https://doi.org/10.22491/2357-9730.101299

Equívocos estatísticos: perguntas que você sempre quis fazer, mas nunca teve coragem

Statistical misconceptions: questions you've always wanted to ask, but never dared

Rogério Boff Borges[1], Vanessa Bielefeldt Leotti[1,2], Aline Castello Branco Mancuso[1], Stela Maris de Jezus Castro[1,2], Vânia Naomi Hirakata[1], Suzi Alves Camey[1,2]

RESUMO

Este artigo visa elucidar algumas dúvidas enfrentadas ou equívocos estatísticos cometidos por pesquisadores de diversas áreas. São explanados os temas: "tradução não é validação", "análise fatorial exploratória ou confirmatória", "nem todo estudo com dois grupos tem delineamento caso-controle", "teste ou ajuste de Bonferroni", "tamanho de amostra para teste de hipóteses e/ou para intervalo de confiança", e "testes ou dados paramétricos". A abordagem é realizada em uma linguagem acessível ao público leigo, utilizando exemplos e sugerindo referências para aprofundar o conhecimento.

Palavras-chave: *Equívocos estatísticos; validação; tradução; caso-controle; Bonferroni; análise fatorial; tamanho de amostra; teste paramétrico*

ABSTRACT

This article aims to answer some questions and elucidate statistical misconceptions of researchers from different fields. The following topics are addressed: "translation is not validation", "exploratory or confirmatory factor analysis", "not every study with two groups is a case-control study", "Bonferroni test or adjustment", "sample size for testing hypotheses and/or for confidence intervals", and "parametric data or tests". The topics are explained in lay terms, using examples and suggesting references to advance knowledge.

Keywords: *Statistical misconceptions; validation; translation; case-control; Bonferroni; factor analysis; sample size; parametric test*

Clin Biomed Res. 2020;40(1):63-70

1 Unidade de Bioestatística, Grupo de Pesquisa e Pós-graduação (GPPG), Hospital de Clínicas de Porto Alegre (HCPA). Porto Alegre, RS, Brasil.

2 Departamento de Estatística, Instituto de Matemática e Estatística, Universidade Federal do Rio Grande do Sul (UFRGS). Porto Alegre, RS, Brasil.

Autor correspondente:
Rogério Boff Borges
l-bioestatistica@hcpa.edu.br
Hospital de Clínicas de Porto Alegre (HCPA)
Rua Ramiro Barcelos, 2350.
90035-007, Porto Alegre, RS, Brasil.

TRADUÇÃO NÃO É VALIDAÇÃO!

No contexto de mensuração de traços latentes (variáveis que não podem ser medidas diretamente, por exemplo: intensidade de sintomas depressivos, severidade da ansiedade, nível de qualidade de vida, estresse pré-operatório, etc), faz-se necessária a criação de instrumentos de medida (muitas vezes conhecidos como questionários ou testes) compostos de variáveis observáveis (usualmente chamadas de itens) que são expressões de facetas destes traços. Um exemplo clássico de instrumento de medida de um traço latente é o Inventário de Ansiedade de Beck (BAI)[1], onde o traço latente de interesse é a severidade da ansiedade. O BAI é composto por 21 itens que correspondem a 21 sintomas observáveis de ansiedade, isto é, cada item é uma variável a qual o indivíduo responde em uma escala ordinal onde 0 significa que ele não apresentou aquele sintoma (Absolutamente não), 1 significa que o indivíduo apresentou aquele sintoma levemente (Não me incomodou muito), 2 significa que o indivíduo apresentou aquele sintoma moderadamente (Foi desagradável, mas pude suportar), e onde 3 significa que o indivíduo apresentou o sintoma de forma severa (Quase não suportei). O escore de severidade da ansiedade varia de zero (o indivíduo responde "Absolutamente não" para todos os 21 itens)

a 63 (o indivíduo responde “Quase não suportei” para todos os itens).

