tripleC 1(1): 1-10, 2003
ISSN 1726-670X
http://www.triple-c.at

Spencer-Brown vs. Probability and Statistics: Entropy Analysis of Subjective Randomness

Julio Michael Stern

Dept.of Applied Mathematics, Institute of Mathematics and Statistics, University of São Paulo. Rua do Matão 1010, Cidade Universitária. 05508-090, São Paulo, Brazil. jstern@ime.usp.br

Abstract: *This article analyses the role of entropy in Bayesian Statistics, focusing on its use as a tool for detection, recognition and validation of eigen-solutions. "Objects as eigen-solutions" is a key metaphor of the cognitive constructivism epistemological framework developed by the philosopher Heinz von Foerster. Special attention is given to some objections to the concepts of probability, statistics and randomization posed by George Spencer-Brown, a figure of great influence in the field of radical constructivism.*

Keywords: Bayesian Statistics, Cognitive constructivism, Eigen-solutions, Maximum entropy, Objective-subjective complementarity, Objectice inference, Randomization, Subjective randomness.

Acknowledgement: The author is grateful for the support of the Department of Applied Mathematics of the Institute of Mathematics and Statistics of the University of São Paulo, FAPESP - Fundação de Amparo à Pesquisa do Estado de São Paulo, and CNPq - Conselho Nacional de Desenvolvimento Científico e Tecnológico (grant PQ-306318-2008-3). The author is also grateful for the help of several of his professional collegues.

1. Introduction

In several already published articles, I defend the use of Bayesian Statistics in the epistemological framework of cognitive constructivism. In particular, I show how the FBST -The Full Bayesian Significance Test for precise hypotheses - can be used as a tool for detection, recognition and validation of eigen-solutions, see (Borges and Stern, 2007), (Pereira et al., 2008), and (Stern, 2003, 2004, 2006, 2007a, 2007b, 2008a, 2008b, 2009). "Objects as eigen-solutions" is a key metaphor of cognitive constructivism as developed by the Austrian-American philosopher Heinz von Foerster, see (Foerster, 2003).

In Statistics, specially in the design of statistical experiments, Randomization plays a role which is in the very core of objective-subjective complementarity, a concept of great significance in the epistemological framework of cognitive constructivism as well as in the theory of Bayesian statistics. The pivotal role of randomization in a well designed statistical experiment is that of a decoupling operation used to sever illegitimate functional links, thus avoiding spurious associations, breaking false influences, separating confounding variables, etc, see (Stern, 2008a) and (Colla and Stern, 2008).

The use of randomization in Statistics is an original idea of Charles Saunders Peirce and Joseph Jastrow, see (Peirce and Jastrow, 1884) and (Hacking, 1988). Randomization is now a standard requirement for many scientific studies. In (Stern 2007b, 2008a) I consider the position of C.S.Peirce as a forerunner of cognitive constructivism, based on the importance, relevance and coherence of his philosophical and scientific work. Among his several contributions, the introduction of randomization in statistical design stands indubitably out. In future articles, I hope to further expand the analysis of the role of Bayesian Statistics in cognitive constructivism and provide other interesting applications. I shall herein analyze some objections to the concepts of probability, statistics and randomization posed by George Spencer-Brown, a figure of great influence in the field of radical constructivism.

In what follows, section 2 corresponds to the first part of this article's title and elaborates upon "the case of Spencer-Brown vs. probability and statistics". Corresponding to the second part of the title, section 3 provides the "the testimony of entropy on subjective randomness". Section 4 offers an objective perspective on randomness,

through an entropy based informational analysis. Section 5 presents our final conclusions.

2. The Case of Spencer-Brown Against Probability and Statistics

In (Spencer-Brown, 1953a, 1953b, 1957), Spencer-Brown analyzed some apparent paradoxes involving the concept of randomness, and concluded that the language of probability and statistics was inappropriate for the practice of scientific inference. In subsequent work, (Spencer-Brown, 1969), he reformulates classical logic using only a generalized *nor* operator (marked *not-or*, unmarked *or*), that he represents à la mode of Charles Saunders Peirce or John Venn, by a graphical boundary or distinction mark, see (Carnielli, 2009), (Edwards, 2004), (Kauffmann, 2001, 2006), (Meguire, 2003), (Peirce, 1880) and (Sheffer, 1913).

Making (or arbitrating) distinctions is, according to Spencer-Brown, the basic (if not the only) operation of human knowledge, an idea that has either influenced or been directly explored by several authors in the radical constructivist movement. The following quotations, from (Spencer-Brown, 1957, p.23,p.66,p.105), are typical arguments used by Spencer-Brown in his rejection of probability and statistics:

"Retroactive reclassification of observations in one of the scientist's most important tools, and we shall meet it again when we consider statistical arguments."

"We have found so far that the concept of probability used in statistical science is meaningless in its own terms; but we have found also that, however meaningful it might have been, its meaningfulness would nevertheless have remained fruitless because of the impossibility of gaining information from experimental results, however significant. This final paradox, in some ways the most beautiful, I shall call the Experimental Paradox."

"The essence of randomness has been taken to be absence of pattern. But what has not hitherto been faced is that the absence of one pattern logically demands the presence of another. It is a mathematical contradiction to say that a series has no pattern; the most we can say is that it has no pattern that anyone is likely to look for. The concept of randomness bears meaning only in relation to the observer: If two observers habitually look for different kinds of pattern they are bound to disagree upon the series which they call random."

Several authors concur, at least in part, with my opinion about Spencer-Brown's technical analysis of probability and statistics, see (Flew, 1959), (Falk and Konold, 2005), (Good, 1958)a and (Mundle, 1959). In Section 3, I carefully explain why I disagree with it. In some of my arguments, which are are based on information theory and the notion of entropy, I dissent from Spencer-Brown's interpretation of measures of order-disorder in sequential signals. In (Atkins 1984), (Attneave, 1959), (Dugdale, 1996), (Krippendorff, 1986) and (Tarasov, 1988) some of the basic concepts in this area are reviewed with a minimum of mathematics.

