# 9. Inferência

## Por que o meu professor de estatística achou que eu podia ter colado

NA PRIMAVERA do meu último ano de faculdade, me inscrevi para o curso de estatística. Na época, eu não era particularmente apaixonado por estatística ou pela maioria das disciplinas com base matemática, mas tinha prometido ao meu pai que faria o curso se pudesse faltar à escola durante dez dias para viajar com a família para a União Soviética. Assim, basicamente fiz o curso em troca da viagem. E isso acabou sendo ótimo, porque gostei de estatística muito mais do que poderia imaginar e visitei a União Soviética na primavera de 1988. Quem pensaria que em breve o país não existiria mais na sua forma comunista?

Essa história é de fato relevante para o capítulo; a questão é que não me dediquei ao curso de estatística durante o período letivo como deveria. Entre outras responsabilidades, também estava redigindo uma tese de encerramento que devia ser entregue mais ou menos na metade do ano. Tínhamos provas regulares no curso de estatística, muitas das quais ignorei ou fui reprovado. Estudei um pouco mais para o exame do primeiro semestre e me saí razoavelmente bem – literalmente. Mas poucas semanas antes do fim do ano, aconteceram duas coisas. Primeira, acabei a minha tese, o que me rendeu muito mais tempo livre. Segunda, percebi que estatística não era nem de perto tão difícil quanto eu imaginava. Comecei a estudar o livro de estatística e fazer os trabalhos do começo do ano. Tirei A no exame final.

Foi aí que o meu professor de estatística, cujo nome esqueci há muito tempo, me chamou em sua sala. Não lembro exatamente o que ele disse, mas foi algo do tipo "Você realmente se saiu muito melhor no exame final do que no do primeiro semestre". Porém, ele não havia me chamado para me parabenizar ou para reconhecer que eu havia finalmente começado a me esforçar na matéria. Havia uma acusação implícita (embora não explícita) nessa convocação; ele de fato esperava que eu explicasse por que me saí tão melhor no exame final do que no do meio do ano. Em suma, o sujeito desconfiou que eu pudesse ter colado. Agora que tenho lecionado por muitos anos, consigo entender melhor essa linha de pensamento. Em quase todo curso que dei, há um surpreendente grau de correlação entre o desempenho de um aluno no exame do primeiro semestre e no exame final. É *sim* muito inusitado um aluno tirar uma nota abaixo da média no meio do ano e depois estar entre os melhores da classe no exame final.

Expliquei que tinha acabado a minha tese e começado a levar o curso a sério (fazendo coisas tipo ler os capítulos recomendados dos livros-textos e fazer dever de casa). Ele pareceu satisfeito com a explicação, e eu saí da sala, ainda um pouco inquieto com a acusação implícita.

Acredite ou não, essa historinha incorpora muito do que você precisa saber sobre inferência estatística, inclusive seus pontos fortes e fragilidades potenciais. *A estatística não pode provar nada com certeza*. Em vez disso, o poder da inferência estatística deriva de observar algum padrão ou resultado e então usar a probabilidade para determinar a explicação mais provável para aquele resultado. Imagine que um jogador estranho chegue a uma cidade e lhe ofereça

uma aposta: ele ganha US$1.000 se tirar seis num único lance de um dado; você ganha US$500 se der qualquer outra coisa – uma aposta muito boa do seu ponto de vista. Ele então vai e tira dez vezes seguidas o seis, levando US$10 mil de você.

Uma explicação possível é que ele tenha tido sorte. Uma explicação alternativa é que de algum modo ele trapaceou. A probabilidade de tirar seis dez vezes seguidas com um dado honesto é de aproximadamente um em 60 milhões. Você não pode provar que ele trapaceou, mas no mínimo deveria ter examinado o dado.

É claro que a explicação mais provável nem sempre é a explicação certa. Coisas extremamente raras acontecem. Linda Cooper é uma mulher da Carolina do Sul que foi atingida por um raio quatro vezes.[1] (A Administração Federal de Controle de Emergências estima a probabilidade de ser atingido por um raio uma única vez como um em 600 mil.) A companhia de seguros de Linda Cooper não pode negar-lhe cobertura simplesmente porque seus ferimentos são estatisticamente improváveis. Voltando ao meu exame de estatística no curso de graduação, o professor tinha um motivo razoável para estar desconfiado. Ele viu um padrão que era bastante improvável; é exatamente assim que investigadores identificam cola em exames padronizados e é como a Comissão de Valores Mobiliários e Câmbio pega operações com informação privilegiada. Mas um padrão improvável é apenas um padrão improvável a não ser que seja corroborado por evidência adicional. Mais adiante neste capítulo discutiremos erros que podem surgir quando a probabilidade nos conduz pelo caminho errado.

Por enquanto, devemos apreciar que a inferência estatística emprega dados para abordar questões importantes. Será que uma droga nova é efetiva no tratamento de doenças cardíacas? Será que celulares provocam câncer? Por favor, perceba que não estou alegando que a estatística pode *responder* a esse tipo de pergunta de forma inequívoca; em vez disso, a inferência nos diz o que é provável e o que é improvável. Pesquisadores não podem provar que uma droga nova é efetiva no tratamento de doenças cardíacas, mesmo quando possuem dados de uma experiência clínica cuidadosamente controlada. Afinal, é muito possível que haja uma variação aleatória nos resultados dos pacientes nos grupos de tratamento e controle, variação esta que não esteja relacionada com a nova droga. Se 53 em cem pacientes que tomam a nova medicação para doenças cardíacas mostraram acentuada melhora em comparação com 49 pacientes em cem que tomaram um placebo, não poderíamos concluir de imediato que a nova droga é efetiva. Esse é um resultado que pode ser explicado facilmente pelas variações casuais entre os dois grupos, e não pela nova droga.

Mas, em vez disso, suponha que 91 em cada cem pacientes recebendo a nova medicação mostrem uma acentuada melhora, em comparação com 49 em cem do grupo de controle. Ainda é possível que esse resultado impressionante não esteja relacionado com a nova droga; os pacientes no grupo de tratamento podem ser particularmente afortunados ou ter uma capacidade de recuperação rápida. *Mas essa é agora uma explicação bem menos provável.* Na linguagem formal da inferência estatística, pesquisadores provavelmente concluiriam o seguinte: (1) se a droga experimental não tem efeito, raramente veríamos uma variação de resultados dessa dimensão entre aqueles que recebem a droga e aqueles que tomam placebo. (2) Portanto, é muito improvável que a droga não tenha efeito positivo. (3) A explicação alternativa – e mais provável – para o padrão de dados observados é que a droga experimental tenha efeito positivo.

A inferência estatística é o processo pelo qual os dados falam conosco, possibilitando-nos tirar conclusões significativas. Essa é a recompensa! O foco da estatística não é fazer uma miríade de cálculos matemáticos rigorosos; o foco é adquirir compreensão de fenômenos sociais significativos. A inferência estatística é na realidade o casamento de dois conceitos que já discutimos: dados e probabilidade (com uma pequena ajuda do teorema do limite central). Neste capítulo, tomei um importante atalho metodológico. Todos os exemplos partirão do pressuposto de que estamos trabalhando com amostras grandes, adequadamente extraídas. Essa premissa significa que o teorema do limite central se aplica, e que a média e o desvio padrão para qualquer amostra serão aproximadamente os mesmos que a média e o desvio padrão para a população da qual a amostra é retirada. Ambas as coisas facilitam os nossos cálculos.

A inferência estatística não depende dessa premissa simplificadora, mas os diversos artifícios metodológicos para lidar com amostras pequenas ou dados imperfeitos muitas vezes atrapalham a compreensão do quadro maior. O propósito aqui é apresentar o poder da inferência estatística e explicar como ela funciona. Uma vez entendido isso, é fácil aprofundar a complexidade.

UMA DAS FERRAMENTAS mais comuns da inferência estatística é o teste de hipóteses. Na verdade, já introduzi esse conceito – só que sem a terminologia rebuscada. Conforme observado anteriormente, a estatística sozinha não pode *provar* nada; em vez disso, usamos a inferência estatística para aceitar ou rejeitar explicações com base na sua relativa probabilidade. Para ser mais preciso, qualquer inferência estatística começa com uma hipótese nula implícita ou explícita. Essa é a nossa premissa de partida, que será rejeitada ou não com base em análise estatística subsequente. Se rejeitamos a hipótese nula, então geralmente aceitamos alguma hipótese alternativa que seja mais consistente com os dados observados. Por exemplo, num tribunal a premissa de partida, ou hipótese nula, é que o réu é inocente. A tarefa da promotoria é persuadir o juiz ou o júri a rejeitar essa premissa e aceitar a hipótese alternativa, ou seja, que o réu é culpado. Como questão de lógica, a hipótese alternativa é uma conclusão que precisa ser verdade se é para rejeitar a hipótese nula. Consideremos alguns exemplos:

Hipótese nula: essa nova droga experimental não é mais efetiva em prevenir a malária do que um placebo.

