EDITORIAL

JORNAL **VASCULAR** BRASILEIRO
ISSN 1677-7301 (Online)

# P-valor e dimensão do efeito em estudos clínicos e experimentais

*P-value and effect-size in clinical and experimental studies*

Anna Carolina Miola[1], Hélio Amante Miot[1]

**Como citar:** Miola AC, Miot HA. P-valor e dimensão do efeito em estudos clínicos e experimentais. J Vasc Bras. 2021;20:e20210038. https://doi.org/10.1590/1677-5449.210038

A natureza complexa dos sistemas biológicos faz com que muitos experimentos apresentem certa variabilidade amostral. Ainda, grande parte das intervenções biomédicas promove efeitos moderados e sem um evidente gradiente dose-resposta. Contudo, ao passo que se emprega a estatística para concluir quanto à diferença entre amostras, a maior variabilidade das medidas e a modesta diferença entre grupos comprometem o poder analítico (erro tipo II). Esse detalhe exige uma cuidadosa interpretação do p-valor (significância estatística) e da dimensão do efeito na inferência resultante de estudos de comparação entre grupos, apesar desses conceitos se aplicarem também a análises de correlação, concordância, sobrevivência, testes diagnósticos, entre outros[1-5].

Segundo a estatística frequentista, duas ou mais amostras podem ser originárias de uma mesma população, porém, apresentam certa variabilidade em algumas características. Quanto mais similares forem as amostras, maior a chance de terem a mesma natureza; por outro lado, amostras que se apresentam de forma muito diferente têm menor chance de terem sido selecionadas ao acaso, dentro da mesma população. Os estatísticos desenvolveram uma série de modelos matemáticos que estimam a probabilidade de que amostras pertençam a uma mesma população e que suas diferenças constatadas no experimento tenham ocorrido ao acaso. De forma geral, o p-valor de um teste estatístico retorna à probabilidade teórica de que valores mais extremos do que os encontrados sejam frutos do acaso, desde que os grupos testados sejam realmente iguais ($H_0$ verdadeira)[6,7].

Cabe ao pesquisador definir o ponto de corte a partir do qual ele considera, para o p-valor, uma probabilidade baixa o suficiente para assumir que os grupos sejam diferentes. A decisão desse nível de significância (nível α), assim como a direção da análise (uni ou bicaudal), devem ser baseadas em princípios teóricos e definidas previamente à análise. Isso é de fundamental importância, porque toda escolha de um ponto de corte pode sacrificar conclusões derivadas de resultados muito próximos a esse limite. Por exemplo, não se deve sobrevalorizar $p = 0,04$ em detrimento de $p = 0,06$, quando o ponto de corte escolhido for $p < 0,05$[8].

Em testes de comparação entre grupos, o p-valor é influenciado pela diferença entre as médias (ou proporções), mas também pela variância dos dados e pela dimensão da amostra. A Figura 1 mostra três situações diferentes, em que se comparam amostras com variação nos desvios-padrão e tamanho amostral. Amostras com mesma média e desvio padrão apresentam p-valores diferentes, de acordo com o tamanho amostral (Figura 1 A *vs*. B). Já amostras com a mesma média e tamanho amostral apresentam p-valores distintos se diferirem apenas quanto ao desvio padrão (Figura 1 A *vs*. C).

Convencionalmente, pesquisadores adotam níveis de significância na faixa de 5% ($p \leq 0,05$) para a análise de pequenas amostras ($n < 50$) e, com isso, assumem o risco de o resultado encontrado ocorrer ao acaso em pelo menos uma vez a cada 20 execuções do experimento[9]. A adoção de níveis de significância mais restritos (por exemplo, $p < 0,01$) aumenta a reprodutibilidade dos estudos, porém, deve penalizá-los com maiores erros do tipo II. Contudo, como o tamanho amostral e o número de variáveis envolvidas na análise (número de comparações)

[1] Universidade Estadual Paulista – UNESP, Faculdade de Medicina – FMB, Departamento de Infectologia, Dermatologia, Diagnóstico por Imagem e Radioterapia, Botucatu, SP, Brasil.

