Significância com confiança?

Mário B. Wagner
Doutor em Epidemiologia (Universidade de Londres)
Professor Adjunto, Departamento de Medicina Social,
Faculdade de Medicina, Universidade Federal do Rio Grande do Sul

Fonte:
Jornal de Pediatria 1998; **74**:343-346.

Resumo
Objetivos: Discutir brevemente aspectos da inferência estatística em medicina, salientando as limitações da significância obtida nos testes de hipótese da escola estatística clássica de Neyman-Pearson, e a utilidade das estimativas de efeito por intervalos de confiança. Métodos: Revisão de diversos livros de epidemiologia, bioestatística e artigos selecionados. Resultados: Existem diversos testes de significância estatística que são freqüentemente encontrados em artigos publicados na literatura médica. Os resultados desses testes podem levar a conclusões enganosas se não forem adequadamente interpretados. Uma vez que a significância estatística não assegura relevância aos achados do estudo, sugere-se a utilização de medidas de associação e seus respectivos intervalos de confiança para a avaliação da significância clínica. Conclusão: O rótulo "estatisticamente significativo" é potencialmente enganoso e tem estado sob constantes críticas por parte de vários estatísticos e epidemiologistas durante os últimos anos. Para que se possa avaliar "significância" é igualmente ou mais importante do que o valor P que se estime o tamanho da associação envolvida. Isto pode ser obtido através de medidas de associação como o risco relativo e seus intervalos de confiança.

Abstract
Objectives: To briefly discuss aspects of statistical inference in medicine, pointing out limitations of the traditional statistical hypothesis testing under the Neyman-Pearson school, and the advantages of effect estimation using confidence intervals. Methods: Review of a number of epidemiology and biostatistics textbooks and selected articles. Results: There are a number of statistical significance tests which are frequently found in articles published in the medical literature. The results of these tests, however, can be misleading if not properly interpreted. Since statistical significance does not ensure relevance to study findings, measures of association and related confidence intervals are suggested for the proper consideration of clinical significance. Conclusion: The label of "statistically significant" finding is potentially misleading and has been under constant criticism by many statisticians and epidemiologists during recent years. In order to evaluate "significance" it is equally or more important than the P value to estimate the size of the association involved. This can be achieved by using measures of association such as the relative risk and related confidence intervals.

Introdução

Nos artigos publicados em revistas médicas, além da freqüente utilização do método epidemiológico moderno são empregadas técnicas estatísticas (algumas bastante avançadas), tanto no planejamento e condução dos estudos, como na análise e interpretação dos resultados. Este fenômeno ocorre de tal forma, que o clínico geral interessado em manter-se atualizado com a literatura não pode mais dar-se ao luxo de deixar os aspectos epidemiológicos e estatísticos dos estudos aos "*experts*", sob pena de prejudicar o entendimento global dos artigos que lê.

Termos como "estatisticamente significativo" ou expressões do tipo $P < 0,001$, são freqüentemente encontrados em artigos médicos referindo-se aos resultados dos estudos. Muitas vezes as interpretações desses termos e expressões deixam a desejar e podem até induzir o leitor a conclusões incorretas. Este artigo pretende discutir brevemente alguns aspectos da inferência estatística na pesquisa em medicina, das limitações da chamada "significância" obtida nos testes de hipótese da escola estatística clássica de Neyman-Pearson, e da utilidade da estimativa de efeito por intervalos de confiança.

Testes de significância estatística

Diferentes tipos de variáveis requerem diferentes tipos de análises. As variáveis (fatores ou características que medimos nos pacientes) podem assumir quatro níveis de medida: nominal, ordinal, intervalar e de razão. As variáveis nos níveis de medida nominal e ordinal são freqüentemente também chamadas de variáveis qualitativas, enquanto que aquelas nos níveis intervalar e de razão são consideradas variáveis quantitativas. Maiores detalhes sobre as definições dos níveis de medida das variáveis podem ser encontrados em livros texto de bioestatística[1,2] ou em recente publicação do *Jornal de Pediatria*[3].

De uma forma geral, variáveis quantitativas são analisadas através de testes paramétricos e as qualitativas por testes não-paramétricos[2]. Os testes paramétricos recebem este nome devido aos seus parâmetros fundamentais: a média e o desvio padrão. Além disso, os testes paramétricos partem do pressuposto que os dados a serem analisados seguem um padrão de distribuição conhecido como curva Normal ou curva de Gauss. Já os testes não-paramétricos são baseados em outra abordagem estatística a qual não se baseia na média nem no desvio padrão. Geralmente, em testes não-paramétricos os dados são classificados em postos ou posições (ranks) e comparados sem que haja a necessidade de seguirem um padrão específico de distribuição. Por esta razão os testes não-paramétricos são também conhecidos como testes de distribuição livre[4].

