

REVISÃO SISTEMÁTICA E METASSÍNTESE DE ARTIGOS SOBRE TESTES DE SIGNIFICÂNCIA: UM TEXTO PARA PROFISSIONAIS DE EDUCAÇÃO FÍSICA

Claudio Damião Rosa^{1*} Talisson Santos Chaves²

¹ Instituto Federal do Norte de Minas Gerais (IFNMG), Araçuaí, Minas Gerais;

² Instituto Federal Baiano (IFBaiano), Itapetinga, Bahia

Resumo: Testes de significância continuam sendo o principal meio de analisar a ocorrência de diferenças entre grupos ou associação entre variáveis na população, utilizando dados amostrais. Isso inclui, mas não está limitado, a área da Educação Física. Diversos textos foram publicados em português sobre esses testes, contudo, nenhuma síntese desses textos foi realizada até o momento. Tal síntese favorece uma compreensão mais ampla, do que a leitura isolada de cada texto, sobre a interpretação de testes de significância. Diante disso, realizamos uma revisão sistemática e metassíntese de 26 artigos em português que tiveram como objetivo principal providenciar informações úteis à interpretação de testes de significância. As buscas, realizadas em 15 de abril de 2021 e atualizadas em 28 de agosto de 2023, ocorreram no Google Acadêmico cujo é especialmente adequado para identificação de textos em português. Dois autores participaram na seleção dos estudos em texto completo e na síntese qualitativa dos dados. Com base nessa metassíntese, (1) apresentamos conceitos fundamentais para interpretação de testes de significância, (2) discutimos as vantagens e desvantagens de utilizar valores arbitrários de p (por exemplo, $p < 0,05$), (3) explicamos como interpretar estimativas e intervalos de confiança e (4) apontamos uma relevante lacuna na literatura. Pesquisadores frequentemente interpretam valores de p , de acordo com pontos de corte arbitrários (comumente 0,05), ao invés de considerar a estimativa de efeito ou associação e seu intervalo de confiança. Tal abordagem ignora informações relevantes para decisão sobre a importância prática do achado e sua imprecisão (referente à inferência para população). Desse modo, nossa recomendação é que, quando possível, se determine antes da realização do estudo o que são efeitos ou associações importantes e que se interprete a estimativa do efeito ou associação e seu intervalo de confiança.

Palavras-chave: análise estatística; metodologia; revisão sistemática

* Claudio Damião Rosa. claudio2008ilheus@hotmail.com; Instituto Federal do Norte de Minas Gerais (IFNMG); Campus Araçuaí, 39.606-270, Araçuaí, Minas Gerais, Brasil.

SYSTEMATIC REVIEW AND METASYNTHESIS OF PAPERS ON SIGNIFICANCE TESTS: A TEXT FOR PHYSICAL EDUCATION PROFESSIONALS

Abstract: Significance tests continue to be the main means of analyzing the occurrence of differences between groups or associations between variables in the population, using sample data. Several texts have been published in Portuguese about these tests, however, no synthesis of these texts has been carried out to date. Such a synthesis favors a broader understanding, than the isolated reading of each text, regarding the interpretation of significance tests. Given this, we carried out a systematic review and metasyntesis of 26 articles in Portuguese whose main objective was to provide useful information for the interpretation of significance tests. The searches, carried out on April 15, 2021, and updated on August 28, 2023, took place on Google Scholar, which is especially suitable for identifying texts in Portuguese. Two authors participated in the selection of full-text studies and in the qualitative synthesis of data. Based on this metasyntesis, (1) we present fundamental concepts for interpreting significance tests, (2) we discuss the advantages and disadvantages of using arbitrary p values (for example, $p < 0.05$), (3) we explain how to interpret estimates and confidence intervals, and (4) we point out a relevant gap in the literature. Researchers often interpret p-values according to arbitrary cut-off points (commonly 0.05), rather than considering the estimated effect or association and its confidence interval. This approach ignores information relevant to decisions about the practical importance of the finding and its imprecision (referring to inference for the population). Therefore, our recommendation is that, when possible, determine before carrying out the study what are important effects or associations and interpret the estimate of the effect or association and its confidence interval.

Key words: statistical analysis; methodology; systematic review

Introdução

Realizar um teste de uma hipótese continua sendo o principal meio de analisar se há diferenças entre grupos ou correlação entre variáveis na população, utilizando dados amostrais (1–5). Nesse contexto, o valor encontrado na amostra é uma estimativa do valor a ser encontrado na população, isto é, o parâmetro. No caso do popular teste *t* de *Student*, por exemplo, a hipótese nula considera que não há diferença na população, entre os grupos comparados utilizando dados da amostra (2). Como um exemplo da área da Educação Física, Rocha et al. (6) realizaram testes de hipóteses para analisar a associação entre comportamento sedentário, características sociodemográficas, ocupacionais, estilo de vida e condições clínicas. O teste de hipótese é comumente chamado de teste de significância e o seu resultado pode ser classificado como *estatisticamente significativo* quando o valor de *p* é inferior a determinado critério, arbitrariamente determinado, frequentemente 0,05. Quando o valor de *p* é superior a esse critério diz-se que o resultado é *estatisticamente não significativo*.

Apesar de pouco informativo, quando utilizado de forma isolada, alguns pesquisadores focam no valor de *p* ao interpretar os resultados de testes de significância (2). Baker (7) explica que um valor de *p* = 0,05 significa que, se a hipótese nula for verdadeira, e todas as outras suposições feitas sobre os dados forem válidas, a chance de se ter obtido um resultado tão ou mais extremo que o observado é de 5%. Desse modo, o valor de *p* refere-se a uma probabilidade. Em outras palavras, quanto menor o valor de *p*, mais improvável que a hipótese nula seja verdadeira e, conseqüentemente, menores são as chances de incorrer em erro ao se aceitar a hipótese alternativa (5). Quanto a interpretação do teste de significância, é relevante entender os fatores que influenciam o valor de *p* (por exemplo, tamanho da amostra) e perceber que comumente há resultados mais informativos a serem interpretados conjuntamente ao valor de *p* (2). Recentemente, um artigo publicado no periódico científico *Nature* recomendou que pesquisadores interpretem a magnitude da estimativa do efeito ou associação e os limites de seu intervalo de confiança (1).

