

Estimando a prevalência de uma doença a partir de um teste diagnóstico

Lucas Emanuel Resck Domingues
Lucas Machado Moschen
Victor Bitarães

Escola de Matemática Aplicada (EMAp)
Fundação Getúlio Vargas

29/04/2020

Introdução

Suponha que desejamos estimar a proporção $\theta \in (0, 1)$ de indivíduos infectados com um determinado patógeno em uma população. Suponha ainda que dispomos de um teste laboratorial, que produz o resultados $r = \{-, +\}$ indicando se o indivíduo (y_i) é livre (0) ou infectado (1). Se o teste fosse perfeito, poderíamos escrever a probabilidade de observar $y = \sum_{i=1}^n y_i$ testes positivos em n testes realizados como¹

$$\Pr(y \mid \theta, n) = \binom{n}{y} \theta^y (1 - \theta)^{n-y}. \quad (1)$$

Infelizmente, o teste não é perfeito, acertando o diagnóstico com probabilidades fixas da seguinte forma²

$$\Pr(r = + \mid y_i = 0) := 1 - u, \quad (2)$$

$$\Pr(r = - \mid y_i = 1) := 1 - v, \quad (3)$$

de modo que agora, assumindo $u + v > 1$, escrevemos³

$$\begin{aligned} \Pr(r = + \mid \theta, u, v) &= \Pr(r = + \mid y_i = 0) \Pr(y_i = 0) + \Pr(r = + \mid y_i = 1) \Pr(y_i = 1) \\ &= \Pr(r = + \mid y_i = 0) \Pr(y_i = 0) + (1 - \Pr(r = - \mid y_i = 1)) \Pr(y_i = 1) \\ &= (1 - u)(1 - \theta) + (1 - (1 - v))\theta \\ &= 1 - u + \theta(u + v - 1) \end{aligned} \quad (4)$$

e podemos reescrever a probabilidade em~(1):

$$\Pr(y \mid \theta, n, u, v) = \binom{n}{y} [1 - u + \theta(u + v - 1)]^y [u - \theta(u + v - 1)]^{n-y}. \quad (5)$$

¹Porquê?

²Naturalmente, $u, v \in (0, 1)$, levando em conta a restrição $u + v > 1$.

³Exercício bônus: mostre porquê.

Problemas

a) Escolha e justifique uma distribuição *a priori* para θ – lembre-se que neste exercício u e v são fixos;

Resposta:

Uma distribuição de probabilidade plausível para o parâmetro θ é dada pela Distribuição Beta. É uma distribuição suficientemente flexível, dado que pode assumir desde a forma da uniforme, até formato de uma normal a uma exponencial. Como estamos falando de uma priori, entretanto, precisamos escolher os parâmetros a e b . Então teremos:

$$\xi(\theta) = \frac{1}{B(a, b)} \theta^{a-1} (1 - \theta)^{b-1}$$

Segundo **este artigo**, os hiperparâmetros podem ser escolhidos em um painel de experts, usando resultados de estudos anteriores. Nesse caso, a ideia é aproximar a moda da distribuição para a prevalência acreditada por alguns epidemiologistas e, também, aproximar o desvio padrão como uma parte, um quarto, por exemplo, dos limites inferior e superior.

É importante ressaltar a importância da escolha dos hiperparâmetros, já que eles influenciam bastante o resultado, já que podem, por exemplo, impossibilitar a ocorrência de alguns prevalências.

Se considerarmos que não temos informação anterior à observação, escolher uma priori não informativa é algo bem interessante. No artigo de **Jeffreys** é construído um método baseado na informação de Fisher para encontrar essa priori. Intuitivamente, a ideia é minimizar o impacto da priori na nossa posteriori, e, nesse caso, a distribuição dela vai se aproximar à estimativa da verossimelhança máxima. Assim, $p(\theta) \propto \sqrt{\det I(\theta)}$. Como estamos em um caso unidimensional:

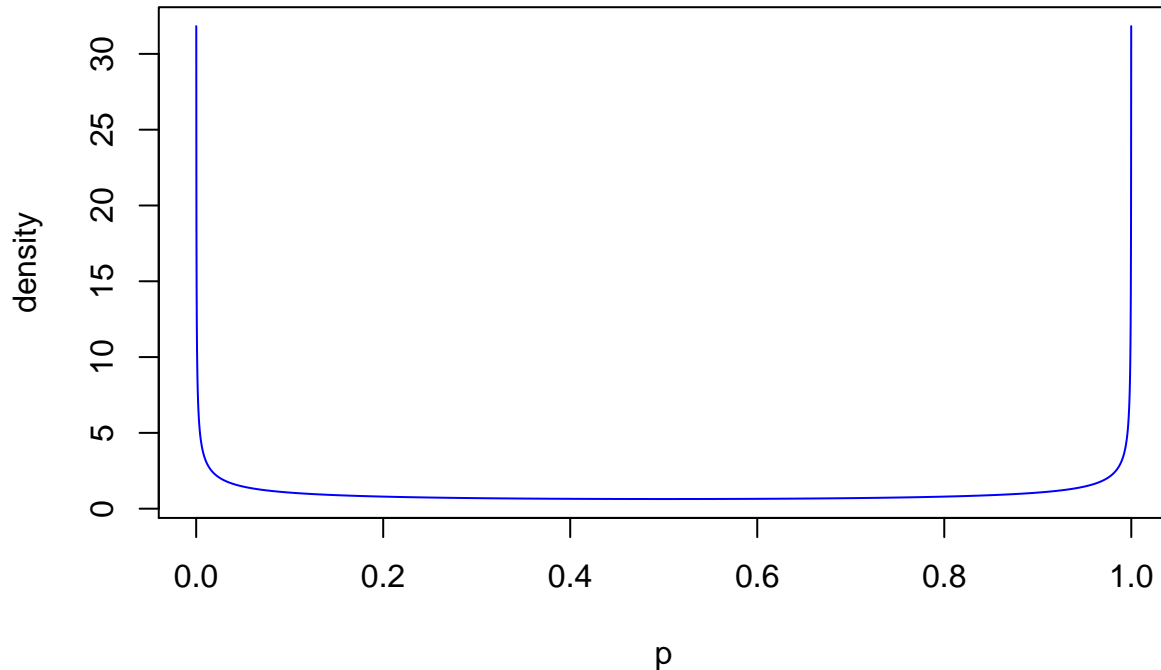
$$\begin{aligned} p(\theta) \propto (I(\theta))^{\frac{1}{2}} &= (E[(\frac{d}{d\theta} \log f(y|\theta, u, v))^2])^{\frac{1}{2}} \\ &= (E[(\frac{d}{d\theta} \log f(y|\theta, u, v))^2])^{\frac{1}{2}} \\ &= \left(E \left[\left(\frac{y(u+v-1)}{1-u+\theta(u+v-1)} + \frac{(y-n)(u+v-1)}{u-\theta(u+v-1)} \right)^2 \right] \right)^{\frac{1}{2}} \\ &= \left((u+v-1)^2 E \left[\left(\frac{y-n+un-n\theta(u+v-1)}{(1-u+\theta(u+v-1))(u-\theta(u+v-1))} \right)^2 \right] \right)^{\frac{1}{2}} \\ &= \left((u+v-1)^2 E \left[\left(\frac{y-n+un-n\theta(u+v-1)}{(1-u+\theta(u+v-1))(u-\theta(u+v-1))} \right)^2 \right] \right)^{\frac{1}{2}} \\ &= \frac{(u+v-1)}{(1-u+\theta(u+v-1))(u-\theta(u+v-1))} E[(y-n+un-n\theta(u+v-1))^2]^{\frac{1}{2}} \\ &= \frac{(u+v-1)}{(1-u+\theta(u+v-1))(u-\theta(u+v-1))} (Var[y] + E[y-n(1-u+\theta(u+v-1))]^2)^{\frac{1}{2}} \\ &= \frac{(u+v-1)(n(1-u+\theta(u+v-1))(u-\theta(u+v-1)))^{\frac{1}{2}}}{(1-u+\theta(u+v-1))(u-\theta(u+v-1))} \\ &= \frac{n^{\frac{1}{2}}(u+v-1)(1-u+\theta(u+v-1))^{\frac{1}{2}}(u-\theta(u+v-1))^{\frac{1}{2}}}{(1-u+\theta(u+v-1))(u-\theta(u+v-1))} \\ &= n^{\frac{1}{2}}(u+v-1)(1-u+\theta(u+v-1))^{-\frac{1}{2}}(u-\theta(u+v-1))^{-\frac{1}{2}} \end{aligned}$$

Assim $p(\theta) \propto f(1-u+\theta(u+v-1), \frac{1}{2}, \frac{1}{2}) = \frac{f(\theta, \frac{1}{2}, \frac{1}{2})}{u+v-1}$, onde f é a função densidade da distribuição Beta. Concluimos $p(\theta)$ tem distribuição *Beta*(0.5, 0.5).

