Distributions of the k-major index in random words and permutations

Wilbur Li
Clements High School

under the direction of
Cesar Cuenca
Massachusetts Institute of Technology

January 15, 2016

Abstract

The most noted permutation statistics, inversion number and major index, have been studied extensively in recent decades. Many results found regarding their joint distribution and normality have been found. We focus on the k-major index, a natural interpolation between the inversion number and major index, which exhibits elegant properties of equidistribution. We generalize normality and equidistribution to words with fixed alphabet size and large length. We then draw analogs from the fluctuation of the k-major index on a random, fixed permutation or word as k varies to standard Brownian motion. We prove a fundamental similarity of covariance independence, and find other connections. Finally, we observe limiting shapes of joint distributions of other pairs of potentially promising statistics.

1 Introduction

The study of permutation statistics arose in the early 20th century and has seen extensive research in recent years [1, 2]. Permutation statistics have interesting applications in combinatorics and representation theory, where some statistics are generalized to Weyl groups of root systems other than type A [3].

A permutation statistic is a map $S_n \to \mathbb{N}$. Consider a permutation $w \in S_n$; we often write it in one-line notation as $w = (w_1\, w_2 \cdots w_n)$. Let the *descent set* be $\mathrm{Des}(w) = \{i | w_i > w_{i+1}\}$. With this, we look at the descent number, the inversion number, and the major index, denoted by $\mathrm{des}(w), \mathrm{inv}(w), \mathrm{maj}(w)$, respectively. They are given by

$$\begin{aligned} \mathrm{des}(w) &= \#\{i | w_i > w_{i+1}\}; \\ \mathrm{inv}(w) &= \#\{(i,j) | i < j \text{ and } w_i > w_j\}; \\ \mathrm{maj}(w) &= \sum_{w_i > w_{i+1}} i. \end{aligned}$$

For example, the permutation $w = (1\ 4\ 3\ 5\ 2)$ has $\mathrm{Des}(w) = \{2, 4\}$, so $\mathrm{des}(w) = 2,\ \mathrm{inv}(w) = 3,\ \mathrm{maj}(w) = 6$.

Some completed work includes the computation by Corteel et al [4] of relative distributions between descent number and major index, and also the proof that cycle lengths over a random permutation are Poisson distributed [3].

Two of the most significant statistics, major index and inversion number, have been studied extensively [1, 2, 4, 5]. Major MacMahon [6] first identified the major index in 1913, and then he showed it to be equidistributed with the inversion number. Explicitly, with common notation $[n]_p = \frac{1-p^{n+1}}{1-p}$, $[n]_p! = [n]_p \cdots [1]_p$; the identities hold

$$\sum_{w \in S_n} p^{\mathrm{inv}(w)} = \sum_{w \in S_n} p^{\mathrm{maj}(w)} = [n]_p!. \tag{1}$$

In the present work, we focus on the natural analogues of major index and inversion number in the set of words with a fixed alphabet and fixed length. More generally, in this setting we shall study the k-major index, a statistic that interpolates between the major index and inversion number. Specifically, the 1-major index is the major index and the n-major index is the inversion number. The k-major index is equidistributed with the inversion number for all $1 \leq k \leq n$. Moreover, it has had interesting applications in representation theory; importantly, it has been used to provide a combinatorial proof of Macdonald positivity [5].

We introduce notation and review existing work in Section 2. In Section 3, we prove a Central Limit Theorem for the k-major statistic in words. In Section 4, we discuss k-major covariances as they converge across different lengths of permutations and different k's. We observe and conjecture on connections between k-major and Brownian motion.

2 Background

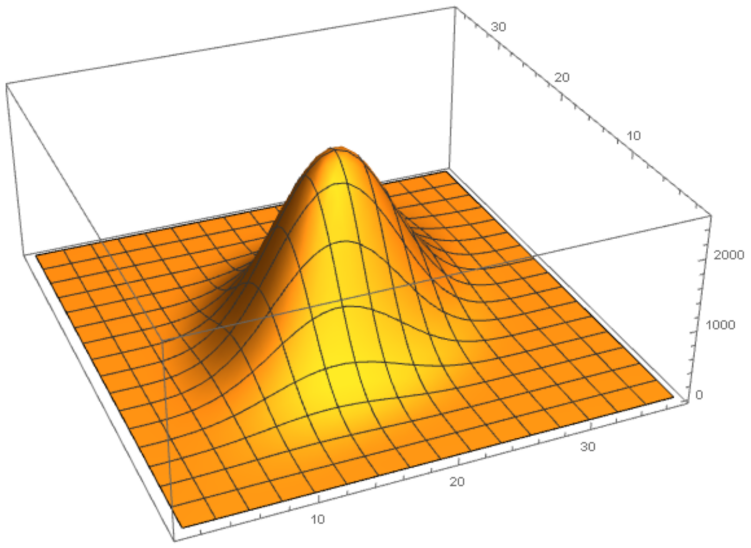


Figure 1: Joint frequency graph. The x-axis represents inversion number and the y-axis, the major index, while the z-values indicate the number of permutations in S_9 with those statistic values.

Results on distributions of statistics, such as the generating functions in equation (1), were studied early on. Work on the joint distribution of statistics began with Foata [7] who discovered in 1968 a bijection ϕ that proves $\text{inv}(\phi(w)) = \text{maj}(w)$ for any $w \in S_n$. Recently, in 2011, a study by Baxter and Zeilberger [8] on the asymptotics of the joint distribution of the inversion number and major index showed that the joint distribution of these statistics converges to a bivariate normal distribution in the limit (see Figure 1).

Our setting is distinct from the previous work mentioned above. We work with words.

Definition 2.1. An ***alphabet*** $\mathbf{a}$ is a nonempty set of letters $\{a_1, \ldots, a_l\}$ with a complete ordering $a_1 < \cdots < a_l$. A ***word*** w over an alphabet $\mathbf{a}$ is a string $(w_1 \cdots w_n)$ of letters such that $w_i \in \mathbf{a}$ for all i. This contrasts with a ***multiset permutation*** of length n, which is a rearrangement of a given multiset $\{1^{m_1}, \ldots, l^{m_l}\}$, where the letter j appears m_j times in the permutation, and $\sum_{i=1}^{l} m_i = n$.

