From count to measure, from Shannon to differential entropy

Claude Elwood Shannon[1], working on digital communications, proposed a magnitude defined on discrete random variables which was called entropy, *H*. *H* is a good indicator of uncertainty. This magnitude is well known and deeply studied.

The attempt of extend this magnitude to continuous random variables lead us to differential entropy, which is rejected as magnitude with the same meaning that *H*. The main drawback of differential entropy is that it is not invariant to variable transformations, so any differential entropy could be negative and What does negative entropy mean? Nothing, because *H*=0 means no uncertainty.

Let me review the uniform distribution. This pfd (probability distribution function) has two versions, one continuous and discrete the other, but both shares the same mathematical formalism. This distribution assigns the same probability to each possible outcome of the uniform variable.

Let *X* be an uniform random variable, let *R*=max(*X*)-min(*X*), then the entropy is $H_U = \log(R)$ If you want to know if this entropy is differential or Shannon you must see the nature of the number *R*. If it is not integer or if it has units then the entropy must be differential, if the number is integer and dimensionless there is a good chance for a discrete variable, and the meaning of entropy changes according to this choice. Or you arrives to the same scenario when you model the problem and decides if one varible is best suited for continuous or discrete representation. Here something is not working properly.

Suppose you have a sample with *n* IID uniform discrete variables. The *H* associated to this sample is $\log(R^n)$ and this could be interpreted as the log of the volume of the space of possibles outcomes, the volume of the uniform uncertainty.

Suppose you have a discrete uniform variable, but *R* is big and it would be useful express the outcome in some new unit so 1 new unit worths *w* old units and the new parameter is $R_{new} = \frac{R}{w}$ so $H_U = \log(R_{new}) = \log(R) - \log(w)$ And the catastrophic negative entropy arrives if $w > R$. It was expected, we go from integers to rationals, facing to the differential entropy.

At last, suppose a continuous uniform variable, measured in some unit. For the purpose of get rid off the dimensions of the argument of log, it is divided by 1 unit. This is a standard trick with the only consequence that the result of the computation is somehow referred to this unit . Then, if $H=0$ the length of $R=1$ unit, the magnitude of the entropy is 0 referred to units, and the uniform uncertainty occupies a volume of 1 unit.

[1] SHANNON, C. E. A Mathematical Theory of Communication. *The Bell System Technical Journal* **27,** 379–423, 623–656 (1948).

The practical consecuence is: All the outcomes of this random variable have the same number at the units (and frozen digits all over to the left). But nothing is say about the digits to the right of units. There is a volume of uniform uncertainty of magnitude 1 unit.

If $R=\frac{1}{10}$ then $H=-\log(10)$ and the practical consecuence is that the first digit at the right of the decimal separator contributes its entropy, wich is supposed 0 in the example, that is a fixed digit.

You can repeat this procedure an infinite number of times, making the volume of uniform uncertainty $R=0$ and $H=-\infty$, then you reach absolute certainty.

So the value $H=0$ is "equidistant" from the absolute certainty ($H=-\infty$) and the absolute uncertainty ($H=\infty$).

From a practical point of view $H=0$ means "I am sure about the units, but not about tenths", and a negative value of entropy could be interpreted as "I am sure about this number up to the n-th decimal place, but not about the others".

This interpretation of entropy has consequences

- Entropy is a continous measure of uncertainty, referred to the unitary scale (which is chosen by the observer for continuous variables, and has a "natural" scale, counts, for discrete variables)
- The underpinning of a not parametric method for the estimation (mesure) of H from real samples.This is necessary because the parametric estimation of H heavily depends on the pdf attributed to data.
- It may be that the problem of measuring uncertainty and the problem of choosing the right scale for the measure are inseparable and that they are, in some way, solved at once by H