I also disapprove some of Spencer Brown's proposed methodologies to detect "relevant" event sequences, that is, his criteria to "mark distinct patterns" in empirical observations. My objections have a lot in common with the standard caveats against *ex post facto* "fishing expeditions" for interesting outcomes, or simple *post hoc* "sub-group analysis" in experimental data banks. This kind of retroactive or retrospective data analyses is considered a questionable statistical practice, and pointed as the culprit of many misconceived studies, misleading arguments and mistaken conclusions. The literature on statistical methodology for clinical trials has been particularly keen in warning against this kind of practice. See (Tribble, 2008) and (Wang et al., 2007) for two interesting papers addressing this specific issue and published in high impact medicine journals less than a year before I wrote this text. When consulting for pharmaceutical companies or advising in the design of statistical experiments, I often find it useful to quote Conan Doyle's Sherlock Holmes, in The Adventure of Wisteria Lodge:

"Still, it is an error to argue in front of your data. You find yourself insensibly twisting them around to fit your theories."

Finally, I am also suspicious or skeptical about the intension behind some applications of Spencer-Brown's research program, including the use of extrasensory empathic perception for coded message communication, exercises on object manipulation using paranormal powers, etc. Unable to reconcile his psychic research program with statistical science, Spencer-Brown had no regrets in disqualifying the later, as he clearly stated in the prestigious scientific journal *Nature*, see (Spencer-Brown, 1953b,p.594):

[On telepathy:] *"Taking the psychical research*

data (that is, the residuum when fraud and incompetence are excluded), I tried to show that these now threw more doubt upon existing presuppositions in the theory of probability than in the theory of communication."

[On psychokinesis:] *"If such an 'agency' could thus 'upset' a process of randomizing, then all our conclusions drawn through the statistical tests of significance would be equally affected, including the conclusions about the 'psychokinesis' experiments themselves. (How are the target numbers for the die throws to be randomly chosen? By more die throws?) To speak of an 'agency' which can 'upset' any process of randomization in an uncontrollable manner is logically equivalent to speaking of an inadequacy in the theoretical model for empirical randomness, like the luminiferous ether of an earlier controversy, becomes, with the obsolescence of the calculus in which it occurs, a superfluous term."*

Spencer-Brown's conclusions, including his analysis of probability, were considered to be controversial (if not unreasonable or extravagant) even by his own colleagues at the Society of Psychical Research, see (Scott, 1958), (Soal et al., 1958). It seems that current research in this area, even not being free (or afraid) of criticism, has abandoned the path of naïve confrontation with statistical science, see (Atmanspacher, 2005) and (Ehm, 2005). For additional comments, see (Henning, 2006), (Kaptchuk and Kerr, 2004), (Utts, 1991), and (Wassermann, 1955).

Curiously, Charles Saunders Peirce and his student Joseph Jastrow, who introduced the idea of randomization in statistical trials, also struggled with some of the very same dilemmas faced by Spencer-Brown, namely, the eventual detection of distinct patterns or seemingly ordered (sub)strings in a long random sequence. Peirce and Jastrow did not have at their disposal the heavy mathematical artillery I have quoted in the previous paragraphs. Nevertheless, as experienced explorers that are not easily lured, when traveling in desert sands, by the mirage of a misplaced oasis, these intrepid pioneers were able to avoid the conceptual pitfalls that lead Spencer-Brown so far astray. For more details see (Stern, 2008a), (Hacking, 1988), (Peirce and Jastrow, 1884) and (Bonassi et al., 2008, 2009) and (Dehue, 1997).

As stated in the introduction, the cognitive constructivist framework can be supported by the FBST, a non-decision theoretic formalism drawn from Bayesian statistics. The FBST was conceived as a tool for validating objective knowledge of eigen-solutions and, as such, can be easily integrated to the epistemological framework of cognitive constructivism in scientific research practice. Contrasting our distinct views of cognitive constructivism, it is not at all surprising that I have come to conclusions concerning the use of probability and statistics, and also to the relation between probability and logic, that are fundamentally different from those of Spencer-Brown.

3. Pseudo, Quasi and Subjective Randomness

The focus of the present section are the properties of "natural" and "artificial" random sequences. The implementation of probabilistic algorithms require good random number generators, (RNGs). These algorithms include: numerical integration methods such as Monte Carlo or Markov Chain Monte Carlo (MCMC); evolutionary computing and stochastic optimization methods such as genetic programming and simulated annealing; and also, of course, the efficient implementation of randomization methods.

The most basic random number generator replicates i.i.d. (independent and identically distributed) random variables uniformly distributed in the unit interval, $[0, 1[$. From this basic uniform generator one gets a uniform generator in the d-dimensional unit box, $[0, 1[^d$, and, from the later, non-linear generators for many other multivariate distributions, see (Hammersley and Handscomb, 1964) and (Ripley, 1987).

Historically, the technology of random number generators was developed in the context of Monte Carlo methods. The nature of Monte Carlo algorithms makes them very sensitive to correlations, auto-correlations and other statistical properties of the random number generator used in its implementation. Hence, in this context, the statistical properties of "natural" and "artificial" random sequences came to close scrutiny. For the aforementioned historical and technological reasons, Monte Carlo methods are frequently used as a benchmark for testing the properties of these generators. Hence, although Monte Carlo methods proper lie outside the scope of this article, we shall keep them as a standard application benchmark in our discussions.

The clever ideas and also the caveats of engineering good random number generators

are in the core of many paradoxes found by Spencer-Brown. The objective of this section is to explain the basic ideas behind these generators and, in so doing, avoid the conceptual traps and pitfalls that took Spencer-Brown analyses so much off course.