Contudo, interpretar dicotomicamente (*estatisticamente significativo versus estatisticamente não significativo*) o valor de *p* continua uma prática comum em diversas áreas do conhecimento (1,2,4,5). Ainda, pouco se sabe sobre as orientações disponíveis em artigos publicados em português sobre a interpretação de testes de significância. Artigos em português são importantes porque esse idioma é falado em 10 países, totalizando uma população mundial de centenas de milhões de falantes (8). A síntese desses artigos tem diversos benefícios: (a) favorece uma compreensão abrangente acerca das orientações, disponíveis em português, sobre a interpretação de testes de significância; (b) reduz o risco de replicação não intencional de conteúdo já existente sobre testes de significância; (c) permite identificar lacunas na literatura; e, (d) pode auxiliar na seleção de textos por professores universitários que ministram disciplinas nas quais o teste de significância é um conteúdo relevante.

Desse modo, pretendemos com o presente estudo responder a seguinte questão de pesquisa “Quais informações estão disponíveis em português, em periódicos revisados por pares, sobre a interpretação de testes de significância?”. Nessa direção, nosso objetivo foi sintetizar informações sobre a interpretação de testes de significância contidas em artigos publicados em português.

Método

Delineamento do Estudo

Esse estudo consiste em uma revisão sistemática da literatura com a realização de uma síntese qualitativa do conteúdo dos artigos incluídos, isto é, metassíntese (10–12). Seguindo orientações do PRISMA-P (*Preferred reporting items for systematic review and meta-analysis protocols*) disponíveis em Shamseer et al. (13), um protocolo de revisão foi desenvolvido abrangendo a questão de pesquisa, critérios de inclusão, estratégia de busca e análise. Esse protocolo foi registrado no *Open Science Framework* (OSF) antes do início das

buscas. Decisões sobre a inclusão ou exclusão dos artigos em texto completo foram checadas por um coautor e a análise do conteúdo dos artigos incluídos foi realizada por dois autores independentemente. Esses autores alcançaram consenso sobre que estudos incluir na metassíntese e sobre a análise de conteúdo.

CrITÉRIOS de Inclusão

Foram incluídos artigos científicos publicados em periódicos revisados por pares que tiveram como objetivo principal providenciar informações úteis à interpretação de testes de significância. Isso inclui informações que favorecem o entendimento do que é hipótese nula e alternativa, valor de p , erros do tipo I e II, intervalos de confiança e estimativas de associação ou efeito no contexto de testes de significância. A intenção principal foi buscar artigos que apresentassem uma visão geral sobre a interpretação de testes de significância e não artigos com foco em temas específicos. Neste sentido, excluímos artigos com foco em testes e técnicas estatísticas específicas (p.ex., teste t ou ANOVA), em pressupostos (p.ex., normalidade) destes testes e em medidas de efeito específicas (p.ex., d de Cohen). Também foram excluídos artigos com foco em estatística Bayesiana. Nós consideramos um teste de significância, uma análise estatística que avalia a probabilidade de um dado resultado ter ocorrido, considerando-se a hipótese nula verdadeira.

Estratégia de Busca

As buscas foram realizadas na base de dados Google Acadêmico e os termos de busca estão organizados na Tabela 1. O Google Acadêmico favorece a identificação de artigos em português e é tão abrangente quanto uma combinação de diversas bases de dados (14). Adicionalmente, as referências dos artigos incluídos e nossos arquivos (por exemplo, pasta digital no computador) foram consultados em busca de artigos elegíveis.

Informações para a análise de conteúdo foram agrupadas com base em categorias determinadas *a priori*, a saber: (a) conceitos relevantes para interpretação de testes de significância; (b) interpretação do valor de p ; (c) vantagens e desvantagens de análises focadas em limiares arbitrários de p (por exemplo, $p \leq 0,05$); (d) interpretação de estimativas de associação e efeito e de seus intervalos de confiança; e, (e) evidências sobre a interpretação de testes de significância em artigos publicados em periódicos brasileiros.

:

Tabela 1 – Termos de busca utilizados no Google Acadêmico em 15 de abril de 2021 (resultados até 2021) e em 28 de agosto de 2023 (resultados de 2021 a 2023)

Termo de busca	Resultados até 2021	Resultados 2021 a 2023
allintitle: "Teste de hipóteses"	57	8
allintitle: "testes de significância"	16	1
allintitle: "valor de p"	12	2
allintitle: "intervalo de confiança"	52	5
allintitle: "intervalos de confiança"	90	9
allintitle: "significância estatística"	13	2
allintitle: "tamanho do efeito"	5	6
allintitle: "tamanhos do efeito"	5	0
allintitle: "magnitude do efeito"	9	1
allintitle: "Nível de significância"	16	1
allintitle: valor-p	12	4
allintitle: I OR II OR III "Erro do tipo"	4	2
allintitle: I OR II OR III "Erro tipo"	18	4
allintitle: "poder estatístico"	6	1
allintitle: "valores de p"	1	0

Extração dos Dados e Análise do Conteúdo

Os dois autores desse estudo leram todos os estudos incluídos de forma independente. O primeiro autor criou um arquivo do Word para cada estudo incluído e arquivos específicos para cada categoria em que o conteúdo dos artigos foi agrupado. Ainda, o primeiro autor escreveu a primeira versão do texto cujo foi editado pelo segundo autor até que os dois alcançaram consenso sobre quais informações seriam apresentadas e de que forma. Informações para a análise de conteúdo foram agrupadas com base em categorias determinadas a priori, a saber: (a) conceitos relevantes para interpretação de testes de significância; (b) interpretação do valor de p ; (c) vantagens e desvantagens de análises focadas em limiares arbitrários de p (por exemplo, $p \leq 0,05$); (d) interpretação de estimativas de associação e efeito e de seus intervalos de confiança; e, (e) evidências sobre a interpretação de testes de significância em artigos publicados em periódicos brasileiros.

