Frequência do Uso Adequado dos Testes Estatísticos nos Artigos Originais Publicados na Revista Brasileira de Anestesiologia entre janeiro de 2008 e dezembro de 2009

Fabiano Timbó Barbosa [1], Diego Agra de Souza [2]

Resumo: Barbosa FT, Souza DA – Frequência do Uso Adequado dos Testes Estatísticos nos Artigos Originais Publicados na Revista Brasileira de Anestesiologia entre janeiro de 2008 e dezembro de 2009.

Justificativa e objetivos: A realização de uma análise estatística é necessária para uma avaliação adequada do artigo original por parte do leitor, possibilitando-lhe melhor visualização e compreensão dos resultados. O objetivo desta pesquisa foi determinar a frequência do uso adequado dos testes estatísticos de hipóteses presentes nos artigos originais publicados na Revista Brasileira de Anestesiologia no período entre janeiro de 2008 e dezembro de 2009.

Métodos: Foram selecionados artigos originais publicados na Revista Brasileira de Anestesiologia entre janeiro de 2008 a dezembro de 2009. O uso dos testes estatísticos foi avaliado como apropriado quando a seleção dos testes foi adequada para variáveis contínuas e categóricas e para testes paramétricos e não paramétricos; houve descrição do fator de correção quando se relatou o uso de múltiplas comparações; foi mencionado o uso específico de um teste estatístico para a análise de uma variável.

Resultados: Foram selecionados 76 artigos originais, com um total de 179 testes estatísticos de hipóteses. A frequência dos testes estatísticos mais utilizados foi: 20,11% para o qui-quadrado, 19,55% para o teste *t* de student, 10,05% para o teste de ANOVA e 9,49% para o teste exato de Fisher. A frequência de uso adequado dos testes estatísticos de hipóteses foi de 56,42% (IC 95% 49,16% a 63,68%), de uso inadequado 13,41% (IC 95% 8,42% a 18,40%), ocorrendo resultado inconclusivo em 30,16% (IC 95% 23,44% a 36,88%).

Conclusões: A frequência do uso adequado dos testes estatísticos utilizados nos artigos originais publicados na Revista Brasileira de Anestesiologia entre janeiro de 2008 e dezembro de 2009 foi de 56,42%.

Unitermos: ANESTESIOLOGIA: publicação; ESTATÍSTICA: interpretação de dados; METODOLOGIA CIENTÍFICA: estatística.

[Rev Bras Anestesiol 2010;60(5): 528-536] ©Elsevier Editora Ltda.

INTRODUÇÃO

Os leitores de revistas científicas devem fazer uma interpretação crítica do delineamento e da condução da pesquisa, assim como realizar análise estatística nos testes empregados em cada pesquisa para, subsequentemente, interpretar seus resultados [1]. A literatura vem demonstrando que os clínicos, principalmente aqueles que não têm uma educação formal em epidemiologia e bioestatística, têm um entendimento pobre dos testes estatísticos e uma habilidade limitada para interpretar os resultados dos estudos publicados na forma de artigos originais nos periódicos [2].

É necessário realizar uma análise estatística no artigo original para que o leitor tenha condições de melhor visualizar e compreender os resultados, assim como entender como os dados da pesquisa foram tratados, embora nem sempre ela seja obrigatória, uma vez que alguns artigos originais provêm de pesquisas qualitativas ou de estudos meramente descritivos. É importante que a análise estatística seja selecionada e realizada adequadamente, a fim de validar os resultados encontrados em cada pesquisa. Outras revistas científicas já realizaram análise de seu material, havendo interesse por parte dos editores em aprimorar suas publicações [3-6].

O objetivo desta pesquisa foi determinar a frequência do uso adequado dos testes estatísticos de hipóteses presentes nos artigos originais publicados na Revista Brasileira de Anestesiologia no período entre janeiro de 2008 e dezembro de 2009.

