Rev Bras Cir Cardiovasc 2008; 23(3): 396-399

Leitura crítica dos dados estatísticos em trabalhos científicos

Critical reading of the statistical data in scientific studies

Mário José da CONCEIÇÃO, TSA[1]

RBCCV 44205-1006

Resumo

Objetivos: **A Estatística é ferramenta valorizada no testemunho da validade das conclusões dos trabalhos científicos. O objetivo dessa revisão foi apresentar alguns conceitos relacionados com os cálculos estatísticos que são fundamentais para a leitura e o pensamento críticos diante da literatura médica.**

Conteúdo: **Em geral, os autores apresentam os resultados de seus estudos na forma de gráficos, quadros e tabelas com dados quantitativos, acompanhados de estatísticas descritivas (médias, desvios-padrão, medianas) e quase sempre mencionando os testes estatísticos realizados. Após revisão, em inúmeros desses estudos, será difícil encontrar valor atribuível ao teste estatístico. Assim, fica ao leitor a tarefa de avaliar a adequação das informações e buscar as evidências contrárias aos possíveis erros que poderiam ameaçar a validade das conclusões.**

Conclusões: **Muitas vezes, pelo exame do desenho do estudo, observa-se o excessivo peso dado aos cálculos estatísticos como fatores definitivos, provas irrefutáveis, de conclusões discutíveis, quando não equivocadas.**

Descritores: **Análise estatística. Interpretação estatística de dados. Métodos e procedimentos estatísticos. Projetos de pesquisa /estatística & dados numéricos.**

Abstract

Objectives: **Statistics are a valuable tool that validates the conclusions of scientific works. The objective of this review was to present some concepts related to statistic calculations that are fundamental for the critical reading and analysis of medical literature.**

Contents: **In general, authors present the results of their studies as charts, boxes, and tables with quantitative data, along with descriptive statistics (means, standard deviations, medians), and almost always mention the statistic tests used. After reviewing several studies, it was difficult to find the value attributed to the statistical test. Thus, it is up to the reader to evaluate the adequacy of the information, and to search for evidence that contradict possible mistakes that could threaten the validity of their conclusion.**

Conclusions: **Examining the design of the studies one observes that, in many of them, excessive importance is given to statistical calculations as definitive factors, irrefutable evidence of arguable, or equivocal, conclusions.**

Descriptors: **Statistical analysis. Data interpretation, statistical. Statistical methods and procedures. Research design/statistics & numerical data.**

INTRODUÇÃO

A Estatística, ou Bioestatística como se convencionou chamá-la, quando aplicada às ciências biológicas, é uma ferramenta valorizada no testemunho da validade das conclusões dos trabalhos científicos. Em geral, os autores apresentam os resultados de seus estudos na forma de gráficos, quadros e tabelas com dados quantitativos, acompanhados de estatísticas descritivas (médias, desvios padrão, medianas) e quase sempre mencionando os testes estatísticos realizados. Os resultados desses testes são apresentados como valores de p. Após revisão, em inúmeros

1. Professor de Técnicas Cirúrgicas e Anestésicas – FURB – Blumenau – SC; Membro dos Conselhos Editoriais da Revista Brasileira de Anestesiologia, Pediatric Anesthesia e Regional Anesthesia and Pain Medicine; Co-Responsável pelo CET Integrado da SESSC – Florianópolis, SC

Endereço para correspondência:
Dr. Mário José da Conceição
Rua Germano Wendhausen, 32/401 - 88015-460 Florianópolis, SC
E-mail: marioconceicao@uol.com.br

Artigo publicado originalmente na Revista Brasileira de Anestesiologia Vol. 58, No 3, Maio-Junho, 2008.
Reproduzido com autorização dos Editores.
Descritores e Referências adaptados para as Normas da RBCCV.

desses estudos, será difícil encontrar valor atribuível ao teste estatístico. Assim, fica ao leitor a tarefa de avaliar a adequação das informações e buscar as evidências contrárias aos possíveis erros que poderiam ameaçar a validade das conclusões.

Milhares de trabalhos científicos são publicados anualmente em centenas de periódicos dedicados a divulgação de pesquisas oriundas da anestesiologia ou de áreas correlatas.

