Rev Bras Anestesiol
2008; 58: 3: 260-266

Leitura Crítica dos Dados Estatísticos em Trabalhos Científicos

Critical Reading of the Statistical Data in Scientific Studies

Mário José da Conceição, TSA[1]

RESUMO

Conceição MJ — Leitura Crítica dos Dados Estatísticos em Trabalhos Científicos.

JUSTIFICATIVA E 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.*

Unitermos: ESTATÍSTICA: testes; METODOLOGIA CIENTÍFICA: estatística, projeto.

SUMMARY

Conceição MJ — Critical Reading of the Statistical Data in Scientific Studies.

BACKGROUND AND 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.*

Key Words: SCIENTIFIC METHODOLOGY: statistics, project; STATISTICS: tests.

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 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

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

*Apresentado (**Submitted**) em 12 de março de 2007*
*Aceito (**Accepted**) para publicação em 19 de fevereiro de 2008*

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

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 de Variância (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 deter-

minado 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 I, 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 a tabela I, $p < 0,05$ 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". 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 nutri-

Tabela I – Apresentada em sua Versão Original e Traduzida

Table I – Characteristics of Patients who Received Mivacurium after Atracurium (Group AM), Cisatracurium (Group CM) or Mivacurium (Group MM)

	Group AM	Group CM	Group MM
Age (yr)	5.4 (2.3 -12.5)	6.0 (2.3 - 12.0)	5.8 (2.6 - 12.9)
Weight (kg)	20.0 (10.3 - 40.0)	21.0 (13.5 - 55.0)	23.0 (14.0 - 56.0)

Data are presented as median (ranges).
N = 15 per group.
There were no statistically significant between-group differences.

Tabela I – 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.

ção 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 I.

Critical Reading of the Statistical Data in Scientific Studies

Mário José da Conceição, TSA, M.D.

INTRODUCTION

Statistics, or Biostatistics, as it is conventionally called when applied to biological sciences, is a valuable tool to validate the conclusions of scientific works. 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 they mention the statistical tests used in the analysis. Results of those tests are presented as values of "p". After reviewing several studies, it is difficult to find the value given to the statistical tests used. Therefore, it is up to the reader to evaluate the adequacy of the information presented and look for evidence that contradict the possible mistakes that could threaten the validity of the conclusions.

Thousands of scientific works dedicated to the divulgation of studies in the field of anesthesia, and correlated fields, are published every year in hundreds of journals. Biostatistics is used by the majority of those studies, including both basic sciences and clinical studies, to validate their conclusions. Examining the design of the study, one observes that excessive importance is given to statistical calculations as definitive factors, irrefutable evidence of arguable, and even mistaken, conclusions. The objective of this review was to present some concepts related with statistical calculations that are fundamental for the critical reading and analysis of medical literature.

The mistake of the equivalent test

Browsing through the pages of medical journals one will find on the Methods section of several articles the insistent presence of $p > 0.05$ or $p < 0.05$, which mean statistically non-significant and significant, respectively. On finding a $p > 0.05$ or < 0.05, the author fundaments all the importance of his/her study on the result of this calculation, and it is done in a way he/she considers brilliant, and concludes that the phenomenon or fact being studied exists (or does not). Apparently, this problem has worsened with the advent of computers and its charts, which include several statistical programs, facilitating considerably those calculations. This has made several statistical analyses available to authors. However, not every author is prepared to use them properly.

Morphine is a potent dose-dependent respiratory depressant, regardless whether it is administered intravenously or in the neuroaxis; an unquestionable truth, at least up to now. As an example, consider a study in which the summary of its method is as follows: two groups of patients treated with fixed doses of morphine administered in the neuroaxis.

Postoperatively, they are transferred to two different places: one group is transferred to the regular ward while the other group goes to the intensive care unit. The objective of the study was to evaluate the development of respiratory depression in patients treated with morphine and the difference between both groups. The study presented a result of $p > 0.05$, i.e., without statistically significant differences between both groups. Based on this result, the authors concluded that patients treated with morphine administered in the neuroaxis, are not at risk for respiratory depression. A $p > 0.05$, "without statistically significant differences" suggests lack of evidence of an effect. When one reads "statistically significant differences were not observed between both groups", one is not facing the complete information. The high value of p does not mean absence of an effect, as the authors wrongly concluded. It only means that the data was not enough to establish the need of postoperative observation of those patients. In other articles (very common on articles in English), for space purposes or any other reason, the authors omit the word "statistics" and write: "differences between groups were not observed" or "significant differences were observed between both groups". Differences of 5% can be clinically significant, but not statistically significant. Going back to the example of morphine, if only one patient had developed respiratory depression requiring ventilatory support, clinically it is highly significant, for obvious reasons. When reading a conclusion based on values of "p" (higher or lower), the reader should interpret only "statistical differences between groups". Incorrect, if not dangerous, would be to assume that there was equivalency between groups for a certain clinical occurrence observed [1].

