PERSPECTIVAS

Artigo convidado
Versão traduzida
DOI: http://dx.doi.org/10.1590/S0034-759020190609

PLUS ÇA CHANGE, PLUS C'EST LA MÊME CHOSE

Em 1572, a análise de um único ponto de dados, a Supernova de Tycho, revelou que a abóboda celeste é inconstante, ao contrário do paradigma aceito na época (Wootton, 2015). Menos de 40 anos depois, em 1610, Galileu Galilei publicou suas sensacionais descobertas em *Sidereus Nuncius*, um curto tratado que demonstrou a existência de estrelas não vistas a olho nu e a natureza da Via Láctea (Galilei, 1610). Desde então, análises de dados têm sido fundamentais em pesquisas científicas, e há vários exemplos de seu uso na solução de problemas importantes e difíceis. Aos 24 anos, Carl Friedrich Gauss (1809) usou mínimos quadrados para prever corretamente a posição do planeta anão Ceres em 1801, ao aparecer por trás do brilho do Sol. Uma análise espacial simples identificou a fonte do surto de cólera em Broad Street, Londres, em 1854 (Snow, 1855). Entre 1856 e 1863, uma estimativa criteriosa de frequências permitiu a descoberta das regras básicas da hereditariedade de características físicas em plantas por Gregor Mendel (1866). No final da década de 1940, um estudo abrangente e retrospectivo liderado por Richard Doll e A. Bradford Hill (1950) demonstrou a forte ligação entre tabagismo e câncer de pulmão.

O constante desenvolvimento e refinamento de novos métodos estatísticos após 1880 por Galton, Student, Fisher e outros pesquisadores originou diversas aplicações de métodos de análise de dados na indústria e nos negócios. Ideias e métodos desenvolvidos e popularizados por Shewhart, Deming e outros intelectuais tornaram o controle estatístico de qualidade parte integrante do processo de fabricação industrial. Esse processo incorporou o uso de modernos modelos experimentais após a Segunda Guerra Mundial. O poder de computação de baixo custo, a coleta automatizada de dados e o desenvolvimento de *softwares* simples e versáteis de análise de dados, especialmente o R, que é abrangente e gratuito, expandiram significativamente as aplicações estatísticas. Consequentemente, nasceu a era do *Big Data* (grande volume de dados). Assumindo que grande parte do que é publicado é verdadeiro, o *Big Data* resolverá problemas críticos em áreas tão diferentes quanto diagnósticos médicos, avaliação de crédito, previsão do tempo e reconhecimento facial. Teremos produtos e serviços muito melhores, e uma compreensão muito mais profunda de processos físicos e culturais, como resultado da análise de conjuntos de dados cada vez maiores e uso de métodos robustos em computadores modernos.

Embora fazer previsões seja difícil, tenho certeza de que os principais problemas que enfrentaremos ao aplicar métodos de análise estatística a problemas de negócios na era do *Big Data* serão os mesmos problemas com os quais lidamos há décadas. Os dados analisados continuarão sendo informações, ideias e conclusões. Os dados são usados para contar uma história e analisados com métodos analíticos. Os três grandes desafios da análise de dados, em ordem de importância, são a sobrevalorização da significância estatística, a falta de reprodutibilidade e deixar de fornecer respostas exatas a perguntas erradas.

A significância estatística é vista por muitos como o padrão-ouro. Em termos simples, significância estatística é a probabilidade de que um conjunto de observações seja o resultado de

FLAVIO BARTMANN[1]
fcb2122@columbia.edu
ORCID: 0000-0002-9308-3049

[1]Columbia University, School of International and Public Affairs, Nova York, NY, Estados Unidos da América

 CC BY

flutuação aleatória. A obtenção de um valor estatístico pequeno, geralmente menor que 5%, indica a presença de fatores que explicam o comportamento não aleatório. O cálculo do valor-P (a medida usual de significância) envolve um processo complexo de desenho de estudo, seleção de modelos estatísticos, e coleta e análise de dados. O nível de significância postulado é preciso apenas se todas as condições estatísticas forem satisfeitas e o modelo estatístico escolhido fornecer uma interpretação adequada dos dados. No entanto, esse raramente é o caso. As situações são desafiadoras no contexto comercial em que os modelos usados podem ser muito complexos. Esses modelos complexos são tipicamente modelos de regressão com diversas variáveis explicativas em que as relações com a resposta são consideradas lineares, e o cálculo da significância estimada dos coeficientes é frequentemente irrelevante. Coleta de dados e delineamento experimental inadequados aumentam a complexidade. O problema é tão sério que um grande movimento estabelecido para reduzir ou mesmo eliminar os valores-P reuniu muito apoio entre os estatísticos (McShane, Gal, Gelman, Robert e Tackett, 2019).

