Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Estatística

Mestrado em Estatística
Inferência Bayesiana
1^o semestre de 2020

O Princípio de Verossimilhança

Gustavo L. Gilardoni

14 de Janeiro de 2022

Nestas notas será introduzido o *Princípio de Verossimilhança* e discutida a sua relação com os Princípios de *Condicionalidade* e de *Suficiência*. Vamos ver também que nenhum dos métodos usuais da Inferência clássica satisfaz o Princípio de Verossimilhança. Finalmente, será discutido o Princípio da Probabilidade Inversa de Laplace, desenvolvido na segunda metade do século XVIII, que na próxima unidade servirá como motivação para a introdução do Paradigma Bayesiano.

0 Introdução

A *Inferência Estatística* consiste no estudo de ferramentas usadas para se fazer afirmações sobre um modelo a partir da observação de quantidades aleatórias supostamente geradas por esse modelo. Enquanto a maioria dos cursos de Inferência Estatística aprofundam em assuntos tais como *estimação* e *testes de hipóteses* e *Método de Máxima Versossimilhança*, raramente discutem princípios gerais que deveriam guiar o processo inferencial.

Nessa perspectiva, Estas notas tem por objetivo apresentar e discutir o assim chamado *Princípio de Verossimilhança* (PV) e a sua relação com os possivelmente mais intuitivos *Princípio de Condicionalidade* (PC) e *Princípio de Suficiência* (PS). Os três princípios são apresentados propondo uma reflexão sobre alguns exemplos extremamente simples.

Cada um dos três princípios são apresentados informalmente nas Seções 1–3, onde também discute-se porque as medidas centrais do paradigma frequêntista, tais como p-valores ou variâncias amostrais de estimadores, violam o PV.

A Seção 4 segue o trabalho fundamental de Allan Birnbaum (1962). A partir do conceito de *evidência* associado a um par (experimento,observação) define-se formalmente os três princípios mencionados acima de forma a permitir o enunciado e a prova do teorema fundamental que asevera a equivalência entre o PV e a conjunção dos PC e PS. Finalmente, na Seção 5 apresentamos brevemente o paradigma da *probabilidade inversa*, formulado por Laplace em 1774 (veja, por exemplo, Hald (2007)) e o paradigma Bayesiano, que respeitam o PV. Finalmente, algumas conclusões são apresentadas na Seção 6 e um apêndice contem código **R** usado na elaboração das figuras e no cálculo dos exemplos.

Além do artigo fundamental de Birnbaum (1962) mencionado acima, o aluno interessado em aprofundar o assunto pode (deve!) consultar a monografia de Berger and Wolpert (1988). Entre outras referências, o assunto também é tratado no texto de Robert (2007).

Finalmente, antes de terminar esta seção, uma advertência: nada do escrito a seguir está certo nem está errado. O objetivo da apresentação é simplesmente provocar uma reflexão sobre algumas questões fundamentais.

1 O Princípio de Verossimilhança

Exemplo 1. Uma moeda foi lançada obtendo-se a seguinte sequencia de *caras* e *coroas*, codificadas respectivamente como "1"e "0":

$$1\,1\,0\,1\,1\,1\,1\,0\,1\,1\,1\,0 \tag{1}$$

Deseja-se testar a hipótese $H_a : \theta > 0.5$, onde θ é a probabilidade de cara num único lançamento dessa moeda. □

Abaixo seguem duas possíveis respostas. Você acha que alguma delas está certa? Pode imaginar alguma outra?

Resposta 1.1. Supondo que os lançamentos são idênticos e independentes, como tem 12 lançamentos, o número de caras X segue uma distribuição Binomial com $n = 12$ e probabilidade de sucesso θ, isto é, $P_\theta(X = x) = \binom{12}{x} \theta^x (1-\theta)^{12-x}$ $(x = 0, 1, \ldots, 12)$. Como a distribuição binomial é uma família com razão de verossimilhança monótona e a hipótese $\theta > 0.5$ é unilateral, existe um teste PUM (poder uniformemente máximo), cuja região crítica tem a forma $X \geq c$, onde c e o nível de significância α satisfazem que $P_{\theta=0.5}(X \geq c) = \alpha$. Alternativamente, considerando que foram observadas $x = 9$ caras, o *p-valor* do teste é

$$P_{\theta=0.5}(X \geq 9) = \sum_{x=9}^{12} \binom{12}{x} (0.5)^x (1-0.5)^{12-x} \doteq 0.0730 \tag{2}$$

(se não quiser fazer a conta manualmente, simplesmente escreva "`1-pbinom(8,12,0.5)`" numa sessão do **R**). Esse p-valor de 7.3% mostra evidência moderada contra $H_0 : \theta \leq 0.5$. Por exemplo, H_0 seria rejeitada para $\alpha = 0.10$ mas não seria rejeitada para $\alpha = 0.05$.

Resposta 1.2. Como o último lançamento foi coroa, podemos pensar que a moeda foi lançada até obter a terceira coroa. Nesse caso teríamos que o número de caras Y até obter a terceira coroa segue uma distribuição binomial negativa com $k = 3$ fracassos e probabilidade de sucesso θ, isto é $P_\theta(Y = y) = \binom{y+2}{y} \theta^y (1-\theta)^3$ para $y = 0, 1, \ldots$. Como a distribuição binomial negativa tem razão de verossimilhança monótona, o teste PUM para $H_a : \theta > 0.5$ tem região crítica $Y \geq c$. Portanto, considerando que foi observado $y = 9$ (i.e. 9 caras até a terceira coroa), o p-valor do teste é

$$P_{\theta=0.5}(Y \geq 9) = \sum_{y=9}^{\infty} \binom{y+2}{y} (0.5)^y (1-0.5)^3 = 1 - \sum_{y=0}^{8} \binom{y+2}{y} (0.5)^y (1-0.5)^3 \doteq 0.0327 \tag{3}$$

(novamente, se não escreva "`1-pnbinom(8,3,0.5)`" quiser fazer a conta manualmente, numa sessão do **R**). Esse p-valor de 3.27% mostra evidência forte contra a $H_0 : \theta \leq 0.5$. Comparando com a resposta 1.1, H_0 seria rejeitada para $\alpha = 0.10$ e **também** para $\alpha = 0.05$.

