Prova 2 - Inferência Estatística

Henrique Ennes

4/Set/2021

Regras:

- A prova vale 100 pontos distribuídos igualmente nas 5 questões.
- O aluno tem um prazo de 24h para solucionar a prova a mão e fazer upload da solução.
- A prova é individual e com consulta ao livro, notas de aula e outros documentos do curso.
- Dúvidas sobre a prova são feitas e respondidas pelo fórum da disciplina. Perguntas são feitas durante as primeiras 12h de prova, leia a prova cuidadosamente.
- O aluno que desejar entregar a prova digitada, terá um acréscimo de 24h para tal.
- Seja cuidadoso na entrega: questões cuja solução estiver desorganizada, difícil de ler ou compreender o que foi apresentado, receberão nota ZERO.
- Questões escritas a mão entregues após 24h, terão nota ZERO, sem direito a revisão.
- Os códigos computacionais das questões 4 e 5 devem ser uploaded separadamente em arquivo texto simples. Falha em observar esta regra implica em ZERO no item correspondente.
- Caso o código não replique o resultado apresentado, a nota será ZERO na questão inteira.
- O código pode ser uploaded para uma conta git e o link submetido como entrega, porém a data da última atualização deve ser condizente com a entrega da prova.
- Será aceita apenas uma submissão da solução após as primeiras 24h.

Introdução do problema:

Até então, nós sempre assumimos que a família de distribuições que gerou nossos dados é conhecida. Entretanto, na prática isso raramente é verdade e o que fazemos é assumirmos que o modelo escolhido é uma boa aproximação para nossos dados.

Dito isso, uma pergunta natural a se fazer é se os resultados que temos para o Estimador de Máxima Verossimilhança ainda são corretos e relevantes, mesmo no contexto de termos escolhido um modelo inicial errado. É isso que vamos descobrir nessa prova!

Hipóteses do problema:

1. Temos uma amostra i.i.d. $X_1, X_2, \ldots, X_n$ dada por uma função continua de densidade $g : \mathbb{R} \to \mathbb{R}$.

2. Tomamos como modelo base uma família paramétrica de funções de densidade $\mathcal{F} = \{f(\cdot, \theta) : \theta \in \Theta\}$, com $\Theta \subset \mathbb{R}$ um compacto e tal que $f(\cdot, \theta)$ é contínua para todo $\theta \in \Theta$. Note que **não** estamos assumindo que $g \in \mathcal{F}$.

3. A função da amostra $\widehat{\theta}_n$ dada por

$$\widehat{\theta}_n = \arg\max_{\theta \in \Theta} \left\{ \frac{1}{n} \sum_{i=1}^{n} \log f(X_i, \theta) \right\}$$

   existe e é mensurável, para todo $n \in \mathbb{N}$.

4. Temos que $\mathbb{E} \log g(X)$ existe e $|\log f(x, \theta)| \leq m(x)$ para todo $\theta \in \Theta$, onde m é integrável com respeito a densidade g, isto é,

$$\int m(x) g(x) dx < \infty.$$

5. A distância de Kullback-Leibler $D(g \| f(\cdot, \theta))$ dada por

$$D(g \| f(\cdot, \theta)) = \mathbb{E}_g \log \left(\frac{g(X)}{f(X, \theta)} \right)$$

   tem um único minimizador θ_* em Θ.

Exercício 1:

1. **(2 pontos)** O que podemos dizer sobre o limite abaixo com relação à sua convergência:
$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \log f(X_i, \theta).$$

2. **(2 pontos)** Mostre que
$$\mathbb{E}_g \log f(X, \theta) = \mathbb{E}_g \log g(X) - D(g \| f(\cdot, \theta)).$$

3. **(14 pontos)** Supondo que as hipóteses 1 até 5 são válidas, mostre que o estimador $\widehat{\theta}_n$ é consistente para θ^*.

4. **(2 pontos)** Caracterize a relação entre $f(\cdot, \theta_*)$ e g em termos da distância de Kullback-Leibler?

Solução (1)**.** Note que $\frac{1}{n}\sum_{i=1}^{n} \log f(X_i, \theta)$ é uma média amostral de $f(X_i, \theta)$ que, visto que são funções contínuas das variáveis i.i.d. $\{X_1, ..., X_n\}$ são, elas mesmas, variáveis i.i.d.. Assim, sendo o valor esperado de cada uma dessas i.i.d. $\mathbb{E} \log f(X_i, \theta)$, pela lei fraca dos grandes números, temos que $\sum_{i=1}^{n} \log f(X_i, \theta)$ converge em probabilidade para $\mathbb{E} \log f(X_i, \theta)$. Assim, o limite fica
$$\lim \frac{1}{n} \sum_{i=1}^{n} \log f(X_i, \theta) = \mathbb{E} \log f(X_i, \theta).$$

Solução (2)**.** Comece notando a existência de $\mathbb{E}_g \log f(X, \theta)$. Isso pois, na hipótese 4, assumimos que que $|\log f(X, \theta)| \leq m(x)$, onde $m(x)$ é integrável. Consequentemente, $\mathbb{E}_g |f(X, \theta)| < \infty$, assim, $\log f(X, \theta)$ é integrável. Agora, usando que $D(g \| f(\cdot, \theta)) = \mathbb{E}_g \log \left(\frac{g(X)}{f(X,\theta)} \right)$, tem-se por propriedades básicas de logaritmos, pela linearidade da esperança e a existência de $\mathbb{E}_g \log g(X)$, também garantida pela hipótese 4
$$D(g \| f(\cdot, \theta)) = \mathbb{E}_g \log \left(\frac{g(X)}{f(X, \theta)} \right) = \mathbb{E} \log g(X) - \mathbb{E} \log f(X),$$
o que claramente leva ao resultado sugerido, isto é
$$\mathbb{E}_g \log f(X, \theta) = \mathbb{E}_g \log g(X) - D(g \| f(\cdot, \theta)).$$

