Lista 3 - Métodos de reamostragem e integração numérica

► Métodos de reamostragem

★ **Observação**: quando necessário, utilizar a semente 12345.

1. Utilize a transformação Box-Muller para gerar uma amostra de 30 números aleatórios provenientes de uma distribuição normal com $\mu = 20$ e $\sigma = 1$. Com esses dados, mostre que se $\hat{\theta} = \bar{x}$, a expressão

$$\hat{se} = \sqrt{\frac{1}{n(n-1)} \sum_{i=1}^{n} (y_i - \bar{y})^2}$$

   é igual à estimativa jackknife do erro padrão (1) da nota de aula 3.

2. Considere uma amostra de 15 escolas de direto nos quais foram observados a pontuação média da escola em um teste de direito nacional (LSAT) e a média de notas de graduação da escola(GPA). Com dados do arquivo `law_school.csv`, pede-se:

   a) defina θ como o coeficiente de correlação entre LSAT e GPA e determine o valor da estimativa $\hat{\theta}$. (*Resposta*: `0.7763745`)

   b) calcule a estimativa jackknife do erro padrão e o vício para o coeficiente de correlação dos dados da escola de direito. As seis primeiras réplicas jackknife são:
   `[1] 0.8929471 0.7637068 0.7549984 0.7760968 0.7313197 0.7799687`

   c) calcule a estimativa *bootstrap* do erro padrão para o coeficiente de correlação, utilizando $B = 500$.
   Sugestão: A amostra é formada por um par de observações. Por isso, faça:
   ```
   x <- seq(1,n,1)
   i <- sample(x,n,replace=TRUE)
   aboot <- dat[i,]
   ```
   em que `aboot` é uma amostra *bootstrap*.
   As seis primeiras réplicas *bootstrap* são:
   `[1] 0.8481334 0.6894070 0.8359762 0.7663279 0.7541281 0.7452616`

   d) compare os resultados dos itens *b* e *c*.

3. Em 12 lotes de manteiga de amendoim apresentaram resíduos de aflatoxina em partes por bilhão de 4,94; 5,06; 4,53; 5,07; 4,99; 5,16; 4,38; 4,43; 4,93; 4,72; 4,92 e 4,96. Para esses dados:

   a) calcule a média e seu erro padrão (a partir do padrão desvio) e a mediana.

   b) calcule a estimativa jackknife da média e mediana com o seus erros padrões.

   c) utilizando $B = 1000$ reamostras, calcule a estimativa *bootstrap* da média e mediana com o seus erros padrões.

   d) Compare os resultados do item a, b e c.

4. Gerar 100 amostras $Y_1, \ldots, Y_{20}$ provenientes de uma população com distribuição $N(1, 1)$, utilizando a transformação Box-Muller. Então, supondo que o parâmetro de interesse seja $\theta = \text{var}(\bar{Y}) = \frac{\sigma^2}{n}$, pede-se:

a) para cada amostra, calcular a estimativa jackknife e *bootstrap* de $\hat{\theta}$, ou seja, a média das réplicas. "Guardar" as estimativas em um objeto do tipo matriz de ordem 100×2. Considere $B = 1000$ reamostragens. As seis primeiras média das réplicas são:

```
             jack       boot
[1,] 0.03799378 0.03574179
[2,] 0.02432807 0.02316751
[3,] 0.02639213 0.02505389
[4,] 0.03269027 0.03124289
[5,] 0.03163722 0.03023826
[6,] 0.07847741 0.07491616
```

b) calcular a média e o desvio padrão das estimativas da variância de $\bar{Y}$ sobre as 100 amostras. Qual apresentou melhor aproximação e menor desvio?

c) construir um histograma com as estimativas jackknife e *bootstrap* de modo que a visualização seja um histograma do lado do outro (utilize: `par(mfrow=c(1,2))`). Insira, também, uma linha vertical utilizando a função `abline(v = , col = )` para indicar o valor exato do parâmetro θ e uma outra linha vertical para indicar a média das 100 estimativas. Altere a cor das linhas e insira uma legenda, utilizando a função `legend(x,y, c('Exato','Média das réplicas'), lty=c(1,1), col=c(,), bty='n')`, em que x e y indicam a posição de onde deve ser colocada a legenda.

5. Determine a distribuição acumulada empírica e a função distribuição acumulada:

a) Suponha que uma espécie de híbrido de milho foi desenvolvida por melhoramento genético e sendo assim tem taxa de germinação em 83%. Seja Y o número de sementes germinadas em um total de 10 sementes segue uma distribuição binomial. Nesse contexto, de 20 placas de petri contendo 10 sementes desse híbrido de milho foi o observado quantas sementes germinaram e os resultados foram:
`9 5 8 8 6 9 7 7 8 8 10 9 10 9 7 8 8 7 10 9`

b) Suponha a concentração bacteriana no leite de uma região segue uma distribuição logística com função densidade de probabilidade dada por

$$f(y) = \frac{\exp\left[-\frac{(y-\mu)}{\sigma}\right]}{\sigma\left\{1+\exp\left[-\frac{(y-\mu)}{\sigma}\right]\right\}^2}, \quad y \in \mathbb{R},$$

em que $\mu = 440$ ufc/ml e $\sigma = 113$ ufc/ml. Então, dessa região foram coletadas 10 amostras de leite e os resultados da análise laboratorial (em ufc/ml) foram de:
`186.0 168.6 656.2 572.2 631.6 387.9 565.9 539.7 441.5 570.7`

