PAPER • **OPEN ACCESS**

On the comparison of diversity of parts of a distribution

To cite this article: Rajeev Rajaram *et al* 2023 *J. Phys. Commun.* **7** 075006

Journal of Physics Communications

PAPER

On the comparison of diversity of parts of a distribution

Rajeev Rajaram[1,*], Nathan Ritchey[1] and Brian Castellani[2]

[1] Department of Math. Sci., Kent State University, United States of America
[2] Durham Research Methods Centre, Durham University, Durham, United Kingdom
[*] Author to whom any correspondence should be addressed.

E-mail: rrajaram@kent.edu

Keywords: ecological diversity, evolutionary diversity, mathematical diversity, Hill numbers, Shannon entropy

OPEN ACCESS

RECEIVED
14 May 2023

REVISED
16 June 2023

ACCEPTED FOR PUBLICATION
20 July 2023

PUBLISHED
1 August 2023

Abstract

The literature on diversity measures, regardless of the metric used (e.g., Gini-Simpson index, Shannon entropy) has a notable gap: not much has been done to connect these measures back to the shape of the original distribution, or to use them to compare the diversity of parts of a given distribution and their relationship to the diversity of the whole distribution. As such, the precise quantification of the relationship between the probability of each type p_i and the diversity D in non-uniform distributions, both among parts of a distribution as well as the whole, remains unresolved. This is particularly true for Hill numbers, despite their usefulness as 'effective numbers'. This gap is problematic as most real-world systems (e.g., income distributions, economic complexity indices, rankings, ecological systems) have unequal distributions, varying frequencies, and comprise multiple diversity types with unknown frequencies that can change. To address this issue, we connect case-based entropy, an approach to diversity we developed, to the shape of a probability distribution; allowing us to show that the original probability distribution g_1, the case-based entropy curve g_2 and the $c_{\{1,k\}}$ versus the $c_{\{1,k\}}$*ln $A_{\{1,k\}}$ curve g_3, which we call the *slope of diversity*, are one-to-one (or injective), i.e., a different probability distribution g_1 gives a different curve for g_2 and g_3. Hence, a different permutation of the original probability distribution g_1(that leads to a different shape) will uniquely determine the graphs g_2 and g_3. By proving the injective nature of our approach, we will have established a unique way to measure the degree of uniformity of parts as measured by D_P/c_P for a given part P of the original probability distribution, and also have shown a unique way to compute the D_P/c_P for various shapes of the original distribution and (in terms of comparison) for different curves.

1. The challenge of measuring diversity

Within the natural and mathematical sciences, mathematical diversity refers to the measurement and quantification of diversity within a given system or population using mathematical principles and tools. Over the last several decades, while a considerable literature has developed around measuring diversity, there remains a key challenge: regardless of the metric used, such as Gini-Simpson index, Shannon entropy, or Hill numbers, has a notable gap, in that not much has been done to connect these measures back to the shape of the original distribution, or to use them to compare the diversity of parts of a given distribution and their relationship to the diversity of the whole distribution (Jost 2006, Leinster and Cobbold 2012, Chao and Jost 2015, Hsieh *et al*, 2016, Pavoine *et al*, 2016, Jost 2019).

This gap is especially evident in the case of Hill numbers ${}^qD$, which provide a way to measure diversity for a distribution by providing the species richness value or the number of types for an equivalent uniform distribution that has the same entropy as the original distribution (MacArthur 1965, Hill 1973, Peet 1974, Jost 2006, Gaggiotti *et al*, 2018, Jost 2019).

1.1. Hill numbers

Hill numbers (MacArthur 1965, Hill 1973, Peet 1974), provide a comprehensive framework to capture different aspects of diversity. Hill numbers incorporate the concepts of richness, evenness, and dominance into a single

numerical index. These indices allow for the comparison and ranking of diverse systems, such as ecological communities, species populations, or even mathematical databases.

Hill numbers are characterized by a parameter q that favors types with lower or higher frequencies, depending on whether $0 < q < 1$ or $q > 1$, respectively. When $q = 1$, ${}^{1}D$ weighs each type proportional to its relative frequency, ultimately resulting in e^{H}, where H is the Shannon entropy of the distribution. A more recent book on an axiomatic approach to defining and characterizing diversity can be found in (Leinster 2021).

The interpretation of Hill numbers and mathematical diversity depends on the specific context in which they are applied. In ecology, Hill numbers are often used to characterize biodiversity in ecological communities. They provide a way to summarize the distribution of species abundances and assess the relative importance of rare versus common species.

When analyzing species data using Hill numbers ${}^{q}D$, the values obtained can be interpreted as follows:

1. Hill number with $q = 0$: This represents the species richness, which counts the number of unique species present in the community. A higher value indicates greater species richness.
2. Hill number with $q = 1$: This is known as the exponential of the Shannon entropy and reflects both species richness and evenness. It captures the distribution of abundances among species, with higher values indicating a more even distribution.
3. Hill number with $q = 2$: Also referred to as the inverse Simpson index, it emphasizes the dominance of abundant species. A lower value indicates greater dominance of a few dominant species, while a higher value suggests a more equitable distribution of abundances among species.
4. Hill number with $q \to \infty$: This represents the effective number of species, which accounts for both richness and evenness. It quantifies the diversity as if the species were equally abundant. A higher value signifies a more diverse community.

Interpreting Hill numbers in other contexts depends on the application and the specific definition of diversity being used. For example, in mathematical datasets, Hill numbers can be employed to assess the diversity of numerical values, patterns, or structures. In this case, higher Hill numbers indicate a greater variety and complexity in the dataset.

Overall, Hill numbers provide a unified framework to measure and interpret diversity by incorporating multiple dimensions of richness, evenness, and dominance. They enable researchers to compare and quantify diversity across different systems, identify patterns of variation, and evaluate the impacts of disturbances or interventions on diversity.

1.2. The challenge

Despite the usefulness of Hill numbers as 'effective numbers,' the exact relationship between the probability of each type in a distribution and the Hill number itself remains unexplored. Furthermore, the original notion of diversity that was due to Hill and Jost is actually insensitive to permutations i.e., if we rearrange the probabilities of the original distribution g_1, then the diversity of the entire distribution will remain unchanged.

These issues are particularly problematic since most real-world systems have unequal distributions, varying frequencies, and comprise multiple diversity types with unknown frequencies that can change. Such systems include income distributions, economic complexity indices, ecological systems, species diversity, and ranking systems, from genes and exposomic biological assays to measures of economic and health inequality. An excellent example is the Gini coefficient. Despite being one of the most widely used measures of economic inequality, it has several serious flaws. For our purposes, the most important is that it provides the same coefficient for different income distributions, such that several countries can have different income distributions but the same Gini index. As this example hopefully illustrates, the Gini index and other measures of diversity struggle with the precise quantification of the relationship between the probability of each type p_i and the diversity D in non-uniform distributions, both among parts of a distribution as well as the whole. As a result, while highly important, this issue remains unresolved.

1.3. Purpose of current study

To address this gap, we will explicitly connect case-based entropy, an approach to diversity that we developed, to the shape of a probability distribution. We made initial progress on this gap in (Rajaram and Castellani 2020) by proving an interesting result relating the probabilities p_i in a distribution with K types (including J types whose frequencies can be changed) and the total diversity D_K. In the current paper, we will show that the case-based entropy curve g_2 and the $c_{\{1,k\}}$ versus the $c_{\{1,k\}}{}^{*}\ln A_{\{1,k\}}$ curve g_3, which we call the *slope of diversity* are one-to-one (or injective), i.e., a different probability distribution g_1 gives a different curve for g_2 and g_3. Hence, a

Table 1. General dataset with complexity types x_i each having a probability p_i and a frequency f_i.

X	P	F
x_1	p_1	f_1
x_2	p_2	f_2
x_3	p_3	f_3
⋮	⋮	⋮
x_J	p_J	f_J
⋮	⋮	⋮
x_K	p_K	f_K

different permutation of the original probability distribution g_1 (that leads to a different shape) will uniquely determine the graphs g_2 and g_3. By proving the injective nature of our approach, we will have established a unique way to measure the degree of uniformity of parts as measured by D_P/c_P and also have shown a unique way to compute the D_P/c_P for various shapes of the original distribution.

As our case study, we will consider a general probability distribution with a random variable X as shown in table 1 (signifying different types or categories), where x_i denotes the $i-th$ type, with probability p_i and frequency f_i. We note that the random variable X under study can be quantitative as well as qualitative. For our case study, we will ask the following question: Can we establish a relationship (direct or indirect) between the probabilities p_i and the case-based entropy curve C_c as a function of the cumulative probability c? More specifically, what if any, is a relationship between the shape of the case-based entropy curve (C_c versus c) and the original probability distribution shown in table 1?

2. A formal introduction to diversity

Diversity is commonly used as a measure to assess the 'richness' or number of types in a distribution and its 'evenness,' or equal probability of occurrence among diversity types, as reported by several studies (MacArthur 1965, Hill 1973, Peet 1974, Jost 2006). This definition of diversity is based on the intuition that if all types in the distribution occur with the same probability, diversity should be equal to the number of types K. Conversely, any deviation from uniformity in probabilities will always result in a lower diversity value.

Definition 2.1. (Shannon Diversity corresponding to $q = 1$ for Hill numbers) Given an ordered set of types numbered as $i \in \mathbf{N}$ and their corresponding probabilities p_i, the diversity of the entire distribution ${}^1D_K$ is defined as the number of equiprobable types needed to yield the same value of Shannon entropy H.

Shannon entropy is defined as below:

$$H_K = -\sum_{l=1}^{K} p_l \ln(p_l). \tag{1}$$

It was shown (MacArthur 1965, Hill 1973, Peet 1974, Jost 2006, Rajaram and Castellani 2016) that definition 2.1 implies that the total diversity ${}^1D_K$ is given by:

$${}^1D_K = e^H = \prod_{l=1}^{K} \frac{1}{p_l^{p_l}};\ p_l = \frac{f_l}{\sum_{k=1}^{K} f_k}. \tag{2}$$

Furthermore, we denote the diversity of the first i types (or partial diversity) as ${}^1D_{\{1,i\}}$, where $i = 1,..,K$. The partial diversity up to the first i types is given by:

$${}^1D_{\{1,i\}} = \prod_{l=1}^{i} \frac{1}{(p_{l_{\{1,i\}}})^{p_{l_{\{1,i\}}}}};\ p_{l_{\{1,i\}}} = \frac{p_l}{\sum_{k=1}^{i} p_k} = \frac{f_l}{\sum_{k=1}^{i} f_k}. \tag{3}$$

We note that equations (2) and (3) can be rewritten in terms of the frequencies f_i as below. We will continue to use the modified equation (4) in our exposition.

$$ {}^{1}D_K = \frac{\sum_{l=1}^{K} f_l}{\prod_{j=1}^{K} f_j^{\left(\frac{f_j}{\sum_{l=1}^{K} f_l}\right)}};^{1} D_{\{1,i\}} = \frac{\sum_{l=1}^{i} f_l}{\prod_{j=1}^{i} f_j^{\left(\frac{f_j}{\sum_{l=1}^{i} f_l}\right)}}. \tag{4} $$

In this paper, we have two objectives:

1. We make a case for the ratio D_P/c_P i.e., diversity of a part to its cumulative probability as a way to measure the degree of uniformity of the part P, and also show a way to compute this ratio for arbitrary parts from the graph of the slope of diversity curve ($c_{\{1,k\}}$ versus $c_{\{1,k\}}{}^{*}\ln A_{\{1,k\}}$). This will prove to be an important way to measure the extent of uniformity of parts of a distribution.
2. We prove some results that relate the case-based entropy curve i.e. $c_{\{1,k\}}$ versus $C_{\{1,k\}}$ to the original probability distribution, again, by using the graph of slope of diversity curve $c_{\{1,k\}}$ versus $c_{\{1,k\}}{}^{*}\ln A_{\{1,k\}}$. This will close the gap of relating the Hill numbers back to the shape of the original distribution.

The paper is organized as follows: In section 3 we lay down the foundation towards using the ratio D_P/c_P as a means to compare the degree of uniformity of parts of a distribution. In section 4, we show a way to compute D_P/c_P for parts of a given distribution using a new curve that plots $c_{\{1,k\}}$ versus $c_{\{1,k\}}{}^{*}\ln A\{1,\ k\}$, which we call *slope of diversity*. In section 5, we prove some results related to comparing the ratio D_P/c_P for different parts of a distribution. In section 6 we relate the case-based entropy curve to the original probability distribution given in table 1. In section 7 we use the geometric distribution as an example to demonstrate some of our results. In section 8, conclude the paper with some remarks on the results.

3. The ratio $\frac{D_P}{c_P}$ for parts P of a distribution

We recall the following two important 'parts-to-whole' formulae that were proved in (Rajaram and Castellani 2020).

