Charles Wheelan

# Estatística

## O que é, para que serve, como funciona

Tradução:
George Schlesinger

Revisão técnica:
Jairo Nicolau
*Professor titular do Departamento de Ciências Políticas da UFRJ*

ZAHAR

*Para Katrina*

Título original:
*Naked Statistics*
(*Stripping the Dread from the Data*)

Tradução autorizada da primeira edição americana,
publicada em 2013 por W.W. Norton & Company,
de Nova York, Estados Unidos



Grafia atualizada respeitando o novo
Acordo Ortográfico da Língua Portuguesa

Preparação: Rosa L. Peralta
Revisão: Carolina Rodrigues, Eduardo Monteiro
Indexação: Gabriella Russano | Capa: Estúdio Insólito

CIP-Brasil. Catalogação na publicação
Sindicato Nacional dos Editores de Livros, RJ

Wheelan, Charles
W57e Estatística: o que é, para que serve, como funciona/Charles Wheelan; tradução George Schlesinger. – 1.ed. – Rio de Janeiro: Zahar, 2016.
il.

Tradução de: Naked statistics (stripping the dread from the data)
Apêndice
Inclui bibliografia e índice
ISBN 978-85-378-1512-0

1. Estatística. I. Schlesinger, George. II. Título.

CDD: 519.5
15-26679 CDU: 519.2

# Introdução

*Por que eu detestava cálculo, mas adoro estatística*

Nunca tive uma boa relação com a matemática. Não gosto dos números pelos números em si nem me impressiono com fórmulas rebuscadas que não têm aplicação no mundo real. No ensino médio eu desgostava particularmente de cálculo pela simples razão de que ninguém jamais se deu ao trabalho de me dizer por que eu precisava aprender aquilo. Qual é a área sob uma parábola? Quem se importa?

Um dos momentos decisivos da minha vida ocorreu durante meu último ano no colégio, no fim do primeiro semestre do curso de Advanced Placement em cálculo.* Embora estivesse concentrado para o exame final, devo admitir que estava menos preparado do que deveria. (Eu havia sido aceito na minha primeira opção universitária algumas semanas antes, o que drenara a já pouca motivação que eu tinha para o curso.) Quando comecei a fazer o exame, as questões me pareceram completamente estranhas. Não quero dizer que estava tendo dificuldades em resolvê-las. Quero dizer que eu nem mesmo reconhecia o que estava sendo pedido. Para mim, não era nenhuma novidade estar despreparado para as provas, mas, parafraseando Donald Rumsfeld, eu geralmente sabia o que não sabia. Aquele exame parecia ainda mais incompreensível que o normal. Folheei então as páginas por alguns momentos e acabei de certa maneira me rendendo. Fui até a frente da classe, onde a minha professora de cálculo, que

* Advanced Placement (colocação avançada) é um programa instituído pelas autoridades educacionais americanas e canadenses oferecendo currículo e exames de nível universitário para estudantes do ensino médio. As faculdades americanas muitas vezes destinam vagas e créditos para alunos que obtêm as melhores notas nas provas desse programa. (N.T.)

chamarei de Carol Smith, estava supervisionando o exame. "Sra. Smith", eu disse, "não reconheço grande parte do que está sendo pedido no teste."

Basta dizer que a sra. Smith não gostava de mim muito mais do que eu gostava dela. Sim, posso admitir agora que às vezes eu usava meus limitados poderes de presidente da associação de alunos para marcar assembleias de toda a escola justamente para que a aula da sra. Smith fosse cancelada. Sim, meus amigos e eu chegamos a mandar flores de "um admirador secreto" para a sra. Smith durante uma aula só para podermos cair na risada no fundo da sala enquanto ela olhava ao redor envergonhada. E, sim, eu parei de fazer qualquer dever de casa assim que entrei na faculdade.

Logo, quando fui até a sra. Smith no meio do exame e disse que a matéria não me parecia familiar, ela foi, por assim dizer, pouco solidária. "Charles", disse em voz alta, ostensivamente para mim, mas dirigindo-se às filas de carteiras para se certificar de que toda a classe ouvisse, "se você tivesse estudado, a matéria lhe pareceria mais familiar." Era um ponto inquestionável.

