Artigo de revisão *DOI: 10.5902/2179460X13195*

Ciência e Natura, Santa Maria, v. 36 Ed. Especial, 2014, p. 14–22
Revista do Centro de Ciências Naturais e Exatas - UFSM
ISSN impressa: 0100-8307 ISSN on-line: 2179-460X

CIÊNCIA e NATURA 35 anos

Uma medida de evidência alternativa para testar hipóteses gerais

S-value: an alternative measure of evidence for testing general null hypotheses

Alexandre G. Patriota*[1]

[1]Instituto de Matemática e Estatística, Universidade de São Paulo

Resumo

Ferramentas estatísticas são comumente aplicadas para testar hipóteses de interesse no meio científico. Usualmente, utiliza-se o p-valor como um termômetro de discrepância entre hipótese proposta e os dados observados: quanto menor o p-valor, maior a discrepância da hipótese de interesse para explicar o comportamento dos dados observados. Este trabalho discute algumas "inconsistências" do p-valor quando empregado para testar hipóteses aninhadas. É apresentada uma nova medida de evidência que cumpre certas relações lógicas que o p-valor não satisfaz. Novas regras de aceitação e rejeição de hipóteses gerais são analisadas sob a nova metodologia. Este artigo intenta apresentar o problema de testar hipóteses científicas de forma simples e acessível a alunos de (pós)graduação em estatística e áreas correlatas.

Palavras-chave: *Coerência, Estatística Clássica, Medida de evidência, P-valor, S-valor, Teste de hipótese.*

Abstract

In the classical paradigm, the famous p-value is employed for testing scientific hypotheses. It is used as a discrepancy thermometer between the proposed hypothesis and the observed data*: the smaller is the p-value, the larger is the discrepancy of the hypothesis of interest to explain the* data *behavior. In this paper, we discuss some known "inconsistencies" of this measure when employed for testing nested hypotheses. It is also reviewed a recent measure of evidence, called s-value, that satisfies certain logical relations not met by p-values. New rules of acceptation and rejection are investigated under this new approach. This paper intents to present the subject as simple and accessible as possible to (under)graduate students of statistics and related fields.*

Keywords: *Coherence; Classical statistics; Evidence measure; Hypothesis testing; P-value; S-value.*

*patriota@ime.usp.br.
Recebido: 21/01/2014 Revisado: 14/05/2014

1 Introdução

Testar se uma hipótese de interesse pode ser estatisticamente rejeitada é uma das práticas mais importantes e corriqueiras no meio científico. Em geral, os pesquisadores envolvidos no estudo elaboram uma hipótese de interesse e desejam testá-la utilizando ferramentas e procedimentos estatísticos. Tal hipótese é formada por uma proposição científica traduzida em termos matemáticos. A proposição científica é aquela que pode, de alguma forma, ser refutada, i.e., deve ser possível engendrar um argumento que evidencie sua falsidade. Caso seja impossível elaborar argumentos racionais (ou empíricos) para falsear uma determinada proposição, então tal proposição não é científica. Por exemplo, a proposição "Fadas curam câncer" não é científica, pois não é possível engendrar um argumento que permita verificar sua falsidade (Fadas não são acessíveis em nosso mundo observado). Por outro lado, a proposição "O medicamento A é, em média, mais eficaz do que o medicamento B" é científica, pois pode-se pensar em um experimento para refutá-la.

Usualmente, definem-se duas hipóteses, a saber: H_0, contendo a proposição q, e H_1, contendo a negação da proposição q. Por exemplo, suponha que o interesse consiste em comparar o poder de cura de dois medicamentos A e B, dessa forma pode-se formar as seguintes hipóteses:

> H_0 : "os medicamentos A e B têm o mesmo efeito médio"
> H_1 : "os medicamentos A e B **não** têm o mesmo efeito médio".

Diz-se que H_0 é a hipótese nula e H_1 a hipótese alternativa. Há duas escolas que defendem procedimentos diferentes para testar tais hipóteses, a saber: procedimento de Neyman-Pearson (Neyman and Pearson, 1933) e procedimento Fisheriano (Fisher, 1935). O primeiro permite aceitar a hipótese nula ou aceitar a hipótese alternativa (rejeitar a hipótese nula seria equivalente a aceitar a hipótese alternativa e vice-versa), ao passo que o segundo permite apenas rejeitar a hipótese nula. No segundo procedimento, dois tipos de conclusões para o teste podem ser enunciadas: "há evidência suficiente para rejeitar H_0" ou "**não** há evidência suficiente para rejeitar H_0", sendo que a última conclusão não implica na aceitação de H_0. Em ambos os casos as decisões dependerão do resultado experimental e do tipo de hipóteses envolvidas. Por exemplo, suponha que

> H_0 : "todos os cisnes são brancos"

e que foi observado de fato todos os cisnes brancos na amostra coletada. Note que não existe evidência para rejeitar H_0, porém essa hipótese não deve ser aceita, pois não se pode observar a totalidade dos cisnes em todos os tempos. Por outro lado, a hipótese nula

> H_0' : "todos os cisnes que estão, neste momento, no zoológico de São Paulo são brancos"

pode ser aceita, bastaria verificar se todos os cisnes do zoológico de São Paulo são de fato brancos. A hipótese H_0 é bem mais restrita que H_0', pois na segunda apenas restringiu-se a "cor branca" para os cisnes que estão no zoológico de São Paulo. Em termos gerais tem-se o seguinte: se o universo de possibilidades é aberto ou infinito não-enumerável e a hipótese nula contém uma proposição significativamente mais restritiva que a hipótese alternativa, então será mais apropriado apenas coletar evidências para rejeitá-la, as evidências nunca seriam significativas para aceitar a hipótese nula (para mais detalhes veja Popper, 1989). Este procedimento se adequa ao método Fisheriano. No entanto, se o universo de possibilidades é fechado e a hipótese nula contém uma proposição cuja restritividade é similar ao da proposição definida na hipótese alternativa, então o procedimento de Neyman-Pearson poderá ser mais útil.

