A Friendly Amendment of the Theil Index and Consequent Test for Across and Within Theil Inequality

Kobi Abayomi*, and William Darity, Jr.†

August 6, 2010

Abstract

Theil's Index is a version of Shannon's Entropy familiar to econometricians as a measure of distributional irregularity or inequality. The standard version of Theil's index is misspecified and thus unsuitable for statistical use: the commonly used partitioning of across and within contributions of disjoint groups generates a misleading interpretation of across group inequality. We explore an adjustment of Theil's index, in view of Shannon's original axiomatization, and suggest a natural test statistic for inequality across and within groups. We illustrate the method on a well known survey on US household wealth and income.

1 Introduction

Theil's index ([1]) is a *measure of inequality* — often defined as one of a class of functions that are increasing on unequal arguments. Members of this class include the *Lorenz curve* and *Gini's index* (see [13]) in this class.

Theil's index is a version of Shannon's Entropy — familiar as a statistic that quantifies the magnitude of *information* in a communication channel ([9]). In Shannon's original setup, the entropy measures the granularity of Markov Process, i.e. the distribution of probability mass over the states of the process. More generally, Shannon's entropy is a function on a probability distribution, and thus a function on a simplex.

We adjust the Theil index by revisiting Shannon's initial construction — in view of the *within* and *across* group partitioned representation — and offer a test for inequality on the term by term representation. We illustrate the methodology on a well known longitudinal

*School of Industrial Engineering - Statistics Group; Georgia Institute of Technology, Atlanta, GA.
†Department of Public Policy, Duke University, Durham, NC.

data set: the University of Michigan Health and Retirement Survey. We find statistically significant inequality via the observed Theil indexes, both across and within black and white race categories, on wealth and as well as income.

1.1 Theil's Index

Thiel's index, on a population of n total individuals, is commonly defined as:

$$T = \sum_{j=1}^{m} p_j \; R_j \; log_b \; R_j + \sum_{j=1}^{m} p_j \; R_j \; T_j \tag{1}$$

with

$$T_j = {n_j}^{-1} \sum_{i \in g_j} r_{ij} \; log_b \; r_{ij}, \tag{2}$$

m disjoint groups, $g_1, ..., g_m$, each with n_j members - $n = \sum_j n_j$ (see [7] and [14]). Each of R_j is the 'share' of the underlying random variable — say $X_1, ..., X_n$ for a population — apportioned to group j; p_j is a probability of choosing group j, the relative cardinality of group j; r_{ij} is the conditional share of X for individual i, given membership in group g_j. The logarithmic base is usually the natural number e; here we parameterize it as b. Our affixes to the standard notation are: (i) group membership j to within group share r_i; and (ii) $b > 0$ as a parameter.

1.2 Shannon's Entropy

Shannon defined the measure:

$$H(p_1, ..., p_n) = -\sum_{i=1}^{n} p_i \; log \; p_i, \tag{3}$$

where p_i is the probability of a system being in cell i of a 'phase' space, as a quantification of the uncertainty of a realization from a discrete Markov process. Shannon's idea was to define a measure via the associated probabilities of occurrence of a Markov process $\mathcal{X}$ with states $X_1, ..., X_n$: $\mathbb{P}(\mathcal{X}(\omega) = X_i) = p_i$.

Shannon's axioms for the measure H are: (i) H should be continuous in the p_i's; H should be a monotonic increasing function of n if all of the states of $\mathcal{X}$ are equally likely; (iii) H should be equivalent for repartitioning of probability space, i.e. for changes of measure, *in particular for conditioning.* Shannon's proof is that the axioms yield:

$$H(p_1, ..., p_n) = -K \sum_{i=1}^{n} p_i \; log \; p_i, \tag{4}$$

with K a constant. Shannon was informal (in fact he references Frechet, [15], for a 'detailed treatment'); the form of H is a direct consequence of the Chapman-Kolmogorov equations for discrete Markov processes.

The imposition that the measure be consistent via conditioning yields the logarithmic sum: it is Shannon's insight to define the measure as the *entropy* of a set of probabilities $\mathbf{p} = (p_1, ..., p_n)$ and to label $H(X)$ as the entropy of the 'chance' variable X. Thus $H(X)$ is a label for $H(\mathbf{p})$ when $X \sim \mathbf{p}$. This distinction — that H is a function on the probability simplex, and not on the space of X — is sublimated somewhat by Theil's use of Shannon's measure on income 'shares' ([1], [2], and [5]) and, perhaps, fully suppressed by later authors.

Notice that we discuss Theil's index (and Shannon's entropy) in the manner of sample statistics - in particular, as U statistics, see [16] - by eliding a deeper conversation on T as estimators (or estimators of functions of) population parameters. This is partly in adherence to style, the index is commonly not discussed as a parameter estimate in the econometrics literature, and partly substantive: our amendment to T can be derived solely via arguments on an expected value of the sampled T, with minimal assumptions on an underlying probability space.

