Principles for Mining Summaries:

Theorems and Proofs

Robert J. Hilderman and Howard J. Hamilton

Department of Computer Science, University of Regina

Regina, Saskatchewan, Canada S4S 0A2

E-mail: {hilder,hamilton}@cs.uregina.ca

Abstract

An important problem in the area of data mining is the development of effective measures of interestingness for ranking discovered knowledge. In this paper, we investigate the problem of ranking the interestingness of summaries, where a single dataset can be generalized in many different ways and to many levels of granularity. We propose five principles that any measure must satisfy to be considered useful for ranking summaries generated from databases, and perform a comparative analysis of sixteen well-known diversity measures to determine those which satisfy the proposed principles. The sixteen diversity measures have previously been utilized in various disciplines, such as information theory, statistics, ecology, and economics. Their use for ranking summaries is a new application area. The objective of this work is to gain some insight into the behaviour that can be expected from each of the measures in practice. From these analyses, we conclude that six of the measures are most useful within the context of ranking summaries generated from databases.

1 Introduction

An important problem in the area of data mining is the development of effective measures of interestingness for ranking discovered knowledge. Numerous techniques for determining the interestingness of discovered knowledge have recently been reported in the literature. For example, in In [3], it is shown that the most interesting rules reside along a support/confidence border. A technique is described in [18] that discovers interesting rules by eliminating those that are not interesting. And an extensive survey of recently proposed interestingness measures is described in [9].

In this paper, we investigate the problem of ranking the interestingness of summaries generated from databases, where a single dataset can be generalized in many different ways and to many levels of granularity. We propose five principles that any diversity measure must satisfy to be considered useful for ranking summaries generated from databases. We perform a comparative analysis of sixteen diversity measures to determine those that satisfy the proposed principles. This theory provides a foundation for an intuitive understanding of the term "interestingness" when used within this context. Finally, we analyze the distribution of the index values generated by each of the measures. The objective of this work is to gain some insight into the behaviour that can be expected from each of the measures in practice. Since each of the measures represents an alternative definition of diversity, the choice of which measure to use may make a difference. Thus, our principled approach has more general applicability beyond the analysis of the sixteen diversity measures described here. That is, when choosing any candidate interestingness measure for ranking summaries, it should be determined which of the five principles are satisfied, and then using this knowledge, judge the suitability of the candidate interestingness measure for the intended application.

The remainder of the paper is organized as follows. In Section 2, we review previously published work related to diversity measures, in general, and the use of diversity measures for ranking the interestingness of summaries, in particular. In Section 3, we review our data mining technique for generating summaries from databases. In Section 4 provide a brief overview of sixteen diversity measures empirically evaluated in previous work. In Section 5, we develop a theory of interestingness for diversity measures used to rank summaries. In Section 6, we identify those diversity measures that satisfy the proposed principles. We conclude in Section 7.

2 Related Work

Diversity measures have previously been utilized in many disciplines. Although the terminology varies depending upon the application, there are numerous research areas where the concept of diversity has been considered useful. In ecology, various measures of diversity have been proposed and studied to aid in understanding the variability of populations of organisms within different types of habitat [13]. Diversity measures have also been used by economists and social scientists to study the distribution of income between different socioeconomic groups and geographical regions [2]. In information theory, diversity measures are used to measure the information content in messages [19]. Diversity measures have been used to describe the linguistic differences between the inhabitants of neighboring geographic regions [12]. Other applications may be found in software engineering [15] and bibliometrics [16]. More general treatments attempt to define the concept of diversity and develop a related theory of diversity measurement [20].

Utilizing diversity measures to rank the interestingness of summaries generated from databases is a natural and useful extension into a new application area. We introduced this use of diversity measures in [8, 7], where well-known diversity measures from information theory, statistics, ecology, and economics were evaluated. An empirical analysis found that highly ranked, concise summaries provided a reasonable starting point for further analysis of discovered knowledge. It has also been shown that for selected sample datasets, the order in which some of the measures rank summaries is highly correlated. However, the rank ordering of summaries can vary substantially when different measures are used. Other work focusing on the analysis of related algorithms and data structures is reported in [10].

3 Background

Let a *summary* S be a relation defined on the columns $\{(A_1, D_1), (A_2, D_2), \ldots, (A_n, D_n)\}$, where each (A_i, D_i) is an attribute-domain pair. Also, let $\{(A_1, v_{i1}), (A_2, v_{i2}), \ldots, (A_n, v_{in})\}$, $i = 1, 2, \ldots, m$, be a set of m unique tuples, where each (A_j, v_{ij}) is an attribute-value pair and each v_{ij} is a value from the domain D_j associated with attribute A_j. One attribute A_k is a derived attribute, called *Count*, whose domain D_k is the set of positive integers, and whose value v_{ik} for each attribute-value pair (A_k, v_{ik}) is equal to the number of tuples which have been aggregated from the base relation (i.e., the unconditioned data present in the original database). A sample summary is shown in Table 1.

Table 1: A sample summary

Office	*Quantity*	*Amount*	*Count*
West	8	$200.00	4
East	11	$275.00	3

A summary such as the one shown in Table 1 can be generated using *attribute-oriented generalization* (AOG) [6] and *domain generalization graphs* (DGGs) [10]. For example, a sample DGG for the *Office* attribute is shown in Figure 1. In Figure 1, each node in the DGG is a partition of the domain values that can be used to describe the attribute and the arcs connecting each pair of nodes defines a generalization relation based upon AOG.

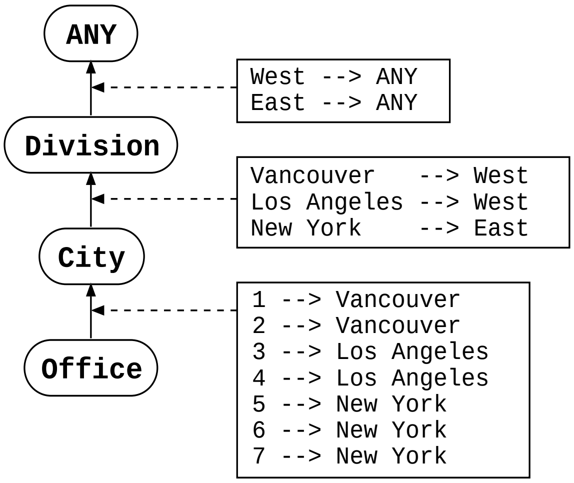

Figure 1: A DGG for the *Office* attribute

AOG summarizes the information in a database by replacing specific attribute values with more general concepts according to user-defined taxonomies. For example, the domain for the *Office* attribute is represented by the *Office* node. Increasingly more general descriptions of the domain values are represented by the *City*, *Division*, and *ANY* nodes. In Figure 1, the generalization relations consists of table lookups (other generalization relations besides table lookups are possible [17], but we restrict our discussion to table lookups for the sake of simplicity and clarity). The table associated with the arc between the *Office* and *City* nodes defines the mapping of the domain values of the *Office* node to the domain values of the *City* node.

That is, the values 1 and 2 are mapped to *Vancouver*, 3 and 4 are mapped to *Los Angeles*, and 5, 6, and 7 are mapped to *New York*. The table associated with the arc between the *City* and *Division* nodes can be described similarly. The table associated with the arc between the *Division* and *ANY* nodes maps all values in the *Division* domain to the special value *ANY*. So, the domain values for the *Office* node correspond to the most specific representation of the data, the domain values for the *City* and *Division* nodes correspond to a more general representation of the data, and the *ANY* node corresponds to the most general representation of the data.

The summary in Table 1 can be generated from the sales transaction database in Table 2, and corresponds to the *Division* node of the DGG in Figure 1. For each tuple in the sales transaction database, we simply substitute the appropriate value from the lookup table associated with the arc being traversed, and keep a count of the number of tuples that have been aggregated from the base relation and save it in the derived *Count* attribute. We also aggregate any numeric attributes that have been selected for inclusion in the summaries being generated (here we assume the *Office*, *Quantity*, and *Amount* attributes have been selected). These steps are repeated as each arc in the DGG is traversed, resulting in the generation of a new and unique summary following each traversal. The generalization space (i.e., the set of all possible summaries that can be generated) for the DGG shown in Figure 1 consists of $4 - 1 = 3$ summaries, corresponding to the *City*, *Division*, and *ANY* nodes.

Table 2: A sales transaction database

Office	*Shape*	*Size*	*Colour*	*Quantity*	*Amount*
2	round	small	white	2	\$50.00
5	square	medium	black	3	\$75.00
3	round	large	white	1	\$25.00
7	round	large	black	4	\$100.00
1	square	x-large	white	3	\$75.00
6	round	small	white	4	\$100.00
4	square	small	black	2	\$50.00

When there are multiple DGGs associated with an attribute, meaning knowledge about the attribute can be expressed in different ways, a multi-path DGG can be constructed from the single-path DGGs. For example, the DGG shown on the right side of Figure 2 is a multi-path DGG that has been constructed from the two single-path DGGs shown on the left. There are now two possible arcs that can be traversed from the *City* node, from which summaries can be generated corresponding to the *Division* and *Country* nodes.

Here we assume that common names used to describe the nodes in the single-path DGGs represents the same partition of the domain, and that the arc connecting adjacent nodes describes the same generalization relation. For example, in Figure 2, since we assume that the *ANY*, *City*, and *Office* nodes in the single-path DGGs on the left represent the same partition of the domain and that the arcs describe the same generalization relation, then the like-named nodes and arcs can be combined in the multi-path DGG. The generalization space for the multi-path DGG consists of $5 - 1 = 4$ summaries, corresponding to the *City*, *Division*, *Country*, and *ANY* nodes. In contrast, the generalization space for the two single path DGGs consists of $2 \times (4 - 1) = 6$ summaries, containing duplicates of those summaries corresponding to the *City* and *ANY* nodes.

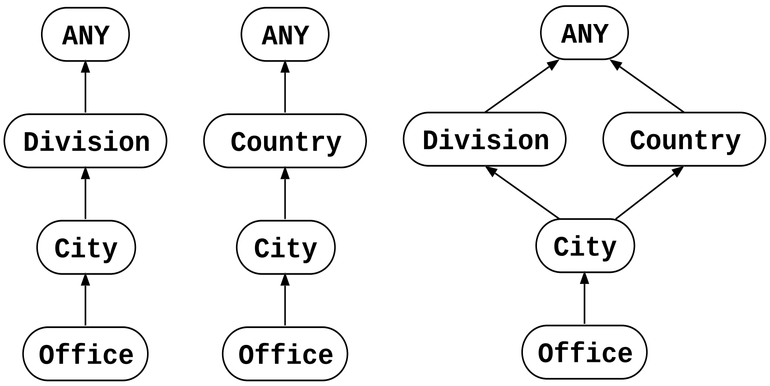

Figure 2: A multi-path DGG for the *Office* attribute

So far, we have only been concerned with the summaries generated from a database where a single attribute is associated with a DGG. Of course, we may want to generate summaries from databases where multiple attributes are associated with DGGs. In this situation, known as *multi-attribute generalization*, a set of individual attributes can be considered to be a single attribute (called a *compound attribute*) whose domain is the cross-product of the individual attribute domains. A summary generated from the cross-product domain for the compound attribute corresponds to a unique combination of nodes from the DGGs associated with the individual attributes, where one node is selected from the DGG associated with each attribute. The complexity of the DGGs is a primary factor determining the number of summaries that can be generated, and depends only on the number of nodes in the generalization space, it is not dependent upon the number of tuples in the original input relation or the number of attributes selected for generalization.

4 Objective Interestingness Measures

When the number of attributes to be generalized is large or the DGGs associated with a set of attributes are complex, many summaries may be generated. A user may then be required to evaluate each one to determine whether it contains an interesting result. Given hundreds, or even thousands of summaries, it is simply not feasible to consider trying to identify all of the interesting summaries using a manual technique. What is needed is an effective measure of interestingness to assist in the interpretation and evaluation of the discovered knowledge. The development of such measures is currently an active research area in KDD. Such measures are broadly classified as either objective or subjective. *Objective measures* are based upon the structure of discovered patterns, such as the frequency with which combinations of items appear in sales transactions [1]. *Subjective measures* are based upon user beliefs or biases regarding relationships in the data, such as an approach utilizing Bayes Theorem to revise prior beliefs [14], or an approach utilizing templates to describe interesting patterns [11].

The tuples in a summary generated from a database are unique, and therefore, can be considered to be a population with a structure that can be described by some frequency or probability distribution. Here, we review sixteen diversity measures described in [7], and shown in Figure 3, that evaluate the frequency or probability distribution of the values in the derived *Count* attribute to assign a single real-valued index that represents its interestingness relative to other summaries generated from the same database. The measures are well-known measures of dispersion, dominance, inequality, and concentration that have previously been successfully applied in several areas of the social, ecological, information, and computer sciences.

The variables used to describe the diversity measures are defined, as follows. Let m be the total number of tuples in a summary. Let n_i be the value contained in the derived *Count* attribute for tuple t_i. Let $N = \sum_{i=1}^{m} n_i$ be the total count. Let p be the actual probability distribution of the tuples based upon the values n_i. Let $p_i = n_i/N$ be the actual probability for tuple t_i. Let q be a uniform probability distribution of the tuples. Let $\bar{u} = N/m$ be the count for tuple t_i, $i = 1, 2, \ldots, m$ according to the uniform distribution q. Let $\bar{q} = 1/m$ be the probability for tuple t_i, for all $i = 1, 2, \ldots, m$ according to the uniform distribution q. Let r be the probability distribution obtained by combining the values n_i and $\bar{u}$. Let $r_i = (n_i + \bar{u})/2N$, be the probability for tuples t_i, for all $i = 1, 2, \ldots, m$ according to the distribution r.

$$I_{Variance} = \frac{\sum_{i=1}^{m} (p_i - \bar{q})^2}{m-1}$$

$$I_{Simpson} = \sum_{i=1}^{m} p_i^2$$

$$I_{Shannon} = -\sum_{i=1}^{m} p_i \log_2 p_i$$

$$I_{Total} = m * I_{Shannon}$$

$$I_{Max} = \log_2 m$$

$$I_{McIntosh} = \frac{N - \sqrt{\sum_{i=1}^{m} n_i^2}}{N - \sqrt{N}}$$

$$I_{Lorenz} = \bar{q} \sum_{i=1}^{m} (m - i + 1) p_i$$

$$I_{Gini} = 0.5 \left(\sum_{i=1}^{m} \sum_{j=1}^{m} |p_i \bar{q} - p_j \bar{q}| \right)$$

$$I_{Berger} = \max(p_i)$$

$$I_{Schutz} = \frac{\sum_{i=1}^{m} |p_i - \bar{q}|}{2m\bar{q}}$$

$$I_{Bray} = \frac{\sum_{i=1}^{m} \min(n_i, \bar{u})}{N}$$

$$I_{Whittaker} = 1 - \left(0.5 \sum_{i=1}^{m} |p_i - \bar{q}| \right)$$

$$I_{Kullback} = \log_2 m - \left(\sum_{i=1}^{m} p_i \log_2 \frac{p_i}{\bar{q}} \right)$$

$$I_{MacArthur} = \left(-\sum_{i=1}^{m} r_i \log_2 r_i \right) - \left(0.5 \left(-\sum_{i=1}^{m} p_i \log_2 p_i \right) + \log_2 m \right)$$

$$I_{Theil} = \frac{\sum_{i=1}^{m} |p_i \log_2 p_i - \bar{q} \log_2 \bar{q}|}{m\bar{q}}$$

$$I_{Atkinson} = 1 - \left(\prod_{i=1}^{m} \frac{p_i}{\bar{q}} \right)^{\bar{q}}$$

Figure 3: Sixteen-diversity-measures

5 Principles of Interestingness

We now develop a theory of interestingness through the mathematical formulation of five principles that must be satisfied by any acceptable diversity measure for ranking the interestingness of summaries generated from databases. We study functions f of m variables, $f(n_1, \ldots, n_m)$, where f denotes a general measure of diversity, m and each n_i (n_i assumed to be non-zero) are as defined in the previous section, and $(n_1, \ldots, n_m)$ is a vector corresponding to the values in the derived *Count* attribute for some arbitrary summary. We begin by specifying two fundamental principles.

Minimum Value Principle (P1). Given a vector $(n_1, \ldots, n_m)$, where $n_i = n_j$, $i \neq j$, for all i, j, $f(n_1, \ldots, n_m)$ attains its minimum value.

P1 specifies that the minimum interestingness should be attained when the tuple counts are all equal (i.e., uniformly distributed).

Maximum Value Principle (P2). Given a vector $(n_1, \ldots, n_m)$, where $n_1 = N - m + 1$, $n_i = 1$, $i = 2, \ldots, m$, and $N > m$, $f(n_1, \ldots, n_m)$ attains its maximum value.

P2 specifies that the maximum interestingness should be attained when the tuple counts are distributed as unevenly as possible.

The behaviour of a measure relative to satisfying both *P1* and *P2* is significant because it reveals an important characteristic about its fundamental nature as a measure of diversity. A measure of diversity can generally be considered either a *measure of concentration* or a *measure of dispersion*. A measure of concentration can be viewed as the opposite of a measure of dispersion, and we can convert one to the other via simple transformations. For example, if g corresponds to a measure of dispersion, then we can convert it to a measure of concentration f, where $f = \max(g) - g$. Here we only consider measures of concentration (i.e., those satisfying P1 and P2), so measures of dispersion (i.e., those not satisfying P1 and P2 in a particular way, and to be discussed in the next section) were transformed into measures of concentration prior to our analysis.