Theorem 3.1. *Given a probability distribution similar to table 1, the diversity of the entire distribution* ${}^{q}D_K$ *for some complex system or dataset, and the diversities of disjoint parts* ${}^{q}D_{P_i}$ *and their respective cumulative probabilities* c_{P_i} *are related as follows:*

$$ {}^{1}D_K = \prod_{P_i \in \mathcal{P}} \left(\frac{{}^{1}D_{P_i}}{c_{P_i}}\right)^{c_{P_i}}, \tag{5} $$

and

$$ {}^{q}D_K = \left(\sum_{P_i \in P} c_{P_i} \left(\frac{{}^{q}D_{P_i}}{c_{P_i}}\right)^{(1-q)}\right)^{\frac{1}{1-q}}. \tag{6} $$

We note that equations (5) and (6) are simply the weighted geometric and arithmetic means (of order $1-q$) respectively of the ratio $\left(\frac{{}^{q}D_{P_i}}{c_{P_i}}\right)$. We also note that ${}^{1}D_K = \lim_{q\to 1}^{q} D_K$. The following corollary can be easily proved.

Corollary 3.1. *Given a probability distribution similar to table 1, let the part* $P = \bigcup_i P_i$ *be a disjoint union of sub-parts* P_i. *Then, the diversity of the part* ${}^{q}D_P$ *and the diversities of disjoint sub-parts* ${}^{q}D_{P_i}$ *and their respective cumulative probabilities* c_{P_i} *are related as follows:*

$$ \left(\frac{{}^{1}D_P}{c_P}\right)^{c_P} = \prod_{P_i \in P} \left(\frac{{}^{1}D_{P_i}}{c_{P_i}}\right)^{c_{P_i}}, \tag{7} $$

and

$$ c_P \left(\frac{{}^{q}D_P}{c_P}\right)^{1-q} = \sum_{P_i \in P} c_{P_i} \left(\frac{{}^{q}D_{P_i}}{c_{P_i}}\right)^{(1-q)}. \tag{8} $$

Proof. The proof is obtained by re-normalizing the probability of the type j in part P_i as $\tilde{p}_j = \frac{p_j}{c_{P_i}}$ and using the formulas 5 and 6 in a recursive fashion.

Remark 3.1. If we consider each part P_i in the derivation of the above theorem to be exactly one type i.e., $P_i = \{i\}\ \forall\, i = 1,\ldots,K$, then $D_{P_i} = 1\ \forall\, i = 1,\ldots,K$ and equation (5) reduces to equation (2).

Remark 3.2. We can restrict ourselves to a portion of the distribution starting from $l = 1$ to say $l = k$. Then theorem 3.1 is true for the restriction for any sub-partition $\mathcal{P}_k$ of such a restriction. In this case, the probabilities will have to be re-normalized as $p_l = p_{l\{1,k\}}$ and $c_l = c_{l\{1,k\}}$.

We can re-imagine the given probability distribution (or, as we will see, any of its parts as well) as an abstract uniform distribution having the same entropy (or conditional entropy of parts, if we are looking at parts). Hence, there is a close relationship between the diversity of a distribution (or its part) to uniformity. In both theorem 3.1 and corollary 3.1 we notice the occurrence of the ratio $\frac{D_{P_i}}{c_{P_i}}$ repeatedly, giving us a sense that this ratio should play an important part in comparing the degree of uniformity or closeness to uniform distribution among parts of a given distribution.

From this point in the paper, we choose to focus our results for $q = 1$ for the Hill number ${}^qD$ to show our results since the weight given to each type is proportional to the abundance of the type if $q = 1$. Accordingly, we will omit the left superscript of 1 while writing the diversity D.

3.1. An abstract visualization of the part *P* of a distribution

It is well known (MacArthur 1965, Hill 1973, Peet 1974, Jost 2006, Gaggiotti *et al*, 2018, Jost 2019) that a non-uniform distribution with a diversity of D_K can be abstractly redrawn as a uniform distribution with D_K number of types each having a probability of $\frac{1}{D_K}$. The abstract uniform distribution has the same Shannon entropy as the original distribution, and hence has the same degree of probabilistic uncertainty as the original distribution. The D_K number of types for the abstract equivalent distribution may no longer be an integer. We call the D_K types as *Shannon Equivalent Equiprobable (SEE)* types.

In a similar way, we consider a part of a distribution P with diversity D_P and cumulative probability c_P, where $P = \{k_1, k_2\}$ is the part between indices k_1 and k_2 for example. For this part P, we can associate an equivalent abstract uniform distribution which has D_P number of SEE types, each of which has a probability of $\frac{c_P}{D_P}$. This abstract equivalent uniform distribution has the same entropy as conditional entropy of the original distribution given the part P. In other words, it has the same degree of uncertainty as the part P.

Hence, we have the following: Given a distribution consisting of disjoint parts P_i with diversity D_{P_i} and cumulative probability c_{P_i} so that $\bigcup_i P_i$ is the entire distribution, each part P_i can be redrawn as an abstract uniform distribution with D_{P_i} number of SEE types each having a probability of $\frac{c_{P_i}}{D_{P_i}}$. We also have that the abstract equivalent has the same entropy as the conditional entropy of the original distribution given the part P_i and its total cumulative probability will also be equal to c_{P_i}. We will refer to this abstract equivalent uniform distribution henceforth as the SEE equivalent of the part P_i. More generally, as shown in figure 1, each of the disjoint parts P_i of a given distribution can be equivalently replaced with an abstract uniform distribution each having a diversity of D_{P_i} and a cumulative probability of c_{P_i}. As we will see, this is a very important equivalence that allows us to compare the uniformity of the abstract equivalent SEE types of the original parts instead of the original parts themselves. Comparing the latter is much easier because the abstract equivalent SEE types are uniformly distributed even though the original parts themselves may not be uniform.

3.2. The case for using the ratio D_P/c_P to compare degrees of uniformity

Given the conclusion from the last section, it is clear that comparing degrees of uniformity of parts of a distribution boils down to comparing degrees of uniformity of the abstract SEE (*Shannon Equivalent Equiprobable*) equivalents of its parts.

We look at an example of a distribution where the SEE equivalents of three parts I, II, and III are shown as in figure 2. We assume that these three parts are the SEE equivalent types of three parts of a given distribution. The probability values are fictitious and are used to make an important point i.e., the ratio D_P/c_P for each abstract equivalent SEE part (and hence the same ratio for the original part itself) is a measure of how much more or less uniformly distributed a given part is compared to other parts, thereby showing that the ratio D_P/c_P is a relative measurement of degree of uniformity of parts of the original distribution.

It is easy to calculate

$$D_I = 10, \quad D_{II} = 20, \quad \text{and} \quad D_{III} = 30.$$

So,

$$\frac{D_I}{c_I} = \frac{10}{30/100} = 100/3, \quad \frac{D_{II}}{c_{II}} = \frac{20}{40/100} = 50, \quad \text{and} \quad \frac{D_{III}}{c_{III}} = \frac{30}{30/100} = 100,$$

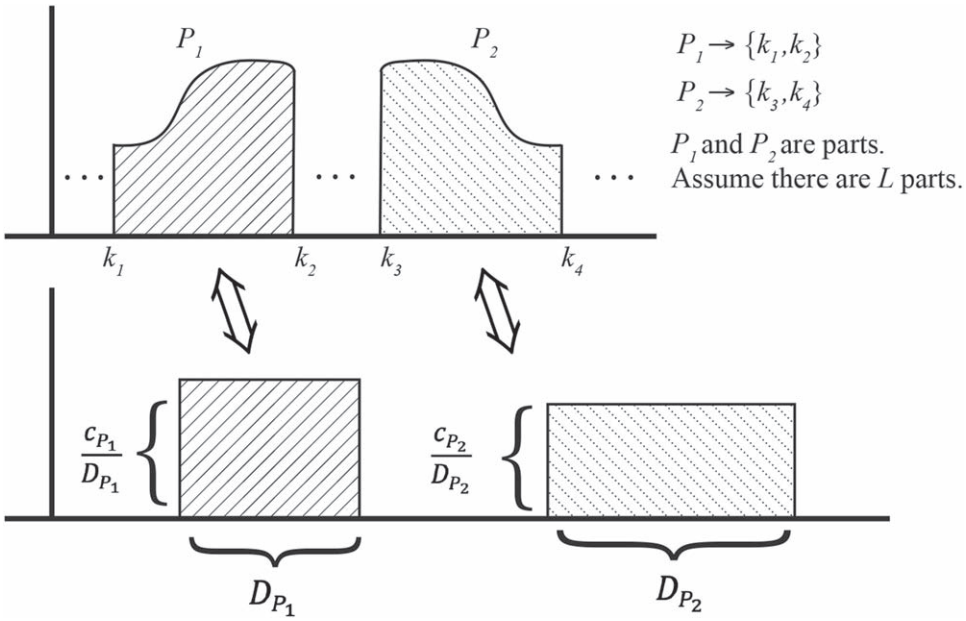

Figure 1. Parts of a general distribution along with their abstract SEE equivalents.

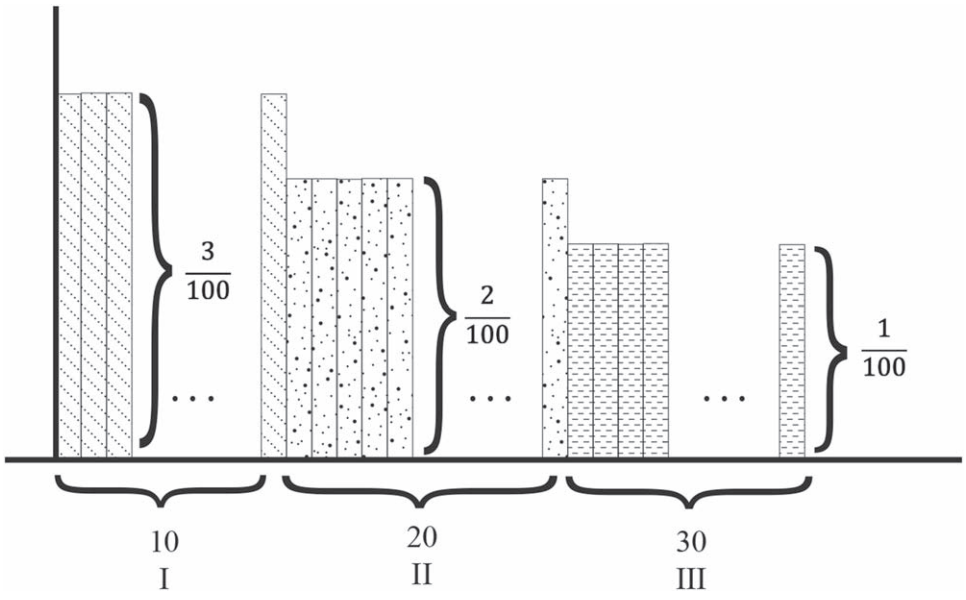

Figure 2. Example showing comparison of equivalent uniform parts.

and the total diversity from theorem 3.1 is

$$\begin{aligned} D &= \left(\frac{D_I}{c_I}\right)^{C_I}\left(\frac{D_{II}}{c_{II}}\right)^{C_{II}}\left(\frac{D_{III}}{c_{III}}\right)^{C_{III}} \\ &= \left(\frac{100}{3}\right)^{30/100}(50)^{40/100}(100)^{30/100} \\ &\approx 54.51. \end{aligned}$$

So, in part I, 33.33 SEE types are assigned per unit cumulative frequency, in part II, 50 SEE types are assigned and in part III, 100 SEE types are assigned per unit cumulative frequency. Thus, part III has the most number of SEE types per unit cumulative frequency followed by part II, followed by part I.

The main point is this: SEE types are uniformly distributed. If we are talking about the entire distribution (which has a cumulative probability of 1), then the diversity of the entire distribution (in this case $D = 54.51$) is an indication of the extent of uniformity of the entire distribution. Hence, the diversity D of entire distributions can be compared to indicate

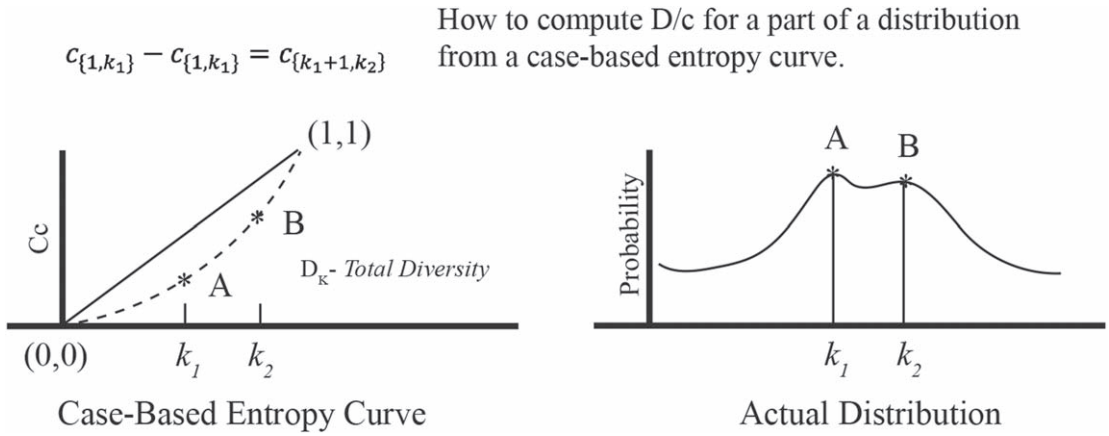

Figure 3. Case-based entropy of a part.

their degrees of uniformity. However, if we are talking about parts of a distribution (where the cumulative probability of the given part is no longer equal to 1 but $c < 1$), then the number of SEE types per unit cumulative frequency (which is the same as D_P/c_P) is a better indication of the degree of uniformity. We note that $D_P/c_P = D$ if $c_P = 1$ and hence D_P/c_P as a measure of degree of uniformity, is a generalization of the total diversity D but for parts of a distribution.