Power of the sample

Since it is not feasible to study all individuals affected by the same phenomenon, one uses a group of individuals chosen from said population to represent it. This is called sample. Very often "p" is greater than 0.05 simply because the number of individuals in the study (sample) is too small. How many times one has read: "thirty patients randomly etc…" In fact, Brazilian authors love "randomized" and "randomization". Statisticians call this a type II error; i.e., when one does not detect, in a given sample, the phenomenon studied when it does exist[1]. Several post-graduate thesis work with small samples due to the short time available until the end of the post-graduation course and the amount of data that has to be gathered to write the thesis. The probability that a study will detect the phenomenon studied when it exists is called "power". Power depends on group variability, size of the sample, the true nature of the phenomenon being observed, and the level of significance. A good clinical study should inform the calculated power of the sample, so the reader can evaluate "non-statistically significant" results. It would be reasonable to think that respiratory depression, after the administration of morphine in the neuroaxis, did not manifest in 30 patients, but it could have developed on the 32nd patient if the study sample had 35 patients. The power of the sample is defined by a percentage. A sample can be 40% or 99% reliable to detect a phenomenon. Do not trust large samples. This is a common mistake in scientific studies: the author(s) think that a huge sample (for example, 5,000 cases) allows him to infer absolute results. Bigger is not always better when it comes to sample size. Therefore, the author should, before starting the study, carefully plan the size of the study sample, to make sure it is appropriate for his objectives. Gathering 10,000 cases of anything is absolutely inappropriate [1]; and the result of all this effort? None.

Choosing the wrong statistical program

Statistical packages available in the market, or those associated with the charts included in computers, cannot prevent the researcher from using the wrong model or indicate the limitations of the program. For example, how many times, in the medical literature, has the Bonferroni test been used to validate the Analysis of Variance (ANOVA)? The Bonferroni test, or the Dunn's test for multiple hypotheses, dispenses ANOVA and was not idealized for *post hoc* (after the fact) comparisons, but for *a priori* tests. The wrong program can generate $p < 0.05$. When reading clinical studies, one must pay attention when complex statistical tests indicate certain effects that simpler tests reject. It is necessary to understand whether the author describes carefully the model used (and why) or simply refers to an automatic method of selecting variables. It is not enough to mention the parameters that fed his program without the guarantee that he verified whether they were allocated correctly [2].

Evidence originated by several studies

One single article is not enough to make a decision about a phenomenon. It is very common to find several studies on the same subject with different conclusions. One study might present statistically significant differences, attesting the existence of a specific phenomenon, while two or three other studies present the opposite conclusion. Those observations might be a consequence of mistakes incurred. As mentioned before, values of $p > 0.05$ do not guarantee the equivalence, but it indicates the lack of evidence of a statistically significant difference. To infer that the number of studies, pro and against the evidence, define the problem can also be a mistake. A comparison among studies might align, on the same level, studies that are not appropriate or whose method was not properly planned. Multicenter studies, which combine data from different places, are more trustworthy. Statistically speaking, the advantage of multicenter studies, when compared with a single study, lies in the reduced confidence interval for a phenomenon [2]. In this context, one can argue the power of metanalysis to validate clinical observations. Experts diverge on this theme. However, a metanalysis of small samples is hardly the same as a large clinical assay resulting from a multicentric study. Besides, metanalysis do not substitute well-planned clinical observations.

Table I – Presented in its Original Version

Table I – Characteristics of Patients who Received Mivacurium after Atracurium (Group AM), Cisatracurium (Group CM) or Mivacurium (Group MM)

	Group AM	Group CM	Group MM
Age (yr)	5.4 (2.3 -12.5)	6.0 (2.3 - 12.0)	5.8 (2.6 - 12.9)
Weight (kg)	20.0 (10.3 - 40.0)	21.0 (13.5 - 55.0)	23.0 (14.0 - 56.0)

Data are presented as median (ranges).
N = 15 per group.
There were no statistically significant between-group differences.