6. Utilizando os resultados do item 5, pede-se:

a) Construa o gráfico da distribuição acumulada empírica numericamente e analiticamente para os itens a e b. Verifique se as figuras são iguais.

b) Sobreponha a curva da função distribuição acumulada sobre o gráfico da distribuição acumulada empírica para os itens a e b. Altere a cor da curva da função de distribuição acumulada.
Observação:

- Lembre-se que para variável aleatória discreta, a função de distribuição acumulada é um gráfico de escada.
- Adapte a função `random_lg` do exercício 10 da lista 2 para incluir μ e σ. Então, utilize essa função para gerar os números aleatórios para construir a curva da função de distribuição acumulada da logística, utilizando a função `lines( )`.

7. Suponha que três ratos, companheiros de ninhada, tenham pesos 82, 107 e 93 g.

   a) Qual é o peso médio dos ratos?

   b) Quantas amostras *bootstrap* dessa amostra original são possíveis? Liste as amostras *boostrap* possíveis.

   c) Calcule a média de cada amostra *bootstrap* possível.

   d) Calcule a média das médias das amostras *bootstrap*. Como isso se compara com o média da amostra original?

   e) Quais são os valores alto e baixo das amostras *bootstrap*?

8. De acordo com o Conselho Nacional de Segurança nos Transportes dos EUA, o número de acidentes aéreos por ano de 1983 a 2006 foi 23, 16, 21, 24, 34, 30, 28, 24, 26, 18, 23, 23, 36, 37, 49, 50, 51, 56, 46, 41, 54, 30, 40 e 31. Para esses dados:

   a) calcule a média e seu erro padrão (a partir do padrão desvio) e a mediana.

   b) utilizando $B = 1000$ reamostras, calcule a estimativa *bootstrap* da média e mediana com o seus erros padrões.

   c) Compare os resultados do item a e b.

9. Gerar números aleatórios de tamanho $n = 5, 25, 50$ provenientes de uma distribuição exponencial com $\lambda = 1$, utilizando o método da transformação inversa.

   a) Para cada tamanho de amostra, calcular $B = 200$ réplicas *bootstrap* da média, $\bar{Y} = \sum_{i=1}^{n} Y_i$. Para isso, "guardar" as réplicas *bootstrap* da média em um objeto do tipo matriz de ordem $B \times n$. As seis primeiras réplicas para cada tamanho de amostra são

```
            n5       n25       n50
[1,] 1.528684 1.557244 1.0004369
[2,] 1.796669 1.127018 1.0211849
[3,] 1.515081 1.261102 0.7908425
[4,] 1.989562 1.210476 0.8607398
[5,] 1.665783 1.088807 0.8660756
[6,] 1.173901 1.398150 1.3907778
```

   b) Construir um histograma com as réplicas *bootstrap* de cada tamanho da amostra de modo que a visualização seja um histograma do lado do outro. Insira, também, uma linha vertical para indicar a média exata da distribuição exponencial e uma outra linha vertical para indicar a média das B réplicas *bootstrap*. Altere a cor das linhas e insira uma legenda.

   c) Construir um histograma com as réplicas *bootstrap* de cada tamanho da amostra e sobreponha a densidade exata de $\bar{Y}$, segundo o Teorema Central do Limite (TCL). A visualização também deve ser um histograma do lado do outro.

d) O que os itens b e c indicam? Discuta se o *bootstrap* está fazendo algo que a resposta do TCL não faz.

10. Gerar números aleatórios de tamanho $n = 10, 20, 50$ provenientes de uma distribuição normal com $\mu = 0$ e $\sigma^2 = 1$, utilizando o método da transformação Box-Muller. Repetir os mesmos itens do exercício anterior. Nesse caso, as seis primeiras réplicas de cada amostra são:

```
              n10         n20          n50
[1,] -0.13735184 -0.03725691 -0.109022104
[2,]  0.07816394  0.03956557  0.052162199
[3,]  0.12051320  0.27743992  0.196233248
[4,]  0.61851299  0.10992104  0.001387132
[5,] -0.10444496  0.17877908  0.002900164
[6,]  0.03057153  0.59495399  0.002115665
```

► Integração numérica

1. (Módulo de probabilidade - Lara) Dada a função $f(x) = \frac{3}{4}x(2-x), \quad 0 < x < 2$. Calcule a $P(X > 1)$, utilizando a quadratura de Gauss-Legendre para $n = 1, 2, 3, 4$ e 5. A partir de qual n a quadratura converge para o valor exato de 1/2?

2. (Módulo de probabilidade - Lara) Numa população, a expectativa de vida, em anos, tem distribuição exponencial com parâmetro $\lambda = 1/70$. Pede-se:

   (a) determine $E(X)$, utilizando a quadratura de Gauss-Laguerre para $n = 1, 2, 3, 4$ e 5. A partir de qual n a quadratura converge para o valor exato de 70 anos?

