Journal of Integer Sequences, Vol. 29 (2026),
Article 26.1.5

Integer Sequences for Diversity Statistics

Bradley K. Moon and Noah A. Rosenberg
Department of Biology
Stanford University
Stanford, CA 94305
USA
bradmoon@stanford.edu
noahr@stanford.edu

Abstract

Consider a discrete set of objects and a sample of size N taken with replacement from the set, producing a list of counts of the objects that corresponds to a partition of N. Two statistics that are commonly used for measuring the "diversity" of the sample are the Gini-Simpson index and the Shannon index. We study the number of possible values that these indices can take across all possible partitions of the sample size N as N increases. The two statistics are highly correlated over the set of partitions of N. However, the number of possible values that the Shannon index can take (A383683) far exceeds the number of possible values of the Gini-Simpson index (A069999), with the latter growing quadratically and the former growing faster than every polynomial.

1 Introduction

How does one measure the "diversity" of a set of objects drawn from a finite set of classes? The question arises in many contexts, including ecology [10], genetics [12], health services research [16], and the social sciences [5]. Consider a discrete set D of objects $\{1, 2, \ldots, I\}$, with $I \geq 1$, and N draws from D with replacement, producing counts $(n_1, n_2, \ldots, n_I)$ with $\sum_{i=1}^{I} n_i = N$. The ordered vector of counts is a *sample*. A *diversity index* is a real-valued function that is computed from a sample and that seeks to capture the "diversity" of the sample. Typically, a diversity index has its lowest value if $n_i = N$ for some value of i (each draw is the same as all the others), and its highest value if $n_i = 1$ for N distinct values of i (each draw is distinct from all the others).

Two frequently used indices of diversity are the *Gini-Simpson index* $H = 1-\sum_{i=1}^{I}(n_i/N)^2$, and the *Shannon index* $S = -\sum_{i=1}^{I}(n_i/N)\log(n_i/N)$ [9, pp. 39, 94]. The Gini-Simpson index measures the probability that two objects randomly drawn with replacement from a probability distribution with frequencies n_i/N are distinct. In population genetics, it is known as *heterozygosity*, as it computes the probability that a genotype is heterozygous—that it contains two distinct alleles—if the allele frequencies in an allele frequency distribution have values $n_1/N, n_2/N, \ldots, n_I/N$. The Shannon index is the information-theoretic measure of the entropy of the probability distribution specified by $(n_1/N, n_2/N, \ldots, n_I/N)$. Across all possible samples of size N, the Gini-Simpson index has the property that its lowest value is 0 if $n_i = N$ for some i, and if $I \geq N$, its highest value is $1 - \frac{1}{N}$ if $n_i = 1$ for N distinct values of i. The Shannon index also has its lowest value of 0 if $n_i = N$ for some i. If $I \geq N$, its highest value is $\log N$ if $n_i = 1$ for N distinct values of i.

Considering only the counts in a sample and not the object labels $1, 2, \ldots, I$—that is, disregarding the order of entries in its vector—a sample is simply a partition of N. Provided that the number of distinct objects I satisfies $I \geq N$, the set of all possible samples of size N corresponds to the partitions of N, where we understand each partition of N to be a non-decreasing sequence of positive integers summing to N [1, Chapter 1] (appending zeroes to extend partitions to possess length I where necessary).

Multiple partitions of N can potentially produce the same value of the Gini-Simpson index. For example, for $N = 6$, partitions $(4, 1, 1)$ and $(3, 3)$ both produce $H = \frac{1}{2}$. Similarly, multiple partitions can produce the same value of the Shannon index. With $N = 8$, partitions $(4, 1, 1, 1, 1)$ and $(2, 2, 2, 2)$ both produce $S = 2\log 2$. Because distinct partitions map to the same value of a diversity statistic, the value of the statistic does not necessarily indicate that the underlying sample possesses a specific feature, such as a specific value of its largest entry, or a specific number of nonzero entries.

In analyses of discrete data, the Gini-Simpson and Shannon indices are used in similar ways, and they are seen to have high empirical correlation coefficients. Considering all 627 partitions of $N = 20$, the Pearson correlation coefficient [6, p. 117] of the Gini-Simpson and Shannon indices is approximately 0.927. As we will see here, however, despite their conceptual similarity, as N grows, the number of possible values of the Shannon index far exceeds the number of possible values of the Gini-Simpson index. We study the sequences describing the number of possible values of the Gini-Simpson index and the number of possible values of the Shannon index. In particular, we review results concerning the sequence for the Gini-Simpson index and discuss new results for the sequence for the Shannon index. We also briefly discuss implications for the use of diversity statistics.