Observe that standard permutations are just multiset permutations with $l = n$ and all $m_i = 1$. Without loss of generality, we use the ordered alphabet $\mathbf{a} = \{1, \ldots, l\}$ hereinafter.

The joint normality for permutations was generalized by Thiel in 2013 [2] to multiset permutations. These joint distribution proofs used a limit approximation of mixed moments. The proof assigned X_n to be the z-score for inversion number centralized on its own mean and standard deviation over the sample space of S_n, and assigned Y_n as the equivalent measure for major index. It was shown that as $n \to \infty$, $\mathbb{E}[X_n^r Y_n^s]$ approaches the joint moments of $\mathbb{E}[N_1^r N_2^s]$, where N_1, N_2 are two independent normal random variables, thus proving that X_n, Y_n are asymptotically independently normal.

In 2008, Sami Assaf [5] introduced the k-major index, a value that interpolates between the inversion number and major index, by combining the k-inversion set and the k-descent set, defined below.

Definition 2.2. For a word w, we define the ***d-descent set***, $\text{Des}_d(w)$, as the set of pairs of

indices $(i, i+d)$, such that $w_i > w_{i+d}$. Also, the ***k-inversion set***, $\mathrm{Inv}_k(w)$ is the union of all the d-descent sets, for $d < k$, $\mathrm{Inv}_k(w) = \bigcup_{d<k} \mathrm{Des}_d(w)$.

With these ingredients, we introduce the k-major index statistic.

Definition 2.3. Given a word w, let the ***k-major index***, or just k-major, of w be

$$\mathrm{maj}_k(w) = |\mathrm{Inv}_k(w)| + \sum_{(i,i+k)\in \mathrm{Des}_k(w)} i.$$

Notice that the 1-major index is the major index defined in Section 1. On the other hand, the n-major index is the inversion number.

We denote by $\mathbb{P}[V]$ the probability of an event V and by $\mathbb{1}_V$ the indicator variable of V.

Definition 2.4. Two statistics f and g are ***equidistributed*** over the set of words Ω, if for all m, we have $\mathbb{P}[f(w) = m] = \mathbb{P}[g(w) = m]$ for a word $w \in \Omega$ chosen uniformly at random.

Assaf [5] showed in 2008 that any k-major index is equidistributed with the inversion number, over any set of multiset permutations. In the next section, we consider the distribution of $\mathrm{maj}_k(w)$, for a fixed k. We shall sample a random word from the set W_n of words of length n and letters from the alphabet $\mathbf{a} = \{1, \ldots, l\}$. Then $|W_n| = l^n$.

3 Fixed k Distributions

3.1 Equidistribution

Proposition 3.1. *If two statistics f, g are equidistributed over all multisets of size n, they are equidistributed over the set of words W_n.*

Proof. Let M be the family of all multisets $\{1^{m_1}, \ldots, l^{m_l}\}$ where $\sum_{i=1}^{l} m_i = n$. Note that together, the multiset permutations of each $T \in M$ account for every possible word of length

n, since each word has n letters from the alphabet $\{1, \ldots, l\}$. For any m,

$$\begin{aligned}\mathbb{P}[f(w) = m] &= \sum_{T \in M} \mathbb{P}[f(w) = m | w \in S_T] \mathbb{P}[w \in S_T] \\ &= \sum_{T \in M} \mathbb{P}[g(w) = m | w \in S_T] \mathbb{P}[w \in S_T] \\ &= \mathbb{P}[g(w) = m].\end{aligned}$$

Here S_T denotes the set of multiset permutations of multiset T. □

With Assaf's conclusions [5] on the equidistribution of all k-major indices over multisets of size n, we can apply Proposition 3.1 to say each k-major index is equidistributed identically over the set of words of length n.

In the next section, we prove a Central Limit Theorem for the k-major statistic on words. As a first step, we find here explicit expressions for the mean μ_n and variance σ_n^2 of the inversion statistic inv (and therefore of any k-major statistic). In a word w, let $X_{i,j} = \mathbb{1}_{w_i > w_j}$ be the indicator random variables for the pair (i, j). Observe that $\sum_{1 \le i < j \le n} X_{i,j} = \mathrm{inv}(w)$.

Proposition 3.2. *The distribution of* $\mathrm{inv}(w)$ *over* $w \in W_n$ *has mean* $\mu_n = \frac{l-1}{2l}\binom{n}{2}$ *and variance* $\sigma_n^2 = \frac{l^2-1}{72l^2} n(n-1)(2n+5)$.

Proof. Each $X_{i,j}$ has a probability $\binom{l}{2}/l^2 = \frac{l-1}{2l}$ of being 1, dependent only on the elements at indices i, j. This means $\mathbb{E}[X_{i,j}] = \frac{l-1}{2l}$, and so

$$\mathbb{E}[\mathrm{inv}(w)] = \sum_{i<j} \mathbb{E}[X_{i,j}] = \binom{n}{2} \frac{l-1}{2l}.$$

The variance of a Bernoulli variable with parameter p is $p(1-p)$, so

$$\begin{aligned}
\mathrm{Var}(X_{i,j}) &= \frac{l-1}{2l}\frac{l+1}{2l} \\
\mathrm{Var}(\mathrm{inv}(w)) &= \mathrm{Var}\left(\sum_{i<j} X_{i,j}\right) \\
&= \sum_{i<j} \mathrm{Var}(X_{i,j}) + 2 \sum_{(i,j)<_L(i',j')} \mathrm{Cov}(X_{i,j}, X_{i',j'}) \\
&= \binom{n}{2}\frac{l-1}{2l}\frac{l+1}{2l} + 2 \sum_{(i,j)<_L(i',j')} \mathrm{Cov}(X_{i,j}, X_{i',j'}).
\end{aligned}$$

