Artigos de Revisão

Epidemiologia explicada – O valor de prova (p)

Francisco Botelho[1, 2], Carlos Silva[1, 3], Francisco Cruz[1, 3]

1 – Serviço de Urologia do Hospital de S. João, Porto
2 – Serviço de Higiene e Epidemiologia, Faculdade de Medicina da Universidade do Porto, Porto
3 – Faculdade de Medicina da Universidade do Porto

Resumo

Uma investigação quase sempre detecta diferenças entre grupos. Será a diferença devida ao acaso ou podemos rejeitar esta hipótese (a hipótese nula) e concluir que as diferenças são reais? Esta revisão faz uma introdução à filosofia por detrás dos testes de hipóteses (significância) e ao valor de p. Apresenta as linhas de orientação para uma correcta interpretação dos valores de p e são apresentados os erros mais frequentes nesta área. Os métodos para o cálculo desta variável estatística são apresentados nos casos com que os investigadores mais frequentemente se deparam.

Abstract

The result of a single experiment will almost always show some difference between groups. Is the difference due to chance, or can we reject this (the null) hypothesis and conclude that there is a true difference? The present review introduces the general philosophy behind hypothesis (significance) testing and *p* values. Guidelines for the interpretation of *p* values are also provided, along with some of the common pitfalls. The methods for calculation of this statistical variable are presented in the most frequents scenarios for the researchers.

Porque surgiu a necessidade da utilização do valor de prova (p) na investigação científica?

A análise de dados da larga maioria dos estudos investigacionais inclui comparações entre dois ou mais grupos de indivíduos. Estas comparações podem ser as mais diversas: sobrevida de doentes tratados com fármaco A ou B; percentagem de cura em indivíduos com e sem determinado factor de prognóstico; presença de determinado factor de risco em doentes e não doentes.

Desde cedo os investigadores aperceberam-se que, sobretudo quando se utilizam amostras pequenas, os resultados podiam ser influenciados por factores aleatórios. Por exemplo, ao lançar cinco vezes um dado com a mão direita e cinco vezes com a mão esquerda, obtêm-se provavelmente valores superiores com uma mão do que com outra. Isto não se deve a uma maior predisposição de uma mão em obter resultados mais elevados, mas apenas ao efeito dos factores aleatórios. Repetindo a experiência os resultados seriam provavelmente diferentes.

Assim, quando se comparam dois grupos, como se sabe se as diferenças encontradas se devem ou não ao acaso? A verdade é que nunca se pode ter a certeza. A resposta da estatística a esta questão foi arranjar fórmulas matemáticas, cujo

resultado forneça uma ideia da probabilidade dos resultados poderem ser devidos ao acaso. Esta solução deixa ao critério dos investigadores o considerarem essa probabilidade relevante ou não. Essa probabilidade é traduzida pelo valor de prova (usualmente designado por "p").

O que é o p e qual o seu significado?

O valor de prova (p) é uma estimativa da probabilidade de se obter resultados iguais ou mais extremos (isto é, com maiores diferenças entre os grupos), partindo do pressuposto que não existem diferenças entre os grupos em análise relativamente à variável em estudo[1]. Este pressuposto é designado em termos estatísticos por hipótese nula. Dito de uma forma simplista é a probabilidade de obter esse resultado, se na realidade não existissem diferenças entre os grupos em análise.

Num estudo hipotético, doentes submetidos ao tratamento A apresentam uma sobrevida superior em 6 meses aos tratados com o tratamento B. O valor de p respectivo foi 0,08. Este valor significa que, se não existissem diferenças de eficácia entre os dois tratamentos, a probabilidade de um dos tratamentos obter diferenças de sobrevida iguais ou superiores a 6 meses, num estudo semelhante, seria de 8%.

De forma errónea, o valor de p é frequentemente interpretado como medida da magnitude da associação entre duas variáveis. Na verdade, apenas informa a probabilidade que uma associação, identificada no estudo, seja um achado falso-positivo decorrente do acaso. Outra interpretação errada frequente é a de que corresponde à probabilidade de os resultados estarem correctos. Contudo, para os resultados serem correctos, não podem ser influenciados por erros aleatórios (avaliados pelo valor de p), nem por viéses e confundimento (cuja existência não é passível de avaliação por nenhuma fórmula estatística).

Como valorizar o valor de p?

Como traduz uma probabilidade, o valor de p varia entre 0 e 1. Quanto mais próximo de 0, maior a probabilidade da diferença encontrada entre os grupos não ser devido a factores aleatórios. Antes de iniciar a análise dos resultados o investigador deve reflectir sobre qual o valor de p que considera suficientemente baixo para poder desprezar eventuais erros aleatórios, rejeitando a hipótese nula. Este valor de corte é designado por nível de significância.

Por consenso, é habitual considerar-se que caso o valor de p seja inferior a 0,05, a probabilidade dos resultados serem devidos a erros aleatórios é suficientemente baixa para ser desprezada e assim considerarem-se os resultados como "estatisticamente significativos". Por oposição, caso o valor seja superior a 0,05, consideram-se os resultados como "estatisticamente não significativos".

