Óscar Felgueiras
Universidade
do Porto
olfelgue@fc.up.pt

PROBABILIDADES ENGANADORAS

Laplace definiu probabilidade como sendo o quociente do número de casos favoráveis pelo número de casos totais. Embora este conceito de probabilidade seja intuitivo e natural, por vezes pode transmitir uma ideia enganadora e veremos dois casos em que isso acontece.

As probabilidades são hoje em dia amplamente utilizadas e relativamente familiares à comunidade em geral. Convém por isso que haja um certo cuidado na interpretação que se pode fazer de determinada probabilidade, não só quanto ao que ela quer dizer, mas também quanto àquilo que não quer dizer.

1. PARADOXO DE SIMPSON

Imaginemos que o leitor tem de decidir entre recorrer ao Hospital A ou ao Hospital B, e que tem ao seu dispor a informação da seguinte tabela:

	Pacientes	Mortes	Tx. Mortalidade
Hospital A	1000	200	200/1000=20%
Hospital B	1000	100	100/1000=10%

Tendo por base o conhecimento de que a taxa de mortalidade no Hospital B (10%) é inferior à do Hospital A (20%), é natural que esteja mais inclinado a optar pelo Hospital B dado que a probabilidade de um paciente sobreviver no Hospital B é maior do que no Hospital A. No entanto, suponha que estão também disponíveis os registos da condição dos pacientes à entrada em cada um dos hospitais e respetivas estatísticas. Considere a seguinte tabela relativamente aos pacientes classificados em *estado estável* ou em *estado crítico*:

		Pacientes	Mortes	Tx. Mortalidade
Estado Estável	Hosp. A	600	12	12/600=2%
	Hosp. B	900	36	36/900=4%
Estado Crítico	Hosp. A	400	188	188/400=47%
	Hosp. B	100	64	64/100=64%

Note que a tabela inicial consiste simplesmente na soma dos dados apresentados nesta segunda. Seria esta informação discriminada o suficiente para mudar de ideias quanto ao hospital a escolher?

A novidade que aqui surge é o facto de o Hospital A ter uma taxa de mortalidade mais baixa tanto nos pacientes estáveis como nos críticos. Por outro lado, a taxa de mortalidade total mais baixa continua a ser a do Hospital B. O problema é que a taxa de mortalidade total é enganadora por ter na sua formação grupos de pacientes com características heterogéneas e em diferentes proporções. Os pacientes em estado crítico recorrem mais ao Hospital A (400 contra 100 no Hospital B) e os que estão estáveis recorrem mais ao Hospital B (900 contra 600 no Hospital A). Isto leva a suspeitar de que o estado dos pacientes está a pesar na distribuição pelos dois hospitais. Assim, o estado dos pacientes é um fator de confundimento que deve ser tido em conta numa análise criteriosa da taxa de mortalidade. A taxa de mortalidade total mais alta no Hospital A acaba por

ser um reflexo da maior proporção de pacientes críticos para pacientes estáveis nesse hospital e não uma consequência do insucesso no tratamento dos pacientes.

O fenómeno exibido neste exemplo foi identificado por Edward Simpson em 1951[1], e batizado em 1972 por Colin Blyth[2] como paradoxo de Simpson. Ele reflete uma propriedade relativamente elementar de números que satisfaçam simultaneamente as seguintes condições:

$$\frac{a}{A} < \frac{b}{B}, \frac{c}{C} < \frac{d}{D}, \frac{a+c}{A+C} > \frac{b+d}{B+D}.$$

Mais exemplos e detalhes podem ser encontrados a partir da página de Internet do projeto ALEA[3]. Para uma visualização interativa recomendamos o projeto *Visualizing Urban Data* da UC Berkeley [4], assim como um *applet* em GeoGebra criado por Alexander Bogomolny[5].

2. SEQUÊNCIAS DE FACES E COROAS

No lançamento ao ar de uma moeda equilibrada supomos que as probabilidades de sair face (F) ou coroa (C) são iguais. Fixem-se duas sequências finitas de faces e coroas. Lança-se sucessivamente a moeda até que saia uma das sequências fixadas. Tomando por exemplo as sequências FF e CF, qual delas terá mais probabilidade de ocorrer primeiro?

