SAÚDE E AMBIENTE
V.9 • N.1 • 2022 - Fluxo Contínuo

ISSN Digital: **2316-3798**
ISSN Impresso: **2316-3313**
DOI: **10.17564/2316-3798.2022v9n1p230-239**

INTERFACES CIENTÍFICAS

A SIGNIFICÂNCIA ESTATÍSTICA E O USO DO P-VALOR EM PESQUISAS NA SAÚDE: RECOMENDAÇÕES PRÁTICAS

STATISTICAL SIGNIFICANCE AND THE USE OF P-VALUE IN HEALTH RESEARCH: PRACTICAL RECOMMENDATIONS

SIGNIFICACIÓN ESTADÍSTICA Y EL USO DEL VALOR P EN LA INVESTIGACIÓN EN SALUD: RECOMENDACIONES PRÁCTICAS

Dalson Britto Figueiredo Filho[1]
Lucas Silva[2]

RESUMO

O p-valor pode ser definido como uma probabilidade que informa o nível de incompatibilidade dos dados observados com um modelo teórico esperado. Por essa razão, atua como um dos principais parâmetros de significância estatística de pesquisas empíricas. Contudo, a utilização incorreta, associada a problemas como viés de publicação e ausência de padrões específicos de reprodutibilidade, tem gerado problemas em áreas do conhecimento. Objetivo do trabalho é discutir aspectos ligados à importância do p-valor nas pesquisas empíricas. O estudo é teórico-reflexivo, baseado nas recomendações da Associação Americana de Estatística sobre a correta interpretação do p-valor. Além disso, discutimos o papel da significância estatística a partir de uma perspectiva empírica. Em particular, o p-valor: (1) não informa a probabilidade de que a hipótese nula é verdadeira; (2) não indica que os resultados foram produzidos aleatoriamente; (3) não estima o tamanho do efeito observado; (4) não mensura a importância substantiva dos resultados; (5) nunca deve ser interpretado sozinho; (6) não deve ser interpretado quando os pressupostos de seu cálculo forem violados e (7) não pode ser interpretado quando se trabalha com a população. A discussão crítica sobre a utilização de testes de significância é sinal de maturidade estatística. Contudo, os pesquisadores não podem decidir sobre como utilizar o p-valor antes de compreenderem integralmente o seu papel na pesquisa empírica.

PALAVRAS-CHAVE

Análise Estatística. Interpretação Estatística de Dados. P-valor. Análise Quantitativa. Viés Estatístico.

ABSTRACT

The p-value can be defined as a probability that informs the level of incompatibility of the observed data with an expected theoretical model. For this reason, it acts as one of the main parameters of statistical significance of empirical research. However, incorrect use, associated with problems such as publication bias and the absence of specific reproducibility standards, has generated problems in areas of knowledge. The objective of this work is to discuss aspects related to the importance of p-value in empirical research. The study is theoretical-reflective, based on the recommendations of the American Statistical Association on the correct interpretation of the p-value. Furthermore, we discuss the role of statistical significance from an empirical perspective. In particular, the p-value: (1) does not report the probability that the null hypothesis is true; (2) does not indicate that the results were produced randomly; (3) it does not estimate the size of the observed effect; (4) it does not measure the substantive importance of results; (5) should never be interpreted alone; (6) it should not be interpreted when the assumptions of its calculation are violated and (7) it cannot be interpreted when working with the population. Critical discussion about the use of significance tests is a sign of statistical maturity. However, researchers cannot decide how to use the p-value until they fully understand its role in empirical research.

