A entropia de Shannon: uma abordagem axiomática

Shannon's entropy: an axiomatic approach

José Carlos Magossi
Universidade Estadual de Campinas (UNICAMP), Faculdade de Tecnologia (FT), Divisão de Telecomunicações, Limeira, SP, Brasil
https://orcid.org/0000-0001-8985-6507, magossi@ft.unicamp.br

Antônio César da Costa Barros
Universidade Estadual de Campinas (UNICAMP), Faculdade de Tecnologia (FT), Doutorando no Programa de Pós-Graduação em Tecnologia (PPGT), Limeira, SP, Brasil
https://orcid.org/0000-0002-4822-9459, cesar.matema@gmail.com

Informações do Artigo

Como citar este artigo

MAGOSSI, José Carlos; BARROS, Antônio César da Costa. A entropia de Shannon: uma abordagem axiomática. **REMAT: Revista Eletrônica da Matemática**, Bento Gonçalves, RS, v. 7, n. 1, p. e3013, 26 de maio 2021. DOI: https://doi.org/10.35819/remat2021v7i1id4756

Histórico do Artigo
Submissão: 16 de novembro de 2020.
Aceite: 19 de março de 2021.

Palavras-chave
Axiomas
Entropia
Comunicação
Teoria de Shannon
Equação Funcional

Resumo

A palavra "entropia" surgiu no ano de 1864, nos trabalhos de Termodinâmica de Rudolf Clausius. Em 1948, Claude E. Shannon utiliza esse mesmo nome para designar uma medida de informação em seu modelo matemático de comunicação, fundamentado nos conceitos de emissor, receptor, canal, ruído, redundância, codificação e decodificação. Com a medida de informação $H(X) = -C\sum_{i=1}^{n} p_i \log p_i$, a entropia de Shannon, tornase possível analisar a capacidade do canal de comunicação e investir em tratamentos de dados, dando origem ao que se chama atualmente de Teoria da Informação. Além dos aspectos operacionais e tecnológicos da teoria de Shannon, que revelam a era digital, as abordagens matemáticas acerca da fórmula $H(X)$ acabam por revelar também uma vertente voltada às caracterizações de medidas de informação. Entende-se que uma exposição didática da dedução matemática da fórmula de entropia de Shannon, com base em um conjunto de axiomas, seja interessante não somente no sentido pedagógico, mas também para o entendimento da teoria de Shannon. Mostra-se, desse modo, que essa fórmula está imersa em um contexto matemático bem definido (um sistema com axiomas e equações funcionais), que permite, com alterações nos axiomas, definir novas medidas de informação.

Abstract
The word “entropy” arose in the year 1864, in the works of thermodynamics by Rudolf Clausius. In 1948, Claude E. Shannon uses that same name to designate a measure of information in his mathematical model of communication, based on the concepts of emitter, receiver, channel, noise, redundancy, coding and decoding. With the measure of information $H(X) = -C\sum_{i=1}^{n} p_i \log p_i$, the Shannon entropy, it becomes possible to analyze the capacity of the communication channel and invest in data processing, giving rise to what is currently called Information Theory. In addition to the operational and technological aspects of theory of Shannon, that reveal the digital age, from mathematical approaches about the formula $H(X)$, also end up revealing a tendency focused on the characterization of information measures. It is understood that an exposure didactic of mathematical deduction of the formula from Shannon entropy, based on a group of axioms, not only is interesting in the pedagogical sense, but also for understanding theory of Shannon. It thereby show, that this formula is immersed in a well-defined mathematical context (a system with axioms and functional equations), allowing, with changes in the axioms, defining new measures of information.

Keywords
Axioms
Entropy
Communication
Theory of Shannon
Functional Equations

1 Introdução

Rudolf Clausius, no ano de 1864, cunhou o termo **entropia** para representar transformação de energia em seus estudos, no que se denomina hoje em dia de Termodinâmica (CLAUSIUS, 1864). O uso desse termo viria a ocorrer também na mecânica estatística. Mesmo que o termo entropia tenha sido utilizado em outros contextos (MAGOSSI; PAVIOTTI, 2019), Claude E. Shannon, em 1948 (SHANNON, 1948), decide utilizá-lo para representar uma **medida de informação** em seu modelo matemático de comunicação. O objetivo de Shannon era o de desenvolver melhorias na comunicação elétrica (PIERCE, 1961). Com esse viés, o artigo de Shannon impacta fortemente nos processos de comunicação entre máquinas, tanto é que possibilitou o surgimento da Teoria da Informação (SHANNON; WEAVER, 1949), com inúmeras aplicações em diversas áreas do conhecimento, conforme consta, por exemplo, em Cover e Thomas (1991). Por um lado, as aplicações advindas da teoria de Shannon configuram o que se chama de era digital e podem ser observadas em CD’s, internet, TV digital, arquivos ZIP, compactador de dados etc. Por outro, entende-se que a vertente matemática relativa à dedução axiomática da fórmula $H(X)$ de entropia acaba por revelar estruturas matemáticas associadas ao conceito de medida de informação. Em seu artigo de 1948 (SHANNON, 1948, p. 393), Shannon expõe três propriedades que julga suficientes para deduzir sua

fórmula de entropia de forma única, levando-se em conta que $p_i > 0$ (para um conjunto discreto de probabilidades $p_1, p_2, \ldots, p_i, \ldots, p_n$) e $\sum_{i=1}^{n} p_i = 1$. São elas:

> 1. H deve ser contínua em p_i.
> 2. Se todos os p_i forem iguais, $p_i = \frac{1}{n}$, então H deve ser uma função de n monotônica crescente. Com eventos igualmente prováveis, há mais escolha, ou incerteza, quando há mais eventos possíveis.
> 3. Se uma escolha for decomposta em duas escolhas sucessivas, a H original deve ser a soma ponderada dos valores individuais de H (...). (SHANNON, 1948, p. 393, tradução dos autores).

Logo em seguida Shannon (1948, p. 393, tradução dos autores) escreve:

> A única H satisfazendo as três hipóteses acima é da forma
>
> $$H = -K \sum_{i=1}^{n} p_i \log p_i$$
>
> onde K é uma constante positiva.

