Itens Essenciais em Bioestatística

Ângela Tavares Paes

São Paulo, SP

Provavelmente, desde o século XVII [1], as ciências da saúde vêm recorrendo à estatística como instrumento para a análise de fenômenos biológicos. Neste longo período, muitos conceitos e mudanças surgiram nos dois campos do conhecimento. Por um lado, os estatísticos começaram a desenvolver técnicas, motivados principalmente pela sua aplicação, por outro, os médicos passaram a dar ênfase à mensuração como estratégia de análise científica e, assim, a medicina progressivamente sofisticou suas análises quantitativas.

Nas últimas décadas, esta progressão foi vertiginosa apoiada pelo crescimento da atividade científica em todos os campos e pela revolução tecnológica representada particularmente pelos computadores eletrônicos.

Estimulada pelos desafios das ciências da saúde, a estatística respondeu tão vigorosamente que uma nova disciplina, a bioestatística, emergiu em seu meio, organizando um referencial teórico próprio e gerando uma grande variedade de conceitos, métodos e técnicas de análise. Para ser abordado, este universo tão vasto requer, inequivocamente, um estudo por partes, o que se busca neste artigo com a seleção de alguns itens para discussão.

Nível descritivo *(p-value)*

Na grande maioria dos artigos científicos as conclusões são baseadas nele; atualmente são raros os que não o utilizam; pesquisadores são ávidos para o revelarem, mas afinal o que é o "valor do p"?

O "valor do p" ou *p-value* é conhecido na estatística como nível descritivo e está associado ao que chamamos de testes de hipóteses [2]. Portanto, para falar sobre isso é necessário uma breve introdução aos principais conceitos.

Podemos definir como hipóteses questões levantadas relacionadas ao problema em estudo e que, se respondidas, podem ajudar a solucioná-lo. O papel fundamental da hipótese na pesquisa científica é sugerir explicações para os fatos. Uma vez formuladas as hipóteses, estas devem ser comprovadas ou não através do estudo com a ajuda de testes estatísticos. Num teste estatístico são formuladas duas hipóteses chamadas hipótese nula (H_0) e hipótese alternativa(H_1). Hipótese nula é aquela que é colocada à prova, enquanto que hipótese alternativa é aquela que será considerada como aceitável, caso a hipótese nula seja rejeitada.

A grosso modo, nos problemas mais simples da área médica, a hipótese nula está associada à uma igualdade entre médias ou proporções que podem indicar a não associação (independência) entre fatores de interesse. Por exemplo, num estudo sobre fatores de risco para doenças cardiovasculares, uma hipótese nula poderia ser "a proporção de doentes cardiovasculares entre hipertensos é igual à proporção entre não hipertensos" ou "a chance da doença é a mesma para hipertensos e não hipertensos". Isto implicaria em dizer que "não existe associação entre hipertensão e doença cardiovascular". Outro exemplo, desta vez considerando igualdade de médias, pode ser descrito por um estudo sobre tempo de recuperação de pacientes transplantados. Supondo que desejamos comparar três procedimentos cirúrgicos diferentes, uma possível hipótese seria "o tempo médio de recuperação é o mesmo nos três procedimentos cirúrgicos", ou seja "o tipo de procedimento cirúrgico não influencia no tempo de recuperação do paciente".

Todo teste de hipótese possui erros associados a ele. Um dos mais importantes é chamado "erro do tipo I" que corresponde à rejeição da hipótese nula quando esta for verdadeira. No exemplo da doença cardiovascular, a probabilidade do erro do tipo I seria a probabilidade de concluir que há associação quando na verdade não há, ou seja, concluir uma associação que não existe (que é devida ao acaso). No exemplo do tempo de recuperação, o erro do tipo I corresponderia a dizer que o tipo de procedimento cirúrgico influencia no tempo de recuperação quando na realidade o tempo médio é o mesmo nos três procedimentos. A probabilidade do erro do tipo I chama-se nível de significância e é expressa através da letra grega α. Os níveis de significância usualmente adotados são 5%, 1% e 0,1%.