If we think of each type having the same width(amount) of money, then part III has the most amount of Shannon Equivalent Equiprobable (SEE) money per person i.e., 100 multiplied by the width of each bin. In fact, part III has twice as much SEE money per person compared to part II. Since $\frac{D_{III}}{c_{III}}$ is the largest, part III should be treated as the most uniformly distributed, followed by part II followed by part I. In fact, we also note that $D_{III} > D$ and hence part III is actually more uniformly distributed compared to the entire distribution itself. Similarly $D_I < D$ and $D_{II} < D$ mean that parts I and II are less uniformly distributed compared to the entire distribution. Finally, the diversity D of the entire distribution is a measure of degree of uniformity of the entire distribution and theorem 3.1 and corollary 3.1 say that the degree of uniformity of the entire distribution is a weighted geometric mean of the degrees of uniformity of the parts of the distribution. In other words, the total diversity D is a weighted geometric mean of $\frac{D_I}{c_I}$, $\frac{D_{II}}{c_{II}}$ and $\frac{D_{III}}{c_{III}}$.

We used the example above to lay down the intuition for why D_P/c_P is a good measure of the degree of uniformity. Now we consider a general case where a part of a distribution from say index k_1 to k_2 has diversity given by $D_{\{k_1,k_2\}}$ and a cumulative frequency of $c_{\{k_1,k_2\}}$ is given, as shown in figure 1. The ratio $\frac{D_{\{k_1,k_2\}}}{c_{\{k_1,k_2\}}}$ is the amount of SEE types or bins that are assigned per unit cumulative frequency and can be used to measure degree of uniformity in the distribution between parts. The higher the D_P/c_P ratio for a part, the more diversity per unit cumulative frequency compared to another part that has a lower value. Furthermore, this ratio is actually a true proportion as far as the part $\{k_1, k_2\}$ is concerned, since the redrawn SEE equivalent is actually uniform, and hence any portion P_j of this part will contain exactly $\frac{D_{\{k_1,k_2\}}}{c_{\{k_1,k_2\}}} \times c_{P_j}$ number of SEE types. The same intuition that we built using the example above holds true for the general case as well i.e., the ratio $\frac{D_{\{k_1,k_2\}}}{c_{\{k_1,k_2\}}}$ will indicate the number of SEE types per unit cumulative frequency and hence is a measure of the degree of uniformity of the part of the distribution from index k_1 to k_2. We focus on computing the ratio D_P/c_P for a given part P next. Figure 3 below shows how a part of the distribution of the type $\{k_1, k_2\}$ is mapped between the case-based entropy curve and the original distribution.

4. Computing the ratio D_P/c_P for a part of a distribution

A Lorentz curve by the name of case-based entropy was introduced to compare distributions in (Rajaram and Castellani 2016). The case-based entropy of a part $P = \{1, k\}$ is defined as $C_{\{1,k\}} = \frac{D_{\{1,k\}}}{D_K}$, where $D_{\{1,k\}}$ is the diversity of the part P and D_K is total diversity of K types. It is clear from the last section, that the ratio D_P/c_P for a part P is a way to measure the degree of uniformity of the distribution in the part P. In this section, we show how we can use the case-based entropy curve to compute the ratio D_P/c_P for a given part P of a distribution.

We consider the parts denoted by indices $P_1 = \{1, k_1\}$ and $P_2 = \{1, k_2\}$. We know the following:

$$C_{\{1,k_1\}} = \frac{D_{\{1,k_1\}}}{c_{\{1,k_1\}} \cdot D_K} \quad \text{and} \quad C_{\{1,k_2\}} = \frac{D_{\{1,k_2\}}}{c_{\{1,k_2\}} \cdot D_K}.$$

From equation (7) and dividing by the total diversity D_K, we have the following:

$$\left(\frac{D_{\{1,k_2\}}}{c_{\{1,k_2\}} \cdot D_K}\right)^{c_{\{1,k_2\}}} = \left(\frac{D_{\{1,k_1\}}}{c_{\{1,k_1\}} \cdot D_K}\right)^{c_{\{1,k_1\}}} \left(\frac{D_{\{k_1+1,k_2\}}}{c_{\{k_1+1,k_2\}} \cdot D_K}\right)^{c_{\{k_1,k_2\}}}.$$

We define the slopes of the secants on the case-based entropy curve joining the points $(0, 0)$ and $(c_{\{1,k\}}, C_{\{1,k\}})$ by $A_{\{1,k\}}$. In other words $A_{\{1,k\}} = \frac{D_{\{1,k\}}}{D_K c_{\{1,k\}}}$. Using this, the above equation can be rewritten as:

$$A_{\{1,k_2\}}^{c_{\{1,k_2\}}} = A_{\{1,k_1\}}^{c_{\{1,k_1\}}} A_{\{k_1+1,k_2\}}^{c_{\{k_1+1,k_2\}}}.$$

So,

$$A_{\{k_1+1,k_2\}}^{c_{\{k_1+1,k_2\}}} = \frac{A_{\{1,k_2\}}^{c_{\{1,k_2\}}}}{A_{\{1,k_1\}}^{c_{\{1,k_1\}}}},$$

$$\left(\frac{D_{\{k_1+1,k_2\}}}{c_{\{k_1+1,k_2\}} \cdot D_k}\right)^{c_{\{k_1,k_2\}}} = D_K^{c_{\{k_1+1,k_2\}}} \left(\frac{A_{\{1,k_2\}}^{c_{\{1,k_2\}}}}{A_{\{1,k_1\}}^{c_{\{1,k_1\}}}}\right),$$

and solving for $\frac{D_{\{k_1+1,k_2\}}}{c_{\{k_1+1,k_2\}}}$ we have

$$\frac{D_{\{k_1+1,k_2\}}}{c_{\{k_1+1,k_2\}}} = \frac{D_K A_{\{1,k_2\}}^{\left(\frac{c_{\{1,k_2\}}}{c_{\{k_1+1,k_2\}}}\right)}}{A_{\{1,k_1\}}^{\left(\frac{c_{\{1,k_1\}}}{c_{\{k_1+1,k_2\}}}\right)}}.$$

Note that D_K is the total diversity and it is also known separately. In fact, everything on the right hand side of the above equation is known.

Solving for $A_{\{1,k_2\}}$ in the last equation above, we have:

$$A_{\{1,k_2\}} = A_{\{1,k_1\}}^{\left(\frac{c_{\{1,k_1\}}}{c_{\{1,k_2\}}}\right)} \cdot A_{\{k_1+1,k_2\}}^{\left(\frac{c_{\{k_1+1,k_2\}}}{c_{\{1,k_2\}}}\right)},$$

where

$$\frac{c_{\{1,k_1\}}}{c_{\{1,k_2\}}} + \frac{c_{\{k_1+1,k_2\}}}{c_{\{1,k_2\}}} = 1.$$

We now take the natural logarithm of both sides of the above equation to obtain a logarithmic interpolation forumla. That is,

$$\ln(A_{\{1,k_2\}}) = \frac{c_{\{1,k_1\}}}{c_{\{1,k_2\}}} \ln(A_{\{1,k_1\}}) + \frac{c_{\{k_1+1,k_2\}}}{c_{\{1,k_2\}}} \ln(A_{\{k_1+1,k_2\}}), \text{ or}$$

$$c_{\{1,k_2\}} \ln(A_{\{1,k_2\}}) = c_{\{1,k_1\}} \ln(A_{\{1,k_1\}}) + c_{\{k_1+1,k_2\}} \ln(A_{\{k_1+1,k_2\}}), \text{ or}$$

$$\ln(A_{\{k_1+1,k_2\}}) = \frac{c_{\{1,k_2\}} \ln(A_{\{1,k_2\}}) - c_{\{1,k_1\}} \ln(A_{\{1,k_1\}})}{(c_{\{1,k_2\}} - c_{\{1,k_1\}})}.$$

If we plot a graph of $c_{\{1,k\}}$ versus $c_{\{1,k\}} \cdot \ln(A_{\{1,k\}})$, then the above formula is the slope of the secant line of the curve joining the points A and B in figure 4. We note that this curve starts at $(0, 0)$ and ends at $(1, 0)$. We name this curve as the *slope of diversity* curve.

In figure 4, points A and B have the following coordinates:

$$A \rightarrow (c_{\{1,k_1\}},\ c_{\{1,k_1\}} \ln(A_{\{1,k_1\}})),$$
$$B \rightarrow (c_{\{1,k_2\}},\ c_{\{1,k_2\}} \ln(A_{\{1,k_2\}})).$$

Then $\ln(A_{\{k_1+1,k_2\}})$ is the slope of the line joining A and B.

Let $S_{\{k_1,k_2\}}$ be the slope of the line joining $(c_{\{1,k_1\}},\ c_{\{1,k_1\}} \ln(A_{\{1,k_1\}}))$ and $(c_{\{1,k_2\}},\ c_{\{1,k_2\}} \ln(A_{\{1,k_2\}}))$. Also, taking exponentials, we have:

$$\frac{D_{\{k_1+1,k_2\}}}{c_{\{k_1+1,k_2\}}} = D_K \cdot e^{S_{\{k_1,k_2\}}}. \tag{9}$$

Then we have the following equivalence:

$$\frac{D_{\{k_1,k_2\}}}{c_{\{k_1,k_2\}}} \underset{>}{\overset{<}{=}} \frac{D_{\{k_3,k_4\}}}{c_{\{k_3,k_4\}}} \Leftrightarrow D_K e^{S_{\{k_1-1,k_2\}}} \underset{>}{\overset{<}{=}} D_K e^{S_{\{k_3-1,k_4\}}} \Leftrightarrow S_{\{k_1-1,k_2\}} \underset{>}{\overset{<}{=}} S_{\{k_3-1,k_4\}}.$$

This means, as shown in figure 5, that the ordering of the slopes of secants of parts of the slope of diversity curve preserves the same ordering of the ratios of D_P/c_P for the corresponding parts in the original distribution. This means that the $c_{\{1,k\}}$ versus $c_{\{1,k\}}$*$\ln(A_{\{1,k\}})$ (slope of diversity) curve is a way to measure the relative degree of uniformity of parts in the original distribution.

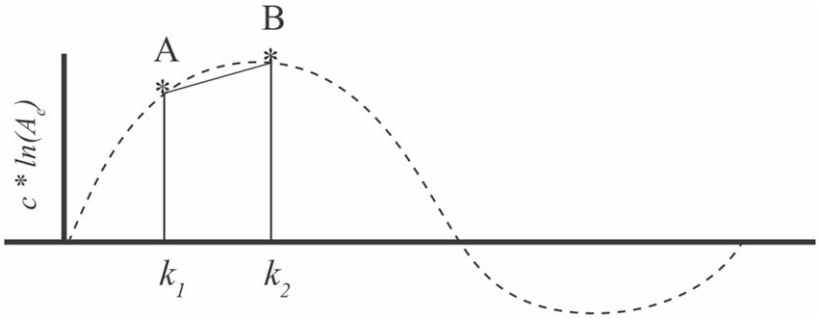

Figure 4. Points to compute secant of c versus c*ln A_c *(Slope of Diversity)* curve.

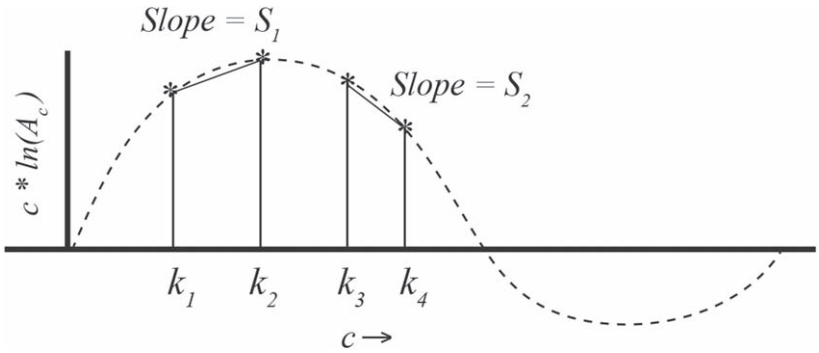

Figure 5. Comparing slopes on the slope of diversity curve.

Remark 4.1. We note that in the graph of slope of diversity, we use the convention that the index $k = 0$ corresponds to the point $(0,\ 0)$ since none of the types in the distribution are included yet. And the point corresponding to $k = 1$ is $\left(p_1,\ \frac{-\ln(p_1 D)}{p_1}\right)$. Hence, the slope of the secant joining $k = 0$ and $k = 1$ is $S_{\{0,1\}} = -\ln(p_1 D)$.