Então bati em retirada de volta para minha carteira. Após alguns minutos, Brian Arbetter, um aluno de cálculo muito mais comprometido que eu, foi até a frente da classe e cochichou algo para a sra. Smith. Ela cochichou de volta e então aconteceu uma coisa verdadeiramente extraordinária. "Classe, preciso da atenção de vocês", a sra. Smith anunciou. "Parece que eu lhes dei o exame do segundo semestre por engano." Já estávamos bem adiantados no horário do teste, de modo que o exame inteiro precisou ser cancelado e remarcado.

Não posso descrever a minha euforia. Parti para a vida, casei-me com uma mulher encantadora, tivemos três filhos saudáveis. Publiquei livros e visitei lugares como o Taj Mahal e o Angkor Wat. Ainda assim, o dia em que a minha professora de cálculo levou o troco é um dos cinco momentos mais formidáveis da minha vida. (O fato de eu quase ter sido reprovado no exame final substitutivo não diminuiu em praticamente nada essa maravilhosa experiência.)

O incidente do exame de cálculo conta muito do que você precisa saber sobre a minha relação com a matemática – mas não tudo. Curiosamente,

no ensino médio eu adorava física, embora a física se apoie fortemente nesse mesmo cálculo que eu me recusava a fazer na aula da sra. Smith. Por quê? *Porque a física tem um propósito claro.* Lembro-me muito bem do meu professor de física no colégio mostrando-nos, durante o campeonato mundial de beisebol, como podíamos usar a fórmula básica da aceleração para estimar a que distância fora rebatida uma bola de *home run*.* Isso é bacana – e a mesma fórmula tem muitas outras aplicações socialmente significativas.

Na faculdade, eu me interessei especialmente pela probabilidade, mais uma vez porque ela me permitia compreender fascinantes situações da vida real. Hoje reconheço que não era a matemática que me incomodava nas aulas de cálculo, e sim ninguém nunca ter me explicado seu sentido. Se você não é fascinado pela elegância da fórmula em si – o que, sem dúvida, eu não sou –, então o cálculo não passa de fórmulas mecânicas e enfadonhas, pelo menos do jeito que me foi ensinado.

Isto me leva para a estatística (que, para os propósitos deste livro, inclui a probabilidade). Eu adoro estatística. Ela pode ser usada para explicar tudo, desde testes de DNA até a idiotice de jogar na loteria. A estatística pode nos ajudar a descobrir os fatores associados a doenças cardíacas e câncer, bem como identificar fraudes em testes padronizados. A estatística pode até nos ajudar a ganhar jogos de programas de TV. Na minha infância, havia um programa famoso chamado *Let's Make a Deal*, com seu igualmente famoso apresentador, Monty Hall. Todo dia no fim do programa, um jogador bem-sucedido ficava junto com Monty diante de três portas: porta n.1, porta n.2 e porta n.3. Monty Hall explicava ao jogador que havia um prêmio altamente desejável atrás de uma das portas – algo como um carro novo – e uma cabra atrás das outras duas. A ideia era simples e direta: o jogador escolhia uma das portas e ficava com o conteúdo atrás dessa porta.

Quando cada jogador ou jogadora ficava diante das portas com Monty Hall, tinha uma chance em três de escolher a porta que seria aberta para

* *Home run*: rebatida em que a bola não consegue ser reposta em jogo antes que o rebatedor consiga dar a volta inteira até a última base. (N.T.)

revelar o valioso prêmio. Mas *Let's Make a Deal* tinha um truque, que tem deleitado os estatísticos desde então (e deixado todo mundo estarrecido). Depois que o jogador escolhia uma porta, Monty Hall abria uma das duas restantes, sempre revelando uma cabra. Digamos que o jogador tivesse escolhido a porta n.1. Monty abria então a porta n.3; ali estaria parada a cabra em pleno palco. Duas portas ainda estavam fechadas, as portas n.1 e 2. Se o prêmio valioso estivesse atrás da n.1, o competidor ganharia; se estivesse atrás da n.2, perderia. Mas é aí que as coisas ficavam interessantes: Monty virava-se para o jogador e perguntava se ele gostaria de mudar de ideia e trocar de porta (da n.1 para a n.2, neste caso). Lembre-se, ambas as portas ainda estavam fechadas, e a única informação nova que o competidor tinha recebido era que uma cabra havia aparecido atrás de uma das portas que ele não tinha escolhido.