Solução (3)**.** Relembre que $\widehat{\theta}_n$ é consistente para θ_* se $\widehat{\theta}_n$ converge em probabilidade para θ_*, onde θ_* é o único minimizador de $D(g \| f(\cdot, \theta))$. Utilizando que $\widehat{\theta}_n = \arg\max_{\theta \in \Theta} \left\{ \frac{1}{n} \sum_{i=1}^{n} \log f(X_i, \theta) \right\}$ e assumindo que podemos comutar o limite com argmax (o que é aceitável, visto toda f contínua)
$$\lim_{n\to\infty} \widehat{\theta}_n = \lim_{n\to\infty} \arg\max_{\theta \in \Theta} \left\{ \frac{1}{n} \sum_{i=1}^{n} \log f(X_i, \theta) \right\} = \arg\max_{\theta \in \Theta} \lim_{n\to\infty} \left\{ \frac{1}{n} \sum_{i=1}^{n} \log f(X_i, \theta) \right\}.$$

Entretanto, por 1.1,

$$\lim_{n\to\infty}\left\{\frac{1}{n}\sum_{i=1}^{n}\log f(X_i,\theta)\right\} = \mathbb{E}\log f(X_i)$$

e por 1.2,

$$\mathbb{E}\log f(X_i) = \mathbb{E}_g \log g(X) - D(g\|f(\cdot,\theta)).$$

Assim,

$$\lim_{n\to\infty}\widehat{\theta}_n = \arg\max_{\theta\in\Theta}\mathbb{E}_g\log g(X) - D(g\|f(\cdot,\theta)).$$

Porém, g é independente do parâmetro θ, que controla apenas f. Assim, $\mathbb{E}_g \log g(X)$ é constante em θ e o valor de máximo de $\mathbb{E}_g \log g(X) - D(g\|f(\cdot,\theta))$ pode ser apenas atingindo variando-se $D(g\|f(\cdot,\theta))$. Porém, isso ocorre exatamente quando $D(g\|f(\cdot,\theta))$ é mínimo, isso é, quando $D(g\|f(\cdot,\theta)) = D(g\|f(\cdot,\theta_*))$. Ou seja, θ_* maximiza $\mathbb{E}_g \log g(X) - D(g\|f(\cdot,\theta))$ e $\lim_{n\to\infty}\widehat{\theta}_n = \theta_*$. Como todos os resultados aqui utilizados (1.1 e 1.2) estão corretos em probabilidade, segue que $\widehat{\theta}_n$ converge em probabilidade para θ_*, segue que $\widehat{\theta}_n$ é consistente.

Solução (4)**.** Como θ_* minimiza a distância de Kullback-Leiber, segue que $D(g\|f(\cdot,\theta))$, tem-se que

$$\mathbb{E}_g \log f(X,\theta_*) = \mathbb{E}_g \log g(X) - \min D(g\|f(\cdot,\theta)) = \max[\mathbb{E}_g \log g(X) - D(g\|f(\cdot,\theta))].$$

Exercício 2:

1. **(15 pontos)** Admitindo as hipóteses 1 até 5, além das hipóteses de regularidade usuais (ver hipóteses do teorema 9.14 do livro do Keener), mostre que
$$\sqrt{n}(\widehat{\theta}_n - \theta_*) \Rightarrow N\left(0, \frac{B}{A^2}\right),$$
onde
$$A = -\int \frac{d^2 \log f(x,\theta)}{d\theta^2}\Big|_{\theta=\theta_*} g(x)dx$$
e
$$B = \int \left(\frac{d \log f(x,\theta)}{d\theta}\Big|_{\theta=\theta_*}\right)^2 g(x)dx$$

2. **(5 pontos)** Suponha que $g \in \mathcal{F}$, isto é, que o modelo que escolhemos é correto. O que acontece com aos valores A, B?

Solução (1)**.** Vamos assumir que 1. θ_* é interior ao espaço de parâmetros; 2. as duas primeiras derivadas de $\log f(X_i, \theta)$ são contínuas em torno de θ_*; 3. $\hat{\theta}_n$ é consistente para θ_*; 4. $\mathbb{E}_g \frac{d \log f(X_i,\theta)}{d\theta}\Big|_{\theta=\theta_*} = 0$[1]; 5. $\mathbb{E}_g \frac{d^2}{d\theta^2} \log f(X_i, \theta)$ existe para todo θ próximo a θ_*[2].

Vamos começar deixando $\overline{W}_n(\theta)$ ser $\overline{W}_n(\theta) = \frac{1}{n}\sum_{i=1}^n \log f(X_i, \theta)$. Notamos que, por Taylor
$$\overline{W}_n'(\hat{\theta}_n) = \overline{W}_n'(\theta_*) + (\hat{\theta}_n - \theta_*)\overline{W}_n''(\tilde{\theta}_n),$$
onde $\tilde{\theta}_n$ é um valor entre $\hat{\theta}_n$ e θ_*. Além disso, como $\hat{\theta}_n$ maximiza $\frac{1}{n}\sum_{i=1}^n \log f(X_i, \theta)$
$$\overline{W_n}'(\hat{\theta_n}) = \frac{d}{d\theta}\frac{1}{n}\sum_{i=1}^n \log f(X_i,\theta)\Big|_{\theta=\theta_*} = 0$$
e então
$$\sqrt{n}(\hat{\theta}_n - \theta_*) = -\sqrt{n}\frac{\overline{W}_n'(\theta_*)}{\overline{W}_n''(\tilde{\theta_n})}. \tag{1}$$

Pelo Teorema Central do Limite, como $\overline{W}_n'$ são i.i.d. para o parâmetro θ_*,

[1]Note que apesar de parecer uma hipótese absurda a primeira vista, 4 faz sentido de se esperar, isso pois $\hat{\theta}_n$ converge para θ_* e temos, para todo n, $\frac{d \log f(X_i,\theta)}{d\theta}\Big|_{\theta=\hat{\theta}_n} = 0$ pela condição de maximização de θ_n. Além disso, o livro assume essa condição.

