ESTATISTICA EM MEDICINA: *P-VARICAÇÃO*

Basílio de Bragança Pereira

Palavras Chave: Intervalos de confiança. Probabilidade. Sensibilidade. Especificidade. Teste diagnóstico. Valor-p.

STATISTICS IN MEDICINE: P-VALUE

Estatístico
Professor Titular, Instituto de Matemática - COPPE, Universidade Federal do Rio de Janeiro

Recebido em: 19/07/95
Aceito em: 20/12/95

INTRODUÇÃO

A interação de trabalho entre o médico e o estatístico têm permitido a observação de que os testes de significância, o Valor-p e os intervalos de confiança são usados freqüentemente de forma errada. A razão parece ser: 1) falta de entendimento destes conceitos; 2) a resposta que o pesquisador procura não pode ser obtida utilizando esses conceitos. Por outro lado, existe atualmente uma grande cobrança na arena científica e nos jornais médico-científicos no sentido da obtenção de "significância" nos resultados dos estudos comparativos. Isto produz um efeito talvez mais nefasto. Diversas pesquisas e experimentos bem realizados não são publicados porque os resultados não foram "significantes" estatisticamente[1]. Salsburg[2] compara esta situação a um fanatismo religioso, onde Salvação corresponde à publicação em um jornal de prestígio (que produz frutos: aumento de salário, convites, etc.). O ritual é a busca do Valor-p (outro nome para 0,05!). O milagre é a obtenção do Valor p < 0,05. O padre a quem se procura aconselhamento é o Estatístico que muitas vezes faz perguntas "irrelevantes" como: Porque e como você realizou este experimento? Não obtendo resposta convincente e não entendendo a necessidade urgente de salvação, começa uma discussão teológica com o seguidor.
O objetivo deste artigo é tentar esclarecer os conceitos estatísticos usando exemplos médicos da literatura, apresentar uma solução alternativa e fornecer sugestões para o pesquisador na área médica de como proceder ao analisar seu experimento utilizando a estatística.

INTERPRETAÇÃO INCORRETA DO VALOR-P

A verificação da falta de entendimento do significado do Valor-p, tem sido testado em turmas de pós-graduação de Medicina e Engenharia usando os seguintes questionários de Diamond e Forrester [1] e Freeman [3] respectivamente.

Questionário 1 - (Diamond e Forrester)

O que você concluiria se um experimento clínico bem planejado, realizado para verificar o efeito de um certo tratamento, resultou em uma resposta benéfica (p < 0,05)?

a) de acordo com este resultado, as chances são menos de 5% de que a terapia não tem efeito;
b) as chances são menos de 5% em obter este resultado se a terapia não tem efeito;
c) as chances são menos de 5% de não ter obtido esse resultado se a terapia tem efeito;
d) nenhuma acima.

Questionário 2 - (Freeman)

Um experimento controlado, realizado para determinar a eficácia de um novo tratamento, conclui que o mesmo é significativamente melhor que placebo (p < 0,05). Qual das seguintes afirmações você prefere?

a) foi provado que o tratamento é melhor que placebo;
b) se o tratamento não tem efeito, há menos de 5% de chance de se obter tal resultado;
c) o efeito observado do tratamento é tão grande que há menos de 5% de chance do tratamento não s[illegible]lhor que placebo;
d) realmente não sei o que é Valor-p e não quero adivinhar.

A conclusão obtida com as aplicações destes questionários coincide com a dos autores. A resposta correta em ambos é b) mas em geral mais de 50% das pessoas respondem incorretamente e todos tem dificuldades de distinguir a diferença entre as escolhas.