Deveria ele trocar?

A resposta é sim. Por quê? Leia o Capítulo 5½.

O PARADOXO DA ESTATÍSTICA é que ela está em toda parte – desde médias de rebatidas até pesquisas presidenciais –, embora a disciplina em si seja considerada desinteressante e inacessível. Muitos livros e aulas de estatística são excessivamente carregados de matemática e jargão. Acredite, os detalhes técnicos são cruciais (e interessantes), mas é apenas grego se você não entender intuitivamente. E você pode nem dar importância para a intuição se não estiver convencido de que existe um motivo para aprendê-la. Cada capítulo deste livro promete responder à pergunta básica que fiz (em vão) para a minha professora de cálculo no colégio: *qual é o objetivo disto?*

Este livro é sobre a intuição. É breve em matemática, equações e gráficos. Quando eles forem usados, prometo que terão um propósito claro e elucidativo. Por outro lado, o livro é pródigo em exemplos para convencer você de que existem excelentes motivos para aprender essa matéria. *A estatística pode ser realmente interessante, e a maior parte dela nem é tão difícil.*

A ideia para este livro nasceu não muito tempo depois da minha infeliz experiência na aula de cálculo da sra. Smith. Ingressei na faculdade para

estudar economia e políticas públicas. Antes mesmo de o programa começar, fui enviado (sem surpresa) para o "campo da matemática", junto com a maioria dos meus colegas, para que nos preparássemos para os rigores que viriam em seguida. Durante três semanas, aprendemos matemática o dia todo numa sala de aula sem janelas, num porão (de verdade).

Num desses dias, tive algo muito próximo de uma epifania de carreira. Nosso instrutor tentava nos ensinar as circunstâncias nas quais a soma de uma série infinita converge para um valor finito. Acompanhe meu raciocínio por um minuto porque esse conceito já vai ficar claro. (Neste instante, você provavelmente está se sentindo como eu me sentia naquela sala sem janelas.) Uma série infinita é um padrão de números que continua indefinidamente, tal como 1 + ½ + ¼ + ⅛… As reticências significam que o padrão continua até o infinito.

Essa é a parte que estávamos tendo mais dificuldade de entender. O instrutor tentava nos convencer, usando alguma prova que há muito tempo já esqueci, de que uma série de números pode continuar para sempre e mesmo assim pode redundar (aproximadamente) em um número finito. Um dos meus colegas de classe, Will Warshauer, não aceitava nada daquilo, apesar da impressionante prova matemática. (Para ser honesto, eu mesmo estava um pouco cético.) Como pode a soma de algo infinito resultar em alguma coisa finita?

Aí tive uma inspiração, ou, mais precisamente, a intuição do que o instrutor estava tentando explicar. Virei-me para Will e expliquei-lhe o que eu tinha acabado de elaborar na minha cabeça. Imagine que você tenha se posicionado a dois metros de uma parede.

Agora avance metade da distância até a parede (um metro), de modo que você se encontre a um metro dela.

Dessa distância de um metro, percorra novamente metade da distância (½ metro). E, a partir desse ½ metro, repita o movimento (aproxime-se ¼ de metro, ou 25 centímetros). Depois repita outra vez (mova-se ⅛ de metro, ou 12,5 centímetros). E assim por diante.

Gradualmente, você vai chegando bem perto da parede. (Por exemplo, quando estiver a ⅟₁₀₂₄ de um centímetro, você andará metade dessa distân-

cia, ou mais 1/2048 de um centímetro.) Mas jamais chegará à parede, porque, por definição, cada movimento fará você percorrer apenas a metade da distância restante. Em outras palavras, você chegará infinitamente perto da parede, mas nunca a alcançará. Se medirmos a sua distância em metros, a série poderá ser descrita como 1 + ½ + ¼ + ⅛…

Aí está o insight: mesmo que você continue se aproximando indefinidamente – com cada movimento percorrendo a metade da distância restante até a parede –, a distância total que você irá percorrer jamais poderá ser maior que dois metros, que é a distância do seu ponto de partida até a parede. Para propósitos matemáticos, a distância total que você percorre pode ser aproximada para dois metros, o que acaba sendo muito conveniente para propósitos de cálculo. Um matemático diria que a soma da série infinita 1m + ½m + ¼m + ⅛m… converge para dois metros, que é o que o nosso instrutor estava tentando nos ensinar naquele dia.

