Arquivos de Ciências do Esporte
Archives of Sport Sciences

Ponto de Vista

Quanto vale o valor-p?

How much is the p-value worth?

Carlos Alberto Mourão Júnior

Resumo
Esse texto aborda o conceito e a utilidade do valor-p, bem como discute seus usos e abusos na bioestatística.

Palavras-Chave: bioestatística, valor-p, metodologia, análise quantitativa, biometria.

Abstract
This text addresses the concept and usefulness of p-value, as well as discusses its uses and abuses in biostatistics.

Keywords: biostatistics, p-value, methodology, quantitative analysis, biometry.

Afiliação dos autores

Universidade Federal de Juiz de Fora, Juiz de Fora, Minas Gerais, Brasil.

Autor correspondente

Universidade Federal de Juiz de Fora, Instituto de Ciências Biológicas Departamento de Fisiologia, CEP: 36036-900, Juiz de Fora, Minas Gerais, Brasil.
e-mail: camouraojr@gmail.com

Conflito de interesses

Os autores declararam não haver conflito de interesses.

Processo de arbitragem

Recebido: 26/08/2019 Aprovado: 16/09/2019

Introdução

É fato que o valor-p é o conceito mais usado (e mal-usado) da bioestatística. Iniciarei esse texto explicando o significado do valor-p, mesmo para os leitores não iniciados em estatística.

Para tanto, lançarei mão de analogias e exemplos que ilustram o que o valor-p significa. Usarei inicialmente "definições" que, embora sejam conceitualmente afastadas do rigor matemático, servem bem para promover a compreensão do que é o valor-p e para que ele serve.

Começarei, para efeito didático, com uma analogia. Quando se leva um réu suspeito de um homicídio a júri, é porque se acredita que ele seja culpado. Entretanto, como é bem mais grave prender um inocente do que deixar solto um culpado, mesmo que acreditemos fortemente na culpa do réu, temos que partir do princípio da presunção da inocência (*in dubio pro reo*).

Ou seja, mesmo acreditando que ele é culpado, é necessário considerá-lo, inicialmente, como inocente, até que se consiga reunir um corpo suficiente de provas em contrário. Essa situação ilustra o fato de que, apesar de querermos provar a culpa, começamos assumindo a inocência.

Agora imagine que eu queira mostrar que um medicamento recém-lançado é mais eficaz que um medicamento já conhecido. Por exemplo, pretendo mostrar que o medicamento A é mais eficiente que o medicamento B para reduzir os níveis de colesterol no sangue. Então dou o medicamento A para um grupo (grupo experimental [grupo A]) e o medicamento B para outro grupo (grupo controle [grupo B]). Da mesma maneira que no caso do júri, quero mostrar que o efeito de A é diferente de B, mas, como quero minimizar o risco de aceitar erroneamente a hipótese de que A e B sejam diferentes (pois isso implicaria modificar uma conduta terapêutica equivocadamente) então, mesmo acreditando, por hipótese, que o efeito dos medicamentos A e B seja diferente, eu começo partindo da premissa de que o efeito dos medicamentos A e B é igual (tecnicamente chamamos a hipótese de que os parâmetros entre grupos são iguais de hipótese nula). Então, o teste estatístico que iremos aplicar vai partir da hipótese de que não existe diferença (hipótese nula) e então irá calcular a chance de que a hipótese nula seja verdadeira. Essa probabilidade é o valor-p. Embora essa definição não seja matematicamente precisa – como veremos mais adiante (quando dermos a definição rigorosa e formal do valor-p) –, ela serve para dar uma noção bastante clara do que seja o valor-p, para fins práticos.

Por convenção, quando o valor-p é menor do que 0,05 (i.e., 5%), isso significa que temos menos de 5% de chance de estarmos errados, quando achamos que os medicamento A e B têm efeitos diferentes. Por exemplo, suponhamos que após fazermos um teste estatístico para testar diferença de médias, encontremos um valor-p de 0,02. Isso significa que temos somente 2% de evidências a favor da hipótese de que os efeitos de A e B são iguais, então rejeitamos tal hipótese (hipótese nula), assumindo o risco de 2% de estarmos errados, e passamos a prescrever o medicamento A ao invés do B. Em outras palavras, se o advogado só consegue reunir 2% de evidências a favor da inocência do réu, então o júri o considera culpado (desde que assuma como aceitável o risco de 2% de estar errado).

