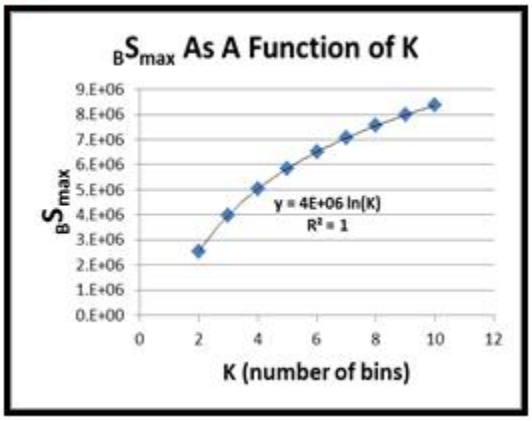
NOTE TO FILE:
Garvin H Boyle
Dated: R2:180119

Entropy in a Histogram, and in ABMs

Frontispiece

Figure 01 – Table of values – $_BS_{max}$ as K varies

A	K	α	$_BS_{max}$
3628800	1	3628800	0.00000
3628800	2	1814400	2515284.62976
3628800	3	1209600	3986628.81662
3628800	4	907200	5030562.09360
3628800	5	725760	5840298.11141
3628800	6	604800	6501899.37618
3628800	7	518400	7061274.24634
3628800	8	453600	7545825.91877
3628800	9	403200	7973230.01610
3628800	10	362880	8355555.32853
	3628800		



Substituting equations [12] and [15] into equation [14] I get:

$$ {}_BS_{index} = \frac{{}_BS}{{}_BS_{max}} = \frac{\left[\ln(A!) - \sum_{i=1}^{K}\left(ln(a_i!)\right)\right]}{A\ln(K)} \qquad [16] $$

If I wish, in resolving the ln/ln division, I can write this as:

$$ {}_SS_{index} = -\sum_{i=1}^{K}(p_i \log_K(p_i)) \qquad [28] $$

Table of Contents

1 - References

A. A. Drăgulescu and V.M. Yakovenko (2000) "Statistical mechanics of money", Eur. Phys. J. B 17, 723-729.
B. 140218 Email from Yakovenko R1.pdf
C. 140409 NTF Discussion With Dr Yakovenko R1.pdf
D. 150527 PPR - Definition of EI R17.pdf
E. 180118 NTF Custom Functions in Excel R3.pdf
F. 180210 NTF Shannon Vs Boltzmann R3.pdf
G. 180214 NTF Entropy and Units of Measure R5.pdf
H. 180214 NTF FactLn() and GammaLn() R1.pdf
I. V.M. Yakovenko (2010) "Statistical mechanics approach to the probability distribution of money", arXiv:1007.5074v1 [q-fin.ST] 28 July 2010.
J. V.M. Yakovenko (2010) "Statistical mechanics of money, debt, and energy consumption", arXiv: 1008.2179v1 [q-fin.ST] 12 Aug 2010.
K. V.M. Yakovenko (2012) "Applications of statistical mechanics to economics: Entropic origins of the probability distributions of money, income, and energy consumption", arXiv: 1204.6483v1 [q-fin.ST] 29 Apr 2012.
L. 180115 XLS Entropy In a Histogram R2.xlsx
M.

2 - Background

This series of diary notes is a rework of a set of notes partially completed in 2013-2014. In 2000 Dr Victor Yakovenko and his student (Drăgulescu) published a set of eight capital exchange models which have come to be known as the BDY model (for Benatti-Drăgulescu-Yakovenko). Later, on his website, Dr Yakovenko had produced a demonstration of rising entropy in his BDY models, and I decided to do the same in my own agent-based models (ABMs). This was all with the goal of understanding the role of the Maximum Entropy Principle (MEP) and Maximum Entropy Production Principle (MEPP) in ABMs such as ModEco or PSoup. That study and the associated diary notes were set aside for a while as I studied Odum's MPP. Now, in 2018, I want to review, update, and complete this set of diary notes.

Ref A is the paper in which I first saw the BDY models described, and Ref B is a series of email in which I discuss entropy with a friend, with comments from Dr Yakovenko. I had opportunity in April 2014 to meet with Dr Yakovenko in Toronto, and Ref C captures my discussion with him, and a set of thoughts that occurred to me over the four hours as I drove home again.