O procedimento Fisheriano é o mais amplamente utilizado e a quantidade estatística relacionada com este procedimento é p-valor (ou valor p). Esta quantidade é utilizada como termômetro para rejeitar uma hipótese: quanto menor for o seu valor mais evidência de que H_0 é falsa. Usualmente rejeita-se a hipótese H_0 se o p-valor for menor do que um ponto de corte, α (e.g., $\alpha = 0.05$), caso seja maior diz-se que não há evidências para rejeitar H_0. Um dos objetivos deste trabalho é divulgar uma terceira alternativa em que é possível tomar três tipos de decisões, a saber: aceitar H_0 (equivalente a rejeitar H_1), aceitar H_1 (equivalente a rejeitar H_0) e não rejeitar nenhuma das duas hipóteses. Observe que quando os dados trazem informação ambígua sobre as hipóteses não se deve nem aceitar, nem rejeitar H_0, neste caso o mais adequado seria coletar mais dados ou outras informações.

2 Definição do p-valor

Há várias metodologias para o cálculo do p-valor. Essencialmente, fixa-se o experimento que fornecerá os dados experimentais, a seguir propõe-se uma estatística T, função dos dados experimentais, tal que: quanto maior o valor observado de T, maior é a discrepância da hipótese nula para explicar o comportamento dos dados observados. Em escrita corrida:

> supondo que H_0 é de fato compatível com os dados observados, então o p-valor é a probabilidade de ser observado em outro experimento uma estatística T, no mínimo,

> tão extrema quanto a que foi observada no experimento atual.

Para representar matematicamente a sentença acima, deve-se notar que o evento $\{T > t\}$ pode ser interpretado como "ocorrer em outro experimento uma estatística T pelo menos tão extrema quanto a que foi observada". Portanto, um valor da estatística observada demasiado alto reflete em uma probabilidade pequena de que algo ainda mais extremo ocorra em um próximo experimento, caso a hipótese seja compatível com os dados. Assim, faz pleno sentido rejeitar a hipótese nula quando o p-valor for suficientemente pequeno. De maneira informal, o p-valor é matematicamente definido por

$$p = P(T > t;\ \text{sob } H_0),$$

em que t é o valor observado da estatística T e "sob H_0" significa que a medida de probabilidade utilizada é aquela que foi especificada na hipótese nula. Vale ressaltar que essa é uma definição informal que mantém certa simplicidade, porém, como todas as definições informais, sua interpretação pode gerar muitas controvérsias. Inicialmente deve-se formalizar a notação e reescrever as hipóteses em termos matemáticos.

2.1 Definição Formal

Para definir formalmente um p-valor, deve-se inicialmente escrever o modelo estatístico utilizando a trinca

$$(\mathcal{X}, \mathcal{F}, \mathcal{P}),$$

em que $\mathcal{X} \subseteq \mathbb{R}^n$ é o espaço amostral (o conjunto que contém todas as possíveis observações do experimento), $\mathcal{F}$ a σ-álgebra de Borel no $\mathbb{R}^n$ e $\mathcal{P}$ uma família de medidas de probabilidade possíveis para explicar o comportamento dos dados provenientes do experimento em questão. Quando a família de probabilidades puder ser escrita $\mathcal{P} = \{P_\theta,\ \theta \in \Theta \subseteq \mathbb{R}^k\}$ com $k < \infty$, diz-se que o modelo estatístico é paramétrico. O problema de testar hipóteses pode ser escrito em termos de subconjuntos da família $\mathcal{P}$, assim, na segunda etapa, deve-se especificar a família de medidas de probabilidade que estão descritas na hipótese nula, ou seja, a família restrita $\mathcal{P}_{H_0} \subset \mathcal{P}$ contemplando todas as medidas de probabilidade que representam a hipótese nula H_0. Por exemplo, suponha que o modelo estatístico contém as medidas $\mathcal{P} = \{P_0, P_1, P_2, P_3\}$ e o interesse consiste em testar se a medida que gera os dados é P_0, então escreve-se $H_0 : P \equiv P_0$. Neste caso, o p-valor é então definido por

$$p = P_0(T > t),$$

ou seja, sob H_0 tem-se que a medida que gera os dados é P_0 e portanto deve-se utilizá-la no cálculo do p-valor. Suponha que o interesse está em verificar se a medida que gera os dados é P_0 ou P_1, então escreve-se $H_0 : P \in \mathcal{P}_{H_0}$, sendo $\mathcal{P}_{H_0} = \{P_0, P_1\}$, neste caso há duas medidas para definir o p-valor, deve-se escolher então a que gera a maior probabilidade:

$$p = \max\{P_0(T > t), P_1(T > t)\},$$

dessa forma o p-valor terá um valor conservador, pois se em pelo menos uma medida o evento $\{T > t\}$ tem probabilidade alta de ocorrer não se deveria rejeitar a hipótese nula. Por outro lado, se para ambas medidas o evento $\{T > t\}$ tem probabilidade baixa de ocorrer, então a hipótese nula pode ser considerada implausível.

Usualmente, quando a família de medidas é paramétrica, i.e., existe um espaço finito dimensional Θ tal que a família de probabilidade seja identificada por $\mathcal{F} = \{P_\theta;\ \theta \in \Theta\}$, então pode-se representar a hipótese nula utilizando a seguinte notação paramétrica $H_0 : \theta \in \Theta_0$, sendo $\mathcal{P}_{H_0} = \{P_\theta;\ \theta \in \Theta_0\}$. Neste caso o p-valor é então definido seguindo a mesma regra anterior

$$p = \sup_{\theta \in \Theta_0} P_\theta(T > t).$$

O conjunto Θ_0 contém todos os valores $\theta \in \Theta$ que satisfazem a restrição imposta pela hipótese nula, e.g., se $\Theta = \{(\theta_1, \theta_2);\ \theta_1, \theta_2 \in \mathbb{R}\}$ e a hipótese nula contém a proposição "θ_1 e θ_2 são iguais", então $\Theta_0 = \{(\theta_1, \theta_2) \in \Theta;\ \theta_1 = \theta_2\}$. Neste exemplo, a hipótese nula pode ser escrita de maneira equivalente, a saber: $H_0 : \theta_1 = \theta_2$.