Hipótese alternativa: essa nova droga experimental pode ajudar a prevenir a malária.

Os dados: um grupo escolhido aleatoriamente recebe a nova droga experimental e um grupo de controle recebe um placebo. No final de certo período de tempo, o grupo que recebe a droga experimental tem muito menos casos de malária que o grupo de controle. Esse seria um resultado extremamente improvável se a droga experimental não tivesse impacto medicinal. Como resultado, *rejeitamos* a hipótese nula de que a nova droga não tem impacto (além do de um placebo) e aceitamos a alternativa lógica, que é a nossa hipótese alternativa. Essa nova droga experimental pode ajudar a prevenir a malária.

Essa abordagem metodológica é estranha o bastante para justificar mais um exemplo. Outra vez, note que a hipótese nula e a hipótese alternativa são complementos lógicos. Se uma é verdadeira, a outra não é. Ou, se rejeitamos uma afirmação, devemos aceitar a outra.

Hipótese nula: tratamento para abuso de substâncias químicas para detentos não reduz sua taxa de reincidência após deixarem a prisão.

Hipótese alternativa: tratamento para abuso de substâncias químicas para detentos reduzirá

sua probabilidade de reincidência depois de soltos.

Os dados (hipotéticos): detentos foram aleatoriamente divididos em dois grupos; o grupo de "tratamento" recebe tratamento para abuso de substâncias e o grupo de controle não recebe. (Trata-se de uma dessas ocasiões bacanas em que o grupo de tratamento realmente recebe tratamento!) Após cinco anos, ambos os grupos têm índices similares de reincidência. Nesse caso, *não podemos rejeitar* a hipótese nula.[a] Os dados não nos deram razão para descartar a nossa premissa inicial de que o tratamento para abuso de substâncias químicas não é uma ferramenta efetiva para impedir ex-infratores de voltar à prisão.

Pode parecer contraintuitivo, mas pesquisadores muitas vezes criam uma hipótese nula na esperança de poder rejeitá-la. Em ambos os exemplos anteriores, o "sucesso" da pesquisa (achar uma nova droga para a malária ou reduzir a reincidência de prisão) envolvia rejeitar a hipótese nula. Os dados tornaram isso possível em apenas um dos casos (a droga para malária).

NUM TRIBUNAL, o limiar para rejeitar uma premissa de inocência é a avaliação qualitativa de o réu ser "culpado além de uma dúvida razoável". Cabe ao juiz ou ao júri definir o que exatamente isso significa. A estatística abriga a mesma ideia básica, mas "culpado além de uma dúvida razoável" é definido quantitativamente. Os pesquisadores em geral perguntam: se a hipótese nula for verdadeira, qual é a probabilidade de observar esse padrão de dados por puro acaso? Usando um exemplo familiar, pesquisadores na área médica podem indagar: se essa droga experimental não tem efeito sobre doenças cardíacas (nossa hipótese nula), qual é a probabilidade de 91 pacientes em cem que tomam a droga mostrarem melhora em comparação com apenas 49 em cem pacientes tomando placebo? Se os dados sugerem que a hipótese nula é extremamente improvável – como no exemplo médico –, então devemos rejeitá-la e aceitar a hipótese alternativa (de que a droga é efetiva no tratamento de doenças cardíacas).

Nessa vertente, vamos revisitar o escândalo de fraude padronizada em Atlanta mencionado em diversos pontos do livro. Os resultados dos testes em Atlanta chamaram atenção primeiro pela alta quantidade de respostas com rasuras "errado para certo". Obviamente, estudantes rasuram respostas o tempo todo durante esse tipo de exame. E alguns grupos de alunos podem ter sido particularmente sortudos em suas mudanças, sem que houvesse necessariamente qualquer fraude envolvida. Por esse motivo, a hipótese nula é que os resultados dos testes padronizados para qualquer distrito escolar são legítimos e que quaisquer padrões irregulares de rasuras são meramente produto do acaso. Sem dúvida não queremos punir alunos ou administradores porque uma proporção inusitadamente alta de alunos resolveu fazer mudanças sensatas em suas folhas de respostas nos minutos finais de um importante exame estadual.

Mas "inusitadamente alta" não chega nem perto de descrever o que aconteceu em Atlanta. Algumas classes tinham folhas de respostas nas quais a quantidade de rasuras errado-para-certo representavam de vinte a cinquenta desvios padrões acima da norma estadual. (Para pôr isso em perspectiva, lembre-se de que a maioria das observações numa distribuição geralmente cai dentro de dois desvios padrões em relação à média.) Então, qual a probabilidade de que os estudantes de Atlanta tenham apagado quantidades maciças de respostas erradas e as substituído por respostas certas por uma simples questão de acaso? O funcionário que analisou os dados descreveu a probabilidade de ocorrência do padrão de Atlanta sem fraude como

aproximadamente igual à chance de ter 70 mil pessoas comparecendo a um jogo de futebol americano no Georgia Dome sendo que todas têm mais de dois metros de altura.[2] Isso pode acontecer? Sim. É provável? Nem tanto.

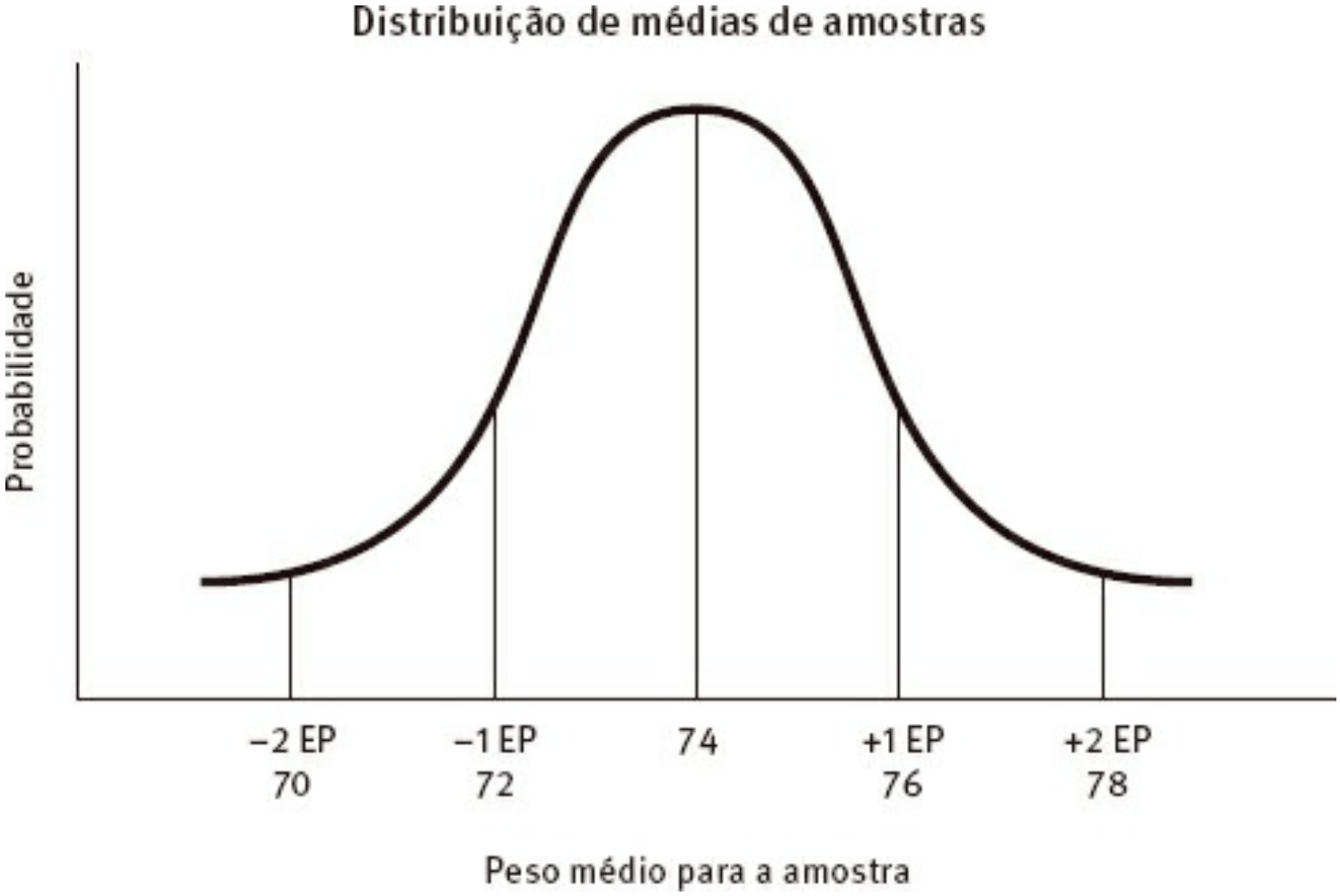
Os funcionários da Geórgia ainda não podiam condenar ninguém por contravenção, da mesma forma que meu professor não pôde (e não devia) me expulsar da escola porque a nota do meu exame final em estatística estava fora de sincronia com a nota do primeiro semestre. *Os funcionários de Atlanta não podiam provar que estava havendo uma fraude*. Podiam, porém, rejeitar a hipótese nula de que os resultados eram legítimos. E podiam fazê-lo com um "alto grau de confiança", o que significa que o padrão observado era quase impossível entre alunos normais fazendo um teste. Portanto, aceitaram explicitamente a hipótese alternativa, a de que estava ocorrendo alguma falcatrua. (Imagino, no entanto, que eles tenham empregado um termo mais oficial.) Investigações subsequentes de fato revelaram os "rasuradores fantasmas". Houve relatos de professores mudando respostas, divulgando respostas, permitindo a alunos de baixo desempenho copiar de alunos de alto desempenho e até mesmo apontando respostas quando parados junto às carteiras dos alunos. A fraude mais escandalosa envolvia um grupo de professores que organizou um encontro animado com muita pizza no fim de semana durante o qual repassaram as folhas de exame e mudaram as respostas dos alunos.