[2]O livro, ao demonstrar o Teorema 9.14, assume uma hipótese mais fraca, isto é, que para todo θ interior ao espaço de parâmetros, $\|I_{[\theta-\epsilon,\theta+\epsilon]}\overline{W}''\|_\infty < \infty$. Entretanto, essa generalização do resultado aqui a ser mostrado não traz nenhuma intuição nova ao problema apenas alguns cuidados com tecnicidades que pouco importam para nós, visto que em todos os casos em que precisaremos deste resultado, nossa hipótese mais fraca é suficiente.

$$
\begin{aligned}
\sqrt{n}\left(\overline{W}_n'(\theta_*) - \mathbb{E}_g W_n'(\theta_*)\right) &= \sqrt{n}\left(\overline{W}_n'(\theta_*) - \mathbb{E}_g \frac{d}{d\theta}\log f(X_i,\theta)\Big|_{\theta=\theta_*}\right) \\
&= \sqrt{n}\left(\overline{W}_n'(\theta_*)\right) \\
&\implies N\left(0, \mathrm{Var}\frac{d}{d\theta}\log f(X_i,\theta)\Big|_{\theta=\theta_*}\right) \\
&= N\Bigg(0, \mathbb{E}\Bigg(\frac{d}{d\theta}\log f(X_i,\theta)\Big|_{\theta=\theta_*} \\
&\quad - \mathbb{E}_g \frac{d}{d\theta}\log f(X_i,\theta)\Big|_{\theta=\theta_*}\Bigg)^2\Bigg) \\
&= N\left(0, \mathbb{E}\left(\frac{d}{d\theta}\log f(X_i,\theta\Big|_{\theta=\theta_*}\right)^2\right) \\
&= N\left(0, \int \left(\frac{d\log f(x,\theta)}{d\theta}\Big|_{\theta=\theta_*}\right)^2 g(x)dx\right) \\
&= N(0,B).
\end{aligned}
$$

Além disso, como $\hat{\theta}_n$ é consistente para θ_* e $\tilde{\theta}_n$ está entre $\hat{\theta}_n$ e θ_*, tem-se que como $W_n''(\hat{\theta}_n) \to W_n''(\theta_*)$ em distribuição, pois W_n'' é assumido contínua perto de θ_*, $W_n''(\tilde{\theta}_n) \to W_n''(\theta_*)$ também em distribuição e, assim

$$
\mathbb{E}_g W_n''(\tilde{\theta}_n) \to \mathbb{E}_g W_n''(\theta_*) = \int_g g(x)\frac{d^2\log f(x,\theta)}{d\theta^2}\Big|_{\theta=\theta_*} g(x)dx = -A.
$$

Assim, utilizando essas duas convergências em (1)

$$
\sqrt{n}(\hat{\theta}_n - \theta_*) = -\sqrt{n}\frac{\overline{W}_n'(\theta_*)}{\overline{W}_n''(\tilde{\theta_n})} \implies -\frac{N(0,B)}{-A} = N\left(0, \frac{B}{A^2}\right).
$$

Solução (2)**.** Suponha $g_\theta(X) = f(X,\theta)$ e os dados são gerados por $g_{\theta_0}(X) = g(X)$ para algum θ_0 no espaço de parâmetros. Então, temos que

$$
D(g, f(\cdot,\theta)) = \mathbb{E}_{\theta_0}\log\frac{g(X,\theta_0)}{f(X,\theta)},
$$

o que, de acordo com o Lema 9.8 (página 156), é positivo para todo parâmetro $\theta \neq \theta_0$ e, como para $\theta = \theta_0$, $\log\frac{g(x,\theta_0)}{f(x,\theta_0)} = 0$ em todo x, $D = 0$ se $\theta = \theta_0$. Isso

implica que $\theta_0 = \theta_*$. Em particular,

$$\begin{aligned} A &= -\int \frac{d^2 \log f(x,\theta)}{d\theta^2}\Big|_{\theta=\theta_*} g(x)dx \\ &= -\int \frac{d^2 \log g_\theta(x)}{d\theta^2}\Big|_{\theta=\theta_0} g(x)dx \\ &= -\mathbb{E}_{\theta_0}\left(\frac{d^2 \log g_\theta(x)}{d\theta^2}\right) \\ &= I(\theta_0), \end{aligned}$$

provido g com regularidade suficiente. De modo similar

$$\begin{aligned} B &= \int \left(\frac{d \log f(x,\theta)}{d\theta}\Big|_{\theta=\theta_*}\right)^2 g(x)dx \\ &= \int \left(\frac{d \log g_\theta(x)}{d\theta}\Big|_{\theta=\theta_0}\right)^2 g(x)dx \\ &= \mathbb{E}_{\theta_0}\left(\frac{d \log g_\theta(x)}{d\theta}\right)^2 \\ &= I(\theta_0). \end{aligned}$$

Assim, se testarmos exatamente $g = f$, temos que $A = B$. Na verdade, esse resultado é ainda mais fundamental. Isso pois, dado, $g = f$, temos que $\hat{\theta}_n$ é o argumento que maximiza $\frac{1}{n}\sum_{i=1}^n \log g(X_i, \theta)$, ou seja, como multiplicação por constantes não interfere no próximo de maximização, temos que $\hat{\theta}_n$ é o maximizador de verossimilhança de θ_0 e assim, pelo Teorema 9.14 e ao relembrar que $\theta_* = \theta_0$

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \implies N(0, I(\theta_0)^{-1}),$$

o que só é condizente com o resultado da 2.1 se $\frac{B}{A^2} = I(\theta_0)^{-1}$, o que é realmente o caso dado $A = B = I(\theta_0)$.