Skewness Principle (P3). Given a vector $(n_1, \ldots, n_m)$, where $n_1 = N - m + 1$, $n_i = 1$, $i = 2, \ldots, m$, and $N > m$, and a vector $(n_1 - c, n_2, \ldots, n_m, n_{m+1}, \ldots, n_{m+c})$, where $n_1 - c > 1$ and $n_i = 1$, $i = 2, \ldots, m + c$, $f(n_1, \ldots, n_m) > f(n_1 - c, n_2, \ldots, n_m, n_{m+1}, \ldots, n_{m+c})$.

P3 specifies that a summary containing m tuples, whose counts are distributed as unevenly as possible, will be more interesting than a summary containing $m + c$ tuples, whose counts are also distributed as unevenly as possible.

Permutation Invariance Principle (P4). Given a vector $(n_1, \ldots, n_m)$ and any permutation $(i_1, \ldots, i_m)$ of $(1, \ldots, m)$, $f(n_i, \ldots, n_m) = f(n_{i_1}, \ldots, n_{i_m})$.

P4 specifies that every permutation of a given distribution of tuple counts should be equally interesting. That is, interestingness is not a labeled property, it is only determined by the distribution of the counts.

Transfer Principle (P5). Given a vector $(n_1, \ldots, n_m)$ and $0 < c < n_j$, $f(n_1, \ldots, n_i + c, \ldots, n_j - c, \ldots, n_m) > f(n_1, \ldots, n_i, \ldots, n_j, \ldots, n_m)$.

P5, adapted from [4], specifies that when a strictly positive transfer is made from the count of one tuple to another tuple whose count is greater, then interestingness increases.

6 Theorems and Proofs

In this section, we address the problem of determining those measures that satisfy the proposed principles. Summarized results are shown in Table 3. In Table 3, the *P1* to *P5* columns describe those measures that satisfy the proposed principles according to mathematical proof. The *Concentration Order* and *Dispersion Order* columns describe those measures that satisfy the proposed principles when all of the measures are considered to be measures of concentration and measures of dispersion, respectively (to be discussed in detail below). The *Aggregate Order* columns are an aggregation of the *Concentration Order* and *Dispersion Order* columns, and describe those measures of concentration and transformed measures of dispersion that satisfy the proposed principles. A measure that satisfies a principle is indicated by the *bullet* symbol (i.e., $\bullet$). A measure that satisfies a principle in both the *Concentration Order* and *Dispersion Order* columns is indicated by the *plus* symbol (i.e., $+$) in the *Aggregate Order* columns. A measure that fails to satisfy a principle in either of the *Concentration Order* or *Dispersion Order* columns is indicated by the *times* symbol (i.e., $\times$) in the *Aggregate Order* columns.

Table 3: Measures satisfying the proposed principles

Measure	*Concentration Order*					*Dispersion Order*					*Aggregate Order*				
	P1	*P2*	*P3*	*P4*	*P5*	*P1*	*P2*	*P3*	*P4*	*P5*	*P1*	*P2*	*P3*	*P4*	*P5*
$I_{Variance}$	•	•	•	•	•				•		•	•	•	+	•
$I_{Simpson}$	•	•	•	•	•				•		•	•	•	+	•
$I_{Shannon}$				•		•	•	•	•	•	•	•	•	+	•
I_{Total}				•		•	•	•	•	•	•	•	•	+	•
I_{Max}				•				•	•		×	×	•	+	×
$I_{McIntosh}$				•		•	•	•	•	•	•	•	•	+	•
I_{Lorenz}			•			•	•			•	•	•		×	•
I_{Gini}	•	•		•	•			•	•		•	•		+	•
I_{Berger}	•	•	•	•					•		•	•	•	+	×
I_{Schutz}	•	•		•				•	•		•	•		+	×
I_{Bray}			•	•		•	•		•		•	•		+	×
$I_{Whittaker}$			•	•		•	•		•		•	•		+	×
$I_{Kullback}$				•		•	•	•	•	•	•	•	•	+	•
$I_{MacArthur}$	•	•		•	•			•	•		•	•		+	•
I_{Theil}	•			•				•	•		•	×		+	×
$I_{Atkinson}$	•	•		•	•			•	•		•	•		+	•

In the *Concentration Order* columns, higher index values generated by a measure are considered to be more interesting. This corresponds to an ordering of index values where the lowest value is associated with a uniform distribution of the counts (according to P1), and the highest value is associated with the most uneven distribution of the counts (according to P2). That we consider higher index values to be more interesting is just an arbitrary starting point for our analysis and will be shown to be an assumption with no significant consequences. In the *Concentration Order* columns, we see that $I_{Variance}$ satisfies both *P1* and *P2*. We also see that $I_{Shannon}$ does not satisfy *P1* and *P2*. Similarly, in the *Dispersion Order* columns, lower index values generated by a measures are considered to be more interesting. This corresponds to an ordering of index values where the highest value is associated with a uniform distribution of the counts (contrary to P1), and the lowest value is associated with the most uneven distribution of the counts (contrary to P2). Now in the *Dispersion Order* columns, we see that $I_{Shannon}$ satisfies both *P1* and *P2*. We also see that $I_{Variance}$ does not satisfy *P1* and *P2*. So, it appears that the order in which summaries are ranked by $I_{Shannon}$ is opposite to the order in which they are ranked by $I_{Variance}$ (at least for the summaries represented by a uniform and the most uneven distribution of the counts). Other measures are seen to behave in a similar manner. Within the context of our analysis, a measure that satisfies both *P1* and *P2*, such as $I_{Variance}$ does in the *Concentration Order* columns, is considered to be a measure of concentration, while a measure that satisfies both *P1* and *P2*, such as $I_{Shannon}$ does in the *Dispersion Order* columns, is considered to be a measure of dispersion.

We could have done our analysis by considering only measures of dispersion. In that case, we could have assumed that lower index values in the current *Concentration Order* columns were to be considered more interesting. Those measures satisfying the proposed principles would then have been identical to those shown in the current *Dispersion Order* columns. Similarly, in the *Dispersion Order* columns, we could have assumed that higher index values were to be considered more interesting. Those measures satisfying the proposed principles would then have been identical to those shown in the current *Concentration Order* columns. We then would have used a transformation function h, similar to f, to convert selected measures of concentration into measures of dispersion. The end result is that those measures satisfying the proposed principles in the *Aggregate Order* columns would have been identical to those currently shown. So the initial assumption that higher values are to be considered more interesting is of no consequence in our analysis. That is, in the proofs that follow, where we show $f(x) > f(y)$ for measures of concentration, we show equivalently that $f(x) < f(y)$ for measures of dispersion.

We now state theorems and proofs for each measure satisfying a principle in the *Adapted Order* columns of Table 3. We also state theorems for those measures that fail to satisfy particular principles, but leave the proofs as an exercise for the reader (vectors are suggested that will clearly demonstrate the failure). At the beginning of each section, we restate the relevant principle for reader convenience.

6.1 Minimum Value Principle

Minimum Value Principle (P1). Given a vector $(n_1, \ldots, n_m)$, where $n_i = n_j$, $i \neq j$, for all i, j, $f(n_1, \ldots, n_m)$ attains its minimum value.

Theorem 6.1. $I_{Variance}$ satisfies P1.

Proof. We need to show that

$$\sum_{i=1}^{m} \left(\frac{n_i}{N} - \frac{1}{m} \right)^2 = 0.$$

The values in the vector $(n_1, \ldots, n_m)$ are uniformly distributed, so $n_1 = \ldots = n_m$, $n_1/N = \ldots = n_m/N$, and all $p_i = n_i/N$ are equal, for all i. Now

$$\bar{q} = \frac{\sum_{i=1}^{m} \left(\frac{n_i}{N}\right)}{m} = \frac{m\left(\frac{n_i}{N}\right)}{m} = \frac{n_i}{N},$$

and

$$\bar{q} = \frac{\sum_{i=1}^{m} \left(\frac{n_i}{N}\right)}{m} = \frac{1}{m},$$

so it is shown that $n_i/N = 1/m$, for all i. Now,

$$\sum_{i=1}^{m} \left(\frac{n_i}{N} - \frac{1}{m} \right)^2 = \sum_{i=1}^{m} \left(\frac{1}{m} - \frac{1}{m} \right)^2 = \sum_{i=1}^{m} 0 = 0,$$

and it is proved.

Theorem 6.2. $I_{Simpson}$ satisfies P1.

Proof. We need to show that when the values in the vector $(n_1, \ldots, n_m)$ are uniformly distributed, $f(n_1, \ldots, n_m) = 1/m$. That is, we need to show that

$$\sum_{i=1}^{m} \left(\frac{n_i}{N}\right)^2 = \frac{1}{m}.$$

We also need to show that for any vector $(n_1 + c, n_2 - d_2, \ldots n_m - d_m)$ (i.e., a vector whose values are not uniformly distributed), where at least one $d_i \neq 0$, $d_i < n_i$, and $\sum_{i=2}^{m} d_i = c$, $f(n_1+c, n_2-d_2, \ldots, n_m-d_m) > 1/m$. That is, we need to show that

$$\left(\frac{n_1+c}{N}\right)^2 + \sum_{i=2}^{m}\left(\frac{n_i-d_i}{N}\right)^2 > \frac{1}{m}.$$

Now, we have $n_i/N = 1/m$, so

$$\sum_{i=1}^{m} p_i^2 = \sum_{i=1}^{m}\left(\frac{n_i}{N}\right)^2 = \sum_{i=1}^{m}\left(\frac{1}{m}\right)^2 = m\left(\frac{1}{m}\right)^2 = m\left(\frac{1}{m^2}\right) = \frac{1}{m},$$

and the first part is shown. For a vector $(n_1 + c, n_2 - d_2, \ldots, n_m - d_m)$, whose values are not uniformly distributed, showing that

$$\left(\frac{n_1+c}{N}\right)^2 + \sum_{i=2}^{m}\left(\frac{n_i-d_i}{N}\right)^2 > \frac{1}{m}$$

is equivalent to showing that

$$\left(\frac{n_1+c}{N}\right)^2 + \sum_{i=2}^{m}\left(\frac{n_i-d_i}{N}\right)^2 - \frac{1}{m} > 0.$$

Since the vector $(n_1 + c, n_2 - d_2, \ldots n_m - d_m)$ is not uniformly distributed, n_1/N is greater than the average $1/m$, so let $1/m = \frac{n_1}{N} - k$. Substituting for $1/m$ in the above inequality yields

$$\left(\frac{n_1+c}{N}\right)^2 + \sum_{i=2}^{m}\left(\frac{n_i-d_i}{N}\right)^2 - \frac{n_1}{N} + k > 0.$$

Expanding the left side of the above inequality, and putting it over the common denominator N^2 yields

$$n_1^2 + 2n_1c + c^2 + \sum_{i=2}^{m}(n_i - d_i)^2 - n_1N + kN^2 > 0.$$

Since $N = n_1 + c + \sum_{i=2}^{m}(n_i - d_i)$, substituting for N in the above inequality yields

$$2n_1 \sum_{i=2}^{m} d_i - 2\sum_{i=2}^{m} n_i d_i + \left(\sum_{i=2}^{m} d_i\right)^2 + \sum_{i=2}^{m} n_i^2 + \sum_{i=2}^{m} d_i^2 + kn_1^2 + 2n_1 k \sum_{i=2}^{m} n_i - n_1 \sum_{i=2}^{m} n_i$$
$$+k\left(\sum_{i=2}^{m} n_i\right)^2 + k\left(\sum_{i=2}^{m} d_i\right)^2 > 0.$$

Now two of the positive terms in the above inequality are greater than the two negative terms, specifically,

$$2n_1 \sum_{i=2}^{m} d_i > 2\sum_{i=2}^{m} n_i d_i$$

and

$$2n_1 k \sum_{i=2}^{m} n_i > n_1 \sum_{i=2}^{m} n_i.$$

The five remaining terms are all positive, so their sum is positive, the second part is shown, and it is proved.

Theorem 6.3. $I_{Shannon}$ satisfies P1.

Proof. We need to show that

$$-\sum_{i=1}^{m} \left(\frac{n_i}{N}\right) \log_2 \left(\frac{n_i}{N}\right) > -\left(\left(\frac{n_1+c}{N}\right) \log_2 \left(\frac{n_1+c}{N}\right) + \sum_{i=2}^{m} \left(\frac{n_i-d_i}{N}\right) \log_2 \left(\frac{n_i-d_i}{N}\right)\right).$$

Multiplying both sides by -1 yields

$$\sum_{i=1}^{m} \left(\frac{n_i}{N}\right) \log_2 \left(\frac{n_i}{N}\right) < \left(\frac{n_1+c}{N}\right) \log_2 \left(\frac{n_1}{N}\right) + \sum_{i=2}^{m} \left(\frac{n_i-d_i}{N}\right) \log_2 \left(\frac{n_i-d_i}{N}\right).$$

Now the values in the vector $(n_1, \ldots, n_m)$ are uniformly distributed, so $n_1 = \ldots = n_m$. Substituting n_1 for n_i yields

$$\sum_{i=1}^{m} \left(\frac{n_1}{N}\right) \log_2 \left(\frac{n_1}{N}\right) < \left(\frac{n_1+c}{N}\right) \log_2 \left(\frac{n_1}{N}\right) + \sum_{i=2}^{m} \left(\frac{n_1-d_i}{N}\right) \log_2 \left(\frac{n_1-d_i}{N}\right).$$

Factoring $1/N$ out of both sides of the above inequality yields

$$\sum_{i=1}^{m} n_1 \log_2 \left(\frac{n_1}{N}\right) < (n_1+c) \log_2 \left(\frac{n_1}{N}\right) + \sum_{i=2}^{m} (n_1-d_i) \log_2 \left(\frac{n_1-d_i}{N}\right).$$

Using the difference property of logarithms to decompose the above inequality yields

$$\sum_{i=1}^{m} n_1 \log_2 n_1 - \sum_{i=1}^{m} n_1 \log_2 N <$$

$$n_1 \log_2(n_1 + c) + c \log_2(n_1 + c) - n_1 \log_2 N - c \log_2 N + \sum_{i=2}^{m} n_1 \log_2(n_1 - d_i)$$

$$- \sum_{i=2}^{m} d_i \log_2(n_1 - d_i) - \sum_{i=2}^{m} n_1 \log_2 N + \sum_{i=2}^{m} d_i \log_2 N.$$

Canceling terms yields

$$\sum_{i=1}^{m} n_1 \log_2 n_1 < (n_1 + c) \log_2(n_1 + c) + \sum_{i=2}^{m} (n_1 - d_i) \log_2(n_1 - d_i).$$

The right side of the above inequality can be written as

$$(n_1 + c) \log_2(n_1 + c) + (n_1 - d_2) \log_2(n_1 - d_2) + \ldots + (n_1 - d_m) \log_2(n_1 - d_m).$$

Since the function $x \log_2 x$ is a convex function, we have

$$(n_1 + c) \log_2(n_1 + c) + (n_1 - d_2) \log_2(n_1 - d_2) + \ldots + (n_1 - d_m) \log_2(n_1 - d_m) \geq$$

$$(n_1 + c + n_1 - d_2 + \ldots + n_1 - d_m) \log_2(n_1 + c + n_1 - d_2 + \ldots + n_1 - d_m).$$

Now $c = \sum_{i=2}^{m} d_i$, so substituting for the d_i's on the right side of the above inequality yields

$$(n_1 + c + n_1 - d_2 + \ldots + n_1 - d_m) \log_2(n_1 + c + n_1 - d_2 + \ldots + n_1 - d_m) =$$

$$(mn_1) \log_2(mn_1),$$

or equivalently

$$(n_1 + c + n_1 - d_2 + \ldots + n_1 - d_m) \log_2(n_1 + c + n_1 - d_2 + \ldots + n_1 - d_m) =$$

$$\sum_{i=1}^{m} n_1 \log_2(mn_1).$$

We now show that

$$\sum_{i=1}^{m} n_1 \log_2 n_1 < \sum_{i=1}^{m} n_1 \log_2(mn_1).$$

Using the addition property of logarithms to decompose the above inequality yields

$$\sum_{i=1}^{m} n_1 \log_2 n_1 < \sum_{i=1}^{m} n_1 \log_2 m + \sum_{i=1}^{m} n_1 \log_2 n_1.$$

Canceling terms yields

$$0 < \sum_{i=1}^{m} n_1 \log_2 m,$$

and it is proved.

Theorem 6.4. I_{Total} satisfies P1.

Proof. We need to show that

$$m\left(-\sum_{i=1}^{m}\left(\frac{n_i}{N}\right)\log_2\left(\frac{n_i}{N}\right)\right) >$$
$$m\left(-\left(\left(\frac{n_1+c}{N}\right)\log_2\left(\frac{n_1+c}{N}\right)+\sum_{i=2}^{m}\left(\frac{n_i-d_i}{N}\right)\log_2\left(\frac{n_i-d_i}{N}\right)\right)\right).$$

Canceling terms and multiplying both sides of the above inequality by -1 yields

$$\sum_{i=1}^{m}\left(\frac{n_i}{N}\right)\log_2\left(\frac{n_i}{N}\right) < \left(\frac{n_1+c}{N}\right)\log_2\left(\frac{n_1}{N}\right)+\sum_{i=2}^{m}\left(\frac{n_i-d_i}{N}\right)\log_2\left(\frac{n_i-d_i}{N}\right),$$

which is proved in Theorem 6.3.