KEYWORDS

Statistical Analysis; Statistical Interpretation of Data; p-value; Quantitative analysis; Statistical Bias

RESÚMEN

El valor p se puede definir como una probabilidad que informa el nivel de incompatibilidad de los datos observados con un modelo teórico esperado. Por ello, actúa como uno de los principales parámetros de significación estadística de la investigación empírica. Sin embargo, el uso incorrecto, asociado a problemas como el sesgo de publicación y la ausencia de estándares específicos de reproducibilidad, ha generado problemas en áreas de conocimiento. El objetivo de este trabajo es discutir aspectos relacionados con la importancia del p-valor en la investigación empírica. El estudio es teórico-reflexivo, basado en las recomendaciones de la American Statistical Association sobre la correcta interpretación del p-valor. Además, discutimos el papel de la significación estadística desde una perspectiva empírica. En particular, el valor p: (1) no informa la probabilidad de que la hipótesis nula sea verdadera; (2) no indica que los resultados se produjeron al azar; (3) no estima el tamaño del efecto observado; (4) no mide la importancia sustantiva de los resultados; (5) nunca debe interpretarse solo; (6) no debe interpretarse cuando se violan los supuestos de su cálculo y (7) no puede interpre-

tarse cuando se trabaja con la población. La discusión crítica sobre el uso de pruebas de significación es un signo de madurez estadística. Sin embargo, los investigadores no pueden decidir cómo usar el valor p hasta que comprendan completamente su papel en la investigación empírica.

PALABRAS CLAVE

Análisis Estadístico; Interpretación Estadística de Datos; valor p; Análisis cuantitativo; Sesgo estadístico

1 INTRODUÇÃO

Em nove de junho de 2016, a Associação Americana de Estatística (American Statistical Association – ASA) publicou um *statement* sobre o papel do p-valor na pesquisa científica (WASSERSTEIN; LAZAR, 2016). A produção do documento foi motivada, entre outras razões, pelo uso incorreto da significância estatística em diversas áreas do conhecimento (STODDART, 2016). Problemas como viés de publicação e ausência de padrões específicos de reprodutibilidade também contribuíram para a elaboração da cartilha oficial.

Alguns periódicos estão debatendo a necessidade de abandonar a significância estatística dos trabalhos (HALSEY, 2019; NATURE, 2019). A principal justificativa é de que o p-valor sozinho não fornece evidências confiáveis a respeito de um determinado modelo ou hipótese de pesquisa (STODDART, 2016).

Afinal, para que serve o p-valor? Este artigo sumariza as principais recomendações da ASA sobre a correta interpretação do p-valor. Em particular, defendemos que: o p-valor (1) não informa a probabilidade de que a hipótese nula é verdadeira; (2) não indica que os resultados foram produzidos aleatoriamente; (3) não estima o tamanho do efeito; (4) não mensura a importância substantiva dos resultados; (5) não deve ser interpretado sozinho quando houver outras ferramentas disponíveis; (6) não deve ser interpretado quando os pressupostos de seu cálculo forem violados e (7) não pode ser interpretado quando se trabalha com a população.

2 MÉTODOS

Trata-se de um estudo teórico-reflexivo baseado nas recomendações da ASA sobre o papel do p--valor na pesquisa científica. Além disso, discutiu-se o papel da significância estatística a partir de uma perspectiva empírica a partir de experiências pedagógicas.

3 RESULTADOS E DISCUSSÃO

Para a cartilha oficial da ASA, "*a p-value is the probability under a specified statistical model that a statistical summary of the data would be equal to or more extreme than its observed value*" (WASSERSTEIN; LAZAR, 2016, p. 131). Do ponto de vista histórico, é atribuída a Ronald Fisher sua criação, em 1925 (FISHER, 1992). Operacionalmente, o p-valor pode ser definido como uma probabilidade que informa o nível de incompatibilidade dos dados observados com um modelo teórico esperado.

Assim como na Matemática, os testes estatísticos procuram provar por contradição (PEREIRA, 1995). Usualmente, tem-se duas hipóteses rivais: a nula (H_0) e a alternativa (H_a). Por exemplo, em um estudo sobre gênero e renda, a hipótese nula assume que homens e mulheres recebem o mesmo salário, enquanto a alternativa postula que a renda dos homens, em média, é maior.

