How to Gauge Inequality and Fairness: A Complete Description of All Decomposable Versions of Theil Index

Saeid Tizpaz-Niari, Olga Kosheleva, and Vladik Kreinovich

Abstract In general, in statistics, the most widely used way to describe the difference between different elements of a sample if by using standard deviation. This characteristic has a nice property of being decomposable: e.g., to compute the mean and standard deviation of the income overall the whole US, it is sufficient to compute the number of people, mean, and standard deviation over each state; this state-by-state information is sufficient to uniquely reconstruct the overall standard deviation. However, e.g., for gauging income inequality, standard deviation is not very adequate: it provides too much weight to outliers like billionaires, and thus, does not provide us with a good understanding of how unequal are incomes of the majority of folks. For this purpose, Theil introduced decomposable modifications of the standard deviation that is now called Theil indices. Crudely speaking, these indices are based on using logarithm instead of the square. Other researchers found other another decomposable modifications that use power law. In this paper, we provide a complete description of all decomposable versions of the Theil index. Specifically, we prove that the currently known functions are the only one for which the corresponding versions of the Theil index are decomposable – so no other decomposable versions are possible. A similar result was previously proven under the additional assumption of linearity; our proof shows that this result is also true in the general case, without assuming linearity.

Saeid Tizpaz-Niari
Department of Computer Science, University of Texas at El Paso, 500 W. University
El Paso, Texas 79968, USA, e-mail: saeid@utep.edu

Olga Kosheleva
Department of Teacher Education, University of Texas at El Paso, 500 W. University
El Paso, Texas 79968, USA, e-mail: olgak@utep.edu

Vladik Kreinovich
Department of Computer Science, University of Texas at El Paso, 500 W. University
El Paso, Texas 79968, USA, e-mail: vladik@utep.edu

1 Formulation of the Problem

Need to gauge inequality and fairness: a general problem. While more and more decisions are made using AI, it becomes more and more important to make sure that these decisions are fair, and that these decisions do not result in an increase in inequality. To be able to do that, we need to gauge inequality and fairness.

In general, if we have n people with incomes $x_1, \ldots, x_n$, then, based of these values, we can find the average income

$$\overline{x} \stackrel{\text{def}}{=} \frac{x_1 + \ldots + x_n}{n}.$$

How can we gauge the inequality, the fact that some incomes differ from the average: some are larger and some are smaller?

Statistics' answer to this question. In statistics, the usual way to measure the deviation from the mean is by using sample variance

$$V = \frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - \overline{x})^2$$

or its square root, standard deviation $\sigma \stackrel{\text{def}}{=} \sqrt{V}$; see, e.g., [2].

The standard deviation σ provides the *absolute* value of the deviation, in the same monetary units as the income itself. To get the *relative* values – e.g., in percents – we can divide σ by the mean $\overline{x}$. The resulting ratio is known as the *coefficient of variation*

$$CV \stackrel{\text{def}}{=} \frac{\sigma}{\overline{x}} = \frac{\sqrt{\frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - \overline{x})^2}}{\overline{x}}.$$

We can somewhat simplify this expression if we divide both the numerator and the denominator by the mean $\overline{x}$. Then we get the following simpler expression:

$$CV = \sqrt{\frac{1}{n} \cdot \sum_{i=1}^{n} (r_i - 1)^2},$$

where we denoted

$$r_i \stackrel{\text{def}}{=} \frac{x_i}{\overline{x}}.$$

Good news: standard deviation is decomposable. Often – e.g., after each census – we need to process a lot of information. Nowadays, it is possible – and done – to move the information about all 300 million people to a single location and to process this information. But even now, this is not computationally easy. In the past, it was not possible.

However, it was always possible to compute the mean and standard deviation for income – and for all other characteristics – since these characteristics are *decomposable* in the following sense.

