Quando dois e dois não são quatro

Dois e dois são sempre quatro. Mas o *quatro* tanto pode ser obtido pela soma «dois mais dois» como pode resultar da soma «um mais três». Parece impossível distingui-lo. O problema, contudo, tem uma tremenda importância prática em estatística.

Tudo terá começado em 1919, quando dois cientistas políticos norte-americanos, William Ogburn e Inez Goltra, publicaram um estudo sobre o comportamento do voto das mulheres recém-recenseadas no estado de Oregon. Os dois investigadores possuíam apenas os totais dos votos e não lhes era possível separá-los por sexos. «Apesar de o processo de votação não permitir contar os votos das mulheres», escreveram, «questionamo-nos sobre a possibilidade de resolver indirectamente este problema.» O método que elegeram foi a correlação dos resultados distritais com a percentagem de mulheres votantes em cada distrito. Assim, em distritos com maior peso de mulheres, atribuíram o afastamento da média aos votantes femininos. Como os próprios investigadores reconheceram, o seu método era falível, pois

podiam perfeitamente ser os homens, nos distritos de maior peso feminino, a orientar diferentemente o seu voto.

O problema de reconstruir comportamentos individuais a partir de dados agregados veio a ser conhecido como o problema da *inferência ecológica* — pois é a ecologia que se ocupa das relações entre os elementos e o seu ambiente —, mas poucos passos fundamentais foram dados para a sua solução.

Três décadas mais tarde, o sociólogo norte-americano William Robinson publicou um trabalho que marcou decisivamente a metodologia das ciências sociais. No essencial, Robinson mostrou que os métodos à data existentes não permitiam reconstruir dados parcelares a partir de dados agregados e popularizou a expressão «falácia ecológica» para descrever as ilações ilegítimas que podiam ser efectuadas dessa maneira. O trabalho de Robinson pôs em causa variadas correntes de investigação sociológica. Os estudos de geografia política, florescentes em França, na Alemanha e nos Estados Unidos, praticamente estancaram ao ser posta em causa a validade dos métodos então seguidos.

O problema da inferência ecológica, no entanto, manteve-se como uma questão premente da estatística aplicada. As questões em estudo são demasiado importantes para que os investigadores aceitem a falta de uma resposta. Costuma-se citar, como exemplo premente, a tentativa de compreender o sucesso eleitoral e político dos nazis no princípio dos anos 30, o que obriga a destrinçar os grupos e classes que apoiaram a subida de Hitler ao poder. Para tal estudo, os sociólogos têm vindo a basear-se nos dados dos círculos eleitorais, para os quais apenas existem resultados agregados.

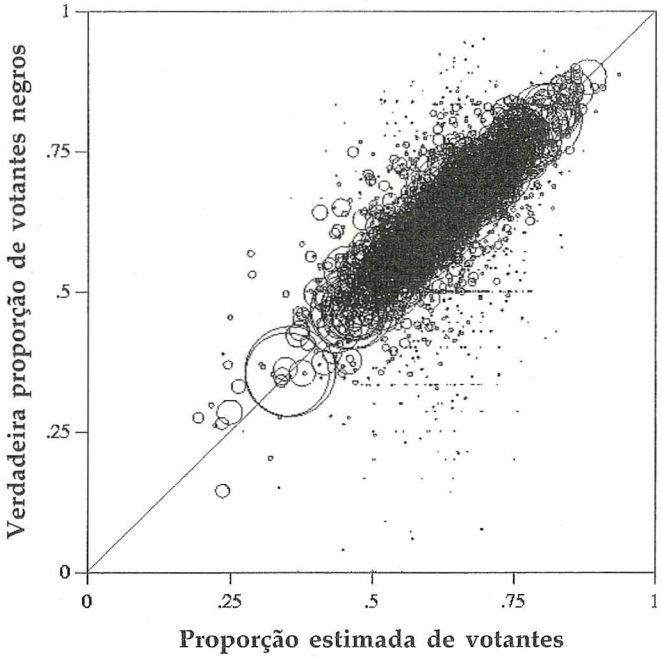
Um outro exemplo premente da importância da inferência ecológica é dado pela epidemiologia. Sabe-se muitas vezes

qual o total de elementos afectados, mas desconhecem-se os bairros em que a população é mais atingida. Os dados estão agregados nos hospitais, mas, nos países menos desenvolvidos, é sempre muito difícil tratá-los de maneira a localizar rapidamente as zonas onde a epidemia mais se desenvolve. Um método eficiente de comparação dos dados agregados com as informações parcelares existentes — por exemplo, em alguns centros de saúde mais bem organizados — poderá detectar a origem da epidemia e ajudar a salvar a vida a muitas pessoas.