MÉTODO

Esta pesquisa foi submetida ao Comitê de Ética em Pesquisa da Universidade Estadual de Ciências da Saúde de Alagoas, que dispensou avaliação por se tratar de pesquisa que utiliza dados de domínio público. O termo de consentimento esclarecido não se aplica a esse tipo de pesquisa. Os gastos

Recebido da Universidade Estadual de Ciências da Saúde de Alagoas.

1. Mestre em Ciências pela Universidade Federal de Alagoas, Professor da disciplina Bases da Técnica Cirúrgica e Anestésica pela Universidade Federal de Alagoas
2. Estudante de medicina da Universidade Estadual de Ciências da Saúde de Alagoas, doutorando

Submetido em 4 de maio de 2010.
Aprovado para publicação em 16 de maio de 2010.

Endereço para correspondência:
Dr. Fabiano Timbó Barbosa
Comendador Palmeira, 113, ap. 202
Farol
57051-150 – Maceió, AL, Brasil
Tel: (82) 9983-2054
E-mail: fabianotimbo@yahoo.com.br

inerentes a esta pesquisa foram de responsabilidade dos próprios autores. Tratou-se de um estudo observacional transversal executado no período de janeiro a março de 2010.

O critério de inclusão foi: artigo publicado na Revista Brasileira de Anestesiologia entre janeiro de 2008 e dezembro de 2009. Foram excluídos outros tipos de artigo que não fossem o artigo original, tais como: artigos de revisão, informações clínicas, relatos de caso, artigos diversos, editoriais e cartas ao editor. O artigo era considerado original quando apresentava em sua descrição os relatos de um método de pesquisa, de apenas um ou de um conjunto de resultados e da interpretação e discussão dos resultados encontrados. O período de 2008 a 2009 foi escolhido por apresentar os artigos originais mais recentes na época de execução desta pesquisa.

A variável primária desta pesquisa foi a frequência do emprego adequado dos testes estatísticos de hipóteses utilizados na avaliação dos resultados. As variáveis secundárias foram: frequência do uso dos testes estatísticos, frequência do relato do valor exato de "p" nos resultados, frequência da presença da estatística descritiva (média, moda, mediana, desvio-padrão, amplitude, variância, erro-padrão, percentil e quartil), frequência do uso de análise de tabelas de contingência (qui-quadrado, teste exato de Fisher, McNemar e teste Z), frequência do uso dos testes avançados de estatística (regressão logística, regressão de Cox, modelo linear univariado e multivariado), frequência de artigos originais com emprego correto dos testes estatísticos, frequência do uso de intervalo de confiança, descrição de hipótese e descrição do cálculo do tamanho da amostra.

A utilização do teste estatístico foi considerada adequada quando:

- A seleção dos testes foi adequada a variáveis contínuas e categóricas e para testes paramétricos e não paramétricos.
- Houve descrição do fator de correção quando se relatou o uso de múltiplas comparações.
- Mencionou-se o uso específico de um teste estatístico para a análise de uma variável.

A análise dos testes foi inconclusiva quando:

- Não foi possível avaliar se a distribuição de variáveis contínuas era normal ou assimétrica.
- Os valores de "p" eram relatados, mas não havia especificações de quais testes haviam sido empregados para cada variável descrita nos resultados.
- Citavam-se o uso de testes e o valor de alfa previamente, porém nos resultados nem o valor de "p" nem os testes foram citados.

Se os dados assumissem a distribuição normal, um teste paramétrico seria considerado corretamente empregado, mas quando esse critério não era atingido considerava-se correto o uso de teste não paramétrico. A distribuição dos dados era considerada normal quando o autor do artigo original analisado relatava que a variável assumira distribuição normal; quando houve descrição da utilização dos testes de Kolmogorov-Smirnov, de Shapiro-Wilk e do teste da normalidade de D'Agostino-Pearson para analisar a distribuição dos dados da variável; pela observação da relação entre a média e o desvio-padrão; pela relação do cálculo do coeficiente de variação; e pela análise de gráficos demonstrados no artigo. O modelo de regressão linear foi considerado apropriado quando utilizado para variáveis contínuas. O uso de testes não paramétricos foi considerado adequado para as variáveis categóricas.