A diferença entre as duas respostas radica essencialmente no que é conhecido na literatura estatística como *regra de parada* . Enquanto na resposta 1.1 imaginamos (sabemos?) que o experimento terminaria após o décimo-segundo lançamento da moeda, na resposta 1.2 imaginamos que terminaria após a terceira coroa. É importante notar que enquanto a sequência de caras e coroas (1) poderia ter sido o resultado tanto do experimento binomial com $n = 12$ quanto do binomial negativo com $k = 3$, muitos possíveis resultados do primeiro não poderiam ocorrer no segundo e vice-versa Por exemplo, a sequência `(1,1,0,1,1,1,1,0,1,1,1,1)` poderia ter sido o resultado do experimento binomial com $n = 12$ mas não do binomial negativo com $k = 3$; a sequência `(0,0,1,0)` poderia ter sido o resultado do experimento binomial negativo com $k = 3$ mas não do binomial com $n = 12$.

Independentemente de como foi o delineamento experimental, binomial com $n = 12$, binomial negativo com $k = 3$ ou qualquer outro que deu origem à sequência (1), o

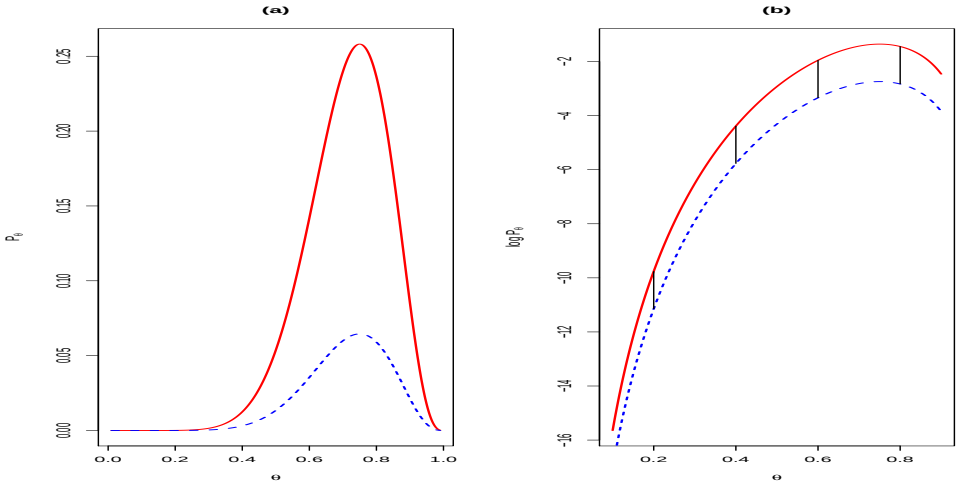


Figura 1: **(a)** verossimilhanças e **(b)** log verossimilhanças para as Respostas 1.1 (binomial com $n = 12$ e $x = 9$, em vermelho) e 1.2 (binomial negativa com $k = 3$ e $y = 9$, em azul).

exemplo apresenta uma questão fundamental no sentido de questionar se a regra de parada deveria influir na inferência a ser realizada. Nesse sentido, veja que o único que efetivamente aconteceu com a moeda foi a sequência (1). Qualquer regra de parada que tenha sido usada está exclusivamente na cabeça do experimentador. Mais precisamente, neste caso, a diferença entre as duas possíveis regras de parada está no que **não** aconteceu e não no que aconteceu!

O Princípio de Verossimilhança (PV) afirma que, dados dois experimentos E_f e E_g que consistem em observar respectivamente quantidades aleatórias X com densidade $f_\theta(x)$ e Y com densidade $g_\theta(y)$, e se x e y são dois possíveis resultados experimentais tais que, existe uma constante $c = c(x, y)$, que não depende do parâmetro θ, para a qual a igualdade $f_\theta(x) = cg_\theta(y)$ vale para todo θ, então a evidência obtida para se fazer inferência sobre θ após observar $X = x$ no experimento E_f é a mesma que aquela obtida após observar $Y = y$ no experimento E_g. Em português corrente, se dois resultados para dois experimentos fornecem verossimilhanças proporcionais, então fornecem a mesma evidência.

Voltando ao exemplo, no experimento binomial temos que $P_\theta(X = 9) = \binom{12}{9} \theta^9 (1-\theta)^3 = 220\,\theta^9 (1-\theta)^3$, enquanto no experimento binomial negativo $P_\theta(Y = 9) = \binom{11}{2} \theta^9 (1-\theta)^3 = 55\,\theta^9 (1-\theta)^3$ (veja a Figura 1). Como as duas funções de verossimilhança são proporcionais, o PV afirma que a evidência obtida após observar $x = 9$ no experimento binomial é igual à obtida após observar $y = 9$ no binomial negativo.

Se a evidência deveria ser, de acordo ao PV, a mesma, não faria sentido ter dois p-valores diferentes em (2) e (3). Ou, em outras palavras, o cálculo de p-valores como medidas de evidência estatística viola o PV. Podemos explicar simplesmente o motivo: Pelo PV, toda a evidência deveria estar contida na função de verossimilhança $\theta \rightarrow$

$L_x(\theta) = f_\theta(x)$, onde x é **somente** aquele que foi efetivamente observado. Possíveis resultados experimentais que **não** ocorreram não fornecem evidência sobre θ. Por esse motivo as vezes fala-se informalmente que o PV assevera a irrelevância de resultados não observados. Veja que, nos cálculos dos p-valores (2) e (3), além dos valores observados $x = 9$ e $y = 9$, entram também $x = 10, 11$ e 12 em (2) e $y = 10, 11, \ldots$ em (3).