TESTE DIAGNÓSTICO E TESTE DE HIPÓTESE

Em testes estatísticos, copiamos a estratégia matemática de provar por contradição. Começando com uma hipótese H_0 que se quer rejeitar, supomos que H_0 é verdadeiro. Desenvolvendo argumentos de forma correta, se chegamos a uma contradição, então, a hipótese H_0 deve ser falsa. Em estatística, copiamos este enfoque, mas em vez de atingir uma contradição, observamos um resultado improvável. Especificamente, começando com uma hipótese nula (por exemplo, que não existe diferença entre dois tratamentos), observamos o resultado de um experimento bem planejado. A seguir, verificamos quão provável é o resultado observado no estudo, supondo não haver diferença entre os tratamentos. Se calculamos através de um procedimento de teste estatístico que o resultado do estudo é improvável, temos então duas alternativas: 1) não há diferença entre os tratamentos, e o que ocorreu foi um resultado muito improvável; 2) há diferença entre os tratamentos (isto é, a premissa inicial era falsa) e o que ocorreu foi um evento muito provável. A decisão mais sensata é considerar a segunda alternativa como a verdadeira.
A distinção entre *teste de hipótese e teste de significância* é que no primeiro especificamos, além de uma hipótese nula, uma hipótese alternativa de interesse específico; no segundo, somente a hipótese nula é de interesse.
Para entender o que Valor-p realmente significa, seguimos Diamond e Forrester [1] e apresentamos o exemplo de teste diagnóstico e a relação com teste de hipótese.
Consideremos a Tabela 1 em que um grupo de pacientes é classificado de acordo com a presença ou ausência de uma doença e do resultado positivo ou negativo de um

teste diagnóstico. A partir daí podemos definir as seguintes quantidades:

Tabela 1 - Resultado do teste diagnóstico de acordo com a presença ou ausência de doença

	DOENÇA		
Teste	**Presente**	**Ausente**	**Total**
Positivo (T^+)	Correto Positivo a	Falso Positivo c	T^+ Positivo a + c
Negativo (T^-)	Falso Negativo b	Correto Negativo d	T^- Negativo b + d
	D-Doentes a + b	A- Ausência c + d	N = a + b + c + d

Sensibilidade - S = $\frac{a}{a+b}$ é a probabilidade do teste ser positivo, quando o paciente é sabidamente doente. Donde S = P(T^+ / D), sendo P (maiúsculo) = probalidade

Especificidade - E = $\frac{d}{c+d}$ é a probabilidade do teste ser negativo, quando o paciente tem ausência da doença. Donde E = P (T^- / A).

Prevalência da Doença = $\frac{a+b}{N}$ é a probabilidade de uma pessoa da população estudada ser doente. Denotado por P(D).

Positividade do teste = $\frac{a+c}{N}$ é a probalidade do teste dar positivo. Denotado por P(T^+).

Negatividade do Teste = $\frac{b+d}{N}$ é a probabilidade do teste dar negativo. Denotado por P(T^-).

Valor de Previsão Positivo - VPP = $\frac{a}{a+c}$

é a probabilidade do teste ser correto quanto ele é positivo. Denotado por P(D / T^+).

Valor de Previsão Negativo - VPN = $\frac{d}{b+d}$

é a probabilidade do teste ser correto quando ele é negativo. Denotado por P(A / T^-).

Para que o teste diagnóstico tenha algum valor é necessário que S + E > 1. Se esta soma se aproxima de 2 o teste é ideal. Se S está próximo de 1 significa que a doença pode praticamente ser excluída entre as pessoas com teste negativo, isto é P(T^- / D) ≈ 0 (Já que praticamente não existem resultados falso-negativos do teste).
A situação anterior é análoga à situação de um teste de hipótese em que desejamos testar a hipótese H_o: D, contra a alternativa H_1:A, com base no resultado amostral T^+ (aceita H_o) ou T^- (rejeita H_o).
A Tabela 2 resume as decisões e erros de um teste de hipótese. Vemos, portanto, que ao realizarmos um teste estatístico podemos incorrer em dois tipos de erro: 1) o erro do tipo I, isto é, rejeitar a hipótese nula quando ela é verdadeira; 2) o erro do tipo II, isto é, aceitar a hipótese alternativa quando ela é falsa. A probabilidade do erro do tipo I é denotada por α e denominado nível de significância do teste. Seu valor é comparado em um teste ao Valor-p ou nível de significância observado da estatística de teste. Por outro lado podemos também tomar duas decisões corretas: 1) aceitar a hipótese nula quando verdadeira; 2) rejeitar a hipótese alternativa quando falsa. Suas probabilidades de ocorrência são 1 - α, denominado nível de confiança do teste, e 1 - β, denominado potência do teste, respectivamente.

