LIETUVOS MATEMATIKOS RINKINYS
Proc. of the Lithuanian Mathematical Society, Ser. A
Vol. 61, 2020, pages 7–12
https://doi.org/10.15388/LMR.2020.22469

# Effective bounds of the variance of statistics on multisets of necklaces

**Arvydas Karbonskis, Eugenijus Manstavičius**

*Institute of Mathematics, Vilnius University*
Naugarduko 24, LT-03225 Vilnius
E-mail: arvydas.karbonskis@mif.stud.vu.lt

Received October 28, 2020; published online February 18, 2021

**Abstract.** The variance of a linear statistics on multisets of necklaces is explored. The upper and lower bounds with optimal constants are obtained.

**Keywords:** Turán–Kubilius inequality; polynomials over finite field; additive function

**AMS Subject Classification:** 60C05

## 1 Introduction and results

Let $(\mathcal{P}, \|\cdot\|)$ be an initial set of weighted objects and

$$\pi(j) := \big|\{p \in \mathcal{P} :\ \|p\| = j\}\big| < \infty$$

for every $j = 1, 2, \ldots$ . Examine the set $\mathcal{G}$ with the extended weight function $\|\cdot\|$ of multisets comprised of $p \in \mathcal{P}$. Namely, $a \in \mathcal{G}$ if $a = \{p_1, \ldots, p_r\}$ and $\|a\| = \|p_1\| + \cdots + \|p_r\|$ including the empty multiset $\emptyset$ of weight 0. Then

$$m(n) := |\mathcal{G}_n| := \big|\{a \in \mathcal{G} :\ \|a\| = n\}\big| = \sum_{\ell(\bar{k})=n} \prod_{j=1}^{n} \binom{\pi(j) + k_j - 1}{k_j},$$

where $\ell(\bar{k}) = 1k_1 + \cdots + nk_n$ if $\bar{k} = (k_1, \ldots, k_n) \in \mathbb{N}_0^n$ and $n \in \mathbb{N}_0 := \mathbb{N} \cup \{0\}$.

In the present paper, we deal with the multisets for which $m(n) = q^n$, where $q \geqslant 2$ is an arbitrary natural number. If $q$ is a prime power, then $\mathcal{G}$ may be interpreted as $\mathbb{F}_q^*[t]$, the set of monic polynomials over a finite field $\mathbb{F}_q$. Then $\mathcal{P}$ is the subset of irreducible polynomials. For an arbitrary such $q$, there exist combinatorial constructions, called *multisets of necklaces* satisfying $m(n) = q^n$ (see, [1, Example 2.12, p. 43]). For multisets, we have the following relations

$$\pi(n) = \frac{1}{n}\sum_{d|n} q^{n/d}\mu(d), \quad q^n = \sum_{d|n} d\pi(d), \tag{1}$$

where in the summations, $d$ runs over natural divisors of $n$ and $\mu(d)$ stands for the Möbius function. The equalities are equivalent to the formal power series relation

$$\sum_{n=0}^{\infty} q^n x^n = \frac{1}{1-qx} = \prod_{j=1}^{\infty} \left(1 - x^j\right)^{-\pi(j)}.$$

Take an $a \in \mathcal{G}_n$ uniformly at random, that is, sample it with probability $\nu_n(\{a\}) = q^{-n}, \quad n \in \mathbb{N}$ and $\nu_0(\{\emptyset\}) = 1$. If $k_j(a) \geqslant 0$ is the number of elements $p_i$ in $a \in \mathcal{G}_n$ of weight $j$, then $\bar{k}(a) = \big(k_1(a), \dots, k_n(a)\big)$ is the structure vector of $a \in \mathcal{G}_n$ satisfying $\ell(\bar{k}(a)) = n$. Its distribution is

$$\nu_n\big(\bar{k}(a) = \bar{s}\big) = \mathbf{1}\big\{\ell(\bar{s}) = n\big\} q^{-n} \prod_{j=1}^{n} \binom{\pi(j) + s_j - 1}{s_j}, \tag{2}$$

where $\bar{s} = (s_1, \dots, s_n) \in \mathbb{N}_0^n$ and $\mathbf{1}\{\cdot\}$ stands for the indicator function.