Quando se precisa medir um traço latente, faz-se necessária a validação do instrumento de medida utilizado na população que se deseja estudar. Contextos nos quais se faz necessária a validação de um instrumento de medida são:

- O instrumento é inédito, isto é, foi criado pelos próprios pesquisadores;
- O instrumento já existe no idioma de origem do estudo (por exemplo, português do Brasil), ou seja, já foi criado por outros pesquisadores do mesmo país, mas não foi validado para a população que se quer estudar. Por exemplo, o instrumento foi criado e validado em uma população do nordeste do Brasil, mas você precisa utilizá-lo em uma população de outra região brasileira que supõe-se ter um comportamento diferente da primeira – neste caso, faz-se necessária a validação do instrumento para seu uso na nova população de interesse;
- O instrumento já existe, mas em um idioma diferente daquele do estudo.

O processo de validação de um instrumento de medida de um traço latente engloba basicamente duas partes distintas: a parte correspondente a medidas de fidedignidade (ou confiabilidade) e a parte correspondente às validades.

A fidedignidade é uma característica dos escores resultantes da aplicação do instrumento de medida. Por exemplo, no caso do BAI, o escore resultante é aquele conhecido como Escore Total, onde valores próximos de zero indicam um grau mínimo de ansiedade e valores próximos de 63 indicam ansiedade severa. A fidedignidade se baseia na consistência (um escore ser consistente, significa que seus resultados não se alteram se o instrumento de medida for aplicado nos mesmos indivíduos em diferentes momentos ou aplicado por diferentes avaliadores em um mesmo indivíduo e mesmo momento do tempo) e na precisão destes escores. Resumindo, fidedignidade é a qualidade dos escores resultantes da aplicação do instrumento de medida, sugerindo que eles são suficientemente consistentes e livres de erros de mensuração (precisos), de modo que sejam úteis[2].

Já a validade, pode ser definida como o grau em que todas as evidências acumuladas corroboram de forma que os escores resultantes da aplicação de um instrumento de medida estejam mensurando exatamente aquele traço latente que se pretendia medir. A validade de um instrumento de medida começa no momento em que se pensa em construí-lo e subsiste durante todo o processo de elaboração, aplicação, correção e interpretação dos resultados[3]. Os autores divergem quanto aos tipos de validade, mas, de um modo geral, pode-se avaliar três tipos de validade: a validade de conteúdo, a validade de critério e a validade de construto. Mais informações sobre medidas de fidedignidade e tipos de validade podem ser encontradas em Urbina[2] e Raymundo[3].

A Figura 1 ilustra o processo de validação de um instrumento de medida. Observe que na etapa 2, quando o instrumento de medida está validado em outro idioma que não o do pesquisador, torna-se necessário fazer a tradução dos itens deste instrumento para o idioma do mesmo. No entanto, o fato de traduzir o instrumento de medida não significa que o mesmo continua validado, pois ainda precisa passar pela etapa 3, que é aquela que de fato vai coletar as evidências de que o instrumento de medida segue medindo o traço latente pretendido e que os escores resultantes da aplicação do mesmo são consistentes e precisos.