Um detalhe importante que deve ser notado é que, para se ter um teste com boas propriedades, a estatística T deve depender da hipótese nula (e da hipótese alternativa, pois a disjunção das duas hipóteses precisa formar o modelo irrestrito). Ou seja, se há duas hipóteses nulas diferentes, tem-se então duas estatísticas do teste diferentes, portanto formalmente dever-se-ia escrever T_{H_0} ou T_{Θ_0} em vez de T; isso evitaria interpretações equivocadas, muito comuns em toda a literatura estatística. Para mais detalhes e discussões sobre a definição precisa de p-valores, veja Patriota (2013).

Note que não existe conceito algum de distribuição de probabilidade condicional envolvido na definição do p-valor, tem-se apenas uma família de probabilidades restrita à hipótese nula. Há alguma confusão nesse ponto específico, mesmo entre estatísticos de formação: alguns defendem que o p-valor é uma probabilidade condicional (dado H_0), porém basta notar que nenhum procedimento probabilístico é utilizado para escolher a medida P_0 (ou a família da hipótese nula $\mathcal{P}_{H_0}$). Dizer que o p-valor é uma probabilidade condicional é equivalente a dizer que existe um espaço de probabilidade subjacente para a família de medidas de probabilidade, porém este espaço probabilístico não é necessário na definição do p-valor. A imposição de tal espaço probabilístico causa problemas teóricos e interpretativos,

pois neste caso a família $\mathcal{P}$ estaria sujeita às restrições que a teoria de probabilidades impõe sobre os conjuntos (e.g., problemas de mensurabilidade, medida zero para subconjuntos de dimensão inferior a dimensão de $\mathcal{P}$ e assim por diante).

3 Inconsistências do p-valor

O p-valor nos dá uma medida de evidência contra uma hipótese específica. Quanto menor o p-valor, mais evidências de que a hipótese nula não se adéqua aos dados observados. Como comentado anteriormente, a estatística T, fundamental no cálculo do p-valor, depende fortemente da hipótese nula (e da hipótese alternativa). Portanto um cuidado especial deve-se ter ao comparar p-valores de hipóteses diferentes, pois duas hipóteses diferentes induzem duas medidas diferentes no cálculo do p-valor. Esta característica gera conflitos de interpretação com diferentes graus de gravidade, para ilustrar dois dos conflitos considere duas populações representadas pelas variáveis[1] X e Y, com esperanças $E(X) = \mu_1$ e $E(Y) = \mu_2$. Considere as seguintes hipóteses nulas

$$H_{01} : \mu_1 = \mu_2,$$
$$H_{02} : \mu_1 = \mu_2 = b,$$
$$H_{03} : \mu_1 = b,$$
$$H_{04} : \mu_2 = b$$

sendo b uma constante real qualquer. Considere as seguintes conclusões:

1. Há evidência para rejeitar H_{02}, mas não há evidência para rejeitar as hipóteses H_{03} e H_{04} separadamente.

2. Há evidência para rejeitar H_{01}, mas não há evidência para rejeitar H_{02}.

Para o primeiro caso, tem-se teoricamente que $H_{02} \equiv H_{03} \wedge H_{04}$, ou seja, a hipótese H_{02} é equivalente às hipóteses H_{03} e H_{04} simultaneamente. Neste caso, espera-se que se H_{02} é rejeitada, então pelo menos uma das hipóteses H_{03} ou H_{04} seja rejeitada. Esta situação é particularmente simples de explicar: no teste da hipótese H_{02} considera-se em geral os dois conjuntos de dados, enquanto que para cada hipótese isolada considera-se usualmente apenas um dos conjuntos de dados e exclui-se o outro. Dessa forma, os testes marginais (H_{03} e H_{04}) não consideram uma possível dependência entre X e Y; vale ressaltar que quanto maior for a dependência maiores serão as diferenças de conclusões. A interpretação adequada: "há indícios contra a hipótese de que as médias são iguais a b", pois os testes marginais não consideram a informação completa. Desmembrar a hipótese conjunta em pedaços marginais pode induzir uma perda de informação pertinente, esse tipo de problema ocorre amiúde na prática. Portanto, testes conjuntos são em geral preferíveis a testes marginais.

Para o segundo caso, tem-se teoricamente que $H_{02} \to H_{01}$, ou seja, a hipótese H_{02} implica H_{01}. Neste caso, rejeitar H_{01} implica em rejeitar H_{02}. O problema agora é diferente, ambas hipóteses são conjuntas e consideram toda a informação dos dados. Nesta seção, dar-se-á um exemplo de conjuntos de dados em que este segundo caso ocorre. Considere $X \sim N(\mu_1, 1)$ e $Y \sim N(\mu_2, 1)$ independentes para simplificar o problema. Assuma que uma amostra de tamanho $n = 100$ foi retirada e as médias amostrais são $\bar{x} = 2{,}14$ e $\bar{y} = 1{,}84$. A estatística utilizada em todos os testes será

$$T = -2\log\left(\frac{L_{\widehat{\theta}_0}(X)}{L_{\widehat{\theta}}(X)}\right) = 2(\ell_{\widehat{\theta}}(X) - \ell_{\widehat{\theta}_0}(X)),$$