3.1. Random Number Generators

The concept of randomness is usually applied to a variable or a process (to be generated or observed) involving some uncertainty. The following definition is presented at (Hammersley and Handscomb, 1964,p.10):

"A random event is an event which has a chance of happening, and probability is a numerical measure of that chance."

Monte Carlo, and several other probabilistic algorithms, require a random number generator. With the last definition in mind, engineering devices based on sophisticated physical processes have been built in the hope of offering a source of "true" random numbers. However, these special devices were cumbersome, expensive, not portable nor universally available, and often unreliable. Moreover, practitioners soon realized that simple deterministic sequences could successfully be used to emulate a random generator, as stated in the following quotes (our emphasis) at (Hammersley and Handscomb, 1964,p.26) and (Ripley, 1987,p.15):

"For electronic digital computer it is most convenient to calculate a sequence of numbers one at a time as required, by a completely specified rule which is, however, so devised that no **reasonable** *statistical test will detect any significant departure from randomness. Such a sequence is called* pseudorandom*. The great advantage of a specified rule is that the sequence can be exactly reproduced for purposes of computational checking."*

"A sequence of pseudorandom *numbers (U_i) is a deterministic sequence of numbers in $[0,1]$ having the same* **relevant** *statistical properties as a sequence of random numbers."*

Many deterministic random emulators used today are Linear Congruential Pseudo-Random Generators (LCPRG), as in the following example:

$$x_{i+1} = (ax_i + c) \bmod m \ ,$$

where the multiplier a, the increment c and the modulus m should obey the conditions: (i) c and m are relatively prime; (ii) $a-1$ is divisible by all prime factors of m; (iii) $a-1$ is a multiple of 4 if m is a multiple of 4. LCPRG's are fast and easy to implement if m is taken as the computer's word range, 2^s, where s is the computer's word size, typically $s = 32$ or $s = 64$. The LCPRG's starting point, x_0, is called the seed. Given the same seed the LCPG will reproduce the same sequence, a very convenient feature for tracing, debugging and verifying application programs.

However, LCPRG's are not an universal solution. For example, it is trivial to devise some statistics that will be far from random, see (Marsaglia, 1968). There the importance of the words **reasonable** and **relevant** in the last quotations becomes clear: For most practical applications these statistics are irrelevant. LCPRG's can also exhibit very long range auto-correlations and, unfortunately, these are more likely to affect long simulated time series required in some special applications. The composition of several LCPRG's by periodic seed refresh may mitigate some of these difficulties, see (Ripley 1987). LCPRG's are also not appropriate to some special applications in cryptography, see (Boyar, 1989). Current state of the art generators are given in (Matsumoto et al., 1992, 1998).

3.2. Chance is Lumpy - Random and Quasi-Random Generators

"Chance is Lumpy" is Robert Abelson's First Law of Statistics, stated in (Abelson, 1995,p.xv).

The probabilistic expectation is a linear operator, that is, $E(Ax+b) = AE(x)+b$, where x in random vector and A and b are a determined matrix and vector. The Covariance operator is defined as $\mathrm{Cov}(x) = E((x-E(x)) \otimes (x-E(x)))$. Hence, $\mathrm{Cov}(Ax+b) = A\mathrm{Cov}(x)A'$. Therefore, given n i.i.d. scalar variables, $x_i \,|\, \mathrm{Var}(x_i) = \sigma^2$, the variance of their mean, $m = (1/n)\mathbf{1}'x$, is given by σ^2/n. So, in this case, the mean value converge to the expected value at a rate of $1/\sqrt{(n)}$.

Quasi-random sequences are deterministic sequences built not to emulate random sequences, as pseudo-random sequences do, but to achieve faster convergence rates. For d-dimensional quasi-random sequences, an appropriate measure of fluctuation, called discrepancy, only grows at a rate of $\log(n)^d$, hence growing much slower than $\sqrt{(n)}$. Therefore, the convergence rate corresponding to quasi-random sequences, $\log(n)^d/n$, is much faster than the one corresponding to (pseudo) random sequences,

$\sqrt(n)/n$. Figure 1 allows the visual comparison of typical (pseudo) random (left) and quasi-random (right) sequences in $[0,1[^2$. By visual inspection we see that the points of the quasi-random sequence are more "homogeneously scattered" that is, they do not "clump together", as the point of the (pseudo) random sequence often do.

Let us consider an axis-parallel rectangles in the unit box,

$$R = [a_1, b_1[\times [a_2, b_2[\times \ldots [a_d, b_d[\subseteq [0,1[^d \ .$$

The discrepancy of the sequence $s_{1:n}$ in box R, and the overall discrepancy of the sequence are defined as

$$D(s_{1:n}, R) = n\text{Vol}(R) - |s_{1:n} \cap R| \ ,$$

$$D(s_{1:n}) = \sup_{R \in [0,1[^d} |D(s_{1:n}, R)| \ .$$

It is possible to prove that the discrepancy of the Halton-Hammersley sequence, defined next, is of order $O(\log(n)^{d-1})$, see (Matousek, 1991, ch.2).

Halton-Hammersley sets: Given $d-1$ distinct prime numbers, $p(1), p(2), \ldots p(d-1)$, the i-th point, x^i, in the Halton-Hammersley set, $\{x^1, x^2, \ldots x^n\}$, is for $i = 1 : n-1$,

$$i = a_0 + p(k)a_1 + p(k)^2 a_2 + p(k)^3 a_3 + \ldots \ ,$$

$$r_{p(k)}(1) = \frac{a_0}{p(k)} + \frac{a_1}{p(k)^2} + \frac{a_2}{p(k)^3} + \ldots \ .$$

$$x^i = \left[i/n, r_{p(1)}(i), r_{p(2)}(i), \ldots r_{p(d-1)}(i)\right]' \ ,$$

That is, the $(k+1)$-th coordinate of x^i, $x^i_{k+1} = r_{p(k)}(i)$, is obtained by the bit (or digit) reversal of i written in $p(k)$-ary or base $p(k)$ notation.