Here we used the fact that the variance of a sum is the sum of the variances plus the sum of twice the mutual covariances. The notation $(i,j) <_L (i',j')$ uses the lexicographical ordering L of the ordered pairs: comparing i and i', then j and j' if necessary. Now we compute using $\mathrm{Cov}(X,Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$, where X, Y are arbitrary random variables. Note that if $\{i,j\} \cap \{i',j'\} = \emptyset$, then $X_{i,j}$ and $X_{i',j'}$ are independent, since all indices are chosen separately. Therefore we consider three cases of dependence, or cases where indices overlap.
Case 1 $i = i'$ **and** $j < j'$: This configuration appears $\binom{n}{3}$ times in a word of length n, since we choose the set of indices $\{i, j, j'\}$:

$$\begin{aligned}
\mathrm{Cov}(X_{i,j}, X_{i,j'}) &= \mathbb{E}[(X_{i,j} - \frac{l-1}{2l})(X_{i,j'} - \frac{l-1}{2l})] \\
&= \mathbb{E}[X_{i,j}X_{i,j'}] - \frac{l-1}{2l}\left(\mathbb{E}[X_{i,j}] + \mathbb{E}[X_{i,j'}]\right) + \left(\frac{l-1}{2l}\right)^2.
\end{aligned}$$

We evaluate the first and middle terms of this expression

$$\begin{aligned}
\mathbb{E}[X_{i,j}X_{i,j'}] &= \frac{1}{l^3}\sum_{t=1}^{l}(t-1)^2 \\
&= \frac{(l-1)l(2l-1)}{6l^3} = \frac{(l-1)(2l-1)}{6l^2}; \qquad (2) \\
\mathbb{E}[X_{i,j}] &= \mathbb{E}[X_{i,j'}] = \frac{l-1}{2l}.
\end{aligned}$$

We therefore conclude

$$\mathrm{Cov}(X_{i,j}, X_{i,j'}) = \frac{(l-1)l(2l-1)}{6l^3} - 2\left(\frac{l-1}{2l}\right)^2 + \left(\frac{l-1}{2l}\right)^2 = \frac{l^2-1}{12l^2}.$$

Case 2 $i < i'$ **and** $j = j'$: This is symmetric with Case 1, considering a reversal. This configuration occurs in $\binom{n}{3}$ ways, and takes covariance $\frac{l^2-1}{12l^2}$.

Case 3 $j = i'$: This again occurs $\binom{n}{3}$ times. We perform the same expansion,

$$\begin{aligned}\mathrm{Cov}(X_{i,j}, X_{j,j'}) &= \mathbb{E}[X_{i,j}X_{j,j'}] - \frac{l-1}{2l}\left(\mathbb{E}[X_{i,j}] + \mathbb{E}[X_{j,j'}]\right) + \left(\frac{l-1}{2l}\right)^2 \\ &= \mathbb{E}[X_{i,j}X_{j,j'}] - \frac{l-1}{2l}\left(\frac{l-1}{l}\right) + \left(\frac{l-1}{2l}\right)^2.\end{aligned}$$

We have $\mathbb{E}[X_{i,j}X_{j,j'}] = \frac{1}{l^3}\binom{l}{3}$. This comes out to $\mathrm{Cov}(X_{i,j}, X_{j,j'}) = \frac{1-l^2}{12l^2}$. Combining these cases, we have

$$\begin{aligned}\mathrm{Var}(\mathrm{inv}(w)) &= \binom{n}{2}\frac{l^2-1}{4l^2} + 2\binom{n}{3}\left(\frac{l^2-1}{12l^2} + \frac{l^2-1}{12l^2} - \frac{l^2-1}{12l^2}\right) \\ &= \frac{l^2-1}{72l^2}n(n-1)(2n+5).\end{aligned}$$

□

3.2 Central Limit Theorem

In this section, we prove a CLT for the k-major statistic on words. Our main tool is a variant of the CLT, similar to the one proposed by Lomnicki and Zaremba [9] for triangular arrays of random variables; refer to Appendix A for the original statement.

Proposition 3.3. *Let $V_{i,k}$ with $i = 1, 2, \ldots; k = i+1, i+2, \ldots$ be random variables satisfying*

1. *Any two finite sets of variables $V_{i,k}$ have the property that no value taken by either index of any element of one set appears among the values of the indices of the elements*

of the other set, then these two sets are mutually independent;

2. $\mathbb{E}[V_{i,k}] = 0$ *and* $\mathbb{E}[V_{i,k}^2] = \mu_2$, *a constant for all* $i = 1, 2, \ldots;\ k = i+1, i+2, \ldots;$

3. $\mathbb{E}[V_{i,k}V_{i,j}] = \mathbb{E}[V_{i,j}V_{k,j}] = c$ *and* $\mathbb{E}[V_{i,k}V_{k,j}] = -c$, *where* c *is another constant, for all integer triples* $i < k < j$;

4. *The random variables* $V_{i,k}$ *have collectively bounded moments up to some order* m.

Now, given

$$\bar{V}_N = \frac{2}{N(N-1)} \sum_{i=1}^{N-1} \sum_{k=i+1}^{N} V_{i,k},$$

We have, for any $r \leq m$,

$$\lim_{N\to\infty} N^{r/2}\, \mathbb{E}(\bar{V}_N^r) = \begin{cases} \frac{2^{r/2} r!}{(r/2)!} \left(\frac{c}{3}\right)^{r/2} & \text{if } r \text{ is even;} \\ 0 & \text{if } r \text{ is odd.} \end{cases}$$

Proof. The proof of our statement is identical to the proof of Lomnicki and Zaremba's theorem, except for one adjustment. The original statement requires $\mathbb{E}[V_{i,k}V_{i,j}] = \mathbb{E}[V_{i,j}V_{k,j}] = \mathbb{E}[V_{i,k}V_{k,j}] = c$, while our statement requires instead that $\mathbb{E}[V_{i,k}V_{k,j}] = -c$, but the rest is the same. By following the proof of their theorem, it is not hard to see that we obtain a similar formula for the (normalized) limit of moments of $\bar{V}_N$, with $\frac{c}{3}$ instead of c. □

Remark. In the statement of the proposition of their theorem, Lomnicki and Zaremba omitted the requirement $\mathbb{E}[V_{i,k}V_{k,j}] = c$, but it is clear from their proof that this is required.