2 Relationship between the Gini-Simpson and Shannon indices

We first illustrate that the Gini-Simpson and Shannon indices have a close relationship. In particular, we show that it is possible to find bounds on the Shannon index given knowledge of only the Gini-Simpson index and the number of distinct objects I in the set D. For this computation, we work with a vector of nonnegative real numbers p_i with $\sum_{i=1}^{I} p_i = 1$. The Gini-Simpson index is $1 - \sum_{i=1}^{I} p_i^2$, and the Shannon index is $-\sum_{i=1}^{I} p_i \log p_i$. The analysis generalizes beyond the case in which the nonnegative real numbers $(p_1, p_2, \ldots, p_I)$ can be written $(n_1/N, n_2/N, \ldots, n_I/N)$ for integers $n_1, n_2, \ldots, n_I$ and N, with $N \leq I$.

Proposition 1 ([15, Thm. 2.7])**.** *Let $I \geq 2$. For all length-I vectors $\mathbf{p} = (p_1, p_2, \ldots, p_I)$ of nonnegative real numbers with $\sum_{i=1}^{I} p_i = 1$ and at least two nonzero entries, writing $J = \sum_{i=1}^{I} p_i^2$ and $M = \max(p_1, p_2, \ldots, p_I)$, the following inequality holds:*

$$\frac{1}{\lceil J^{-1} \rceil}\left(1 + \sqrt{\frac{\lceil J^{-1} \rceil J - 1}{\lceil J^{-1} \rceil - 1}}\right) \leq M \leq \frac{1}{I}\Big(1 + \sqrt{(IJ-1)(I-1)}\Big).$$

It is convenient to define functions

$$M_{\min}(J) = \frac{1}{\lceil J^{-1} \rceil}\left(1 + \sqrt{\frac{\lceil J^{-1} \rceil J - 1}{\lceil J^{-1} \rceil - 1}}\right), \tag{1}$$

$$M_{\max}(J) = \frac{1}{I}\Big(1 + \sqrt{(IJ-1)(I-1)}\Big). \tag{2}$$

These functions are monotonically increasing on $[\frac{1}{I}, 1)$ ([15, Lemma 2.9]).

Proposition 2 ([3, Cor. 3.16])**.** *Let $I \geq 2$. For all length-I vectors $\mathbf{p} = (p_1, p_2, \ldots, p_I)$ of nonnegative real numbers with $\sum_{i=1}^{I} p_i = 1$ and at least two nonzero entries, writing $S = -\sum_{i=1}^{I} p_i \log p_i$ and $M = \max(p_1, p_2, \ldots, p_I)$, the following inequality holds:*

$$\lfloor M^{-1} \rfloor M \log \frac{1}{M} + (1 - \lfloor M^{-1} \rfloor M) \log\left(\frac{1}{1 - \lfloor M^{-1} \rfloor M}\right) \leq S \leq M \log \frac{1}{M} + (1-M)\log\left(\frac{I-1}{1-M}\right).$$

Next, note that the functions in the lower and upper bounds in Proposition 2 are monotonically decreasing functions of M on the permissible domain $M \in [\frac{1}{I}, 1)$. To do so, write

$$S_{\min}(M) = \lfloor M^{-1} \rfloor M \log \frac{1}{M} + (1 - \lfloor M^{-1} \rfloor M) \log\left(\frac{1}{1 - \lfloor M^{-1}} M\rfloor\right), \tag{3}$$

$$S_{\max}(M) = M \log \frac{1}{M} + (1-M)\log\left(\frac{I-1}{1-M}\right). \tag{4}$$

For the monotonicity of the upper bound, we have $dS_{\max}(M)/dM = -\log\big(\frac{M(I-1)}{1-M}\big)$. Because $M \geq \frac{1}{I}$, $I \geq 2$, and $M < 1$, we have $\frac{M(I-1)}{1-M} \geq 1$, and $dS_{\max}(M)/dM \leq 0$ with equality if and only if $M = \frac{1}{I}$, so that $S_{\max}(M)$ is monotonically decreasing on $[\frac{1}{I}, 1)$.

For the lower bound, it suffices to verify that $S_{\min}(M)$ is decreasing on intervals $E_k = (\frac{1}{k+1}, \frac{1}{k})$ for $1 \leq k \leq I-1$, where $\lfloor M^{-1} \rfloor$ has the fixed value k, as $S_{\min}(M)$ is continuous at the interval boundaries $M = \frac{1}{2}, \frac{1}{3}, \ldots, \frac{1}{I-1}$ (and at $M = \frac{1}{I}$). On intervals E_k, we have that $dS_{\min}(M)/dM = -k\log(\frac{M}{1-kM})$, a negative quantity as $kM < 1$ and $\frac{M}{1-kM} > 1$ on E_k.