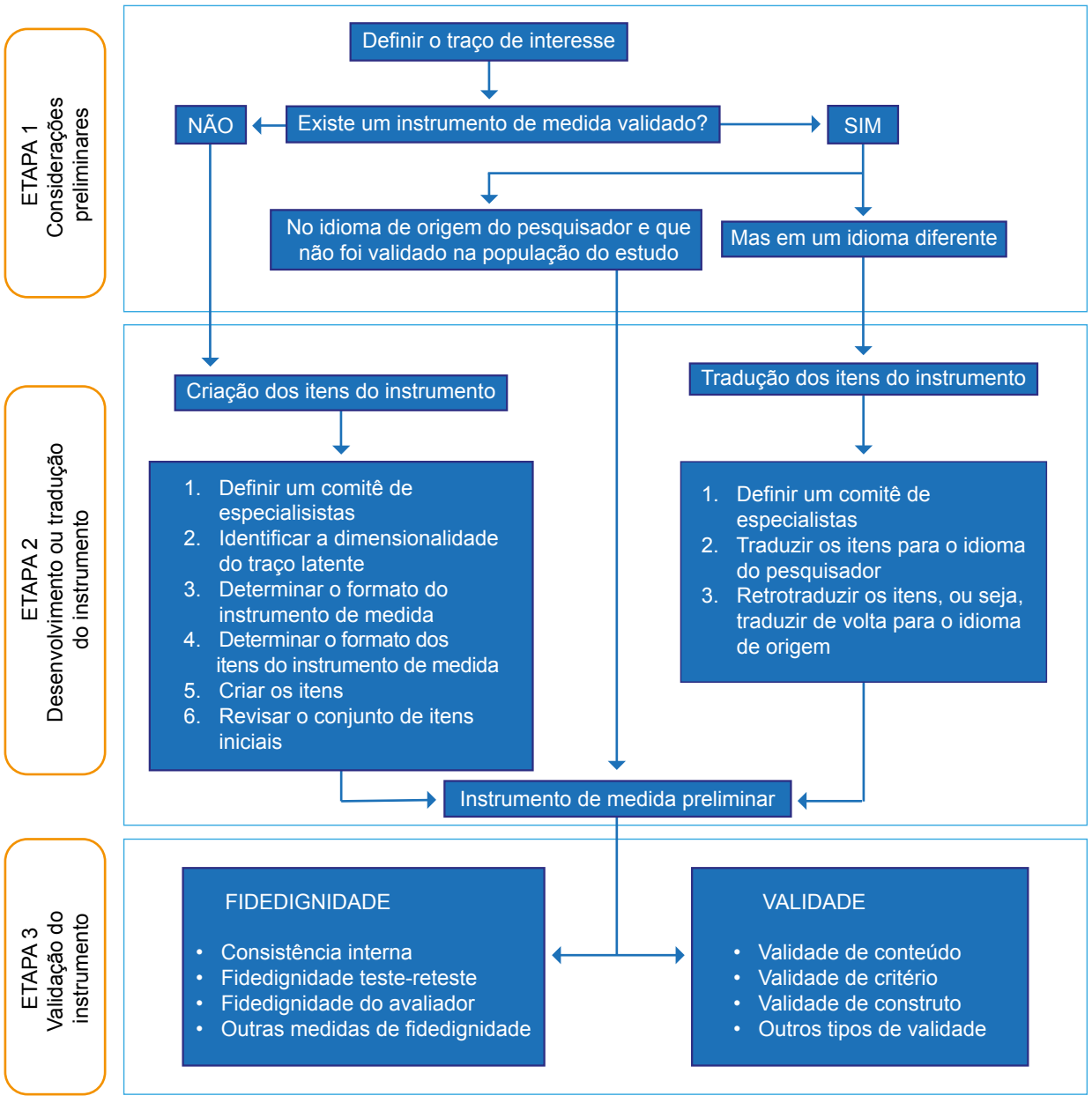

Figura 1: Ilustração de um processo de validação de um instrumento de medida.
Fonte: adaptado de Tsang et al.[4]

ANÁLISE FATORIAL EXPLORATÓRIA OU CONFIRMATÓRIA?

Análise fatorial é o nome dado a uma classe de métodos estatísticos multivariados, que não pressupõe a existência de variáveis dependentes e independentes. Tem como objetivo principal analisar o padrão de correlações existentes dentro de um conjunto grande de variáveis e verificar se existem dimensões latentes comuns, também chamadas de fatores. Portanto, utilizamos uma análise fatorial, quando desejamos avaliar, dentre diversas variáveis quais são as que se "agrupam" apresentando um padrão semelhante de correlações, mas que são relativamente independentes de outro(s) conjunto(s) de variável(is). De uma forma bem específica, como resultado de uma análise fatorial, o que se pretende obter é uma quantidade menor de variáveis do que o questionário inicial, mas que preserve ao máximo a explicação (variabilidade) das informações iniciais[5,6]. A análise fatorial pode ser realizada de duas formas: exploratória ou confirmatória.

Na análise fatorial exploratória, o agrupamento das variáveis irá ocorrer tendo ou não uma estrutura subjacente, ou seja, de forma mais "natural", sem

nenhuma restrição *a priori* nem sobre as correlações existentes entre as variáveis observadas, nem sobre o número de fatores existentes[5,6]. O número de fatores a serem extraídos é exatamente o mesmo número de variáveis existentes, e todos são compostos por uma combinação ponderada (com pesos) de cada uma das variáveis. No entanto, os fatores são compostos seguindo algumas regras definidas: os pesos do primeiro fator são constituídos de forma a agregar a quantidade máxima de variabilidade entre as pontuações de todos os sujeitos. Do segundo fator em diante, os fatores são determinados de forma a priorizar as variáveis que não tiveram pesos importantes no(s) fator(es) anterior(es) e explicando o máximo de variabilidade, e que seja não correlacionado (relativamente independente) do(s) fator(es) anterior(es)[7]. Por exemplo, Senna et al.[8] analisaram a satisfação da mulher com a amamentação. Através de um instrumento contendo 30 itens, os autores realizaram uma análise fatorial exploratória e verificaram que haviam três fatores (em algumas áreas também é chamado de domínios) associados à essa satisfação: (1) prazer e realização do papel materno, (2) físico, emocional e social materno e (3) crescimento, desenvolvimento e satisfação infantil.