Ref D is a draft paper in which I develop a measure of entropy as exhibited in operating ABMs similar to the BDY models. This series of NTFs is being (re-)written in support of that document.

Ref E is a technical diary note describing how to write a custom function in MS Excel. This skill is needed to pursue a study of entropy in ABMs.

Ref F is a diary note in which I use the combinatorial multinomial coefficient and the basic form of "Stirling's Approximation" of ln(A!) to derive Shannon's equation for entropy [1] from Boltzmann's equation [2], as described briefly in Yakovenko's papers at Refs I, J and K.

I then have two definitional formulae that can be used to calculate the entropy of any histogram (preferably of a conserved quantity) in an agent-based model. One (equation [1]) is based directly on the work of Shannon, and the other (equation [2]) is based on the work of Boltzmann.

Based on Shannon:

$${}_{S}S = {}_{S}C \times \left[-\sum_{i=1}^{k} (p_i \times \ln[p_i])\right]$$	[1]

Based on Boltzmann:

$${}_{B}S = {}_{B}C \times [k_B \ln(\Omega)]$$	[2]

Where ${}_{S}C$ and ${}_{B}C$ are scaling constants each associated with what I am calling the "Shannon regime" of equations, and the "Boltzmann regime". In the Boltzmann regime, ${}_{B}C$ would be Boltzmann's constant. I am not sure what ${}_{S}C$ would be in the Shannon regime.

In analytical terms, these two equations are related by first replacing Ω, the multiplicity of microstates, with the multinomial coefficient $\Omega = A!/\prod_{i=1}^{K}[a_i!]$, and second, by approximating the function ln(x!) with x ln(x) – x, a simple version of Stirling's approximation for ln(x!). This results in the analytical expression in equation [3], where the (≈) symbol is meant to imply equality over most of the domain of interest x ∈ [0, ∞].

Boltzmann in terms of Shannon:

$${}_{B}S \approx \left[\frac{{}_{B}C}{{}_{S}C}\right] \times A \times {}_{S}S$$	[3]

Unfortunately, these formulae diverge as x approaches zero, and that is the portion of the domain of most interest to me when studying ABMs. I have therefore decided, somewhat arbitrarily, to use the equation that comes from the Boltzmann regime when calculating entropy in ABMs. But, for the purposes of this diary note, I will still look at both of equations [1] and [2].

One confusing aspect of the Ref F note is the units of measure for the two formulae for entropy. In the Ref G diary note I examine the nature of the units of measure used for entropy in both thermodynamic theory and in information theory and come to two conclusions: (1) that all units

of measure for entropy are dimensionless numbers; and (2) that the formula I used for calculating entropy in an ABM is most closely associated with nats (units of measure from information theory) when the number of agents is very large, but diverges from nats when the number of agents is of more practical size. I therefore name my units of measure hnats, and the prepended 'h' stands for 'histogram'.

In Ref H I try to closely examine the two functions ln(x!) and GammaLn(x) that are intimately connected to my definition of entropy in histograms, and in agent-based models.

All of this was preparatory work needed to undertake the work on this diary note, and it is all background to the Ref D draft paper.

Ref L is an MS Excel spreadsheet associated with this diary note.

3 - Purpose

The purpose of this diary note is to present a consolidated description of how to calculate the entropic index of some conserved quantity within an agent-based model. I am going to do this using the two different base equations for entropy in each of the two "regimes" of thought explored at Ref F xxx.

4 - Discussion

In this diary note I am going to develop the equations for S_{max} in each of the two regimes of thought explored in the Ref F diary note, and, following that, develop the equations for S_{index} in each of the two regimes. I do this noting that it is my intention to only use the Boltzmann regime in the future, but I want to explore the connection between the two at the same time. Why? Because, when (1) the number of agents in a model is large and (2) when the histograms are not sparsely populated, then the distortion introduced by Stirling's approximation disappears and the two regimes of thought should be identical in practical outcome.

4.1 - S_{max}?