Tabela 2 - Decisões e erros de teste de hipótese

	REALIDADE	
Decisão do Teste	**H_o: D - Verdadeiro**	**H_1: A - Verdadeiro**
Aceita H_o (T^+)	Decisão Correta Probabilidade: 1- α	Erro Tipo II Probabilidade: β,
Aceita H_1 (T^-) (Rejeita H_o)	Erro Tipo I Probabilidade: α	Decisão Correta Probabilidade: 1- β,

T^+ = teste positivo; T^- = teste negativo

A Tabela 3 mostra as analogias entre teste diagnóstico e teste de hipóteses. Vemos, portanto, a correspondência entre as probabilidades associadas a um teste estatístico de hipótese e as taxas que ocorrem em um teste diagnostico. A tabela apresenta também a correspondência entre as terminologias utilizadas pelo médico (em um teste diagnostico) e pelo estatístico (em um teste de hipóteses).

Tabela 3 - Analogias: Teste Diagnóstico x Testes de Hipóteses

Taxa	Símbolo	Teste Diagnóstico	Teste Hipótese
Correto Positivo	S = P(T^+ / D)	Sensibilidade	1 - α Nível de confiança
Correto Negativo	E = P(T^- / A)	Especificidade	1 - β, Potência
Falso Positivo	P(T^+ / A)	1-Especificidade	β, Erro Tipo II
Falso Negativo	P(T^- / D)	1-Sensibilidade	α Erro Tipo I, Valor-p, Nível de significância observado

Portanto, pela Tabela 3 sabemos agora que o Valor-p nada mais é que o equivalente a 1-sensibilidade, ou seja, P(T^- / D).

É claro que para o médico esta informação não é muito nítida nem de muita utilidade. Na realidade o que o médico deseja (e na maioria das vezes pensa que o Valor-p significa) é a taxa **P(D/T),** ou seja, 1 - **P(A/T),** ou ainda, **1 - VPN.** Esta é a probabilidade do paciente estar doente sabendo que o seu resultado no teste diagnóstico é negativo, ou seja, a probabilidade da hipótese H_o : D ser verdadeira quando a regra de decisão T manda rejeitar H_o. Ao discutirmos mais adiante o Teorema de Bayes mostraremos um outro argumento estatístico que nos permite obter as taxas de interesse clínico.

VALOR-P E TAMANHO DA AMOSTRA

Podemos considerar que todo Valor-p, por exemplo 0,041, indica igual evidência contra a hipótese, independentemente da hipótese e do contexto dos dados?
Considere um experimento em que todos os pacientes recebem ambos os tratamentos A e B e são solicitados a explicar suas preferências. Os quatros conjuntos de dados são:

Dados I: 15 pacientes preferem A, 5 preferem B (**r** = 15/20 = 0,75);
Dados II: 114 pacientes preferem A, 86 preferem B (**r** = 0,57);
Dados III: 1046 pacientes preferem A, 954 preferem B (**r** = 0,523);
Dados IV: 1001445 pacientes preferem A, 998555 preferem B (r = 0,5007).

Estes conjuntos de dados produzem um Valor-p de 0,041, obtido de um teste de proporção para a hipótese nula r = 1/2 (que indica igual preferência por A e B) (r é o valor verdadeiro, **r** é o valor calculado dos dados), informando assim que o resultado obtido em cada conjunto (indicando maior preferência por A) é significativamente diferente da hipótese nula (que indica igual preferência por A e B).
Porém, estes dados não são igualmente convincentes sobre a maior preferência por A. Os dados I provavelmente serão rejeitados por ser a amostra muito pequena, embora indiquem uma preferência para A de 75%. Os dados IV indicam de forma quase conclusiva que as preferências são iguais, isto é, r = 1/2 (preferência por A de 50,07%). Portanto, o Valor-p de 0,041 não pode ser tomado como evidência independente do contexto e do tamanho da amostra.
Freeman[3] mostra outro exemplo de um teste de média de uma distribuição normal, que um valor p de 0,01 em uma amostra de tamanho 100 apresenta menor evidência contra a hipótese do que um valor p de 0,05 em uma amostra do tamanho 10.
Por isso, o médico deve ter sempre em mente, ao analisar resultados de experimentos, que em medicina é preciso diferenciar significância estatística de significância biológica.