Ou seja, Shannon entendeu que seus axiomas deveriam ser plausíveis no sentido operacional e suficientes para se obter $H(X)$ como um teorema com base em seus axiomas.

> Esse teorema, e as hipóteses requeridas para sua demonstração, não são, de forma alguma, necessários para a presente teoria. É dada, principalmente, para conceder uma certa plausibilidade a algumas de nossas definições posteriores. A real justificativa dessas definições, entretanto, residirá nas suas implicações (SHANNON, 1948, p. 393, tradução dos autores).

Shannon deixa clara a importância de sua fórmula no quesito implicações, nas melhorias esperadas nos processos de comunicação, ou seja, pode-se dizer, na sua relação com "tecnologia". Por exemplo, é razoável supor que uma mensagem escolhida entre dez mensagens possíveis transmite uma menor quantidade de informação que uma mensagem escolhida entre um milhão de mensagens possíveis (PIERCE, 1961). Ou, sob outra ótica, é razoável supor que a incerteza associada a um evento simples, com um único componente, $X = \{x_1\}$, seja menor do que a incerteza associada a um evento maior $X = \{x_1, x_2, \ldots, x_n\}$. Nesse caso, é plausível estimar que encontrar um livro numa biblioteca é bem menos incerto do que encontrar um livro em dez bibliotecas. Assim, é relevante indicar que a função $H(X)$ procurada deva ser monotônica crescente. Além disso, $H(X)$ deve ser contínua em p_i, isto é, para pequenas alterações nas probabilidades p_i obtém-se pequenas alterações na medida de incerteza em $H(X)$. Khinchin (1957), Faddeev (1956), Aczél e Daróczy (1975), por exemplo, forneceram, por outro lado, demonstrações matemáticas de $H(X)$, diferentes da exposta por Shannon, e auxiliaram no surgimento de uma área matemática voltada à **Caracterização de Medidas de Informação** (ACZÉL; DARÓCZY, 1975, EBANKS; SAHOO;

SANDER, 1998). Com isso, investigam-se não somente as diferentes deduções da fórmula $H(X)$ de Shannon, como também as possíveis alterações no sistema axiomático que a produz como teorema. Isso abre espaço para questionamentos sobre a relação entre fórmulas, axiomas e sua importância operacional, diga-se, prática. Torna-se então recorrente o questionamento sobre quais axiomas são "relevantes" na dedução da fórmula $H(X)$, cuja ideia subjacente ao termo "relevante" vem de que cada axioma esteja em consonância com uma determinada aplicação a ser utilizada (ACZÉL; FORTE; NG, 1974, ACZÉL; DARÓCZY, 1975, ACZÉL, 1984a, ACZÉL, 1984b). Pode-se dizer, grosso modo, que um questionamento interessante é o de saber se o conceito de informação pode, ou não, ser separado da fórmula de $H(X)$ de Shannon (INGARDEN; URBANIK, 1962). Nessa mesma esteira, neste artigo, opta-se por escrever sobre a dedução matemática da fórmula de entropia de Shannon, sem pretensões de ineditismo, e não sobre as consequências tecnológicas advindas dela. O foco é muito mais de uma exposição didática, que apresente num único texto, a dedução de $H(X)$, sem que um número excessivo de passagens matemáticas seja deixado a cargo do leitor, do que uma exposição que exiba a relação entre "informação" e tecnologia. Nota-se, com base na literatura sobre teoria e medidas da informação, que a relação entre o conceito de informação e a fórmula de Shannon ainda é exaustivamente investigada. Assim, uma exposição didática, sobre a parte formal de $H(X)$, não deixa de ser um artigo de divulgação científica, aliada a uma exposição matemática, com vistas à relação, sob a ótica da teoria de Shannon, entre os conceitos de informação e entropia.

Na sequência, expõe-se a demonstração da fórmula de entropia $H(X)$, com base em axiomas extraídos de Ash (1990, p. 8), os quais remetem-se a Shannon (1948). Doravante, neste texto, considera-se X uma variável randômica discreta:

$$X = \begin{pmatrix} x_1 & x_2 & x_3 & \dots & x_n \\ p_1 & p_2 & p_3 & \dots & p_n \end{pmatrix},$$

em que cada $p_i > 0$ e $\sum_{i=1}^{n} p_i = 1$. Segundo Shannon, $H(X)$ é uma fórmula que representa a incerteza associada a X, **a entropia de** X.

2 Demonstração de que $H(X) = -C\sum_{i=1}^{M} p_i \log p_i$

Robert Ash, em seu livro *Information Theory* (ASH, 1990), expõe a axiomática utilizada por Shannon (1948) e deduz a fórmula 1 de entropia, como única possível. Ele se fundamenta, tal como em Shannon (1948), nos seguintes axiomas:

Axioma 2.1. $H\left(\frac{1}{M}, \frac{1}{M}, \ldots, \frac{1}{M}\right) = f(M)$ *é uma função monotônica crescente em* M $(M = 1, 2, \ldots)$;

Axioma 2.2. $f(ML) = f(M) + f(L) \qquad (M, L = 1, 2, \ldots)$;

Axioma 2.3. $H(p_1, p_2, \ldots, p_M) =$

$$= H(p_1 + p_2 + \ldots + p_r, p_{r+1} + \ldots + p_M) + (p_1 + p_2 + \ldots + p_r) H\left(\frac{p_1}{\sum_{i=1}^{r} p_i}, \ldots, \frac{p_r}{\sum_{i=1}^{r} p_i}\right)$$

$$+ (p_{r+1} + p_{r+2} + \ldots + p_M) H\left(\frac{p_{r+1}}{\sum_{i=r+1}^{M} p_i}, \ldots, \frac{p_n}{\sum_{i=r+1}^{M} p_i}\right) \qquad (r = 1, 2, \ldots, M-1);$$

Axioma 2.4. $H(p, 1-p)$ *é uma função de* p*, contínua.*

A única função que satisfaz os axiomas 2.1, 2.2, 2.3 e 2.4 é

$$H(X) = H(p_1, p_2, \ldots, p_M) = -C \sum_{i=1}^{M} p_i \log p_i. \tag{1}$$

A demonstração é elaborada em duas partes. Na primeira parte, mostra-se que, se a fórmula $H(X)$ é verdadeira, então os axiomas 2.1, 2.2, 2.3 e 2.4 são verdadeiros. Na segunda, mostra-se que, se os axiomas 2.1, 2.2, 2.3 e 2.4 são verdadeiros, então, como consequência, se tem a fórmula $H(X)$.