Exercício 3:

Considere os estimadores $\widehat{A}$ e $\widehat{B}$ dos valores A e B do item anterior, dados por

$$\widehat{A}_n = -\frac{1}{n}\sum_{i=1}^{n} \frac{d^2 \log f(X_i,\theta)}{d\theta^2}\Big|_{\theta=\theta_*}$$

e

$$\widehat{B}_n = \frac{1}{n}\sum_{i=1}^{n} \left(\frac{d \log f(X_i,\theta)}{d\theta}\Big|_{\theta=\theta_*}\right)^2.$$

1. **(5 pontos)** O que podemos dizer sobre o limite em probabilidade de

$$\frac{\widehat{B_n}}{\widehat{A_n}^2}.$$

2. **(10 pontos)** Suponha que admitimos um modelo exponencial para os dados, com parâmetro $\lambda > 0$. Calcule o estimador

$$\frac{\widehat{B_n}}{\widehat{A_n}^2}$$

   nesse caso.

3. **(5 pontos)** O que podemos dizer sobre a convergência em distribuição do estimador de MLE no item anterior? Compare com o caso em que sabemos que a distribuição real dos dados é de fato exponencial.

Solução (1). Note que $\frac{d^2 \log f(X_i,\theta)}{d\theta^2}\Big|_{\theta=\theta_*}$ e $\left(\frac{d \log f(X_i,\theta)}{d\theta}\Big|_{\theta=\theta_*}\right)^2$ são i.i.d. em X_i e, como $\hat{A}_n$ e $\hat{B}_n$ são médias amostrais dessas i.i.d., segue pela Lei Fraca que $\widehat{A}_n \to \mathbb{E}_g \frac{d^2 \log f(X_i,\theta)}{d\theta^2}\Big|_{\theta=\theta_*} = A$ e $\widehat{B}_n \to \mathbb{E}_g \left(\frac{d \log f(X_i,\theta)}{d\theta}\Big|_{\theta=\theta_*}\right)^2 = B$ em probabilidades, dado essas frações e o limite existem (i.e. $\widehat{B}_n$ e seu limite são não nulos). Então, por Slutsky, em distribuição

$$\frac{\widehat{B_n}}{\widehat{A_n}^2} \implies \frac{\mathbb{E}_g \left(\frac{d \log f(X_i,\theta)}{d\theta}\Big|_{\theta=\theta_*}\right)^2}{\left(\mathbb{E}_g \frac{d^2 \log f(X_i,\theta)}{d\theta^2}\Big|_{\theta=\theta_*}\right)^2} = \frac{B}{A^2}.$$

Porém, como ambos os termos são constantes (i.e. não são funções em X_i) a convergência em distribuição implica convergência em probabilidade.

Em especial, se $g = f$ como discutido no Problema 2.2, essa convergência em probabilidade é para $I(\theta_*)^{-1}$.

Solução (2). Comece notando que, para $f(x_i, \lambda) = \lambda \exp(-\lambda x_i)$, temos

$$\begin{aligned}
\log f(x_i, \lambda) &= \log \lambda - \lambda x_i \\
\frac{d \log f(x_i, \lambda)}{d\lambda} &= \frac{1}{\lambda} - x_i \\
\frac{d^2 \log f(x_i, \lambda)}{d\lambda^2} &= -\frac{1}{\lambda^2}.
\end{aligned}$$

Assim,

$$\widehat{A_n} = -\frac{1}{n} \sum_{i=1}^{n} \frac{d^2 \log f(x_i, \lambda)}{d\lambda^2}\Big|_{\lambda=\lambda_*} = -\frac{1}{n} \sum_{i=1}^{n} -\frac{1}{\lambda_*^2} = \frac{1}{\lambda_*^2}$$

e

$$\begin{aligned}
\widehat{B_n} &= \frac{1}{n} \sum_{i=1}^{n} \left(\frac{d \log f(X_i, \theta)}{d\theta}\Big|_{\theta=\theta_*} \right)^2 \\
&= \frac{1}{n} \sum_{i=1}^{n} \left(\frac{1}{\lambda_*} - x_i \right)^2 \\
&= \frac{1}{n} \sum_{i=1}^{n} \left(\frac{1}{\lambda_*^2} - 2\frac{x_i}{\lambda_*} + x_i^2 \right) \\
&= \frac{1}{\lambda_*^2} - 2\frac{\overline{X}}{\lambda_*} + \overline{X^2}.
\end{aligned}$$

Finalmente,

$$\frac{\widehat{B_n}}{\widehat{A_n}^2} = \lambda_*^2 - 2\overline{X}\lambda_*^3 + \overline{X^2}\lambda_*^4.$$

Solução (3). Em geral, pela Lei Fraca, $\overline{X} \to \mathbb{E}_g X_i$ e $\overline{X^2} \to \mathbb{E}(X_i^2)$ em probabilidade, assim

$$\frac{\widehat{B_n}}{\widehat{A_n}^2} \to \lambda_*^2 - 2\mathbb{E}_g(X_i)\lambda_*^3 + \mathbb{E}_g(X_i^2)\lambda^4.$$

