(X_1,\dots,X_m)|^2 < \infty$*. Then*

$$\frac{U_n - \binom{n}{m}\mu}{n^{m-1/2}} \xrightarrow{\mathrm{d}} N(0,\sigma^2),$$

where

$$\mu = \mathbb{E}\,f(X_1,\dots,X_m)$$

and

$$\sigma^2 \geq 0.$$

(Explicit formula, but omitted today.)

Also joint normal limits for several f. (Cramér-Wold device.)

Degenerate cases

If $\sigma^2 = 0$, then we get non-normal limits with another normalization. Typically an infinite sum of squares of normal variables. (Higher degeneracies lead to higher-degree polynomials.)

Degenerate cases

If $\sigma^2 = 0$, then we get non-normal limits with another normalization. Typically an infinite sum of squares of normal variables. (Higher degeneracies lead to higher-degree polynomials.)

Sometimes $\sigma^2 = 0$ follows from some symmetry property.

In most applications, $\sigma^2 > 0$ (and thus asymptotic normality), but this can be surprisingly difficult to show.

Hoeffding's proof

Hoeffding's proof is based on a projection method:
Assume $\mathbb{E}\, f(X_1, \ldots, X_m) = 0$. Define

$$f_i(X_i) = \mathbb{E}\big[f(X_1, \ldots, X_m) \mid X_i\big].$$

Approximate $f(X_1, \ldots, X_m)$ by $f_1(X_1) + \cdots + f_m(X_m)$. The resulting sum is asymptotically normal by the standard central limit theorem. (Use triangular arrays in the asymmetric case.)
The error has small variance and can be ignored.
QED

Corollary of proof.

$$\sigma^2 = 0 \iff f_i(X_i) = 0 \quad \text{a.s.} \quad \text{for every } i = 1, \ldots, m.$$

Application: patterns in random words

Consider a random string $\Xi_n = \xi_1 \cdots \xi_n$ consisting of n i.i.d. random letters from a finite alphabet $\mathcal{A}$, and consider the number of occurences of a given word $\mathbf{w} = w_1 \cdots w_\ell$ as a *subsequence*; to be precise, an *occurrence* of $\mathbf{w}$ in Ξ_n is an increasing sequence of indices $i_1 < \cdots < i_\ell$ in $[n] = \{1, \ldots, n\}$ such that

$$\xi_{i_1}\xi_{i_2}\cdots\xi_{i_\ell} = \mathbf{w}, \qquad \text{i.e., } \xi_{i_k} = w_k \text{ for every } k \in [\ell]. \tag{1}$$

Flajolet, Szpankowski and Vallée (2006) proved (by different methods) that $N_n(\mathbf{w})$ is asymptotically normal as $n \to \infty$.

Application: patterns in random words

Consider a random string $\Xi_n = \xi_1 \cdots \xi_n$ consisting of n i.i.d. random letters from a finite alphabet $\mathcal{A}$, and consider the number of occurences of a given word $\mathbf{w} = w_1 \cdots w_\ell$ as a *subsequence*; to be precise, an *occurrence* of $\mathbf{w}$ in Ξ_n is an increasing sequence of indices $i_1 < \cdots < i_\ell$ in $[n] = \{1, \ldots, n\}$ such that

$$\xi_{i_1}\xi_{i_2}\cdots\xi_{i_\ell} = \mathbf{w}, \qquad \text{i.e., } \xi_{i_k} = w_k \text{ for every } k \in [\ell]. \tag{1}$$

Flajolet, Szpankowski and Vallée (2006) proved (by different methods) that $N_n(\mathbf{w})$ is asymptotically normal as $n \to \infty$.

We have

$$N_n(\mathbf{w}) = \sum_{i_1 < \cdots < i_m} f\left(\xi_{i_1}, \ldots, \xi_{i_m}\right) = U_n(f).$$

for an indicator function f. The result thus follows from Hoeffding's theorem.

Application: patterns in a permutation

Let $\mathfrak{S}_n$ be the set of permutations of $[n] := \{1, \ldots, n\}$.

If $\tau = \tau_1 \cdots \tau_k \in \mathfrak{S}_k$ and $\pi = \pi_1 \cdots \pi_n \in \mathfrak{S}_n$, then an *occurrence* of τ in π is a subsequence $\pi_{i_1} \cdots \pi_{i_k}$, with $1 \leq i_1 < \cdots < i_k \leq n$, that has the same relative order as τ. τ is called a *pattern*.