Formalmente, o nível descritivo (p) é definido como o "menor nível de significância (α) que pode ser assumido para se rejeitar H_0", porém esta interpretação não é simples até mesmo para os estatísticos. Considerando, de maneira muito generalizada, que os pesquisadores ao rejeitarem a hipótese nula costumam dizer que existe "significância estatística" ou que o resultado é "estatisticamente significante", poderíamos definir o nível descritivo (p) como a "probabilidade mínima de erro ao concluir que existe significância estatística".

Instituto Dante Pazzanese de Cardiologia – São Paulo
Correspondência: Ângela Tavares Paes – Laboratório de Epidemiologia e Estatística – Av. Dr. Dante Pazzanese, 500 – 04012-180 – São Paulo, SP
Recebido para publicação em 9/6/98
Aceito em 11/7/98

É importante ressaltar que o nível de significância (α) é um valor arbitrado previamente pelo pesquisador, enquanto que o nível descritivo (p) é calculado de acordo com os dados obtidos. Fixado α e calculado o "p", a pergunta é: "será que posso dizer com segurança que o resultado é estatisticamente significante?". Para responder à esta questão é necessário avaliar se a probabilidade de erro é "aceitável" ou não, isto é, se o "valor do p" é pequeno o suficiente para concluir que existe "significância estatística" dentro de uma margem de erro tolerável. Mas saber "o que é pequeno ou grande" depende do nível de significância adotado, portanto a decisão do pesquisador sempre estará baseada na comparação entre os dois valores. Se o valor do p for menor que o nível de significância (α) deve-se concluir que o resultado é significante pois o erro está dentro do limite fixado. Por outro lado, se o valor de p for superior à α significa que o menor erro que podemos estar cometendo ainda é maior do que o erro máximo permitido, o que nos levaria a concluir que o resultado é não significante pois o risco de uma conclusão errada seria acima do que se deseja assumir. Segue abaixo um esquema que resume a regra de decisão descrita.

p-value $< \alpha \Rightarrow$ rejeito $H_0 \Rightarrow$ diferença
p-value $> \alpha \Rightarrow$ não rejeito $H_0 \Rightarrow$ igualdade

p-value	α	p-value
resultado "estatisticamente significante"		resultado "não significante"

A grande vantagem de se utilizar o nível descritivo é a possibilidade de "quantificar" a significância, ou seja, no lugar de uma resposta do tipo "sim ou não" temos a informação de "quanto". Considere os exemplos da tabela abaixo:

Resposta usual	***p-value***
(*) ou p < 0,05	p=0,0002 e p=0,048
n.s.	P=0,085 e p=0,987

Note que no 1º exemplo os dois resultados são "significantes", porém o valor de 0,0002 expressa uma significância muito maior do que 0,048. Além disso, este último valor é muito próximo ao nível usual de 5%, o que pode causar dúvidas ou ressalvas na tomada de decisão. No 2º exemplo temos dois resultados não significantes. O 2º valor (0,987) praticamente não expressa significância estatística nenhuma pois o erro é de quase 100%. Já o 1º (0,085) embora não seja significante ao nível de 5% é um valor bastante indicativo. Portanto, é muito mais valioso e informativo expressar as conclusões através do valor exato do p em vez de apenas menor ou maior que o nível de significância (α) fixado.

Intervalos de confiança

Médias, medianas, modas são chamadas estimativas pontuais pois correspondem a um único valor que estima características de um grupo sob estudo. Existem também as estimativas por intervalos que são expressas por um limite inferior e um superior entre os quais acredita-se estar o verdadeiro valor do parâmetro. Por exemplo, num estudo em pacientes hipertensos pode-se dizer que a pressão arterial (PA) média é 87, 5 variando de 85,7 a 89,3 (intervalo de confiança = [85,7; 89,3]).

Uma das utilidades dos intervalos é dar a idéia da dispersão ou variabilidade das estimativas. Um intervalo muito grande indica que a estimativa calculada não é tão acurada quanto outra com intervalo menor, ou seja, quanto maior a amplitude do intervalo menor a confiabilidade da estimativa.