The Halton-Hammersley set is a generalization of van der Corput set, built in the bidimensional unit square, $d = 2$, using the first prime number, $p = 2$. The following example, from (Hammersley and Handscomb, 1964,p.33) and (Guenther, 2003,p.117), builds the 8-point van der Corput set, expressed in binary and decimal notation.

```
function x= corput(n,b)
% size n base b v.d.corput set
m=floor(log(n)/log(b));
u=1:n;  D=[];
for i=0:m
  d= rem(u,b);
  u= (u-d)/b;
  D= [D; d];
end
x=((1./b').^(1:(m\ma1)))*D;
```

Decimal		Binary	
i	$r_2(i)$	i	$r_2(i)$
1	0.5	1	0.1
2	0.21	10	0.01
3	0.75	11	0.11
4	0.125	100	0.001
5	0.625	101	0.101
6	0.375	110	0.011
7	0.875	111	0.111
8	0.0625	1000	0.0001

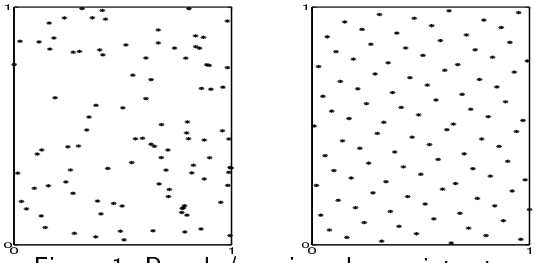

Figure 1: Pseudo/quasi-random point sets

Quasi-random sequences, also known as low-discrepancy sequences, can substitute pseudo-random sequences in some applications of Monte Carlo methods, achieving higher accuracy with less computational effort, see (Merkel, 2005), (Okten, 1999) and (Sen et al., 2006). Nevertheless, since by design the points of a quasi-random sequence tend to avoid each other, strong (negative) correlations are expected to appear. In this way, the very reason that can make quasi-random sequences so helpful, can ultimately impose some limits to their applicability. Some of these problems are commented in (Morokoff, 1998,p.766):

"First, quasi-Monte Carlo methods are valid for integration problems, but may not be directly applicable to simulations, due to the correlations between the points of a quasi-random sequence. ... A second limitation: the improved accuracy of quasi-Monte Carlo methods is generally lost for problems of high dimension or problems in which the integrand is not smooth."

3.3. Subjective Randomness Paradoxes

When asked to look at patterns like those in Figure 1, many subjects perceive the quasi-random set as "more random" than the (pseudo) random set. How can this paradox be explained? This was the topic of many psychological studies in the field of subjective randomness. The quotation in the next paragraph is from one of these studies, (Falk and Konold, 1997,p.306), emphasis are ours:

"One major source of confusion is the fact that randomness involves two distinct ideas: **process** *and* **pattern**, *(Zabell, 1992). It is natural to think of randomness as a process that generates unpredictable outcomes, this is a stochastic process according to (GellMann, 1994). Randomness of a* **process** *refers to the* **unpredictability** *of the individual event in the series (Lopes, 1982, 1987). This is what Spencer Brown 9Spencer-Brown, 1957) calls* **primary randomness**. *However, one usually determines the randomness of the process by means of its output, which is supposed to be* **patternless**. *This kind of randomness refers, by definition, to a sequence. It is labeled* **secondary randomness** *by Spencer Brown. It requires that all symbol types, as well as all ordered pairs (diagrams), ordered triplets (trigrams)... n-grams in the sequence be equiprobable. This definition could be valid for any n only in infinite sequences, and it may be approximated in finite sequences only up to ns much smaller than the sequence's length. The entropy measure of randomness is based on this definition, see ch.1 and 2 of (Attneave, 1959).*

These two aspects of randomness are closely related. We ordinarily expect outcomes generated by a random process to be patternless. Most of them are. Conversely, a sequence whose order is random supports the hypothesis that it was generated by a random mechanism, whereas sequences whose order is not random cast doubt on the random nature of the generating process."

Spencer-Brown was intrigued by the apparent incompatibility of the notions of primary and secondary randomness. The apparent collision of these two notions generates several interesting paradoxes, taking Spencer-Brown to question the applicability of the concept of randomness in particular and probability and statistical analysis in general, see (Spencer-Brown, 1953a, 1953b, 1957) and also (Flew, 1959), (Good, 1958), (Mundle, 1959), (Henning, 2006), (Kaptchuk, 2004), (Utts, 1991), (Wassermann, 1955) and (Tversky, 1971). In fact, several subsequent psychological studies were able to confirm that, for many subjects, the intuitive or common-sense perception of primary and secondary randomness are quite discrepant. However, a careful mathematical analysis can reconcile the two notions of randomness. These topics are discussed in this section.

The relation between the joint and conditional entropy for a pair of random variables,

$$H(i,j) = H(j) + H(i\,|\,j) = H(i) + H(j\,|\,i)\ ,$$

motivates the definition of first, second and higher order entropies, see section 4. These are defined over the distribution of words of size m in a string of letters from an alphabet of size a.

$$H_1 = \sum\nolimits_j p(j) \log p(j)\ ,$$

$$H_2 = \sum\nolimits_{i,j} p(i)p(j\,|\,i) \log p(j\,|\,i)\ ,$$

$$H_3 = \sum\nolimits_{i,j,k} p(i)p(j\,|\,i)p(k\,|\,i,j) \log p(k\,|\,i,j)\ \ldots$$

It is possible to use these entropy measures to assess the disorder or lack of pattern in a given finite sequence, using the empirical probability distributions of single letters, pairs, triplets, etc. However, in order to have a significant empirical distribution of m-plets, any possible m-plet must be well represented in the sequence, that is, the word size, m, is required to be very short relative to the sequence log-size, that is, $m << \log_a(n)$.

