Z. Wahrscheinlichkeitstheorie verw. Geb. 5, 217—225 (1966)

Exponential Entropy as a Measure of Extent of a Distribution

L. L. CAMPBELL*

Received September 1, 1964 / December 17, 1964

1. Introduction

It has long been recognized that entropy is, in some sense, a measure of spread or extent of the associated distribution. This idea is investigated from several points of view in the present paper. The basic result is that e^H is a measure of extent, where H is SHANNON's [1] entropy in natural units. In addition, it will be shown that RÉNYI's [2, 3] entropy of order α is also connected in a natural way with measures of spread.

The measures of extent which are considered here are all, in one way or another, generalizations of the notion of range of a distribution. Indeed, for a uniform distribution, the exponential entropy e^H is just the range. By the "range" of a distribution we mean the number of points in the sample space if the sample space is discrete, and the length of the interval on which the probability density function is different from zero if the sample space is the real line. This and other terminology will be defined more precisely in Section 2.

In Section 3 we examine the idea of range and its generalizations from a fairly direct and simple point of view. In Section 4 we look at the probabilities of sequences of independent events and see how they can provide a measure of extent. This approach is closely related to the development of the coding theorems of information theory. In Section 5 we consider optimum ways of digitally recording the outcome of a random experiment. This leads to another interpretation of e^H as a measure of spread. In Section 6 some known properties of entropy are presented in a form which supports the interpretation of exponential entropy given here. Finally, in Section 7, the connection between concentration of the probability distribution and size of the measure of extent is investigated.

2. Definitions and Preliminary Remarks

In order to avoid difficulties connected with the question of existence of product measures and Radon-Nikodym derivatives it will be assumed throughout the paper that all measures which occur are σ-finite. Two measures which are absolutely continuous with respect to each other will be called *equivalent*.

Let $(\Omega, \mathscr{A}, P)$ be a probability space and let ν be another measure on $\mathscr{A}$ which is equivalent to P. Then the Radon-Nikodym derivatives $dP/d\nu$ and $d\nu/dP = (dP/d\nu)^{-1}$ exist. The measure ν in our considerations will usually be a "natural" measure on Ω. For example, if Ω is a discrete space with points $x_1, x_2, \ldots$, ν might be counting measure which assigns to each set a measure equal to the number of

* This research was supported in part by the Defence Research Telecommunications Establishment (DRB) under Contract No. CD.DRB/313002 with Queen's University.

points in this set. In this case $dP/d\nu$ at the point x_i is just $P(x_i)$. As another example suppose that Ω is the Euclidean space E_n and that the probability measure P is determined by a probability density function. Then a natural choice for ν is Lebesgue measure on E_n. With this choice of ν, $dP/d\nu$ is just the probability density function.

We will say that the probability distribution is *uniform* (with respect to ν) if $dP/d\nu$ is constant on Ω except for a set of probability zero.

By the *range* of the distribution (with respect to ν) we mean $\nu(\Omega)$. This use of the word "range" differs slightly from some other usages which are connected with a random variable on Ω. Frequently "range" means either the range space of the random variable or the size of this range space.

In analogy with the notation of HARDY, LITTLEWOOD, and PÓLYA [4] we define the mean of order t of the probability distribution by

$$M_t[\nu, P] = \left(\int \left[\frac{d\nu}{dP}\right]^t dP\right)^{1/t} \quad (t \neq 0) \tag{1}$$

and

$$M_0[\nu, P] = \exp\left(\int \ln\left[\frac{d\nu}{dP}\right] dP\right). \tag{2}$$

Here and subsequently, all integrals are over Ω. It is easily shown that, when the integrals exist,

$$\lim_{t\to 0} M_t[\nu, P] = M_0[\nu, P].$$

In addition, we define the entropy of the distribution (with respect to ν) by

$$H[\nu, P] = \ln M_0[\nu, P] \tag{3}$$

and the entropy of order α by

$$I_\alpha[\nu, P] = \ln M_{1-\alpha}[\nu, P]. \tag{4}$$

It follows that

$$\lim_{\alpha\to 1} I_\alpha[\nu, P] = H[\nu, P].$$

Equations (3) and (4) suggest the name "exponential entropy" for M_t which is used in the title of this paper.