In my various readings about entropy, I admit that I have only come across the concept of "maximum entropy" in reference to closed systems. With that in mind, I am uncertain how these ideas apply to open systems. So, all of the discussion in this diary note is meant explicitly to apply to closed systems, but its applicability to open systems has yet to be determined.

In a closed system, in which some quantity is conserved in all transactions and merely transferred from agent to agent, then a history of the collective ownership of that quantity can be constructed as a time series of histograms, and there is a maximum possible value of entropy associated with each such time series. An economic version of the second law of thermodynamics would tell us that the entropy of such a time series should rise towards that maximum value and hover in that vicinity. But, there are several design characteristics of an ABM and of the entropy measurement regime that might affect the actual value of entropy calculated for any histogram, or for any time series of histograms. Using "money" as the conserved quantity, as an example, these design characteristics include:

- Maximum debt allowed per agent;
- Maximum wealth allowed per agent;
- The number of wealth bins in each histogram;
- The width of the bins in each histogram;
- The number of empty bins beyond the lower and upper bounds on the wealth of agents;
- How edge effects are handled.

Rather than trying to address all possible variations for all possible ABMs at one, I need some simplifying assumptions. Then, when I get answers for one type of ABM, under one set of assumptions, I may be able to expand the ideas to apply to others. Since I know that some of these assumptions work for "Model I" of the EiLab application, I will start there.

I was introduced to Yakovenko's so-called BDY models (see Ref A) when I was looking for an explanation of the entropic origins of the distribution of wealth in my ModEco models. Eventually a friend asked me to reproduce the BDY models from Ref A, and I did that in models A through H of the EiLab C++ application. Those models have a single constraint on the boundary of wealth – a minimum wealth allowed per agent. I decided to build a variant on the BDY model with this one simple addition, an upper bound on wealth. So, instead of a single boundary, I had two boundaries – so I called it a doubly-bounded BDY model. It is "Model I" of EiLab.

In a closed system, the entropy will be at a maximum value when all bins of the histogram contain precisely the same number of agents. That is to say, when the distribution of the quantity has a uniform distribution. Practical constraints on the operation of the model may prevent the entropy of the system from ever approaching this maximum value, and that is when you get entropy-induced distributions that are other than uniform. Yakovenko has written a number of papers about such non-uniform distributions of wealth. However, when you add the second upper boundary, the model may then approach a uniform distribution as it runs. This, then, is where I need to start in trying to understand the existence and role of S_{max} for some time series of histograms that may be produced by an ABM.

For the following discussion:

- Let A be the number of agents in the ABM. A is an integer greater than zero. The number of agents is conserved, so A is constant.
- Each agent can hold wealth in units of \$1 – no fractional dollars being considered.
- Assume that the total wealth of all agents, W, is conserved throughout a run of the model. I.e. W is constant.
- Let W_{min} be the minimum wealth any agent can hold, and let W_{max} be the maximum wealth any agent can hold.
- Let K be the number of equally sized (\$1) non-overlapping bins that cover the range of possible wealth of agents $[W_{min}, W_{max}]$. K is a constant. The bins can be enumerated using the variable i, having values from 1 to K, and the number of agents in each bin is a_i.

We then get the equation:

$$A = \sum_{i=1}^{K} (a_i)$$	[4]

In any histogram, for given A and K, entropy is at a maximum when the distribution is uniform and the number of agents in each bin is equal to A/K agents per bin. Let α be defined as:

$$\alpha \equiv \frac{A}{k}$$	[5]

4.2 - $_{B}S_{max}$ – Boltzmann Regime

The multinomial coefficient is:

$$\Omega = \frac{A!}{\prod_{i=1}^{K}[a_i!]}$$	[6]

Inserting this into equation [2] we get the expression for entropy of a histogram in the "Boltzmann regime", as discussed in Ref F.