Existem vários métodos para expressar intervalos, sendo exemplos o valor máximo e o valor mínimo e os intervalos de percentis, como o intervalo 25% - 75%. O mais conhecido e talvez o mais correto seja o "intervalo de confiança"[3] que permite incorporar uma probabilidade de erro. Esta probabilidade de erro é inferida a partir de um conhecimento do modelo de distribuição de freqüências do fenômeno estudado. O modelo que mais habitualmente se ajusta à ocorrência de fenômenos biológicos é o de distribuição normal, cujo intervalo de confiança envolve para sua construção o conhecimento da variância (que permite o cálculo do desvio padrão). Os intervalos podem ser construídos com diferentes coeficientes de confiança, sendo em geral mais utilizados os intervalos de confiança de 95% ou 99%. A cada coeficiente corresponde um valor crítico da distribuição, que é uma medida de distância da estimativa pontual que se expressa em unidades de desvios padrão. Abaixo descrevemos informalmente a fórmula geral dos intervalos de confiança.

IC = estimativa pontual $\pm$ valor crítico da distribuição $*$ desvio padrão da estimativa

Quando se constrói um intervalo de confiança para se descrever a variabilidade de uma medida, o desvio padrão utilizado é o desvio padrão da medida em questão. Quando se constrói um intervalo de confiança para valores possíveis para uma estimativa pontual, por exemplo, uma média, o desvio padrão utilizado é uma estimativa de desvio padrão para uma suposta série de medidas de médias. Este é um caso especial de desvio padrão que recebe o nome de erro padrão da média.

Exemplos:

		Intervalo de confiança de 95% (Valor crítico ≈ 2)	
PAD média	87,53	[85,72	89,34]
	Estimativa pontual = média	Estimativa pontual - valor crítico x erro padrão	Estimativa pontual + valor crítico x erro padrão

		Intervalo de confiança de 95% (Valor crítico ≈ 2)	
% hipertensos	23,15% Estimativa pontual = proporção	[19,85% Estimativa pontual – valor crítico x erro padrão	26,45%] Estimativa pontual + valor crítico x erro padrão

É comum em artigos médicos os valores de medidas estarem expressos na forma de médias mais ou menos desvio padrão, como por exemplo 87,5±4,8. A reação natural do leitor é subtrair e somar este valor e interpretar como limites de intervalo de confiança. Porém, este cálculo corresponde a se criar um intervalo de confiança de 65%, correspondente ao valor crítico de 1 desvio padrão, e leva a um intervalo muito menor do que o habitual intervalo de confiança de 95%, cujo valor crítico é de aproximadamente 2.

Além de informar sobre a variabilidade/dispersão de estimativas pontuais, os intervalos de confiança podem também expressar a "significância estatística" dos testes referentes às comparações. Por exemplo, num teste de comparação de duas médias, um intervalo de confiança para a diferença entre as médias que contém o valor zero indica que a diferença não é significante, ou seja, que não existe diferença entre as médias. Já em uma comparação de proporções em que se deseja estimar o risco relativo, a ausência de significância se dá quando o intervalo para o risco relativo contém o valor 1, pois isto indica que as duas proporções podem ser iguais.

Exemplos:

Comparação entre duas médias				
média 1	média 2	dif.	IC 95% p/dif.	Obs.
53,4	62,8	-9,4	[7,6 ; 11,2]	não contém o valor 0 (∴médias diferentes)
45,9	47,2	-1,3	[-3,3 ; 0,30]	contém o valor 0 (∴médias iguais)

Estimação de risco relativo		
odds ratio	IC 95% p/odds	Obs.
2,13	[1,58 ; 2,68]	não contém o valor 1 (∴há risco)
1,25	[0,48 ; 2,02]	contém o valor 1 (∴não há risco)

Uma maneira bastante eficiente de resumir resultados de comparações de médias é analisar os intervalos de confiança graficamente. O exemplo a seguir refere-se à comparação entre pacientes de três faixas etárias (até 55 anos, 55 a 70 anos e mais de 70 anos) com relação à PA diastólica média. Note que entre os dois primeiros grupos existe uma grande superposição dos intervalos enquanto que o intervalo correspondente ao grupo dos pacientes com mais de 70 anos não se sobrepõe, ou seja, o limite inferior deste intervalo ainda é maior que o limite superior dos demais. Isto implica que provavelmente um teste estatístico para comparar as três médias indicaria que não existe diferença entre os dois primeiros grupos e que pacientes do 3º grupo tem PA média maior que nos outros dois.