Em particular, se $g = f$, como vimos, $\mathbb{E}_g(X_i) = \mathbb{E}_{\theta_*} X_i = \frac{1}{\theta_*}$ e $\mathbb{E}_g(X_i^2) = \mathbb{E}_{\theta_*} X_i^2 = \frac{2}{\theta_*^2}$. Assim

$$\frac{\widehat{B_n}}{\widehat{A_n}^2} \to \lambda_*^2 - 2\lambda_*^2 + 2\lambda^2 = \lambda_*^2 = \left(-\mathbb{E}_{\lambda_*} \frac{d^2 \log g}{d\lambda^2} \right)^{-1} = I(\theta_*)^{-1}.$$

De fato, isso condiz com o previsto no Problema 2.2. Relembrando que $\hat{\theta}_n$ é um estimador de máxima verossimilhança, se $f = g$, como vimos, em distribuição

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \implies N(0, I(\theta_*)^{-1}),$$

o que só é possível, quando comparamos com

$$\sqrt{n}(\widehat{\theta}_n - \theta_*) \Rightarrow N\left(0, \frac{B}{A^2}\right)$$

e com $\frac{\widehat{B}_n}{\widehat{A}_n^2} \to \frac{A}{B^2}$ se $\frac{\widehat{B_n}}{\widehat{A_n}^2} \to I(\theta_*)^{-1}$.

Exercício 4:

Suponha que queremos testar a hipótese $H_0 : s(\theta_*) = 0$ versus $H_1 : s(\theta) \neq 0$, onde $s : \mathbb{R} \to \mathbb{R}$ é continuamente diferenciável em toda parte.

1. **(10 pontos)** Desenvolva o teste de Wald de forma apropriada no contexto em que estamos, isto é, que o modelo escolhido pode não ser o correto.
2. **(10 pontos)** Suponha que a distribuição verdadeira dos dados é normal padrão truncada nos positivos e que o modelo exponencial da questão anterior é utilizado. Desejamos realizar o teste para $s(\theta) = \theta - \theta^*$.
   - desenvolva um experimento Monte Carlo (5000 replicações) que permita avaliar o erro tipo I do teste, a nível 1%, 5% e 10%, em amostras de tamanho 10, 50, 100, 200 e 500.
   - desenvolva um experimento Monte Carlo que permita avaliar o erro tipo II do teste. Para simplificar, vamos fazer $H_0 : \theta - 1.1\,\theta^* = 0$.
   - Produza uma tabela apresentando os resultados e os desvios padrão das estimativas.
   - Uma das entregas é o código para replicação do experimento. Ele deve gerar as tabelas e gráficos e será executado. O código deve ser entregue em formato texto simples (.R, .py, .c, .f, .jl, ...) e comentado, não é aceito um notebook ou outros formatos equivalentes.

Solução (1)**.** Relembre que pelo Problema 2.1,

$$\sqrt{n}(\widehat{\theta}_n - \theta_*) \Rightarrow N\left(0, \frac{B}{A^2}\right)$$

Vamos aqui assumir s uma função real-real, tal que $s(\theta) = 0$ se e somente se $\theta \in \Omega_0$, onde Ω_0 é a hipótese H_0. Utilizando-se do método delta, se assumirmos $\theta_* \in \Omega_0$

$$\sqrt{n}(s(\widehat{\theta}_n) - s(\theta_*)) = \sqrt{n}(s(\widehat{\theta}_n)) \Rightarrow N\left(0, \frac{B}{A^2}(s'(\theta_*))^2\right).$$

Pelo mesmo argumento usado pelo livro ao derivar o teste Wald (página 361), isto é, que a convergência de $\hat{\theta}_n$ para θ_* continua válida ao escrevermos a fórmula abaixo

$$\frac{ns(\widehat{\theta}_n)^2}{\frac{B}{A^2}(s'(\theta_*))^2} \implies \chi^2_1,$$

pois s reduz o espaço parametral de 1 para 0 graus de liberdade, sendo assim, χ^2_1 é uma chi-quadrada com um grau de liberdade.

Note porém que esse resultado sofre de uma deficiência. Em geral, não sabemos g, tornando impossível calcular A e B diretamente. Entretanto, usando

o que vimos na questão 3.1, isto é[3], $\frac{\widehat{B_n}}{\widehat{A_n}^2} \implies \frac{B}{A^2}$

$$\frac{ns(\widehat{\theta}_n)^2}{\frac{\hat{B}}{\hat{A}^2}(s'(\theta_*))^2} \implies \frac{ns(\widehat{\theta}_n)^2}{\frac{B}{A^2}(s'(\theta_*))^2} \implies \chi^2_1.$$

Ainda assim, continuamos com uma dependência em g por meio da necessidade de encontrar θ_*, tanto para avaliar as derivadas presentes em s, A e B. Usando novamente que $\hat{\theta}_n \implies \theta_*$, podemos, pelo mesmo argumento,

$$\frac{ns(\widehat{\theta}_n)^2}{\frac{\hat{B}(\hat{\theta}_n)}{\hat{A}^2(\hat{\theta}_n)}(s'(\hat{\theta}_n))^2} \implies \frac{ns(\widehat{\theta}_n)^2}{\frac{\hat{B}}{\hat{A}^2}(s'(\theta_*))^2} \implies \chi^2_1,$$

onde definimos

$$\widehat{A}_n(\hat{\theta}_n) = -\frac{1}{n}\sum_{i=1}^{n} \frac{d^2 \log f(X_i, \theta)}{d\theta^2}\Big|_{\theta=\hat{\theta}_n}$$

e

$$\widehat{B}_n(\hat{\theta}_n) = \frac{1}{n}\sum_{i=1}^{n} \left(\frac{d \log f(X_i, \theta)}{d\theta}\Big|_{\theta=\hat{\theta}_n}\right)^2,$$

isto é

$$\frac{\widehat{B}_n(\hat{\theta}_n)}{\widehat{A}_n^2(\hat{\theta}_n)} = \hat{\theta}_n^2 - 2\overline{X}\hat{\theta}_n^3 + \overline{X^2}\hat{\theta}_n^4$$