2.1 **Primeira parte.** Se vale $H(X)$, então valem os axiomas 2.1, 2.2, 2.3 e 2.4.

Proposição 2.1 (Axioma 2.1)**.** *(KHINCHIN, 1957, p. 9-10) A função*

$$H(p_1, p_2, \ldots, p_M) = -C \sum_{i=1}^{M} p_i \log p_i = -C(p_1 \log p_1 + p_2 \log p_2 + \ldots + p_M \log p_M)$$

é monotônica crescente para M *eventos equiprováveis.*

Demonstração: Supõe-se que para cada i, $p_i = \frac{1}{M}$, levando-se em conta que são M eventos equiprováveis. Assim, a função H pode ser entendida como sendo de uma única variável, qual seja,

a variável M. Dessa forma, seja

$$f(M) = H\left(\frac{1}{M}, \frac{1}{M}, \ldots, \frac{1}{M}\right) = -C\underbrace{\left(\frac{1}{M}\log\frac{1}{M} + \ldots + \frac{1}{M}\log\frac{1}{M}\right)}_{M\ vezes}.$$

$$f(M) = -CM\left(\frac{1}{M}\log\frac{1}{M}\right) = -C\log M^{-1} = C\log M.$$

Como $f(M+1)$ implica $(M+1)$ argumentos, tem-se que

$$f(M+1) = H\left(\frac{1}{M+1}, \ldots, \frac{1}{M+1}\right) = -C\underbrace{\left(\frac{1}{M+1}\log\frac{1}{M+1} + \ldots + \frac{1}{M+1}\log\frac{1}{M+1}\right)}_{(M+1)\ vezes}.$$

$$f(M+1) = -C(M+1)\frac{1}{M+1}\log(M+1)^{-1} = C\log(M+1).$$

Portanto,

$$f(M+1) = C\log(M+1).$$

Tem-se então que $\log M < \log(M+1)$, haja vista que $0 < 1 \Rightarrow M < M+1$, desde que a base do logaritmo seja maior que 1. Assim, ao multiplicar ambos os lados por $C > 0$, tem-se $C\log M < C\log(M+1)$. Portanto, $f(M) < f(M+1)$. Assim, $f(M)$ é uma função monotônica crescente. ■

Proposição 2.2 (Axioma 2.2)**.** *Assume-se que* $H(X) = -C\sum_{i=1}^{M} p_i \log p_i$. *Se*

$$H\left(\frac{1}{M}, \frac{1}{M}, \ldots, \frac{1}{M}\right) = f(M),$$

então $f(M.N) = f(M) + f(N)$.

Demonstração: Com base na proposição 2.1, para eventos equiprováveis,

$$H\left(\frac{1}{M}, \frac{1}{M}, \ldots, \frac{1}{M}\right) = C\log M \quad \text{e} \quad H\left(\frac{1}{N}, \frac{1}{N}, \ldots, \frac{1}{N}\right) = C\log N.$$

Nesse caso, nota-se que

$$f(M.N) = \underbrace{H\left(\frac{1}{M.N}, \frac{1}{M.N}, \ldots, \frac{1}{M.N}\right)}_{M.N\ \text{vezes}} = C\log(M.N) = C\log M + C\log N.$$

Assim, $f(M) + f(N) = C\log M + C\log N = C\log(M.N) = f(M.N)$. ■

Proposição 2.3 (Axioma 2.3)**.** *Se vale* $H(X)$, *então vale o axioma de agrupamento.*

Demonstração:

$$H(p_1+p_2+\ldots+p_r,p_{r+1}+\ldots+p_M)+(p_1+p_2+\ldots+p_r)H\left(\frac{p_1}{\sum_{j=1}^{r}p_j},\ldots,\frac{p_r}{\sum_{j=1}^{r}p_j}\right)$$

$$+(p_{r+1}+p_{r+2}+\ldots+p_M)H\left(\frac{p_{r+1}}{\sum_{j=r+1}^{M}p_j},\ldots,\frac{p_M}{\sum_{j=r+1}^{r}p_j}\right)$$

$$=-C\left(\sum_{j=1}^{r}p_j\right)\log\left(\sum_{j=1}^{r}p_j\right)-C\left(\sum_{j=r+1}^{M}p_j\right)\log\left(\sum_{j=r+1}^{M}p_j\right)$$

$$+\left(\sum_{j=1}^{r}p_j\right)\left[-C\sum_{i=1}^{r}\frac{p_i}{\sum_{j=1}^{r}p_j}\log\frac{p_i}{\sum_{j=1}^{r}p_j}\right]+\left(\sum_{j=r+1}^{M}p_j\right)\left[-C\sum_{i=r+1}^{M}\frac{p_i}{\sum_{j=r+1}^{M}p_j}\log\frac{p_i}{\sum_{j=r+1}^{M}p_j}\right]$$

$$=-C\left(\sum_{j=1}^{r}p_j\right)\log\left(\sum_{j=1}^{r}p_j\right)-C\left(\sum_{j=r+1}^{M}p_j\right)\log\left(\sum_{j=r+1}^{M}p_j\right)$$

$$+\left[-C\sum_{i=1}^{r}p_i\log\frac{p_i}{\sum_{j=1}^{r}p_j}\right]+\left[-C\sum_{i=r+1}^{M}p_i\log\frac{p_i}{\sum_{j=r+1}^{M}p_j}\right]$$

$$=-C\sum_{j=1}^{r}p_j\log\sum_{j=1}^{r}p_j-C\sum_{j=r+1}^{M}p_j\log\sum_{j=r+1}^{M}p_j$$

$$+\left[-C\sum_{i=1}^{r}p_i\log p_i+C\sum_{i=1}^{r}p_i\log\sum_{j=1}^{r}p_j\right]+\left[-C\sum_{i=r+1}^{M}p_i\log p_i+C\sum_{i=r+1}^{M}p_i\log\sum_{j=r+1}^{M}p_j\right]$$

$$=-C\sum_{i=1}^{M}p_i\log p_i=H(p_1,p_2,\ldots,p_M)$$

■

Proposição 2.4 (Axioma 2.4)**.** *Se $H(X)$ é verdade, então $H(p_1,p_2,\ldots,p_M)$ é uma função contínua para cada variável p_i no intervalo $(0,1]$.*