Theorem 6.5. $I_{McIntosh}$ satisfies P1.

Proof. We need to show that

$$\frac{N-\sqrt{\sum_{i=1}^{m} n_i^2}}{N-\sqrt{N}} > \frac{N-\sqrt{(n_1+c)^2+\sum_{i=2}^{m}(n_i-d_i)^2}}{N-\sqrt{N}}.$$

Canceling terms and multiplying both sides of the above inequality by -1 yields

$$\sum_{i=1}^{m} n_i^2 < (n_1 + c)^2 + \sum_{i=2}^{m} (n_i - d_i)^2.$$

The values in the vector $(n_1, \ldots, n_m)$ are uniformly distributed, so $n_1 = \ldots = n_m$. Substituting n_1 for n_i in the above inequality and expanding the terms on the right side yields

$$mn_1^2 < n_1^2 + 2n_1c + c^2 + \sum_{i=2}^{m} (n_1^2 - 2n_1d_i + d_i^2).$$

Distributing the summation term on the right side of the above inequality yields

$$mn_1^2 < n_1^2 + 2n_1c + c^2 + \sum_{i=2}^{m} n_1^2 - \sum_{i=2}^{m} 2n_1d_i + \sum_{i=2}^{m} d_i^2.$$

Since $c = \sum_{i=2}^{m} d_i$, substituting for c in the above inequality yields

$$mn_1^2 < n_1^2 + \sum_{i=2}^{m} 2n_1d_i + \left(\sum_{i=2}^{m} d_i\right)^2 + \sum_{i=2}^{m} n_1^2 - \sum_{i=2}^{m} 2n_1d_i + \sum_{i=2}^{m} d_i^2.$$

Canceling terms yields

$$0 < \left(\sum_{i=2}^{m} d_i\right)^2 + \sum_{i=2}^{m} d_i^2,$$

and it is proved.

Theorem 6.6. I_{Lorenz} satisfies P1.

Proof. We need to show that

$$\frac{1}{m}\sum_{i=1}^{m}(m-i+1)\left(\frac{n_i}{N}\right) > \frac{1}{m}\left(\sum_{i=1}^{m-1}(m-i+1)\frac{n_i - d_i}{N} + \frac{n_m + c}{N}\right).$$

The values in the vector $(n_1, \ldots, n_m)$ are uniformly distributed, so $n_1 = \ldots = n_m$. Substituting n_m for n_i

and distributing the summation on the right side yields

$$\frac{1}{m}\sum_{i=1}^{m-1}(m-i+1)\left(\frac{n_m}{N}\right)+\frac{n_m}{N}>$$
$$\frac{1}{m}\left(\sum_{i=1}^{m-1}(m-i+1)\frac{n_m}{N}-\sum_{i=1}^{m-1}(m-i+1)\frac{d_i}{N}+\frac{n_m}{N}+\frac{c}{N}\right).$$

Canceling terms yields

$$0>-\sum_{i=1}^{m-1}(m-i+1)\frac{d_i}{N}+\frac{c}{N}.$$

Since $c=\sum_{i=1}^{m-1}d_i$, substituting for c in the above inequality and moving $-\sum_{i=1}^{m-1}(m-i+1)\frac{d_i}{N}$ to the left side yields

$$\sum_{i=1}^{m-1}(m-i+1)\frac{d_i}{N}>\sum_{i=1}^{m-1}\frac{d_i}{N}.$$

Factoring $1/N$ out of both sides of the above inequality yields

$$\sum_{i=1}^{m-1}(m-i+1)d_i>\sum_{i=1}^{m-1}d_i,$$

which is obviously true, and it is proved.

Theorem 6.7. I_{Gini} satisfies P1.

Proof. We need to show that

$$\frac{\sum_{i=1}^{m}\sum_{j=1}^{m}\left|\left(\frac{n_i}{N}\right)\left(\frac{1}{m}\right)-\left(\frac{n_j}{N}\right)\left(\frac{1}{m}\right)\right|}{2}=0.$$

The values in the vector $(n_1,\ldots,n_m)$ are uniformly distributed, so $n_i/N=1/m$,

$$\begin{aligned}\frac{\sum_{i=1}^{m}\sum_{j=1}^{m}\left|\left(\frac{n_i}{N}\right)\left(\frac{1}{m}\right)-\left(\frac{n_j}{N}\right)\left(\frac{1}{m}\right)\right|}{2} &= \frac{\sum_{i=1}^{m}\sum_{j=1}^{m}\left|\left(\frac{1}{m}\right)\left(\frac{1}{m}\right)-\left(\frac{1}{m}\right)\left(\frac{1}{m}\right)\right|}{2}\\ &= \frac{\sum_{i=1}^{m}\sum_{j=1}^{m}|0|}{2}=0,\end{aligned}$$

and it is proved.

Theorem 6.8. I_{Berger} satisfies P1.

Proof. We need to show that when the values in the vector $(n_1, \ldots, n_m)$ are uniformly distributed, $f(n_1, \ldots, n_m) = n_1/N$. That is, we need to show that

$$\max\left(\frac{n_i}{N}\right) = \frac{n_1}{N}.$$

We also need to show that for any vector $(n_1 + c, n_2 - d_2, \ldots n_m - d_m)$ (i.e., a vector whose values are not uniformly distributed), where at least one $d_i \neq 0$, $d_i < n_i$, and $\sum_{i=2}^{m} d_i = c$, $f(n_1+c, n_2-d_2, \ldots, n_m-d_m) > n_1/N$. That is, we need to show that

$$\max\left(\frac{n_i}{N}\right) = \frac{n_1 + c}{N} > \frac{n_1}{N}.$$

The values in the vector $(n_1, \ldots, n_m)$ are uniformly distributed, so $n_1 = \ldots = n_m$, $n_1/N = \ldots = n_m/N$, all $p_i = n_i/N$ are equal, for all i, $\max(p_i) = p_1 = n_1/N$, and the first part is shown. We now show that

$$\frac{n_1 + c}{N} > \frac{n_1}{N}.$$

Factoring $1/N$ out of both sides of the above inequality yields

$$n_1 + c > n_1.$$

Canceling n_1 from both sides yields

$$c > 0.$$

Since $c = \sum_{i=2}^{m}$, and at least one $d_i > 0$, the above inequality is obviously true, the second part is shown, and it is proved.

Theorem 6.9. I_{Schutz} satisfies P1.

Proof. We need to show that

$$\frac{\sum_{i=1}^{m} \left|\frac{n_i}{N} - \frac{1}{m}\right|}{2m\left(\frac{1}{m}\right)} = 0.$$

The values in the vector $(n_1, \ldots, n_m)$ are uniformly distributed, so $n_i/N = 1/m$,

$$\frac{\sum_{i=1}^{m} \left|\frac{n_i}{N} - \frac{1}{m}\right|}{2m\left(\frac{1}{m}\right)} = \frac{\sum_{i=1}^{m} \left|\frac{1}{m} - \frac{1}{m}\right|}{2} = \frac{\sum_{i=1}^{m} |0|}{2} = \frac{0}{2} = 0,$$

and it is proved.

Theorem 6.10. I_{Bray} satisfies P1.

Proof. We need to show that

$$\frac{\sum_{i=1}^{m} min(n_i, \bar{u})}{N} = 1,$$

which is equivalent to showing that

$$\sum_{i=1}^{m} min(n_i, \bar{u}) = N.$$

The values in the vector $(n_1, \ldots, n_m)$ are uniformly distributed, so each $n_i = \bar{u}$, for all i. Substituting for $\bar{u}$ in the above equality yields

$$\sum_{i=1}^{m} min(\bar{u}, \bar{u}) = N,$$

which is equivalent to

$$m\bar{u} = N.$$

Now $N = m\bar{u}$, so substituting for N in the above equality yields

$$N = N,$$

and it is proved.

Theorem 6.11. $I_{Whittaker}$ satisfies P1.

Proof. We need to show that

$$1 - \left(0.5 \sum_{i=1}^{m} \left|\frac{n_i}{N} - \frac{1}{m}\right|\right) > 1 - \left(0.5 \left(\left|\frac{n_1 + c}{N} - \frac{1}{m}\right| + \sum_{i=2}^{m} \left|\frac{n_i - d_i}{N} - \frac{1}{m}\right|\right)\right).$$

Canceling terms and multiplying both sides of the above inequality by -1 yields

$$\sum_{i=1}^{m} \left| \frac{n_i}{N} - \frac{1}{m} \right| < \left| \frac{n_1 + c}{N} - \frac{1}{m} \right| + \sum_{i=2}^{m} \left| \frac{n_i - d_i}{N} - \frac{1}{m} \right| .$$

The values in the vector $(n_1, \ldots, n_m)$ are uniformly distributed, so $n_i/N = 1/m$. Substituting for n_i/N in the above inequality yields

$$0 < \left| \frac{n_1 + c}{N} - \frac{1}{m} \right| + \sum_{i=2}^{m} \left| \frac{n_i - d_i}{N} - \frac{1}{m} \right| .$$

Since all the terms on the right side of the above inequality are positive. their sum is positive, and it is proved.

Theorem 6.12. $I_{Kullback}$ satisfies P1.

Proof. We need to show that

$$log_2 m - \sum_{i=1}^{m} \left(\frac{n_i}{N} \right) \log_2 \left(\frac{m n_i}{N} \right) >$$
$$log_2 m - \left(\left(\frac{n_1 + c}{N} \right) \log_2 \left(\frac{m(n_1 + c)}{N} \right) + \sum_{i=2}^{m} \left(\frac{n_i - d_i}{N} \right) \log_2 \left(\frac{m(n_i - d_i)}{N} \right) \right) .$$

Canceling terms and multiplying both sides of the above inequality by -1 yields

$$\sum_{i=1}^{m} \left(\frac{n_i}{N} \right) \log_2 \left(\frac{m n_i}{N} \right) <$$
$$\left(\frac{n_1 + c}{N} \right) \log_2 \left(\frac{m(n_1 + c)}{N} \right) + \sum_{i=2}^{m} \left(\frac{n_i - d_i}{N} \right) \log_2 \left(\frac{m(n_i - d_i)}{N} \right) .$$

The values in the vector $(n_1, \ldots, n_m)$ are uniformly distributed, so $n_1 = \ldots = n_m$. Substituting n_1 for n_i yields

$$\sum_{i=1}^{m} \left(\frac{n_1}{N} \right) \log_2 \left(\frac{m n_1}{N} \right) <$$
$$\left(\frac{n_1 + c}{N} \right) \log_2 \left(\frac{m(n_1 + c)}{N} \right) + \sum_{i=2}^{m} \left(\frac{n_1 - d_i}{N} \right) \log_2 \left(\frac{m(n_1 - d_i)}{N} \right) .$$

Factoring $1/N$ out of both sides of the above inequality yields

$$\sum_{i=1}^{m} n_1 \log_2 \left(\frac{mn_1}{N}\right) < (n_1 + c)\log_2 \left(\frac{m(n_1+c)}{N}\right) + \sum_{i=2}^{m}(n_1 - d_i)\log_2 \left(\frac{m(n_1-d_i)}{N}\right).$$

Using the difference property of logarithms to decompose the above inequality yields

$$\begin{aligned}
&\sum_{i=1}^{m} n_1 \log_2(mn_1) - \sum_{i=1}^{m} n_1 \log_2 N < \\
&\quad (n_1+c)\log_2(m(n_1+c)) - n_1 \log_2 N - c\log_2 N + \sum_{i=2}^{m}(n_1-d_i)\log_2(m(n_1-d_i)) \\
&\quad -\sum_{i=2}^{m} n_1 \log_2 N + \sum_{i=2}^{m} d_i \log_2 N.
\end{aligned}$$

Canceling terms and using the addition property of logarithms to further decompose the above inequality yields

$$\begin{aligned}
&\sum_{i=1}^{m} n_1 \log_2 m + \sum_{i=1}^{m} n_1 \log_2 n_1 < \\
&\quad n_1 \log_2 m + c\log_2 m + (n_1+c)\log_2(n_1+c) + \sum_{i=2}^{m} n_1 \log_2 m - \sum_{i=2}^{m} d_i \log_2 m \\
&\quad +\sum_{i=2}^{m}(n_1-d_i)\log_2(n_1-d_i).
\end{aligned}$$

Canceling terms yields

$$\sum_{i=1}^{m} n_1 \log_2 n_1 < (n_1+c)\log_2(n_1+c) + \sum_{i=2}^{m}(n_1-d_i)\log_2(n_1-d_i),$$

which is proved in Theorem 6.3.

Theorem 6.13. $I_{MacArthur}$ satisfies P1.

Proof. We need to show that

$$-\sum_{i=1}^{m}\left(\frac{n_i+\bar{u}}{2N}\right)\log_2\left(\frac{n_i+\bar{u}}{2N}\right) - \frac{\left(-\sum_{i=1}^{m}\left(\frac{n_i}{N}\right)\log_2\left(\frac{n_i}{N}\right)\right)+\log_2 m}{2} = 0.$$

The values in the vector $(n_1, \ldots, n_m)$ are uniformly distributed, so $n_i = \bar{u}$. Substituting for $\bar{u}$ in the above equality yields

$$-\sum_{i=1}^{m} \left(\frac{2n_i}{2N}\right) \log_2 \left(\frac{2n_i}{2N}\right) - \frac{\left(-\sum_{i=1}^{m} \left(\frac{n_i}{N}\right) \log_2 \left(\frac{n_i}{N}\right)\right) + \log_2 m}{2} = 0.$$

Also, the values in the vector $(n_1, \ldots, n_m)$ are uniformly distributed, so $n_i/N = 1/m$. Substituting for n_i/N in the above equality yields

$$-\sum_{i=1}^{m} \left(\frac{1}{m}\right) \log_2 \left(\frac{1}{m}\right) - \frac{\left(-\sum_{i=1}^{m} \left(\frac{1}{m}\right) \log_2 \left(\frac{1}{m}\right)\right) + \log_2 m}{2} = 0.$$

Now $\sum_{i=1}^{m} 1/m = 1$, so substituting for $\sum_{i=1}^{m} 1/m$ in the above equality yields

$$-\log_2 \left(\frac{1}{m}\right) - \frac{\left(-\log_2 \left(\frac{1}{m}\right)\right) + \log_2 m}{2} = 0.$$

Using the difference property of logarithms, we have $\log_2 1/m = \log_2 1 - \log_2 m$, and $\log_2 1 = 0$, so decomposing the above equality and canceling terms yields

$$\log_2 m - \frac{\log_2 m + \log_2 m}{2} = \log_2 m - \frac{2 \log_2 m}{2} = \log_2 m - \log_2 m = 0,$$

and it is proved.

Theorem 6.14. I_{Theil} satisfies P1.

Proof. We need to show that

$$\sum_{i=1}^{m} \left| \left(\frac{n_i}{N}\right) \log_2 \left(\frac{n_i}{N}\right) - \left(\frac{1}{m}\right) \log_2 \left(\frac{1}{m}\right) \right| = 0.$$

The values in the vector $(n_1, \ldots, n_m)$ are uniformly distributed, so $n_i/N = 1/m$. Substituting for n_i/N in the above equality yields

$$\sum_{i=1}^{m} \left| \left(\frac{1}{m}\right) \log_2 \left(\frac{1}{m}\right) - \left(\frac{1}{m}\right) \log_2 \left(\frac{1}{m}\right) \right| = \sum_{i=1}^{m} |0| = 0,$$

and it is proved.

Theorem 6.15. $I_{Atkinson}$ satisfies P1.

Proof. We need to show that

$$1 - \left(\prod_{i=1}^{m} \frac{\left(\frac{n_i}{N}\right)}{\left(\frac{1}{m}\right)} \right)^{\frac{1}{m}} = 0.$$

The values in the vector $(n_1, \ldots, n_m)$ are uniformly distributed, so $n_i/N = 1/m$. Substituting for n_i/N in the above equality yields

$$1 - \left(\prod_{i=1}^{m} \frac{\left(\frac{1}{m}\right)}{\left(\frac{1}{m}\right)} \right)^{\frac{1}{m}} = 1 - \left(\prod_{i=1}^{m} 1 \right)^{\frac{1}{m}} = 1 - (1)^{\frac{1}{m}} = 1 - 1 = 0,$$

and it is proved.

Theorem 6.16. I_{Max} does not satisfy P1.

Proof. Left as an exercise for the reader. Let $N = 50$, $m = 2$, $n = (n_1, n_2) = (25, 25)$, and $n' = (n'_1, n'_2) = (49, 1)$.

6.2 Maximum Value Principle

Maximum Value Principle (P2). Given a vector $(n_1, \ldots, n_m)$, where $n_1 = N - m + 1$, $n_i = 1$, $i = 2, \ldots, m$, and $N > m$, $f(n_1, \ldots, n_m)$ attains its maximum value.

Theorem 6.17. $I_{Variance}$ satisfies P2.