O importante é que eu convenci Will. E convenci a mim mesmo. Não consigo me recordar da matemática que prova que a soma de uma série infinita pode convergir para um número finito, mas isso aí eu sempre posso procurar na internet. E quando o fizer, provavelmente fará sentido. Pela minha experiência, a intuição torna a matemática e outros detalhes técnicos mais compreensíveis – mas não necessariamente o contrário.

O objetivo deste livro é tornar mais intuitivos e acessíveis os conceitos estatísticos mais importantes, não só para aqueles de nós obrigados a estudá-los em salas de aula sem janelas, mas para qualquer pessoa interessada no extraordinário poder dos números e dados.

Agora, tendo acabado de demonstrar que as ferramentas centrais da estatística são menos intuitivas e acessíveis do que deveriam ser, vou fazer uma afirmação aparentemente contraditória: a estatística pode ser *extremamente acessível* no sentido de que qualquer um com dados e um computador pode executar procedimentos estatísticos sofisticados usando apenas algumas teclas. O problema é que, se os dados forem pobres, ou se as técnicas estatísticas forem usadas de maneira inadequada, podemos chegar a conclusões

bastante enganosas e até mesmo potencialmente perigosas. Considere a seguinte manchete hipotética de uma notícia na internet: *pessoas que fazem pequenas pausas no trabalho estão muito mais propensas a morrer de câncer.* Imagine essa manchete surgindo do nada na sua tela enquanto você está navegando pela web. De acordo com um estudo em tese impressionante com 36 mil funcionários de escritório (um conjunto de dados enorme!), os funcionários que relataram sair do escritório para pausas regulares de dez minutos durante o dia de trabalho eram 41% mais propensos a desenvolver câncer nos cinco anos seguintes do que os funcionários que não saem do escritório durante o dia de trabalho. Obviamente precisamos agir diante de achados como esse – talvez algum tipo de campanha nacional de conscientização para impedir pausas curtas durante o serviço.

Ou talvez precisemos apenas pensar com mais clareza sobre o que muitos funcionários fazem durante o intervalo de dez minutos. Minha experiência profissional sugere que muitos desses funcionários que relatam sair do escritório para pausas curtas se amontoam na frente da entrada do prédio para fumar (criando uma nuvem de fumaça através da qual o resto de nós precisa passar para entrar ou sair). Eu inferiria que são talvez os cigarros, e não os intervalos breves no expediente, a causa do câncer. Inventei esse exemplo apenas para ser particularmente absurdo, mas posso garantir que muitas abominações estatísticas na vida real são quase tão absurdas uma vez que forem desconstruídas.

A estatística é como uma arma de alto calibre: útil quando usada de forma correta e potencialmente desastrosa em mãos erradas. Este livro *não vai* fazer de você um especialista em estatística; ele *vai* lhe ensinar a ter suficiente cuidado e respeito pela área para que você não cometa o equivalente estatístico de explodir a cabeça de alguém com um tiro.

Este não é um livro-texto, o que é libertador em termos dos tópicos que devem ser cobertos e das maneiras como podem ser explicados. *O livro foi planejado para introduzir os conceitos estatísticos de maior relevância para a vida cotidiana.* Como os cientistas concluem que algo provoca câncer? Como funcionam as pesquisas de opinião (e o que pode dar errado)? Quem "mente com estatística", e como se faz isso? Como a sua empresa

de cartão de crédito usa os dados sobre o que você anda comprando para prever qual a probabilidade de você deixar de efetuar um pagamento? (É sério, eles podem fazer uma coisa dessas.)

Se você quer entender os números por trás da notícia e apreciar o extraordinário (e crescente) poder dos dados, este é o material de que você precisa. No final, espero ter persuadido você da observação feita pela primeira vez pelo matemático e escritor sueco Andrejs Dunkels: é fácil mentir com estatística, mas é difícil dizer a verdade sem ela.