Quanto menor o p-valor, maior é a incompatibilidade entre os resultados observados e a hipótese nula, assumindo que ela é verdadeira. Esse nível de incompatibilidade gera suspeição sobre a plausibilidade da hipótese nula. Portanto, quanto maior a incompatibilidade entre o valor observado e o valor esperado, mais confiante o pesquisador estará em rejeitar a hipótese nula. Em nosso exemplo, um p-valor muito pequeno indicaria que os dados observados são incompatíveis com a hipótese nula de igualdade de rendimentos entre homens e mulheres.

Mas quão pequeno deve ser o p-valor para rejeitar a hipótese nula? Cuidado: não existe patamar objetivo de rejeição. A comunidade científica utiliza três principais critérios: 1%, 5% e 10%. Em particular, o vocábulo "significância estatística" geralmente vem acompanhado de p-valor<0,05. Esse limiar é arbitrário e reflete a preferência pessoal de Fisher (1992). Gelman e Stern (2006) argumentam que a diferença entre um resultado significativo e um não significativo em si pode não ser estatisticamente significante. Isso porque a inclusão/exclusão de uma variável e/ou a inserção/remoção de alguns casos pode alterar o nível de significância observado e, dessa forma, modificar radicalmente as conclusões de pesquisa.

Depois de apresentar o conceito, o próximo passo é aprender para que serve o p-valor. Para fixar a compreensão, iremos apresentar como não o interpretar. O Quadro 1 sumariza essas recomendações.

Quadro 1 – Sete coisas que se deve saber sobre o p-valor

1	O p-valor não indica a probabilidade de que a hipótese nula é verdadeira
2	O p-valor não indica que os resultados foram produzidos aleatoriamente
3	O p-valor não indica o tamanho do efeito observado
4	O p-valor não mensura a importância substantiva dos resultados observados
5	O p-valor não deve ser interpretado sozinho quando houver outras ferramentas disponíveis
6	O p-valor não deve ser interpretado quando os pressupostos de seu cálculo forem violados
7	O p-valor não deve ser interpretado quando se trabalha com a população

Fonte: elaborado pelos autores a partir de Wasserstein e Lazar (2016), Stoddart (2016) e Pereira (1995).

3.1 O P-VALOR NÃO INDICA A PROBABILIDADE DE QUE A HIPÓTESE NULA É VERDADEIRA

De acordo com Goodman (2008), esse é o equívoco mais comum na interpretação do p-valor. Um valor de p de 0,01 não indica que a hipótese nula tem uma probabilidade de 1% de estar correta. Isso porque o cálculo do p-valor já assume que a hipótese nula é verdadeira, logo, não pode indicar se ela é correta ou não. Pela nossa experiência pedagógica, esse erro é muito comum entre alunos de graduação e pós-graduação, mas também acomete pesquisadores mais experientes. Por exemplo, Diamond e Forrester (1979), a partir de uma amostra de 24 médicos cardiologistas, reportam que 50% foram incapazes de identificar corretamente o significado do p-valor em um questionário com múltiplas escolhas.

3.2 O P-VALOR NÃO INDICA QUE OS RESULTADOS FORAM PRODUZIDOS ALEATORIAMENTE

Esse equívoco é recorrente e pode ser interpretado como uma extensão do erro anterior (GREENLAND; POOLE, 2013). Por exemplo, um valor de p de 0,06 não significa que existe uma probabilidade de 6% de que os resultados observados foram produzidos ao acaso. Esse problema ocorre porque os pesquisadores utilizam o p-valor para inferir a respeito da veracidade da hipótese nula ou sobre o papel da aleatoriedade na produção do resultado observado. O p-valor não é nenhum dos dois. De acordo com a ASA, "*a p-value provides one approach to summarizing the incompatibility between a particular set of data and a proposed model for the data*"[3] (WASSERSTEIN; LAZAR, 2016, p. 131).