1.3 Revisiting Theil *à* Shannon

This discrepancy is important: 'information', as defined by Shannon, is in units which correspond to b: the choice of base for the logarithm and K the 'normalizing' constant. The natural choice for Shannon's measure is $b = 2$ — information 'bits' are units on the base 2 'alphabet'. This 'alphabet' is the codification of the states of the random process $\mathcal{X}$, thus the measure H is in terms of units which are consonant to the natural granularity, or specification, of the associated random process.

In Theil's specification

$$T = n^{-1} \sum_{i=1}^{n} r_i \; log \; r_i \tag{5}$$

where r_i is the 'share' of income for individual i, an individual i must then be a state of a Markov process. The 'income share' r_i, then, is not the value of an associated random variable but *must* be a probability of occurrence. Theil suggests an income share r_i can

be interpreted as 'the chance a random dollar [from the budget] will be spent on the *ith* [person]' [1].

In practice, i.e on data $\mathbf{x} = (x_1, ..., x_n)$, x_i is set to be the income of individual i and

$$r_i = \frac{x_i}{\overline{x}} \tag{6}$$

is the fraction of income for individual i with respect to sample mean $\overline{x}$. Theil's construction suggests

$$H = -\sum_{i=1}^{n} \frac{x_i}{n\overline{x}} \log_b \frac{x_i}{n\overline{x}} \tag{7}$$

is used as Shannon's entropy. This yields $\mathbf{p} = (p_1, ..., p_n) = (\frac{x_1}{n\overline{x}}, ..., \frac{x_n}{n\overline{x}}) = \mathbf{r}/n$ on a simplex, however, the interpretation of Theil's index is inconsistent with Shannon's design of the measure on a probability space.

Since

$$\frac{log_{b_0}\, t}{log_{b_1}\, t} = K, \ \forall t \tag{8}$$

say, K can serve as a conversion between information units b_0 and b_1. This feature is elided from Theil's construction with the loss of Shannon's constant.[1]. This vagueness — the loss of the constant and the arbitrary specification of the logarithmic base — vitiates the typical use of the index as a composition of across and within group inequalities. While increasing, or decreasing K will not erase this phenomena, K is a placeholder for the change of base and as such is analogous to choosing the correct b.

2 Partitioning T

Consider this rewrite of (1),

$$T = \sum_G \frac{n_g}{n} \frac{\overline{X}_g}{\overline{X}} log_b \frac{\overline{X}_g}{\overline{X}} + \sum_G \frac{\overline{X}_g}{\overline{X}} \frac{1}{n_g} \sum_g \frac{X_{ig}}{\overline{X}_g} log_b \frac{X_{ig}}{\overline{X}_g} \tag{9}$$

with: X_{ig} the observed income for individual i, in group g; $\overline{X}_g$, the sample mean for group g; and $\sum_G$ a sum over all groups and $\sum_g$ a sum over group g. The first term on the right hand side

[1] To be fair, Theil recognized the use of the constant but considered it arbitrary ([1])

$$Across = \sum_{G} \frac{n_j}{n} \frac{\overline{X}_g}{\overline{\overline{X}}} \, log_b \frac{\overline{X}_g}{\overline{\overline{X}}} \tag{10}$$

is the measure of the *across* or *between* group inequality; the second term

$$Within = \sum_{G} \frac{\overline{X}_g}{\overline{\overline{X}}} \frac{1}{n_g} \sum_{g} \frac{X_{ig}}{\overline{\overline{X}}_g} log_b \frac{X_{ig}}{\overline{\overline{X}}_g} \tag{11}$$

the *within* group inequality.

Assume all incomes are positive and continuously distributed, the group and overall means exist, and the observations are independent and identically distributed - modulo group memberships $g \in G$, each of size n_g, with $n = \sum_G n_g$:

$$X \sim \mu < \infty, \; X, \mu > 0 \tag{12}$$

and thus

$$\overline{X} \sim \mu_g < \infty, \; \overline{X}, \mu_g > 0 \tag{13}$$

Inasmuch as Theil's index is used to measure inter-group income disparity - see [14] and [12] - the underlying parameters of interest must be the group and overall means: the collection on G of μ_g and μ. We can interpret T in analogy to the well known Analysis of Variance (ANOVA) setup for test of mean differences or treatment effects. In the ANOVA setup we address the hypothesis of group mean difference via a partitioning of sums of squares ; T is used to address group inequality differences via log sums or decomposed entropies. See Sen [17] for an excellent discussion.

The ANOVA theory is extremely well developed (see [18] for example); theory for the partitioning of T, via entropies by construction, is incomparably inferior. We believe this is partly due to the disconnection between the statistical entropy/signal processing literature and applied research using T. Additionally, and importantly, the weaker assumptions of the T inequality setup - only that distributions exist, contrasted with the reliance on normality for ANOVA - make statistical inference much more difficult . Recent work, promisingly, begins to address this disconnect, see [19].

The deduction of the complete probability distribution for T is a non-trivial task — however, we can assert some important characteristics about the decomposition of T using only the linearity of the expectation, minimal assumptions on the finiteness and positivity of incomes X and their exchangeability via our discussion of T as a U-statistic.