Resultados e Discussão

De um total de quase 400 relatos identificados, 26 foram considerados elegíveis para o presente estudo (Figura 1). Descrevemos esses 26 artigos brevemente na Tabela 2. O ano de publicação dos artigos incluídos variou de 1984 a 2023, sendo divulgados em periódicos científicos de diversas áreas do conhecimento, incluindo: medicina, educação física, educação, agricultura, estudos comportamentais, e fisioterapia. Curiosamente, nenhum dos estudos incluídos em nossa revisão de literatura foram publicados em um periódico de destaque na área da psicologia; área conhecida por sua preocupação com métodos de pesquisa e estatística. A seguir sintetizamos o conteúdo dos artigos incluídos com base nas categorias determinadas *a priori*.

Conceitos Relevantes para Interpretação de Testes de Significância

Diversos conceitos foram descritos pelos artigos incluídos e podem ser considerados relevantes para a interpretação de testes de significância. Muitos conceitos foram descritos por mais de um artigo. Por exemplo, Hirakata et al. (15) explica que a estatística do teste “mede o grau de concordância entre uma amostra e a hipótese nula”. Nas palavras de Frei (16), a estatística do teste “mede a diferença do que foi observado entre a média amostral e o valor indicado na Hipótese Nula, a suposta média populacional”. Na Tabela 3, organizamos uma definição unificada de cada conceito que julgamos relevante para interpretação de testes de significância.

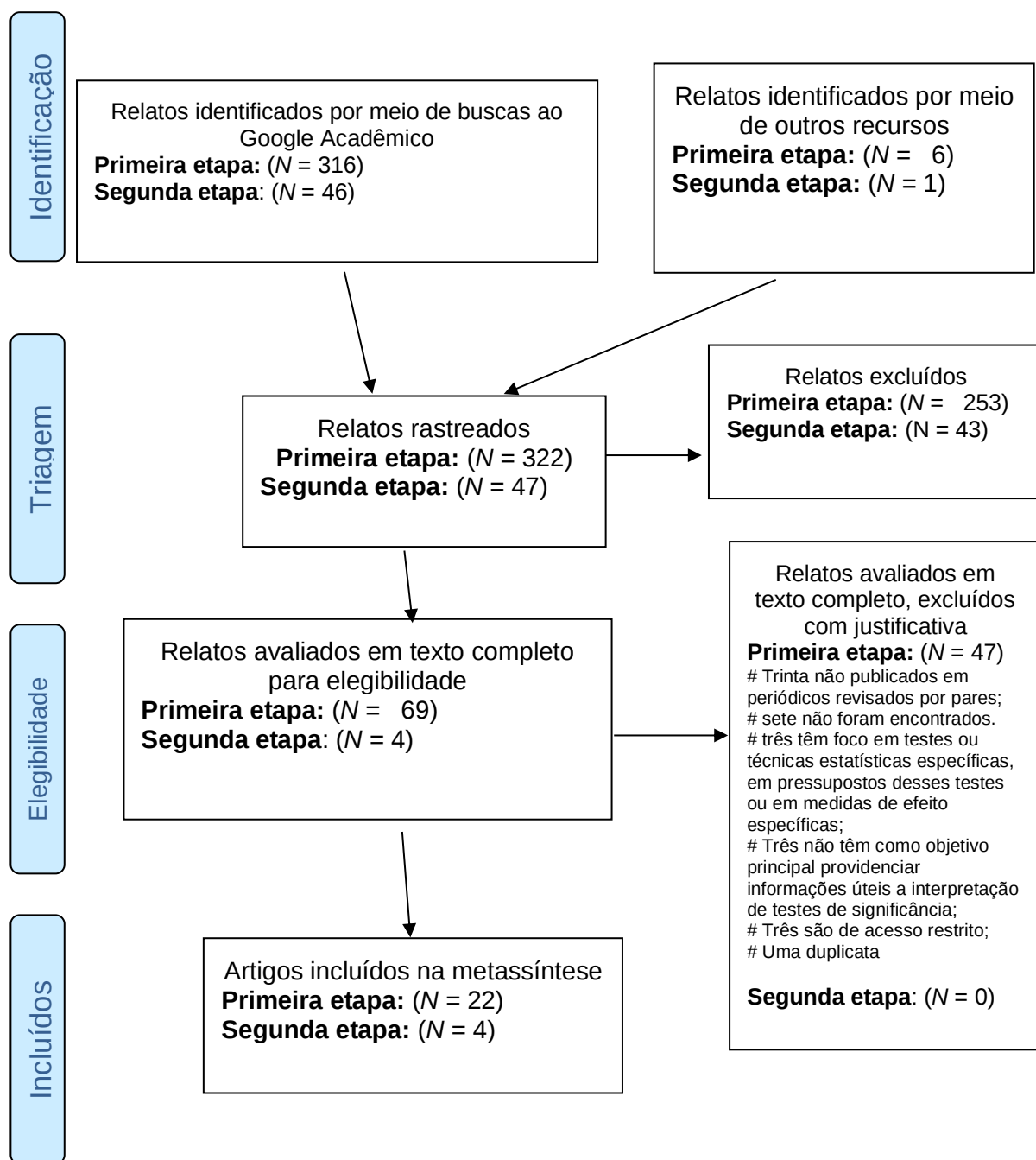


Figura 1 - Fluxograma Ilustrando o Processo de Busca e Seleção dos Artigos Incluídos na Metassíntese