Theorem 3.4. *For any* $1 \leq k \leq N$, *the* k*-major statistic is asymptotically normally distributed over words* $w \in W_N$. *Explicitly,*

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{\mathrm{maj}_k(w) - \mu_n}{\sigma_n} \leq x\right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-s^2/2} ds,$$

where μ_n, σ_n^2 are the mean and variance of the permutation $w \in S_n$ that is chosen uniformly at random.

Proof. It suffices to show that $\text{inv}(w)$ is normally distributed, since Theorem 3.1 guarantees that every other $\text{maj}_k(w)$ will be identically distributed for any k.

We use the same indicators $X_{i,j}$ as in Proposition 3.2. The inversion number is then a sum of these $\binom{n}{2}$ identically distributed random variables.

Consider an infinite sequence $S = s_1, s_2, \ldots$ of letters from the alphabet $\mathbf{a} = \{1, \ldots, l\}$. Each s_i is defined independently, with $\mathbb{P}[s_i = t] = \frac{1}{l}$ for any letter $1 \leq t \leq l$. For an integer N, let w be the word representing the first N letters of the sequence S, and let $Z_N = \frac{\text{inv}(w) - \mu_N}{\sigma_N}$, where Z_N depends solely on S and N. We show that as $N \to \infty$, the Z_N's approach $\mathcal{N}(0,1)$ in distribution as we consider all sequences S.

We use the variation of the Central Limit Theorem for the dependent variables stated in Proposition 3.3. We consider the adjusted variables $Y_{i,j} = X_{i,j} - \frac{l-1}{2l}$. Here we check that all four conditions are satisfied.

1. Two variables in the sum $Y_{i,j}, Y_{k,h}$ are independent when $\{i,j\} \cap \{k,h\} = \emptyset$. This holds since the indices i, j, k, h are each determined independently;

2. $\mathbb{E}[Y_{i,k}] = 0$, since $\mathbb{E}[X_{i,k}] = \frac{l-1}{2l}$ by Proposition 3.2. For each $i < k$, we compute $\mu_2 = \mathbb{E}[Y_{i,k}^2] = \text{Var}(Y_{i,k}) + \mathbb{E}[Y_{i,k}]^2 = \frac{l-1}{2l}\frac{l+1}{2l}$;

3. We show $\mathbb{E}[Y_{i,k}Y_{i,j}] = \mathbb{E}[Y_{i,j}Y_{k,j}] = c$ and $\mathbb{E}[Y_{i,k}Y_{k,j}] = -c$ when $i < k < j$. We evaluate the left side and show it is constant as i, j, k vary. We are given three fixed indices $i < k < j$, and we have $\mathbb{E}[Y_{i,k}Y_{i,j}] = \mathbb{E}[(X_{i,k} - \frac{l-1}{2l})(X_{i,j} - \frac{l-1}{2l})] = \mathbb{E}[X_{i,k}X_{i,j}] - \frac{l-1}{2l}(\mathbb{E}[X_{i,k}] + \mathbb{E}[X_{i,j}]) + \left(\frac{l-1}{2l}\right)^2 = \mathbb{E}[X_{i,k}X_{i,j}] - \left(\frac{l-1}{2l}\right)^2$. We take from (2) in Proposition 3.2 that $\mathbb{E}[X_{i,k}X_{i,j}] = \frac{(l-1)(2l-1)}{6l^2}$. This means $\mathbb{E}[Y_{i,k}Y_{i,j}] = \frac{(l-1)(2l-1)}{6l^2} - \left(\frac{l-1}{2l}\right)^2 = \frac{l^2-1}{12l^2}$. Observe that $\mathbb{E}[Y_{i,j}Y_{k,j}]$ can be computed similarly. Now $\mathbb{E}[X_{i,k}X_{k,j}] = \binom{l}{3}/l^3$, so from above $\mathbb{E}[Y_{i,k}Y_{k,j}] = \binom{l}{3}/l^3 - \left(\frac{l-1}{2l}\right)^2 = -\frac{l^2-1}{12l^2}$, so we fulfill conditions with $c = \frac{l^2-1}{12l^2}$;

4. $|Y_{i,j}| < 1$ for all $i < j$. Then for any m, and any $m_1, \ldots, m_k < m$, we have $\mathbb{E}[|Y_{i_1,j_1}^{m_1} \cdots Y_{i_k,j_k}^{m_k}|] < 1$ for all $m_1, \ldots, m_k \in \mathbb{N}$.

With all conditions satisfied, Proposition 3.3 implies that for any r,

$$\lim_{N\to\infty} N^{r/2}\, \mathbb{E}(\bar{Y}_N^r) = \begin{cases} \frac{2^{r/2} r!}{(r/2)!} \left(\frac{c}{3}\right)^{r/2} & \text{if } r \text{ is even;} \\ 0 & \text{if } r \text{ is odd.} \end{cases}$$

Now

$$\begin{aligned} \bar{Y}_N &= \frac{1}{\binom{N}{2}} \sum_{i<j\le N} \left(X_{i,j} - \frac{l-1}{2l} \right) \\ &= \frac{1}{\binom{N}{2}} (\mathrm{inv}(w) - \mu_N). \end{aligned}$$

Using this interpretation of $\bar{Y}_N$, we have

$$\begin{aligned} N^{r/2}\mathbb{E}[\bar{Y}_N^r] &= \mathbb{E}\left[\left(\frac{\mathrm{inv}(w) - \mu_N}{\sigma_N} \frac{N^{1/2}\sigma_N}{\binom{N}{2}} \right)^r \right] \\ &= \mathbb{E}\left[\left(\frac{\mathrm{inv}(w) - \mu_N}{\sigma_N} \frac{\sigma_N}{\sqrt{N}(N-1)/2} \right)^r \right] \\ &= \mathbb{E}[Z_N^r] \left(\frac{2\sigma_N}{N^{3/2} - N^{1/2}} \right)^r. \end{aligned}$$