Já a análise fatorial confirmatória trabalha com a ideia de testar uma hipótese de uma estrutura (conhecida ou esperada) subjacente. Geralmente, esta etapa da análise é feita após um amplo estudo da estrutura da combinação das variáveis. A análise fatorial confirmatória é um tipo de análise que pode ser realizada através do métodos de Modelagem de Equações Estruturais. A partir de um modelo teórico proposto (com variáveis observadas e não observadas, as latentes), a estrutura de correlações entre todas elas é testada e uma série de estatísticas que servem para verificar o ajuste do modelo aos dados coletados é calculada, permitindo, assim, que o pesquisador saiba se o modelo proposto se adequa aos dados. Geralmente essa análise é realizada na Etapa 3, de validação do instrumento, discutida na secção anterior. Por exemplo, Terra et al.[9] realizaram uma análise fatorial confirmatória em dados de aplicação do *Parental Bonding Instrument*, ajustando diferentes estruturas fatoriais ao conjunto de dados. Nesse caso uma análise fatorial confirmatória é aplicável pois havia estudos prévios propondo estruturas de correlação dos itens e os autores desejavam verificar se essas estruturam se mantém em seu conjunto de dados.

A recomendação, em geral, é de que a análise fatorial confirmatória seja feita depois da exploratória. Não é recomendado que a análise fatorial confirmatória seja realizada com os resultados obtidos na exploratória e utilizando a mesma base de dados. Neste caso, deve-se partir o banco em 2 partes, sendo que numa delas será feita a análise fatorial exploratória e seus resultados serão testados através da análise confirmatória na outra parte do banco de dados. No artigo de Ang[10], por exemplo, a autora descreve que criou um inventário de relacionamento de estudante-professor e utilizou, numa amostra inicial, a análise fatorial exploratória para descobrir se as questões que ela coletou de outros trabalhos da literatura se correlacionavam de forma a compor fatores consistentes. Após, numa outra amostra, chamada estudo 2, ela testou a estrutura fatorial obtida no estudo 1, através de uma análise fatorial confirmatória.

NEM TODO ESTUDO COM DOIS GRUPOS TEM DELINEAMENTO CASO – CONTROLE

Primeiramente é necessário estabelecer em qual contexto o termo "caso – controle" está sendo aplicado. O delineamento epidemiológico caso – controle recebe esse nome devido ao procedimento de seleção das unidades amostrais, na qual são selecionadas de acordo com a presença (casos) ou ausência (controle) do desfecho de interesse, conforme já discutido em Nunes et al.[11]. Sendo assim, nesse delineamento, o termo "caso – controle" se refere ao desfecho e à forma com que os dados foram coletados. Por exemplo, Severo et al.[12] avaliou fatores de risco para quedas em pacientes adultos hospitalizados, os pacientes foram selecionados partindo do desfecho, sofreu queda (caso) vs não sofreu queda (controle), e então foram avaliados os fatores de risco.

Um outro contexto no qual esse termo pode ser aplicado, é em um delineamento ensaio clínico, no qual, no início do estudo, nenhuma das unidades amostrais têm o desfecho e são alocadas de forma aleatória no grupo que receberá o tratamento (caso) ou no grupo placebo (controle). Nesta situação, o termo "caso – controle" é utilizado para descrever os grupo de comparação (variável independente) e não o desfecho em si. Por exemplo, Tres et al.[13], descreve os resultados de um estudo na qual pacientes com diabetes mellitus tipo 2 foram alocados, de maneira aleatória, em grupos que que receberam diacereina 50 mg (casos) ou placebo (controle) por 12 semanas, para avaliar a variação dos níveis de hemoglobina glicada.

Para mais exemplos dos principais delineamentos e suas diferenças, consulte Nunes et al.[11].

