Some New (Old) Directions in Entropy, Information and Measurement of Income Inequality

Ravi Kanbur
www.kanbur.dyson.cornell.edu

Keynote, LAMES/LACEA,; Puebla, Mexico; 8 November, 2019

Introduction

- In this presentation I revisit an earlier literature which attempted to link entropy and information theory to measurement of income inequality.
- There was a flurry of activity in this area some decades ago—Theil (1967), Cowell (1980), Shorrocks (1983), etc. But the engagement between the two literatures appears to have subsided since then.
- We all use generalized entropy indices, but "entropy" is just part of a label, rather than indicating a deep engagement with the roots of entropy and information theory.
- I believe a renewed engagement has much to offer us, at least as a complement to the (now) conventional inequality measurement literature.

- In the first half of the talk I will take up the roots of entropy and information theory.
- In the second half I will look at applications to inequality measurement.
- Rather than focus on one specific aspect of entropy or information theory, I will highlight a number of different perspectives briefly, in the hope of triggering interest in one or other approach to the issue.

Inequality Measurement

- Let $x = (x_1, x_2, \ldots, x_K)$ be a vector of non-negative incomes for $i = 1,2,\ldots K$ individuals.

- Let total income be X and mean income be $\mu = X/N$.

- $\bar{\mu} = (\mu, \mu, \ldots, \mu)$ be the equal distribution vector.

- The problem of inequality measurement can then be stated as the problem of quantifying the divergence between the x and $\bar{\mu}$.

- A conventional approach is to specify a social welfare function on x and use this to quantify the divergence, for example through an “equally distributed equivalent” level of income and its distance from the mean, a la Atkinson.
- Another, related, conventional approach is to specify desirable normative properties and derive classes of inequality measures which satisfy these properties.
- Properties include scale independence and Principle of Transfers.
- Decomposability as a property is also prominent, although we might question in what precise sense this has normative content.

- This is the conventional approach.
- I would like in this presentation to consider an alternative perspective, which is somewhat more statistical in nature.
- Consider the following thought experiment. From the distribution of dollars across individuals, $x = (x_1, x_2, \ldots, x_K)$, a single dollar is drawn at random. How easy is it to trace back which individual the dollar came from?
- The extreme cases provide some intuition. If all dollars belonged to one individual, tracing back is easy. At the other end, if dollars were equally distributed, tracing back would in some sense be the most difficult. In between is in between.

- I believe that this way of looking at things—the ease or difficulty of tracing a randomly drawn dollar back to its source—leads to an interesting parallel/complement to the standard normative frame we use in inequality measurement.
- It is worth exploring, and it leads us to an engagement with the fundamentals of Entropy and of Information Theory which takes us beyond the prevalent use of “Generalized Entropy” as a mere label for a formula which all of us use in applied work.
- Such engagement was present in the past (eg in Theil), but seems to have diminished considerably in the current literature.

Entropy

- But why “Entropy”? Where did that come from?
- For this we need to go first to Statistical Physics and then to Information Theory to understand the implications of a remarkable formula.
- And we will encounter the work of two geniuses—Ludwig Boltzmann and Claude Shannon—and the remarkable unity across very different fields of science.
- Throughout, as we delve into statistical physics and binary coding, bear in mind the central thought experiment of tracing a randomly drawn dollar to the individual, or the “bin”, from which it came.

Entropy Definition

- Let a discrete probability distribution be defined as

$$p_k;\ k = 1, 2, \dots, K$$

$$p_k \ \geq 0\ ,\ \sum_k p_k = 1$$

- The Entropy of the distribution is defined as

$$\mathrm{H} = -\sum_k p_k \log\ p_k$$

- (note, by convention $0 \log 0 = 0$).

Entropy and Statistical Physics

- Let there be K energy levels and X (identical) gas molecules (or particles). What is the distribution of particles across energy levels?
- One way to approach this is “Newtonian”. Write down the equation of motion of each particle, and all possible interactions between particles. Possible in principle, but infeasible in practice—huge number of molecules (10 to the power 23 etc).
- The genius of Ludwig Boltzmann (1844-1906) was to formulate this as a statistical problem, giving rise to the field of statistical physics.