Theorem 3. *Let $I \geq 2$. For all length-I vectors $\mathbf{p} = (p_1, p_2, \ldots, p_I)$ of nonnegative real numbers with $\sum_{i=1}^{I} p_i = 1$ and at least two nonzero entries, if the Gini-Simpson index $H = 1 - \sum_{i=1}^{I} p_i^2$ is given, $0 < H \leq 1 - \frac{1}{I}$, then the Shannon index $S = -\sum_{i=1}^{I} p_i \log p_i$ satisfies*

$$S_{\min}\big(M_{\max}(1-H)\big) \leq S \leq S_{\max}\big(M_{\min}(1-H)\big),$$

where functions $M_{\min}$, $M_{\max}$, $S_{\min}$, and $S_{\max}$ are defined in Eqs. (1)-(4).

Proof. Because $S_{\min}(M)$ is monotonically decreasing for $M \in [\frac{1}{I}, 1)$, to find a lower bound for S, we can set M to its largest possible value given $H = 1 - J$, or $M_{\max}(1-H)$. Similarly, because $S_{\max}(M)$ is also monotonically decreasing for $M \in [\frac{1}{I}, 1)$, to find an upper bound for S, we can set M to its smallest possible value given $H = 1 - J$, or $M_{\min}(1-H)$. □

For the case of $I = 40$, Figure 1 plots the bounds from Theorem 3 on the Shannon index in relation to the Gini-Simpson index. The bounds are loose, but the plot nevertheless illustrates the close relationship between the Shannon and Gini-Simpson indices.

3 Numerical comparisons

Although the Gini-Simpson and Shannon indices are closely related in their numerical values computed for samples, as illustrated via the bounds in Section 2, we will see that the numbers of *discrete* values of the two indices have different behavior as the sample size N increases.

Table 1 provides the numbers of values of the Gini-Simpson and Shannon indices across all possible partitions of N for small values of N. Even for small N, we see that the number of possible values of the Gini-Simpson index is substantially lower than the number of possible values of the Shannon index—which is in turn less than the number of partitions. Sequence numbers refer to the On-Line Encyclopedia of Integer Sequences (OEIS) [13].

We next demonstrate a faster rate of increase of the number of values of the Shannon index relative to the number of values of the Gini-Simpson index, as N increases. Let $\gamma(N)$ denote the number of possible values of the Gini-Simpson index across all samples of size N. Let $\sigma(N)$ denote the corresponding number of possible values of the Shannon index.

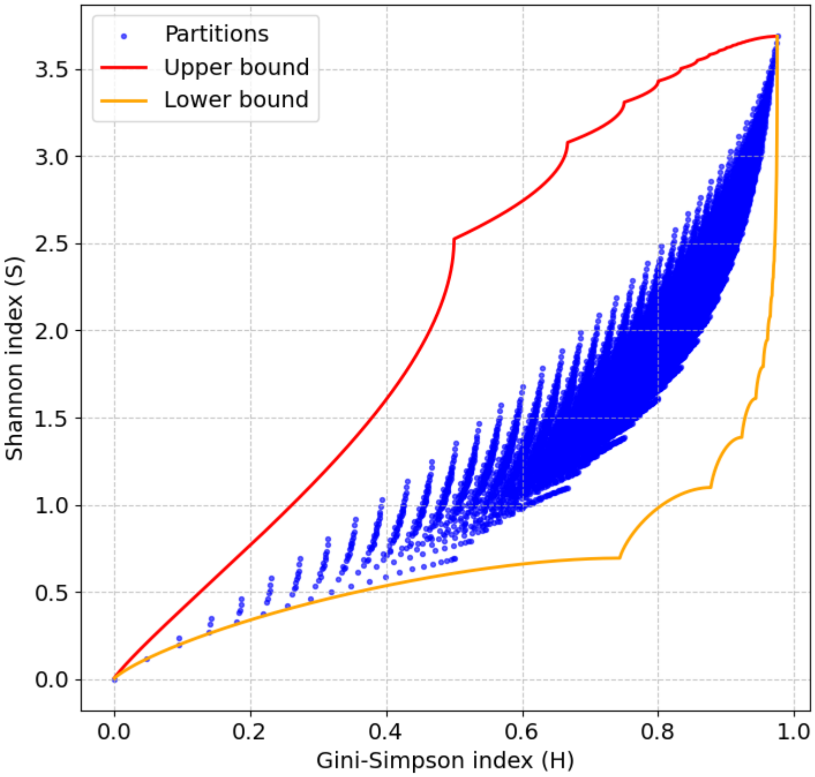

Figure 1: Upper and lower bounds on the Shannon index in relation to the Gini-Simpson index, for the case of $I = 40$. The bounds follow Theorem 3. The points show (H, S) for the 37,338 partitions of 40.