Demonstração: Nesse caso, leva-se em conta que

$$\sum_{i=1}^{M} p_i = p_1 + p_2 + \ldots + p_{M-1} + p_M = 1.$$

$$H(p_1, p_2, \ldots, p_M) = -\sum_{i=1}^{M} p_i \log p_i = -\left(p_1 \log p_1 + p_2 \log p_2 + \ldots + p_M \log p_M\right)$$

$$= -(p_1 \log p_1 + p_2 \log p_2 + \ldots + p_{M-1} \log p_{M-1} +$$

$$+\underbrace{(1 - p_1 - p_2 - \ldots - p_{M-1})}_{p_M} \log \underbrace{(1 - p_1 - p_2 - \ldots - p_{M-1})}_{p_M}).$$

Nota-se que $p_1, p_2, \ldots, p_{M-1}$ e também $1 - p_1 - p_2 - \ldots - p_{M-1}$ são variáveis contínuas em $(0, 1]$. Soma e produto de funções contínuas ainda é uma função contínua. Também, a função logarítmica é uma função contínua, bem como o logaritmo de uma função contínua. Portanto, para cada variável p, $H(p, 1-p)$ é uma função contínua em $(0, 1]$. ■

2.2 **Segunda parte.** Se valem os axiomas 2.1, 2.2, 2.3 e 2.4, então vale $H(X)$.

Convém previamente demonstrar algumas proposições que serão utilizadas na demonstração da fórmula $H(X)$ de entropia (ASH, 1990, REZA, 1961).

Proposição 2.5. *(ASH, 1990, p. 9) Para todos os inteiros positivos M e n, tem-se:*

$$f(M^n) = nf(M).$$

Demonstração: A demonstração é feita por indução no valores de n presentes na fórmula

$$f(M^n) = nf(M). \tag{2}$$

A fórmula vale para $n = 1$, haja vista que

$$f(M^1) = f(M) = 1.f(M) = f(M).$$

Assume-se que a fórmula 2 é verdadeira para um $n = k$ qualquer, ou seja, vale, por hipótese, que $f(M^k) = kf(M)$. Mostra-se que vale para $n = k + 1$. De fato,

$$f(M^{k+1}) = f(M^k.M^1).$$

Pelo axioma 2.2, tem-se que

$$f(M^{k+1}) = f(M^k.M^1) = f(M^k) + f(M).$$

Pela hipótese da indução, $f(M^k) = kf(M)$. Logo,

$$f(M^{k+1}) = f(M^k.M^1) = f(M^k) + f(M) = kf(M) + f(M) = (k+1)f(M).$$

Portanto, $f(M^{k+1}) = (k+1)f(M)$. ■

Proposição 2.6. *Se $M > 1$ é um inteiro fixo, então, para qualquer r inteiro positivo existe $k \in \mathbb{N}$, tal que, conforme Ash (1990),*

$$M^k \leq 2^r < M^{k+1}.$$

Demonstração: Sejam $x = \frac{r}{\log_2 M}$ e $k = \lfloor x \rfloor$. Então,

$$k \leq \frac{r}{\log_2 M} < k+1 \Rightarrow k \log_2 M \leq r < (k+1)\log_2 M \Rightarrow \log_2 M^k \leq r < \log_2 M^{k+1}.$$

Como $r = \log_2 2^r$, então

$$\log_2 M^k \leq r < \log_2 M^{k+1} \Rightarrow \log_2 M^k \leq \log_2 2^r < \log_2 M^{k+1} \Rightarrow M^K \leq 2^r < M^{k+1}.$$

Portanto, para $M > 1$,

$$M^k \leq 2^r < M^{k+1}.$$

■

Proposição 2.7. *(ASH, 1990, p. 9) Se C é um número positivo e $M = 1, 2, 3, \ldots$, então*

$$f(M) = C \log M. \tag{3}$$

Demonstração: Consideram-se os casos em que $M = 1$ e o caso em que $M > 1$.

$M = 1$. Para $M = 1$, mostra-se que, na fórmula 3, se deve ter que $f(1) = C \log 1 = 0$. Com o auxílio do axioma 2.2, tem-se:

$$f(1) = \underbrace{f(1.1) = f(1) + f(1)}_{\text{axioma 2.2}}.$$

Portanto, se $f(1) = f(1) + f(1)$, então $f(1) = 0$. E isso satisfaz a fórmula 3 para $M = 1$.

$M > 1$. Se r é um inteiro positivo qualquer e $M > 1$ é um inteiro positivo fixo, então, conforme demonstrado na proposição 2.6, existe um $k \in \mathbb{N}$ tal que

$$M^k \leq 2^r < M^{k+1}.$$

Segue do axioma 2.1 que, como f é uma função crescente[1], então

$$f(M^k) \leq f(2^r) < f(M^{k+1}).$$

Pela proposição 2.5, tem-se que

$$kf(M) \leq rf(2) < (k+1)f(M).$$

Ao dividir por $r \neq 0$, tem-se

$$\frac{kf(M)}{r} \leq \frac{rf(2)}{r} < \frac{(k+1)f(M)}{r} =$$
$$\frac{kf(M)}{r} \leq f(2) < \frac{(k+1)f(M)}{r}.$$

Ao dividir por $f(M) \neq 0$, tem-se

$$\frac{kf(M)}{f(M)r} \leq \frac{f(2)}{f(M)} < \frac{(k+1)f(M)}{f(M)r}.$$

Portanto,

$$\frac{k}{r} \leq \frac{f(2)}{f(M)} < \frac{(k+1)}{r}. \tag{4}$$

Como a função logarítmica é uma função crescente, então é possível aplicar o logaritmo[2] na desigualdade $M^k \leq 2^r < M^{k+1}$ e obter $\log M^k \leq \log 2^r < \log M^{k+1}$, que é equivalente a $k \log M \leq r \log 2 < (k+1) \log M$. Ao dividir os componentes da desigualdade por $\log M$, tem-se

$$k \leq \frac{r \log 2}{\log M} < (k+1).$$

Ao dividir os componentes da desigualdade por r, tem-se

$$\frac{k}{r} \leq \frac{\log 2}{\log M} < \frac{(k+1)}{r}. \tag{5}$$

[1] A função f é crescente se para $x < y$ se tem que $f(x) < f(y)$ (BARTLE; SHERBERT, 2011, p. 153).
[2] A notação $\log M$ representa o logaritmo numa base qualquer maior do que 1.