Outras medidas clássicas também violam o PV. Por exemplo, quando se diz que um estimador $\hat{\theta}(X)$ é não-viesado nos referimos ao fato que $\int \hat{\theta}(x)\, f_\theta(x)\, dx \equiv \theta$, mas no cálculo dessa integral considera-se $f_\theta(x)$ para todo x. Da mesma forma, quando falamos da variância de um estimador, nos referimos a um momento calculado com respeito a todos os possíveis resultados do experimento, portanto violando também o PV.

inclusive os que não ocorreram no experimento

PV: se as verossimilhanças são proporcionais, a evidência não deveria ser diferente.

2 O Princípio de Condicionalidade

Exemplo 2. Num laboratório existem duas balanças. As leituras das duas balanças são normalmente distribuídas e não-viesadas, mas enquanto o erro da balança A é 0.1g, o da balança B é 0.9g (sim, a balança A é nove vezes mais precisa que a B). Mais precisamente, quando um objeto cujo peso real é μ é pesado na balança A, ela retorna uma variável aleatória $X \sim \text{Normal}(\mu, \sigma^2 = 0.01)$, enquanto a balança B retorna $Y \sim \text{Normal}(\mu, \sigma^2 = 0.81)$.

Um pesquisador deseja calcular um intervalo de confiança para o peso de um certo objeto. Para isso, realiza o seguinte experimento: primeiro lança uma moeda não viesada, se o resultado for cara ele pesa o objeto na balança A, se for coroa o pesa na balança B.

Suponha que o resultado do lançamento da moeda foi cara e a balança (A, pois o lançamento foi cara!) retornou um peso de 3.7g. A questão é como construir um IC para μ com, digamos, 95% de confiança. □

Abaixo seguem duas possíveis respostas. Você acha que alguma delas está certa? Imagina alguma alternativa? (Veja que os dois intervalos fornecidos abaixo são bem diferentes: o da Resposta 2.1 é muito mais preciso.)

Resposta 2.1. Como a moeda foi cara, sabemos que o objeto foi pesado na balança A, cujo erro padrão é 0.1g. Logo, usando o IC usual $\bar{x} \pm z_{\alpha/2}\, \sigma/\sqrt{n}$ com $n = 1$, $\bar{x} = x = 3.7$, $\sigma = 0.1$ e $\alpha = 0.05$, temos que o IC desejado é $3.7 \pm (1.96)\,(0.1) = (3.504; 3.896)g$.

Resposta 2.2. Seja Z o resultado de uma pesada. Segue que Z é uma *mistura* da $\text{Normal}(\mu, \sigma^2 = 0.01)$ e da $\text{Normal}(\mu, \sigma^2 = 0.81)$ com pesos $w_1 = w_2 = 0.5$ (veja a Figura 2). Outra forma de dizer isso é como segue. Seja M uma quantidade aleatória que toma os valores A e B com probabilidades 0.5. Logo $[Z|M = A] = X \sim \text{Normal}(\mu, \sigma^2 = 0.01)$, enquanto $[Z|M = B] = Y \sim \text{Normal}(\mu, \sigma^2 = 0.81)$. Pelas regras dos momentos condicionais, $\text{E}(Z) = \text{E}[\text{E}(Z|M)] = 0.5\,\text{E}(Z|M = A) + 0.5\,\text{E}(Z|M = B) = 0.5\mu + 0.5\mu = \mu$, enquanto $\text{Var}(Z) = \text{E}[\text{Var}(Z|M)] + \text{Var}[\text{E}(Z|M)] = 0.5(0.01) + 0.5(0.81) + \text{Var}(\mu) = 0.41 + 0 = 0.41 \doteq (0.640)^2$.

Numa primeira aproximação, poderíamos então usar a fórmula $\bar{z} \pm z_{\alpha/2}\, \sigma/\sqrt{n}$ com $n = 1$, $\bar{z} = z = 3.7$, $\sigma \doteq 0.640$ e $z_{\alpha/2} = 1.96$, e o intervalo desejado seria $3.7 \pm$

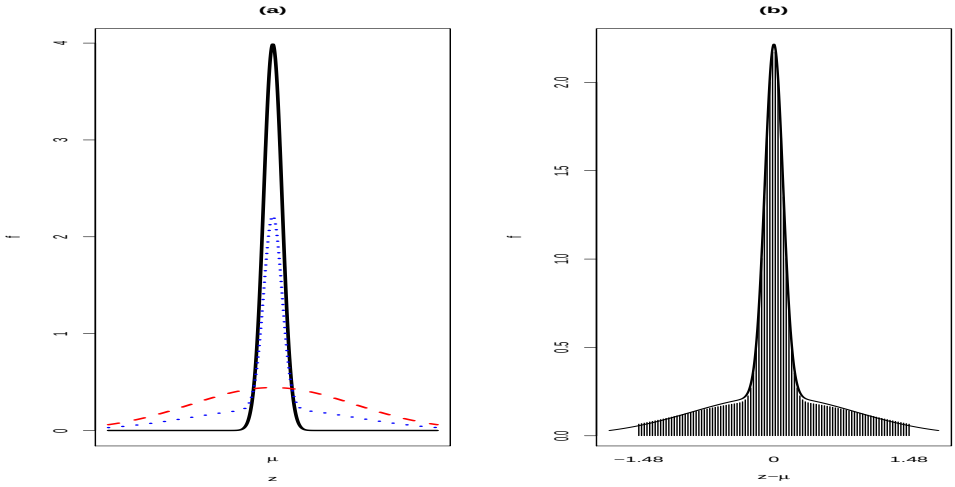


Figura 2: **(a)** densidades da distribuição N(μ,$\sigma^2 = 0.01$) (preto), N(μ,$\sigma^2 = 0.81$) (vermelho) e da mistura dessas duas com pesos $w_1 = w_2 = 0.5$ (azul); **(b)** densidade da mistura das Normais (0,$\sigma^2 = 0.01$) e (0,$\sigma^2 = 0.81$) com pesos $w_1 = w_2 = 0.5$ (verde); com os correspondentes percentis $v_{0.025} \doteq -1.480$ e $v_{0.975} \doteq 1.480$ (verifique que a área riscada é igual a 0.95).