TESTE OU AJUSTE DE BONFERRONI, QUANDO UTILIZAR?

O ajuste de Bonferroni, ou correção de Bonferroni, utilizado em muitas análises estatísticas, é uma correção do valor de alfa, tal que $\alpha_{bonferroni} = \alpha/k$, onde alfa (a) é o nível de significância global do experimento e k é o

número de comparações a serem realizadas (definido previamente)[14]. Esta correção costuma ser empregada em estudos que incluem muitos testes de hipótese, dado a premissa de que se for testado o suficiente, encontrar-se-á inevitavelmente algo significativo[15]. Bender e Lange[16] discutem quando e como utilizar o ajuste para múltiplos testes. Segundo os autores não há uma resposta simples e única, mas dependerá de qual taxa de erro o pesquisador quer controlar: a individual, de cada teste, ou a experimental (global). No entanto, recomenda-se que o seu uso seja restrito a um número pequeno de comparações.

Já o teste de Bonferroni, comumente utilizado nas comparações *post-hoc* de análises como ANOVA (*Analysis of Variance*) e GEE (*Generalized Estimating Equation*), por exemplo, refere-se ao ajuste do nível de significância do teste LSD de Fisher, especificamente. O teste LSD de Fisher fixa o erro tipo I (α) por comparação, mas se diversas comparações são feitas ao nível de significância , o nível de significância global fica muito alto. Por isso Bonferroni propôs um ajuste no nível de significância do teste LSD, garantido assim um nível global para o experimento. O teste LSD de Fisher ajustado levou, então, o nome de "teste de Bonferroni".

Neste contexto de comparações múltiplas, além do teste de Bonferroni, diversos outros testes têm sido propostos na literatura[17,18], com diferentes indicações. O Quadro 1 lista alguns dos testes mais utilizados e quando são indicados. Dentre estes, Callegari-Jacques[19] elenca três como os mais populares: Tukey, Student-Newman-Keuls (SNK) e Bonferroni.

Conforme o Quadro 1, os testes podem ser diferenciados por sua indicação e também pela taxa de erro que se deseja controlar, individual ou experimental. Os testes com taxa de erro individual controlam o nível de significância (α) isoladamente, por par de comparação, já os testes com taxa de erro experimental garantem o nível de significância (α) estabelecido para todo o conjunto de comparações. É importante considerar a taxa de erro experimental ao efetuar múltiplas comparações porque as chances de cometer um erro do tipo I para uma série de comparações é maior do que para uma comparação separada.

Os testes que controlam o nível de significância para experimentos são ditos conservadores, porque rejeitam a hipótese de igualdade de médias com baixa probabilidade. Esses testes têm, portanto, menor poder. Em contraposição, os testes que controlam o nível de significância por comparação são mais liberais, porque rejeitam a hipótese da nulidade com mais facilidade, mas têm grande poder, pois nível de significância e poder do teste crescem juntos[20,21].

Quando aplicado nas mesmas condições, o teste de Bonferroni irá produzir intervalos de confiança maiores que o teste de Tukey. E, por isso, em algumas situações, o teste de Bonferroni pode ser muito conservador (fraco), isto é, a taxa de erro do experimento fica menor do que o nível de significância estabelecido (α). Em geral, é dito que o LSD é o mais liberal, o Bonferroni é o mais conservador e o Tukey seria um meio termo[22]. Girardi et al.[23] compararam o nível de significância e o poder de cinco testes de comparação, destacando o teste de Duncan com maior poder, seguido por Tukey e Bonferroni. Contudo, cabe salientar que todos os procedimentos para a comparação de médias têm vantagens e desvantagens. Não existe um teste definitivamente "melhor" que todos os outros.

Quadro 1: Testes para Comparações Múltiplas.

