ANO XX — N.º 74-75 **GAZETA DE MATEMÁTICA** MARÇO/JUNHO — 1959

EDITOR: *Gazeta de Matemática, Lda.* ADMINISTRADOR: *A. Sá da Costa*

REDACTORES: *J. Gaspar Teixeira, J. Morgado e J. da Silva Paulo*

Composto na Tipografia Matemática, Lda. — Rua Diário de Notícias, 134 - 1.º - Esq. — Telef. 29449 — LISBOA 2

A definição de entropia em cálculo das probabilidades

por **J. J. Dionísio**

O propósito deste artigo é expor como se introduz no Cálculo das Probabilidades o conceito de entropia. Colocar-nos-emos naturalmente no caso mais simples das distribuições discretas e a isso se limitarão as nossas considerações. Seguiremos para o efeito a primeira das memórias de A. I. KHINCHIN editadas pela casa Dover de Nova York sob o título *Mathematical Foundations of Information Theory*, editadas também em Berlim (Deutscher Verlag der Wissenschaften) acompanhadas de trabalhos de outros autores, sob a epígrafe *Arbeiten zur Informationstheorie.*

Suponhamos que o resultado de uma experiência A, em lugar de ser perfeitamente determinado, comporta certo número n de modalidades, a designar por $A_1, \cdots, A_n$. Suponhamos mais que uma das modalidades se produz com certeza (cada vez que se realiza a experiência) e que, além disso, uma só tem lugar, o que quer dizer que as n modalidades são *incompatíveis* duas a duas.

Sendo possível prescrever aos acontecimentos $A_1, \cdots, A_n$ probabilidades $p_1, \cdots, p_n$, respectivamente, tem-se nas condições anteriores

$$p_1 + p_2 + \cdots + p_n = 1$$

Dizemos então que A é um *sistema completo de acontecimentos* e escrevemos

$$A = \begin{pmatrix} A_1 & \cdots & A_n \\ p_1 & \cdots & p_n \end{pmatrix}$$

Assim, no lançamento de um dado homogéneo, o aparecimento A_1 de um número de pontos igual ou inferior a 4 e o acontecimento contrário A_2 (aparecimento de 5 ou 6 pontos) constituem um sistema completo A, tendo-se

$$A = \begin{pmatrix} A_1 & A_2 \\ \frac{2}{3} & \frac{1}{3} \end{pmatrix}$$

pois que é $\frac{2}{3} = 4 \times \frac{1}{6}$ a probabilidade de surgir qualquer das faces com 1, 2, 3 ou 4 pontos e $\frac{1}{3} = 2 \times \frac{1}{6} = 1 - \frac{2}{3}$ a probabilidade de A_2.

Como o resultado da realização da experiência A não é unìvocamente determinado, segue-se que a cada sistema completo de $n > 1$ acontecimentos anda ligada uma *indeterminação* ou *incerteza* quanto ao resultado de uma experência que venha a efectuar-se. E por outro lado, uma vez realizada esta, o

que era incerteza é agora *informação* — o saber-se qual a modalidade que teve lugar. E compreende-se que, quanto maior a incerteza inicial, mais ampla a informação final.

Ao sistema

$$\begin{pmatrix} A_1 \\ 1 \end{pmatrix}$$

pode dizer-se que corresponde uma incerteza nula, já que o resultado é unìvocamente determinado; e por isso mesmo é nula a informação obtida, pois conhecíamos de antemão esse mesmo resultado.

Dos sistemas

$$A = \begin{pmatrix} A_1 & A_2 & A_3 \\ 0{,}01 & 0{,}02 & 0{,}97 \end{pmatrix}$$

e

$$B = \begin{pmatrix} B_1 & B_2 \\ 0{,}5 & 0{,}5 \end{pmatrix}$$

é evidente que corresponde ao primeiro uma incerteza *menor* do que aquela contida no segundo: é quase certo que A aparece sob a forma A_3 (97 vezes em cada 100) ao passo que é igualmente provável o realizar-se B sob a forma B_1 ou sob a forma B_2.