2.1 Within and Across

First, we assert that the expectation of the *within* term is greater than zero.

Theorem 2.1. $E[Within] \geq 0$

Proof.

$$E[Within] = E[E[\sum_{G} \frac{\overline{X}_g}{\overline{\overline{X}}} \frac{1}{n_g} \sum_{g} \frac{X_{ig}}{\overline{X}_g} log_b \frac{X_{ig}}{\overline{X}_g} | G = g]] \tag{14}$$

which yields

$$E[Within] = E[\sum_{G} \frac{1}{n_g} \frac{\overline{X_g}}{\overline{\overline{X}_g}} E[\frac{1}{\overline{\overline{X}}} \sum_{g} X_{ig} log_b \frac{X_{ig}}{\overline{X}_g} | G = g]] \tag{15}$$

and then, by conditioning on $G = g$

$$E[Within] = E[\sum_{G} \frac{1}{\overline{\overline{X}} n_g} \sum_{g} E[X_{ig} log_b \frac{X_{ig}}{\overline{X}_g} | G = g]] \tag{16}$$

$$= E[\sum_{G} \frac{1}{\overline{\overline{X}} n_g} E[\sum_{g} X_{ig} log_b \frac{X_{ig}}{\overline{X}_g}]] \tag{17}$$

$$\geq E[\sum_{G} \frac{1}{\overline{\overline{X}} n_g} E[n_g \overline{X}_g log_b \frac{n_g \overline{X}_g}{n_g \overline{X}_g}]] \tag{18}$$

$$\geq E[\sum_{G} \frac{1}{\overline{\overline{X}} n_g} E[0]] = 0 \tag{19}$$

with (18) because of the log-sum inequality.

□

This illustrates that, for any - arbitrary - choice of log base b, the expectation of the within term of T is strictly positive as it is a sum of strictly positive numbers.

Now for the expectation of the across term.

For conciseness let: $\alpha_g = \frac{n_g}{n} \frac{\overline{X}_g}{\overline{\overline{X}}}$, these are ratios of group income to overall income; and $\beta_g^b = log_b \frac{\overline{X}_g}{\overline{\overline{X}}}$, the log ratios of group mean income to overall mean income.

We continue with the claim that the expectation of the across term is bounded above by the expectation of the sum of the log ratios of group mean incomes to overall mean incomes, β_g^b

Theorem 2.2. $E[Across] \leq E[\sum_G \beta_g^b]$

Proof.
Write:

$$E[Across] = E[\sum_G \frac{n_g}{n} \frac{\overline{X}_g}{\overline{X}} log_b \frac{\overline{X}_g}{\overline{X}}] = E[\sum_G \alpha_g \beta_g^b] \tag{20}$$

and note that $\sum_G \alpha_g = 1$, and $\alpha_g \geq 0$, by assumption, $\forall g$. Then,

$$E[Across] = E[\sum_G \alpha_g \beta_g^b] \tag{21}$$

$$\leq E[\sum_G \alpha_g \sum_G \beta_g^b] = E[1 \cdot \sum_G \beta_g^b] \tag{22}$$

$$= E[\sum_G \beta_g^b] \tag{23}$$

□

Thus the across term of T (see 1, 9, 10 and 11 - above) is bounded above by $E[\sum_g \beta_g^b]$. Illustrations of this bound (as a function of b - the choice of log base - and $\overline{x}_g, g \in G$, the collection of group (sample) means) highlight a disparate effect on T when it is partitioned as across and within group inequalities.

Notice that for $b < 1$, $\beta_g^b < 1$ decrease the across term; for $b < 1$ $\beta_g^b < 1$ increase the across term. Note that $\beta_g^b < 1$ for groups having less than the population average. See Figure 1.

We call a "singular" collection G one where $\overline{X}_g = \overline{X}$, $\forall g$. For these collections - where all groups have the same average - $\beta_g^b = 1$, $\forall g$, trivially.

Consider, without loss of generality, any non-singular collection β_g^b, ordered increasingly: setting b marks the "gradient", say, for the sum of the collection β_g^b - or the rate of descent (ascent) of (23). Our remark is that b can be chosen such that $E[\sum_G \beta_g^b] < 0$.

We illustrate this via a brief lemma:

Lemma 2.3. *For any non-singular collection of groups, G, ordered without loss of generality, $\exists\, t$ such that*

$$\beta_{g_t}^b < 0 < \beta_{g_{t+1}}^b \tag{24}$$

Proof.

To see this, take $b > 1$ and take any $X,\ X > 0$, by the minimal assumptions in (12) and (13), draw and order m samples $X_1, ...X_m$. With $\overline{X} = m^{-1} \sum X_i$, $\exists\ t$ such that

$$X_t < \overline{X} < X_{t+1} \tag{25}$$

since otherwise $X_{(s)} > \overline{X},\ \forall\ s$ or $X_{(s)} < \overline{X},\ \forall\ s$, in obvious contradiction. Then (25) implies (26), below

$$\frac{\overline{X}_{g_t}}{\overline{X}} < 1 < \frac{\overline{X}_{g_{t+1}}}{\overline{X}} \tag{26}$$

and thus (24), in our notation, when $b > 1$.