We evaluate the limit for part of this product

$$\begin{aligned} \lim_{N\to\infty} \left(\frac{2\sigma_N}{N^{3/2} - N^{1/2}} \right)^r &= \lim_{N\to\infty} \left(\frac{2\sqrt{\frac{l^2-1}{72l^2} N(N-1)(2N+5)}}{N^{3/2} - N^{1/2}} \right)^r \\ &= \lim_{N\to\infty} \left(2\sqrt{\frac{l^2-1}{72l^2}} \frac{\sqrt{N(N-1)(2N+5)}}{N^{3/2} - N^{1/2}} \right)^r. \end{aligned}$$

The inner right fraction converges to $\sqrt{2}$ as we take this limit, and we are left with

$$\begin{aligned}\lim_{N\to\infty}\left(\frac{2\sigma_N}{N^{3/2}-N^{1/2}}\right)^r &= \left(2\sqrt{\frac{l^2-1}{36l^2}}\right)^r \\ &= \left(\frac{\sqrt{l^2-1}}{3l}\right)^r.\end{aligned}$$

Because $l > 1$, this quantity is positive and finite. We proceed to use the limit from the conclusion, so for even r we have

$$\left(\lim_{N\to\infty}\mathbb{E}[Z_N^r]\right)\left(\frac{\sqrt{l^2-1}}{3l}\right)^r = \frac{2^{r/2}r!}{(r/2)!}\left(\frac{c}{3}\right)^{r/2}.$$

This means for even r,

$$\begin{aligned}\lim_{N\to\infty}\mathbb{E}[Z_N^r] &= \frac{2^{r/2}r!}{(r/2)!}\left(\frac{c}{3}\right)^{r/2}\left(\frac{3l}{\sqrt{l^2-1}}\right)^r \\ &= \frac{2^{r/2}r!}{(r/2)!}\left(\frac{l^2-1}{36l^2}\right)^{r/2}\left(\frac{3l}{\sqrt{l^2-1}}\right)^r \\ &= \frac{r!}{2^{r/2}(r/2)!}.\end{aligned}$$

For odd r, we conclude

$$\begin{gathered}\left(\lim_{N\to\infty}\mathbb{E}[Z_N^r]\right)\left(\frac{\sqrt{l^2-1}}{3l}\right)^r = 0 \\ \lim_{N\to\infty}\mathbb{E}[Z_N^r] = 0.\end{gathered}$$

The r-moments of the Z_N's exhibit the limit behavior $\mathbb{E}[Z_n^r] \to \frac{r!}{2^{r/2}(r/2)!}$ for even r and $\mathbb{E}[Z_n^r] \to 0$ for odd r. Knowing that the moments of these random variables approach the moments of the normal distribution, we apply a convergence theorem such as Theorem 30.2 in *Probability and Measure* by Billingsley [10]. It follows that

$$\lim_{N\to\infty} \mathbb{E}[Z_N^r] = \mathbb{E}[\mathcal{N}(0,1)^r], \text{ convergence of moments, implies,}$$

$$\lim_{N\to\infty} \mathbb{P}[Z_N \leq x] = \mathbb{P}[\mathcal{N}(0,1) \leq x], \text{ convergence in distribution.}$$

The values of Z_N take on a normal distribution centered at 0 with variance 1. This shows that the inversion statistic is normally distributed across all words with mean μ_n and standard deviation σ_n, and therefore the k-major index is normally distributed across words for all individual k. □

4 Varying the k Parameter

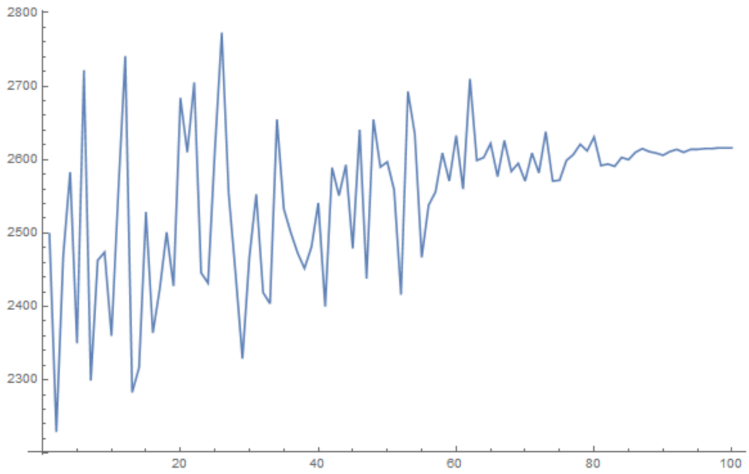


Figure 2: Given a fixed permutation $w \in S_{100}$, we plot k on the x-axis and a corresponding $\text{maj}_k(w)$ on the y-axis. Observe the gradually stabilizing, seemingly random graph.

If we analyze the value of $\text{maj}_k(w)$ on a fixed $w \in W_n$ as k ranges from 1 to n, the k-major index varies widely and then stabilizes near $\text{inv}(w)$, as shown in Figure 2. We find that this k-major progression has much in common with Brownian motion.

Let M_w be the function defined by $M_w(k) = \text{maj}_k(w)$ (w is fixed) and k varies from 1 to n. Define $C(h,k) = \mathbb{E}[\text{maj}_h(w)\,\text{maj}_k(w)]$, where $w \in S_n$ is random, for all relevant h, k.

Theorem 4.1. *Consider the probability space S_n, where each $w \in S_n$ is equally likely. The covariance* $\mathrm{Cov}(\mathrm{maj}_h(w), \mathrm{maj}_k(w))$ *depends only on* $\min(h, k)$. *In other words,* $\mathrm{Cov}(\mathrm{maj}_h(w), \mathrm{maj}_k(w)) = \mathrm{Cov}(\mathrm{maj}_{h+1}(w), \mathrm{maj}_k(w))$ *for all* $h > k$.

Proof. By definition,

$$\mathrm{Cov}(\mathrm{maj}_h(w), \mathrm{maj}_k(w)) = \mathbb{E}[\mathrm{maj}_h(w)\,\mathrm{maj}_k(w)] - \mathbb{E}[\mathrm{maj}_h(w)]\,\mathbb{E}[\mathrm{maj}_k(w)].$$

Observe that all k-major statistics are equidistributed, so they all have the same mean. Thus to prove the theorem, it suffices to show that $\mathbb{E}[\mathrm{maj}_h(w)\,\mathrm{maj}_k(w)]$ is constant if h varies over the numbers $k, k+1, \ldots$.

