A Época de Ouro — 1970/1980

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.

O rápido incremento do poder computacional verificado a partir da segunda metade do século XX, sobretudo com o aparecimento dos computadores pessoais na década de 1970, acompanhado pelo desenvolvimento de diversos programas específicos de aplicação da Estatística, teve um impacto substancial na evolução da Estatística, tornando as suas aplicações mais acessíveis, céleres e eficazes, bem como permitindo o desenvolvimento (e o reaparecimento) de metodologias que, sem o computador, são impraticáveis. Deste modo, verificou-se um aumento significativo na investigação realizada nesta área e, por conseguinte, na quantidade de trabalhos publicados, no aparecimento de novas revistas científicas, bem como na diversificação de áreas com metodologias específicas para a aplicação da Estatística.

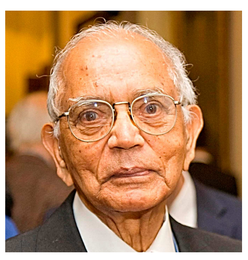
A evolução dos computadores e dos programas de Estatística

A aplicação de muitas técnicas de Estatística, que hoje rotineiramente utilizamos, seriam impossíveis ou, pelo menos, bastante morosas de concretizar caso não pudéssemos recorrer à capacidade de cálculo dos computadores. Por este motivo, a evolução dos computadores teve um impacto significativo na evolução da Estatística ao ponto de Frank Yates, já em 1966, considerar que os computadores são os responsáveis pela "*Second Revolution in Statistics*" (considerando que a primeira revolução, realizada principalmente por Karl Pearson e Fisher, deveu-se ao uso da máquina de calcular). Por exemplo, na determinação dos valores dos coeficientes de uma regressão linear aplicando o método dos mínimos quadrados, técnica que já era aplicada no século XIX, é necessário determinar a inversa de matrizes que, se não tiverem uma dimensão diminuta, é uma tarefa árdua se for efetuada sem o recurso a um computador. Deste modo, as técnicas de análise multivariada, apesar de algumas já existirem antes do aparecimento dos computadores, tinham uma aplicação quase nula e, por este motivo, eram praticamente desconhecidas. Por conseguinte, a análise estatística era quase restrita ao caso univariado e, mesmo neste caso, as metodologias aplicadas eram muitas vezes restritas a modelos simplificados, passíveis de serem aplicados (sem o recurso ao computador). Desta forma, a utilização do computador é indispensável nas aplicações da Estatística, quer se o número de observações for razoável, quer na aplicação de algumas metodologias que são impraticáveis sem o recurso às suas capacidades de cálculo.

Na década de 1970 assistiu-se a uma enorme evolução dos computadores, nomeadamente com a criação dos primeiros computadores pessoais, pois além de se tornarem cada vez mais pequenos, em tamanho e peso, bem como acessíveis (monetariamente), verificou-se um aumento significativo da sua capacidade computacional bem como de armazenamento de informação (grandes quantidades de dados).

Para a Estatística beneficiar da evolução dos computadores era imprescindível o desenvolvimento de programas que possibilitassem a utilização das principais metodologias de aplicação da Estatística. Os primeiros programas de Estatística a serem desenvolvidos foram o **BMDP** (*BioMeDical Package*), que começou a ser desenvolvido em 1957 na Universidade da Califórnia, Los Angeles, na área das aplicações em saúde, sendo a sua versão original distribuída a partir de 1961. Em 1968 surge a primeira versão do **SPSS** (inicialmente denominado *Statistical Package for the Social Sciences*), desenvolvido na Universidade de Chicago para aplicações às ciências sociais. Quase simultaneamente surge o **SAS** (*Statistical Analysis System*) na Universidade da Carolina do Norte, sendo destinado para aplicações na área da agronomia. Na década de 70 surgem diversos programas, tais como o **SUDAAN**, desenvolvido em 1970, o **SOUPAC** em 1971, o **Minitab** e o **SAP** em 1972, o **S** em 1976, entre muitos outros. Atualmente há uma enorme variedade de programas disponíveis, gratuitos ou pagos, que permitem a aplicação das principais técnicas estatísticas.

O desenvolvimento das aplicações

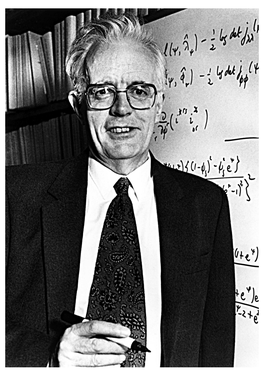

Calyampudi Rao

Entre 1940 e 1970 verificaram-se avanços fundamentais na Estatística Matemática, destacando-se o papel de Calyampudi Rao, como ilustra a desigualdade de Cramér-Rao, o teorema de Rao-Blackwell ou o teste de *scores* de Rao. Outros investigadores influentes nestas décadas foram Maurice Bartlett, David Blackwell, Harold Cramér, Erich Lehmann, Dennis Lindley, Henry Scheffé, entre muitos outros.

No início da década de 1970 as bases da Estatística Matemática estavam praticamente assentes. Deste modo, assistiu-se sobretudo a uma grande evolução das aplicações da Estatística, com a introdução de novas metodologias com o objetivo de resolução de problemas concretos.