We get the proposition for $b < 1$ by realizing that the groups may be ordered without loss of generality.

□

For $b < 1$, terms $\beta^b_{g_s},\ s \leq t$ are positive and $\beta^b_{g_s},\ s > t$ are negative contributions to the across term 23; for $b > 1$ the change is from negative to positive. A choice of b fixes (arbitrarily) the particular t, modulo the underlying distribution of the collection $\overline{X}_g$ and (likewise) β^b_g.

Consider these propositions:

Theorem 2.4.

(I)

$$b \leq \min_G(\overline{X}_g/\overline{X}) \implies E(\sum_G \beta^b_g) \geq 0 \tag{27}$$

(II)

$$b \geq \max_G(\overline{X}_g/\overline{X}) \implies E(\sum_G \beta^b_g) \leq 0 \tag{28}$$

with equality, in both conclusions, for singular collections.

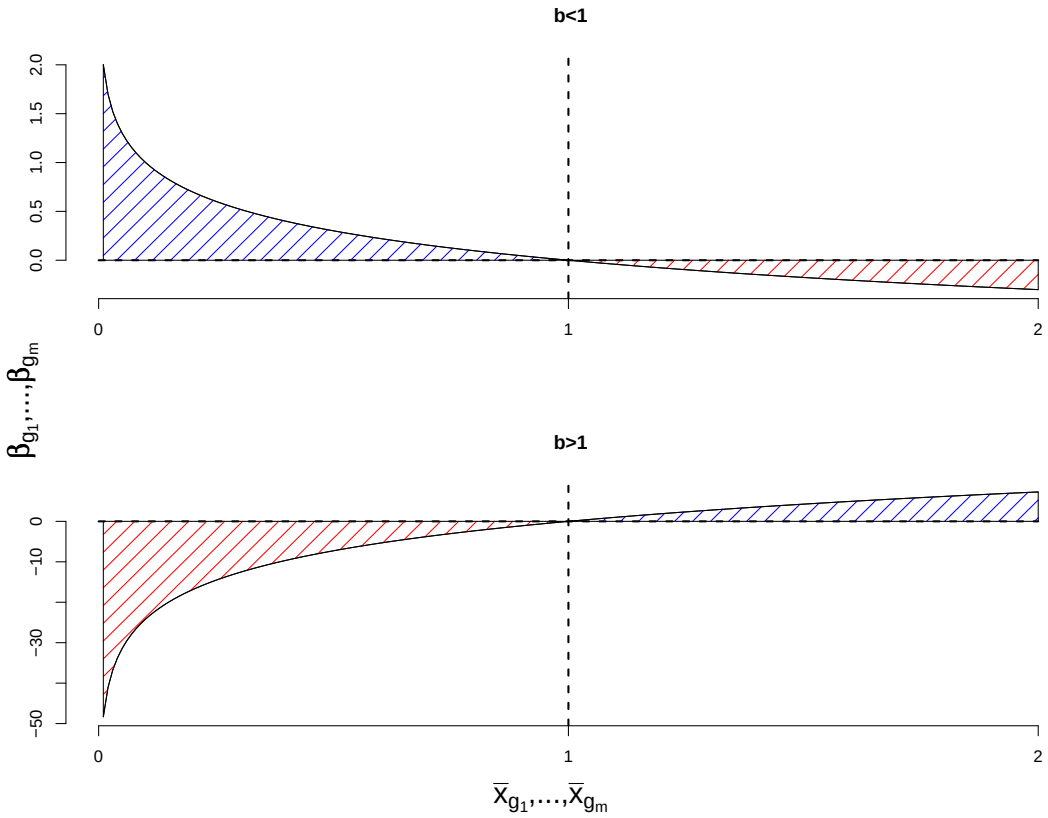

Figure 1: Illustration of β_g vs. $\overline{x}_g$ for log base $b < 1$ and $b > 1$. The contribution to the across term of Theil's index - T, equation (10), is concave up or down by choice of b. Consider a collection of groups $g_1, ..., g_m$ ordered by group mean $\overline{x}_g$, without loss of generality. The panels illustrate the magnitude of the contribution to the expectation in equation (23) for a collection of groups with average incomes (uniformly, say) distributed within (two, say) multiples of the population average. Setting b sets the inflection for increasing/decreasing (or decreasing/increasing) contributions to the across term. A choice of b dictates the expectation of the *across* part of Theil's index.

Proof.

Take $b < 1$, case (I) and recall notation $\beta_g^b = log_b \frac{\overline{X}_g}{\overline{X}}$.

Fix $b \leq \min_G(\overline{X}_g/\overline{X})$, thus $b \leq \overline{X}_1/\overline{X}$.