Solução (2)**.** Usamos para esse caso $s(\theta) = \theta - \theta_0$ para qualquer θ_0 interior ao espaço de parâmetros. Assim, $s'(\theta_*) = 1$ e, para achar o valor teórico de θ_*, minimizaremos $D(g\|f(\cdot,\theta))$, onde $g(x) = \frac{2}{\sqrt{2\pi}}\exp(-x^2/2)$ e $f(x,\theta)) = \theta\exp(-\theta x)$. Assim

$$\begin{aligned}\mathbb{E}_g \log\left(\frac{g(X)}{f(X,\theta)}\right) &= \mathbb{E}_g \log\left(\frac{2}{\sqrt{2\pi}}\right) - \mathbb{E}_g\frac{X^2}{2} - \mathbb{E}_g \log\theta + \theta\mathbb{E}_g X \\ &= \frac{2}{\sqrt{2\pi}} - \frac{1}{2} - \log\theta + \theta\sqrt{\frac{2}{\pi}}.\end{aligned}$$

Para minimizar, igualamos a derivada da expressão acima a zero, o que leva a

$$\theta_* = \sqrt{\frac{\pi}{2}} \approx 1.25.$$

De fato, podemos ver na Figura 1 que o $\hat{\theta}_n$ não só converge para esse valor de θ_*, mas também que a aproximação desses dois valores melhora a medida que n cresce, assim como descrito no Problema 2.1.

[3]Note, novamente, que esse argumento provavelmente merece uma formalização maior, mas assumindo (como estamos) que estamos trabalhando apenas com funções contínuas, nossa nível de perigo não é assim tão grande, afinal. De fato, como veremos na próxima questão, os resultados aqui mostrados possuem boas implicações empíricas então, sejamos, um pouco utilitaristas pelo bem da Matemática.

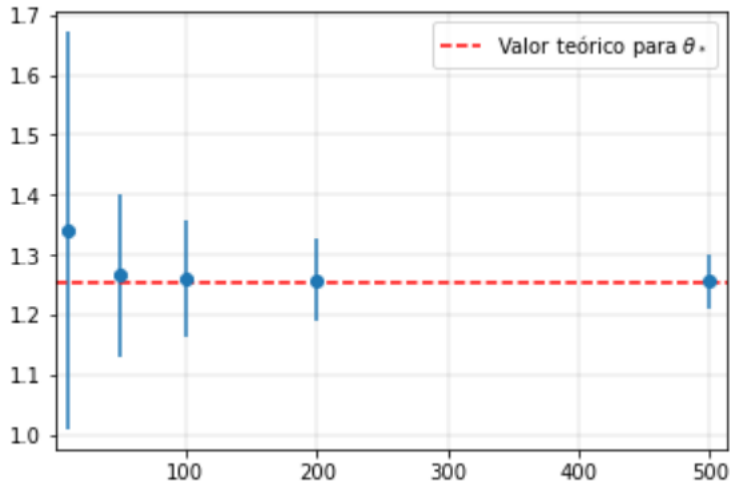

Figura 1: Valores de $\hat{\theta}_n$ para distintos tamanhos de amostragem n, comparado com o valor teórico esperado de $\theta_* = \sqrt{\frac{\pi}{2}}$.

Nosso teste utilizando o valor teórico fica

$$T_w = \frac{ns(\theta_n)^2}{\frac{\hat{B}}{\hat{A}^2}}$$

e, para um nível α, rejeitamos H_0 se e somente se $T_w > q$, onde q é o quantil $1-\alpha$ da distribuição χ^2_1. Assim, podemos estimar o Erro Tipo I do teste, para diferentes valores de n, ao calcular a probabilidade de, dado $\theta_0 = \theta_*$, rejeitar H_0 como simplesmente, a probabilidade de T_w ser maior que q. De modo similar, calculamos o Erro Tipo II ao assumir $\theta_0 \neq \theta_*$ (em particular, assumiremos $\theta_0 = 1.1 \times \theta_*$), como a probabilidade de T_w ser menor que q. Na Figura 2, vemos quais são as médias de Erros para valores de n entre 10, 50, 100, 200 e 500, com testes de nível $\alpha = 10\%$, 5% e 1%.

Entretanto, como mencionado, o cálculo de $\frac{\hat{B}}{\hat{A}^2}$ utilizado para gerar a tabela da Figura 2 requer conhecimento prévio de g, o que nem sempre é possível. Para isso, vamos utilizar os valores empíricos de $\hat{A}_n(\hat{\theta}_n)$ e $\hat{B}_n(\hat{\theta}_n)$, com as mesmas quantidades de θ_n obtidas ao criar a Figura 1. Note que os resultados continuam bem parecidos com o caso em que usamos o minimizador teórico.

Tamanho do Dataset	10%	5%	1%
10	0.1514	0.1264	0.1004
50	0.1012	0.0680	0.0478
100	0.1058	0.0560	0.0330
200	0.1032	0.0590	0.0300
500	0.1060	0.0556	0.0264

Tamanho do Dataset	10%	5%	1%
10	0.8902	0.9234	0.9468
50	0.7858	0.8858	0.9532
100	0.6256	0.7696	0.8564
200	0.4080	0.5510	0.6530
500	0.1062	0.1622	0.2290

Figura 2: Médias de Erros Tipo I (esquerda) e Tipo II (direita) utilizando-se o valor teórico de θ_*. Note que, como esperado, diminuir o Erro Tipo I implica em aumento no Erro Tipo II para um n fixo.

