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

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

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

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

How to cite: Miola AC, Miot HA. P-value and effect-size in clinical and experimental studies. J Vasc Bras. 2021;20:e20210038. https://doi.org/10.1590/1677-5449.210038

The complex nature of biological systems causes a certain degree of sample variation in many experiments. Moreover, most biomedical interventions promote moderate effects that do not have an obvious dose-response slope. As a result, when statistics are used to determine the difference between samples, the combination of large measurement variations and modest differences between groups compromises their analytical power (type II error). This means it is imperative to interpret *p*-values (statistical significance) and effect sizes with great care when making inferences from the results of studies that make comparisons between groups, although these concepts are also applicable to analyses of correlation, agreement, survival, and diagnostic tests, among others.[1-5]

According to frequentist statistics, two or more samples may be drawn from the same population, but nevertheless show a certain variability in some of their characteristics. The greater the similarity between the samples, the greater the likelihood that they will be of the same nature; while the flip side is that samples that are very different will be less likely to have been chosen at random, from within the same population. Statisticians have developed a series of mathematical models that estimate the probability that samples belong to the same population and the differences observed between them in an experiment have occurred by chance. As a general rule, the *p*-value of a statistical test reflects the theoretical probability that values more extreme than those observed are the result of chance, as long as the groups tested are truly equal (H_0 is true).[6,7]

It is the researchers' responsibility to define a cutoff point beyond which they can consider that the *p*-value denotes a low enough probability that the groups can be assumed to be different. The choice of this significance level (level α) and the decision on the direction of analysis (one-tailed or two-tailed), should be based on theoretical principles and should be defined before the analysis. This is of fundamental importance, because every cutoff point chosen has the potential to sacrifice conclusions derived from results very close to this limit. For example, if the cutoff point chosen is $p < 0.05$, $p = 0.04$, it is overvalued in detriment to $p = 0.06$.[8]

In tests comparing groups, the *p*-value is influenced by the difference between the means (or proportions), but also by the variance of the data and by the dimensions of the sample. Figure 1 illustrates three different situations, in which samples with variation in standard deviations and sample size are compared. Samples with the same mean and standard deviation have different *p*-values, depending on the sample sizes (Figure 1 A *vs*. B). In turn, samples with the same mean and sample size have different *p*-values if they differ only in terms of their standard deviation (Figure 1 A *vs*. C).

By convention, researchers adopt significance levels in the region of 5% ($p \leq 0.05$) for analysis of small samples ($n < 50$) and, by so doing, accept the risk that the result observed occurs by chance at least once in every 20 times the experiment is run.[9] Adoption of more stringent significance levels (for example, $p < 0.01$) increases the reproducibility of studies, but penalizes them with larger type II errors. However, since the sample size and the number of variables involved in the analysis (number of comparisons) influence the *p*-value, this should be carefully weighed up when choosing the significance level. Use of very large samples ($n > 1,000$) makes finding low *p*-values by

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

Financial support: None.
Conflicts of interest: No conflicts of interest declared concerning the publication of this article.
Submitted: March 05, 2021. Accepted: April 15, 2021.

The study was carried out at Departamento de Infectologia, Dermatologia, Diagnóstico por Imagem e Radioterapia, Faculdade de Medicina (FMB), Universidade Estadual Paulista (UNESP), Botucatu, SP, Brazil.

chance more likely, so it is recommended that more stringent significance levels be used, such as $p \leq 0.001$. Modern genetic experiments simultaneously compare thousands of variables, making detection of small *p*-values by chance more likely, so it is recommended that significance levels of the order of $p < 5x10^{-8}$ should be adopted.[10,11]

The *p*-values produced by a statistical test should be reported as their exact values, with a number of decimal places compatible with the magnitude that is being evaluated. For example, $p = 0.032$ should be reported, rather than $p < 0.05$ or $p = 0.032016$.[12,13] Increasing the number of decimal places is not proof that the results are more important or reliable. Moreover, marginal *p*-values, that are borderline to the significance level (for example, $p = 0.067$), should not be interpreted as a "trend" to rejection of the null hypothesis, since expanding the sample does not guarantee that the difference between groups will be maintained.[14]

It is, therefore, important that the *p*-value should not be used as a measure of the validity of a result or of the strength of an association.[15] Neither should *p*-values larger than the significance level (for example, $p > 0.1$) be interpreted as showing that the samples are identical.[7] One additional measure for understanding the relationship between the groups sampled is provided by estimators of effect size.[16]

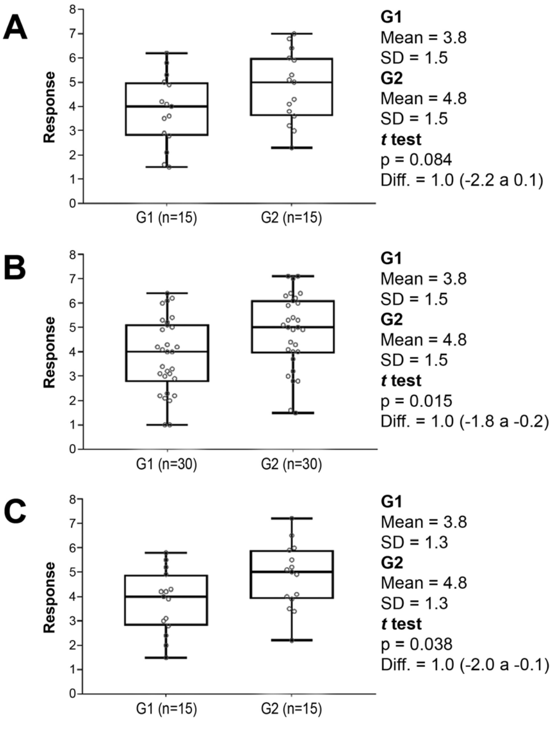

Figure 1. Hypothetical examples of (bidirectional) comparisons between two treatment groups (G1 and G2), all with the same means and medians. (A) Sample with 15 participants per group (p = 0.08); (B) Sample with 30 participants per group and the same standard deviation as in example A (p = 0.02); (C) Sample with 15 participants per group and a smaller standard deviation than example A (p = 0.04).