Example: 31425 is an occurence of 213 in 31425

Let $\mathrm{occ}_\tau(\pi)$ be the number of occurrences of τ in π.
For example, $\mathrm{occ}_{21}(\pi)$ is the number of inversions in π.
(Kendall's τ, again.)

Application: patterns in a permutation

Let $\mathfrak{S}_n$ be the set of permutations of $[n] := \{1, \dots, n\}$.

If $\tau = \tau_1 \cdots \tau_k \in \mathfrak{S}_k$ and $\pi = \pi_1 \cdots \pi_n \in \mathfrak{S}_n$, then an *occurrence* of τ in π is a subsequence $\pi_{i_1} \cdots \pi_{i_k}$, with $1 \le i_1 < \cdots < i_k \le n$, that has the same relative order as τ. τ is called a *pattern*.

Example: 31425 is an occurence of 213 in 31425

Let $\operatorname{occ}_\tau(\pi)$ be the number of occurrences of τ in π.
For example, $\operatorname{occ}_{21}(\pi)$ is the number of inversions in π.
(Kendall's τ, again.)

Let $\boldsymbol{\pi} = \boldsymbol{\pi}_n$ be a random permutation of length n, drawn uniformly from all permutations in $\mathfrak{S}_n$.

Bóna (2007) proved that for any fixed τ, the number of occurrences $\operatorname{occ}_\tau(\boldsymbol{\pi}_n)$ is asymptotically normal.

We can generate $\boldsymbol{\pi}_n$ by taking a sequence $(X_i)_1^n$ of i.i.d. random variables with a uniform distribution $X_i \sim U(0,1)$, and then replacing the values $X_1, \ldots, X_n$, in increasing order, by $1, \ldots, n$.

We can generate $\boldsymbol{\pi}_n$ by taking a sequence $(X_i)_1^n$ of i.i.d. random variables with a uniform distribution $X_i \sim U(0,1)$, and then replacing the values $X_1, \dots, X_n$, in increasing order, by $1, \dots, n$.

Then, the number $N_n(\tau)$ of occurrences of a fixed permutation $\tau = \tau_1 \cdots \tau_\ell$ in $\boldsymbol{\pi}_n$ is given by the U-statistic $U_n(f)$ with

$$f(x_1, \dots, x_\ell) := \prod_{1 \le i < j \le \ell} \mathbf{1}\{x_i < x_j \iff \tau_i < \tau_j\}. \qquad (2)$$

Thus, Bóna's theorem follows from Hoeffding's.
(Janson, Nakamura and Zeilberger 2015)

Degenerate cases, again

The U-statistics in the applications above are non-degenerate, except in trivial cases. Some linear combinations are degenerate.

Example

$\mathrm{occ}_{123}(\boldsymbol{\pi})+\mathrm{occ}_{231}(\boldsymbol{\pi})+\mathrm{occ}_{312}(\boldsymbol{\pi})-\mathrm{occ}_{132}(\boldsymbol{\pi})-\mathrm{occ}_{213}(\boldsymbol{\pi})-\mathrm{occ}_{321}(\boldsymbol{\pi})$.

(Fisher and Lee, Nonparametric measures of angular-angular association, 1982)

In fact, the space of non-trivial linear combinations of $\mathrm{occ}_\tau(\boldsymbol{\pi})$, $\tau \in \mathfrak{S}_k$, has dimension $k! - 1$. The space of normal limits has dimension $(k-1)^2$, so the space of degenerate linear combinations has dimension $k! - 1 - (k-1)^2$. See further Even-Zohar (2020) and Even-Zohar, Lakrec and Tessler (2021) (random words).

Variations: vincular and constrained patterns

A *vincular pattern* in a permutation is a pattern where some entries are marked, and we only count occurrences where a marked entry is adjacent to the next one.

Example

The vincular pattern 2^*13 counts triples $(i, i+1, j)$ with $i+1 < j$ and $\pi_{i+1} < \pi_i < \pi_j$.

The number of occurences is asymptotically normal (Hofer, 2016).

In particular, marking every element means that we count only substrings (consecutive patterns) $\pi_i \pi_{i+1} \cdots \pi_{i+m-1}$ that have the right order. (Bóna 2010).

Variations: vincular and constrained patterns

A *vincular pattern* in a permutation is a pattern where some entries are marked, and we only count occurrences where a marked entry is adjacent to the next one.