Tamanho do Dataset	10%	5%	1%
10	0.1076	0.0394	0.0146
50	0.1132	0.0682	0.0378
100	0.1002	0.0602	0.0354
200	0.1084	0.0560	0.0260
500	0.0970	0.0452	0.0228

Tamanho do Dataset	10%	5%	1%
10	0.7314	0.8178	0.9002
50	0.6690	0.7554	0.8280
100	0.5762	0.6804	0.7544
200	0.3862	0.5092	0.6194
500	0.1178	0.1822	0.2576

Figura 3: Médias de Erros Tipo I (esquerda) e Tipo II (direita) utilizando-se somente valores teórico de θ_*, calculados como médias de $\hat{\theta}_n$.

Exercício 5:

Vimos que a Identidade da Matriz de Informação deixa de ser válida sob erro de especificação do modelo, isto é, se $g \notin \mathcal{F}$. Esta observação permite criação de mecanismos para verificar se o modelo está corretamente especificado.

1. **(5 pontos)** Proponha uma função $s(\theta)$ para o teste de Wald desenvolvido anteriormente que permita testar se o modelo está corretamente especificado ou não.

2. **(5 pontos)** explique como podemos usar este teste para construir um intervalo de confiança para $s(\theta)$ e como podemos testar a hipótese.

3. **(10 pontos)**

    - Gere 30 observações de uma distribuição gamma com parâmetros $\alpha = 1.1$ e $\beta = .5$:

$$g(x; \alpha, \beta) = \frac{\beta^\alpha x^{\alpha-1} e^{-\beta x}}{\Gamma(\alpha)} \quad \text{para } x > 0 \quad \alpha, \beta > 0.$$

    - Suponha que usamos o modelo exponencial com parâmetro λ para os dados.
    - Construa um intervalo bootstrap para $s(\lambda)$ com 5000 amostras bootstrap e determine se rejeitamos ou não a hipótese nula do modelo estar corretamente especificado ao nível de 5%.
    - Explique o algoritmo utilizado.
    - Uma das entregas é o código para replicação do experimento que será executado. O código deve ser entregue em formato texto simples (.R, .py, .c, .f, .jl, ...) e comentado, não é aceito um notebook ou outros formatos equivalentes.

Solução (1). Como vimos na Questão 2.2, se $g(X) = f(X, \theta_0)$, então $A = B = I(\theta_0)$. Vamos usar esse fato como uma medida de igualdade entre g e f, isto é, deixaremos $s(\theta) = A - B + \theta - \theta_0$. Assim, $s(\theta_*) = 0$ se $A = B$ e $\theta_* = \theta_0$, ou seja, somente se acertamos o modelo e o parâmetro θ_0. Assim, assumindo que podemos comutar a derivada em θ com as integrais

$$\begin{aligned}
s'(\theta) &= \frac{dA}{d\theta} - \frac{dB}{d\theta} + \frac{d(\theta - \theta_0)}{d\theta} \\
&= -\int \frac{d^3 \log f(x,\theta)}{d\theta^3}\Big|_{\theta=\theta} g(x)dx - \int \frac{d}{d\theta}\left(\frac{d \log f(x,\theta)}{d\theta}\right)^2\Big|_{\theta=\theta} g(x)dx + 1 \\
&= -\int \left(\frac{d^3 \log f(x,\theta)}{d\theta^3} + \frac{d}{d\theta}\left(\frac{d \log f(x,\theta)}{d\theta}\right)^2\right)\Big|_{\theta=\theta} g(x)dx + 1 \\
&= -\int \left(\frac{d^3 \log f(x,\theta)}{d\theta^3} + 2\frac{d \log f(x,\theta)}{d\theta}\frac{d^2 \log f(x,\theta)}{d\theta^2}\right)\Big|_{\theta=\theta} g(x)dx + 1,
\end{aligned}$$

de modo que o teste de Wald fica

$$\frac{n\left(-\int\left(\frac{d^2\log f(x,\theta)}{d\theta^2}+\left(\frac{d\log f(x,\theta)}{d\theta}\right)^2\right)\Big|_{\theta=\theta_*}g(x)dx+\hat{\theta}_n-\theta_0\right)^2}{\frac{\hat{B}}{\hat{A}^2}\left[-\int\left(\frac{d^3\log f(x,\theta)}{d\theta^3}+2\frac{d\log f(x,\theta)}{d\theta}\frac{d^2\log f(x,\theta)}{d\theta^2}\right)\Big|_{\theta=\theta_*}g(x)dx+1\right]^2}\implies\chi_1^2.$$

Novamente, porém, requeremos novamente saber a distribuição g de antemão. Para uma versão totalmente empírica

$$T_w=\frac{n\left(-\frac{1}{n}\sum_{i=1}^n\left(\frac{d^2\log f(x,\theta)}{d\theta^2}+\left(\frac{d\log f(x,\theta)}{d\theta}\right)^2\right)\Big|_{\theta=\hat{\theta}_n}+\hat{\theta}_n-\theta_0\right)^2}{\frac{\hat{B}(\hat{\theta}_n)}{\hat{A}^2(\hat{\theta}_n)}\left[-\frac{1}{n}\sum_{i=1}^n\left(\frac{d^3\log f(x,\theta)}{d\theta^3}+2\frac{d\log f(x,\theta)}{d\theta}\frac{d^2\log f(x,\theta)}{d\theta^2}\right)\Big|_{\theta=\hat{\theta}_n}+1\right]^2}\implies\chi_1^2.$$