We are interested in the distribution with respect to $\nu_n$ of the linear statistics

$$h(\bar{c}) := h(\bar{c}, a) = c_1 k_1(a) + \dots + c_n k_n(a), \quad \bar{c} = (c_1, \dots, c_n) \in \mathbb{R}^n. \tag{3}$$

The number of components in $a$ is such a function, namely, it equals $k_1(a) + \dots + k_n(a)$. We refer to [1] for more sophisticated examples.

The present paper is devoted to the variance of $h(\bar{c})$ which is a sum of dependent random variables (r.vs) as the relation $h(\bar{j}, a) = \ell(\bar{k}(a)) = n$ for each $a \in \mathcal{G}_n$ shows. Estimating it, we propose an approach to overcome technical obstacles stemming from dependence.

In the sequel, the expectations and variances with respect to $\nu_n$ will be denoted by $\mathbf{E}_n$ and $\mathbf{V}_n$ while, when the probability space $(\Omega, \mathcal{F}, P)$ is not specified, we will respectively use the notation $\mathbf{E}$ and $\mathbf{V}$. The summation indexes $i, j, l, k, m, m_1$ and $m_2$ will be natural numbers.

**Theorem 1** *If $\bar{c} \in \mathbb{R}^n$ and $n \in \mathbb{N}_0$, then*

$$\mathbf{V}_n h(\bar{c}) = \sum_{1 \leqslant jk \leqslant n} c_j^2 \pi(j) k q^{-jk} - \sum_{\substack{il+jk>n \\ il \leqslant n, jk \leqslant n}} c_i c_j \pi(i)\pi(j) q^{-il-jk}. \tag{4}$$

The sketch of the proof is given at the beginning of Section 2.

It is known [1] that, for a fixed $j$, the r.v. $k_j(a)$ converges in distribution to the r.v. $\gamma_j$ distributed according the negative binomial law $NB(\pi(j), q^{-j})$. If $\{\gamma_1, \gamma_2, \dots\}$ are mutually independent, define the statistics $Y_n = c_1\gamma_1 + \dots + n\gamma_n$. We shall see that the first sum on the right-hand side in (4) is close to $\mathbf{V}Y_n$; therefore, estimating $\mathbf{V}_n h(\bar{c})$, we use the following quadratic forms:

$$B_n(\bar{c}) := \sum_{1 \leqslant jk \leqslant n} c_j^2 \pi(j) k q^{-jk}, \qquad R_n(\bar{c}) = \sum_{m \leqslant n} m q^{-2m} \Big(\sum_{j|m} c_j \pi(j)\Big)^2.$$

**Theorem 2** *If $n \geqslant 2$, then*

$$\mathbf{V}_n h(\bar{c}) \leqslant B_n(\bar{c}) + \frac{1}{2} R_n(\bar{c}). \tag{5}$$

*The inequality becomes an equality for*

$$c_j = c_j^* := \frac{3}{\pi(j)} \sum_{d|j} d q^d \mu\left(\frac{j}{d}\right) - (2n+1)j, \quad 1 \leqslant j \leqslant n. \tag{6}$$

**Corollary 1** *If $n \geqslant 2$ and $\bar{c} \neq \bar{0}$, then*

$$\mathbf{V}_n h(\bar{c}) < \frac{3}{2} B_n(\bar{c}) < \left(\frac{3}{2} - \frac{q-1}{q} n q^{-n}\right) \mathbf{V} Y_n. \tag{7}$$

The inequalities are trivial for functions proportional to $h(\bar{j}, a) = n$ if $a \in \mathcal{G}_n$, because of $\mathbf{V}_n h(\bar{j}) = 0$ then. A shift of $\bar{c}$ eliminates this inconvenience. Observe that either of $B_n(\bar{c} - t\bar{j})$ and $R_n(\bar{c} - t\bar{j})$ attain their minimums in $t \in \mathbb{R}$ at

$$t = t_c := \frac{2}{(n+1)n} \sum_{m \leqslant n} m q^{-m} \sum_{j|m} c_j \pi(j).$$

**Theorem 3** *If $n \geqslant 3$, then*

$$B_n(\bar{c} - t_c\bar{j}) - \frac{1}{3} R_n(\bar{c} - t_c\bar{j}) \leqslant \mathbf{V}_n h(\bar{c}) \leqslant B_n(\bar{c} - t_c\bar{j}) + \frac{1}{2} R_n(\bar{c} - t_c\bar{j}). \tag{8}$$