Assim, para a análise de dados quantitativos os testes preferenciais são os testes paramétricos. Esses testes são considerados mais poderosos (maior capacidade de detectar diferenças) do que os testes não-paramétricos e por isso, sempre que possível devem ser utilizados em dados quantitativos. Os pressupostos básicos para a utilização dos testes paramétricos são três: (a) dados quantitativos, (b) padrão de distribuição compatível com a curva Normal e (c) homogeneidade de variâncias (homocedasticidade) entre os grupos a serem comparados. São exemplos de testes paramétricos: teste *t* de Student, coeficiente "r" de correlação de Pearson e análise de variância (ANOVA).

Já os testes não-paramétricos não possuem pressupostos específicos. São utilizados na análise de dados qualitativos e são os substitutos nas situações onde os testes clássicos (paramétricos) não podem ser utilizados. São exemplos de testes não-paramétricos: teste U de Mann-Whitney, coeficiente de correlação de Spearman, teste de Kruskal-Wallis e teste de qui-quadrado.

Na escolha de um teste estatístico, além do tipo de variável envolvida na análise, é importante avaliar se os dados são oriundos de observações pareadas ou independentes. De uma forma geral, dados pareados são gerados quando ocorrem observações seriadas no mesmo indivíduo ou pareamento no delineamento do estudo. Dados independentes referem-se a grupos diferentes de indivíduos nas comparações.

Finalmente, deve-se considerar quantos grupos serão comparados (2 grupos ou 3+ grupos) e quantas variáveis estarão envolvidas simultaneamente na análise (bivariada ou multivariada). O Quadro 1 apresenta uma lista de alguns dos testes estatísticos freqüentemente utilizados na pesquisa médica.

Quadro 1: Testes e técnicas estatísticas freqüentemente encontradas em artigos publicados na literatura médica

ANÁLISES BIVARIADAS

Comparação de Grupos

Tipo de observação	2 grupos			3 ou + grupos		
	Razão/Intervalar	Ordinal	Nominal	Razão/Intervalar	Ordinal	Nominal
Independente	t Student independente	U de Mann-Whitney	χ^2; RR; exato de Fisher	ANOVA de um critério (oneway)	teste de Kruskal-Wallis	χ^2
Pareada	t Student pareado	T de Wilcoxon	χ^2 de McNemar; RR	ANOVA de dois critérios (two-way)	teste de Friedman	teste de Cochran

Correlação de variáveis

Nível de medida	
Razão/Intervalar	Ordinal
Coeficiente de Pearson	Coeficientes de Spearman ou Kendall

Concordância entre observadores

Coeficiente Kappa

ANÁLISES AVANÇADAS

Nome	Variável dependente (desfecho)	Variáveis independentes (fatores em estudo)	Comentários
Análise Estratificada	dicotômica	dicotômicos ou politômicos	Possibilta o controle do efeito de alguns fatores simultaneamente. Fácil de conduzir, mas rapidamente ineficiente a medida que mais fatores são considerados.
Modelos Multivariados			Os modelos multivariados são mais eficientes do que a análise estratificada para o manejo simultaneo de diversos fatores
Regressão linear	quantitativa	quantitativas ou dicotômicas	Modelo de origem. Estima coeficientes angulares "b".
Regressão logística	dicotômica	dicotômicas, politômicas ou quantitativas	Estima *odds ratios* ajustados para o efeito de diversos fatores simultaneamente.
Regressão de Cox	tempo para ocorrência de um evento	dicotômicas, politômicas ou quantitativos	Estima riscos relativos (via densidade de incidência) ajustados para o efeito de diversos fatores simultaneamente.

Significância estatística versus intervalo de confiança

A avaliação do papel do acaso (ou efeito da variabilidade amostral) vem sendo feita tradicionalmente em medicina através de testes de hipótese ou testes de significância. A significância pode ser avaliada pelos mais variados testes estatísticos (veja alguns exemplos no Quadro 1) e é geralmente expressa através do chamado valor P.

Segundo a teoria estatística, o valor P é contínuo, varia entre 0 e 1, e representa a compatibilidade dos dados observados com a hipótese nula, ou seja, a hipótese de que não há associação entre desfecho e fator em estudo. Apesar da escala contínua do valor P, muitos pesquisadores insistem em classificar os resultados dos testes estatísticos em uma dicotomia do tipo "sim" ou "não". Desta forma, os resultados dos estudos são considerados significativos quando $P \leq 0{,}05$ e não significativos quando $P > 0{,}05$. E, segundo esta abordagem, valores $P = 0{,}06$ e $P = 0{,}60$ são absurdamente tratados da mesma forma: como resultados não significativos.