No exemplo de Atlanta, pudemos rejeitar a hipótese nula de "não fraude" porque o padrão dos resultados dos testes era absurdamente improvável na ausência de trapaças. Mas o quanto a hipótese nula deve ser implausível para podermos rejeitá-la e recorrer a alguma explicação alternativa?

Um dos limiares mais comuns utilizados por pesquisadores para rejeitar uma hipótese nula é 5%, geralmente escrito em forma decimal: 0,05. Essa probabilidade é conhecida como nível de significância e representa o limite superior para a probabilidade de observação de algum padrão de dados se a hipótese nula fosse verdadeira. Acompanhe meu raciocínio por um momento, porque na realidade não é tão complicado assim.

Pensemos sobre um nível de significância de 0,05. Podemos rejeitar uma hipótese nula no nível 0,05 se houver uma chance menor do que 5% de obter um resultado no mínimo tão extremo quanto o que observamos se a hipótese nula fosse verdadeira. Um exemplo simples pode deixar isso muito mais claro. Detesto ter que fazer isso com você, porém imagine mais uma vez que você foi encarregado de achar um ônibus perdido (em parte por causa dos seus valentes esforços no capítulo anterior). Só que agora você está trabalhando em período integral para os pesquisadores do estudo Changing Lives, e eles lhe deram alguns dados excelentes para ajudar nas informações para o seu trabalho. Cada ônibus operado pelos organizadores do estudo tem aproximadamente sessenta passageiros, então podemos tratar os passageiros de qualquer ônibus como uma amostra aleatória tirada da população total do Changing Lives. Você é despertado de manhã cedo pela notícia de que um ônibus na área de Boston foi sequestrado por um grupo terrorista pró-obesidade.[b] Sua missão é descer de um helicóptero sobre o teto do ônibus em movimento, esgueirar-se para dentro pela saída de emergência e aí determinar furtivamente se os passageiros são os participantes do estudo Changing Lives, apenas baseado nos seus pesos. (Falando sério, não é mais implausível que as tramas dos filmes de ação, sendo muito mais educativo.)

Quando o helicóptero decola da base de comando, você recebe uma metralhadora, várias granadas, um relógio que também funciona como câmera de alta resolução e os dados que calculamos no capítulo anterior sobre o peso médio e o erro padrão para amostras tiradas dos participantes do Changing Lives. Qualquer amostra aleatória de sessenta participantes terá um peso médio esperado de 74 quilos e um desvio padrão de dezesseis quilos, uma vez que estes são a média e o desvio padrão para todos os participantes do estudo (a população). Com esses dados, podemos calcular o erro padrão para a média da amostra: $s/\sqrt{n} = 16/\sqrt{60} = 16/7{,}75 = 2{,}1$. No controle da missão, a seguinte distribuição é escaneada para dentro da sua retina direita, de modo que você possa consultá-la depois de penetrar no ônibus em movimento e pesar secretamente todos os passageiros.




Como mostra a distribuição acima, devemos esperar que aproximadamente 95% de todas as amostras de sessenta pessoas tiradas dos participantes do Changing Lives tenham peso médio dentro de dois erros padrões em relação à média da população, ou aproximadamente entre setenta quilos e 78 quilos.[c] Inversamente, apenas cinco vezes em cem uma amostra de sessenta pessoas escolhidas aleatoriamente entre os participantes do Changing Lives teria uma peso médio maior que 78 quilos e menor que setenta quilos. (Você está conduzindo o que é conhecido como teste de hipótese de "duas caudas" – ou "bicaudal"; a diferença entre este e um teste de hipótese de "uma cauda" – ou "unicaudal" – será coberta no apêndice no fim deste capítulo.) Os seus orientadores na força-tarefa de contraterrorismo decidiram que 0,05 é o nível de significância

para a sua missão. Se o peso médio dos sessenta passageiros no ônibus sequestrado for acima de 78 ou abaixo de setenta, então você rejeitará a hipótese nula de que o ônibus contém participantes do Changing Lives, aceitará a hipótese alternativa de que o ônibus contém sessenta pessoas que se dirigem para outro lugar e aguardará novas ordens.

Você tem sucesso em pousar e entrar no ônibus em movimento e secretamente pesar todos os passageiros. O peso médio para essa amostra de sessenta pessoas é 62 quilos, o que cai a mais de dois erros padrões abaixo da média. (Outra pista importante é que todos os passageiros são crianças vestindo camisetas do "Acampamento de Hóquei Glendale".)

Pelas instruções da sua missão, você pode rejeitar a hipótese nula de que aquele ônibus contém uma amostra aleatória de sessenta participantes do estudo Changing Lives a um nível de significância 0,05. Isso significa que (1) o peso médio no ônibus cai numa faixa que esperaríamos observar apenas cinco vezes em cem se a hipótese nula fosse verdadeira e aquele fosse realmente um ônibus cheio de passageiros do Changing Lives; (2) você pode rejeitar a hipótese nula no nível de significância 0,05; e (3) em média, 95 vezes em cem você terá rejeitado corretamente a hipótese nula, e cinco vezes em cem você estará errado, o que no caso significa que você concluiu que aquele *não* é o ônibus do Changing Lives, quando na verdade é. Essa amostra do pessoal do Changing Lives simplesmente acontece de ter um peso médio particularmente alto ou baixo em relação à média geral dos participantes do estudo.

A missão ainda não acabou. Sua superiora no controle da missão (papel desempenhado por Angelina Jolie na versão cinematográfica deste exemplo) lhe pede para calcular o valor-p para o seu resultado. O valor-p é a probabilidade específica de obter um resultado no mínimo tão extremo quanto o que você observou se a hipótese nula for verdadeira. O peso médio dos passageiros desse ônibus é 62 quilos, o que corresponde a 5,7 erros padrões abaixo da média dos participantes do estudo. A probabilidade de se obter um resultado pelo menos tão extremo se essa fosse realmente uma amostra de participantes do Changing Lives é de menos de 0,0001. (Num documento de pesquisa, isso seria registrado como $p<0,0001$.) Completada sua missão, você salta do ônibus em movimento e pousa a salvo no assento do passageiro de um conversível passando pela pista adjacente.

[Esta história também tem um final feliz. Quando os terroristas próobesidade ficam sabendo mais sobre o Festival Internacional da Salsicha na cidade, concordam em abandonar a violência e trabalhar pacificamente para promover a obesidade expandindo e divulgando festivais de salsicha ao redor do mundo.]

SE O NÍVEL DE SIGNIFICÂNCIA de 0,05 parece um tanto arbitrário, é porque ele de fato é. Não existe um único limiar estatístico padronizado para rejeitar uma hipótese nula. Tanto 0,01 como 0,1 também são limiares razoavelmente comuns para fazer o tipo de análise descrito acima.

Obviamente, rejeitar uma hipótese nula no nível 0,01 (o que significa que há menos de uma chance em cem de observar um resultado nessa faixa se a hipótese nula fosse verdadeira) carrega mais peso estatístico do que rejeitar a hipótese nula no nível 0,1 (o que significa que há menos de uma chance em dez de observar esse resultado se a hipótese nula fosse verdadeira). Os prós e contras dos diferentes níveis de significância serão discutidos mais adiante neste capítulo. Por enquanto, o importante é que, quando podemos rejeitar uma hipótese nula com um nível de

significância razoável, os resultados são ditos "estatisticamente significativos".