*Both inequalities are sharp.*

**Corollary 2** *If $n \geqslant 3$ and $\bar{c} \neq \alpha\bar{j}$ for every $\alpha \in \mathbb{R}$, then*

$$\frac{2}{3} B_n(\bar{c} - t_c\bar{j}) < \mathbf{V}_n h(\bar{c}) < \frac{3}{2} B_n(\bar{c} - t_c\bar{j}).$$

The proofs of the last two theorems presented in Section 2 are built upon the ideas and auxiliary results obtained in [4], [2] and [5].

## 2 Proofs

We firstly recall known facts about random multisets which can be found in [3] and [1, Section 2.3]. Let $\bar{\gamma}^{(x)} = (\gamma_1^{(x)}, \gamma_2^{(x)}, \ldots)$ be the infinite dimensional vector of independent r.vs having the negative binomial distributions $NB(\pi(j), x^j)$, namely,

$$P(\gamma_j^{(x)} = m) = \binom{\pi(j) + m - 1}{m} (1 - x^j)^{\pi(j)} x^{jm}, \quad m = 0, 1, \ldots$$

where $0 < x \leqslant q^{-1}$. Then $\gamma_j^{(q^{-1})} = \gamma_j$ which has been introduced in Introduction. For convenience, we extend $\bar{k}(a)$ to $\bar{k}(a) := (k_1(a), \ldots, k_n(a), 0, \ldots)$ and use infinite dimensional vectors. Set $\theta^{(x)} = 1\gamma_1^{(x)} + \cdots + n\gamma_n^{(x)} + (n+1)\gamma_{n+1}^{(x)} + \cdots$ The latter r.v. is well defined if $0 < x < q^{-1}$, since the condition of the Boreli–Cantelli lemma is satisfied:

$$\sum_{j=1}^{\infty} P\big(\gamma_j^{(x)} \neq 0\big) = \sum_{j=1}^{\infty} \big(1 - \big(1 - x^j\big)^{\pi(j)}\big) < \infty.$$

**Lemma 1** *If* $\bar{s} = (s_1, \ldots, s_j, s_{j+1}, \ldots) \in \mathbb{N}_0^\infty$ *and* $0 < x < q^{-1}$, *then*

$$\nu_n\big(\bar{k}(a) = \bar{s}\big) = P(\bar{\gamma}^{(x)} = \bar{s} \big| \theta^{(x)} = n).$$

*Proof.* Actually, this is Lemma 2.2 in [3] stated there for $\mathbb{F}_q[t]$. The details remain the same in the more general case. □

**Lemma 2** *For a functional* $\Psi : \mathbb{N}_0^\infty \to \mathbb{R}$ *such that* $\mathbf{E}|\Psi(\bar{\gamma}^{(x)})| < \infty$, *we have*

$$\mathbf{E}\Psi(\bar{\gamma}^{(x)}) = (1 - qx)\left(\Psi(\bar{0}) + \sum_{n=1}^{\infty} \mathbf{E}_n \Psi\big(\bar{k}(a)\big) q^n x^n\right), \quad 0 < x < q^{-1}.$$

*Proof.* Apply Lemma 1 in the double averaging as follows:

$$\begin{aligned} \mathbf{E}\Psi(\bar{\gamma}^{(x)}) &= \sum_{n=0}^{\infty} \mathbf{E}\big(\Psi\big(\bar{\gamma}^{(x)}\big) \big| \theta^{(x)} = n\big) P\big(\theta^{(x)} = n\big) \\ &= \sum_{n=0}^{\infty} \mathbf{E}_n \Psi\big(\bar{k}(a)\big) P\big(\theta^{(x)} = n\big). \quad \square \end{aligned}$$

*Proof of Theorem 1.* It is straightforward. Applying the last lemma for the relevant $\Psi$, one can easily find the needed mixed moments of $k_j(a)$, $1 \leqslant j \leqslant n$, and further, the variance of the linear combination $h(a)$. □

To prove Theorems 2 and 3, we will apply the following lemmas concerning particular matrices and quadratic forms.