Ora o que se pretende é justamente examinar a possibilidade de definir por cada sistema completo de acontecimentos A uma função $H(A)$ que de algum modo se comporte como a *medida da sua indeterminação* ou, que é o mesmo, como a *medida da informação* em que se traduz a realização de cada experiência.

A expressão analítica da função $H(A)$ dependerá, é claro, das variáveis $p_1, \cdots, p_n$ e só dependerá destas, uma vez que os valores que elas tomem em cada caso definem completamente o sistema A do ponto de vista probabilístico.

E a exigência de que a função

$$H(A) = H(p_1, p_2, \cdots p_n)$$

satisfaça o requisito anterior — medir a indeterminação ou a informação contida no sistema — pode reformular-se de uma maneira mais concreta através de certas propriedades a que ela deverá obedecer.

Quais sejam algumas dessas propriedades não é difícil descortinar.

Primeira. Se algum p_k iguala a unidade, o que obriga ao anulamento dos restantes, deve ter-se

$$H(A) = 0$$

porquanto não subsiste indeterminação alguma em tal hipótese, como já observámos.

Segunda. H deve ser uma função simétrica das n variáveis $p_1, \cdots, p_n$ de que depende. Quer dizer, efectuada uma permutação qualquer de $p_1, \cdots, p_n$ na expressão analítica de H, não deve esta alterar-se.

Tal exigência decorre tão sòmente de que na consideração de A é imaterial a ordem por que se enumerem as respectivas modalidades $A_1, \cdots, A_n$.

Terceira. H deve atingir um valor máximo quando as variáveis tomam todas o mesmo valor

$$p_1 = p_2 = \cdots = p_n$$

o qual será $\frac{1}{n}$, dado que é $p_1 + \cdots + p_n = 1$.

Por outras palavras,

$$H\left(\frac{1}{n}, \cdots, \frac{1}{n}\right)$$

deverá ser o máximo valor de H.

Já observámos, com efeito, que a incerteza é máxima quando todos os acontecimentos $A_1, \cdots, A_n$ são igualmente prováveis.

Quarta. Consideremos dois sistemas completos de acontecimentos

$$A = \begin{pmatrix} A_1 \cdots A_m \\ p_1 \cdots p_m \end{pmatrix}$$

e

$$B = \begin{pmatrix} B_1 \cdots B_n \\ q_1 \cdots q_n \end{pmatrix}$$

e suponhamos que eles são *independentes*, isto é, que a realização de cada A_k não predispõe absolutamente para a realização de nenhum B_r, e reciprocamente.

A partir de A e B construamos um novo sistema completo de $m \times n$ acontecimentos

$$C = (A \text{ e } B) = \begin{pmatrix} A_1 \text{ e } B_1 & A_1 \text{ e } B_2 \cdots A_m \text{ e } B_n \\ r_{11} & r_{12} \qquad\qquad r_{mn} \end{pmatrix}$$

A independência de A e B traduz-se em

$$r_{ij} = p_i q_j$$

pelo que, na verdade,

$$\sum_{ij} r_{ij} = \sum_{ij} p_i q_j = \sum_i p_i \sum_j q_j = 1$$

A experiência C consiste na realização das experiências A e B e cada modalidade de C não é mais do que um resultado de A com outro de B:

$$C_{ij} = (A_i \text{ e } B_j)$$

Posto isto, impomos à função H a condição

(1) $$H(A \text{ e } B) = H(A) + H(B)$$

a qual significa que a informação contida em C é a *soma* das informações contidas em A e B.

Se A e B não fossem independentes, se estivéssemos, por exemplo, perante a situação extrema

$$m = n, A_i \text{ implica } B_i (i = 1, \cdots, n)$$

então é evidente que a realização de B, quando precedida da realização de A, não comportaria informação alguma: seria em tal caso $H(A \text{ e } B) = H(A)$.

