Simulação e a Recriação de Fenómenos (Pseudo)aleatórios

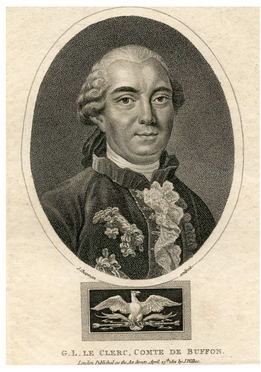
Rui Santos *rui.santos@ipleiria.pt*, Escola Superior de Tecnologia e Gestão do Instituto Politécnico de Leiria, CEAUL — Centro de Estatística e Aplicações da Universidade de Lisboa

Trabalho financiado por Fundos Nacionais através da Fundação para a Ciência e a Tecnologia, no âmbito do projeto PEst-OE/MAT/UI0006/2011.

A simulação, com recurso às potencialidades dos atuais computadores, permite observar o comportamento de fenómenos aleatórios repetidos um grande número de vezes e, através dos teoremas de convergência estocástica, caracterizar probabilisticamente esses fenómenos.

A origem da simulação

A realização repetida de uma experiência aleatória e a determinação da frequência relativa de determinado acontecimento, com o objetivo de determinar um valor aproximado da probabilidade desse acontecimento, pode ser encontrada de uma forma isolada nos séculos XVIII ou XIX.

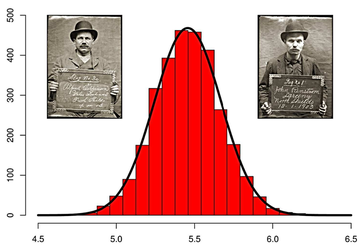


Buffon

Se considerarmos como simulação o uso de uma ferramenta que nos permita gerar dados aleatórios ou pseudoaleatórios, com vista à visualização empírica de um determinado resultado matemático, podemos situar a sua origem em meados do século XVIII. De facto, nesta época, conde de Buffon (1701−1788) efetuou 2048 conjuntos de lançamentos de moedas na procura de uma resposta para o famoso *paradoxo de São Petersburgo*. Outro problema que suscitou a curiosidade dos matemáticos para a sua verificação empírica foi o *problema das agulhas de Buffon*.

Se usarmos uma definição de simulação mais restrita, e considerarmos que o seu objetivo é aprender mais sobre um determinado processo que se está a estudar, então a origem da sua utilização será reportada apenas ao último quarto do século XVIII, com os trabalhos de Erastus de Forest (1834−1888), George Darwin (1809−1882) e Francis Galton (1822−1911).

Todavia, a sua origem é usualmente atribuída a William Gosset (1876−1937), mais conhecido por Student, que, através de 3000 observações (cartões com as medidas antropométricas do dedo médio de 3000 criminosos), gerou 750 amostras escolhidas aleatoriamente de dimensão 4 para analisar o comportamento da média amostral em amostras pequenas.



Independentemente da definição de simulação que adotemos, esta consiste num processo artificial que imita o comportamento de um fenómeno aleatório, permitindo obter uma descrição aproximada das suas caraterísticas. A partir da década de 1940 a simulação obteve uma enorme divulgação devido ao aparecimento do *método de simulação de Monte Carlo* bem como dos computadores eletrónicos, permitindo resolver problemas cuja resolução teórica é complexa ou cuja experimentação real é condicionada por aspetos económicos, temporais ou espaciais.

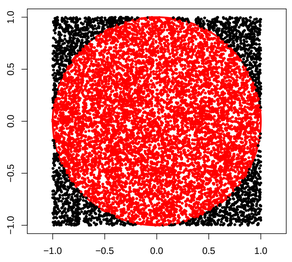
Von Neumann

Segundo Nicholas Metropolis (1915−1999), Stanislaw Ulam (1909−1984) ter-se-á lembrado das técnicas estatísticas de amostragem e do facto de estas haverem sido muitas vezes rejeitadas devido à morosidade e complexidade dos cálculos que lhes estavam associados, vendo no novo computador ENIAC (*Electronic Numerical Integrator and Computer*) uma forma de *ressuscitar* essas técnicas e discutindo esta ideia com John von Neumann (1903−1957).

Ulam

O método de simulação de Monte Carlo foi desenvolvido por Ulam para estudar o coeficiente de difusão dos neutrões dos átomos de urânio para a construção da bomba atómica e pôde ser implementado com os meios computacionais do ENIAC. Contudo, Enrico Fermi (1901−1954) já tinha desenvolvido, de forma independente, o método quinze anos antes, mas não publicou nenhum artigo sobre o tema. Ulam e Metropolis publicam, em 1949, o primeiro artigo sobre o método de simulação de Monte Carlo (nome proposto por Metropolis pelo facto de estarem a trabalhar em projetos secretos — Projeto Manhattan, e Monte Carlo ser uma das capitais mundiais do jogo).

