

Short note

Entropy and Effective Support Size

Marian Grendar

Department of Mathematics, FPV UMB, Tajovskeho 40, 974 01 Banska Bystrica, Slovakia
Institute of Measurement Science, Bratislava, Slovakia
Institute of Mathematics and Computer Science, Banska Bystrica, Slovakia
E-mail: marian.grendar@savba.sk

Received: 5 May 2006 / Accepted: 10 August 2006 / Published: 21 August 2006

Abstract: Notion of Effective size of support (Ess) of a random variable is introduced. A small set of natural requirements that a measure of Ess should satisfy is presented. The measure with prescribed properties is in a direct (exp-) relationship to the family of Rényi's α -entropies which includes also Shannon's entropy H . Considerations of choice of the value of α imply that $\exp(H)$ appears to be the most appropriate measure of Ess.

Entropy and Ess can be viewed thanks to their log/exp relationship as two aspects of the same thing. In Probability and Statistics the Ess aspect could appear more basic than the entropic one.

Keywords: Rényi's entropy; Shannon's entropy; support, interpretation, Probability, Statistics.

MSC 2000 codes: 94A17

1 Introduction

Interpretation of Shannon's entropy $H(p)$ is usually developed in context of an experiment where the entropy is described as a measure of uncertainty; cf. [6], [5], [7]. Motivated by a simple (and well-known) observation that $\exp(H(p))$ is equal to the size of support of the underlying random variable for the uniform distribution, in this short note we introduce concept of Effective size of support (Ess). Measure of Ess should satisfy a small set of natural requirements. The class of Ess measures $\mathbb{S}(\cdot, \alpha) = (\sum_{i=1}^m p_i^\alpha)^{\frac{1}{1-\alpha}}$ which satisfy the requirements is in a direct relationship to the family of Rényi's α -entropies which includes as its special case also Shannon's entropy. We address the issue of selecting the value of α such that the corresponding $\mathbb{S}(\cdot, \alpha)$ would be the most appropriate measure of Ess. Unlike to entropy, Ess has an obvious meaning. From the point of view of Probability or Statistics, Ess can be seen as a more natural concept than entropy.

2 Effective size of support

Let X be a discrete random variable which can take on values from a finite set $\mathcal{X}$ of m elements, with probabilities specified by the probability mass function (pmf) p . The support of X is a set $\mathcal{S}(p(X)) \triangleq \{p : p_i > 0, i = 1, 2, \dots, m\}$. Let $|\mathcal{S}(p(X))|$ denote the size of the support.

While pmf $p = [0.5, 0.5]$ makes both outcomes equally likely, the following pmf $q = [0.999, 0.001]$ characterizes a random variable that can take on almost exclusively only one of two values. However, both p and q have the same size of support. This motivates a need for a quantity that could measure size of support of the random variable in a different way, so that the random variable can be placed in the range $[1, m]$ according to its pmf. We will call the new quantity/measure the effective support size (Ess), and denote it by $\mathbb{S}(p(X))$; $\mathbb{S}(p)$ or $\mathbb{S}(X)$, for short. The example makes it obvious that $\mathbb{S}(\cdot)$ should be such that $\mathbb{S}(q)$ will be close to 1, while to p it should assign value $\mathbb{S}(p) = 2$.

3 Properties of Ess

Ess should have certain properties, dictated by common sense.

P1) $\mathbb{S}(p)$ should be continuous, symmetric function (i.e., invariant under exchange of p_i, p_j , $i, j = 1, \dots, m$).

P2) $\mathbb{S}(\delta_m) = 1 \leq \mathbb{S}(p_m) \leq \mathbb{S}(u_m) = m$; where u_m denotes the uniform pmf on m -element support, δ_m denotes an m -element pmf with probability concentrated at one point, p_m denotes a pmf¹ with $|\mathcal{S}(p)| = m$.

P3) $\mathbb{S}([p_m, 0]) = \mathbb{S}(p_m)$.

P4) $\mathbb{S}(p(X, Y)) = \mathbb{S}(p(X))\mathbb{S}(p(Y))$, if X and Y are independent random variables.