em que $L_\theta(X)$ é a função de verossimilhança, $\ell_\theta(X) = \log(L_\theta(X))$, $\widehat{\theta} = \arg\max_{\theta\in\Theta} \ell_\theta(X)$ e $\widehat{\theta}_0 = \arg\max_{\theta\in\Theta_0} \ell_\theta(X)$ são os estimadores de máxima verossimilhanças irrestrito e restrito a Θ_0, respectivamente. A estatística T será chamada simplesmente de estatística da razão de verossimilhanças (apesar de na realidade ser -2 vezes o logaritmo da razão de verossimilhanças). Para qualquer modelo que satisfaça condições de regularidade mínimas, tem-se que T converge para uma qui-quadrado com $r = \dim(\Theta) - \dim(\Theta_0)$ graus de liberdade (isso é válido sempre que o logaritmo da função de verossimilhança é côncavo, Θ_0 for um conjunto sem singularidades e a dimensão de Θ_0 for menor que a de Θ). Aqui, pela normalidade dos dados, a distribuição dessa estatística é exata.

Considere $H_{01} : \mu_1 = \mu_2$, levando-se em conta a normalidade dos dados pode-se utilizar a estatística da razão de verossimilhanças, que possui propriedades assintóticas ótimas (Bahadur and Raghavachari, 1972; Mudholkar and Chaubey, 2009). Assim, pode-se mostrar que a estatística da razão de verossimilhanças para testar essa hipótese específica é

$$T_1 = \frac{n}{2}(\bar{X} - \bar{Y})^2.$$

Pode-se mostrar ainda que, restrita à hipótese H_{01}, T_1 tem distribuição χ^2 com 1 grau de liberdade (denote por χ^2_1). Portanto, se o tamanho amostral for $n = 100$, $\bar{x} = 2{,}14$ e $\bar{y} = 1{,}84$, tem-se $t_1 = 4{,}50$ e o p-valor $p_1 = P(\chi^2_1 > t_1) = 0{,}03$. Ou seja, Há evidência para rejeitar que $\mu_1 = \mu_2$, a 5% de significância estatística.

> Note que foi observado evidência para rejeitar que $\mu_1 = \mu_2$ (p-valor = 0,03), portanto, para os mesmos dados observados, deve-se teoricamente rejeitar, em especial, que $\mu_1 = \mu_2 = 2$.

[1] Variáveis aleatórias estão em maiúsculo e os valores observados das variáveis aleatórias em minúsculo.

Considere agora a hipótese nula $H_{02} : \mu_1 = \mu_2 = 2$. Sob as mesmas suposições anteriores (normalidade dos dados e variância unitária) tem-se que o logaritmo da estatística da razão de verossimilhanças é

$$T_2 = n[(\bar{X} - 2)^2 + (\bar{Y} - 2)^2].$$

Pode-se mostrar ainda que, restrita à hipótese H_{02}, T_2 tem distribuição χ^2 com 2 graus de liberdade (denote por χ^2_2). Portanto, considerando os mesmos valores amostrais ($n = 100$, $\bar{x} = 2{,}14$ e $\bar{y} = 1{,}84$), tem-se $t_2 = 4{,}52$ e o p-valor $p_2 = P(\chi^2_2 > t_2) = 0{,}10$. Ou seja, não encontra-se evidência para rejeitar $\mu_1 = \mu_2 = 2$, a 5% de significância estatística.

> Note que **não** foi observado evidência para rejeitar que $\mu_1 = \mu_2 = 2$ (p-valor = 0,10).

Em resumo, deve-se ter cuidado ao utilizar p-valores em hipóteses aninhadas. Deve-se ter em mente que: *se houver evidências para rejeitar uma hipótese H_0, não se pode afirmar que há evidências para rejeitar qualquer outra hipótese H_0' que implica H_0.* Um algoritmo que gera conjuntos de dados com este problema é descrito em `http://eranraviv.com/blog/on-p-value/`. Mais detalhes podem ser estudados em Schervish (1996), Izbicki *et al.* (2012) e Patriota (2013). Na próxima seção é discutida uma medida de evidência em que tais problemas não ocorrem.

4 Nova proposta de medida de evidência

No artigo "*A classical measure of evidence for general null hypotheses*" é proposto uma nova medida de evidência que está livre de contradições lógicas. Considere a seguinte hipótese geral $H_0 : \theta \in \Theta_0$, em que $\Theta_0 \subseteq \Theta$. A ideia central da proposta é criar a menor região de confiança possível para o vetor θ tal que inclua pelo menos um ponto da borda de Θ_0. A Figura 1 ilustra a ideia para $\Theta = \mathbb{R}^2$, em que $\hat{\theta}$ é a estimativa de máxima verossimilhança (EMV), a região elipsoidal marcada com o símbolo "$\times$" em seu interior é a menor região de confiança que inclui pelo menos um ponto da borda de Θ_0 (região disforme marcada com pontos em seu interior). Para cada região de confiança tem-se associado um coeficiente de confiança γ e um nível de significância $\alpha = 1 - \gamma$, a definição da medida de evidência é simplesmente o nível de significância associado a menor região de confiança que inclui pelo menos um ponto do espaço Θ_0. Este nível de significância será chamado de s-valor e sua definição formal segue na sequência.