Fonte de financiamento: Nenhuma.
Conflito de interesse: Os autores declararam não haver conflitos de interesse que precisam ser informados.
Submetido em: Março 05, 2021. Aceito em: Abril 15, 2021.

O estudo foi realizado no Departamento de Infectologia, Dermatologia, Diagnóstico por Imagem e Radioterapia, Faculdade de Medicina (FMB), Universidade Estadual Paulista (UNESP), Botucatu, SP, Brasil.

influenciam o p-valor, isso deve ser cuidadosamente ponderado na decisão do nível de significância. O emprego de amostras vultosas (n > 1.000) favorece o encontro ocasional de p-valores pequenos, sendo recomendado utilizar níveis de significância mais restritos, como $p \leq 0,001$. As modernas explorações genéticas comparam, simultaneamente, milhares de variáveis, favorecendo o encontro casual de p-valores diminutos, sendo recomendados níveis de significância da ordem de $p < 5x10^{-8}$.[10,11]

Os p-valores resultantes de um teste estatístico devem ser apresentados como sua medida exata e com um número de decimais compatível com a grandeza que se propõe avaliar. Por exemplo, deve-se referir p = 0,032 em vez de p < 0,05 ou de p = 0,032016[12,13]. O acréscimo de decimais não é contraprova de maior importância ou fidedignidade dos resultados. Ainda, p-valores marginais ao nível de significância (por exemplo, p = 0,067) não devem ser interpretados como uma "tendência" para rejeitar a hipótese nula, uma vez que a ampliação da amostra não garante que a diferença entre os grupos seja mantida[14].

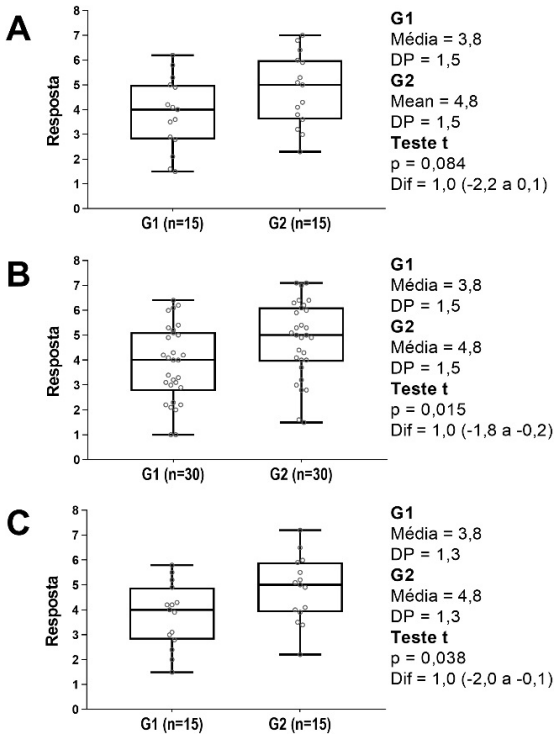


**Figura 1.** Exemplos hipotéticos de comparações (bidirecionais) de dois grupos de tratamento (G1 e G2), todos com mesma média e mediana. (A) Amostra com 15 participantes por grupo (p = 0,08); (B) Amostra com 30 participantes por grupo e mesmo desvio padrão que o Exemplo A (p = 0,02); (C) Amostra com 15 participantes por grupo e menor desvio padrão que o exemplo A (p = 0,04).

É, pois, importante que o p-valor não seja utilizado como medida de validade de um resultado ou da força de uma associação[15]. Tampouco p-valores maiores que o nível de significância (por exemplo, p > 0,1) devem ser interpretados como identidade entre as amostras[7]. Uma medida adicional para a compreensão da relação entre os grupos amostrados são os estimadores chamados de dimensão do efeito[16].