Mas eu tenho aspirações ainda mais arrojadas que essa. Acho que você poderá realmente vir a gostar de estatística. As ideias subjacentes são fabulosamente interessantes e relevantes. A chave é separar as ideias importantes dos herméticos detalhes técnicos que possam atrapalhar. *Esta é a estatística.*

# 4. Correlação

*Como a Netflix sabe quais filmes eu gosto?*

A Netflix insiste que vou gostar do filme *Bhutto*, um documentário que oferece uma "visão em profundidade e às vezes incendiária da vida e da trágica morte da ex-primeira-ministra paquistanesa Benazir Bhutto". Provavelmente vou gostar do filme. (Eu o adicionei ao "Minha lista".) As recomendações da Netflix às quais assisti no passado foram incríveis. E quando eles recomendam um filme a que já assisti, costuma ser um de que eu realmente gostei.

Como a Netflix faz isso? Será que existe alguma gigantesca equipe de estagiários na sede da corporação que usou uma combinação do Google e entrevistas com a minha família e amigos para determinar que eu poderia gostar de um documentário sobre uma ex-primeira-ministra paquistanesa? É claro que não. A Netflix simplesmente domina algumas estatísticas sofisticadas. *A Netflix não me conhece*. Mas conhece os filmes dos quais gostei no passado (porque eu os avaliei). Usando essa informação, junto com as avaliações de outros clientes e um computador potente, a Netflix pode fazer previsões incrivelmente acuradas sobre as minhas preferências.

Adiante voltarei ao algoritmo específico da Netflix para fazer essas escolhas; por enquanto, o ponto importante é que tudo está baseado em correlação. A Netflix recomenda filmes que são semelhantes a outros filmes de que gostei; e também recomenda filmes que foram muito bem avaliados por outros clientes cujas avaliações são similares às minhas. *Bhutto* foi recomendado por causa das cinco estrelas com que avaliei dois outros documentários, *Enron: os mais espertos da sala* e *Sob a névoa da guerra*.

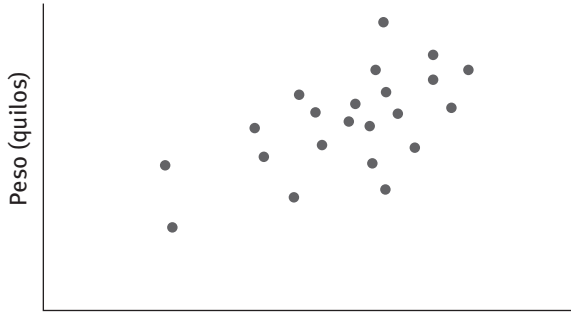
A correlação mede o grau em que dois fenômenos estão relacionados entre si. Por exemplo, existe uma correlação entre temperaturas de verão

e venda de sorvete. Quando uma sobe, a outra sobe também. Duas variáveis têm correlação positiva se uma variação numa delas é associada a uma variação da outra no mesmo sentido, tal como a relação entre altura e peso. Pessoas mais altas pesam mais (em média); pessoas mais baixas pesam menos. Uma correlação é negativa se uma variação positiva numa das variáveis está associada a uma variação negativa na outra, tal como a relação entre exercício e peso.

O aspecto traiçoeiro nesses tipos de associações é que nem toda observação se encaixa no padrão. Às vezes pessoas mais baixas pesam mais que pessoas mais altas. Às vezes pessoas que não se exercitam são mais magras que pessoas que se exercitam o tempo todo. Ainda assim, existe uma relação significativa entre altura e peso, bem como entre exercício e peso.

Se fôssemos colocar num gráfico de dispersão as alturas e pesos de uma amostra aleatória de americanos adultos, seria de esperar ver algo do seguinte tipo:

**Gráfico de dispersão de altura e peso**




Se fôssemos criar um gráfico de dispersão entre exercício (medido em minutos por semana de exercício intensivo) e peso, seria de esperar uma correlação negativa, com os que se exercitam mais tendendo a pesar menos. Mas um padrão consistindo em pontos dispersos numa página é uma ferramenta um tanto tosca. (Se a Netflix tentasse me fazer recomendações de filmes com um gráfico das avaliações de milhares de filmes por milhões de clientes, os resultados soterrariam a sede debaixo de gráficos de disper-

são.) Em vez disso, o poder da correlação como ferramenta estatística é que podemos encapsular uma associação entre duas variáveis numa única estatística descritiva: o coeficiente de correlação.