Let $C = \mathbb{E}[\mathrm{maj}_{k+1}(w)\,\mathrm{maj}_k(w)]$. We want to show that each value of h in the expectation will evaluate to C.

We do so by induction. Given $h > k$, we assume $\mathbb{E}[\mathrm{maj}_h(w)\,\mathrm{maj}_k(w)] = C$. Given indices $i < j$, we use indicator $X_{i,j}$ to evaluate $w_j > w_i$, i.e., whether the letter at position j is greater than the one at i. Now, given a permutation w, observe

$$\begin{aligned}
\mathrm{maj}_h &= \sum_{s-r<h} X_{r,s} + \sum_i i X_{i,i+h} \\
\mathrm{maj}_{h+1} &= \sum_{s-r<h+1} X_{r,s} + \sum_i i X_{i,i+h+1} \\
\mathrm{maj}_{h+1} - \mathrm{maj}_h &= \sum_r (1-r) X_{r,r+h} + \sum_s s X_{s,s+h+1}.
\end{aligned}$$

Note that in summing these indicator products, we incorporate every pair in the valid range. Call this last difference expression δ. This means

$$\begin{aligned}
\mathbb{E}[\mathrm{maj}_{h+1}(w)\,\mathrm{maj}_k(w)] &= \mathbb{E}[(\mathrm{maj}_h(w) + \delta)\,\mathrm{maj}_k(w)] \\
&= \mathbb{E}[\delta\,\mathrm{maj}_k(w)] + C.
\end{aligned}$$

We hope to show that $\mathbb{E}[\delta \operatorname{maj}_k(w)] = 0$.

Lemma 4.2. *Given any fixed c, k, and $h > k$,*

$$\mathbb{E}[X_{c,c+k} \sum sX_{s,s+h+1}] = \mathbb{E}[X_{c,c+k} \sum (r-1)X_{r,r+h}],$$

as w ranges over S_n. Note that there are $n-h-1$ nonzero terms on either side.

Proof. We consider the sums in pairs. Any given p produces a pair of elements consisting of $L_p = \mathbb{E}[X_{c,c+k}\, pX_{p,p+h+1}]$ on the LHS and $R_p = \mathbb{E}[X_{c,c+k}\, pX_{p+1,p+h+1}]$ on the RHS. Notice that showing $\sum L_p = \sum R_p$ for $1 \le p \le n-h-1$ would adequately determine the relation.

The sets $F = \{c, c+k\}$ and $T = \{p, p+1, p+h+1\}$ can have at most 1 collision, unless $k = 1$ (which we will consider separately). Consider the values of p for which there are 0 collisions between F, T. Then the multiplied indicators are independent and $L_p = R_p = \frac{1}{4}p$. These pairs can all be disregarded, since they are equal between LHS and RHS.

Next, consider when $F \cap T = \{p+h+1\}$. Then either $c+k = p+h+1$ or $c = p+h+1$. The probability distributions for both of these cases are identical for L_p and R_p, so this case can, too, be disregarded.

We are left with four cases of $|F \cap T| = 1$: $p+1 = c,\ p = c,\ p+1 = c+k,\ p = c+k$. We show that the sums of L_p and R_p are equivalent over these scenarios.

Case 1: $p+1 = c$. The expectance $\mathbb{E}[X_{c,c+k}X_{p,p+h+1}] = \frac{1}{4}$, so $L_p = \frac{1}{4}(c-1)$. Meanwhile, $\mathbb{E}[X_{c,c+k}X_{p+1,p+h+1}] = \frac{1}{3}$, so $R_p = \frac{1}{3}(c-1)$.

Case 2: $p = c$. We have $\mathbb{E}[X_{c,c+k}X_{p,p+h+1}] = \frac{1}{3}$ and therefore $L_p = \frac{1}{3}c$. On the other hand, $\mathbb{E}[X_{c,c+k}X_{p+1,p+h+1}] = \frac{1}{4}$, so $R_p = \frac{1}{4}c$.

Case 3: $p+1 = c+k$. The expectance $\mathbb{E}[X_{c,c+k}X_{p,p+h+1}] = \frac{1}{4}$, so $L_p = \frac{1}{4}(c+k-1)$. Meanwhile, $\mathbb{E}[X_{c,c+k}X_{p+1,p+h+1}] = \frac{1}{6}$, so $R_p = \frac{1}{6}(c+k-1)$.

Case 4: $p = c+k$. We have $\mathbb{E}[X_{c,c+k}X_{p,p+h+1}] = \frac{1}{6}$ and therefore $L_p = \frac{1}{6}(c+k)$. On the other hand, $\mathbb{E}[X_{c,c+k}X_{p+1,p+h+1}] = \frac{1}{4}$, so $R_p = \frac{1}{4}(c+k)$.

Summing over all of these, we have

$$\sum L_p = c + \frac{5k}{12} - \frac{1}{2} = \sum R_p.$$

This proves the equality of the overall expression as p ranges over all $1 \leq p \leq n-h-1$. The same computation results for $k = 1$, with the alteration that Cases 2 & 3 are combined. The lemma is proved. □

Note that

$$\mathrm{maj}_k = \sum X_{i,i+1} + \sum X_{i,i+2} + \cdots + \sum X_{i,i+k-1} + \sum i\, X_{i,i+k}.$$

Also note that, from Lemma 4.2, for any c', k', t' where $h' > k'$,

$$\mathbb{E}[t'\, X_{c',c'+k'} \sum s'\, X_{s',s'+h'+1}] = \mathbb{E}[t'\, X_{c',c'+k'} \sum (r'-1)\, X_{r',r'+h'}].$$

Applying this equality numerous times for k' taking on $1, \ldots, k$ and c' becoming $1, \ldots, n-k$, we have the sum of equations that yield

$$\mathbb{E}[\mathrm{maj}_k \sum s\, X_{s,s+h+1}] = \mathbb{E}[\mathrm{maj}_k \sum (r-1)\, X_{r,r+h}].$$