Desde que as amostras representem adequadamente uma população (coleta aleatorizada), suas estatísticas podem estimar parâmetros dessa mesma população, permitindo realizar inferências sobre o comportamento das variáveis estudadas. A dimensão do efeito é um indicador que quantifica a diferença entre as amostras, e a estimativa do seu intervalo de confiança de 95% (IC95%) dimensiona a incerteza do comportamento do parâmetro na população de origem, retornando uma informação mais valiosa que o p-valor quanto ao real comportamento do fenômeno estudado[17,18].

A Tabela 1 apresenta os principais indicadores de dimensão de efeito utilizados em estudos epidemiológicos e que devem acompanhar o p-valor nos resultados de testes estatísticos, contudo, o significado independente de cada um deles ultrapassa o escopo do texto[19]. Há, ainda, outros estimadores de dimensão de efeito, mais empregados em estudos experimentais, cuja interpretação é menos intuitiva; entre eles, estão o coeficiente "d" de Cohen, $R^2$, o ômega e o "eta" quadrado ($\omega^2$ e $\eta^2$), que podem requerer suporte de um estatístico experiente[18,20].

Todo teste estatístico deve ser apresentado (e interpretado) de acordo com o p-valor, uma dimensão do efeito, e seu IC95%[12,13,21,22]. Um experimento que resulte em grande dimensão de efeito e p-valor = 0,06 é certamente mais relevante que um resultado que exiba pequena dimensão do efeito e $p < 0,01$[23-25].

Um estudo recente que avaliou a efetividade de meias de compressão na melhora do edema ocupacional resultou em $p < 0,0001$[26], todavia, a indisponibilidade dos valores de redução como dimensão do efeito (por exemplo, redução do diâmetro vespertino do tornozelo, ou escore VEINES) dificulta a interpretação dos dados e sua inferência visando a indicação clínica.

Por outro lado, especialmente, em amostragens mais vultosas, o encontro de p-valores reduzidos pode não representar em um efeito clinicamente sensível que leve à mudança de paradigmas médicos. Na importante revisão sistemática de Martinez-Zapata et al.[27] sobre ventonônicos em insuficiência venosa, foi sugerida a superioridade de drogas venotônicas devido a sua significância estatística ($p < 0,05$), porém, a dimensão do efeito encontrada resultou em uma redução média de apenas 4,27 mm (IC95% 2,93–5,61 mm) na circunferência do tornozelo de 2.010 participantes (15 estudos), o

**Tabela 1.** Principais dimensões de efeito de acordo com o tipo do estudo epidemiológico.

| Tipo de estudo | Dimensão do efeito |
|---|---|
| Diagnóstico | Sensibilidade, especificidade, valor preditivo positivo (ou negativo), razão de verossimilhança, área sob a curva ROC |
| Ecológico | Coeficientes de correlação (r ou rho) |
| Caso-controle | Razão de chances, razão de prevalência |
| Sobrevivência | Razão de risco |
| Ensaio clínico/coorte | Risco relativo, risco atribuível, redução do risco relativo, redução absoluta do risco, número necessário para o tratamento (ou para dano), diferença absoluta entre os grupos (percentual ou médias). |

ROC = característica de operação do receptor.

que, apesar de verdadeiro, não representa um benefício evidente para o paciente com edema dos membros inferiores.

Excepcionalmente, pode haver uma discreta divergência entre a amplitude da dimensão de efeito e o p-valor, por exemplo, como um resultado de risco relativo 0,70 (IC95% 0,36–1,01) e p-valor = 0,045, porém, isso não deve ser considerado um erro, já que são estimativas oriundas de cálculos diferentes e que tendem a convergir com o aumento amostral.