Notice as well

$$1 \geq \beta_{g_1}^b > \beta_{g_2}^b > \cdots > \beta_{g_t}^b > 0 > \beta_{g_{t+1}}^b > \cdots > \beta_{g_m}^b \tag{29}$$

by lemma 2.3.
Then rewrite (23)

$$E[\sum_G \beta_g^b] = E[\sum_G log_b \frac{\overline{X}_g}{\overline{\overline{X}}}] \tag{30}$$

$$= E[log_b(\frac{\prod_G \overline{X}_g}{\overline{X}^m})] \tag{31}$$

by property of the logarithm.
By the Cauchy-Schwarz inequality,

$$\sqrt[m]{\prod_G \overline{X}_g} \leq \overline{X} \tag{32}$$

which implies

$$\prod_G \overline{X}_g \leq \overline{X}^m \tag{33}$$

Thus, recalling $b < 1$ and the property of the logarithm (see Figure 1).

$$log_b(\frac{\prod_G \overline{X}_g}{\overline{X}^b}) \geq 0 \tag{34}$$

so

$$E[\sum_G \beta_g^b] = E[log_b(\frac{\prod_G \overline{X}_g}{\overline{X}^b})] \geq 0 \tag{35}$$

For case (II), fix $b \geq \max_G(\overline{X}_g/\overline{X})$, thus $b > 1$.
Notice, again via lemma 2.3

$$1 \leq \beta_{g_m}^b < \beta_{g_{m-1}}^b < \cdots < \beta_{g_t+1}^b < 0 < \beta_{g_t}^b < \cdots < \beta_{g_1}^b \tag{36}$$

Since $b > 1$ rewrite

$$log_b(\frac{\prod_G \overline{X}_g}{\overline{X}^b}) \leq 0 \tag{37}$$

and

$$E[\sum_G \beta_g^b] = E[log_b(\frac{\prod_G \overline{X}_g}{\overline{X}^b})] \leq 0 \tag{38}$$

□

This demonstrates that the bound for the expected value of the across partition of Theil's index can be made positive or negative, arbitrarily, depending upon choice of log base b. Theil himself is indeterminate on the importance of b; in [4] he mentions it in passing as a normalizing constant and suggests use of the natural number e. Later authors follow this convention.

This oversight is a problem when the index is used to decompose or aggregate inequality over several groups. With just one group, say - as $T \in [0, log_b(n)]$ - T is bounded by $log_b n$. If $b = n$ then the index is on $[0, 1]$, b can incautiously be considered a scaling parameter as in equation (8) above. Recent work, see [20] and [21], investigating statistical properties of the non-decomposed index T ignores - without loss - the role of this scaling parameter.

The value of b is non-ignorable, however, when we desire inference on inter-group differences in inequality.

3 A T-test for inequality

Consider a hypothesis testing setup for the inequality of the income distribution in a population: for any $b > 0$ the expected value of Theil's index T is zero. Note that this the value of T for a uniform distribution of incomes, and notice that the incomes are exchangeable in the sense that T is equivalent for equal deviations from equality/uniformity at the upper and lower ends of the distribution.

In this setup we can treat the observed value of T, as a test statistic for the hypothesis of inequality:

$$\begin{aligned} H_0 &: \mathcal{T} = 0 \\ &vs. \\ H_a &: \mathcal{T} > 0 \end{aligned} \tag{39}$$

with $\mathcal{T}$ the true (or population) value of Theil's index.

Martinez-Camblor (see [21] suggest that the observed value, or estimator, T is asymptotically normal about T. Using this finding yields

$$p - value = \mathbb{P}_{H_0:\mathcal{T}=0,b}\left(Z > \frac{T}{s.e.(T)}\right) \tag{40}$$

with Z a standard normal Gaussian random variable and $s.e.(\hat{T})$ the standard error of the estimator, estimated in [21] via bootstrapping.

This setup is insufficient for partitioned versions of the index. Consider the partitioning $T = T_a + T_w$ into across and within terms under an arbitrary $b > 0$: we have demonstrated that values of b exist where $E(T_a) < 0$ but $E(T_w) > 0$, *even under a null hypothesis* of $H_0 : \mathcal{T} = 0$.

To see this simply recognize that the hypothesis $H_0 : \mathcal{T} = 0$ is composite over the set $\{\mathcal{T}_a = -\mathcal{T}_b\}$ with $\mathcal{T}_a = 0 = \mathcal{T}_b$ only a special case. This is to say that overall income inequality is zero for all cases where across group inequality is opposite to within group inequality. We can say that overall income inequality is zero when across and within group inequalities are zero, but not the converse.

We suggest simultaneous testing of across and within group inequality:

$$\begin{aligned} H_0 : \mathcal{T}_a = 0 = \mathcal{T}_w \\ vs. \\ H_a : \mathcal{T}_a > 0; \mathcal{T}_w \neq 0 \end{aligned} \tag{41}$$

Given an appropriate b, this is

$$\begin{aligned} H_0 : \mathcal{T}_a = 0 = \mathcal{T}_w \\ vs. \\ H_a : \mathcal{T}_a > 0; \mathcal{T}_w > 0 \end{aligned} \tag{42}$$

The significance level of the test can be set via a Bonferroni type correction for multiple hypothesis, or more conservatively:

$$p - value = \max\left\{\mathbb{P}_{H_0:\mathcal{T}_a=0,b}\left(Z > \frac{T_a}{s.e.(T_a)}\right), \mathbb{P}_{H_0:\mathcal{T}_w=0,b}\left(Z > \frac{T_w}{s.e.(T_w)}\right)\right\} \tag{43}$$

with Z a standard normal Gaussian random variable, as before, by Gaussianity of linear transforms. The standard errors can be estimated separately, again via bootstrap, for each term: $s.e.(T_a), s.e.(T_w)$. Note that the standard error estimated without decomposition, i.e. as $s.e.(T)$ on T, yields an upper bound for the variance of decomposition $T_a + T_w$; it is unclear if the across and within terms are (even linearly) independent under reasonable assumptions on the underlying incomes X and grouping G.