This is my standard equation for the definition of entropy of a histogram, **as it stood at the end of discussion in Ref F** (I am going to adjust the scaling factor BC as I work through this):

$$_{B}S = {}_{B}C \times \left[\ln(A!) - \sum_{i=1}^{K}\left(ln(a_i!)\right)\right]$$	[7]

where $_{B}C$ is a dimensionless scaling constant.
Substituting equation [5] into equation [7] we get:

$$_{B}S_{max} = {}_{B}C \times \left[\ln(A!) - \sum_{i=1}^{K}\left(ln(\alpha!)\right)\right]$$	[8]

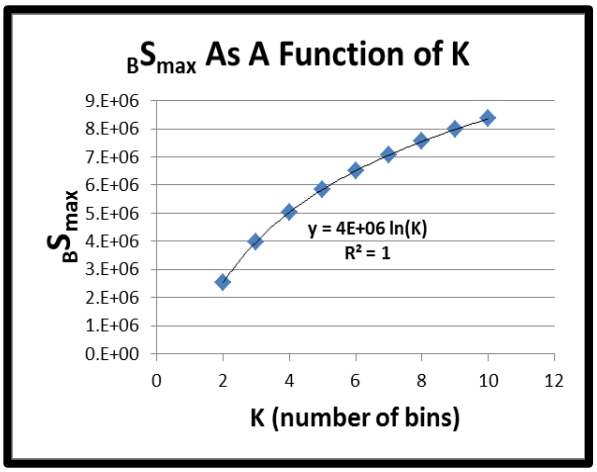
Which resolves to:

$$_{B}S_{max} = {}_{B}C \times [\ln(A!) - K\ln(\alpha!)]$$	[9]

At the Ref L spreadsheet I explored the implications of this formula, as shown in the following figures. In Figure 01 I show a table in which A was held constant, and K was allowed to vary. I chose a value for A=10! such that A/K is always a whole integer, and so the value of α! can be calculated exactly using the factorial function, and ln(α!) is therefore easily calculated.

Figure 01 – Table of values – $_BS_{max}$ as K varies

A	K	α	$_BS_{max}$
3628800	1	3628800	0.00000
3628800	2	1814400	2515284.62976
3628800	3	1209600	3986628.81662
3628800	4	907200	5030562.09360
3628800	5	725760	5840298.11141
3628800	6	604800	6501899.37618
3628800	7	518400	7061274.24634
3628800	8	453600	7545825.91877
3628800	9	403200	7973230.01610
3628800	10	362880	8355555.32853
	3628800		



The table and graph in Figure 01 lead to two interesting facts, only one of which is obvious. First, note that the rising value of $_BS_{max}$ has a trend line (generated from MS Excel using "least squares" techniques) that is logarithmic with an R_2 value of precisely 1. This means that if I divide $_BS_{max}$ by ln(K) I should get a constant that is independent of K. My dimensionless scaling factor now has some use. Suppose I make the scaling factor $_BC$ equal to 1/ln(K) then $_BS_{max}$ should be constant, independent of K. Making that change to equation [9] gives me:

$$_BS_{max} = \frac{1}{\ln(K)} \times [\ln(A!) - K\ln(\alpha!)] \qquad [10]$$

When the divisor is applied to the two terms of the numerator, this resolves to:

$$_BS_{max} = \frac{\ln(A!) - K\ln(\alpha!)}{\ln(K)} = \log_K(A!) - K\log_K(\alpha!) \qquad [11]$$

But, there is another interesting thing to be found here. When I use the MS Excel spreadsheet to divide the computed values of $_BS_{max}$ by ln(K), the constant that I get is very very close to A.

See Figure 02 for a version of the above table of values with BSmax / ln(K) added, and the difference between that number and A.

Figure 02 – Table of values – $_BS_{max}$ as K varies – modulated by 1/ln(K)

A	K	α	$_BS_{max}$	$_BS_{max}$/ln(K)	Diff
3628800	1	3628800	0.00000	#DIV/0!	#DIV/0!
3628800	2	1814400	2515284.62976	3628788.66	11.3382
3628800	3	1209600	3986628.81662	3628785.93	14.0691
3628800	4	907200	5030562.09360	3628783.49	16.5073
3628800	5	725760	5840298.11141	3628781.24	18.7551
3628800	6	604800	6501899.37618	3628779.13	20.8654
3628800	7	518400	7061274.24634	3628777.13	22.8698
3628800	8	453600	7545825.91877	3628775.21	24.7892
3628800	9	403200	7973230.01610	3628773.36	26.6382
3628800	10	362880	8355555.32853	3628771.57	28.4276
	3628800				