Como a distância entre $\frac{k}{r}$ e $\frac{k+1}{r}$ é $\frac{1}{r}$, pode-se então afirmar, com base nas desigualdades 4 e 5, que

$$\left|\frac{\log 2}{\log M} - \frac{f(2)}{f(M)}\right| < \frac{1}{r}.$$

Levando-se em conta que o valor de M é o valor de um inteiro fixo e que o valor de r é fornecido, ou seja, é arbitrário, é possível analisar o caso em que $r \to \infty$. Nesse caso, tem-se que $\frac{1}{r} \to 0$. Com isso, obtém-se a igualdade $\frac{\log 2}{\log M} = \frac{f(2)}{f(M)}$. Isto é,

$$f(M) = \frac{f(2)}{\log 2} \log M.$$

Assumindo-se que $\frac{f(2)}{\log 2}$ é uma constante, digamos C, tem-se então a fórmula procurada:

$$f(M) = C \log M.$$

E isso completa a demonstração dessa proposição. ■

Proposição 2.8. *(ASH, 1990, p. 10) Se C é um número positivo e p é um número racional qualquer, então*

$$H(p, 1-p) = -C[p \log p + (1-p)\log(1-p)].$$

Demonstração: Seja $p = \frac{r}{s}$ um número racional em que r e s são inteiros positivos. Pelo axioma 2.3, é possível partilhar os s componentes em $r + (s-r)$ componentes, ou seja, escrever $f(s) = H\left(\frac{1}{s}, \ldots, \frac{1}{s}\right)$ como

$$f(s) = H\left(\underbrace{\frac{1}{s}, \frac{1}{s}, \ldots, \frac{1}{s}}_{r \text{ componentes}}, \underbrace{\frac{1}{s}, \frac{1}{s}, \frac{1}{s}, \ldots, \frac{1}{s}}_{(s-r) \text{ componentes}}\right).$$

Agora, pelo axioma 2.3, tem-se que

$$f(s) = H\left(\frac{r}{s}, \frac{s-r}{s}\right) + \frac{r}{s} f(r) + \frac{s-r}{s} f(s-r).$$

Pela proposição 2.7, aplicada em

$$f(s) = H\left(\frac{r}{s}, \frac{s-r}{s}\right) + \frac{r}{s} f(r) + \frac{s-r}{s} f(s-r),$$

e levando-se em conta que $p = \frac{r}{s}$, tem-se que $C \log s = H(p, 1-p) + Cp \log r + C(1-p)\log(s-r)$. Obtém-se então $H(p, 1-p) = -C[p \log r - \log s + (1-p)\log(s-r)]$, que vale

$$H(p, 1-p) = -C[p \log r - p \log s + p \log s - \log s + (1-p)\log(s-r)],$$

$$H(p, 1-p) = -C[p \log r - p \log s - (1-p) \log s + (1-p) \log(s-r)],$$

que é equivalente a $H(p, 1-p) = -C\left[p \log \frac{r}{s} + (1-p) \log \frac{s-r}{s}\right]$, isto é,

$$H(p, 1-p) = -C[p \log p + (1-p) \log(1-p)].$$

E a proposição está demonstrada. ■

Proposição 2.9. *(ASH, 1990, p. 10) Se C é um número positivo e $p \in (0, 1]$ é um número real, então*

$$H(p, 1-p) = -C[p \log p + (1-p) \log(1-p)].$$

Demonstração: Assume-se, como hipótese, a proposição 2.8. Seja $\{r_n\}$ uma sequência, em $\mathbb{R}$, de números racionais cujo limite é o número real p, ou seja, $\lim\limits_{n\to\infty} \{r_n\} = p$.

Assim, $H(p, 1-p) = H(\lim\limits_{n\to\infty} \{r_n\}, 1 - \lim\limits_{n\to\infty} \{r_n\})$. Como H é contínua,[3]

$$= \lim_{n\to\infty} H(\{r_n\}, 1 - \{r_n\}) = \lim_{n\to\infty} [-C[\{r_n\} \log\{r_n\} + (1 - \{r_n\}) \log(1 - \{r_n\})]$$

$$= -C[\lim_{n\to\infty} \{r_n\} \log[\lim_{n\to\infty} \{r_n\}] + (1 - \lim_{n\to\infty} \{r_n\}) \log(1 - [\lim_{n\to\infty} \{r_n\}])]$$

$$= -C[p \log p + (1-p) \log(1-p)].$$

■

Teorema 2.10. *Ao se assumir os axiomas 2.1, 2.2, 2.3 e 2.4, deduz-se, como teorema, a fórmula*

$$H(p_1, p_2, \ldots, p_M) = -C \sum_{i=1}^{M} p_i \log p_i. \tag{6}$$

Demonstração: A demonstração é feita por indução em M. Para o caso em que $M > 2$, lança-se mão do axioma 2.3, para mostrar que

$$H(p_1, p_2, \ldots, p_M) = H(p_1 + p_2 + \ldots + p_{M-1}, p_M) +$$

$$(p_1 + p_2 + \ldots + p_{M-1}) H\left(\frac{p_1}{\sum\limits_{i=1}^{M-1} p_i}, \ldots, \frac{p_{M-1}}{\sum\limits_{i=1}^{M-1} p_i}\right) + p_M H(1).$$

[3]Conforme consta em Bartle e Sherbert (2011, p. 130-133), sob certas condições, se f e g são funções contínuas de números reais e k é uma constante, então $(f+g)$, $(f-g)$, (fg), (kf) e $(g \circ f)$ são funções contínuas em $a \in \mathbb{R}$. Assim, $\lim\limits_{x\to a}(\log f(x)) = \log(\lim\limits_{x\to a} f(x))$.