O coeficiente de correlação tem duas características fabulosamente atraentes. A primeira, por razões matemáticas que foram relegadas ao apêndice, trata-se de um número único que varia de –1 a 1. Uma correlação de 1, muitas vezes descrita como correlação perfeita, significa que qualquer alteração em uma das variáveis está associada com uma alteração equivalente na outra variável no mesmo sentido.

Uma correlação de –1, ou correlação negativa perfeita, significa que toda alteração em uma variável está associada a uma alteração equivalente na outra variável em sentido oposto.

Quanto mais perto de 1 ou –1 estiver a correlação, mais forte a associação. Uma correlação de 0 (ou próxima a 0) significa que as variáveis não têm associação significativa entre si, como a relação entre o número do sapato e os resultados em exames escolares.

A segunda característica atraente do coeficiente de correlação é que ele não está ligado a nenhuma unidade. Podemos calcular a correlação entre altura e peso – mesmo que a altura seja medida em centímetros e o peso em quilogramas. Podemos até calcular a correlação entre a quantidade de televisores que alunos do ensino médio têm em suas casas e seus resultados em exames escolares, e eu lhes asseguro que será positiva. (Falarei mais sobre essa relação daqui a pouco.) O coeficiente de correlação faz uma coisa aparentemente milagrosa: reduz uma complexa bagunça de dados medidos em unidades diferentes (como o nosso gráfico de dispersão de altura e peso) numa única e elegante estatística descritiva.

Como?

Mantendo o hábito, pus a fórmula mais comum para se calcular o coeficiente de correlação no apêndice ao final do capítulo. Essa não é uma estatística que você vai calcular à mão. (Depois de você inserir os dados, um programa básico como o Microsoft Excel calcula a correlação entre as duas variáveis.) Ainda assim, intuitivamente não é tão difícil. A fórmula para calcular o coeficiente de correlação faz o seguinte:

1. Calcula a média e o desvio padrão para ambas as variáveis. Se nos ativermos ao exemplo de altura e peso, saberíamos então a altura média das pessoas na amostra, o peso médio das pessoas na amostra e o desvio padrão tanto para a altura como para o peso.
2. Converte todos os dados de modo que cada observação seja representada por sua distância da média (seu desvio padrão). Acompanhe meu raciocínio; não é tão complicado. Suponha que a altura média na amostra seja de 170 centímetros (com um desvio padrão de dez centímetros); e que o peso médio seja de 75 quilos (com um desvio padrão de cinco quilos). Agora suponha que você tenha 182 centímetros de altura e pese 71 quilos. Podemos dizer também que sua altura é 1,2 desvios padrões acima da média em altura [$^{(180-165)}/_{10}$], e seu peso 0,8 desvios padrões abaixo da média, ou –0,8 para fins de fórmula [$^{(71-75)}/_{5}$]. *Sim, é incomum alguém estar acima da média em altura e abaixo da média em peso, mas já que você pagou um bom dinheiro pelo livro, achei que deveria pelo menos fazer você alto e magro.* Note que a sua altura e peso, anteriormente em centímetros e quilos, foram reduzidos a 1,2 e −0,8. É isso que faz as unidades sumirem.
3. Aqui eu libero minhas mãos e deixo o computador fazer o serviço. A fórmula calcula então a relação entre altura e peso de todos os indivíduos da amostra, medidos pelas unidades-padrão. Quando os indivíduos da amostra são altos, digamos 1,5 ou dois desvios padrões acima da média, o que tende a acontecer com seus pesos *medidos em desvios padrões da média para o peso*? E quando os indivíduos estão perto da média em termos de altura, quais são seus pesos, medidos em unidades de desvio padrão?