Não é difícil ver que a única maneira de FF ocorrer antes de CF é sair FF logo nos primeiros dois lançamentos, acontecimento que tem probabilidade 1/4 de acontecer. Isto porque a partir do momento em que saísse C, automaticamente iria ser CF a sair primeiro. Logo, a probabilidade de CF sair antes de FF é a complementar, $1 - 1/4 = 3/4$. O quociente entre estas duas probabilidades designa-se por *odds*. Neste caso concreto, diz-se que CF *bate* FF com *odds* 3/1 por este valor ser maior do que 1.

Podemos também definir o tempo médio de espera de uma sequência, $E(\cdot)$, como uma função que associa a cada sequência finita o número de lançamentos esperados até que ela saia. Assim, se quisermos determinar um caso elementar, $E(\mathrm{F})$, podemos começar por observar que

$$E(\mathrm{F}) = 1 + \frac{1}{2}E(\mathrm{F}|\mathrm{F}) + \frac{1}{2}E(\mathrm{F}|\mathrm{C})$$

onde $E(\mathrm{F}|\mathrm{F})$ e $E(\mathrm{F}|\mathrm{C})$ denotam o número de lançamentos esperados para que saia F, a partir do momento em que tenha saído F e C, respetivamente. Claro que $E(\mathrm{F}|\mathrm{F}) = 0$ e $E(\mathrm{F}|\mathrm{C}) = E(\mathrm{F})$, pelo que

$$E(\mathrm{F}) = 1 + \frac{1}{2}E(\mathrm{F}),$$

daí resultando que

$$E(\mathrm{F}) = 2.$$

É óbvio também que $E(\mathrm{C}) = 2$. Estamos agora em condições de calcular $E(\mathrm{CF})$ e $E(\mathrm{FF})$. Assim,

$$\begin{aligned} E(\mathrm{CF}) &= 1 + \frac{1}{2}E(\mathrm{CF}|\mathrm{F}) + \frac{1}{2}E(\mathrm{CF}|\mathrm{C}) \\ &= 1 + \frac{1}{2}E(\mathrm{CF}) + \frac{1}{2}E(\mathrm{F}), \end{aligned}$$

e consequentemente $E(\mathrm{CF}) = 4$. Por outro lado,

$$\begin{aligned} E(\mathrm{FF}) &= 1 + \frac{1}{2}E(\mathrm{FF}|\mathrm{F}) + \frac{1}{2}E(\mathrm{FF}|\mathrm{C}) \\ &= 1 + \frac{1}{2}(1 + \frac{1}{2}E(\mathrm{FF}|\mathrm{C})) + \frac{1}{2}E(\mathrm{FF}) \\ &= \frac{3}{2} + \frac{3}{4}E(\mathrm{FF}), \end{aligned}$$

o que implica que $E(\mathrm{FF}) = 6$. Observando que $E(\mathrm{CF}) = 4 < 6 = E(\mathrm{FF})$ e que CF bate FF com *odds* 3/1, poderia parecer possível estabelecer uma associação entre haver um menor número de lançamentos esperados até sair CF e o facto de CF bater FF. Algo que seria um engano, como veremos de seguida.

Não é preciso procurar muito para se encontrar um contraexemplo. Tomando FF e FC, tem-se que $E(\mathrm{FC}) = 4 < 6 = E(\mathrm{FF})$. No entanto, nenhuma destas sequências bate a outra. Basta notar de que cada vez que sai um F existe uma probabilidade igual de se completar uma das sequências FF ou FC.

Analisando sequências de tamanho 3, dá-se um caso bastante inesperado. Verifica-se que FFC bate FCC (*odds* 2/1), FCC bate CCF (*odds* 3/1), CCF bate CFF (*odds* 2/1) e CFF bate FFC (*odds* 3/1). Isto mostra a não transitividade da relação de uma sequência bater outra.

[1] Simpson, E. "The interpretation of interaction in contingency tables". *Journal of the Royal Statistical Society, Series B*. 13, 238–241, 1951.

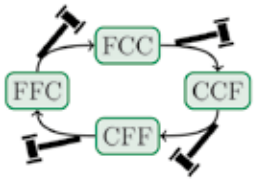
[2] Blyth, C. "On Simpson's paradox and the sure-thing principle". *Journal of the American Statistical Association*. 67, 364–366, 1972.