Assume-se, como hipótese de indução, que essa fórmula é válida para todo os valores de 1 até $M-1$. Nota-se que $p_M H(1) = p_M f(1) = p_M.0 = 0$. Assim,

$$H(p_1, p_2, \ldots, p_M) = -C[(p_1 + p_2 + \ldots + p_{M-1}) \log(p_1 + p_2 + \ldots + p_{M-1}) + p_M \log p_M]$$

$$-C(p_1 + \ldots + p_{M-1}).\left[\frac{p_1}{\sum_{i=1}^{M-1} p_i} \log\left(\frac{p_1}{\sum_{i=1}^{M-1} p_i}\right) + \ldots + \frac{p_{M-1}}{\sum_{i=1}^{M-1} p_i} \log\left(\frac{p_{M-1}}{\sum_{i=1}^{M-1} p_i}\right)\right] + p_M(0)$$

$$= -C\left[\left(\sum_{i=1}^{M-1} p_i\right) \log\left(\sum_{i=1}^{M-1} p_i\right) + p_M \log p_M\right] - C\left[\sum_{i=1}^{M-1} p_i \log p_i - \left(\sum_{i=1}^{M-1} p_i\right) \log \sum_{i=1}^{M-1} p_i\right]$$

$$= -C \sum_{i=1}^{M} p_i \log p_i.$$

Isto é,

$$H(p_1, p_2, \ldots, p_M) = -C \sum_{i=1}^{M} p_i \log p_i.$$

E a demonstração está completa. ■

Ou seja, com a primeira parte (2.1) e segunda parte (2.2), mostrou-se que a única função que satisfaz os axiomas 2.1, 2.2, 2.3 e 2.4 é $H(p_1, p_2, \ldots, p_M) = -C \sum_{i=1}^{M} p_i \log p_i$.

2.3 Informação de um evento simples

A demonstração da fórmula $H(X)$ é independente da base do logaritmo, haja vista que sempre é possível lançar mão da clássica fórmula de mudança de base em logaritmos, $\log_b a = \frac{\log_c a}{\log_c b}$.

Algumas unidades de medida da entropia, em Teoria da Informação, são expressas de acordo com a base do logaritmo (RIOUL, 2018, p. 27) :

- Se a entropia $H(X)$ é expressa em **bits**, então o logaritmo é escrito na base 2 e a entropia máxima é um bit.
- Se a entropia $H(X)$ é expressa em **decits**, ou como se diz também **Hartley**, então a base do logaritmo é 10 e a entropia máxima é um decit.

- Se a entropia $H(X)$ é expressa em **nats**, então o logaritmo é expresso na base e e a entropia máxima é um nat.

A fórmula $H(X)$ refere-se à incerteza de um sistema completo de eventos probabilísticos. No caso de um único evento, um evento simples, é possível caracterizar a função $I(p) = -\log p$. Nesse caso, $I(p)$ pode ser lido como a informação associada a um único evento.

Teorema 2.11. *Para $p \in (0, 1]$, a função $I(p) = -\log_2 p$ é a única que satisfaz os seguintes axiomas (ASH, 1990, ACZÉL; DARÓCZY, 1975):*

Axioma I1) $I(pq) = I(p) + I(q) \qquad 0 < p \leq 1, \quad 0 < q \leq 1;$

Axioma I2) $I(p)$ *é uma função contínua e decrescente em* $p \in (0, 1]$;

Axioma I3) $I(\frac{1}{2}) = 1.$

Demonstração:

a) Seja $f(n) = I\left(\frac{1}{n}\right)$. Então, ao assumir o **Axioma I1**, tem-se que $f(mn) = f(m) + f(n)$. Assim,

$$f(m.n) = I\left(\frac{1}{mn}\right) = \underbrace{I\left(\frac{1}{m}.\frac{1}{n}\right) = I\left(\frac{1}{m}\right) + I\left(\frac{1}{n}\right)}_{\text{Axioma I1}} = f(m) + f(n).$$

b) Seja agora $m < n$. Assim, $\frac{1}{n} < \frac{1}{m}$. Como, pelo **Axioma I2**, I é decrescente, então

$$m < n \Rightarrow \frac{1}{n} < \frac{1}{m} \Rightarrow I\left(\frac{1}{m}\right) < I\left(\frac{1}{n}\right) \Rightarrow f(m) < f(n).$$

Conclui-se então que f é uma função crescente que satisfaz as seguintes condições:

A1) $f(mn) = f(m) + f(n)$;

A2) $m < n \Rightarrow f(m) < f(n)$.

A função f então satisfaz o axioma 2.1 e o axioma 2.2. Logo, pela proposição 2.7, é possível concluir que $f(n) = -C \log n$, onde $C > 0$ e a notação $\log n$ indica um logaritmo numa base b maior do que 1.

c) Seja agora $p = \frac{r}{s} \in \mathbb{Q}$. Segue-se que

$$I\left(\frac{1}{s}\right) = \underbrace{I\left(\frac{r}{s}.\frac{1}{r}\right) = I\left(\frac{r}{s}\right) + I\left(\frac{1}{r}\right)}_{\text{Axioma I1}}.$$

Portanto,

$$I\left(\frac{1}{s}\right) = I\left(\frac{r}{s}\right) + I\left(\frac{1}{r}\right) \Rightarrow I\left(\frac{r}{s}\right) = I\left(\frac{1}{s}\right) - I\left(\frac{1}{r}\right) \Rightarrow I\left(\frac{r}{s}\right) = f(s) - f(r)$$

$$\Rightarrow I\left(\frac{r}{s}\right) = C\log s - C\log r = C\log\frac{s}{r}.$$

Logo,

$$I(p) = I\left(\frac{r}{s}\right) = C\log\frac{s}{r} = -C\log\frac{r}{s} = -C\log p.$$

