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

Saeid Tizpaz-Niari, Olga Kosheleva, and Vladik Kreinovich

2Olga Kosheleva and Vladik Kreinovich
University of Texas at El Paso, 500 W. University
El Paso, Texas 79968, USA
saeid@utep.edu, olgak@utep.edu, vladik@utep.edu
1. Need to gauge inequality and fairness: a general problem

- More and more decisions are made using AI.
- So, 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

$$
\overline{x} \overset{\text{def}}{=} \frac{x_1 + \cdots + x_n}{n}.
$$

- How can we gauge the inequality, the fact that some incomes differ from the average: some are larger and some are smaller?
2. **Statistics’ answer to this question**

- In statistics, the usual way to measure the deviation from the mean is by using sample variance
  \[ V = \frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - \bar{x})^2. \]

- Practitioners also use variance’s square root, standard deviation
  \[ \sigma \overset{\text{def}}{=} \sqrt{V}. \]

- The standard deviation \( \sigma \) 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 \( \sigma \) by the mean \( \bar{x} \).

- The resulting ratio is known as the *coefficient of variation*
  \[ CV \overset{\text{def}}{=} \frac{\sigma}{\bar{x}} = \sqrt{\frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - \bar{x})^2}. \]
3. Statistics’ answer to this question (cont-d)

- 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:

$$CV = \sqrt{\frac{1}{n} \cdot \sum_{i=1}^{n} (r_i - 1)^2}, \text{ where we denoted } r_i \overset{\text{def}}{=} \frac{x_i}{\overline{x}}.$$
4. **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 being 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 the standard deviation for income – and for all other characteristics.
- The reason is that these characteristics are *decomposable* in the following sense.
5. **Good news: standard deviation is decomposable (cont-d)**

- Suppose that the sample $N = \{1, \ldots, n\}$ is divided into several disjoint sub-samples $N = N_1 \cup \ldots \cup N_m$, so that
  \[ N_j \cap N_{j'} = \emptyset \text{ for all } j \neq j'. \]
- Suppose that for each sub-sample $N_j$, we know the number of elements $n_j$ and the mean $\bar{x}_j = \frac{1}{n_j} \cdot \sum_{i \in N_j} x_i$.
- Then this information is sufficient to compute the overall mean, there is no need to also use the original values $x_i$. 

6. Good news: standard deviation is decomposable (cont-d)

- 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: \( n = n_1 + \ldots + n_m \).
7. Good news: standard deviation is decomposable (cont-d)

- 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 $n_j \cdot \bar{x}_j$.

- Thus, we have
  \[ \sum_{i=1}^{n} x_i = \sum_{j=1}^{m} n_j \cdot \bar{x}_j. \]

- So,
  \[ \bar{x} = \frac{n_1 \cdot \bar{x}_1 + \ldots + n_m \cdot \bar{x}_m}{n_1 + \ldots + n_m}. \]

- 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.
8. Good news: standard deviation is decomposable (cont-d)

- Indeed, it is known that the variance can be represented in the equivalent form $V = M - (\bar{x})^2$, where $M$ denotes the second sample moment

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

- If we know the mean $\bar{x}$ and the variance $V$, we can therefore reconstruct the second moment as $M = V + (\bar{x})^2$.

- Similarly, if we know the variance $V_j$ and mean $\bar{x}_j$ for each sub-sample, we can reconstruct each sub-sample second moment as

$$M_j = V_j + (\bar{x}_j)^2.$$ 

- Similarly to the case of the mean, we can reconstruct the overall second moment by using the following formula:

$$M = \frac{n_1 \cdot M_1 + \ldots + n_m \cdot M_m}{n_1 + \ldots + n_m}.$$
9. **Good news: standard deviation is decomposable (cont-d)**

- Then, we can reconstruct $V$ as $M - (\bar{x})^2$.
- Because of the decomposability property, it is possible to easily process the census results as follows.
- For each town, we compute the three values – the number of people, the mean, and the variance – over this town.
- Then, for each state:
  - we take the three values corresponding to each town from this state, and
  - we combine them into the three values corresponding to the state.
- Finally:
  - we take the three values corresponding to each state, and
  - we combine them into the desired three values describing the overall census results.
10. 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.