Eis o que isso significa na vida real. Quando você lê no jornal que as pessoas que comem vinte bolinhos de farelo de trigo por dia têm taxas menores de câncer de cólon do que pessoas que não comem quantidades prodigiosas de farelo de trigo, a pesquisa acadêmica subjacente provavelmente verificou algo do tipo: (1) em algum grande conjunto de dados, os pesquisadores determinaram que indivíduos que comiam pelo menos vinte bolinhos de farelo de trigo por dia tinham uma incidência menor de câncer de cólon do que indivíduos que não relatavam comer tanto farelo. (2) A hipótese nula dos pesquisadores foi que comer bolinhos de farelo de trigo não tem impacto no câncer de cólon. (3) A disparidade dos resultados de câncer de cólon entre aqueles que comiam montes de farelo e aqueles que não comiam não podia ser explicada facilmente pelo puro acaso. Mais especificamente, se comer bolinhos de farelo de trigo não tem real ligação com câncer de cólon, a probabilidade de se ter uma diferença tão grande na incidência de câncer entre comedores e não comedores de farelo de trigo por mero acaso é inferior a algum limiar, tal como 0,05. (Esse limiar deve ser estabelecido pelos pesquisadores *antes* de fazerem sua análise estatística para evitar a escolha posterior de um limiar conveniente para fazer com que os resultados pareçam significativos.) (4) O artigo acadêmico provavelmente contém uma conclusão dizendo algo nesta linha: "Encontramos uma ligação estatisticamente significativa entre o consumo diário de vinte ou mais bolinhos de farelo de trigo e uma redução na incidência de câncer de cólon. Esses resultados são significativos no nível 0,05."

Quando mais tarde eu ler a respeito do estudo no *Chicago Sun-Times* enquanto tomo meu café da manhã de ovos com bacon, a manchete provavelmente será mais direta e interessante: "Vinte bolinhos de farelo de trigo por dia ajudam a prevenir o câncer de cólon." No entanto, a manchete do jornal, embora muito mais interessante que o artigo acadêmico, poderá também apresentar uma séria inacurácia. O estudo na realidade não alega que comer bolinhos de trigo reduz o risco de o indivíduo ter câncer de cólon; apenas mostra uma correlação negativa entre o consumo desses bolinhos e a incidência de câncer de cólon num grande conjunto de dados. Essa associação estatística não é suficiente para provar que tais bolinhos *causam* a melhora no resultado da saúde. Afinal, o tipo de pessoa que come bolinhos de farelo de trigo (especialmente vinte por dia!) pode fazer várias outras coisas que reduzem o risco de câncer, tais como comer menos carne vermelha, exercitar-se regularmente, fazer exames regulares para detectar câncer, e assim por diante. (Esse é o "viés do usuário saudável" do Capítulo 7.) Será que podemos atribuir esses resultados à ação dos bolinhos de farelo ou a outros comportamentos ou atributos pessoais compartilhados por pessoas que comem uma porção de bolinhos de farelo de trigo por dia? Essa distinção entre correlação e causalidade é crucial para uma interpretação adequada dos resultados estatísticos. Revisitaremos mais adiante no livro essa ideia de que "correlação não equivale a causalidade".

Devo também ressaltar que a significância estatística não diz nada a respeito do *tamanho* dessa associação. Pessoas que comem montes de bolinhos de farelo de trigo podem ter uma incidência menor de câncer de cólon, mas quanto menor? A diferença nos índices de câncer de cólon para comedores e não comedores de farelo de trigo pode ser trivial; a constatação de uma significância estatística significa apenas que o efeito observado, por menor que seja, não é provável de ocorrer por coincidência. Suponha que você se depare com um estudo bem planejado que descobriu uma relação positiva estatisticamente significativa entre comer uma

banana antes dos exames escolares e obter uma nota mais alta na parte de matemática do exame. Uma das primeiras perguntas que você deseja fazer é: qual é o tamanho desse efeito? Poderia facilmente ser 0,9 ponto; num teste com um escore médio de quinhentos, esse número não muda a vida de ninguém. No Capítulo 11, voltaremos a essa distinção crucial entre *tamanho* e *significância* quando se trata de interpretar resultados estatísticos.

Nesse meio-tempo, a descoberta de que "não há associação estatisticamente significativa" entre duas variáveis quer dizer que qualquer relação entre as duas variáveis pode ser razoavelmente explicada apenas pelo acaso. O *New York Times* publicou uma denúncia bombástica sobre empresas de tecnologia oferecendo a preço baixo programas que elas alegam melhorar o desempenho de alunos, quando os dados sugerem outra coisa.[3] Segundo o artigo, a Carnegie Mellon University vende um programa chamado Cognitive Tutor com esta temerária alegação: "Currículos matemáticos revolucionários. Resultados revolucionários." Todavia, uma avaliação do Cognitive Tutor conduzida pelo Departamento de Educação dos Estados Unidos concluiu que o produto "não tem efeitos discerníveis" nos resultados dos exames de alunos do ensino médio. (O *Times* sugere que a campanha de marketing apropriada deveria ser "Currículos matemáticos indistintos. Resultados não provados".) Na verdade, um estudo de dez produtos de softwares programados para melhorar o domínio de matérias como matemática ou leitura descobriu que nove deles "não têm efeitos estatisticamente significativos nos resultados dos exames". Em outras palavras, pesquisadores federais não podem descartar o mero acaso como causa de qualquer variação no desempenho de estudantes que usam esses produtos e estudantes que não usam.

DEIXE-ME FAZER uma pausa para lembrar a você por que tudo isso tem importância. Uma matéria no *Wall Street Journal* em maio de 2011 trazia a manchete: "Descoberta relação entre autismo e tamanho do cérebro". Essa é uma descoberta importante, pois as causas do transtorno do espectro autista permanecem vagas. A primeira frase da matéria do *Wall Street Journal*, que resumia um artigo publicado na revista *Archives of General Psychiatry*, reporta: "Crianças com autismo têm cérebros maiores do que crianças sem o distúrbio, e o crescimento parece ocorrer antes dos dois anos de idade, segundo um novo estudo divulgado na segunda-feira."[4] Com base em exames de imagem do cérebro conduzidos em 59 crianças com autismo e 38 crianças sem autismo, pesquisadores na Universidade da Carolina do Norte reportaram que crianças com autismo têm cérebros que são até 10% maiores que os das crianças da mesma idade sem autismo.

Eis a questão médica relevante: existe uma diferença fisiológica nos cérebros de crianças pequenas que têm o espectro do autismo? Se sim, essa descoberta pode levar a uma melhor compreensão do que causa o distúrbio e como ele pode ser tratado ou prevenido.

E eis a questão estatística relevante: os pesquisadores podem fazer inferências abrangentes sobre o espectro do autismo em geral que estejam baseadas num estudo de um grupo aparentemente pequeno de crianças com autismo (59) e um grupo de controle ainda menor (38) – meros 97 sujeitos ao todo? A resposta é sim. Os pesquisadores concluíram que a probabilidade de observar as diferenças no tamanho total do cérebro que descobriram em suas duas amostras seria meramente de duas chances em mil ($p = 0,002$) se de fato não houvesse diferença real no

tamanho do cérebro entre crianças com e sem autismo na população geral.

Fui atrás do estudo original na *Archives of General Psychiatry*.[5] Os métodos usados por esses pesquisadores não são mais sofisticados que os conceitos que cobrimos até aqui. Vou lhe proporcionar um rápido passeio pelos fundamentos desse resultado social e estatisticamente significativo. Primeiro, você deve reconhecer que cada grupo de crianças, as 59 com autismo e as 38 sem autismo, constitui uma amostra razoavelmente grande extraída das respectivas populações – crianças com e sem autismo. As amostras são grandes o bastante para que se aplique o teorema do limite central. Se você já tentou bloquear o conteúdo do último capítulo na sua cabeça, vou lembrá-lo do que diz o teorema do limite central: (1) as médias de amostras para qualquer população estarão distribuídas aproximadamente numa distribuição normal em torno da média real da população; (2) devemos esperar que a média e o desvio padrão da amostra sejam aproximadamente iguais à média e ao desvio padrão da população de onde a amostra é retirada; e (3) aproximadamente 68% das médias das amostras se situam dentro de um erro padrão em relação à média da população, aproximadamente 95%, dentro de dois erros padrões em relação à média da população, e assim por diante.

Numa linguagem menos técnica, tudo isso quer dizer que qualquer amostra deve se parecer bastante com a população da qual é tirada; embora cada amostra seja diferente, seria relativamente raro que a média de uma amostra adequadamente retirada se desvie bastante da média para a população relevante subjacente. Similarmente, seria de esperar também que duas amostras tiradas da mesma população se parecessem bastante entre si. Ou, pensando na situação de modo um pouco diferente, se temos duas amostras com médias extremamente desiguais, a explicação mais provável é que venham de populações diferentes.