- Consider a particular configuration ${}^{x_k}/_{X} = p_k$
- How many different ways can this configuration arise, bearing in mind that the particles are indistinguishable so all that matters is the number of particles in each energy “bin”? This is a multinomial combinatorics problem and the answer is:

$$W = \frac{X!}{x_1!\, x_2! \ldots\ldots x_K!}$$

- W is known as the “multiplicity”. The basic argument is that the greater the multiplicity of a configuration, the more likely it is to be observed.
- Using Stirling’s approximation large X, $\log(X!) \sim XlogX - X$, it can be shown that

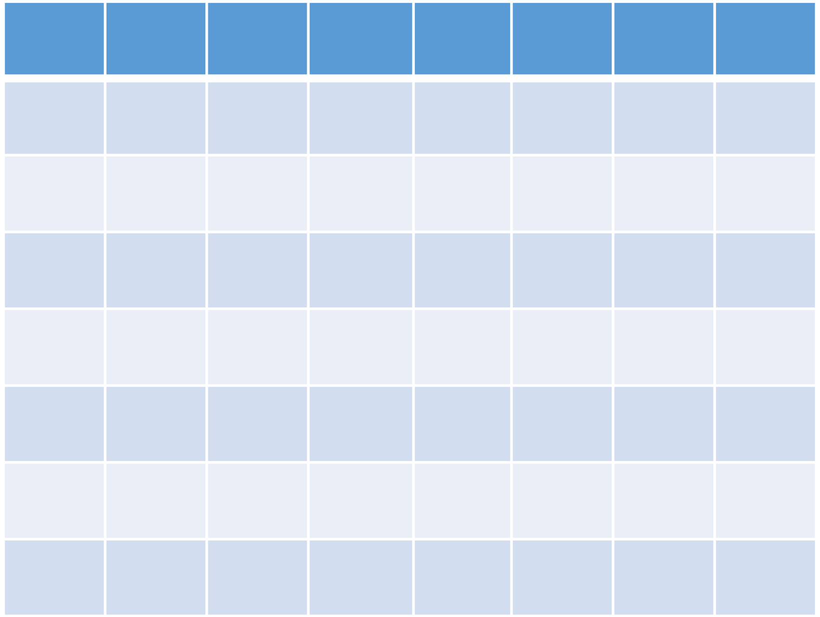
$$\log W = -X \sum_{k} p_k \log\, p_k = X * H$$

- In other words, the Entropy H of a distribution is (proportional to) a measure of how likely it is that the configuration will be observed.
- This is “Boltzmann Entropy”.


S = k. log W
LVDWIG
BOLTZMANN
Dr PHIL. PAULA
BOLTZMANN
ARTHUR
BOLTZMANN
LVDWIG
BOLTZMANN
HENRIETTE
BOLTZMANN

Entropy and Yes/No Questions

- Now change tack completely. Consider a chess board. In one of the 64 squares is a prize, but we don't know which one. It is equally likely that the prize is in any square ie with probability 1/64.



- We don’t know where the prize is but there is a machine which will always give truthful Yes/No answers to a grid location question. Is it here? Is it here? Etc.
- What is the smallest number of questions, ***in expectation***, that will get us to the prize?
- Clearly the wrong thing to do is to ask: Is it in square 1? Is it in square 2? Etc. Or to do it randomly square by square. You might get lucky, but you might not.
- Some reflection gives a more efficient set of questions to pose to a truthful “Yes/No” answer machine.

- Is it in the first four columns? Is it in the first two of the four columns where it is? Is it in the first one of the two columns where it is? Is it in the first four rows of this column? Is it in the first two of the these four rows? Is it the first of these two rows?
- Thus six questions get us to the prize for sure.

- But 6 is an interesting number.