4 Number of possible values of the Gini-Simpson index

We begin by noting that across partitions of N, the number of possible values of the Gini-Simpson index, $1 - \sum_{i=1}^{I}(n_i/N)^2 = 1 - (1/N^2)\sum_{i=1}^{I} n_i^2$, is equal to the number of possible values of $\sum_{i=1}^{I} n_i^2$, the sum of squares of the entries.

4.1 Upper bound

An upper bound on $\gamma(N)$ can be obtained by noting that for each partition of N, we have $N = \sum_{i=1}^{I} n_i \leq \sum_{i=1}^{I} n_i^2 \leq (\sum_{i=1}^{I} n_i)^2 = N^2$. The quantity $\gamma(N)$ is therefore bounded above by the number of integers in the interval $[N, N^2]$.

	Number of possible values		
N	Partitions, $p(N)$	Gini-Simpson index, $\gamma(N)$	Shannon index, $\sigma(N)$
	(OEIS A000041)	(OEIS A069999)	(OEIS A383683)
1	1	1	1
2	2	2	2
3	3	3	3
4	5	5	5
5	7	7	7
6	11	9	11
7	15	13	15
8	22	18	21
9	30	21	29
10	42	27	39
11	56	34	52
12	77	39	68
13	101	46	89
14	135	54	116
15	176	61	149
16	231	72	189
17	297	83	240
18	385	92	298
19	490	106	373
20	627	118	455
21	792	130	562
22	1002	145	690
23	1255	162	837
24	1575	176	1014
25	1958	193	1227
26	2436	209	1480
27	3010	226	1772
28	3718	246	2110
29	4565	265	2516
30	5604	284	2980
31	6842	308	3522
32	8349	330	4147
33	10143	352	4879
34	12310	375	5729
35	14883	402	6688
36	17977	426	7797
37	21637	453	9082
38	26015	480	10546
39	31185	508	12225
40	37338	538	14114
41	44583	570	16303
42	53174	598	18771
43	63261	631	21585
44	75175	661	24760
45	89134	694	28355
46	105558	730	32456
47	124754	765	37042
48	147273	800	42230
49	173525	835	48091
50	204226	872	54612

Table 1: The numbers of possible values of the Gini-Simpson index and Shannon index across all partitions of N for small N.

Further, $\sum_{i=1}^{I} n_i^2$ has the same parity as N: if N is odd, then a partition of N has an odd number of odd parts, and the sum of the squares of the parts is odd, whereas if N is even, a partition of N has an even number of odd parts, and the sum of the squares of the parts is even. The number of integers in $[N, N^2]$ with the same parity as N is $(N^2 - N + 2)/2$. We have the following proposition.

Proposition 4 ([8, p. 1770])**.** *The number of possible values of the Gini-Simpson index across all samples of size N, $\gamma(N)$, is bounded above by $\frac{1}{2}N^2 - \frac{1}{2}N + 1$.*

4.2 Lower bound

We discuss three progressively tighter results concerning the lower bound on $\gamma(N)$, drawing on different ideas. The first result considers that the partitions of N can be placed in a partial order [11, p. 199]. The largest element is $(N, 0, \ldots, 0)$, the smallest is $(1, 1, \ldots, 1)$, and partition P_1 is greater than (distinct) partition P_2 if P_1 majorizes P_2: if, when P_1 and P_2 are written with entries in non-increasing order, each partial sum of entries in P_1 is greater than or equal to the corresponding partial sum for P_2.

By Karamata's inequality [11, p. 156] with the strictly convex function $f(x) = x^2$, distinct partitions P_1, P_2 for which P_1 majorizes P_2 have distinct sums of squares for their parts, with P_1 possessing the larger value. The number of distinct values for the sum of squares of entries in partitions of N is therefore bounded below by the length of the longest chain of partitions of N—the largest subset of the partitions that can be totally ordered. Each entry in the chain has a distinct sum of squares.

Write N in its unique decomposition into two parts satisfying $N = \binom{m+1}{2} + r$ with $0 \leq r \leq m$ for integers m and r. The length of the longest chain of partitions then equals $\frac{1}{3}m(m^2 + 3r - 1) + 1$ [7, p. 9]. This quantity bounds from below the number of distinct values of the Gini-Simpson index as a function of N; starting at $N = 1$, it has initial entries 1, 2, 3, 5, 7, 9, 12, 15, 18, 21 (OEIS A006463 plus 1). The sequence A006463(N) has asymptotic growth $\frac{2\sqrt{2}}{3}N^{3/2}$ [7, p. 9].