INTERVALOS DE CONFIANÇA: OUTRA FONTE DE MÁ INTERPRETAÇÃO

Não sabendo o que é Valor-p (ou nível de significância observado) também fica difícil interpretar corretamente intervalos de confiança. No sentido estatístico clássico o intervalo de confiança significa que, hipoteticamente, se uma série de estudos idênticos fosse realizado repetidamente com diferentes amostras da mesma população, e um intervalo de 95% de confiança para as diferenças entre as médias fosse obtido em cada estudo, ao se aumentar o número de estudos hipotéticos se observaria que 95% desses intervalos incluiriam a verdadeira diferença entre as médias das populações.
Podemos agora relacionar intervalos de confiança com testes de hipótese. Um intervalo de confiança de nível 1 - p é um intervalo que contém os valores que seriam aceitos em um teste estatístico com probabilidade de erro do tipo I igual a p. Alternativamente, se quisermos testar uma hipótese especificada (por exemplo, diferença nula) basta calcular o intervalo de confiança 100 (1 -p)% e aceitar a hipótese (diferença = 0) se a mesma estiver contida no intervalo, e rejeitar caso contrário.
Para exemplificar, consideremos um estudo que comparou as pressões arteriais de 100 homens diabéticos com as de 100 homens não diabéticos, obtendo-se uma diferença de 6,0mmHg entre as médias das pressões sistólicas dos dois grupos, sendo o erro padrão da diferença entre as médias de 2,5 mmHg. O intervalo de 95% de confiança para a diferença das médias das duas populações é obtido pela fórmula $x \pm 1{,}97\ \sigma x$, onde x = média e σx = erro padrão de x, isto é, $6{,}0 \pm 1{,}97 \times 2{,}5$ mmHg, isto é, o intervalo de 1,1 a 10,9. Este resultado não significa que a diferença entre as médias das populações de diabéticos e não diabéticos está entre 1,1 mmHg e 10,9 mmHg com 95% de chance, como em geral é interpretado erroneamente. Na realidade, significa que se realizarmos vários estudos sobre diabetes, e em cada estudo calcularmos um intervalo de confiança usando a fórmula $x \pm 1{,}97\ \sigma x$, em 95% desses estudos o intervalo calculado conteria a verdadeira diferença entre as médias. Portanto, neste particular estudo temos apenas uma confiança de 95% de que o intervalo contém a diferença, já que é apenas um experimento, e é razoável supor que o intervalo contém a diferença entre as médias.

UMA SOLUÇÃO: INFERÊNCIA BAYESIANA

Os índices resultantes da Tabela 3 (sensibilidade, especificidade, Valor-p) tem limitações sérias. Não ajudam, por exemplo, ao clínico quando este recebe um paciente com resultado positivo do teste diagnóstico e precisa decidir se o paciente está ou não doente, ou até mesmo se deve solicitar mais exames complementares para ter mais certeza na decisão. Ou seja, o que interessa mais ao médico conhecer para os testes diagnósticos são as seguintes taxas: Valor de Previsão Positivo (VPP), Valor de Previsão Negativo (VPN), 1- VPP e 1- VPN.
Estas probabilidades podem ser obtidas através do

Teorema de Bayes que na terminologia médica é escrito:

$$VPP = \frac{\text{sensibilidade x prevalência}}{\text{sensibilidade x prevalência + (1 - especificidade) x (1 - prevalência)}}$$

$$= \frac{\text{sensibilidade x prevalência}}{\text{positividade}}$$

$$VPN = \frac{\text{especificidade x (1 - prevalência)}}{\text{(1 - sensibilidade) x prevalência + especificidade x (1 - prevalência)}}$$

$$= \frac{\text{especificidade x (1 - prevalência)}}{\text{negatividade}}$$

Assim, VPP = P(D / T^+) é a *probabilidade a posteriori* de doença após um diagnóstico positivo de um teste e VPN = P(A / T^-) é a *probabilidade a posteriori* de ausência de doença após um diagnóstico negativo.

