J Theor Probab (2007) 20: 309–325
DOI 10.1007/s10959-007-0070-5

# Normal Approximations for Descents and Inversions of Permutations of Multisets

**Mark Conger · D. Viswanath**

Received: 18 August 2005 / Revised: 17 September 2006 /
Published online: 26 April 2007


**Abstract** Normal approximations for descents and inversions of permutations of the set $\{1, 2, \ldots, n\}$ are well known. We consider the number of inversions of a permutation $\pi(1), \pi(2), \ldots, \pi(n)$ of a multiset with $n$ elements, which is the number of pairs $(i, j)$ with $1 \le i < j \le n$ and $\pi(i) > \pi(j)$. The number of descents is the number of $i$ in the range $1 \le i < n$ such that $\pi(i) > \pi(i + 1)$. We prove that, appropriately normalized, the distribution of both inversions and descents of a random permutation of the multiset approaches the normal distribution as $n \to \infty$, provided that the permutation is equally likely to be any possible permutation of the multiset and no element occurs more than $\alpha n$ times in the multiset for a fixed $\alpha$ with $0 < \alpha < 1$. Both normal approximation theorems are proved using the size bias version of Stein's method of auxiliary randomization and are accompanied by error bounds.

**Keywords** Descents · Inversions · Multisets · Stein's method

## 1 Introduction

Let $\pi(1), \pi(2), \ldots, \pi(n)$ be a permutation of the multiset $\{1^{n_1}, 2^{n_2}, \ldots, h^{n_h}\}$ with $n_1 + \cdots + n_h = n$. The number of inversions, denoted $\mathrm{inv}(\pi)$, is defined as the number of pairs $(i, j)$ with $1 \le i < j \le n$ and $\pi(i) > \pi(j)$. The number of descents, denoted $\mathrm{des}(\pi)$, is the number of positions $i$ with $1 \le i < n$ and $\pi(i) > \pi(i + 1)$. Assume that $\pi$ is uniformly distributed. In this article, we use Stein's method to prove normal approximations with error bounds for $\mathrm{inv}(\pi)$ and $\mathrm{des}(\pi)$.

This work was supported by a research fellowship from the Sloan Foundation.

M. Conger · D. Viswanath (✉)
Department of Mathematics, University of Michigan, 530 Church Street, Ann Arbor,
MI 48109, USA
e-mail: divakar@umich.edu

M. Conger
e-mail: mconger@umich.edu

In the special case where $\pi$ is a uniformly distributed permutation of the set $\{1, 2, \ldots, n\}$, the distributions of both $\mathrm{inv}(\pi)$ and $\mathrm{des}(\pi)$ admit simple descriptions. The distribution of $\mathrm{inv}(\pi)$ is equal to that of the sum $X_1 + \cdots + X_{n-1}$, where the random variables $X_i$, $1 \le i \le n-1$, are independent with $X_i$ uniformly distributed over the set $\{0, 1, \ldots, i\}$. To obtain the distribution of $\mathrm{des}(\pi)$, we need the sum $X_1 + \cdots + X_{n-1} + X_n$, where the $X_i$ are independent and uniformly distributed in the interval $[0, 1]$. The probability that this sum lies in the interval $[d, d+1]$ equals the probability that $\mathrm{des}(\pi)$ equals $d$. According to Knuth [13], the first of these two results was noticed by O. Rodriguez in 1839. The result about $\mathrm{des}(\pi)$ was alluded to by Barton and Mallows [2]. An elegant proof is due to Stanley [18].

Normal approximations to $\mathrm{des}(\pi)$ and $\mathrm{inv}(\pi)$ in this special case can be obtained using these results and standard versions of the central limit theorem. The bounds

$$\left| P\left( \frac{\mathrm{des}(\pi) - (n-1)/2}{\sqrt{(n+1)/12}} \le x \right) - \Phi(x) \right| \le \frac{C}{\sqrt{n}}, \tag{1.1a}$$

$$\left| P\left( \frac{\mathrm{inv}(\pi) - \frac{1}{2}\binom{n}{2}}{\sqrt{n(n-1)(2n+5)/72}} \le x \right) - \Phi(x) \right| \le \frac{C}{\sqrt{n}}, \tag{1.1b}$$

where $C$ is a constant and $\Phi$ is the standard normal distribution, were proved using the method of exchangeable pairs [17, 20] by Fulman [8]. Other proofs of (1.1a) using Stein's method are sketched in [4] and [9].

From the survey by Barton and Mallows [2], it appears that the asymptotic normality of a quantity closely related to $\mathrm{des}(\pi)$, where $\pi$ is a uniformly distributed permutation of the set $\{1, \ldots, n\}$, was stated by Bienyamé in 1874 (Bull. Soc. Math. France, vol. 2, pp. 153–154). Bienyamé was interested in statistical applications. So were Levene and Wolfowitz [14] who stated that *runs* were widely used in quality control and in the study of economic time series. Runs are the monotone segments within a sequence of numbers and are closely related to descents. An early proof of the asymptotic normality of descents, which is implied by (1.1a), is due to Wolfowitz [21].

Let $\{1^{n_1}, 2^{n_2}, \ldots, h^{n_h}\}$ be a multiset, where $n_a$, $1 \le a \le h$, are positive integers. Let $n = n_1 + n_2 + \cdots + n_h$ be the number of elements of the multiset. Let $\alpha$ be a fixed number in $(0, 1)$. We assume that $n_a \le \alpha n$ for $1 \le a \le h$. Let $\pi$ be a uniformly distributed permutation of this multiset. We consider $\mathrm{inv}(\pi)$ and $\mathrm{des}(\pi)$ in this more general situation. The bounds that we obtain for the errors in the normal approximations to these quantities depend upon $\alpha$ and become infinite as $\alpha \to 1$. Let $h : R \to R$ be a bounded and piecewise continuously differentiable function and let $\beta = \max(1/2, \alpha)$. We use the size-bias version of Stein's method introduced by Baldi, Rinott and Stein [1] and prove that, for $n$ large enough,

$$\left| \mathrm{E}h\left( \frac{\mathrm{inv}(\pi) - \mu}{\sigma} \right) - \Phi h \right| \le C \left( \frac{\|h\|}{\beta(1-\beta)(\beta(1-\beta)n^{1/2} - C_1 n^{-1/2})} + \frac{\|Dh\|}{(\beta(1-\beta)n^{1/3} - C_2 n^{-2/3})^{3/2}} \right),$$

where $C$, $C_1$, and $C_2$ are some positive constants, $\Phi h$ is the expectation of $h$ with respect to the standard normal distribution, and $\mu$ and $\sigma^2$ are the mean and variance of $\mathrm{inv}(\pi)$, respectively. If $\alpha \geq 1/2$, then $\beta = \alpha$. Therefore the bound above diverges as $\alpha \to 1$. We prove a similar result for $\mathrm{des}(\pi)$.

Bounds such as the one given in the previous paragraph require $h$ to be continuous. Goldstein [9] has proved a normal approximation theorem that holds for nonsmooth $h$. We use that theorem to prove that

$$\left|\mathrm{P}\left(\frac{\mathrm{inv}(\pi)-\mu}{\sigma} \leq x\right) - \Phi(x)\right| \leq C(\beta)/\sqrt{n}$$

and that

$$\left|\mathrm{P}\left(\frac{\mathrm{des}(\pi)-\mu}{\sigma} \leq x\right) - \Phi(x)\right| \leq C(\beta)/\sqrt{n}.$$

These results are contained in Theorems 2.12 and 2.16 of this paper. The quantity $C(\beta)$ diverges when $\alpha \to 1$. As before $\beta = \max(1/2, \alpha)$. When the $n$ elements of the multiset are distinct, with $n \geq 2$, we may use $\alpha = 1/n$ and $\beta = 1/2$. Therefore the results stated above imply 1.1a and 1.1b.