Suppose that the sample $N = \{1, \ldots, n\}$ is divided into several disjoint subsamples $N = N_1 \cup \ldots \cup N_m$, so that $N_j \cap N_{j'} = \emptyset$ for all $j \neq j'$. Suppose that for each sub-sample N_j, we know the number of elements n_j and the mean

$$\overline{x}_j = \frac{1}{n_j} \cdot \sum_{i \in N_j} x_i.$$

Then this information is sufficient to compute the overall mean, there is no need to use the original values x_i. Indeed, the overall mean is the ratio of the overall sum and the overall size n of the sample. The overall size n of the sample can be obtained by adding the numbers of elements in each sub-sample: $n = n_1 + \ldots + n_m$. The sum of all x_i's can be obtained by adding the sums corresponding to all sub-samples:

$$\sum_{i=1}^{n} x_i = \sum_{j=1}^{m} \sum_{i \in N_j} x_i.$$

Each such sub-sample sum is equal to $n_j \cdot \overline{x}_j$. Thus, we have

$$\sum_{i=1}^{n} x_i = \sum_{j=1}^{m} n_j \cdot \overline{x}_j,$$

and, thus:

$$\overline{x} = \frac{n_1 \cdot \overline{x}_1 + \ldots + n_m \cdot \overline{x}_m}{n_1 + \ldots + n_m}.$$

Similarly, if for each sub-sample, we know the number of elements, the mean, and the variance over each sub-sample, then we can reconstruct the overall number of elements, mean, and the variance. Indeed, it is known that the variance can be represented in the equivalent form $V = M - (\overline{x})^2$, where M denotes the second sample moment

$$M \stackrel{\text{def}}{=} \sum_{i=1}^{n} x_i^2.$$

If we know the mean $\overline{x}$ and the variance V, we can therefore reconstruct the second moment as $M = V + (\overline{x})^2$. Similarly, if we know the variance V_j and mean $\overline{x}_j$ for each sub-sample, we can reconstruct each sub-sample second moment as $M_j = V_j + (\overline{x}_j)^2$. Similarly to the case of the mean, we can reconstruct the overall second moment by using the following formula:

$$M = \frac{n_1 \cdot M_1 + \ldots + n_m \cdot M_m}{n_1 + \ldots + n_m}.$$

Then, we can reconstruct V as $M + (\overline{x})^2$.

Because of the decomposability property, it is possible to easily process the census results as follows:

- for each town, we compute the 3 values – number of people, mean, and variance – over this town;
- then, for each state, we take the 3 values corresponding to each town from this state, and combine them into 3 values corresponding to the state;
- finally, we take the 3 values corresponding to each state, and combine them into the desired 3 values describing the overall census results.

Limitations of variance. The problem with variance – or, equivalently, with coefficient of variation – is that it places too much weight on huge differences $x_i - \bar{x}$. For example, for incomes, it provides a good understanding of how many billionaires and multi-millionaires we have, but their contribution hides an important information of how equal or how unequal are the incomes of the vast majority of people.

It is therefore desirable to come up with decomposable measures that are less dependent on the some outliers.

Theil index: original forms. Computation of the coefficient of variation is based on computing the sample mean and the value

$$I \stackrel{\text{def}}{=} \frac{1}{n} \cdot \sum_{i=1}^{n} f(r_i), \tag{1}$$

for a smooth function $f(r) = (r-1)^2$. A natural idea is to use the same formula (1), but with some other smooth – e.g., twice differentiable – function $f(r)$.

In the 1960s, a Dutch econometrician Henri Theil noticed that if we choose $f(r) = \ln(r)$ or $f(r) = r \cdot \ln(r)$, then we still get decomposable indices: in the sense that:

- if we know the number of elements n_j, mean $\bar{x}_j$, and the value I_j of the index (1) for each sub-sample,
- then, based on this information, we can uniquely reconstruct the number of elements n, the mean $\bar{x}$, and the index I corresponding the whole sample; see, e.g., [5].

Other decomposable versions of the Theil index. Other decomposable versions of the Theil index were described in the paper [4]: they correspond to $f(r) = r^\alpha$ for any a (or, to be more precise, to $f(r) = r^\alpha - 1$).

Natural question. A natural question is: to provide a complete description of all smooth functions $f(r)$ for which the index (1) is decomposable.

What is known. A partial answer to this question was given in [3] – under an additional condition that the formula describing of the index of the sample in terms of indices of sub-samples is linear. It turns out that, under this assumption, the already proposed functions $\ln(r)$, $r \cdot \ln(r)$, and r^a are the only ones for which the corresponding index is decomposable.