Proposition 5 ([15, Section 6.2.1])**.** *The number of possible values of the Gini-Simpson index across all samples of size N, $\gamma(N)$, is bounded below by $\frac{1}{3}m(m^2 + 3r - 1) + 1$, where (m, r) gives the unique decomposition $N = \binom{m+1}{2} + r$ with $0 \leq r \leq m$. The lower bound grows with $\frac{2\sqrt{2}}{3}N^{3/2}$.*

The number of possible values $\gamma(N)$ of the Gini-Simpson index across partitions grows polynomially; the growth has lower bound $\frac{2\sqrt{2}}{3}N^{3/2}$ and upper bound $\frac{1}{2}N^2$.

A second, tighter lower bound on $\gamma(N)$ follows from a result of Winkler [19].

Proposition 6. *The number of possible values of the Gini-Simpson index across all samples of size N, $\gamma(N)$, is bounded below by $(\lfloor \frac{1}{5}N^2 \rfloor - N)/2$.*

Lemma 2 of Winkler [19] demonstrated that for each m with $N \leq m \leq \frac{1}{5}N^2$ and $m \equiv N \pmod 2$, there exists a partition of N for which the sum of squares of the parts is m.

The number of integers m in $[N, \frac{1}{5}N^2]$ with the same parity as N is bounded below by $(\lfloor \frac{1}{5}N^2 \rfloor - N)/2$. Taken together, Propositions 4 and 6 show that the number of possible values $\gamma(N)$ of the Gini-Simpson index across partitions grows quadratically.

The third, stronger result for the lower bound can be obtained via a result from Reznick [14]. In particular, Reznick [14] investigated the largest integer $A(N)$ for which each integer $N+2j$ with j in $\{0, 1, 2, \ldots, A(N)\}$ can be represented as the sum of the squares of the parts of a partition of N. The quantity $A(N)+1$ provides a lower bound on $\gamma(N)$, as sums of squares of the parts of partitions of N traverse the values $N+2j$ for all integers $j \in \{0, 1, 2, \ldots, A(N)\}$.

The theorem on p. 201 of [14] demonstrates that $A(N) \sim \frac{1}{2}N^2 - \sqrt{2}N^{3/2}$. In other words, the number of distinct values of $\gamma(N)$ is bounded below by a quantity that, in the leading term, is asymptotic to $\frac{1}{2}N^2$. With Proposition 4, we conclude not only that the number of possible values $\gamma(N)$ of the Gini-Simpson index across partitions grows quadratically, but that the growth constant can be specified: $\gamma(N) \sim \frac{1}{2}N^2$.

Proposition 7. *The number of possible values of the Gini-Simpson index across all samples of size N, $\gamma(N)$, satisfies $\gamma(N) \sim \frac{1}{2}N^2$.*

Table 2 gives the sequence of values of $A(N)$ for small N, where it can be seen that $A(N)$ is near the asymptotic approximation $\lfloor \frac{1}{2}N^2 - \sqrt{2}N^{3/2} \rfloor$. The table also shows $N + 2A(N)$, which provides the largest integer $N + 2j$ such that all integers $\{N, N+2, \ldots, N+2j\}$ can be written as the sum of squares of the parts of a partition of N.

5 Number of possible values of the Shannon index

Whereas the growth of the number of possible values $\gamma(N)$ of the Gini-Simpson index across partitions of N grows quadratically, we will see that the number of possible values $\sigma(N)$ of the Shannon index across partitions of N grows faster than polynomially.

The Shannon index $-\sum_{i=1}^{I}(n_i/N)\log(n_i/N)$ can be rewritten $\log N - \frac{1}{N}\sum_{i=1}^{N} n_i \log n_i$. The number of possible values of the Shannon index across partitions of N is equal to the number of possible values of $\sum_{i=1}^{I} n_i \log n_i$ across partitions of N.

5.1 Upper bound

The number of possible values of the Shannon index across samples of size N is trivially bounded above by the number of partitions of size N, or $p(N)$. The Hardy-Ramanujan asymptotic formula for the partition function specifies [1, eq. 5.1.2]

$$p(N) \sim \frac{1}{4\sqrt{3}N} e^{\pi\sqrt{\frac{2}{3}}\sqrt{N}}. \tag{5}$$

An upper bound for $p(N)$ applicable for all $N \geq 1$ is $p(N) \leq e^{\pi\sqrt{\frac{2}{3}}\sqrt{N}}$ [2, Theorem 14.5].