The generating function of the number of permutations of a multiset with a given number of inversions is a rational function. Using this generating function, Diaconis [7] has shown that the asymptotic distribution of $\mathrm{inv}(\pi)$, where $\pi$ is uniformly distributed over permutations of a multiset, is normal. Theorem 2.12 about $\mathrm{inv}(\pi)$ is accompanied by an error bound of the correct order, which is $O(1/\sqrt{n})$, and the dependence of the error bound on $\alpha$ is also explicitly shown in our theorem. The generating function for the number of permutations of a multiset with a given number of descents, which is related to Foata's correspondence, was found by MacMahon [13, 15]. However, normal approximations to this quantity, such as the approximation given in Theorem 2.16, do not seem to be available.

Segments of $\pi(1), \ldots, \pi(n)$ between successive descents, or runs, are in ascending order. Knuth [13] has stated that runs are important in the study of sorting algorithms because runs are segments that are already in sorted order. Among the applications of descents and inversions to the study of sorting algorithms, multiway merging with replacement selection merits special mention. In this sorting method, the given sequence is first split into runs and the runs are merged together. Our results are pertinent to sorting algorithms if the keys used for sorting are allowed to repeat. For example, Theorem 2.16 about $\mathrm{des}(\pi)$ gives an idea of how many runs to expect if multiway merging is used on a sequence of records with repeated keys.

Descents and inversions have been used as test statistics in the special case where $\pi$ is a permutation of $\{1, 2, \ldots, n\}$. As already mentioned, early work on runs and descents was stimulated by statistical applications. Of the ten empirical tests for the randomness of a sequence of distinct numbers discussed by Knuth [12], one is based on runs and descents. Taking our results into account, inversions and descents can be used to test if a given permutation of a multiset of numbers is random. There are other ways to test if a given permutation of a multiset of numbers is random. If a permutation passes $n$ empirical tests for randomness but fails the $n+1$st, it is not random. Therefore having a greater number of empirical tests available makes for more robust testing [12].

DNA sequences are strings of the four letters $A$, $C$, $G$, and $T$. It is now well known that these sequences are far from random [11]. It has even been suggested that these sequences are similar to human languages [16]. Some commonly used compression algorithms such as the Lempel-Ziv method fail to compress typical DNA sequences however [11]. The entropy estimates of DNA sequences given in [11] and [16] proceed by dividing the sequence into blocks in some way. For instance, blocks of 6 consecutive letters are considered in [11]. These entropy estimates show that DNA sequences are not random.

In Sect. 3, we report the descents and inversions of the 19th chromosome of the human genome mainly as an illustration. We consider all 24 possible orderings of $A$, $C$, $G$, and $T$. With respect to each of these orderings, a calculation of descents and inversions shows that the number of descents and inversions of the DNA sequence departs from the mean by a large multiple of the standard deviation. It may be of some interest that this method of showing the DNA sequence to be far from random considers only single letters without dividing them into blocks.

Although we consider all possible orderings of $A$, $C$, $G$, and $T$, it must be noted that the molecular weights of the corresponding compounds implies the order $C < T < A < G$. This is as natural as any order one can hope to find among four physical objects.

Our interest in permutations of multisets was provoked by their connection to riffle shuffles of decks with repeated cards [5].

We do not give explicit numerical constants in our Theorems 2.12 and 2.16 about descents and inversions of permutations of multisets. It is worth noting that explicit numerical constants are not given for most of the detailed examples in Stein's book [20], and all the examples in the papers by Baldi, Rinott, and Stein [1], by Goldstein and Rinott [10], and by Rinott and Rotar [17]. Furthermore, even the asymptotic normality for descents of permutations of multisets implied by Theorem 2.16 is a new result, and so is Theorem 2.12 which shows the dependence of the bounds for normal approximation on the size of the multiset and the parameter $\alpha$ that characterizes the multiset.

## 2 Descents and Inversions of Permutations of Multisets

If $W \geq 0$ is a non-negative and integrable random variable, the distribution of $W^*$ is said to be $W$-size biased, if $\mathrm{E}(Wf(W)) = \mathrm{E}W\mathrm{E}(f(W^*))$ for all continuous functions $f$ for which the expectation on the left hand side of the equality exists.

Stein's method [19, 20] refers to the use of auxiliary randomization to find normal approximations to the distribution of some random variables. In the theorem below, the auxiliary randomization requires the construction of $W^*$ which must be $W$-size biased. The theorem below can be found in [1], but we follow its formulation in [10].

**Theorem 2.1** *Let $W$ be a non-negative random variable with $\mathrm{E}W = \mu$ and $\mathrm{Var}(W) = \sigma^2$. Let $W^*$ be jointly defined with $W$ such that its distribution is $W$-size biased. Let $h$ be a function from $R$ to $R$ such that $h$ is continuous and its derivative $Dh$ is piecewise continuous. Then*

$$\left|\mathrm{E}h\left(\frac{W-\mu}{\sigma}\right) - \Phi h\right| \leq 2\|h\|\frac{\mu}{\sigma^2}\sqrt{\mathrm{Var}(\mathrm{E}(W^*-W|W))} + \|Dh\|\frac{\mu}{\sigma^3}\mathrm{E}(W^*-W)^2,$$

*where $\Phi h$ is the expectation of $h$ with respect to the standard normal distribution and $\|\cdot\|$ is the supremum norm.*

When $h$ is the indicator function of the half line $(-\infty, x]$, the following theorem found in [9] applies. Its proof uses a smoothing inequality and other techniques found in [17].

**Theorem 2.2** *Let $W$ be a non-negative random variable with $\mathrm{E}W = \mu$ and $\mathrm{Var}(W) = \sigma^2$. Let $W^*$ be jointly defined with $W$ such that its distribution is $W$-size biased. Let $|W^* - W| \le B$ and let $A = B/\sigma$. Let $B \le \sigma^{3/2}/\sqrt{6\mu}$. Then*

$$\left|P\left(\frac{W-\mu}{\sigma} \le x\right) - \Phi(x)\right| \le 0.4A + \frac{\mu}{\sigma}(64A^2 + 4A^3) + \frac{23\mu}{\sigma^2}\sqrt{\mathrm{Var}(\mathrm{E}(W^* - W|W))},$$

*where $\Phi$ is the standard normal distribution.*

In Theorems 2.1 and 2.2 above, we added the superscript $*$ to $W$ to denote a random variable with the $W$-size biased distribution. In the lemma below, random variables $X_i$ with the superscript $*$ do not necessarily have the $X_i$-size biased distribution. Here and later, our convention is to use the superscript $*$ when random variables are constructed as a part of the size biasing procedure. This notation is due to [1].

The construction of size biased variables in this paper will be based on the following lemma found in [1] and [10].

**Lemma 2.3** *Let $W = X_1 + X_2 + \cdots + X_n$, where each $X_i$ is a non-negative random variable with finite mean. Let $I$ be a random variable which is independent of the $X_i$ and which satisfies $\mathrm{P}(I = i) = \mathrm{E}X_i / \sum_{j=1}^n \mathrm{E}X_j$. Define $W^*$ as $W^* = X_1^* + X_2^* + \cdots + X_n^*$, where for given $I$ $X_I^*$ has the $X_I$-size biased distribution and*

$$\mathrm{P}((X_1^*, X_2^*, \dots, X_n^*) \in A | I = i, X_i^* = x) = \mathrm{P}((X_1, X_2, \dots, X_n) \in A | X_i = x). \quad (2.1)$$

*Then $W^*$ has the $W$-size biased distribution.*

Whenever Lemma 2.3 is applied here, we will find $X_i$ are 0-1 valued random variables, and the size biased distribution for such variables is concentrated at 1. Therefore, for our purposes, (2.1) can be written as $\mathrm{P}((X_1^*, X_2^*, \dots, X_n^*) \in A | I = i) = \mathrm{P}((X_1, X_2, \dots, X_n) \in A | X_i = 1)$.

Let $\pi$ be a uniformly distributed permutation of the multiset $\{1^{n_1}, 2^{n_2}, \dots, h^{n_h}\}$ and $n = n_1 + n_2 + \cdots + n_h$. Each $n_a$, $1 \le a \le h$, is a positive integer. The symbols $i$, $j$, $k$, $l$, with and without numerical subscripts, are used to index the set $\{1, 2, \dots, n\}$. The symbols $a, b, c, d$ are used to index the set $\{1, 2, \dots, h\}$. We also assume $n_a \le \alpha n$ for $1 \le a \le h$ and for some $\alpha$ in $(0, 1)$, $n \ge 4$, and $h \ge 2$.