The remaining question. The remaining question is: what will happen in the general case, when we do not make this linearity assumption?

What we do in this paper. In this paper, we provide a complete description of functions $f(r)$ for which the index (1) is decomposable – without making an additional assumption that decomposability is described by a linear formula. It turns out that even without this assumption, no other cases of decomposability appear. In other words, the already proposed functions $\ln(r)$, $r \cdot \ln(r)$, and r^a are the only ones for which the corresponding index is decomposable.

To summarize:

- it is known that these functions lead to a decomposable index;
- we prove that, vice versa, if a function leads to a decomposable index, then this function is of one of the three above types – and thus, no other function leads to a decomposable index.

2 Main result

Definition 1. *We say that a twice differentiable function $f(r)$ is* decomposable *if the following property holds: For each tuple $x_1, \ldots, x_n$ of positive real numbers, and for each subdivision of the set $N = \{1, \ldots, n\}$ into disjoint subsets $N = N_1 \cup \ldots \cup N_m$ (i.e., subsets for which $N_j \cap N_{j'} = \emptyset$ for all $j \neq j'$):*

- *if for each subset j, we know the number of elements n_j in this subset, the mean $\overline{x}_j$ over this subset, and the corresponding value I_j of the index (1):*

$$I_j = \frac{1}{n_j} \cdot \sum_{i \in N_j} f\left(\frac{x_i}{\overline{x}_j}\right),$$

- *then, based on this information, we can uniquely reconstruct the overall number of elements n, the overall mean $\overline{x}$, and the overall value I of the index (1):*

$$I = \frac{1}{n} \cdot \sum_{i=1}^{n} f\left(\frac{x_i}{\overline{x}}\right).$$

Our main result is to describe all decomposable functions. To formulate our result, we need the following auxiliary definition.

Definition 2. *We say that two functions $f(r)$ and $g(r)$ are* equivalent *if there exist real numbers $a \neq 0$, b and c for which, for all r, we have $g(r) = a \cdot f(r) + b \cdot x + c$.*

Comment. If two functions are equivalent, then the corresponding indices $I(g)$ and $I(f)$ are related as follows:

$$I(g) = \frac{1}{n} \cdot \sum_{i=1}^{n} g(r_i) = a \cdot \frac{1}{n} \cdot \sum_{i=1}^{n} f(r_i) + b \cdot \sum_{i=1}^{n} r_i + c \cdot \sum_{i=1}^{n} 1,$$

where we denoted

$$r_i \stackrel{\text{def}}{=} \frac{x_i}{\overline{x}}.$$

Here,

$$\sum_{i=1}^{n} r_i = \sum_{i=1}^{n} \frac{x_i}{\overline{x}} = \frac{1}{\overline{x}} \cdot \sum_{i=1}^{n} x_i = \frac{x_1 + \ldots + x_n}{\dfrac{x_1 + \ldots + x_n}{n}} = n,$$

and

$$\sum_{i=1}^{n} 1 = n.$$

Thus,

$$I_g = a \cdot I_f + b \cdot n + c \cdot n.$$

So, if we know I_f, we can uniquely reconstruct I_g, and vice versa. In this sense, the indices corresponding to functions $f(r)$ and $g(r)$ are indeed equivalent.

Now, we are ready to formulate our main result.

Theorem. *A function $f(r)$ is decomposable if and only if it is equivalent to one of the following functions:* $\ln(r)$*,* $r \cdot \ln(r)$*, and* r^{α} *for some* α*.*

3 Proof

1°. It is known that functions from all the three types are decomposable. So, to complete our proof, we need to prove that if a function $f(r)$ is decomposable, then it is indeed equivalent to one of the functions listed in the formulation of out theorem.

Comment. In this proof, we use several mathematical ideas from [1].

2°. If a subset N_j consists of a single element x, then its number of elements is $n_j = 1$, its mean is $\overline{x}_j = x$, and – since here $x_i = \overline{x}$ – its index (1) is equal to $f(1)$.