Se a distância de uma variável em relação à média tende a ser amplamente consistente com a distância da outra variável em relação à média (por exemplo, pessoas distantes da média em termos de altura, em qualquer um dos dois sentidos, também tendem a estar distantes da média no mesmo sentido em termos de peso), então seria de esperar uma forte correlação positiva.

Se a distância em relação à média de uma das variáveis tende a corresponder a uma distância similar em relação à média da segunda variável

*no sentido oposto* (por exemplo, pessoas bem acima da média em termos de exercício tendem a estar bem abaixo da média em termos de peso), então devemos esperar uma forte correlação negativa.

Se duas variáveis não tendem a se desviar da média segundo nenhum padrão significativo (por exemplo, número do sapato e exercício), então devemos esperar uma correlação pequena ou nula.

Você sofreu intensamente nesta seção; voltaremos já, já para o aluguel de filmes. Antes de retornarmos à Netflix, porém, vamos refletir sobre outro aspecto da vida em que a correlação é relevante: o Teste de Raciocínio SAT. Conhecido antigamente nos Estados Unidos como Teste de Aptidão Acadêmica (SAT, na sigla em inglês), trata-se de um exame padronizado composto de três partes – matemática, leitura crítica e redação – cujo objetivo é mensurar a capacidade acadêmica e predizer o desempenho universitário. É claro que há motivo razoável para se perguntar (especialmente aqueles que não gostam de testes padronizados): não é para isso que serve o ensino médio? Por que um exame de quatro horas é tão importante quando os funcionários encarregados da admissão universitária têm acesso a *quatro anos* de notas tiradas no ensino médio?

A resposta para essas perguntas encontra-se camuflada nos Capítulos 1 e 2. Notas do ensino médio são uma estatística descritiva imperfeita. Um aluno que tira notas medíocres enquanto enfrenta uma programação difícil com aulas de matemática e ciências pode ter maior capacidade e potencial acadêmico do que um aluno no mesmo colégio com notas melhores em matérias menos desafiadoras. Obviamente há discrepâncias potenciais ainda maiores de uma escola para outra. Segundo o College Board, que produz e administra o SAT, o teste foi criado para "democratizar o acesso ao ensino superior para todos os estudantes". Muito justo. O SAT fornece uma medida padronizada de capacidade que pode ser facilmente comparada entre todos os alunos que se candidatam ao ensino superior. *Mas será que é uma boa medida de capacidade?* Se queremos um critério que possa ser comparado facilmente entre estudantes, poderíamos também mandar os

alunos de último ano correrem um tiro de cem metros, que é mais barato e mais fácil do que administrar o SAT. O problema, obviamente, é que a performance num tiro de cem metros não tem correlação com desempenho acadêmico. Obter os dados é fácil; só que eles simplesmente não nos revelam nada de significativo.

Então, qual é a qualidade da informação obtida pelo SAT? Infelizmente para futuras gerações de alunos do ensino médio, o SAT faz um trabalho razoavelmente bom em predizer as notas de primeiro ano de faculdade. O College Board publica as correlações relevantes. Numa escala de 0 (absolutamente nenhuma correlação) a 1 (correlação perfeita), a correlação entre a média de notas no ensino médio e a média de notas no primeiro ano da faculdade é 0,56. (Para dar alguma perspectiva a esse número, a correlação entre altura e peso para homens adultos nos Estados Unidos é aproximadamente 0,4.) A correlação entre o placar composto do SAT (leitura crítica, matemática e redação) e a média das notas do primeiro ano universitário também é 0,56.[1] Esse resultado parece argumentar a favor da eliminação do SAT, pois o teste parece não dar resultados melhores na previsão do desempenho universitário do que as notas do ensino médio. Na verdade, o melhor previsor de todos é uma combinação do SAT e da média do ensino médio, que tem uma correlação de 0,64 com as notas do primeiro ano universitário. Sinto muito por ter que dizer isso.

Um ponto crucial nesta discussão geral é que correlação não implica causalidade; uma associação positiva ou negativa entre duas variáveis não significa necessariamente que uma variação numa delas esteja causando a variação na outra. Por exemplo, anteriormente aludi a uma provável correlação positiva entre os resultados do SAT de um aluno e a quantidade de televisores que sua família possui. Isso não significa que pais superansiosos possam aumentar o placar dos testes de seus filhos comprando cinco aparelhos de televisão adicionais para a casa. E provavelmente tampouco significa que assistir muito à televisão seja bom para o desempenho acadêmico.