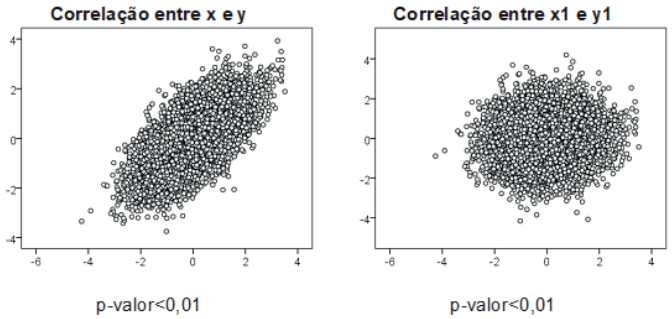
3.3 O P-VALOR NÃO INDICA O TAMANHO DO EFEITO OBSERVADO

Pela nossa experiência pessoal, esse erro é especialmente recorrente em congressos profissionais. Em geral, ouve-se a expressão de que um coeficiente de regressão foi "altamente significativo" como sinônimo da magnitude do efeito. Essa interpretação também é errada. O p-valor pequeno não necessariamente implica na presença de um efeito grande/importante e p-valor alto também não é sinônimo de falta de importância ou ausência de efeito. Por exemplo, em amostras muito pequenas, apenas efeitos de grande magnitude tendem a ser detectados. Em amostras excessivamente grandes, qualquer efeito será detectado, independentemente do tamanho.

Para fixar a compreensão desse ponto simulamos uma amostra com 10.000 observações e quatro variáveis com diferentes níveis de associação. Tecnicamente, a correlação entre x e y é de 0,666, enquanto a correlação entre x1 e y1 é de 0,05. A Figura 1 ilustra a dispersão dos respectivos pares.

3 "um valor-p fornece uma abordagem para resumir a incompatibilidade entre um determinado conjunto de dados e um modelo proposto para os dados"

Figura 1 – Diferentes níveis de correlação



Fonte: elaborado pelos autores

Mesma significância, relações totalmente diferentes. Essa é a lição que devemos aprender a partir da inspeção gráfica dos resultados da simulação. A cartilha da ASA afirma categoricamente que "*scientific conclusions and business policy decisions should not be based only whether a p-value passes a specific threshold*"[4] (WASSERSTEIN; LAZAR, 2016, p. 131). Infelizmente, muitos pesquisadores ainda utilizam o patamar arbitrário de 0,05 para classificar os resultados de forma binária: significativo ou não-significativo. Esse procedimento é errado e "*leads to considerable distortion of the scientific progress*"[5]. Não é o tamanho do p-valor que informa a contribuição de um determinado trabalho científico. Resultados significativos são tão importantes quanto aqueles que não rejeitam a hipótese nula (MEHLER *et al.*, 2019; MEHTA, 2019).

3.4 O p-valor não mensura a importância substantiva dos resultados observados

De acordo com a cartilha da ASA, "*statistical significance is not equivalent to scientific, human, or economic significance*"[6] (WASSERSTEIN; LAZAR, 2016, p. 132). Não devemos confundir significância estatística com significância substantiva (MOORE *et al.*, 2017). Um p-valor de 0,001 não é melhor, nem mais importante do que um de 0,1. Além disso, a magnitude do p-valor é fortemente influenciada pelo tamanho da amostra. Dessa forma, se a quantidade de observações é extremamente grande, o p-valor tende a diminuir, independentemente do tamanho do efeito ou da diferença existente entre os grupos (HAIR *et al.*, 2009). No limite, com uma amostra excessivamente grande, qualquer diferença será estatisticamente significativa a despeito de sua magnitude.