Proof. We need to show that

$$\left(\frac{n_1 + c}{N} - \frac{1}{m}\right)^2 + \sum_{i=2}^{m} \left(\frac{n_i - d_i}{N} - \frac{1}{m}\right)^2 > \left(\frac{n_1}{N} - \frac{1}{m}\right)^2 + \sum_{i=2}^{m} \left(\frac{n_i}{N} - \frac{1}{m}\right)^2,$$

which is equivalent to showing that

$$\left(\frac{n_1+c}{N}-\frac{1}{m}\right)^2+\sum_{i=2}^{m}\left(\frac{n_i-d_i}{N}-\frac{1}{m}\right)^2-\left(\frac{n_1}{N}-\frac{1}{m}\right)^2-\sum_{i=2}^{m}\left(\frac{n_i}{N}-\frac{1}{m}\right)^2>0.$$

Expanding the left side of the above inequality, putting it over the common denominator Nm, and canceling terms yields

$$2n_1m^2c-2mcN+m^2c^2+\sum_{i=2}^{m}(2md_iN-2n_im^2d_i+m^2d_i^2)>0.$$

Since $c=\sum_{i=2}^{m}d_i$, substituting for c in the above inequality yields

$$2n_1m^2\sum_{i=2}^{m}d_i-2mN\sum_{i=2}^{m}d_i+m^2\left(\sum_{i=2}^{m}d_i\right)^2+2mN\sum_{i=2}^{m}d_i-2m^2\sum_{i=2}^{m}n_id_i+m^2\sum_{i=2}^{m}d_i^2>0.$$

Canceling terms and factoring m^2 out of the left side of the above inequality yields

$$2n_1\sum_{i=2}^{m}d_i+\left(\sum_{i=2}^{m}d_i\right)^2-2\sum_{i=2}^{m}n_id_i+\sum_{i=2}^{m}d_i^2>0.$$

Now one of the positive terms is greater than the negative term, specifically,

$$2n_1\sum_{i=2}^{m}d_i>2\sum_{i=2}^{m}n_id_i.$$

The two remaining terms are both positive, their sum is positive, and it is proved.

Theorem 6.18. $I_{Simpson}$ satisfies P2.

Proof. We need to show that

$$\left(\frac{n_1+c}{N}\right)^2+\sum_{i=2}^{m}\left(\frac{n_i-d_i}{N}\right)^2>\left(\frac{n_1}{N}\right)^2+\sum_{i=2}^{m}\left(\frac{n_i}{N}\right)^2,$$

which is equivalent to showing that

$$\left(\frac{n_1+c}{N}\right)^2+\sum_{i=2}^{m}\left(\frac{n_i-d_i}{N}\right)^2-\left(\frac{n_1}{N}\right)^2-\sum_{i=2}^{m}\left(\frac{n_i}{N}\right)^2>0.$$

Expanding the left side of the above inequality, putting it over the common denominator N^2, and canceling terms yields

$$2n_1c + c^2 + \sum_{i=2}^{m}(n_i^2 - 2n_id_i + d_i^2) - \sum_{i=2}^{m} n_i^2 > 0.$$

Since $c = \sum_{i=2}^{m} d_i$, substituting for c in the above inequality yields

$$2n_1\sum_{i=2}^{m} d_i + \left(\sum_{i=2}^{m} d_i\right)^2 + \sum_{i=2}^{m} n_i^2 - 2\sum_{i=2}^{m} n_id_i + \sum_{i=2}^{m} d_i^2 - \sum_{i=2}^{m} n_i^2 > 0.$$

Canceling terms yields

$$2n_1\sum_{i=2}^{m} d_i + \left(\sum_{i=2}^{m} d_i\right)^2 - 2\sum_{i=2}^{m} n_id_i + \sum_{i=2}^{m} d_i^2 > 0.$$

Now one of the positive terms is greater than the negative term, specifically,

$$2n_1\sum_{i=2}^{m} d_i > 2\sum_{i=2}^{m} n_id_i.$$

The two remaining terms are both positive, their sum is positive, and it is proved.

Theorem 6.19. $I_{Shannon}$ satisfies P2.

Proof. We need to show that

$$-\left(\left(\frac{n_1}{N}\right)\log_2\left(\frac{n_1}{N}\right) + \sum_{i=1}^{m}\left(\frac{n_i}{N}\right)\log_2\left(\frac{n_i}{N}\right)\right) >$$
$$-\left(\left(\frac{n_1+c}{N}\right)\log_2\left(\frac{n_1+c}{N}\right) + \sum_{i=2}^{m}\left(\frac{n_i-d_i}{N}\right)\log_2\left(\frac{n_i-d_i}{N}\right)\right).$$

Multiplying both sides by -1 yields

$$\left(\frac{n_1}{N}\right)\log_2\left(\frac{n_1}{N}\right) + \sum_{i=1}^{m}\left(\frac{n_i}{N}\right)\log_2\left(\frac{n_i}{N}\right) <$$
$$\left(\frac{n_1+c}{N}\right)\log_2\left(\frac{n_1+c}{N}\right) + \sum_{i=2}^{m}\left(\frac{n_i-d_i}{N}\right)\log_2\left(\frac{n_i-d_i}{N}\right).$$

Factoring $1/N$ out of both sides of the above inequality yields

$$
\begin{aligned}
&n_1 \log_2\left(\frac{n_1}{N}\right) + \sum_{i=1}^{m} n_i \log_2\left(\frac{n_i}{N}\right) < \\
&\quad (n_1 + c)\log_2\left(\frac{n_1 + c}{N}\right) + \sum_{i=2}^{m}(n_i - d_i)\log_2\left(\frac{n_i - d_i}{N}\right),
\end{aligned}
$$

which is equivalent to

$$
\begin{aligned}
&n_1 \log_2\left(\frac{n_1}{N}\right) + \sum_{i=1}^{m} n_i \log_2\left(\frac{n_i}{N}\right) < \\
&\quad n_1 \log_2\left(\frac{n_1 + c}{N}\right) + c\log_2\left(\frac{n_1 + c}{N}\right) + \sum_{i=2}^{m} n_i \log_2\left(\frac{n_i - d_i}{N}\right) - d_i \log_2\left(\frac{n_i - d_i}{N}\right).
\end{aligned}
$$

Using the difference property of logarithms to decompose the above inequality yields

$$
\begin{aligned}
&n_1 \log_2 n_1 - n_1 \log_2 N + \sum_{i=1}^{m} n_i \log_2 n_i - \sum_{i=1}^{m} n_i \log_2 N < \\
&\quad n_1 \log_2(n_1 + c) + c\log_2(n_1 + c) - n_1 \log_2 N - c\log_2 N + \sum_{i=2}^{m} n_i \log_2(n_i - d_i) \\
&\quad - \sum_{i=2}^{m} d_i \log_2(n_i - d_i) - \sum_{i=2}^{m} n_i \log_2 N + \sum_{i=2}^{m} d_i \log_2 N.
\end{aligned}
$$

Now $c = \sum_{i=2}^{m} d_i$, so substituting for c on the right side of the above inequality and canceling terms yields

$$
n_1 \log_2 n_1 + \sum_{i=1}^{m} n_i \log_2 n_i < (n_1 + c)\log_2(n_1 + c) + \sum_{i=2}^{m}(n_i - d_i)\log_2(n_i - d_i),
$$

which can be written as

$$
\begin{aligned}
&n_1 \log_2 n_1 + n_2 \log_2 n_2 + \ldots + n_m \log_2 n_m < \\
&\quad (n_1 + c)\log_2(n_1 + c) + (n_2 - d_2)\log_2(n_2 - d_2) + \ldots + (n_m - d_m)\log_2(n_m - d_m).
\end{aligned}
$$

Since the function $x \log_2 x$ is a convex function, we have

$$(n_1+c)\log_2(n_1+c)+(n_2-d_2)\log_2(n_2-d_2)+\ldots+(n_m-d_m)\log_2(n_m-d_m) \geq$$
$$(n_1+c+n_2-d_2+\ldots+n_m-d_m)\log_2(n_1+c+n_2-d_2+\ldots+n_m-d_m).$$

Now $c=\sum_{i=2}^{m} d_i$, so substituting for the d_i's on the right side of the above inequality yields

$$(n_1+c+n_1-d_2+\ldots+n_1-d_m)\log_2(n_1+c+n_1-d_2+\ldots+n_1-d_m)=$$
$$(n_1+n_2+\ldots+n_m)\log_2(n_1+n_2+\ldots+n_m).$$

We now show that

$$n_1\log_2 n_1+n_2\log_2 n_2+\ldots+n_m\log_2 n_m<(n_1+n_2+\ldots+n_m)\log_2(n_1+n_2+\ldots+n_m).$$

Distributing $(n_1+n_2+\ldots+n_m)$ over the log on the right side of the above inequality yields

$$n_1\log_2 n_1+n_2\log_2 n_2+\ldots+n_m\log_2 n_m<$$
$$n_1\log_2(n_1+n_2+\ldots+n_m)+n_2\log_2(n_1+n_2+\ldots+n_m)+\ldots$$
$$+n_m\log_2(n_1+n_2+\ldots+n_m),$$

which is obviously true, and it is proved.

Theorem 6.20. I_{Total} satisfies P2.

Proof. We need to show that

$$m\left(-\sum_{i=1}^{m}\left(\frac{n_i}{N}\right)\log_2\left(\frac{n_i}{N}\right)\right)>$$
$$m\left(-\left(\left(\frac{n_1+c}{N}\right)\log_2\left(\frac{n_1+c}{N}\right)+\sum_{i=2}^{m}\left(\frac{n_i-d_i}{N}\right)\log_2\left(\frac{n_i-d_i}{N}\right)\right)\right).$$

Canceling terms and multiplying both sides of the above inequality by -1 yields

$$\sum_{i=1}^{m}\left(\frac{n_i}{N}\right)\log_2\left(\frac{n_i}{N}\right) < \left(\frac{n_1+c}{N}\right)\log_2\left(\frac{n_1}{N}\right) + \sum_{i=2}^{m}\left(\frac{n_i-d_i}{N}\right)\log_2\left(\frac{n_i-d_i}{N}\right),$$

which is proved in Theorem 6.19.

Theorem 6.21. $I_{McIntosh}$ satisfies P2.

Proof. We need to show that

$$\frac{N-\sqrt{(n_1+c)^2+\sum_{i=2}^{m}(n_i-d_i)^2}}{N-\sqrt{N}} < \frac{N-\sqrt{n_1^2+\sum_{i=2}^{m}n_i^2}}{N-\sqrt{N}}.$$

Canceling terms and multiplying both sides of the above inequality by -1 yields

$$(n_1+c)^2+\sum_{i=2}^{m}(n_i-d_i)^2 > n_1^2+\sum_{i=2}^{m}n_i^2.$$

Expanding the terms on the left side of the above inequality yields

$$n_1^2+2n_1c+c^2+\sum_{i=2}^{m}(n_i^2-2n_id_i+d_i^2) > n_1^2+\sum_{i=2}^{m}n_i^2.$$

Distributing the summation on the left side of the above inequality yields

$$n_1^2+2n_1c+c^2+\sum_{i=2}^{m}n_i^2-\sum_{i=2}^{m}2n_id_i+\sum_{i=2}^{m}d_i^2 > n_1^2+\sum_{i=2}^{m}n_i^2.$$

Since $c=\sum_{i=2}^{m}d_i$, substituting for c in the above inequality yields

$$n_1^2+\sum_{i=2}^{m}2n_1d_i+\left(\sum_{i=2}^{m}d_i\right)^2+\sum_{i=2}^{m}n_i^2-\sum_{i=2}^{m}2n_id_i+\sum_{i=2}^{m}d_i^2 > n_1^2+\sum_{i=2}^{m}n_i^2.$$

Canceling terms yields

$$2\sum_{i=2}^{m}(n_1-n_i)d_i+\left(\sum_{i=2}^{m}d_i\right)^2+\sum_{i=2}^{m}d_i^2 > 0.$$

Since all terms on the left side of the above inequality are positive, their sum is positive, and it is proved.

Theorem 6.22. I_{Lorenz} satisfies P2.

Proof. We need to show that

$$\frac{1}{m}\left(\sum_{i=1}^{m-1}(m-i+1)\frac{n_i}{N}+\frac{n_m}{N}\right) > \frac{1}{m}\left(\sum_{i=1}^{m-1}(m-i+1)\frac{n_i-d_i}{N}+\frac{n_m+c}{N}\right).$$

Canceling terms and distributing the summation on the right side of the above inequality yields

$$\sum_{i=1}^{m-1}(m-i+1)\frac{n_i}{N}+\frac{n_m}{N} > \sum_{i=1}^{m-1}(m-i+1)\frac{n_i}{N}-\sum_{i=1}^{m-1}(m-i+1)\frac{d_i}{N}+\frac{n_m}{N}+\frac{c}{N}.$$

Canceling terms yields

$$0 > -\sum_{i=1}^{m-1}(m-i+1)\frac{d_i}{N}+\frac{c}{N}.$$

Since $c=\sum_{i=1}^{m-1} d_i$, substituting for c in the above inequality, and moving $-\sum_{i=1}^{m-1}\frac{d_i}{N}$ to the left side yields

$$\sum_{i=1}^{m-1}(m-i+1)\frac{d_i}{N} > \sum_{i=1}^{m-1}\frac{d_i}{N}.$$

Factoring $1/N$ out of both sides of the above inequality yields

$$\sum_{i=1}^{m-1}(m-i+1)d_i > \sum_{i=1}^{m-1} d_i,$$

which is obviously true, and it is proved.

Theorem 6.23. I_{Gini} satisfies P2.

Proof. To simplify the proof of this theorem, we use an equivalent form of the Gini index described in [5], where it is given as

$$\frac{2\left(\frac{m+1}{2}-\sum_{i=1}^{m} i\left(\frac{n_i}{N}\right)\right)}{m}.$$

We need to show that

$$\frac{2\left(\frac{m+1}{2}-\left(\frac{n_1+c}{N}+\sum_{i=2}^{m} i\left(\frac{n_i-d_i}{N}\right)\right)\right)}{m} > \frac{2\left(\frac{m+1}{2}-\left(\frac{n_1}{N}+\sum_{i=2}^{m} i\left(\frac{n_i}{N}\right)\right)\right)}{m}.$$

Canceling terms and multiplying both sides of the above inequality by -1 yields

$$n_1 + c + \sum_{i=2}^{m} i(n_i - d_i) < n_1 + \sum_{i=2}^{m} in_i.$$

Distributing the summation on the left side of the above inequality, canceling terms and moving $-\sum_{i=2}^{m} id_i$ to the right side yields

$$c < \sum_{i=2}^{m} id_i.$$

Since $c = \sum_{i=2}^{m} d_i$, substituting for c in the above inequality yields

$$\sum_{i=2}^{m} d_i < \sum_{i=2}^{m} id_i,$$

which is obviously true, and it is proved.

Theorem 6.24. I_{Berger} satisfies P2.

Proof. We need to show that

$$\frac{n_1 + c}{N} > \frac{n_1}{N},$$

which is proved in Theorem 6.8.

Theorem 6.25. I_{Schutz} satisfies P2.

Proof. We need to show that

$$\left|\frac{N - m + 1}{N} - \frac{1}{m}\right| + \sum_{i=2}^{m} \left|\frac{1}{N} - \frac{1}{m}\right| > \sum_{i=1}^{m} \left|\frac{n_i}{N} - \frac{1}{m}\right|,$$

which is equivalent to showing that

$$\left|\frac{N - m + 1}{N} - \frac{1}{m}\right| + \sum_{i=2}^{m} \left|\frac{1}{N} - \frac{1}{m}\right| > \sum_{i=1}^{k} \left|\frac{n_i}{N} - \frac{1}{m}\right| + \sum_{i=k+1}^{m} \left|\frac{n_i}{N} - \frac{1}{m}\right|.$$

If we assume that k is chosen such that $n_i/N > 1/m$, for $i \leq k$, and $n_i/N < 1/m$, for $i > k$, then the above inequality is equivalent to

$$\left(\frac{N-m+1}{N}-\frac{1}{m}\right)+\sum_{i=2}^{m}\left(\frac{1}{N}-\frac{1}{m}\right)>\sum_{i=1}^{k}\left(\frac{n_i}{N}-\frac{1}{m}\right)-\sum_{i=k+1}^{m}\left(\frac{n_i}{N}-\frac{1}{m}\right).$$

Now for every $k \in \{1,\ldots,m\}$,

$$\sum_{i=1}^{k}\left(\frac{n_i}{N}-\frac{1}{m}\right)=-\sum_{i=k+1}^{m}\left(\frac{n_i}{N}-\frac{1}{m}\right),$$

so substituting for $-\sum_{i=k+1}^{m}(n_i/N-1/m)$ in the above inequality yields

$$2\left(\frac{N-m+1}{N}-\frac{1}{m}\right)>2\sum_{i=1}^{k}\left(\frac{n_i}{N}-\frac{1}{m}\right).$$

Canceling terms and putting both sides of the above inequality over the common denominator Nm yields

$$Nm-m^2+m-N>\sum_{i=1}^{k}mn_i-Nk,$$

which is equivalent to

$$Nm+Nk-N>\sum_{i=1}^{k}mn_i+m^2-m.$$

Since $N=mn_k$ (i.e., n_k is the average), substituting for N in the above inequality and factoring m out of both sides yields

$$mn_k+kn_k-n_k>\sum_{i=1}^{k}n_i+m-1,$$

which is equivalent to

$$mn_k+kn_k+1>\sum_{i=1}^{k}n_i+m+n_k.$$

Since $mn_k=N=\sum_{i=1}^{m}n_i=\sum_{i=1}^{k}n_i+\sum_{i=k+1}^{m}n_i$, substituting for mn_k in the above inequality yields

$$\sum_{i=1}^{k}n_i+\sum_{i=k+1}^{m}n_i+kn_k+1>\sum_{i=1}^{k}n_i+m+n_k.$$

Canceling $\sum_{i=1}^{k} n_i$ from both sides of the above inequality, and moving n_k to the left side, yields

$$\sum_{i=k+1}^{m} n_i + (k-1)n_k + 1 > m,$$

which can be written as

$$n_k + n_k + \ldots + n_k + n_{k+1} + n_{k+2} + \ldots + n_m + 1 > m.$$

The left side of the above inequality contains m terms, consisting of $(k-1)$ n_k terms, $(m-k)$ n_i terms, and 1. The vector $(n_1, \ldots, n_m)$ is not a uniform distribution, so $n_k > 1$, each $n_i \geq 1$, for $i = k+1, \ldots, m$, the sum of these terms is greater than m, and it is proved.