We have that, $S_{\{k_1-1,k_2\}} > S_{\{k_3-1,k_4\}}$ means that

$$\frac{D_{\{k_1,k_2\}}}{c_{\{k_1,k_2\}}} > \frac{D_{\{k_3,k_4\}}}{c_{\{k_3,k_4\}}}.$$

For example, the part $\{k_1, k_2\}$ has more SEE types per unit cumulative frequency than the part $\{k_3, k_4\}$ and hence, is more uniformly distributed compared to the part $\{k_3, k_4\}$. Alternatively, we could form the ratio $\frac{S_{\{k_1-1,k_2\}}}{S_{\{k_3-1,k_4\}}}$, which tells us how much more or less uniformly distributed the part $\{k_1, k_2\}$ is relative to the part $\{k_3, k_4\}$. Hence, with the *slope of diversity* curve, we have definitively created a quantitative way to compare the degree of uniformity of parts of a given distribution using the slopes of its secants.

Having established the importance of the ratio D_P/c_P as a way to measure the degree of uniformity of a part of a distribution, and created a way to compute D_P/c_P, we now explore ways to compare the ratio D_P/c_P for different parts of a distribution, and its relationship to the original distribution.

5. Some results related to comparing D_P/c_P for parts

In this section, we summarize our findings about comparison of the D_P/c_P ratio of parts in the form of theorems.

Theorem 5.1. *Let* P_1 *and* P_2 *be two parts of a probability distribution like in table 1. Then we have the following equivalence:*

$$A_{P_1} \underset{>}{\overset{<}{=}} A_{P_2} \Leftrightarrow \frac{D_{P_1}}{c_{P_1}} \underset{>}{\overset{<}{=}} \frac{D_{P_2}}{c_{P_2}}.$$

Proof. Follows by definition of A_{P_1} and A_{P_2}.

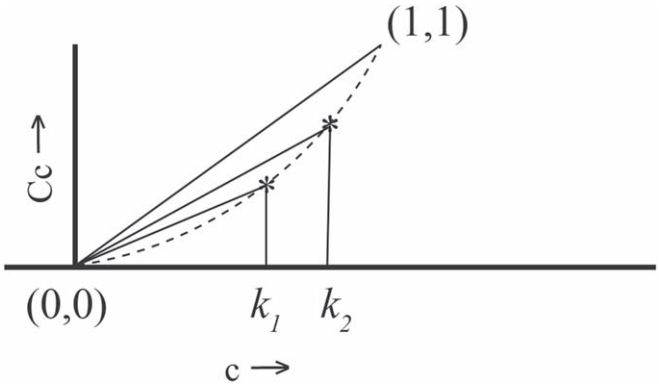

Figure 6. Comparing A's to relate the D_P/c_P ratios.

Theorem 5.1 is illustrated in figure 6. The key point here is to note that for $P_1 = \{1, k_1\}$ and $P_2 = \{1, k_2\}$, A_{P_1} and A_{P_2} are slopes of secants on the case-based entropy curve between the points $(0, 0)$ and (c_{P_1}, C_{P_1}) and the points the points $(0, 0)$ and (c_{P_2}, C_{P_2}) respectively. So, the degree of uniformity of parts that look like $P_1 = \{1, k_1\}$ and $P_2 = \{1, k_2\}$ can be directly compared from slopes of secants from the original case-based entropy curve.

However, if the parts of the type $P_1 = \{k_1, k_2\}$ and $P_2 = \{k_3, k_4\}$ where $k_1 \neq 1$ and $k_2 \neq 1$, then the case-based entropy curve cannot be directly used to compare their degrees of uniformity. For these types of general partitions that don't start at the index 1, we need to plot slope of diversity curve like in figure 5. We summarize this development in the previous section in the form of the following theorem:

Theorem 5.2. *Let $P_1 = \{k_1, k_2\}$ and $P_2 = \{k_3, k_4\}$ be general disjoint parts of a probability distribution like in table 1. Then we have the following equivalence:*

$$S_{\{k_1-1,k_2\}} \begin{matrix} \leq \\ = \\ > \end{matrix} S_{\{k_3-1,k_4\}} \Leftrightarrow \frac{D_{\{k_1,k_2\}}}{c_{\{k_1,k_2\}}} \begin{matrix} \leq \\ = \\ > \end{matrix} \frac{D_{\{k_3,k_4\}}}{c_{\{k_3,k_4\}}}.$$

Proof. Already explained in the previous section. We refer to figure 5 for an illustration as well.

Remark 5.1. We note that theorem 5.1 is subsumed by theorem 5.2. This is because we could choose partitions of the form $P_1 = \{1, k_1\}$ and $P_2 = \{1, k_2\}$ and obtain the same ordering of the D_P/c_P ratios for the parts P_1 and P_2 in theorem 5.2. For such a partition $\{1, k\}$, the slope $S_{\{0,k\}}$ is the slope of the line joining $(0, 0)$ and $(c_{\{1,k\}}, c_{\{1,k\}}{}^*\ln A_{\{1,k\}})$. Hence, we have the following:

$$S_{\{0,k_1\}} \begin{matrix} \leq \\ = \\ > \end{matrix} S_{\{0,k_2\}} \Leftrightarrow \frac{c_{\{1,k_1\}}{}^*\ln(A_{\{1,k_1\}})}{c_{\{1,k_1\}}} \begin{matrix} \leq \\ = \\ > \end{matrix} \frac{c_{\{1,k_2\}}{}^*\ln(A_{\{1,k_2\}})}{c_{\{1,k_2\}}} \Leftrightarrow A_{\{1,k_1\}} \begin{matrix} \leq \\ = \\ > \end{matrix} A_{\{1,k_2\}},$$

which shows that theorem 5.1 is subsumed by theorem 5.2.

We now state and prove an explicit relationship between the probabilities in the original distribution and the slope of diversity curve. We note that this is the first time that the diversity of a distribution is directly related to the individual probabilities in the distribution, thereby establishing a connection between the diversity and the shape of the original distribution.

Theorem 5.3. *Given a probability distribution like in table 1, we have the following:*

$$\frac{1}{p_k} = D_K e^{S_{\{k-1,k\}}}. \tag{10}$$

Proof. We already showed the following in equation (9):

$$\frac{D_{\{k_1,k_2\}}}{c_{\{k_1,k_2\}}} = D_K \cdot e^{S_{\{k_1-1,k_2\}}}.$$

Now, let's choose $k_1 = k_2 = k$. Then, $D_{\{k_1,k_2\}} = D_{\{k,k\}} = 1$ and $c_{\{k_1,k_2\}} = c_{\{k,k\}} = p_k$. This implies that equation (9) becomes

$$\frac{1}{p_k} = D_K e^{S_{\{k-1,k\}}},$$

or

$$p_k = \frac{1}{D_K} e^{-S_{\{k-1,k\}}}.$$

This proves the Theorem.

Remark 5.2. We can alternatively explicitly show the relationship in the above theorem as follows:

$$\begin{aligned} S_{\{k-1,k\}} &= \frac{c_{\{1,k\}} \ln(A_{\{1,k\}}) - c_{\{1,k-1\}} \ln(A_{\{1,k-1\}})}{c_{\{1,k\}} - c_{\{1,k-1\}}} \\ &= \frac{c_{\{1,k\}} \ln\left(\frac{D_{\{1,k\}}}{D_K c_{\{1,k_1\}}}\right) - c_{\{1,k-1\}} \ln\left(\frac{D_{\{1,k-1\}}}{D_K c_{\{1,k-1\}}}\right)}{p_k} \\ &= \ln\left[\frac{\left(\frac{D_{\{1,k\}}}{D_K c_{\{1,k\}}}\right)^{c_{\{1,k\}}/p_k}}{\left(\frac{D_{\{1,k-1\}}}{D_K c_{\{1,k-1\}}}\right)^{c_{\{1,k-1\}}/p_k}}\right] \\ &= \ln\left[\frac{1}{D_K} \cdot \left(\frac{1}{p_k}\right)^{p_k/p_k}\right] \\ &= \ln\left[\frac{1}{D_K \cdot p_k}\right], \end{aligned}$$

and exponentiating both sides gives us the result of theorem 5.3.

This means that we can completely reconstruct the original distribution simply by looking at slopes of the form $S_{\{k-1,k\}}$ for all $k = 1,\ldots,K$, and computing

$$p_k = \frac{e^{-S_{\{k-1,k\}}}}{D_K}.$$

This is a key reconstruction result, which is illustrated by the following graph:

Figure 7 shows that there is a one-to-one (injective) correspondence between the original distribution and the case-based entropy curve via the slope of diversity curve. This is a new result.

It also means that two different distributions will give two different case based entropy curves that are unique to the shape of each distribution. It also means that two different distributions will give two different slope of diversity curves as well.

Theorem 5.4. *Given a probability distribution like in table 1, let $\mathcal{G}_1$ be the set of graphs of the original probability distribution, $\mathcal{G}_2$ be the set of graphs of the corresponding case-based entropy curves, and $\mathcal{G}_3$ be the set of graphs of the corresponding slope of diversity curves, with g_1, g_2 and g_3 denoting elements (graphs) in $\mathcal{G}_1$, $\mathcal{G}_2$ and $\mathcal{G}_3$ respectively. In addition, let $T_{j\to k}$ be the map from the graph $\mathcal{G}_j$ to the graph $\mathcal{G}_k$ where $j, k = 1, 2, 3$. Then we have the following:*

$$T_{j\to k}\colon \mathcal{G}_j \xrightarrow{\sim} \mathcal{G}_k \tag{11}$$

is injective (or one-to-one).

Remark 5.3. We note that since the number of points on the original distribution curve, the case-based entropy curve and the slope of diversity curve are equal, the map $T_{j\to k}\colon \mathcal{G}_j \to^{\sim} \mathcal{G}_k$ is defined to be the natural map between the points in the same order as they appear from left to right.

Proof.

(1) $T_{1\to 2}$: Let $g_1^a, g_1^b \in \mathcal{G}_1$. We will show below that $T_{1\to 2}(g_1^a) = T_{1\to 2}(g_1^b)$ implies that $g_1^a = g_1^b$

$$\begin{aligned} &T_{1\to 2}(g_1^a) = T_{1\to 2}(g_1^b) \Rightarrow (c_{\{1,k\}}^a, C_{\{1,k\}}^a) = (c_{\{1,k\}}^b, C_{\{1,k\}}^b)\ \forall k \\ &\Rightarrow c_{\{1,k\}}^a = c_{\{1,k\}}^b\ \forall k \Rightarrow p_k^a = p_k^b\ \forall k. \\ &\text{Hence, } T_{1\to 2}(g_1^a) = T_{1\to 2}(g_1^b) \Rightarrow g_1^a = g_1^b. \end{aligned}$$

This shows that the map $T_{1\to 2}$ is injective.

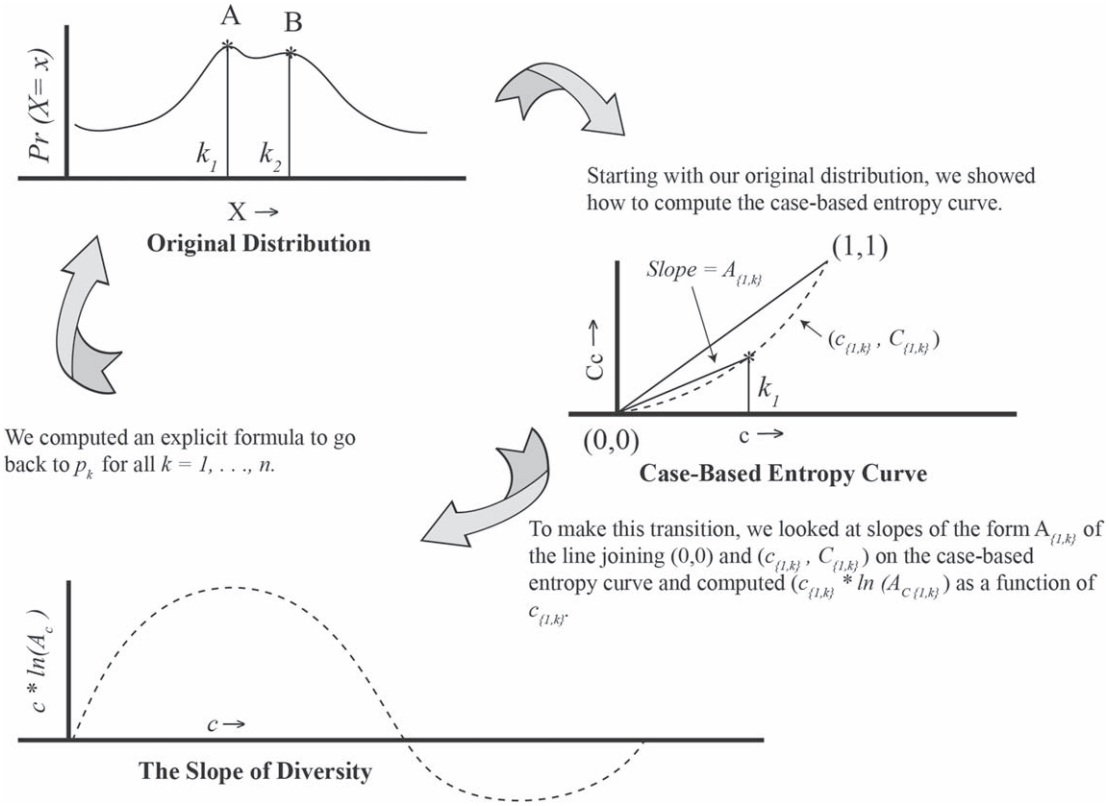

Figure 7. Reconstructing the original distribution.