Now, some distortion must come into the calculations of $_BS_{max}$ due to the necessary use of Stirling's approximation for all calculations for which α is greater than 170. That would be all of them. But, in all cases the modulated value of $_BS_{max}$ is within 0.001% of the value of A. This leads to the final equation for $_BS_{max}$:

$$ {}_BS_{max} = A \qquad [12]$$

Now, I do not have the mathematical skill to show the analytical arguments that would prove that A is the correct answer. But, practically speaking, this seems to be true. The small difference seems to be described by equation [13] with R^2 value of 0.9998, again based on a 'least squares' type of trend line produced by MS Excel.

$$ Diff = -0.0616K^2 + 2.7283K + 8.7822 \qquad [13]$$

4.3 - S_{index}?

I arbitrarily define the entropic index to be calculated using the following formula.

$$ S_{index} \equiv \frac{S}{S_{max}} \qquad [14]$$

It only makes sense to define an entropic index in those instances for which there exists a maximum value of calculated entropy S_{max}. Assuming that such a thing is possible, there are, it seems, two types of maximum entropy that might be considered as the benchmark on which the index is built. Consider a time-series of histograms produced by a run of some ABM. We can identify two possible circumstances:

- **Pragmatic maximum possible** – the configuration of the model does not allow for a uniform distribution of the conserved quantity to appear, so the largest possible value for entropy that can be achieved falls short of the theoretical maximum that is possible. For example, in most of the BDY models studied by Dr Yakovenko, a uniform distribution is not possible, so the maximum entropy that can be exhibited is less than the theoretical maximum.
- **Theoretical maximum possible** – the configuration of the model enables the possibility of a uniform distribution of the conserved quantity, and the theoretical maximum can be achieved. Mode I of EiLab is able to show a uniform distribution, if the right amount of money is put into the model, and then the theoretical maximum is achievable.

Using the pragmatic maximum as the benchmark, every model closed model will evolve to have an entropic index close to 1. But then I will not be able to evaluate the economic activity of one model against another. It is better to use the theoretical maximum as the benchmark for the index. Then most models will evolve to operate at some index less than 1, and only the most effective models will evolve to operate at an index of 1.

What I have in mind here is a measure of the ability of an economic ABM to self-organize to operate at maximum entropy. My assumption is that those that function at an index close to 1 are more effective engines at turning a distribution of money into economic work that circulates money and brings about economic benefits. So, it seems to me that having a concept of a theoretical S_{max} is a good thing to develop. For the 'Boltzmann regime' ${}_{B}S_{max} = A$ would seem to be the theoretical maximum.

4.4 - ${}_{B}S_{index}$ – Boltzmann Regime

So, I am ready to define the index for the Boltzmann regime, but first I am going to redefine ${}_{B}S$ to include the new dimensionless scaling factor ${}_{B}C = 1/\ln(K)$.

So the entropy of a histogram ${}_{B}S$ is now calculated using this formula (compare with equation [7] above):

$$ {}_{B}S = \frac{1}{\ln(K)} \times \left[\ln(A!) - \sum_{i=1}^{K} \left(ln(a_i!) \right) \right] \qquad [15] $$

Substituting equations [12] and [15] into equation [14] I get:

$$ {}_{B}\boldsymbol{S}_{\boldsymbol{index}} \equiv \frac{{}_{B}\boldsymbol{S}}{{}_{B}\boldsymbol{S}_{\boldsymbol{max}}} = \frac{\left[\ln(\boldsymbol{A}!) - \sum_{\boldsymbol{i}=1}^{\boldsymbol{K}} \left(\boldsymbol{ln}(\boldsymbol{a_i}!) \right) \right]}{\boldsymbol{A} \ln(\boldsymbol{K})} \qquad \textbf{[16]} $$

4.5 - $_SS_{max}$ – Shannon Regime

This is my Shannon-regime equation for the definition of entropy of a histogram, **as it stood at the end of discussion in Ref F** (I am also going to adjust the scaling factor $_SC$ as I work through this):

$$_SS = {}_SC \times \left[-A \sum_{i=1}^{K} (p_i \ln(p_i))\right] \quad [17]$$

where $_SC$ is a dimensionless scaling factor. This is the equivalent of equation [7] for the Boltzmann regime.