Theorem 6.26. I_{Bray} satisfies P2.

Proof. We need to show that

$$\frac{1}{N}\left(\frac{N}{m} + \sum_{i=2}^{m} 1\right) < \frac{1}{N}\left(\frac{N}{m} + \sum_{i=2}^{k} \frac{N}{m} + \sum_{i=k+1}^{m} (n_i + d_i)\right).$$

Canceling terms yields

$$\sum_{i=2}^{m} 1 < \sum_{i=2}^{k} \frac{N}{m} + \sum_{i=k+1}^{m} (n_i + d_i),$$

which can be written as

$$\sum_{i=2}^{k} 1 + \sum_{i=k+1}^{m} 1 < \sum_{i=2}^{k} \frac{N}{m} + \sum_{i=k+1}^{m} (n_i + d_i).$$

Now $N/m > 1$,

$$\sum_{i=2}^{k} \frac{N}{m} > \sum_{i=2}^{k} 1$$

$n_i + d_i > 1$,

$$\sum_{i=k+1}^{m} (n_i + d_i) > \sum_{i=k+1}^{m} 1,$$

and it is proved.

Theorem 6.27. $I_{Whittaker}$ satisfies P2.

Proof. We need to show that

$$1-\left(0.5\sum_{i=1}^{m}\left|\frac{n_i}{N}-\frac{1}{m}\right|\right)>1-\left(0.5\left(\left|\frac{N-m+1}{N}-\frac{1}{m}\right|+\sum_{i=2}^{m}\left|\frac{1}{N}-\frac{1}{m}\right|\right)\right).$$

Canceling terms and multiplying both sides of the above inequality by - 1 yields

$$\sum_{i=1}^{m}\left|\frac{n_i}{N}-\frac{1}{m}\right|<\left|\frac{N-m+1}{N}-\frac{1}{m}\right|+\sum_{i=2}^{m}\left|\frac{1}{N}-\frac{1}{m}\right|,$$

which is proved in Theorem 6.25.

Theorem 6.28. $I_{Kullback}$ satisfies P2.

Proof. We need to show that

$$\begin{aligned}log_2m-&\left(\left(\frac{n_1}{N}\right)\log_2\left(\frac{mn_1}{N}\right)+\sum_{i=2}^{m}\left(\frac{n_i}{N}\right)\log_2\left(\frac{mn_i}{N}\right)\right)> \\ log_2m-&\left(\left(\frac{n_1+c}{N}\right)\log_2\left(\frac{m(n_1+c)}{N}\right)+\sum_{i=2}^{m}\left(\frac{n_i-d_i}{N}\right)\log_2\left(\frac{m(n_i-d_i)}{N}\right)\right).\end{aligned}$$

Canceling terms and multiplying both sides of the above inequality by -1 yields

$$\begin{aligned}&\left(\frac{n_1}{N}\right)\log_2\left(\frac{mn_1}{N}\right)+\sum_{i=2}^{m}\left(\frac{n_i}{N}\right)\log_2\left(\frac{mn_i}{N}\right)< \\ &\quad\left(\frac{n_1+c}{N}\right)\log_2\left(\frac{m(n_1+c)}{N}\right)+\sum_{i=2}^{m}\left(\frac{n_i-d_i}{N}\right)\log_2\left(\frac{m(n_i-d_i)}{N}\right).\end{aligned}$$

Factoring $1/N$ out of both sides of the above inequality yields

$$\begin{aligned}&n_1\log_2\left(\frac{mn_1}{N}\right)+\sum_{i=2}^{m}n_1\log_2\left(\frac{mn_1}{N}\right)< \\ &\quad(n_1+c)\log_2\left(\frac{m(n_1+c)}{N}\right)+\sum_{i=2}^{m}(n_i-d_i)\log_2\left(\frac{m(n_i-d_i)}{N}\right),\end{aligned}$$

which is equivalent to

$$
\begin{aligned}
& n_1 \log_2 \left(\frac{mn_1}{N}\right) + \sum_{i=2}^{m} n_1 \log_2 \left(\frac{mn_1}{N}\right) < \\
& \quad n_1 \log_2 \left(\frac{m(n_1+c)}{N}\right) + c \log_2 \left(\frac{m(n_1+c)}{N}\right) + \sum_{i=2}^{m} n_i \log_2 \left(\frac{m(n_i-d_i)}{N}\right) \\
& \quad - \sum_{i=2}^{m} d_i \log_2 \left(\frac{m(n_i-d_i)}{N}\right).
\end{aligned}
$$

Using the difference property of logarithms to decompose the above inequality yields

$$
\begin{aligned}
& n_1 \log_2 mn_1 - n_1 \log_2 N + \sum_{i=2}^{m} n_i \log_2 (mn_i) - \sum_{i=2}^{m} n_i \log_2 N < \\
& \quad (n_1+c) \log_2 (m(n_1+c)) - n_1 \log_2 N - c \log_2 N + \sum_{i=2}^{m} n_i \log_2 (m(n_i-d_i)) \\
& \quad - \sum_{i=2}^{m} d_i \log_2 (m(n_i-d_i)) - \sum_{i=2}^{m} n_i \log_2 N + \sum_{i=2}^{m} d_i \log_2 N.
\end{aligned}
$$

Now $c = \sum_{i=2}^{m} d_i$, so substituting for c on the right side of the above inequality and canceling terms yields

$$
\begin{aligned}
& n_1 \log_2 mn_1 + \sum_{i=2}^{m} n_i \log_2 (mn_i) < \\
& \quad n_1 \log_2 (m(n_1+c)) + c \log_2 (m(n_1+c)) + \sum_{i=2}^{m} n_i \log_2 (m(n_i-d_i)) \\
& \quad - \sum_{i=2}^{m} d_i \log_2 (m(n_i-d_i)).
\end{aligned}
$$

Using the addition property of logarithms to further decompose the above inequality yields

$$
\begin{aligned}
& n_1 \log_2 m + n_1 \log_2 n_1 + \sum_{i=2}^{m} n_i \log_2 m + \sum_{i=2}^{m} n_i \log_2 n_i < \\
& \quad n_1 \log_2 m + n_1 \log_2 (n_1+c) + c \log_2 m + c \log_2 (n_1+c) + \sum_{i=2}^{m} n_i \log_2 m \\
& \quad + \sum_{i=2}^{m} n_i \log_2 (n_i-d_i) - \sum_{i=2}^{m} d_i \log_2 m - \sum_{i=2}^{m} d_i \log_2 (n_i-d_i).
\end{aligned}
$$

Again substituting for c on the right side of the above inequality, and canceling terms yields

$$n_1 \log_2 n_1 + \sum_{i=2}^{m} n_i \log_2 n_i < (n_1 + c) \log_2(n_1 + c) + \sum_{i=2}^{m}(n_i - d_i) \log_2(n_i - d_i),$$

which is proved in Theorem 6.3.

Theorem 6.29. $I_{MacArthur}$ satisfies P2.

Proof. Let the vector $(a, \ldots, a)$ containing m a's be the uniform distribution. Then the vector obtained by combining the vectors $(n_1 + c, n_2 - d_2, \ldots, n_m - d_m)$ and $(a, \ldots, a)$ is the vector $(n_1 + c + a, n_2 - d_2 + a, \ldots, n_m - d_m + a)$, and the vector obtained by combining the vectors $(n_1, \ldots, n_m)$ and $(a, \ldots, a)$ is the vector $(n_1 + a, \ldots, n_m + a)$. We need to show that

$$\begin{aligned}
&-\left(\left(\frac{n_1+c+a}{2N}\right)\log_2\left(\frac{n_1+c+a}{2N}\right) + \sum_{i=2}^{m}\left(\frac{n_i-d_i+a}{2N}\right)\log_2\left(\frac{n_i-d_i+a}{2N}\right)\right) \\
&-\left(\frac{-\left(\left(\frac{n_1+c}{N}\right)\log_2\left(\frac{n_1+c}{N}\right) + \sum_{i=2}^{m}\left(\frac{n_i-d_i}{N}\right)\log_2\left(\frac{n_i-d_i}{N}\right)\right) + \log_2 m}{2}\right) > \\
&-\left(\left(\frac{n_1+a}{2N}\right)\log_2\left(\frac{n_1+a}{2N}\right) + \sum_{i=2}^{m}\left(\frac{n_i+a}{2N}\right)\log_2\left(\frac{n_i+a}{2N}\right)\right) \\
&-\left(\frac{-\left(\left(\frac{n_1}{N}\right)\log_2\left(\frac{n_1}{N}\right) + \sum_{i=2}^{m}\left(\frac{n_i}{N}\right)\log_2\left(\frac{n_i}{N}\right)\right) + \log_2 m}{2}\right),
\end{aligned}$$

which is equivalent to showing that

$$\begin{aligned}
&-\left(\frac{n_1+c+a}{2N}\right)\log_2\left(\frac{n_1+c+a}{2N}\right) - \sum_{i=2}^{m}\left(\frac{n_i-d_i+a}{2N}\right)\log_2\left(\frac{n_i-d_i+a}{2N}\right) \\
&+\left(\frac{n_1+c}{2N}\right)\log_2\left(\frac{n_1+c}{N}\right) + \sum_{i=2}^{m}\left(\frac{n_i-d_i}{2N}\right)\log_2\left(\frac{n_i-d_i}{N}\right) - \frac{\log_2 m}{2} > \\
&-\left(\frac{n_1+a}{2N}\right)\log_2\left(\frac{n_1+a}{2N}\right) - \sum_{i=2}^{m}\left(\frac{n_i+a}{2N}\right)\log_2\left(\frac{n_i+a}{2N}\right) + \left(\frac{n_1}{2N}\right)\log_2\left(\frac{n_1}{N}\right) \\
&+\sum_{i=2}^{m}\left(\frac{n_i}{2N}\right)\log_2\left(\frac{n_i}{N}\right) - \frac{\log_2 m}{2}.
\end{aligned}$$

Canceling terms, factoring $1/2N$ out of both sides of the above inequality, and moving those terms preceded by a minus sign to the other side of the inequality yields

$$
\begin{aligned}
&(n_1+a)\log_2\left(\frac{n_1+a}{2N}\right)+\sum_{i=2}^{m}(n_i+a)\log_2\left(\frac{n_i+a}{2N}\right)+(n_1+c)\log_2\left(\frac{n_1+c}{N}\right)\\
&+\sum_{i=2}^{m}(n_i-d_i)\log_2\left(\frac{n_i-d_i}{N}\right)>\\
&\quad(n_1+c+a)\log_2\left(\frac{n_1+c+a}{2N}\right)+\sum_{i=2}^{m}(n_i-d_i+a)\log_2\left(\frac{n_i-d_i+a}{2N}\right)\\
&\quad+n_1\log_2\left(\frac{n_1}{N}\right)+\sum_{i=2}^{m}n_i\log_2\left(\frac{n_i}{N}\right).
\end{aligned}
$$

Using the difference property of logarithms to decompose the above inequality yields

$$
\begin{aligned}
&(n_1+a)\log_2(n_1+a)-n_1\log_2(2N)-a\log_2(2N)+\sum_{i=2}^{m}(n_i+a)\log_2(n_i+a)\\
&-\sum_{i=2}^{m}n_i\log_2(2N)-\sum_{i=2}^{m}a\log_2(2N)+(n_1+c)\log_2(n_1+c)-n_1\log_2 N-c\log_2 N\\
&+\sum_{i=2}^{m}(n_i-d_i)\log_2(n_i-d_i)-\sum_{i=2}^{m}n_i\log_2 N+\sum_{i=2}^{m}d_i\log_2 N>\\
&\quad(n_1+c+a)\log_2(n_1+c+a)-n_1\log_2(2N)-c\log_2(2N)-a\log_2(2N)\\
&\quad+\sum_{i=2}^{m}(n_i-d_i+a)\log_2(n_i-d_i+a)-\sum_{i=2}^{m}n_i\log_2(2N)+\sum_{i=2}^{m}d_i\log_2(2N)\\
&\quad-\sum_{i=2}^{m}a\log_2(2N)+n_1\log_2 n_1-n_1\log_2 N+\sum_{i=2}^{m}n_i\log_2 n_i-\sum_{i=2}^{m}n_i\log_2 N.
\end{aligned}
$$

Canceling terms yields

$$
\begin{aligned}
&(n_1+a)\log_2(n_1+a)+\sum_{i=2}^{m}(n_i+a)\log_2\left(\frac{n_i+a}{2N}\right)+(n_1+c)\log_2\left(\frac{n_1+c}{N}\right)\\
&+\sum_{i=2}^{m}(n_i-d_i)\log_2\left(\frac{n_i-d_i}{N}\right)>\\
&\quad(n_1+c+a)\log_2\left(\frac{n_1+c+a}{2N}\right)+\sum_{i=2}^{m}(n_i-d_i+a)\log_2\left(\frac{n_i-d_i+a}{2N}\right)\\
&\quad+\sum_{i=2}^{m}n_i\log_2 n_i.
\end{aligned}
$$

To show that the above inequality is true, we need to show that

$$(n_1 + a)\log_2(n_1 + a) + (n_1 + c)\log_2(n_1 + c) > (n_1 + c + a)\log_2(n_1 + c + a)$$

and

$$\sum_{i=2}^{m}(n_i + a)\log_2(n_i + a) + \sum_{i=2}^{m}(n_i - d_i)\log_2(n_i - d_i) >$$
$$\sum_{i=2}^{m}(n_i - d_i + a)\log_2(n_i - d_i + a) + \sum_{i=2}^{m} n_i \log_2 n_i.$$

We now show that

$$(n_1 + a)\log_2(n_1 + a) + (n_1 + c)\log_2(n_1 + c) > (n_1 + c + a)\log_2(n_1 + c + a).$$

Since the function $x\log_2 x$ is a convex function, we have

$$(n_1 + a)\log_2(n_1 + a) + (n_1 + c)\log_2(n_1 + c) \geq$$
$$(n_1 + c + n_1 + a)\log_2(n_1 + c + n_1 + a),$$

so obviously

$$(n_1 + a)\log_2(n_1 + a) + (n_1 + c)\log_2(n_1 + c) > (n_1 + c + a)\log_2(n_1 + c + a),$$

and the first part is shown. We now show that

$$\sum_{i=2}^{m}(n_i + a)\log_2(n_i + a) + \sum_{i=2}^{m}(n_i - d_i)\log_2(n_i - d_i) >$$
$$\sum_{i=2}^{m}(n_i - d_i + a)\log_2(n_i - d_i + a) + \sum_{i=2}^{m} n_i \log_2 n_i.$$

The above inequality can be written as

$$
\begin{aligned}
&(n_2+a)\log_2(n_2+a)+(n_3+a)\log_2(n_3+a)+\ldots+(n_m+a)\log_2(n_m+a)\\
&+(n_2-d_2)\log_2(n_2-d_2)+(n_3-d_3)\log_2(n_3-d_3)+\ldots\\
&+(n_m-d_m)\log_2(n_m-d_m)>\\
&\quad(n_2-d_2+a)\log_2(n_2-d_2+a)+(n_3-d_3+a)\log_2(n_3-d_3+a)+\ldots\\
&\quad+(n_m-d_m+a)\log_2(n_m-d_m+a)+n_2\log_2 n_2+n_3\log_2 n_3+\ldots+n_m\log_2 n_m.
\end{aligned}
$$

Since the function $x\log_2 x$ is a convex function, we have

$$
\begin{aligned}
&(n_2+a)\log_2(n_2+a)+(n_3+a)\log_2(n_3+a)+\ldots+(n_m+a)\log_2(n_m+a)\\
&+(n_2-d_2)\log_2(n_2-d_2)+(n_3-d_3)\log_2(n_3-d_3)+\ldots\\
&+(n_m-d_m)\log_2(n_m-d_m)\geq\\
&\quad(n_2+a+n_3+a+\ldots+n_m+a+n_2-d_2+n_3-d_3+\ldots+n_m-d_m)\\
&\quad\log_2(n_2+a+n_3+a+\ldots+n_m+a+n_2-d_2+n_3-d_3+\ldots+n_m-d_m),
\end{aligned}
$$

which can be written as

$$
\begin{aligned}
&(n_2+a)\log_2(n_2+a)+(n_3+a)\log_2(n_3+a)+\ldots+(n_m+a)\log_2(n_m+a)\\
&+(n_2-d_2)\log_2(n_2-d_2)+(n_3-d_3)\log_2(n_3-d_3)+\ldots\\
&+(n_m-d_m)\log_2(n_m-d_m)\geq\\
&\quad(n_2+n_3+\ldots+n_m+(n_2-d_2+a)+(n_3-d_3+a)+\ldots\\
&\quad+(n_m-d_m+a))\log_2(n_2+n_3+\ldots+n_m+(n_2-d_2+a)+(n_3-d_3+a)+\ldots\\
&\quad+(n_m-d_m+a)).
\end{aligned}
$$

Distributing $(n_2+n_3+\ldots+n_m+(n_2-d_2+a)+(n_3-d_3+a)+\ldots+(n_m-d_m+a))$ over the log in the above inequality yields

$$(n_2+a)\log_2(n_2+a)+(n_3+a)\log_2(n_3+a)+\ldots+(n_m+a)\log_2(n_m+a)$$

$$+(n_2-d_2)\log_2(n_2-d_2)+(n_3-d_3)\log_2(n_3-d_3)+\ldots$$

$$+(n_m-d_m)\log_2(n_m-d_m)\geq$$

$$(n_2\log_2(n_2+n_3+\ldots+n_m+(n_2-d_2+a)+(n_3-d_3+a)+\ldots$$

$$+(n_m-d_m+a))+n_3\log_2(n_2+n_3+\ldots+n_m+(n_2-d_2+a)$$

$$+(n_3-d_3+a)+\ldots+(n_m-d_m+a))+\ldots+n_m\log_2(n_2+n_3+\ldots+n_m$$

$$+(n_2-d_2+a)+(n_3-d_3+a)+\ldots+(n_m-d_m+a))+(n_2-d_2+a)\log_2(n_2$$

$$+n_3+\ldots+n_m+(n_2-d_2+a)+(n_3-d_3+a)+\ldots+(n_m-d_m+a))$$

$$+(n_3-d_3+a)\log_2(n_2+n_3+\ldots+n_m+(n_2-d_2+a)+(n_3-d_3+a)+\ldots$$

$$+(n_m-d_m+a))+\ldots+(n_m-d_m+a))\log_2(n_2+n_3+\ldots+n_m$$

$$+(n_2-d_2+a)+(n_3-d_3+a)+\ldots+(n_m-d_m+a))$$

so obviously

$$(n_2+a)\log_2(n_2+a)+(n_3+a)\log_2(n_3+a)+\ldots+(n_m+a)\log_2(n_m+a)$$

$$+(n_2-d_2)\log_2(n_2-d_2)+(n_3-d_3)\log_2(n_3-d_3)+\ldots$$

$$+(n_m-d_m)\log_2(n_m-d_m)>$$

$$n_2\log_2 n_2+n_3\log_2 n_3+\ldots+n_m\log_2 n_m+(n_2-d_2+a)\log_2(n_2-d_2+a)$$

$$+(n_3-d_3+a)\log_2(n_3-d_3+a)+\ldots+(n_m-d_m+a)\log_2(n_m-d_m+a),$$

the second part is shown, and it is proved.