(2) Let $g_2^a, g_2^b \in \mathcal{G}_2$. We will show below that $T_{2\to 3}(g_2^a) = T_{2\to 3}(g_2^b)$ implies that $g_2^a = g_2^b$.

$$T_{2\to 3}(g_2^a) = T_{2\to 3}(g_2^b) \Rightarrow (c^a_{\{1,k\}}, c^a_{\{1,k\}} * \ln(A^a_{\{1,k\}})) = (c^b_{\{1,k\}}, c^b_{\{1,k\}} * \ln(A^b_{\{1,k\}}))\ \forall k.$$

$$\Rightarrow c^a_{\{1,k\}} = c^b_{\{1,k\}} \text{ and } c^a_{\{1,k\}} * \ln(A^a_{\{1,k\}}) = c^b_{\{1,k\}} * \ln(A^b_{\{1,k\}})\ \forall k$$

$$\Rightarrow A^a_{\{1,k\}} = A^b_{\{1,k\}} \Rightarrow \frac{D^a_{\{1,k\}}}{D_K c^a_{\{1,k\}}} = \frac{D^b_{\{1,k\}}}{D_K c^b_{\{1,k\}}} \Rightarrow \frac{D^a_{\{1,k\}}}{D_K} = \frac{D^b_{\{1,k\}}}{D_K}\ \forall k$$

$$C_{\{1,k_1\}^a} = C_{\{1,k_2\}^b}\ \forall k.$$

$$\text{Hence, } T_{2\to 3}(g_2^a) = T_{2\to 3}(g_2^b) \Rightarrow g_2^a = g_2^b.$$

This shows that the map $T_{2\to 3}$ is injective.

(3) Let $g_3^a, g_3^b \in \mathcal{G}_3$. We will show below that $T_{3\to 1}(g_3^a) = T_{3\to 1}(g_3^b)$ implies that $g_3^a = g_3^b$.

$$T_{3\to 1}(g_3^a) = T_{3\to 1}(g_3^b) \Rightarrow p_k^a = p_k^b\ \forall \Rightarrow c_k^a = c_k^b\ \forall k$$

$$\text{Also, } D^a_{\{1,k\}} = c^a_{\{1,k\}} \prod_{i=1}^{k} \left(\frac{1}{p_i^a}\right)^{\frac{p_i^a}{c^a_{\{1,k\}}}} = c^b_{\{1,k\}} \prod_{i=1}^{k} \left(\frac{1}{p_i^b}\right)^{\frac{p_i^b}{c^b_{\{1,k\}}}} = D^b_{\{1,k\}}\ \forall k$$

$$\text{Hence, } A^a_{\{1,k\}} = \frac{D^a_{\{1,k\}}}{D_K c^a_{\{1,k\}}} = \frac{D^b_{\{1,k\}}}{D_K c^b_{\{1,k\}}} = A^b_{\{1,k\}}\ \forall k$$

$$\text{And hence, } (c^a_{\{1,k\}}, c^a_{\{1,k\}} * \ln(A^a_{\{1,k\}}) = (c^b_{\{1,k\}}, c^b_{\{1,k\}} * \ln(A^b_{\{1,k_2\}})\ \forall k.$$

$$\text{Hence, } T_{3\to 1}(g_3^a) = T_{3\to 1}(g_3^b) \Rightarrow g_3^a = g_3^b.$$

This shows that the map $T_{3\to 1}$ is injective.

Remark 5.4. We note that the inverse of an injective map is also injective. Hence, we could have shown that the inverses of the maps $T_{j\to k}$ are injective, and that would have also proved theorem 5.4. The key to any or all of those proof steps is that both coordinates should match for two points to be equal, and that forces the uniqueness of points because the equality of one of the coordinates (typically the x) leads to an equality of indices, probabilities or cumulative probabilities.

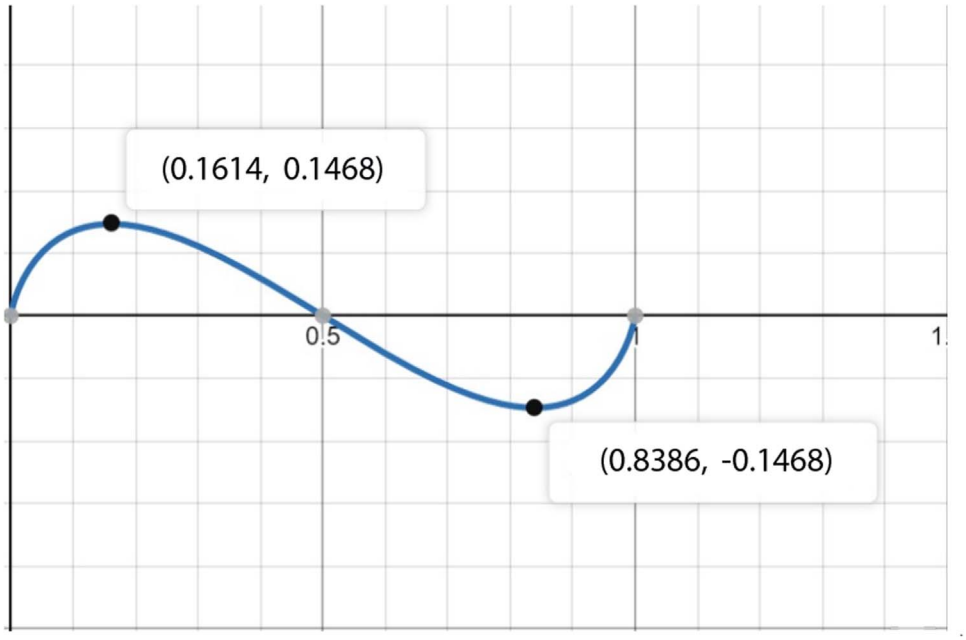

Figure 8. Graph of $c_{\{1,K\}}$ versus $c_{\{1,k\}}{}^{*}\ln A_{\{1,k\}}$ for the infinite geometric distribution.

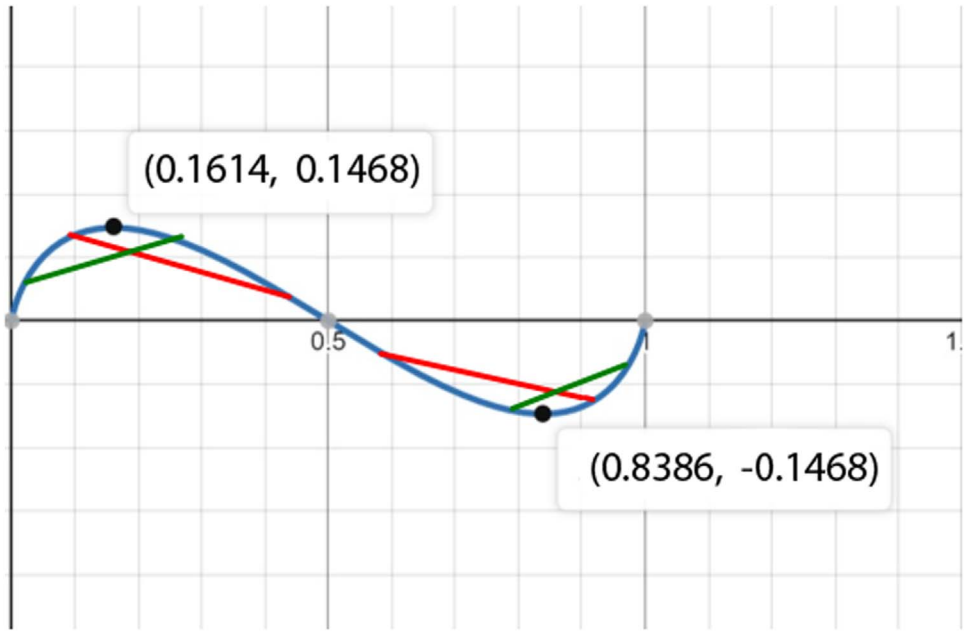

Figure 9. Graph of $c_{\{1,K\}}$ versus $c_{\{1,k\}}{}^{*}\ln A_{\{1,k\}}$ for the infinite geometric distribution showing secant lines on either side of $c = 0.5$ that have the same slope.

Remark 5.5. Theorem 5.4 is a significant improvement compared to the original notion of diversity as introduced by (Hill 1973) and (Jost 2006) in its own right. This is because, the Hill numbers ${}^{q}D$ are insensitive to rearrangements in the original distribution. In other words, any permutation of the probabilities in the original distribution will lead to the same value for the diversity ${}^{q}D$. This might be alright for qualitative distributions such as for species in a forest, since in such a context, we are only interested in the diversity of the distribution modulo permutations. However, the shape of the original probability distribution becomes extremely important in the context of a quantitative distribution such as for income, where we are interested in quantifying and comparing the amount of inequality (or degree of uniformity) that exists in different parts of the distribution. Given that the graph of $c_{\{1,k\}}$ versus $c_{\{1,k\}}{}^{*}\ln A_{\{1,k\}}$ or slope of diversity $g_3 \in \mathcal{G}_3$ is extremely useful to directly read off the D_P/c_P ratios (which measure the degree of uniformity of parts) by looking at slopes of secants, such

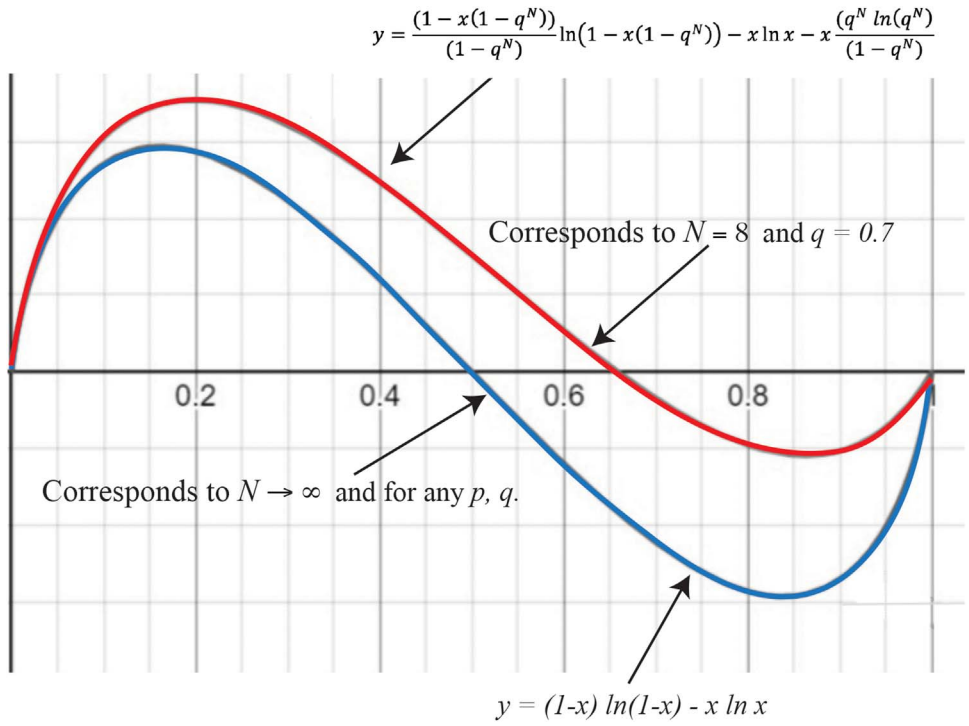

Figure 10. Graph of $c_{\{1,K\}}$ versus for the truncated finite (in red) and the infinite geometric distribution (in blue).

comparisons can be easily made from the graph of $g_2 \in \mathcal{G}_2$ indirectly by using $g_3 \in \mathcal{G}_3$.In this context, establishing the injectivity of the maps $T_{j\to k}$ is an important step that allows us go back and forth between the graphs $\mathcal{G}_j$ and $\mathcal{G}_k$. In essence, we have shown in theorem 5.4 that the shape of the original probability distribution $g_1 \in \mathcal{G}_1$ uniquely determines the case-based entropy curve $g_2 \in \mathcal{G}_2$ and the slope of diversity curve $g_3 \in \mathcal{G}_3$, and vice-versa.