Há um recente movimento acadêmico para a completa abolição do p-valor e do termo "estatisticamente significante" nas publicações científicas, em preferência pela representação exclusiva da dimensão de efeito de um teste, por ser mais informativa e permitir a generalização dos resultados[28]. Realmente, estudos que baseiam suas conclusões unicamente no p-valor são mais susceptíveis à não reprodutibilidade, além de estimularem os pesquisadores a perseguirem a significância estatística em detrimento à relevância do resultado (*p-hacking*)[23,28-31]. Contudo, esse ainda é um movimento incipiente entre os pesquisadores, e uma campanha para a interpretação correta do p-valor analisado em combinação com a dimensão do efeito constitui uma alternativa mais acertada que sua abolição[32,33].

Finalmente, as comparações entre grupos podem ser avaliadas de forma uni ou bidirecional (uni/bicaudal). Convenciona-se chamar de estudo de diferença quando avaliamos se o comportamento de uma variável pode ser superior ou inferior entre as amostras. Entretanto, muitas avaliações são, por natureza, unidirecionais, como a comparação do número de casos de uma doença entre vacinados e não vacinados; ou em testes de não inferioridade entre duas terapias[34]. Nesses exemplos, não faz parte da hipótese de pesquisa a possibilidade de que o resultado seja contemplado de forma bidirecional. O emprego de análises unicaudais, todavia, não é consensual entre os epidemiologistas, porque, apesar de apresentarem maior poder estatístico e demandarem menor amostragem, aumentam a chance de erro tipo I[35-37]. Tais análises exigem supervisão de um estatístico experiente para o cálculo do p-valor e do IC95% unicaudais.

Enquanto a dimensão do p-valor pode informar ao leitor se há algum efeito significativo, o mesmo não revela a extensão do impacto desse efeito nas variáveis estudadas[38]. Portanto, o pesquisador deve estar atento aos resultados dos testes estatísticos, no sentido de que sua interpretação deva contemplar o p-valor em conjunto com a dimensão do efeito, especialmente estimada pelo seu intervalo de confiança de 95%, já que o significado pragmático do experimento é uma informação independente da sua significância estatística.

## REFERÊNCIAS


1. Miot HA. Análise de concordância em estudos clínicos e experimentais. J Vasc Bras. 2016;15(2):89-92. http://dx.doi.org/10.1590/1677-5449.004216. PMid:29930571.
2. Polo TCF, Miot HA. Use of ROC curves in clinical and experimental studies. J Vasc Bras. 2020;19:e20200186. http://dx.doi.org/10.1590/1677-5449.200186.
3. Miot HA. Correlation analysis in clinical and experimental studies. J Vasc Bras. 2018;17(4):275-9. http://dx.doi.org/10.1590/1677-5449.174118. PMid:30787944.
4. Schober P, Bossers SM, Schwarte LA. statistical significance versus clinical importance of observed effect sizes: what do P values and confidence intervals really represent? Anesth Analg. 2018;126(3):1068-72. http://dx.doi.org/10.1213/ANE.0000000000002798. PMid:29337724.
5. Miot HA. Análise de sobrevivência em estudos clínicos e experimentais. J Vasc Bras. 2017;16(4):267-9. http://dx.doi.org/10.1590/1677-5449.001604. PMid:29930659.
6. Concato J, Hartigan JA. P values: from suggestion to superstition. J Investig Med. 2016;64(7):1166-71. http://dx.doi.org/10.1136/jim-2016-000206. PMid:27489256.
7. Wasserstein RL, Lazar NA. The ASA statement on p-values: context, process, and purpose. Am Stat. 2016;70(2):129-33. http://dx.doi.org/10.1080/00031305.2016.1154108.
8. Miot HA. Tamanho da amostra em estudos clínicos e experimentais. J Vasc Bras. 2011;10(4):275-8. http://dx.doi.org/10.1590/S1677-54492011000400001.
9. Halsey LG, Curran-Everett D, Vowler SL, Drummond GB. The fickle P value generates irreproducible results. Nat Methods.