$(1.96)\,(0.640) = (2.446; 4.954)g$.

O problema com esse IC é que o uso de $z_{\alpha/2} = 1.96$ vem da tabela Normal, e mistura de Normais não é Normal! Uma solução mais apropriada, **sempre nessa linha de pensamento**, passa por observar que a quantidade aleatória $V = Z - \mu$, cuja distribuição é a de uma mistura das Normais (0,$\sigma^2 = 0.01$) e (0,$\sigma^2 = 0.81$) com pesos $w_1 = w_2 = 0.5$ [ver Figura 2 (b)], é uma quantidade *pivotal* para μ (i.e. é monótona em μ e a sua distribuição não depende de μ). Como a distribuição de $V = Z - \mu$ é simétrica com respeito a $v = 0$ e é fácil calcular numericamente que os percentis 2.5 e 97.5% dessa distribuição são respectivamente $v_{0.025} \doteq -1.480$ e $v_{0.975} \doteq 1.480$ (verifique!), um IC mais apropriado seria então $z \pm 1.480 = 3.7 \pm 1.480 = (2.22; 5.18)g$.

Observação. Como já foi dito, na linha de pensamento da Resposta 2.2, o segundo IC apresentado é mais apropriado que o primeiro, embora para efeitos desta discussão qual deles usar não é muito relevante. Veja que qualquer um deles é muito menos preciso que o IC $\mu \in (3.504; 3.896)$g que foi obtido na Resposta 2.1.

Observação. Se lançar uma moeda para decidir qual balança usar pareceria ser tosco, imagine uma situação onde o pesquisador pesará o objeto na balança A se ela não estiver sendo usada por outros pesquisadores, e que a chance disso ocorrer é 50%. O efeito será o mesmo do lançamento da moeda.

Se no enunciado do exemplo fosse informado somente que o objeto foi pesado na balança A, sem nenhuma referência à moeda, qualquer referência à balança B é desnecessária e é claro que a resposta deveria ser o IC $(3.504; 3.896)$g da Resposta 2.1. Por outro lado, se a moeda fosse lançada para escolher a balança, **mas o resultado do**

lançamento fosse omitido do resultado do experimento (isto é, sabemos somente que o resultado da pesada foi 3.7g, mas não qual das duas balanças foi usada), o IC $(2.22; 5.18)$g da Resposta 2.2 seria apropriado. A reflexão proposta pelo exemplo basicamente pergunta se o fato da balança A ter sido escolhida com o lançamento da moeda, **cuja probabilidade de cara não depende de** μ, deve influir na escolha do intervalo. Em outras palavras, se sabemos que a balança B **não** foi usada, por quê a sua precisão deveria ser levada em conta para se fazer inferência sobre μ? Se no enunciado a precisão da balança B fosse mudada para, digamos, 0.5 ou 10 ou ainda 1000g, o IC da resposta 2.2 deveria ser também mudado. Mas como sabemos que a balança B não foi usada, pareceria que a sua precisão deveria ser irrelevante para se fazer inferência sobre μ.

O PC afirma que, quando um experimento é escolhido por algum mecanismo aleatório cuja distribuição de probabilidade **não** depende do parâmetro em estudo, somente o experimento que foi efetivamente realizado é o seu resultado são relevantes para se fazer inferência sobre esse parâmetro. No caso do exemplo proposto no começo desta seção, ele afirma que o fato da balança A ter sido escolhida pelo lançamento da moeda é irrelevante para fazer inferência sobre μ. Portanto, de acordo ao princípio de condicionalidade, o IC usando a mistura de normais proposto na resposta 2.2 não deveria ser usado. Note porém que o PC não implica que deveríamos usar o IC da resposta 2.1, somente que qualquer IC ou outro método inferencial **não** deveriam levar em conta o lançamento da moeda.

3 O Princípio de Suficiência

.

É o menos controverso dos três princípios discutidos aqui. O princípio de suficiência postula que a evidência resultante de observar $X = x$ num dado experimento é a mesma que a de observar $T(X) = T(x)$, onde T é um estatístico suficiente. Por exemplo, suponha que num experimento observa-se uma amostra de tamanho $n = 10$ da distribuição Normal com média desconhecida μ e variância conhecida $\sigma^2 = 1$. Nesse caso, temos a mesma evidência pelo fato de conhecer que o resultado do experimento foi $(x_1, \ldots, x_{10})$ do que se somente soubéssemos $\bar{x} = (10)^{-1} \sum_{i=1}^{10} x_i$, visto que o estatístico $\bar{X} = (10)^{-1} \sum_{i=1}^{10} X_i$ é suficiente.

Para efeito do princípio podemos usar qualquer uma das duas definições usuais de suficiência: (1) O estatístico $T(X)$ é suficiente se a distribuição condicional de X dado que $T(X) = t$ não depende do parâmetro θ para (quase) todo t ou (2) O estatístico $T(X)$ é suficiente se e somente se existem duas funções k e h tais que $f_\theta(x) = k(x)\, h(T(x); \theta)$, onde $k(x)$ depende da observação x mas não do parâmetro θ enquanto $h(T(x); \theta)$ depende do parâmetro θ e do valor da estatística $T(x)$ (esta última definição é usualmente conhecida como Lema da Fatoração de Neyman).