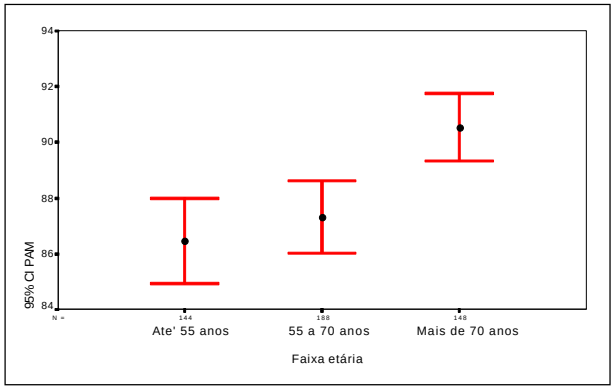


Gráfico I - Comparação entre três grupos de idade em relação à pressão arterial média. Fonte: Instituto Dante Pazzanese de Cardiologia – pacientes submetidos à cirurgia cardíaca entre 1993 e 1994

Podemos analisar também através de gráficos, intervalos de confiança para riscos relativos. Neste caso, costuma-se traçar a linha que passa pelo valor 1 e são considerados significativos os riscos relativos correspondentes aos intervalos que não cruzarem com esta linha, isto é, que não contiverem o valor 1.

O exemplo a seguir é baseado em um estudo cujo objetivo é identificar fatores de risco para impotência sexual em pacientes cardiopatas [4]. Observe que, entre as variáveis

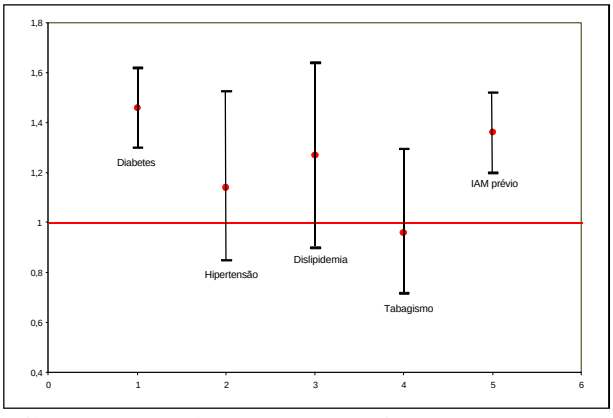


Gráfico II - Avaliação dos fatores de risco para impotência sexual.

investigadas, apenas os intervalos referentes a diabetes e infarto agudo do miocárdio prévio não passam pela linha do 1, ou seja, a influência desses fatores na chance de impotência sexual é considerada significante a um nível de 5%.

Relevância clínica x significância estatística

Não se pode acreditar cegamente em tudo que os testes estatísticos mostram [5]. O que o médico deve se perguntar ao interpretar os resultados de uma pesquisa é "os resultados obtidos são relevantes do ponto de vista clínico?".

Muitas vezes um resultado "estatisticamente significante" pode não ser " clinicamente importante". Por exemplo, um teste de comparação de médias pode detectar uma diferença de 2mmHg na PA como sendo "altamente significante" apesar desta diferença não ter nenhuma implicação clínica. Portanto, a importância em termos biológicos não deve ser julgada pelos estatísticos, mas sim pelos profissionais da área em que a pesquisa está sendo feita.

A figura a seguir [6] mostra cinco tipos de resultados, expressos em termos de intervalos de confiança, que exemplificam diferentes situações. A linha horizontal cheia corresponde ao valor que, se presente no intervalo, indica que o resultado não é estatisticamente significante - intervalo de confiança da diferença incluindo o valor zero. Por exemplo, num teste estatístico que compara duas médias, o intervalo referente à diferença entre elas não deve conter o valor zero.