6. Results relating the case-based entropy curve with the original probability distribution

In this section, we present some results that relate the case-based entropy curve i.e. $c_{\{1,k\}}$ versus $C_{\{1,k\}}$ curve to the original probability distribution. The importance of these results stems from the fact that for the first time, we have a connection between the variation of diversity of a given distribution in table 1 as measured by the case-based entropy curve $c_{\{1,k\}}$ versus $C_{\{1,k\}}$ and the slope of diversity curve $c_{\{1,k\}}$ versus the $c_{\{1,k\}}{}^*\ln A_{\{1,k\}}$ curve, and the probabilities p_k in the original distribution. This will achieve the objective of connecting the Hill numbers from (Jost 2006, Leinster and Cobbold 2012, Chao and Jost 2015, Hsieh *et al*, 2016, Pavoine *et al*, 2016, Jost 2019) to the shape of the original distribution.

Theorem 6.1. *Given a probability distribution like in table 1, the average case-based entropy per cumulative probability A_{P_i} are related as follows:*

$$\prod_{P_i\in\mathcal{P}} (A_{P_i})^{c_{P_i}} = 1. \tag{12}$$

Proof. Divide equation (5) by D_K and rewriting using the fact that $\sum_{P_i\in\mathcal{P}} c_{P_i} = 1$ we get:

$$1 = \frac{1}{D_K}\prod_{P_i\in\mathcal{P}}\left(\frac{D_{P_i}}{c_{P_i}}\right)^{c_{P_i}} = \prod_{P_i\in\mathcal{P}}\left(\frac{D_{P_i}}{D_K c_{P_i}}\right)^{c_{P_i}} = \prod_{P_i\in\mathcal{P}} (A_{P_i})^{c_{P_i}}.$$

Corollary 6.1. *Given a probability distribution like in table 1, and a part* P *with a disjoint partition given by $\cup_i P_i$, the average case-based entropy per cumulative probability* A_P *and A_{P_i} are related as follows:*

$$(A_P)^{c_P} = \prod_{P_i\in\mathcal{P}} (A_{P_i})^{c_{P_i}}. \tag{13}$$

Proof. This follows from dividing equation (7) in corollary 3.1 by the total diversity D_K.

Theorem 6.2. *Given a probability distribution like in table 1, we have the following:*

$$[C_{\{1,k\}} \underset{>}{\overset{<}{=}} c_{\{1,k\}}] \Leftrightarrow \left[\frac{D_{\{1,k\}}}{D_{\{(k+1),K\}}} \underset{>}{\overset{<}{=}} \frac{c_{\{1,k\}}}{c_{\{(k+1),K\}}}\right]. \tag{14}$$

W. efirst rewrite equation (5) for the partition given by $P_1 = \{1,\ k\}$ and $P_2 = \{(k+1),\ K\}$ to get the following:

$$D_K = \left(\frac{D_{\{1,k\}}}{c_{\{1,k\}}}\right)^{c_{\{1,k\}}}\left(\frac{D_{\{(k+1),K\}}}{c_{\{(k+1),K\}}}\right)^{c_{\{(k+1),K\}}} \Rightarrow C_{\{1,k\}} = \frac{D_{\{1,k\}}}{D_K} = c_{\{1,k\}}\left(\frac{\frac{D_{\{1,k\}}}{D_{\{(k+1),K\}}}}{\frac{c_{\{1,k\}}}{c_{\{(k+1),K\}}}}\right)^{c_{\{(k+1),K\}}}.$$

Hence, we have the following which proves the theorem:

$$\frac{C_{\{1,k\}}}{c_{\{1,k\}}} \underset{>}{\overset{<}{=}} 1 \Leftrightarrow \left(\frac{\frac{D_{\{1,k\}}}{D_{\{(k+1),K\}}}}{\frac{c_{\{1,k\}}}{c_{\{(k+1),K\}}}}\right)^{c_{\{(k+1),K\}}} \underset{>}{\overset{<}{=}} 1 \Leftrightarrow \left(\frac{\frac{D_{\{1,k\}}}{D_{\{(k+1),K\}}}}{\frac{c_{\{1,k\}}}{c_{\{(k+1),K\}}}}\right) \underset{>}{\overset{<}{=}} 1 \Leftrightarrow \frac{D_{\{1,k\}}}{D_{\{(k+1),K\}}} \underset{>}{\overset{<}{=}} \frac{c_{\{1,k\}}}{c_{\{(k+1),K\}}}.$$

Remark 6.1. Rearranging equation (14), we have

$$[C_{\{1,k\}} \underset{>}{\overset{<}{=}} c_{\{1,k\}}] \Leftrightarrow \left[\frac{D_{\{1,k\}}}{D_{\{(k+1),K\}}} \underset{>}{\overset{<}{=}} \frac{c_{\{1,k\}}}{c_{\{(k+1),K\}}}\right] \Leftrightarrow \left[\frac{D_{\{1,k\}}}{c_{\{1,k\}}} \underset{>}{\overset{<}{=}} \frac{D_{\{(k+1),K\}}}{c_{\{(k+1),K\}}}\right]. \tag{15}$$

This means that

$$[A_{\{1,k\}} \underset{>}{\overset{<}{=}} 1] \Leftrightarrow \left[\frac{D_{\{1,k\}}}{c_{\{1,k\}}} \underset{>}{\overset{<}{=}} \frac{D_{\{(k+1),K\}}}{c_{\{(k+1),K\}}}\right].$$

This means that if the average case-based entropy per unit cumulative frequency i.e., the slope of the line joining the points $(0, 0)$ and $(c_{\{1,k\}}, C_{\{1,k\}})$ on the case-based entropy curve is less than 1, then the portion of the original probability distribution with indices $\{1, k\}$ is less uniformly distributed compared to the portion of the original probability distribution with indices $\{k+1, K\}$.

Theorem 6.3. *Given a probability distribution like in table 1, we have the following for any fixed* k*:*

$$C_{\{1,k\}} \underset{>}{\overset{<}{}} c_{\{1,k\}} \Leftrightarrow C_{\{(k+1),K\}} \underset{<}{\overset{>}{}} c_{\{(k+1),K\}}. \tag{16}$$

Proof.

$$D_K = \left(\frac{D_{\{1,k\}}}{c_{\{1,k\}}}\right)^{c_{\{1,k\}}}\left(\frac{D_{\{(k+1),K\}}}{c_{\{(k+1),K\}}}\right)^{c_{\{(k+1),K\}}} \Rightarrow \left(\frac{D_{\{1,k\}}}{D_K c_{\{1,k\}}}\right)^{c_{\{1,k\}}}\left(\frac{D_{\{(k+1),K\}}}{D_K c_{\{(k+1),K\}}}\right)^{c_{\{(k+1),K\}}} = 1$$

$$\Rightarrow \left(\frac{C_{\{1,k\}}}{c_{\{1,k\}}}\right)^{c_{\{1,k\}}}\left(\frac{C_{\{(k+1),K\}}}{c_{\{(k+1),K\}}}\right)^{c_{\{(k+1),K\}}} = 1.$$

From the last equation above, this means the following:

$$C_{\{1,k\}} \underset{>}{\overset{<}{}} c_{\{1,k\}} \Leftrightarrow \frac{C_{\{1,k\}}}{c_{\{1,k\}}} \underset{>}{\overset{<}{}} 1 \Leftrightarrow \left(\frac{C_{\{1,k\}}}{c_{\{1,k\}}}\right)^{c_{\{1,k\}}} \underset{>}{\overset{<}{}} 1 \Leftrightarrow \left(\frac{C_{\{(k+1),K\}}}{c_{\{(k+1),K\}}}\right)^{c_{\{(k+1),K\}}} \underset{<}{\overset{>}{}} 1$$

$$\Leftrightarrow \left(\frac{C_{\{(k+1),K\}}}{c_{\{(k+1),K\}}}\right) \underset{<}{\overset{>}{}} 1 \Leftrightarrow C_{\{(k+1),K\}} \underset{<}{\overset{>}{}} c_{\{(k+1),K\}}.$$

This proves the theorem.

Theorem 6.4. *Given a probability distribution like in table 1, and given a fixed index* k*, we have the following:*

$$(D_{\{1,k\}} p_{k\{1,k\}})^{p_{k\{1,k\}}} = \left(\frac{A_{\{1,(k-1)\}}}{A_{\{1,k\}}}\right)^{c_{(k-1)\{1,k\}}} \text{ for any } k. \tag{17}$$

Proof. We rewrite equation (5) by using remark 3.2 restricting ourselves to a partition $\mathcal{P}_1 = \{1,\ldots,(k-1)\}$, $\mathcal{P}_2 = \{k\}$ of the index set $\{1,\ldots,k\}$. Now, we rewrite equation (5) for this new partition using the fact that $p_{k\{1,k\}} + c_{(k-1)\{1,k\}} = 1$ as below:

$$D_{\{1,k\}} = \left(\frac{D_{\{1,(k-1)\}}}{c_{(k-1)\{1,k\}}}\right)^{c_{(k-1)\{1,k\}}} \frac{1}{(p_{k\{1,k\}})^{p_{k\{1,k\}}}}$$

$$\Rightarrow (D_{\{1,k\}}p_{k\{1,k\}})^{p_{k\{1,k\}}} = \left(\frac{D_{\{1,(k-1)\}}}{D_{\{1,k\}}c_{(k-1)\{1,k\}}}\right)^{c_{(k-1)\{1,k\}}} = \left(\frac{A_{\{1,(k-1)\}}}{A_{\{1,k\}}}\right)^{c_{(k-1)\{1,k\}}}.$$

Remark 6.2. We remark that the following variation of equation (17) is true as well and the proof follows the same lines and using the fact that $p_{k\{k,K\}} + (1 - c_k)_{\{k,K\}} = 1$:

$$D_{\{k,K\}} = \left(\frac{D_{\{(k+1),K\}}}{(1 - c_k)_{\{k,K\}}}\right)^{(1-c_k)_{\{k,K\}}} \frac{1}{(p_{k\{k,K\}})^{p_{k\{k,K\}}}}$$

$$\Rightarrow (D_{\{k,K\}}p_{k\{k,K\}})^{p_{k\{k,K\}}} = \left(\frac{D_{\{(k+1),K\}}}{D_{\{k,K\}}(1 - c_k)_{\{k,K\}}}\right)^{(1-c_k)_{\{k,K\}}} = \left(\frac{A_{\{(k+1),K\}}}{A_{\{k,K\}}}\right)^{(1-c_k)_{\{k,K\}}}. \quad (18)$$

Theorem 6.5. *Given a probability distribution like in table 1, we have the following:*

$$C_{\{1,k\}} = c_{\{1,k\}}\ \forall k = 1,..,K \Leftrightarrow p_k = \frac{1}{D_K}\ \forall k = 1,..,K. \quad (19)$$

Proof.

$$C_{\{1,k\}} = c_{\{1,k\}}\ \forall k = 1,..,K \Leftrightarrow A_{\{1,k\}} = 1\ \forall k = 1,\ldots,K.$$

We use theorem 6.4 to get the following:

$$\Leftrightarrow (D_{\{1,k\}}p_{k\{1,k\}})^{p_{k\{1,k\}}} = 1 \Leftrightarrow p_{k\{1,k\}} = \frac{1}{D_{\{1,k\}}}$$

$$\Leftrightarrow p_k = \frac{c_k}{D_{\{1,k\}}} = \frac{1}{D_K}\frac{D_K c_k}{D_{\{1,k\}}} = \frac{1}{A_{\{1,k\}}D_K} = \frac{1}{D_K}\ \forall k = 1,\ldots,K.$$

In other words, the case-based entropy curve is a straight line joining (0, 0) and (1, 1) if and only if the original distribution is uniform.

Theorem 6.6. *Given a probability distribution like in table 1, we have the following:*

$$C_{\{1,k\}} < c_{\{1,k\}}\ \forall k > L \text{ for some } L \Leftrightarrow p_k < \frac{1}{D_K A_{\{1,k\}}}\ \forall k > M \text{ for some } M \quad (20)$$

Proof. First, we note that

$$C_{\{1,k\}} < c_{\{1,k\}} \Leftrightarrow A_{\{1,k\}} < 1\ \forall k > L.$$

We also note that $A_{\{1,K\}} = 1$. This means that there exists some $M > L$ so that

$$A_{\{1,M\}} < A_{\{1,(M+1)\}} < \ldots < A_{\{1,(k-1)\}} < A_{\{1,K\}} = 1,$$

or in other words, the sequence $\{A_{\{1,k\}}\}_{k=M}^{K}$ is an increasing sequence that converges to 1, for some $M > L$. Now, we have the following from equation (17):

$$A_{\{1,k\}} < 1\,\forall k > M \text{ for some } M > L$$

$$\Rightarrow (D_{\{1,k\}}p_{k\{1,k\}})^{p_{k\{1,k\}}} = \left(\frac{A_{\{1,(k-1)\}}}{A_{\{1,k\}}}\right)^{c_{k-1\{1,k\}}} < 1 \Rightarrow p_{k\{1,k\}} < \frac{1}{D_{\{1,k\}}} \Rightarrow p_k < \frac{c_k}{D_{\{1,k\}}}$$

$$= \frac{1}{D_K}\frac{D_K c_k}{D_{\{1,k\}}} = \frac{1}{A_{\{1,k\}}D_K}\ \forall k > M$$

Going the other direction, we have:

$$p_k < \frac{1}{A_{\{1,k\}}D_K}\forall k > M \Rightarrow (D_{\{1,k\}}p_{k\{1,k\}}) < 1\ \forall k > M$$

$$\Rightarrow (D_{\{1,k\}}p_{k\{1,k\}})^{p_{k\{1,k\}}} < 1\ \forall k > M \Rightarrow \left(\frac{A_{\{1,(k-1)\}}}{A_{\{1,k\}}}\right)^{c_{(k-1)\{1,k\}}} < 1\ \forall k > M$$

$$\Rightarrow \left(\frac{A_{\{1,(k-1)\}}}{A_{\{1,k\}}}\right) < 1 \Rightarrow A_{\{1,(k-1)\}} < A_{\{1,k\}}\ \forall k > M.$$

Since $A_{\{1,K\}} = 1$, we have that

$$A_{\{1,k\}} < 1\ \forall k > M + 1 \Rightarrow C_{\{1,k\}} < c_{\{1,k\}}\ \forall k > M + 1,$$

and the sequence $\{A_{\{1,k\}}\}_{k=M+1}^{K}$ is an increasing sequence that converges to 1. This proves the the theorem.