Portanto, $I(p) = -C\log p$, para todo número racional p.

d) Para o caso em que $p \in \mathbb{R}$, segue-se tal como na proposição 2.8. Seja $\{r_n\}$ uma sequência, em $\mathbb{R}$, de números racionais cujo limite é um número real p, ou seja,

$$\lim_{n\to\infty}\{r_n\} = p.$$

Assim,

$$I(p) = \underbrace{I(\lim_{n\to\infty}\{r_n\}) = \lim_{n\to\infty} I(\{r_n\})}_{I \text{ é uma função contínua}} = \lim_{n\to\infty}(-C\log\{r_n\})$$

$$= -C\log\left(\lim_{n\to\infty}\{r_n\}\right) = -C\log p.$$

e) Tem-se então que $I(p) = -C\log_2 p$. Pelo **axioma I3**, $I(\frac{1}{2}) = 1$, logo

$$I\left(\frac{1}{2}\right) = -C\log_2\frac{1}{2} = 1 \Rightarrow -C\log_2 2^{-1} = 1 \Rightarrow C\log_2 2 = 1 \Rightarrow C = 1.$$

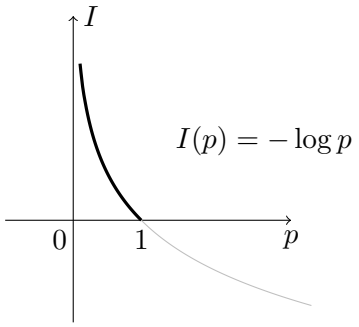
Portanto,

$$I(p) = -\log_2 p.$$

A demonstração está completa. ■

Conforme exposto na Figura 1, pode-se observar que, para $0 < p \leq 1$, quanto mais incerto é o evento (mais p tende a 0), mais se tem de aumento de informação. Por outro lado, quanto mais certo é o evento (p tende a 1), menos informação se tem.

Figura 1: Gráfico da "informação" $I(p)$.



Fonte: Elaboração dos autores (2020).

3 Um exemplo ilustrativo

Considera-se a constante $C = 1$ e $H(X) = -\sum_{i=1}^{n} p_i \log p_i$ com logaritmo na base 2. Além disso, seja uma variável randômica discreta X:

$$X = \begin{pmatrix} A & B & C \\ p_1 & p_2 & p_3 \end{pmatrix},$$

em que $p_1 = \frac{1}{3}, p_2 = \frac{2}{5}$ e $p_3 = \frac{4}{15}$ e $\sum_{i=1}^{3} p_i = \frac{1}{3} + \frac{2}{5} + \frac{4}{15} = 1$. Nesse caso, nota-se, para logaritmos na base 2, que:

$$H(X) = -\sum_{i=1}^{3} p_i \log p_i = -p_1 \log p_1 - p_2 \log p_2 - p_3 \log p_3$$

e que

$$H\left(\frac{1}{3}, \frac{2}{5}, \frac{4}{15}\right) = -\frac{1}{3}\log\frac{1}{3} - \frac{2}{5}\log\frac{2}{5} - \frac{4}{15}\log\frac{4}{15} \approx 1,5655 \tag{7}$$

representa a incerteza associada à variável discreta X.

Supõe-se que haja uma partição de X em subeventos:

$$X = \begin{pmatrix} A & M \\ p_1 & p_2 + p_3 \end{pmatrix},$$

em que $M = B \cup C$:

$$M = \begin{pmatrix} B & C \\ p_2 & p_3 \end{pmatrix}.$$

Assim, pelo axioma 2.3, tem-se que

$$H(p_1,p_2,p_3) = H(p_1,p_2+p_3) + p_1 H\left(\frac{p_1}{\sum_{i=1}^{1} p_i}\right) + (p_2+p_3)H\left(\frac{p_2}{\sum_{i=2}^{3} p_i}, \frac{p_3}{\sum_{i=2}^{3} p_i}\right).$$

Como $p_1 = \frac{1}{3}, p_2 = \frac{2}{5}$ e $p_3 = \frac{4}{15}$, segue-se que

$$H\left(\frac{1}{3},\frac{2}{5},\frac{4}{15}\right) = H\left(\frac{1}{3},\frac{10}{15}\right) + \frac{1}{3}H\left(\frac{\frac{1}{3}}{\frac{1}{3}}\right) + \frac{10}{15}H\left(\frac{\frac{2}{5}}{\frac{10}{15}}, \frac{\frac{4}{15}}{\frac{10}{15}}\right);$$

$$H\left(\frac{1}{3},\frac{2}{5},\frac{4}{15}\right) = H\left(\frac{1}{3},\frac{10}{15}\right) + \frac{1}{3}H(1) + \frac{10}{15}H\left(\frac{3}{5},\frac{2}{5}\right).$$

Como $H(1) = -1\log 1 = 0$, tem-se que

$$H\left(\frac{1}{3},\frac{2}{5},\frac{4}{15}\right) = H\left(\frac{1}{3},\frac{10}{15}\right) + \frac{10}{15}H\left(\frac{3}{5},\frac{2}{5}\right).$$

em que

a) $H\left(\frac{1}{3},\frac{10}{15}\right) = -\frac{1}{3}\log\frac{1}{3} - \frac{10}{15}\log\frac{10}{15} \approx 0,9182;$

b) $\frac{10}{15}H\left(\frac{3}{5},\frac{2}{5}\right) = \frac{10}{15}\left(-\frac{3}{5}\log\frac{3}{5} - \frac{2}{5}\log\frac{2}{5}\right) = -\frac{2}{5}\log\frac{3}{5} - \frac{4}{15}\log\frac{2}{5} \approx 0,6473.$

Portanto,

$$H\left(\frac{1}{3},\frac{10}{15}\right) + \frac{10}{15}H\left(\frac{4}{15},\frac{2}{5}\right) = -\frac{1}{3}\log\frac{1}{3} - \frac{10}{15}\log\frac{10}{15} - \frac{2}{5}\log\frac{3}{5} - \frac{4}{15}\log\frac{2}{5} \approx 1,5655. \quad (8)$$

Assim, as equações 7 e 8 são iguais, isto é:

$$H\left(\frac{1}{3},\frac{2}{5},\frac{4}{15}\right) = H\left(\frac{1}{3},\frac{10}{15}\right) + \frac{10}{15}H\left(\frac{4}{15},\frac{2}{5}\right).$$

Conclui-se então que a incerteza associada a um evento geral se mantém a mesma ao particionar o evento em subeventos. Isso significa que a incerteza não pode aumentar ao particionar um evento em subeventos.

