Distribuição de probabilidade conjunta

- Uma *distribuição de probabilidade conjunta* é uma distribuição de probabilidade num produto Cartesiano (de dois ou mais conjuntos).
- Por exemplo, dados conjuntos $S = \{s_1, \ldots, s_M\}$ e $T = \{t_1, \ldots, t_N\}$, uma função $P : S \times T \to [0, 1]$, $(s_i, t_j) \mapsto p_{ij}$ com $\sum_{i=1}^{M} \sum_{j=1}^{N} p_{ij} = 1$.
  A sua entropia é dada por

$$H(P) = -\sum_{i=1}^{M} \sum_{j=1}^{N} p_{ij} \log(p_{ij}).$$

- Se tivermos distribuições de probabilidades P_S e P_T, em S e T, respetivamente, podemos definir uma distribuição de probabilidade conjunta em $S \times T$ por

$$P(s_i, t_j) = P_S(s_i) P_T(t_j),$$

a que chamaremos a *distribuição combinada*.

- Nem todas as distribuições de probabilidade conjuntas são desta forma.
Por exemplo, $P(s_i, t_j) = 1/3$ para $i, j \in \{1, 2\}$ com $(i, j) \neq (2, 2)$.

Axiomatização da entropia

- Vamos ver que a entropia de uma distribuição de probabilidade, como foi definida, é a única função que goza de certas propriedades naturais.
- Para o efeito, vamos começar por pensar na entropia como uma medida da incerteza do resultado da escolha de um evento do espaço de amostragem.
- Representemos por $H(p_1, \ldots, p_n)$ a entropia de uma distribuição de probabilidade de uma variável aleatória dada por $P(X = s_i) = p_i$ ($i = 1, \ldots, n$).

- No caso de uma distribuição uniforme, tomemos $f(M) = H(\frac{1}{M}, \ldots, \frac{1}{M})$ e notemos que nos parece intuitivamente natural que a incerteza seja tanto maior quanto maior for M, ou seja que a função f seja estritamente crescente.

- Por outro lado, se considerarmos duas distribuições de probabilidade uniformes, sobre espaços de amostragem com M e N pontos, a distribuição combinada, sobre um espaço com MN pontos, é também uma distribuição uniforme.
- Em média, esperamos que, conhecido o valor da variável aleatória correspondente à primeira distribuição de probabilidade, ou seja, se for removida a incerteza $f(M)$, a incerteza que resta na distribuição combinada seja o valor da incerteza $f(N)$ da segunda distribuição de probabilidade.
- Por outras palavras, é razoável pretender que $f(MN) = f(M) + f(N)$.

- Consideremos agora uma distribuição de probabilidade finita arbitrária $(p_1, \ldots, p_N)$, correspondendo ao espaço de amostragem $S = \{s_1, \ldots, s_N\}$.
- Os eventos $A = \{s_1, \ldots, s_r\}$ e $B = \{s_{r+1}, \ldots, s_N\}$ têm então as probabilidades respetivas $p' = p_1 + \cdots + p_r$ e $p'' = p_{r+1} + \cdots + p_N$.
    - Como se relacionam as incertezas $H(p_1, \ldots, p_N)$ e $H(p', p'')$?

- Escolhido aleatoriamente um elemento de S, ele poderá pertencer a A ou a B, respetivamente com as probabilidades p' e p'', sendo a incerteza média correspondente $H(p', p'')$.
  - No primeiro caso, a probabilidade de se tratar de um dado $s_i \in A$, será então a probabilidade condicionada p_i/p' e a incerteza será em média $H(p_1/p', \ldots, p_r/p')$.
  - No segundo caso, a probabilidade de se tratar de um dado $s_i \in B$, será então a probabilidade condicionada p_i/p'' e a incerteza será em média $H(p_{r+1}/p'', \ldots, p_N/p'')$.
- É portanto de esperar que $H(p_1, \ldots, p_N) - H(p', p'')$ seja a média pesada daqueles dois valores, nomeadamente $p'H(p_1/p', \ldots, p_r/p') + p''H(p_{r+1}/p'', \ldots, p_N/p'')$.