A simulação de Monte Carlo consiste num processo de simulação computacional que gera números pseudoaleatórios (números gerados a partir de algoritmos conhecidos e, como tal, conhecido o número inicial e o algoritmo utilizado é possível determinar os números que irão ser gerados) e determina que proporção desses números verifica uma ou mais propriedades.



Simulação do valor de π

Simulação e convergência estocástica

Através de números (pseudo)aleatórios (distribuição uniforme no intervalo $[0,1]$) podemos gerar réplicas de qualquer distribuição de probabilidade recorrendo ao Teorema da Transformação Uniformizante (através de transformações da distribuição uniforme). Deste modo, podemos simular n réplicas independentes de uma variável aleatória X com função de distribuição $\mathbf{F}_X$ (uma amostra aleatória de dimensão n da característica X). Os resultados de convergência estocástica asseguram, sob determinadas condições, que, para uma dimensão n elevada, podemos através da observação de n réplicas inferir ou obter valores aproximados para algumas caraterísticas de X, razão pela qual os resultados assintóticos são fundamentais como elo de ligação entre a Probabilidade e a Estatística. Por este motivo, a Lei dos Grandes Números (convergência quase certa da média amostral para a média populacional), o Teorema Limite Central (convergência em distribuição da média amostral para a distribuição gaussiana), o Teorema de Glivenko-Cantelli (convergência uniforme da função de distribuição empírica para a função de distribuição teórica) e o Teorema do Limite Extremal (convergência em distribuição dos extremos para a distribuição generalizada de valores extremos), entre outros resultados de convergência, são o fundamento da utilização da simulação.

O papel fundamental da simulação na Estatística

A simulação, com a atual capacidade de cálculo dos computadores, é uma das principais ferramentas do final do século XX e início do XXI, como ilustram duas metodologias das principais abordagens à Inferência Estatística, a metodologia *Bootstrap* da visão frequencista e o método de Monte Carlo via Cadeias de Markov (MCMC – *Markov Chain Monte Carlo*) da visão bayesiana.

A dedução das distribuições amostrais de funções $h(X_1, X_2, \cdots, X_n)$ da amostra aleatória $(X_1, X_2, \cdots, X_n)$, é muitas vezes uma tarefa complexa. Contudo, conhecendo a distribuição das variáveis X_i e a função h podemos recorrer à simulação de amostras para analisar o comportamento de $h(X_1, X_2, \cdots, X_n)$. No caso de a distribuição de X_i ser desconhecida, Bradley Efron (1938−) propôs, em 1979, a utilização da metodologia de reamostragem *Bootstrap*, em que, a partir da amostra observada $(x_1, x_2, \cdots, x_n)$, se simulam m amostras independentes com reposição e mesma dimensão n (cada elemento destas amostras corresponde à escolha aleatória de um elemento da amostra observada). Deste modo, com m elevado, podemos caracterizar o comportamento de $h(X_1, X_2, \cdots, X_n)$, nomeadamente estimar a sua distribuição com base na sua função de distribuição empírica (observada nas amostras simuladas). Apesar de este método ter sido proposto como uma metodologia não paramétrica, há já diversas versões paramétricas, bem como versões que permitam incluir dependência entre as variáveis. Esta metodologia desempenha atualmente um papel de destaque na Estatística frequencista.

Na visão bayesiana, a obtenção analítica de distribuições *a posteriori* é, na maioria das situações, extremamente complexo. Deste modo, na determinação de tais distribuições tornou-se imprescindível o recurso à simulação, nomeadamente ao método de Monte Carlo via Cadeias de Markov (MCMC). Este método é baseado nos conceitos de Cadeia de Markov de ordem 1 (a distribuição de uma observação depende unicamente da informação da observação anterior), no algoritmo de simulação de Metropolis (deduzido em 1953 com o objetivo de caracterizar as propriedades da distribuição de Boltzmann, utilizada na Física Estatística para descrever alguns comportamentos de partículas) e no método de amostragem de Gibbs (nome de uma variante do algoritmo de Metropolis deduzida pelos irmãos Donald Geman (1943−) e Stuart Geman (1949−) em 1984, criada para ser aplicada na recuperação de imagens digitais, supondo que as alterações sofridas em cada ponto dependem apenas de uma vizinhança desse ponto (distribuição de Gibbs) onde é utilizada uma estrutura bayesiana). Por fim, em 1990, Alan Gelfand (1945−) e Adrian Smith (1946−) introduzem esta metodologia na Estatística. Desde então o número de aplicações desta metodologia de simulação aumentou exponencialmente, nomeadamente na Estatística bayesiana onde permitiu a obtenção de distribuições *a posteriori*, tornando a aplicação da metodologia bayesiana mais acessível e, por conseguinte, mais utilizada na atualidade.