4 O Teorema de Birnbaum

Em 1962 Allan Birnbaum publico um artigo seminal no qual mostrou que o PV era equivalente ao PC e PS. Para isso ele define um conceito de evidência $\text{Ev}(E_f, x)$ associado a um experimento E_f no qual se observa uma quantidade aleatória $X \sim f_\theta$ e a observação $X = x$. Essa evidência pode ser qualquer medida, incluso clássica como um intervalo de confiança, p-valor, um par (estimativa;erro de estimação) etc. A formulação é muito ampla e a única relação necessária para essa evidência é a igualdade. Isto é, para a formulação de Birnbaum somente precisamos dizer quando que $\text{Ev}(E_f, x)$ é igual à $\text{Ev}(E_g, y)$, mas não precisamos falar por exemplo de "mais"ou "menos"evidência. Com esse conceito de evidência definimos agora formalmente os PV, PC e PS. Salientamos que o parâmetro θ é o mesmo em todos os experimentos que seguem, como na discussão do PV acima, onde θ era a probabilidade de cara da mesma moeda do Exemplo 1, independente se a regra de parada determina uma distribuição Binomial ou Binomial Negativa ou qualquer outra.

Definição 1 (Princípio de Verossimilhança). Sejam E_f e E_g dois experimentos nos quais observam-se respectivamente $X \sim f_\theta(x)$ e $Y \sim g_\theta(y)$, e suponha que x e y são dois resultados para os quais existe uma constante $c = c(x, y)$, que não depende de θ, tal que $f_\theta(x) = cg_\theta(y)$ para todo θ. Então $\text{Ev}(E_f, x) = \text{Ev}(E_g, y)$.

Definição 2 (Princípio de Condicionalidade). Sejam E_f e E_g dois experimentos quaisquer e defina um experimento composto E^c que consiste em lançar uma moeda não-viesada e escolher realizar E_f se o resultado do lançamento for cara ou realizar E_g se for coroa. O resultado desse experimento composto é reportado como o par (E, z), onde $E = E_f$ e $z = x$ se o lançamento foi cara e o resultado do experimento realizado foi $X = x$, ou $E = E_g$ e $z = y$ se foi coroa e observou-se $Y = y$. Então

$$\text{Ev}(E^c, (E, z)) = \text{Ev}(E, z)\,. \tag{4}$$

Observação. Como mencionado anteriormente no exemplo das balanças, é essencial aqui que o resultado do experimento composto seja o par formado pelo experimento efetivamente realizado e o seu resultado. Isto é, após realizar o experimento E^c, sabemos qual dos experimentos E_f ou E_g foi efetivamente realizado.

Definição 3 (Princípio de Suficiência). Seja E_f um experimento que consiste em observar $X \sim f_\theta(x)$ e suponha que o estatístico $S(X)$ é suficiente. Então, se x e y são dois possíveis resultados do experimento para os quais $S(x) = S(y)$, segue que $\text{Ev}(E_f, x) = \text{Ev}(E_f, y)$

Observação. Os PC e PS enunciados aqui são versões menos restritivas que resultam suficientes para provar a equivalência com o PV no próximo teorema. Versões mais fortes desses princípios são possíveis. Por exemplo, no PC poderiamos considerar a escolha entre mais de dois experimentos e/ou com mecanismos aleatórios com probabilidades desiguais sempre, é claro, que as probabilidades com que são escolhidos os experimentos não dependa de θ. → Ou seja, depende só de X

Teorema 1. *O PV vale se, e somente se, os PC e PS valem.*

Demonstração. Faremos a prova somente no caso discreto. O caso continuo requer formalismos de Teoria da Medida que fogem ao escopo da disciplina.

Começamos pela parte mais fácil, que é mostrar que PV implica PC e PS. Suponha então que o PV vale e considere o experimento composto como no PC. A verossimilhança associada a $(E^c, (E, z))$ é

$$P_\theta^c(E, z) = P_\theta^c(E)\, P_\theta^c(z|E) = \begin{cases} \frac{1}{2}\, f_\theta(x) & \text{se } E = E_f,\ z = x \\ \\ \frac{1}{2}\, g_\theta(y) & \text{se } E = E_g,\ z = y \end{cases}, \tag{5}$$

enquanto a verossimilhança associada a (E, z) é

$$P_\theta(E, z) = \begin{cases} f_\theta(x) & \text{se } E = E_f,\ z = x \\ \\ g_\theta(y) & \text{se } E = E_g,\ z = y \end{cases}. \tag{6}$$

Como (5) e (6) são proporcionais (alias, com $c = \frac{1}{2}$), segue pelo PV que $\text{Ev}(E^c, (E, z)) = \text{Ev}(E, z)$, logo o PC também vale.

Para mostrar que o PV implica o PS, considere no PV o caso em que os dois experimentos são iguais (i.é. $f_\theta \equiv g_\theta$ e $X \overset{d}{=} Y$), e suponha que T é suficiente. Se $T(x) = T(y)$, segue pelo Lema da Fatoração que $f_\theta(x) = k(x)\, h(T(x); \theta) = [k(x)/k(y)]\, k(y)\, h(T(y); \theta) = [k(x)/k(y)]\, f_\theta(y)$, o que mostra que a verossimilhança associada ao par (E_f, x) é proporcional àquela associada a (E_f, y), com constante $c = k(x)/k(y)$. Logo, pelo PV, segue que $\text{Ev}(E_f, x) = \text{Ev}(E_f, y)$

Finalmente, para provar o recíproco, isto é, que o PC e o PS conjuntamente implicam o PV, considere que o PC e o PS valem, e sejam (E_f, x_0) e (E_g, y_0) tais que $f_\theta(x_0) = c g_\theta(y_0)$.Precisamos mostrar então que $\text{Ev}(E_f, x_0) = \text{Ev}(E_g, y_0)$

Considere o experimento composto E^c. Pelo PC sabemos que (4) vale para todo (E, Z), portanto para os casos particulares que $(E, z) = (E_f, x_0)$ e que $(E, z) = (E_g, y_0)$ temos que

$$\begin{cases} \text{Ev}(E^c, (E_f, x_0)) = \text{Ev}(E_f, x_0) \ \text{ e} \\ \\ \text{Ev}(E^c, (E_g, y_0)) = \text{Ev}(E_g, y_0) \end{cases}. \tag{7}$$