3.5 O p-valor não deve ser interpretado sozinho quando houver outras ferramentas disponíveis

4 "conclusões científicas e decisões de política de negócios não devem ser baseadas apenas se um valor-p ultrapassa um limite específico"

5 "leva a uma distorção considerável do progresso científico"

6 "significado estatístico não é equivalente ao significado científico, humano ou econômico"

Essa recomendação é ponto pacífico na literatura (GOODMAN, 2008; GREENLAND; POOLE, 2013; WASSERSTEIN; LAZAR, 2016). Por exemplo, antes de inferir a respeito da natureza de uma determinada distribuição, deve-se examinar o histograma ou *boxplot* dos dados. Similarmente, antes de concluir a respeito da significância de uma correlação, deve-se examinar o gráfico de dispersão. Ele se aplica a qualquer teste de significância já que diferentes elementos podem induzir um resultado significativo falso (falso positivo) ou omitir um resultado significativo verdadeiro (falso negativo). Por exemplo, em análise de séries temporais, ninguém deve (ou deveria) concluir nada a partir da significância das estimativas sem primeiramente examinar a distribuição dos dados e os gráficos de autocorrelação.

A interpretação correta do p-valor também depende da transparência dos resultados. Tecnicamente, existem várias formas de produzir um resultado estatisticamente significativo (*P-hacking*). Os Planos de Pré-Análise e os registros prévios dos estudos científicos reduzem a probabilidade de encontrar falsos positivos. Uma das formas de combater o viés de publicação é exigir dos autores não apenas os dados, mas o passo a passo das análises estatísticas. No original, "*valid scientific conclusions based on p-values and related statistics cannot be drawn without at least knowing how many and which analyses were conducted, and how those analyses were selected for reporting*"[7] (WASSERSTEIN; LAZAR, 2016, p. 132).

3.6 O P-VALOR NÃO DEVE SER INTERPRETADO QUANDO OS PRESSUPOSTOS DE SEU CÁLCULO FOREM VIOLADOS

Algumas violações são mais graves do que outras. E alguns testes são mais robustos do que outros. Antes de atribuir muita importância ao p-valor, o pesquisador deve analisar criticamente em que medida os dados utilizados respeitam os pressupostos do seu cálculo. Por exemplo, assume-se que as observações foram aleatoriamente selecionadas e são independentes entre si. Portanto, faz pouco sentido examinar o p-valor em uma amostra por conveniência (FIGUEIREDO FILHO *et al.*, 2014), já que amostras não aleatórias tendem a produzir estimativas enviesadas dos parâmetros populacionais. Por outro lado, sempre que os pressupostos forem devidamente respeitados, o p-valor tende a cumprir o seu papel de informar o nível de incompatibilidade dos dados observados com um modelo teórico esperado.

3.7 O P-VALOR NÃO DEVE SER INTERPRETADO QUANDO SE TRABALHA COM A POPULAÇÃO

Esse é um ponto controverso entre Frequentistas e Bayesianos (GELMAN; STERN, 2006). A utilização de amostras se justifica pela economia de tempo e recursos. E, se forem corretamente coletadas, produzem estimativas confiáveis a respeito dos parâmetros populacionais. Para Hair e colaboradores (2009), a utilização do censo populacional torna a inferência estatística desnecessária já que qualquer efeito ou diferença observada, por menor que seja, existe na população.

O papel da Estatística é utilizar amostras para realizar inferências válidas para a população. Se a amostra é igual a população, não existe necessidade de estimação já que os parâmetros populacionais já são conhecidos. Imagine o censo de uma população em que o salário dos homens é R$

7 "conclusões científicas válidas baseadas em valores p e estatísticas relacionadas não podem ser tiradas sem pelo menos saber quantas e quais análises foram conduzidas, e como essas análises foram selecionadas para relatório"

100,00 e as mulheres recebem R$ 99,99. Não faz sentido questionar se a diferença é estatisticamente significativa, pois não existe incerteza a respeito dos parâmetros populacionais. Não existe estimação na população, apenas na amostra.