Outro exemplo ainda é oferecido pelo *marketing*. Conhece-se muitas vezes o resultado final de uma campanha publicitária e sabe-se também qual a distribuição da população alvo em termos etários e de rendimento. No entanto, é habitualmente muito caro realizar inquéritos que permitam destrinçar as faixas etárias e sociais que melhor responderam à campanha. E esse conhecimento será essencial para um eficiente plano de *marketing*.

Até agora, os métodos existentes têm tido pouco sucesso. Costumam citar-se exemplos caricatos, como o de um estudo de sociólogos israelitas que, ao procurar prever o número de eleitores fiéis ao Partido Trabalhista, estimou um *número negativo de votantes!* Ou o exemplo de uma empresa de sondagens eleitorais norte-americana que concluiu que 120 por cento dos negros do estado do Louisiana tinham votado a favor dos democratas!

Gary King, um estatístico e cientista político que investiga e lecciona em Harvard, encontrou novas soluções para o problema da inferência ecológica. O seu método é muito mais complexo que os usuais procedimentos multivariados, pois é não linear. O algoritmo começa por analisar as unidades mais pequenas que é possível obter.



Comparando as suas estimativas com resultados posteriormente obtidos, Gary King obteve uma notável concordância com a realidade. O gráfico mostra os 3262 círculos eleitorais do Louisiana com bolas proporcionais ao número de votantes por distrito. Quase todos os elementos se encontram ao longo da diagonal, indicando que a fracção estimada e a fracção real de votantes são praticamente idênticas

A partir daí calcula limites lógicos para cada subgrupo. Se, por exemplo, foram mil os votantes num determinado candidato, o número de mulheres que votaram nesse candidato não pode ser menor que zero nem maior que mil. Estes limites, que parecem triviais, introduzem não linearidades no instrumental estatístico. O passo seguinte do algoritmo é a estimação de um valor mais verosímil, que maximize a correlação das estimativas para cada subgrupo

com os dados parcelares e fragmentários existentes sobre alguns dos subgrupos. Finalmente, essas estimativas são comparadas com o que se conhece de alguns subgrupos e corrigidas.

O método é evidentemente bastante complexo e requer todo um livro para ser devidamente explicado[1]. O que importa é que Gary King testou o algoritmo em mais de 16 mil casos e as suas estimativas revelaram um notável ajustamento à realidade. A Associação de Ciência Política Norte-Americana (APSA) atribuiu-lhe o Prémio Gosnell pelo «melhor trabalho metodológico» do ano e a National Science Foundation (NSF) dos Estados Unidos não podia ter sido mais entusiástica. «Espera-se que a solução de Gary King venha a contribuir para uma análise de dados mais precisa», disse Frank Scioli, director da fundação, «e que isso leve a decisões políticas mais bem fundamentadas e a uma melhor compreensão da economia e da sociedade.»

---

[1] O trabalho de Gary King foi publicado no livro *A Solution to the Ecological Inference Problem*, Princeton University Press, 1997. O autor colocou ainda na Internet os programas de computador que permitem a aplicação do seu método. Esses programas, que correm em DOS, Windows ou na linguagem GAUSS, estão disponíveis, gratuitamente, no endereço *http://gking.harvard.edu*.

A MATEMÁTICA DAS COISAS : DO PAPEL A4 AOS ATACADORES DE SAPATOS, DO GPS ÀS RODAS DENTADAS / NUNO CRATO

AUTOR(ES):	Crato Nuno 1952-; Santos José Carlos, ed. lit.; Valente Guilherme, ed. lit.
EDIÇÃO:	4o ed.
PUBLICAÇÃO:	Lisboa : Gradiva 2008
DESCR. FÍSICA:	245 p. : il. ; 23 cm
COLECÇÃO:	Temas de Matemática / José Carlos Santos / Guilherme Valente ; 6
ISBN:	978-989-616-241-2