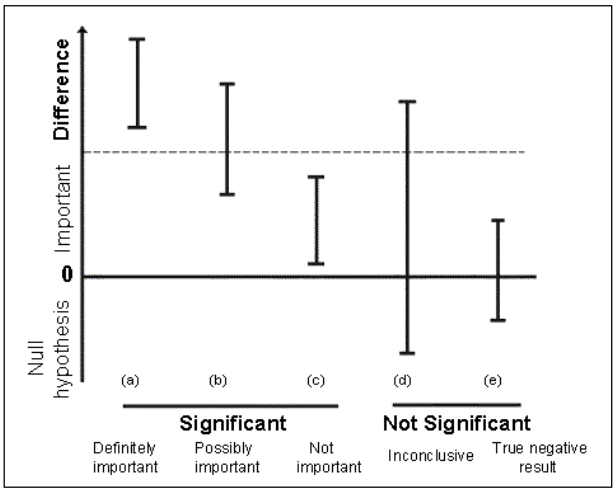
A linha pontilhada representa o valor a partir do qual a diferença é considerada importante na prática. No exemplo da PA, o médico pode optar por considerar relevante apenas as diferenças acima de 5 ou 10mmHg.

Assim, são considerados estatisticamente significantes os intervalos que não cruzam com a linha cheia. No entanto, somente aqueles que ultrapassam o limite da relevância clínica (linha horizontal pontilhada) é que devem ser de fato considerados relevantes.

Desta forma podemos analisar as cinco situações representadas na figura. A situação (a) mostra um resultado considerado definitivamente importante, pois além de significante, todo o intervalo está acima do que é considerado clinicamente relevante. A situação (b) corresponde a um resultado que é estatisticamente significante mas cujo intervalo cruza com a linha da relevância, ou seja, não se pode afirmar com certeza que o resultado seja relevante mas este é possivelmente um resultado importante.

O 3º caso (c) traduz um dos erros mais freqüentes em pesquisas na área médica. Mostra um resultado que apesar de significante do ponto de vista estatístico, não tem relevância clínica nenhuma pois o limite superior do intervalo ainda é inferior ao valor considerado importante. Muitas vezes isto não é levado em consideração pelo clínico que toma conclusões baseadas nos resultados estatísticos sem avaliar a real importância da diferença detectada.

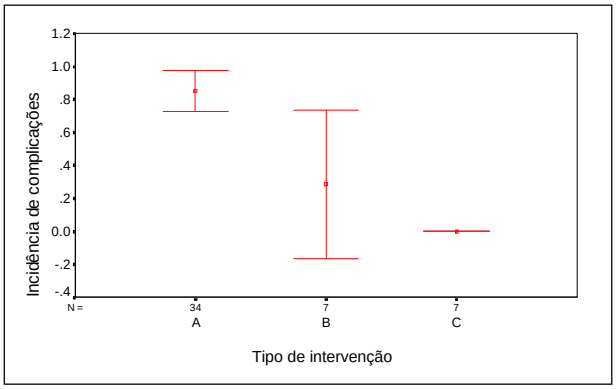
Os dois últimos casos correspondem a resultados não significantes. Como já foi observado no item anterior intervalos com grande amplitude correspondem a estimativas pouco precisas. Isto acontece na situação (d) que mostra



Intervalos de confiança e as cinco possíveis interpretações sobre significância estatística e importância prática [6].

um intervalo bastante impreciso. Isto torna o resultado inconclusivo pois o verdadeiro valor pode estar tanto abaixo do valor estatístico quanto acima do clínico. Já o último caso representa um resultado que chamamos de verdadeiro negativo, pois ambas as conclusões coincidem.