Solução (2)**.** Com o teste Wald T_w descrito no problema anterior, podemos criar um intervalo de confiança (assimptótico) para $s(\theta_*)=0$, isto é, para $f=g$. Isso pois, se quisermos construir uma região S de parâmetros θ_0 em que a probabilidade de $f=g$ é $1-\alpha$, apenas utilizamos o teste em que se rejeita H_0 se $T_w>q$, onde q é o $1-\alpha$ quantil de χ_1^2, pois, dada a convergência de T_w para χ_1^2, T_w terá assim probabilidade $1-\alpha$ de ser maior que q. Assim, deixamos $S=\{\theta_0|T_w(\theta_0)\leq q\}$, isto é, a região em que aceitamos $H_0: A=B, \theta_*=\theta_0$.

Solução (3)**.** Começamos novamente notando que o valor teórico de θ_* pode ser achado ao minimizarmos $D(g\|f(\cdot,\theta))$, onde

$$\begin{aligned}D(g\|f(\cdot,\theta))&=\mathbb{E}_g\log\frac{g(x)}{f(x\theta)}\\&=\mathbb{E}_g\log\frac{\beta^\alpha x^{\alpha-1}e^{-\beta x}}{\Gamma(\alpha)\theta\exp(-x\theta)}\\&=\alpha\log\beta+(\alpha-1)\mathbb{E}_g\log X-\beta\mathbb{E}_gX-\Gamma(\alpha)-\log\theta+\theta\mathbb{E}_gX,\end{aligned}$$

que, novamente é minimizado ao igualar a derivada em θ a zero, gerando, pela independência de g em

$$\theta_*=\frac{\beta}{\alpha}\approx 0.455.$$

Novamente, podemos testar se $\hat{\hat{\theta}}_n\approx\theta_*$, onde, agora, $\hat{\hat{\theta}}_n$ é a média bootstrap, isto é

$$\hat{\hat{\theta}}_n=\frac{1}{B}\sum_{i=1}^B\arg\max_{\theta\in\Theta}\left\{\frac{1}{n}\sum_{i=1}^n\log f(X_i,\theta)\right\}.$$

De fato, utilizando-se o método bootstrap com amostragem em 5000, conseguimos um valor de $\hat{\hat{\theta}}_n=0.435\pm0.063$, o que está de acordo, empiricamente, com o valor teórico esperado de $\theta_*=0.455$.

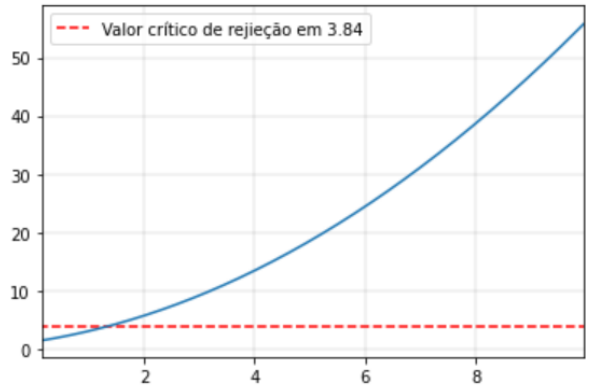

Figura 4: Valor do teste de Wald desenvolvido. A região de rejeição corresponde aos parâmetros $\lambda = \theta_0$ superiores à curva vermelha. Assim, nossa região de $1 - \alpha$ de confiança ficou, nesse experimento, $(0.1, 1.310)$.

Para achar os valores no teste T_w, notemos que

$$\begin{aligned}
\log f(x_i, \theta) &= \log \theta - \theta x_i \\
\frac{d \log f(x_i, \theta)}{d\theta} &= \frac{1}{\theta} - x_i \\
\frac{d^2 \log f(x_i, \theta)}{d\theta^2} &= -\frac{1}{\theta^2} \\
\frac{d^3 \log f(x_i, \theta)}{d\theta^3} &= \frac{2}{\theta^3}.
\end{aligned}$$

Assim, nosso teste empírico fica

$$T_w = \frac{n(\overline{X^2} - \frac{2\overline{X}}{\theta_*} + \hat{\theta}_n - \theta_0)^2}{\frac{\hat{B}(\hat{\theta}_n)}{\hat{A}^2(\hat{\theta}_n)} \left(\frac{2\overline{X}}{\theta_*^2} + 1 \right)^2} \implies \chi_1^2.$$

Para nosso caso específico, $n = 30$ (número de amostras distintas) e calculamos $\hat{\theta}_n$ usando o método de bootstrap apresentado acima. Assim, variando $\lambda = \theta_0$, podemos medir se $T_w(\lambda)$ ultrapassa o valor crítico, que no nosso caso é o $1 - 0.05 = 9.5$ quantil da distribuição χ_1^2, isto é, 3.84. O resultado dessa análise, variando λ de 0.1 a 10 é mostrado na Figura 4. Nos nossos testes, o intervalo de confiança $1 - \alpha$ ficou $(0.1, 1.310)$. Note porém, que como estamos trabalhando com um número de dados independentes n bem baixo, esse teste não é estatisticamente muito estável, no sentido que, pouco pode-se dizer sobre a distribuição g de modo geral (o que, no fundo, não deve ser o objetivo desse exercício, pois estamos supondo conhecimento nenhum sobre g, sendo os dados gerados nossa única informação propriamente disponível). Assim, toda vez que realizamos o teste, podemos encontrar intervalos bem distintos, dependendo apenas em qual os valores de dados X_i sorteados.