Remark 6.3. In particular, if the sequence of average case-based entropies per unit frequency given by $\{A_{\{1,k\}}\}_{k=M+1}^{K}$ increases to 1 very fast (or very slow), then the probabilities given by $\{p_k\}_{k=M+1}^{K}$ will form a right tail sequence of decreasing probabilities that decreases quickly (or slowly).

Theorem 6.7. *Given a probability distribution like in table 1, we have the following:*

$$C_{\{1,k\}} > c_{\{1,k\}}\ \forall k < L \text{ for some } L \Leftrightarrow p_k < \frac{1}{D_K A_{\{k,K\}}}\ \forall k < M \text{for some} M. \tag{21}$$

F. irst we note from remark 6.2, theorem 6.1 and theorem 6.3 that

$$C_{\{1,k\}} > c_{\{1,k\}} \Leftrightarrow C_{\{(k+1),K\}} < c_{\{(k+1),K\}} \Leftrightarrow A_{\{(k+1),K\}} < 1\ \forall k < L.$$

We also note that $A_{\{1,K\}} = 1$. This means that there exists some $M > L$ so that

$$A_{\{M,K\}} < A_{\{(M-1),K\}} < \ldots < A_{\{2,K\}} < A_{\{1,K\}} = 1,$$

or in other words, the sequence $\{A_{\{(M-k),K\}}\}_{k=0}^{(M-1)}$ is an increasing sequence that converges to 1, for some $M > L$. Now, we have the following from remark 6.2:

$$A_{\{(M-k),K\}} < 1\, \forall k < M \ \text{ for some } M > L$$

$$\Rightarrow (D_{\{k,K\}} p_{k\{k,K\}})^{p_{k\{k,K\}}} = \left(\frac{A_{\{(k+1),K\}}}{A_{\{k,K\}}}\right)^{(1-c_k)_{\{k,K\}}} < 1 \Rightarrow p_{k\{k,K\}} < \frac{1}{D_{\{k,K\}}}$$

$$\Rightarrow p_k < \frac{(1-c_{(k-1)})}{D_{\{k,K\}}} = \frac{D_K(1-c_{(k-1)})}{D_K D_{\{k,K\}}} = \frac{1}{D_K A_{\{k,K\}}}\ \forall k < M$$

Going the other direction, we have:

$$p_k < \frac{1}{A_{\{(k+1),K\}} D_K}\, \forall k < M \Rightarrow (p_{k\{(k+1),K\}} D_{\{(k+1),K\}}) < 1\ \forall k < M.$$

$$\Rightarrow (p_{k\{(k+1),K\}} D_{\{(k+1),K\}})^{p_{k\{(k+1),K\}}} < 1\ \forall k < M \Rightarrow \left(\frac{A_{\{(k+2),K\}}}{A_{\{(k+1),K\}}}\right)^{(1-c_{(k+1)})_{\{(k+1),K\}}} < 1\ \forall k > M$$

$$\Rightarrow \left(\frac{A_{\{(k+2),K\}}}{A_{\{(k+1),K\}}}\right) < 1 \Rightarrow A_{\{(k+2),K\}} < A_{\{(k+1),K\}}\ \forall k < M.$$

Since $A_{\{1,K\}} = 1$, we have that

$$A_{\{(k+1),K\}} < 1\ \forall k < M+1 \Rightarrow C_{\{(k+1),K\}} < c_{\{(k+1),K\}}\ \forall k < M+1,$$

$$\Rightarrow C_{\{1,k\}} > c_{\{1,k\}}\ \forall k < M+1$$

and the sequence $\{A_{\{(M+1-k),K\}}\}_{k=0}^{M}$ is an increasing sequence that converges to 1. This proves the theorem.

Remark 6.4. In particular, if the sequence of average case-based entropies per unit frequency given by $\{A_{\{(M-k),K\}}\}_{k=0}^{(M-1)}$ increases to 1 very fast (or very slow), then the probabilities given by $\{p_k\}_{k=1}^{M}$ will form a left tail sequence of increasing probabilities that increases quickly (or slowly).

Remark 6.5. We note that theorems 6.6 and 6.7 while giving us a sense of existence of tails on the right and left ends of the original probability distributions, only give us conservative estimates on what exactly those probabilities could be. We will see below that we can improve this significantly using the $c_{\{1,k\}}$ versus $c_{\{1,k\}}$*ln $A_{\{1,k\}}$ (slope of diversity) curve.

Theorem 6.8. *Given a probability distribution like in table 1, then we have the following:*

$$p_i \underset{>}{\overset{<}{=}} p_{i+1} \Leftrightarrow S_{\{i-1,i\}} \underset{<}{\overset{>}{=}} S_{\{i,i+1\}}. \tag{22}$$

Proof. We know that

$$\frac{1}{p_k} = D_K e^{S_{\{k-1,k\}}}.$$

Now, if

$$p_i \underset{>}{\overset{<}{=}} p_{i+1}, \tag{23}$$

then

$$\frac{1}{p_i} \begin{matrix} > \\ = \\ < \end{matrix} \frac{1}{p_{i+1}} \Leftrightarrow e^{S_{\{i-1,i\}}} \begin{matrix} > \\ = \\ < \end{matrix} e^{S_{\{i,i+1\}}} \Leftrightarrow S_{\{i-1,i\}} \begin{matrix} > \\ = \\ < \end{matrix} S_{\{i,i+1\}}.$$

This proves the Theorem.

Remark 6.6. If $S_{\{i-1,i\}} < S_{\{i,i+1\}}$ for all $i = k,\ldots,K-1$ then $p_i \geqslant p_{i+1}$ for all $i = k,\ldots,K$ leading to a right tail. Also, if $S_{\{i-1,i\}} > S_{\{i,i+1\}}$ for all $i = 1,\ldots,k-1$, then $p_i \leqslant p_{i+1}$ for all i=1,…,k leading to a left tail of probabilities. Equation (23) explicitly dictates how sharply the probabilities decay at either tail, depending on how sharply the slopes $S_{\{k-1,k\}}$ decay in the $c_{\{1,k\}}$ versus $c_{\{1,k\}}$*ln $A_{\{1,k\}}$ curve. Hence, this vastly improves theorems 6.6 and 6.7.

7. Geometric distribution example

In this section, we use the geometric distribution as an example to illustrate some of the results related to the $c_{\{1,k\}}$ versus $c_{\{1,k\}}$*ln $c_{\{1,k\}}$ (slope of diversity) curve. Since most of the results from the $c_{\{1,k\}}$ versus $C_{\{1,k\}}$ curve (the case-based entropy curve) are subsumed by the $c_{\{1,k\}}$ versus $c_{\{1,k\}}$*ln $c_{\{1,k\}}$ curve, we only focus on demonstrating the usefulness of the latter using the geometric distribution.

We consider a geometric random variable $X = 1, 2, 3,...$ and define

$$p_i = P(X = i) = pq^{(i-1)},$$
$$H = \frac{-p\ln(p) - q\ln(q)}{p}, \text{ and}$$
$$D = \frac{1}{pq^{q/p)}}.$$

The entropy H and the diversity D can be easily calculated.

We truncate up to $i = K$ and re-normalize the probabilities $\hat{p}_i$ (so that they add up to 1) to get

$$\hat{p}_i = \frac{pq^{(i-1)}}{(1-q^K)},$$
$$c_{\{1,K\}} = 1 - q^K,$$
$$H_{\{1,K\}} = H + \frac{Kq^K}{(1-q^K)}\ln(q), \text{ and}$$
$$D_{\{1,K\}} = Dq^{\frac{Kq^K}{(1-q^k)}} = D(q^k)^{q^K/(1-q^K)}.$$

Notice that $q^K = 1 - c_{\{1,K\}}$, so $D_{\{1,K\}} = D(1-c_{\{1,K\}})^{(1-c_{\{1,K\}})/c_{\{1,K\}}}$. Thus,

$$C_{\{1,K\}} = \frac{D_{\{1,K\}}}{D} = (1-c_{\{1,K\}})^{(1-c_{\{1,K\}})/c_{\{1,K\}}}$$
$$A_{\{1,K\}} = \frac{D_{\{1,K\}}}{D \cdot c_{\{1,K\}}} = \frac{(1-c_{\{1,K\}}}{c_{\{1,K\}}}$$
$$\ln(A_{\{1,K\}}) = \frac{1-c_{\{1,K\}}}{c_{\{1,K\}}}\ln(1-c_{\{1,K\}}) - \ln(c_{\{1,K\}})$$
$$c_{\{1,K\}}\ln(A_{\{1,K\}}) = (1-c_{\{1,K\}})\ln(1-c_{\{1,K\}}) - c_{\{1,K\}}\ln(c_{\{1,K\}}).$$

So, interestingly enough, for the geometric distribution with infinite support, we have an explicit expression for $c_{\{1,K\}}\ln(A_{\{1,K\}})$ as a function of $c_{\{1,K\}}$, and it is independent of p or q.

Let $\hat{K}$ be the index corresponding to $c_{\{1,\hat{K}\}} = 0.5$. Then,

$$(0.5)\ln(A_{\{1,\hat{K}\}} = 0.5\ln(0.5) - 0.5\ln(0.5) = 0.$$

This implies that $A_{\{1,\hat{K}\}} = 1$ and that $\frac{D_{\{1,\hat{K}\}}}{c_{\{1,\hat{K}\}}} = D$. This also implies that $A_{\{\hat{K}+1,\infty\}} = 1$ and $\frac{D_{\{\hat{K}+1,\infty\}}}{c_{\{\hat{K}+1,\infty\}}} = D$.

So, no matter the choice of p and q, all geometric distributions satisfy the property that $c_{\{1,\hat{K}\}} = 0.5$ splits the distribution into two parts, both of which have the same number of SEE types equal to the diversity of the entire distribution as shown by figure 8.

So, the SEE types of the first half up to $c = 0.5$ and the second half from $c = 0.5$ to $c = 1$ are equal and equal to D, which is the total diversity of the entire distribution. That's an interesting result since the number of types $\hat{K}$ corresponding to $c_{\{1,\hat{K}\}} = 0.5$ is finite which means there are infinitely many types from $c = 0.5$ to $c = 1$, and yet

$$\frac{D_{c_{\{1,\hat{k}\}}}}{c_{\{1,\hat{K}\}}} = \frac{D_{c_{\{\hat{k}+1,\infty\}}}}{c_{\{\hat{K}+1,\infty\}}} = D.$$

In fact, much more is true. Let $y = c_{\{1,K\}}\ln(A_{c_{\{1,K\}}})$ and $x = c_{\{1,k\}}$. Then, $y = f(x) = (1-x)\ln(1-x) - x\ln(x)$ is symmetric about $x = 0.5$ so that $f(0.5+t) = -f(0.5-t)$ for $0 < t < 0/5$.

This means that for every secant line to the left of $x = 0.5$, we can find another matching secant line to the right of $x = 0.5$ with the same slope, as the figure illustrates. That is, for any subset of types $\{k_1, k_2\}$ to the left of $x = 0.5$, there is an equivalent subset of types $\{k_3, k_4\}$ to the right of $x = 0.5$ (possibly with more types) that have the same number of SEE types as shown in figure 9.

Since $f(c) = -f(1-c)$ and $f(x) = (1-x)\ln(1-x) - x\ln(x)$, we know that

$$\begin{aligned} S_1 &= \frac{f(c_1) - f(c_2)}{(c_1 - c_2)} \\ &= \frac{-f(-f(1-c_1) - (-f(1-c_2))}{(c_1 - c_2)} \\ &= \frac{f(1-c_2) - f(1-c_1)}{(1-c_2) - (1-c_1)} \\ &= S_2, \end{aligned}$$

So, in fact the $\frac{D}{c}$ value (number of SEE types) for the part of the distribution from (c_1, c_2) is equal to the $\frac{D}{c}$ value from $(1-c_1, 1-c_2)$. Note that if $c_1 < c_2 < 0.5$ then $0.5 < (1-c_2) < (1-c_1)$ is on the other side of $c = 0.5$.