Por outro lado, existe também a situação inversa. Um resultado que não seja "estatisticamente significante" pode ser muito importante, não devendo ser desconsiderado. O exemplo abaixo corresponde a um estudo que examinou complicações pós-operatórias em três tipos de intervenção cirúrgica para uma patologia congênita rara das vias hepáticas [7].O teste de comparação de médias indicou que não havia diferença significante, porém pode-se notar que o procedimento do tipo "C" registra uma freqüência de ocorrência de complicações muito menor que a dos outros grupos. Provavelmente, uma eventual diferença significante entre os grupos não pode ser registrada devido ao pequeno número de observações, mas por se tratar de doença rara não resta ao investigador a opção de aguardar um maior número de casos para uma reanálise de seus dados. Seria insensato, e provavelmente antiético, ignorar as evidências disponíveis de superioridade da intervenção do tipo "C". Ainda que os resultados tenham sido obra do acaso como sugere a análise estatística, ao médico compete orientar sua prática de acordo com os conheci-



mentos que dispõe, ainda que se mantenha alerta para eventuais mudanças à luz de novas evidências.

Portanto, não se deve fechar os olhos para um resultado que não seja significativo, é preciso levar em conta também a importância do objeto que está sendo estudado.

O tamanho da amostra

Uma pergunta muito freqüente em estudos na área é "qual deve ser o tamanho da minha amostra?". Esta é uma questão delicada e muitas vezes polêmica. Considere o seguinte exemplo: se nos fosse perguntado "quanto de dinheiro preciso levar para as minhas férias?" a resposta imediata seria "depende". Depende do lugar que deseja ir, quanto tempo pretende ficar, quantas pessoas, qual o meio de transporte e, é claro, entre outros detalhes, qual o dinheiro disponível. Da mesma forma, arbitrar um tamanho adequado de amostra envolve conhecimento da natureza das medidas realizadas, do plano de análise, do nível de erro aceitável para estimativas etc.

Há com freqüência uma ênfase excessiva ao cálculo do tamanho de amostra em detrimento da concepção cuidadosa de um plano amostral [8], que são as estratégias a serem adotadas para garantir que a amostra a ser estudada seja representativa do universo real do fenômeno a ser estudado. Os vícios [9] de seleção, de detecção, de exposição, de informação ou de memória não serão prevenidos por qualquer definição de tamanho de amostra, mas sim por um plano amostral cuidadoso. O tamanho da amostra vai depender da viabilidade de coleta de dados, que envolve principalmente tempo, custos e disponibilidade de casos para serem estudados. Isto não significa que o cálculo de tamanho de amostra seja dispensável. O que desejamos salientar aqui é que ele deve ser utilizado como planejamento, isto é, como parte de um estudo bem delineado onde ele não substitua o compromisso do investigador de analisar a representatividade dos casos estudados, seja qual for o número a ser observado.

Uma das vantagens de se calcular corretamente o tamanho da amostra é a possibilidade de economia. Por exemplo, um estudo bem planejado pode, a partir de uma amostra não muito grande, obter as mesmas conclusões de um estudo que envolveu uma amostra muito maior por não ter sido previamente planejado.

Entretanto, o cálculo do tamanho da amostra não garante um resultado significante [10]. É conveniente planejar o tamanho da amostra para que se possa ter amostras grandes o suficiente para detectar diferenças importantes (amostras muito pequenas podem deixar que diferenças importantes passem desapercebidas). Por outro lado, amostras exageradamente grandes além de elevar o custo do estudo, podem tornar diferenças clinicamente irrelevantes em estatisticamente significativas.

Para o planejamento do tamanho da amostra o investigador precisa estabelecer algumas definições como: tipo de estudo que pretende realizar (ex. estudo de prevalência, ensaio clínico, coorte, caso-controle); o tipo de medida que deve utilizar (ex. medidas contínuas, categorizadas, prevalência, incidência); o tipo de análise (ex. diferenças entre médias, diferença entre proporções, cálculo de risco); a margem de erro que pode assumir para o estudo (ex. o nível de significância e o poder do teste estatístico que pretende aplicar).

Estes conceitos podem ser melhor esclarecidos na *homepage* do Laboratório de Epidemiologia e Estatística (www.lee.dante.br) que apresenta um serviço que calcula tamanhos de amostra para alguns dos desenhos de pesquisa médica/biológica mais freqüentes, além de oferecer textos de apoio para compreensão de cada item envolvido no cálculo e referências bibliográficas para orientarem interessados num estudo autônomo.

Que técnica estatística utilizar?