Theorem 6.30. $I_{Atkinson}$ satisfies P2.

Proof. We need to show that

$$1-\left(\frac{\left(\frac{N-m+1}{N}\right)}{\left(\frac{1}{m}\right)}\prod_{i=2}^{m}\frac{\left(\frac{1}{N}\right)}{\left(\frac{1}{m}\right)}\right)^{\frac{1}{m}} > 1-\left(\prod_{i=1}^{m}\frac{\left(\frac{n_i}{N}\right)}{\left(\frac{1}{m}\right)}\right)^{\frac{1}{m}}.$$

Canceling terms and multiplying both sides by -1 yields

$$N-m+1<\prod_{i=1}^{m}n_i,$$

which is equivalent to

$$N+1<\prod_{i=1}^{m}n_i+m.$$

Since $N=\sum_{i=1}^{m}n_i$, substituting for N in the above inequality yields

$$\sum_{i=1}^{m}n_i+1<\prod_{i=1}^{m}n_i+m.$$

The vector $(n_1,\ldots,n_m)$ is not a uniform distribution, so $n_1>1$ and $\sum_{i=1}^{m}n_i<\prod_{i=1}^{m}n_i$. Also, $1<m$, so the sum of the terms on the left side of the above inequality is less than the sum of the terms on the right side, and it is proved.

Theorem 6.31. I_{Max} and I_{Theil} do not satisfy P2.

Proof. Left as an exercise for the reader. Let $N=50$, $m=2$, $n=(n_1,n_2)=(49,1)$, and $n^{'}=(n_1^{'},n_2^{'})=(25,25)$.

6.3 Skewness Principle

Skewness Principle (P3). Given a vector $(n_1,\ldots,n_m)$, where $n_1=N-m+1$, $n_i=1$, $i=2,\ldots,m$, and $N>m$, and a vector $(n_1-c,n_2,\ldots,n_m,n_{m+1},\ldots,n_{m+c})$, where $n_1-c>1$ and $n_i=1$, $i=2,\ldots,m+c$, $f(n_1,\ldots,n_m)>f(n_1-c,n_2,\ldots,n_m,n_{m+1},\ldots,n_{m+c})$.

Conjecture 6.32. $I_{Variance}$ satisfies P3.

Proof. To be completed as part of future research.

Theorem 6.33. $I_{Simpson}$ satisfies P3.

Proof. We need to show that

$$\left(\frac{n_1}{N}\right)^2 + \sum_{i=2}^{m}\left(\frac{1}{N}\right)^2 > \left(\frac{n_1 - k}{N}\right)^2 + \sum_{i=2}^{m+k}\left(\frac{1}{N}\right)^2.$$

Expanding both sides of the above inequality yields

$$\frac{n_1^2}{N^2} + \frac{m-1}{N^2} > \frac{n_1^2 - 2n_1k + k^2}{N^2} + \frac{m+k-1}{N^2}.$$

Canceling terms yields

$$\frac{2n_1k}{N^2} > \frac{k^2+k}{N^2}.$$

Factoring k/N^2 out of both sides of the above inequality yields

$$2n_1 > k + 1.$$

Now $k \geq 1$ and $n_1 > k$, so $2n_1 > k+1$, and it is proved.

Theorem 6.34. $I_{Shannon}$ satisfies P3.

Proof. We need to show that

$$-\left(\left(\frac{n_1}{N}\right)\log_2\left(\frac{n_1}{N}\right) + \sum_{i=2}^{m}\left(\frac{1}{N}\right)\log_2\left(\frac{1}{N}\right)\right) <$$
$$-\left(\left(\frac{n_1 - c}{N}\right)\log_2\left(\frac{n_1 - c}{N}\right) + \sum_{i=2}^{m+c}\left(\frac{1}{N}\right)\log_2\left(\frac{1}{N}\right)\right).$$

Multiplying both sides of the above inequality by -1 and factoring $1/N$ out of both sides yields

$$n_1\log_2\left(\frac{n_1}{N}\right) + (m-1)\log_2\left(\frac{1}{N}\right) > (n_1 - c)\log_2\left(\frac{n_1 - c}{N}\right) + (m+c-1)\log_2\left(\frac{1}{N}\right),$$

which is equivalent to

$$n_1 \log_2 \left(\frac{n_1}{N}\right) > (n_1 - c) \log_2 \left(\frac{n_1 - c}{N}\right) + c \log_2 \left(\frac{1}{N}\right).$$

Using the difference and addition properties of logarithms yields

$$\begin{aligned} & n_1 \log_2 n_1 - n_1 \log_2 N > \\ & \quad n_1 \log_2 (n_1 - c) - n_1 \log_2 N - c \log_2 (n_1 - c) + c \log_2 N + c \log_2 1 - c \log_2 N. \end{aligned}$$

Canceling terms yields

$$n_1 \log_2 n_1 > (n_1 - c) \log_2 (n_1 - c),$$

which is obviously true, and it is proved.

Theorem 6.35. I_{Total} satisfies P3.

Proof. We need to show that

$$\begin{aligned} & m\left(-\left(\left(\frac{n_1}{N}\right)\log_2\left(\frac{n_1}{N}\right) + \sum_{i=2}^{m}\left(\frac{1}{N}\right)\log_2\left(\frac{1}{N}\right)\right)\right) < \\ & \quad m\left(-\left(\left(\frac{n_1 - c}{N}\right)\log_2\left(\frac{n_1 - c}{N}\right) + \sum_{i=2}^{m+c}\left(\frac{1}{N}\right)\log_2\left(\frac{1}{N}\right)\right)\right). \end{aligned}$$

Canceling terms and multiplying both sides of the above inequality by -1 yields

$$\left(\frac{n_1}{N}\right)\log_2\left(\frac{n_1}{N}\right) + \sum_{i=2}^{m}\left(\frac{1}{N}\right)\log_2\left(\frac{1}{N}\right) > \left(\frac{n_1 - c}{N}\right)\log_2\left(\frac{n_1 - c}{N}\right) + \sum_{i=2}^{m+c}\left(\frac{1}{N}\right)\log_2\left(\frac{1}{N}\right),$$

which is proved in Theorem 6.34.

Theorem 6.36. I_{Max} satisfies P3.

Proof. We need to show that

$$\log_2 m < \log_2 (m + c).$$

Now $c > 0$, so the above inequality is obviously true, and it is proved.

Theorem 6.37. $I_{McIntosh}$ satisfies P3.

Proof. We need to show that

$$\frac{N - \sqrt{n_1^2 + \sum_{i=2}^{m} 1^2}}{N - \sqrt{N}} < \frac{N - \sqrt{(n_1 - k)^2 + \sum_{i=2}^{m+k} 1^2}}{N - \sqrt{N}}.$$

Canceling terms and multiplying both sides of the above inequality by -1 yields

$$n_1^2 + m - 1 > (n_1 - k)^2 + (m + k - 1).$$

Canceling terms yields

$$0 > -2n_1 k + k^2 + k.$$

Moving $-2n_1 k$ to the left side of the above inequality and factoring k out of both sides yields

$$2n_1 > k + 1.$$

Now $k \geq 1$ and $n_1 > k$, so $2n_1 > k + 1$, and it is proved.

Theorem 6.38. I_{Lorenz} satisfies P3.

Proof. We need to show that

$$\left(\frac{1}{m}\right)\left(\frac{n_1 m}{N} + \sum_{i=2}^{m}(m - i + 1)\left(\frac{1}{N}\right)\right) >$$
$$\left(\frac{1}{m+k}\right)\left(\frac{(n_1 - k)m}{N} + \sum_{i=2}^{m+k}(m - i + 1)\left(\frac{1}{N}\right)\right).$$

Since

$$\sum_{i=2}^{m}(m - i + 1)\left(\frac{1}{N}\right) = \frac{(m+1)m}{N} + \frac{3(m+1)}{2N} - \frac{(m+1)^2}{2N} - \frac{2m}{N} - \frac{1}{N}$$

and

$$\sum_{i=2}^{m+k}(m - i + 1)\left(\frac{1}{N}\right) = \frac{(m+k+1)m}{N} + \frac{3(m+k+1)}{2N} - \frac{(m+k+1)^2}{2N} - \frac{2m}{N} - \frac{1}{N},$$

substituting for the summation terms in the above inequality yields

$$\left(\frac{1}{m}\right)\left(\frac{n_1 m}{N}+\frac{(m+1)m}{N}+\frac{3(m+1)}{2N}-\frac{(m+1)^2}{2N}-\frac{2m}{N}-\frac{1}{N}\right)>$$

$$\left(\frac{1}{m+k}\right)\left(\frac{(n_1-k)m}{N}+\frac{(m+k+1)m}{N}+\frac{3(m+k+1)}{2N}-\frac{(m+k+1)^2}{2N}-\frac{2m}{N}-\frac{1}{N}\right).$$

Canceling terms yields

$$\frac{2n_1k+3mk+k^2}{2N(m+k)}>\frac{2k}{2N(m+k)}.$$

Factoring $k/(2N(m+k))$ out of both sides of the above inequality yields

$$2n_1+3m+k>2.$$

Now $n_1 \geq 2$, $m>1$, and $k>0$, so the above inequality is obviously true, and it is proved.

Theorem 6.39. I_{Berger} satisfies P3.

Proof. We need to show that

$$\frac{n_1}{N}>\frac{n_1-k}{N}.$$

Factoring $1/N$ out of both sides of the above inequality yields

$$n_1>n_1-k.$$

Canceling n_1 from both sides of the above inequality and moving k to the left side yields

$$k>0,$$

and it is proved.

Theorem 6.40. $I_{Kullback}$ satisfies P3.

Proof. We need to show that

$$log_2 m - \left(\left(\frac{n_1}{N}\right) \log_2 \left(\frac{mn_1}{N}\right) + \sum_{i=2}^{m} \left(\frac{1}{N}\right) \log_2 \left(\frac{m}{N}\right) \right) <$$
$$log_2 m - \left(\left(\frac{n_1 - c}{N}\right) \log_2 \left(\frac{m(n_1 - c)}{N}\right) + \sum_{i=2}^{m+c} \left(\frac{1}{N}\right) \log_2 \left(\frac{m}{N}\right) \right).$$

Canceling terms and multiplying both sides of the above inequality by -1 yields

$$\left(\frac{n_1}{N}\right) \log_2 \left(\frac{mn_1}{N}\right) + \sum_{i=2}^{m} \left(\frac{1}{N}\right) \log_2 \left(\frac{m}{N}\right) >$$
$$\left(\frac{n_1 - c}{N}\right) \log_2 \left(\frac{m(n_1 - c)}{N}\right) + \sum_{i=2}^{m+c} \left(\frac{1}{N}\right) \log_2 \left(\frac{m}{N}\right).$$

Factoring $1/N$ out of both sides of the above inequality yields

$$n_1 \log_2 \left(\frac{mn_1}{N}\right) + (m-1) \log_2 \left(\frac{m}{N}\right) > (n_1 - c) \log_2 \left(\frac{m(n_1 - c)}{N}\right) + (m + c - 1) \log_2 \left(\frac{m}{N}\right).$$

Canceling terms yields

$$n_1 \log_2 \left(\frac{mn_1}{N}\right) > (n_1 - c) \log_2 \left(\frac{m(n_1 - c)}{N}\right) + c \log_2 \left(\frac{m}{N}\right).$$

Using the difference and addition properties of logarithms yields

$$n_1 \log_2 m + n_1 \log_2 n_1 - n_1 \log_2 N >$$
$$n_1 \log_2 m - c \log_2 m + (n_1 - c) \log_2 (n_1 - c) - n_1 \log_2 N + c \log_2 N$$
$$+ c \log_2 m - c \log_2 N.$$

Canceling terms yields

$$n_1 \log_2 n_1 > (n_1 - c) \log_2(n_1 - c),$$

which is obviously true, and it is proved.

Theorem 6.41. I_{Lorenz}, I_{Gini}, I_{Schutz}, I_{Bray}, $I_{Whittaker}$, $I_{MacArthur}$, I_{Theil}, and $I_{Atkinson}$ do not satisfy P3.

Proof. Left as an exercise for the reader. Let $N = 4$, $m = 2$, $c = 1$, $n = (n_1, n_2) = (3, 1)$, and $n^{'} = (n_1^{'}, n_2^{'}, n_3^{'}) = (2, 1, 1)$.

6.4 Permutation Invariance Principle

Permutation Invariance Principle (P4). Given a vector $(n_1, \ldots, n_m)$ and any permutation $(i_1, \ldots, i_m)$ of $(1, \ldots, m)$, $f(n_i, \ldots, n_m) = f(n_{i_1}, \ldots, n_{i_m})$.

Theorem 6.42. All measures except the I_{Lorenz} measure satisfy P4.

Proof. That this is true is immediately obvious due to the fact that the I_{Lorenz} measure not only uses the value of the summation index for referencing the values contained in the *Count* attribute, it also uses the value of the summation index in the calculation of the values that it generates. In contrast, the other measures use the value of the summation or multiplication index exclusively for referencing the values contained in the *Count* attribute.

6.5 Transfer Principle

Transfer Principle (P5). Given a vector $(n_1, \ldots, n_m)$ and $0 < c < n_j$, $f(n_1, \ldots, n_i + c, \ldots, n_j - c, \ldots, n_m) > f(n_1, \ldots, n_i, \ldots, n_j, \ldots, n_m)$.

Theorem 6.43. $I_{Variance}$ satisfies P5.

Proof. We need to show that

$$\sum_{i=1}^{j-1}\left(\frac{n_i}{N}-\frac{1}{m}\right)^2+\left(\frac{n_j+c}{N}-\frac{1}{m}\right)^2+\sum_{i=j+1}^{k-1}\left(\frac{n_i}{N}-\frac{1}{m}\right)^2+\left(\frac{n_k-c}{N}-\frac{1}{m}\right)^2$$
$$+\sum_{i=k+1}^{m}\left(\frac{n_i}{N}-\frac{1}{m}\right)^2>$$
$$\sum_{i=1}^{j-1}\left(\frac{n_i}{N}-\frac{1}{m}\right)^2+\left(\frac{n_j}{N}-\frac{1}{m}\right)^2+\sum_{i=j+1}^{k-1}\left(\frac{n_i}{N}-\frac{1}{m}\right)^2+\left(\frac{n_k}{N}-\frac{1}{m}\right)^2$$
$$+\sum_{i=k+1}^{m}\left(\frac{n_i}{N}-\frac{1}{m}\right)^2,$$

which after canceling terms is equivalent to showing that

$$\left(\frac{n_j+c}{N}-\frac{1}{m}\right)^2+\left(\frac{n_k-c}{N}-\frac{1}{m}\right)^2>\left(\frac{n_j}{N}-\frac{1}{m}\right)^2+\left(\frac{n_k}{N}-\frac{1}{m}\right)^2.$$

Expanding both sides of the above inequality yields

$$\frac{n_j^2+2n_jc+2c^2+n_k^2-2n_kc}{N^2}-\frac{2(n_j+n_k)}{Nm}+\frac{2}{m^2}>\frac{n_j^2+n_k^2}{N^2}-\frac{2(n_j+n_k)}{Nm}+\frac{2}{m^2}.$$

Canceling terms yields

$$\frac{2(n_jc+c^2-n_kc)}{N^2}>0,$$

which is equivalent to

$$\frac{2(n_jc+c^2)}{N^2}>\frac{2n_kc}{N^2}.$$

Factoring $2c/N^2$ out of both sides of the above inequality yields

$$n_j+c>n_k.$$

Now $n_j>n_k$ and $c>0$, so the above inequality is obviously true, and it is proved.