2015;12(3):179-85. http://dx.doi.org/10.1038/nmeth.3288. PMid:25719825.

10. Clarke GM, Anderson CA, Pettersson FH, Cardon LR, Morris AP, Zondervan KT. Basic statistical analysis in genetic case-control studies. Nat Protoc. 2011;6(2):121-33. http://dx.doi.org/10.1038/nprot.2010.182. PMid:21293453.
11. Barsh GS, Copenhaver GP, Gibson G, Williams SM. Guidelines for genome-wide association studies. PLoS Genet. 2012;8(7):e1002812. http://dx.doi.org/10.1371/journal.pgen.1002812. PMid:22792080.
12. Indrayan A. Reporting of Basic Statistical Methods in Biomedical Journals: Improved SAMPL Guidelines. Indian Pediatr. 2020;57(1):43-8. http://dx.doi.org/10.1007/s13312-020-1702-4. PMid:31937697.
13. Lang TA, Altman DG. Basic statistical reporting for articles published in biomedical journals: the "Statistical Analyses and Methods in the Published Literature" or the SAMPL Guidelines. Int J Nurs Stud. 2015;52(1):5-9. http://dx.doi.org/10.1016/j.ijnurstu.2014.09.006. PMid:25441757.
14. Ferreira JC, Patino CM. What does the p value really mean? J Bras Pneumol. 2015;41(5):485. http://dx.doi.org/10.1590/S1806-37132015000000215. PMid:26578145.
15. Wasserstein RL, Schirm AL, Lazar NA. Moving to a world beyond "p< 0.05". Am Stat. 2019;73(Supl 1):1-19. http://dx.doi.org/10.1080/00031305.2019.1583913.
16. Lee DK. Alternatives to P value: confidence interval and effect size. Korean J Anesthesiol. 2016;69(6):555-62. http://dx.doi.org/10.4097/kjae.2016.69.6.555. PMid:27924194.
17. McGough JJ, Faraone SV. Estimating the size of treatment effects: moving beyond p values. Psychiatry (Edgmont). 2009;6(10):21-9. PMid:20011465.
18. Nakagawa S, Cuthill IC. Effect size, confidence interval and statistical significance: a practical guide for biologists. Biol Rev Camb Philos Soc. 2007;82(4):591-605. http://dx.doi.org/10.1111/j.1469-185X.2007.00027.x. PMid:17944619.
19. Coutinho ES, Cunha GM. Conceitos básicos de epidemiologia e estatística para a leitura de ensaios clínicos controlados. Br J Psychiatry. 2005;27(2):146-51. http://dx.doi.org/10.1590/S1516-44462005000200015.
20. Conboy JE. Algumas medidas típicas univariadas da magnitude do efeito. Anal Psicol. 2003;21(2):145-58. http://dx.doi.org/10.14417/ap.29.
21. Schulz KF, Altman DG, Moher D, CONSORT Group. CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. PLoS Med. 2010;7(3):e1000251. http://dx.doi.org/10.1371/journal.pmed.1000251. PMid:20352064.
22. von Elm E, Altman DG, Egger M, Pocock SJ, Gotzsche PC, Vandenbroucke JP, et al. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. PLoS Med. 2007;4(10):e296. http://dx.doi.org/10.1371/journal.pmed.0040296. PMid:17941714.
23. Nuzzo R. Scientific method: statistical errors. Nature. 2014;506(7487):150-2. http://dx.doi.org/10.1038/506150a. PMid:24522584.
24. Matthews JN, Altman DG. Statistics notes. Interaction 2: Compare effect sizes not P values. BMJ. 1996;313(7060):808. http://dx.doi.org/10.1136/bmj.313.7060.808. PMid:8842080.
25. Fleischmann M, Vaughan B. Commentary: statistical significance and clinical significance - A call to consider patient reported outcome measures, effect size, confidence interval and minimal clinically important difference (MCID). J Bodyw Mov Ther. 2019;23(4):690-4. http://dx.doi.org/10.1016/j.jbmt.2019.02.009. PMid:31733748.