Imagine a seguinte situação: um paciente chega ao consultório médico e já na recepção preenche uma ficha sobre os principais sintomas que o levaram ao consultório. A recepcionista registra os dados no computador e depois de alguns segundos devolve ao paciente uma receita com o nome dos medicamentos que ele deve usar, sem sequer ter sido examinado pelo médico. Isto só seria possível se existisse um tratamento padrão associado a cada diagnóstico que tornasse desnecessário qualquer exame complementar sobre características do paciente (sexo, idade), medicação prévia, outros sintomas, etc.

Da mesma forma que a prática da Medicina não é completamente objetiva, a da Estatística também não. Não existem "receitas prontas" para tratar doentes, assim como não existem fluxogramas que indiquem as técnicas estatísticas que devem ser utilizadas em cada caso. O que existem são "práticas comuns" que podem ser aplicadas ou não, dependendo das condições do estudo. Por exemplo, um médico pode optar por um medicamento alternativo no lugar de um mais comum, devido às condições do paciente (pode ser um paciente idoso cujo medicamento não é recomendado). Portanto, o que pretendemos ressaltar é que cada caso deve envolver uma análise particular, assim como em um exame clínico, de forma que a escolha da técnica seja feita com critério e cuidado.

Conclusões

"Where shall I begin, please your Majesty?" asked the White Rabbit.

"Begin at the begining and go on till you come to the end: then stop". Said the King of Hearts. Lewis Carol. *Alice in Wonderland: Alice's evidence*.

A interdisciplinaridade da ciência moderna convida, senão exige, que profissionais de diferentes filiações acadêmicas colaborem para a produção do conhecimento. Em ciências da saúde, médicos e estatísticos precisam buscar conciliar seus conhecimentos para uma colaboração adequada [11]. O presente artigo é um modesto esforço neste

sentido. Longe de se propor como referencial didático, busca tentativamente sugerir uma agenda preliminar de tópicos para uma uniformização conceitual. A escolha dos itens abordados não foi aleatória, mas fruto do que nos sugerem alguns anos de convivência com profissionais de saúde. Foram preteridos itens relativos às estratégias de análise (por exemplo, os testes estatísticos mais utilizados em estudos médicos), mas como ensina em parábolas de contos infantis o matemático Lewis Carol, comecemos pelo começo.

Referências

1. Graunt J - Natural and political observations mentioned in a following index, and made upon bills of mortality. In: Wilcox WF (ed) - Natural and Political Observations Made upon Bills of Mortality by John Graunt. Baltimore: Johns Hopkins Press, 1937.
2. Matthews DE, Farewell VT - Using and Understanding Medical Statistics. New York: Karger, 1988: 17-19.
3. Intervalos de confiança. In: Fonseca JS, Martins GA - Curso de Estatística. São Paulo: Atlas, 1982: 162-75.
4. Izukawa, NM - Impotência sexual arteriogênica em pacientes portadores de arterosclerose das artérias coronárias. São Paulo, 1997 (dados originais, processamento fictício para criação de exemplo).
5. Feinstein AR - Why clinical epidemiology. Clin Research 1973; 20: 821-5.
6. Armitage P, Berry G - Statistical Inference. In: Statistical Methods in Medical Research. 3rd ed. Oxford: Blackwell, 1994: 98-9.
7. Herman P - Litíase intra-hepática primária: resultados do tratamento cirúrgico e análise dos fatores prognósticos. Tese de Doutorado apresentada à FMUSP em 1997. (dados reais, processamento fictício).
8. Hulley SB, Gove S, Browner WS, Cummings SR - Choosing the study subjects: specification and sampling. In: Hulley SB, Cummings SR - Designing Clinical Research. Baltimore: Williams Wilkins, 1988: 18-29.
9. Rothman KJ - Modern Epidemiology. Boston: Little Brown, 1986: 83-94
10. Vieira S - Metodologia Científica para a Área de Saúde. São Paulo: Sarvier, 1984: 77-82.
11. Feinstein AR - Stochastic significance, consistency, apposite data, and some other remedies for the intellectual pollutants of statistical vocabulary. Clin Pharmacol Ther 1977; 22: 113-23.