Example

The vincular pattern 2^*13 counts triples $(i, i+1, j)$ with $i+1 < j$ and $\pi_{i+1} < \pi_i < \pi_j$.

The number of occurences is asymptotically normal (Hofer, 2016).

In particular, marking every element means that we count only substrings (consecutive patterns) $\pi_i \pi_{i+1} \cdots \pi_{i+m-1}$ that have the right order. (Bóna 2010).

More general constraints: gaps at most d, or exactly d. Such constraints were studied for patterns in random words by Flajolet et al (2006).

Vincular and more general constrained patterns correspond to constrained U-statistics, where the sum is restricted to certain m-tuples.

Remark. This is an instance of the large class of *incomplete U-statistics* introduced by Blom (1976).

Theorem

Hoeffding's theorem extends to constrained U-statistics. I.e., they are asymptotically normal.

Proof by example.

In the example 2*13 of a vincular pattern above, let again $\boldsymbol{\pi} \in \mathfrak{S}_n$ be constructed from i.i.d. $(X_i)_1^n$. Define $Y_i := (X_i, X_{i+1}) \in \mathbb{R}^2$. Then

$$\operatorname{occ}_\tau(\boldsymbol{\pi}) = \sum_{i,j:i+1<j} f(Y_i, Y_j)$$

for a suitable f. This is, up to a negligible error (viz., terms with $j = i + 1$), a U-statistic of order 2 based on (Y_i).

□

Proof by example.

In the example 2^*13 of a vincular pattern above, let again $\boldsymbol{\pi} \in \mathfrak{S}_n$ be constructed from i.i.d. $(X_i)_1^n$. Define $Y_i := (X_i, X_{i+1}) \in \mathbb{R}^2$. Then

$$\mathrm{occ}_\tau(\boldsymbol{\pi}) = \sum_{i,j:i+1<j} f(Y_i, Y_j)$$

for a suitable f. This is, up to a negligible error (viz., terms with $j = i + 1$), a U-statistic of order 2 based on (Y_i).
However, the sequence (Y_i) is not i.i.d. !

□

Proof by example.

In the example 2*13 of a vincular pattern above, let again $\boldsymbol{\pi} \in \mathfrak{S}_n$ be constructed from i.i.d. $(X_i)_1^n$. Define $Y_i := (X_i, X_{i+1}) \in \mathbb{R}^2$. Then

$$\mathrm{occ}_\tau(\boldsymbol{\pi}) = \sum_{i,j:i+1<j} f(Y_i, Y_j)$$

for a suitable f. This is, up to a negligible error (viz., terms with $j = i+1$), a U-statistic of order 2 based on (Y_i).
However, the sequence (Y_i) is not i.i.d. !

No problem!

□

Proof by example.

In the example 2^*13 of a vincular pattern above, let again $\boldsymbol{\pi} \in \mathfrak{S}_n$ be constructed from i.i.d. $(X_i)_1^n$. Define $Y_i := (X_i, X_{i+1}) \in \mathbb{R}^2$. Then

$$\operatorname{occ}_\tau(\boldsymbol{\pi}) = \sum_{i,j:i+1<j} f(Y_i, Y_j)$$

for a suitable f. This is, up to a negligible error (viz., terms with $j = i + 1$), a U-statistic of order 2 based on (Y_i).
However, the sequence (Y_i) is not i.i.d. !

No problem!
The sequence is 1-dependent, and this is enough for the central limit theorem (Orey, 1958), and Hoeffding's proof can be modified. (Janson, 2022+)
(In general m-dependence is enough.) □

Degenerate cases

New possibilities for degeneracy with vincular patterns.

Example

$$\mathrm{occ}_{1*3*2}(\pi) + \mathrm{occ}_{2*3*1}(\pi) - \mathrm{occ}_{2*1*3}(\pi) - \mathrm{occ}_{3*1*2}(\pi) \in \{0, \pm 1\}.$$

The possibilities are not completely explored!

Degenerate cases

New possibilities for degeneracy with vincular patterns.

Example

$$\mathrm{occ}_{1*3*2}(\pi) + \mathrm{occ}_{2*3*1}(\pi) - \mathrm{occ}_{2*1*3}(\pi) - \mathrm{occ}_{3*1*2}(\pi) \in \{0, \pm 1\}.$$

The possibilities are not completely explored!