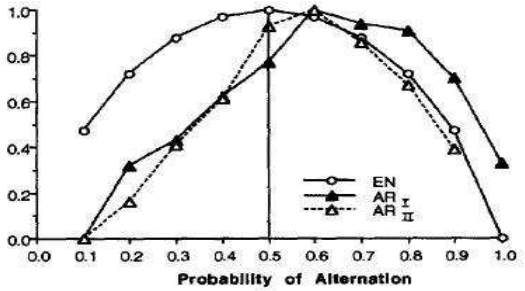

Figure 2: Pixel alternation - H_2-entropy, EN vs. apparent randomness, AR.

In the article (Falk and Konold, 1997), Figure 2 displays the typical perceived or apparent randomness of Boolean (0-1) bit sequences, represented as black-and-white pixel grids, versus the second order entropy of the same strings, see also (Attneave, 1959). Clearly, there is a remarkable bias of the apparent randomness relative to the entropic measure.

This effect is known as the *gambler's fallacy* when betting on *cool spots*. It consists of expecting the random sequence to "compensate" finite average fluctuations from expected values. This effect is also described in (Falk and Konold, 1997,p.303): *" When people invent superfluous explanations because they perceive patterns in random phenomena, they commit what is known in statistical parlance as Type I error. The other way of going awry, known as Type H error, occurs when one dismisses stimuli showing some regularity as random. The numerous randomization studies in which participants generated too many*

alternations and viewed this output as random, as well as the judgments of overalternating sets as maximally random in the perception studies, were all instances of type II error in research results."

Of course, some gamblers exhibit the opposite behavior, preferring to bet on *hot spots*, expecting the same fluctuations to reoccur. These effects are the consequence of a perceived coupling, by a negative or positive correlation or other measure of association, between non overlapping segments that are in fact supposed to be decoupled, uncorrelated or have no association, that is, to be independent. For a statistical analysis, see (Bonassi et al., 2008, 2009). A possible psychological explanation of the gambler's fallacy is given by the constructivist theory of Jean Piaget, see (Piaget, 1951), (Falk and Konold, 1997,p.316 in which any "lump" in the sequence is (miss) perceived as non-random order:

"In analogy to Piaget's operations, which are conceived as internalized actions, perceived randomness might emerge from hypothetical action, that is, from a thought experiment in which one describes, predicts, or abbreviates the sequence. The harder the task in such a thought experiment, the more random the sequence is judged to be."

The same hierarchical decomposition scheme used for higher order conditional entropy measures can be adapted to measure the disorder or patternless of a sequence, relative to a given subject's model of "computer" or generation mechanism. In the case of a discrete string, this generation model could be, for example, a deterministic or probabilistic Turing machine, a fixed or variable length Markov chain, etc. It is assumed that the model is regulated by a code, program or vector parameter, θ, and outputs a data vector or observed string, x. The hierarchical complexity measure of such a model emulates the Bayesian prior and conditional likelihood decomposition, $H(p(\theta, x)) = H(p(\theta)) + H(p(x \mid \theta))$, that is, the total complexity is given by the complexity of the program plus the complexity of the output given the program. This is the starting point for several complexity models, like Andrey Kolmogorov, Ray Solomonoff and Gregory Chaitin's computational comlexity models, Jorma Rissanen's Minimum Description Length (MDL), and Chris Wallace and David Boulton's Minimum Message Length (MML). All these alternative complexity models can also be used to successfully reconcile the notions of primary and secondary randomness, showing that they are asymptotically equivalent, see (Chaitin, 1975, 1988), (Kac, 1983), (Kapur, 1989), (Rissanen, 1989) and (Wallace, 2005).

4. Entropy and Some Generalizations

Entropy is the cornerstone concept of the preceding section, used as a central idea in the understanding of order and disorder in stochastic processes. Entropy is the key that allowed us to unlock the mysteries and solve the paradoxes of subjective randomness, making it possible to reconcile the notions of unpredictability of stochastic process and patternless of randomly generated sequences. Similar entropy based arguments reappear, in more abstract, subtle or intricate forms, in the analysis of technical aspects of Bayesian statistics like, for example, the use of prior and posterior distributions and the interpretation of their informational content. This section gives a short review covering the definition of entropy, its main properties, and some of its most important uses in mathematical statistics.

The origins of the entropy concept lay in the fields of Thermodynamics and Statistical Physics, but its applications have extended far and wide to many other phenomena, physical or not. The entropy of a probability distribution, $H(p(x))$, is a measure of uncertainty (or impurity, confusion) in a system whose states, $x \in \mathcal{X}$, have $p(x)$ as probability distribution. We follow closely the presentation in the following references. For the basic concepts, see (Dugdale, 1996), (Csiszar, 1974), (Khinchin, 1953) and (Renyi, 1961, 1970).

4.1. Boltzmann-Gibbs-Shannon Entropy

If $H(p(x))$ is to be a measure of uncertainty, it is reasonable that it should satisfy the following list of requirements. For the sake of simplicity, we present the theory in finite spaces.

1) If the system has n possible states, $x_1, \ldots x_n$, the entropy of the system with a given distribution, $p_i \equiv p(x_i)$, is a function H such that $H = H_n(p_1, \ldots, p_n)$.

2) H is a continuous function.

3) H is a function symmetric in its arguments.

4) The entropy is unchanged if an impossible state is added to the system, that is, $H_n(p_1, \ldots p_n) = H_{n+1}(p_1, \ldots p_n, 0)$.

5) The system's entropy is minimal and null when the system is fully determined, that is, $H_n(0, \ldots, 0, 1, 0, \ldots 0) = 0$.