Eis aqui um rápido exemplo intuitivo. Suponha que a sua hipótese nula seja que jogadores profissionais de basquete masculino tenham a mesma altura média que o resto da população masculina adulta. Você seleciona ao acaso uma amostra de cinquenta jogadores de basquete profissionais e uma amostra de cinquenta homens que não jogam basquete profissional. Suponha que a altura média da sua amostra de jogadores seja de 1,98 metro e a altura média dos não jogadores seja de 1,75 metro (uma diferença de 23 centímetros). Qual é a probabilidade de observar uma diferença tão grande na altura média entre as duas amostras se de fato não houver diferença na altura média entre jogadores de basquete profissionais e todos os outros homens na população geral? A resposta não técnica: muito, muito, muito baixa.[d]

O artigo da pesquisa sobre autismo tem a mesma metodologia básica. O artigo compara diversas medições de tamanhos de cérebro entre as amostras de crianças. (As medições do cérebro foram feitas com imagens de ressonância magnética aos dois anos e mais uma vez entre os quatro e cinco anos.) Vou me ater apenas a uma medição, o volume total do cérebro. A hipótese nula dos pesquisadores presumivelmente foi que não há diferenças anatômicas nos cérebros de crianças com autismo e sem autismo. A hipótese alternativa é que os cérebros de crianças com transtorno do espectro autista são fundamentalmente diferentes. Tal descoberta ainda deixaria uma porção de perguntas, mas apontaria uma direção para futuras investigações.

Nesse estudo, as crianças com transtorno do espectro autista tinham um volume cerebral médio de 1.310,4 centímetros cúbicos; enquanto as crianças no grupo de controle tinham um volume cerebral médio de 1.238,8 centímetros cúbicos. Portanto, a diferença no volume cerebral

médio entre os dois grupos é de 71,6 centímetros cúbicos. Qual é a probabilidade desse resultado se de fato não houver diferença no tamanho médio do cérebro na população em geral entre crianças que têm transtorno do espectro autista e crianças que não têm?

Você deve se lembrar do último capítulo que podemos criar um erro padrão para cada uma das nossas amostras: $s/\sqrt{n}$, onde *s* é o desvio padrão da amostra e n é o número de observações. O artigo da pesquisa nos fornece esses números. O erro padrão para o volume cerebral total das 59 crianças na amostra com transtorno do espectro autista é treze centímetros cúbicos; o erro padrão para o volume cerebral total na amostra de 38 crianças do grupo de controle é dezoito centímetros cúbicos. Você se lembrará de que o teorema do limite central nos diz que, para 95 amostras em cem, a média da amostra vai se situar dentro de dois erros padrões da média real da população, num ou noutro sentido.

Como resultado, podemos inferir a partir da nossa amostra que 95 vezes em cem o intervalo de 1.310,4 centímetros cúbicos ± 26 (que corresponde a dois erros padrões) conterá o volume cerebral médio para *todas* as crianças com transtorno do espectro autista. Essa expressão é chamada de intervalo de confiança. Podemos dizer com 95% de confiança que a faixa de 1.284,4 até 1.336,4 centímetros cúbicos contém o volume cerebral total médio para crianças na população geral com transtorno do espectro autista.

Usando a mesma metodologia, podemos dizer com 95% de confiança que o intervalo de 1.238,8 ± 36, ou entre 1.202,8 e 1.274,8 centímetros cúbicos, incluirá o volume cerebral médio para crianças na população geral que não tem transtorno do espectro autista.

Sim, aqui há um bocado de números. Talvez você tenha acabado de arremessar o livro para o outro lado da sala.[e] Se não o fez, ou se foi pegar o livro de volta, o que você deve notar é que os nossos intervalos de confiança não se sobrepõem. O *limite inferior* do nosso intervalo de confiança de 95% para o tamanho cerebral médio de crianças com autismo na população geral (1.284,4 centímetros cúbicos) ainda é mais alto que o *limite superior* para o intervalo de confiança de 95% para o tamanho cerebral médio de crianças pequenas sem autismo na população (1.274,8 centímetros cúbicos), como ilustra o diagrama a seguir.


Essa é a primeira pista de que pode haver uma diferença anatômica subjacente nos cérebros de crianças pequenas com transtorno do espectro autista. Contudo, é apenas uma pista. Todas essas inferências baseiam-se em dados referentes a menos de cem crianças. Talvez tenhamos somente amostras excêntricas.

Um último procedimento estatístico pode materializar tudo isso. Se estatística fosse um evento olímpico como patinação artística, esta seria a apresentação final, após a qual os fãs eufóricos jogam buquês de flores sobre o gelo. Podemos calcular a probabilidade exata de observar uma diferença de médias no mínimo tão grande (1.310,4 versus 1.238,8 centímetros cúbicos) se realmente não houver diferença no tamanho do cérebro entre crianças com transtorno do espectro autista e todas as outras na população geral. Podemos achar um valor-p para a diferença observada entre as médias.

Para evitar que você volte a jogar o livro através da sala, pus a fórmula no apêndice deste capítulo. A intuição é simples e direta. Se pegarmos duas amostras grandes da mesma população, seria de esperar que tenham médias bastante similares. Na verdade, o nosso melhor palpite é que tenham médias idênticas. Por exemplo, se fôssemos selecionar cem jogadores da NBA e eles tivessem uma altura média de 1,98 metro, então eu esperaria que outra amostra aleatória de cem jogadores da NBA tivesse uma média próxima a 1,98 metro. Ok, talvez as duas amostras tivessem uma diferença de quatro ou cinco centímetros. Mas é menos provável que as médias das duas amostras tenham uma diferença de dez centímetros – e menos provável ainda que tenham uma diferença de quinze ou vinte centímetros. Acontece que podemos calcular um erro padrão para a diferença entre as médias das duas amostras; esse erro padrão nos dá uma medida da dispersão que podemos esperar, em média, quando subtraímos a média de uma amostra da média da outra. (Como eu disse antes, a fórmula está no apêndice do capítulo.) O importante é que podemos usar esse erro padrão para calcular a probabilidade de que ambas as amostras provenham da mesma população. Eis como funciona:

1. Se duas amostras são tiradas da mesma população, o nosso melhor palpite é que a diferença entre suas médias seja zero.
2. O teorema do limite central nos diz que, em amostras repetidas, a *diferença entre duas médias* estará distribuída aproximadamente como uma distribuição normal. (Cá entre nós, você já está adorando o teorema do limite central ou não?)
3. Se as duas amostras realmente provêm da mesma população, então, em cerca de 68 casos em cem, a diferença entre as médias das duas amostras estará dentro de um erro padrão de zero. E, em cerca de 95 casos em cem, a diferença entre as médias das duas amostras estará dentro de dois erros padrões de zero. E em 99,7 casos em cem, a diferença estará dentro de três erros padrões de zero – que acaba sendo o que motiva a conclusão do artigo de pesquisa sobre autismo com o qual começamos.

Conforme observado anteriormente, a diferença no tamanho cerebral médio entre a amostra de crianças com transtorno do espectro autista e o grupo de controle é de 71,6 centímetros cúbicos. O erro padrão para esta diferença é 22,7, o que significa que a diferença entre as médias das duas amostras é maior do que três erros padrões a partir de zero; um resultado tão (ou mais) extremo seria esperado apenas duas vezes em mil se essas amostras são tiradas de uma população idêntica.

No artigo publicado na revista *Archives of General Psychiatry*, os autores reportam um valor-p de 0,002, conforme mencionei antes. Agora você sabe de onde surgiu esse valor!

APESAR DE TODAS as maravilhas da inferência estatística, existem também algumas armadilhas significativas. Elas derivam do exemplo que introduziu este capítulo: meu desconfiado professor de estatística. O poderoso processo de inferência estatística baseia-se na probabilidade, não em algum tipo de certeza cósmica. Não queremos mandar gente para a cadeia pelo equivalente a tirar dois *royal flushes*[f] seguidos; isso *pode* acontecer, mesmo que a pessoa não esteja trapaceando. Como resultado, temos um dilema fundamental quando se trata de qualquer tipo de teste de hipótese.

Essa realidade estatística ganhou repercussão em 2011 quando o *Journal of Personality and Social Psychology* preparava-se para publicar um artigo acadêmico que, à primeira vista, era parecido com milhares de outros artigos acadêmicos.[6] Um professor de Cornell propôs explicitamente uma hipótese nula, conduziu um experimento para testá-la e então rejeitou-a a um nível de significância de 0,05 com base nos resultados experimentais. O resultado causou um grande alvoroço, tanto em círculos científicos como nos principais veículos de mídia como o *New York Times*.

Basta dizer que artigos no *Journal of Personality and Social Psychology* em geral não atraem grandes manchetes jornalísticas. O que exatamente tornou esse estudo tão controverso? O pesquisador em questão estava testando a capacidade humana de exercitar percepção extrassensorial, ou PES. A hipótese nula era que a PES não existe; a hipótese alternativa era que seres humanos têm sim poderes extrassensoriais. Para estudar essa questão, o pesquisador recrutou uma grande amostra de participantes para examinar duas "cortinas" postadas numa tela de computador. Um programa colocava aleatoriamente uma foto erótica atrás de uma ou outra cortina. Em tentativas repetidas, os participantes do estudo foram capazes de escolher a cortina com a foto erótica 53% das vezes, enquanto a probabilidade diz que isso aconteceria apenas 50% das vezes. Por causa do grande tamanho da amostra, o pesquisador pôde rejeitar a hipótese nula de que a percepção extrassensorial não existe e, em vez disso, aceitar a hipótese alternativa de que a percepção extrassensorial pode possibilitar às pessoas pressentir eventos futuros. A decisão de publicar o artigo foi amplamente criticada com argumentos de que um único evento estatisticamente significativo pode com facilidade ser produto do acaso, sobretudo quando não há nenhuma outra evidência para corroborar ou mesmo explicar o achado. O *New York Times* sintetizou as críticas: "Alegações que desafiam quase toda lei da ciência são por definição extraordinárias e, portanto, requerem evidências extraordinárias. A negligência de levar isso em consideração – como análises convencionais em ciências sociais fazem – faz com que muitos achados pareçam bem mais significativos do que realmente são."