4.1 Definição formal do s-valor

Para evitar as inconsistências relatadas do p-valor, faz-se necessário definir uma medida de evidência que utilize a mesma métrica para qualquer subconjunto de Θ. Isso será feito utilizando regiões de confiança baseadas na razão de verossimilhanças, pois Sprott (2000) apresenta exemplos de conjuntos de confiança fora do espaço paramétrico quando tais regiões não são baseadas na razão de verossimilhanças. Considerando que X representa a amostra e que $\widehat{\theta} \in \Theta$, a região de confiança com nível de significância α é definida por

$$\Lambda_\alpha(X) = \{\theta \in \Theta : 2(\ell_{\widehat{\theta}}(X) - \ell_\theta(X)) \leq F_\alpha\},$$

em que F_α é tal que $F(F_\alpha) = 1 - \alpha$ com F sendo uma função de distribuição acumulada estritamente contínua da variável aleatória $T_\theta = 2(\ell_{\widehat{\theta}}(X) - \ell_\theta(X))$ que não depende de θ. Nem sempre a função de distribuição exata da variável T_θ será contínua ou independente de θ, nestes casos pode-se utilizar a função de distribuição acumulada assintótica da variável T_θ. Vale lembrar que a distribuição acumulada exata de T_θ é dada por

$$F_\theta(t) = P_\theta(T_\theta \leq t).$$

No cálculo da região de confiança é assumido que $F_\theta \equiv F$ não depende de θ. Sempre que as condições de regularidade estiverem satisfeitas para a estatística da razão de verossimilhanças, a função F pode ser definida como a distribuição acumulada da qui-quadrado com k graus de liberdade, sendo k a dimensão de Θ. Para evitar excesso de notação será considerado apenas $\Lambda_\alpha \equiv \Lambda_\alpha(X)$. Note que a região de confiança Λ_α nem sempre será uma elipse como mostrado na Figura 1, isso ocorrerá para a distribuição normal com variância conhecida.

O s-valor, medida de evidência para testar a hipótese geral $H_0 : \theta \in \Theta_0$, é então definido matematicamente por

$$s(X,\Theta_0) = \max\{0, \sup\{\alpha \in (0,1) : \ \Lambda_\alpha \cap \Theta_0 \neq \varnothing\}\}. \quad (1)$$

Observe que o s-valor depende tanto da amostra X quanto do subconjunto Θ_0 da hipótese nula, por simplicidade de notação será considerado apenas $s(\Theta_0) \equiv s(X, \Theta_0)$, quando o conjunto Θ_0 não for relevante utilizar-se-á apenas $s \equiv s(\Theta_0)$. Note que valores altos de s indicam que existe pelo menos um ponto em Θ_0 que está perto do EMV, enquanto que valores pequenos de s indicam que todos os pontos em Θ_0 estão longe do EMV. Para uma forma mais simples de calcular o s-valor, veja a fórmula (2) da próxima seção.

4.1.1 Propriedades do s-valor

Algumas propriedades importantes do s-valor são listadas abaixo:

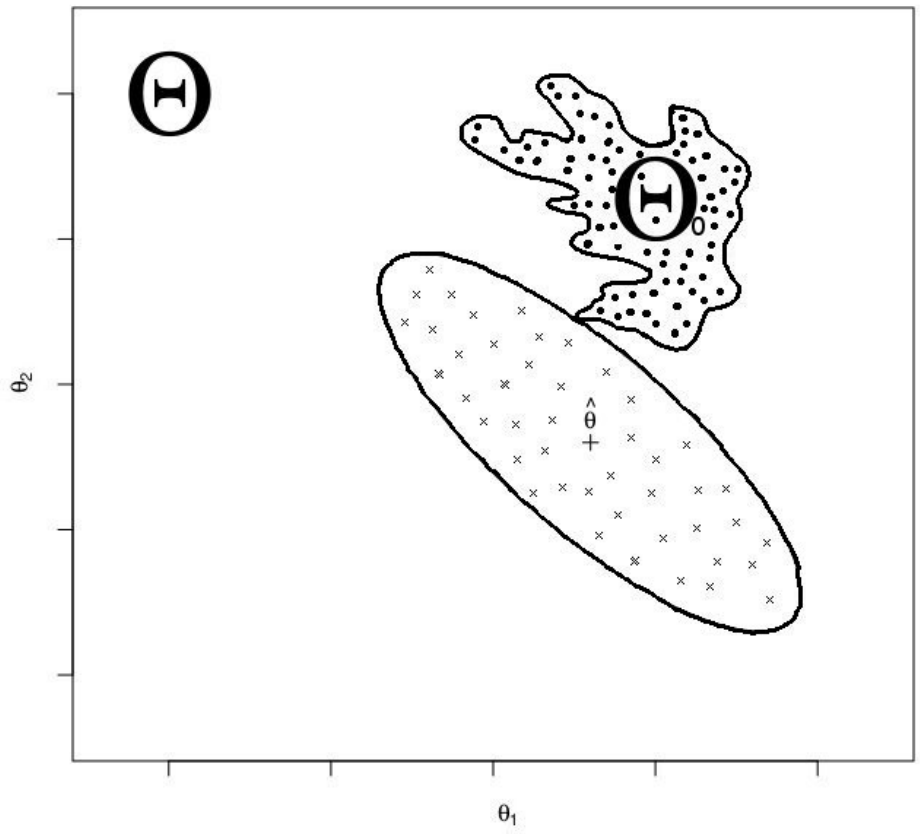

Figura 1: Menor região de confiança que intercepta pelo menos um ponto de Θ_0

1. $s(\varnothing) = 0$ e $s(\Theta) = 1$
2. Se $\Theta_1 \subset \Theta_2$, então $s(\Theta_1) \leq s(\Theta_2)$
3. Para quaisquer $\Theta_1, \Theta_2 \subset \Theta$, tem-se $s(\Theta_1 \cup \Theta_2) = \max\{s(\Theta_1), s(\Theta_2)\}$
4. Seja $\Theta_1 \subset \Theta$, então $s(\Theta_1) = \sup_{\theta \in \Theta_1} s(\{\theta\})$
5. $s(\Theta_1) = 1$ ou $s(\Theta_1^c) = 1$. Se $\widehat{\theta} \in \overline{\Theta_1}$ (fecho de Θ_1), então $s(\Theta_1) = 1$, se $\widehat{\theta} \in \overline{\Theta_1^c}$, então $s(\Theta_1^c) = 1$.