Maneuvering the terms, we have

$$\mathbb{E}[\mathrm{maj}_k \sum s\, X_{s,s+h+1}] + \mathbb{E}[\mathrm{maj}_k \sum (1-r)\, X_{r,r+h}] = 0$$
$$\mathbb{E}[\mathrm{maj}_k (\sum s\, X_{s,s+h+1} + \sum (1-r)\, X_{r,r+h})] = \mathbb{E}[\mathrm{maj}_k\, \delta] = 0.$$

This is what we need to complete the induction. Therefore

$$C = \mathbb{E}[\mathrm{maj}_{h+1}\, \mathrm{maj}_k] = \cdots = \mathbb{E}[\mathrm{maj}_n\, \mathrm{maj}_k], \text{ and so, as seen,}$$

$$\mathrm{Cov}(\mathrm{maj}_{h+1}, \mathrm{maj}_k) = \cdots = \mathrm{Cov}(\mathrm{maj}_n, \mathrm{maj}_k).$$

Theorem 4.1 is proved. □

Suspecting the same claim for words rather than permutations, we verify that for $n \leq 8$ and a range of l, identical results hold true.

Conjecture 4.3. *The covariance between two major indices h, k as we range over all words depends only on $\min(h, k)$; i.e., for some k, the value $\mathrm{Cov}(\mathrm{maj}_h(w), \mathrm{maj}_k(w))$ is constant for all $h > k$, considering all $w \in W_n$, with W_n the set of words of length n.*

Moving back to the case of permutations, Theorem 4.1 shows that for $h > k$, covariance $\mathrm{Cov}(\mathrm{maj}_h(w), \mathrm{maj}_k(w))$ depends only on k, so we can express this covariance as $C(k)$. If we plot the adjusted $C(k)$ as k ranges on $1 \leq k \leq n-1$, we obtain the plot in Figure 3; note that we scale all domain and range values to keep them on $[0, 1]$.

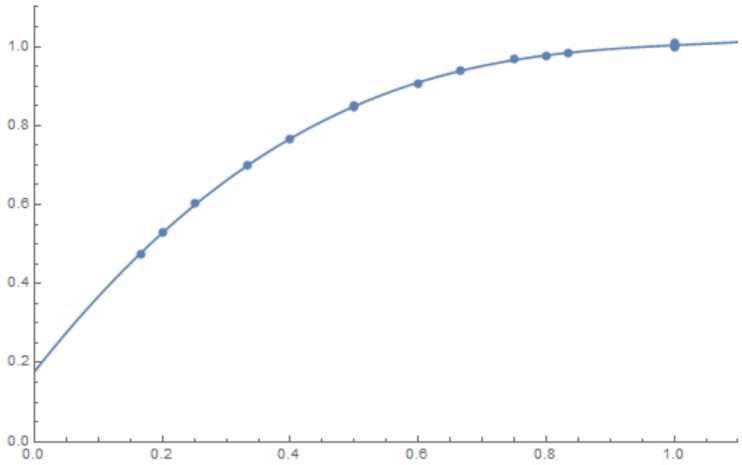


Figure 3: For several values of n, we plot the points $\left(\frac{k}{n-1}, \frac{C(k)}{\sigma_n}\right)$ for $1 \leq k \leq n$.

Conjecture 4.4. *As n approaches infinity, the adjusted points $\left(\frac{k}{n-1}, \frac{C(k)}{\sigma_n}\right)$ with $1 \leq k \leq n$ approach a quartic polynomial function.*

The regression quartic is drawn through the points in Figure 3, with very high correlation.

We make some observations that connect to Brownian motion. Let us enumerate properties of this stochastic process $\mathbf{B} = \{B_t | t \in [0, \infty)\}$:

1. $B_0 = 0$ and B_t is normally distributed with mean 0 and variance t, for each $t \geq 0$;
2. $\mathbf{B}$ has stationary increments, the distribution of $B_t - B_s$ is the same as B_{t-s};
3. $\mathbf{B}$ has independent increments;
4. B_t is almost surely continuous.

Surprisingly, this type of motion shares characteristics with M_w. Namely, (1) Each $M_w(k)$ is asymptotically normally distributed as seen in [5] and then Theorem 3.4; (2) Both $\mathrm{Cov}(B_h, B_k)$ and $\mathrm{Cov}(M_w(h), M_w(k))$ are determined by $\min(h, k)$ by Theorem 4.1. We ask the following.

Problem 4.5. *We can define a stochastic process X on $[0, 1]$ by letting $X_t := M_w(\lfloor nt \rfloor)$. Does this process converge to the Brownian motion (after suitable renormalization) on the interval $[0, 1]$, in some sense? We have in mind some statement like that of a simple random walk converging to the Brownian motion.*

5 Conclusion

In this paper we look at common permutation statistics in new ways, extending them to words and characterizing novel trends. We focus on the k-major index, a meaningful interpolation between the inversion number and the major index. We extend earlier results of normality to words, and identify the parameters of these distributions. We identify possible relations between k-major and Brownian motion by varying the k-index for fixed w in $\mathrm{maj}_k(w)$. We observe similarities and then prove a fundamental mutual property of constant covariance.

6 Acknowledgements

I want to thank my mentor, Cesar Cuenca of the MIT Math Department, for guiding my studies and directing my work. Thanks to Dr Tanya Khovanova for her insight and suggestions in reviewing the paper. Also thanks to Prof. David Jerison, and Prof. Ankur Moitra for organizing the math students at RSI. Thanks to my research tutor Dr John Rickert for his patience and insight in reviewing my work. Thanks to Prof. Richard Stanley for the suggestion to work on the asymptotics of permutation statistics. Thanks to Pavel Galashin for his helpful conversations on the k-major distribution. Thanks to Noah Golowich and Shashwat Kishore for revising my paper. Thanks to Girishvar Venkat for reading over my outlines. Thanks to the Research Science Institute held by the Center for Excellence in Education at the Massachusetts Institute of Technology for providing facilities, coordinating students and mentors, and encouraging science students through its summer program.