Uma resposta para esse tipo de absurdo poderia ser estabelecer um limiar mais rigoroso para definir a significância estatística, tal como 0,001.[g] Mas isso cria um problema em si. Escolher o nível apropriado de significância estatística envolve uma inerente escolha e suas consequências.

Se o nosso ônus de prova para rejeitar a hipótese nula for baixo demais (por exemplo, 0,1), vamos nos perceber rejeitando periodicamente a hipótese nula quando de fato ela é verdadeira (como eu desconfio ter sido o caso no estudo de PES). Em jargão estatístico, isso é conhecido como erro Tipo I. Considere o exemplo de um tribunal americano, onde a hipótese nula é que o réu não é culpado e o limiar para rejeitar a hipótese nula é "culpado além de uma dúvida razoável". Suponha que relaxemos esse limiar para algo como "um forte palpite de que o sujeito

fez aquilo". Isso irá assegurar que mais criminosos acabem indo para a cadeia – e também mais pessoas inocentes. Num contexto estatístico, equivale a ter um nível de significância relativamente baixo, como 0,1.

Bem, uma chance em dez não é algo extremamente improvável. Considere esse desafio no contexto de aprovar uma nova droga para o câncer. Para cada dez drogas que aprovamos com esse ônus de prova estatística relativamente baixo, uma delas não funciona realmente e mostrou resultados promissores nos testes apenas por acaso. (Ou, no exemplo do tribunal, para cada dez réus considerados culpados, um deles era na realidade inocente.) Um erro Tipo I envolve rejeitar equivocadamente uma hipótese nula. Embora a terminologia seja um tanto contraintuitiva, isso também é conhecido como "falso positivo". Eis um meio de conciliar o jargão: quando você vai ao médico e faz exames para detectar alguma doença, a hipótese nula é de que você não tenha a doença. Se os resultados do laboratório podem ser usados para rejeitar a hipótese nula, diz-se que você testou positivo. E se você testou positivo e na realidade não está doente, então é um falso positivo.

Em todo caso, quanto menor o ônus estatístico para rejeitar a hipótese nula, mais provável que a rejeição aconteça. Obviamente, preferiríamos não aprovar drogas para o câncer ineficazes nem mandar réus inocentes para a cadeia.

Mas aqui há uma tensão. Quanto mais alto o limiar para rejeitar a hipótese nula, mais provável é que fracassemos em rejeitar uma hipótese nula que deveria ser rejeitada. Se exigirmos cinco testemunhas oculares para condenar todo réu criminoso, então uma porção de réus culpados será erroneamente solta. (É claro que menos inocentes irão para a cadeia.) Se adotarmos o nível de significância 0,001 nos testes clínicos para todas as novas drogas para câncer, então de fato minimizaremos a aprovação de drogas ineficazes. (Há apenas uma chance em mil de rejeitar erradamente a hipótese nula de que a droga seja mais efetiva que um placebo.) Todavia introduzimos o risco de não aprovar muitas drogas efetivas porque colocamos o sarrafo da aprovação muito alto. Esse é conhecido como erro Tipo II, ou falso negativo.

Que tipo de erro é pior? Depende das circunstâncias. A questão mais importante é que você reconheça a escolha e as consequências. Não existe "almoço grátis" em estatística. Considere as seguintes situações não estatísticas, todas elas envolvendo uma escolha entre erros Tipo I e Tipo II.

1. Filtros de spam. A hipótese nula é que qualquer mensagem de e-mail específica *não* é spam. O seu filtro de spam busca indícios que podem ser usados para rejeitar a hipótese nula para qualquer e-mail específico, tais como enormes listas de distribuição ou expressões do tipo "aumento de pênis". Um erro Tipo I seria excluir uma mensagem que não seja realmente spam (um falso positivo). Um erro Tipo II seria deixar passar pelo filtro um spam para sua caixa de entrada (um falso negativo). Se pesarmos os custos de deixar de receber uma mensagem importante em relação aos custos de receber ocasionais mensagens sobre vitaminas à base de ervas, a maioria das pessoas provavelmente tenderia a permitir erros Tipo II. Um filtro de spam idealmente projetado deveria requerer um grau relativamente alto de certeza antes de rejeitar a hipótese nula de que uma mensagem para você seja legítima e bloqueá-la.
2. Detecção de câncer. Temos numerosos testes para detecção precoce de câncer, tais como

mamografias (para câncer de mama), teste de PSA (câncer de próstata) e até mesmo exames de ressonância magnética computadorizada de corpo inteiro para qualquer coisa que pareça suspeita. A hipótese nula para qualquer um que passe por um exame de detecção é que não haja câncer presente. A premissa sempre tem sido de que um erro Tipo I (um falso positivo que acabe não sendo nada) é muito mais preferível a um erro Tipo II (um falso negativo que deixa de diagnosticar um câncer). Historicamente, a tendência em relação a exames de detecção de câncer tem sido oposta à do exemplo do spam. Médicos e pacientes estão dispostos a tolerar uma quantidade razoável de erros Tipo I (falsos positivos) para evitar a possibilidade de um erro Tipo II (falhar num diagnóstico de câncer). Mais recentemente, os especialistas em políticas de saúde pública começaram a questionar essa visão por causa dos elevados custos e sérios efeitos colaterais associados com falsos positivos.

3. Captura de terroristas. Nem um erro Tipo I nem um erro Tipo II é aceitável nessa situação, e é por isso que a sociedade continua debatendo sobre o equilíbrio apropriado entre combater o terrorismo e proteger as liberdades civis. A hipótese nula é que um indivíduo não é terrorista. Como no contexto do crime comum, não queremos cometer um erro Tipo I e mandar gente inocente para a prisão de Guantánamo. Contudo, num mundo com armas de destruição em massa, deixar livre mesmo um único terrorista (erro do Tipo II) pode ser literalmente catastrófico. É por isso – quer você aprove ou não – que os Estados Unidos mantêm suspeitos de terrorismo na prisão de Guantánamo com base em menos evidências do que seria exigido para condená-los numa corte criminal comum.

A inferência estatística não é mágica nem infalível, mas é uma ferramenta extraordinária para dar sentido ao mundo. Podemos adquirir grande percepção de muitos fenômenos da vida apenas determinando a explicação mais provável. A maioria de nós faz isso o tempo todo (por exemplo, “Penso que o aluno de faculdade desmaiado no chão cercado de latas de cerveja bebeu demais” em vez de “Penso que o aluno de faculdade desmaiado no chão cercado de latas de cerveja foi envenenado por terroristas”).

A inferência estatística apenas formaliza o processo.

# APÊNDICE AO CAPÍTULO 9

## Cálculo do erro padrão para uma diferença de médias

Fórmula para comparar duas médias:

$$\frac{\bar{x} - \bar{y}}{\sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}}}$$

→ o numerador fornece o valor da diferença entre as médias

→ o denominador fornece o erro padrão para uma diferença entre as médias das duas amostras

onde:

$\bar{x}$ = média da amostra $x$

$\bar{y}$ = média da amostra $y$

$s_x$ = desvio padrão para a amostra $x$

$s_y$ = desvio padrão para a amostra $y$

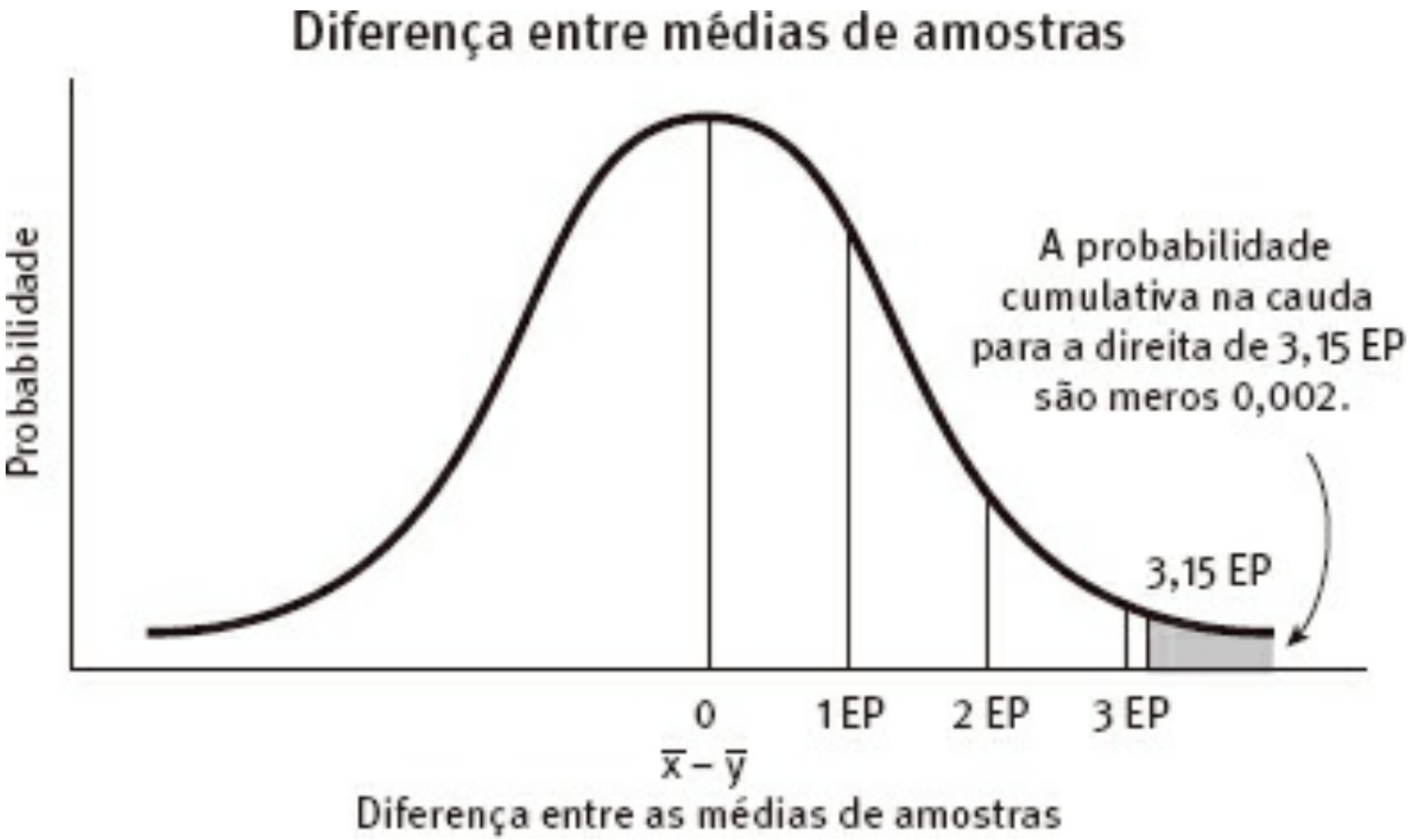
$n_x$ = número de observações na amostra $x$

$n_y$ = número de observações na amostra $y$

Nossa hipótese nula é que as médias das duas amostras são iguais. A fórmula acima calcula a diferença observada nas médias em relação ao tamanho do erro padrão para a diferença nas médias. Mais uma vez, apoiamo-nos fortemente na distribuição normal. Se as médias da população subjacente forem verdadeiramente iguais, então seria de esperar que a diferença nas médias das amostras seja menor que um erro padrão cerca de 68% das vezes; menos de dois erros padrões, cerca de 95% das vezes; e assim por diante.

No exemplo do autismo apresentado no capítulo, a diferença da média entre as duas amostras foi de 71,6 centímetros cúbicos, com um erro padrão de 22,7. A razão dessa diferença observada é de 3,15, o que significa que as duas amostras têm médias separadas por mais de três erros padrões. Como foi observado no capítulo, a probabilidade de se obter amostras com tal diferença de médias se as populações subjacentes tiverem a mesma média é muito, muito pequena. Especificamente, a probabilidade de observar uma diferença de médias que seja 3,15 erros

padrões ou mais é de 0,002.




## Teste de hipótese de uma ou duas caudas (uni ou bicaudal)

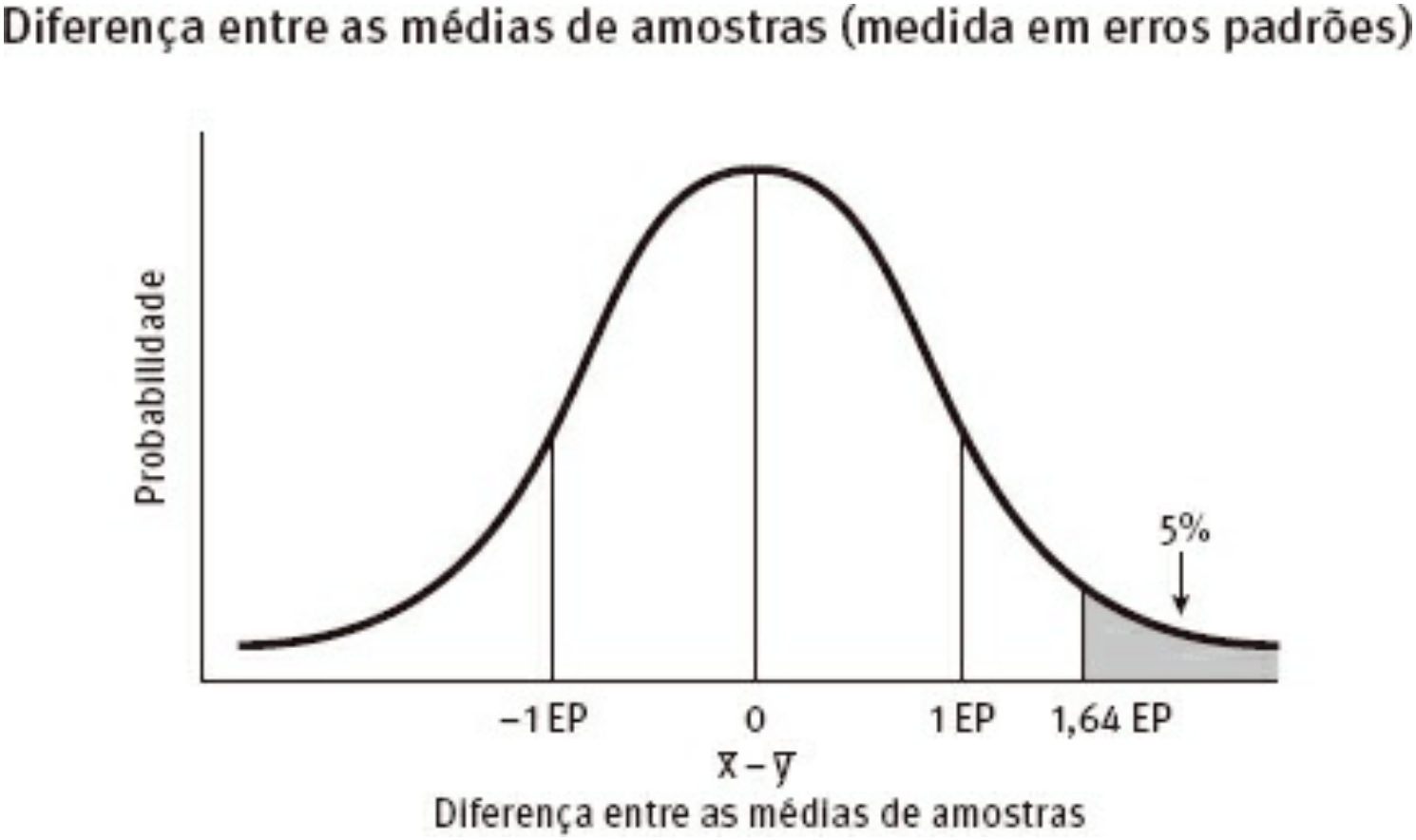
Este capítulo introduziu a ideia de usar amostras para testar se jogadores de basquete profissional *têm a mesma altura* que a população geral. Deixei de lado um detalhe. A nossa hipótese nula é que os jogadores de basquete têm a mesma altura que os homens na população geral. O que deixei de lado é que temos duas hipóteses alternativas possíveis.

Uma hipótese alternativa é que os jogadores de basquete profissional têm uma altura diferente da população masculina como um todo; eles podem ser mais altos que os outros homens da população ou mais baixos. Essa foi a abordagem que você adotou quando entrou no ônibus sequestrado e pesou os passageiros para determinar se eram participantes do estudo Changing Lives. Você podia rejeitar a hipótese nula de os passageiros do ônibus serem participantes do estudo se o peso médio deles fosse significativamente superior à média geral dos participantes do Changing Lives *ou* se fosse significativamente inferior (como acabou sendo o caso). A nossa segunda hipótese alternativa é que os jogadores de basquete profissional são em média mais altos que os outros homens da população. Nesse caso, o conhecimento anterior que temos sobre essa questão nos diz que os jogadores de basquete não podem ser mais baixos que a população geral. A distinção entre as duas hipóteses alternativas determinará se fazemos um teste de hipótese unicaudal ou um teste de hipótese bicaudal.

Em ambos os casos, vamos supor que faremos um teste com nível de significância 0,05.