As a special case, let ν be counting measure and let Ω be a finite space with points $x_1, \ldots, x_N$ and with $P(x_i) = p_i$. Then

$$H[\nu, P] = -\sum_{i=1}^{N} p_i \ln p_i$$

and

$$I_\alpha[\nu, P] = \frac{1}{1-\alpha} \ln\left(\sum_{i=1}^{N} {p_i}^\alpha\right).$$

Except for the use of natural logarithms instead of logarithms to the base 2, H and I_α are respectively the entropy of SHANNON [1] and the entropy of order α of RÉNYI [2, 3].

If ν is Lebesgue measure, H and I_α are SHANNON's and RÉNYI's entropies for a continuous distribution. If ν is another probability measure, $-H[\nu, P]$ is the "information gain" of RÉNYI [3] and the "directed divergence" of KULLBACK [5].

It should also be remarked that our notation differs from that of ACZÉL and DARÓCZY [6, 7] who define a quantity M_α in such a way that $I_\alpha = -\log M_\alpha$. RÉNYI [1, 2], ACZÉL, and DARÓCZY [6, 7, 8] have studied axiomatic characterizations of I_α and M_α.

Some elementary properties of the mean M_t will now be noted. First of all,

$$M_1[\nu, P] = \int \frac{d\nu}{dP}\, dP = \nu(\Omega)\,, \tag{5}$$

so that the mean of order one is the range. Second, if the distribution is uniform, the derivative $d\nu/dP$ is a constant which is easily seen to be $\nu(\Omega)$. In this case it follows that

$$M_t[\nu, P] = \nu(\Omega)\,. \tag{6}$$

Finally, it can be shown [4] that when $s < t$,

$$M_s[\nu, P] \leqq M_t[\nu, P]\,, \tag{7}$$

with equality if and only if the probability distribution is uniform.

3. Generalized Range

The range, $\nu(\Omega)$, is perhaps the most elementary measure of extent of a distribution. The disadvantage of the range is that it assigns the same weight to sets of low probability as to sets of high probability. In consequence, the range is often infinite for simple probability distributions. As noted above, we can write

$$\nu(\Omega) = \int \frac{d\nu}{dP}\, dP\,.$$

A more useful measure of extent might be obtained by modifying the integrand in such a way that the effect of small values of $dP/d\nu$ is decreased. However, in making such a modification we should preserve homogeneity in ν. That is, if the measure ν is replaced by $c\nu$, where c is a positive number, the measure of extent should also be multiplied by c. Clearly $M_t[\nu, P]$ satisfies these requirements for $t < 1$. Also, in view of (5) and (7), it follows that

$$M_t[\nu, P] \leqq \nu(\Omega) \qquad (t \leqq 1)\,. \tag{8}$$

Thus $M_t[\nu, P] (t < 1)$ has some desirable properties for a measure of extent.

The mean $M_0[\nu, P]$ is the mean which is least affected by variations in the value of $dP/d\nu$ and on this ground merits some special consideration as a measure of extent.

4. Sequences of Independent Events

Frequently we are concerned with the number of possible outcomes if an experiment is performed many times, rather than with the number of possible outcomes if the experiment is performed once. Suppose the same experiment is performed m times ($m \geqq 1$), that the performances are independent, and that the

probabilities of outcomes each time are given by the measure P. We are then concerned with events in the m-fold direct product of Ω with itself. Let P_m and ν_m denote the corresponding product measures. If the outcomes of the individual experiments are $y_1, y_2, \ldots, y_m$ the Radon-Nikodym derivative is

$$\frac{d\nu_m}{dP_m}(y_1, \ldots, y_m) = \frac{d\nu}{dP}(y_1)\frac{d\nu}{dP}(y_2)\cdots\frac{d\nu}{dP}(y_m).$$

When the distribution is uniform, we have

$$\frac{d\nu_m}{dP_m} = \left(\frac{d\nu}{dP}\right)^m = [\nu(\Omega)]^m.$$

This suggests that $(d\nu_m/dP_m)^{1/m}$ might approximate some useful measure of range even for non-uniform distributions.