- It is the value of $\mathrm{H} = -\sum_k p_k \log p_k$ when $p_k = 1/64$ and

the logarithm is to the base 2

- Base 2 because we are in the binary world of “Yes/No”, “Off/On”, “0/1”

- Is 6 the smallest number of questions which, in expectation, will take us to the prize (reveal the information)? The answer is yes.
- In fact we have a remarkable general theorem due to another genius, Claude Shannon.
- Take the general case where the probability of finding the prize in square k is p_k and we have any number of squares, $k = 1,2, \dots\dots K$
- The smallest number of questions needed, in expectation, to find the prize is none other than the Entropy of the distribution
- $\mathrm{H} = -\sum_k p_k \log p_k$

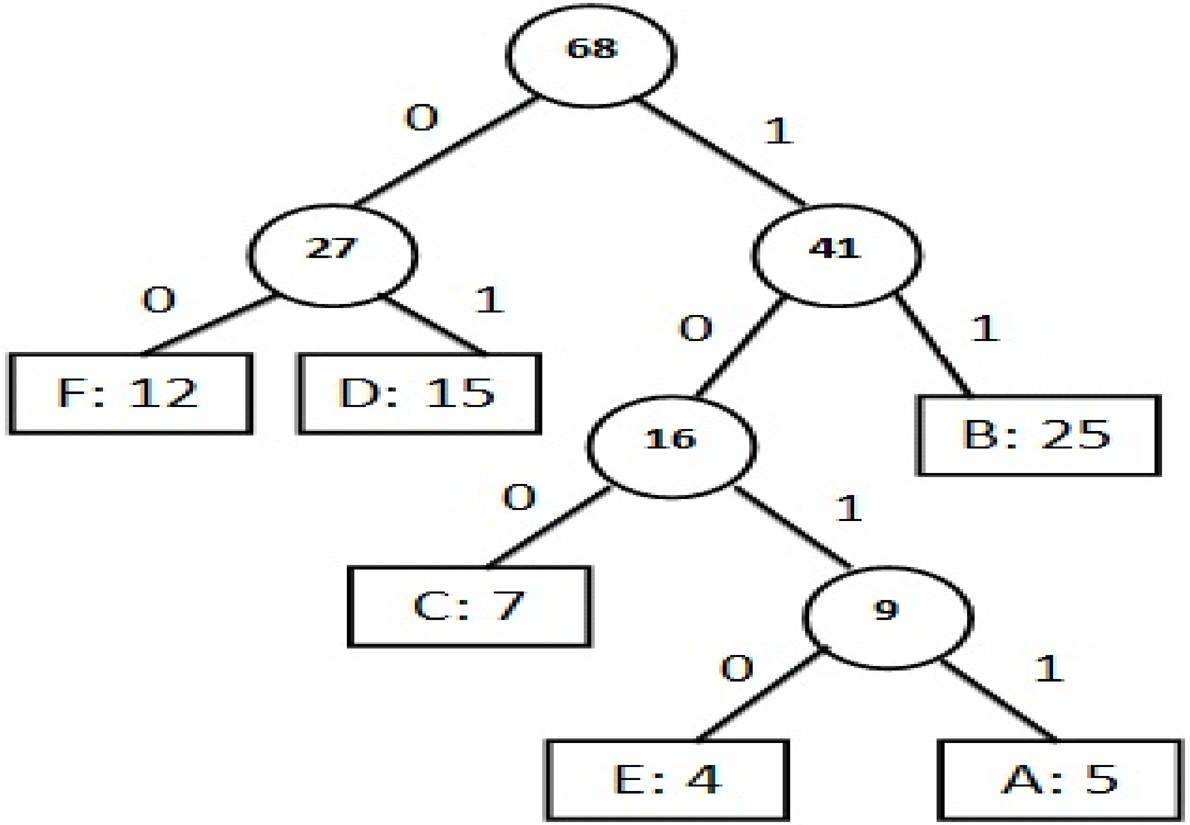
Entropy and Coding

- In fact Shannon proved this in the context of finding an efficient binary coding for messages. (It was 1948 and he worked for Bell Labs).
- Shannon, 1948: “A Mathematical Theory of Communication”
- Suppose we want to assign a binary code (ie made up of a string of 0’s and 1’s) for each letter of the alphabet to send messages through “on/off” switches—the dawn of the digital age.
- How should we do this if we want to save on code length—the number of 0’s and 1’s?