Em vista disso, deve-se salientar que "significância estatística" envolve uma questão de probabilidade de que exista ou não uma associação entre desfecho e fator em estudo (*qualquer* associação de *qualquer* tamanho). Assim, se a amostra for pequena, associações moderadas ou até mesmo fortes podem ser consideradas não significativas (seriam então associações *in*significantes?). Por outro lado, em grandes estudos epidemiológicos, até mesmo fracas associações atingem com facilidade a marca de $P = 0{,}001$.

Com isso, pretende-se deixar claro que existe uma grande diferença entre significância estatística e significância (relevância) clínica. A significância estatística refere-se *exclusivamente* ao fato da associação observada ser, na verdade, diferente de zero. Conseqüentemente, a significância estatística *nada informa sobre o tamanho ou importância clínica da associação*.

Uma vez que a significância estatística pode levar a interpretações inadequadas dos achados, o que se deve fazer então? Diversos autores entre epidemiologistas e estatísticos[5-8] tem argumentado que sempre que possível deve ser feita uma avaliação cautelosa das diferenças absolutas e/ou relativas observadas entre os grupos de estudo. É sugerido que sejam calculadas medidas de associação tipo risco relativo (também são aceitas abordagens alternativas como diferenças de médias, proporções ou outros índices) de forma que se possa estimar a força da associação.

Para a avaliação do papel do acaso (variabilidade amostral) sugere-se o uso de intervalos de confiança para as diferenças ou risco relativo. Rothman & Greenland[9] questionam a aparente arbitrariedade do nível de significância (α) fixado em 0,05 e argumentam que em situações onde temos uma associação forte outros níveis (p.e., $\alpha = 0{,}10$) podem ser utilizados, obtendo-se intervalos de confiança de 90%.

Quando comparada com o valor P, a abordagem via intervalo de confiança na avaliação de uma associação apresenta pelo menos três vantagens. O valor P, como representante da significância estatística, informa simplesmente a compatibilidade dos dados com a hipótese em teste. Já o intervalo de confiança, por sua vez, informa simultaneamente: (a) uma estimativa da magnitude da associação (p.e., risco relativo); (b) a variabilidade desta estimativa, através da amplitude dos limites inferior e superior do intervalo; e (c) a compatibilidade dos dados com a hipótese em teste. (Para a interpretação de intervalos de confiança de medidas de associação, veja Wagner & Callegari-Jacques[10]).

Considerações finais

O uso quase que compulsivo de testes de significância é generalizado na literatura médica. O achado "significativo" ou um rótulo de $P < 0{,}05$ é interpretado de forma ingênua por muitos médicos e pesquisadores como um certificado do tipo "ISO 9002" de qualidade de seus dados. Este tipo de visão vem sendo duramente criticada por diversos grupos de epidemiologistas e estatísticos[5-8] que acreditam ter ocorrido uma supervalorização do famoso teste de hipótese da

escola estatística clássica de Neyman-Pearson. Assim, sugere-se que na avaliação de "significância" seja calculada uma estimativa do tamanho ou força do efeito através de medidas de associação (p.e., risco relativo) com o cálculo adicional de intervalos de confiança para determinar o papel da variabilidade amostral.

Referências

1. Kirkwood BR. *Essentials of medical statistics*. Oxford: Blackwell Scientific Publications, 1988.
2. Altman DG. *Practical Statistics for Medical Research*. London: Chapman & Hall, 1991.
3. Wagner MB. Aspectos básicos da descrição e sumarização de informações em medicina. *Jornal de Pediatria* 1998; **74:** 71-76.
4. Siegel S. *Estatística Não-Paramétrica*. São Paulo: McGraw-Hill, 1975.
5. Gardner MA & Altman DG. Confidence intervals rather than P values: estimation rather than hypothesis testing. *Br Med J* 1986; **292:** 746-750.
6. Pocock SJ, Hughes MD, Lee RJ. Statistical problems in the reporting of clinical trials. *N Engl J Med* 1987; **317:** 426-432.
7. Greenland S. Randomization, statistics, and casual inference. *Epidemiology* 1990; **1:** 421-429.
8. Goodman SN. P values, hypothesis tests, and likelihood: implications for epidemiology of a neglected historical debate. *Am J Epidemiol* 1993; **137:** 485-496.
9. Rothman KJ & Greenland S. Approaches to Statistical Analysis. In: *Modern Epidemiology*, edited by Rothman KJ & Greenland S. Philadelphia: Lippincot-Raven, 1998, pp. 181-200.
10. Wagner MB & Callegari-Jacques SM. Medidas de associação em estudos epidemiológicos: risco relativo e odds ratio. *Jornal de Pediatria* 1998; **74:** 247-251.