4 Considerações finais

A demonstração matemática da fórmula de entropia de Shannon associa-se ao conceito de medida de informação, que é, em essência, o núcleo da Teoria da Informação. As possíveis indagações sobre as demonstrações matemáticas de medidas de informação revelam as caracterizações de medidas de informação, que exibem a riqueza da estrutura subjacente à fórmula de entropia, e abrem espaço para questionamentos acerca da noção de "informação" poder, ou não, ser separada da fórmula $H(X)$ e do conceito de probabilidade (INGARDEN; URBANIK, 1962). Não há como negar as inúmeras interpretações científicas presentes nos axiomas de Shannon e na fórmula $H(X)$ (ACZÉL; DARÓCZY, 1975, EBANKS; SAHOO; SANDER, 1998). O objetivo de uma exposição pedagógica da demonstração da fórmula de entropia de Shannon é auxiliar o leitor, estudante, das áreas de engenharia, Teoria da Informação etc., na compreensão das nuances dessa fórmula. Entende-se que exposições como a deste texto possam estimular, tal como a geometria plana de Euclides[4], o desenvolvimento de novos caminhos na Teoria da Informação.

Cabe observar que, neste texto, optou-se pelo estudo da fórmula $H(X) = -C\sum_{i=1}^{M} p_i \log p_i$ de Shannon, sob a ótica de modelos discretos em vez da fórmula $H(\mathscr{X}) = -\int_{-\infty}^{+\infty} p(x)\log p(x)dx$ para modelos contínuos, haja vista o objetivo ser o de investigar equações funcionais e os axiomas indicados por Shannon para deduzir $H(X)$. Enquanto que, no caso discreto, X é uma variável aleatória discreta, sendo p_i um valor da probabilidade associada a $X = x_i$, no caso contínuo tem-se $p(x)$ como uma função de densidade de probabilidade da variável aleatória absolutamente contínua $\mathscr{X}$. Em contraste com os modelos discretos em que $H(X) \geq 0$, para modelos contínuos $H(\mathscr{X})$ pode ser negativa, positiva ou arbitrariamente grande (REZA, 1961, p. 268, ASH, 1990, p. 236, COVER; THOMAS, 2006, p. 247). Na literatura científica, ainda há ricas discussões acerca das propriedades e interpretações resultantes das possíveis interseções entre os conceitos de entropia discreta $H(X)$ (SHANNON, 1948, p. 393) e entropia diferencial $H(\mathscr{X})$ (SHANNON, 1948, p. 628) que, entre outras aplicações, acabam também por iluminar dois importantes problemas em engenharia da comunicação, quais sejam, compressão de dados e razão de transmissão.

[4]Nota-se que geometrias não-euclidianas surgiram com base na alteração do axioma das paralelas de Euclides.

Agradecimentos

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001.

Referências

ACZÉL, J.; DARÓCZY, Z. **On measures of information and their characterizations**. v. 115, New York: Academic Press, 1975.

ACZÉL, J.; FORTE, B.; NG, C. T. Why the Shannon and Hartley entropies are “natural”. **Advances in Applied Probability**, v. 6, n. 1, p. 131-146, 1974. DOI: https://doi.org/10.2307/1426210.

ACZÉL, J. Measuring information beyond communication theory: Some probably useful and some almost certainly useless generalizations. **Information processing & management**, v. 20, n. 3, p. 383-395, 1984a.

ACZÉL, J. Measuring information beyond communication theory: Why some generalized information measures may be useful, others not. Survey paper. **Aequationes Mathematicae**, v. 27, p. 1-19, 1984b.

ASH, R. B. **Information Theory**. New York: Dover Publications, 1990.

BARTLE, R. G.; SHERBERT, D. R. **Introduction to Real Analysis**. 4. ed, New York: John Wiley & Sons, 2011.

CLAUSIUS, R. **Abhandlungen über die mechanische Wärmetheorie**. Braunschweig: Druck Und Verlag Von Friedrich Vieweg Und Sohn, 1864.

COVER, T. M.; THOMAS, J. A. **Elements of Information Theory**. 1. ed. John Wiley & Sons, 1991, 2. ed., NJ, 2006.

EBANKS, B.; SAHOO, P.; SANDER, W. **Characterizations of Information Measures**. Singapura: World Scientific, 1998.

FADDEEV, D. K. On the Concept of Entropy of a Finite Probability Scheme. Originalmente publicado em Russo em **Uspekhi Matematicheskikh Nauk**, v. 11, n. 1 (67), p. 227-231, 1956.

INGARDEN, R. S.; URBANIK, K. Information without probability. **Colloquium Mathematicum**, v. IX, p. 132-150, 1962.

KHINCHIN, A. I. **Mathematical Foundations of Information Theory**. Trad. R. A. Silverman; M. D. Friedman. New York: Dover Publications, 1957.

MAGOSSI, J. C.; PAVIOTTI, J. R. Incerteza em Entropia. **Revista Brasileira de História da Ciência**, Rio de Janeiro, v. 12, n. 1, p. 84-96, jan/jun 2019.

PIERCE, J. R. **An Introduction to Information Theory – Symbols, Signals and Noise**. New York: Dover Publications, 1961.

REZA, F. M. **An Introduction to Information Theory**. New York: McGraw-Hill, 1961.

RIOUL, O. **Teoria da Informação e da Codificação**. Trad. José Carlos Magossi. Campinas: Editora da Unicamp, 2018.

SHANNON, C. E. A Mathematical Theory of Communication. **The Bell System Technical Journal**, v. 27, p. 379-423, p. 623-656, jul. 1948.

SHANNON, C. E.; WEAVER, W. **The Mathematical Theory of Communication**. Urbana, Illinois: University of Illinois Press, 1949.