- The answer surely depends on the frequency with which the letter appears in the language. More frequent letters should get the shorter code.
- Here is an example illustrating one such method of coding—the Huffman Coding Algorithm. Let there be 6 letters with the following frequency.

Letter	A	B	C	D	E	F
Frequency	5	25	7	15	4	12

- Combine the two lowest frequencies into one. Treating this as one unit, combine again the two lowest frequencies into one. And so on till we reach the total frequency of all letters.
- This produces a tree diagram with left branches assigned label zero and left branch assigned label 1.
- We now have a binary code assignment for each letter as follows:



68
0
1
27
41
0
1
0
1
F: 12
D: 15
16
B: 25
0
1
C: 7
9
0
1
E: 4
A: 5

- Note:
- The actual Yes/No questions are also now specified in the diagram.
- The questions structure is equivalent to the coding structure.
- The lower frequency elements have longer codes.
- (Not obvious)—length of code is inversely related to the log of frequency to base 2.

- Thus the Huffman algorithm assigns a binary code to each “letter”/”event”/”square” etc.
- Given the frequencies of the letters, the procedure gives an expected code length, or the expected number of questions to get to a letter.

- The Huffman Algorithm is good, giving a relatively low expected code length. But this code length is not the shortest.
- In his remarkable theorem Shannon showed what the shortest possible code length, or the smallest number of questions, in expectation, is:

- $\mathrm{H} = -\sum_k p_k \log p_k$

- Or, Entropy!

- Loosely speaking, this is because the “depth” of the tree to a leaf from the top is proportional to the log (to the base 2) of the frequency.
- Thus the optimal code length for each outcome k is proportional to $-\log\ p_k$; the higher the probability, the shorter the optimal code length.
- Hence the expected code length is the expectation of $\log\ p_k$, or

- $\mathrm{H} = -\sum_k p_k \log\ p_k$

Entropy and Information

- Let there be K possible events; k = 1, 2, …..K.
- Let the probability of event k be given by p_k
- What is the “uncertainty” of the above set up?
- Depends on what we mean by “uncertainty”.

- In an intuitive sense, if one of the events had probability of 1 and the others were of probability zero, we know for sure what is happening and uncertainty is low.
- Note that in this case $\mathrm{H} = -\sum_k \, p_k \log \, p_k \; = 0$.
- Again in an intuitive sense, the most uncertain case is where all probabilities are equal ie $p_k = {}^1/_K$. Anything is equally likely.
- Note that in this case entropy $\mathrm{H} = -\sum_k \, p_k \log \, p_k \; = \log K$. This rises with K, the more equi-probable states there are, the greater is the uncertainty.

- So it looks like $\mathrm{H} = -\sum_k p_k \log p_k$ has something to do with the uncertainty of the distribution p_k.
- But what exactly? It was the genius of Shannon to show the exact connection through an axiomatization.
- Let $\mathrm{H}(p_1, p_2, \ldots p_K)$ be any function that claims to represent the uncertainty or "information content" of a probability distribution $(p_1, p_2, \ldots p_K)$.
- Shannon proposed three axioms for such a function, two of which are straightforward, with a third which has the real meat.

- Axiom 1: H is a continuous function of its arguments.
- Axiom 2: If all the probabilities are equal then H is an increasing function of the number of states or events, K.