The above considerations lead us to consider

$$R_m = E\left[\left(\frac{d\nu_m}{dP_m}\right)^{1/m}\right] \tag{9}$$

as a measure of the extent of a distribution. Here, E denotes the mathematical expectation. Because the outcomes are independent and identically distributed, we have

$$R_m = \left(E\left[\left(\frac{d\nu}{dP}\right)^{1/m}\right]\right)^m = M_t[\nu, P] \qquad \left(t = \frac{1}{m}\right). \tag{10}$$

Letting $m \to \infty$ we obtain the measure of extent

$$R_\infty = \lim_{m\to\infty} R_m = M_0[\nu, P]. \tag{11}$$

From (5) and (7) it will be seen that $\{R_m\}$ is a sequence which decreases monotonically from the range, R_1, to $M_0[\nu, P]$.

It is interesting to note that the approach to the coding theorems of information theory used by KHINCHIN [9], for example, utilizes the result that

$$E\left[\ln\left(\frac{d\nu_m}{dP_m}\right)^{1/m}\right] = H[\nu, P].$$

Equation (10) implies, on the other hand, that

$$\ln E\left[\left(\frac{d\nu_m}{dP_m}\right)^{1/m}\right] = I_\alpha[\nu, P] \qquad (\alpha = 1 - m^{-1}).$$

This effect of interchanging the logarithm and expectation has been pointed out by RÉNYI [10] in a somewhat different context.

5. Digital Representation of Events

In many situations digital data-processing systems are used to record and compute with the outcomes of random experiments. Frequently the experimenter will have some freedom in the choice of a way to record events. We wish to develop a way of comparing two different ways of recording the same event.

Suppose that a set A has the size $\nu(A) \neq 0$. The smaller A is, the more digits must be recorded to specify A exactly. In many common situations, the number of digits necessary to specify A will be approximately $-\log \nu(A)$, or possibly

$-\log \nu(A) + \textit{constant}$. Now suppose that there is another method of recording A which assigns to A the size $\mu(A) \neq 0$. This could occur either by assigning the measures $\nu(A)$ and $\mu(A)$ to A directly, or by mapping the set A into two other sets of sizes $\nu(A)$ and $\mu(A)$. The difference in the number of digits required in the two recording schemes is proportional to $\log \mu(A) - \log \nu(A)$.

If Ω were a finite space with elementary events $A_1, A_2, \ldots$, the average difference in number of digits in the two recording schemes would be proportional to

$$\sum P(A_i) \log \frac{\mu(A_i)}{\nu(A_i)}.$$

If Ω is not finite, we can form finer and finer partitions of Ω and let the above expression pass to a limit. If the measures P, μ, ν possess reasonable properties, this limit would be

$$\Delta = \int \log\left(\frac{d\mu}{d\nu}\right) dP.$$

Thus Δ can be regarded as some measure of the difference in storage requirements between the two recording methods associated with the measures μ and ν.

If $\Delta = 0$, the two methods of recording events require the same average amount of digital storage space. This suggests the following.

Definition 1. Let $(\Omega, \mathscr{A}, P)$ be a probability space and let μ and ν be two equivalent measures on $\mathscr{A}$. The measures μ and ν are *digitally equivalent* if the following integral exists and

$$\int \ln\left(\frac{d\mu}{d\nu}\right) dP = 0. \tag{12}$$

If ν is some natural measure which is equivalent to P we have seen that $\nu(\Omega)$ is a possible measure of the extent of Ω. If μ is some other measure which is digitally equivalent to ν, then $\mu(\Omega)$ is another possible measure of the extent of Ω. This suggests

Definition 2. The *intrinsic extent* of the probability space $(\Omega, \mathscr{A}, P)$ with respect to the measure ν is

$$K[\nu] = \inf \mu(\Omega),$$

where the infimum is taken over the class of all measures which are digitally equivalent to ν.