For a single constrained pattern count in a random word or permutation, the U-statistic is not degenerate except in trivial cases. This is not trivial to show. (I have a general theorem that can be used; Janson 2022+)

Other permutation classes

Let $\mathfrak{S}' \subset \mathfrak{S}$ be a class of permutations and consider $\mathrm{occ}_\tau(\boldsymbol{\pi}_n)$ where now $\boldsymbol{\pi}_n$ is uniformly random in $\mathfrak{S}'_n := \mathfrak{S}' \cap \mathfrak{S}_n$.

One important case: Let $\mathfrak{S}' := \mathfrak{S}(\tau_1, \dots, \tau_k)$, the set of permutations in $\mathfrak{S}$ that avoid $\tau_1, \dots, \tau_k$, i.e., $\mathrm{occ}_{\tau_i}(\pi) = 0$ for every τ_i.

Other permutation classes

Let $\mathfrak{S}' \subset \mathfrak{S}$ be a class of permutations and consider $\mathrm{occ}_\tau(\boldsymbol{\pi}_n)$ where now $\boldsymbol{\pi}_n$ is uniformly random in $\mathfrak{S}'_n := \mathfrak{S}' \cap \mathfrak{S}_n$.

One important case: Let $\mathfrak{S}' := \mathfrak{S}(\tau_1, \ldots, \tau_k)$, the set of permutations in $\mathfrak{S}$ that avoid $\tau_1, \ldots, \tau_k$, i.e., $\mathrm{occ}_{\tau_i}(\pi) = 0$ for every τ_i.

In some cases, it is possible to find an encoding of the permutations in the class $\mathfrak{S}'$ such that the number of occurrences of a pattern τ can be written as a U-statistic.

Possible only for some permutation classes!

Block decompositions of permutations

If $\tau \in \mathfrak{S}_m$ and $\tau \in \mathfrak{S}_n$, their (direct) *sum* $\tau \oplus \tau \in \mathfrak{S}_{m+n}$ is defined by letting τ act on $[m+1, m+n]$ in the natural way; more formally, $\tau \oplus \tau = \pi \in \mathfrak{S}_{m+n}$ where $\pi_i = \tau_i$ for $1 \leq i \leq m$, and $\pi_{j+m} = \tau_j + m$ for $1 \leq j \leq n$.

A permutation $\pi \in \mathfrak{S}_*$ is *decomposable* if $\pi = \tau \oplus \tau$ for some $\tau, \tau \in \mathfrak{S}_*$, and *indecomposable* otherwise; we also call an indecomposable permutation a *block*.

It is easy to see that any permutation $\pi \in \mathfrak{S}_*$ has a unique decomposition $\pi = \pi_1 \oplus \cdots \oplus \pi_\ell$ into indecomposable permutations (blocks) $\pi_1, \ldots, \pi_\ell$; we call these the *blocks of* π.

Example: $\{231, 312\}$-avoiding permutations

Theorem

Let $\tau \in \mathfrak{S}(231, 312)$ *have b blocks. Then, for a random* $\boldsymbol{\pi}_n \in \mathfrak{S}_n(231, 312)$,

$$\frac{\mathrm{occ}_\tau(\boldsymbol{\pi}_n) - n^b/b!}{n^{b-1/2}} \xrightarrow{\mathrm{d}} N(0, \gamma^2)$$

for some constant γ^2.

Example The number of inversions.

$$\frac{\mathrm{occ}_{21}(\boldsymbol{\pi}_n) - n}{n^{1/2}} \xrightarrow{\mathrm{d}} N(0, 6).$$

Proof.

- $\pi \in \mathfrak{S}(231, 312) \iff$ each block is decreasing: $\ell(\ell-1)\cdots 21$ [Simion and Schmidt, 1995].
- If the block lengths of $\boldsymbol{\pi}_n$ are $\ell_1, \ldots, \ell_m$, and the block lengths of τ are $s_1, \ldots, s_b$, then

$$\mathrm{occ}_\tau(\boldsymbol{\pi}_n) = \sum_{i_1 < \cdots < i_b} \prod_{j=1}^{b} \binom{\ell_{i_j}}{s_j}.$$

- If the block lengths of $\boldsymbol{\pi}_n$ are $\ell_1, \ldots, \ell_m$, then $\sum_i \ell_i = n$, and $(\ell_1, \ldots, \ell_m)$ is a uniformly random composition of n. Thus, the block lengths $\ell_1, \ldots, \ell_m$ can be realized as the first elements, up to sum n, of an i.i.d. sequence $L_1, L_2, \ldots$ of random variables with a Geometric $\mathrm{Ge}(1/2)$ distribution. I.e., define $N(n) := \max\{k : \sum_1^k L_i \geq n\}$. Then the block lengths can be taken as $(L_1, \ldots, L_{N(n)})$ (with the last term truncated if necessary).