Thanks to my sponsors, Dr and Mrs Nathan J. Waldman, Drs Gang and Cong Yu, Mr Zheng Chen and Ms Chun Wang, and Mrs Cynthia Pickett-Stevenson for supporting me in attending the RSI program.

References

[1] A. M. Garsia and I. Gessel. Permutation statistics and partitions. *Adv. Math.*, 31:288–305, 1979.

[2] M. Thiel. The inversion number and the major index are asymptotically jointly distributed on words. *arXiv*, 1302.6708, 2013.

[3] A. Granville. Cycle lengths in a permutation are typically poisson. *J. Combin.*, 13, 2006.

[4] S. Corteel, I. M. Gessel, C. D. Savage, and H. S. Wilf. The joint distribution of descent and major index over restricted sets of permutations. *Ann. Comb.*, 11:375–386, 2007.

[5] S. H. Assaf. A generalized major index statistic. *Sem. Lotharingien de Combinatoire*, 60, 2008.

[6] P. A. MacMahon. The indices of permutations and the derivation therefrom of functions of a single variable associated with the permutations of any assemblage of objects. *Amer. J. Math*, 35:314–321, 1913.

[7] D. Foata. On the netto inversion number of a sequence. *Proc. Amer. Math. Soc.*, 19:236–240, 1968.

[8] A. Baxter and D. Zeilberger. The number of inversions and the major index of permutations are asymptotically joint-independently-normal. *arXiv*, 1004.1160, 2011.

[9] Z. A. Lomnicki and S. K. Zaremba. A further instance of the central limit theorem for dependent random variables. *Math. Nachr.*, 66:490–494, 1957.

[10] P. Billingsley. *Probability and Measure*. John Wiley & Sons, Inc., 3rd edition, 1995.

Appendix A Original variant of CLT

We reproduce the original version of the Central Limit Theorem described by Lomnicki and Zaremba [9]. Let $V_{i,k}$ with $i = 1, 2, \ldots; k = i+1, i+2, \ldots$ be random variables subject to the following conditions

1. Any two finite sets of variables $V_{i,k}$ have the property that no value taken by either index of any element of one set appears among the values of the indices of the elements of the other set, then these two sets are mutually independent;

2. $\mathbb{E}[V_{i,k}] = 0$ and $\mathbb{E}[V_{i,k}^2] = \mu_2$, a constant for all $i = 1, 2, \ldots;\ k = i+1, i+2, \ldots$;

3. $\mathbb{E}[V_{i,k}V_{i,j}] = \mathbb{E}[V_{i,j}V_{k,j}] = c$, another constant, for all natural number triples $i < k < j$;

4. The random variables $V_{i,k}$ have collectively bounded moments up to some order m.

Now, given

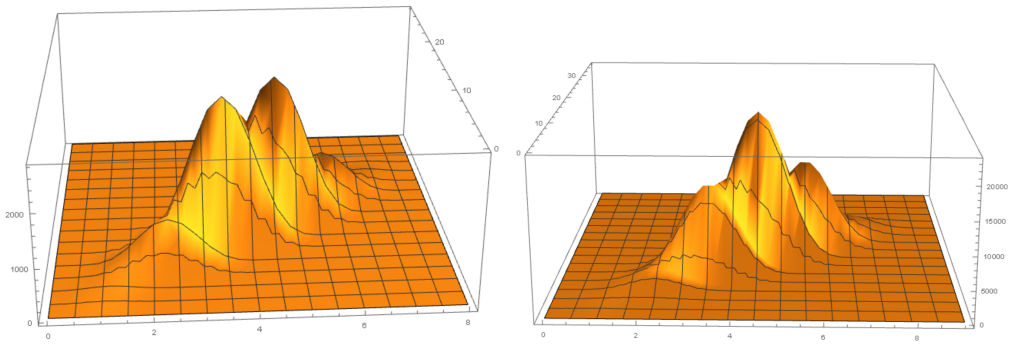
$$\bar{V}_N = \frac{2}{N(N-1)} \sum_{i=1}^{N-1} \sum_{k=i+1}^{N} V_{i,k},$$

we have, for any $r \leq m$,

$$\lim_{N\to\infty} N^{r/2}\, \mathbb{E}(\bar{V}_N^r) = \begin{cases} \frac{2^{r/2} r!}{(r/2)!}\, c^{r/2} & \text{if } r \text{ is even;} \\ 0 & \text{if } r \text{ is odd.} \end{cases}$$

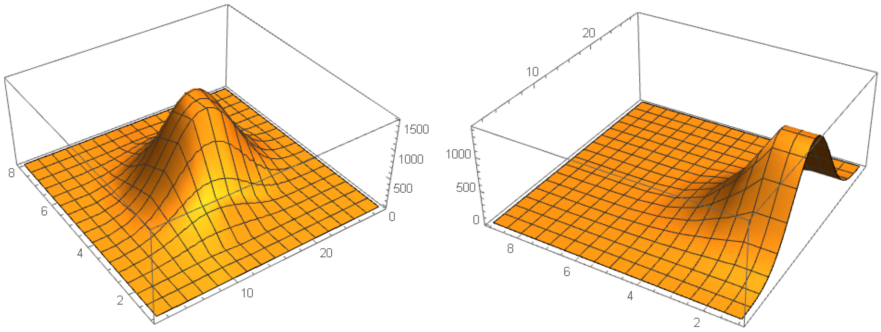
Appendix B Qualitative Comparisons

A broad search of other pairs of statistics leaves a couple smooth surfaces to examine. Comparing the descent and major indices results in Figure 4. For n even, two twin peaks with lower adjacent peaks result, while for odd n, the limit shape becomes a three-peak distribution with three significant peaks, one central, and two side. Figure 5b shows a decreasing plot for fixed major index, and this may be related to the fact that the cycle length in permutations is Poisson distributed [3].



(a) $n = 8$ (b) $n = 9$

Figure 4: Frequency plots of descent number and major index.



(a) Exceedance and inversion number for $n = 8$. (b) Fixed point and major index for $n = 8$.

Figure 5: Frequency plots.