The intrinsic extent can often be evaluated with the aid of

Theorem 1. *Let* $0 < M_0[\nu, P] < \infty$, *where* M_0 *is defined by* (2). *Then*

$$K[\nu] = M_0[\nu, P]. \tag{13}$$

This theorem follows easily from a theorem given by HOFFMAN [11] in his derivation of SZEGÖ's theorem. However, a direct proof is short. Let $H = H[\nu, P] = \ln M_0[\nu, P]$. If μ is digitally equivalent to ν we have

$$\begin{aligned}
\mu(\Omega) &= \int \frac{d\mu}{d\nu} \frac{d\nu}{dP} dP \\
&= e^H \int \exp\left(\ln \frac{d\mu}{d\nu} + \ln \frac{d\nu}{dP} - H\right) dP \\
&\geqq e^H \int \left(1 + \ln \frac{d\mu}{d\nu} + \ln \frac{d\nu}{dP} - H\right) dP = e^H.
\end{aligned}$$

The inequality follows from the inequality $e^u \geqq 1 + u$ and the last equality follows from (12), the definition of H, and the fact that $P(\Omega) = 1$. Thus for any μ which is digitally equivalent to ν, $\mu(\Omega) \geqq M_0[\nu, P]$. Now consider the particular measure $\mu_1 = M_0[\nu, P]P$. It is easily shown that μ_1 is digitally equivalent to ν and that $\mu_1(\Omega) = M_0[\nu, P]$. This completes the proof.

It follows from (5) and (7) that $K[\nu] \leqq \nu(\Omega)$ with equality if and only if the distribution is uniform with respect to ν.

Theorem 1 is closely connected with the coding theorem for a noiseless channel [12]. An essential part of the coding theorem states that the minimum of $-\sum p_i \ln q_i$ is $H = -\sum p_i \ln p_i$ when the minimization is performed subject to the constraint that $\sum q_i = 1$. Theorem 1 states that the minimum of $\sum q_i$ subject to the constraint $\sum p_i \ln q_i = 0$ is e^H.

6. Special Properties

In this Section we mention some special properties of $M_0[\nu, P]$ which support the interpretation as a measure of extent. First the interpretation will be illustrated by a few special distributions.

If Ω is a set of N points, each having probability N^{-1}, and ν is counting measure, then $M_0 = N$. If Ω is the interval $a \leqq x \leqq b$, if the probability measure is determined by the constant density function $(b - a)^{-1}$, and if ν is Lebesgue measure, then $M_0 = b - a$.

The normal or Gaussian probability distributions in one and two dimensions also provide interesting illustrations. If Ω is the real line, ν is Lebesgue measure, and P is the normal probability distribution with mean m and standard deviation σ, we have $M_0 = (2\pi e)^{1/2}\sigma$. Thus the extent is proportional to the usual measure of spread, σ. If Ω is the plane, ν is Lebesgue measure in the plane, and the probability distribution is the distribution of two normal random variables with standard deviations σ_1 and σ_2 and correlation coefficient ϱ, then

$$M_0 = 2\pi e \sigma_1 \sigma_2 \sqrt{1 - \varrho^2}.$$

When the random variables are independent, so that $\varrho = 0$, the extent is just the product of the extents of the two associated one dimensional distributions.

There is another property of entropy which has a natural interpretation in the present context. Let X be a random variable with density function $f(x)$ and entropy (with respect to Lebesgue measure)

$$H_x = -\int_{-\infty}^{\infty} f(x) \ln f(x)\, dx.$$

Let $Y = aX$ be a new random variable with density function $|a|^{-1} f(a^{-1}y)$. The entropy of Y is given by

$$H_y = H_x + \ln|a|.$$

In terms of the extents this equation becomes

$$M_0[\nu, P_y] = |a|\, M_0[\nu, P_x].$$

That is, the extent has been multiplied by the scale factor $|a|$, as would be expected.