O cálculo do tamanho da amostra revelou a necessidade de se analisarem 76 artigos originais considerando a frequência do uso adequado dos testes de hipóteses de 70%, uma precisão absoluta de 10% e um nível de significância de 5% [7]. Utilizou-se uma estatística descritiva por meio da frequência simples e do intervalo de confiança de 95% para cada ponto estimado.

RESULTADO

Foram selecionados e analisados 76 artigos a partir dos mais recentes, abrangendo os volumes 59 e 58 da Revista Brasileira de Anestesiologia. Neles, encontrou-se um total de 179 testes de hipóteses. Os resultados das variáveis primárias e secundárias encontram-se nas Tabelas I e II.

Tabela I – Frequência de Uso dos Testes Estatísticos

Frequência dos testes, métodos estatísticos e métodos de regressão		
	Percentual (%)	Absoluto
Teste χ^2	20,11	36
Teste *t*	19,55	35
ANOVA	10,06	18
Fisher	9,50	17
Mann-Whitney	7,82	14
Kruskal-Wallis	6,70	12
Wilcoxon	3,91	7
Kolmogorov-Smirnov	3,35	6
Regressão linear múltipla	2,79	5
Correlação de Spearman	1,68	3
ANOVA rep	1,68	3
Regressão logística	1,12	2
Tukey	1,12	2
Curva de aprendizagem	1,12	2
Alfa de Cronbach	1,12	2
Scheffé	1,12	2
Student-Newman-Keuls	1,12	2
Mood	1,12	2
Friedman	0,56	1
Regressão linear simples	0,56	1
Kaplan-Meier	0,56	1
Curva CUSUM	0,56	1
Shapiro-Wilk	0,56	1
Bartlett	0,56	1
Kappa	0,56	1
Teste L	0,56	1
Log-Rank	0,56	1

Tabela II – Resultado das Variáveis Primárias e Secundárias: emprego de testes estatísticos de hipóteses

Frequência de uso dos testes estatísticos de hipóteses			
	Valor absoluto	Valor relativo	IC 95%
Adequado	101	56,42%	49,16% – 63,68%
Inadequado	24	13,41%	8,42% – 18,40%
Inconclusivo	54	30,16%	23,44% – 36,88%
Descrição do cálculo do tamanho da amostra			
	Valor absoluto	Valor relativo	IC 95%
Sim	20	26,32%	16,42% – 36,22%
Não	56	73,68%	63,78% – 83,58%
Descrição da hipótese da pesquisa			
	Valor absoluto	Valor relativo	IC 95%
Sim	8	10,53%	3,63% – 17,43%
Não	68	89,47%	82,57% – 96,37%
Descrição do valor de “p”			
	Valor absoluto	Valor relativo	IC 95%
Sim	63	82,89%	74,42% – 91,36%
Não	13	17,11%	8,64% – 25,58%
Emprego do valor de IC			
	Valor absoluto	Valor relativo	IC 95%
Sim	10	13,16%	5,56% – 20,76%
Não	66	86,84%	79,24 – 94,44%

A estatística descritiva esteve presente em todos os artigos pesquisados. Os artigos que só utilizaram estatística descritiva totalizaram 10,52% (8/76).

Levando-se em conta cada artigo original individualmente: 30,26% (23/76) apresentaram todos os métodos estatísticos utilizados de forma adequada, 22,36% (17/76) apresentaram todos os testes empregados incorretamente e 28,94 (22/76) apresentaram resultados inconclusivos. É preciso, também, mencionar que houve 0,39% (3/76) de artigos originais com testes estatísticos considerados corretos, associados a testes estatísticos inconclusivos, e 0,39% (3/76) com testes incorretos e inconclusivos.

DISCUSSÃO

Os três passos a serem considerados para definir qual o melhor teste a ser empregado para a análise estatística dos dados são: analisar a pergunta contida na pesquisa, determinar o nível de mensuração dos dados e definir o melhor delineamento de pesquisa a ser utilizado para elucidar o fenômeno ou os dados da população de interesse para a pesquisa [7]. Quando um teste estatístico é empregado de forma inadequada, os resultados encontrados podem não ser reprodutíveis nas populações.