A explicação mais lógica para tal correlação seria que pais com elevado nível de educação podem se dar ao luxo de ter uma porção de aparelhos de televisão e tendem a ter filhos cujos resultados nos testes estão acima da média. Tanto televisores como resultados de testes são provavelmente causados por uma terceira variável, que é a educação dos pais. Não posso provar a correlação entre esses aparelhos na casa e resultados do SAT. (O College Board não fornece esses dados.) No entanto, posso provar que alunos de famílias mais ricas têm em média escores no SAT mais altos do que alunos de famílias menos ricas. Segundo o College Board, alunos com renda familiar acima de US$200 mil têm um placar médio no SAT de matemática de 586, em comparação com um placar médio de 460 para alunos com renda familiar de US$20 mil ou menos.[2] Ao mesmo tempo, também é provável que famílias com renda superior a US$200 mil tenham mais televisores em suas (múltiplas) casas do que famílias com renda de US$20 mil ou menos.

Comecei a escrever este capítulo muitos dias atrás. Desde então, tive a oportunidade de assistir ao documentário *Bhutto*, um filme excepcional sobre uma família excepcional. As sequências originais, que começam com a partilha da Índia e do Paquistão em 1947 e vão até o assassinato de Benazir Bhutto em 2007, são extraordinárias. A voz de Bhutto é muito bem intercalada ao longo do filme na forma de discursos e entrevistas. Em todo caso, dei cinco estrelas ao filme, que é praticamente o que a Netflix previu.

No nível mais básico, a Netflix está explorando o conceito de correlação. Primeiro, eu avalio um conjunto de filmes. A Netflix compara minhas avaliações com as de outros clientes para identificar aqueles cujas avaliações estejam altamente correlacionadas com as minhas. Esses clientes tendem a gostar dos filmes que eu gosto. Uma vez estabelecido isso, a Netflix pode recomendar filmes que receberam alta avaliação de clientes de mentalidade semelhante à minha, mas que eu ainda não assisti.

Esse é o "quadro geral". A metodologia real é muito mais complexa. Na verdade, a Netflix lançou em 2006 um concurso no qual membros do público foram convidados a projetar um mecanismo que melhorasse as recomendações existentes da empresa em pelo menos 10% (o que significa que o sistema ficaria 10% mais acurado em predizer como um cliente avaliaria um filme depois de assistir). O vencedor ganharia US$1 milhão.

Todo indivíduo ou equipe que se inscreveu para o concurso recebeu "dados de treinamento" consistindo em mais de 100 milhões de avaliações de 18 mil filmes por 480 mil clientes Netflix. Um conjunto separado de 2,8 milhões de avaliações foi "retido", o que significa que a Netflix sabia como os clientes tinham avaliado esses filmes, mas os participantes do concurso não. Os competidores foram julgados com base na acurácia com que seus algoritmos previam as avaliações reais dos clientes para esses filmes retidos. Durante três anos, milhares de equipes de mais de 180 países submeteram propostas. Havia duas exigências para participar: primeira, o vencedor deveria licenciar o algoritmo para a Netflix; segunda, o vencedor tinha de "descrever ao mundo como você fez e por que funciona".[3]

Em 2009, a Netflix anunciou o vencedor: uma equipe de sete pessoas composta de estatísticos e cientistas da computação dos Estados Unidos, Áustria, Canadá e Israel. Sinto muito, não posso descrever o sistema ganhador, nem mesmo no apêndice. O artigo explicando o sistema tem 92 páginas.* Eu fico impressionado com a qualidade das recomendações da Netflix. Ainda assim, o sistema é apenas uma supervariação rebuscada do que as pessoas vêm fazendo desde a aurora do cinema: achar alguém com gosto semelhante e pedir uma recomendação. Você tende a gostar do que eu gosto, e não gostar do que eu não gosto, então, o que acha do novo filme do George Clooney?

Essa é a essência da correlação.

---

* Você pode lê-lo em: http://www.netflixprize.com/assets/GranPrize2009_BPC_PragmaticTheory.pdf.