Mais importante do que classificar um resultado como estatisticamente significativo ou não, é fornecer o valor de p e deixar outros valorizar essa probabilidade de forma crítica. Assim, não é correcto referir apenas $p<0,05$ ou $p>0,05$, devendo-se sempre indicar o valor exacto do p[2].

Um resultado "estatisticamente significativo" significa que a probabilidade da diferença encontrada entres os grupos ser devida a factores aleatórios é mínima e por isso pode ser desprezada. Isto não significa que o resultado seja correcto. Qualquer resultado, para além de poder ser afectado por erros aleatórios, pode ser afectados por erros sistemáticos (viéses e factores de confundimento). A discussão destes vai para além do âmbito deste artigo, mas erros cometidos no planeamento do estudo, recolha das informações ou análise dos dados não são avaliados pelo valor de p. Estes erros apenas podem ser avaliados pelo investigador ou leitor de um artigo científico que reflicta de forma crítica sobre os métodos utilizados.

Como regra, o valor de p é tanto menor quanto maiores as diferenças entre os grupos e quanto maior o tamanho amostral. Por isso, estudos científicos com tamanhos amostrais da ordem de grandeza dos milhares de participantes, podem detectar de forma estatisticamente significativa pequenas diferenças entre grupos. Se estas diferenças são "clinicamente" significativas ou não, é uma avaliação subjectiva, dependendo de múltiplos outros factores. Portanto, resultados estatisticamente significativos podem não ter qualquer relevo clínico[3].

Inversamente, um estudo com tamanho amostral pequeno pode detectar diferenças importantes e reais entre grupos, mas não conseguir obter significância estatística. Assim, quando um resultado não é estatisticamente significativo,

Dados	Nº de grupos comparados	Teste de Significância frequentemente utilizados
Paramétricos*	2	t-student
	>2	ANOVA
Não paramétricos	2	Mann-Whitney U
	>2	Kruskal-Wallis
Emparelhados		t-sudent emparelhado

Tabela 1 – Principais testes de significância utilizados para comparar médias/medianas

* o principal critério para podermos considerar os dados paramétricos é que a que a variável dependente tenha uma distribuição normal no histograma

pode dever-se a duas situações: na realidade não existe diferença entre os grupos ou existe diferença entre os grupos mas o estudo não teve capacidade (designada poder estatístico), para a demonstrar. Na verdade porém, nunca podemos ter a certeza de qual das duas hipóteses é a correcta.

O valor de p não é, contudo, a única forma de avaliar a probabilidade de existência de erros aleatórios. O Intervalo de Confiança de 95% também permite inferir se uma medida de associação é estatisticamente significativa. A definição e interpretação do Intervalo de Confiança está fora do âmbito deste artigo. Como regra, considera-se o resultado estatisticamente significativo, se o valor 1 não estiver contido no intervalo. Por exemplo, Freedland[4] reporta que a obesidade está associada a carcinoma prostático referindo um Odds Ratio de 1.98, com respectivo Intervalo de Confiança de 95% a variar entre 1.17 a 3.32. Isto significa que quem é obeso tem uma probabilidade de apresentar carcinoma prostático 98% superior aos não obesos (informação obtida do Odds Ratio) e que este resultado é estatisticamente significativo (já que o valor 1 não está contido no intervalo de confiança). O Intervalo de Confiança é mais informativo que o valor p. Permite, para além da avaliação da significância estatística de um resultado, a avaliação directa do grau de precisão da estimativa calculada.

Como se calcula o valor p?

O valor p é calculado através de diferentes testes matemáticos cuja escolha depende dos dados e análise pretendida. Se se comparar proporções (percentagens de cura, por exemplo) geralmente usa-se o teste qui-quadrado (χ^2) ou uma das suas variantes. Para comparação de médias ou medianas entre grupos diferentes (valor médio na escala do IPSS, por exemplo) existem vários testes que podem ser realizados, como os descritos na tabela 1.

Existem ainda outras técnicas estatísticas que são utilizadas em situações particulares e que também fornecem o valor p: Teste de Pearson ou Spearman (quando se avalia a associação de duas variáveis contínuas por correlação), Regressão Multivariada (quando se pretende ajustar para factores de confundimento), Regressão Cox ou Log Rank Test (quando utilizamos Análise de Sobrevivência).

Todos estes testes, actualmente disponíveis em diversos pacotes de software estatístico, estão muito acessíveis ao utilizador comum. No entanto, da sua má utilização podem resultar erros importantes, pelo que quem não tiver conhecimentos e prática nesta área deverá procurar ajuda especializada.

Bibliografia

1. Whitley E, Ball J. Statistics review 3: hypothesis testing and P values Crit Care 2002; 6 (3): 222-225
2. Cleophas TJ, Blume J, Peipert JF. Clinical trials: Renewed attention to the interpretation of the P values – review. What your statistician never told you about P-values. Am J Ther 2004; 11 (4): 317-322
3. Blume J, Peipert JF. What your statistician never told you about P-values. J Am Assoc Gynecol Laparosc 2003; 10 (4): 439-444
4. Freedland SJ, Wen J, Wuerstle M, Shah A, Lai D, Moalej B, et al. Obesity is a significant risk factor for prostate cancer at the time of biopsy. Urology 2008; 72 (5): 1102-1105