6) The entropy is maximal when all states are equally probable, that is, $\{ \frac{1}{n}\mathbf{1} \} = \arg\max H_n$.

7) Maximal entropy increases with the number of states, i.e., $H_{n+1}\left(\frac{1}{n+1}\mathbf{1}\right) > H_n\left(\frac{1}{n}\mathbf{1}\right)$.

8) Entropy is an extensive quantity, i.e., given two independent systems, with distributions p e q, the entropy of the composite system is additive, i.e., $H_{nm}(r) = H_n(p) + H_m(q)$, $r_{i,j} = p_i\, q_j$.

The Boltzmann-Gibbs-Shannon measure of entropy, $H_n(p) = -I_n(p) = -\sum_{i=1}^{n} p_i \log(p_i) = -\,\mathsf{E}_i \log(p_i)$, $0 \log(0) \equiv 0$, satisfies requirements (1) to (8), and is the most usual measure of entropy. In Physics it is usual to take the logarithm in Napier base, while in Computer Science it is usual to take base 2 and in Engineering it is usual to take base 10. The opposite of the entropy, $I(p) = -H(p)$, the Neguentropy, is a measure of Information available about the system.

For the Boltzmann-Gibbs-Shannon entropy we can extend requirement 8, and compute the composite Neguentopy even without independence. Writing $q_j^i = \mathsf{Pr}(j \mid i)$, we get,

$$I_{nm}(r) = I_n(p) + \sum_{i=1}^{n} p_i\, I_m(q^i) \ .$$

If we add this last identity as item number 9 in the former list of requirements, we have a characterization of Boltzmann-Gibbs-Shannon entropy, see (Khinchin, 1953), (Renyi, 1970).

Like many important concepts, this measure of entropy was discovered and re-discovered several times in different contexts, and sometimes the uniqueness and identity of the concept was not immediately recognized. A well known anecdote refers the answer given by von Neumann, after Shannon asked him how to call a "newly" discovered concept in Information Theory. As reported by Shannon in (Tribus, 1971,p.180):

"My greatest concern was what to call it. I thought of calling it information, but the word was overly used, so I decided to call it uncertainty. When I discussed it with John von Neumann, he had a better idea. Von Neumann told me, You should call it entropy, for two reasons. In the first place your uncertainty function has been used in statistical mechanics under that name, so it already has a name. In the second place, and more important, nobody knows what entropy really is, so in a debate you will always have the advantage."

4.2. Csiszar's Divergence

In order to check that requirement (6) is satisfied, we can use (with $q \propto 1$) the following lemma:

Lemma: Shannon Inequality
If p and q are two distributions over a system with n possible states, and $q_i \neq 0$, then the Information of p Relative to q, $I_n(p,q)$, is positive, except if $p = q$, when it is null,

Proof: If φ is a convex function, Jensen inequality holds, $\mathsf{E}(\varphi(x)) \geq \varphi(\mathsf{E}(X))$. Taking $\varphi(t) = t\ln(t)$ and $t_i = p_i/q_i$, we get $\mathsf{E}_q(t) = 1$, and $I_n(p,q) = \sum q_i t_i \log t_i \geq 1 \log(1) = 0$.

Shannon's inequality motivates the use of the Relative Information as a measure of (non symmetric) "distance" between distributions. In Statistics this measure is known as the Kullback-Leibler distance. The denominations Directed Divergence or Cross Information are used in Engineering. The proof of Shannon inequality motivates the following generalization of divergence:

Definition: Csiszar's φ-divergence.
Given a convex function φ, where $0\varphi(0/0) = 0$ and $0\varphi(c/0) = c \lim_{t\to\infty} \varphi(t)/t$,

$$d_\varphi(p,q) = \sum_{i=1}^{n} q_i\, \varphi\left(\frac{p_i}{q_i}\right) \ .$$

For example, we can define the absolute and the quadratic divergence as
$Ab(p,q) = \sum \frac{|p_i - q_i|}{q_i}$, for $\varphi(t) = |t-1|$; and
$\chi^2(p,q) = \sum \frac{(p_i - q_i)^2}{q_i}$, for $\varphi(t) = (t-1)^2$.

5. Final Remarks

The objections raised by Spencer-Brown against probability and statistics, analyzed in sections 1 and 2, are somewhat simplistic and stereotypical, possibly explaining why they had little influence outside a small circle of admirers, most of them related to the radical constructivism movement. However, arguments very similar to those used to demystify Spencer-Brown's misconceptions and elucidate its misunderstandings, reappear in more subtle or abstract forms in the analysis of far more technical matters like, for example, the use and interpretation of prior and posterior distributions in Bayesian statistics.

In this article, entropy is presented as a cornerstone concept for the precise analysis and a key idea for the correct understanding of several important topics in probability and statistics. This understanding should help to clear the way for establishing Bayesian statistics as a preferred toll for scientific inference in mainstream cognitive constructivism.

References

Abelson,R.P. *Statistics as Principled Argument*. LEA: Hillsdale, NJ, *1995*.

Atkins,P.W. *The Second Law*. The Scientific American Books: NY, *1984*.

Atmanspacher,H. Non-Physicalist Physical Approaches. Guest Editorial. *Mind and Matter, 2005*, 3, 2, 3-6.

Attneave,E. *Applications of Information Theory to Psychology: A summary of basic concepts, methods, and results*. Holt, Rinehart and Winston: NY, *1959*.

Bonassi,F.V.; Stern,R.B.; Wechsler,S. The Gambler's Fallacy: A Bayesian Approach. *MaxEnt2008, AIP Conf.Proc. 1073*,8-15.

Bonassi,F.V.; Nishimura,R.; Stern,R.B. In Defense of Randomization: A Subjectivist Bayesian Approach. *MaxEnt 2009, AIP Conf.Proc, 1193*, 32-39.