Define $X_{ij}$, for $i < j$, as 1 if $\pi(i) > \pi(j)$ and as 0 otherwise. Some facts about the joint distribution of $X_{ij}$ will be necessary. Denote the probabilities

$$\mathrm{P}(X_{ij} = 1) \quad \text{with } i < j,$$

$$\begin{aligned}
&\mathrm{P}(X_{ij_1}=1, X_{ij_2}=1) \quad \text{with } i<j_1 \text{ and } i<j_2,\\
&\mathrm{P}(X_{i_1j}=1, X_{i_2j}=1) \quad \text{with } i_1<j \text{ and } i_2<j,\\
&\mathrm{P}(X_{ik}=1, X_{kj}=1) \quad \text{with } i<k<j,\\
&\mathrm{P}(X_{i_1j_1}=1, X_{i_2j_2}=1) \quad \text{with } i_1<j_1, i_2<j_2, \text{ and } (i_1,j_1)\neq(i_2,j_2)
\end{aligned}$$

by $p_1$, $p_2$, $p_3$, $p_4$, and $p_5$, respectively. Elementary arguments can be used to deduce formulas, such as $p_1=\sum_{a<b} n_a n_b/(n(n-1))$ and $p_4=\sum_{a<b<c} n_a n_b n_c/(n(n-1)(n-2))$, for $p_1$, $p_2$, $p_3$, $p_4$, and $p_5$. From such formulas, we deduce

$$\begin{aligned}
p_1 &= \frac{n^2-\sum_a n_a^2}{2n(n-1)},\\
p_2+p_3+p_4 &= \frac{5n^3/6-n^2+(-3n/2+1)\sum_a n_a^2+(2/3)\sum_a n_a^3}{n(n-1)(n-2)},\\
p_4 &= \frac{n^3/6-(n/2)\sum_a n_a^2+(1/3)\sum_a n_a^3}{n(n-1)(n-2)},\\
p_5 &= \frac{n^4/4-n^3+n^2/2+(1/4)\big(\sum_a n_a^2\big)^2+(-n^2/2+2n-1/2)\sum_a n_a^2-\sum_a n_a^3}{n(n-1)(n-2)(n-3)}.
\end{aligned} \tag{2.2}$$

The formulas in (2.3) will be used to derive expressions for $\mathrm{Var}(\mathrm{inv}(\pi))$ and $\mathrm{Var}(\mathrm{des}(\pi))$.

The assumption $n_a \le \alpha n$, for some $\alpha\in(0,1)$, is used in the two lemmas below. The lemmas, however, are worded in terms of $\beta=\max(1/2,\alpha)$ and use the weaker assumption $n_a\le\beta n$ for $\beta\in[1/2,1)$. In both the lemmas the assumption $n_a\le\beta n$ implies $h\ge 2$.

**Lemma 2.4** *Assume $\beta\in[1/2,1)$, $n_a\ge 0$ for all $a$, and $\sum_a n_a=n$. If $n_a\le\beta n$ for $1\le a\le h$, then $2\beta(1-\beta)n^2\le n^2-\sum_a n_a^2\le n^2$ and $3\beta(1-\beta)n^3\le n^3-\sum_a n_a^3\le n^3$.*

*Proof* To lower bound $n^2-\sum_a n_a^2$, note that $x\ge y>0$, $\delta>0$, and $y-\delta\ge 0$ imply $(x+\delta)^2+(y-\delta)^2>x^2+y^2$. Thus for a given sum $x+y$, the quantity $x^2+y^2$ increases when the difference $x-y$ is increased. Thus given $\sum_a n_a=n$ and the constraints $n_a\ge 0$, the quantity $\sum_a n_a^2$ is increased whenever two positive numbers are chosen from $n_a$, $1\le a\le h$, and the lesser of them is decreased and the greater increased by the same amount. Therefore, under the constraints $n_a\le\beta n$, $\sum_a n_a^2$ is maximum when $n_1=\beta n$, $n_2=(1-\beta)n$, and $n_a=0$ for $a>2$. The lower bound for $n^3-\sum_a n_a^3$ is also obtained when $n_1=\beta n$, $n_2=(1-\beta)n$, and $n_a=0$ for $a>2$. The upper bounds are trivial. □

Concerning the lemma below, it is worth noting that $\beta^4-4\beta^4+4\beta-1=(1-\beta)^2(\beta^2+2\beta-1)>0$ for $\beta\in[1/2,1)$.

**Lemma 2.5** *Assume $\beta \in [1/2, 1)$, $n_a \geq 0$ for all $a$, and $\sum_a n_a = n$. If $n_a \leq \beta n$ for $1 \leq a \leq h$,*

$$(\beta^4 - 4\beta^2 + 4\beta - 1)n^4 \leq n^4/3 + \left(\sum_a n_a^2\right)^2 - (4n/3)\sum_a n_a^3 \leq n^4/3.$$

*Proof* The upper bound follows from the inequality $n\sum_a n_a^3 \geq (\sum_a n_a^2)^2$.

We prove the lower bound assuming $\beta > 1/2$. The proof for $\beta = 1/2$ can be obtained with minor changes. The proof will make careful use of the Kuhn-Tucker conditions as explained in [3, Theorem 9.2-3].

We attempt to minimize $J(n_1, n_2, \dots, n_h) = (\sum_a n_a^2)^2 - (4n/3)\sum_a n_a^3$ subject to the affine constraints $\sum_a n_a = n$, $-n_a \leq 0$, and $n_a - \beta n \leq 0$, where the last two constraints hold for $1 \leq a \leq h$. We assume $n_1 \geq n_2 \geq \cdots \geq n_h \geq 0$ without loss of generality.

Let $DJ$ be the gradient vector whose $a$th entry is

$$\frac{\partial J}{\partial n_a} = 4n_a \sum_b n_b^2 - 4nn_a^2 = 4n_a \sum_b n_b(n_b - n_a).$$

The sum of the entries of $DJ$ must be 0 because the term $4n_a n_b(n_b - n_a)$ in $\partial J/\partial n_a$ is canceled by the term $4n_a n_b(n_a - n_b)$ in $\partial J/\partial n_a$. If there exists an $a$ such that $n_1 > n_a > 0$, then the first entry of $DJ$ must be strictly negative and therefore some other entry must be strictly positive.

Let $u \in R^h$ be the vector with all entries equal to 1. Let $v_a \in R^h$ be the vector with its $a$th entry equal to $-1$ and all other entries equal to 0. Let $w_1 = -v_1$. Note that $n_a - \beta n = 0$ is possible only if $a = 1$ as we have assumed $\beta > 1/2$ and $n_1 \geq n_2 \geq \cdots$.

Suppose $(n_1, n_2, \dots, n_h)$ is a local minimum of $J$. The Kuhn-Tucker conditions require that it must be possible to make all entries of $DJ$ zero by adding multiples of certain vectors. We are always allowed to add any real multiple of $u$ because the constraint $\sum_a n_a = n$ is always in force. We are allowed to add a positive multiple of $v_a$ if and only if $n_a = 0$ because the constraint $-n_a \leq 0$ can then be violated by making an infinitesimal change to $n_a$. We are allowed to add a positive multiple of $w_1$ if and only if $n_1 - \beta n = 0$ by a similar reason.