Para usar o PS, no experimento E^c defina o estatístico

$$S(E, z) = \begin{cases} (E_g, y_0) & \text{se } E = E_f,\ z = x_0 \\ \\ (E, z) & \text{caso contrário} \end{cases} \tag{8}$$

(isto é, $S(E, z)$ é a identidade com a única excepção que $S(E_g, y_0) = (E_f, x_0)$). Para mostrar que $S(E, z)$ é suficiente, defina

$$k(E, z) = \begin{cases} \frac{1}{2}\, c & \text{se } E = E_f,\ z = x_0 \\ \\ \frac{1}{2} & \text{caso contrário} \end{cases}$$

e

$$h(S(E,z);\theta) = \begin{cases} g_\theta(x) & \text{se } S(E,z) = (E_g, y) \text{ para algum } y \\ \\ f_\theta(y) & \text{se } S(E,z) = (E_f, x) \text{ para algum } x \end{cases},$$

de forma que a verossimilhança (5) é igual ao produto $k(E,z)\, h(S(E,z);\theta)$ e a suficiência segue pelo Lema da Fatoração. [A igualdade $P^c(E,z) = k(E,z)\, h(S(E,z);\theta)$ é trivial se $S(E,z) \neq (E_g, y_0)$; por outro lado $S(E,z) = (E_g, y_0)$ em dois casos, (i) quando $(E,z) = (E_g, y_0)$ e (ii) $(E,z) = (E_f, x_0)$; no caso (i) temos $P^c(E_g, y_0) = \frac{1}{2} g_\theta(y_0) = k(E_g, y_0)\, h(S(E_g, y_0);\theta)$; enquanto no caso (ii) temos que $P^c(E_f, x_0) = \frac{1}{2} f_\theta(x_0) = \frac{1}{2}\, c\, g_\theta(x_0) = k(E_f, x_0)\, h(S(E_g, y_0);\theta)$.]

Finalmente, como S é suficiente e $S(E_g, y_0) = S(E_f, x_0) = (E_f, x_0)$, segue pelo PS que $\text{Ev}(E^c, (E_f, x_0)) = \text{Ev}(E^c, (E_g, y_0))$. Isto junto com (7) permite concluir que $\text{Ev}(E_f, x_0) = \text{Ev}(E_g, y_0)$, como queríamos provar. □

5 Laplace e o paradigma da Probabilidade Inversa

Considerando que o PS é amplamente aceito, negar o PV levaria a ter que negar o PC, o que, como discutido acima, teria consequências pouco intuitivas. Ainda, como aceitar o PV leva a desistir da maioria das medidas frequentistas, devemos nos perguntar como seria possível fazer inferência sobre θ usando tão somente a informação na função de verossimilhança $\theta \mapsto f_\theta(x)$, onde x é fixo e igual ao valor que foi efetivamente observado.

É sugestivo que possivelmente o primeiro paradigma estatístico enunciado precisamente como tal age em conformidade com o PV. Laplace em 1774 formula o conceito de *probabilidade inversa*, que enuncia da seguinte forma:

> Se um evento pode ser a consequência de n causas diferentes, a probabilidade da existência dessas causas dado o evento estão para cada outra como as probabilidades do evento dadas as causas, de forma que a probabilidade de cada causa é igual a probabilidade do evento dada essa causa dividida pela soma das probabilidades do evento dada cada uma das causas.

(Laplace, 1774, traduzido de Hald, 2007, pág. 35). Laplace não oferece nenhuma justificativa para esse princípio, para ele (em 1774!) é uma afirmação obvia. Denote por $X = x$ o evento e suponha que θ é uma variável aleatória tomando valores em $\{\theta_1, \ldots, \theta_n\}$. Em notação moderna o paradigma da probabilidade inversa de Laplace significa que

$$\frac{P(\theta = \theta_i | X = x)}{P(\theta = \theta_j | X = x)} = \frac{P(X = x | \theta = \theta_i)}{P(X = x | \theta = \theta_j)}$$

e

$$P(\theta = \theta_i | X = x) = \frac{P(X = x | \theta = \theta_i)}{\sum_{i=j}^{n} P(X = x | \theta = \theta_j)}\,. \tag{9}$$

Modernamente, esse resultado seria obtido *via* o Teorema de Bayes se assumimos que $P(\theta = \theta_1) = \cdots = P(\theta = \theta_n) = n^{-1}$, i.e. que as n diferentes causas são equiprováveis

antes de observar a ocorrência do evento $X = x$. A equação (9) explicaria então como atualizar as probabilidades das diferentes *causas* ($\theta = \theta_i$) **após** observarmos que $X = x$.

A extensão para o caso continuo é imediata. Segundo Laplace devemos considerar o parâmetro θ como sendo uma v.a. com densidade

$$p(\theta|X = x) = \frac{f_\theta(x)}{\int_\Theta f_\theta(x)\, d\theta}\,. \tag{10}$$

Veja que (i) a integral no denominador acima é com respeito à θ, de forma que somente a verossimilhança $f_\theta(x)$ para o x observado é necessária para calcular (10) e (ii) se duas verossimilhanças são proporcionais, a constante de proporcionalidade c vai cancelar no numerador e denominador de (10). Dessa forma, o paradigma da probabilidade inversa não viola o PV.

Resposta 1.3. Considere novamente o Exemplo 1. Independentemente do fato do experimento ter parado após $n = 12$ lançamentos ou $k = 3$ coroas, de acordo ao paradigma da probabilidade inversa devemos tratar θ como sendo uma v.a. com densidade

$$\pi(\theta) = \frac{\theta^9\,(1-\theta)^3}{\int_0^1 \theta^9\,(1-\theta)^3\, d\theta} = 2860\,\theta^9\,(1-\theta)^3 \tag{11}$$

para $0 < \theta < 1$ [o denominador acima pode ser calculado usando integração por partes reiteradas vezes — o que é bastante trabalhoso —, ou lembrando que a distribuição Beta com $\alpha = 10$ e $\beta = 4$ tem densidade igual a $\Gamma(14)\,\theta^9\,(1-\theta)^3/[\Gamma(10)\,\Gamma(4)] = 13!\,\theta^9\,(1-\theta)^3/[9!\,3!] = 2860\,\theta^9\,(1-\theta)^3$]. Em outras palavras, após observar 9 caras e 3 coroas, independente da regra de parada, o método da probabilidade inversa leva a tratar θ como uma v.a. com distribuição Beta($\alpha = 10$,$\beta = 4$) (veja a Figura 3; note que a densidade Beta(10,4) é simplesmente proporcional à quaisquer uma das duas verossimilhanças da Figura 1 (a), de forma que o único que muda é a escala no eixo das ordenadas).