[3] *http://www.alea.pt/html/statofic/html/dossier/doc/ActivAlea02.pdf*

[4] *http://vudlab.com/simpsons/*

[5] *http://www.cut-the-knot.org/Curriculum/Algebra/SimpsonParadox.shtml*

Saliente-se que todas estas quatro sequências têm um número esperado de lançamentos igual a 8.



Considerando sequências de tamanho 4, verifica-se que CFCF bate FCFF com *odds* 9/5 ainda que $E(\text{CFCF}) = 20 > 18 = E(\text{FCFF})$. Isto significa que apesar de esperarmos que CFCF ocorra em média depois de FCFF, na verdade CFCF tem maior probabilidade de ocorrer antes de FCFF!

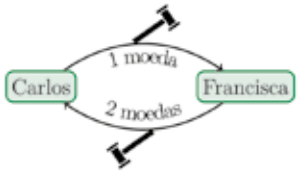
Em particular, suponha-se que o Carlos e a Francisca apostam um contra o outro na sequência que acreditam que vai sair primeiro ao lançarem consecutivamente uma moeda: o Carlos aposta em CFCF e a Francisca em FCFF. Como CFCF bate FCFF com *odds* 9/5, então o Carlos tem probabilidade 9/14 de ganhar.

Por outro lado, imaginemos que a Francisca, talvez por perceber que parte em desvantagem, propõe uma subtil alteração de regras: continuarem ambos a apostar nas mesmas sequências, mas jogando com duas moedas. Ou seja, cada um dos jogadores lança a sua própria moeda e observa o número de lançamentos que tem de fazer até sair a sua sequência escolhida. Se houver empate o jogo reinicia-se.

Nesta nova forma de jogar, a Francisca vai poder beneficiar de estar a apostar numa sequência que tem menor tempo médio de espera e de haver independência entre a saída das duas sequências. As probabilidades de vitória para cada jogador podem ser calculadas usando, por exemplo, cadeias de Markov.

A tabela apresentada de seguida contém o resumo das probabilidades dos diferentes resultados caso os jogadores decidam jogar com base na sequência das respetivas moedas ou com base na sequência de uma só moeda. Chega-se assim à conclusão de que a alteração de regras torna a Francisca a mais provável vencedora do jogo.

	Carlos	Francisca	Empate
1 moeda	$\frac{9}{14}(64.29\%)$	$\frac{5}{14}(35.71\%)$	0 (0%)
2 moedas	$\frac{4187}{9228}(45.37\%)$	$\frac{78293}{152345}(51.39\%)$	$\frac{251362}{7769319}(3.24\%)$



A abordagem sobre sequências de moedas aqui feita tem como fio condutor o artigo de 1974 de Martin Gardner[6], mais tarde incluído em livro[7]. Ela baseia-se num jogo chamado Penney-Ante proposto por Walter Penney[8] em 1969. Para uma breve introdução a estas e outras questões relacionadas com probabilidades, recomendo a TED Talk de Peter Donnely[9]. Para uma exploração mais detalhada do assunto, ver os artigos de Raymond Robertson[10] e de Yutaka Nishiyama[11]. Este último apresenta uma explicação particularmente acessível de como calcular odds de uma sequência bater outra.

---

[6] Gardner, M. "Mathematical games", *Scientific American*, 231(4), 120-125, 1974.

[7] Gardner, M. *Time Travel and Other Mathematical Bewilderments*}, New York: W. H. Freeman, 64-66, 1988.

[8] Penney, W. "Problem 95. Penney-Ante". *J. Recr. Math.* 2, 241, 1969.

[9] *http://www.youtube.com/watch?v=kLmzxmRcUTo*

[10] Nickerson, R. "Penney Ante: Counterintuitive Probabilities in Coin Tossing". *The UMAP Journal.* 28(4).

[11] Y. Nishiyama. "Pattern matching probabilities and paradoxes as a new variation on Penney's coin game". *Intl Jrnl of Pure and Applied Math.* 59(3) 357-366, 2010 disponível em *http://www.ijpam.eu/contents/2010-59-3/10/10.pdf*