Finally, there are two results due to HIRSCHMAN [13] which are significant in this context. The first relates M_0 to the moments and the second shows that there is an uncertainty relation similar to the uncertainty relation for standard deviations. Let $f(x)$ be a probability density function and let

$$m(a, q) = [\int_{-\infty}^{\infty} |x - a|^q f(x)\, dx]^{1/q}.$$

If ν is Lebesgue measure and P is the probability measure associated with $f(x)$, one of HIRSCHMAN's inequalities becomes

$$M_0[\nu, P] \leqq 2\, q^{(1-q)/q}\, e^{1/q}\, \Gamma(q^{-1})\, m(a, q).$$

For the second result, let

$$\Psi(x) \sim \int_{-\infty}^{\infty} \Phi(y)\, e^{2\pi i x y}\, dy$$

and let

$$\int_{-\infty}^{\infty} |\Psi|^2\, dx = \int_{-\infty}^{\infty} |\Phi|^2\, dy = 1.$$

Let ν be Lebesgue measure and let P_1 and P_2 be the probability measures associated with the density functions $|\Psi|^2$ and $|\Phi|^2$. In our notation, HIRSCHMAN's second result is that

$$M_0[\nu, P_1]\, M_0[\nu, P_2] \geqq 1.$$

This result is an analogue of the uncertainty principle of quantum mechanics which makes a similar assertion about the product of two standard deviations.

7. The Significance of Small Values of M_t *

If the standard deviation of a random variable is small it is a consequence of the TCHEBICHEV inequality that most of the probability distribution of the random variable is concentrated in some small interval around the mean. If $M_t[\nu, P]$ is to be interpreted as a measure of extent one would expect similarly that a small value of $M_t[\nu, P]$ should imply that most of the probability measure is concentrated on a set of small ν-measure. For $0 < t \leqq 1$ this is true. For $t = 0$ this statement is not necessarily true. However, if $M_0[\nu, P]$ is small, it is possible to demonstrate the existence of a set E such that $\nu(E)/P(E)$ is small. Thus, at least some of the probability measure is concentrated on a set of relatively small ν-measure.

First we derive a pair of inequalities which are somewhat analogous to the TCHEBICHEV inequality. Let

$$E_c = \left\{\omega : \omega \in \Omega,\ \frac{dP}{d\nu}(\omega) \geqq c\right\}$$

and let E_c^* be the complement of E_c. Then, for $0 < t \leqq 1$,

$$c^{-t}\, P(E_c^*) = \int_{E_c^*} c^{-t}\, dP \leqq \int_{E_c^*} \left(\frac{d\nu}{dP}\right)^t dP \leqq (M_t[\nu, P])^t.$$

* The author is indebted to Professor A. RÉNYI for pointing out the importance of the question which is treated in this section and for suggesting some of the results which are presented here.

Thus,

$$P(E_c) \geqq 1 - (c\,M_t)^t. \tag{14}$$

Also,

$$c^{1-t}\,\nu(E_c) \leqq \int_{E_c} \left(\frac{dP}{d\nu}\right)^{1-t} d\nu = \int_{E_c} \left(\frac{d\nu}{dP}\right)^{t} dP \leqq (M_t)^t.$$

Thus

$$\nu(E_c) \leqq \frac{(c\,M_t)^t}{c}. \tag{15}$$

From the definition of M_t it follows that $d\nu/dP \leqq M_t$ on some non-void set. Hence, as long as $c < (M_t)^{-1}$, the set E_c is not void. Thus, when $0 < t \leqq 1$, if M_t is small there exists a set of small ν-measure whose probability is close to one.