Tabela 2 - Descrição Breve dos 26 Artigos Incluídos na Metassíntese

Estudo	Descrição
(17)	Descreve o conceito de tamanho de efeito e o relaciona com significância estatística. Também comenta sobre a relação entre tamanho da amostra, significância prática e significância estatística. Aponta que a utilização exclusiva do nível de significância para analisar e discutir os achados não é suficiente.
(5)	Discorre sobre a interpretação do valor de p , discutindo os seguintes pontos sobre esse 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.
(4)	Descreve conceitos relevantes para interpretação de testes de significância como valor de p . Explica as falácias da probabilidade inversa, replicação, tamanho do efeito, da significância clínica ou prática. Trata da importância dada ao teste de significância e da incapacidade de muitos profissionais da saúde de interpretá-lo.
(18)	Comenta sobre o pragmatismo de utilizar valores arbitrários de p . Explica porque estudos negativos ($p > 0,05$) não comprovam inexistência de uma associação ou efeito e porque estudos negativos e positivos ($p < 0,05$) não são necessariamente contraditórios. Apresenta sugestões sobre o reporte e interpretação do valor de p .
(19)	Descreve limitações associadas ao teste da hipótese nula sustentadas no valor de p . Discute as implicações de focar na significância prática ao invés da significância estatística. Apresenta proposta para utilização de técnicas de inferências baseadas na magnitude do efeito.
(16)	Avalia os efeitos do uso de planilha eletrônica interativa para simular eventos na aprendizagem do conceito estatístico inferencial do valor de p . Apresenta informações sobre como o valor de p é calculado, o que ele é, o que um valor de p pequeno significa e a relação entre rejeição da hipótese nula e “aceitação” da hipótese alternativa.
(15)	Discute questões referentes à metodologia de um teste de hipótese, tamanho de efeito, erros do tipo I, II e III, valor de p e poder estatístico.
(20)	Explica os conceitos de valor de p e hipótese nula. Apresenta informações ligadas a interpretação do ponto de corte 0,05 e ao valor exato de p . Comenta sobre a diferença entre significância estatística e significância clínica. Ilustra a relação entre valor- p e tamanho da amostra.
(21)	Apresenta os conceitos de valor de p e erros do tipo I e II. Comenta sobre o uso do ponto de corte 0,05, sobre as suposições associadas ao teste de significância e sobre a importância de se considerar intervalos de confiança e estimativas de efeito em associação ao valor de p .

Estudo	Descrição
(22)	Argumenta que o valor de p diz respeito apenas a probabilidade de rejeitar a hipótese nula. Comenta que na área da biomedicina é comum omitir estimativas de efeito e intervalos de confiança. Discute aspectos negativos da utilização do limiar $p < 0,05$ e aponta a interpretação de estimativas de efeito e sua imprecisão como abordagem mais apropriadas, no momento.
(23)	Comenta sobre o que é p e como interpretá-lo. Explica a diferença entre significância clínica e estatística, assim como alerta para o problema de se interpretar intervalos de confiança dicotomicamente.
(24)	Descreve método para comparação gráfica entre a diferença em médias de dois grupos. Esse método considera 95% do intervalo de confiança das médias e de sua diferença.
(25)	Comenta sobre a diferença entre significância estatística e significância prática. Aponta a relevância de considerar medidas de efeito e intervalos de confiança. Menciona alguns tipos de medidas de efeito e alerta sobre o uso de regras para sua interpretação.
(26)	Apresenta os conceitos de poder e tamanho do efeito. Comenta sobre a importância de reportar e interpretar estimativas de efeito e sobre a diferença entre significância clínica e estatística.
(27)	Apresenta o conceito de diversos termos relevantes. Discute os limites de interpretações baseadas no valor de p . Explica a interpretação de diversas estimativas de efeito para comparação de dois grupos. Comenta o uso de intervalos de confiança.
(28)	Utiliza um exemplo para ilustrar o que é o valor de p e como interpretá-lo. Explica como interpretar o ponto de corte de 0,05 e o valor exato de p . Diferencia significância estatística e significância clínica. Comenta sobre a interpretação equivocada de um valor de p estatisticamente não significativo como ausência de efeito.
(29)	Diferencia significância estatística de significância prática. Crítica decisões baseadas em limiares arbitrários de p , esclarecendo como o valor de p não deve ser interpretado. Sugere focar em estimativas de efeito e seus intervalos de confiança.
(30)	Explica o que é o intervalo de confiança e como interpretá-lo. Comenta sobre a vantagem do uso do intervalo de confiança sobre a interpretação exclusiva do valor de p . Diferencia população de origem de população de estudo.
(31)	Apresenta informações para a interpretação do valor de p e do ponto de corte 0,05. Aponta a diferença entre significância clínica e significância estatística e comenta sobre a importância de se reportar estudos que encontraram um valor de p superior a 0,05.
(32)	Descreve a lógica da tomada de decisão em testes de hipótese, conceitos como hipótese e hipótese nula, erros do tipo I e II, poder do teste e magnitude do efeito. Comenta sobre a diferença entre significância estatística e significância clínica.
(33)	Analisa erros cometidos por estudantes de Engenharia ao resolver questões referentes ao teste de hipótese. Aponta a importância de entender os conceitos relacionados ao teste de hipótese, tais como amostra, população, e erro do tipo II. Organiza o teste de hipótese em seis etapas.

Estudo	Descrição
(34)	Explica conceitos como população, parâmetro, estimativa e intervalo de confiança. Ilustra como a estimativa pode diferir do parâmetro e a relação entre intervalo de confiança, tamanho da amostra e nível de confiança. Explica o processo de decisão no teste de hipótese e aponta a necessidade de se apresentar as estimativas, intervalos de confiança e os valores exatos de p .
(35)	Discute a interpretação livre do valor de p ao invés de um ponto de corte delimitado a priori. Explica os conceitos de nível de significância, erro do tipo I e II, poder, e valor de p . Aborda o processo de formulação de hipóteses.
(36)	Esclarece conceitos estatísticos usando exemplos da literatura médica. Oferece alternativas ao teste de significância e fornece sugestões relacionadas a análise de dados.
(37)	Visa difundir a utilização de nível de significância de 10% para testes estatísticos de ensaios florestais. Descreve os conceitos de teste estatístico, hipótese de nulidade, e erros do tipo I e II. Comenta sobre a relação entre níveis de significância e erros do tipo II. Sugere que o ideal para melhorar a precisão dos dados é considerar um número ótimo de repetições e parcelas.
(38)	Sugere a utilização de níveis de significância de acordo com características do experimento. Descreve os conceitos de erro do tipo I, II, e III. Explica como interpretar diferenças estatisticamente significantes e valores prováveis na população com base nos dados da amostra.