We remark that our adjustment of b does not suggest that across and within terms are equal under the alternative - in fact T_w often dominates $T_a$under our recommendation - but *the expectation of both is known and specified under the null hypothesis of inequality* and *both are restricted non-negative in consonance with Theil's and Shannon's design.*

4 Illustration

We offer an illustration using the University of Michigan's Health and Retirement Survey (HRS) data [22]. The data are a battery of responses from a well known longitudinal study; Elemech (see [12]) in particular investigates intra-group inequality for white, black and latino Americans.

Using data from the 2000, 2002, 2004, 2006 and 2008 waves of the survey we calculate the Theil indexes on wealth and income household totals, grouping on ethnicity. We chose to ignore Hispanic/Latino classification for this illustration. This grouping yielded 19580, 18167, 20134, 18469 and 724 observations, by increasing year.[2] We used responses to total net worth and total income.

We case-wise deleted values from the data that were less than zero - by assumption here and elsewhere Theil's index is calculated on positive quantities. Additionally, after removing cases with negative values in either wealth or income, we shifted all responses by one. This allowed us to calculate the index on all remaining values - preventing infinite logarithms of zero. We point out that these two modifications of the data may have an appreciable effect on the measurement of inequality: the deletion of negative values - in particular wealth - may affect the across and within measurement of black inequality, and the difference - measured via logarithm - between zero and one may be also obscure the effect of inter vs. intra group inequality.

[2]The vast majority of responses to ethnic category - approximately 17000 - in the 2008 data were missing. We case-wise deleted the non-responses; the indexes are calculated from the few remaining values.

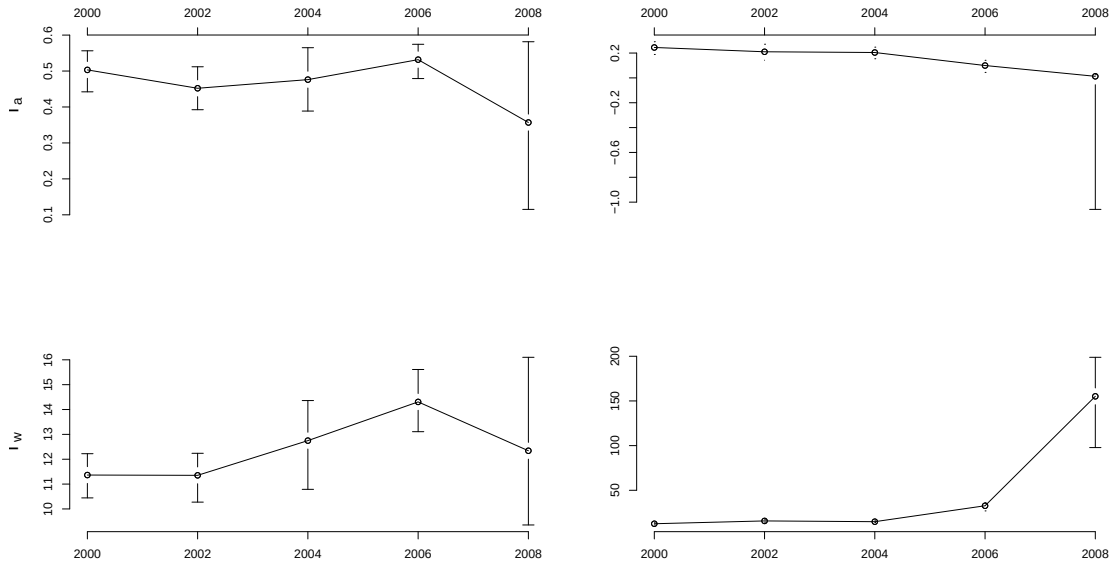

Figure 2: Illustration of Theil's index calculated on wealth - left hand column - and income - right hand column - using the University of Michigan's Health and Retirement Survey (HRS) data: 2000, 2002, 2004, 2006, and 2008. The upper row is the across term, the lower row is the within term. Both terms are fixed by log base $b = \min(\overline{x}_g/\overline{x})$ the ratio of the poorer (black) group sample mean to the overall mean. The small number of observations in 2008 (n=724) inflate the (estimate of) standard error - via ordinary bootstrap. Across group inequality appears to be stable or decreasing from 2000-2008; within group inequality appears to be increasing from 2000-2006. Confidence bars are at 95 percent significance.