If $t = 0$, the inequalities (14) and (15) are still true, but they are trivial. Let

$$E = \left\{\omega : \omega \in \Omega, \frac{d\nu}{dP}(\omega) \leqq M_0\right\}.$$

Since

$$\ln M_0 = \int_{\Omega} \ln \frac{d\nu}{dP}\, dP,$$

E is not void and $P(E) > 0$. Then

$$\nu(E) = \int_{E} \frac{d\nu}{dP}\, dP \leqq M_0\, P(E).$$

Hence

$$\frac{\nu(E)}{P(E)} \leqq M_0. \tag{16}$$

Therefore, if M_0 is small, there exists a set whose ν-measure is small relative to its probability measure.

The fact that the whole probability measure is not necessarily concentrated on a set of small ν-measure when M_0 is small is easily shown by an example. Let ν be Lebesgue measure and let P_n be the probability measure which has the density function

$$f_n(x) = \begin{cases} n & 0 < x < \dfrac{1}{2n} \\ 1 & \dfrac{1}{2} < x < 1 \\ 0 & \text{otherwise}. \end{cases}$$

Then

$$M_0[\nu, P_n] = n^{-1/2}.$$

For large n, M_0 is small, while the distribution is not concentrated on a small set. However, there is a set on which the distribution is relatively highly concentrated. For, if E is the interval $(0, 1/2n)$,

$$\frac{P(E)}{\nu(E)} = n.$$

It was remarked earlier that the range M_1 is not always a good measure of extent because it is not sensitive to variations in the probability density. It appears that M_0 is too sensitive to the presence of sharp peaks in the density. Possibly M_t for some intermediate value of t will prove to be more useful than either M_0 or M_1.

References

[1] SHANNON, C. E.: A mathematical theory of communication. Bell System techn. J. **27**, 379—423 and 623—656 (1948).

[2] RÉNYI, A.: On measures of entropy and information. Proc. Fourth Berkeley Sympos. math. Statist. Probab. **1**, 547—561 (1961).

[3] — Wahrscheinlichkeitsrechnung, mit einem Anhang über Informationstheorie. Berlin: Deutscher Verlag der Wissenschaften 1962.

[4] HARDY, G. H., J. E. LITTLEWOOD, and G. PÓLYA: Inequalities. Cambridge: Cambridge Univ. Press 1934.

[5] KULLBACK, S.: Information theory and statistics. New York: Wiley 1959.

[6] ACZÉL, J., and Z. DARÓCZY: Charakterisierung der Entropien positiver Ordnung und der Shannonschen Entropie. Acta math. Acad. Sci. Hungar. **14**, 95—121 (1963).

[7] — — Sur la caractérisation axiomatique des entropies d'ordre positif, y comprise l'entropie de Shannon. C. r. Acad. Sci., (Paris) **257**, 1581—1584 (1963).

[8] DARÓCZY, Z.: Über die gemeinsame Charakterisierung der zu den nicht vollständigen Verteilungen gehörigen Entropien von Shannon und von Rényi. Z. Wahrscheinlichkeitstheorie verw. Geb. **1**, 381—388 (1963).

[9] KHINCHIN, A. I.: The entropy concept in probability theory. Uspehi Mat. Nauk 8, 3—20 (1953). Translations in A. I. KHINCHIN: Mathematical foundations of information theory. New York: Dover 1957, and in Arbeiten zur Informationstheorie I, 2nd edition. Berlin: Deutscher Verlag der Wissenschaften 1961.

[10] RÉNYI, A.: On the foundations of information theory. Address presented at the 34th Session of the International Statistics Institute, Ottawa, Canada, 1963.

[11] HOFFMAN, K.: Banach spaces of analytic functions. P. 48. Englewood Cliffs: Prentice-Hall 1962.

[12] FEINSTEIN, A.: Foundations of information theory. Pp. 17—20. New York: McGraw-Hill 1958.

[13] HIRSCHMAN, I. I.: A note on entropy. Amer. J. Math. **79**, 152—156 (1957).

Department of Mathematics
Queen's University
Kingston, Ontario, Canada