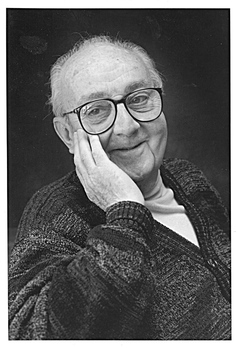
John Tukey destaca a importância da análise exploratória dos dados em 1977, introduzindo, por exemplo, o diagrama de extremos e quartis que hoje ensinamos no ensino básico. Na década de 1980 publica, com David Hoaglin e Frederik Mosteler, duas obras influentes que divulgaram a utilização da análise exploratória de dados bem como da análise robusta (uma área que teve um grande desenvolvimento a partir da década de 1970).

Com o crescente número de publicações baseadas em análises de dados houve a necessidade de desenvolver técnicas que permitam analisar simultaneamente um conjunto de estudos independentes, para sintetizar e combinar as conclusões individuais, bem como obter resultados mais credíveis. K. Pearson, Fisher, Yates, Cochran e Finney já tinham realizado estudos com objetivos análogos, mas somente em 1976 surge o termo meta-análise numa aplicação à psicologia de Gene Glass. Em 1974 Peter Elwood e Archie Cochrane realizam o primeiro ensaio clínico aleatório para avaliar se a aspirina diminui a mortalidade após infarto do miocárdio, não tendo obtido resultados significativos, pelo que integraram novos dados na investigação utilizando técnicas de meta-análise. A publicação dos resultados em 1980 difundiu estas metodologias na Medicina.


David Cox

A análise de sobrevivência, com origem no século XVII na demografia e ciências atuariais, trabalha sobretudo com dados censurados com o objetivo de analisar o tempo que decorre até a ocorrência de um determinado acontecimento de interesse, podendo ser aplicada em medicina, biologia, demografia, controlo de qualidade, entre muitas outras áreas. Sir David Cox, célebre pela transformação Box-Cox definida em 1964 com o objetivo de estabilizar a variância em estudos econométricos e em sucessões cronológicas, introduz em 1972 o modelo de riscos proporcionais (também denominado modelo de Cox) para avaliar o efeito das covariáveis.

George Box, de quem é famosa a frase "*Essentially all models are wrong, but some are useful*", e Gwilym Jenkins propõem em 1970 uma metodologia para a aplicação do modelo ARIMA (auto-regressivo integrado de média móvel) na modelação de séries cronológicas. A metodologia de Box-Jenkins permite estabilizar a série e estimar os parâmetros do modelo com base nos coeficientes de autocorrelação e autocorrelação parcial. Em 1982 Robert Engle propõe o modelo ARCH (auto-regressivo com heterocedasticidade condicional) e em 1986 Tim Bollerslev propõe o modelo ARCH generalizado (GARCH) para modelar a volatilidade dos mercados financeiros.


George Box

Em 1972 John Nelder e Robert Wedderburn apresentam os Modelos Lineares Generalizados (GLM), uma generalização da regressão linear, permitindo obter um algoritmo para a obtenção da estimativa de máxima verosimilhança de diversos modelos de regressão, tais como a regressão linear ou a logística. Em 1986 surgem os Modelos Aditivos Generalizados (GAM) propostos por Trevor Hartie e Robert Tibshirani, uma extensão dos GLM permitindo que algumas covariáveis tenham um efeito não linear no preditor associado à variável resposta. Atualmente estas metodologias desempenham um papel capital em muitas aplicações da Estatística.

Robert Wedderburn define em 1974 o método de estimação por quasi-verossimilhança aplicável em situações de sobredispersão. Em 1977 Arthur Dempster, Nan Laird e Donald Rubin propõem o algoritmo EM (*expectation-maximization*), um método iterativo para determinar o estimador de máxima verossimilhança ou de máxima densidade *a posteriori* quando o modelo depende de variáveis latentes (não observadas de forma direta) ou existem dados omissos. No mesmo ano Kimiko Bowman e Leonard Shenton investigam a aplicação do método da máxima verossimilhança em pequenas amostras.

Em 1949 Maurice Quenouille cria a metodologia de reamostragem *Jackknife*, sendo desenvolvido por Tukey em 1958. Em 1979 Bradley Efron propõe a utilização da metodologia de reamostragem *Bootstrap*, com o objetivo de deduzir "what Jackknife is an approximation to". Em 1984 os irmãos Donald Geman e Stuart Geman desenvolvem o método de amostragem de Gibbs que está na base dos métodos de Monte Carlo via Cadeias de Markov (MCMC) cujo uso intensivo pela comunidade estatística se deve ao seminal artigo de Alan Gelfand e Adrian Smith de 1990. No mesmo ano surge a metodologia *Data Mining* (Mineração de Dados) com a necessidade de análise de grandes quantidades de dados, com o objetivo de detetar padrões ou estruturas nos dados, outra área na qual o computador é imprescindível.

De facto, a evolução do computador tornou a aplicação da Estatística bastante acessível, de tal modo que na atualidade a Estatística desempenha um papel fundamental em praticamente todas as áreas do conhecimento, sendo o seu domínio indispensável em muitas profissões. Por conseguinte, o desenvolvimento da literacia estatística é considerado fundamental e, por este motivo, os conceitos basilares de Estatística já ocupam um lugar de destaque no ensino básico e secundário.