- Axiom 3. H satisfies the following “composition law”. Group the first two events into a single union event, of course with probability $\mathrm{m} = (p_1 + p_2)$. The conditional probabilities of event 1 and event 2, given that the union event, are of course $({}^{p_1}/_m, {}^{p_2}/_m)$. The axiom says that the information content of the overall distribution must be consistent with the information content of the union and conditional distributions

$$H(p_1, p_2, p_3, \dots p_K) = H(m, p_3, \dots p_K) + mH({}^{p_1}/_m, {}^{p_2}/_m)$$

- Shannon’s remarkable theorem is that these three axioms tie down the H function to

$$\mathrm{H} = -C \sum_{k} p_k \log\, p_k$$

where C is a constant.

- There are many other, perhaps more intuitive, routes to the functional form. Here is one.
- Let $I(p_k)$ be the “information content” or “surprise value” of an event k occurring. If the probability of an event is very high (the sun always rises in the East), then the surprise value of that event actually happening is very low. On the other hand if an event is very unlikely (it rains the Sahara), then the surprise value of that event happening is high.
- So $I(p_k)$ should be decreasing in its argument.

- Now consider the information value of two independent events k and l occurring together. The probability of the intersection event, since is the product of the two constituent probabilities, $p_k p_l$.
- It seems reasonable to specify that the surprise value of this joint event, since it is made up of two independent events, is the sum of the surprise content of the constitutive events.
- $I(p_k p_l) = I(p_k) + I(p_l)$

- The only function which satisfies these requirements (up to a constant C) is the log function:
- $I(p_k) = -\log p_k$
- If the surprise content (or information value) of a single event k is given by the above, then the surprise content of the whole distribution can be specified as the expectation of $I(p_k)$. So we are back to:
- $\sum_k p_k I(p_k) = -\sum_k p_k \log p_k = H$
- In other words, Entropy!

Information and Messages

- Suppose we get a message which tells us that the probability of event k has gone from p_k to q_k. What is the informational change? Clearly it is by how much the information value of the event occurring has changed:
- $I(p_k) - I(q_k) = -\log(p_k) - [-\log(q_k)] = \log({}^{q_k}/_{p_k})$
- Suppose the message now transforms each of the k probabilities. For each event there is an informational change. The overall change, in expectation, taken with respect to the new probabilities, is
- $D(q_k \,||\, p_k) = \sum_k q_k \log(q_k / p_k)$

- It can be shown that this is non-negative, with zero if and only if the two probability distributions are identical.

- This entity is known as the “Kullback-Leibler divergence”, or “informational divergence”, or “relative entropy” between the distribution q and the distribution p.

- $D(q_k \,||\, p_k) = \sum_k q_k \log(q_k / p_k)$

$= -\sum_k q_k \log(p_k) - (-\sum_k q_k \log(q_k))$

= H(q, p) – H(q)

“Cross Entropy of q and p” minus “Entropy of q”

- What is cross-entropy?
- Different ways of looking at it, but think of the coding perspective.
- We know there is an optimal code for probability distribution q, and one for p.
- What if we apply the optimal code for p to q? Clearly it is sub-optimal. The code length will be longer—the number of questions will be more than the smaller optimal number.
- Cross-entropy H(q, p) is the mistake we make in using the information from p to encode q. This perspective also turns out to be important for inequality measurement.

Applications to Inequality (I)
Classic Theil

- Let us start simply with the classic exposition by Theil.
- Let q be the actual vector of income shares:
- $q = (x_1/K\mu, x_2/K\mu, \ldots..)$
- Let p be the vector of equal income shares:
- $p = (1/K, 1/K, \ldots\ldots)$

- The Entropy of q is:
- $\mathrm{H} = -\sum_k q_k \log q_k$
- This is the smallest number of Yes/No questions (in expectation) needed to trace the origin of a dollar drawn at random.
- Similarly, the Entropy of p is the smallest number of Yes/No questions (in expectation) needed to trace the origin of a dollar drawn at random from an equal distribution.
- The fall in the number of questions needed is a measure of the degree of inequality.

- A little manipulation shows, as is well known, that this difference is the Theil index of inequality:
- $T = (1/K)\ \Sigma(x_i/\mu) \log (x_i/\mu)$
- The standard literature arrives at this through requiring decomposability of a particular type (along with scale independence and Principle of Transfers).
- The normative import, as opposed to operational convenience, of decomposability is not immediately obvious.
- But the “number of questions” interpretation of Theil might provide a stronger normative foundation for this commonly used measure.