A classificação dos artigos originais levando em conta o cálculo do tamanho da amostra evidenciou que 73,69% dos textos analisados apresentaram-se sem a descrição desse cálculo. O tamanho da amostra apresenta relação inversa com o valor de “p” encontrado pelos testes estatísticos, portanto amostras muito grandes tendem a apresentar baixos valores de “p” e vice-versa, enquanto amostras muito pequenas podem não evidenciar diferenças clinicamente significantes [8]. O tamanho adequado da amostra também permite estimar gastos e minimizar a aplicação de intervenções em um número maior do que o necessário de pacientes para a comprovação da hipótese da pesquisa [9]. Os autores desta pesquisa não procuraram avaliar o efeito dos resultados relatados nos artigos originais na prática clínica da anestesiologia, mas a aplicação adequada do teste estatístico para as variáveis apresentadas pelos autores dos artigos. O julgamento acerca da validade dos resultados relatados nos artigos originais deve ser realizado pelos leitores dos artigos, mas o cálculo do tamanho da amostra é um item que impõe qualidade à pesquisa executada, portanto, quando presentes no relato do artigo original, os resultados apresentados podem ter maior crédito. Não relatar o cálculo do tamanho da amostra não deve ser confundido com aplicar inadequadamente um teste estatístico. Quando um artigo retrata resultados sem significância estatística, isso não significa necessariamente que o efeito clínico pesquisado não exista, mas sim que o estudo talvez não tenha apresentado poder estatístico suficiente para captá-lo, por isso percebem-se, com muita frequência, frases nos artigos originais das mais variadas áreas do conhecimento enfocando indiretamente a importância desse cálculo, como: “mais estudos são necessários” ou “a amostra foi pequena para captar a diferença”.

O uso adequado dos testes estatísticos na amostra selecionada não superou os 70% assumidos na hipótese desta pesquisa e que se basearam na literatura médica internacional [7]. Os fatores que podem justificar esse achado se devem à consideração de que a maior parte dos erros no uso dos testes percebidos nesta pesquisa se deveu à utilização de teste *t* de *Student* para amostras pequenas, nas quais os autores desta pesquisa não conseguiram perceber que os dados tivessem assumido uma distribuição normal, e pelo uso de um teste paramétrico quando seria mais apropriado utilizar um teste não paramétrico. O resultado encontrado nesta pes-

quisa não retira seu crédito na comunidade científica, pois a média do uso adequado em revistas internacionais pode não chegar a 30% [3-6].

A análise da frequência do uso dos testes estatísticos evidenciou que o teste *t* de *Student* foi o teste paramétrico mais utilizado nos artigos originais que usaram testes estatísticos de hipóteses. Além disso, deixou claro que a estatística descritiva esteve presente em todos os artigos originais. Esses resultados encontrados corroboram outras pesquisas dentro ou fora do âmbito da terapia intensiva que demonstraram que o teste *t* de *Student* e a estatística descritiva são o tratamento estatístico mais utilizado nas pesquisas [3-7,9]. A condição de se analisarem dois grupos independentes é uma prática comum nas pesquisas da área médica e isso pode justificar a maior frequência do uso do teste *t* de *Student* [10]. A estatística descritiva serve para organizar e sumarizar os dados e representa o ponto final nas pesquisas de cunho descritivo e o ponto inicial em algumas pesquisas antes da realização dos testes de hipóteses [11]. A estatística descritiva auxilia na caracterização das populações e facilita a percepção do leitor quanto às diferenças ou semelhanças existentes.