- However, 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.
11. Theil index: original forms

- Computation of the coefficient of variation is based on computing the sample mean and the value \( I \overset{\text{def}}{=} \frac{1}{n} \cdot \sum_{i=1}^{n} f(r_i) \).

- For \( CV \), we use a smooth function \( f(r) = (r - 1)^2 \).

- A natural idea is to use the same formula, 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.
12. Theil index: original forms (cont-d)

- Namely, if we know the number of elements $n_j$, the mean $\bar{x}_j$, and the value $I_j$ of the index 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.

- Similar formulas:
  - with a properly defined individual discrimination instead of the income,
  - have been successfully used to detect (and debug) fairness defects in socioeconomic applications of deep neural networks.
13. Other decomposable versions of the Theil index

- Other decomposable versions of the Theil index were found: they correspond to \( f(r) = r^\alpha - 1 \) for any \( \alpha \).

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

- A partial answer to this question was given in 1980 – 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 \( r^\alpha \) are the only ones for which the corresponding index is decomposable.
14. Other decomposable versions of the Theil index (cont-d)

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

- In this talk:
  - we provide a complete description of functions $f(r)$ for which the index $I$ 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.
15. Other decomposable versions of the Theil index (cont-d)

- In other words, the already proposed functions $\ln(r)$, $r \cdot \ln(r)$, and $r^\alpha$ 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;
  - thus, no other function leads to a decomposable index.
16. Main definition

- 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 $N = N_1 \cup \ldots \cup N_m$.

- If for each subset $j$, we know the number of elements $n_j$ in this subset,
  - the mean $\bar{x}_j$ over this subset, and
  - the corresponding value $I_j$ of the index $I$: $I_j = \frac{1}{n_j} \cdot \sum_{i \in N_j} f \left( \frac{x_i}{\bar{x}_j} \right)$,

- then, based on this information, we can uniquely reconstruct:
  - the overall number of elements $n$, the overall mean $\bar{x}$, and
  - the overall value $I$ of the index $I$: $I = \frac{1}{n} \cdot \sum_{i=1}^{n} f \left( \frac{x_i}{\bar{x}} \right)$.
17. Main result

- Our main result is to describe all decomposable functions.
- To formulate our result, we need the following auxiliary definition.
- 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
  \[ g(r) = a \cdot f(r) + b \cdot x + c. \]
- If two functions are equivalent, then the corresponding indices $I(g)$ and $I(f)$ are related as follows:
  \[ I(g) = \frac{1}{n} \cdot \sum_{i=1}^{n} g(r_i) = a \cdot \frac{1}{n} \cdot \sum_{i=1}^{n} f(r_i) + b \cdot \sum_{i=1}^{n} r_i + c \cdot \sum_{i=1}^{n} 1. \]
- Here, we denoted $r_i \overset{\text{def}}{=} \frac{x_i}{x}$. 
18. Main result (cont-d)

- Here, \( \sum_{i=1}^{n} 1 = n \) and
  \[
  \sum_{i=1}^{n} r_i = \sum_{i=1}^{n} \frac{x_i}{\bar{x}} = \frac{1}{\bar{x}} \cdot \sum_{i=1}^{n} x_i = \frac{x_1 + \ldots + x_n}{\frac{x_1 + \ldots + x_n}{n}} = n.
  \]

- 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.

- 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 \( r^\alpha \) for some \( \alpha \).
19. Proof

- 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 our theorem.
- If a subset $N_j$ consists of a single element $x$, then:
  - its number of elements is $n_j = 1$,
  - its mean is $\bar{x}_j = x$,
  - and - since here $x_i = \bar{x}$ - its index $I$ is equal to $f(1)$.
- Let us pick some natural number $n \geq 4$.
- Let us consider a tuple $(x_1, \ldots, x_n, x)$ consisting of $n + 1$ positive numbers for which the sum of the first $n$ elements is equal to $n$. 
20. Proof (cont-d)

- For this tuple, the mean value is equal to
  \[
  \overline{x} = \frac{x_1 + \ldots + x_n + x}{n + 1} = \frac{n + x}{n + 1}.
  \]

- Let us denote \( \lambda \overset{\text{def}}{=} \frac{1}{\overline{x}} = \frac{n + 1}{n + x} \).