Theorem 6.44. $I_{Simpson}$ satisfies P5.

Proof. We need to show that

$$\sum_{i=1}^{j-1}\left(\frac{n_i}{N}\right)^2+\left(\frac{n_j+c}{N}\right)^2+\sum_{i=j+1}^{k-1}\left(\frac{n_i}{N}\right)^2+\left(\frac{n_k-c}{N}\right)^2+\sum_{i=k+1}^{m}\left(\frac{n_i}{N}\right)^2>$$
$$\sum_{i=1}^{j-1}\left(\frac{n_i}{N}\right)^2+\left(\frac{n_j}{N}\right)^2+\sum_{i=j+1}^{k-1}\left(\frac{n_i}{N}\right)^2+\left(\frac{n_k}{N}\right)^2+\sum_{i=k+1}^{m}\left(\frac{n_i}{N}\right)^2,$$

which after canceling terms is equivalent to showing that

$$\left(\frac{n_j+c}{N}\right)^2+\left(\frac{n_k-c}{N}\right)^2>\left(\frac{n_j}{N}\right)^2+\left(\frac{n_k}{N}\right)^2.$$

Expanding both sides of the above inequality yields

$$\frac{n_j^2+2n_jc+c^2}{N^2}+\frac{n_k^2+2n_kc+c^2}{N^2}>\frac{n_j^2}{N^2}+\frac{n_k^2}{N^2}.$$

Canceling terms yields

$$\frac{2n_jc}{N^2}+\frac{2c^2}{N^2}-\frac{2n_kc}{N^2}>0,$$

which is equivalent to

$$\frac{2n_jc}{N^2}+\frac{2c^2}{N^2}>\frac{2n_kc}{N^2}.$$

Factoring $2c/N^2$ out of both sides of the above inequality yields

$$n_j+c>n_k.$$

Now $n_j > n_k$ and $c > 0$, so the above inequality is obviously true, and it is proved.

Theorem 6.45. $I_{Shannon}$ satisfies P5.

Proof. We need to show that

$$-\left(\sum_{i=1}^{j-1}\left(\frac{n_i}{N}\right)\log_2\left(\frac{n_i}{N}\right)+\left(\frac{n_j}{N}\right)\log_2\left(\frac{n_j}{N}\right)+\sum_{i=j+1}^{k-1}\left(\frac{n_i}{N}\right)\log_2\left(\frac{n_i}{N}\right)\right.$$
$$\left.+\left(\frac{n_k}{N}\right)\log_2\left(\frac{n_k}{N}\right)+\sum_{i=k+1}^{m}\left(\frac{n_i}{N}\right)\log_2\left(\frac{n_i}{N}\right)\right)>$$
$$-\left(\sum_{i=1}^{j-1}\left(\frac{n_i}{N}\right)\log_2\left(\frac{n_i}{N}\right)+\left(\frac{n_j+c}{N}\right)\log_2\left(\frac{n_j+c}{N}\right)+\sum_{i=j+1}^{k-1}\left(\frac{n_i}{N}\right)\log_2\left(\frac{n_i}{N}\right)\right.$$
$$\left.+\left(\frac{n_k-c}{N}\right)\log_2\left(\frac{n_k-c}{N}\right)+\sum_{i=k+1}^{m}\left(\frac{n_i}{N}\right)\log_2\left(\frac{n_i}{N}\right)\right).$$

Multiplying both sides of the above inequality by -1 and canceling terms yields

$$\left(\frac{n_j}{N}\right)\log_2\left(\frac{n_j}{N}\right)+\left(\frac{n_k}{N}\right)\log_2\left(\frac{n_k}{N}\right)<\left(\frac{n_j+c}{N}\right)\log_2\left(\frac{n_j+c}{N}\right)+\left(\frac{n_k-c}{N}\right)\log_2\left(\frac{n_k-c}{N}\right).$$

Factoring $1/N$ out of both sides of the above inequality yields

$$n_j\log_2\left(\frac{n_j}{N}\right)+n_k\log_2\left(\frac{n_k}{N}\right)<(n_j+c)\log_2\left(\frac{n_j+c}{N}\right)+(n_k-c)\log_2\left(\frac{n_k-c}{N}\right).$$

Using the difference property of logarithms to decompose the above inequality yields

$$n_j\log_2 n_j-n_j\log_2 N+n_k\log_2 n_k-n_k\log_2 N<$$
$$(n_j+c)\log_2(n_j+c)-(n_j+c)\log_2 N+(n_k-c)\log_2(n_k-c)-(n_k-c)\log_2 N.$$

Canceling terms yields

$$n_j\log_2 n_j+n_k\log_2 n_k<(n_j+c)\log_2(n_j+c)+(n_k-c)\log_2(n_k-c).$$

Since the function $x \log_2 x$ is a convex function, we have

$$(n_j + c)\log_2(n_j + c) + (n_k - c)\log_2(n_k - c) \geq (n_j + c + n_k - c)\log_2(n_j + c + n_k - c),$$

which is equivalent to

$$(n_j + c)\log_2(n_j + c) + (n_k - c)\log_2(n_k - c) \geq (n_j + n_k)\log_2(n_j + n_k).$$

We now show that

$$n_j \log_2 n_j + n_k \log_2 n_k < (n_j + n_k)\log_2(n_j + n_k).$$

Distributing $(n_j + n_k)$ over the log on the right side of the above inequality yields

$$n_j \log_2 n_j + n_k \log_2 n_k < n_j \log_2(n_j + n_k) + n_k \log_2(n_j + n_k),$$

which is obviously true, and it is proved.

Theorem 6.46. I_{Total} satisfies P5.

Proof. We need to show that

$$\begin{aligned}
&m\left(-\left(\sum_{i=1}^{j-1}\left(\frac{n_i}{N}\right)\log_2\left(\frac{n_i}{N}\right) + \left(\frac{n_j}{N}\right)\log_2\left(\frac{n_j}{N}\right) + \sum_{i=j+1}^{k-1}\left(\frac{n_i}{N}\right)\log_2\left(\frac{n_i}{N}\right)\right.\right.\\
&\left.\left.+\left(\frac{n_k}{N}\right)\log_2\left(\frac{n_k}{N}\right) + \sum_{i=k+1}^{m}\left(\frac{n_i}{N}\right)\log_2\left(\frac{n_i}{N}\right)\right)\right) > \\
&\quad m\left(-\left(\sum_{i=1}^{j-1}\left(\frac{n_i}{N}\right)\log_2\left(\frac{n_i}{N}\right) + \left(\frac{n_j + c}{N}\right)\log_2\left(\frac{n_j + c}{N}\right) + \sum_{i=j+1}^{k-1}\left(\frac{n_i}{N}\right)\log_2\left(\frac{n_i}{N}\right)\right.\right.\\
&\quad\left.\left.+\left(\frac{n_k - c}{N}\right)\log_2\left(\frac{n_k - c}{N}\right) + \sum_{i=k+1}^{m}\left(\frac{n_i}{N}\right)\log_2\left(\frac{n_i}{N}\right)\right)\right).
\end{aligned}$$

Canceling terms and multiplying both sides of the above inequality by -1 yields

$$\sum_{i=1}^{j-1}\left(\frac{n_i}{N}\right)\log_2\left(\frac{n_i}{N}\right)+\left(\frac{n_j}{N}\right)\log_2\left(\frac{n_j}{N}\right)+\sum_{i=j+1}^{k-1}\left(\frac{n_i}{N}\right)\log_2\left(\frac{n_i}{N}\right)$$
$$+\left(\frac{n_k}{N}\right)\log_2\left(\frac{n_k}{N}\right)+\sum_{i=k+1}^{m}\left(\frac{n_i}{N}\right)\log_2\left(\frac{n_i}{N}\right)<$$
$$\sum_{i=1}^{j-1}\left(\frac{n_i}{N}\right)\log_2\left(\frac{n_i}{N}\right)+\left(\frac{n_j+c}{N}\right)\log_2\left(\frac{n_j+c}{N}\right)+\sum_{i=j+1}^{k-1}\left(\frac{n_i}{N}\right)\log_2\left(\frac{n_i}{N}\right)$$
$$+\left(\frac{n_k-c}{N}\right)\log_2\left(\frac{n_k-c}{N}\right)+\sum_{i=k+1}^{m}\left(\frac{n_i}{N}\right)\log_2\left(\frac{n_i}{N}\right),$$

which is proved in Theorem 6.45.

Theorem 6.47. $I_{McIntosh}$ satisfies P5.

Proof. We need to show that

$$\frac{N-sqrt\sum_{i=1}^{j-1}n_i^2+(n_j+c)^2+\sum_{i=j+1}^{k-1}n_i^2+(n_k-c)^2+\sum_{i=k+1}^{m}n_i^2}{N-\sqrt{N}}<$$
$$\frac{N-sqrt\sum_{i=1}^{j-1}n_i^2+n_j^2+\sum_{i=j+1}^{k-1}n_i^2+n_k^2+\sum_{i=k+1}^{m}n_i^2}{N-\sqrt{N}}.$$

Canceling terms and multiplying both sides of the above inequality by -1 yields

$$(n_j+c)^2+(n_k-c)^2>n_j^2+n_k^2.$$

Expanding the terms in the left side yields

$$n_j^2+2n_jc+c^2+n_k^2-2n_kc+c^2>n_j^2+n_k^2.$$

Canceling terms yields and moving $-2n_kc$ to the right side yields

$$2n_jc+2c^2>2n_kc.$$

Factoring $2c$ out of both sides of the above inequality yields

$$n_j + c > n_k.$$

Now $n_j > n_k$ and $c > 0$, so the above inequality is obviously true, and it is proved.

Theorem 6.48. I_{Lorenz} satisfies P5.

Proof. We need to show that

$$\frac{1}{m}\left(\sum_{i=1}^{j-1}(m-i+1)\left(\frac{n_i}{N}\right)+(m-j+1)\left(\frac{n_j}{N}\right)+\sum_{i=j+1}^{k-1}(m-i+1)\left(\frac{n_i}{N}\right)\right.$$
$$\left.+(m-k+1)\left(\frac{n_k}{N}\right)+\sum_{i=k+1}^{m}(m-i+1)\left(\frac{n_i}{N}\right)\right)>$$
$$\frac{1}{m}\left(\sum_{i=1}^{j-1}(m-i+1)\left(\frac{n_i}{N}\right)+(m-j+1)\left(\frac{n_j-c}{N}\right)+\sum_{i=j+1}^{k-1}(m-i+1)\left(\frac{n_i}{N}\right)\right.$$
$$\left.+(m-k+1)\left(\frac{n_k+c}{N}\right)+\sum_{i=k+1}^{m}(m-i+1)\left(\frac{n_i}{N}\right)\right).$$

Canceling terms yields

$$(m-j+1)\left(\frac{n_j}{N}\right)+(m-k+1)\left(\frac{n_k}{N}\right)>(m-j+1)\left(\frac{n_j-c}{N}\right)+(m-k+1)\left(\frac{n_k+c}{N}\right).$$

Expanding terms on the right side of the above inequality yields

$$(m-j+1)\left(\frac{n_j}{N}\right)+(m-k+1)\left(\frac{n_k}{N}\right)>$$
$$(m-j+1)\left(\frac{n_j}{N}-(m-j+1)\frac{c}{N}\right)+(m-k+1)\left(\frac{n_k}{N}+(m-k+1)\frac{c}{N}\right).$$

Canceling terms yields

$$0>-(m-j+1)\left(\frac{c}{N}\right)+(m-k+1)\left(\frac{c}{N}\right).$$

Moving $-(m-j+1)\left(\frac{c}{N}\right)$ to the left side yields

$$(m-j+1)\left(\frac{c}{N}\right) > (m-k+1)\left(\frac{c}{N}\right).$$

Factoring c/N out of both sides of the above inequality and canceling terms yields

$$-j > -k$$

which is equivalent to

$$j < k,$$

which is obviously true, and it is proved.

Theorem 6.49. I_{Gini} satisfies P5.

Proof. To simplify the proof of this theorem, we again use the equivalent form of the Gini index introduced in the proof of Theorem 6.23. We need to show that

$$\frac{2\left(\frac{m+1}{2}-\left(\sum_{i=1}^{j-1} i\left(\frac{n_i}{N}\right)+j\left(\frac{n_j+c}{N}\right)+\sum_{i=j+1}^{k-1} i\left(\frac{n_i}{N}\right)+k\left(\frac{n_k-c}{N}\right)+\sum_{i=k+1}^{m} i\left(\frac{n_i}{N}\right)\right)\right)}{m} >$$

$$\frac{2\left(\frac{m+1}{2}-\left(\sum_{i=1}^{j-1} i\left(\frac{n_i}{N}\right)+j\left(\frac{n_j}{N}\right)+\sum_{i=j+1}^{k-1} i\left(\frac{n_i}{N}\right)+k\left(\frac{n_k}{N}\right)+\sum_{i=k+1}^{m} i\left(\frac{n_i}{N}\right)\right)\right)}{m}.$$

Canceling terms from both sides of the above inequality yields

$$-j\left(\frac{n_j+c}{N}\right)-k\left(\frac{n_k-c}{N}\right) > -j\left(\frac{n_j}{N}\right)-k\left(\frac{n_k}{N}\right),$$

which is equivalent to

$$-\frac{jc}{N}+\frac{kc}{N} > 0.$$

Factoring c/N out of the left side of the above inequality yields and moving $-j$ to the right side yields

$$k > j,$$

which is obviously true, and it is proved.

Theorem 6.50. $I_{Kullback}$ satisfies P5.

Proof. We need to show that

$$\begin{aligned}
&log_2 m - \left(\sum_{i=1}^{j-1} \left(\frac{n_i}{N}\right) \log_2\left(\frac{mn_i}{N}\right) + \left(\frac{n_j}{N}\right) \log_2\left(\frac{mn_j}{N}\right) + \sum_{i=j+1}^{k-1} \left(\frac{n_i}{N}\right) \log_2\left(\frac{mn_i}{N}\right) \right. \\
&\left. + \left(\frac{n_k}{N}\right) \log_2\left(\frac{mn_k}{N}\right) + \sum_{i=k+1}^{m} \left(\frac{n_i}{N}\right) \log_2\left(\frac{mn_i}{N}\right) \right) > \\
&\quad log_2 m - \left(\sum_{i=1}^{j-1} \left(\frac{n_i}{N}\right) \log_2\left(\frac{mn_i}{N}\right) + \left(\frac{n_j + c}{N}\right) \log_2\left(\frac{m(n_j + c)}{N}\right) \right) \\
&\quad + \sum_{i=j+1}^{k-1} \left(\frac{n_i}{N}\right) \log_2\left(\frac{mn_i}{N}\right) + \left(\frac{n_k - c}{N}\right) \log_2\left(\frac{m(n_k - c)}{N}\right) \\
&\quad + \sum_{i=k+1}^{m} \left(\frac{n_i}{N}\right) \log_2\left(\frac{mn_i}{N}\right).
\end{aligned}$$

Multiplying both sides of the above inequality by -1 and canceling terms yields

$$\begin{aligned}
&\left(\frac{n_j}{N}\right) \log_2\left(\frac{mn_j}{N}\right) + \left(\frac{n_k}{N}\right) \log_2\left(\frac{mn_k}{N}\right) < \\
&\quad \left(\frac{n_j + c}{N}\right) \log_2\left(\frac{m(n_j + c)}{N}\right) + \left(\frac{n_k - c}{N}\right) \log_2\left(\frac{m(n_k - c)}{N}\right).
\end{aligned}$$

Factoring $1/N$ out of both sides of the above inequality yields

$$\begin{aligned}
&n_j \log_2\left(\frac{mn_j}{N}\right) + n_k \log_2\left(\frac{mn_k}{N}\right) < \\
&\quad (n_j + c) \log_2\left(\frac{m(n_j + c)}{N}\right) + (n_k - c) \log_2\left(\frac{m(n_k - c)}{N}\right).
\end{aligned}$$

Using the difference and addition properties of logarithms to decompose the above inequality yields

$$n_j \log_2 m + n_j \log_2 n_j - n_j \log_2 N + n_k \log_2 m + n_k \log_2 n_k - n_k \log_2 N <$$
$$n_j \log_2 m + c \log_2 m + (n_j + c) \log_2 (n_j + c) - n_j \log_2 N - c \log_2 N + n_k \log_2 m$$
$$-c \log_2 m + (n_k - c) \log_2 (n_k - c) - n_k \log_2 N + c \log_2 N.$$

Canceling terms yields

$$n_j \log_2 n_j + n_k \log_2 n_k < (n_j + c) \log_2 (n_j + c) + (n_k - c) \log_2 (n_k - c),$$

which is proved in Theorem 6.45.

Theorem 6.51. $I_{MacArthur}$ satisfies P5.