- Hence, up to a negligble error (from the last block),

$$\mathrm{occ}_\tau(\pi_n) = \sum_{1 \le i_1 < \cdots < i_b \le N(n)} \prod_{j=1}^{b} \binom{L_{i_j}}{s_j}.$$

This is a U-statistic, based on the i.i.d. sequence (L_i).

But the sum is up to the random $N(n)$ and not to a fixed n.

- Hence, up to a negligble error (from the last block),

$$\mathrm{occ}_\tau(\pi_n) = \sum_{1 \le i_1 < \cdots < i_b \le N(n)} \prod_{j=1}^{b} \binom{L_{i_j}}{s_j}.$$

  This is a U-statistic, based on the i.i.d. sequence (L_i).

  But the sum is up to the random $N(n)$ and not to a fixed n.
- No problem!

- Hence, up to a negligble error (from the last block),

$$\mathrm{occ}_\tau(\pi_n) = \sum_{1 \le i_1 < \cdots < i_b \le N(n)} \prod_{j=1}^{b} \binom{L_{i_j}}{s_j}.$$

  This is a U-statistic, based on the i.i.d. sequence (L_i).

  But the sum is up to the random $N(n)$ and not to a fixed n.

- No problem!
  Renewal theory shows that Hoeffding's proof can be adapted.
  (Janson, 2018)

Example: $\{231, 312, 321\}$-avoiding permutations

Theorem

Let $\tau \in \mathfrak{S}(231, 312, 321)$ have b blocks. Then, for a random $\boldsymbol{\pi}_n \in \mathfrak{S}_n(231, 312, 321)$,

$$\frac{\mathrm{occ}_\tau(\boldsymbol{\pi}_n) - \mu n^b}{n^{b-1/2}} \xrightarrow{\mathrm{d}} N(0, \gamma^2)$$

for some constants μ, γ.

Example The number of inversions. $\tau = 21$. $b = 1$. A calculation yields $\mu = (3 - \sqrt{5})/2$ and $\gamma^2 = 5^{-3/2}$.

$$\frac{\mathrm{occ}_{21}(\boldsymbol{\pi}_n) - \frac{3-\sqrt{5}}{2} n}{n^{1/2}} \xrightarrow{\mathrm{d}} N\left(0, 5^{-3/2}\right).$$

Proof.

- $\pi \in \mathfrak{S}(231, 312, 321) \iff$ each block is of the type 1 or 21. [Simion and Schmidt, 1995].
- Thus π is determined by its sequence of block lengths $\ell_1, \ldots, \ell_m$ with $\ell_i \in \{1, 2\}$ and $\sum_i \ell_i = n$.
- Let $p := (\sqrt{5} - 1)/2$, the golden ratio, so that $p + p^2 = 1$. Let $X_1, X_2, \ldots$ be an i.i.d. sequence of random variables with

$$\mathbb{P}(X_i = 1) = p, \qquad \mathbb{P}(X_i = 2) = p^2.$$

  Let $S_k := \sum_{i=1}^{k} X_i$ and $N(n) := \min\{k : S_k \geq n\}$. Then, the sequence $\ell_1, \ldots, \ell_B$ of block lengths of a uniformly random permutation $\boldsymbol{\pi}_n \in \mathfrak{S}(231, 312, 321)$ has the same distribution as $(X_1, \ldots, X_{N(n)})$ conditioned on $S_{N(n)} = n$.
  Consequently, $\mathrm{occ}_\tau(\boldsymbol{\pi}_n)$ can be expressed as a U-statistic based on $X_1, \ldots, X_{N(n)}$, conditioned as above.

- This is almost as in the preceding case.

  But in this case, we also condition on the event $S_{N(n)} = n$, i.e., that some sum S_k exactly equals n.

- This is almost as in the preceding case.

  But in this case, we also condition on the event $S_{N(n)} = n$, i.e., that some sum S_k exactly equals n.

- No problem!

- This is almost as in the preceding case.

  But in this case, we also condition on the event $S_{N(n)} = n$, i.e., that some sum S_k exactly equals n.

- No problem!