N	$\lfloor \frac{1}{2}N^2 - \sqrt{2}N^{3/2} \rfloor$	$A(N)$	$N + 2A(N)$
		(OEIS A381811)	(OEIS A383682)
1	-1	0	1
2	-3	1	4
3	-3	1	5
4	-4	3	10
5	-4	4	13
6	-3	4	14
7	-2	7	21
8	-1	13	34
9	2	13	35
10	5	18	46
11	8	25	61
12	13	25	62
13	18	32	77
14	23	32	78
15	30	40	95
16	37	49	114
17	45	52	121
18	53	62	142
19	63	73	165
20	73	85	190
21	84	102	225
22	96	112	246
23	108	127	277
24	121	133	290
25	135	160	345
26	150	166	358
27	166	166	359
28	182	184	396
29	199	203	435
30	217	208	446
31	236	228	487
32	255	249	530
33	276	271	575
34	297	294	622
35	319	322	679
36	342	343	722
37	366	373	783
38	390	376	790
39	416	376	791
40	442	403	846
41	469	431	903
42	497	490	1022
43	525	521	1085
44	555	521	1086
45	585	553	1151
46	616	592	1230
47	648	620	1287
48	681	655	1358
49	715	662	1373
50	750	662	1374

Table 2: The largest integer $A(N)$ for which each integer $N+2j$ with j in $\{0,1,2,\ldots,A(N)\}$ can be represented as the sum of the squares of a partition of N, and the associated integer $N+2A(N)$. The integer $A(N)$ has asymptotic equivalence to $\lfloor \frac{1}{2}N^2 - \sqrt{2}N^{3/2} \rfloor$.

Proposition 8. *For all $N \geq 1$, the number of possible values of the Shannon index across all samples of size N, $\sigma(N)$, is bounded above by $e^{\pi\sqrt{\frac{2}{3}}\sqrt{N}}$.*

This upper bound can be improved. For each integer $m \geq 2$, we can find two partitions of $4m$ with equal Shannon indices: $(m, m, 2, 2, \ldots, 2)$, where the number of copies of 2 is m, and $(2m, 1, 1, \ldots, 1)$, with $2m$ copies of 1. Both have $\sum_{i=1}^{I} n_i \log n_i$ equal to $2m \log m + 2m \log 2$. In particular, partitions $(2, 2, 2, 2)$ and $(4, 1, 1, 1, 1)$ of 8 have the same value, $8 \log 2$. As a result, among the partitions of $N \geq 9$, we have at least $p(N-8)$ duplicate values of the Shannon index: each partition of $N-8$ together with $(2, 2, 2, 2)$ shares its Shannon index with the same partition of $N-8$ together with $(4, 1, 1, 1, 1)$. The number of values of the Shannon index, $\sigma(N)$, is therefore bounded above by $p(N) - p(N-8)$. However, it is not the upper bound but rather the lower bound that provides the key result that the number of values of the Shannon index far exceeds that of the Gini-Simpson index as N increases.

5.2 Lower bound

Our main result is that the number of possible values of the Shannon index grows faster than polynomially.

Theorem 9. *For all polynomials $q(N)$,*

$$\lim_{N\to\infty} \frac{\sigma(N)}{|q(N)|} = \infty.$$

Proof. First, the argument that underlies Proposition 5 in Section 4.2 applies to the strictly concave $f(x) = -x \log x$: two distinct partitions P_1 and P_2 of N for which P_1 majorizes P_2 possess distinct Shannon indices, with P_2 now possessing the larger value. The number of distinct values for the Shannon index is therefore bounded below by the length of the longest chain in the partial order on partitions. Letting $c_1 = \frac{2\sqrt{2}}{3} - \epsilon_1$ for ϵ_1 small, we have $\sigma(N) > c_1 N^{3/2}$ for all sufficiently large $N > N_1$.

Enumerate all the prime numbers $p_1 < p_2 < \cdots < p_k$ in the interval $(\frac{N}{2}, \frac{3N}{4})$. Each partition of N that includes one of these primes p_i cannot include another, p_j, as $p_i + p_j > N$. A partition that includes p_i is formed from p_i and a partition of $N - p_i$. Therefore, the total number of values of the Shannon index formed by partitions that include a prime in $(\frac{N}{2}, \frac{3N}{4})$ is $\sum_{i=1}^{k} \sigma(N - p_i)$, and

$$\sigma(N) \geq \sum_{i=1}^{k} \sigma(N - p_i) \geq \sum_{i=1}^{k} \sigma(N - p_k) \geq k\, \sigma\big(\lfloor N/4 \rfloor\big). \tag{6}$$