3°. Let us pick some natural number $n \geq 4$, and let us consider a $(x_1, \ldots, x_n, x)$ tuple consisting of $n+1$ positive numbers for which the sum of the first n elements is equal to n. For this tuple, the mean value is equal to

$$\overline{x} = \frac{x_1 + \ldots + x_n + x}{n+1} = \frac{n+x}{n+1}. \tag{2}$$

Let us denote

$$\lambda \stackrel{\text{def}}{=} \frac{1}{\overline{x}} = \frac{n+1}{n+x}. \tag{3}$$

Then, by definition of λ, we have $r_i = \lambda \cdot x_i$ and thus, the index (1) is equal to

$$I = \frac{1}{n+1} \cdot \left(\sum_{i=1}^{n} f(\lambda \cdot x_i) + f(\lambda \cdot x) \right). \tag{4}$$

Let us consider a subdivision of this tuple into two subsets: $N_1 = \{1,\ldots,n\}$ and $N_2 = \{n+1\}$. For the subset N_1, the mean $\bar{x}_1$ is equal to 1, so we have $r_i = x_i$ and thus, the index (1) is described by the formula

$$I_1 = \frac{1}{n}\cdot\sum_{i=1}^{n} f(x_i). \tag{5}$$

For the subset N_2, according to Part 2 of this proof, the mean is $\bar{x}_2 = x$ and the index is $I_2 = f(1)$.

For this subdivision, decomposability means that:

- the value I – as described by the formula (4) – and thus, the value

$$S_\lambda \stackrel{\text{def}}{=} \sum_{i=1}^{n} f(\lambda\cdot x_i), \tag{6}$$

  which is equal to $(n+1)\cdot I - f(\lambda\cdot x)$,
- is uniquely determined if we know I_1 and the mean $\bar{x}_1$ – i.e., equivalently, the sum

$$S_1 \stackrel{\text{def}}{=} \sum_{i=1}^{n} f(x_i), \tag{7}$$

  which is equal to $n\cdot I_1$, and if we know the sum $x_1+\ldots+x_n = n$.

In other words, if for some other tuple $y_1,\ldots,y_n$, we have the same values of the sum S_1 and of the sum of elements, we should have the same value of the quantity S_λ. So, the following property must be satisfied:

- if we have

$$f(y_1)+\ldots+f(y_n) = f(x_1)+\ldots+f(x_n) \tag{8}$$

  and

$$y_1+\ldots+y_n = x_1+\ldots+x_n, \tag{9}$$

- then we should have

$$f(\lambda\cdot y_1)+\ldots+f(\lambda\cdot y_n) = f(\lambda\cdot x_1)+\ldots+f(\lambda\cdot x_n). \tag{10}$$

This property must hold for any λ that can be obtained by formula (3). For every positive $\lambda < 1+1/n$, we can find an $x>0$ for which the formula (3) holds: namely, we can take

$$x = \frac{n+1}{\lambda} - n.$$

Thus, the above property must be satisfied for all λ between 0 and $1+1/n$. Let us use this property to describe the function $f(r)$.

4°. Let us consider the values y_i that are close to x_i, i.e., for which $y_i = x_i + s\cdot d_i$ for some small s. For such y_i, we have

$$f(y_i) = f(x_i+s\cdot d_i) = f(x_i)+f'(x_i)\cdot s\cdot d_i + o(s).$$

Thus, the equality (8) takes the form

$$f'(x_1)\cdot s\cdot d_1+\ldots+f'(x_n)\cdot s\cdot d_n+o(s)=0.$$

If we divide both sides by s, by get an equivalent equality

$$f'(x_1)\cdot d_1+\ldots+f'(x_n)\cdot d_n+o(1)=0. \tag{11}$$

Similarly, the equalities (9) and (10) take the form

$$d_1+\ldots+d_n+o(1)=0, \tag{12}$$

and

$$f'(\lambda\cdot x_1)\cdot d_1+\ldots+f'(\lambda\cdot x_n)\cdot d_n+o(1)=0. \tag{13}$$