- Then, by definition of \( \lambda \), we have \( r_i = \lambda \cdot x_i \) and thus, the index \( I \) is equal to
  \[
  I = \frac{1}{n + 1} \cdot \left( \sum_{i=1}^{n} f(\lambda \cdot x_i) + f(\lambda \cdot x) \right).
  \]

- 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 \( \overline{x}_1 \) is equal to 1, so we have \( r_i = x_i \) and thus, the index \( I \) is described by the formula
  \[
  I_1 = \frac{1}{n} \cdot \sum_{i=1}^{n} f(x_i).
  \]

- For the subset \( N_2 \), as we showed, the mean is \( \overline{x}_2 = x \) and the index is \( I_2 = f(1) \).
21. Proof (cont-d)

- For this subdivision, decomposability means that:
  - the value $I$ is uniquely determined
  - if we know $I_1$ the sample size $n$, the mean $\bar{x}_1$, and the values $\lambda$ and $x$.

- The value $S_{\lambda} \overset{\text{def}}{=} \sum_{i=1}^{n} f(\lambda \cdot x_i)$ is equal to $(n + 1) \cdot I - f(\lambda \cdot x)$.

- Since $x$ is uniquely determines by $\lambda$, we can conclude that the value $S_{\lambda}$ is also uniquely determined by values $\lambda$, $I_1$, $n$, and $\bar{x}_1$.

- One can check that, once we know $n$, then knowing $I_1$ is equivalent to knowing $S_1 \overset{\text{def}}{=} \sum_{i=1}^{n} f(x_i)$, since $S_1$ is equal to $n \cdot I_1$.

- Similarly, once we know $n$, knowing $\bar{x}_1$ is equivalent to knowing the sum $x_1 + \ldots + x_n = n$. 
22. Proof (cont-d)

- In other words, decomposability means that 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 $n$ elements $y_i$,
  - we should have the same value of the quantity $S_\lambda$.

- So, the following property must be satisfied:
  - if we have $f(y_1) + \ldots + f(y_n) = f(x_1) + \ldots + f(x_n)$ and
    $$y_1 + \ldots + y_n = x_1 + \ldots + x_n = n,$$
  - then we should have
    $$f(\lambda \cdot y_1) + \ldots + f(\lambda \cdot y_n) = f(\lambda \cdot x_1) + \ldots + f(\lambda \cdot x_n).$$

- This property must hold for any $\lambda$ that can be obtained by the above formula.
23. Proof (cont-d)

- For every positive $\lambda < 1 + 1/n$, we can find an $x > 0$ for which this formula holds: namely, we can take $x = \frac{n + 1}{\lambda} - n$.

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

- Let us consider the values $y_i$ that are close to $x_i$, i.e., for which, for some values $d_i$, we have $y_i = x_i + s \cdot d_i$ 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 of the sums 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 $s$, we get an equivalent equality
  
  \[ f'(x_1) \cdot d_1 + \ldots + f'(x_n) \cdot d_n + o(1) = 0. \]
24. Proof (cont-d)

- Similarly, both given equalities take the form

\[
d_1 + \ldots + d_n + o(1) = 0, \quad \text{and} \quad f'(\lambda \cdot x_1) \cdot d_1 + \ldots + f'(\lambda \cdot x_n) \cdot d_n + o(1) = 0.
\]

- So, in the limit \( s \to 0 \), we conclude that for each vector \( d = (d_1, \ldots, d_n) \):
  - if we have \( f'(x_1) \cdot d_1 + \ldots + f'(x_n) \cdot d_n = 0 \) and \( d_1 + \ldots + d_n = 0 \),
  - then we should have \( f'(\lambda \cdot x_1) \cdot d_1 + \ldots + f'(\lambda \cdot x_n) \cdot d_n = 0 \).

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

\[
b \overset{\text{def}}{=} (f'(\lambda \cdot x_1), \ldots, f'(\lambda \cdot x_n)).
\]

- The scalar product of two vectors is 0 if and only if these two vectors are orthogonal.
25. Proof (cont-d)

- 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 earlier.

- 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 \(b = A \cdot a + B \cdot e\).