Teste	Indicação	Taxa de Erro Alfa (α)	Poder
Bonferroni	Para comparar todos os pares.	Experimental.	Menos poderoso que Tukey, porém mais conservador que Sidak.
Sidak	Para comparar todos os pares.	Experimental.	Conservador, mas um pouco menos que Bonferroni.
Tukey	Para comparar todos os pares.	Experimental.	Menos poderoso que LSD Fisher, porém mais poderoso da categoria.
Duncan*	Para comparar todos os pares.	Individual.	O mais poderoso da categoria.
LSD de Fisher	Para comparar todos os pares.	Individual.	O mais poderoso depois de Duncan.
Dunnett	Comparação entre controle e os demais grupos.	Experimental.	Mais poderoso da categoria.
Games Howell	Usado quando não se pressupõe igualdade de variâncias.	Experimental.	Mais poderoso da categoria, porém menos poderoso que Tukey.
MCB de Hsu	Para comparar o grupo com a maior ou menor média aos outros grupos.	Experimental.	Mais poderoso que Tukey.
Scheffe	Para testar contrastes (grupos de médias) pré-selecionados.	Experimental.	Menos poderoso que Tukey e Bonferroni.

***** Utiliza uma abordagem bayesiana.

TAMANHO DE AMOSTRA PARA TESTE DE HIPÓTESES E/OU PARA INTERVALO DE CONFIANÇA?

Em primeiro lugar, é necessário esclarecer que o cálculo de tamanho de amostra é uma etapa necessária para estudos que tem como objetivo fazer inferência estatística. Quando se deseja obter conclusões estatísticas sobre todo um conjunto de unidades de interesse (a população) a partir apenas de um subconjunto de unidades (a amostra), recorre-se à inferência estatística. Assim, quando se faz a avaliação de todas as unidades de interesse, ou seja, quando se realiza um censo, não é necessário fazer cálculo de tamanho amostral prévio a coleta de dados.

O cálculo do tamanho da amostra deve ocorrer no momento de planejamento de um estudo com coleta de dados primários[24]. Esta etapa é muito importante, pois apenas com o cálculo apropriado é que se pode tomar conclusões precisas e acuradas a partir dos dados coletados[25]. É necessário que o objetivo primário do estudo já esteja definido, pois o cálculo será procedido com base nesse objetivo.

A inferência estatística pode ser dividida em duas grandes áreas: a estimação e os testes de hipóteses. Ao fixar o objetivo primário do estudo, o pesquisador precisa buscar reconhecer se este objetivo vai ser atingido estatisticamente via estimação de intervalo de confiança ou via testes de hipóteses.

Na estimação, deseja-se obter uma estimativa de alguma quantidade populacional de interesse, que é chamada de parâmetro (um proporção de pacientes acometidos por uma doença, por exemplo), com alguma precisão e algum nível de confiança estabelecidos pelo pesquisador e utilizados no cálculo de tamanho de amostra. Tal precisão também pode ser chamada de margem de erro e é definida como metade da amplitude de um intervalo de confiança.

Já no teste de hipóteses, há uma questão específica a ser respondida sobre um ou mais parâmetros (exemplos: se a proporção de doentes é igual a um valor de referência; se as médias da pressão arterial de pacientes tratados com anti-hipertensivo e não tratados são iguais) e baseado nessa questão constrói-se duas hipóteses estatísticas, nula e alternativa, que se contradizem. Para calcular o tamanho amostral com tal objetivo, será necessário, entre outras informações, que o pesquisador estabeleça o nível de significância e o poder desejados para o teste. Tais conceitos foram explicados em Hirakata et al.[26].

É importante observar que é possível que um tamanho amostral calculado para a estimação de um parâmetro não conduza a um poder aceitável no teste de hipóteses sobre este mesmo parâmetro. Da mesma forma, um tamanho de amostra calculado para teste de hipótese, mesmo com poder aceitável, pode gerar um intervalo de confiança bastante amplo.

Por exemplo, para testar a diferença do coeficiente de correlação de Pearson em relação ao valor zero (correlação nula), assumindo uma correlação de 0,5, um nível de significância de 5% e poder de 80%, é necessário pesquisar 30 unidades. Já para a estimação, desse mesmo coeficiente de correlação de Pearson, é necessário definir a amplitude que se deseja para o intervalo de confiança. Quanto menor a amplitude (maior precisão), maior o tamanho amostral necessário. A tabela 1 mostra o tamanho de amostra necessário para estimar o coeficiente de correlação de Pearson em diferentes cenários de amplitude dos intervalo de confiança, com 95% de nível de confiança, mantendo a correlação esperada de 0,5.

Tabela 1: Exemplos de tamanhos amostrais necessários para estimação do coeficiente de correlação de Pearson com 95% de confiança e correlação esperada de 0,5.