4 CONCLUSÕES

O p-valor é a mais comum e importante medida de incerteza utilizada na pesquisa científica. Por outro lado, no entanto, é também a estatística mais mal interpretada. A discussão crítica sobre a utilização de testes de significância é sinal de maturidade estatística. Contudo, os pesquisadores não podem decidir sobre como utilizar o p-valor antes de compreenderem integralmente o seu papel na pesquisa empírica. Apesar de utilizar como referencial teórico-analítico um documento produzido por uma das maiores associações científicas sobre o assunto, a ausência de uma análise sistemática da literatura é a principal limitação do trabalho. Apesar disso, esperamos, com este artigo, difundir o debate sobre significância estatística no meio acadêmico nacional e facilitar a compreensão do p-valor na pesquisa científica.

REFERÊNCIAS

DIAMOND, G. A.; FORRESTER, J. S. Analysis of probability as an aid in the clinical diagnosis of coronary-artery disease. **New Engl J Med**, v. 300, n. 24, p. 1350-1358, 1979.

FIGUEIREDO FILHO, D. B. *et al*. Reply on the comments on when is statistical significance not significant? **Braz Politic Sci Rev**, v. 8, n. 3, p. 141-150, dez. 2014.

FISHER, R. A. Statistical methods for research workers. **In**: KOTZ, S.; JOHNSON, N. L. (ed.). **Breakthroughs in statistics: methodology and distribution. Springer Series in Statistics.** New York, NY: Springer, 1992.

GELMAN, A.; STERN, H. The difference between "significant" and "not significant" is not itself statistically significant. **Am Stat**, v. 60, n. 4, p. 328-331, 2006.

GOODMAN, S. A Dirty Dozen: twelve p-value misconceptions. **Semin Hematol**, v. 45, n. 3, p. 135-140, 2008.

GREENLAND, S.; POOLE, C. Living with p values: resurrecting a Bayesian perspective on frequentist statistics. **Epidemiol**, v. 24, n. 1, p. 62-68, 2013.

HAIR, J. F. *et al*. **Análise multivariada de dados**. Porto Alegre: Bookman Editora, 2009.

HALSEY, L. G. The reign of the p-value is over: what alternative analyses could we employ to fill the power vacuum? **Biol Lett**, v. 15, n. 5, p. e20190174, 2019.

MEHLER, D. *et al*. Appreciating the significance of non-significant findings in Psychology. **J Eur Psychol Student**, v. 10, n. 4, p. 1-7, 2019.

MEHTA, D. Highlight negative results to improve science. **Nature**, 4 out. 2019.

MOORE, D. S. *et al*. **Introduction to the practice of statistics**. 9th Ed. New York: WH Freeman, 2017.

NATURE. It's time to talk about ditching statistical significance. **Nature**, v. 567, n. 7748, p. 283-283, mar. 2019.

PEREIRA, B. Estatística em medicina: p-varicação. **Rev SOCERJ**, v. 8, n. 3, p. 73-78, 1995.

STODDART, C. Is there a reproducibility crisis in science? **Nature**, 25 mai. 2016.

WASSERSTEIN, R. L.; LAZAR, N. A. The ASA statement on p-values: context, process, and purpose. **Am Stat,** v. 70, n. 2, p. 129-133, 2016.

Recebido em: 18 de Fevereiro de 2022
Avaliado em: 22 de Julho de 2022
Aceito em: 10 de Agosto de 2022

A autenticidade desse artigo pode ser conferida no site https://periodicos.set.edu.br

1 Doutor em Ciência Política; Professor Associado I, Departamento de Ciência Política, Programa de Pós-graduação em Ciência Política, Universidade Federal de Pernambuco. ORCID: 0000-0001-6982-2262.
E-mail: dalson.figueiredofo@ufpe.br

2 Bacharel em Ciência Política e Acadêmico de Medicina, Universidade Estadual de Ciências da Saúde do Estado de Alagoas. ORCID: 0000-0002-5013-6278.
E-mail: lucas.silva@academico.uncisal.edu.br