Observe que as taxas clinicamente úteis VPP e VPN, (que aferem a acurácia preditiva do teste) tem a desvantagem de depender fortemente da prevalência da doença ou probabilidade a priori, o que muitas vezes não é conhecida. É importante lembrar que uma prevalência obtida de um estudo não pode ser tomada universalmente. Por outro lado a sensibilidade e a especificidade, e portanto também 1-sensibilidade = Valor-p, embora não acessando a acurácia de um teste de forma clinicamente útil, tem a vantagem de não serem afetadas pela prevalência.

Para ilustrar, consideremos o seguinte exemplo de Soares[5]. Uma das mais difundidas tecnologias para detectar a presença do vírus HIV é o teste ELISA, comercializado por vários laboratórios. Um deles reportou em seus testes preliminares uma sensibilidade de 95% e uma especificidade de 99%. A Tabela 4 apresenta as probabilidades a posteriori de falso-positivo e falso-negativo para diferentes prioris (prevalência).

Tabela 4 - Teste ELISA

Piori	Posteriori	
Prevalência	P(A/T^+) **Falso Positivo %**	P(D/T^-) **Falso Negativo**
1/milhão	99.95	5.01×10^{-8}
1/100.00	99.53	5.01×10^{-7}
1/10.000	95.46	5.01×10^{-6}
1/1000	67.79	1×10^{-4}
1/500	51.23	1×10^{-4}
1/200	29.53	2.5×10^{-4}
1/100	17.25	5.1×10^{-4}
1/50	9.37	10.2×10^{-4}

Considerando que a prevalência da AIDS é pequena, os resultados mostram que no uso em larga escala do teste (mais de 500 indivíduos) grande parte dos indivíduos com resultados positivos serão falsos positivos e por outro lado poucos doentes não serão detectados, isto é, poucos falsos negativos. Portanto, o clínico não se preocupará com o paciente se o teste der negativo, e indicará outro teste alternativo se o teste for positivo.

Estes dados são uma aplicação particular da Inferência Bayesiana, que fornece ao pesquisador a resposta que ele deseja. Suponhamos que estamos interessados na diferença de médias das duas populações de diabéticos discutidos na seção prévia. A inferência Bayesiana transforma a crença a priori que o médico tem sobre a diferença Θ, expressa pela probabilidade a priori $P(\Theta)$, usando a verossimilhança $L(t/\Theta)$ de um experimento que observa t relacionado a Θ. Usando o teorema de Bayes, obtém-se a probabilidade a posteriori $P(\Theta/t)$:

$$P(\Theta / t) \propto L(t / \Theta)\, P(\Theta) \quad (\propto = \text{proporcional a})$$

No exemplo da diabetes visto anteriormente, o clínico, com sua experiência, poderia especificar que acredita que a priori a diferença deve estar entre - 5 e 20 mmHg no máximo. Isto pode ser expresso por uma probabilidade uniforme, pois ele não tem preferência por qualquer valor neste intervalo, isto é

$$P(\Theta) = \frac{1}{25} \qquad -5 < \Theta < 20$$

Realizado o experimento verificou-se que a verossimilhança para a diferença das médias $L(t/\Theta)$ era da forma de uma distribuição normal com média 6 mmHg e erro padrão de 2,5 mmHg. Neste caso pode-se demonstrar que a diferença das média Θ tem distribuição a posteriori $P(\Theta/t)$ também normal com mesma média e erro padrão, isto é que a diferença das médias tem distribuição a posteriori normal. Obtida a distribuição a posteriori podemos utilizar suas características como a média, a moda e a mediana como estimadores Bayesianos, bem como construir intervalos Bayesianos e testes de hipóteses Bayesianos.