**Lemma 3** *Let* $U = ((u_{ij}))$, $i, j \leqslant n$, *be the symmetric matrix with the entries*

$$u_{ij} = \mathbf{1}\{i + j > n\}(ij)^{-1/2}.$$

*The spectrum of* $U$ *is the set* $\{1, -1/2, 1/3, \ldots, (-1)^{n-1}/n\}$. *The eigenvectors corresponding to the first three eigenvalues are proportional to* $\bar{e}_r = (e_{r1}, \ldots, e_{rn})$, *where* $r = 1, 2, 3$ *and, for* $j \leqslant n$,

$$e_{1j} = \sqrt{j}, \quad e_{2j} = (3j - 2n - 1)\sqrt{j}, \quad e_{3j} = \big(10j^2 - 6(2n+1)j + 3n^2 + 3n + 2\big)\sqrt{j}.$$

*Proof.* This is the byproduct of works [4] and [2]. □

Afterwards, let $\bar{e}_r$, $1 \leqslant r \leqslant n$, be the orthogonal basis of $\mathbb{R}^n$ comprised of the eigenvectors of $U$ and $\bar{x}'$ means the transposed vector $\bar{x}$.

**Lemma 4** *If* $b_m \in \mathbb{R}$ *and* $1 \leqslant m \leqslant n$ *and* $n \geqslant 2$, *then*

$$-\frac{1}{2} \sum_{1 \leqslant m \leqslant n} m b_m^2 \leqslant \sum_{\substack{1 \leqslant m_1, m_2 \leqslant n \\ m_1 + m_2 > n}} b_{m_1} b_{m_2} \leqslant \sum_{1 \leqslant m \leqslant n} m b_m^2. \tag{9}$$

*If* $n \geqslant 3$ *and*

$$\sum_{m \leqslant n} m b_m = 0, \tag{10}$$

*then*

$$\sum_{\substack{1\leqslant m_1,m_2\leqslant n\\ m_1+m_2>n}} b_{m_1}b_{m_2} \leqslant \frac{1}{3}\sum_{1\leqslant m\leqslant n} mb_m^2. \tag{11}$$

*Moreover, each bound in* (9) *and* (11) *are achieved, respectively, for* $b_m = e_{rm}/\sqrt{m}$*, where* $r = 2, 1, 3$ *and* $e_{rm}$ *have been defined in Lemma* 3*.*

*Proof.* Inequalities (9) are seen from Lemma 3 after the substitution $b_m = x_m/\sqrt{m}$, $m \leqslant n$, since the extreme eigenvalues are 1 and $-1/2$.

After the same substitution, we further examine the quadratic form with the matrix $U$. Condition (10) reckons the subspace of vectors $\bar{x} = (x_1, \dots, x_n)$ satisfying $x_1 + \dots + x_j\sqrt{j} + \dots + x_n\sqrt{n} = \bar{x}\cdot\bar{e}_1' = 0$. This subspace is spanned over the first eigenvector. In other words, under (10), only the form values obtained in the subspace $L \subset \mathbb{R}^n$ spanned over the vectors $\bar{e}_2, \dots, \bar{e}_n$ count. Hence

$$\max_{\bar{x}\in L}\|\bar{x}\|^{-2}\bar{x}U\bar{x}' \leqslant \max_{2\leqslant r\leqslant n}(-1)^{r-1}/r = 1/3.$$

Returning to $b_m$, from this we obtain inequality (11). □

*Proof of Theorem 2.* After grouping the summands, expression (4) can be rewritten as follows:

$$V_n h(\bar{c}) = B_n(\bar{c}) - \sum_{\substack{m_1,m_2\leqslant n\\ m_1+m_2>n}} \Big(q^{-m_1}\sum_{i|m_1} c_i\pi(i)\Big)\Big(q^{-m_2}\sum_{j|m_2} c_j\pi(j)\Big).$$

Now evidently estimate (5) follows from Lemma 4 with

$$b_m = q^{-m}\sum_{j|m} c_j\pi(j), \quad m \leqslant n.$$

Moreover, it becomes an equality if we take $c_j = c_j^*$ satisfying

$$q^{-m}\sum_{j|m} c_j^*\pi(j) = 3m - 2n - 1,$$

which by the Möbius inversion formula and (1) may be rewritten as (6). □

*To prove* the first assertion of Corollary 1, it suffices to estimate the inner sum in $R_n(\bar{c})$, namely,

$$\Big(\sum_{j|m} c_j\pi(j)\Big)^2 \leqslant \sum_{j|m}\frac{c_j^2\pi(j)}{j}\cdot\sum_{j|m} j\pi(j) = \sum_{j|m}\frac{c_j^2\pi(j)}{j}\cdot q^m.$$