Amplitude do intervalo de confiança	n
0,2	219
0,3	99
0,4	57
0,5	37
0,56	30

Conforme apresentado na Tabela 1, as mesmas 30 unidades calculadas para o teste de hipótese exemplificado anteriormente resultarão em uma amplitude do intervalo de confiança de 0,56. No entanto, caso se deseje uma estimativa mais precisa, com uma amplitude de 0,30, por exemplo, será necessário pesquisar 99 unidades.

Assim, é importante que o pesquisador tenha claro qual é o objetivo primário do estudo e se esse objetivo será atendido via estimação ou teste de hipóteses. Nos casos em que ambas as opções estão contempladas nos objetivos, basta calcular os tamanhos amostrais para cada situação e tomar como tamanho amostral final o maior deles.

Referências recomendadas sobre cálculo de tamanho de amostra são Hulley et al.[27], Ryan[28] e Siqueira[29].

"TESTES PARAMÉTRICOS" OU "DADOS PARAMÉTRICOS"

A distribuição de frequência de uma variável aleatória pode ser descrita através de uma função massa/densidade de probabilidade e visualizada através de histogramas, por exemplo. No entanto,

nem sempre é possível identificar qual a verdadeira função massa/ densidade de probabilidade associada ao conjunto de dados. Na prática é assumido que a variável adere à uma distribuição teórica conhecida através de informações observáveis, como a distribuição de frequência observada, análise descritiva e experiência do pesquisador.

Hirakata et al.[26] discutem a definição e as etapas para a realização de um teste de hipóteses. Uma dessas etapas é a escolha, entre os testes previamente definidos no projeto, se será utilizada a abordagem paramétrica ou não, de acordo com a distribuição atribuída à variável. Os testes paramétricos pressupõem que a distribuição dos dados a serem analisados é conhecida, enquanto os testes não paramétricos não fazem tal suposição. Logo, ser ou não ser paramétrico é um atributo do teste estatístico e não dos dados.

É conveniente, sempre que plausível, assumir que os dados aderem à uma distribuição conhecida, principalmente à distribuição normal, devido à variedade de testes disponíveis com esse pressuposto.