Uma vez que a distribuição 11 foi obtida, parece natural basear a nossa decisão com respeito as hipóteses $H_0 : \theta \leq 0.5$ vs. $H_a : \theta > 0.5$ nas suas correspondentes probabilidades. Elas são respectivamente $P(\theta \leq 0.5|x) \doteq 0.046$ e $P(\theta > 0.5|x) \doteq 1 - 0.046 = 0.954$, de forma que podemos afirmar que após observar $x = 9$ caras e $k = 3$ coroas, **independente da regra de parada**, a hipótese alternativa é aproximadamente 20 ($\approx 0.954/0.046$) vezes mais provável que a hipótese nula

Observação. Acima, a notação $P(\cdot|x)$ foi usada para indicar simplesmente que a inferência depende somente do valor de X que foi efetivamente observado. As probabilidades das duas hipóteses acima podem ser calculadas digitando `pbeta(0.5,10,4)` e `1-pbeta(0.5,10,4)` numa sessão do **R**). □

É importante notar que o paradigma da probabilidade inversa permite outros tipos de inferência para θ de forma bastante natural. Por exemplo, se neste contexto fosse desejado um intervalo de confiança para θ, é razoável pensar por exemplo no intervalo determinado pelos percentis 2.5 e 97.5% da distribuição Beta($\alpha = 10$,$\beta = 4$), isto é o

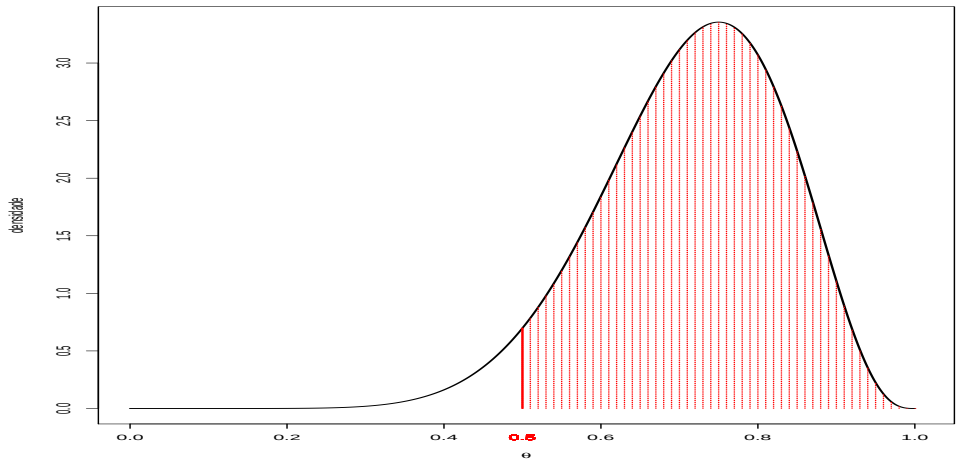


Figura 3: Densidade da distribuição Beta($\alpha = 10$,$\beta = 4$). A área riscada ($\doteq 0.954$) é a probabilidade de θ ser maior ou igual a 0.5.

intervalo $(0.462, 0.910)$ [no **R**, digite `qbeta(0.025,10,4);qbeta(0.975,10,4)`]. Alternativamente, um estimador pontual de θ poderia ser alguma medida de tendência central da distribuição (11), tal como a média $E(\theta|x) = \alpha/(\alpha+\beta) = 10/14 \doteq 0.714$ ou a mediana 0.725 [`qbeta(0.5,10,4)`)].

O método da probabilidade inversa não está livre de críticas. Por exemplo, quando o parâmetro θ pode assumir infinitos valores, nada garante em geral que a integral (ou soma, no caso discreto) no denominador de (10) será finita (é claro que $\int f_\theta(x)\,dx = 1$, mas isso não garante que $\int f_\theta(x)\,d\theta$ seja finita para o x observado). Possivelmente mais importante, o método não é invariante com respeito a reparametrizações, no sentido que ele pode levar a respostas diferentes quando aplicado ao parâmetro θ ou a uma transformação $\eta = g(\theta)$—veja o Exercício 5 da Lista 1).

6 Conclusões

Foi mostrado que não aceitar o PV implica negar também pelo menos o PC ou o PS, o que em exemplos muito simples levaria a agir de forma pouco intuitiva. Como as medidas usuais da inferência frequentista ou clássica não respeitam o PV, surge imediatamente a pergunta se seria possível formular algum paradigma que agisse respeitando o PV. Nesse sentido, na seção 5 descrevemos o princípio da probabilidade inversa de Laplace, que funciona em completo acordo com o PV.

Referências

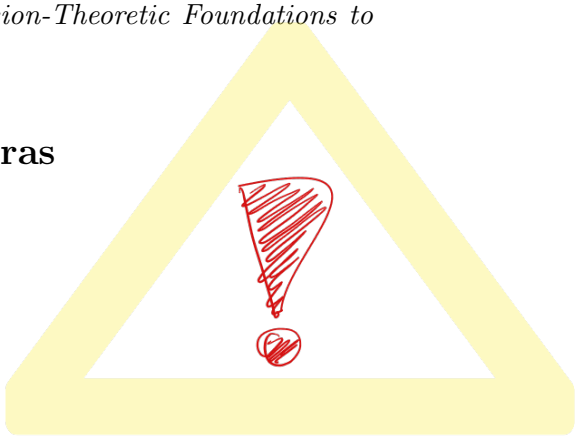
Berger, J. O. and R. O. Wolpert (1988). *The Likelihood Principle* (2nd Ed.). Haywood, CA: The Institute of Mathematical Statistics.