Tabela 3 - Conceitos Relevantes para Interpretação de Testes de Significância Descritos pelos Artigos Incluídos, em Ordem Alfabética

Conceito	Descrição
Amostra ou população de estudo	Um ou mais indivíduos, eventos, ou objetos selecionados de um conjunto de indivíduos, eventos ou objetos de interesse.
Erro do tipo I ou alfa	Rejeitar a hipótese nula quando ela é verdadeira.
Erro do tipo II ou beta	Não rejeitar a hipótese nula quando ela é falsa.
Estatística de teste	Mede a diferença entre uma quantidade observada na amostra e o esperado sob a suposição de que a hipótese nula é verdadeira.
Estimativa ou estatística	Qualquer medida numérica calculada a partir de um conjunto de observações de uma amostra. Por exemplo, média ou proporção.
Inferência estatística	A generalização para a população das observações efetuadas em amostras.
Intervalo de confiança	A interpretação correta de um intervalo de confiança é baseada na noção hipotética de considerar os resultados que seriam obtidos se o estudo fosse repetido várias vezes. Se um estudo fosse repetido infinitamente e, em cada ocasião, fosse calculado um intervalo de confiança de 95%, então 95% desses intervalos conteriam o verdadeiro efeito ou associação.
Hipótese	Em estatística, uma hipótese é uma afirmação sobre o valor de um parâmetro de determinada população.
Hipótese alternativa ou hipótese experimental	Representa uma afirmação de que o parâmetro de interesse difere daquele definido na hipótese nula. Deve ser formulada de acordo com aquilo que o pesquisador deseja corroborar.
Hipótese nula ou hipótese de nulidade ou hipótese probanda	Comumente refere-se a afirmativa matemática de ausência de diferença entre parâmetros ou ausência de associação entre variáveis na população.
Nível de significância	É utilizado como um ponto de corte na probabilidade de se cometer um erro ao tomar-se a decisão estatística de rejeitar a hipótese nula (erro tipo I).
Parâmetro	Qualquer medida numérica calculada a partir de todos os indivíduos de uma população. Por exemplo, média ou proporção.
Poder estatístico	Probabilidade de rejeitar a hipótese nula se ela for falsa.
População de interesse	Conjunto completo de indivíduos, eventos, ou objetos que são de interesse para o pesquisador.
Significância estatística	Está relacionada com a possibilidade de conclusões baseadas em dados da amostra serem atribuídas ao erro amostral.

Conceito	Descrição
Significância prática ou clínica	Está relacionada com a utilidade dos resultados. Por exemplo, resultados que mudam a prática profissional e beneficiam a saúde das pessoas podem ser considerados como de significância prática.
Tamanho, magnitude, ou dimensão do efeito	É uma medida que quantifica a magnitude de um fenômeno (por exemplo, efeito ou associação).
Teste de hipóteses	Método de averiguação sobre a probabilidade de uma afirmação. Um teste de hipóteses é uma regra de decisão para aceitar ou rejeitar uma hipótese, com base nas informações fornecidas pelos dados coletados em uma amostra e, por isso, envolve um risco de afirmar algo errado.
Valor de p	Probabilidade de a estimativa observada ou um valor mais extremo ter ocorrido, considerando que a hipótese nula e as suposições feitas são verdadeiras.

Interpretação do Valor de p

Como explicado na introdução e definido na Tabela 3, o valor de p indica a probabilidade da estimativa observada na amostra ou um valor mais extremo ocorrer, considerando a hipótese nula (H_0) e outras suposições sobre os dados como verdadeiras. Quanto mais improvável o resultado, menor o valor de p e mais confiança temos de que H_0 não é verdadeira. No entanto, resultados raros podem ocorrer, havendo o risco de se concluir erroneamente que H_0 não é verdadeira (15,18). Por exemplo, suponhamos que a H_0 do nosso teste é que não há diferença entre dois grupos quanto ao peso na população. Nesse caso, H_0 supõe que a diferença entre os grupos na população é igual a 0.

Se encontrarmos um valor de $p = 0,04$, ao realizarmos o teste de significância (por exemplo, teste t), a probabilidade de que a diferença entre grupos observada na amostra ou superior a ela ter ocorrido é de 4 em 100 ou 4%, considerando H_0 verdadeira. Nesse caso, com base na interpretação usual dos testes de significância, seria constatado uma diferença estatisticamente significativa entre os grupos, concluindo-se que existe uma diferença na população. Contudo, pode ser que a diferença observada na amostra ocorreu ao acaso, mesmo sendo improvável; afinal, acontecimentos improváveis também ocorrem. Quando observamos um valor de p abaixo do limiar de significância (por exemplo, 0,05) e a H_0 é verdadeira (i.e., não existe diferenças na população), diz-se que ocorreu um erro do tipo I.

Outro ponto importante para compreender a interpretação de p é perceber que um valor de p superior a 0,05, ou outro limiar comumente utilizado, não é evidência de ausência de associação ou ausência de diferenças entre grupos (18). Valores de p maiores que 0,05 indicam que o resultado obtido é mais provável do que a chance de 5 em 100, considerando a H_0 verdadeira. Por exemplo, se o valor de p é 0,30, a chance de o resultado observado ou mais extremo ter ocorrido, sendo H_0 verdadeira, é de 30 em 100. Na interpretação usual de testes de significância, H_0 não é rejeitada quando o valor de p é superior a 0,05. Uma análise que produz tal valor de p é dita estatisticamente não significativa.

Quando existe uma diferença ou associação (H_0 é falsa), mas a H_0 não é rejeitada, é dito que um erro do tipo II ocorreu. Esse entendimento de p evidencia as limitações de se utilizar limiares arbitrários para decidir se uma associação ou efeito existe; o que comumente ocorre na prática ao interpretar resultados como estatisticamente significantes ou estatisticamente não significantes. No entanto, pode-se ponderar que a decisão sobre a utilização ou não de limiares arbitrários de p para interpretação dos resultados deve ser tomada com base nas vantagens e desvantagens dessa abordagem em comparação a outras

alternativas. A principal alternativa a interpretação de valores arbitrários de p é a interpretação da estimativa de efeito ou associação e de seu intervalo de confiança (1,2).