Mas, por que precisamos aplicar um teste estatístico de inferência e calcular um valor-p para decidirmos se há ou não diferença entre dois grupos? Por uma razão muito simples: como não teríamos como testar os medicamentos A e B em todos os pacientes do mundo que têm níveis altos de colesterol, então selecionamos duas amostras possíveis e avaliamos estatisticamente se podemos ou não rejeitar a hipótese nula.

Frente ao que foi dito, podemos agora propor o seguinte desenho experimental para testar se o medicamento A é mais eficiente em reduzir o colesterol que o medicamento B. Escolhemos aleatoriamente duas amostras (A e B) com 30 pacientes cada. Para os 30 sujeitos do grupo A (grupo experimental) ministramos o medicamento A e para os 30 sujeitos do grupo B (grupo controle) ministramos o medicamento B. Suponhamos que, antes de começar o estudo, a média do valor de colesterol de ambos os grupos fosse equivalente. Após três meses, usando os medicamentos, comparamos as médias do colesterol de ambos os grupos e verificamos que, no grupo A, a média do colesterol era de 230 mg/dl, enquanto no grupo B a média do colesterol era de 237 mg/dl. Logo, houve uma diferença de 7 mg/dl, ou seja, o medicamento A foi capaz de reduzir o colesterol em 7 unidades. Mas isso é suficiente para concluir que o medicamento A é mais eficaz? Não, porque esse resultado pode ter sido obra do acaso. Para ficarmos seguros desse resultado, seria necessário que ele se repetisse em vários pares de amostras sorteados aleatoriamente. Se sorteássemos mais algumas dezenas de amostras (no caso, com 30 sujeitos em cada uma) e uma diferença igual ou maior do que 7 aparecesse, já ficaríamos mais confiantes na superioridade do medicamento B. Isso ocorre porque a repetição de um evento gera confiança acerca de sua ocorrência. Se lançamos uma moeda três vezes e obtemos três caras, dá para afirmar que a moeda é viciada? Não, isso pode ser obra do acaso. Mas se lançamos a mesma moeda cem vezes, e obtemos cem caras, a probabilidade de que isso seja obra do acaso é insignificante. Nesse caso, apostaríamos nossas fichas na hipótese de a moeda ser viciada.

No exemplo dos medicamentos A e B, é impossível saber com certeza se a diferença encontrada se deu por acaso. Mas é possível medir a probabilidade de o acaso explicar essa diferença de 7 unidades. Essa probabilidade pode ser representada pelo valor-p. Suponhamos que para fazer a comparação da média de colesterol entre os dois grupos, utilizamos um teste estatístico de comparação de médias (o teste-t), e que o mesmo nos forneceu um valor-p de 0,01. Isso nos sugere que a probabilidade de que a diferença encontrada tem uma chance de apenas 1% de ser por acaso. Ou seja, se escolhêssemos 100 pares de amostras de 50 sujeitos e comparássemos as médias em apenas um par de amostras, a diferença teria sido menor que 7 unidades. Então, concluímos que o medicamento A é superior ao B, correndo o risco de apenas 1% de estarmos equivocados. Repito: nenhuma dessas "definições" do valor-p são matematicamente formais e rigorosas, porém elas são bastante didáticas para se entender intuitivamente o que o valor-p representa.

Um fato digno de nota é que, ao contrário do que muitos imaginam, o valor-p é muito afetado pelo tamanho das amostras que se scolhe. Com efeito, se eu for aumentando o tamanho da amostra, menor será o valor-p, ainda que a diferença entre as médias se mantenha constante! Vamos tentar entender isso a partir de um exemplo concreto.

Imaginemos um estudo hipotético realizado num berçário, cujo objetivo tenha sido verificar se o tamanho dos bebês, ao nascer, era diferente entre os sexos. Para tanto, foram selecionados 10 meninos e 10 meninas, e verificou-se que a média do tamanho dos meninos foi de 49,3 cm com desvio-padrão de 2,4 cm. Nas meninas, a média da altura foi de 48,5 com um desvio-padrão de 2,5 cm. Portanto, nessa amostra (20 crianças) a diferença entre as médias foi de 0,8 cm. Em seguida, foi aplicado um teste-t (para verificar a probabilidade desse resultado ter sido ao acaso) cujo resultado foi $p = 0,51$. Nesse caso, havia o risco de 51% de erro, ao se afirmar que os tamanhos de meninos e meninas eram diferentes. Como o risco aceitável (por mera convenção) é de $p < 0,05$, considerou-se que não havia diferença estatisticamente significante entre o tamanho dos meninos e das meninas. Agora, fazendo a mesma análise (fiz essas simulações usando o aplicativo GraphPad Instat versão 3.06) e variando apenas o tamanho da amostra (n), mantendo as mesmas médias e desvios-padrão, encontramos o seguinte: se $n = 40$ então $p = 0,34$; se $n = 60$ então $p = 0,24$; se $n = 100$ então $p = 0,12$. Mas, quando faço com $n = 200$ (ou seja, 100 meninos e 100 meninas) já tenho um valor-p estatisticamente significante (nesse caso, $p = 0.03$). Esses exemplos podem ser observados na Figura 1.