A análise da frequência do uso dos testes estatísticos evidenciou, ainda, que os testes mais comuns foram o teste *t* de *Student* e o teste do qui-quadrado. O teste *t* de *Student* é um teste paramétrico que serve para avaliar a média de dois grupos quando os dados assumem distribuição normal [10]. O teste do qui-quadrado é realizado para avaliar as proporções [7]. Uma limitação na análise do uso adequado dos testes para tabelas de contingência encontrada nesta pesquisa foi a dificuldade de se perceber em qual situação foram utilizados o teste do qui-quadrado e o teste exato de Fisher, pois havia descrição do uso de ambos no método de alguns artigos, mas os resultados não expressavam em qual variável fora utilizado um ou outro teste. Descrições do tipo "foi utilizado o teste do qui-quadrado" ou "o teste exato de Fisher foi utilizado quando apropriado" impossibilitaram a análise do uso adequado. O relato de uma descrição mais clara acerca do uso de cada teste deveria ser encorajada aos autores dos artigos originais, por que facilitaria a interpretação dos resultados pelos leitores da revista analisada, assim como a percepção acerca da validação dos resultados.

O relato do valor exato de "p" foi evidenciado em 81,57% dos artigos originais que utilizaram teste estatístico. O valor de "p" demonstra a magnitude da significância estatística, porém o pesquisador deve demonstrar a importância clínica do resultado encontrado [9,12]. Utilizar apenas o valor referencial de "p" descrito na seção "métodos" para relatar o resultado em um artigo original prejudica a análise crítica deste artigo, por isso resultados seguidos das expressões $p > 0,05$ ou $p < 0,05$ devem ser evitados.

A descrição do intervalo de confiança esteve presente em 13,15% dos artigos originais analisados. É mais prático apresentar amostras estatísticas como estimativas do resultado que deveria ser obtido se toda a população fosse estudada, porém a falta de precisão que resulta do grau de variabilidade do fator estudado e o limitado tamanho do estudo podem influenciar os resultados [13]. Melhor estimativa do resultado pode ser mostrada pelo intervalo de confiança [13]. Esse intervalo pode ser visto como um sumário de resultados para alguns testes estatísticos e se revela mais informativo do que o resultado com relação à hipótese nula [14]. O intervalo de confiança ainda apresenta a vantagem de apresentar a significância estatística, demonstrando uma faixa de valores em que o verdadeiro valor populacional pode estar levando em conta determinado nível de confiança [13,14]. É muito mais vantajoso para o leitor apresentar os resultados do valor de "p", assim como os valores do intervalo de confiança, do que apresentar apenas uma dessas medidas, o que torna mais lógica a interpretação dos resultados.

Um artigo publicado na década de 1980 demonstrou que aproximadamente metade dos artigos publicados na área médica utilizou inadequadamente os testes estatísticos, sendo o teste *t* de *Student* o maior responsável pelos erros [15]. Algumas regras foram estipuladas para que os leitores possam estimar se os métodos estatísticos foram aplicados adequadamente. São elas: conhecer a diferença entre desvio-padrão e erro-padrão da média, entender o significado do valor de "p" e reconhecer um erro comum no uso do teste *t*. O uso do desvio-padrão mostra o quão distante os valores encontrados estão da média, pois, ao se somar e subtrair o valor de um desvio-padrão da média, tem-se a distribuição de 68% dos dados. O uso do erro-padrão demonstra uma homogeneidade de dados que talvez não seja real. O valor de "p" representa a probabilidade de um resultado ter ocorrido ao acaso, mesmo que não exista na população que deu origem à amostra. O teste *t* deve ser utilizado para a comparação de duas médias, e não para duplas de várias médias, pois isso aumenta a chance de se encontrarem resultados clinicamente importantes ao acaso.

A frequência do uso adequado dos testes estatísticos utilizados nos artigos originais publicados na Revista Brasileira de Anestesiologia entre janeiro de 2008 e dezembro de 2009 foi de 56,42%.