  More renewal theory shows that Hoeffding's proof can be adapted to this case too. (Janson, 2018)

Example: Forest permutations = $\{321, 3412\}$-avoiding

If π is a permutation of $[n]$, then its *permutation graph* G_π is the graph with an edge ij for each inversion (i, j) in π.

Acan and Hitczenko (2016) define π to be a *tree permutation* [*forest permutation*] if G_π is a tree [forest].

$$\{\text{forest permutations}\} = \mathfrak{S}(321, 3412).$$

A permutation is a forest permutation $\iff$ every block is a tree permutation.

Define a random tree permutation (of random length) $\boldsymbol{\tau}$ such that, for every tree permutation τ,

$$\mathbb{P}(\boldsymbol{\tau} = \tau) = p^{|\tau|},$$

with $p = (3 - \sqrt{5})/2$ chosen such that $\sum_{\tau} \mathbb{P}(\boldsymbol{\tau} = \tau) = 1$.

Let $\boldsymbol{\tau}_1, \boldsymbol{\tau}_2, \dots,$ be i.i.d. random tree permutations with this distribution. Let $S_k := \sum_{i=1}^{k} |\boldsymbol{\tau}_k|$, the total length of the k first, and let $N(n) := \min\{k : S_k \geq n\}$. Then, conditioned on $S_{N(n)} = n$, the sum $\boldsymbol{\pi} := \boldsymbol{\tau}_1 \oplus \cdots \oplus \boldsymbol{\tau}_{N(n)}$ is a uniformly distributed forest permutation of length n.

Let $\tau = \tau_1 \oplus \ldots, \oplus \tau_b$ be a forest permutation, decomposed into tree permutations τ_i. Then, up to a small error,

$$\mathrm{occ}_{\tau}(\boldsymbol{\pi}) = \sum_{i_1 < \cdots < i_b} \prod_{j=1}^{b} \mathrm{occ}_{\tau_j}(\boldsymbol{\tau}_{i_j}).$$

This is a U-statistic based on the i.i.d. sequence $(\boldsymbol{\tau}_i)$.

Theorem

For a random forest permutation

$$\frac{\mathrm{occ}_{\tau}(\boldsymbol{\pi}_n) - \mu n^b}{n^{b-1/2}} \xrightarrow{\mathrm{d}} N(0, \gamma^2)$$

for some constants μ, γ.

Proof.

Hoeffding's theorem, with renewal theory modifications as above. □

Example: Random tree permutations

Theorem

For a random tree permutation $\boldsymbol{\pi}_n$ of length n, and a tree permutation τ,

$$\frac{\mathrm{occ}_\tau(\boldsymbol{\pi}_n) - \mu n^b}{n^{b-1/2}} \xrightarrow{\mathrm{d}} N(0, \gamma^2)$$

for some $b \geq 1$ and constants μ, γ.

Proof. ∎

Example: Random tree permutations

Theorem

For a random tree permutation $\boldsymbol{\pi}_n$ of length n, and a tree permutation τ,

$$\frac{\mathrm{occ}_\tau(\boldsymbol{\pi}_n) - \mu n^b}{n^{b-1/2}} \xrightarrow{\mathrm{d}} N(0, \gamma^2)$$

for some $b \geq 1$ and constants μ, γ.

Proof.

Uses a coding of tree permutations by a sequence of runs of 0's or 1's, which again permits $\mathrm{occ}_\tau(\boldsymbol{\pi}_n)$ to be written as a U-statistic.
This time we have to take a vincular U-statistic, and also use renewal theory as above.
Hence the two variations of U-statistics are combined.
Hoeffding's proof can still be adapted. (Janson, 2022+) □

References

Svante Janson: *Gaussian Hilbert Spaces* (Chapter 11), Cambridge Univ. Press, Cambridge, UK, 1997.

Svante Janson: Renewal theory for asymmetric U-statistics. *Electron. J. Probab.* **23** (2018), Paper No. 129, 27 pp.

Svante Janson: Asymptotic normality for m-dependent and constrained U-statistics, with applications to pattern matching in random strings and permutations. *Advances in Applied Probability*, to appear.

Svante Janson: The number of occurrences of patterns in a random tree or forest permutation. arXiv:2203.04182

Svante Janson, Brian Nakamura & Doron Zeilberger: On the asymptotic statistics of the number of occurrences of multiple permutation patterns. *Journal of Combinatorics* **6** (2015), no. 1-2, 117–143.

And references to other authors in the references above.