So, in the limit $s\to 0$, we conclude that for each vector $d=(d_1,\ldots,d_n)$, if we have

$$f'(x_1)\cdot d_1+\ldots+f'(x_n)\cdot d_n=0 \tag{14}$$

and

$$d_1+\ldots+d_n=0, \tag{15}$$

then we should have

$$f'(\lambda\cdot x_1)\cdot d_1+\ldots+f'(\lambda\cdot x_n)\cdot d_n=0. \tag{16}$$

Each of the expressions (14)-(16) is a scalar (dot) product of the vector d with, correspondingly, vectors $a\stackrel{\text{def}}{=}(f'(x_1),\ldots,f'(x_n))$, $e\stackrel{\text{def}}{=}(1,\ldots,1)$, and

$$b\stackrel{\text{def}}{=}(f'(\lambda\cdot x_1),\ldots,f'(\lambda\cdot x_n)).$$

The scalar product of two vectors is 0 if and only if these two vectors are orthogonal. Thus, we conclude that every vector d which is orthogonal to both a and e is also orthogonal to b. This means that the vector b must belong to the plane generated by a and e – otherwise, we can decompose b it into its projection $\text{pr}_{a,e}(b)$ to this plane and the remainder $R=b-\text{pr}_{a,e}(b)$ which is orthogonal to this plane. This remainder is orthogonal to a and e but not to b – which would contradict to the property described in the beginning of this paragraph.

The fact that b belongs to the plane generated by the vector a and b means that, for some real numbers A and B, we have $b=A\cdot a+B\cdot e$, i.e., in terms of vector components, that

$$f'(\lambda\cdot x_i)=A\cdot f'(x_i)+B. \tag{17}$$

5°. In general, the coefficient A and B depend on λ and on the initial tuple $(x_1,\ldots,x_n)$. So, strictly speaking, we should write

$$f'(\lambda\cdot x_i)=A(\lambda,x_1,\ldots,x_n)\cdot f'(x_i)+B(\lambda,x_1,\ldots,x_n). \tag{18}$$

Let us show that A and B do not depend on x_1. Indeed, for $i = 2$ and $i = 3$, we get

$$f'(\lambda \cdot x_2) = A(\lambda, x_1, \ldots, x_n) \cdot f'(x_2) + B(\lambda, x_1, \ldots, x_n); \tag{19}$$

$$f'(\lambda \cdot x_3) = A(\lambda, x_1, \ldots, x_n) \cdot f'(x_3) + B(\lambda, x_1, \ldots, x_n). \tag{20}$$

Subtracting (20) from (19) and dividing both sides of the resulting equation by the coefficient at $A(\lambda, x_1, \ldots, x_n)$, we conclude that

$$A(\lambda, x_1, \ldots, x_n) = \frac{f'(\lambda \cdot x_2) - f'(\lambda \cdot x_3)}{f'(x_2) - f'(x_3)}. \tag{21}$$

So, unless the derivative is a constant – and thus, the function $f(x)$ is linear – the right-hand side depends only on λ, x_2, and x_3 (and does not depend on x_1) and thus, the left-hand side – i.e., the value $A(\lambda, x_1, \ldots, x_n)$ – should also only depend on x_2 and x_3. So, we will write $A(\lambda, x_1, \ldots, x_n) = A(\lambda, x_2, x_3)$.

From (19), we can now conclude that

$$B(\lambda, x_1, \ldots, x_n) = f'(\lambda \cdot x_2) - A(\lambda, x_2, x_3) \cdot f'(x_2). \tag{22}$$

Here too, the right-hand side does not depend on x_1, so the left-hand side – which is $B(\lambda, x_1, \ldots, x_n)$ – also should not depend on x_1. Thus, indeed, neither A nor B depend on x_1. Similarly, we can conclude that A and B cannot depend on x_2, on x_3, etc. – i.e., that A and B only depend on λ. So, the formula (18) takes the following form:

$$F(\lambda \cdot r) = A(\lambda) \cdot F(r) + B(\lambda), \tag{23}$$

where we denoted $F(r) \stackrel{\text{def}}{=} f'(r)$, and the formulas (21) and (22) take the form

$$A(\lambda) = \frac{f'(\lambda \cdot x_2) - f'(\lambda \cdot x_3)}{f'(x_2) - f'(x_3)} \tag{24}$$