26. Agle CG, Sá CKC, Amorim DS Fo, Figueiredo MAM. Avaliação da efetividade do uso de meias de compressão na prevenção do edema ocupacional em cabeleireiras. J Vasc Bras. 2020;19:e20190028. http://dx.doi.org/10.1590/1677-5449.190028.
27. Martinez-Zapata MJ, Vernooij RW, Simancas-Racines D, et al. Phlebotonics for venous insufficiency. Cochrane Database Syst Rev. 2020;11:CD003229. PMid:33141449.
28. Ioannidis JP. Why most published research findings are false. PLoS Med. 2005;2(8):e124. http://dx.doi.org/10.1371/journal.pmed.0020124. PMid:16060722.
29. John LK, Loewenstein G, Prelec D. Measuring the prevalence of questionable research practices with incentives for truth telling. Psychol Sci. 2012;23(5):524-32. http://dx.doi.org/10.1177/0956797611430953. PMid:22508865.
30. Nature. Journals unite for reproducibility [editorial]. Nature. 2014;515:7. http://dx.doi.org/10.1038/515007a. PMid:25373636.
31. Amrhein V, Korner-Nievergelt F, Roth T. The earth is flat (p > 0.05): significance thresholds and the crisis of unreplicable research. PeerJ. 2017;5:e3544. http://dx.doi.org/10.7717/peerj.3544. PMid:28698825.
32. Gao J. P-values - a chronic conundrum. BMC Med Res Methodol. 2020;20(1):167. http://dx.doi.org/10.1186/s12874-020-01051-6. PMid:32580765.
33. Ioannidis JPA. What have we (not) learnt from millions of scientific papers with P values? Am Stat. 2019;73(1):20-5. http://dx.doi.org/10.1080/00031305.2018.1447512.
34. Pinto VF. Estudos clínicos de não-inferioridade: fundamentos e controvérsias. J Vasc Bras. 2010;9(3):145-51. http://dx.doi.org/10.1590/S1677-54492010000300009.
35. Streiner DL. Statistics Commentary Series: Commentary #12-One--Tailed and Two-Tailed Tests. J Clin Psychopharmacol. 2015;35(6):628-9. http://dx.doi.org/10.1097/JCP.0000000000000423. PMid:26479225.
36. Ringwalt C, Paschall MJ, Gorman D, Derzon J, Kinlaw A. The use of one- versus two-tailed tests to evaluate prevention programs. Eval Health Prof. 2011;34(2):135-50. http://dx.doi.org/10.1177/0163278710388178. PMid:21138911.
37. Ludbrook J. Should we use one-sided or two-sided P values in tests of significance? Clin Exp Pharmacol Physiol. 2013;40(6):357-61. http://dx.doi.org/10.1111/1440-1681.12086. PMid:23551169.
38. Sullivan GM, Feinn R. Using effect size-or why the P value is not enough. J Grad Med Educ. 2012;4(3):279-82. http://dx.doi.org/10.4300/JGME-D-12-00156.1. PMid:23997866.

Correspondência
Hélio Miot
Universidade Estadual Paulista – UNESP, Faculdade de Medicina – FMB, Departamento de Infectologia, Dermatologia, Diagnóstico por Imagem e Radioterapia
Campus Universitário de Rubião Jr., S/N
CEP 18618-000 - Botucatu (SP), Brasil
Tel.: (14) 3811-6015
E-mail: heliomiot@gmail.com

Informações sobre os autores
ACM - Dermatologista, Mestre e PhD, Faculdade de Medicina, Universidade Estadual Paulista (FMB-UNESP), Campus de Botucatu.
HAM - Dermatologista, PhD, Faculdade de Medicina, Universidade de São Paulo (FM-USP); Livre-docente, Faculdade de Medicina (FMB-UNESP), Campus de Botucatu.