- In terms of vector components, this means that

\[ f'(\lambda \cdot x_i) = A \cdot f'(x_i) + B \]
26. **Proof (cont-d)**

- In general, the coefficient $A$ and $B$ depend on $\lambda$ and on the initial tuple $(x_1, \ldots, x_n)$.
- So, strictly speaking, we should write
  \[ f'(\lambda \cdot x_i) = A(\lambda, x_1, \ldots, x_n) \cdot f'(x_i) + B(\lambda, x_1, \ldots, x_n). \]
- Let us show that $A$ and $B$ do not depend on $x_1$.
- Indeed, for $i = 2$ and $i = 3$, we get
  \[ f'(\lambda \cdot x_2) = A(\lambda, x_1, \ldots, x_n) \cdot f'(x_2) + B(\lambda, x_1, \ldots, x_n); \]
  \[ f'(\lambda \cdot x_3) = A(\lambda, x_1, \ldots, x_n) \cdot f'(x_3) + B(\lambda, x_1, \ldots, x_n). \]
- Let us subtract the second equation from the first one and divide both sides of the resulting equation by the coefficient at $A(\lambda, x_1, \ldots, x_n)$.
- Then, we conclude that
  \[ A(\lambda, x_1, \ldots, x_n) = \frac{f'(\lambda \cdot x_2) - f'(\lambda \cdot x_3)}{f'(x_2) - f'(x_3)}. \]
27. Proof (cont-d)

- So:
  - unless the derivative is a constant – and thus, the function \( f(x) \)
    is linear,
  - the right-hand side depends only on \( \lambda, x_2, \) and \( x_3 \) (and does not
    depend on \( x_1 \)).

- 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 the equality describing \( f'(\lambda \cdot x_3) \), we can now conclude that
  \[
  B(\lambda, x_1, \ldots, x_n) = f'(\lambda \cdot x_2) - A(\lambda, x_2, x_3) \cdot f'(x_2).
  \]

- 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 \( x_1 \).
28. Proof (cont-d)

- 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.
- So $A$ and $B$ only depend on $\lambda$:
  \[ F(\lambda \cdot r) = A(\lambda) \cdot F(r) + B(\lambda). \]
- Here, we denoted $F(r) \overset{\text{def}}{=} f'(r)$.
- The formulas for $A$ and $B$ take the form
  \[ A(\lambda) = \frac{f'(\lambda \cdot x_2) - f'(\lambda \cdot x_3)}{f'(x_2) - f'(x_3)} \quad \text{and} \quad B(\lambda) = f'(\lambda \cdot x_2) - A(\lambda) \cdot f'(x_2). \]
- We assumed that the function $f(r)$ is twice differentiable.
- Thus, its derivative $F(r) = f'(r)$ is differentiable.
- So, due to the above formulas for $A$ and $B$, the functions $A(\lambda)$ and $B(\lambda)$ are differentiable too.
29. **Proof (cont-d)**

- So, we can differentiate both sides of the formula for $F(\lambda \cdot r)$ by $\lambda$, and get $x \cdot F'(\lambda \cdot r) = A'(\lambda) \cdot F(r) + B'(\lambda)$.

- In particular, for $\lambda = 1$, we get $r \cdot \frac{dF}{dr} = a_0 \cdot F + b_0$.

- Here we denoted $a_0 \overset{\text{def}}{=} A'(1)$ and $b_0 \overset{\text{def}}{=} B'(1)$.

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

- 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 $r \cdot \ln(r)$.

- If $a_0 \neq 0$, then for $G \overset{\text{def}}{=} F + b_0/a_0$, we get $\frac{dG}{a_0 \cdot G} = \frac{dr}{r}$. 

30. Proof (cont-d)

- 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 $r^{\alpha}$ for $\alpha = a_0 + 1$.

- The theorem is proven.
31. Acknowledgments

This work was supported in part by:

- National Science Foundation grants 1623190, HRD-1834620, HRD-2034030, and EAR-2225395;
- AT&T Fellowship in Information Technology;
- program of the development of the Scientific-Educational Mathematical Center of Volga Federal District No. 075-02-2020-1478, and
- a grant from the Hungarian National Research, Development and Innovation Office (NRDI).