Borges,W.; Stern,J.M. The Rules of Logic Composition for the Bayesian Epistemic e-Values. *Logic Journal of the IGPL, 2007, 15, 5-6*, 401-420.

Boyar,J. Inferring Sequences Produced by Pseudo-Random Number Generators. *Journal of the ACM, 1989, 36, 1*, 129-141.

Carnielli,W. Formal Polynomials and the Laws of Form. In *Dimensions of Logical Concepts*, Béziau,J.Y.; Costa-Leite,A. eds. Coleção CLE, 54, UNICAMP: Campinas, Brazil, *2009*.

Chaitin,G.J. Randomness and Mathematical Proof. *Scientific American, 1975, 232*, 47-52.

Chaitin,G.J. Randomness in Arithmetic. *Scientific American, 1988, 259*, 80-85.

Colla,E.; Stern,J.M. Sparse Factorization Methods for Inference in Bayesian Networks. *AIP Conf.Proc, 1073*, 136-143.

Csiszar,I. Information Measures. *7th Prage Conf.of Information Theory, 1974, 2*, 73-86.

Dehue,T. Deception, Efficiency, and Random Groups: Psychology and the Gradual Origination of the Random Group Design. *Isis, 1997, 88, 4*, 653-673.

Dugdale,J.S. *Entropy and Its Physical Meaning*. Taylor and Francis: London, *1996*.

Edwards,A.W.F. *Cogwheels of the Mind. The Story of Venn Diagrams*. Johns Hopkins University Press, *2004*.

Ehm,W. Meta-Analysis of Mind-Matter Experiments: A Statistical Modeling Perspective. *Mind and Matter, 2005, 3, 1*, 85-132.

Falk,R.; Konold,C. Making Sense of Randomness: Implicit Encoding as a Basis for Judgment. *Psychological Review, 1997, 104, 2*, 301-318.

Falk,R.; Konold,C. Subjective Randomness. *Encyclopedia of Statistical Sciences, 2005, 13*, 8397-8403.

Flew,A. Probability and Statistical Inference by G.Spencer-Brown (review). *The Philosophical Quarterly, 1959, 9, 37*, 380-381.

Foerster, H.von. *Understanding Understanding: Essays on Cybernetics and Cognition*. Springer Verlag: NY, *2003*.

Gell'Mann,M. *The Quark and the Jaguar: Adventures in the Simple and the Complex*. Freeman, NY, *1994*.

Good,I.J. Probability and Statistical Inference by G.Spencer-Brown (review). *The British Journal for the Philosophy of Science, 1958, 9, 35*, 251-255.

Günther,M.; Jüngel,A. *Finanzderivate mit MATLAB. Mathematische Modellierung und numerische Simulation* . Vieweg Verlag: Wiesbaden, *2003*, 117.

Hacking,I. Telepathy: Origins of Randomization in Experimental Design. *Isis, 1988, 79, 3*, 427-451.

Hammersley,J.M.; Handscomb,D.C. *Monte Carlo Methods*. Chapman and Hall: London, *1964*.

Henning,C. *Falsification of Propensity Models by Statistical Tests and the Goodness-of-Fit Paradox*. Technical Report, Department of Statistical Science, University College, London. *2006*.

Kac,M. What is Random? *American Scientist, 1983, 71*, 405-406.

Khinchin,A.I. *Mathematical Foundations of Information Theory*. Dover: NY, *1953*.

Kaptchuk,T.J., Kerr,C.E. Commentary: Unbiased Divination, Unbiased Evidence, and the Patulin Clinical Trial. *International Journal of Epidemiology, 2004, 33*, 247-251.

Kapur,J.N. *Maximum Entropy Models in Science and Engineering*. John Wiley: New Delhi, *1989*.

Kauffman,L.H. The Mathematics of Charles Sanders Peirce. *Cybernetics and Human Knowing, 2001, 8*, 79-110.

Kauffman,L.H. *Laws of Form: An Exploration in Mathematics and Foundations. 2006*.
`http://www.math.uic.edu/~kauffman/Laws.pdf`

Krippendorff,K. *Information Theory: Structural Models for Qualitative Data*. Quant.Applic.in the Social Sciences. v.62, *1986*.

Lopes,L.L. Doing the Impossible: A Note on Induction and the Experience of Randomness. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 1982, 8*, 626-636.

Lopes,L.L.; Oden,G.C. Distinguishing Between Random and Nonrandom Events. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 1987, 13*, 392-400.

Marsaglia,G. Random Numbers Fall Mainly in the Planes. *Proceedings of the National Academy of Sciences, 1968, 61*, 25-28.

Matoušek,J. *Geometric Discrepancy*. Springer: Berlin, *1991*.

Matsumoto,M.; Nishimura,T. Mersenne Twister: A 623-dimensionally Equidistributed Uniform Pseudorandom Number Generator. *ACM Trans.Model.Comput Simul., 1998, 8*, 3-30.

Matsumoto,M.; Kurita,Y. Twisted GFSR Generators. *ACM Trans.Model.Comput.Simul.. 1992,2*, 179-194; *1994,4*, 254-266.

Meguire,P. Discovering Boundary Algebra: A Simple Notation for Boolean Algebra and the Truth Functions. *Int. J. General Systems, 2003, 32*, 25-87.

Merkel,R. *Analysis and Enhancements of Adaptive Random Testing*. Swinburne University. Melburne, Australia, *2005*.

Morokoff,W.J. Generating Quasi-Random Paths for Stochastic Processes. *SIAM Review, 1998, 40, 4,* 765-788.

Mundle,C.W.K. Probability and Statistical Inference by G.Spencer-Brown (review). *Philosophy, 1959, 34, 129,* 150-154.

Peirce,C.S.; Jastrow,J. On small Differences of Sensation. *Memoirs of the National Academy of Sciences, 1884, 3,* 75-83.