Proof. Let the vector $(a, \ldots, a)$ containing m a's be the uniform distribution. Then the distribution obtained by combining the vectors $(n_1 + c, n_2 - d_2, \ldots, n_m - d_m)$ and $(a, \ldots, a)$ is the vector $(n_1 + c + a, n_2 - d_2 + a, \ldots, n_m - d_m + a)$, and the distribution obtained by combining the vectors $(n_1, \ldots, n_m)$ and $(a, \ldots, a)$ is the vector $(n_1 + a, \ldots, n_m + a)$. We need to show that

$$\begin{aligned}
&-\left(\sum_{i=1}^{j-1} \left(\frac{n_i + a}{2N}\right) \log_2 \left(\frac{n_i + a}{2N}\right) + \left(\frac{n_j + c + a}{2N}\right) \log_2 \left(\frac{n_j + c + a}{2N}\right)\right.\\
&+ \sum_{i=j+1}^{k-1} \left(\frac{n_i + a}{2N}\right) \log_2 \left(\frac{n_i + a}{2N}\right) + \left(\frac{n_k - c + a}{2N}\right) \log_2 \left(\frac{n_k - c + a}{2N}\right)\\
&\left.+ \sum_{i=k+1}^{m} \left(\frac{n_i + a}{2N}\right) \log_2 \left(\frac{n_i + a}{2N}\right)\right)\\
&-\left(-\left(\sum_{i=1}^{j-1} \left(\frac{n_i}{N}\right) \log_2 \left(\frac{n_i}{N}\right) + \left(\frac{n_j + c}{N}\right) \log_2 \left(\frac{n_j + c}{N}\right)\right.\right.\\
&+ \sum_{i=j+1}^{k-1} \left(\frac{n_i}{N}\right) \log_2 \left(\frac{n_i}{N}\right) + \left(\frac{n_k - c}{N}\right) \log_2 \left(\frac{n_k - c}{N}\right)\\
&\left.\left.+ \sum_{i=k+1}^{m} \left(\frac{n_i}{N}\right) \log_2 \left(\frac{n_i}{N}\right)\right) + \log_2 m\right) \Big/ 2 >
\end{aligned}$$

$$
\begin{aligned}
&-\left(\sum_{i=1}^{j-1}\left(\frac{n_i+a}{2N}\right)\log_2\left(\frac{n_i+a}{2N}\right)+\left(\frac{n_j+a}{2N}\right)\log_2\left(\frac{n_j+a}{2N}\right)\right.\\
&+\sum_{i=j+1}^{k-1}\left(\frac{n_i+a}{2N}\right)\log_2\left(\frac{n_i+a}{2N}\right)+\left(\frac{n_k+a}{2N}\right)\log_2\left(\frac{n_k+a}{2N}\right)\\
&\left.+\sum_{i=k+1}^{m}\left(\frac{n_i+a}{2N}\right)\log_2\left(\frac{n_i+a}{2N}\right)\right)\\
&-\left(-\left(\sum_{i=1}^{j-1}\left(\frac{n_i}{N}\right)\log_2\left(\frac{n_i}{N}\right)+\left(\frac{n_j}{N}\right)\log_2\left(\frac{n_j}{N}\right)\right.\right.\\
&+\sum_{i=j+1}^{k-1}\left(\frac{n_i}{N}\right)\log_2\left(\frac{n_i}{N}\right)+\left(\frac{n_k}{N}\right)\log_2\left(\frac{n_k}{N}\right)\\
&\left.\left.+\sum_{i=k+1}^{m}\left(\frac{n_i}{N}\right)\log_2\left(\frac{n_i}{N}\right)\right)+\log_2 m\right)\Big/ 2,
\end{aligned}
$$

which is equivalent to

$$
\begin{aligned}
&-\sum_{i=1}^{j-1}\left(\frac{n_i+a}{2N}\right)\log_2\left(\frac{n_i+a}{2N}\right)-\left(\frac{n_j+c+a}{2N}\right)\log_2\left(\frac{n_j+c+a}{2N}\right)\\
&-\sum_{i=j+1}^{k-1}\left(\frac{n_i+a}{2N}\right)\log_2\left(\frac{n_i+a}{2N}\right)-\left(\frac{n_k-c+a}{2N}\right)\log_2\left(\frac{n_k-c+a}{2N}\right)\\
&-\sum_{i=k+1}^{m}\left(\frac{n_i+a}{2N}\right)\log_2\left(\frac{n_i+a}{2N}\right)+\sum_{i=1}^{j-1}\left(\frac{n_i}{2N}\right)\log_2\left(\frac{n_i}{N}\right)+\left(\frac{n_j+c}{2N}\right)\log_2\left(\frac{n_j+c}{N}\right)\\
&+\sum_{i=j+1}^{k-1}\left(\frac{n_i}{2N}\right)\log_2\left(\frac{n_i}{N}\right)+\left(\frac{n_k-c}{2N}\right)\log_2\left(\frac{n_k-c}{N}\right)+\sum_{i=k+1}^{m}\left(\frac{n_i}{2N}\right)\log_2\left(\frac{n_i}{N}\right)-\frac{\log_2 m}{2}>\\
&-\sum_{i=1}^{j-1}\left(\frac{n_i+a}{2N}\right)\log_2\left(\frac{n_i+a}{2N}\right)-\left(\frac{n_j+a}{2N}\right)\log_2\left(\frac{n_j+a}{2N}\right)\\
&-\sum_{i=j+1}^{k-1}\left(\frac{n_i+a}{2N}\right)\log_2\left(\frac{n_i+a}{2N}\right)-\left(\frac{n_k+a}{2N}\right)\log_2\left(\frac{n_k+a}{2N}\right)\\
&-\sum_{i=k+1}^{m}\left(\frac{n_i+a}{2N}\right)\log_2\left(\frac{n_i+a}{2N}\right)+\sum_{i=1}^{j-1}\left(\frac{n_i}{2N}\right)\log_2\left(\frac{n_i}{N}\right)+\left(\frac{n_j}{2N}\right)\log_2\left(\frac{n_j}{N}\right)\\
&+\sum_{i=j+1}^{k-1}\left(\frac{n_i}{2N}\right)\log_2\left(\frac{n_i}{N}\right)+\left(\frac{n_k}{2N}\right)\log_2\left(\frac{n_k}{N}\right)+\sum_{i=k+1}^{m}\left(\frac{n_i}{2N}\right)\log_2\left(\frac{n_i}{N}\right)-\frac{\log_2 m}{2}.
\end{aligned}
$$

Canceling terms, factoring $1/2N$ out of both sides of the above inequality, and moving terms preceded by a minus sign to the other side of the inequality yields

$$
\begin{aligned}
&(n_j+c)\log_2\left(\frac{n_j+c}{N}\right)+(n_k-c)\log_2\left(\frac{n_k-c}{N}\right)+(n_j+a)\log_2\left(\frac{n_j+a}{2N}\right)\\
&+(n_k+a)\log_2\left(\frac{n_k+a}{2N}\right)>\\
&\quad(n_j+c+a)\log_2\left(\frac{n_j+c+a}{2N}\right)+(n_k-c+a)\log_2\left(\frac{n_k-c+a}{2N}\right)+n_j\log_2\left(\frac{n_j}{N}\right)\\
&\quad+n_k\log_2\left(\frac{n_k}{N}\right).
\end{aligned}
$$

Using the difference property of logarithms to decompose the above inequality yields

$$
\begin{aligned}
&(n_j+c)\log_2(n_j+c)-n_j\log_2 N-c\log_2 N+(n_k-c)\log_2(n_k-c)-n_k\log_2 N\\
&+c\log_2 N+(n_j+a)\log_2(n_j+a)-n_j\log_2 2N-a\log_2 2N+(n_k+a)\log_2(n_k+a)\\
&-n_k\log_2 2N-a\log_2 2N>\\
&\quad(n_j+c+a)\log_2(n_j+c+a)-n_j\log_2 2N-c\log_2 2N-a\log_2 2N\\
&\quad+(n_k-c+a)\log_2(n_k-c+a)-n_k\log_2 2N+c\log_2 2N-a\log_2 2N+n_j\log_2 n_j\\
&\quad-n_j\log_2 N+n_k\log_2 n_k-n_k\log_2 N.
\end{aligned}
$$

Canceling terms yields

$$(n_j+c)\log_2(n_j+c)+(n_k-c)\log_2(n_k-c)+(n_j+a)\log_2(n_j+a)+(n_k+a)\log_2(n_k+a)>$$

$$(n_j+c+a)\log_2(n_j+c+a)+(n_k-c+a)\log_2(n_k-c+a)+n_j\log_2 n_j+n_k\log_2 n_k.$$

To show that the above inequality is true, we need to show that

$$(n_j+c)\log_2(n_j+c)+(n_j+a)\log_2(n_j+a)>$$

$$(n_j+c+a)\log_2(n_j+c+a)+n_j\log_2 n_j$$

and

$$(n_k + a)\log_2(n_k + a) + (n_k - c)\log_2(n_k - c) >$$

$$n_k \log_2 n_k + (n_k - c + a)\log_2(n_k - c + a).$$

We now show that

$$(n_j + c)\log_2(n_j + c) + (n_j + a)\log_2(n_j + a) >$$

$$(n_j + c + a)\log_2(n_j + c + a) + n_j \log_2 n_j.$$

Since the function $x \log_2 x$ is a convex function, we have

$$(n_j + c)\log_2(n_j + c) + (n_j + a)\log_2(n_j + a) \geq (n_j + c + n_j + a)\log_2(n_j + c + n_j + a),$$

which is equivalent to

$$(n_j + c)\log_2(n_j + c) + (n_j + a)\log_2(n_j + a) \geq$$

$$n_j \log_2(n_j + c + n_j + a) + (n_j + c + a)\log_2(n_j + c + n_j + a),$$

so obviously

$$(n_j + c)\log_2(n_j + c) + (n_j + a)\log_2(n_j + a) > n_j \log_2 n_j + (n_j + c + a)\log_2(n_j + c + a),$$

and the first part is shown. We now show that

$$(n_k + a)\log_2(n_k + a) + (n_k - c)\log_2(n_k - c) >$$

$$n_k \log_2 n_k + (n_k - c + a)\log_2(n_k - c + a).$$

Since the function $x \log_2 x$ is a convex function, we have

$$(n_k + a)\log_2(n_k + a) + (n_k - c)\log_2(n_k - c) \geq (n_k + a + n_k - c)\log_2(n_k + a + n_k - c),$$

which is equivalent to

$$(n_j + c)\log_2(n_j + c) + (n_j + a)\log_2(n_j + a) \geq$$
$$n_k \log_2(n_k + a + n_k - c) + (n_k - c + a)\log_2(n_k + a + n_k - c),$$

so obviously

$$(n_j + c)\log_2(n_j + c) + (n_j + a)\log_2(n_j + a) > n_k \log_2 n_k + (n_k - c + a)\log_2(n_k - c + a),$$

the second part is shown, and it is proved.

Theorem 6.52. $I_{Atkinson}$ satisfies P5.

Proof. We need to show that

$$1 - \left(\prod_{i=1}^{j-1} \frac{\left(\frac{n_i}{N}\right)}{\left(\frac{1}{m}\right)} \frac{\left(\frac{n_j+c}{N}\right)}{\left(\frac{1}{m}\right)} \prod_{i=j+1}^{k-1} \frac{\left(\frac{n_i}{N}\right)}{\left(\frac{1}{m}\right)} \frac{\left(\frac{n_k-c}{N}\right)}{\left(\frac{1}{m}\right)} \prod_{i=k+1}^{m} \frac{\left(\frac{n_i}{N}\right)}{\left(\frac{1}{m}\right)} \right)^{\frac{1}{m}} >$$

$$1 - \left(\prod_{i=1}^{j-1} \frac{\left(\frac{n_i}{N}\right)}{\left(\frac{1}{m}\right)} \frac{\left(\frac{n_j}{N}\right)}{\left(\frac{1}{m}\right)} \prod_{i=j+1}^{k-1} \frac{\left(\frac{n_i}{N}\right)}{\left(\frac{1}{m}\right)} \frac{\left(\frac{n_k}{N}\right)}{\left(\frac{1}{m}\right)} \prod_{i=k+1}^{m} \frac{\left(\frac{n_i}{N}\right)}{\left(\frac{1}{m}\right)} \right)^{\frac{1}{m}}.$$

Canceling terms from both sides of the above inequality yields

$$-((n_j + c)(n_k - c)) > -n_j n_k,$$

Expanding the left side of the above inequality yields

$$-n_j n_k + n_j c - n_k c + c^2 > -n_j n_k.$$

Canceling terms and moving $-n_k c$ to the right side of the above inequality yields

$$n_j c + c^2 > n_k c.$$

Factoring c out of both sides of the above inequality yields

$$n_j + c > n_k.$$

Now $n_j > n_k$ and $c > 0$, so the above inequality is obviously true, and it is proved.

Theorem 6.53. I_{Max}, I_{Berger}, I_{Schutz}, I_{Bray}, $I_{Whittaker}$, and I_{Theil} do not satisfy P5.

Proof. Left as an exercise for the reader. Let $N = 7$, $m = 3$, $n = (n_1, n_2, n_3) = (4, 2, 1)$, and $n^{'} = (n_1^{'}, n_2^{'}, n_3^{'}) = (3, 2, 2)$.

7 Conclusion

The use of diversity measures for ranking the interestingness of summaries is a new and promising application area. Here we proposed five principles that interestingness measures should satisfy when used to rank the interestingness of summaries generated from databases. The theoretical results show six measures satisfy all of the proposed principles. These include $I_{Variance}$, $I_{Simpson}$, $I_{Shannon}$, I_{Total}, $I_{McIntosh}$, and $I_{Kullback}$. The nine remaining measures did not perform as well, failing to satisfy at least one of the proposed principles.

References

[1] R. Agrawal and R. Srikant. Fast algorithms for mining association rules. In *Proceedings of the 20th International Conference on Very Large Databases (VLDB'94)*, pages 487–499, Santiago, Chile, September 1994.

[2] A.B. Atkinson. On the measurement of inequality. *Journal of Economic Theory*, 2:244–263, 1970.

[3] R.J. Bayardo and R. Agrawal. Mining the most interesting rules. In *Proceedings of the Fifth International Conference on Knowledge Discovery and Data Mining (KDD'99)*, pages 145–154, San Diego, California, August 1999.

[4] H. Dalton. The measurement of the inequality of incomes. *Economic Journal*, 30:348–361, 1920.

[5] L. Egghe and R. Rousseau. Transfer principles and a classification of concentration measures. *Journal of the American Society for Information Science*, 42(7):479–489, 1991.

[6] J. Han, Y. Cai, and N. Cercone. Data-driven discovery of quantitative rules in relational databases. *IEEE Transactions on Knowledge and Data Engineering*, 5(1):29–40, February 1993.

[7] R.J. Hilderman and H.J. Hamilton. Heuristic measures of interestingness. In J. Zytkow and J. Rauch, editors, *Proceedings of the Third European Conference on the Principles of Data Mining and Knowledge Discovery (PKDD'99)*, pages 232–241, Prague, Czech Republic, September 1999.

[8] R.J. Hilderman and H.J. Hamilton. Heuristics for ranking the interestingness of discovered knowledge. In N. Zhong and L. Zhou, editors, *Proceedings of the Third Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD'99)*, pages 204–209, Beijing, China, April 1999.

[9] R.J. Hilderman and H.J. Hamilton. Knowledge discovery and interestingness measures: A survey. Technical Report CS 99-04, Department of Computer Science, University of Regina, October 1999.

[10] R.J. Hilderman, H.J. Hamilton, and N. Cercone. Data mining in large databases using domain generalization graphs. *Journal of Intelligent Information Systems*, 13(3), November 1999.

[11] M. Klemettinen, H. Mannila, P. Ronkainen, H. Toivonen, and A.I. Verkamo. Finding interesting rules from large sets of discovered association rules. In N.R. Adam, B.K. Bhargava, and Y. Yesha, editors, *Proceedings of the Third International Conference on Information and Knowledge Management*, pages 401–407, Gaitersburg, Maryland, 1994.

[12] S. Lieberson. An extension of Greenberg's linguistic diversity measures. *Language*, 40:526–531, 1964.

[13] A.E. Magurran. *Ecological diversity and its measurement.* Princeton University Press, 1988.

[14] B. Padmanabhan and A. Tuzhilin. A belief-driven method for discovering unexpected patterns. In *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining (KDD'98)*, pages 94–100, New York, New York, August 1998.

[15] D. Partridge and W. Krzanowski. Software diversity: practical statistics for its measurement and exploitation. *Information and Software Technology*, 39:707–717, 1997.

[16] A.D. Pratt. A measure of class concentration in bibliometrics. *Journal of the American Society for Information Science*, 28:285–292, 1977.

[17] D.J. Randall, H.J. Hamilton, and R.J. Hilderman. Temporal generalization with domain generalization graphs. *International Journal of Pattern Recognition and Artificial Intelligence*. To appear.

[18] S. Sahar. Interestingness via what is not interesting. In *Proceedings of the Fifth International Conference on Knowledge Discovery and Data Mining (KDD'99)*, pages 332–336, San Diego, California, August 1999.

[19] C.E. Shannon and W. Weaver. *The mathematical theory of communication*. University of Illinois Press, 1949.

[20] M.L. Weitzman. On diversity. *The Quarterly Journal of Economics*, pages 363–405, May 1992.