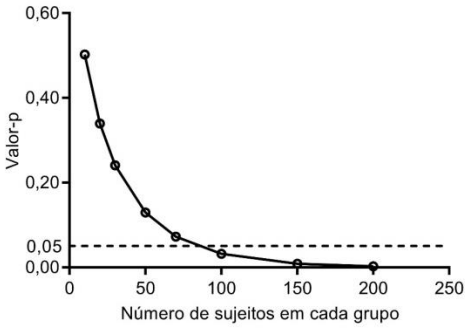

Figura 1. Variação do valor-p em função do tamanho da amostra.

Podemos ver claramente na Figura 1 que o valor-p vai decrescendo de forma exponencial na medida em que aumentamos o tamanho da amostra, mesmo que a diferença entre as médias tenha se mantido constante (no caso, a diferença de tamanho médio entre os meninos e meninas foi de 0.8 cm). Isso mostra claramente que, se quisermos encontrar um $p < 0,05$, basta aumentarmos o tamanho da amostra. De fato, é isso que muitos pesquisadores fazem. Por exemplo, em estudos multicêntricos, onde o n geralmente é maior do que 1.000 sujeitos, qualquer diferença entre grupos, por mais irrisória e insignificante que seja, vai apresentar um $p < 0,001$ (considerado "extremamente significante", estatisticamente falando).

Ademais, o fato de encontrar um valor-p estatisticamente significante não nos fala nada a respeito da magnitude da diferença entre as médias. O valor-p apenas nos diz qual é a probabilidade de a diferença encontrada (seja ela qual for) ter sido obra do acaso. Nada mais do que isso. O valor-p, repito, não nos fala nada a respeito da relevância da diferença encontrada. Afinal, uma mínima diferença de tamanho entre meninos e meninas (menos de 1 cm) se mostrou estatisticamente não significante para um $n = 20$, mas a mesma diferença se mostrou estatisticamente significante para um $n = 200$.

Portanto, para sabermos se uma dada diferença se deu por acaso, usamos o valor-p, mas, para sabermos se tal diferença é ou não relevante, teremos que usar o conhecimento teórico e o bom senso. De fato, nos exemplos anteriores (independentemente do valor-p), uma diferença de apenas 7 mg/dl no colesterol é suficiente para reduzir o risco de doenças? Uma diferença de tamanho de 0,8 cm (menos de 1 cm) nos bebês é algo relevante? Em ambos os casos, mesmo diante de diferenças pífias, podemos ter um $p < 0,05$, ou seja, um resultado estatisticamente significante, mas clinicamente insignificante, irrelevante, até desprezível.

O problema é que o fetiche criado pelo valor-p e a ditadura do $p < 0,05$ passaram a ser critério para nortear tomadas de decisões e também se tornaram condição necessária para se conseguir publicar ou não os resultados de uma pesquisa, num mundo em que o produtivismo acadêmico privilegia quantidade de publicações em detrimento de sua relevância. Deixou-se, há muito, o bom senso de lado, a ciência passou a se resumir na busca desenfreada por um $p < 5\%$, a qualquer custo, como se isso fosse garantia de alguma coisa. Esse mal-uso do valor-p vem sendo denunciado por pesquisadores diversos e também por estatísticos, na literatura científica atual[1-12].

Com toda a certeza, usar o valor-p para tomar decisões pode nos levar a correr riscos inaceitáveis, tais como substituir erroneamente um medicamento novo por um tradicional, modificar erroneamente padrões dietéticos, decidir erroneamente que um medicamento pode causar câncer, dizer erroneamente que o uso de vitaminas e hormônios pode alterar a longevidade, substituir erroneamente um procedimento cirúrgico por outro etc.