A esmagadora maioria, tanto nos estudos destinados a ciência básica, quanto naqueles de pesquisa clínica, usa a bioestatística para referendar suas conclusões. Pelo exame do desenho do estudo, observa-se o excessivo peso dado aos cálculos estatísticos como fatores definitivos, provas irrefutáveis, de conclusões discutíveis, quando não equivocadas.O objetivo dessa revisão foi apresentar alguns

conceitos relacionados com os cálculos estatísticos que são fundamentais para a leitura e pensamento críticos diante da literatura médica.

O erro da prova de equivalência

Na rápida passagem pelas páginas das revistas encontra-se na seção de Método dos vários artigos a insistente aparição do $p > 0,05$ ou o $p < 0,05$ significando, não-significativo ou significativo, sob o ponto de vista estatístico, respectivamente.

Ao encontrar o valor > 0,05 ou < 0,05, o autor passa a colocar todo o peso de seu estudo no valor desse cálculo e, de forma que julga brilhante, conclui pela existência (ou não) de determinado fenômeno, ou fato, estudado.

Aparentemente o problema piorou com o advento dos computadores e suas planilhas, facilitando sobremaneira esses cálculos e já com os mais diversos programas estatísticos embutidos. Essa facilidade colocou ao alcance dos autores a possibilidade de várias análises estatísticas. Nem sempre, porém, todos os autores estão preparados para seu uso adequado.

A morfina é um potente depressor da respiração, dependente de dose, seja por via venosa ou aplicada ao neuroeixo, verdade indiscutível, pelo menos até esse momento. Tome-se como exemplo estudo cujo resumo do método é o seguinte: dois grupos de pacientes tratados com morfina no neuroeixo, doses fixas, são deslocados no pós-operatório para dois lugares diferentes: um dos grupos segue para enfermaria comum, enquanto os pacientes do outro grupo são encaminhados para unidade de terapia intensiva. O objetivo do estudo foi avaliar a depressão respiratória nos pacientes tratados com morfina e a diferença entre os grupos.

O resultado apresentado foi $p > 0,05$, portanto sem diferença estatística entre os grupos. Os autores, com base nisso, concluem que pacientes tratados com morfina no neuroeixo não correm risco de depressão respiratória. Ora, o valor de $p > 0,05$ "sem diferença estatística significativa" sugere a falta de evidência de um efeito. Quando se lê "não houve diferença estatística entre os dois grupos" está-se diante de informação incompleta. O grande valor para p a rigor não significa ausência de efeito como erroneamente os autores concluíram. Apenas que os dados foram insuficientes para estabelecer a necessidade de vigilância no pós-operatório daqueles pacientes. Noutros artigos (comuníssimo nos de língua inglesa) por economia de espaço, ou outro motivo qualquer, os autores omitem o termo "estatística" e escrevem: "não houve diferença entre os grupos" ou "houve diferença significativa entre os grupos".

Diferenças de 5% podem ser significativas do ponto de vista clínico sem, no entanto, o serem do ponto de vista estatístico. No exemplo da morfina, se apenas um dos pacientes tivesse apresentado depressão respiratória, necessitando de suporte ventilatório, do ponto de vista clínico isso seria altamente significativo por motivos óbvios.

Ao ler uma conclusão baseada em valores para p (maior ou menor) o leitor deverá tão somente interpretar "diferença estatística entre os dois grupos". Incorreto, quando não perigoso, será assumir que houve equivalência entre os grupos para determinada ocorrência clínica observada [1].

O poder da amostra

Como seria inviável estudar todos os indivíduos atingidos pelo mesmo fenômeno, se retira dessa população um grupo de indivíduos que passarão a representá-la. A isso, chama-se amostra. Inúmeras vezes, p é maior do que 0,05, simplesmente porque o número de indivíduos estudados (amostra) é muito pequeno. Quantas vezes já foi lido: "trinta pacientes aleatoriamente etc...". Aliás, autores nacionais adoram a tal "randomização" e até "aleatorização". Estatísticos chamam a isso de erro tipo II. Isto é, quando não se detecta, em determinada amostra, o fenômeno observado, mas ele de fato existe [1]. Muitas teses universitárias trabalham com amostras pequenas pela falta de tempo entre o prazo para o término da pós-graduação e a colheita dos dados para escrevê-la. A probabilidade do estudo de detectar o fenômeno estudado, quando ele existe, chama-se "poder".