- Como último requisito, assumimos que a função de p dada por $H(p, 1-p)$ seja contínua.
- Chegamos assim aos seguintes *axiomas* para a incerteza $H(p_1, \ldots, p_N)$:
  1. A função $f(M) = H(\frac{1}{M}, \ldots, \frac{1}{M})$ é estritamente crescente.
  2. $f(MN) = f(M) + f(N)$.
  3. Sendo $p' = p_1 + \cdots + p_r$ e $p'' = p_{r+1} + \cdots + p_N$, tem-se

$$\begin{aligned} H(p_1, \ldots, p_N) = H(p', p'') &+ p' H(p_1/p', \ldots, p_r/p') \\ &+ p'' H(p_{r+1}/p'', \ldots, p_N/p'') \end{aligned}$$

     para qualquer $r \in \{1, \ldots, N-1\}$.
  4. A função $H(p, 1-p)$ é contínua em p.

Teorema 1.3 *As únicas funções* $H(p_1, \ldots, p_N)$ *que satisfazem os axiomas acima são as funções da forma*

$$H(p_1, \ldots, p_N) = -C \sum_{i=1}^{N} p_i \log p_i,$$

onde $C > 0$ *é uma constante e a base do logaritmo é arbitrária (mas fixada).*

Prova. É um exercício verificar que as funções do enunciado satisfazem os axiomas.
Para mostrar que não há outros modelos para os nossos axiomas, podemos sem perda de generalidade considerar logaritmos de base e, pois isso corresponde a uma simples alteração da constante C.

(a) Comecemos por notar que o axioma 2 garante que $f(N^k) = kf(N)$.

(b) Vejamos como deduzir que se tem necessariamente $f(N) = C \ln N$ ($N = 1, 2, \ldots$), onde C é uma constante positiva.
Como $f(1) = f(1 \cdot 1) = f(1) + f(1)$ pelo axioma 2, temos certamente $f(1) = 0$, pelo que a nossa fórmula é verificada para $N = 1$.
Supondo $N > 1$ e dado um inteiro positivo r, seja $k = \lfloor \frac{r \ln 2}{\ln N} \rfloor$, ou seja k é o único inteiro positivo tal que $N^k \leqslant 2^r < N^{k+1}$.
Pelo axioma 1 resulta que $f(N^k) \leqslant f(2^r) < f(N^{k+1})$, o que por (a) equivale a $kf(N) \leqslant rf(2) < (k+1)f(N)$, ou seja ainda $k \leqslant \frac{rf(2)}{f(N)} < k + 1$.
Logo tem-se $\left| \frac{\ln 2}{\ln N} - \frac{f(2)}{f(N)} \right| < \frac{1}{r}$ o que, valendo para todo o inteiro positivo r, garante que o lado esquerdo da desigualdade é nulo.
Concluímos portanto que $f(N) = C \ln N$, onde $C = f(2)/\ln 2$ é uma constante positiva pelo axioma 1.

(c) Podemos agora calcular $H(p, 1-p)$ quando p é um número racional positivo através da fórmula

$$H(p, 1-p) = -C\left(p \ln p + (1-p)\ln(1-p)\right). \tag{2}$$

De facto, se $p = r/s$, onde r e s são inteiros positivos, então o axioma 3 garante-nos que

$$\begin{aligned} f(s) &= H(\underbrace{\frac{1}{s}, \ldots, \frac{1}{s}}_{r}, \underbrace{\frac{1}{s}, \ldots, \frac{1}{s}}_{s-r}) \\ &= H(\frac{r}{s}, \frac{s-r}{s}) + \frac{r}{s} f(r) + \frac{s-r}{s} f(s-r). \end{aligned}$$

Em, face de (b), obtemos

$$C \ln s = H(p, 1-p) + Cp \ln r + C(1-p)\ln(s-r),$$

donde resulta que

$$\begin{aligned} H(p, 1-p) &= -C\left(p \ln r - \ln s + (1-p)\ln(s-r)\right) \\ &= -C\left(p \ln r - p \ln s + p \ln s - \ln s + (1-p)\ln(s-r)\right) \\ &= -C\left(p \ln p + (1-p)\ln(1-p)\right). \end{aligned}$$

(d) Em face do axioma 4, deduzimos que a fórmula (2) é válida para todo o $p \in [0,1]$ pois os racionais positivos naquele intervalo formam um conjunto denso e as funções de ambos os lados da nossa fórmula são contínuas.