Applications to Inequality (II)
Theil and Informational Divergence

- If q is the actual distribution of income shares and p is the distribution of equal income shares, then it can also be shown by easy manipulation that:
- $D(q_k \,||\, p_k) = \sum_k q_k \log(q_k / p_k) = (1/K)\, \Sigma(x_i/\mu) \log (x_i/\mu) = T$
- In other words, the Theil index of inequality is the Kullback-Leibler Divergence between the actual income shares and equal income shares.
- This is true mathematically. But think now of an information theory based interpretation of the Theil index.

- Recall that $D(q_k \,||\, p_k\,)$ is the "surprise" when the prior is p but it turns out that the actual is q.
- The axiomatic basis of this "surprise" is to be found in Shannon's formulation of information axioms.
- Recall that $D(q_k \,||\, p_k\,)$ also has interpretation a measure of "coding error" when we are using the optimal binary code ie the optimal sequence of "Yes/No" questions, for the equal distribution BUT we are applying it (incorrectly) for the actual, unequal distribution.
- These interpretations will bear further exploration and investigation.

Application to Inequality III
Theil and Testing for Fairness

- Start with the notion of surprise at an actual unequal distribution when the prior is that of an equal distribution.
- This perspective can be used to develop a test for "fairness" as follows.
- Imagine a helicopter which has a basket full of X dollars to drop, one by one, on to a population of K individuals.
- Let the probability that a dollar sticks to individual k be p_k
- Then p_k= 1/K is a specification of "equality of opportunity" in this world where no other attributes of individuals are specified.

- After the helicopter drop, we observe an actual distribution of dollars across individuals:
- $q_k;\ k = 1, 2, \ldots, K$
- On the basis of these observations we would like to test for the hypothesis that the process was fair, in other words,
- p_k= 1/K for all K.
- Consider the Likelihood Ratio test for this null hypothesis. It can be shown that the LR test statistic for this hypothesis under the multinomial process of the helicopter drop is in fact proportional to the Theil Index of Inequality:
- LR $\propto$ T

- This argument, and many extensions and variations on the theme, are presented in Kanbur and Snell, Economic Journal, 2019.
- The central take away is that the Theil Index of Inequality can be interpreted as a test statistic for the hypothesis of fairness, for a specified income generating processes and a specified notion of fairness within that process.
- Inequality indices, in this perspective, are no longer (just) ex post measures of welfare loss from inequality. They can be turned to use as test statistics for fairness.

Application to Inequality IV
Priors Other Than Equal Distribution

- Return to the idea of "surprise" as between a prior p and a posterior q.
- Up to now we have been using the equal distribution as the prior, p_k= 1/K
- But why should the equal distribution be the prior or the norm?
- Recent developments in the literature have highlighted that some unequal distributions could equally well be the norm, for example if they exhibit "equality of opportunity", in the sense that higher income as the result of higher "effort" is normatively justifiable, while income difference as the result of "circumstances" are not.

- Thus we need to consider the surprise at, or divergence of, the actual distribution relative to a prior distribution which is not necessarily the equal distribution.
- But we already have the framework in which to do this, since the Kullback-Leibler divergence measure is axiomatized for a general prior p and a general posterior q.
- If we can specify the norm or the prior with reference to specific ethical principles, then the divergence machinery, and its interpretations, can be brought into play.

- Going back to the thought experiment of tracing a dollar drawn at random back to its source through a series of Yes/No questions, the divergence measures how many more questions we will need using a process optimized for the norm distribution but (erroneously) used for the actual distribution.
- Of course this still leaves open the issue of what the norm vector should be. Applications of the Roemer (1998) approach have set the norm vector to be one where each individual income is replaced by the mean income of the circumstance group (eg race/gender, etc) to which the individual belongs.
- Hufe, Kanbur and Peichl (2018) have augmented this to include poverty into considerations of fairness. And there are many other variations.