REFERÊNCIAS / REFERENCES

01. Windish DM, Hout SJ, Green ML – Medicine residentes' understanding of the biostatistics and results in the medical literature. JAMA, 2007;298:1010-1022.
02. Wullf HR, Anderson B, Brandenhoff P et al. – What do doctors know about statistics? Stat Med, 1987;6:3-10.
03. Avram MJ, Shanks CA, Dykes MH et al. – Statistical methods in anesthesia articles: an evaluation of two American journals during two six-month periods. Anesth Analg, 1985;64:607-611.
04. Hokanson JA, Luttman DJ,Weiss GB – Frequency and diversity of use of statistical techniques in oncology journals. CancerTreat Rep, 1986;70:589-594.
05. Cardiel MH, Goldsmith CH – Type of statistical techniques in rheumatology and internal medicine journals. Rev Invest Clin, 1995;47:197-201.
06. Huang W, LaBerge JM, Lu Y et al. – Research publications in vascular and interventional radiology: research topics, study designs, and statistical methods. J Vasc Interv Radiol, 2002;13:247-255.
07. Kurichi JE, Sonnad SS – Statistical Methods in the Surgical Literature. J Am Coll Surg, 2006;202:476-484.
08. Cavalcanti AB, Akamine N, Sousa JMA – Avaliação Crítica da Literatura. em: Knobel E – Condutas no Paciente Grave. 3ª ed. São Paulo, Atheneu, 2006;2635-2647.

09. Barbosa FT, Jucá MJ – Avaliação da qualidade dos ensaios clínicos aleatórios em anestesia publicados na revista brasileira de anestesiologia no período de 2005 a 2008. Rev Bras Anestesiol, 2009;59:223-233.
10. Gaddis GM, Gaddis ML – Introduction to biostatistics: part 4, statistical inference techniques in hypothesis testing. Ann Emerg Med, 1990;19:820-825.
11. McHugh ML – Descriptive statistics, part I: level of measurement. J Spec Pediatr Nurs,, 2003;8:35-37.
12. Gonçalves GP, Barbosa FT, Barbosa LT et al. – Avaliação da qualidade dos ensaios clínicos aleatórios em terapia intensiva. Rev Bras Ter Intensiva, 2009;21:45-50.
13. Gardner MJ, Altman DG – Confidence intervals rather than P values: estimation rather than hypothesis testing. BMJ, 1886;292:746-750.
14. Thompson WG – Statistical criteria in the interpretation of epidemiologic data. Am J Publ Health, 1987;77:191-194.
15. Glantz SA – Biostatistics: how to detect, correct and prevent errors in the medical literature. Circulation, 1980;61:1-7.

Resumen: Barbosa FT, Souza DA – Frecuencia del Uso Adecuado de los Test Estadísticos en los Artículos Originales Publicados en la Revista Brasileña de Anestesiología entre enero de 2008 y diciembre de 2009.

Justificativa y objetivos: La realización de un análisis estadístico se hace necesario para una evaluación pertinente del artículo original por parte del lector, ayudándolo a obtener una mejor visualización y comprensión de los resultados. El objetivo de esta investigación fue determinar la frecuencia del uso adecuado de los test estadísticos de hipótesis presentes en los artículos originales publicados en la Revista Brasileña de Anestesiología, entre enero de 2008 y diciembre de 2009.

Métodos: Se seleccionaron artículos originales publicados en la Revista Brasileña de Anestesiología entre enero de 2008 a diciembre de 2009. El uso de los test estadísticos se evaluó como apropiado cuando: la selección de los test fue satisfactoria para las variables continuas y categóricas y para el test paramétrico y no paramétrico; hubo una descripción del factor de corrección cuando se relató el uso de múltiples comparaciones; fue mencionado el uso específico de un test estadístico para el análisis de una variable.

Resultados: Se seleccionaron 76 artículos originales, con un total de 179 test estadísticos de hipótesis. La frecuencia de los test estadísticos más utilizados fue: 20,11% para el Chi-Cuadrado, 19,55%, para el test *t* de *Student*, 10,05% para el test de ANOVA y 9,49% para el test exacto de Fisher. La frecuencia de uso adecuado de los test estadísticos de hipótesis fue de un 56,42% (IC 95% 49,16% a 63,68%), de uso inadecuado, 13,41% (IC 95% 8,42% a 18,40%), con un resultado sin conclusiones en un 30,16% (IC 95% 23,44% a 36,88%).

Conclusiones: La frecuencia del uso adecuado de los test estadísticos utilizados en los artículos originales publicados en la Revista Brasileña de Anestesiología entre enero de 2008 y diciembre de 2009, fue de un 56,42%.