As propriedades acima mostram que s é uma medida possibilística definida nos subconjuntos do espaço paramétrico Θ (a definição de medida possibilística está contemplada nos itens 1 e 3 das propriedades acima). Há uma relação interessante entre o p-valor, calculado usando a estatística da razão de verossimilhanças, e o s-valor para uma hipótese nula simples (i.e., $H_0 : \theta = \theta_0$). Neste caso, sempre que F for estritamente crescente e contínua, pode-se mostrar que para cada hipótese nula tem-se a seguinte relação

$$s = 1 - F(F_{H_0}^{-1}(1 - p)),$$

em que $p = P_{H_0}(T_{H_0} > t) = 1 - F_{H_0}(t)$ com F_{H_0} sendo a função distribuição acumulada (estritamente crescente e contínua), calculada sob a hipótese nula H_0, da estatística da razão de verossimilhanças T_{H_0}. Com esta fórmula pode-se facilmente calcular o s-valor dado um p-valor. Basta saber qual a função acumulada utilizada no cálculo do p-valor e qual a função F utilizada no cálculo do s-valor. Supondo condições de regularidade para a verossimilhança (logaritmo estritamente côncavo) e para a função F (estritamente crescente e contínua), pode-se calcular o s-valor da seguinte maneira:

$$s(\Theta_0) = 1 - F(T_{\widehat{\theta}_0}), \tag{2}$$

em que $T_\theta = 2(\ell_{\widehat{\theta}}(X) - \ell_\theta(X))$ e $\widehat{\theta}_0 = \arg\max_{\theta \in \Theta_0} \ell_\theta(X)$ é a estimativa de máxima verossimilhança restrito a Θ_0 (veja o Lemma 3.2 de Patriota, 2013). A Fórmula (2)

destaca a diferença entre p-valor e s-valor, note que para o cálculo do p-valor a distribuição acumulada varia de acordo com a hipótese nula, i.e., $p = 1 - F_{H_0}(t)$, enquanto que o s-valor utiliza a distribuição acumulada para criar regiões de confiança para o vetor θ, i.e., $s = 1 - F(t)$, em que t é a estatística da razão de verossimilhanças observada. Um procedimento de maximização geral pode ser implementado supondo que o espaço paramétrico sob a hipótese nula é da forma $\Theta_0 = \{\theta \in \Theta : \ h(\theta) \leq 0, \ g(\theta) = 0\}$ com $h : \Theta \to \mathbb{R}^{k_1}$ e $g : \Theta \to \mathbb{R}^{k_2}$ funções que impõe as restrições da hipótese nula, esse procedimento será desenvolvido em pesquisas futuras.

Vale a pena ressaltar que tanto o p-valor quanto o s-valor dependem da estatística da razão de verossimilhanças T_{H_0}, e esta por sua vez varia dependendo do conjunto Θ_0 especificado. Porém, no s-valor não se utiliza a distribuição de probabilidade induzida por T_{H_0} e este é o motivo principal para que não ocorra conflitos lógicos sobre as relações entre subconjuntos de Θ.

Com respeito a decisão sobre aceitação ou rejeição da hipótese nula utilizando o s-valor, considere $H_0 : \theta \in \Theta_0$ com $\Theta_0 \subset \Theta$. Define-se a função Φ, o "grau de crença" em subconjuntos de Θ, da seguinte forma:

$$\Phi(\Theta_0) = \langle s(\Theta_0), s(\Theta_0^c) \rangle,$$

em que $\Phi(\Theta_0) = \langle 1{,}0 \rangle$ é o valor maximal (aceita-se imediatamente H_0, pois o s-valor indicará consistência total com Θ_0 e inconsistência total com Θ_0^c) e $\Phi(\Theta_0) = \langle 0{,}1 \rangle$ é o valor minimal (rejeita-se imediatamente H_0, pois o s-valor indicará inconsistência total com Θ_0 e consistência total com Θ_0^c).

O valor de ignorância total com respeito a aceitação/rejeição de H_0 ocorre quando $\Phi(\Theta_0) = \langle 1{,}1 \rangle$ (i.e., o EMV está na borda de $\overline{\Theta_0}$ indicando compatibilidade com as duas hipóteses). Excluindo-se os casos patológicos, na prática observa-se algo entre o valor minimal e o valor maximal que dependerá da estimativa de máxima verossimilhança, ou seja,

- $\Phi(\Theta_0) = \langle 1, b \rangle$, com $b \in (0{,}1]$, se $\widehat{\theta} \in \Theta_0$. Neste caso, se b for suficientemente pequeno poder-se-á aceitar H_0.
- $\Phi(\Theta_0) = \langle a, 1 \rangle$, com $a \in (0{,}1]$, se $\widehat{\theta} \in \Theta_0^c$. Neste caso, se a for suficientemente pequeno pod-se-á rejeitar H_0.

Note que se o EMV pertencer a Θ_0, então faz sentido aceitar H_0, porém não faz sentido rejeitá-la. Se por outro lado a estimativa pertencer a Θ_0^c, então faz sentido rejeitar H_0, porém não faz sentido aceitá-la. Suponha por exemplo que o interesse consiste em testar se as médias μ_1 e μ_2 são positivas e observa-se médias amostrais ($n = 1000$) $\bar{x} = 1000$ e $\bar{y} = 7000$ com variâncias $s_x^2 = 02$ e $s_y^2 = 13$. Neste caso faz pleno sentido aceitar a hipótese de que as médias populacionais são positivas. No mesmo exemplo anterior, suponha que agora o interesse consiste em testar se as médias μ_1 e μ_2 são ambas negativas e observa-se as mesmas médias e variâncias amostrais. Neste caso faz pleno sentido rejeitar a hipótese de que as médias populacionais são negativas. Por outro lado, se as médias amostrais estiverem perto de pelo menos um dos eixos do plano cartesiano, então não se poderá concluir com tanta certeza sobre aceitação e rejeição de que ambas as médias são positivas/negativas.