By the prime number theorem [2, Chapter 4], as N grows large, the number of primes k in $(\frac{N}{2}, \frac{3N}{4})$ satisfies $k \sim \frac{N/4}{\log(N/4)}$. Hence, for a constant $c_2 = 1 - \epsilon_2$ with ϵ_2 small, for sufficiently large $N > N_2$,

$$\sigma(N) \geq c_2 \bigg(\frac{N/4}{\log(N/4)}\bigg)\, \sigma\big(\lfloor N/4 \rfloor\big). \tag{7}$$

Suppose for contradiction that σ grows polynomially. Then there exists a polynomially growing function $q(N) \sim c_3 N^d$ with a positive coefficient c_3 and exponent $d \geq \frac{3}{2}$, such that $\lim_{N\to\infty} \big(\sigma(N)/q(N)\big) = c_4 < \infty$, where $c_4 > 0$. Then

$$\lim_{N\to\infty} \frac{\sigma(N)}{q(N)} \geq \lim_{N\to\infty} c_2 \left(\frac{N/4}{\log(N/4)}\right) \frac{\sigma\big(\lfloor N/4 \rfloor\big)}{q(N)}. \tag{8}$$

We then have

$$\lim_{N\to\infty} \frac{\sigma\big(\lfloor N/4 \rfloor\big)}{q(N)} = \lim_{N\to\infty} \frac{\sigma\big(\lfloor N/4 \rfloor\big)}{q(\lfloor N/4 \rfloor)} \frac{q(\lfloor N/4 \rfloor)}{q(N)} = \frac{c_4}{4^d}$$

from the fact that $q(N)$ is asymptotically equivalent to $c_3 N^d$. It follows that the right-hand limit in Eq. (8) is infinite, as $\lim_{N\to\infty}(\frac{N}{4}/\log\frac{N}{4}) = \infty$, contradicting the assumption that the left-hand limit is finite. □

A stronger lower bound than Theorem 9 can be obtained by consideration of prime partitions, partitions that consist only of primes. Decomposing each n_i in the sum $\sum_{i=1}^{I} n_i \log n_i$ for a partition of N by its prime factorization, $\sum_{i=1}^{I} n_i \log n_i$ can be written $\sum_{i=1}^{k} a_i \log p_i$, where $(p_1, p_2, \ldots, p_k)$ now represent all the primes less than or equal to N, and the a_i are nonnegative integers. For distinct prime numbers p_i and p_j, the quantity $\log p_i$ cannot be an integer multiple of $\log p_j$, so that if the coefficients $(a_1, a_2, \ldots, a_k)$ differ for a pair of partitions of N, then the partitions must have different values for the Shannon index.

Consider two distinct *prime* partitions of N, P_1 and P_2. Because $P_1 \neq P_2$, in the representations of the Shannon indices for P_1 and P_2, there must exist some prime number p_i whose associated coefficient a_i differs between the two partitions. The Shannon indices for P_1 and P_2 therefore differ. The number of prime partitions of N (OEIS A000607) provides a lower bound on the number of distinct values for the Shannon index across partitions of N.

Letting $p^*(N)$ denote the number of prime partitions of N, an asymptotic expression for $p^*(N)$ is [4, 17, 18]

$$p^*(N) \sim \frac{1}{2N^{3/4}(3\log N)^{1/4}} e^{2\pi\sqrt{\frac{1}{3}}\sqrt{\frac{N}{\log N}}}. \tag{9}$$

Hence, because $\sigma(N)$, the number of values of the Shannon index, is bounded below by the number of prime partitions $p^*(N)$, the quantity $\sigma(N)$ grows not only faster than polynomially, but at least as fast as the exponentially growing expression in Eq. (9).

We can observe in Table 1 that the number of possible values $\sigma(N)$ of the Shannon index, while growing faster than the quadratically growing number of possible values $\gamma(N)$ of the Gini-Simpson index, appears to decrease as a fraction of the number of partitions $p(N)$.

6 Discussion

The Gini-Simpson and Shannon indices are related, in the sense that for a sample, the Shannon index lies in a narrow range given the Gini-Simpson index (Theorem 3, Figure

1). However, across samples, the indices differ in their properties. Whereas the number of distinct values of the Gini-Simpson index across samples of size N grows with $O(N^2)$ (Section 4), the number of distinct values of the Shannon index grows faster than polynomially, at least as fast as the (exponentially growing) number of prime partitions (Section 5.2).