   (b) para um indivíduo escolhido ao acaso, determine a probabilidade de viver pelo menos até os 75 anos, utilizando a quadratura de Gauss-Legendre para $n = 1, \ldots, 10$. (0,3425189)

   (c) construa um gráfico de pontos (`pch=16`) da quantidades de nós da quadratura (eixo da abcissas) pela estimativa da probabilidade (eixo das ordenadas) do item b. Insira uma linha horizontal, utilizando a função `abline(h=,col=, lty=2)` para indicar o valor exato da probabilidade.

   (d) a probabilidade de ele morrer antes dos 75, sabendo-que o indivíduo acabou de completar 60 anos, ou seja, ele pode viver mais do que 60 anos. Utilize a quadratura de Gauss-Legendre para $n = 1, \ldots, 10$. (0,1928823)

3. (Módulo de probabilidade - Lara) A observação dos pesos, X, de um grande número de espigas de milho mostrou que essa variável é normalmente distribuída com $\mu = 90g$ e $\sigma^2 = 49g^2$. Num programa de melhoramento, entre outras características, uma linhagem, para continuar no programa, deve satisfazer $78g < X < 104g$.

   a) Calcule $E(X)$, utilizando a quadratura de Gauss-Hermite para $n = 1, 2, 3, 4$ e 5. A partir de qual n a quadratura converge para o valor exato de 90g?

   b) Determine a probabilidade de uma linhagem qualquer continuar num programa de melhoramento em relação ao peso? Utilize a quadratura de Gauss-Legendre para $n = 1, \ldots, 10$. (0,9340117)

   c) Construa um gráfico de pontos da quantidades de nós da quadratura pela estimativa da probabilidade do item b. Insira uma linha horizontal para indicar o valor exato da probabilidade.

4. (Módulo de probabilidade - Lara) Suponha novamente concentração bacteriana no leite de uma região do exercício 5b. Agora considere, $\mu = 440,8$ ufc/ml e $\sigma = 113,5$ ufc/ml. Então,

   a) calcule $\int_{-\infty}^{\infty} f(x)dx$, utilizando a quadratura de Gauss-Hermite com n igual a uma sequência de 10 até 250 de 10 em 10 valores. A partir de qual n a estimativa obtida pela quadratura converge para o verdadeiro valor de 1? (Faça uma transformação de variáveis: $z = \frac{x-440,8}{113.5}$)

   b) calcule a integral em a utilizando a quadratura de Gauss-Laguerre com n igual a uma sequência de 10 até 250 de 10 em 10 valores. Lembre-se que a distribuição é simétrica.

   c) Qual quadratura de Gauss utiliza menos pontos para resolver a integral definida em a?

   d) se uma amostra de leite desta região é selecionada para análise, determine a probabilidade da concentração bacteriana não ultrapassar 250 ufc/ml utilizando a quadratura de Gauss-Legendre com $n = 1, \ldots, 10$. (0,15696). **Sugestão**: Como a distribuição é simétrica, $P(X < 250) = 0,5 - P(250 < X < 440,8)$.

5. Como se justifica a utilização de uma específica quadratura de Gauss para resolver as integrais definidas dos itens 1 a 5?

6. Seja X uma variável aleatória com função densidade de probabilidade dada por

$$f(x) = \frac{2}{9} x \exp\left[-\left(\frac{x}{3}\right)^2\right], \quad x > 0.$$

   a) Qual quadratura de Gauss pode ser utilizada para calcular a integral

$$\int_0^{\infty} f(x)dx?$$

   Justifique a resposta.

   b) Calcule a integral do item a utilizando $n = 1, \ldots, 10$ pontos.

7. Utilizando a aproximação de Laplace, determine os valores da função gama dada por

$$\Gamma(\lambda) = \int_0^{\infty} x^{\lambda-1} e^{-x} dx,$$

   considerando λ uma sequência de 10 valores entre 2 e 5 (inclusive). Compare os valores com os valores da função `gamma(`λ`)` do `R` por meio de um gráfico de pontos.

8. Considere a seguinte função

$$f(x) = \frac{1}{x^2 + 1}.$$

   a) Qual quadratura de Gauss pode ser utilizada para calcular a integral $\int_{-1}^{2} f(x)dx$? Justifique a resposta.

   b) Utilizando a aproximação de Laplace e a quadratura de Gauss, determine o valor da integral do item a. Compare os resultados, considerando $n = 1, 2, 3, 4$ e 5 pontos.

9. Seja X uma variável aleatória com função densidade de probabilidade dada por

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}}, \quad x \in \mathbb{R}.$$

a) Qual quadratura de Gauss pode ser utilizada para calcular a integral $\int_{-\infty}^{\infty} f(x)dx$? Justifique a resposta.

b) Utilizando a aproximação de Laplace e a quadratura de Gauss, calcule a integral do item a. Para a quadratura utilize $n = 1, 2, 3, 4$ e 5 pontos e compare os resultados.