and

$$B(\lambda) = f'(\lambda \cdot x_2) - A(\lambda) \cdot f'(x_2). \tag{25}$$

6°. We assumed that the function $f(r)$ is twice differentiable. Thus, its derivative $F(r) = f'(r)$ is differentiable and so, due to formulas (24) and (25), the functions $A(\lambda)$ and $B(\lambda)$ are differentiable too. So, we can differentiate both sides of (23) by λ, and get $x \cdot F'(\lambda \cdot r) = A'(\lambda) \cdot F(r) + B'(\lambda)$. In particular, for $\lambda = 1$, we get

$$r \cdot \frac{dF}{dr} = a_0 \cdot F + b_0.$$

We can separate the variables if we multiply both sides by dr and divide both sides by r and $a_0 \cdot F + b_0$. Then, we get

$$\frac{dF}{a_0 \cdot F + b_0} = \frac{dr}{r}. \tag{26}$$

7°. If $a_0 = 0$, then, integrating both sides, we get $b_0^{-1} \cdot F = \ln(r) + C$, where C is an integration constant, i.e., $F(x) = f'(r) = b_0 \cdot \ln(r) + b_0 \cdot C$. Integrating again, we conclude that $f(r) = b_0 \cdot r \cdot \ln(r) + \text{const} \cdot x + \text{const}$, i.e., that $f(r)$ is equivalent to $r \cdot \ln(r)$.

8°. If $a_0 \neq 0$, then for $G \stackrel{\text{def}}{=} F + b_0/a_0$, we get

$$\frac{dG}{a_0 \cdot G} = \frac{dr}{r}.$$

Integrating both sides, we get $a_0^{-1} \cdot \ln(G) = \ln(r) + C$, so $\ln(G) = a_0 \cdot \ln(r) + a_0 \cdot C$. Applying $\exp(x)$ to both sides, we conclude that $G(r) = \text{const} \cdot r^{a_0}$, thus $F(r) = G(r) - b_0/a_0$ has the form $f'(r) = F(r) = \text{const} \cdot r^{a_0} + \text{const}$.

To get $f(r)$, we need to integrate one more time. When $a_0 = -1$, we get $f(r) = \text{const} \cdot \ln(r) + \text{const} \cdot r + \text{const}$, i.e., we get a function equivalent to $\ln(r)$. When $a_0 \neq -1$, we get $f(r) = \text{const} \cdot r^{a_0+1} + \text{const} \cdot x + \text{const}$, i.e., we get a function equivalent to r^α for $\alpha = a_0 + 1$.

The theorem is proven.

Acknowledgments

This work was supported in part by the National Science Foundation grants 1623190 (A Model of Change for Preparing a New Generation for Professional Practice in Computer Science), HRD-1834620 and HRD-2034030 (CAHSI Includes), EAR-2225395 (Center for Collective Impact in Earthquake Science C-CIES), and by the AT&T Fellowship in Information Technology.

It was also supported by a grant from the Hungarian National Research, Development and Innovation Office (NRDI).

References

1. O. Kosheleva, "Symmetry-group justification of maximum entropy method and generalized maximum entropy methods in image processing", In: G. J. Erickson, J. T. Rychert, and C. R. Smith (eds.), *Maximum Entropy and Bayesian Methods*, Kluwer, Dordrecht, 1998, pp. 101–113.
2. D. J. Sheskin, *Handbook of Parametric and Nonparametric Statistical Procedures*, Chapman and Hall/CRC, Boca Raton, Florida, 2011.
3. A. F. Shorrocks, "The class of additively decomposable inequality measures", *Econometrica*, 1980, Vol. 48, No. 3, pp. 613–625.

4. T. Speicher, H. Heidari, N. Grgic-Hlaca, K. P. Gummadi, A. Singla, A. Weller, and M. B. Zafar, "A unified approach to quantifying algorithmic unfairness: measuring individual & group unfairness *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data KDD 18*, London, UK, August 19–23, 2018, pp. 2239–2248, https://doi.org/10.1145/3219819.3220046
5. H. Theil, *Economics and Information Theory*, North Holland, Amsterdam, 1967.