REFERÊNCIAS

1. Beck AT, Epstein N, Brown G, Steer RA. An inventory for measuring clinical anxiety: psychometric properties. *J Consult Clin Psychol*. 1988;56(6):893-7.
2. Urbina S. *Essentials of psychological testing*. Hoboken: John Wiley & Sons; 2014.
3. Raymundo VP. Construção e validação de instrumentos: um desafio para a Psicolinguística. *Letras Hoje* [Internet]. 2009 [citado 2020 fev 7];44(3):86-93. Disponível em: http://revistaseletronicas.pucrs.br/ojs/index.php/fale/article/view/5768
4. Tsang S, Royse CF, Terkawi AS. Guidelines for developing, translating, and validating a questionnaire in perioperative and pain medicine. *Saudi J Anaesth*. 2017;11(Suppl 1):S80-9.
5. Hair JF, Anderson RE, Tatham RL, Black WC. *Análise Multivariada de Dados*. 5a. Porto Alegre: Bookman; 2005.
6. Tabachnick BG, Fidell LS. *Using multivariate statistics*. New York: Harper & Row; 1983.
7. Streiner DL, Norman GR. *Health measurement scales*. 3a ed. New York: Oxford University Press; 2005.
8. Senna AFK, Giugliani C, Lago JCA, Bizon AMBL, Martins ACM, Oliveira CAV, et al. Validação de instrumento para avaliação da satisfação da mulher com a amamentação para a população brasileira. *J Pediatr*. 2020;96(1):84-91.
9. Terra L, Hauck S, Fillipon AP, Sanchez P, Hirakata V, Schestatsky S, et al. Confirmatory Factor Analysis of the Parental Bonding Instrument in a Brazilian Female Population. *Aust N Z J Psychiatry*. 2009;43(4):348-54.
10. Ang RP. Development and validation of the Teacher-Student Relationship Inventory using exploratory and confirmatory factor analysis. *J Exp Educ*. 2005;74(1):55-73.
11. Nunes LN, Camey SA, Guimarães LSP, Mancuso ACB, Hirakata VN. Os principais delineamentos na Epidemiologia. *Clin Biomed Res* [Internet]. 2013 [citado 2020 fev 7];33(2):178-83. Disponível em: https://seer.ufrgs.br/hcpa/article/view/42338
12. Severo IM, Kuchenbecker RS, Vieira DFVB, Lucena AF, Almeida MA, Severo IM, et al. Fatores de risco para quedas em pacientes adultos hospitalizados: um estudo caso-controle. *Rev Latino-\Am Enfermagem* [Internet]. 2018 [citado 2020 março 11];26:e3016. Disponível em: http://www.scielo.br/scielo.php?script=sci_abstract&pid=S0104-11692018000100332&lng=en&nrm=iso&tlng=pt
13. Tres GS, Fuchs SC, Piovesan F, Koehler-Santos P, Pereira FS, Camey S, et al. Effect of Diacerein on Metabolic Control and Inflammatory Markers in Patients with Type 2 Diabetes Using Antidiabetic Agents: A Randomized Controlled Trial. *J Diabetes Res* [Internet]. 2018 [citado 2020 março 11];2018. Disponível em: https://www.ncbi.nlm.nih.gov/pmc/articles/ PMC5902058/
14. Dean A, Voss D, Draguljić D. *Design and Analysis of Experiments* [Internet]. Cham: Springer; 2017 [citado 2020 abr 30]. Disponível em: http://link.springer.com/10.1007/978-3-319-52250-0
15. Bland JM, Altman DG. Multiple significance tests: the Bonferroni method. *BMJ*. 1995;310(6973):170.
16. Bender R, Lange S. Adjusting for multiple testing--when and how? *J Clin Epidemiol*. 2001;54(4):343-9.
17. Zar JH. *Biostatistical analysis*. Upper Saddle River: Prentice Hall; 1999.
18. Cardellino RA, Siewerdt F. Utilização correta e incorreta dos testes de comparação de médias. *Rev Soc Bras Zoot* [Internet]. 1992 [citado 2020 dezembro 23];21(6):985-95. Disponível em: http://files.giselemoreira.webnode.com/200000368-7d6c07f643/Utiliza%C3%A7%C3%A3o%20incorreta%20testes%20m%C3%A9dias.pdf
19. Callegari-Jacques SM. *Bioestatística*: princípios e aplicações. Porto Alegre: Artmed; 2003.
20. Conagin A, Barbin D. Bonferroni's and Sidak's modified tests. *Sci Agric* (Piracicaba, Braz.). 2006;63(1):70-6.
21. Shingala MC. Comparison of post hoc tests for unequal variance. *Int J New Tech Sci Eng*. 2015;2(5):22-33.
22. Fukushi RK. *Análise de Variância (ANOVA)* [Internet]. 2016 [citado 2019 dezembro 20]. Disponível em: https://rstudio-pubs-static.s3.amazonaws.com/201742_ba0f209e7e2c47619342c0112d616e7a.html
23. Girardi LH, Cargnelutti Filho A, Storck L. Erro tipo I e poder de cinco testes de comparação múltipla de médias. *Rev Bras Biom*. 2009;27(1):23-36.
24. Hickey GL, Grant SW, Dunning J, Siepe M. Statistical primer: sample size and power calculations—why, when and how?. *Eur J Cardiothorac Surg*. 2018;54(1):4-9.

25. Nayak B. Understanding the relevance of sample size calculation. *Indian J Ophthalmol*. 2010;58(6):469-70.

26. Hirakata VN, Mancuso ACB, Castro SMJ. Teste de hipóteses: perguntas que você sempre quis fazer, mas nunca teve coragem. *Clin Biomed Res* [Internet]. 2019 [citado 2019 setembro 27];39(2):181-5. Disponível em: https://seer.ufrgs.br/hcpa/article/view/93649

27. Hulley SB, Cummings SR, Browner WS, Grady D, Hearst N, Newman TB, et al. *Delineando a pesquisa clínica*: uma abordagem epidemiológica. Porto Alegre: Artmed; 2003.

28. Ryan TP. *Sample size determination and power*. Hoboken: John Wiley & Sons; 2013.

29. Siqueira AL. *Dimensionamento de amostra para estudos na área da saúde*. Belo Horizonte: Folium Editorial; 2017.

Recebido: 25 mar, 2020
Aceito: 25 mar, 2020