É absolutamente natural a imposição à função H das quatro propriedades acima descritas. Passamos a verificar que elas são satisfeitas pela função

(2) $$H(p_1, \cdots, p_n) = -\sum_{k=1}^{n} p_k \log p_k .$$

Reconhece-se logo que ela se anula sempre que algum p_k iguale a unidade (ficando por conseguinte nulos os restantes p_k; toma-se $0 \log 0 = 0$).

A simetria da função (2) também é evidente.

Verifiquemos a quarta propriedade. Temos sucessivamente

$$\begin{aligned} H(A \text{ e } B) &= -\sum_{ij} r_{ij} \log r_{ij} \\ &= -\sum_{ij} p_i q_j \log (p_i q_j) \\ &= -\sum_{ij} p_i q_j (\log p_i + \log q_j) \\ &= -\Big(\sum_j q_j\Big) \sum_i p_i \log p_i - \Big(\sum_i p_i\Big) \sum_j q_j \log q_j \\ &= H(A) + H(B) \end{aligned}$$

Veremos até, dentro em pouco, que a função (2) obedece a uma lei de que esta quarta propriedade é apenas um caso particular — o da independência de A e B.

Com este fim e também com o de verificarmos a terceira propriedade, teremos necessidade de utilizar uma desigualdade que é satisfeita por todas as funções $f(x)$ que *admitem segunda derivada não-negativa* num certo intervalo:

(3) $$f''(x) \geqq 0$$

A desigualdade em questão é a seguinte

(4_n) $$\begin{aligned} &f(\alpha_1 x_1 + \cdots + \alpha_n x_n) \leqq \\ &\leqq \alpha_1 f(x_1) + \cdots + \alpha_n f(x_n) \end{aligned}$$

válida para todos os números $x_1, \cdots, x_n$ daquele intervalo e para quaisquer $\alpha_1, \cdots, \alpha_n$ não-negativos submetidos à condição

(5) $$\alpha_1 + \cdots + \alpha_n = 1$$

Em particular, tomando

$$\alpha_1 = \cdots = \alpha_n = \frac{1}{n}$$

obtém-se de (4_n)

$$(6_n)\quad f\left(\frac{x_1+\cdots+x_n}{n}\right)\leq\frac{f(x_1)+\cdots+f(x_n)}{n}$$

isto é, o valor da função calculado para a média aritmética das abcissas não excede a média aritmética dos valores que a função toma nas mesmas abcissas.

Não é difícil provar, em sentido inverso, que toda a função *contínua* que verifique (6_n) também verifica (4_n). E, ainda, que *qualquer* função que satisfaça a desigualdade (6_n) para $n=2$ igualmente a satisfaz para todos os outros valores de n (noutros termos, (6_2) implica (6_n)). O leitor pode consultar a este respeito a obra clássica de HARDY, LITTLEWOOD and POLYA, *Inequalities* (Cambridge).

Função que verifique (4_n) chama-se *convexa*. Provemos pois que *é convexa toda a função com segunda derivada não-negativa*, isto é, que (3) implica (4_n).

Fazendo

$$(7)\qquad x_o=\sum_{i=1}^{n}\alpha_i x_i$$

temos, usando a fórmula de TAYLOR com resto de segunda ordem de LAGRANGE,

$$f(x_k)=f(x_o)+(x_k-x_o)\,f'(x_o)+\\+\frac{1}{2}(x_k-x_o)^2\,f''(\xi)$$

com ξ entre x_o e x_k.

A terceira parcela do segundo membro é não-negativa em virtude da condição (3). Desprezando-a fica

$$f(x_k)\geq f(x_o)+(x_k-x_o)f'(x_o)$$

Multipliquemos ambos os membros por α_k, façamos k variar de 1 a n e somemos membro a membro as n desigualdades assim obtidas. Resulta

$$\sum_k\alpha_k f(x_k)\geq f(x_o)\sum_k\alpha_k+f'(x_o)\sum_k\alpha_k x_k-\\-x_o f'(x_o)\sum_k\alpha_k$$

donde, usando (5) e (7),

$$\sum_k\alpha_k f(x_k)\geq f\left(\sum_k\alpha_k\,x_k\right)$$

que é precisamente (4_n), como desejávamos.

Após esta digressão de análise elementar encontramo-nos já em condições de verificar que a função (2) tem a propriedade de máximo.

Em primeiro lugar,

$$(8)\qquad f(x)=x\log x$$

satisfaz (3) no intervalo $(0,\infty)$, pois que

$$f''(x)=\frac{1}{x}$$

Aplicando então (6_n) a $f(x)$ dada por (8), tomando

$$x_k=p_k\qquad(k=1,\cdots,n)$$

fica

$$\frac{\sum p_k}{n}\log\frac{\sum p_k}{n}\leq\frac{\sum p_k\log p_k}{n}$$

ou

$$(9)\qquad \log\frac{1}{n}\leq\sum_k p_k\log p_k$$

Como

$$H\left(\frac{1}{n},\cdots,\frac{1}{n}\right)=-\sum_{k=1}^{n}\frac{1}{n}\log\frac{1}{n}=-\log\frac{1}{n}$$

a desigualdade (9) vem a ser, como desejávamos, o mesmo que

$$H\left(\frac{1}{n},\cdots,\frac{1}{n}\right)\geq H(p_1,\cdots,p_n)$$

Voltemo-nos para a generalização, a que já aludimos, da quarta propriedade. Rejeitamos a independência de A e B, pelo que agora não é já $r_{ij}=p_i q_j$ mas sim, mais geralmente,

$$(10)\qquad r_{ij}=p_i q_{ij}$$

onde q_{ij} significa *a probabilidade de realização de* B_j *quando se tenha realizado* A_i.

Calculando $H(C)$ temos

$$H(C) = -\sum r_{ij} \log r_{ij} =$$
$$-\sum p_i q_{ij} (\log p_i + \log q_{ij})$$

ou

$$(11) \quad H(C) = -\sum_i (p_i \log p_i) \sum_j q_{ij} + \\ + \sum_i p_i \left(-\sum_j q_{ij} \log q_{ij} \right)$$

Mas é

$$(12) \quad \sum_j q_{ij} = 1$$

porquanto a realização de A_i acompanha-se com certeza da realização de algum dos acontecimentos $B_1, \cdots, B_n$, o que dá

$$q_{i1} + \cdots + q_{in} = \text{prob. de } (A_i \text{ e } B_1) + \cdots \\ \cdots + \text{prob. de } (A_i \text{ e } B_n) = 1$$

Na segunda parcela do segundo membro de (11) figura a função que designamos por $H_i(B)$,

$$H_i(B) = -\sum_j q_{ij} \log q_{ij}$$

e na verdade ela não é mais do que a função H calculada para o sistema B *na hipótese de ter sido realizada a experiência* A *com o resultado* A_i.

Considerando estes números $H_i(B)$ $(i = 1, \cdots, m)$ como valores de uma nova função, manifestamente atingidos com as probabilidades respectivas p_i, o valor médio de tal função será dado por

$$(13) \quad H_A(B) = \sum_i p_i H_i(B)$$

que é, como qualquer valor médio, a soma dos produtos dos valores da função pelas correspondentes probabilidades.

Atendendo a (12) e (13) a igualdade (11) passa a escrever-se

$$(14) \quad H(A \text{ e } B) = H(A) + H_A(B)$$

Esta a relação que procurávamos. É de facto uma generalização de (1) pois (1) é o seu caso particular em que, por ser

$$q_{ij} = q_j$$

(independência de A e B) se obtém

$$H_A(B) = H(B)$$

como o leitor fàcilmente verifica.