Figure 2 plots the across and within observed Theil indexes, by year. The small number of data in 2008 inflates the estimate of the standard error - obvious from the illustration by the wider confidence bars around the estimate. In general the observed estimate of across group inequality from 2000-2008 appears to be decreasing for both wealth and income. However - in the lower panel of figure 2 - the within group inequality sum appears to (mainly) increase from 2000-2008. The 2008 estimate may be unreliable.

Year	Across	Within
2000	12.2	19.8
2002	7	22.1
2004	10	30.8
2006	4.0	13.9
2008	.0034	5.45

Table 1: Test statistics, Z_o, for T_a, Theil's index for calculated for income on the HRS data from 2000-2008. Compare each with a quantile from the standard normal distribution - for either of $H_0 : T_a, T_w = 0$, $Z_o \geq 1.64$ implies significance at the .05 level.

Year	Across	Within
2000	25	23.6
2002	15	18.6
2004	9.4	13.7
2006	20.4	22.16
2008	2.97	7.38

Table 2: Test statistics Z_o, for T_w, Theil's index for calculated for wealth on the HRS data from 2000-2008. Compare each with a quantile from the standard normal distribution - for either of $H_0 : T_a, T_w = 0$, $Z_o \geq 1.64$ implies significance at the .05 level.

Across group wealth and income inequality appears to be decreasing from 2000-2008, yet across group wealth inequality seems to be noticeably greater than income inequality. Over the same period within group inequality, via both wealth and income, appears to be increasing. This agrees with contemporary research (see [14] and [12]). While differences across groups appear to be lessening - within racial group stratification is worsening. As well, it is important to notice that within group income disparity - the lower right panel in figure 2 - is larger in magnitude and change.

5 Discussion

We have illustrated a straightforward explanation for large differences in magnitude of observed values for partitions of Theil's index of inequality. The issue is related to the loss of the positive constant K in Theil's version of Shannon's entropy. K accounts for the choice of logarithmic base b. In Shannon's paper the natural choice for b is 2 for a binary variable.

One issue is the indistinguishability between the random variable, any sigma algebra (set of events) and associated probabilities. In fact, while Shannon's original measure was derived as function on a probability space and the log arose from first principles, there *is* an arbitrariness that can be exploited by choosing the logarithmic basis. Shannon himself states (see [9]):

> *The choice of a logarithmic base corresponds to the choice of a unit for measuring information.*

The choice of log base is analogous to the *minimal* distance measure on the probability space: in Shannon's seminal paper the basic distance is *Hamming* or the number of unequal positions in equal length strings[3]. In Shannon's setup (for telecommunication), then, base 2 is the appropriate distance measure - for binary strings.

Logs arrive *naturally* as the limits in binomial exponentiation and *algorithmically* as tools for transforming multiplication into addition and for idempotent differentiation (and integration) in calculus [10]. In this setting the use of the logarithm allows for independence - i.e. factorization of probability distribution - to be additive.

We illustrate the inequality of the partitioning of Theil's index as an artifact of choice of logarithmic base b. In light of this demonstration we suggest conscious fixing of the logarithmic index to guarantee both terms - across and within - are positive. This additional specification is consonant with the design of the index as an increasing measure of distance from equality - or uniformity of income distribution - with a maximum at $log_b n$.

It is unclear if the across and within partitions are independent under broad distributional assumptions for incomes. As well, in most cases n is much bigger than the min or max beta - usually x are in dollars, not millions of dollars or some other fraction. We make our comments about the partitioning of the index using the least restrictive assumptions on the underlying probability distribution of the incomes and groupings.

Using our recommendation then, the b of the index is the percent income of the poorest or richest group - and then the bound is the geometric mean of the entire population if everyone's income was that of the poorest or richest group.[4]

In this setting - with only the broadest probabilistic/distributional assumptions - we suggest a simple test for each of the terms of the index via their asymptotic normality in their role as estimators. In this way standard errors can be easily and quickly calculated

[3] *Minimal* in the sense of minimal complete: the minimal distance which distinguishes unique elements.

[4] Lastly, there is support in the literature for this change of logarithmic base technique. Rocke and Durbin investigate the *started log* (adding a constant in the log argument) and log-linear hybridization (log above a cutoff and linear below) [8] as instances of what are called generalized logarithm) transformations. And of course Box-Cox [6] and Tukey [11] have outlined general families of transforms.

via ordinary bootstrap and the overall hypothesis test of distributional inequality can be conservatively considered as the join of both individual tests.

We believe our approach introduces a necessary deeper inspection of the statistical properties of Theil's index - in particular - and formal hypothesis testing procedures on inequality. We comment that our paper places Theil's contribution, in particular, in a statistical econometric context - and more generally suggests a greater role for statistical analysis on the wider class of indices of inequality.

Consider the following (nested) tests, just via Theil's index:

All groups are equal This is equivalent to the ordinary ANOVA setup for equality of group means $\overline{X}_g$ *not* via Theil's log ratios.

Each group is equally unequal This is to say that $T_g = T \ \forall g \in G$. A natural test statistic is T_a the weighted sum of the across group inequalities.