Let us first consider the type of local minimum where the Kuhn-Tucker conditions can be satisfied without adding a positive multiple of $w_1$. Suppose $n_1 > n_a > 0$ for some $a$ for such a local minimum. Then the first entry of $DJ$ is strictly negative and some other entry is strictly positive. If the positive entry is $\partial J/\partial n_b$, then $n_b$ must be nonzero and therefore $n_b > 0$. Such a $DJ$ cannot be made zero by adding a multiple of $u$ and positive multiples of $v_c$ corresponding to $n_c = 0$. The only way to make the $b$th entry of $DJ$ equal to 0 is by adding a negative multiple of $u$. But this means the first entry remains negative and nonzero, and the only way to make it 0 is by adding a multiple of $w_1$ which is not allowed by assumption. Therefore any local minimum of this type must have $n_1 = n_2 = \cdots = n_s = n/s$ and $n_c = 0$ for $c > s$, where $2 \leq s \leq h$. The value of $J$ at such a point is $-n^4/(3s^2)$. Since $s \geq 2$,

$$J \geq -n^4/12 \tag{2.3}$$

at any local minimum of this type.

We next consider the type of local minimum where it is necessary to add a positive multiple of $w_1$ to satisfy the Kuhn-Tucker conditions. At such a local minimum $n_1 = \beta n$ and $n_1 > n_2 \geq \cdots \geq n_h \geq 0$. Suppose $n_h = 0$. Then the first entry of $DJ$ is strictly negative, some other entry is strictly positive, and the last entry is 0. We cannot make all those three entries zero by adding a real multiple of $u$, a positive multiple of $w_1$, and a positive multiple of $v_h$ to $DJ$. Thus $n_h > 0$. Next suppose that the $a$th entry of $DJ$ is not equal to the $b$th entry of $DJ$ for some $a, b > 1$. It is impossible to make the 1st, $a$th, and $b$th entries of $DJ$ zero by adding a multiple of $u$ and a positive multiple of $w_1$. Therefore, all entries of $DJ$ except the first must be equal. The expression for $\partial J/\partial n_a$ given above is quadratic in $n_a$. Thus we may conclude that at any local minimum of this type $n_1 = \beta n$ and $n_2, \ldots, n_h$ can take on at most two different values. Although the argument assumed $h \geq 3$, the conclusion holds when $h = 2$ as well. When $h = 2$ and a positive multiple of $w_1$ is added to DJ to satisfy the Kuhn-Tucker conditions, we must have $n_1 = \beta n$ and $n_2 = (1-\beta)n$.

We now consider the value of $J$ assuming that $n_1 = \beta n$, that $x$ of the $n_a$s equal $n_x$, that $y$ of the $n_a$s equal $n_y$, and that $xn_x + yn_y = n(1-\beta)$. We also assume that $x$ is a positive integer, that $y$ is a non-negative integer, that $x \geq y$, and of course that $n_x$ and $n_y$ are non-negative. Then

$$J = (\beta^4 n^4 - 4\beta^3 n^4/3) + (xn_x^2 + yn_y^2)^2 + 2\beta^2 n^2 (xn_x^2 + yn_y^2) - (4n/3)(xn_x^3 + yn_y^3),$$

which we will think of as a sum of four terms. If follows from elementary inequalities that the minimum of $xn_x^2 + yn_y^2$ under the given constraints is $n^2(1-\beta)^2/(x+y)$, and that the minimum of $-(4n/3)(xn_x^3 + yn_y^3)$ occurs when $n_x = n(1-\beta)$, $x = 1$, and $n_y = 0$. We can minimize each of the four terms of $J$ separately to obtain

$$J \geq \beta^4 n^4 - 4\beta^3 n^4/3 - 4(1-\beta)^3 n^4/3. \tag{2.4}$$

The value of $J$ at any local minimum of the type discussed in the previous paragraph must either equal or exceed the lower bound in (2.4).

So far, we have proved that the value of $J$ at a local minimum satisfies the lower bound given by either (2.3) or (2.4), depending upon the type of the local minimum. For $\beta \in [1/2, 1)$, $\beta^4 - 4\beta^3/3 - 4(1-\beta)^3/3 < -1/12$ by an elementary argument. Therefore the lower bound for $J$ given by (2.4) holds at all local minima and the lower bound for $J + n^4/3$ stated in the lemma is proved. □

### 2.1 Inversions of Permutations of Multisets

Let $W = \sum_{i<j} X_{ij}$. Then $W = \mathrm{inv}(\pi)$. We assume that $\pi$ is uniformly distributed over permutations of the multiset $\{1^{n_1}, 2^{n_2}, \ldots, h^{n_h}\}$.

**Lemma 2.6** *Let $\mu = \mathrm{E}W$ and $\sigma^2 = \mathrm{Var}(W)$. Then*

$$\mu = \frac{n^2 - \sum_a n_a^2}{4} \quad \text{and} \quad \sigma^2 = \binom{n}{3} \frac{n^5 - n^2 \sum_a n_a^3}{6n^2(n-1)^2(n-2)} + O(n^2).$$

*Proof* Since $\mu = \binom{n}{2} p_1$, where $p_1 = \mathrm{E}X_{ij}$, and $p_1$ is given by (2.3), the expression for $\mu$ in the lemma must hold.

We first show that

$$\sigma^2 = \binom{n}{2}(p_1 - p_1^2) + 2\binom{n}{3}(p_2 + p_3 + p_4 - 3p_1^2) + 6\binom{n}{4}(p_5 - p_1^2), \quad (2.5)$$

where the $p_i$ are given by (2.3). If $\mathrm{Var}(W)$ with $W = \sum_{i<j} X_{ij}$ is written as a sum of variances and covariances of the $X_{ij}$, there are $\binom{n}{2}$ variance terms each of which is equal to $p_1 - p_1^2$. There are $\binom{n}{3}$ terms of the form $2\mathrm{Covar}(X_{ij_1}, X_{ij_2})$ with $i < j_1 < j_2$ and each of those is equal to $2(p_2 - p_1^2)$. We can account for terms of the form $2\mathrm{Covar}(X_{i_1 j}, X_{i_2 j})$ with $i_1 < i_2 < j$ and of the form $2\mathrm{Covar}(X_{ik}, X_{kj})$ with $i < k < j$ similarly. Thus far we have explained the first two terms of (2.5). All the other terms in the expansion of $\mathrm{Var}(W)$ are of the form $2\mathrm{Covar}(X_{i_i j_1}, X_{i_2 j_2})$ with $i_1 < j_1$, $i_2 < j_2$, and $(i_1, j_1) < (i_2, j_2)$ in lexicographic order. The last term of (2.5) follows if we note that the number of such terms is $\binom{3n}{4}$.

The expression for $\sigma^2$ in the lemma is deduced using (2.3), (2.5), and the two inequalities $\sum_a n_a^2 < n^2$ and $\sum_a n_a^3 < n^3$. □

We now turn to the construction of the size biased variable $W^*$ required by Theorems 2.1 and 2.2. Let $I$ be uniformly distributed over all pairs $(i, j)$ with $1 \le i < j \le n$ and let it be independent of $\pi$. Let $J = (a, b)$, for $h \ge a \; > b \ge 1$, with probability $n_a n_b / \sum_{c<d} n_c n_d$, and let $J$ be independent of both $\pi$ and $I$. Now $\pi^*$ is constructed from $\pi$, $I$, and $J$ as follows. If $I = (i, j)$ and $\pi(i) > \pi(j)$, then $\pi^* = \pi$. If $I = (i, j)$, $\pi(i) \le \pi(j)$ and $J = (a, b)$, $\pi^*$ is constructed in the following steps:

1. Let $i^*$ and $j^*$ be uniformly distributed over the sets $\{i|\pi(i) = a\}$ and $\{j|\pi(j) = b\}$, respectively. They must be independent of each other and all other random variables.
2. If $\{i, j\} \cap \{i^*, j^*\} = \phi$, or $i = i^*$, $j \ne j^*$, or $i \ne i^*$, $j = j^*$, exchange $\pi(i)$ with $\pi(i^*)$ and $\pi(j)$ with $\pi(j^*)$ to get $\pi^*$.
3. If $i = j^*$, $j = i^*$, exchange $\pi(i)$ and $\pi(j)$ to get $\pi^*$.
4. If $i = j^*$, $j \ne i^*$, then $\pi^*(i) = \pi(i^*)$, $\pi^*(j) = \pi(j^*) = \pi(i)$, $\pi^*(i^*) = \pi(j)$, and $\pi^*(k) = \pi(k)$ if $k \ne i, j, i^*$.
5. If $i \ne j^*$, $j = i^*$, then $\pi^*(i) = \pi(i^*) = \pi(j)$, $\pi^*(j) = \pi(j^*)$, $\pi^*(j^*) = \pi(i)$, and $\pi^*(k) = \pi(k)$ for $k \ne i, j, j^*$.