O poder depende da variabilidade nos grupos, do tamanho da amostra, da verdade do fenômeno a ser observado e do nível de significância. Um bom trabalho de investigação clínica deve informar o poder calculado da amostra, de tal forma que o leitor possa avaliar resultados "não-significativos pelos cálculos estatísticos". Seria bastante razoável se pensar que a depressão respiratória, após o uso da morfina no neuroeixo, deixou de se manifestar em 30 pacientes, mas bem poderia ter atingido o 32º paciente, se a amostra fosse de 35 pacientes. O poder da amostra é definido por porcentagem.

Uma amostra pode ser 40% confiável na detecção do fenômeno, ou 99% confiável. Desconfie sempre de grandes

amostras. Esse é um erro muito comum em trabalhos científicos: o(s) autor(es) pensa(m) que uma amostra enorme (5 mil casos, por exemplo) lhe(s) garante o direito de inferir resultados absolutos. Maior nem sempre significa melhor em termos de tamanho da amostra.

Assim, o autor precisa antes de iniciar a pesquisa planejar com cautela o tamanho da sua amostra, para que ele seja apropriado ao seu intento. Levantar seus 10 mil casos de qualquer coisa é absolutamente inapropriado [1], resultado de todo esse esforço? Nenhum.

A escolha errada do programa estatístico

Os pacotes estatísticos disponíveis no mercado, ou embutidos nas planilhas dos computadores, são incapazes de evitar que o pesquisador utilize o modelo errado, ou apontar as limitações do programa. Por exemplo, quantas vezes já se encontrou na literatura médica aplicação do teste de Bonferroni para validar resultados obtidos por Análise deVariância (ANOVA)? O teste de Bonferroni, ou teste de comparações múltiplas de Dunn, dispensa a ANOVA e não foi idealizado para comparações post-hoc (depois disso) e sim comparações a priori. Programa errado pode gerar $p < 0,05$.

Na leitura de pesquisas clínicas é preciso atenção redobrada quando testes estatísticos complexos indicam um determinado efeito que testes mais simples recusam. É preciso procurar entender se o autor descreve com cuidado o modelo utilizado (e o porquê), ou simplesmente refere um método automático para seleção de variáveis. Torna-se insuficiente o autor listar as variáveis que alimentaram seu programa, sem a garantia de que verificou se elas foram alocadas de forma correta [2].

Evidências oriundas de vários estudos

Um único artigo é insuficiente para decidir sobre um fenômeno observado. É bastante comum encontrar-se vários artigos sobre o mesmo assunto apresentando conclusões diferentes. Um único artigo pode apresentar diferença estatística, atestando a existência de determinado fenômeno estudado, entretanto dois ou três outros concluem exatamente pelo contrário.

Essas observações podem advir de erros. Valores de $p > 0,05$, como já mencionado, não são garantia de equivalência, mas falta de evidência de uma diferença estatística. Deduzir que o número de artigos, a favor e contra a evidência, define o problema, pode ser também em erro. A comparação entre estudos pode colocar lado a lado trabalhos que foram inapropriados ou com método mal planejado. Mais confiáveis são os estudos multicêntricos, combinando dados de diferentes locais. Pela ótica estatística, o trunfo dos estudos multicêntricos está em reduzir o intervalo de confiança para um determinado fenômeno observado, quando comparado com um único estudo [2]. Nesse contexto discute-se o poder das metanálises para validar observações clínicas. As opiniões entre especialistas divergem.

Entretanto, aparentemente uma metanálise, de pequenas amostras, dificilmente será o mesmo que um grande ensaio clínico advindo de estudo multicêntrico. Além do mais, metanálises não substituem observações clínicas bem planejadas.

Equilíbrio entre grupo-controle e grupo de estudo

A maioria dos estudos clínicos, em nossa área, inicia a descrição dos resultados apresentando comparação de características básicas entre dois grupos: sexo, idade, peso e estado físico, ao que chamam de "dados demográficos". A intenção do autor é mostrar aos leitores que os dois grupos são equilibrados. Com muita freqüência se acrescenta o valor de p para testar a diferença entre os dois grupos. Mesmo assim pode haver equívocos. Há diferenças entre pacientes dos grupos que poderão interferir nos resultados [3].