Vantagens e Desvantagens de Interpretar Valores Arbitrários de p

Nesta seção discutimos as vantagens e desvantagens de se interpretar o valor de p ou limiares arbitrários (por exemplo, 0,05), ao invés da estimativa de efeito ou associação e seu intervalo de confiança.

Vantagem: Os defensores do uso de valores arbitrários de p como 0,05 comumente apontam como vantagem a objetividade dessa abordagem (por exemplo, 18,35). De acordo com esses autores, se um nível de significância não for determinado *a priori*, pesquisadores podem ser influenciados por suas crenças e emoções durante a interpretação dos resultados.

Desvantagens: A primeira desvantagem de interpretar limiares arbitrários ou o próprio valor de p ao invés da estimativa e seu intervalo de confiança é que os primeiros são menos informativos sobre a importância prática dos achados (5,17,19,20,22,24,27,30). Por exemplo, por si só um $p = 0,01$ não diz nada sobre a relevância de um resultado. Em uma amostra grande, diferenças irrelevantes entre grupos podem produzir valores de p pequenos. Por outro lado, saber que o grupo tratamento possui o dobro das chances (estimativa) de recuperação do que o grupo controle é bem mais informativo sobre a relevância prática dos achados.

A segunda desvantagem é que o valor de p por si só não é informativo sobre a imprecisão dos resultados. Olhar para o intervalo de confiança da estimativa permite entender melhor quão confiante pode-se estar de que o resultado observado na amostra se aplica a população de interesse.

A terceira desvantagem refere-se à dificuldade de compreender o valor de p (4,5,16,17,21,33). Frei (16), por exemplo, argumenta que “existe uma grande variedade de interpretações do valor de p , e a maioria delas é incorreta”. Na mesma direção, Figueiredo Filho e Silva (5) afirmam que o valor de p é a “estatística mais mal interpretada” em pesquisas científicas. Por outro lado, intervalos de confiança de 95% podem ser (por simplicidade) interpretados como indicando um intervalo dentro do qual podemos ter 95% de certeza de que o verdadeiro efeito ou associação está (ver definição estritamente correta na Tabela 3) (39).

Essas vantagens e desvantagens não impedem o pesquisador de considerar todas as informações disponíveis para a interpretação dos seus achados, por exemplo, valor de p , estimativa e intervalo de confiança. De fato, esta é a abordagem sugerida por parte dos autores dos artigos incluídos nessa metassíntese (5,17,21,23, por exemplo, 25,27,28,34,36).

Interpretação de Estimativas de Efeito ou Associação e de seus Intervalos de Confiança

Considerando que a estimativa é mais informativa sobre a relevância prática dos achados e o intervalo de confiança é mais informativo sobre a imprecisão (em termos de inferência) do que o valor de p , nesta categoria/seção organizamos informações relevantes para a interpretação da estimativa e de seu intervalo de confiança. Termos comumente utilizados para se referir à estimativa são tamanho, dimensão ou magnitude do efeito (32). Algumas vezes, a estimativa calculada será de simples interpretação (por exemplo, morto/vivo), outras vezes, será importante decidir sobre as características da estimativa. Por exemplo, em uma intervenção com pessoas diagnosticadas com depressão, pode-se avaliar uma redução na pontuação total de uma escala de depressão ou avaliar quantos participantes tiveram uma redução substancial nos sintomas (p.ex. 50%), ou ainda quantos participantes se recuperaram da depressão (40). Não é impossível, é claro, que todas essas estimativas de efeito sejam reportadas. Cabe aos responsáveis pela divulgação dos achados decidirem quais estimativas são mais acuradas e interpretáveis. Quanto mais informações forem fornecidas sobre os achados, mais abrangente poderá ser a avaliação deles.

No que concerne à interpretação de estimativas, é importante considerar a diferença entre significância prática (ou importância) e significância estatística. Consideremos, por exemplo, que há interesse em se aplicar na prática profissional uma intervenção que produziu uma redução média de um segundo (estimativa) no tempo da prova de 100 metros de atletas

profissionais. Dado que um segundo, nesse contexto, pode determinar o resultado da prova, podemos dizer que a estimativa tem significância prática. Como essa redução foi observada em uma amostra é importante sabermos a probabilidade de se observar uma redução na população de interesse e, possivelmente, em um atleta específico. Suponhamos que o intervalo de confiança de 95% dessa estimativa é de 0,5 a 1,5 segundos. Nesse caso, sabemos que esse resultado também é estatisticamente significativo (considerando $\alpha = 0,05$), pois, a porção inferior do intervalo de confiança não cruzou o zero. Se o resultado fosse estatisticamente não significativo, saberíamos que uma redução de zero também é compatível (relativamente provável) com os achados da pesquisa. O intervalo de 0,5 a 1,5 também indica que uma redução de 0,5 ou 1,5 segundos são compatíveis com os achados da pesquisa. Se estiver pensando em um atleta específico, será importante avaliar como a amostra da pesquisa é comparável a esse atleta, em aspectos que poderiam influenciar os efeitos da intervenção.

É relevante observar que a importância dos achados é dependente do contexto específico do estudo (25,27). Por exemplo, a mesma diferença de um segundo no tempo da prova de 100 metros pode não ter tanta importância para uma pessoa que está se preparando para uma prova física de um concurso quanto para um atleta de alto rendimento. Nesse contexto, é recomendado que os pesquisadores definam antes da realização do estudo o que pode ser considerado um efeito ou associação importante (19,29).

Quanto à interpretação de intervalos de confiança, intervalos menores são mais precisos porque incluem uma menor amplitude de valores prováveis para a população do que intervalos maiores (27). Além disso, os valores mais prováveis para a população encontram-se mais próximos da estimativa (23). Por exemplo, considerando a estimativa de 1 segundo e o intervalo 0,5 a 1,5, sabemos que o valor 0,7 é mais provável de ser o valor da população do que o valor 0,5 porque 0,7 está mais próximo de 1 (a estimativa) do que 0,5.