Finally, $W^* = \sum_{i<j} X_{ij}^*$, where $X_{ij}^*$ is 1 if $\pi^*(i) > \pi^*(j)$ and 0 otherwise.

We prove below that $W^*$ has the $W$-size biased distribution. If $\pi$ were a uniformly distributed permutation of $\{1, 2, \ldots, n\}$, it would be enough to exchange $\pi(i)$ and $\pi(j)$ if $\pi(i) < \pi(j)$ to get $\pi^*$. The resulting $W^*$ would have the $W$-size biased distribution. However, since we are dealing with a multiset here, $\pi(i) = \pi(j)$ is also a possibility. The construction of $\pi^*$ given above is not as simple mainly because this possibility has to be dealt with.

The following lemma is needed to prove that $W^*$ has the $W$-size biased distribution. Subtraction and union of multisets have the obvious meanings in the statement of the lemma. The lemma is stated without proof.

**Lemma 2.7** *Let $\pi$ be a uniformly distributed permutation of the multiset $\{1^{n_1}, 2^{n_2}, \ldots, h^{n_h}\}$. If one $a$ out of $n_a$ possible choices is chosen uniformly from $\pi$ and changed to $b$, the resulting permutation is a uniformly distributed permutation of the multiset $(\{1^{n_1}, 2^{n_2}, \ldots, h^{n_h}\} - \{a\}) \cup \{b\}$. Similarly, if one of $n_a$ $a$s and one of $n_b$ $b$s are picked uniformly and independently from $\pi$ and changed to $c$ and $d$, respectively, then the resulting permutation is a uniformly distributed permutation of a possibly new multiset.*

**Lemma 2.8** *The random variable $W^*$ has the $W$-size biased distribution.*

*Proof* By Lemma 2.3, it is enough to show that $\mathrm{P}(\pi^* \in A|I = (i,j)) = \mathrm{P}(\pi \in A|\pi(i) > \pi(j))$. Now

$$\begin{aligned}\mathrm{P}(\pi^* \in A|I = (i,j)) &= \mathrm{P}(\pi \in A|\pi(i) > \pi(j))\mathrm{P}(\pi(i) > \pi(j)) \\ &\quad + \mathrm{P}(\pi^* \in A|\pi(i) \le \pi(j), I = (i,j))\mathrm{P}(\pi(i) \le \pi(j)).\end{aligned}$$

The first term in the right hand side of the equation above is not conditioned on $I$ because $\mathrm{P}(\pi^* \in A|\pi(i) > \pi(j), I = (i,j)) = \mathrm{P}(\pi \in A|\pi(i) > \pi(j), I = (i,j))$, by the construction of $\pi^*$, and because $\pi$ is independent of $I$. Thus, if we can show $\mathrm{P}(\pi^* \in A|\pi(i) \le \pi(j), I = (i,j)) = \mathrm{P}(\pi \in A|\pi(i) > \pi(j))$, the proof will be complete.

The proof is completed by the sequence of equalities below and the explanation that follows them.

$$\begin{aligned}&\mathrm{P}(\pi^* \in A|\pi(i) \le \pi(j), I = (i,j)) \\ &\quad = \sum_{a>b} \mathrm{P}(\pi^* \in A|\pi(i) \le \pi(j), I = (i,j), J = (a,b)) \\ &\qquad \times \mathrm{P}(\pi(i) = a, \pi(j) = b|\pi(i) > \pi(j)) \\ &\quad = \sum_{a>b} \mathrm{P}(\pi \in A|\pi(i) = a, \pi(j) = b, I = (i,j), J = (a,b)) \\ &\qquad \times \mathrm{P}(\pi(i) = a, \pi(j) = b|\pi(i) > \pi(j)) \\ &\quad = \mathrm{P}(\pi \in A|\pi(i) > \pi(j)).\end{aligned}$$

The first equality is true because $J$ is independent of $\pi$ and $I$, and $\mathrm{P}(J = (a,b)) = P(\pi(i) = a, \pi(j) = b|\pi(i) > \pi(j))$. The construction of $\pi^*$ from $\pi, I, J$ and Lemma 2.7 imply the second equality. More specifically, we note that Lemma 2.7 implies that given $\pi(i) \le \pi(j)$, $I = (i,j)$ and $J = (a,b)$, the arrangement $\pi^*(1), \pi^*(2), \ldots, \pi^*(n)$ with the $i$th and the $j$th numbers *struck out* is a uniformly distributed permutation of the multiset $\{1^{n_1}, 2^{n_2}, \ldots, h^{n_h}\} - \{a,b\}$. □

We now focus on finding a useful upper bound for $\mathrm{Var}(\mathrm{E}(W^* - W|\pi))$. Given a sequence of numbers $s_1, s_2, \ldots, s_p$, we throw $q$ and $r$ into the same set if and only if $s_q = s_r$. In this way, we get a partition of $\{1, 2, \ldots, p\}$ into sets, and we may arrange the sets of the partition so that the values of $s_q$ for $q$ in the set increase. We refer to such an ordered partition of $\{1, 2, \ldots, p\}$ as the *relative order* of $s_1, s_2, \ldots, s_p$.

For our purpose, it is sufficient to note that the number of possible relative orders is bounded by $2^p p!$.

**Lemma 2.9** *Let $P_1$ be the probability that $\pi(1), \pi(2), \dots, \pi(p)$ occur in a certain relative order when $\pi$ is a uniformly distributed permutation of $\{1^{n_1}, 2^{n_2}, \dots, h^{n_h}\}$, and let that probability be $P_2$ if $\pi$ is a uniformly distributed permutation of the multiset $\{1^{n'_1}, 2^{n'_2}, \dots, h^{n'_h}\}$. Assume that $n_a \ge n'_a$, $\sum_a (n_a - n'_a) \le 5$. We allow $n'_a = 0$. If $p \le 5$ then $|P_1 - P_2| \le C/n$ for some constant $C$.*

*Proof* The proof is obtained by writing down formulas for $P_1$ and $P_2$. We show the proof for the relative order $\pi(1) < \pi(2) < \dots < \pi(p)$.

Let $n' = \sum_a n'_a$. The probability $P_1$ is given by

$$\frac{\sum n_{a_1} n_{a_2} \dots n_{a_p}}{n(n-1)\dots(n-p+1)}, \tag{2.6}$$

where the sum is taken over $1 \le a_1 < a_2 < \dots < a_p \le h$. The formula for $P_2$ is obtained by adding a prime to all the $n$s in (2.6). Now

$$P_1 - P_2 = \frac{\sum n_{a_1} n_{a_2} \dots n_{a_p} - n'_{a_1} n'_{a_2} \dots n'_{a_p}}{n(n-1)\dots(n-p+1)} - P_2 \left( 1 - \frac{n'(n'-1)\dots(n'-p+1)}{n(n-1)\dots(n-p+1)} \right),$$

$0 \le n - n' \le 5$, and

$$\begin{aligned} &n_{a_1} n_{a_2} \dots n_{a_p} - n'_{a_1} n'_{a_2} \dots n'_{a_p} \\ &\quad \le n_{a_1} n_{a_2} \dots n_{a_p} ((n_{a_1} - n'_{a_1})/n_{a_1} + \dots + (n_{a_p} - n'_{a_p})/n_{a_p}), \end{aligned}$$

together imply $|P_1 - P_2| \le C/n$. □

**Lemma 2.10** *Let $f(\pi(1), \pi(2), \dots, \pi(p))$ and $g(\pi(p+1), \pi(p+2), \dots, \pi(p+q))$ be functions that depend only upon the relative order of their argument lists. Assume that $|f|$, $|g|$, $p$, and $q$ are all upper bounded by 5. If $\pi$ is a uniformly distributed permutation of the multiset $\{1^{n_1}, 2^{n_2}, \dots, h^{n_h}\}$, then*

$$\left| \mathrm{Covar}(f(\pi(1), \pi(2), \dots, \pi(p)), g(\pi(p+1), \pi(p+2), \dots, \pi(p+q))) \right| \le C/n$$

*for some constant $C$.*

*Proof* It is enough to consider $f$ and $g$ to be indicator functions that are 1 for a certain relative order of their argument lists and 0 for all other relative orders. All other $f$ and $g$ are linear combinations of a constant number of indicator functions with coefficients that are bounded by constants.