Each group contributes equally to inequality measure This is more general version of the first hypothesis setup where groupwise inequality is weighted by group size. A possible test statistic is

$$\sum_G \frac{n_g}{n} \frac{\overline{X}_g}{\overline{X}} log_b(\frac{\overline{X}_g}{\overline{X}})$$

which should converge to a constant.

Inequality across groups is equivalent to inequality within groups Under this hypothesis, this ratio

$$\sum_G \frac{n_g}{n} \frac{\overline{X}_g}{\overline{X}} log_b \frac{\overline{X}_g}{\overline{X}} / \sum_G \frac{\overline{X}_g}{\overline{X}} \frac{1}{n_g} \sum_g \frac{\overline{X}_{ig}}{\overline{X}_g} log_b \frac{X_{ig}}{\overline{X}_g}$$

is the ratio of two positive quantities: for well chosen b the ratio is one.

6 Acknowledgments

Kobi Abayomi recognizes and thanks Susan Holmes of the Stanford University Statistics Department. Much of the research work for this paper was completed during a VIGRE Summer Fellowship for Junior Faculty, Summer 2008.

Kobi Abayomi also thanks: Santanu Dey at ISYE Georgia Tech and Sebastian Pokutta of Technische Universität Darmstadt for several priceless hints on log sum inequalities, Makram Talih at CUNY Biostats for pointing our recent work on limit theorems for T.

References

[1] Theil, H. (1965) , "The Information Approach to Demand Analysis," *Econometrica*, **33**, 67-87.

[2] Theil, H. (1967) "The Information Approach to the Aggregation of Input-Output Tables", **49**, 4, 451-462.

[3] Theil, H. (1969) "On the Use of Information Theory in Concepts in the Analysis of Financial Statements." *Management Science*, Vol. **15**, No. **9**, pp. 459-480.

[4] Theil, H (1967) *Economics and Information Theory*. Chicago: Rand McNally and Company. Amsterdam: North Holland Publishing Company.

[5] Theil, H. (1973) "A New Index Number Formula", *The Review of Economics and Statistics* **55**, 4, 498-502.

[6] Box, G.E.P and Cox, D.R. (1964) "An analysis of transformations." *Journal of the Royal Statistics Society, Series B*. **26**, pp. 211-252

[7] Conceicao, P Galbraith, J (2000) "Construction of Long and Dense Time-Series of Inequality Using the Thiel Index." *Eastern Economic Journal*, **26**, 1.

[8] Rocke, D. and Durbin, B. (2003) "Approximate variance-stabilizing transformations for gene-expression microarray data." *Bioinformatics*. **19**, 8. pp 966-972.

[9] Shannon, C.E. (1948) "A Mathematical Theory of Communication." *The Bell System Technical Journal*. **26**, pp. 379-423

[10] Spivak, Michael (1994) Calculus. Publish or Perish. Houston, Texas.

[11] Tukey, J.W. (1964) "On the comparative anatomy of transforms." *Annals of Mathematical Statistics*, **28**. pp. 602-632.

[12] Elmelech Y. (2006) "Determinants of Intra-Group Wealth Inequality Among Whites, Blacks, and Latinos." In Nembhard J. and N. Chiteji (Eds.). Wealth Accumulation and Communities of Color in the United States. University of Michigan Press: 91-112.

[13] Gastwirth, J. (1972) "The estimation of the Lorenz Curve and Gini Index." *The Review of Economics and Statistics*, **54**, 306-316.

[14] Darity, William Jr. and Deshpande, Ashwini (2000) "Tracing The Divide: Intergroup Disparity Across Countries." *Eastern Economic Journal*, **26**, 1. 75-85.

[15] Frechet, M. (1952) *Methode des fonctions arbitraires, theorie des evenements en chaine dans le cas d'un nombre fini d'etats possibles.* Paris. Gauthier-Villars.

[16] Hoeffding, W. (1948) "A class of statistics with asymptotically normal distribution." *Annals of Mathematical Statistics*, **19**, pp. 294-325.

[17] Sen, Amartya (1997) *On Economic Inequality.* Clarendon Press, Oxford.

[18] Wu, J and Hamada, M. (2009) *Experiments: Planning, Analysis and Optimization* Wiley. New Jersey.

[19] Borell, L and Talih, M (2010) "A symmetric, entropy-based, relative and quasi-absolute measure of health disparities: An example using dental caries in US children and adolescents." *In Review.*

[20] Biewen, M and Jenkins, S (2006) "Variance Estimation for Generalized Entropy and Atkinson Inequality Indices: The Complex Survey Data Case." *Oxford Bulletin of Economics and Statistics.* **68**, 3. 371-383.

[21] Martinez-Camblor, P. (2007) "Central Limit Theorems for S-Gini and Theil Inequality Coefficients." *Revista Colombiana de Estadística.* **30**, 2. 287-300.

[22] Health and Retirement Study, public use dataset. Produced and distributed by the University of Michigan with funding from the National Institute on Aging (grant number NIA U01AG009740). Ann Arbor, MI, (2010).