7.1. Truncated geometric distribution

Now, we look at the truncated geometric distribution: $x = 1,\ldots,N$. The probabilities are normalized to add up to 1 as below:

$$p_i = \frac{pq^{(i-1)}}{(1-q^N)} = P(X = i)$$

Let's concentrate on the part from $\{1, K\}$. Then we have

$$\begin{aligned} c_{\{1,K\}} &= \sum_{i=1}^{K} p_i \\ &= \frac{p}{(1-q^N)} \sum_{i=1}^{K} q^{(i-1)} \\ &= \frac{p(1-q^K)}{(1-q^N)(1-q)} \\ &= \left(\frac{1-q^K}{1-q^N}\right). \end{aligned}$$

We again normalize the probabilities for the part $\{1, K\}$ to get $\hat{p}_i = \frac{pq^{(i-1)}}{(1-q^K)}$, for $i = 1,\ldots,K$. So, all of the formulas for $\{1, K\}$ are the same as the $\{1, K\}$ formulas from the previous section. Thus, for $\{1, N\}$, we replace all formulas from $\{1, K\}$ with $K = N$ as follows:

$$\begin{aligned} D &= \frac{p\ln(p) - q\ln(q)}{p} = \ln\left(\frac{p}{q^{q/p}}\right); \\ D_{\{1,N\}} &= Dq^{Nq^N/(1-q^N)}; \\ D_{\{1,K\}} &= Dq^{Kq^K/(1-q^K)}; \text{ and} \\ C_{\{1,K\}} &= \frac{D_{\{1,K\}}}{D_{\{1,N\}}} = q^{\left\{\frac{Kq^K}{(1-q^K)} - \frac{Nq^N}{(1-q^N)}\right\}}. \end{aligned}$$

Recall that

$$\begin{aligned} c_{\{1,K\}} &= \frac{1-q^K}{1-q^N}, \\ 1 - q^K &= c_{\{1,K\}}(1-q^N), \text{ and} \\ q^K &= 1 - c_{\{1,K\}}(1-q^N). \end{aligned}$$

So,

$$C_{\{1,K\}} = \frac{[1 - c_{\{1,K\}}(1 - q^N)]^{\frac{1-c_{\{1,K\}}(1-q^N)}{c_{\{1,K\}}(1-q^N)}}}{(q^N)^{(q^N/(1-q^N)}}$$

and

$$\begin{aligned} A_{\{1,K\}} &= \frac{C_{\{1,K\}}}{c_{\{1,K\}}} \\ &= \frac{[1 - c_{\{1,K\}}(1 - q^N)]^{\frac{1-c_{\{1,K\}}(1-q^N)}{c_{\{1,K\}}(1-q^N)}}}{c_{\{1,K\}}\,(q^N)^{(q^N/(1-q^N)}} \ln(A_{\{1,K\}}) \\ &= \left(\frac{1 - c_{\{1,K\}}(1 - q^N)}{c_{\{1,K\}}(1 - q^N)}\right) \ln(1 - c_{\{1,K\}}(1 - q^N)) - \ln(c_{\{1,K\}}) - \frac{q^N}{(1 - q^N)} \ln(q^N). \end{aligned}$$

Thus,

$$c_{\{1,K\}} \ln(A_{\{1,K\}}) = \frac{1 - c_{\{1,K\}}(1 - q^N)}{(1 - q^N)} \ln(1 - c_{\{1,K\}}(1 - q^N)) - c_{\{1,K\}} \ln(c_{\{1,K\}}) - c_{\{1,K\}} \frac{q^N}{(1 - q^N)} \ln(q^N).$$

Hence, for the truncated and normalized geometric distribution, the formula for $c_{\{1,K\}}$*ln $A_{\{1,K\}}$ is not independent of p and q. The blue graph in figure 10 corresponds to N approaching infinity, for any p and q. The red graph corresponds to $N = 8$ and $q = 0.7$.

8. Conclusion

Given the real-world challenges of measuring diversity we had two objectives for this study. First, to introduce and justify the ratio D_P/c_P as a measure of degree of uniformity of a part of a given distribution in table 1. Second, to prove results that concretely link the case-based entropy curve and the original probability distribution (via the slope of diversity curve), thereby (for the first time), establishing an explicit and concrete link between the diversity of parts of a distribution and the original probabilities themselves. We have achieved both objectives in this paper, and also demonstrated how to compute some of the quantities such as the $c_{\{1,k\}}$ versus $c_{\{1,k\}}$*ln $A_{\{1,k\}}$ curve for the geometric distribution, which we call the *slope of diversity*.

These two results are an important step towards concretely comparing and contrasting the degrees of uniformity of parts of a given probability distribution within and across different distributions, given that most real-world systems have unequal distributions, varying frequencies, and comprise multiple diversity types with unknown frequencies that can change. Such systems, as we mentioned in the introduction, include income distributions, economic complexity indices, ecological systems, species diversity, and ranking systems, from genes and exposomic biological assays to measures of economic and health inequality. For example, returning to the Gini coefficient from the introduction, our approach allows for several advances. First, because our approach does not conflate different distributions with the same coefficient, we can provide a unique case-based entropy or slope of diversity curve for each and every income distribution.

Second, we can also provide, for any given country's income distribution, the precise quantification of the relationship between the probability of each income level (p_i) and the total income diversity D for any country, both among parts of their respective income distribution as well as the whole using the case-based entropy and the slope of diversity curves. The Gini index cannot do that, for example.

Third, we have also established a concrete quantitative means of comparing the degree of uniformity of parts of a distribution. Such a comparison is extremely important in studying the prevalence of inequality (as in the case of incomes, for example) in whole distributions and their parts. A quantifiable measure of the degree of uniformity (or inequality) for quantitative variables such as income and resources, will pave the way to formulate policies that will lead to equity in distribution of resources, and also measure such an achievement by using the D_P/c_P ratio.

Fourth, we have closed the gap that exists in the literature on diversity measures by explicitly relating the diversity of parts of a distribution to the probabilities in the original distribution. We have also shown, to repeat a point, in theorem 5.4 that the shape of the original distribution uniquely determines the diversity of its parts and vice-versa. Furthermore, we have also shown how to explicitly compute the individual probabilities of the original distribution, as in the case of income for example, from the case-based entropy curve. This is a significant step towards linking the concept of diversity to the shape of the original distribution which, as we have commented in remark 5.5, is extremely important in quantifying and locating regions in the original distribution that are more or less unequally distributed.

In a sense, the two objectives of the paper are inter-twined in the following way: Diversity (or the ratio D_P/c_P) is a measure of uniformity of a distribution, and hence we need to justify its use and show that D_P/c_P can be computed for any part of a distribution easily, which was the first objective. Given the distribution of diversity in the form of the case-based entropy curve, computation of D_P/c_P, and its use as a measure of uniformity of a given part of a distribution would be meaningless unless the variation of diversity of parts in a given distribution (in the form of the case-based entropy curve) uniquely determines the original distribution. This was the point of the second objective. In summary, we need to know how to measure and quantify inequality within parts of a distribution, need to know how to compute such a quantification, and we also need to be reassured that there is a one-to-one correspondence between such a computation and the variation of probabilities in the original distribution (i.e., the shape of the original distribution). The last point is important, thinking about such measures as the Gini index, as we do not want two different distributions that have completely different shapes to lead to the same quantification of inequality, as it would then be difficult to pinpoint the original distribution (or its parts) by simply studying the variation of inequality (or diversity).

We conclude by stating that in our future work, we will endeavor to extend the results in this paper to continuous distributions, and also try to apply the results to improve any existing measures of inequality in the context of quantitative distributions, specifically the Gini index.

Data availability statement

No new data were created or analysed in this study.

Appendix. Notation

(1) K: The number of types in a distribution.

(2) D_K: Diversity of the entire distribution i.e., all K types.

(3) D_P: Diversity of the part P

(4) c_P: Sum of probabilities (or cumulative probability) of the part P.

(5) $\mathcal{P}$: An ascending disjoint partition of the set of indices $\{1,..,K\}$ such that every element $P_i \in \mathcal{P}$ satisfies the property that $i < j \Rightarrow \max P_i < \max P_j$. In other words, the partition preserves the ordering of the numbers $\{1,..,K\}$. In particular, the member $\{i,(i+1),..,j\}$ denotes the types in the distribution between indices i and j and will be denoted by $\{i,j\}$.

(6) D_{P_i}: Diversity of the part of the distribution corresponding to indices in $P_i \in \mathcal{P}$.

(7) $c_{P_i} = \sum_{l \in P_i} p_l$: sum of probabilities of types in the part of the distribution corresponding to indices in $P_i \in \mathcal{P}$.

(8) $p_{l\{1,i\}} = \frac{p_l}{c_{\{1,i\}}} = \frac{p_l}{\sum_{k=1}^{i} p_k} = \frac{f_l}{\sum_{k=1}^{i} f_k}$: conditional probabilities for the first i types $l = 1,..,i$. Same definition for p_{lP_i} for any partition $P_i \in \mathcal{P}$ i.e., $p_{lP_i} = \frac{p_l}{c_{P_i}} = \frac{p_l}{\sum_{k \in P_i} p_k} = \frac{f_l}{\sum_{k \in P_i} f_k}$.

(9) $c_{l\{1,i\}} = \frac{c_l}{c_{\{1,i\}}} = \frac{c_l}{\sum_{k=1}^{i} p_k}$: conditional cumulative probabilities for the first i types $l = 1,..,i$. Same definition for c_{lP_i} for any partition $P_i \in \mathcal{P}$ i.e., $c_{lP_i} = \frac{c_l}{c_{P_i}} = \frac{c_l}{\sum_{k \in P_i} p_k}$. So in general, whenever there is a partition P_i as a subscript, it means that we are dividing the probability (or cumulative probability) in the base by c_{P_i}.

(10) $A_{P_i} = \frac{C_{P_i}}{c_{P_i}} = \frac{D_{P_i}}{D_K \cdot c_{P_i}}$: Average case-based entropy per unit cumulative probability of the part of the distribution corresponding to indices in $P_i \in \mathcal{P}$.

(11) $S_{\{k_1,k_2\}}$= slope of the line joining $(c_{\{1,k_1\}},\ c_{\{1,k_1\}} \ln(A_{\{1,k_1\}}))$ and $(c_{\{1,k_2\}},\ c_{\{1,k_2\}} \ln(A_{\{1,k_2\}}))$ on the $c_{\{1,k\}}$ versus $c_{\{1,k\}} \ln(A_{\{1,k\}})$ (or slope of diversity) curve.

(12) $\mathcal{G}_1$: This is the set of all probability distributions like in table 1 with elements denoted by g_1.

(13) $\mathcal{G}_2$: This is the set of all case-based entropy curves with elements denoted by g_2.

(14) $\mathcal{G}_3$: This is the set of all slope of diversity curves with elemetns denoted by g_3.

ORCID iDs

Rajeev Rajaram https://orcid.org/0000-0003-1680-6706

References

Chao A and Jost L 2015 Estimating diversity and entropy profiles via discovery rates of new species *Methods in Ecology and Evolution* **6** 873–82

Gaggiotti O E, Chao A, Peres-Neto P, Chiu C-H, Edwards C, Fortin M-J, Jost L, Richards C M and Selkoe K A 2018 Diversity from genes to ecosystems: a unifying framework to study variation across biological metrics and scales *Evolutionary Applications* **11** 1027–93

Hill M 1973 Diversity and evenness: a unifying notation and its consequences *Ecology* **54** 427–32

Hsieh T, Ma K and Chao A 2016 Inext: an r package for rarefaction and extrapolation of species diversity (hill numbers) *Methods in Ecology and Evolution* **7** 1451–6

Jost L 2006 Entropy and diversity *Oikos* **113** 363–75

Jost L 2019 What do we mean by diversity? the path towards quantification *Mètode Science Studies Journal-Annual Review* **9** 55–61

Leinster T 2021 *Entropy and Diversity: The Axiomatic Approach* (Cambridge, UK: Oxford University Press) (https://doi.org/10.1017/9781108963558)

Leinster T and Cobbold C A 2012 Measuring diversity: the importance of species similarity *Ecology* **93** 477–89

MacArthur R 1965 Patterns of species diversity *Biological Review* **40** 510–33

Pavoine S, Marcon E and Ricotta C 2016 'Equivalent numbers' for species, phylogenetic or functional diversity in a nested hierarchy of multiple scales *Methods in Ecology and Evolution* **7** 1152–63

Peet R 1974 The measurement of species diversity *Annual Review of Ecological Systems* **5** 285–307

Rajaram R and Castellani B 2016 An entropy based measure for comparing distributions of complexity *Physica A: Statistical Mechanics and its Applications* **453** 35–43

Rajaram R and Castellani B 2020 Diversity in complex systems: measuring parts of the distribution to the whole *J. Phys. Commun.* **4** 045008