We state the proof assuming $f$ and $g$ are 1 if their arguments are in strictly increasing order and 0 otherwise. Let $\mathrm{P}(f=1) = P_1$ and $\mathrm{P}(g=1) = P_2$. Then

$$\begin{aligned} \mathrm{P}(fg = 1) = \sum &\mathrm{P}(\pi(1) < \pi(2) < \dots < \pi(p) | \pi(p+1) = a_1, \dots, \pi(p+q) = a_q) \\ &\times \mathrm{P}(\pi(p+1) = a_1, \dots, \pi(p+q) = a_q), \end{aligned}$$

where the sum is over $1 \le a_1 < a_2 < \cdots < a_q \le h$. By the previous Lemma 2.9, each conditional probability in the sum above is $P_1 + O(1/n)$. Therefore, $\mathrm{P}(fg = 1) = P_1 P_2 + O(1/n)$ and $\mathrm{Covar}(f, g) = O(1/n)$. $\square$

**Lemma 2.11**

$$\mathrm{Var}(\mathrm{E}(W^* - W|\pi)) \le \frac{Cn^5}{(n^2 - \sum_a n_a^2)^2}$$

*for some constant* $C$.

*Proof* If $\pi(i) > \pi(j)$, $\mathrm{E}\big(W^* - W\big|\pi, I = (i, j)\big) = 0$. If $\pi(i) \le \pi(j)$,

$$\mathrm{E}(W^* - W|\pi, I = (i, j)) = \frac{1}{\sum_{a>b} n_a n_b} \sum_{i^*, j^*} \sum_{l=1}^{n} \psi_\pi(i, j, i^*, j^*, l),$$

where $(i^*, j^*)$ takes all $\sum_{a>b} n_a n_b$ possible values with $\pi(i^*) > \pi(j^*)$ and $\psi_\pi(i, j, i^*, j^*, l)$ is the change in the number of inversions between position $l$ and positions $i, j, i^*, j^*$ when $\pi(i)$, $\pi(j)$, $\pi(i^*)$, $\pi(j^*)$ are exchanged to construct $\pi^*$. Note that $|\psi_\pi| \le 4$. We now have

$$\mathrm{E}(W^* - W|\pi) = \frac{1}{\binom{n}{2} \sum_{a>b} n_a n_b} \sum_{i,j} \sum_{i^*, j^*} \sum_{l=1}^{n} \psi_\pi(i, j, i^*, j^*, l), \tag{2.7}$$

where $i, j$ take all values satisfying $1 \le i < j \le h$ and $\pi(i) \le \pi(j)$, and where $i^*, j^*$ take values as already indicated.

We use (2.7) to write $\mathrm{Var}(\mathrm{E}(W^* - W|\pi))$ as a sum of variance and covariance terms. The number of variance terms is bounded by $n^5$. The number of covariance terms

$$\mathrm{Covar}(\psi_\pi(i_1, j_1, i_1^*, j_1^*, l_1), \psi_\pi(i_2, j_2, i_2^*, j_2^*, l_2)) \tag{2.8}$$

with $\{i_1, j_1, i_1^*, j_1^*, l_1\} \cap \{i_2, j_2, i_2^*, j_2^*, l_2\} \ne \phi$ is fewer than $25n^9$. Since $|\psi_\pi| \le 4$, the contribution of the variance terms and covariance terms with the property just described is bounded by $16(n^5 + 25n^9)/(\binom{n}{2}\frac{1}{2}(n^2 - \sum_a n_a^2))^2$. We have used $\sum_{a>b} n_a n_b = \frac{1}{2}(n^2 - \sum_a n_a^2)$ to obtain this bound.

Covariance terms of the form (2.8) with $\{i_1, j_1, i_1^*, j_1^*, l_1\} \cap \{i_2, j_2, i_2^*, j_2^*, l_2\} = \phi$ remain to be considered. The number of such terms is fewer than $n^{10}$. Lemma 2.10 can be applied to argue that such covariances are $O(1/n)$ as we may use the fact that $\pi$ is uniformly distributed to assume $i_1, j_1, i_1^*, j_1^*, l_1 = 1, 2, 3, 4, 5$ and $i_2, j_2, i_2^*, j_2^*, l_2 = 6, 7, 8, 9, 10$ with no loss of generality. The proof can now be easily completed. $\square$

**Theorem 2.12** *Let* $\pi$ *be a uniformly distributed permutation of the multiset* $\{1^{n_1}, 2^{n_2}, \ldots, h^{n_h}\}$, *where* $n_a \in Z^+$ *for* $1 \le a \le h$. *Assume that* $\alpha \in (0, 1)$ *is fixed and that* $n_a \le \alpha n$ *for* $1 \le a \le h$. *Let* $\beta = \max(1/2, \alpha)$. *Let* $h : R \to R$ *be a bounded continuous function with bounded piecewise continuous derivative* $Dh$. *Then for*

$n > n_0(\beta)$,

$$\left|\mathrm{E}h\left(\frac{\mathrm{inv}(\pi)-\mu}{\sigma}\right)-\Phi h\right| \le C\left(\frac{\|h\|}{\beta(1-\beta)(\beta(1-\beta)n^{1/2}-C_1 n^{-1/2})} + \frac{\|Dh\|}{(\beta(1-\beta)n^{1/3}-C_2 n^{-2/3})^{3/2}}\right)$$

*where $C$, $C_1$, and $C_2$ are some positive constants, $\Phi h$ is the expectation of $h$ with respect to the standard normal distribution, and $\mu$ and $\sigma^2$ are the mean and variance of* $\mathrm{inv}(\pi)$, *respectively.*

*If $C(\beta)$ is allowed to depend upon $\beta$, we may assert*

$$\left|\mathrm{P}\left(\frac{\mathrm{inv}(\pi)-\mu}{\sigma}\le x\right)-\Phi(x)\right| \le C(\beta)/\sqrt{n}$$

*for some positive constant $C(\beta)$.*

*Proof* Let $W = \mathrm{inv}(\pi)$. By Lemmas 2.4 and 2.6, $\sigma^2 \ge (\beta(1-\beta)/12)n^3 + O(n^2)$ and $\mu \le n^2/4$. By Lemmas 2.4 and 2.11, $\mathrm{Var}(\mathrm{E}(W^* - W|\pi)) \le Cn/(\beta(1-\beta))^2$ for some constant $C$. By construction of the size biased variable $W^*$, $|W^* - W| \le 4n$, and therefore $\mathrm{E}(W^* - W)^2 \le 16n^2$. If we note that $\mathrm{Var}(\mathrm{E}(W^* - W|W)) \le \mathrm{Var}(\mathrm{E}(W^* - W|\pi))$, Theorem 2.1 can be applied to prove the first part of this theorem.

The second part is proved using Theorem 2.2. By construction of $W^*$, $|W^* - W| \le 4n$. Therefore we can take $B = 4n$. The inequality $B \le \sigma^{3/2}/\sqrt{6\mu}$ must hold for large enough $n$ by bounds for $\sigma$ and $\mu$ given above. □

### 2.2 Descents of Permutations of Multisets

Let $W = X_{12} + X_{23} + \cdots + X_{n-1,n}$. Then $W = \mathrm{des}(\pi)$, with $\pi$ uniformly distributed over permutations of the multiset $\{1^{n_1}, 2^{n_2}, \dots, h^{n_h}\}$.