Rejeitaremos a nossa hipótese nula se observarmos uma diferença nas alturas entre as duas amostras que ocorreria cinco vezes em cem ou menos se todos os caras tivessem realmente a mesma altura. Até aqui, tudo bem.

É aqui que as coisas começam a ficar um pouquinho mais matizadas. Quando a nossa hipótese alternativa é que jogadores de basquete são mais altos que outros homens, nós fazemos um *teste de hipótese unicaudal*. Medimos a diferença na altura média entre a nossa amostra de jogadores de basquete e a nossa amostra de homens comuns. Sabemos que se a nossa hipótese nula for verdadeira, então observaremos uma diferença que é de 1,64 erro padrão ou mais apenas em cinco vezes em cem. Nós rejeitamos a nossa hipótese nula se o nosso resultado cair nessa faixa, como mostra o diagrama a seguir.



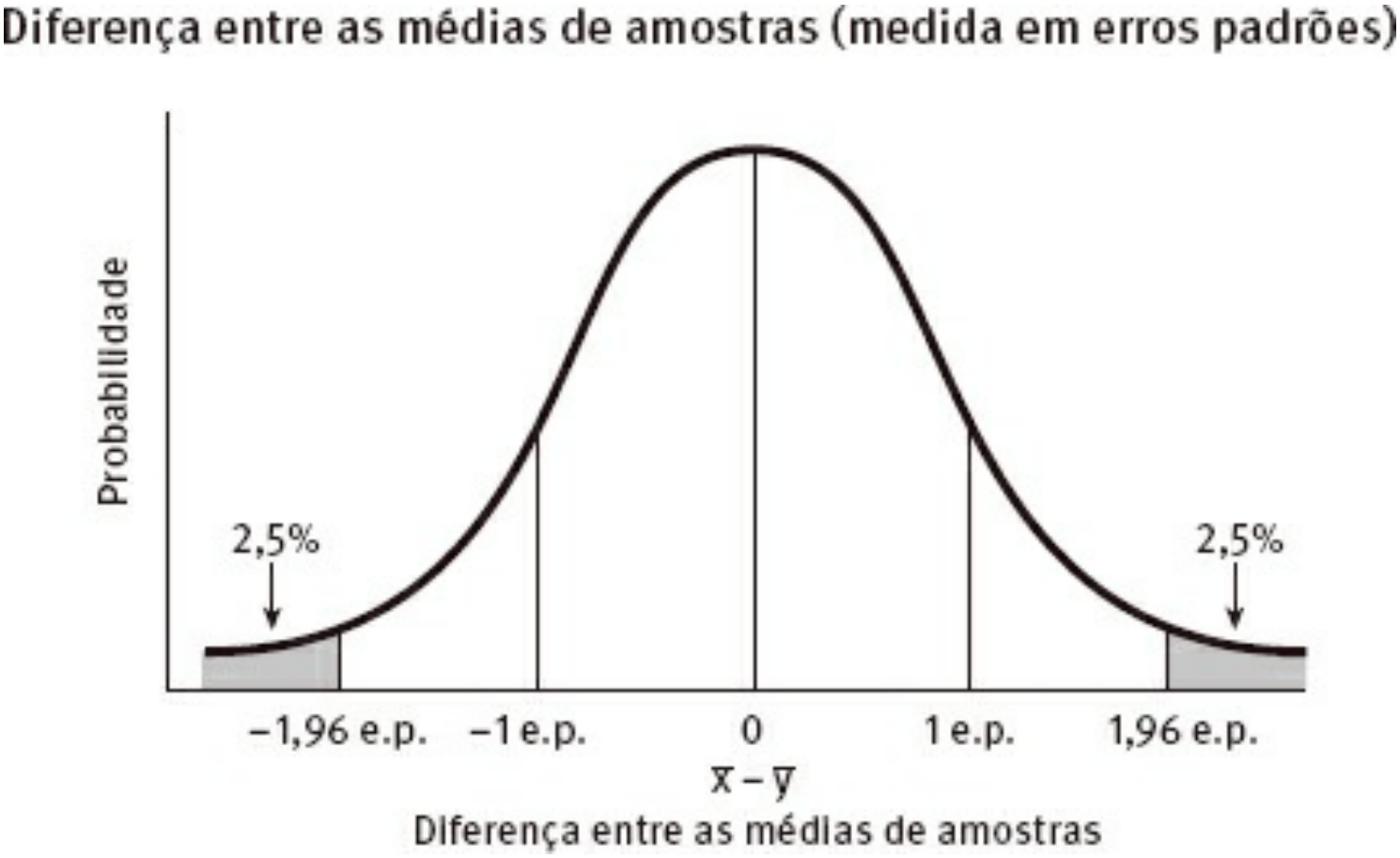

Agora revisitemos a outra hipótese alternativa – de que jogadores de basquete profissional pudessem ser mais altos ou mais baixos que a população geral. Nossa abordagem geral é a mesma. Mais uma vez, rejeitaremos nossa hipótese nula de jogadores de basquete serem da mesma altura que a população geral se obtivermos um resultado que ocorreria apenas cinco vezes em cem ou menos se realmente não houvesse diferença de altura. No entanto, há algo de diferente: precisamos considerar agora a possibilidade de jogadores de basquete serem mais baixos que a população geral. Portanto, rejeitaremos nossa hipótese nula se a nossa amostra de jogadores profissionais tiver uma altura média que seja significativamente superior *ou inferior* que a altura média para a nossa amostra de homens comuns. Isso requer um *teste de hipótese bicaudal*. Os pontos de corte para rejeitar a nossa hipótese nula serão diferentes porque agora precisamos levar em conta a possibilidade de uma diferença grande nas médias de amostras em ambas as direções: positiva ou negativa. Mais especificamente, a faixa na qual rejeitaremos a

nossa hipótese nula foi dividida em duas caudas. Ainda rejeitaremos a nossa hipótese nula se tivermos um resultado que ocorreria apenas 5% das vezes ou menos se os jogadores de basquete tivessem a mesma altura que a população geral; só que agora temos dois jeitos diferentes de poder rejeitar a hipótese nula.

Rejeitaremos a nossa hipótese nula se a altura média para a amostra de jogadores for tão maior que a média para homens comuns que observaríamos esse resultado apenas *2,5 vezes em cem* se os jogadores de basquete tivessem realmente a mesma altura que todo mundo.

E rejeitaremos a nossa hipótese nula se a altura média para a amostra de jogadores for tão menor que a média para homens comuns que observaríamos esse resultado apenas 2,5 vezes em cem se os jogadores de basquete tivessem realmente a mesma altura que todo mundo.

Juntas, essas duas contingências somam 5%, como ilustra o gráfico a seguir.




O julgamento deve informar se o tipo de teste de hipótese mais apropriado para a análise que está sendo conduzida deve ser uni ou bicaudal.

---

[a] Por uma questão de semântica, nós não *provamos* que a hipótese nula é verdadeira (que o tratamento para abuso de substâncias químicas não tem efeito). Ele pode acabar se revelando extremamente efetivo para outro grupo de detentos. Ou talvez muito mais detentos desse grupo teriam sido reincidentes se não tivessem recebido tratamento. Em todo caso, com base nos dados coletados, meramente *falhamos em rejeitar* nossa hipótese nula. Há uma distinção semelhante entre "falhar em rejeitar" uma hipótese nula e aceitar a hipótese nula. Só porque um estudo não

pôde refutar que o tratamento para abuso de substâncias não tem efeito (sim, uma dupla negativa) isso não significa que se deve aceitar que o tratamento para abuso de substâncias seja inútil. Aqui há uma significativa distinção estatística. Dito isso, a pesquisa frequentemente é projetada para respaldar uma política, e os funcionários do sistema carcerário, que precisam decidir onde alocar recursos, podem aceitar razoavelmente a posição de que o tratamento é ineficaz, até que sejam persuadidos do contrário. Aqui, como em tantas outras áreas da estatística, o julgamento tem importância.

[b] Esse exemplo é inspirado em fatos reais. Obviamente muitos detalhes foram modificados por razões de segurança nacional. Não posso nem confirmar nem negar meu próprio envolvimento.

[c] Para ser preciso, 95% de todas as médias de amostras estarão dentro de *1,96 erro padrão* acima ou abaixo da média da população.

[d] Existem duas hipóteses alternativas possíveis. Uma é que os jogadores de basquete profissionais são mais altos que a população masculina geral. A outra é meramente que os jogadores de basquete profissionais tenham uma altura média diferente da população masculina em geral (deixando aberta a possibilidade de haver jogadores de basquete que possam na realidade ser mais baixos que outros homens). Essa distinção tem um pequeno impacto quando se realizam testes de significância e se calculam valores-p. Ela é explicada em textos mais avançados e não é importante para a nossa discussão geral aqui.

[e] Confesso que uma vez rasguei um livro de estatística ao meio por pura frustração.

[f] No pôquer, sequência máxima de cartas de um mesmo naipe. (N.T.)

[g] Outra resposta seria tentar replicar os resultados em estudos adicionais.