Assuming that the samples are adequately representative of a population (randomized collection), their statistics can be used to estimate parameters of that population, enabling inferences to be made about the behavior of the variables studied. Effect size is an indicator that quantifies the difference between samples, and an estimation of its 95% confidence interval (95%CI) provides a measure of the uncertainty of the behavior of that parameter in the population from which the sample was drawn, providing more valuable information about the true behavior of the phenomenon studied than the *p*-value offers.[17,18]

Table 1 lists the most important indicators of effect size used in epidemiological studies, which should be presented together with the *p*-value in the results of statistical tests, although the independent meaning of each of them is beyond the scope of this text.[19] There are other estimators of effect size, which are more often used in experimental studies and which are less intuitive to interpret. These include Cohen's "d" coefficient ", R^2, and omega and "eta" squared (ω^2 and η^2), which may require help from an experienced statistician.[18,20]

Every statistical test should be presented (and interpreted) according to its *p*-value, an effect size, and its 95%CI.[12,13,21,22] An experiment that results in a large effect size and a *p*-value = 0.06 is undoubtedly more relevant than a result with a small effect size but $p < 0.01$.[23-25]

For example, a recent study that assessed the effectiveness of compression stockings for improving occupational edema found a result with $p < 0.0001$.[26] However, the non-availability of reduction values as an effect size (for example, reduction in the diameter of the ankle in the evening, or VEINES scores) makes it difficult to interpret the data and their inferences with a view to clinical use.

Furthermore, particularly when dealing with larger samples, detection of low *p*-values may not indicate a clinically sensitive effect that leads to changes to medical paradigms. In an important systematic review conducted by Martinez-Zapata et al.[27] on the subject of phlebotonics for venous insufficiency, it was suggested that phlebotonics are superior, on the basis of their statistical significance ($p < 0.05$), but the effect size observed was the result of a mean reduction of just 4.27 mm (95%CI 2.93–5.61 mm) in ankle circumference in 2,010 participants (15 studies),

Table 1. Principal measures of effect according to the type of epidemiological study.

Type of study	Effect size
Diagnostic	Sensitivity, specificity, positive (or negative) predictive value, likelihood ratio, area under the ROC curve
Ecological	Correlation coefficients (r or rho)
Case-control	Odds ratio, prevalence ratio
Survival	Hazard ratio
Clinical trial/cohort study	Relative risk, attributable risk, reduction in relative risk, absolute risk reduction, number needed to treat (or to harm), absolute difference between groups (percentages or means).

ROC = receiver operating characteristic.

which, although true, does not indicate an evident benefit for patients with edema of the lower limbs.

Occasionally, there may be a discrete divergence between the amplitude of the effect size and the *p*-value. For example, a relative risk of 0.70 (95%CI 0.36–1.01) and a *p*-value = 0.045. However, this should not be considered an error, since the estimates originate from different calculations and tend to converge as sample sizes increase.

There is a recent academic movement in favor of total abolition of *p*-values and of the term “statistically significant” from scientific publications, giving preference to exclusively reporting the effect size of a test, because it is more informative and allows generalization of results.[28] Undoubtedly, studies that base their conclusions entirely on the *p*-value are more susceptible to non-reproducibility, in addition to encouraging researchers to pursue statistical significance in detriment to the relevance of the result (“p-hacking”).[23,28-31] However, this is still an incipient movement among researchers, since a campaign for correct interpretation of *p*-values analyzed in conjunction with effect sizes is a more correct option than abolishing p-values.[32,33]

Finally, comparisons between groups can be assessed either unidirectionally or bidirectionally (one-tailed or two-tailed). A test is usually called a difference study if we are assessing the behavior of a variable that can be larger or smaller between samples. However, many assessments are by their nature unidirectional, such as a comparison of the number of cases of a disease between people who have been vaccinated and those who have not; or a test of non-inferiority comparing two treatments.[34] In these examples, the possibility that the result could be considered bidirectionally is not part of the research hypothesis. However, use of one-tailed analyses is not consensus among epidemiologists, because, although they have greater statistical power and need smaller sample sizes, they increase the chance of type I error.[35-37] These analyses require supervision by an experienced statistician to calculate the one-tailed *p*-value and 95%CI.

While the size of the *p*-value can inform a reader whether there is a significant effect, it does not reveal the extent of the impact of this effect on the variables studied.[38] Researchers must therefore be cautious about the results of statistical tests, in the sense that the p-value should be interpreted in conjunction with the effect size, in particular as estimated by the 95% confidence interval, since the pragmatic significance of an experiment is an information that is independent of its statistical significance.

REFERENCES

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.

Correspondence
Prof. Hélio Miot
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.: +55 (14) 3811-6015
E-mail: heliomiot@gmail.com

Author informations
ACM - Dermatologist, MSc and PhD, Faculdade de Medicina, Universidade Estadual Paulista (FMB-UNESP), Campus de Botucatu.
HAM - Dermatologist, PhD, Faculdade de Medicina, Universidade de São Paulo (FM-USP); Tenured professor, Faculdade de Medicina (FMB-UNESP), Campus de Botucatu.