Evidências Sobre a Interpretação de Testes de Significância em Periódicos Brasileiros: Uma Lacuna na Literatura

Uma das categorias estabelecidas *a priori* tinha a intenção de organizar evidências sobre a interpretação de testes de significância em artigos publicados em periódicos brasileiros. Infelizmente, nenhum dos artigos incluídos providenciou tais evidências, apesar de alguns autores afirmarem que a interpretação de resultados como estatisticamente significantes ou estatisticamente não significantes é a abordagem dominante (por exemplo, 22,29,32,33). Desse modo, existe uma carência por evidências acerca da interpretação de testes de significância em periódicos nacionais.

Conclusão

No presente trabalho, apresentamos informações relevantes para interpretação de testes de significância, com base, em 26 artigos publicados em português. Pesquisadores frequentemente interpretam valores de p , de acordo com pontos de corte arbitrários (comumente 0,05), ao invés de considerar a estimativa de efeito ou associação e seu intervalo de confiança. Tal abordagem ignora informações relevantes para decisão sobre a importância prática do achado e sua imprecisão (referente à inferência para população). Desse modo, nossa recomendação é que, quando possível, se determine antes da realização do estudo o que são efeitos ou associações importantes e que se interprete a estimativa do efeito ou associação e seu intervalo de confiança conjuntamente com o valor de p (1,2).

Alertamos que nossa intenção com a identificação e síntese dos 26 artigos sobre a interpretação de teste de significância não é recomendar os pesquisadores a seguirem as orientações destes artigos. De fato, alguns artigos defendem visões diferentes sobre a interpretação de testes de significância (por exemplo, sobre o uso de limiares arbitrários de p) e nem todos os artigos são livres de equívocos. Estudos futuros poderão providenciar informações sobre a interpretação de testes de significância em periódicos científicos brasileiros. Pesquisas futuras também podem abordar temas não tratados nesse trabalho como tamanho amostral e ajuste do modelo aos dados.

Conflitos de Interesse

Os autores declaram que não há conflitos de interesse.

Referências

1. Amrhein V, Greenland S, McShane B. Scientists rise up against statistical significance. *Nature* [Internet]. 2019 Mar 20;567(7748):305–7. Available from: <http://www.nature.com/articles/d41586-019-00857-9>
2. Cumming G. The new statistics: Why and how. *Psychol Sci*. 2014 Jan;25(1):7–29.
3. Muff S, Nilsen EB, O'Hara RB, Nater CR. Response to 'Why P values are not measures of evidence' by D. Lakens. *Trends Ecol Evol* [Internet]. 2022 Apr;37(4):291–2. Available from: <https://linkinghub.elsevier.com/retrieve/pii/S0169534722000015>
4. Andreu MF, Ballve LPD, Verdecchia DH, Monzón AM, Carvalho TD de. O valor de p é adequadamente interpretado pelos profissionais de terapia intensiva? Uma pesquisa on-line. *Rev Bras Ter Intensiva*. 2021;33(1):88–95.
5. Figueiredo Filho D, Silva L. A significância estatística e o uso do p-valor em pesquisas na saúde: Recomendações práticas. *Interfaces Científicas - Saúde e Ambient* [Internet]. 2022 Sep 12;9(1):230–9. Available from: <https://periodicos.set.edu.br/saude/article/view/10867>
6. Rocha SV, Sacheto RM, Lessa RS, Squarcin CFR, Dias MF, Santos M de C, et al. COMPORTAMENTO SEDENTÁRIO ENTRE TRABALHADORES DA SAÚDE EM TEMPOS DE PANDEMIA DE COVID-19. *Rev Bras Ciência e Mov*. 2024;31(1).
7. Baker M. Statisticians issue warning over misuse of P values. *Nature* [Internet]. 2016 Mar 7;531:151. Available from: <http://www.nature.com/doifinder/10.1038/nature.2016.19503>
8. Oliveira GM de. Política linguística e internacionalização: A língua portuguesa no mundo globalizado do século XXI. *Trab em Linguística Apl*. 2013 Dec;52(2):409–33.
9. Instituto Brasileiro de Opinião Pública e Estatística. 73% dos brasileiros não sabem ler textos em inglês [Internet]. 2012 [cited 2019 May 8]. Available from: <http://www.ibopeinteligencia.com/noticias-e-pesquisas/73-dos-brasileiros-nao-sabem-ler-textos-em-ingles/>
10. Levitt HM, Bamberg M, Creswell JW, Frost DM, Suárez-orozco C. Reporting standards for qualitative primary, qualitative meta-analytic, and mixed methods research in Psychology : The APA Publications and Communications Board Task Force Report. *Am Psychol*. 2018;73(1):26–46.
11. Finfgeld-Connett D. Use of content analysis to conduct knowledge-building and theory-generating qualitative systematic reviews. *Qual Res*. 2014;14(3):341–52.
12. Rosa CD, Delabrida Z. Método experimental e ensaios clínicos: Metassíntese de artigos de revisão publicados em português [Experimental method and clinical trials: Metassynthesis of review articles published in Portuguese]. *Psico* [Internet]. 2021 Dec

31;52(4):e36259.

Available

from:

<https://revistaseletronicas.pucrs.br/index.php/revistapsico/article/view/36259>

13. Shamseer L, Moher D, Clarke M, Gherzi D, Liberati A, Petticrew M, et al. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015: Elaboration and explanation. *BMJ*. 2015 Jan;349:g7647–g7647.

14. Gehanno J-F, Rollin L, Darmoni S. Is the coverage of Google Scholar enough to be used alone for systematic reviews. *BMC Med Inform Decis Mak*. 2013 Dec;13(7).