The Gini-Simpson index has been widely used across fields, in part due to its mathematical simplicity, its potential for moment estimation in a statistical setting, and its natural meaning, such as its interpretation in genetics as the probability that two genetic copies in an organism have distinct types [15, p. 10]. However, our observation that the Shannon index—which is perhaps even more widely used [9, p. 32]—has more distinct values indicates that a value for the Shannon index comes closer to encoding the precise partition used in its calculation than does a value for the Gini-Simpson index. This property is useful in a setting in which the Shannon index and the sample size but not the partition are recorded in a data analysis, as it may often be possible to recover the lost partition.

We can consider extensions of the Gini-Simpson index to higher powers than the square. Let $J_k(\mathbf{p}) = \sum_{i=1}^{I} p_i^k$ for an integer $k \geq 2$, and let $H_k(\mathbf{p}) = 1 - J_k(\mathbf{p})$ [15, Chapter 5]. The number of distinct values of J_k across partitions of N has the same $O(N^{3/2})$ lower bound shown for the Gini-Simpson index in Proposition 5, following the argument of Section 4.2 with the strictly convex $f(x) = x^k$. For the upper bound, the sum of the kth powers of the entries of a partition of N lies in $[N, N^k]$. By the parity argument of Proposition 4, the number of distinct values of J_k is bounded above by $\frac{1}{2}N^k - \frac{1}{2}N + 1$, a degree-$k$ polynomial.

As the number of distinct values of the Shannon index grows faster than polynomially, it grows faster than the number of distinct values of a generalized Gini-Simpson index with a higher power. The problem of determining a precise asymptotic for the number of distinct values of the Shannon index remains open.

7 Acknowledgments

We thank Shoshana Elgart, Stephan Wagner, and a reviewer for suggestions. We acknowledge support from National Science Foundation grant DMS-2450005.

References

[1] G. E. Andrews, *The Theory of Partitions*, Cambridge University Press, 1998.

[2] T. Apostol, *Introduction to Analytic Number Theory*, Springer, 1976.

[3] A. J. Aw and N. A. Rosenberg, Bounding measures of genetic similarity and diversity using majorization, *J. Math. Biol.* **77** (2018), 711–737.

[4] J. Bartel, R. K. Bhaduri, M. Brack, and M. V. N. Murthy, Asymptotic prime partitions into integers, *Phys. Rev. E* **95** (2017), 052108.

[5] F. A. Cowell, *Measuring Inequality*, Oxford University Press, 2011.

[6] Y. Dodge, *The Concise Encyclopedia of Statistics*, Springer, 2008.

[7] C. Greene and D. J. Kleitman, Longest chains in the lattice of integer partitions ordered by majorization, *European J. Combin.* **7** (1986), 1–10.

[8] H. Innan, K. Zhang, P. Marjoram, S. Tavaré, and N. A. Rosenberg, Statistical tests of the coalescent model based on the haplotype frequency distribution and the number of segregating sites, *Genetics* **169** (2005), 1763–1777.

[9] T. Leinster, *Entropy and Diversity: The Axiomatic Approach*, Cambridge University Press, 2021.

[10] A. E. Magurran, *Measuring Biological Diversity*, Blackwell Science, 2004.

[11] A. W. Marshall, I. Olkin, and B. C. Arnold, *Inequalities: Theory of Majorization and Its Applications*, Springer, 2nd edition, 2010.

[12] M. Nei, *Molecular Evolutionary Genetics*, Columbia University Press, 1987.

[13] OEIS Foundation, Inc., *The On-Line Encyclopedia of Integer Sequences*, https://oeis.org, 2025.

[14] B. Reznick, The sum of the squares of the parts of a partition, and some related questions, *J. Number Theory* **33** (1989), 199–208.

[15] N. A. Rosenberg, *Mathematical Properties of Population-Genetic Statistics: Quadratic Forms Most Beautiful*, Princeton University Press, 2025.

[16] N. A. Rosenberg and D. M. Zulman, Measures of care fragmentation: mathematical insights from population genetics, *Health Serv. Res.* **55** (2020), 318–327.

[17] R. C. Vaughan, On the number of partitions into primes, *Ramanujan J.* **15** (2008), 109–121.

[18] R. C. Vaughan, Corrigendum to "On the number of partitions into primes", *Ramanujan J.* **46** (2018), 307–308.

[19] P. Winkler, Mean distance in a tree, *Discrete Appl. Math.* **27** (1990), 179–185.

2020 *Mathematics Subject Classification*: Primary 05A17; Secondary 11P81, 62P10, 94A17.
Keywords: diversity, heterozygosity, majorization, partition, Shannon entropy.

(Concerned with sequences A000041, A000607, A006463, A069999, A381811, A383682, and A383683.)

Received May 20 2025; revised versions received May 21 2025; December 19 2025; February 10 2026. Published in *Journal of Integer Sequences*, February 10 2026.

Return to Journal of Integer Sequences home page