Birnbaum, A. (1962). On the foundations of statistical inference (with discussion). *Journal of the American Statistical Association 57*, 321–372.

Hald, A. (2007). *A History of Parametric Statistical Inference from Bernoulli to Fisher, 1713-1935*. New York: Springer.

Robert, C. P. (2007). *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation*. New York: Springer.

Apêndice: Script R para cálculos e figuras



```
# Exemplo 1: Dados:  S S S F S F S S S S S F

x<-9
n<-12
k<-3
theta.0<-0.5

# Inferencia Classica, delineamento (parada) binomial

dbinom(0:n,n,theta.0) # probabilidades sob H_0
sum(dbinom(x:n,n,theta.0)) # p-valor

# Infer?ncia Classica, delineamento (parada) binomial negativo

dnbinom(0:(2*x),k,theta.0) # probabilidades sob H_0
1-sum(dnbinom(0:(x-1),k,theta.0)) # p-valor

#     Fig. 1 (duas verossimilhanças para exemplo do PV)

par(mfrow=c(1,2))
theta<-seq(from=0.01,to=0.99,length=1000)

plot(theta,220*theta^9*(1-theta)^3,type='l',col='red',
     xlab=expression(theta),ylab=expression(P[theta]),lty=1,lwd=2)
lines(theta,55*theta^9*(1-theta)^3,col='blue',lty=2,lwd=2)
title(main='(a)')

theta<-seq(from=0.1,to=0.9,length=1000)
plot(theta,log(220)+9*log(theta)+3*log(1-theta),type='l',col='red',
     xlab=expression(theta),ylab=expression(paste('log',' ',P[theta])),lty=1,lwd=2)
lines(theta,log(55)+9*log(theta)+3*log(1-theta),col='blue',
      ylim=c(log(55)+9*log(0.1)+3*log(0.9),log(220)+9*log(0.9)+3*log(0.1)),lty=2,lwd=2)
title(main='(b)')
```

```
te<-c(0.2,0.4,0.6,0.8)
for(t in te){
  y.l<-log(55)+9*log(t)+3*log(1-t)
  lines(c(t,t),c(y.l,log(220)+y.l-log(55)))
}

#   Exemplo 2 (PC, misturas)

sigma.a<-0.1
sigma.b<-0.9
nivel<-0.95
z.nivel<-qnorm((1+nivel)/2)

x<-3.7

x-z.nivel*sigma.a;x+z.nivel*sigma.a   # IC, Resposta 2.1

# Calculo dos percentis da mistura

F.mistura<-function(z){0.5*pnorm(z,0,sigma.a)+0.5*pnorm(z,0,sigma.b)-(1+nivel)/2}
mist.upper<-uniroot(f=F.mistura,interval=c(0,3))$root

F.mistura<-function(z){0.5*pnorm(z,0,sigma.a)+0.5*pnorm(z,0,sigma.b)-(1-nivel)/2}
mist.lower<-uniroot(f=F.mistura,interval=c(-3,0))$root

mist.lower;mist.upper

#     Fig. 2 (mistura de va's para o PC)

par(mfrow=c(1,2))

mu<-3
z<-seq(from=-1.8,to=1.8,length=10000)
z<-mu+z
f.a<-dnorm(z,mu,sigma.a)
f.b<-dnorm(z,mu,sigma.b)
f.mist<-0.5*f.a+0.5*f.b

matplot(z,cbind(f.a,f.b,f.mist),type='l',
        ylab='f',xlab='z',xaxt='n',lwd=3,col=c('black','red','blue'))
mtext(expression(mu),side=1,line=1,at=mu)
title(main='(a)')

z<-mu-z
plot(z,f.mist,type='l',
```

```
      ylab='f',xlab=expression(paste('z','-',mu)),xaxt='n',lwd=2)
mtext('0',side=1,at=0,line=1)
title(main='(b)')

mtext(as.character(round(mist.lower,3)),side=1,line=1,at=mist.lower)
mtext(as.character(round(mist.upper,3)),side=1,line=1,at=mist.upper)

z.dots<-seq(from=mist.lower,to=mist.upper,length=100)
for(z in z.dots){
  lines(c(z,z),c(0,0.5*dnorm(z,0,0.1)+0.5*dnorm(z,0,1)),lty=1,lwd=0.3)
}

# Exemplo 1: inferencia sob probabilidade inversa

alfa<-x+1
beta<-n-x+1

1/beta(alfa,beta)  # calculo do reciproco da integral do denominador

1-pbeta(theta.0,alfa,beta)   #  probabilidade da hipótese alternativa
pbeta(theta.0,alfa,beta)     #  probabilidade da hipótese nula

(1-pbeta(theta.0,alfa,beta))/pbeta(theta.0,alfa,beta) # chances relativas

qbeta(0.025,alfa,beta); qbeta(0.975,alfa,beta)  # intervalo de credibilidade para theta, 95%

alfa/(alfa + beta)  # média da distribuição
qbeta(0.5,alfa,beta)   # mediana

#     Fig. 3 (dist. Beta(10,4)

par(mfrow=c(1,1))
theta<-seq(from=0,to=1,length=1000)

plot(theta,dbeta(theta,10,4),type='l',xlim=c(0,1.01),lwd=2,
     xlab=expression(theta),ylab='densidade')
lines(c(theta.0,theta.0),c(0,dbeta(theta.0,alfa,beta)),lwd=1,col='red')

for(t in seq(theta.0,1,0.01)) {
  lines(c(t,t),c(0,dbeta(t,alfa,beta)),lty=2,col='red',lwd=0.5)
mtext(as.character(theta.0),at=theta.0,side=1,line=1,col='red')
}
```