(e) Finalmente, provemos que a função $H(p_1, \ldots, p_N)$ é dada pela fórmula geral do enunciado do teorema, para o que procedemos por indução sobre N, tendo os casos $N = 1, 2$ já sido estabelecidos, pelo que supomos que $N > 2$.
Pelo axioma 3, tomando $q = p_1 + \cdots + p_{N-1}$, obtemos a fórmula

$$H(p_1, \ldots, p_N) = H(q, p_N) + qH(\frac{p_1}{q}, \ldots, \frac{p_{N-1}}{q}) + p_N H(1).$$

Assumindo, como hipótese de indução, que a fórmula em vista vale para uma distribuição de probabilidade envolvendo $N - 1$ termos, deduzimos que

$$\begin{aligned} H(p_1, \ldots, p_N) &= -C(q \ln q + p_N \ln p_N) - Cq \sum_{i=1}^{N-1} \frac{p_i}{q} \ln \frac{p_i}{q} + p_N \cdot 0 \\ &= -C(q \ln q + p_N \ln p_N) - C \left(\sum_{i=1}^{N-1} p_i \ln p_i - q \ln q \right) \\ &= -C \sum_{i=1}^{N} p_i \ln p_i . \square \end{aligned}$$

Distribuições marginais

- Dada uma distribuição de probabilidade conjunta P sobre $S \times T$, as *distribuições marginais* sobre S e T são dadas por

$$P_S(s_i) = \sum_{j=1}^{N} P(s_i, t_j) \quad \text{e} \quad P_T(t_j) = \sum_{i=1}^{M} P(s_i, t_j)$$

- As distribuições marginais dizem-se *independentes* se a distribuição P coincidir com a distribuição combinada.

Teorema 1.4 *Se P é uma distribuição de probabilidade conjunta sobre $S \times T$ e P_S e P_T são as respetivas distribuições marginais em S e T, então*

$$H(P) \leqslant H(P_S) + H(P_T) \tag{3}$$

verificando-se a igualdade sse as distribuições marginais são independentes.

Prova. Temos

$$H(P_S) = -\sum_{i=1}^{M} P_S(s_i)\log P_S(s_i) = -\sum_{i=1}^{M}\sum_{j=1}^{N} P(s_i, t_j)\log P_S(s_i)$$

$$H(P_T) = -\sum_{j=1}^{N} P_T(t_j)\log P_T(t_j) = -\sum_{i=1}^{M}\sum_{j=1}^{N} P(s_i, t_j)\log P_T(t_j)$$

$$H(P_S) + H(P_T) = -\sum_{i=1}^{M}\sum_{j=1}^{N} P(s_i, t_j)\log(P_S(s_i)P_T(t_j))$$

$$H(P) = -\sum_{i=1}^{M}\sum_{j=1}^{N} P(s_i, t_j)\log P(s_i, t_j)$$

Logo $H(P) \leqslant H(P_S) + H(P_T)$ pelo Lema 1.1 pois

$$\sum_{i=1}^{M}\sum_{j=1}^{N} P(s_i, t_j) = 1 \quad \text{e} \quad \sum_{i=1}^{M}\sum_{j=1}^{N} P_S(s_i)P_T(t_j) = 1$$

com igualdade sse $P(s_i, t_j) = P_S(s_i)P_T(t_j)$ para todos os i, j. $\square$

Probabilidade condicional

- Seja S um espaço de amostragem com uma distribuição de probabilidade P e sejam E e F eventos de S.
- A *probabilidade condicional de E dado F* é

$$P(E|F) = \frac{P(E \cap F)}{P(F)}.$$

- Segue que $P(E|F)P(F) = P(E \cap F) = P(F|E)P(E)$ e o Teorema de Bayes:

Teorema 1.5 *Nas condições acima, tem-se*

$$P(E|F) = \frac{P(F|E)P(E)}{P(F)}.\square$$

Distribuições de probabilidade condicionadas

- Seja P uma distribuição de probabilidade conjunta sobre $S \times T$, com S e T como acima, e sejam P_S e P_T as respetivas distribuições marginais.
- A *probabilidade condicional de s_i dado t_j* é

$$P(s_i|t_j) = \frac{P(s_i, t_j)}{P_T(t_j)} = \frac{P(s_i, t_j)}{\sum_{k=1}^{M} P(s_k, t_j)}.$$

- Note-se que, fixado t_j, a função $s_i \mapsto P(s_i|t_j)$ define uma distribuição de probabilidade sobre S,
  a qual se chama a *distribuição de probabilidade condicional dado t_j* e se representa por $P_{S|t_j}$.

- A correspondente entropia é dada pela fórmula

$$H(P_{S|t_j}) = -\sum_{i=1}^{M} P(s_i|t_j)\log P(s_i|t_j).$$

- A média pesada das distribuições $P_{S|t_j}$ é a distribuição marginal P_S:

$$\sum_{j=1}^{N} P_T(t_j)P_{S|t_j}(s_i) = \sum_{j=1}^{N} P(s_i, t_j) = P_S(s_i).$$

- Define-se a *entropia condicionada de P dado T*, que se representa por $H(P_{S|T})$, como sendo o valor esperado das entropias $H(P_{S|t_j})$, ou seja

$$\begin{aligned} H(P_{S|T}) &= \sum_{j=1}^{N} P_T(t_j) H(P_{S|t_j}) \\ &= -\sum_{j=1}^{N} P_T(t_j) \sum_{i=1}^{M} P(s_i|t_j) \log P(s_i|t_j) \\ &= -\sum_{i=1}^{M} \sum_{j=1}^{N} P(s_i, t_j) \log P(s_i|t_j). \end{aligned}$$

Teorema 1.6
$H(P) = H(P_T) + H(P_{S|T}) = H(P_S) + H(P_{T|S})$.

Prova.

$$
\begin{aligned}
H(P) &= -\sum_{i=1}^{M}\sum_{j=1}^{N} P(s_i,t_j)\log P(s_i,t_j) \\
&= -\sum_{i=1}^{M}\sum_{j=1}^{N} P(s_i,t_j)\log(P_T(t_j)P(s_i|t_j)) \\
&= -\sum_{i=1}^{M}\sum_{j=1}^{N} P(s_i,t_j)\log P_T(t_j) - \sum_{i=1}^{M}\sum_{j=1}^{N} P(s_i,t_j)\log P(s_i|t_j) \\
&= -\sum_{j=1}^{N} P_T(t_j)\log P_T(t_j) - \sum_{i=1}^{M}\sum_{j=1}^{N} P(s_i,t_j)\log P(s_i|t_j) \\
&= H(P_T) + H(P_{S|T})
\end{aligned}
$$

sendo a outra igualdade análoga. $\square$

Teorema 1.7 *Tem-se* $H(P_{S|T}) \leqslant H(P_S)$*, com igualdade sse as distribuições marginais* P_S *e* P_T *forem independentes.*

Prova. Pelo Teorema 1.6 tem-se $H(P) = H(P_T) + H(P_{S|T})$.
Pelo Teorema 1.4 tem-se $H(P) \leqslant H(P_S) + H(P_T)$, com igualdade sse P_S e P_T forem independentes.
O resultado segue imediatamente. $\square$

- Em termos informais, ao condicionar T por S reduzimos a entropia, i.e., reduzimos a incerteza.

Fontes de informação

- Como teoria probabilística, a teoria da informação lida em grande parte com sequências aleatórias.
- Noutros contextos, elas podem ser chamadas de séries temporais, processos estocásticos (discretos), sinais.
- Um gerador de tais sequências diz-se uma *fonte de informação*.
- Habitualmente, trata-se de sequências de elementos de um conjunto finito, a que se chama o *alfabeto*.
- Os elementos do alfabeto dizem-se *símbolos* (ou *letras*).

Exemplos

- Deitar uma moeda ao ar e registar o resultado: frente (H) ou verso (T), gera uma sequência aleatória de letras do alfabeto $\{H, T\}$.
- Deitar um dado e registar o resultado que aparece na face do topo, digamos no alfabeto $\{1, 2, 3, 4, 5, 6\}$.
- Os computadores geram sequências de bits, por exemplo que transmitem através de uma rede "ethernet", as quais podem ser vistas como sequências aleatórias no alfabeto $\{0, 1\}$.
- Um texto em português é uma sequência aleatória de letras do alfabeto, de dígitos, de sinais de pontuação, espaços, e outros sinais especiais como €, £, $, &, etc

  Neste exemplo, encontra-se uma correlação não trivial entre símbolos contíguos, esperando-se, por exemplo, encontrar um ã ou um õ a seguir a um ç.

Informação?

- Em que medida é que se trata efetivamente de informação e como se mede a informação?
- Quanto mais imprevisível for um dado símbolo, em certo sentido mais informação ele comporta.
- Um símbolo totalmente previsível é dispensável e não acrescenta nada à informação (a não ser porventura facilidade de leitura).
- Assim, se tivermos uma distribuição de probabilidade que descreve a probabilidade de obter um dado símbolo (no caso de existirem correlações entre símbolos consecutivos, devemos considerar distribuições de probabilidade condicionada),
  a entropia, como medida de incerteza, é portanto também uma medida da quantidade de informação.