Further, using the expression of $\mathbf{V}Y_n$, we just estimate the remainder:

$$\begin{aligned}\mathbf{V}Y_n - B_n(\bar{c}) &= \sum_{j\leqslant n} c_j^2\pi(j)\sum_{k>n/j} kq^{-jk} \geqslant nq^{-n}\sum_{j\leqslant n}\frac{c_j^2\pi(j)}{j}\frac{q^j}{(q^j-1)^2}\cdot\frac{q^j-1}{q^j}\\ &\geqslant nq^{-n-1}(q-1)\mathbf{V}Y_n.\end{aligned}$$

Plugging both estimates into (5), we obtain the first inequality in Corollary 1 with $\leqslant$ instead of $<$. In fact, we obtained the strict inequality since Cauchy's inequality applied in the last step is strict if $\bar{c}$ is not proportional to $\bar{j}$, and in this exceptional case, $\mathbf{V}h(\bar{c}) = 0$. $\square$

*Proof of Theorem 3.* Observe that $\mathbf{V}_n h(\bar{c}) = \mathbf{V}_n\big(h(\bar{c}) - tn\big) = \mathbf{V}_n h(\bar{c} - t\bar{j})$ for every $t \in \mathbb{R}$. Hence the right-hand inequality follows from (5) applied for the shifted statistics.

To get the lower bound of variance, we combine (4) and (11). We start with

$$\mathbf{V}_n h(\bar{c} - t_c\bar{j}) = B_n(\bar{c} - t_c\bar{j}) - \sum_{\substack{m_1, m_2 \leqslant n \\ m_1 + m_2 > n}} \tilde{b}_{m_1}\tilde{b}_{m_2},$$

where

$$\tilde{b}_m = q^{-m} \sum_{j|m} (c_j - t_c j)\pi(j)$$

and $m \leqslant n$. By the definition of $t_c$ the latter sequence satisfies condition (10). Hence by (11),

$$\sum_{\substack{m_1, m_2 \leqslant n \\ m_1 + m_2 > n}} \tilde{b}_{m_1}\tilde{b}_{m_2} \leqslant \frac{1}{3} \sum_{m \leqslant n} m\tilde{b}_m^2 = \frac{1}{3} R_n(\bar{c} - t_c\bar{j}).$$

This and (4) imply the lower bound. Moreover, the latter is sharp since Lemma 4 assures this by a choice of a particular sequence $\tilde{b}_m$, $m \leqslant n$. $\square$

## References


[1] R. Arratia, A.D. Barbour, S. Tavaré. *Logarithmic Combinatorial Structures: A Probabilistic Approach. EMS Monographs in Mathematics.* EMS Publishing House, Zürich, 2003.

[2] Ž. Baronėnas, E. Manstavičius, P. Šapokaitė. A sharp inequality for the variance with respect to the ewens sampling formula. 2019, arXiv:2003.05975v1.

[3] J.C. Hansen. Factorization in $f_q[x]$ and brownian motion. *Combin. Probab. Comput.*, **2**:285–299, 1993.

[4] J. Klimavičius, E. Manstavičius. The Turán–Kubilius inequality on permutations. *Annales Univ. Sci. Budapest., Sect. Comp.*, **48**:45–51, 2018.

[5] E. Manstavičius. Sharp bounds for the variance of linear statistics on random permutations. *Random Struct. Alg.*, 2020. https://doi.org/10.1002/rsa.20951.


REZIUMĖ

**Vėrinių multiaibių statistikos dispersijos efektyvūs įverčiai**

*A. Karbonskis, E. Manstavičius*

Nagrinėjama tiesinės statistikos, apibrėžtos atsitiktinių vėrinių multiaibėje, dispersija. Gauti tikslūs viršutinieji ir apatinieji įverčiai.

*Raktiniai žodžiai*: Turanas–Kubiliaus nelygybė; daugianariai virš baigtinio lauko; priedų funkcija