15. Hirakata VN, Mancuso ACB, Castro SM de J. Teste de hipóteses: perguntas que você sempre quis fazer, mas nunca teve coragem. *Clin Biomed Res [Internet]*. 2019;39(2):181–5. Available from: <https://seer.ufrgs.br/hcpa/article/view/93649>

16. Frei F. Uso interativo de planilha eletrônica para o ensino de Estatística: O caso do valor de p. *Educ Matemática Pesqui [Internet]*. 2019 Sep 2;21(2):187–201. Available from: <https://revistas.pucsp.br/emp/article/view/37784>

17. Flório FM, Zanin L, Santos Júnior LM dos, Meneghim M de C, Ambrosano GMB. Tamanho do efeito em estudos observacionais na área de Saúde Bucal Coletiva: Importância, cálculo e interpretação. *Cien Saude Colet [Internet]*. 2023 Feb;28(2):599–608. Available from: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1413-81232023000200599&tIng=pt

18. Correia LCL, Bagano GO, Melo MHV de. Devemos aposentar a significância estatística? *Brazilian J Pain [Internet]*. 2020;2(3):299–300. Available from: https://www.scielo.br/scielo.php?script=sci_arttext&pid=S2595-31922019000300211&Ing=en&nrm=iso&tIng=en

19. Marcelino R, Pasquarelli BN, Sampaio J. Inferência baseada em magnitudes na investigação em ciências do esporte. A necessidade de romper com os testes de hipótese nula e os valores de p. *Rev Bras Educ Física e Esporte [Internet]*. 2019 Jun 2;33(4):667–76. Available from: <https://www.revistas.usp.br/rbefe/article/view/170602>

20. Mourão Junior CA. Quanto vale o valor-p? *Arq Ciências do Esporte*. 2019;7(2):72–4.

21. Costa AMA da, Oliveira AM de, Petto J. Você entende o valor de P? *Fisioter Bras [Internet]*. 2019 Sep 3;20(4):460–1. Available from: <http://openjournalsolutions.com.br/atlantica/index.php/fisioterapiabrasil/article/view/3217>

22. Feitosa GS. A atenção ao intervalo de confiança nas publicações. *Rev Científica Hosp St Izabel [Internet]*. 2018 May 14;2(2):7–8. Available from: <https://revistacientifica.hospitalsantaizabel.org.br/index.php/RCHSI/article/view/89>

23. Teixeira PM. Sobre o significado da significância estatística. *Acta Med Port*. 2018;31(5):238–40.

24. Marques Junior NK. Estatística para ser aplicada nos estudos das ciências do esporte e da atividade física. *Rev Bras Prescrição e Fisiol do Exerc*. 2018;12(76):564–73.

25. Azevedo PHSM de, Junior DP, Botero JP. Significância estatística: Medidas e interpretações. *Rev Bras Ciência e Mov.* 2016;24(2):206–9.
26. Ferrari R, Caram LM de O, Garcia T, Paiva SAR de, Vale SA do, Tanni SE. Tamanho do efeito. *Pneumol Paul.* 2016;29(1):73–4.
27. Espirito-Santo H, Daniel FB. Calcular e apresentar tamanhos do efeito em trabalhos científicos (1): As limitações do $p < 0,05$ na análise de diferenças de médias de dois grupos. *Rev Port Investig Comport e Soc [Internet]*. 2015 Feb 28;1(1):3–16. Available from: <https://rpics.ismt.pt/index.php/ISMT/article/view/14>
28. Ferreira JC, Patino CM. O que realmente significa o valor-p? *J Bras Pneumol.* 2015;41(5):485.
29. Marcelino R, Sampaio J. Investigação em Ciências do Desporto: Dos testes de hipótese nula à necessidade de interpretações com significância prática e/ou clínica. *Bol SPE.* 2015;28–35.
30. Patino CM, Ferreira JC. Intervalos de confiança: Uma ferramenta útil para estimar o tamanho do efeito no mundo real. *J Bras Pneumol.* 2015;41(6):565–6.
31. Feitosa GS. A incessante procura do “p” e seu significado em ciência médica. *Rev Saúde HSI.* 2015;(3):7.
32. Loureiro LM de J, Gameiro MGH. Interpretação crítica dos resultados estatísticos: para lá da significância estatística. *Rev Enferm Ref [Internet]*. 2011 Mar 1;III Série(nº 3):151–62. Available from: http://www.esenfc.pt/rr/index.php?module=rr&target=publicationDetails&pesquisa=&id_artigo=2224&id_revista=9&id_edicao=35
33. Sebastiani RG, Viali L. Teste de hipóteses: Uma análise dos erros cometidos por alunos de engenharia. *Bolema Bol Educ matemática.* 2011;24(40):835–54.
34. Lunet N, Severo M, Barros H. Inferência estatística: Intervalos de confiança vs. testes de hipóteses. *Rev Port Cir [Internet]*. 2008;(5):33–40. Available from: <https://revista.spcir.com/index.php/spcir/article/view/254>
35. Costa SF. Nível de significância alpha ou valor-p?. *Estud Em Avaliação Educ.* 2002;(25):185–98.
36. Pereira B de B. Estatística em medicina: p-variação. *Rev SOCERJ.* 1995;III(3):73–8.
37. Garcia CH. A utilização do nível de significância de 10% para os testes estatísticos florestais. *Rev Agric.* 1991;66(2):107–22.
38. Gomes FP. Níveis de significância nas análises estatísticas. *Informações Agronômicas.* 1984;(25):8.
39. Schünemann HJ, Vist GE, Higgins JP, Santesso N, Deeks JJ, Glasziou P, et al. Interpreting results and drawing conclusions. In: Higgins JPT, Thomas J, Chandler J,

Cumpston M, Li T, Page MJ, et al., editors. Cochrane Handbook for Systematic Reviews of Interventions. 2nd ed. Chichester: John Wiley & Sons; 2019.

40. Riedel M, Möller H-J, Obermeier M, Schennach-Wolff R, Bauer M, Adli M, et al. Response and remission criteria in major depression – A validation of current practice. J Psychiatr Res [Internet]. 2010 Nov;44(15):1063–8. Available from: <https://linkinghub.elsevier.com/retrieve/pii/S0022395610000750>