O segundo grande problema com a análise de dados é que muitos dados, talvez a maioria, não são reprodutíveis (Ioannidis, 2005). A causa mais comum é uma ou mais falhas no planejamento de um experimento ou pesquisa. Além disso, falhas na coleta de dados, uso de métodos inadequados e plágios são comuns e invalidam estudos. Em pesquisas científicas, os estudos podem ser refeitos, e algum grau de autocorreção pode ser alcançado. Contudo, em ambientes industriais, onde as decisões de negócios são frequentemente orientadas por dados urgentes, a replicação de pesquisas é mais difícil. Um estudo de mercado ou viabilidade mal realizado pode ter consequências onerosas. Um caso típico foi a plataforma Newton da Apple.

O terceiro problema pode ser o mais sério e o mais difícil de resolver. John Tukey (1962) costumava dizer que "uma resposta aproximada à pergunta certa vale muito mais do que uma resposta precisa à pergunta errada." A história clássica contada exaustivamente por estatísticos no mundo inteiro (ou "cientistas de dados", no jargão moderno), mas com mais frequência nos corredores do Departamento de Estatística da Universidade de Columbia, é sobre o reforço de aviões britânicos usados nos bombardeios da Alemanha no final da Segunda Guerra Mundial. Muitos aviões foram danificados pela artilharia antiaérea alemã e, por esse motivo, foram blindados. A preocupação mais importante a ser abordada pelos envolvidos nesse projeto foi quais partes deveriam ser blindadas. Os aviões que retornaram com avarias foram cuidadosamente examinados, e foi decidido blindar as partes que haviam sofrido mais danos pela artilharia. Curiosamente, teria sido um erro fatal se essa decisão tivesse sido tomada. A variável correta que deveria ter sido revisada não estava relacionada aos aviões que foram examinados, mas aos aviões que haviam sido destruídos e não retornaram. Felizmente, Abraham Wald inteligentemente sugeriu que as partes não afetadas nos aviões que retornaram deveriam ter sido reforçadas.

Conjuntos de dados, mesmo aqueles com *terabytes*, são apenas a matéria-prima do conhecimento. Hoje, quase tudo pode ser monitorado e medido, mas o principal desafio continua sendo a capacidade de usar e analisar conjuntos de dados e compreendê-los para fornecer interpretações confiáveis.

NOTA DA REDAÇÃO

Tradução do título: Quanto mais as coisas mudam, mais elas permanecem as mesmas

REFERÊNCIAS

Doll, R., & Hill, A. B. (1950) Smoking and carcinoma of the lung. British Medical J*ournal*, *2*(4682), 739-748. doi:10.1136/bmj.2.4682.739

Galilei, G. (1610). *Sidereus Nuncius*. Venice, IT: Thomam Baglionum.

Gauss, C. F. (1809). *Theoria motus corporum coelestium in sectionibus conicis solem ambientium*. Hamburg, Germany: Friedrich Perthes and I. H. Besser.

Ioannidis, J. P. A. (2005). Why most published research findings are false. PLOS Medicine, 2(8), e124. doi:10.1371/journal.pmed.0020124

McShane, B. B., Gal, D., Gelman, A., Robert, C., & Tackett, J. L. (2019). Abandon statistical significance. The American Statistician, 73, 235-245. doi:10.1080/00031305.2018.1527253

Mendel, G. (1866). Versuche über Plflanzenhybriden. Verhandlungen des naturforschenden Vereines in Brünn, Bd. IV für das Jahr 1865 (pp. 3-47). *Abhandlungen*.

Snow, J. (1855). On the *m*ode of communication of cholera. London, UK: John Churchill.

Tukey, J. W. (1962). The future of data analysis. Annals of Mathematical Statistics, 33, 1-67. doi:10.1214/aoms/1177704711

Wootton, D. (2015). *The invention of science: A new history of the scientific revolution*. London, UK: Allen Lane.