Por exemplo: observe a Tabela 1, extraída de uma análise de efeitos de bloqueadores neuromusculares em crianças. Os autores assumem (e induzem os leitores) que esses grupos são perfeitamente homogêneos. Todavia, nada é mencionado a respeito do estado de nutrição ou hidratação dessas crianças. Aqui $p < 0,05$ foi interpretado como prova inconteste da homogeneidade e que outras variáveis são desprezíveis independentemente do modelo de estudo. Curioso é que a recíproca pode ser verdadeira. Há métodos que utilizam o valor de $p < 0,05$ para provar a necessidade de inclusão de outras variáveis. Voltando à Tabela 1, $p < 0,05$ atesta a hipótese que a distribuição das variáveis não ocorreu por sorte ou de forma arbitrária. Todavia, no método, os autores afirmaram que a distribuição foi aleatória, então foi "por sorte".

Tabela 1. Características dos Pacientes que Receberam Mivacúrio após Atracúrio (Grupo AM), Cisatracúrio (Grupo CM) ou Mivacúrio (Grupo MM)

	Grupo AM	Grupo CM	Grupo MM
Idade (anos)	5,4 (2,3 -12,5)	6,0 (2,3 – 12,0)	5,8 (2,6 – 12,9)
Peso (kg)	20,0 (10,3 – 40,0)	21,0 (13,5 – 55,0)	23,0 (14,0 – 56,0)

Dados estão apresentados na forma de médias.
N =15 por grupo.
Não houve diferença estatística significativa entre os grupos.

O erro aqui está na certeza de que $p < 0,05$ determina as variáveis que devem (ou não) ser incluídas no modelo (sexo, idade, peso) e quais devem ser, com segurança, ignoradas (estado de nutrição, hidratação). Tratando-se de bloqueadores neuromusculares, indiscutivelmente o

estado de nutrição das crianças poderia ter interferido nos resultados, mas o sexo é pouco provável. É comum, entre os autores, pensar seja suficiente dizer apenas que os pacientes "foram selecionados aleatoriamente". No desenho do método algumas variáveis podem ter sido desprezadas [3], bem como modelos com muitas variáveis são de difícil interpretação e utilização.

Todavia, o autor necessita explicar o impacto das variáveis excluídas, nos seus resultados. A isto, chama-se "análise sensível". Os resultados tornam-se convincentes quando apresentados de forma correta. Alguns conselhos editoriais mais rigorosos solicitam, por parte do autor, o que muitas vezes causa indignação por parte desses, o envio dessas informações, inclusive a lista de dados de onde foram extraídos os resultados.

CONCLUSÕES

Se há pretensão de ler de forma crítica artigo científico é necessário apenas conhecimento básico dos princípios de estatística. As seguintes perguntas devem, todavia, ser respondidas:

• O autor forneceu informação quanto às medidas basais dos grupos em estudo?

• O autor usou intervalos de confiança na descrição de resultados, sobretudo se nenhuma evidência foi encontrada?

• Há inconsistência entre informações apresentadas em gráficos e quadros e as informadas e analisadas no texto?

• Os valores de p estão corretamente interpretados?

• O autor utilizou testes de ajustes (Newmann-Keuls, Dunnet e outros) para comparações múltiplas?

• O autor justificou com propriedade o modelo estatístico empregado? Modelos complexos não são necessariamente os corretos. É preciso ficar atento para o problema das comparações múltiplas com muitos testes estatísticos.

Nota de esclarecimento: os artigos publicados, consultados como exemplo de erros, foram omitidos dessas referências, por consideração ética aos autores. Foram consultados periódicos em língua inglesa, espanhola e portuguesa. Além disso, foi utilizada a experiência do autor em revisar artigos enviados para publicação em três periódicos diversos. Pelo mesmo motivo não foi referenciado o artigo no qual foi publicada a Tabela 1.

REFERÊNCIAS

1. Abramson JH. Survey methods in community medicine: epidemiologic studies. 5th ed. New York:Churchill-Livingstone;1999. p.311-25.

2. Avram MJ, Shanks CA, Dykes MH, Ronai AK, Stiers WM. Statistical methods in anesthesia articles: an evaluation of two American journals during two six-month periods. Anesth Analg. 1985;64(6):607-11.

3. Dawson B, Trapp RG. Bioestatística básica e clínica. 3a ed. Rio de Janeiro:McGraw-Hill;2003.