Se usássemos o bom-senso e valorizássemos a magnitude das diferenças encontradas nos estudos ao invés de valorizarmos o valor-p, não cometeríamos tais erros. Se observarmos que uma diferença de 0,8 cm no tamanho de bebês é insignificante do ponto de vista prático, nem nos preocuparíamos se tal diferença é estatisticamente "significante" ou não. Se olhássemos para o óbvio, para aquilo que nossos resultados nos mostram, não cairíamos na cilada de tomar decisões em função do valor-p. Se as revistas científicas ainda tivessem um rastro de bom-senso, jamais tomariam decisões de publicar ou não artigos com base no valor-p. Mas infelizmente o valor-p se tornou um gigante desajeitado e sem controle, um verdadeiro monstro que, após criado, vem causando consequências imprevisíveis e perigosas.

Agora que estou finalizando - só para efeito de informação -, como eu disse antes, embora as explicações que eu dei sobre o valor-p sejam didaticamente válidas, elas não são formalmente (matematicamente) precisas. O conceito formal, caso possa vir a interessar a alguém, é o seguinte. O valor-p é uma probabilidade condicional, calculada a partir da pressuposição de que a hipótese nula é verdadeira, ou seja, um valor-p de 0,05 não significa que haja 95% de chance de que determinada hipótese esteja correta. Em vez disso, significa que se a hipótese nula é verdade, e todas as outras suposições feitas são válidas, há 5% de chance de obter um resultado pelo menos tão extremo quanto o observado[13-15]. Pois bem, essa é a definição formal do valor-p. Concordo que ela é bem indigesta para os não iniciados em estatística. Mas, no frigir dos ovos, para efeitos práticos, o que ela significa é o que tentamos explicar de maneira não rigorosa nesse artigo.

Por fim, o objetivo desse artigo foi justamente "colocar os pingos nos is" e mostrar para que o valor-p serve, e para que ele não serve. Ele é somente uma probabilidade de a diferença encontrada ter sido fruto do acaso. Nada mais do que isso! Ele não é capaz de dizer mais nada além disso. Ele não fala absolutamente nada a respeito da importância ou do significado prático das diferenças encontradas. Conclamamos a comunidade científica a dar menos valor ao valor-p e valorizar mais o bom-senso, antes que esse entre de vez em extinção e arraste a ciência para uma crise de dimensões e consequências inimagináveis.

Referências

1. Hardwicke TE, Ioannidis JPA. Petitions in scientific argumentation: dissecting the request to retire statistical significance. European journal of clinical investigation. 2019.
2. Amrhein V, Greenland S, McShane B. Scientists rise up against statistical significance. Nature. 2019;567(7748):305-7.
3. Benjamin DJ, Berger JO, Johannesson M, Nosek BA, Wagenmakers EJ, Berk R, et al. Redefine statistical significance. Nature human behaviour. 2018;2(1):6-10.
4. Amrhein V, Greenland S. Remove, rather than redefine, statistical significance. Nature human behaviour. 2018;2(1):4.
5. McShane BB, Gal D. Statistical Significance and the Dichotomization of Evidence. Journal of the American Statistical Association. 2017;112(519):885-95.
6. Laber EB, Shedden K. Statistical Significance and the Dichotomization of Evidence: The Relevance of the ASA Statement on Statistical Significance and p-values for Statisticians. J Am Stat Assoc. 2017;112(519):902-4.
7. Briggs WM. The Substitute for p-Values. Journal of the American Statistical Association. 2017;112(519):897-8.
8. Berry D. A p-Value to Die For. Journal of the American Statistical Association. 2017;112(519):895-7.
9. Yaddanapudi LN. The American Statistical Association statement on P-values explained. Journal of anaesthesiology, clinical pharmacology. 2016;32(4):421-3.
10. Wasserstein RL, Lazar NA. The ASA Statement on p-Values: Context, Process, and Purpose. The American Statistician. 2016;70(2):129-33.
11. Baker M. Statisticians issue warning over misuse of P values. Nature. 2016;531(7593):151.
12. Ferreira JC, Patino CM. What does the p value really mean? Jornal brasileiro de pneumologia : publicacao oficial da Sociedade Brasileira de Pneumologia e Tisilogia. 2015;41(5):485.
13. Motulsky H. Essential biostatistics: a nonmathematical approach. New York: Oxford University Press; 2016.
14. Motulsky H. Intuitive biostatistics: a nonmathematical guide to statistical thinking. 2 ed. New York: Oxford University Press; 2010.
15. Bailar JC, Mosteller F. Medical uses of statistics. 2 ed. Boston: NEJM Books; 1992.