p_i is defined here as.

$$p_i \equiv \frac{a_i}{A} \quad [18]$$

Substituting [18] into [17] we get:

$$_SS = {}_SC \times \left[-A \sum_{i=1}^{K} \left(\frac{a_i}{A} \ln\left(\frac{a_i}{A}\right)\right)\right] \quad [19]$$

To find $_SS_{max}$, substituting α for a_i,this becomes:

$$_SS_{max} = {}_SC \times \left[-A \sum_{i=1}^{K} \left(\frac{\alpha}{A} \ln\left(\frac{\alpha}{A}\right)\right)\right] \quad [20]$$

It will require a few careful steps to resolve this. First, move the constant α / A outside of the sum, and cancel the A with A:

$$_SS_{max} = {}_SC \times \left[-A \frac{\alpha}{A} \sum_{i=1}^{K} \left(\ln\left(\frac{\alpha}{A}\right)\right)\right] = {}_SC \times \left[-\alpha \sum_{i=1}^{K} \left(\ln\left(\frac{\alpha}{A}\right)\right)\right] \quad [21]$$

Second, within the ln() function replace α with A / K:

$$ {}_{S}S_{max} = {}_{S}C \times \left[-\alpha \sum_{i=1}^{K} \left(\ln\left(\frac{1}{K}\right) \right) \right] = {}_{S}C \times \left[\alpha \sum_{i=1}^{K} (\ln(K)) \right] \quad [22] $$

Now, resolve the sum of K constants:

$$ {}_{S}S_{max} = {}_{S}C \times [\alpha K \ln(K)] = {}_{S}C \times A \ln(K) \quad [23] $$

Again, as for equation [9] in the Boltzmann regime, this rises proportional to ln(K), so a scaling factor that is the reciprocal of that will cause ${}_{S}S_{max}$ to be constant.

$$ {}_{S}S_{max} = \frac{1}{\ln(K)} \times A \ln(K) = A \quad [24] $$

So, I get the same equation for S_{max} in the Boltzmann regime (equation [12]) and in the Shannon regime (equation [24]). In the Boltzmann regime, there was a non-analytical step in which some intuition had to be applied when interpreting a graph. In the Shannon regime, the results were developed analytically.

4.6 - ${}_{S}S_{index}$ – Shannon Regime

So, again, I am ready to define the index for the Shannon regime, but first I am going to redefine ${}_{S}S$ to include the new dimensionless scaling factor ${}_{S}C = 1/\ln(K)$.

So the entropy of a histogram in the Shannon regime ${}_{S}S$ is now calculated using this formula (compare with equation [17] above):

$$ {}_{S}S = \frac{1}{\ln(K)} \times \left[-A \sum_{i=1}^{K} (p_i \ln(p_i)) \right] \quad [25] $$

Define ${}_{S}S_{index}$ as:

$$ {}_{S}S_{index} \equiv \frac{{}_{S}S}{{}_{S}S_{max}} = \frac{[-A \sum_{i=1}^{K} (p_i \ln(p_i))]}{A \ln(K)} \quad [26] $$

Cancelling the As I get:

$$_{S}S_{index} = \frac{[-\sum_{i=1}^{K}(p_i \ln(p_i))]}{\ln(K)}$$	[27]

If I wish, in resolving the ln/ln division, I can write this as:

$$\boldsymbol{_{S}S_{index} = -\sum_{i=1}^{K}(p_i \log_K(p_i))}$$	**[28]**

5 - Summary

In either regime of calculations, I get a surprisingly compact formula for the entropic index. (See equations [16] and [28].)

In both cases, the maximum entropy possible is A, and the index is the entropy divided by A.

In both cases, the result is independent of the number of bins in the histogram, as long as the number of agents in the model does not change, and as long as the bin widths do not change.

The assumption that K is a constant has been relaxed, as it is of no consequence if it changes. However, the assumptions that A and W are constant has not been relaxed.