The first two properties are obvious. The third one states that extending support by an impossible outcome should leave Ess unchanged. Only the fourth property needs, perhaps, some little discussion. Or, better, an example. Let $p(X) = [1, 1, 1]/3$ and $p(Y) = [1, 1]/2$ and let X be

¹A note on notation: p_m denotes a pmf with m -element support; p_i is i -th component of the pmf.

independent of Y . Then $p(X, Y) = [1, 1, 1, 1, 1, 1]/6$. According to P2), $\mathbb{S}(p(X)) = 3$, $\mathbb{S}(p(Y)) = 2$ and $\mathbb{S}(p(X, Y)) = 6 = \mathbb{S}(p(X))\mathbb{S}(p(Y))$. It is reasonable to require the product relationship to hold for independent random variables with arbitrary distributions.

The properties P1-P4 are satisfied by $\mathbb{S}(p, \alpha) = (\sum_{i=1}^m p_i^\alpha)^{\frac{1}{1-\alpha}}$, where α is a positive real number, different than 1. Note that $\mathbb{S}(\cdot)$ of this form is exp of Rényi's entropy. For $\alpha \rightarrow 1$, $\mathbb{S}(p, \alpha)$ also satisfies P1-P4 and takes the form of $\exp(H(p))$, where $H(p) \triangleq -\sum_{i=1}^m p_i \log p_i$ is Shannon's entropy²; cf. [1]. It is thus reasonable to define $\mathbb{S}(p, \alpha)$ for $\alpha = 1$ this way (with the convention $0 \log 0 = 0$), so that $\mathbb{S}(\cdot)$ then becomes a continuous function of α .

4 Selecting α

The requirements P1-P4 define entire class of measures of effective support size. This opens a problem of selecting α .

It is instructive to begin addressing the problem with a consideration of behavior of $\mathbb{S}(p(X), \alpha)$ at the limit values of α . It can be easily seen that as $\alpha \rightarrow 0$, $\mathbb{S}(p(X), \alpha) \rightarrow |\mathcal{S}(p(X))|$, i.e., the size of the support. Thus, the closer the α to zero, the more $\mathbb{S}(\cdot, \alpha)$ behaves like the standard support size $|\mathcal{S}(p(X))|$.

For $\alpha \rightarrow \infty$, $\mathbb{S}(p(X), \alpha) = \frac{1}{\hat{p}(X)}$, where $\hat{p}(X) = \sup_{i=1,2,\dots,m} p_i$. Thus, the higher the α , the more $\mathbb{S}(\cdot, \alpha)$ judges a pmf solely by its component with the highest value of probability. At the limit, all pmf's with the same $\hat{p}(X)$ are seen as entirely equivalent.

For the sake of illustration, in Table 1, $\mathbb{S}(p, \alpha)$ is given for various two-element pmf's, and $\alpha = 0.001, 0.1, 0.5, 0.9, 1.0, 1.5, 2.0, 10, \infty$.

Table 1: $\mathbb{S}(p, \alpha)$ for $\alpha = 0.001, 0.1, 0.5, 0.9, 1.0, 1.5, 2.0, 10, \infty$ and different p 's.

α	[0.5, 0.5]	[0.6, 0.4]	$\mathbb{S}(p, \alpha)$ [0.7, 0.3]	[0.8, 0.2]	[0.9, 0.1]	[1.0, 0.0]
0.001	2.000000	1.999959	1.999826	1.999554	1.998979	1.000000
0.1	2.000000	1.995925	1.982696	1.956233	1.902332	1.000000
0.5	2.000000	1.979796	1.916515	1.800000	1.600000	1.000000
0.9	2.000000	1.964013	1.856116	1.675654	1.416403	1.000000
1.0	2.000000	1.960132	1.842023	1.649385	1.384145	1.000000
1.5	2.000000	1.941178	1.777878	1.543210	1.275510	1.000000
2.0	2.000000	1.923077	1.724138	1.470588	1.219512	1.000000
10.0	2.000000	1.760634	1.486289	1.281379	1.124195	1.000000
∞	2.000000	1.666666	1.428571	1.250000	1.111111	1.000000

Based on the table, in this simplest case of two-valued random variable we would opt for $\mathbb{S}(\cdot, \infty)$ as the good measure of Ess. However, for larger $|\mathcal{S}|$ this choice becomes less attractive. As it was already noted, $\mathbb{S}(\cdot, \infty) = 1/\hat{p}$ and all pmf's with the same $\hat{p}$ are seen to have the same Ess. For

²In this paper, $\log$ denotes the natural logarithm.

instance, $p = [0.95, 0.05]$ and $q = [0.95, x]$ where x stands for the other remaining 99 components with the value $0.05/99 = 0.0005$, are by $\mathbb{S}(\cdot, \infty)$ judged to have the same Ess, equal to 1.053. Just for a comparison, $\mathbb{S}(p, 1) = 1.220$, while $\mathbb{S}(q, 1) = 1.535$. This undesirable feature of $\mathbb{S}(\cdot, \infty)$ manifests itself even more sharply in the case of continuous random variables.