É óbvio que, havendo interdependência de A e B, o realizar-se A fornece alguma informação a respeito de B. Quer dizer, é de esperar que tenha lugar a desigualdade

$$(15) \quad H_A(B) \leq H(B)$$

e na verdade assim é. Para o demonstrar, retomemos a função (8),

$$f(x) = x \log x$$

e apliquemos-lhe a desigualdade (4_n) com

$$\alpha_i = p_i \qquad x_i = r_{ij}$$

Obtém-se

$$\left(\sum_i p_i r_{ij} \right) \log \left(\sum_i p_i r_{ij} \right) \leq \sum_i p_i r_{ij} \log r_{ij}$$

Mas porque é

$$\sum_i p_i r_{ij} = p_1 r_{1j} + \cdots + p_m r_{mj}$$
$$= \text{prob. de } (A_1 \text{ e } B_j) + \cdots + \text{prob. de } (A_m \text{ e } B_j) \\ = \text{prob. de } B_j = q_j$$

a desigualdade anterior reescreve-se

$$q_j \log q_j \leq \sum_i p_i r_{ij} \log r_{ij}$$

donde, fazendo $j = 1, \cdots, n$ e somando membro a membro,

$$\sum_j q_j \log q_j \leq \sum_i p_i \sum_j r_{ij} \log r_{ij}$$

ou seja

$$H(B) \geq \sum_i p_i H_i(B)$$

que é a desigualdade (15) que assim fica portanto demonstrada.

Em resumo, a função simétrica $H(p_1, \cdots, p_n)$ definida por (2) satisfaz aquele mínimo de requisitos sem os quais não poderia tomar-se como medida da indeterminação ou da informação contida no sistema.

Cabe todavia perguntar se não haverá outras funções que obedeçam às mesmas exigências. A resposta é a seguinte: função simétrica $H(p_1, \cdots, p_n)$ definida e contínua para $0 \leq p_k \leq 1$, com valor máximo para $p_1 = \cdots = p_n = \frac{1}{n}$, que satisfaz a desigualdade (14) — generalização de (1) — e que satisfaz ainda a condição de tomar o mesmo valor para os sistemas

$$\begin{pmatrix} A_1 \cdots A_n \\ p_1 \cdots p_n \end{pmatrix}$$

$$\begin{pmatrix} A_1 \cdots A_n\, I \\ p_1 \quad\; p_n\, 0 \end{pmatrix}$$

onde I designa o acontecimento impossível (de probabilidade nula) — tal função, repetimos, *coincide com a função definida por* (2) *à parte um factor de proporcionalidade*. Para a demonstração remetemos o leitor às memórias que começámos por citar.

À função

$$H(A) = -\sum_k p_k \log p_k$$

foi dado o nome de ENTROPIA, o mesmo, como se sabe, de certa grandeza termodinâmica. Não é difícil surpreender o liame entre os dois conceitos, o probabilístico e o termodinâmico.

O conceito termodinâmico de entropia foi introduzido por CLAUSIUS em 1865 como a grandeza S que satisfaz a relação diferencial

$$dS = \frac{1}{T} dQ \tag{16}$$

em que dQ é a diferencial da quantidade de calor $Q(T)$ do sistema físico considerado, suposto este à temperatura absoluta T. Fixado um estado inicial E_0 do sistema e chamando E ao estado final, mostrou CLAUSIUS que o integral

$$\int dS$$

depende não só de E como também da evolução do sistema; mas que tem um mesmo valor, que é *máximo*, para todas as evoluções *reversíveis*. A esse valor máximo chamou entropia do sistema no estado E.

Depois, BOLTZMAN, em 1896, aplicando a teoria de CLAUSIUS a um sistema gazoso em evolução reversível, provou que a integração da relação (16) dá

$$S = k \log W \tag{17}$$

onde k é uma constante positiva dependente de T, e W é a *probabilidade* do estado considerado do gaz, calculada quanto às posições e velocidades possíveis das moléculas de que se compõe.