Balance between control and study groups

Most clinical studies, in our field, begin the description of results by comparing basic characteristics between two groups: gender, age, weight, and physical status, which are called "demographic data". The intention of the author is to demonstrate to readers that both groups are balanced. Very often, the value of "p" is added to test the difference between both groups. But mistakes can still be made. There are differences among groups of patients that can interfere with the results [3]. For example, observe table I, which was extracted from an analysis of the effects of neuromuscular blockers in children. The authors assume (and also induce the reader) that those groups are perfectly homogenous. However, nothing is mentioned regarding their nutritional state or hydration status. Here, a $p < 0.05$ was interpreted as undeniable proof of the homogeneity of the groups, and that other parameters can be discarded, regardless of the study model. It is curious that the reciprocal can be true. There are methods that use a value of $p < 0.05$ to prove the need to include other parameters. Returning to table I, $p < 0.05$ attests that the distribution of parameters was not luck or arbitrary. However, under methods, the authors stated that distribution was at random; thus, it was "by luck". The mistake here lies in the certainty of the authors that $p < 0.05$ determines parameters that should (or should not) be included in the model (gender, age, weight) and which ones should be safely ignored (nutritional state, hydration). In the case of neuromuscular blockers, the nutritional state of the children could have, undeniably, interfered with the results, but it is probable that gender could not. It is common, among authors, to think that it is enough to mention that patients "were randomly selected". On the design of the method, some parameters could have been ignored [3]; on the other hand, models with too many parameters are difficult to interpret and use. However, the author must explain the impact on the results of the variables excluded. This is called "sensitivity analysis". Results become convincible when properly presented. More rigorous Editorial Boards ask the author to send this information, including the list of parameters from where the results were extracted, which causes indignation in many of them.

CONCLUSIONS

If one intends to read a scientific article critically, he/she needs to know only the basic principles of statistics. However, the following questions should be answered:

- Did the author provide information regarding the mean baseline parameters of the study groups?
- Did the author use confidence intervals on the description of the results, especially when no evidence was found?
- Are there inconsistencies between the information presented on charts and boxes and those in the body of the text?
- Is the interpretation of "p" values correct?
- Did the author use adjustment tests (Newmann-Keuls, Dunnet, and other) for multiple comparisons?
- Did the author justify adequately the statistical model used? Complex models are not necessarily correct. One should be attentive for the problem of multiple comparisons with many statistical tests.

Note: Articles consulted as example of mistakes, were not included in the references due to ethical consideration with the authors. Articles in English, Spanish, and Portuguese were reviewed. Besides, the experience of the author on reviewing articles for publication in three journals was used. For the same reason, the article in which table I was published was not mentioned on the references.

REFERÊNCIAS – *REFERENCES*

01. Abramson JH — Survey Methods in Community Medicine: Epidemiologic Studies, 5th Ed., New York, Churchill-Livingstone, 1999;311-325.
02. Avram M — Statistical methods in anesthesia articles: an evaluation of two American journals during two six-month periods. Anesth Analg, 1985;64:604-611.
03. Dawson B, Trapp RG — Bioestatística Básica e Clínica, 3ª Ed., Rio de Janeiro, McGraw-Hill, 2001;7-20.

RESUMEN

Conceição MJ — Lectura Crítica de los Datos Estadísticos en Trabajos Científicos.

JUSTIFICATIVA Y OBJETIVOS: *La Estadística es la herramienta valorada como testimonio de la validez de las conclusiones de los trabajos científicos. El objetivo de esa revisión fue presentar algunos conceptos relacionados a los cálculos estadísticos que son fundamentales para la lectura y el pensamiento críticos frente a la literatura médica.*

CONTENIDO: *En general los autores presentan los resultados de sus estudios en la forma de gráficos, cuadros y tablas con datos cuantitativos, acompañados de estadísticas descriptivas (promedios, desvíos estándar, medianas) y casi siempre mencionando los tests estadísticos realizados. Después de la revisión, en innumerables estudios, será difícil encontrar un valor atribuible al test estadístico. De esa forma le queda al lector la tarea de evaluar la adecuación de las informaciones y buscar las evidencias contrarias a los posibles errores que podrían amenazar la validez de las conclusiones.*

CONCLUSIONES: *Muchas veces, por medio del examen del diseño del estudio, se observa el excesivo peso dado a los cálculos estadísticos como factores definitivos, pruebas irrefutables, de conclusiones discutibles, cuando no están equivocadas.*