5 Ess in the continuous case

The continuous-case analogue³ of $\mathbb{S}(p, \alpha) = (\sum_{i=1}^m p_i^\alpha)^{\frac{1}{1-\alpha}}$ is $\mathbb{S}_c(f(x), \alpha) \triangleq (\int f^\alpha(x) dx)^{\frac{1}{1-\alpha}}$, where $f(x)$ denotes a density with respect to Lebesgue measure. The continuous-case $\mathbb{S}_c$, though always positive, can – naturally – be smaller than one. And the discrete-case upper bound m is now replaced by ∞ . It is worth stressing that $\mathbb{S}_c$ behaves with respect to shift and scale transformations in the desired manner. Indeed, if $Y = X + a$, then $\mathbb{S}_c(Y, \alpha) = \mathbb{S}_c(X, \alpha)$; if $Y = aX$, then $\mathbb{S}_c(Y, \alpha) = a \mathbb{S}_c(X, \alpha)$.

For the Gaussian $n(\mu, \sigma^2)$ distribution, $\mathbb{S}(\cdot, \alpha) = \frac{\sqrt{2\pi\sigma^2}}{\alpha^{\frac{1}{2(1-\alpha)}}}$; cf. [8]. This for $\alpha \rightarrow \infty$ converges to $\sqrt{2\pi\sigma^2}$, so that for $\sigma^2 = 1$ it becomes $\sqrt{2\pi} = 2.5067$. It is worth comparing with $\mathbb{S}(\cdot, 1) = \sqrt{2e\pi\sigma^2}$ (cf. [9]), which reduces in the case of $\sigma^2 = 1$ to 4.1327. This makes much more sense.

That $\mathbb{S}(\cdot, \infty)$ is not the appropriate measure of Ess can be even more clearly seen in the case of the Exponential distribution. For $\beta e^{-\beta x}$ with $\beta = 1$, $\mathbb{S}(\cdot, \infty) = 1$ while $\mathbb{S}(\cdot, 1) = e$.

6 Adding another property

The above considerations suggest that $\mathbb{S}(\cdot, 1)$ might be the most appropriate of the Ess measures which satisfy the requirements P1-P4. The question is whether there is some other requirement that is reasonable to add to the already employed properties, such that it could narrow down the set of $\mathbb{S}(\cdot, \alpha)$ to $\mathbb{S}(\cdot, 1)$.

To this end, let us consider two random variables X, Y that, in general, might be dependent. It is natural, to extend requirement P4 to the more general setting, by requiring that⁴

$$\text{P4}^*) \quad \mathbb{S}(p(X)) \mathbb{S}(p(Y)) \geq \mathbb{S}(p(X, Y)),$$

with the equality if and only if X and Y are independent.

For $\alpha \neq 1$, it might be in some cases that instead of $\geq$ the opposite relation $<$ holds true. Indeed, consider for instance the following bivariate discrete random variable with pmf $p(X, Y)$

0.2	0.05	0.05	0.3
0.3	0.2	0.2	0.7
0.5	0.25	0.25	$X \setminus Y$

³The relationship between discrete and continuous $\mathbb{S}(\cdot)$ is analogous to that of discrete and differential entropies; cf. [6], [2], [7].

⁴In an earlier version [3] of the paper we considered a different property which involved a notion of Ess for a mean of conditional distributions.

Marginal pmf $p(X)$ has $\mathbb{S}(X, \infty) = 2$, and $\mathbb{S}(Y, \infty) = \frac{10}{7}$. Hence, $\mathbb{S}(X, \infty)\mathbb{S}(Y, \infty) = 2.86$, which is smaller than $\mathbb{S}(p(X, Y), \infty) = \frac{10}{3}$. After a minor change in the joint pmf, such that the marginals remain unchanged, it is possible to satisfy P4*. It is known (cf. [1]) that solely $\mathbb{S}(\cdot, 1)$ always satisfies the natural requirement P4*.