- But whatever the norm vector, an information theoretic interpretation of the divergence between the actual and the norm holds out promise for anchoring normative intuitions, at least in complementary fashion to conventional approaches.

Application to Inequality V
Flipping Prior and Posterior

- Return to the basic Theil argument of divergence between the equal distribution p_k= 1/K as the prior and the actual distribution q as the posterior.
- What if we were to flip things around—think of the actual as prior and the equal distribution as posterior?
- Then it can be shown that the Kullback-Leibler divergence measure is simply the Mean Log Deviation (MLD):
- $D(q_k \,||\, p_k) = (1/K)\, \Sigma \log(\mu/x_i)$

- This is also known as Theil’s second measure, L, because Theil (1967) had derived it in terms of “flipping” the prior and posterior compared to his main index T.
- Despite L being the “second” measure relative to his main index T in Theil (1967), half a century later it is in fact MLD which is becoming the workhorse measure of inequality in applied work.
- This is primarily because MLD has nice “path independence” properties in sub-group decomposition compared to T. This follows from the mathematical fact that the within group component of inequality is a population share weighted sum of sub-group inequalities for MLD, but an income share weighted sum for T.

- This comparison led Shorrocks (1983) to argue that MLD “is the most satisfactory of the decomposable measures.”
- This is undoubtedly the case from an operational point of view. Many of us, myself included, use MLD in applied work for this reason.
- But what about the normative basis of the measures?
- T is the surprise using equality as the base and moving to the actual distribution—the distribution we actually have. MLD is the surprise in moving from the distribution we actually have to a hypothetical equal distribution.
- So in this perspective the choice depends not on decomposability properties per se, but on what is normatively more appropriate as the prior—equality or actual?

- Of course the issue arises no matter what the norm—equality, equality of opportunity, etc. The issue is rather what should be the prior—the norm or the actual?
- If the norm is the prior then T-type measures are appropriate.
- If the actual is the prior, then MLD-type measures are appropriate.
- Magdalou and Knock (2011) and Hufe, Kanbur and Peichl (2019) consider and apply alternative T-type and MLD-type measures.

Application to Inequality VI
Shannon Inequality of Opportunity

- Go back to the asking questions game.
- A dollar is drawn at random from q_k and we want to find the smallest number of Yes/No questions (in expectation) which will get us to the individual source. We know that this magnitude is the entropy of the distribution of dollars across individuals.
- Bu suppose now that before the income questions we can ask one more question, to which the machine will again give us a truthful answer.
- Is the individual a man or a woman?

- The answer to this question will be uninformative if the distributions for males and females are identical.
- But if the distributions differ and we know these distributions, then conditional on the answer to the first question we can design an optimal sequence of questions (the shortest code) for the case of a male dollar versus a female dollar.
- Intuitively, and it can be shown formally, the answer to the first question allows a lower expected number of questions to trace the dollar.
- This is a metric on the informational gain from knowing whether the dollar is male or female. It can also be treated as a metric on the degree of stratification in society.

- Various formal results in information theory now come into play.
- Let there be two random variables X and Y with a joint distribution. Let H(X) be the entropy of the marginal for X and H(Y) the entropy for the marginal for Y.
- Let $H\ (X|Y)$ and $H\ (Y|X)$ be the respective conditional entropies.
- Then:
- $I(X,Y) = H(X) - H(X|Y)$

is called the "mutual information" between X and Y.

- (Note: $I(X,Y) = I(Y,X)$)

- As noted before, knowing Y (the conditioning or “circumstance” variable) reduces the number of questions needed to trace a randomly drawn dollar X to its source.

- Following Dos Santos and Wiener (2019), an information theoretic measure of Inequality of Opportunity, or Shannon Inequality of Opportunity (SIOP), can be defined as:

- $SIOp = \frac{H(X) - H(X|Y)}{H(X)}$

- This can be compared to Roemer Inequality of Opportunity (RIOP), based on between circumstance group MLD, MLD_B:

- $RIOp = \frac{MLD_B}{MLD}$

- SIOp does not depend at all on the support of the random variables X and Y . Transformation of the supports does not change SIOp . But, except for proportionate scaling, $RIOp$ does change with transformations of the support.