Observação 1: se a dimensão de Θ_0 for menor do que a dimensão de Θ, então tem-se sempre o caso $\Phi(\Theta_0) = \langle a, 1 \rangle$ com $a \in (0{,}1]$.

Observação 2: se a dimensão de Θ_0^c for menor do que a dimensão de Θ, então tem-se sempre o caso $\Phi(\Theta_0) = \langle 1, b \rangle$ com $b \in (0{,}1]$.

Observação 3: se a dimensão de Θ_0 for igual a dimensão de Θ_0^c, então o caso dependerá se $\widehat{\theta} \in \Theta_0$ ou $\widehat{\theta} \in \Theta_0^c$ como descrito logo acima.

Em geral, os testes de hipóteses quase sempre consideraram que a dimensão de Θ_0 é menor que a dimensão de Θ. Nos casos em que a dimensão de Θ_0 é igual a dimensão de Θ, o cálculo do p-valor não é simples (Mudholkar and Chaubey, 2009). Note que, neste trabalho, assumem-se hipóteses bem gerais em que Θ_0 pode ter dimensão menor ou até mesmo igual a dimensão de Θ.

Um problema em aberto é encontrar valores b^* e a^* tais que: 1) se $b < b^*$ então aceita-se H_0 e 2) se $a < a^*$ então rejeita-se H_0, levando-se em conta as observações acima. Caso $b \geq b^*$ ou $a \geq a^*$, então diz-se que é necessário mais informação para tomar alguma conclusão sobre aceitação/rejeição de H_0.

4.2 Cálculo do s-valor

O cálculo do s-valor para problemas mais gerais (cumprindo-se as condições de regularidade discutidas nas seções anteriores) deve seguir o seguinte algoritmo:

- Definir a função de verossimilhança para o modelo adotado;
- Encontrar a estimativa de máxima verossimilhança irrestrito, $\widehat{\theta}$. A estimativa de máxima verossimilhança $\widehat{\theta}$ é o ponto de referência para verificar se a afirmação contida na hipótese nula está perto ou longe;
- Definir a hipótese nula e o espaço paramétrico restrito $\Theta_0 \subset \Theta$;

- Encontrar a estimativa de máxima verossimilhança restrita a Θ_0, $\widehat{\theta}_0$;
- Calcular $d_0 = 2(\ell_{\widehat{\theta}}(X) - \ell_{\widehat{\theta}_0}(X))$ sempre que for possível. Caso contrário,

$$d_0 = -2\log\left(\frac{\sup_{\theta\in\Theta_0} L(x,\theta)}{\sup_{\theta\in\Theta} L(x,\theta)}\right);$$

- Encontrar a distribuição F relacionada com a razão de verossimilhanças (pode-se utilizar a qui-quadrado com k graus de liberdade, em que k é a dimensão de Θ). Caso a função F não seja contínua e estritamente crescente, o próximo passo não será válido;
- Se F for contínua e estritamente crescente, então calcular $s(\Theta_0) = 1 - F(d_0)$. Caso contrário, deve-se calcular as regiões de confiança como indica a fórmula (1).

4.3 Aplicação do s-valor

Considere uma amostra independente e identicamente distribuída $X = (X_1, \ldots, X_n)$ em que $X_1 \sim N_2(\theta, \Sigma)$ com $\theta = (\mu_1, \mu_2)^\top$ e Σ uma matriz de variâncias-covariâncias **conhecida**. Neste exemplo o espaço paramétrico é $\Theta = \{(\mu_1, \mu_2) \in \mathbb{R}^2 : \ \mu_1, \mu_2 \in \mathbb{R}\}$ e a função de verossimilhanças é

$$L_\theta(X) = \frac{1}{(2\pi)^n |\Sigma|^{\frac{n}{2}}} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(X_i-\theta)^\top\Sigma^{-1}(X_i-\theta)\right).$$

Pode-se mostrar que o EMV é

$$\widehat{\theta} = \bar{X} = \begin{pmatrix}\bar{X}_1\\ \bar{X}_2\end{pmatrix}$$

e o logaritmo da razão de verossimilhanças é

$$2(\ell_{\widehat{\theta}}(X) - \ell_\theta(X)) = n(\bar{X}-\theta)^\top\Sigma^{-1}(\bar{X}-\theta) \sim \chi^2_2. \quad (3)$$

Portanto, a região de confiança com nível de significância de α% é dada por

$$\Lambda_\alpha = \{\theta \in \Theta : \ n(\bar{X}-\theta)^\top\Sigma^{-1}(\bar{X}-\theta) \leq F_\alpha\}$$

em que F_α é o $(1-\alpha)$-quantil da qui-quadrado com dois graus de liberdade, χ^2_2. Neste exemplo, a função de distribuição acumulada F a ser utilizada é a da χ^2_2. Considere as três hipóteses nulas a seguir:

$$H_{01} : \mu_1 = \mu_2,$$
$$H_{02} : \mu_1 = \mu_2 = 0,$$
$$H_{03} : \mu_1 = \mu_2 = 0.2.$$