**Lemma 2.13** *Let $\mu = \mathrm{E}W$ and $\sigma^2 = \mathrm{Var}(W)$. Then*

$$\mu = \frac{n^2 - \sum_a n_a^2}{2n} \quad \textit{and} \quad \sigma^2 = \frac{n^4/3 + (\sum_a n_a^2)^2 - (4n/3)\sum_a n_a^3}{4n(n-1)^2} + O(1).$$

*Proof* Since $\mu = (n-1)p_1$, where $p_1 = \mathrm{E}X_{ij}$, and $p_1$ is given by (2.3), the expression for $\mu$ in the lemma must hold.

We first show that

$$\sigma^2 = (n-1)(p_1 - p_1^2) + 2(n-2)(p_4 - p_1^2) + (n-2)(n-3)(p_5 - p_1^2), \qquad (2.9)$$

where the $p_i$ are given by (2.3). If $\mathrm{Var}(W)$, with $W = X_{12} + X_{23} + \cdots + X_{n-1,n}$, is written as the sum of variances and covariances of the $X_{i,i+1}$, there are $(n-1)$ variance terms, each equal to $p_1 - p_1^2$. There are $(n-2)$ covariance terms of the form $\mathrm{Covar}(X_{i,i+1}, X_{i+1,i+2})$ each equal to $p_4 - p_1^2$. The remaining covariance terms are all equal to $p_5 - p_1^2$.

The expression for $\sigma^2$ in the lemma is deduced using (2.3), (2.9), and the two inequalities $\sum_a n_a^2 < n^2$ and $\sum_a n_a^3 < n^3$. □

The construction of the size biased variable $W^*$ is the same as the construction for inversions given immediately after Lemma 2.6 with the following differences. The random variable $I$ must be equal to one of $(1,2), (2,3), \ldots, (n-1, n)$ with equal probability. In the construction of $\pi^*$, the symbol $j$ must be replaced everywhere by $i+1$. Finally, $W^* = X_{12}^* + X_{23}^* + \cdots + X_{n,n-1}^*$, where $X_{ij}^*$ is 1 if $\pi^*(i) > \pi^*(j)$ and 0 otherwise.

**Lemma 2.14** *The random variable $W^*$ has the $W$-size biased distribution.*

*Proof* Similar to the proof of Lemma 2.8. □

**Lemma 2.15**

$$\mathrm{Var}(\mathrm{E}(W^* - W|\pi)) \le \frac{Cn^5}{n^2(n^2 - \sum_a n_a^2)^2}$$

*for some constant $C$.*

*Proof* By arguing as in the proof of Lemma 2.11, we get

$$\mathrm{E}(W^* - W|\pi) = \frac{2}{n(n^2 - \sum_a n_a^2)} \sum_i \sum_{i^*, j^*} \psi_\pi(i, i^*, j^*). \tag{2.10}$$

In (2.10), $i$ takes all values such that $\pi(i) \le \pi(i+1)$, $(i^*, j^*)$ takes all values such that $\pi(i^*) > \pi(j^*)$, and $\psi_\pi(i, i^*, j^*) = \mathrm{des}(\pi^*) - \mathrm{des}(\pi)$, where $\pi^*$ is constructed by exchanging $\pi(i), \pi(i+1), \pi(i^*), \pi(j^*)$ as described. Note that $|\psi_\pi| \le 7$.

We use (2.10) to write $\mathrm{Var}(\mathrm{E}(W^* - W|\pi))$ as the sum of variance and covariance terms. There are $O(n^3)$ variance terms of the $\mathrm{Var}(\psi_\pi)$. The number of terms of the form

$$\mathrm{Covar}(\psi_\pi(i_1, i_1^*, j_1^*), \psi_\pi(i_2, i_2^*, j_2^*)), \tag{2.11}$$

where one of the numbers $\{i_1, i_1^*, j_1^*\}$ differs from one of the numbers $\{i_2, i_2^*, j_2^*\}$ by 3 or less in magnitude is $O(n^5)$. The magnitude of such covariance terms and of the variance terms is bounded by 49. The number of covariance terms of the form (2.11) where none of the numbers $\{i_1, i_1^*, j_1^*\}$ differs from any one of the numbers $\{i_2, i_2^*, j_2^*\}$ by 3 or less in magnitude is $O(n^6)$. By Lemma 2.10, the magnitude of such covariance terms is $O(1/n)$. The proof is now easily completed. □

It is worth noting again that $\beta^4 - 4\beta^2 + 4\beta - 1 = (1-\beta)^2(\beta^2 + 2\beta - 1) > 0$ for $\beta \in [1/2, 1)$.

**Theorem 2.16** *Let $\pi$ be a uniformly distributed permutation of the multiset $\{1^{n_1}, 2^{n_2}, \ldots, h^{n_h}\}$, where $n_a \in Z^+$ for $1 \le a \le h$. Assume that $\alpha \in (0,1)$ is fixed and that $n_a \le \alpha n$ for $1 \le a \le h$. Let $\beta = \max(1/2, \alpha)$. Let $h : R \to R$ be a bounded* *continuous function with bounded piecewise continuous derivative $Dh$. Then for $n > n_0(\beta)$,*

$$\left|\mathrm{E}h\left(\frac{\mathrm{des}(\pi)-\mu}{\sigma}\right)-\Phi h\right| \le C\left(\frac{\|h\|}{\beta(1-\beta)\sqrt{n}(\beta^4-4\beta^2+4\beta-1-C_1n^{-1})}\right.$$
$$\left.+\frac{\|Dh\|}{((\beta^4-4\beta^2+4\beta-1)n^{1/3}-C_2n^{-2/3})^{3/2}}\right),$$

*where $C$, $C_1$, and $C_2$ are some positive constants, $\Phi h$ is the expectation of $h$ with respect to the standard normal distribution, and $\mu$ and $\sigma^2$ are the mean and variance of* $\mathrm{des}(\pi)$, *respectively.*

*If $C(\beta)$ is allowed to depend upon $\beta$, we may assert*

$$\left|\mathrm{P}\left(\frac{\mathrm{des}(\pi)-\mu}{\sigma}\le x\right)-\Phi(x)\right| \le C(\beta)/\sqrt{n}$$

*for some positive constant $C(\beta)$.*

*Proof* Let $W = \mathrm{des}(\pi)$. By Lemmas 2.4, 2.5 and 2.13 $\sigma^2 \ge ((\beta^4 - 4\beta^2 + 4\beta - 1)/4)n + O(1)$ and $\mu \le n/2$. By Lemmas 2.4 and 2.15, $\mathrm{Var}(\mathrm{E}(W^* - W|\pi)) \le C/(n\beta^2(1-\beta)^2)$ for some constant $C$. By construction of the size biased variable $W^*$, $|W^* - W| \le 7$, and therefore $\mathrm{E}(W^* - W)^2 \le 49$. If we note that $\mathrm{Var}(\mathrm{E}(W^* - W|W)) \le \mathrm{Var}(\mathrm{E}(W^* - W\big|\pi))$, Theorem 2.1 can be applied to prove the first part of this theorem.

The second part is proved using Theorem 2.2. By construction of $W^*$, $|W^* - W| \le 8$. Therefore we can take $B = 8$. The inequality $B \le \sigma^{3/2}/\sqrt{6\mu}$ must hold for large enough $n$ by bounds for $\sigma$ and $\mu$ given above. □

## 3 Descents and Inversions of the Human Genome

The human genome consists of 24 chromosomes, each of which is a sequence of bases labeled A, C, G, or T. The 19th chromosome has the following counts for the four bases (see [6]):

$$n_A = 14383026, \qquad n_C = 13473774, \qquad n_G = 13506612, \qquad n_T = 14422243.$$

The version of the human genome reported in [6] has 341 gaps. The 19th chromosome has only three gaps in the middle. We ignored these gaps when counting the number of inversions and descents.