A diferença agora é que estes resultados informam ao médico exatamente a resposta que ele deseja:

1) no exemplo da diabetes, o intervalo Bayesiano de 95% de probabilidade é próximo de 1,1 a 10,9. Aqui, a interpretação é que a diferença entre médias está neste intervalo com probabilidade especificada de 95%. Não se recorre à realizações hipotéticas do experimento que não foram feitas e talvez nunca serão. Muitas vezes, intervalos Bayesianos e clássicos coincidem numericamente, mas a interpretação é diferente (compare com o intervalo do exemplo da diabetes);

2) testes de hipóteses Bayesianos para a hipótese H_0: r = 1/2 (igual preferência) no exemplo visto anteriormente, que concluíram que r = 1/2 com probabilidades:
P(r = 1/2, Dados I: n = 20 e **r** = 0,75) = 0,382

P(r = 1/2, Dados II: n = 200 e **r** = 0,57) = 0,637
P(r = 1/2, Dados III: n = 2000 e **r** = 0,523) = 0,846
P(r = 1/2, Dados IV: n = 2000000 e **r** = 0,5007) = 0,994

Portanto, H_o: r = 1/2 (igual preferência) é bem plausível segundo os dados II, III e IV.

Estas probabilidades são obtidas da expressão

$$P(H_0 / \text{Dados}) = \left\{1 + \left[(1+n)^{1/2} \exp\left\{-\frac{n}{n+1} \times \frac{Z_\alpha}{2}\right\}^2\right]^{-1}\right\}^{-1}$$ (onde Z_α é o valor da tabela da distribuição normal correspondente ao valor α, no caso $\alpha = 0{,}041$, $Z\alpha = 1{,}97$) e sua justificativa pode ser vista em Berge e Salke[6]. Observe que aqui a interpretação que o médico em geral tem de testes de hipótese é correta, isto é, a probabilidade da hipótese testada ser verdadeira é fornecida, ao contrário de obtermos uma "confiança" na veracidade da hipótese.

3) podemos também relacionar alguns testes de hipóteses Bayesianos com intervalos Bayesianos. No exemplo da diabetes, como H_o: $\Theta = 0$ (nenhuma diferença) está fora do intervalo Bayesiano de 95% de probabilidade (isto é, 1,1 a 10,9), rejeitaríamos esta hipótese pois ela está numa região com probabilidade menor que 0,05.
A inferência Bayesiana é também alvo de críticas e causa discussões acaloradas entre seus adeptos e os de outras escolas, principalmente os da estatística clássica. Um dos pontos de maior controvérsia é a especificação da priori. Entretanto o assunto está fora do alcance deste artigo. Para maiores detalhes das críticas de parte a parte existe uma extensa bibliografia na literatura estatística.

CONCLUSÕES E SUGESTÕES

1) em publicações científicas apresente sempre que possível não só o Valor-p mas também o intervalo de confiança;

2) o Valor-p e os intervalos de confiança devem ser considerados como mais uma estatística descritiva, assim como as médias, as variâncias, a sensibilidade, a especificidade; as tabelas, etc., a serem fornecidas no texto científico;

3) em experimentos com grandes amostras é aconselhável usar o nível de significância (ou Valor-p) de 0,01 (ou menor) ao invés do usual 0,05;

4) a comunidade médica deve utilizar mais o enfoque Bayesiano, que forneça respostas que o médico procura;
5) o problema da especificação da priori não deve servir de obstáculo ao uso da inferência Bayesiana. Primeiro, o médico sempre tem algumas crenças a priori. Segundo, vale a pena tentar várias prioris (como no caso de prevalência de HIV) e verificar se as conclusões se alteram.

AGRADECIMENTOS

O autor é grato ao Professor Dr. Hélio Migon e ao Dr. Roberto Bassan por sugestões e correções que resultaram em melhora substancial do texto.

REFERÊNCIAS BIBLIOGRÁFICAS

1. Diamond GA, Forrester JS. Clinical trials and statistical verdicts: probable grounds for appeal. Ann Intern Med 1983; 98 : 3 8 5 -3 94

2. Salsburg DS. The religion of statistics as practiced in medical journals. Am Statist 1985; 39 : 220-223.

3. Freeman PR. The role of p-values in analysing trial results (with discussion). Stat Med 1993; 12: 1443-1458

4. Gardner MJ, Altman DG. Confidence intervals rather than p values: estimation rather than hypothesis testing. Br Med J 1986; 292 : 746-750

5. Soares JF. O teste de detecção do vírus da AIDS. Boletim da Associação Brasileira de Estatística 1987; 8: 10-16

6. Berger JO, Selke T. Testing a point null hypothesis: the irreconciability of p values and evidence (with discussion). J Am Stat Assoc 1987; 82 : 112-139.