- If the probability distributions of income are identical across types, SIOp is zero. In this case, RIOp is also zero. However, RIOp is zero whenever the income means of the types are identical.
- Roemer discusses strong and weak notions of equality of opportunity and this is related to the above point.
- A mean preserving spread of income distribution within each type will reduce RIOp, because MLD_B will remain unchanged while overall MLD will increase. will go up. An entropy increasing spread of income distribution within each type which maintains entropy of the income distribution overall will reduce SIOp.

- Comparisons of $SIOp$ and $RIOp$, with particular attention to their comparative normative foundations, are an interesting area for further research.

Generalizations

- Additivity is at the core of Shannon information entropy:
- $I(p_k p_l) = I(p_k) + I(p_l)$
- But what if information is not additive in this way. What if
- $I(p_k p_l) = I(p_k) + I(p_l) + (1 - \propto) I(p_k) I(p_l)$
- This leads to what is known as “Tsallis Entropy”
- $S = \frac{c}{\propto - 1}(1 - \sum_k p_k^{\propto})$
- The exploration of these generalized forms is still in its infancy in microeconomics, although in inequality measurement we do have “generalized entropy” measures of inequality.

Conclusion

- We use the label “generalized entropy measures of inequality” all the time, especially in applied work.
- But these days we seldom engage with the entropy and information theory roots of these measures. We used to do more of this some decades ago.
- There is increasing application of information theoretic concepts in microeconomics, including for example in Sims’s 2003 Nobel Prize winning paper on Rational Inattention, or Matejka and McKay’s recent AER 2015 paper on the foundations of the multinomial logit model.

- I have argued that such (re) engagement will be greatly beneficial to the literature on inequality measurement.
- It will illuminate possible normative foundations of inequality measures, and it will (re) open new (old) directions for research.

Selected References

- Cowell, F. A. 1980. Generalized entropy and the measurement of distributional change. *European Economic Review* 13, 147—159.
- Dos Santos, Paulo, and Noe Wiener. 2019. "Indices of Informational Association and Analysis of Complex Socio-Economic Systems."
- Hufe, Paul, Ravi Kanbur and Andreas Peichl. 2018. "Measuring unfair Inequality: Reconciling Equality of Opportunity and Freedom from Poverty." CEPR Discussion Paper, No. 12989.
- Kanbur, Ravi and Andy Snell. 2019. "Inequality Indices as Tests of Fairness" *Economic Journal*, forthcoming, https://doi.org/10.1111/ecoj.12637
- Golan, Amos. 2018. *Foundations of Info-Metrics: Modeling, Inference, and Imperfect Information*. Oxford University Press.
- Magdalou, B., and R. Nock. 2011. "Income distributions and decomposable divergence measures". *Journal of Economic Theory* 146 (6): 2440–2454.
- Matejka and Alisdair McKay. 2015. "Rational Inattention to Discrete Choices: A New Foundation for the Multinomial Logit Model." *American Economic Review*, 105(1): 272–298
- Roemer, John and Alain Trannoy. 2015. "Equality of opportunity," in A. Atkinson and F. Bourguignon (eds.), *Handbook of Income Distribution, vol.2A* , Amsterdam: Elsevier-NorthHolland,
- Shannon, C. E. 1948. "A Mathematical Theory of Communication." *Bell System Technical Journal* 27: 370-423.
- Shorrocks, Anthony F. 1983. "The Class of Additively Decomposable Inequality Measures." *Econometrica*, Vol. 48, No. 3 (Apr., 1980), pp. 613-625
- Sims, Christopher A. 2003. "Implications of Rational Inattention." *Journal of Monetary Economics* 50 (3): 665–90.
- Theil, Henry. 1967. *Economics and Information Theory*. North Holland.

Thank You!