7 Summary

Shannon's entropy is a key concept of Communication Theory. In Probability and Statistics the entropy is usually interpreted as a measure of uncertainty about realization of a random variable, or as a measure of complexity or uniformness of a probability distribution. Though the entropy is within Probability and Statistics from time to time (and from area to area) blamed for failing to be measure of all the fancy and intangible things, it remains to be a valuable tool.

In this note we introduced⁵ concept of the Effective support size (Ess) of a random variable. There are a few requirements that the measure $\mathbb{S}(p(X))$ of Ess of a probability distribution $p(X)$ should satisfy. The requirements turn to be direct analogues of those placed on entropy; cf. [5], [1]. It thus should not be surprising that they are satisfied⁶ by $\mathbb{S}(p, \alpha) = (\sum_{i=1}^m p_i^\alpha)^{\frac{1}{1-\alpha}}$ which is the exponential of Rényi's entropy.

Since $\mathbb{S}(\cdot, \alpha)$ is in fact a continuum of measures of Ess, it is necessary to find out which of them would be the most appropriate measure(s) of Ess. It seems that $\mathbb{S}(\cdot, 1) = \exp(H(\cdot))$, where $H(\cdot)$ is Shannon's entropy, is the best choice; cf. Sect. 4 and 5. We also argued for expanding the key requirement P4 into a more general requirement P4*. The enhanced set of requirements is satisfied solely by $\mathbb{S}(\cdot, 1)$.

We maintain that from the point of view of Probability and Statistics, Ess is more basic concept than entropy. The two concepts are related together by the exp / log link. Without the link thus for instance knowing that Shannon's entropy of the Gaussian variable is $H(\cdot) = \log \sqrt{2e\pi\sigma^2}$ does not say much. Figuratively speaking, thanks to Ess entropy itself becomes more informative.

Ess adds also a new meaning to the Maximum Entropy method [4]. For instance the classic finding [6] that the Gaussian distribution has the maximal value of Shannon's entropy among all distributions with prescribed second moment can be rephrased as stating that among all such distributions the one with the biggest effective support is the Gaussian distribution.

Acknowledgement Supported by VEGA 1/3016/06 grant. I am grateful to Michael George for posing a problem which has induced this work. Critical comments of three anonymous referees helped to improve both content and presentation – and are gratefully acknowledged.

To George. May 2, 2006.

⁵It is unlikely that something like Ess has not been already spotted. Yet, we are aware only that Cover and Thomas [2] interpret exp of Shannon's entropy of a random sequence as an effective volume of random variable, in the context of their discussion of the Asymptotic Equipartition Property.

⁶In the discrete case. For a discussion of the case of a continuous random variable see Section 5.

References

- [1] Aczél, J. Entropies Old and New (and Both New and Old) and Their Characterizations; in Bayesian Inference and Maximum Entropy Methods in Science and Engineering (G. Erickson and Y. Zhai, eds.) 2004, 119-127.
- [2] Cover, T. M. and Thomas, J. A. Elements of Information Theory; Wiley: New York, 1991, pp. 224-238.
- [3] Grendar, M. Effective Support Size. Online preprint, arXiv:math.ST/0605146. 2006.
- [4] Jaynes, E. T. Information Theory and Statistical Mechanics. Phys. Rev. 1957, 106 & 108, 620 & 171.
- [5] Khinchin, A. I. Mathematical Foundations of Information Theory; Dover: New York, 1957, pp. 9-13.
- [6] Shannon, C. E. The Mathematical Theory of Communication. Bell Syst. Techn. Journ. 1948, 27, 379-423, 623-656.
- [7] Smith, J. D. H. Some Observations on the Concepts of Information-Theoretic Entropy and Randomness. Entropy 2001, 3, 1-11.
- [8] Song, K.-S. Rényi Information, Loglikelihood and an Intrinsic Distribution Measure. Jour. Stat. Inference and Planning 2001, 93, 51-69.
- [9] Verdugo Lazo, A. C. G.; Rathie, P. N. On the Entropy of Continuous Probability Distributions. IEEE Trans. IT 1978, IT:24, 120-122.