From Lemmas 2.6 and 2.13, and their proofs, we find the expected number of descents and inversions to be $\mu_d = 2.0912146861 \times 10^7$ and $\mu_i = 5.8329890505 \times 10^{14}$, respectively. The standard deviations are $\sigma_d = 2.0871959423 \times 10^3$ and $\sigma_i = 6.7231321079 \times 10^{10}$. Data about the 19th chromosome reported in Table 1 can be used to calculate the number of descents and inversions for any ordering of A, C, G, and T. By Theorems 2.12 and 2.16, the number of descents and inversions must have a distribution that is close to the normal distribution if $\pi$ is a uniformly distributed permutation of the bases in the 19th chromosome. The number of descents and

**Table 1** The first table above reports the number of occurrences of $\pi(i) = x$ and $\pi(i+1) = y$. The second table reports the number of occurrences of $\pi(i) = x$ and $\pi(j) = y$, with $i < j$. The permutation $\pi$ corresponds to chromosome 19, and $x$ and $y$ can be A, C, G, or T

| | A | C | G | T |
|---|---|---|---|---|
| A | 4229414 | 2833985 | 4154304 | 3165323 |
| C | 4221129 | 4044958 | 1057112 | 4150574 |
| G | 3423863 | 3180474 | 4056078 | 2846197 |
| T | 2508620 | 3414357 | 4239118 | 4260148 |

| | A | C | G | T |
|---|---|---|---|---|
| A | 103435711266825 | 94175991781325 | 94404662110136 | 103982892949612 |
| C | 99617649978799 | 90771286164651 | 90984870248490 | 100143945584446 |
| G | 99861289457776 | 91000167345198 | 91214277105966 | 100388853252680 |
| T | 103452603097706 | 94178097170636 | 94406787118036 | 104000539364403 |

**Table 2** This table reports the normalized number of descents and inversions of the 19th chromosome, when the orders shown in the first and the fourth columns are considered increasing

| Order | $(\text{des} - \mu_d)/\sigma_d$ | $(\text{inv} - \mu_i)/\sigma_i$ | Order | $(\text{des} - \mu_d)/\sigma_d$ | $(\text{inv} - \mu_i)/\sigma_i$ |
|---|---|---|---|---|---|
| A, C, G, T | 36.13 | −11.64 | C, A, G, T | −628.47 | −92.58 |
| G, A, C, T | −631.23 | −93.03 | A, G, C, T | −981.20 | −11.86 |
| C, G, A, T | −278.50 | −173.74 | G, C, A, T | −1295.83 | −173.96 |
| T, C, A, G | −628.47 | 93.03 | C, T, A, G | −981.20 | 4.29 |
| A, T, C, G | −278.50 | 166.08 | T, A, C, G | 36.13 | 173.96 |
| C, A, T, G | −1295.83 | −3.60 | A, C, T, G | −631.23 | 77.34 |
| A, G, T, C | −628.47 | 76.87 | G, A, T, C | −278.50 | −4.29 |
| T, A, G, C | −981.20 | 173.74 | A, T, G, C | −1295.83 | 165.85 |
| G, T, A, C | 36.13 | 3.60 | T, G, A, C | −631.23 | 92.58 |
| T, G, C, A | −1295.83 | 11.64 | G, T, C, A | −628.47 | −77.34 |
| C, T, G, A | −631.23 | −76.87 | T, C, G, A | −278.50 | 11.86 |
| G, C, T, A | −981.20 | −166.08 | C, G, T, A | 36.13 | −165.85 |

inversions in the 19th chromosome itself is reported in Table 2 for all possible orderings of $A$, $C$, $G$, and $T$ and with suitable normalization. From each line of this table, we may infer that the null hypothesis stating the 19th chromosome to be a random permutation of its bases is very unlikely to hold.

Estimations of the entropy of DNA sequences can be found in [11] and [16]. Those estimates too imply that DNA sequences are far from random. We note that Table 2 assumes the number of $A$s, $C$s, $G$s, and $T$s to be given and computes a statistic to test if their arrangement in a sequence is random. This is a different notion of randomness from that of entropy. For instance, it is possible for a sequence to have $A$ for 90% of its letters which would mean that the sequence can be significantly compressed. Yet the arrangement of the letters could be generated randomly.

In the bounds given by Theorems 2.12 and 2.16 the constants $C(\beta)$ are not determined explicitly. In this example $n$ is greater than $5 \times 10^7$. For the large departures from the mean that are seen in Table 2, it is reasonable to assume that the probabilities of finding such departures, if the sequence were a uniformly distributed permutation, are less than .001. Such a bound is implied in most cases by Chebyshev's inequality. Yet even this is surely an overestimate. For uniformly distributed $\pi$, the probabilities that $\mathrm{des}(\pi)$ and $\mathrm{inv}(\pi)$ depart from their means by a certain amount appear to fall off at least as fast as the bell curve does away from zero. Therefore, for large deviations from the mean, the bounds given by Theorems 2.12 and 2.16 are not accurate and better bounds would be desirable.

**Acknowledgements** The authors thank D. Burns and J. Fulman for helpful discussions, and the referees for their work.

## References

1. Baldi, P., Rinott, Y., Stein, C.: A normal approximation for the number of local maxima of a random function on a graph. In: Probability, Statistics, and Mathematics: Papers in Honor of Samuel Karlin, pp. 59–81. Academic (1989)
2. Barton, D.E., Mallows, C.L.: Some aspects of random sequence. Ann. Math. Stat. **36**, 236–260 (1965)
3. Ciarlet, P.G.: Introduction to Numerical Linear Algebra and Optimization. Cambridge University Press, Cambridge (1989)
4. Conger, M.: A refinement of Eulerian numbers, and the joint distribution of $\pi(1)$ and $\mathrm{Des}(\pi)$ in $s_n$. Ars Comb., to appear. Available at www.arxiv.org
5. Conger, M., Viswanath, D.: Riffle shuffles of decks with repeated cards. Ann. Probab. **34**, 804–819 (2006)
6. International Human Genome Consortium: Finishing the euchromatic sequence of the human genome. Nature **431**, 931–945 (2004)
7. Diaconis, P.: Group Representations in Probability and Statistics. Institute of Mathematical Statistics, USA (1988)
8. Fulman, J.: Stein's method and non-reversible Markov chains. In: Stein's Method: Expository Lectures and Applications, pp. 66–77. Institute of Mathematical Statistics, Hayward (2004)
9. Goldstein, L.: Berry Esseen bounds for combinatorial central limit theorems and pattern occurences, using zero and size biasing. J. Appl. Probab. **42**, 661–683 (2005)
10. Goldstein, L., Rinott, Y.: Multivariate normal approximations by Stein's method and size bias couplings. J. Appl. Probab. **33**, 1–17 (1996)
11. Gusev, V.D., Nemytikov, L.A., Chuzanova, N.A.: On the complexity measure of genetic sequences. Bioinformatics **15**, 994–999 (1999)
12. Knuth, D.E.: Seminumerical Algorithms, The Art of Computer Programming, vol. 2, 3rd edn. Addison-Wesley, Massachusetts (1997)
13. Knuth, D.E.: Searching and Sorting, The Art of Computer Programming, vol. 3, 3rd edn. Addison-Wesley, Massachusetts (1998)
14. Levene, H., Wolfowitz, J.: The covariance matrix of runs up and down. Ann. Math. Stat. **15**, 58–69 (1944)
15. MacMahon, P.A.: Combinatory Analysis, vol. 1. Cambridge University Press, Cambridge (1915)
16. Mantegna, R.N., et al.: Linguistic features of noncoding dna. Phys. Rev. Lett. **15**, 3169–3172 (1994)
17. Rinott, Y., Rotar, V.: On coupling constructions and rates in the CLT for dependent summands with applications to the antivoter model and U-statistics. Ann. Appl. Probab. **7**, 1080–1105 (1997)
18. Stanley, R.: Eulerian partitions of a unit hypercube. In: Aigner, M. (ed.) Higher Combinatorics. Reidel, Dordrecht (1977)
19. Stein, C.: A bound for the error in the normal approximation of the distribution of a sum of dependent random variables. In: Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, vol. 2, pp. 583–602 (1972)
20. Stein, C.: Approximate Computation of Expectations. Institute of Mathematical Statistics, Hayward (1986)
21. Wolfowitz, J.: Asymptotic distribution of runs up and down. Ann. Math. Stat. **15**, 163–172 (1944)