Os espaços paramétricos para as hipóteses nulas acima são, respectivamente: $\Theta_{01} = \{(\mu_1, \mu_2) \in \Theta : \ \mu_1 = \mu_2, \ \mu_1, \mu_2 \in \mathbb{R}\}$, $\Theta_{02} = \{(0,0)\}$ e $\Theta_{03} = \{(0.2,0.2)\}$. Como todos os espaços tem dimensão inferior a dimensão de Θ, tem-se que o s-valor sempre será igual a 1 no espaço paramétrico da hipótese alternativa. Ou seja, a função de crença sempre será do tipo $\Phi(\Theta_0) = \langle a, 1\rangle$ com $a = s(\Theta_0) \in (0,1]$ que deve ser obtido para cada hipótese acima. Note também que os estimadores de máxima verossimilhança restritos a Θ_{01}, Θ_{02} e Θ_{03} são dados, respectivamente, por

$$\widehat{\theta}_{01} = \begin{pmatrix}\bar{X}_1 + \bar{X}_2\\ \bar{X}_1 + \bar{X}_2\end{pmatrix}, \quad \widehat{\theta}_{02} = \begin{pmatrix}0\\0\end{pmatrix} \quad \text{e} \quad \widehat{\theta}_{03} = \begin{pmatrix}0.2\\0.2\end{pmatrix},$$

substituindo estes valores na razão de verossimilhanças (3) obtém-se

$$s(\Theta_{01}) = 1 - F\left(n(\bar{X}-\widehat{\theta}_{01})^\top\Sigma^{-1}(\bar{X}-\widehat{\theta}_{01})\right),$$
$$s(\Theta_{02}) = 1 - F\left(n\bar{X}^\top\Sigma^{-1}\bar{X}\right) \text{ e}$$
$$s(\Theta_{01}) = 1 - F\left(n(\bar{X}-\widehat{\theta}_{03})^\top\Sigma^{-1}(\bar{X}-\widehat{\theta}_{03})\right).$$

Dados: Suponha que uma amostra de tamanho $n = 100$ foi retirada e a média amostral observada $\bar{x} = (\bar{x}_1, \bar{x}_2)^\top = (-0.051, 0.074)^\top$ e $\Sigma = \begin{pmatrix}1 & 0.8\\ 0.8 & 1\end{pmatrix}$. Tem-se que os s-valores e os p-valores são, respectivamente,

Para H_{01}: $s_1(\Theta_{01}) = 0.1420$ e $p_1(\Theta_{01}) = 0.0481$,

Para H_{02}: $s_2(\Theta_{02}) = 0.1410$ e $p_2(\Theta_{02}) = 0.1410$,

Para H_{03}: $s_3(\Theta_{03}) = 0.0197$ e $p_3(\Theta_{03}) = 0.0197$.

Observe que as medidas de evidências tem comportamentos diferentes. O s-valor mostra uma evidência forte contra H_{03}, porém para as hipóteses H_{01} e H_{02} as evidências não se mostram importantes. Por outro lado, o p-valor mostra evidência contra H_{01} e H_{03}, mas não apresenta evidência contra H_{02}. Contudo, note que $\Theta_{02} \subset \Theta_{01}$ e $\Theta_{03} \subset \Theta_{01}$, portanto esperar-se-ía que a medida de evidência fosse menor para H_{02} do que para H_{01}, os resultados acima mostram que a característica esperada ocorre apenas para o s-valor. Neste exemplo específico, $p_2(\Theta_{02}) = s_2(\Theta_{02})$ e $s_3(\Theta_{03}) = p_3(\Theta_{03})$ isso ocorreu pois as distribuições são equivalentes, i.e., $F \equiv F_{H_{02}} \equiv F_{H_{03}}$. De maneira geral, sempre que o vetor de parâmetros for inteiramente especificado na hipótese nula, as duas metodologias coincidirão. Para mais aplicações consulte Patriota (2013).

5 Conclusões e comentários finais

Este trabalho discutiu algumas inconsistências inerentes ao p-valor e apresentou uma nova medida de evidência que as contorna. Esta medida foi previamente publicada

em Patriota (2013). Vale ressaltar que algumas questões ainda permanecem em aberto:

- Encontrar um ponto de corte para elaborar uma conclusão sobre aceitação ou rejeição da hipótese nula (*via* funções de perda, métodos frequentistas, etc);
- Comparar a performance do s-valor com metodologias Bayesianas (fator de Bayes, posterioris);
- Comparar outros tipos de regiões de confiança na construção do s-valor (*via* estatística de Wald, Escore, etc);
- Estudar as implicações filosóficas dessa nova medida na inferência clássica;
- Estudar as propriedades do s-valor quando a função de log-verossimilhança não é côncava;
- Utilizar diferentes funções acumuladas F para obter diferentes interpretações do s-valor;
- Caso seja utilizada a distribuição assintótica (quiquadrado) para representar a função F, pode-se aplicar correções de segunda ordem para melhorar o desempenho das regiões de confiança.

Agradecimentos

Agradeço ao Fábio Mariano Bayer pelo convite para escrever este artigo de divulgação e pelas sugestões importantes que o tornaram mais fluido e de fácil entendimento. Agradeço também a Fernanda Cristiane de Oliveira e ao Jonatas Eduardo Cesar pela leitura cuidadosa e sugestões textuais que auxiliaram na compreensão geral do conteúdo.

Referências

Bahadur, R.R., Baghavachari, M. (1972). Some asymptotic properties of likelihood ratios on general sample spaces, In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, 1*, Univ. California Press, Berkeley, California, 129–152.

Fisher, R.A. (1935). The logic of inductive inference, *Journal of the Royal Statistical Society, 98*, 39–54.

Izbicki, R., Fossaluza, V., Hounie, A.G., Nakano, E.Y., Pereira, C.A. (2012). Testing allele homogeneity: The problem of nested hypotheses, *BMC Genetics*, **13**:103.

Mudholkar, G.S., Chaubey, Y.P. (2009). On defining p-values, *Statistics & Probability Letters, 79*, 1963–1971.

Neyman, J., Pearson, E.S. (1933). On the Problem of the Most Efficient Tests of Statistical Hypotheses, *Philosophical Transactions of the Royal Society of London. Series A, 231*, 289–337.

Patriota, A.G. (2013). A classical measure of evidence for general null hypotheses, *Fuzzy Sets and Systems, 233*, 74–88.

Popper, K.R. (1989). Conjectures and Refutations: The Growth of Scientific Knowledge. London: Routledge.

Schervish, M.J. (1996). P Values: What they are and what they are not, *The American Statistician, 50*, 203–206.

Sprott, D.A. *Statistical Inference in Science*, Springer, New York, 2000.