Permanecendo o sistema isolado, já CLAUSIUS inferira que a entropia do sistema aumenta constantemente. Nos termos de BOLTZMAN, reinterpreta-se esta lei sob a seguinte forma, como mostra a relação (17): *os sistemas isolados evoluem para os estados mais prováveis*. Ou ainda: a indeterminação do estado de um sistema isolado diminue à medida que ele evolue.

Tal conclusão não contradiz as considerações teóricas que antes expusemos e nas quais o crescimento da função entropia $H(p_1, \cdots, p_n)$ significa um aumento da indeterminação. E não há contradição em virtude do sinal *menos* que afecta a definição probabilística mas não a termodinâmica.

Observemos ainda que a definição (2) da função entropia é a definição de um *valor médio*, no caso sujeito o dos n números

$$-\log p_1, \cdots, -\log p_n$$

Por outro lado, prova-se em Cálculo das Probabilidades que, em certas situações, os valores que toma uma variável aleatória aproximam-se do valor médio dessa variável o suficiente para que, do ponto de vista experimental, os possamos com ele identificar.

Assim se esclarece, nas suas linhas gerais, a conexão entre a relação (17) de BOLTZMAN e a definição (2) da função entropia.

Contudo, foi a moderna teoria das telecomunicações e do controle automático que levou o cientista americano C. E. SHANON a introduzir a definição geral de entropia (*A mathematical theory of communication*, Bell System Techn. J., 27, 1948), criando-se assim um novo ramo do Cálculo das Probabilidades que se encontra em pleno desenvolvimento: a teoria da informação.

Princípios fundamentais dos computadores digitais automáticos

por **A. César de Freitas**

O que vai seguir-se não é mais do que dissemos numa série de palestras feitas na Faculdade de Ciências de Lisboa em Dezembro de 1958 e destinadas a divulgar as ideias que estão na base dos mais poderosos meios de cálculo numérico actualmente existentes.

Nessas palestras procurou-se sempre apresentar os assuntos na sua forma mais elementar e sem entrar em aspectos técnicos, para que os princípios fundamentais fossem percebidos pelo maior número possível de ouvintes, a maioria deles inteiramente leigos na matéria. Esses princípios fundamentais, que como se verá são bastante simples, podem ser apreendidos sem grande dificuldade por qualquer pessoa com instrução equivalente à dos nossos cursos secundários.

Evidentemente que, apenas pela leitura destas palestras, ninguém fica apto a construir uma máquina calculadora electrónica por mais rudimentar que ela seja; o que se pretende é que elas permitam fazer uma ideia, mais ou menos precisa, da maneira como trabalha uma das maiores maravilhas inventadas pelo homem, cujas aplicações são cada vez maiores em quase todos os ramos da actividade humana. Por outro lado, para aqueles que tenham um interesse especial pelo assunto, julgamos ter apresentado os elementos suficientes à compreensão de obras mais especializadas. Principalmente para estes inserimos no final alguma bibliografia por ordem crescente de especialização.

1. Generalidades. As máquinas matemáticas podem classificar-se em *máquinas digitais* e *máquinas analógicas*. As primeiras trabalham por *contagem* de acontecimentos discretos, as segundas *medem* grandezas contínuas. Um exemplo de máquina digital é a máquina calculadora vulgar (manual ou eléctrica); a régua de cálculo é uma máquina analógica em que os números são representados por comprimentos.

Mais precisamente o termo *digital*, quando aplicado a uma máquina, implica mecanização das matemáticas dos problemas, uma vez esses problemas sejam postos na forma aritmética; o termo *analógico* implica uma semelhança de propriedades nos aspectos que interessam, e na máquina analógica a analogia baseia-se na identidade (ou seme-