Peirce,C.S. A Boolean Algebra with One Constant. *1880.* In Hartshorne,C.; Weiss,P.; Burks,A. edts. *Collected Papers of Charles Sanders Peirce.* InteLex: Charlottesville, *1992, 4,* 12-20.

Pereira,C.A.B.; Stern,J.M. Evidence and Credibility: Full Bayesian Significance Test for Precise Hypotheses. *Entropy Journal, 1999, 1,* 69-80.

Pereira,C.A.B.; Wechsler,S.; Stern,J.M. Can a Significance Test be Genuinely Bayesian? *Bayesian Analysis, 2008 3, 1,* 79-100.

Piaget,J.; Inhelder,B. *The Origin of the Idea of Chance in Children.* Norton: NY, *1975.*

Renyi,A. On Measures of Entropy and Information. *Proc. 4-th Berkeley Symp. on Math Sats. and Prob. 1961, V-I,* 547-561.

Renyi,A. *Probability Theory.* North-Holland: Amsterdam, *1970.*

Ripley,B.D. *Stochastic Simulation.* Wiley: NY, *1987.*

Rissanen,J. *Stochastic Complexity in Statistical Inquiry.* World Scientific: NY, *1989.*

Scott,C. G.Spencer-Brown and Probability: A Critique. *J.Soc. Psychical Research, 1958,* 39, 217-234.

Sen,S.K.; Samanta,T.; Reese,A. (2006). Quasi Versus Pseudo Random Generators: Discrepancy, Complexity and Integration-Error Based Comparisson. *Int.J. of Innovative Computing, Information and Control,* 2, 3, 621-651.

Sheffer,H.M. A Set of Five Independent Postulates for Boolean Algebras, with Application to Logical Constants. *Trans. Amer. Math. Soc., 1913, 14,* 481-488.

Soal,S.G.; Stratton,F.J.; Thouless,R.H. (1953). Statistical Significance in Psychical Research. *Nature, 1958, 172,* 594.

Spencer-Brown,G. Statistical Significance in Psychical Research. *Nature, 1953, 172,* 154-156.

Spencer-Brown,G. Answer to Soal et al. *Nature, 1953, 172,* 594-595.

Spencer-Brown,G. *Probability and Scientific Inference.* Longmans Green: London, *1957.*

Spencer-Brown,G. *Laws of Form.* Allen and Unwin: London, *1969.*

Stern,J.M. Significance Tests, Belief Calculi, and Burden of Proof in Legal and Scientific Discourse. Laptec-2003, *Frontiers in Artificial Intelligence and its Applications, 2003, 101,* 139–147.

Stern,J.M. Paraconsistent Sensitivity Analysis for Bayesian Significance Tests. SBIA'04, in LNAI, *2004, 3171,* 134–143.

Stern,J.M. Language, Metaphor and Metaphysics: The Subjective Side of Science. Tech.Rep. MAC-IME-USP-06-09. *2006.*

Stern,J.M. Cognitive Constructivism, Eigen-Solutions, and Sharp Statistical Hypotheses. *Cybernetics and Human Knowing, 2007, 14, 1,* 9-36.

Stern,J.M. Language and the Self-Reference Paradox. *Cybernetics and Human Knowing,* 2007, 14, 4, 71-92.

Stern,J.M. Complex Structures, Modularity and Stochastic Evolution. Tech.Rep. IME-USP-MAP-07-01, *2007.*

Stern,J.M. Decoupling, Sparsity, Randomization, and Objective Bayesian Inference. *Cybernetics and Human Knowing, 2008, 15, 2,* 49-68.

Stern,J.M. *Cognitive Constructivism and the Epistemic Significance of Sharp Statistical Hypotheses.* 28th MaxEnt, International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering, *2008.*

Stern,J.M. The Living and Intelligent Universe. MBR09 - The Internaternational Conference on Model-Based Reasoning in Science and Technology, Unicamp, Brazil, *2009.*

Tarasov,L. *The World is Built on Probability.* MIR: Moscow, *1988.*

Tribble,C. Industry-Sponsored Negative Trials and the Potential Pitfalls of Post Hoc Analysis. *Arch.Surg, 2008,143,*933-934.

Tversky,Y; Kahneman,D. Belief in the Law of Small Numbers. *Psychological Bulletin, 1971, 76,* 105-110.

Utts,J. Replication and Meta-Analysis in Parapsychology. with comments by Bayarri,M.J.; Berger,J.; Dawson,R.; Diaconis,P.; Greenhouse,J.B.; Hayman,R.; Morris,R.L. and Mosteller,F. *Statistical Science, 1991, 6, 4,* 363-403.

Wallace,C.S. *Statistical and Inductive Inference by Minimum Message Length.* Springer: NY, *2005.*

Wang,R.; Lagakos,S.W.; Ware,J.H.; Hunter,D.J.; Drazen,J.M. Statistics in Medicine - Reporting of Subgroup Analyses in Clinical Trials *The New England Journal of Medicine, 2007, 357,* 2189-2194.

Wassermann,G.D. Some Comments on the Methods and Statements in Parapsychology and Other Sciences. *The British Journal for the Philosophy of Science, 1955, 6, 22,* 122-140.

Zabell,S.L. The Quest for Randomness and its Statistical Applications. In Gordon,E.; Gordon,S. eds. *Statistics for the Twenty-First Century.* Mathematical Association of America: Washington, DC, *1992,* 139-150.

About the Author(s)

Julio Michael Stern
Full Professor at the Dept.of Applied Mathematics of IME-USP - The Institute of Mathematics and Statistics of the University of São Paulo. President of ISBrA, the Brazilian chapter of ISBA - The International Society of Bayesian Analysis. M.Sc. in Physics from IF-USP, Ph.D. in Operations Reasearch from Cornell University, Liv.Doc. in Computer Science from IME-USP.
ULR: $www.ime.usp.br/\sim jstern$