Vejam os o formato de uma $Beta(0.5, 0.5)$

```
p = seq(from = 0, to = 1, length = 10000)
plot(p, dbeta(p, 0.5, 0.5), ylab="density", type="l", col=4,
     main = "Distribuição a Priori não informativa")
```

Distribuição a Priori não informativa



b) Derive $\Pr(\theta \mid y, n, u, v)$;

Resposta:

Para esse exercício, farei uma pequena simplificação. Sabemos que a Distribuição Beta é conjugada da verossimelhança binomial. Então o formato da Posteriori é conhecido e é, também, uma Distribuição Beta. Nesse caso, utilizarei uma priori para o parâmetro $\omega = 1 - u + \theta(u + v - 1)$. Nesse caso, a verossimelhança é de uma binomial com esse parâmetro e podemos usar $Beta(0.5, 0.5)$ para obter os cálculos da Posteriori. Sabemos, portanto, que $p(\omega \mid y, n, u, v)$ tem distribuição $Beta(0.5 + y, 0.5 + n - y)$

Por fim, $p(\theta \mid y, n, u, v) = \frac{p(\omega \mid y, n, u, v)}{u + v - 1}$

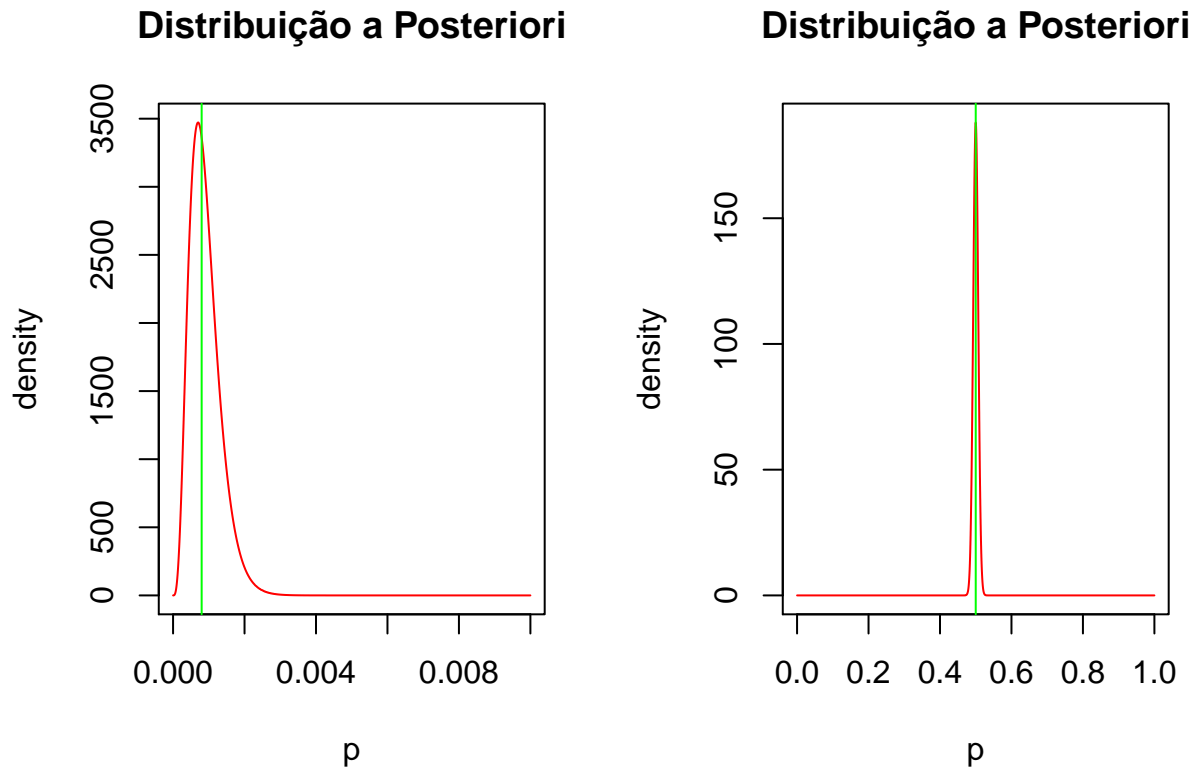
Considere esse exemplo:

```
u <- 0.6
v <- 0.7
y <- 4
n <- 5000

par(mfrow = c(1,2))
p <- seq(from = 0, to = 0.01, length = 10000)
plot(p, dbeta(p, 0.5 + y, 0.5 + n - y)/(u + v - 1), type = 'l', col = "red",
     main = "Distribuição a Posteriori", ylab = "density")
abline(v = y/n, col = "green")

y <- 2500
```

```
p <- seq(from = 0, to = 1, length = 10000)
plot(p, dbeta(p, 0.5 + y, 0.5 + n - y)/(u + v - 1), type = 'l', col = "red",
     main = "Distribuição a Posteriori", ylab = "density")
abline(v = y/n, col = "green")
```



Observe como é concentrada em torno da média observacional, indicada com a linha em verde.

- c) Suponha que $y = 4$ e $n = 5000$. Qual a média *a posteriori* de θ ? Produza intervalos de credibilidade de 80, 90 e 95% para θ .

Resposta:

$$E[\theta | Y = y] = \frac{1}{u + v - 1} \frac{0.5 + y}{0.5 + n - y} = \frac{4.5}{4996.5} \frac{1}{u + v - 1}$$

- d) **Bônus.** Que melhorias você faria neste modelo? Que outras fontes de incerteza estão sendo ignoradas?

Resposta: