
Fundamentos Estatísticos para Ciência dos Dados
Livro de Exercícios, com Soluções

Renato Martins Assunção
Departamento de Ciência da Computação, UFMG

Prefácio

Nas últimas duas décadas, houve uma importante mudança na teoria e na prática ...

Falar sobre os alunos que colaboraram.

Sumário

1	Revisão de Matemática	1
2	Probabilidade Básica	7
3	Variáveis Aleatórias	13
4	Transformação de uma v.a.e Monte Carlo	33
5	Vetores Aleatórios	47
6	Modelos multivariados gaussianos	59
7	Classificação	71
8	Teoremas Limite: LGN e TCL	83
9	Regressão Linear	87
10	Teoria da Estimação Pontual	117
11	Método de Máxima Verossimilhança	131

Capítulo 1

Revisão de Matemática

Estes exercícios visam a uma revisão de fatos básicos de matemática e probabilidade que serão necessários durante a disciplina.

1. *Teoria de Conjuntos*: O objetivo é apenas verificar se você está informado sobre a diferença conceitual entre conjuntos enumeráveis e não-enumeráveis. Não é necessário saber provar que um conjunto é não-enumerável. Diga quais dos conjuntos abaixo é um conjunto enumerável e qual é não-enumerável:

- $\{0, 1, 2\}$
- naturais: $\mathbb{N} = \{0, 1, 2, \dots\}$
- inteiros: $\mathbb{Z} = \{\dots, -2, -1, 0, 1, 2, \dots\}$
- reticulado inteiro no plano: $\{(x, y); x \in \mathbb{Z}, y \in \mathbb{Z}\}$
- racionais: $\mathbb{Q} = \{p/q; q > 0, p \in \mathbb{Z}, q \in \mathbb{N}\}$
- reais: \mathbb{R}
- irracionais: $\mathbb{R} - \mathbb{Q}$.

2. *Propriedades básicas de expoentes*. Identifique abaixo quais igualdades estão corretas:

- $x^a y^a = (xy)^a$
- $\frac{a}{x+y} = \frac{a}{x} + \frac{a}{y}$
- $(x^a)^b = x^{ab}$
- $(x/y)^a = x^a / y^b$
- $(x + y)^a = x^a + y^a$
- $x^a y^b = (xy)^{a+b}$
- $(-x)^2 = -x^2$
- $\sqrt{x^2 + y^2} = |x| + |y|$
- $\frac{x+y}{a} = \frac{x}{a} + \frac{y}{a}$

- Os pontos $(x, y) \in \mathbb{R}^2$ que satisfazem a equação $x^2 + y^2 = 1$ formam ?? no plano real.
- Os pontos que satisfazem a equação $x^2 + y^2 = 4$ formam ??.
- Os pontos que satisfazem a equação $(x - 2)^2 + (y + 1)^2 = 1$ formam ??.
- Os pontos que satisfazem a equação $\left(\frac{x-2}{2}\right)^2 + \left(\frac{y+1}{1}\right)^2 = 1$ formam ??.

4. *Propriedades básicas das funções exp e log.*

- Esboce o gráfico das funções $f(x) = \log(3x + 1)$ e $f(x) = \exp(3x)$. Identifique o maior domínio na reta em que as funções podem ser definidas.
- Obtenha as derivadas $f'(x)$ das duas funções acima.
- Verifique quais das seguintes igualdades são válidas:
 - $\log(xy) = \log(x) + \log(y)$.
 - $\log(x + y) = \log(x) \times \log(y)$.
 - $\exp(x + y) = \exp(x) + \exp(y)$.
 - $\exp(x + y) = \exp(x) \times \exp(y)$.
 - $\log(x/y) = \log(x) - \log(y)$.
 - $\exp(xy) = (\exp(x))^y$.
 - $\exp(xy) = \exp(x) + \exp(y)$.

5. Esboce o gráfico da função $f(x) = \exp(-3(x - 1)^2)$ e obtenha a sua derivada $f'(x)$.

6. *Expansão de Taylor até segunda ordem.* Esta é uma das fórmulas mais úteis em matemática pois ela permite aproximar uma função $f(x)$ que pode ser muito complicada por uma função muito simples, um polinômio de segundo grau. A aproximação de Taylor precisa escolher um ponto de referência x_0 e ela vale para pontos x no entorno de x_0 (este entorno varia de problema para problema). A expansão de Taylor da função f no ponto x_0 próximo de x_0 é dada por

$$f(x) \approx f(x_0) + f'(x_0)(x - x_0) + \frac{1}{2}f''(x_0)(x - x_0)^2$$

Essencialmente, todas as funções que aparecem na prática da análise de dados podem ser aproximadas pela expansão de Taylor.

- Obtenha a expressão aproximada para $f(x) = \exp(x)$ para $x \approx x_0 = 0$.
- Faça um gráfico com as duas funções, $f(x)$ e sua aproximação de Taylor de 2a. ordem, para $x \in (-1, 2)$.
- Repita com $x_0 = 1$: obtenha a expressão aproximada para $f(x) = \exp(x)$ para $x \approx x_0 = 1$.
- Faça um gráfico com as duas funções, $f(x)$ e sua aproximação de Taylor de 2a. ordem, para $x \in (-1, 2)$.

Você deve obter gráficos iguais ao da Figura 1.1.

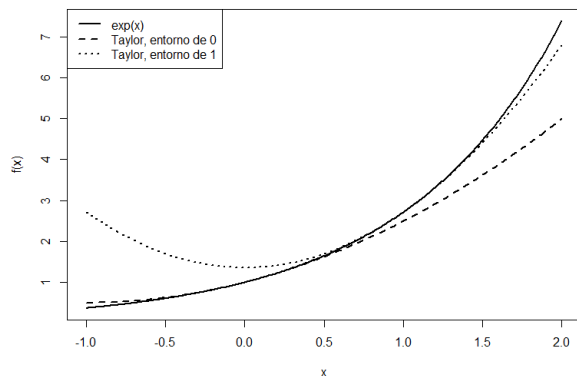


Figura 1.1: Aproximação de Taylor até a segunda ordem de $f(x) = e^x$ em torno de $x_0 = 0$ e em torno de $x_0 = 1$.

7. Na expansão de Taylor, aproximações numa região mais extensa em torno do ponto de referência x_0 podem ser obtidas usando um polinômio de grau mais elevado (o que implica calcular derivadas de ordens mais elevadas):

$$f(x) \approx f(x_0) + f'(x_0)(x-x_0) + \frac{1}{2!}f''(x_0)(x-x_0)^2 + \frac{1}{3!}f'''(x_0)(x-x_0)^3 + \frac{1}{4!}f^{(4)}(x_0)(x-x_0)^4 + \dots$$

Por exemplo, em torno de $x_0 = 0$ e usando a expansão até a 4a. ordem, temos

$$\frac{e^x}{\cos(x)} \approx 1 + x + x^2 + \frac{2x^3}{3} + \frac{x^4}{2}$$

Faça um gráfico de $f(x) = e^x / \cos(x)$ com a aproximação até a segunda ordem (basta usar os primeiros 3 termos acima) e até a quarta ordem para $x \in (-1, 1)$. Você deve obter um gráfico igual ao da Figura 1.2.

8. Considere a seguinte matriz 5×3 contendo dados de 5 apartamentos colocados a venda em BH:

$$\mathbf{X} = \begin{bmatrix} 1 & 153 & 2 \\ 1 & 107 & 1 \\ 1 & 238 & 3 \\ 1 & 179 & 2 \\ 1 & 250 & 4 \end{bmatrix}$$

Cada linha possui dados de um apartamento distinto. A primeira coluna contém apenas o valor constante 1. A segunda coluna mostra a área (em metros quadrados) de cada apto. A terceira coluna mostra o número de quartos do apto.

Seja $\beta = (\beta_0, \beta_1, \beta_3)^t$ um vetor-coluna 3×1 .

- Sejam v_1, \dots, v_k vetores em \mathbb{R}^5 . Verifique que o conjunto das combinações lineares desses vetores forma um sub-espço vetorial do \mathbb{R}^5 .

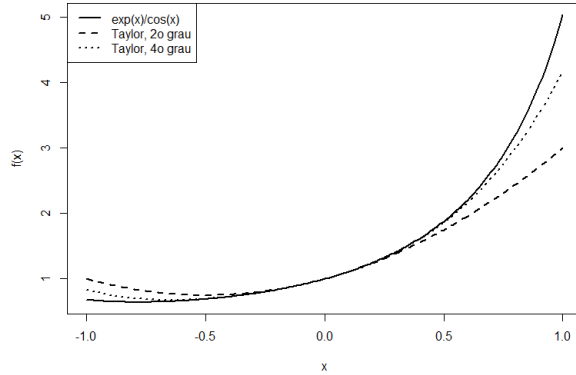


Figura 1.2: Aproximação de Taylor até a segunda e a quarta ordem de $f(x) = e^x/\cos(x)$ em torno de $x_0 = 0$ no intervalo $(-1, 1)$.

- Verifique que é válida a seguinte igualdade: $\mathbf{X}\beta = \beta_0 X_0 + \beta_1 X_1 + \beta_3 X_3$ onde X_0, X_1, X_3 são os vetores-coluna da matriz \mathbf{X} .
 - O conjunto $\mathfrak{M}(\mathbf{X})$ das combinações lineares das colunas de \mathbf{X} é igual a $\mathfrak{M}(\mathbf{X}) = \{\mathbf{X}\beta \mid \beta \in \mathbb{R}^3\}$ e é um sub-espço vetorial do \mathbb{R}^5 . V ou F? Se V, qual a dimensão do sub-espço vetorial $\mathfrak{M}(\mathbf{X})$?
9. Uma manipulação algébrica que é muito comum em estatística envolve uma decomposição de soma de quadrados. Seja $\bar{x} = (x_1, \dots, x_n)/n$ a média aritmética de x_1, \dots, x_n . Verifique que:
- $\sum_i (x_i - \bar{x})^2 = \sum_i x_i^2 - n\bar{x}^2$.
 - Se $a \in \mathbb{R}$, então $\sum_i (x_i - a)^2 = \sum_i (x_i \pm \bar{x} - a)^2 = \sum_i (x_i - \bar{x})^2 + n(\bar{x} - a)^2$.
 - A partir do item anterior, conclua que o valor de $a \in \mathbb{R}$ que minimiza $\sum_i (x_i - a)^2$ é o valor $a = \bar{x}$.
10. Nosso curso precisa usar vários resultados de álgebra de matrizes. Seja $\mathbf{x} = (x_1, \dots, x_n)$ um vetor-coluna $n \times 1$ e \mathbf{A} uma matriz $n \times n$. \mathbf{A}' indica a matriz transposta de \mathbf{A} . As seguintes identidades matriciais são fundamentais em nosso curso. Verifique que elas estão corretas, checando que o lado direito é igual ao lado esquerdo.
- $\mathbf{x}' \mathbf{A} \mathbf{x} = \sum_{i,j} x_i x_j A_{ij}$
 - $\mathbf{x}' \mathbf{x} = \sum_i x_i^2$
 - $\mathbf{x} \mathbf{x}'$ é uma matriz simétrica $n \times n$ com elemento (i, j) dado por $x_i x_j$.
11. O vetor gradiente é a extensão do conceito de derivada para funções de \mathbb{R}^n para \mathbb{R} . Para ser mais concreto, você pode imaginar a altura $f(\mathbf{x}) = f(x_1, x_2)$ de uma superfície f para

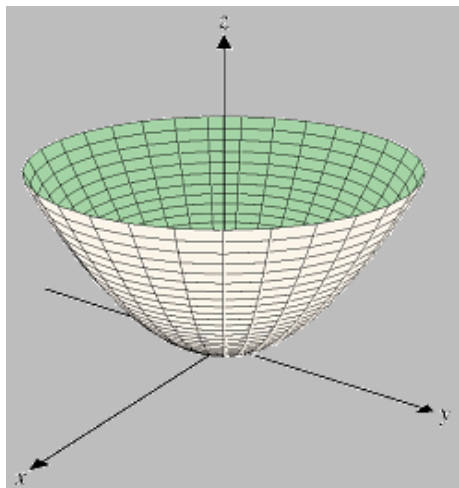


Figura 1.3: Gráfico da função $f(x_1, x_2) = x_1^2 + x_2^2$.

cada posição $\mathbf{x} = (x_1, x_2)$ do plano \mathbb{R}^2 . Seja

$$\begin{aligned} f : \mathbb{R}^n &\longrightarrow \mathbb{R} \\ \mathbf{x} &\longrightarrow f(\mathbf{x}) \end{aligned}$$

A derivada mede o quanto $f(\mathbf{x})$ varia quando \mathbf{x} sofre uma pequena perturbação. Os matemáticos percebem que a quantidade desta variação em $f(\mathbf{x})$ dependia da direção da perturbação com relação a \mathbf{x} ! Isto é, a variação em $f(\mathbf{x})$ depende do ponto \mathbf{x} em que estamos, de quanto nos afastamos desse ponto e, diferente do caso uni-dimensional, da direção em que nos afastamos dele.

Por exemplo, se $f(x_1, x_2) = x_1^2 + x_2^2$, a função f é chamada de parabolóide e seu gráfico pode ser visto na Figura 1.3. Observe que a função f é igual à distância ao quadrado entre o ponto $\mathbf{x} = (x_1, x_2)$ e a origem $\mathbf{0} = (0, 0)$. Portanto, se nos movimentarmos ao longo dos círculos concêntricos, o valor de $f(x_1, x_2)$ não varia e sua derivada deveria ser zero. Isto é, suponha que estamos num ponto $\mathbf{x} = (x_1, x_2)$ qualquer, distante $r = \sqrt{x_1^2 + x_2^2}$ da origem $(0, 0)$. Se nos movimentarmos ligeiramente, *mas ainda mantendo a mesma distância r da origem*, a função f não muda de valor e portanto sua variação é igual a zero. Assim, um ligeiro movimento ao longo da direção tangente ao círculo concêntrico deveria implicar numa derivada iguala zero.

Por outro lado, se nos movimentarmos em outras direções, a variação de f pode ser positiva ou negativa. Por exemplo, se ??

There is a nice way to describe the gradient geometrically. Consider $z = f(x, y) = 4x^2 + y^2$. The surface defined by this function is an elliptical paraboloid. This is a bowl-shaped surface. The bottom of the bowl lies at the origin. The figure below shows the level curves, defined by $f(x, y) = c$, of the surface. The level curves are the ellipses $4x^2 + y^2 = c$. The gradient vector $\langle 8x, 2y \rangle$ is plotted at the 3 points $(\sqrt{1.25}, 0), (1, 1), (0, \sqrt{5})$. As the

- 6 plot shows, the gradient vector at (x, y) is normal to the level curve through (x, y) . As we will see below, the gradient vector points in the direction of greatest rate of increase of $f(x, y)$

12. O conceito de derivada pode ser estendido para funções f de \mathbb{R}^n em \mathbb{R}^m . Seja

$$\begin{aligned} \mathbf{f} : \mathbb{R}^n &\longrightarrow \mathbb{R}^m \\ \mathbf{x} &\longrightarrow \mathbf{f}(\mathbf{x}) = (f_1(\mathbf{x}), \dots, f_m(\mathbf{x})) \end{aligned}$$

onde cada componente $f_j(\mathbf{x})$ é uma função de \mathbb{R}^n para \mathbb{R} . Derivada é matriz $m \times n$??
contrario a gradiente.

13. ?? Integral e derivada num pequeno segmento; $f(x, y_0)$ é uma curva (uma funcao uni-dimensional).

Capítulo 2

Probabilidade Básica

1. Uma função de probabilidade \mathbb{P} satisfaz aos três axiomas de Kolmogorov. Prove as seguintes propriedades derivadas destes três axiomas:

(P1) $\mathbb{P}(A^C) = 1 - \mathbb{P}(A)$

(P2) $0 \leq \mathbb{P}(A) \leq 1$ para todo evento $A \in \mathcal{A}$.

(P3) se $A_1 \subset A_2 \implies \mathbb{P}(A_1) \leq \mathbb{P}(A_2)$

(P4) $\mathbb{P}(\bigcup_{n=1}^{\infty} A_i) \leq \sum_{n=1}^{\infty} \mathbb{P}(A_i)$

(P5) $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$. Esta propriedade é o caso geral de $\mathbb{P}(A \cup B)$

2. Sejam $\Omega = [0, 1]$ e $f(\omega)$ a densidade de probabilidade. Marque V ou F nas afirmações abaixo:

- $\mathbb{P}([a, b]) = b - a$ se $[a, b] \subset [0, 1]$.
- Como a integral de $f(\omega)$ sobre $[0, 1]$ é igual a 1 temos $f(\omega) \leq 1$ para todo $\omega \in [0, 1]$.
- Podemos ter $f(\omega) > 1$ para todo $\omega \in [0, 1]$.
- $f(x)$ pode ser descontínua.
- $f(x)$ não pode ter dois ou mais pontos de máximo.
- $f(x) = 2x$ é uma densidade válida.
- $f(x) = 12(x - 0.5)^2$ é uma densidade válida.
- Não podemos ter $f(x) \rightarrow \infty$ quando $x \rightarrow 1$ porque a densidade deve integrar 1 no intervalo $[0, 1]$.

3. Se $A \subset B$ temos $\mathbb{P}(B|A) = 1 \geq \mathbb{P}(B)$. Assim, a ocorrência de A aumenta a probabilidade de B para seu valor máximo possível, que é 1. Temos certeza que B ocorreu pois B é parte de A . E o contrário? Mostre que se $A \supset B$ podemos concluir que $\mathbb{P}(B|A) \geq \mathbb{P}(B)$. Isto é, se B é parte de A , saber que A ocorreu tende a aumentar a chance de B ocorrer. Intuitivamente, se B for uma grande parte de A devemos ter $\mathbb{P}(B|A) \approx 1$.

- 8 4. No momento do diagnóstico de um câncer de estômago para um paciente qualquer, definimos o evento B como sendo o evento em que o paciente tem pelo menos mais 1 ano de vida. Suponha que $\mathbb{P}(B) = 0.70$. Seja A o evento em que um paciente de câncer de estômago tenha uma autópsia confirmando que o tumor é benigno. Imaginamos que $\mathbb{P}(B|A)$ seja maior que 0.70. Explique como as probabilidades $\mathbb{P}(B)$ e $\mathbb{P}(B|A)$ poderiam ser estimadas com base numa grande amostra de pacientes de câncer de estômago. Que frequências relativas você usaria para estimá-las?
5. Um ponto aleatório (x, y) é escolhido completamente ao acaso no disco de raio unitário centrado na origem. Temos $\Omega = \{(x, y) \in \mathbb{R}^2 : x^2 + y^2 \leq 1\}$ e a probabilidade de um evento $E \subset \Omega$ é dada por $\mathbb{P}(E) = \text{área de } E$. Seja A o evento “distância do ponto escolhido e a origem é menor que $1/2$ ” e B o evento “a coordenada x do ponto escolhido é maior que y ”. Mostre que A e B são independentes.
6. Marque V ou F nas afirmações abaixo:
- Se $A \cap B = \emptyset$ então A e B são eventos independentes.
 - Se $A \subset B$ então A e B são eventos independentes.
 - Se $\mathbb{P}(B|A) = \mathbb{P}(B)$ dizemos que B é independente de A . A ordem dos eventos não importa pois isto implica que $\mathbb{P}(A|B) = \mathbb{P}(A)$ e portanto que B também é independente de A .
 - $\mathbb{P}(B|A) > \mathbb{P}(B)$ se, e somente se, $\mathbb{P}(A \cap B) > \mathbb{P}(B) \times \mathbb{P}(A)$.
7. Mostre que, se $\mathbb{P}(A) = 0$, então A é independente de qualquer outro evento B . Isto faz sentido intuitivamente?
8. Um evento A é independente de si mesmo se, e somente se, $\mathbb{P}(A) = 0$ ou 1. Prove isto em uma linha.
9. Se A e B são independentes então A e B^c também são independentes (e também A^c e B , e ainda A^c e B^c). Prove isto usando que $A = (A \cap B^c) \cup (A \cap B)$.
10. Se A, B, C são eventos mutuamente independentes, mostre que C é independente de $A \cap B$, de $A \cap B^c$, de $A^c \cap B^c$. Isto é, mostre que, por exemplo, $\mathbb{P}(A \cap B^c \cap C) = \mathbb{P}(A)\mathbb{P}(B^c)\mathbb{P}(C)$, etc.
11. Usando o resultado acima, mostre que, se A, B, C são eventos mutuamente independentes, C é independente de $A \cup B$. Dica: escreva $A \cup B$ como uma interseção de conjuntos.
12. Mostre que, se B é um evento 5 vezes mais provável que outro evento A , então $\mathbb{P}(B|A) = 5\mathbb{P}(A|B)$. De maneira geral, mostre que

$$\frac{\mathbb{P}(B|A)}{\mathbb{P}(A|B)} = \frac{\mathbb{P}(B)}{\mathbb{P}(A)}$$

13. Responda V ou F :

- Sensitividade é a probabilidade de que um indivíduo tenha $TESTE+$ e $HIV+$.

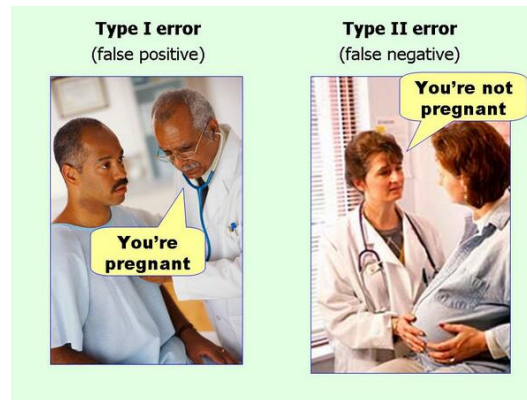


Figura 2.1: Falso positivo e falso negativo

- Sensitividade é a probabilidade de que um indivíduo com $HIV+$ tenha resultado $TESTE+$.
 - Sensitividade é a probabilidade de que um indivíduo com resultado $TESTE+$ seja $HIV+$.
 - Falso positivo ocorre se o teste indica positivo quando o paciente é $HIV+$
14. A Figura 2.1 mostra exemplos de falso positivo e falso negativo. Explique estes erros em termos de probabilidades condicionais $\mathbb{P}(A|B)$ explicando o que são os eventos A e B em cada um dos dois erros.
 15. Leia o delicioso artigo *Chances are* sobre a regra de Bayes do matemático Steven Strogatz no New York Times: <http://opinionator.blogs.nytimes.com/2010/04/25/chances-are/>. Steven é bastante conhecido por seu artigo *Collective dynamics of small-world networks* na revista *Nature* em 1998. Como uma medida do impacto deste artigo, ele foi o mais citado sobre redes entre 1998 e 2008 considerando todas as disciplinas científicas, bem como o sexto mais citado - sobre qualquer tema - em física. Usando a regra de Bayes, verifique que a resposta de 9% para $\mathbb{P}(\text{cancer} | \text{mamografia} +)$ é correta.
 16. Sejam A , B e C três eventos com probabilidade positiva (isto é, $\mathbb{P}(A) > 0$, $\mathbb{P}(B) > 0$ e $\mathbb{P}(C) > 0$). Mostre que

$$\mathbb{P}(A \cap B \cap C) = \mathbb{P}(A|B \cap C)\mathbb{P}(B|C)\mathbb{P}(C)$$

Dica: Faça $B \cap C = D$ e aplique a definição de probabilidade condicional em $\mathbb{P}(A \cap D)$. A seguir, aplique de novo esta definição em $\mathbb{P}(D) = \mathbb{P}(B \cap C)$

17. Mostre que $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$ implica que $\mathbb{P}(A|B) = \mathbb{P}(A)$.
18. Definimos $A \perp B|C$ se $\mathbb{P}(A|B \cap C) = \mathbb{P}(A|C)$. Mostre que $A \perp B|C$ se, e somente se, $\mathbb{P}(A \cap B|C) = \mathbb{P}(A|C) \mathbb{P}(B|C)$.
19. Problema de Monty Hall. Leia a descrição deste problema em http://en.wikipedia.org/wiki/Monty_Hall_problem. Faça um pequeno exercício de simulação em que voce

- 10 repete o jogo um grande número de vezes (100 mil, digamos) e testa as duas possíveis estratégias em cada jogo. Um jogador SEMPRE troca as portas. O outro nunca troca. Verique qual estratégia é mais eficiente (quem vence mais frequentemente)
20. Da Wikipedia: http://en.wikipedia.org/wiki/Three_Prisoners_problem. Three prisoners, A, B and C, are in separate cells and sentenced to death. The governor has selected one of them at random to be pardoned. The warden knows which one is pardoned, but is not allowed to tell. Prisoner A begs the warden to let him know the identity of one of the others who is going to be executed. "If B is to be pardoned, give me C's name. If C is to be pardoned, give me B's name. And if I'm to be pardoned, flip a coin to decide whether to name B or C."
- The warden tells A that B is to be executed. Prisoner A is pleased because he believes that his probability of surviving has gone up from $1/3$ to $1/2$, as it is now between him and C. Prisoner A secretly tells C the news, who is also pleased, because he reasons that A still has a chance of $1/3$ to be the pardoned one, but his chance has gone up to $2/3$. What is the correct answer?
- Faca um estudo de simulação similar ao do problema anterior para achar a resposta de maneira empírica.
21. Da wikipedia: http://en.wikipedia.org/wiki/Boy_or_Girl_paradox. Martin Gardner published one of the earliest variants of the Boy or Girl paradox in Scientific American. He phrased the paradox as follows:
- Mr. Jones has two children. The older child is a girl. What is the probability that both children are girls?
 - Mr. Smith has two children. At least one of them is a boy. What is the probability that both children are boys?
22. Do Buzz de Terence Tao, <https://profiles.google.com/114134834346472219368/buzz/G5DnA8EL7D3>. Another interesting place where one can contrast classical deduction with Bayesian deduction is with regard to taking converses. In classical logic, if one knows that A implies B, one cannot then deduce that B implies A. However, in Bayesian probability, if one knows that the presence of A elevates the probability that B is true, then an observation of B will conversely elevate the prior probability that A is true, thanks to Bayes' formula: if $\mathbb{P}(B|A) > \mathbb{P}(B)$, then $\mathbb{P}(A|B) > \mathbb{P}(A)$.
- Prove este resultado.
23. Usando probabilidade para fazer perguntas delicadas num questionario. Exemplos de perguntas delicadas para as quais queremos saber a probabilidade de uma resposta SIM:
- Você já fumou baseado alguma vez?
 - Você já roubou algum objeto numa loja?
 - Você é a favor do aborto? Etc.
- O ENTREVISTADO rola um dado bem balanceado e NÃO MOSTRA O RESULTADO AO ENTREVISTADOR. Se sair 1, 2 ou 3, o entrevistado responde SIM ou NÃO à pergunta

delicada. Se sair 4, 5 ou 6, ele responde SIM ou NÃO à seguinte pergunta alternativa:¹¹ o último dígito de sua conta bancária (ou de sua identidade) é par?

Quando o entrevistador ouve a resposta (digamos, SIM), ele não sabe a qual das duas perguntas o entrevistado está respondendo, se à delicada ou àquela sobre o dígito.

Suponha que a proporção de entrevistados respondendo SIM ao entrevistador foi 0.32. A amostra é bastante grande de modo que supomos $\mathbb{P}(SIM) \approx 0.32$.

Use a lei de probabilidade total para expandir $\mathbb{P}(SIM)$ em função dos dois resultados possíveis do dado e sugira uma estimativa para a probabilidade $\mathbb{P}(SIM|\text{dado} = 1, 2, 3)$.

24. Suppose that the probability of mothers being hypertensive (high blood pressure) is 0.1 and fathers is 0.2. Find the probability of a child's parents both being hypertensive, assuming both events are independent. Note: we would expect these two events to be independent if the primary determinants of hypertensivity were genetic, however if the primary determinants were environmental then we might expect the two events not to be independent.
25. Num teste de diagnóstico, a sensibilidade é a probabilidade de que o teste seja positivo dado que o indivíduo realmente seja doente: $\mathbb{P}(T+|D+)$. Um sinônimo muito usado para esta probabilidade no contexto de recuperação de informação é *recall*. As afirmativas abaixo foram ouvidas pelo autor em diferentes ocasiões. Explique cada uma das sentenças.
- “Aumentar muito o recall é praticamente não deixar passar um caso positivo”.
 - Uma alta sensibilidade significa uma alta taxa de “verdadeiros positivos”.
 - Em recuperação de informação, recall é a proporção de documentos recuperados que são relevantes. Coloque este problema no arcabouço de sensibilidade (isto é, faça a equivalência com “teste positivo”, “doente”, etc.)
26. Curva ROC ??

Capítulo 3

Variáveis Aleatórias

Esta lista de exercícios visa ao aprendizado de algumas das características das principais distribuições de probabilidade. Você vai se familiarizar com seus principais aspectos visuais e quantitativos, vai aprender a simular estas distribuições no R e a verificar se um conjunto de dados segue uma determinada distribuição usando o teste qui-quadrado e de Kolmogorov.

Vamos aprender um poucos sobre as seguintes distribuições:

- Discretas: binomial, Poisson, geométrica, Pareto-Zipf
- Contínuas: uniforme, gaussiana (ou normal), log-normal, gama, beta, Pareto.

O R possui um conjunto de funções para trabalhar com as principais distribuições de probabilidade. Todas operam com uma sintaxe similar. O primeiro caracter do nome da função identifica o que você quer fazer com ela: gerar números aleatórios, calcular uma probabilidade, uma probabilidade acumulada ou um quantil. Os caracteres seguinte identificam a distribuição.

Por exemplo, se quisermos trabalhar com a distribuição binomial com $n = 10$ repetições e probabilidade de sucesso $\theta = 0.15$ podemos usar:

- `rbinom(13, 20, 0.15)`: gera um conjunto de 13 inteiros aleatórios, cada um deles seguindo uma binomial $\text{Bin}(n = 20, \theta = 0.15)$.
- `dbinom(13, 20, 0.15)`: se $X \sim \text{Bin}(20, 0.15)$, este comando calcula a função de probabilidade $\mathbb{P}(X = 13) = p(13)$ para as v.a's discretas. Podemos passar vetores como argumento. Por exemplo, `dbinom(c(10, 11, 12), 20, 0.15)` retorna o vetor $(\mathbb{P}(X = 10), \mathbb{P}(X = 11), \mathbb{P}(X = 12))$.
- `pbinom(13, 20, 0.15)`: Calcula a função de probabilidade acumulada \mathbb{F} no ponto 13. Isto é, calcula $\mathbb{F}(13) = \mathbb{P}(X \leq 13)$ onde $X \sim \text{Bin}(20, 0.15)$.
- `qbinom(0.20, 20, 0.15)`: Calcula o quantil x associado com a de probabilidade acumulada 0.20. Isto é, calcula o valor de x tal que $\mathbb{F}(x) = \mathbb{P}(X \leq x) = 0.20$. Como X é uma v.a. discreta que acumula probabilidades aos saltos, a probabilidade acumulada até x pode ser apenas aproximadamente igual a 0.20.

As funções correspondentes para uma gaussiana são `rnorm`, `dnorm`, `pnorm`, `qnorm`. Se quisermos trabalhar com uma gaussiana $N(\mu, \sigma^2)$, com valor esperado $\mu = 10$ e $\sigma = 2$:

- `rnorm(100, 10, 2)`: gera um conjunto de 100 valores aleatórios independentes de uma v.a. $X \sim N(10, 2^2)$.
- `dnorm(11.25, 10, 2)`: retorna o valor da densidade $f(x)$ de $N(10, 2)$ no ponto $x = 11.25$. Isto é, retorna $f(11.25)$. O comando `dnorm(c(11.25, 13.15), 10, 2)` retorna um vetor com os valores $(f(11.25), f(13.15))$.
- `pnorm(11.25, 10, 2)`: Calcula a função de probabilidade acumulada no ponto 11.25. Isto é, calcula $\mathbb{F}(11.25) = \mathbb{P}(X \leq 11.25)$ onde $X \sim N(10, 2^2)$.
- `pnorm(0.20, 10, 2)`: Calcula o quantil x tal que $\mathbb{F}(x) = \mathbb{P}(X \leq x) = 0.20$. Como X é uma v.a. contínua que acumula probabilidades continuamente, a probabilidade acumulada até x é exatamente igual a 0.20.

Para uma Poisson, são as seguintes: `rpois`, `dpois`, `ppois` e `qpois`. Para a exponencial, temos `rexp`, `dexp`, `pexp` e `qexp`. Para conhecer todas as distribuições disponíveis no R, digite `?distributions` ou, equivalentemente, `help(distributions)`.

1. Seja $X \sim \text{Bin}(10, 0.4)$. Para obter e plotar (veja Figura 3.1) os valores da função de probabilidade $\mathbb{P}(X = k)$ e da função de probabilidade acumulada $\mathbb{F}(x)$ uso os seguintes comandos:

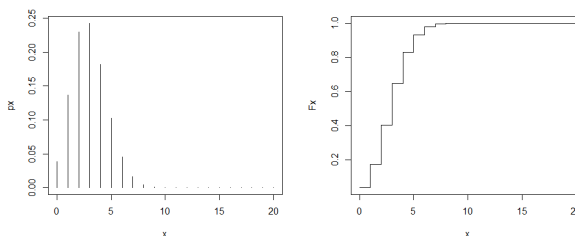


Figura 3.1: Função de probabilidade $\mathbb{P}(X = k)$ (esquerda) e da função de probabilidade acumulada $\mathbb{F}(x)$ (direita) de uma v.a. binomial $\text{Bin}(n = 10, \theta = 0.40)$.

```
x <- 0:10
px <- dbinom(x, 10, 0.40)
par(mfrow=c(1,2)) # janela grafica com uma linha de 2 plots
plot(x, px, type = "h") # para usar linhas verticais at\'} os pontos (x,px)
Fx <- pbinom(x, 10, 0.35)
plot(x, Fx, type = "s") # o argumento "s"
```

- Sua vez agora. Obtenha o gráfico das probabilidades $\mathbb{P}(X = k)$ e da função de probabilidade acumulada $\mathbb{F}(x)$ para uma v.a. $X \sim \text{Bin}(n = 20, \theta = 0.15)$. Em seguida, responda às questões abaixo.
- Qual o valor k em que $\mathbb{P}(X = k)$ é máxima? Quanto é esta probabilidade máxima?
- VISUALMENTE, obtenha uma faixa de valores (a, b) na qual a probabilidade de $X \in (a, b)$ seja próxima de 1. Procure grosseiramente obter a faixa mais estreita possível.
- O valor (teórico) de $\mathbb{E}(X)$ no caso de uma binomial é $n\theta$. Como é o comportamento da função $\mathbb{P}(X = k)$ no entorno deste valor $\mathbb{E}(X)$? Ela tem valores $\mathbb{P}(X = k)$ relativamente altos?
- Confirme esta impressão calculando $\mathbb{P}(a \leq X \leq b)$ usando a função `dnorm` ou `pnorm` do *R*. Por exemplo, se eu quiser $\mathbb{P}(5 \leq X \leq 8)$, uso `sum(dnorm(5:8, 20, 0.15))` ou então `pbinom(8, 20, 0.15) - pbinom(5-0.01, 20, 0.15)`. Porque eu subtraio 0.01 de 5 na chamada da segunda função?
- Use `qbinom` para obter o inteiro k tal que $\mathbb{F}(k) = \mathbb{P}(X \leq k) \approx 0.95$.
- Verifique o valor da probabilidade acumulada exata $\mathbb{F}(k)$ obtida com o inteiro acima usando `pbinom`.
- Gere 1000 valores aleatórios independentes de $X \sim \text{Bin}(n = 20, \theta = 0.15)$. Estes valores caíram, em sua maioria, na faixa que você escolheu mais acima? Qual a porcentagem de valores que caiu na faixa que você escolheu?
- Compare os valores das probabilidades $\mathbb{P}(X = k)$ para $k = 0, \dots, 6$ e as frequências relativas destes inteiros nos 100 valores simulados. São parecidos?

2. Este problema é similar ao anterior, usando agora a distribuição de Poisson.

- Obtenha o gráfico das probabilidades $\mathbb{P}(X = k)$ e da função de probabilidade acumulada $\mathbb{F}(x)$ para uma v.a. $X \sim \text{Poisson}(\lambda)$ usando dois valores: $\lambda = 0.73$ e $\lambda = 10$.
- O valor k em que $\mathbb{P}(X = k)$ é máximo é próximo de $\mathbb{E}(X) = \lambda$?
- Obtenha um intervalo de valores (a, b) , o mais curto possível grosseiramente, para o qual $\mathbb{P}(X \in (a, b)) \approx 1$.
- Usando `ppois` do *R*, calcule $\mathbb{P}(a \leq X \leq b)$.
- Gere 200 valores aleatórios independentes de $X \sim \text{Poisson}(\lambda)$ com os dois valores acima para λ .
- Compare os valores das probabilidades $\mathbb{P}(X = k)$ para $k = 0, \dots, 6$ e as frequências relativas destes inteiros nos 100 valores simulados. São parecidos?

3. Este problema é similar ao anterior, usando agora a distribuição discreta de Pareto, também chamada de distribuição de Zipf. Ver <http://en.wikipedia.org/wiki/Zipf>

`slaw`. A distribuição de Pareto (discreta ou contínua) não está disponível em `R` e não se encontra em alguns pacotes especializados. Entretanto, não é necessário usar estes pacotes já que ela é facilmente simulada ou calculada. Veremos técnicas de simulação Monte Carlo em breve, então apenas aceite por enquanto o algoritmo abaixo.

A distribuição discreta de Pareto possui suporte igual a $\{1, 2, \dots, N\}$ onde N pode ser infinito. Além de N , ela possui um outro parâmetro, $\alpha > 0$. A função massa de probabilidade é dada por

$$\mathbb{P}(X = k) = \frac{C}{k^{1+\alpha}}$$

onde C é uma constante escolhida para que as probabilidades somem 1. Observe que C é dada por

$$\frac{1}{C} = \sum_{k=1}^N \frac{1}{k^{1+\alpha}}$$

Se N for um número finito, não existe uma expressão analítica para esta soma e ela deve ser calculada somando-se os valores. Se N for infinito, a expressão acima é chamada de função ζ de Riemann:

$$\zeta(1 + \alpha) = \sum_{k=1}^{\infty} \frac{1}{k^{1+\alpha}} = \frac{1}{\Gamma(1 + \alpha)} \int_0^{\infty} \frac{x^{\alpha}}{e^x - 1} dx \quad (3.1)$$

(ver http://en.wikipedia.org/wiki/Riemann_zeta_function).

Para alguns valores específicos de α , a função zeta $\zeta(1 + \alpha)$ tem valores conhecidos exatamente. Por exemplo, para $\alpha = 1$ é possível mostrar que

$$\zeta(2) = \sum_{k=1}^{\infty} \frac{1}{k^2} = 1 + \frac{1}{2^2} + \frac{1}{3^2} + \dots = \frac{\pi^2}{6} \approx 1.645$$

Exceto nestes casos particulares, no caso de $N = \infty$, a constante $C = 1/\zeta(1 + \alpha)$ deve ser aproximada numericamente somando-se um número grande de termos da série ou calculando numericamente a integral em (3.1). Por exemplo, para $\alpha = 1/2$, temos $\zeta(1 + 1/2) \approx 2.612$, e para $\alpha = 2$, temos $\zeta(1 + 2) \approx 1.202$.

Tendo um valor para a constante C , podemos plotar os valores de $\mathbb{P}(X = k)$ e também da função de probabilidade acumulada $\mathbb{F}(k)$ já que

$$\mathbb{F}(k) = \mathbb{P}(X \leq k) = \sum_{i=1}^k \mathbb{P}(X = i) = C \sum_{i=1}^k \frac{1}{i^{1+\alpha}}.$$

- Usando os valores $\alpha = 1/2, 1, 2$, obtenha em `R` o gráfico das probabilidades $\mathbb{P}(X = k)$ e da função de probabilidade acumulada $\mathbb{F}(x)$ para uma v.a. $X \sim \text{Zipf}(\alpha)$ com $N = \infty$. Em `R`, não chame a constante de integração de `c` pois este é o nome da função de concatenação de vetores e, como um defeito do `R`, ele não avisa que você está sobrepondo uma função-base crucial. Faça a escala horizontal variar nos inteiros de 1 a 20. Obtenha $\mathbb{F}(x)$ usando o comando `cumsum` que retorna o vetor de somas acumuladas de um vetor.

- Pelo gráfico, as probabilidades parecem cair rápido, talvez exponencialmente. Mas isto não é verdade. O comportamento dessa queda quando k aumenta é a principal razão propriedade que faz com que a distribuição power-law de Pareto (ou Zipf) seja tão importante na prática da análise de dados. Para entender como as probabilidades diminuem em direção a zero a medida que k cresce, obtenha a raz ao entre valores sucessivos de $\mathbb{P}(X = k)$. Isto é, mostre que

$$\frac{\mathbb{P}(X = k + 1)}{\mathbb{P}(X = k)} = \left(\frac{k}{k + 1} \right)^{1+\alpha}$$

Perceba agora que, quando k cresce, $k/(k + 1)$ é sempre menor que 1 mas cada vez mais próximo de 1 e portanto

$$\mathbb{P}(X = k + 1) \approx \mathbb{P}(X = k)$$

se k for bem grande. As duas probabilidades serão pequenas mas quase idênticas. Isto é, a medida que k cresce, as probabilidades decaem muito lentamente, quase nadaquando k for bem grande.

- Quando $\alpha > 0$ crescer, o que você esperar acontecer ao gerar inteiros Zipf com estes α grandes em relação à geração com α apenas ligeiramente maior que 1.
- Faça um gráfico dos pontos $(\log(k), \log(\mathbb{P}(X = k)))$. O resultado é o que você esperava? Usando `abline(log(C), -(1+alpha))`, sobreponha uma reta com intercepto $\log(C)$ e inclinaç ao $-(1 + \alpha)$.
- Chega de análise teórica, vamos simular por MOnTe Carlo alguns valores Zipf agora. A função R abaixo faz isto para você:

```
rzipf = function(nsim = 1, alpha = 1, Cte = 1/1.645)
{
  res = numeric(nsim)
  for(i in 1:nsim){
    x = -1
    k = 1
    F = p = Cte
    U = runif(1)
    while( x == -1){
      if(U < F) x = k
      else{
        p = p * (k/(k+1))^(1+alpha)
        F = F + p
        k = k+1
      }
    }
    res[i] = x
  }
  res
}
```

Por default, a função assume $\alpha = 1$ e fornece também a constante C . Para gerar $nsim = 400$ valores com estes argumentos default, basta digitar `rzipf(400)`. Para gerar 400 valores de uma Zipf com $\alpha = 1/2$ e com a constante $C = 1/2.612$ determinada por este valor de α , basta digitar `rzipf(400, 1/2, 1/2.62)`.

Agora, a tarefa: gere 400 valores de Zipf com $\alpha = 1/2, 1, 2$ (as constantes estão no texto acima). Verifique que apesar da maioria dos valores ficar num intervalo limitado, valores extremamente grandes (relativamente aos demais) são gerados com facilidade. Repita a geração algumas vezes para observar este efeito. Reporte na lista apenas uma dessas repetições.

4. Este problema trata da distribuição gaussiana ou normal, a mais importante distribuição na análise de dados. Ela é uma v.a. contínua com suporte na reta real $\mathbb{R} = (-\infty, \infty)$ e com densidade de probabilidade dependendo de dois parâmetros, μ e σ :

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)$$

Neste exercício você vai se familiarizar com a distribuição gaussiana.

- Divida a tela gráfica 2×2 e desenhe o gráfico das densidades de probabilidade de uma $N(0, 1)$ na posição (1, 1) da janela, uma $N(2, 1)$ na posição (1, 2), uma $N(0, 4)$ na posição (2, 1) e uma $N(2, 4)$ na posição (2, 2).
- Qual o ponto x em que $f(x)$ assume o valor máximo? Este ponto depende de σ ? E a altura $f(x)$ no ponto de máximo, ela depende de σ ?
- No caso da gaussiana, o parâmetro σ controla a variação em torno de μ . Para uma $N(10, 5)$ verifique que área debaixo da densidade entre $10 - 2 \times \sqrt{5}$ e $10 + 2 \times \sqrt{5}$ é aproximadamente igual a 0.95. Use a função `pnorm` para isto. Este é um resultado geral: no caso de uma gaussiana, a chance de observar um valor distante mais de 2σ de do valor esperado e central μ é aproximadamente 0.05.
- Gere 200 valores aleatórios independentes de $X \sim N(\mu, \sigma)$ com μ e σ escolhidos por você. Faça um histograma forçando a área total ser igual a 1 (argumento `prob=T`) e sobreponha a curva da densidade gaussiana que você usou. Eles se parecem?

5. Seja $X \sim \exp(1/3)$. Isto é, $X \sim \exp(\lambda)$ com $\lambda = 1/3$. Isto implica que a densidade $f(x)$ é igual a

$$f(x) = \begin{cases} 0, & \text{se } x < 0 \\ (1/3) \exp(-x/3), & \text{se } x \geq 0 \end{cases}$$

Calcule $E(X)$, $F(x)$ e $P(X > 3)$.

6. X é uma v.a. com distribuição Pareto contínua com parâmetros m e α . Isto é,

$$f_X(x) = \begin{cases} 0 & \text{se } x \leq m \\ c/x^{\alpha+1} & \text{se } x > m \end{cases}$$

onde a constante de integração c é dada por $c = \alpha m^\alpha$.

Calcule $\mathbb{F}(x) = \mathbb{P}(X \leq x)$. Calcule também $\mathbb{E}(X)$ para $\alpha > 1$ (a integral $\mathbb{E}(X)$ não existe se $0 < \alpha \leq 1$).

Para simular 1000 valores de uma Pareto e visualizar os resultados com R , basta digitar:

```
m=1; alpha=1
x = m*(1-runif(1000))^(1/alpha)
par(mfrow=c(1,2))
hist(x); plot(x)
```

Repita estes comandos algumas vezes. Veja como valores muito extremos de X são gerados com facilidade.

7. O arquivo `vadiscreta.txt` possui uma tabela de dados com $n = 200$ itens e $k = 6$ atributos, todos discretos. Podemos assumir que os itens são replicações independentes de um mecanismo aleatório. Queremos encontrar um modelo probabilístico para cada coluna-atributo da tabela.

Para cada uma das colunas, vamos assumir que os $n = 200$ itens são realizações independentes de uma mesma v.a. discreta. As 3 primeiras colunas possuem 5 valores possíveis. Isto é, o suporte das v.a.s é o conjunto $\{1, 2, 3, 4, 5\}$. Embora os valores possíveis sejam os mesmos nos três atributos, as probabilidades associadas são diferentes.

Para o primeiro atributo, acredita-se que os cinco valores sejam igualmente prováveis.

Para o segundo atributo, deseja-se verificar se as probabilidades são similares a outras cinco probabilidades deduzidas de uma teoria. Esta teoria afirma que a chance de observar k decai exponencialmente com k . Isto é, na segunda coluna queremos verificar se temos $\mathbb{P}(X = k) = c\theta^k$ onde $\theta \in (0, 1)$ é uma constante e c é outra constante necessária para que as probabilidades somem 1. Mostre que este modelo implica em ter as razões entre probabilidades sucessivas constantes e iguais a θ :

$$r_k = \frac{\mathbb{P}(X = k+1)}{\mathbb{P}(X = k)} = \theta$$

para $k = 1, 2, 3, 4$.

Para o terceiro atributo, existe uma outra população similar que foi exaustivamente estudada e para a qual encontrou-se o seguinte:

k	1	2	3	4	5
$\mathbb{P}(X = k)$	0.44	0.11	0.12	0.32	0.01

k	0	1	2	3	4	5	6	7	8
<i>OBS</i>	215	1485	5331	10649	14959	11929	6678	2092	342
<i>ESP</i>	165.22	1401.69	5202.65	11034.65	14627.60	12409.87	6580.24	1993.78	264.30
<i>DIF</i>	49.78	83.31	128.35	-385.65	331.40	-480.87	97.76	98.22	77.70

Tabela 3.1: Número k de filhos do sexo masculino em 53680 famílias de tamanho 8. *OBS* são os números observados na Alemanha no século XIX e *ESP* são os números esperados sob o modelo binomial. A linha *DIF* é a diferença entre as linhas *OBS* e *ESP*.

Estime as probabilidades $\mathbb{P}(X = k)$ para $k = 1, \dots, 5$ em cada coluna usando as frequências relativas de cada categoria. A seguir, verifique *informalmente* (usando gráficos ou comparações simples de tabelas de números) se os modelos para cada um dos atributos é compatível com os dados observados.

Não quero que você saia pesquisando para encontrar maneiras ótimas de resolver o problema. Também não estou esperando nem pedindo que você use o teste qui-quadrado ou Kolmogorov.

-
8. Em seu livro clássico *Statistical Methods for Research Workers*, Ronald A. Fisher, o maior gênio estatístico que já existiu, analisa alguns dados referentes ao número de filhos do sexo masculino entre famílias com um número total de filhos igual a 8. Os dados foram coletados por A. Geissler em uma região da Alemanha no período de 1876 a 1885. O livro de Fisher tem uma página na wikipedia, http://en.wikipedia.org/wiki/Statistical_Methods_for_Research_Workers.

É bem conhecido que os nascimentos do sexo masculino são ligeiramente mais numerosos do que os nascimentos do sexo feminino. Suponhamos que a probabilidade de uma criança ser do sexo masculino seja $\theta > 0.5$. Suponha que os 8 nascimentos sucessivos numa família de tamanho 8 sejam independentes. Assuma também que θ é o mesmo para todas as famílias e para os 8 nascimentos ao longo de uma sequência familiar. Então o número X de meninos numa família de tamanho 8 seguiria uma distribuição binomial: $X \sim \text{Bin}(8, \theta)$.

A linha *OBS* na Tabela 3.1 mostra o número de famílias com k filhos do sexo masculino na população de 53680 famílias com exatamente 8 filhos. A linha *ESP* mostra o número esperado de famílias com k meninos dentre os 8 filhos se o modelo binomial se aplicar.

Impressionava que algumas centenas de famílias tivessem todos os seus 8 filhos homens ou todos mulheres. Estas centenas de famílias não representavam um excesso em relação ao que se espera sob o modelo binomial? Se as famílias diferissem não só pelo acaso associado com a distribuição binomial, mas também por uma tendência por parte de alguns pais para produzir homens ou mulheres, os dados não seriam bem ajustados por uma binomial. Imagine, para tomar um exemplo muito extremo, que cada família escolhesse um valor θ para sua probabilidade de gerar meninos. Após escolher “seu” θ , ou a sua “moeda”, cada família a jogasse para cima 8 vezes de acordo com o modelo binomial. Suponha que as famílias retirassem os seus θ 's de uma urna onde houvessem apenas dois tipos de valores e em iguais proporções: $\theta = 0.01$ e $\theta = 0.99$. Neste caso, veríamos nos dados um acúmulo

de famílias nas categorias extremas (com 0 ou 1 ou então com 8 ou 7 filhos), sem muitas famílias com número intermediário de filhos homens.²¹

É claro que não vemos nada tão extremo nos dados acima. Fisher escreve: *The observed series differs from expectation markedly in two respects: one is the excess of unequally divided families; the other is the irregularity of the central values, showing an apparent bias in favour of even values. No biological reason is suggested for the latter discrepancy, which therefore detracts from the value of the data. The excess of the extreme types of family may be treated in more detail by comparing the observed with the expected ...*

- Para verificar sua compreensão do problema, obtenha os números esperados que estão na Tabela 3.1.
- Calcule a estatística qui-quadrado neste problema (você deve obter um valor de $X^2 = 91.87$).
- Qual a distribuição de referência desta estatística?
- Qual o p-valor associado com esta estatística? (DICA: use *pchisq* para obter o p-valor igual a 0.0 (numa aproximação até 15 casas decimais))

-
9. No livro clássico *Statistical Methods for Research Workers* de Ronald A. Fisher, ele apresenta alguns dados de contagens de leveduras de cerveja (fungos) obtidas através da observação humana num microscópio (hemocitômetro). Um área de 1 milímetro quadrado foi dividido em 400 quadrados de área igual e foi contado o número de fungos em cada um deles. A tabela 3.2 mostra quantos quadradinhos tiveram k fungos. Ela também mostra os números esperados segundo um modelo de Poisson.

Obtenha a coluna de números esperados segundo o modelo Poisson. Em seguida, use o teste qui-quadrado para testar se os dados são compatíveis com esta hipótese.

-
10. Este exercício usa o teste de Kolmogorov. Gere 100 valores i.i.d. de uma $N(0,1)$ e teste se eles de fato vem de uma $N(0,1)$ usando o teste de Kolmogorov. Em *R*, você pode usar o comando `ks.test` para isto:

```
x = rnorm(100)
ks.test(x, "pnorm", 0, 1)
```

Repita estes comandos 1000 vezes. Colete os valores da estatística D e do p-valor nestas 1000 simulações. Faça um histograma padronizado dos 1000 valores de D . Qual é o intervalo dentro do qual você pode esperar os valores de D quando o modelo proposto coincide com o processo gerador de dados?

Faça um histograma dos p-valores obtidos. Ele deve ter uma distribuição aproximadamente uniforme no intervalo $(0,1)$. Qual a proporção dos p-valores simulados que ficaram menores que 0.05?

k	Obs	Exp
0	0	3.71
1	20	17.37
2	43	40.65
3	53	63.41
4	86	74.19
5	70	69.44
6	54	54.16
7	37	36.21
8	18	21.18
9	10	11.02
10	5	5.16
11	2	2.19
12	2	0.86
13	0	0.31
14	0	0.10
15	0	0.03
16	0	0.03
Total	400	400

Tabela 3.2: Contagens de leveduras de cerveja em 400 quadrados e valores esperados segundo modelo Poisson.

11. Um exercício simples e importante. Você deve coletar algum conjunto de dados, QUALQUER UM, com pelo menos 100 instâncias e pelo menos dois atributos. Pelo menos um dos atributos deve ser numérico. Os dados podem ser, por exemplo, os tamanhos de seus arquivos pessoais, dados de uma rede de computadores, dados extraídos de textos ou imagens. O fundamental é que você mesmo colete ou obtenha os dados. Não devem ser entregues dados de bases conhecidas e disponíveis na web em repositórios de dados.

Em seguida, você deve tentar ajustar uma distribuição, qualquer uma, a um dos atributos em seus dados usando o teste qui-quadrado ou o teste de Kolmogorov.

?? Vou ficar surpreso se você conseguir ajustar alguma coisa. Portanto, não há nenhuma expectativa da minha parte de que você consiga, nesta altura do curso, fazer um ajuste de alguma distribuição. ??

-
12. Buscar exemplo em redes com binomial - coursera - ???

-
13. Prove que $m = \mathbb{E}(Y)$ minimiza $\mathbb{E}(Y - m)^2$. Dica: soma e subtraia $\mathbb{E}(Y)$, expanda o quadrado e tome esperança de cada termo. A seguir, derive com relação a m .

-
14. O R possui uma função, `ks.test()`, que implementa o teste de Kolmogorov. Suponha que x é um vetor com n valores numéricos distintos. Então `ks.test(x,"pnorm",m,dp)`

testa se x pode vir da distribuição $N(m, dp)$, uma normal (ou gaussiana) com média $\mu = m$ e desvio-padrão $\sigma = dp$. Outras distribuições são possíveis substituindo o string "pnorm": as pré-definidas em R (veja com `?distributions`) ou qualquer outra para a qual você crie uma função que calcula a função distribuição acumulada teórica.

```
> ks.test(x, "pnorm")
```

One-sample Kolmogorov-Smirnov test

```
data: x
D = 0.0805, p-value = 0.876
alternative hypothesis: two-sided
```

A saída de `ks.test()` fornece o valor de $D_n = \max_x |\hat{F}_n(x) - F(x)|$ e o p-valor (a esta altura talvez já tenhamos aprendido o conceito de p-valor). Dissemos em sala que se $\sqrt{n}D_n$ estiver aproximadamente entre 0.4 e 1.4, podemos aceitar o modelo (não há evidência nos dados para rejeitar o modelo). Se $\sqrt{n}D_n > 1.36$, rejeitamos o modelo.

Gere alguns dados com $n = 50$ de uma normal qualquer e use a função `ks.test()` para verificar se o teste rejeita o modelo. Faça o teste de dois modos: use o modelo correto que você usou para gerar seus dados e depois use um modelo diferente deste alterando, por exemplo, o valor de μ ou σ .

15. Implemente em R uma função para calcular o resultado de um teste de Kolmogorov. A função estará restrita a testar apenas o modelo normal com média $\mu = m$ e desvio-padrão $\sigma = dp$ que devem ser fornecidas pelo usuário ou obtidas dos próprios dados (default) usando a média aritmética (comando `mean()`) e o desvio-padrão amostral (raiz da saída do comando `var()`). Não se preocupe em lidar com os casos extremos (usuário fornecer vetor nulo, fornecer vetor com valores repetidos, etc).

Observação importante: pode-se provar que $D_n = \max_x |\hat{F}_n(x) - F(x)|$ assume seu valor em um dos pontos de salto de $\hat{F}_n(x)$, ou imediatamente antes de x_i ou em x_i , onde x_i é um dos valores do vetor.

16. Obtenha duas amostras de dados estatísticos, uma com valores de v.a.'s discretas e outra com dados contínuos, ligados de alguma forma a problemas de seu interesse. Postule um modelo de dados i.i.d. para cada lista com uma distribuição de probabilidade que você acredita que possa se ajustar aos dados. Calcule a qualidade do ajuste da sua distribuição usando o teste qui-quadrado de Pearson. Exponha as dificuldades ou dúvidas práticas ou teóricas não-cobertas no curso que você talvez encontre neste exercício.

17. *Distribuição de Zipf.* Considere um longo texto (ou vários textos juntos). Algumas palavras aparecem pouco, são raras. Outras aparecem com muita frequência. Por exemplo,

no português brasileiro (ver <http://www.linguateca.pt/>), temos a seguinte tabela de frequência (ocorrência aproximada da palavra em cada bloco de texto de um milhão de palavras):

palavra	posto (rank)	frequência (por 1M)
de	1	79607
a	2	48238
ser	27	4033
amor	802	174
chuva	2087	70
probabilidade	8901	12
interativo	14343	6
algoritmo	21531	3

Imagine o seguinte experimento: escolha uma palavra ao acaso do texto. Note que escolhemos do texto, onde algumas palavras aparecem repetidas várias vezes, e não de uma lista de palavras distintas (como num índice, onde cada palavra aparece apenas uma vez). Seja Y a v.a. indicando o posto (ou rank) da palavra escolhida. Por exemplo, se a palavra escolhida é *amor* o valor de Y é 802. Se a palavra escolhida é *de*, o valor de Y é 2. É óbvio que os valores de Y estão concentrados em valores baixos: com maior probabilidade devem ser escolhidas as palavras que aparecem com mais frequência no texto. Qual a distribuição de Y ? Depende da língua? Depende do assunto tratado na coleção de textos? Estudiosos dizem que um modelo de distribuição de probabilidade ajusta-se a uma ampla classe de problemas: a distribuição de Zipf.

Os valores possíveis de Y são iguais a $1, 2, 3, \dots, N$. Às vezes, o número N não é conhecido ou é simplesmente ignorado pois jogamos fora a informação sobre as palavras muito pouco frequentes (com posto muito alto). A distribuição de Zipf diz que

$$\mathbb{P}(Y = k) = \frac{c}{k^\theta} \quad (3.2)$$

onde θ é uma constante que varia de problema para problema e c é a constante de normalização. Isto é, como $1 = \sum_k \mathbb{P}(Y = k)$, teremos

$$c = \frac{1}{\sum_k 1/k^\theta}$$

O fato fundamental na distribuição de Zipf é que as probabilidades decaem de forma polinomial com k . O parâmetro θ costuma ser um valor próximo de 1.

Se tomarmos logaritmo dos dois lados de (3.2), temos

$$\log(\mathbb{P}(Y = k)) = \log(c) - \theta \log(k) = a - \theta \log(k).$$

Assim, no caso de uma distribuição de Zipf, um plot de $\log(\mathbb{P}(Y = k))$ versus $\log(k)$ deveria exibir uma linha reta cuja inclinação seria o negativo do parâmetro θ (tipicamente, -1).

Na tabela acima, temos a frequência n_k em 1 milhão de algumas palavras do português, bem como seu rank. Probabilidades são aproximadamente a frequência relativa de modo que $\mathbb{P}(Y = k) \approx n_k/10^6$ e portanto

$$\log(c) - \theta \log(k) = \log(\mathbb{P}(Y = k)) \approx \log(n_k/10^6)$$

o que implica em

$$\log(n_k) \approx (\log(c) - 6 \log(10)) - \theta \log(k)$$

Assim, para checar se uma distribuição de Zipf ajusta-se aos dados, podemos fazer um scatter-plot dos pontos $(\log(k), \log(n_k))$ e verificar se eles caem aproximadamente ao longo de uma linha reta. Ajustando uma reta (por mínimos quadrados ou regressão linear, por exemplo) podemos encontrar uma estimativa para θ .

- Da Wikipedia: *Zipf's law states that given some corpus of natural language utterances, the frequency of any word is inversely proportional to its rank in the frequency table. Thus the most frequent word will occur approximately twice as often as the second most frequent word, three times as often as the third most frequent word, etc. For example, in the Brown Corpus of American English text, the word "the" is the most frequently occurring word, and by itself accounts for nearly 7% of all word occurrences (69,971 out of slightly over 1 million). True to Zipf's Law, the second-place word "of" accounts for slightly over 3.5% of words (36,411 occurrences), followed by "and" (28,852). Only 135 vocabulary items are needed to account for half the Brown Corpus.* Explique como a equação (3.2) implicaria que *the frequency of any word is inversely proportional to its rank*. A resposta é simples e direta, não tem nada sutil ou complicado aqui.
- Use os dados da tabela acima para fazer o scatter-plot dos pontos $(\log(k), \log(n_k))$. Em R, basta fazer `summary(lm(y ~ x))` onde `y` e `x` são os vetores com $\log(n_k)$ e $\log(k)$, respectivamente. Qual o valor da inclinação? RESP: -0.999.
- Um excelente material sobre Zipf e Pareto: <http://arxiv.org/abs/cond-mat/0412004>

-
18. A distribuição de Poisson é muito usada para modelar dados de contagens. Por exemplo, ela pode ser usada para modelar o número de mortes por certa doença numa região durante um ano, o número de falhas num software descobertas num certo período de desenvolvimento, o número de requisições de um certo recurso numa rede, etc.

Se Y tem distribuição de Poisson então ela possui um número infinito de valores possíveis: $0, 1, 2, \dots$ com probabilidades associadas dadas por $\mathbb{P}(Y = k) = \lambda^k \exp(-\lambda)/k!$ onde $\lambda > 0$ é um parâmetro controlando a forma da distribuição.

- Se o conjunto de valores possíveis é infinito, como é possível explicar que a soma $\sum_k \mathbb{P}(Y = k)$ das probabilidades não seja infinita?
- Mostre que $\sum_k \mathbb{P}(Y = k) = 1$ (consulte qq livro de probab ou a web)

- Mostre que $\mathbb{E}(Y) = \sum_k k\mathbb{P}(Y = k) = \lambda$ (isto mostra qual o significado do parâmetro λ). Pode-se mostrar (não precisa fazer isto) que a variância $\mathbb{V}(Y) = \lambda$. Isto é, no caso Poisson, $\mathbb{E}(Y) = \lambda = \mathbb{V}(Y)$.
- Supondo que $\lambda = 0.3$, use a função `dpois` do R para calcular $\mathbb{P}(Y = 2)$.
- Supondo que $\lambda = 0.3$, use a função `ppois` do R para calcular $\mathbb{P}(Y \geq 3)$.
- Supondo que $\lambda = 0.3$, use a função `rpois` do R para simular 3 mil valores de $Y \sim \text{Poisson}(0.3)$. Com a amostra gerada, use a proporção de vezes em que $Y \geq 3$ para estimar $\mathbb{P}(Y \geq 3) = 0.0036$. Verifique também que a média aritmética dos 300 valores gerados é aproximadamente igual a λ .
- Repita os itens 4 a 6 usando $\lambda = 3$.

19. A distribuição Gama é muito flexível, adotando formas muito distintas dependendo de seus parâmetros α e β (veja na wikipedia). Existe mais de uma forma de parametrizar a distribuição gama. A mais comum (e usada como default pelo R) é aquela em que a densidade de probabilidade de $Y \sim \text{Gamma}(\alpha, \beta)$ é dada por $f(x) = 0$ se $x \leq 0$ e, se $x > 0$, por

$$f(x) = cx^{\alpha-1} \exp(-\beta x).$$

onde c é uma constante para que a área total debaixo da curva seja 1.

- Usando `rgamma` do R, gere 350 valores de uma gama com $\alpha = 9$ e $\beta = 3$. Faça um histograma padronizado destes números gerados e, usando `lines` e `dgamma`, sobreponha a curva densidade. Ficam parecidos?
- Teoricamente, temos $\mathbb{E}(Y) = \alpha/\beta = 9/3 = 3$. Calcule a média aritmética dos 350 números gerados e compare com $\mathbb{E}(Y)$. Eles ser aproximadamente iguais.
- Usando `pgamma`, faça um gráfico da distribuição acumulada teórica de uma gama com $\alpha = 9$ e $\beta = 3$. Sobreponha o gráfico da distribuição acumulada empírica: eles se parecem? O script abaixo foi usado para algo similar num dos slides do curso:

```
set.seed(1)
dados <- rnorm(30, 10, 2) # gera 30 valores de uma N(10,2)
Fn <- ecdf(dados) # calcula a funcao dist acum empirica
plot(Fn, verticals= T, do.p=F, main="", xlab="x", col="blue", xlim=c(3, 16))
x <- seq(3, 16, by=0.1)
y <- pnorm(x, 10, 2) # calcula a acumulada teorica de uma N(10,2)
lines(x,y, col="red")
```

20. Gere $n = 100$ valores aleatórios em $(0, 1)$ e guarde num vetor x . Repita isto e guarde o resultado em y . Faça um gráfico de dispersão (scatterplot) de x versus y (o objetivo é apenas fazer um scatterplot qualquer). Como você poderia gerar y se você quisessem que seus valores dependessem de alguma forma do valor correspondente x ?

21. Seja Y uma v.a. com valor esperado $\mathbb{E}(Y) = \mu$ e variância $\mathbb{V}(Y) = \sigma^2$. Prove a desigualdade de Tchebyshev: $\mathbb{P}(|Y - \mu| \geq k\sigma) \leq 1/k^2$. OBS: Qualquer livro de probabilidade (ou a web) possui a demonstração.
-

22. Aplique a desigualdade de Tchebyshev com $k = 1, 2, 4, 6, 10$. O que acontece com a cota (bound) dado pela desigualdade? Como é o seu decaimento?
-

23. Seja $X \sim \exp(1/3)$. Isto é, $X \sim \exp(\lambda)$ com $\lambda = 1/3$. Calcule $\mathbb{E}(X)$, $\mathbb{V}(X)$, $\mathbb{F}(x)$ e $\mathbb{P}(X > 3)$.

RESP: $\mathbb{E}(X) = 3$; $\mathbb{V}(X) = 3^2$; $\mathbb{F}(x) = 0$ se $x \leq 0$ e $\mathbb{F}(x) = 1 - \exp(-x/3)$, para $x > 0$; $\mathbb{P}(X > 3) = \exp(-3/3) = 0.37$. Veja que, embora o valor esperado $\mathbb{E}(X)$ seja igual a 3, temos $\mathbb{P}(X > 3) < 1/2$. No caso geral, $\mathbb{E}(X) = 1/\lambda$; $\mathbb{V}(X) = 1/\lambda^2$; $\mathbb{F}(x) = 0$ se $x \leq 0$ e $\mathbb{F}(x) = 1 - \exp(-\lambda x)$, para $x > 0$; $\mathbb{P}(X > 3) = \exp(-3\lambda)$.

24. Use o método da transformada inversa com a função $\mathbb{F}(x)$ calculada no exercício anterior para gerar 1000 valores aleatórios de $X \sim \exp(1/3)$. Faça um histograma (normalizado com área 1) dos 1000 valores gerados e sobreponha o gráfico da função densidade de probabilidade (dada por $f(x) = 0$ se $x \leq 0$ e $f(x) = (1/3)\exp(-x/3)$, se $x > 0$). Calcule a média aritmética e compare com o valor teórico $\mathbb{E}(X) = 3$.

Gere uma *segunda* amostra de tamanho 1000 e recalcule a média aritmética. Verifique que o valor teórico $\mathbb{E}(X) = 3$ permanece o mesmo mas que a média aritmética varia de amostra para amostra.

RESP: Como $\mathbb{F}(x) = 1 - \exp(-x/3)$ então $X = -3 \log(1 - U) \sim \exp(1/3)$ se $U \sim U(0, 1)$. Script em R (§ significa comentário):

```
x <- -3*log(1-runif(1000)) # runif(n) gera n v.a.'s U(0,1)
mean(x)
hist(x, prob=T) # normaliza com o argumento prob=

# sobrepondo o grafico da densidade:
grid <- seq(0, 20, length=100) # vetor com pontos no eixo x
y <- dexp(grid, rate=1/3) # densidade exp(1/3) nos pontos de grid

lines(grid, y, type = "l") # adiciona linhas ao grafico anterior

# Podemos tambem calcular diretamente o valor da funcao densidade
hist(x, prob=T) #normaliza com o argumento prob=
y <- 1/3 * exp(-grid/3)
lines(grid, y, type = "l") # adiciona ao grafico anterior

# R tem varias funcoes para gerar v.a.'s de distribuicoes conhecidas
x <- rexp(1000, rate=1/3) # gera 1000 v.a.'s exp(1/3)
x <- rnorm(1000, m=10, sigma=2) # gera 1000 v.a.'s N(10, 2^2)
x <- rgamma(1000, alpha=3, beta =2) # 1000 Gamma(3,2)
```

25. X é uma v.a. com distribuição Pareto com parâmetros m e $\alpha = 2$. Isto é,

$$f_X(x) = \begin{cases} 0 & \text{se } x \leq m \\ c/x^{\alpha+1} & \text{se } x > m \end{cases}$$

onde a constante de integração c é dada por $c = \alpha m^\alpha$. Calcule $\mathbb{F}(x) = \mathbb{P}(X \leq x)$.

Calcule também $\mathbb{E}(X)$ para $\alpha > 1$ (a integral $\mathbb{E}(X)$ não existe se $0 < \alpha \leq 1$).

RESP: $\mathbb{F}(x) = 0$ se $x \leq m$ e $\mathbb{F}(x) = 1 - (m/x)^\alpha$. Temos

$$\mathbb{E}(X) = \int_m^\infty x c/x^{\alpha+1} dx = \int_m^\infty c/x^\alpha dx = c/((1-\alpha)x^{\alpha-1}) = \alpha m/(\alpha-1)$$

26. Usando o método da transformada inversa, gere 1000 valores de uma Pareto com $\alpha = 4$ e $m = 1$. Gere outros 1000 valores de uma Pareto com $\alpha = 2, 1$, e 0.5 . Qual o efeito de diminuir α em direção a zero? Compare $\mathbb{E}(X)$ com os valores gerados nos casos em que $\alpha > 1$.

RESP: Como $\mathbb{F}(x) = 1 - (m/x)^\alpha$ para $x > m$, temos a v.a. $X = \mathbb{F}^{-1}(U) = m/(1-U)^{1/\alpha}$ com distribuição Pareto com parâmetros m e α . Se $m = 1$ e $\alpha = 2$, temos $X = 1/\sqrt{1-U}$. Tomando $\alpha = 1$, temos $X = 1/(1-U)$ e, se $\alpha = 1/2$, temos $X = 1/(1-U)^2$. Usando um script R (Note a ESCALA dos 4 gráficos abaixo e compare $\mathbb{E}(X)$ com os valores gerados nos casos em que $\alpha > 1$):

```
x4 <- 1/(1-runif(1000))^(0.25) # gera 1000 v.a.'s Pareto m=1 e alpha=4
x2 <- 1/sqrt(1-runif(1000)) # gera 1000 v.a.'s Pareto m=1 e alpha=2
x1 <- 1/(1-runif(1000)) # Pareto m=1 e alpha=1
x05 <- 1/(1-runif(1000))^2 # Pareto m=1 e alpha=1/2
par(mfrow=c(2,2)) # divide a janela grafica numa matriz 2x2
plot(x4) # grafico da **sequencia** de valores gerados
plot(x2); plot(x1); plot(x05)
```

27. Verifique como valores extremos são facilmente gerados pela Pareto. Tomando $m = 1$ e $\alpha = 0.5$, calcule $\mathbb{P}(X \leq 10)$. A seguir, calcule $\mathbb{P}(X > 10^k)$ para $k = 2, 3, 4, 5$.

RESP: Com $m = 1$ e $\alpha = 0.5$, $\mathbb{F}(x) = \mathbb{P}(X \leq x) = 1 - 1/\sqrt{x}$ para $x > 0$. Assim, $\mathbb{P}(X \leq 10) = 0.68$. Isto é, 68% dos valores gerados numa simulação serão no máximo iguais a 10. Temos $\mathbb{P}(X \leq 100) = 1/\sqrt{100} = 0.1$. Portanto, 10% dos valores gerados numa simulação devem maiores que 100. Muito maiores? $\mathbb{P}(X \leq 1000) = 0.03$ ou 3% dos valores gerados são maiores que 1000. Temos $\mathbb{P}(X \leq 10000) = 0.01$ e $\mathbb{P}(X \leq 100000) = 0.003$ e finalmente $\mathbb{P}(X \leq 1000000) = 0.001$. Assim, 1 em cada 1000 valores gerados serão maiores que 1 milhão mesmo que a maioria (68%) sejam menores que 10.

28. Durante a IIWW, foram mapeados os locais atingidos por bombas ao sul de Londres.²⁹ A área foi dividida em $n = 576$ quadradinhos, cada um com $0.25km^2$. O número total de bombas que atingiu a região foi 537. Seja X_i o número de bombas no quadradinho i . Vimos em sala que um bom modelo para as contagens X_1, \dots, X_{576} supõem que elas sejam instâncias de uma variável aleatória Poisson(λ).

Neste exercício vamos verificar que o teste qui-quadrado permite eliminar escolhas incorretas para a distribuição de X . Vamos supor que a contagem de bombas ACRESCIDA DE UMA UNIDADE siga a distribuição logarítmica com parâmetro $\theta \in [0, 1]$, que é definida da seguinte forma:

- Valores possíveis: $\{1, 2, 3, \dots\}$
- $\mathbb{P}(Y = k) = -1/\log(1 - \theta) \theta^k/k$ para $k = 1, 2, \dots$

com $\mathbb{E}(Y) = -1/\log(1 - \theta) \theta/(1 - \theta)$.

OBS: Você deve ter notado que somamos 1 a X pois o número de bombas pode ser zero e a distribuição logarítmica começa de 1. Isto é, supomos que $Y = X + 1$ siga a distribuição logarítmica.

- Estime $\mathbb{E}(Y)$ usando a média aritmética $\bar{X} + 1$ e obtenha assim uma estimativa de θ . RESP: $\hat{y} = 1 + 0.9323$ e $\hat{\theta} = 0.696$.
- A seguir, preencha os valores esperados do número de quadrados com k bombas na tabela abaixo. Por exemplo, o número esperado com 0 bombas é dado por

$$576 \times \mathbb{P}(X = 0) = 576 \times \mathbb{P}(Y = 1) = -\frac{1}{\log(1 - \hat{\theta})} \hat{\theta}^1/1 = 576 \times 0.585 = 336.96$$

k	0	1	2	3	4	5 e acima
Obs	229	211	93	35	7	1
Esp	336.96	??	??	??	??	??

Para a última categoria, use $\mathbb{P}(X \geq 5) = 1 - \sum_{j=1}^4 \mathbb{P}(X = j)$.

- Embora seja óbvio que a distribuição logarítmica não se ajusta a estes dados, calcule a estatística qui-quadrado a partir das diferenças entre os valores observados e esperados nesta tabela.
- Obtenha o p-valor com o comando `1-pchisq(qq, df)` onde qq é o valor da estatística qui-quadrado e df é o número de graus de liberdade.

-
29. Use os dados das contagens mensais de cirurgias cardiovasculares infantis em hospitais discutido em sala de aula para verificar se, para cada hospital, podemos assumir que as suas contagens Y_1, Y_2, \dots, Y_n sejam i.i.d. e sigam uma distribuição de Poisson com parâmetro λ . Os dados estão no arquivo `cirurgia.txt`. Use o script `cirurgia.R` para iniciar sua análise. Eu deixei todos os comandos necessários para fazer a análise com o Hospital 1. Considere os seguintes exercícios:

- Faça um teste qui-quadrado para o SÉTIMO hospital da tabela (linha 7). Usando a função `pchisq`, calcule o p-valor associado com a hipótese ou modelo assumido (isto é, i.i.d. $\text{Poisson}(\lambda)$)).
 - EXERCÍCIO OPCIONAL, BONUS (PONTO EXTRA SE ENTREGAR): Generalize o código anterior fazendo uma função em R para executar estes cálculos para cada um dos hospitais da tabela. Você vai precisar considerar a criação das classes de valores que vai variar de hospital para hospital.
 - O teste qui-quadrado é interessante porquê é um teste geneérico, pode ser usado para com qualquer modelo para a distribuição de uma variável aleatória. Entretanto, supondo que a distribuição é uma Poisson, podemos explorar alguns aspectos específicos **desta** distribuição para avaliar se o modelo ajusta-se aos dados observados. Uma dessas formas, puramente visual, é a seguinte:
 - Mostre que, se $Y \sim \text{Poisson}(\lambda)$, então $r_k = \mathbb{P}(Y = k)/\mathbb{P}(Y = k + 1) = (k + 1)/\lambda$.
 - Tomando logaritmo natural dos dois lados da igualdade acima, temos $\log(r_k) = -\log(\lambda) + \log(k + 1)$. Assim, um gráfico de $\log(r_k)$ versus $\log(k + 1)$ deveria mostrar uma linha reta com coeficiente angular 1 e intercepto que vai variar com o valor de λ .
 - Estime r_k pela razão $f_k = n_k/n_{k+1}$ onde n_k é o número de elementos da amostra que são iguais a k . Faça o gráfico de $\log(f_k)$ versus $\log(k + 1)$ e verifique se aparece aproximadamente uma reta de inclinação 1. Faça isto APENAS PARA O HOSPITAL 7 para entregar (se quiser fazer mais, fique à vontade!)
-
- Uma outra forma simples de verificar se o modelo de Poisson ajusta-se aos dados que estão num vetor y é o teste de dispersão:
 - Numa v.a. Poisson, o valor esperado $\mathbb{E}(Y) = \lambda$ é igual à variância da v.a. $\mathbb{V}(Y)$.
 - Estime $\mathbb{E}(Y)$ pela média aritmética m das observações no vetor y (usando `mean(y)` no R).
 - Estime $\mathbb{V}(Y)$ pela variância v da amostra (usando `var(y)` no R).
 - Calcule a razão v/m , que deveria ser próxima de 1 se o modelo Poisson é correto.
 - Como avaliar se v/m está próximo de 1? Se n é o comprimento do vetor de contagens y , pode-se mostrar que $(n - 1)v/m$ segue aproximadamente uma distribuição qui-quadrado (ela aqui de novo) com $n - 1$ graus de liberdade.
 - Usando os dados do HOSPITAL 7, calcule o p-valor deste teste.
-

30. Os primeiros 608 dígitos da expansão decimal do número π tem as seguintes frequências:

k	0	1	2	3	4	5	6	7	8	9
Obs	60	62	67	68	64	56	62	44	58	67
Esp	??	??	??	??	??	??	??	??	??	??

Estes dados são compatíveis com a suposição de que cada dígito é escolhido de forma completamente aleatória? Isto é, de acordo com uma distribuição uniforme discreta sobre os possíveis dígitos?

-
31. Vamos usar a desigualdade de Tchebychev,

$$\mathbb{P}\left(\left|\frac{X - \mu}{\sigma}\right| \leq \delta\right) \leq \frac{1}{\delta^2}$$

para gerar um intervalo de predição para X . Suponha que X possua uma distribuição de probabilidade arbitrária com $\mathbb{E}(X) = \mu = 120$ e $Var(X) = \sigma^2 = 10^2$. Usando a desigualdade de Tchebychev, mostre que o intervalo $(120 \pm 45) = (75, 165)$ deverá conter pelo menos 95% dos dados gerados de X , qualquer que seja a distribuição de X .

Suponha agora que sabemos algo mais sobre a distribuição de X . Este conhecimento adicional reduz substancialmente a incerteza acerca dos valores gerados da distribuição. Agora, usando o comando `qnorm` do R, mostre que o intervalo que conterá 95% dos valores de uma amostra de X é $120 \pm 1.96 * 10 = (100.4, 139.6)$.

32. Gere uma amostra de tamanho $n = 100$ de uma normal com $\mu = 10$ e $\sigma = 1$. Faça o `qqplot` da amostra usando o comando `qqnorm(x)`. Repita isto 10 vezes. O gráfico ficou na forma de uma linha reta todas as vezes?

Refaça este exercício gerando os seus dados de uma distribuição Cauchy com parâmetro de locação $\mu = 10$ e escala $\sigma = 1$: `rcauchy(100, 10, 1)`. Para comparar a distribuição $N(10, 1)$ e a $\text{Cauchy}(10, 1)$, use os seguintes comandos:

```
x <- seq(4, 16, by=0.01)
yc <- dcauchy(x, location = 10, scale = 1)
yn <- dnorm(x, mean = 10, sd = 1)
plot(x, yn, type="l")
lines(x, yc, lty=2)
legend("topright", c("normal", "Cauchy"), lty=1:2)
qqnorm(rcauchy(100, 10, 1))
```

33. No site <http://www.athenasc.com/prob-sup.html> você encontra exercícios SUPLEMENTARES de Bertsekas and Tsitsiklis. Considerando o arquivo relacionado ao capítulo 3, faça os seguintes exercícios: 2, 8, 9, 21(b).
-

34. Obtenha os dados de fragmentos de vidro coletados pela polícia forense do livro *All of Statistics* no site: <http://www.stat.cmu.edu/~larry/all-of-statistics/index.html> Estime a densidade da primeira variável (refractive index) usando um histograma e um estimador de densidade baseado em kernel. Experimente com diferentes bins para o histograma e bandwidths para o kernel. Veja como fazer isto em R no endereço: <http://www.statmethods.net/graphs/density.html>.

- 32 O comando básico para a estimativa de kernel é da forma: `density(x, kernel = "gauss")` onde o kernel gaussiano é o escolhido. Outras opções incluem: "epanechnikov", "rectangular", "triangular", "biweight", "cosine", "optcosine". A escolha default do bandwidth é calculada com a seguinte fórmula (conhecida como a regra de bolo de Silverman):

$$h = 0.9 \min \left\{ \hat{\sigma}, \frac{R}{1.34} \right\} n^{-1/5}$$

onde $R = q_{0.75} - q_{0.25}$ é a diferença entre o 1º e o 3º quartis e que pode ser calculado usando-se, por exemplo, o comando $IQR(x)$. Verifique qual o valor deste bandwidth h de Silverman e, a seguir, calcule a estimativa de kernel usando o dobro e a metade deste valor (isto é, usando $2h$ e $h/2$). Compare os resultados obtidos visualmente. Qual parece representar melhor a densidade $f(x)$ dos dados?

Capítulo 4

Transformação de uma v.a.e Monte Carlo

Transformação de uma v.a.

Em probabilidade e análise de dados, é fundamental trabalhar com transformações de v.a.'s. Isto é, Y é uma v.a. transformada por uma função matemática a partir da v.a. X . Temos $Y = h(X)$. Por exemplo, $Y = h(X) = X^2$ ou $Y = h(X) = \sqrt{X} + \log(X)$. A v.a. Y passa a ter uma distribuição de probabilidade $f(y)$ sobre seus valores possíveis induzida pela distribuição de probabilidade de X (representada pela densidade $f(x)$) e pela transformação h . O cálculo de probabilidades de eventos associados com a ocorrência de $Y = h(X)$ pode ser reduzido ao cálculo de probabilidades de eventos associados com a ocorrência de X .

Por exemplo, seja X o lado de um quadrado aleatório. A v.a. X é selecionada de uma distribuição $\text{Unif}(0, 1)$. Isto é, a probabilidade de que X caia num intervalo (a, b) cotido em $(0, 1)$ é o comprimento $b - a$ do intervalo. A área do quadrado aleatório formado com lado X é a v.a. $Y = X^2$.

Qual a distribuição de Y ? Como os valores possíveis de X formam o intervalo $(0, 1)$, os valores possíveis de $Y = X^2$ também formam o intervalo $(0, 1)$. Entretanto, as probabilidades associadas da v.a. X e de Y são bem diferentes. Enquanto X tem uma distribuição $U(0, 1)$, a distribuição de Y não é uniforme. Por exemplo, o intervalo $(0, 0.1)$ e o intervalo $(0.9, 1)$ possuem probabilidades diferentes sob a distribuição de Y embora tenham o mesmo comprimento. Outra consequência é que, em geral, teremos $\mathbb{E}(Y) = \mathbb{E}(X^2) \neq (\mathbb{E}(X))^2$. Em palavras mais memoráveis: a esperança de $g(X)$ não é g da esperança de X . Ou, em geral, $\mathbb{E}(g(X)) \neq g(\mathbb{E}(X))$.

Outro exemplo relevante: seja T o tempo de vida a partir de um indivíduo adulto a partir de seus 22 anos. A distribuição de T é conhecida pelos atuários a partir de dados estatísticos da idade ao morrer de grandes populações. A distribuição é um bom modelo para este tipo de dado. O indivíduo quer adquirir um seguro de vida que pague 100 mil reais a beneficiários (esposa e filhos) no caso de seu falecimento em qualquer momento do futuro. Qual o valor que a seguradora deve cobrar hoje desse indivíduo para cobrir os seus custos com o pagamento aos

beneficiários? O pagamento é certo mas o momento do pagamento no futuro é aleatório. Vamos ignorar as despesas e lucros da seguradora, queremos apenas cobrir o custo do pagamento do benefício no momento de morte.

Isto complica as coisas pois dinheiro tem valor no tempo. Suponha que a taxa de juros anual estimada para os próximos anos seja δ . Por exemplo, $\delta = 0.05$ significa uma taxa de juros de 5% ao ano. Um capital de C unidades auferindo juros anuais de δ acumula um total de $Ce^{\delta t}$ ao final de t anos. Para pagar 100 mil reais daqui a t anos, este tipo de cálculo financeiro permite à seguradora saber quando deveria cobrar hoje de um indivíduo que vai viver t anos. Para pagar 100 dentro de t anos ela deve cobrar Y de modo que Y posto a juros por t anos forneça 100. Isto é, $Ye^{\delta t} = 100$, ou $Y = 100e^{-\delta t}$. A quantidade $\exp(-\delta t)$ é chamada de fator de desconto e valor necessário hoje para acumular um certo total no futuro é chamado de valor presente deste total futuro. Por exemplo, com $\delta = 0.05$, o valor presente de 100 mil reais a serem pagos daqui a 27 anos é igual a $100e^{-0.05 * 27} = 25.9240$, ou 26 mil reais aproximadamente.

Se o tempo de vida futuro do indivíduo fosse conhecido aos 22 anos de idade, o cálculo do valor presente dos 100 mil reais a serem pagos no futuro seria trivial. Mas sabendo quanto ainda vai viver, o indivíduo não teria interesse em fazer seguro. Se fosse viver por muito tempo, não precisaria fazer seguro. Se fosse viver pouco tempo, teria de pagar praticamente os mesmos 100 mil reais que deseja que sua família receba.

Mas a situação mais realista em que o tempo de vida futuro é desconhecido para cada indivíduo exige que a seguradora pense estatisticamente. De um indivíduo com tempo de vida T , ela precisa cobrar pelo menos o valor presente $Y = 100e^{-\delta T}$. Mas como T é aleatório, o valor presente Y também é aleatório. Isto é, $Y = h(T)$ onde h é a função $h(t) = 100e^{-\delta t}$.

A seguradora calcula o valor esperado de Y e cobra este valor de todos os indivíduos na mesma situação. Isto é, a seguradora vai cobrar $\mathbb{E}(Y) = \mathbb{E}(100e^{-\delta T})$ de cada indivíduo de 22 anos do sexo masculino que procurar seus serviços neste tipo de seguro. Ela sabe que, para alguns dos indivíduos, aqueles que viverão pouco tempo, ela estará cobrando menos do que deveria. Por outro lado, ela sabe que vai cobrar mais do que precisa de um indivíduo que viverá muitos anos, muito além do tempo médio de vida. Pela idéia frequentista de probabilidade, sabemos que o valor presente médio de um grande número de indivíduos é aproximadamente igual ao valor teórico $\mathbb{E}(Y) = \mathbb{E}(100e^{-\delta T})$ e portanto as contas da seguradora devem ficar equilibradas.

$$\mathbb{E}(h(X))$$

O cálculo de $\mathbb{E}(h(X))$ é relativamente simples. Se X for uma v.a. contínua com densidade $f(x)$ temos

$$\mathbb{E}(h(X)) = \int_{\mathbb{R}} h(x)f(x)dx$$

Por exemplo, se $Y = h(X) = X^2$, então

$$\mathbb{E}(h(X)) = \mathbb{E}(X^2) = \int_{\mathbb{R}} x^2 f(x)dx$$

Terminar o cálculo depende de conhecer a densidade $f(x)$ mas agora este cálculo é um problema³⁵ puramente matemático.

Se X for uma v.a. discreta com valores possíveis $\{x_1, x_2, \dots\}$ e probabilidades associadas $\{p(x_1), p(x_2), \dots\}$ temos

$$\mathbb{E}(h(X)) = \sum_{x_i} h(x_i)p(x_i)$$

Por exemplo, se $h(x) = \sqrt{x}$ e se $X \sim \text{Poisson}(\lambda = 2.3)$ então

$$\mathbb{E}(h(X)) = \mathbb{E}(\sqrt{X}) = \sum_{k=0}^{\infty} \sqrt{k} \frac{2.3^k e^{-2.3}}{k!}$$

Não existe fórmula conhecida para esta série, ela tem de ser calculada numericamente.

Distribuição de $Y = h(X)$

Muitas vezes, apenas o valor esperado de $Y = h(X)$ não é suficiente. Podemos querer saber, por exemplo, a probabilidade de ter um valor $Y \leq 2$ ou qualquer outro valor arbitrário: $\mathbb{P}(Y \leq 2)$ ou, em termos mais gerais, $\mathbb{F}_Y(y) = \mathbb{P}(Y \leq y)$ para qualquer y .

Como Y é uma função de X e a distribuição de X é conhecida, devemos ser capazes de obter a distribuição de Y a partir daquela de X e da expressão da função h .

De Murphy, seção 2.6.2, pag. 50, a função distribuição acumulada de $Y = h(X)$ é definida como

$$\mathbb{F}_Y(y) = \mathbb{P}(X \in \{x \text{ tal que } h(x) \leq y\})$$

Por exemplo, se $Y = h(X) = X^2$ então

$$\mathbb{F}_Y(y) = \mathbb{P}(X \in \{x \text{ tal que } x^2 \leq y\}) = \mathbb{P}(X \in (-\sqrt{y}, \sqrt{y})) = F_X(-\sqrt{y}) - F_X(\sqrt{y})$$

para $y > 0$.

Existe uma regra geral no caso de X e Y serem v.a.'s contínuas. Seja $Y = h(X)$. Suponha que no suporte de X a função h seja inversível com $g = h^{-1}$ e portanto $x = g(y)$. A densidade de Y no ponto y é dada por

$$f_Y(y) = f_X(g(y)) \left| \frac{dg(y)}{dy} \right|$$

O segundo fator do lado direito é o valor absoluto do jacobiano dessa transformação.

Probabilidades e variáveis aleatórias indicadoras

Toda probabilidade pode ser vista como a esperança de uma v.a. indicadora. Seja A um evento qualquer e $I_A(\omega)$ uma v.a. binária dada por

$$I_A(\omega) = \begin{cases} 1, & \text{se } \omega \in A \\ 0, & \text{caso contrário} \end{cases}$$

A função é apenas a função indicadora de que o evento A ocorreu. Para cada resultado de experimento ela retorna 1 se A ocorreu e 0 se A não ocorreu.

O evento $[I_A = 1]$ é o conjunto de elementos $\omega \in \Omega$ tais que $I_A(\omega) = 1$. Mas isto é exatamente o conjunto de $\omega \in A$. Ou seja, temos a igualdade $[I_A = 1] = A$ de dois eventos em Ω . Portanto, a probabilidade dos dois eventos é a mesma e $\mathbb{P}(I_A = 1) = \mathbb{P}(A)$.

Como a v.a. I_A é binária temos

$$\mathbb{E}(I_A) = 1 \times \mathbb{P}(I_A = 1) + 0 \times \mathbb{P}(I_A = 0) = \mathbb{P}(I_A = 1) = \mathbb{P}(A)$$

Assim, para todo evento A , podemos escrever sua probabilidade $\mathbb{P}(A)$ como a esperança $\mathbb{E}(I_A)$ da v.a. aleatória indicadora da ocorrência de A . Parece um jeito muito complicado de obter uma probabilidade mas este é um truque muito útil em situações um pouco mais complicadas. A utilidade vem da possibilidade de usar algumas propriedades conhecidas do operador \mathbb{E} , e isto torna às vezes mais fácil de manipular que o operador \mathbb{P} . Uma dessas propriedades é a linearidade da esperança: $\mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y)$.

Eventos podem ser definidos em termos de variáveis aleatórias. Por exemplo,

- $[X = 3] = \{\omega \in \Omega \text{ tais que } X(\omega) = 3\}$
- $[X \leq 3] = \{\omega \in \Omega \text{ tais que } X(\omega) \leq 3\}$
- $[1 \leq X \leq 3] = \{\omega \in \Omega \text{ tais que } X(\omega) \in [1, 3]\}$
- se $S \subset \mathbb{R}$ temos $[X \in S] = \{\omega \in \Omega \text{ tais que } X(\omega) \in S\}$

Assim, podemos estender a ideia de escrever como esperança a probabilidade de um evento associado a uma v.a. Por exemplo, seja A o evento $A = [X = 3]$ associado com a v.a. X . Então

$$\mathbb{P}(X = 3) = \mathbb{P}(A) = \mathbb{E}(I_A) = \mathbb{E}(I_{[X=3]})$$

De maneira geral, se $S \subset \mathbb{R}$ temos $\mathbb{P}(X \in S) = \mathbb{E}(I_{[X \in S]})$.

Vamos adotar outra notação alternativa:

$$I_{[X \in S]} = I_S(X)$$

Veja que $I_S(X)$ é uma v.a. Ela recebe como input um resultado $\omega \in \Omega$ e retorna um valor binário $I_{[X \in S]}(\omega) = I_S(X(\omega))$. Esta v.a. é uma transformação $h(X)$ de X . De fato, temos

$$I_S(X(\omega)) = h(X(\omega)) = \begin{cases} 1, & \text{se } X(\omega) \in S \\ 0, & \text{caso contrário} \end{cases}$$

Portanto, se tivermos a densidade $f(x)$ de uma v.a. contínua X , usando o que aprendemos no

$$\begin{aligned}
\mathbb{P}(X \in S) &= \mathbb{E}(I_S(X)) \\
&= \mathbb{E}(h(X)) \\
&= \int_{\mathbb{R}} h(x)f(x)dx \\
&= \int_{\mathbb{R}} 1_S(x)f(x)dx \\
&= \int_S 1 \times f(x)dx + \int_{\mathbb{R}-S} 0 \times f(x)dx \\
&= \int_S f(x)dx
\end{aligned}$$

Eu sei que parece estarmos dando voltas em torno do mesmo ponto mas, acredite, isto é útil. Por exemplo, suponha que $X \sim N(0, 1)$, uma gaussiana padrão que possui densidade

$$f(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x^2\right)$$

Suponha que, por alguma razão queiramos calcular a probabilidade de que $\exp(-|X|) > |X|$. Manipulação algébrica mostra que $\exp(-|x|) > x$ se, e somente se, $x \in (-a, a)$ onde $a \approx 0.5671$. Assim,

$$\mathbb{P}(\exp(-|X|) > |X|) = \int_{(-a,a)} f(x)dx = \int_{\mathbb{R}} h(x)f(x)dx = \mathbb{E}(h(X))$$

onde $h(X) = I_S(X)$ e S é o evento $[\exp(-|X|) > |X|]$.

Exercícios

- Os primeiros 608 dígitos da expansão decimal do número π tem as seguintes frequências:

k	0	1	2	3	4	5	6	7	8	9
Obs	60	62	67	68	64	56	62	44	58	67
Esp	??	??	??	??	??	??	??	??	??	??

Estes dados são compatíveis com a suposição de que cada dígito é escolhido de forma completamente aleatória? Isto é, de acordo com uma distribuição uniforme discreta sobre os possíveis dígitos?

- O R possui uma função, `ks.test()`, que implementa o teste de Kolmogorov. Suponha que x é um vetor com n valores numéricos distintos. Então `ks.test(x, "pnorm", m, dp)` testa se x pode vir da distribuição $N(m, dp)$, uma normal (ou gaussiana) com média $\mu = m$ e desvio-padrão $\sigma = dp$. Outras distribuições são possíveis substituindo o string "pnorm": as pré-definidas em R (veja com `?distributions`) ou qualquer outra para a qual você crie uma função que calcula a função distribuição acumulada teórica.

One-sample Kolmogorov-Smirnov test

```
data: x
D = 0.0805, p-value = 0.876
alternative hypothesis: two-sided
```

A saída de `ks.test()` fornece o valor de $D_n = \max_x |\hat{F}_n(x) - F(x)|$ e o seu p-valor. Dissemos em sala que se $\sqrt{n}D_n > 1.36$, rejeitamos o modelo. Caso contrário, não há muita evidência nos dados para rejeitar o modelo (não quer dizer que o modelo seja correto, apenas não conseguimos rejeitá-lo).

Gere alguns dados com $n = 50$ de uma normal qualquer e use a função `ks.test()` para verificar se o teste rejeita o modelo. Faça o teste de dois modos: use o modelo correto que você usou para gerar seus dados e depois use um modelo diferente deste alterando, por exemplo, o valor de μ ou σ .

3. Implemente em R uma função para calcular o resultado de um teste de Kolmogorov. A função estará restrita a testar apenas o modelo normal com média $\mu = \mathbf{m}$ e desvio-padrão $\sigma = \mathbf{dp}$ que devem ser fornecidas pelo usuário ou obtidas dos próprios dados (default) usando a média aritmética (comando `mean()`) e o desvio-padrão amostral (raiz da saída do comando `var()`). Não se preocupe em lidar com os casos extremos (usuário fornecer vetor nulo, fornecer vetor com valores repetidos, etc).

Observação importante: pode-se provar que para encontrar $D_n = \max_x |\hat{F}_n(x) - F(x)|$ basta varrer os pontos de salto de $\hat{F}_n(x)$, olhando o valor de $\hat{F}_n(x)$ imediatamente antes de x_i ou no próprio ponto x_i , onde x_i é um dos valores do vetor de dados observados.

4. Seja $Y = h(X)$ onde $X \sim \text{Unif}(0, 1)$. A função $h(x)$ é mostrada no gráfico da Figura 4.1. A partir dessa figura, é possível obter aproximadamente os valores da f.d.a. $\mathbb{F}_Y(y)$ sem fazer nenhum cálculo explícito, apenas no olhometro. Dentre as opções abaixo, decida qual o valor que melhor aproxima $\mathbb{F}_Y(y)$.

- $\mathbb{F}_Y(0.9)$ é aproximadamente igual a: (a) zero, (b) 0.01 (c) 0.5 (d) 0.8 (e) 0.95 (f) um.
- $\mathbb{F}_Y(1.1)$ é aproximadamente igual a: (a) zero, (b) 0.01 (c) 0.5 (d) 0.8 (e) 0.95 (f) um.
- $\mathbb{F}_Y(1.8)$ é aproximadamente igual a: (a) zero, (b) 0.01 (c) 0.5 (d) 0.8 (e) 0.95 (f) um.
- $\mathbb{F}_Y(2.1)$ é aproximadamente igual a: (a) zero, (b) 0.01 (c) 0.5 (d) 0.8 (e) 0.95 (f) um.

5. Transformação de v.a.'s: Seja X o lado de um quadrado aleatório. A v.a. X é selecionada de uma distribuição $\text{Unif}(0, 1)$. A área do quadrado formado com lado X é a v.a. $Y = X^2$.

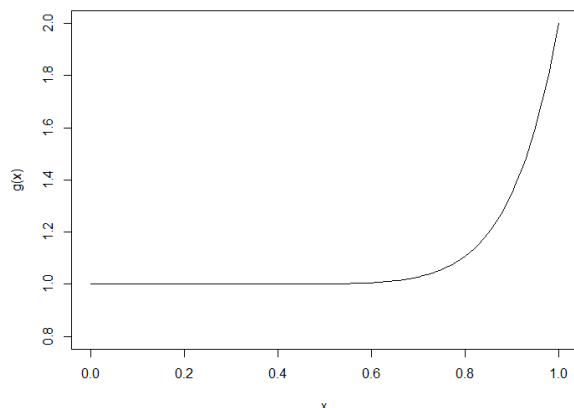


Figura 4.1: Gráfico da função $h(x)$ usada para criar a v.a. $Y = h(X)$ onde $X \sim \text{Unif}(0, 1)$.

- Calcule o comprimento esperado do lado do quadrado $\mathbb{E}(X)$.
- Obtenha também a área esperada $\mathbb{E}(Y)$. É verdade que $\mathbb{E}(Y) = (\mathbb{E}(X))^2$? Ou seja, a área esperada $\mathbb{E}(Y)$ é igual à $(\mathbb{E}(X))^2$, a área de um quadrado cujo lado tem comprimento igual ao comprimento esperado?
- Qual a distribuição de Y ? Isto é, obtenha $F_Y(y)$ para $y \in \mathbb{R}$. item Derive $F_Y(y)$ para obter a densidade $f_Y(y)$ e faça seu gráfico. Qual a região onde mais massa de probabilidade é alocada? O que é mais provável, um quadrado com área menor que 0.1 ou maior que 0.9?

6. Refaça o exercício anterior considerando o volume aleatório $V = X^3$ do cubo aleatório formado com o lado $X \sim \text{Unif}(0, 1)$.

7. Refaça o exercício anterior considerando o volume aleatório $V = (4/3)\pi X^3$ da esfera aleatória formada com o raio $X \sim \text{Unif}(0, 1)$.

8. Supondo X contínua com densidade $f(x)$ e S um subconjunto da reta real. Complete os passos da dedução abaixo preenchendo os locais indicados por ??.

$$\begin{aligned}
 \mathbb{P}(X \in S) &= \int_{??} f(x) dx \\
 &= \int_{??} I_S(x) f(x) dx \\
 &= \mathbb{E}(I_{??}(??))
 \end{aligned}$$

9. Seja $Y = h(X) = 1 + X^{10}$ onde $X \sim U(0, 1)$, uma uniforme no intervalo real $(0, 1)$. Isto é, a probabilidade de que X caia num intervalo (a, b) contido em $(0, 1)$ é o comprimento $b - a$ do intervalo. A Figura 4.1 mostra o gráfico desta transformação.

- Os valores possíveis de Y formam o intervalo $(?, ?)$. Complete os locais marcados com “?”.
- Analisando a Figura 4.1 verifique aonde o intervalo $(0, 0.8)$ no eixo x é levado pela transformação no eixo $y = h(x)$. Faça o mesmo com o intervalo de mesmo comprimento $(0.8, 1)$. Conclua: $\mathbb{P}(Y \in (1.2, 2.0))$ é maior ou menor que $\mathbb{P}(Y \in (0, 1.1))$?
- Sejam os eventos $B = [Y < ?]$ e $A = [X < 1/\sqrt[10]{2}]$, onde $1/\sqrt[10]{2} \approx 0.933$. Os eventos A e B devem ser são iguais. Qual o valor de “?”?
- Considerando a distribuição de Y , calcule $F_Y(y) = \mathbb{P}(Y \leq y)$ para qualquer $y \in \mathbb{R}$ mapeando o evento $[Y \leq 1/2]$ e um evento equivalente $[X \in S]$ e calculando $\mathbb{P}(X \in S)$.
- Derive $F_Y(y)$ para obter a densidade $f(y)$ de Y . Esboce a densidade e com base no gráfico, sem fazer contas, responda: o que é maior, $\mathbb{P}(Y < 1/2)$ ou $\mathbb{P}(Y > 1/2)$?

10. Considere a f.d.a. $\mathbb{F}(x) = \mathbb{P}(X \leq x)$. Quais afirmações abaixo são corretas?

- $\mathbb{F}(x)$ é uma função aleatória.
- Se X é uma v.a. discreta então $\mathbb{F}(x)$ possui saltos em todos os pontos onde X tem massa de probabilidade maior que zero.
- $\mathbb{F}(x)$ mede a probabilidade de X ser menor que média.
- $\mathbb{F}(x)$ é uma função determinística.
- $\mathbb{F}(x)$ só pode ser calculada depois que uma amostra é obtida.
- $\mathbb{F}(x)$ é a mesma função, qualquer que seja a amostra aleatória de X .

11. Exercício para verificar aprendizagem de notação: Seja X_1, X_2, \dots, X_n uma amostra de uma v.a. Considere a f.d.a. empírica

$$\hat{\mathbb{F}}(x) = \frac{1}{n} \text{ no. elementos } \leq x$$

Explique por tque isto é equivalente a escrever

$$\hat{\mathbb{F}}(x) = \frac{\sum_{i=1}^n I_{[X_i \leq x]}}{n} = \frac{\sum_{i=1}^n I_{(-\infty, x]}(X_i)}{n}$$

12. Considere a f.d.a. empírica $\hat{\mathbb{F}}(x) = \sum_i I[X_i \leq x]/n$ baseada numa amostra aleatória de X . Quais afirmações abaixo são corretas?

- $\hat{\mathbb{F}}(x)$ é uma função aleatória.
- Se X é uma v.a. discreta então $\hat{\mathbb{F}}(x)$ possui saltos em todos os pontos onde X tem massa de probabilidade maior que zero.
- $\hat{\mathbb{F}}(x)$ mede a probabilidade de X ser menor que média.
- $\hat{\mathbb{F}}(x)$ é uma função determinística.
- $\hat{\mathbb{F}}(x)$ só pode ser calculada depois que uma amostra é obtida.
- $\hat{\mathbb{F}}(x)$ é a mesma, qualquer que seja a amostra aleatória de X .
- $\mathbb{F}(x)$ é a mesma função, qualquer que seja a amostra aleatória de X .

13. Em finanças, o valor presente (hoje) de um capital c a ser pago daqui a T anos é dado por $V = c \exp(-\delta T)$ onde δ é a taxa de juros anual. Um valor típico é $\delta = 0.04$, o que corresponde a 4% anuais de juros. Imagine que c é o capital a ser pago por uma apólice de seguros a um beneficiário quando um indivíduo falecer. Se T é o tempo de vida futuro (e aleatório) deste indivíduo, $V = c \exp(-\delta T)$ representa o valor atual (presente, no instante da assinatura do contrato da apólice) deste capital futuro e incerto. Para precificar o seguro e estabelecer o prêmio a ser cobrado do segurado, a seguradora precisa calcular o valor esperado $\mathbb{E}(V)$. Supondo que T possui uma distribuição exponencial com parâmetro $\lambda = 1/40$ (ou média igual a 40), obtenha $\mathbb{E}(V)$. OBS: a densidade de uma exponencial com parâmetro λ é dada por

$$f(t) = \begin{cases} 0, & \text{se } t < 0 \\ \lambda e^{-\lambda t}, & \text{se } t \geq 0 \end{cases}$$

14. Discutimos alguns métodos para geração de v.a.'s em sala tais como aceitação/rejeição e transformada inversa. Eles podem ser usados para gerar números aleatórios com quase qualquer distribuição. No entanto, para as distribuições mais importantes, existem técnicas específicas que são melhores do que estas técnicas mais gerais.

A distribuição normal (ou gaussiana) é uma das mais importantes em probabilidade por causa do Teorema Central do Limite. O método mais conhecido e mais usado para gerar gaussianas é o de Box-Mueller: gere duas v.a.'s X_1 e X_2 i.i.d. com distribuição uniforme em $[0, 1]$. Pode-se provar que

$$Y_1 = \sin(2\pi X_1) \sqrt{-2 \ln X_2}$$

e

$$Y_2 = \cos(2\pi X_1) \sqrt{-2 \ln X_2}$$

são gaussianas independentes com $\mu = 0$ e $\sigma = 1$.

Aqui está um código em *R* para gerar `nsim` gaussianas independentes $N(0, 1)$ (com média 0 e variância 1) com este método de Box-Mueller:

```
minharnorm = function(n) sqrt(rexp(n, 0.5)) * cos(runif(n, 0, 2 * pi))
```

Estamos usando que, se $X_2 \sim U(0,1)$, então $-2 \log(X_2) \sim \exp(1/2)$.

Gere 10 mil valores $N(0,1)$ com esta função e crie um histograma padronizado (com área 1, argumento `prob=T`). Compare o histograma dos dados simulados com a densidade gaussiana exata sobrepondo a densidade ao histograma (Crie uma grade e calcule `dnorm`; a seguir use `lines`). São parecidos?

15. Considere a densidade de probabilidade

$$f(x) = \frac{3}{2}\sqrt{x} \quad x \in [0, 1]$$

Escreva uma função em R para gerar v.a.'s com esta distribuição usando (a) o método da transformada inversa, e (b) um método de aceitação-rejeição. Gere 10000 números com cada método e mostre seus resultados num histograma.

16. Podemos usar simulação Monte Carlo para estimar integrais da forma

$$\mathcal{I} = \int_A g(x) dx$$

onde g é uma função qualquer e A é uma região do espaço euclidiano \mathbb{R}^k . Para fazer isto, considere um retângulo k -dimensional D que contenha A e com volume $\text{vol}(D)$. Seja X um vetor uniformemente distribuído no retângulo D (basta gerar uma v.a. uniforme para cada eixo coordenado do retângulo). Seja $Y = g(X)$ se $X \in A$ e $Y = 0$, caso contrário. Pode-se mostrar que

$$\mathbb{E}(Y) = \mathbb{E}(g(X)) = \int_D g(x) \frac{1}{\text{vol}(D)} dx = \int_A g(x) \frac{1}{\text{vol}(D)} dx = \frac{1}{\text{vol}(D)} \mathcal{I}$$

Gere uma grande amostra X_1, X_2, \dots, X_n de v.a.'s i.i.d com a mesma distribuição que X e use a aproximação

$$\frac{\mathcal{I}}{\text{vol}(D)} = \mathbb{E}(Y) \approx \frac{1}{n}(Y_1 + \dots + Y_n)$$

Use esta técnica para estimar a integral dupla

$$\mathcal{I} = \iint_{\Omega} e^{-\sqrt{x^2+y^2}} dx dy$$

na região semicircular Ω definida por

$$x^2 + y^2 \leq 1, \quad x \geq 0.$$

17. Use simulação Monte Carlo para estimar o volume do elipsóide

$$x^2 + \frac{y^2}{4} + \frac{z^2}{16} \leq 1.$$

Você pode assumir que o elipsóide está contido no paralelepípedo $[-1, 1] \times [-2, 2] \times [-4, 4]$.

18. Use o método de transformada inversa para obter uma amostra Monte Carlo de uma distribuição $\exp(\lambda)$ que possui densidade de probabilidade

$$f(x) = \begin{cases} 0, & \text{se } x < 0 \\ \lambda \exp(-\lambda x), & \text{se } x \geq 0 \end{cases}$$

Para isto, use a função de densidade acumulada: $F(x) = 1 - \exp(-\lambda x)$ se $x > 0$.

19. Você já sabe gerar de uma exponencial com $\lambda = 1$ e densidade $g(x)$. Suponha que você queira gerar números aleatórios de uma distribuição Gama com parâmetros $\alpha = 3$ e $\beta = 2$. Isto é, você quer gerar números aleatórios com uma densidade

$$f(x) = \begin{cases} 0, & \text{se } x < 0 \\ 4.5x^2 \exp(-2x), & \text{se } x \geq 0 \end{cases}$$

Mostre que $f(x)/g(x)$ atinge seu ponto de máximo em $x = 2$. Obtenha então o valor da constante c tal que $f(x)/(cg(x)) \leq 1$ para todo x e use esta razão para encontrar c e fazer uma amostragem de $f(x)$ por aceitação-rejeição ($c = 2.44$ deve ser suficiente, nas minhas contas). Qual a porcentagem de valores gerados que foram rejeitados?

20. Uma seguradora possui uma carteira com 50 mil apólices de seguro de vida. Não é possível prever quanto cada pessoa vai viver mas é possível prever o comportamento estatístico dessa massa de segurados. Atuários estudam este fenômeno e já identificaram uma distribuição excelente para o tempo de vida X de adultos: a distribuição de Gompertz que possui densidade de probabilidade $f_X(x)$ dada por:

$$f(x) = Bc^x \exp\left(-\frac{B}{\log(c)}(c^x - 1)\right) = Bc^x S(x)$$

para $x \geq 0$, onde $B > 0$ e $c \geq 1$ são constantes positivas que alteram o formato da função densidade. Evidentemente, $f(x) = 0$ para $x < 0$ pois não existe tempo de vida negativo. A função de distribuição acumulada é igual a

$$\mathbb{F}_X(x) = 1 - e^{-\frac{B(c^x - 1)}{\log(c)}}$$

para $x \geq 0$ e $\mathbb{F}_X(x) = 0$ para $x < 0$. Usando dados recentes de uma seguradora brasileira, podemos tomar $B = 1.02 \times 10^{-4}$ e $c = 1.0855$.

- Com os parâmetros B e c acima, desenhe a curva densidade de probabilidade (use valores x entre 0 e 100 anos).
- SEM FAZER NENHUMA conta, apenas olhando a curva que você gerou, responda:
 - $\mathbb{P}(X < 40)$ é aproximadamente igual a 0.03, 0.10, ou 0.20?
 - Deslize mentalmente um pequeno intervalo de um ano e considere todas as probabilidades do tipo $\mathbb{P}(X \in [k, k + 1])$ onde k é um natural. Qual a idade em que esta probabilidade é aproximadamente máxima: aos $k = 60, 70$ ou 80 anos de idade?
 - O que é maior, a probabilidade de morrer com mais de 100 anos ou de morrer antes de completar 10 anos de idade?
- Inverta a função de distribuição acumulada, mostrando que

$$F^{-1}(u) = \log(1 - \log(c) \log(1 - u)/B) / \log(c)$$

onde $u \in (0, 1)$.

- Use o método da transformada inversa para gerar 50 mil valores independentes de X .
- Com estes números simulados, calcule aproximadamente $\mathbb{P}(X > 80 | X > 50)$. Isto é, calcule aproximadamente a chance de sobreviver pelo menos mais 30 anos dado que chegou a completar 50 anos de idade.
- A seguradora cobra um prêmio de 2 mil reais por uma apólice de seguro de vida que promete pagar 100 mil reais a um beneficiário no momento exato de morte do segurado. A apólice é vendida no momento em que os 50 mil indivíduos nasceram (alterar esta hipótese para que apenas adultos comprem a apólice dá trabalho e não muda o essencial do exercício). Ela coloca o dinheiro rendendo juros de 5% ao ano de forma que dentro de t anos os 2 mil reais terão se transformado em $2 \times \exp(0.05t)$. Se o indivíduo falecer muito cedo, ela terá uma perda financeira. Se ele sobreviver muito tempo, seu prêmio vai acumular juros suficiente para cobrir o pagamento do benefício.

Para a carteira de 50 mil vidas que você gerou, calcule aproximadamente a probabilidade de perda da seguradora usando simulação Monte Carlo.

21. Use Importance Sampling para obter uma amostra de X com a densidade $f(x)$ igual ao do problema 11. Como a densidade $f(x)$ é a de uma gama com parâmetros $\alpha = 3$ e $\beta = 2$ então a esperança $E(X)$ é conhecida analiticamente e é igual a $E(X) = 1.5$. Verifique se a média ponderada da sua amostra tem um valor próximo deste valor.

Use sua amostra para obter valores aproximados das seguintes quantidades que, neste caso simples, podem ser obtidas analiticamente (ou com métodos numéricos bem precisos):

- $DP(X) = \sqrt{\text{Var}(X)} = 0.75$ onde $\text{Var}(X) = E(X - E(X))^2$
- $P(X > 4.3) = E(I[X > 4.3]) = 0.0086$

- $E(10e^{-X}I[2 < X < 3])$ (este número é o valor presente atuarial de um seguro de 45 equipamento que paga 10 unidades se uma certa máquina falhar entre 2 e 3 anos de seu início).

22. No problema anterior, imagine que você não sabe que a constante de normalização da densidade $f(x)$ seja igual a 4.5. Use SIR (Sampling Importance Resampling) para obter uma amostra de $f(x)$ e a seguir estime as mesmas quantidades do problema anterior.

23. No método de aceitação-rejeição queremos amostra de uma densidade-alvo $f(x)$ mas usamos uma amostra retirada de uma densidade $g(x)$ de onde sabemos gerar. Precisamos escolher uma constante M tal que $f(x) \leq Mg(x)$ para todo x . Mostre que $M \geq 1$. DICA: integre dos dois lados da desigualdade.

24. Graças ao Bráulio Veloso (obrigado!), este exercício pede que você use o modelo Bag of Words para classificar alguns textos. O vocabulário foi construído com 12 livros de Ficção, Humor e Religião, com três livros em cada categoria. As stop-words foram eliminadas (as palavras muito comuns mas que não discriminam os textos tais como preposições e artigos).

Existem 4 arquivos. Um deles, `FreqTreinoNoStopWords.csv`, possui a frequência das bases de treino (foram usados 4 livros por categoria). Ele possui 4 campos. O primeiro campo é um string com a palavra. O dicionário está ordenado. As strings que são constituídas somente de números ficaram nas primeiras linhas do arquivo. As demais palavras estão mais abaixo no csv. Os outros 3 campos são as frequências das palavras em cada tipo de livro na seguinte ordem: livros de Ficção, Humor e Religião.

Os outros 3 arquivos são para teste, com um livro por categoria. A resposta correta das categorias das bases de treino é:

- livroTeste0: humor
- livroTeste1: religião
- livroTeste2: Ficção

Ignorando a resposta correta, você deve tentar classificar cada arquivo teste em uma das categorias possíveis usando as ideias discutidas em sala e presentes nos slides.

Capítulo 5

Vetores Aleatórios

1. Considere o conjunto de dados `iris` do R digitando os seguintes comandos:

```
iris
dim(iris)
names(iris)
plot(iris[,3], iris[,4])
plot(iris$Petal.Length, iris$Petal.Width, pch=21,
     bg=c("red","green3","blue")[unclass(iris$Species)])
setosa <- iris[iris$Species == "setosa", 1:4]
plot(setosa)
mean(setosa)
cov(setosa)
cor(setosa)
```

Os dados são uma amostra de exemplos do vetor aleatório $\mathbf{X} = (X_1, X_2, X_3, X_4)$ onde X_1 é o Sepal Length, X_2 é o Sepal Width, X_3 é o Petal Length e X_4 é o Petal Width. Assuma que a distribuição conjunta do vetor \mathbf{X} é uma normal multivariada de dimensão 4 com parâmetros $\boldsymbol{\mu} = (\mu_1, \mu_2, \mu_3, \mu_4)$ e matriz de covariância Σ de dimensão 4×4 . Use os resultados obtidos no R (e apenas DUAS casas decimais) para responder às seguintes questões:

- Forneça uma estimativa para o vetor $\boldsymbol{\mu}$ e para a matriz Σ .
- A partir da matriz de correlações entre os pares de v.a.'s (e do plot de dispersão dos pontos), quais os grupos que são mais correlacionados?
- Obtenha a distribuição do sub-vetor $\mathbf{X}^* = (X_1, X_3)$.
- Obtenha a distribuição CONDICIONAL do sub-vetor $\mathbf{X}^* = (X_1, X_3)$ quando são conhecidos os valores de (X_2, X_4) .
- Obtenha agora distribuição CONDICIONAL do sub-vetor $\mathbf{X}^* = (X_1, X_3)$ quando é conhecido apenas o valor de X_2 .

- Obtenha também distribuição CONDICIONAL do sub-vetor $\mathbf{X} = (X_1, X_3)$ quando é conhecido apenas o valor de X_4 .
 - Comparando as três últimas respostas que você forneceu, qual das duas variáveis isoladamente, X_2 ou X_4 , diminui mais a incerteza acerca de X_3 ? Isto é, se você tivesse de escolher apenas uma delas, X_2 ou X_4 , qual você iria preferir se seu objetivo fosse prever o valor de X_3 ? A resposta é a mesma se o objetivo for prever X_1 ?
 - Considere a melhor preditora para X_3 que você escolheu, dentre X_2 ou X_4 , na questão anterior. Digamos que tenha sido X_4 . Avalie quanto conhecer a outra variável (neste caso, X_2) reduz ADICIONALMENTE a incerteza acerca de X_3 . Isto é, compare $Var(X_3|X_4)$ com $Var(X_3|X_2, X_4)$.
2. Seja $\mathbf{Z} = (Z_1, Z_2, Z_3)$ um vetor de variáveis i.i.d. (independentes e identicamente distribuídas) $N(0, 1)$. Isto é, \mathbf{Z} segue uma distribuição normal multivariada com valor esperado $(0, 0, 0)$ e matriz 3×3 de covariância igual à identidade \mathbf{I} . Você aprendeu a gerar estas v.a.'s na lista anterior.

Queremos agora gerar um vetor aleatório $\mathbf{X} = (X_1, X_2, X_3)$ seguindo uma normal multivariada com valor esperado $\mu = (\mu_1, \mu_2, \mu_3) = (10, 20, -50)$ e com matriz de covariância

$$\Sigma = \begin{bmatrix} 4 & 9 & -14 \\ 9 & 30 & -44 \\ -14 & -44 & 94 \end{bmatrix}$$

Para isto, siga os seguintes passos em R (em matlab, use comandos similares):

- Encontre uma matriz L tal que $\mathbf{L}\mathbf{L}^t = \Sigma$. Uma matriz com esta propriedade é aquela obtida pela decomposição de Cholesky de matrizes simétricas e definidas positivas. Em R, isto é obtido pelo comando `L = t(chol(Sigma))`.
- Gere \mathbf{z} , um vetor 3-dim com v.a.'s iid $N(0, 1)$.
- A seguir, faça

```
x = mu + L %*% z
```

Gere uma amostra de tamanho 2000 dos vetores \mathbf{x} 3-dim e armazene numa matriz `amostra` de dimensão 2000×3 . A seguir, calcule a média aritmética dos 2000 valores de cada coordenada de \mathbf{x} e compare com os três valores do vetor μ . Eles devem ser parecidos.

Usando a amostra, estime os 9 valores da matriz de covariância Σ . Chame esta matriz estimada de S . Verifique que as estimativas são próximas dos valores verdadeiros que você usou para gerar seus dados. Por exemplo, estime o elemento σ_{12} da matriz Σ por

$$s_{12} = \frac{1}{2000} \sum_{i=1}^{2000} (x_{i1} - \bar{x}_1)(x_{i2} - \bar{x}_2)$$

onde \bar{x}_1 e \bar{x}_2 são as médias aritméticas dos 2000 valores observados das v.a.'s 1 e 2. Os termos σ_{jj} da diagonal principal são estimados por

$$s_{jj} = \frac{1}{2000} \sum_{i=1}^{2000} (x_{ij} - \bar{x}_j)^2$$

O comando `cov(x)` calcula a matriz \mathbf{S} diretamente (usando 1999 no denominador, ao invés de 2000). Procure calcular você os termos da matriz \mathbf{S} para ter certeza de que você está entendendo o que estamos fazendo.

3. A matriz Σ é estimada a partir dos dados substituindo o operador teórico e probabilístico \mathbb{E} pela média aritmética dos números específicos da amostra. Assim, σ_{ij} é estimado por sua versão empírica s_{ij} . Qual a diferença entre σ_{ij} e s_{ij} ? Uma maneira de responder a isto é notar que s_{ij} vai ter um valor ligeiramente diferente cada vez que uma nova amostra for gerada, mesmo que a distribuição de probabilidade permaneça a mesma. Já σ_{ij} não vai mudar nunca, é fixo e determinado pela distribuição de probabilidade.

Seja ρ a matriz 3×3 de correlação com elemento

$$\rho_{ij} = \text{Corr}(X_i, X_j) = \frac{\text{Cov}(X_i, X_j)}{\sqrt{\sigma_{ii}\sigma_{jj}}}.$$

Observe que $\rho_{ii} = 1$. Esta matriz ρ é estimada pela matriz \mathbf{R} , cujos elementos são obtidos a partir dos dados da amostra. Assim, ρ_{ij} é estimado por

$$r_{ij} = s_{ij} / \sqrt{s_{ii}s_{jj}}$$

Calcule as matrizes ρ e \mathbf{R} e compare-as.

Este é um dos sentidos que empregamos à expressão *aprendizagem*: usamos os dados observados para aprender (ou inferir) sobre o mecanismo aleatório que gerou estes mesmos dados. Isto é, aprendemos sobre μ e Σ através de (\bar{x}_1, \bar{x}_2) e de \mathbf{S} .

4. Usando a distribuição de \mathbf{X} do problema anterior, seja \mathbf{b} um vetor k -dimensional e \mathbf{C} uma matriz $k \times 3$ formada por constantes. Uma das propriedades da normal multivariada é que a distribuição do vetor $\mathbf{b} + \mathbf{C}\mathbf{X}$ de dimensão k é normal com vetor de médias $\mathbf{b} + \mathbf{C}\mu$ e matriz de $k \times k$ covariância $\mathbf{C}\Sigma\mathbf{C}^t$. Use esta propriedade para obter a distribuição das seguintes variáveis:
 - Distribuição marginal de X_1 , de X_2 e de X_3 .
 - Distribuição de um indicador composto pelas 3 variáveis: $T = 0.4X_1 + 0.3X_2 + 0.3X_3$.
 - Distribuição de um indicador composto pelas 3 variáveis normalizadas: $T = 0.4(X_1 - 10)/2 + 0.3(X_2 - 20)/\sqrt{30} + 0.3(X_3 + 50)/\sqrt{94}$.
 - Distribuição conjunta de $(X_1 - X_2, 4X_1 + 2X_2 - X_3)$.
 - Distribuição conjunta de $(X_1, aX_1 + bX_2 + cX_3)$. onde a, b, c são constantes reais. Em particular, encontre a covariância entre X_1 e o indicador $Y = aX_1 + bX_2 + cX_3$ formado pela combinação linear de X_1, X_2 e X_3 .
5. Considere um vetor $\mathbf{X} = (X_1, \dots, X_p)$ com distribuição normal multivariada. É possível mostrar que, com probabilidade $1 - \alpha$, o vetor aleatório \mathbf{X} deve cair dentro da elipse $D^2 = c$ onde $c = \chi_p^2(\alpha)$ é o quantil $(1 - \alpha)100\%$ de uma distribuição qui-quadrado com p graus de liberdade onde p é a dimensão do vetor \mathbf{X} . No caso particular de um vetor bidimensional, o valor de c associado com a probabilidade $1 - \alpha = 0.95$ é igual a $c = 9.21$ ou $c \approx 9.2$.

- 50 Assim, se $\mathbf{X} = (X_1, X_2)$ estiver fora dessa elipse (isto é, se $D^2 > 9.2$), o ponto pode ser considerado um tanto anômalo ou extremo.

O arquivo stiffness.txt contém dois tipos de medições da rigidez de pranchas de madeira, a primeira aplicando uma onda de choque através da prancha, ea segunda aplicando uma vibração à prancha. Estime o vetor $\mu = (\mu_1, \mu_2)$ e a matriz Σ usando os dados do amostra e a seguir calcula o valor de D^2 para cada ponto da amostra. Qual deles parece extremo? Olhando as duas variáveis INDIVIDUALMENTE seria possível detectar estes pontos extremos?

6. Considere um vetor $\mathbf{X} = (X_1, X_2)$ com distribuição normal bivariada com vetor esperado $\mu = (\mu_1, \mu_2)$ e matriz de covariância

$$\Sigma = \begin{bmatrix} \sigma_{11} & \rho\sqrt{\sigma_{11}\sigma_{22}} \\ \rho\sqrt{\sigma_{11}\sigma_{22}} & \sigma_{22} \end{bmatrix}$$

Usando o resultado dos slides, mostre que a distribuição condicional de $(X_2|X_1 = x_1)$ é $N(\mu_c, \sigma_c^2)$ onde

$$\mu_c = \mu_2 + \rho\sqrt{\frac{\sigma_{22}}{\sigma_{11}}}(x_1 - \mu_1) = \mu_2 + \rho\sqrt{\sigma_{22}}\frac{x_1 - \mu_1}{\sqrt{\sigma_{11}}}$$

e

$$\sigma_c^2 = \sigma_{22}(1 - \rho^2)$$

A partir desses resultados, verifique se as afirmações abaixo são V ou F:

- Saber que o valor $X_1 = x_1$ está dois desvios-padrão acima de seu valor esperado (isto é, $(x_1 - \mu_1)/\sqrt{\sigma_{11}} = 2$) implica que devemos esperar que X_2 também fique dois desvios-padrão acima de seu valor esperado.
 - Dado que $X_1 = x_1$, a variabilidade de X_2 em torno de seu valor esperado é maior se $x_1 < \mu_1$ do que se $x_1 > \mu_1$.
 - Conhecer o valor de X_1 (e assim eliminar parte da incerteza existente) sempre diminui a incerteza da parte aleatória permanece desconhecida (isto é, compare a variabilidade de X_2 condicionada e não-condicionada no valor de X_1).
 - μ_c é uma função linear de x_1 .
7. *Regressão linear e distribuição condicional:* Vamos considerar um modelo (na verdade, mais uma caricatura) de como a renda do trabalho Y de um indivíduo qualquer está associada com o número de anos de estudo X desse mesmo indivíduo. Vamos supor que, para um indivíduo com $X = x$ anos de estudo teremos a renda Y como uma variável aleatória com distribuição normal com esperança $\mathbb{E}(Y|X = x) = g(x) = 300 + 100 * x$ e variância $\sigma^2 = 50^2$. Responda V ou F às afirmações abaixo:
- Se $X = 10$ para um indivíduo (isto é, se ele possui 10 anos de estudo), então a sua renda é uma variável aleatória com distribuição $N(1300, 50^2)$.
 - $\mathbb{E}(Y) = 300 + 100 * x$.
 - $\mathbb{E}(Y|X = x) = 300 + 100 * x$.

- $\mathbb{V}(Y) = 50^2$.
 - $\mathbb{V}(Y|X = x) = 50^2$.
8. Duas variáveis aleatórias contínuas com densidade conjunta $f_{XY}(x, y)$ são *independentes* se, e somente se, a densidade conjunta é o produto das densidades marginais:

$$f_{XY}(x, y) = f_X(x)f_Y(y).$$

Mostre que X e Y são independentes se o vetor (X, Y) seguir uma distribuição uniforme num retângulo $[a, b] \times [c, d]$ e densidade

$$f_{XY}(x, y) = \begin{cases} 1/A, & \text{se } (x, y) \in [a, b] \times [c, d] \\ 0, & \text{caso contrário} \end{cases}$$

onde $A = (b - a)(d - c)$ é a área do retângulo. Para isto, obtenha as marginais $f_X(x)$ e $f_Y(y)$ e mostre que seu produto é igual á densidade conjunta. Verifique também que X e Y seguem distribuições uniformes. Assim, no método de aceitação-rejeição, podemos gerar facilmente dessa densidade uniforme: simplesmente gere X e Y independentemente com distribuições uniformes.

9. Gerar uma amostra aleatória com 300 instâncias do vetor aleatório (X, Y) com densidade

$$f(x, y) = \begin{cases} 0.1 (2 + \sin(2\pi x) + \sin(2\pi y)), & \text{se } (x, y) \in D \\ 0, & \text{caso contrário} \end{cases}$$

O suporte D é um polígono dentro do quadrado $[0, 3] \times [0, 3]$ que pode ser visualizado com estes comandos:

```
poligx = c(1,1,0,1,2,3,2,2,1)
poligy = c(0,1,2,3,3,2,1,0,0)
plot(poligx, poligy, type="l")
```

Gere uma amostra com distribuição uniforme no quadrado, que tem densidade $g(x, y) = 1/9$ no quadrado $[0, 3]^2$, e retenha cada ponto $((x, y)$ gerado com probabilidade $p(x, y) = f(x, y)/(Mg(x, y))$. Verifique que podemos tomar $M \geq 3.6$. Vou usar $M = 3.6$.

Para visualizar a densidade conjunta $f(x, y)$ na região maior do quadrado $[0, 3] \times [0, 3]$, dentro do qual está a região D , digite:

```
f <- function(x,y){ 0.1*(2+sin(2*pi*x)+sin(2*pi*y)) }
eixox = eixoy = seq(0,3,length=101)
z = outer(eixox, eixoy, f)
```

Observe que o suporte de g é maior que o de f . Teoricamente, isto não é problema: pontos gerados dentro quadrado mas fora do polígono D devem ser retidos com probabilidade $p(x, y) = f(x, y)/(Mg(x, y)) = 0/(M/9) = 0$. Isto é, eles devem ser rejeitados com probabilidade 1. Assim, antes mesmo de calcular $p(x, y)$, verifique se cada ponto está

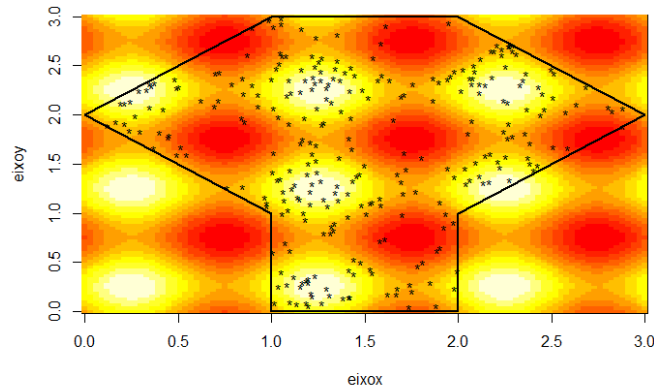


Figura 5.1: Amostra de $f(x, y)$ junto com polígono D .

dentro de D : se não estiver, elimine-o. Se D tiver uma forma muito complicada (como a forma da Lagoa da Pampulha ou o contorno de Minas Gerais), você vai precisar de algoritmos geométricos sofisticados que fazem isto de forma eficiente. No nosso caso, a forma do polígono é muito simples. Complete o código abaixo para obter uma função que testa se $(x, y) \in [0, 3]^2$ está dentro de D (ou proponha um código melhor que o meu):

```
dentroD = function(x,y){
  dentro = F
  if(x >= 1 & x <= 2) dentro = T
  else{
    if((x > 2) & (y >= -1+x) & (y <= 5-x)) dentro = T
    else
      if( ??????
  }
  return(dentro)
}
```

A seguir, faça a amostragem de $f(x, y)$. Você poderá visualizar sua amostra, armazenada na matriz `mat`, junto com a densidade no polígono usando os seguintes comandos, que geram a Figura 5.1:

```
image(eixox, eixoy, z)
lines(poligx, poligy, lwd=2)
points(mat, pch="*")
```

Note como as áreas mais claras são aquelas que estão com maior densidade de pontos aleatórios enquanto as áreas mais vermelhas possuem densidade mais baixa.

10. A função `persp` nativa no R desenha gráficos de perspectiva de uma superfície sobre o plano x-y. Digite `demo(persp)` no console para ter uma ideia do que esta função pode fazer. A seguir, faça você mesmo os gráficos de quatro diferentes funções de densidade de probabilidade $f(x, y)$ de um vetor aleatório bivariado (X, Y) .

- $f(x, y) = (2\pi)^{-1} \exp(-(x^2 + y^2)/2)$, a densidade de uma gaussiana bivariada com variáveis independentes e marginais $X \sim N(0, 1)$ e $Y \sim N(0, 1)$. A constante de integração é igual a $1/(2\pi)$. Faça a superfície considerando a região $[-4, 4] \times [-4, 4]$ do plano (x, y) .
- $f(x, y) = (2\pi\sqrt{0.51})^{-1} \exp\left(-\frac{x^2 + y^2 - 1.4xy}{1.02}\right)$ em $[-4, 4] \times [-4, 4]$. Esta é a densidade de uma gaussiana bivariada de variáveis não-independentes, com correlação $\rho = 0.7$, e com marginais $X \sim N(0, 1)$ e $Y \sim N(0, 1)$.
- $f(x, y) = |\sin(r)|/(44r)$ onde $r = \sqrt{x^2 + y^2}$. Faça a superfície considerando a região $[-10, 10] \times [-10, 10]$ do plano (x, y) dividindo-a em uma grade 30×30 .
- $f(x, y) = 0.5 \exp(-(x/3 + y + \sqrt{xy}/4))$ no retângulo $[0, 6] \times [0, 3]$.
- $f(x, y) = 0.5 * g(x, y) + 0.5 * h(x, y)$ em $[-4, 4] \times [-4, 4]$ onde $g(x, y) = (2\pi)^{-1} \exp(-(x^2 + y^2)/2)$ (como no item 1) e $h(x, y) = 2\pi^{-1} \exp(-4(x-2)^2 - 4(y-2)^2)$

O script a seguir exemplifica como fazer o gráfico de uma densidade bivariada usando a primeira densidade da lista acima.

```
x = seq(-4, 4, by = 0.2)      # usando o parametro by
y = seq(-4, 4, length = 41)   # usando o parametro length

# use outer:
z <- outer(x, y, FUN = function(x,y) exp(-(x^2 + y^2)/2) /(2*pi) )

# outer retorna uma matriz em que na posicao (i,j) temos o valor
# de FUN avaliado com x=x[i] e y=y[j]

# Faca o grafico 3-dim da superficie
persp(x, y, z)

# mudando alguns parametros de persp
persp(x, y, z, theta = 30, phi = 30, expand = 0.5, col = "lightblue")

# mais algumas mudancas
persp(x, y, z, theta = 30, phi = 30, expand = 0.5, col = "lightblue",
      ltheta = 120, shade = 0.75, ticktype = "detailed",
      xlab = "x", ylab = "y", zlab = "densidade f(x,y)")

#####
```



```
# Quando a funcao a ser passada como argumento de FUN for muito longa ou complexa
# basta defini-la separadamente e passar apenas o seu nome como argumento para FUN
# Exemplo:
```

```
f <- function(x,y){
  exp(-(x^2 + y^2)/2) /(2*pi)
}
```

```
z <- outer(x, y, FUN=f)
persp(x, y, z)
```

```
#####
```

```
# Um procedimento alternativo usando a funcao mesh do R:
# crie um reticulado no plano (um grid)
M <- mesh(x, y))
```

```
# com a funcao "with", calcule uma funcao em cada ponto do grid retangular
# o valor de retorno eh uma matriz identica 'a z anterior.
```

```
z <- with (M, exp(-(x^2 + y^2)/2) /(2*pi))
```

```
persp(x, y, z)
```

11. Instale o pacote `plot3D`, criado por Karline Soetaert e baseado na função `persp()`. A vinheta (*vignette*, em inglês) do pacote `plot3d` mostra alguns gráficos muito bonitos. Carregue o pacote e digite os seguintes comandos no console: `exemplo(persp3D)`, `exemplo(Surf3D)` e `exemplo(scatter3D)` para ver exemplos. Além disso, tente este código abaixo para ver o belo histograma tri-dimensional da Figura 5.2. (script adaptado de <http://blog.revolutionanalytics.com/2014/02/3d-plots-in-r.html>).

O dataframe `quakes` fornece informações sobre 1000 terremotos com magnitude maior que 4.0 na escala Richter em torno da ilha Fiji na Oceania a partir de 1964. A longitude e latitude do epicentro desses 1000 eventos são as duas primeiras colunas do dataframe. Podemos ver a posição do epicentro como um vetor aleatório (X, Y) com certa densidade de probabilidade $f(x, y)$.

```
# veja alguma informacao sobre o dataframe
help("quakes")
```

```
dim(quakes) # verifique a dimensao do dataset
```

```
plot(quakes$long, quakes$lat) # scatterplot dos terremotos em lat-long
grid(20,20) # adiciona uma grade 30 x 30 ao plot
```

```
# o histograma tri-dim conta o numero de terremotos em cada celula
```

```

# do grid e levanta uma pilastra cuja altura e' proporcional a esta contagem55
# Este histograma e' uma versao grosseira da densidade de probabilidade f(x,y)
# que gera os pontos aleatorios. Voce ja' consegue imaginar a densidade f(x,y)
# a partir do plot bi-dimencional imaginando alturas maiores nas areas com maior
# densidade de pontos

# agora, o histograma 3-dim
# Carregue o pacote plot3D
require(plot3D)

# particione os eixos x e y
lon <- seq(165.5, 188.5, length.out = 30)
lat <- seq(-38.5, -10, length.out = 30)

# conte o numero de terremotos em cada celula do grid
xy <- table(cut(quakes$long, lon), cut(quakes$lat, lat))

# veja o uso da funcao cut acima.
# aproveite para ler sobre ela pois e' muito util.

?cut

# obtenha o ponto medio em cada celula do grid nos eixos x e y
xmid <- 0.5*(lon[-1] + lon[-length(lon)])
ymid <- 0.5*(lat[-1] + lat[-length(lat)])

# passando argumentos para os parametros de controle da margem da janela grafica
par (mar = par("mar") + c(0, 0, 0, 2))

# O histograma 3D, na versao default. Cores ajudam a visualizar as alturas das barras
hist3D(x = xmid, y = ymid, z = xy)

# Mudando os parametros que controlam o angulo de visao e a posicao do observador
hist3D(x = xmid, y = ymid, z = xy, phi = 5, theta = 25)

# Tirando as cores, pondo rotulos nos eixos e titulo no grafico
hist3D(x = xmid, y = ymid, z = xy, phi = 5, theta = 25, col = "white", border = "black",
      main = "Earth quakes", ylab = "latitude", xlab = "longitude", zlab = "counts")

# Agora uma visao bem mais trabalhada: aumentando o eixo z para criar espaco para os pontos
hist3D(x = xmid, y = ymid, z = xy,
      zlim = c(-20, 40), main = "Earth quakes",
      ylab = "latitude", xlab = "longitude",
      zlab = "counts", bty= "g", phi = 5, theta = 25,
      shade = 0.2, col = "white", border = "black",

```

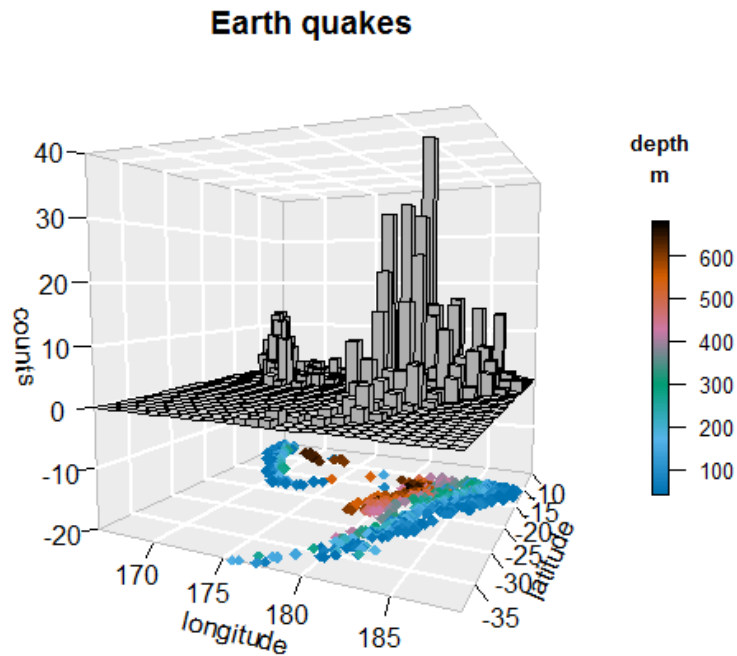


Figura 5.2: Epicentro de terremotos em Fiji com o histograma 3d.

```
d = 1, ticktype = "detailed")
```

```
# Acrescentando os pontos no grafico acima, agora uma visao realmente linda:
with (quakes, scatter3D(x = long, y = lat,
  z = rep(-20, length.out = length(long)),
  colvar = quakes$depth, col = gg.col(100),
  add = TRUE, pch = 18, clab = c("depth", "m"),
  colkey = list(length = 0.5, width = 0.5,
    dist = 0.05, cex.axis = 0.8, cex.clab = 0.8) ))
```

12. Seja $f(x, y) = k(x + y + xy)$ uma densidade de probabilidade do vetor contínuo (X, Y) com suporte na região $[0, 1] \times [0, 1]$. O valor de k é uma constante de normalização que faz a integral ser igual a 1.

- Obtenha k .
- Faça o gráfico 3-dim da densidade conjunta.
- Obtenha a densidade marginal $f_X(x)$ e faça seu gráfico. Avalie esta densidade marginal no ponto $x = 0.2$ e em $x = 0.5$.

- Obtenha a densidade condicional $f_{X|Y}(x|y)$ para x e y genéricos. Se $y = 0.2$, qual é a densidade condicional $f_{X|Y}(x|y = 0.2)$? Repita com $y = 0.9$.
13. Seja $\mathbf{Z} = (Z_1, Z_2, Z_3)$ um vetor de variáveis i.i.d. (independentes e identicamente distribuídas) $N(0, 1)$. Isto é, \mathbf{Z} segue uma distribuição normal multivariada com valor esperado $\boldsymbol{\mu} = (0, 0, 0)$ e matriz 3×3 de covariância igual à identidade \mathbf{I} . Você aprendeu a gerar estas v.a.'s na lista anterior.

Queremos agora gerar um vetor aleatório $\mathbf{X} = (X_1, X_2, X_3)$ seguindo uma normal multivariada com valor esperado $\boldsymbol{\mu} = (\mu_1, \mu_2, \mu_3) = (10, 20, -50)$ e com matriz de covariância

$$\boldsymbol{\Sigma} = \begin{bmatrix} 4 & 9 & -14 \\ 9 & 30 & -44 \\ -14 & -44 & 94 \end{bmatrix}$$

Para isto, siga os seguintes passos em R (em matlab, use comandos similares):

- Encontre uma matriz L tal que $\mathbf{L}\mathbf{L}^t = \boldsymbol{\Sigma}$. Uma matriz com esta propriedade é aquela obtida pela decomposição de Cholesky de matrizes simétricas e definidas positivas. Em R, isto é obtido pelo comando `L = t(chol(Sigma))`.
- Gere \mathbf{z} , um vetor 3-dim com v.a.'s iid $N(0, 1)$.
- A seguir, faça

```
x = mu + L %*% z
```

Gere uma amostra de tamanho 2000 dos vetores \mathbf{x} 3-dim e armazene numa matriz `amostra` de dimensão 2000×3 . A seguir, calcule a média aritmética dos 2000 valores de cada coordenada de \mathbf{x} e compare com os três valores do vetor $\boldsymbol{\mu}$. Eles devem ser parecidos.

Usando a amostra, estime os 9 valores da matriz de covariância $\boldsymbol{\Sigma}$. Chame esta matriz estimada de S . Verifique que as estimativas são próximas dos valores verdadeiros que você usou para gerar seus dados. Por exemplo, estime o elemento σ_{12} da matriz $\boldsymbol{\Sigma}$ por

$$s_{12} = \frac{1}{2000} \sum_{i=1}^{2000} (x_{i1} - \bar{x}_1)(x_{i2} - \bar{x}_2)$$

onde \bar{x}_1 e \bar{x}_2 são as médias aritméticas dos 2000 valores observados das v.a.'s 1 e 2. Os termos σ_{jj} da diagonal principal são estimados por

$$s_{jj} = \frac{1}{2000} \sum_{i=1}^{2000} (x_{ij} - \bar{x}_j)^2$$

O comando `cov(x)` calcula a matriz \mathbf{S} diretamente (usando 1999 no denominador, ao invés de 2000). Procure calcular você os termos da matriz S para ter certeza de que você está entendendo o que estamos fazendo.

A matriz $\boldsymbol{\Sigma}$ é estimada a partir dos dados substituindo o operador teórico e probabilístico \mathbb{E} pela média aritmética dos números específicos da amostra. Assim, σ_{ij} é estimado por

sua versão empírica s_{ij} . Qual a diferença entre σ_{ij} e s_{ij} ? Uma maneira de responder a isto é notar que s_{ij} vai ter um valor ligeiramente diferente cada vez que uma nova amostra for gerada, mesmo que a distribuição de probabilidade permaneça a mesma. Já σ_{ij} não vai mudar nunca, é fixo e determinado pela distribuição de probabilidade.

Seja ρ a matriz 3×3 de correlação com elemento $\rho_{ij} = \text{Corr}(X_i, X_j) = \text{Cov}(X_i, X_j) / \sqrt{\sigma_{ii}\sigma_{jj}}$. Observe que $\rho_{ii} = 1$. Esta matriz ρ é estimada pela matriz \mathbf{R} , cujos elementos são obtidos a partir dos dados da amostra. Assim, ρ_{ij} é estimado por

$$r_{ij} = s_{ij} / \sqrt{s_{ii}s_{jj}}$$

Calcule as matrizes ρ e \mathbf{R} e compare-as.

Este é um dos sentidos em que empregamos à expressão *aprendizagem*: usamos os dados observados para aprender (ou inferir) sobre o mecanismo aleatório que gerou estes mesmos dados. Isto é, aprendemos sobre μ e Σ através de (\bar{x}_1, \bar{x}_2) e de \mathbf{S} .

Capítulo 6

Modelos multivariados gaussianos

Neste capítulo, os exercícios cobrem os assuntos de componentes principais, análise fatorial e análise discriminante (ainda falta este último tópico).

1. Considere os dados da espécie setosa do conjunto de dados `iris` do R digitando os seguintes comandos:

```
iris
dim(iris)
names(iris)
plot(iris[,3], iris[,4])
plot(iris$Petal.Length, iris$Petal.Width, pch=21,
     bg=c("red", "green3", "blue")[unclass(iris$Species)])
setosa <- iris[iris$Species == "setosa", 1:4]
plot(setosa)
mean(setosa)
cov(setosa)
cor(setosa)
```

Os dados são uma amostra de exemplos do vetor aleatório $\mathbf{X} = (X_1, X_2, X_3, X_4)$ onde X_1 é o Sepal Length, X_2 é o Sepal Width, X_3 é o Petal Length e X_4 é o Petal Width. Assuma que a distribuição conjunta do vetor \mathbf{X} é uma normal multivariada de dimensão 4 com parâmetros $\boldsymbol{\mu} = (\mu_1, \mu_2, \mu_3, \mu_4)$ e matriz de covariância Σ de dimensão 4×4 . Use os resultados obtidos no R como estimativas para $\boldsymbol{\mu}$ e Σ , responda às seguintes questões:

- Forneça uma estimativa para o vetor $\boldsymbol{\mu}$ e para a matriz Σ .
- A partir da matriz de correlações entre os pares de v.a.'s (e do plot de dispersão dos pontos), quais os grupos que são mais correlacionados?
- Obtenha a distribuição do sub-vetor $\mathbf{X}^* = (X_1, X_2)$.
- Obtenha a distribuição CONDICIONAL do sub-vetor $\mathbf{X}^* = (X_1, X_2)$ quando são conhecidos os valores de (X_3, X_4) .

- Obtenha agora distribuição **CONDICIONAL** do sub-vetor $\mathbf{X}^* = (X_1, X_2)$ quando é conhecido apenas o valor de X_3 .
- Obtenha também distribuição **CONDICIONAL** do sub-vetor $\mathbf{X}^* = (X_1, X_2)$ quando é conhecido apenas o valor de X_4 .
- Comparando as três últimas respostas que você forneceu, qual das duas variáveis isoladamente, X_3 ou X_4 , diminui mais a incerteza acerca de X_2 ? Isto é, se você tivesse de escolher apenas uma delas, X_3 ou X_4 , qual você iria preferir se seu objetivo fosse prever o valor de X_2 ? A resposta é a mesma se o objetivo for prever X_1 ?
- Considere a melhor preditora para X_2 que você escolheu, dentre X_3 ou X_4 , na questão anterior. Digamos que tenha sido X_4 . Avalie quanto conhecer a outra variável (neste caso, X_3) reduz **ADICIONALMENTE** a incerteza acerca de X_3 . Isto é, compare $Var(X_2|X_4)$ com $Var(X_2|X_3, X_4)$.

2. Considere um vetor $\mathbf{X} = (X_1, X_2)$ com distribuição normal bivariada com vetor esperado $\mu = (\mu_1, \mu_2)$ e matriz de covariância

$$\Sigma = \begin{bmatrix} \sigma_{11} & \rho\sqrt{\sigma_{11}\sigma_{22}} \\ \rho\sqrt{\sigma_{11}\sigma_{22}} & \sigma_{22} \end{bmatrix}$$

Usando o resultado dos slides, mostre que a distribuição condicional de $(X_2|X_1 = x_1)$ é $N(\mu_c, \sigma_c^2)$ onde

$$\mu_c = \mu_2 + \rho\sqrt{\frac{\sigma_{22}}{\sigma_{11}}}(x_1 - \mu_1) = \mu_2 + \rho\sqrt{\sigma_{22}}\frac{x_1 - \mu_1}{\sqrt{\sigma_{11}}}$$

e

$$\sigma_c^2 = \sigma_{22}(1 - \rho^2)$$

A partir desses resultados, verifique se as afirmações abaixo são V ou F:

- Saber que o valor $X_1 = x_1$ está dois desvios-padrão acima de seu valor esperado (isto é, $(x_1 - \mu_1)/\sqrt{\sigma_{11}} = 2$) implica que devemos esperar que X_2 também fique dois desvios-padrão acima de seu valor esperado.
- Dado que $X_1 = x_1$, a variabilidade de X_2 em torno de seu valor esperado é maior se $x_1 < \mu_1$ do que se $x_1 > \mu_1$.
- Conhecer o valor de X_1 (e assim eliminar parte da incerteza existente) sempre diminui a incerteza da parte aleatória permanece desconhecida (isto é, compare a variabilidade de X_2 condicionada e não-condicionada no valor de X_1).
- μ_c é uma função linear de x_1 .

3. Rodar script `acoes.R` com o arquivo `retornos.csv`

4. Neste exercício, você vai gerar alguns vetores gaussianos tri-dimensionais que, de fato vivem em duas dimensões.

```

require(MASS)
nsims=200
Sigma = matrix(c(3,2,2,4),2,2)
pts = mvrnorm(nsims, c(1, 2), Sigma)

pts = cbind(pts, 3*pts[,1]+4*pts[,2])
pairs(pts)

library(scatterplot3d)
scatterplot3d(pts)

library(rgl)
plot3d(pts, col="red", size=3)

A = matrix(c(1, 0, 0, 1, 3, 4), 3, 2, byrow=T)
var.pts = A %*% Sigma %*% t(A)
var.pts

round(cov(pts),2)

eigen(var.pts)
eigen(cov(pts))

```

- Qual os parâmetros μ e Σ da distribuição gaussiana do vetor **pts**?
- Por que o comando `round(cov(pts),2)` não gera exatamente Σ (ignore o erro de aproximação puramente numérico, não estocástico).
- Qual o menor autovalor de Σ ?

Agora, um conjunto de dados simulado que está quase completamente contido num plano do \mathbb{R}^3 .

```

x <- rnorm(1000)
y <- rnorm(1000)
z <- 3 + 1.2*x - 1.2*y + rnorm(1000, sd=0.3)
d2 <- data.frame(x,y,z)
open3d()
plot3d(d2)

```

Isto mostra que não precisamos realmente de 3 dimensões. Esta massa de pontos vive praticamente num espaço de dimensão 2. Este espaço de dimensão 2 é aquele gerado pelas 2 primeiras componentes principais.

Agora, um exemplo com dados reais:

```
?trees # girth = circunferencia
```



```
x= log(trees)
pairs(x)
scatterplot3d(x)
plot3d(x, col="red", size=3)
eigen(cov(x))
```

5. Neste exercício, você vai analisar os dados de uma análise química de vinhos. Você vai ler uma matriz com 178 amostras de diferentes vinhos. Haverá uma linha para cada vinho. A primeira coluna indica o cultivar do vinho (entenda como o tipo de uva usada na fabricação do vinho) tal como Sauvignon Blanc, Cabernet ou Chardonnay (rotulados como 1, 2 ou 3). As 13 colunas seguintes contêm as concentrações de 13 diferentes compostos químicos na amostra.

O objetivo é diferenciar entre os 3 tipos de vinho com base na sua composição química representada pelo vetor 13-dimensional \mathbf{X} . Você precisa criar uma regra para predizer o tipo de vinho (a primeira coluna) a partir das 13 variáveis de composição química. Vamos verificar que, ao invés de usarmos as 13 variáveis, poderemos nos basear em dois índices, os dois primeiros PCAs, que resumem toda a variabilidade simultânea das 13 variáveis.

Estude o script R abaixo. De propósito, ele tem uma quantidade *mínima* de comentários. Procure identificar o que cada linha está fazendo.

WARNING: o help da função `prcomp` é confuso, misturando PCA e análise fatorial nas explicações.

```
arq = "http://archive.ics.uci.edu/ml/machine-learning-databases/wine/wine.data"
wine=read.table(arq, sep=",")

head(wine)
pairs(wine[,2:6])
round(100*cor(wine[,2:14]))
round(apply(wine[,2:14], 2, sd),2)

wine.pca = prcomp(wine[,2:14], scale. = TRUE)
summary(wine.pca)
wine.pca$sdev
sum((wine.pca$sdev)^2)
screeplot(wine.pca, type="lines")

# Barplot das variancias acumuladas
barplot(cumsum(wine.pca$sdev^2)/sum(wine.pca$sdev^2))
# os dois primeiros PCA's explicam aprox 60% da variancia total
# os 5 primeiros explicam aprox 80%
```

```

# Os autovetores
dim(wine.pca$rot)

# O 1o autovetor
wine.pca$rot[,1]

# O 2o autovetor
wine.pca$rot[,2]

# Coordenadas dos pontos ao longo do primeiro componente
fscore1 = wine.pca$x[,1]

# Coordenadas dos pontos ao longo do segundo componente
fscore2 = wine.pca$x[,2]

# plot dos pontos projetados
plot(fscore1, fscore2, pch="*", col=wine[,1]+8)

```

Seja $\mathbf{X}_i = (X_{i1}, \dots, X_{i13})$ a *linha* i da matriz `wine`. Seja $\mathbf{Z}_i = (Z_{i1}, \dots, Z_{i13})$ a *linha* i da matriz `wine` PADRONIZADA. Isto é, $Z_{ij} = (X_{ij} - \bar{x}_j)/s_j$ onde \bar{x}_j é a média aritmética e s_j é o desvio-padrão da coluna j da matriz `wine`. Esta matriz padronizada é obtida com o comando `z = scale(wine[2:14])`:

```

z = scale(wine[2:14])
round(apply(z, 2, mean), 5)
round(apply(z, 2, sd), 5)

```

Vamos considerar \mathbf{Z}_i como um *vetor-coluna* 13-dimensional. Ao invés de usarmos o vetor \mathbf{Z}_i , estamos usando apenas o vetor \mathbf{Y}_i composto pelos dois índices formados pelos dois primeiros componentes principais:

$$\mathbf{Y}_i = \begin{bmatrix} Y_{i1} \\ Y_{i2} \end{bmatrix} = \begin{bmatrix} \mathbf{v}'_1 \mathbf{Z}_i \\ \mathbf{v}'_2 \mathbf{Z}_i \end{bmatrix}$$

onde \mathbf{v}_1 e \mathbf{v}_2 são os dois primeiros autovetores da matriz de correlação de \mathbf{X} .

- Preencha os locais com (??) com os valores numéricos corretos (duas casa decimais apenas):

$$\begin{aligned} Y_{i1} &= (??)Z_{i1} + (??)Z_{i2} + (??)Z_{i3} + \dots + (??)Z_{i,13} \\ Y_{i2} &= (??)Z_{i1} + (??)Z_{i2} + (??)Z_{i3} + \dots + (??)Z_{i,13} \end{aligned}$$

- O último gráfico do acrípt R acima é um plot dos pontos \mathbf{Y}_i dos 178 vinhos. Identifique três regiões do plano Y_1, Y_2 que podem ser usadas para classificar futuras amostras de vinhos em uma das três categorias. Pode apenas esboçar grosseiramente no gráfico a mão livre.



Figura 6.1: Para quem gosta de cerveja, uma Indian Pale Ale de alta qualidade produzida em BH.

- Suponha que uma nova amostra de vinho tem sua composição química medida e encontra-se

$$\mathbf{x} = (13.95, 3.65, 2.25, 18.4, 90.18, 1.55, 0.48, 0.5, 1.34, 10.2, 0.71, 1.48, 587.14)$$

Obtenha seu vetor \mathbf{z} , as suas coordenadas (y_1, y_2) e prediga o seu tipo. Confira sua resposta no final desta lista.

6. Cerveja de excelente qualidade começa a ser produzida no Brasil em pequenas cervejarias e um dos centros mais ativos é a região metropolitana de Belo Horizonte, em especial Nova Lima. Se você gosta, experimente a Kud Kashmir (Figura 6.1).

O arquivo `beer.txt` é um dataset da página de Karl Wuensch, East Carolina University. Uma nova cervejaria está interessada em conhecer o comportamento de escolha do consumidor de cerveja artesanal. Um grupo de 231 consumidores avaliaram a importância de sete qualidades ao decidir se deve ou não comprar uma cerveja. Para cada qualidade, foi dada uma nota numa escala de 0 a 100 para sua importância. As sete qualidades ou variáveis são as seguintes:

- COST: baixo custo por volume (300ml de cerveja)
- SIZE: grande tamanho da garrafa (volume)
- ALCOHOL: alto percentual de álcool da cerveja
- REPUTATION: boa reputação da marca
- COLOR: a cor da cerveja
- AROMA: agradável aroma da cerveja
- TASTE: gosto saboroso da cerveja

A variável SES é uma categoria de status socioeconômico (valores maiores significam status mais elevados). A variável grupo não é explicada, não sei do que se trata. Ignore-a durante o exercício.

O script abaixo executa o seguinte: Leia os dados numa matriz. Use `summary(beer)` (ou olhe os dados na tela) para verificar que existem 11 NAs na variável AROMA. Obtenha a matriz de covariância S das 7 variáveis de qualidade e verifique que os seus desvios-padrão não são muito distintos. Obtenha a matriz de correlação R .

```
beer = as.matrix(read.table("beer.txt", header=T))
summary(beer)
```

```
S = var(beer[,1:7], na.rm=T)
S
sqrt(diag(S)) # sd's not very different
```

```
R = cor(beer[,1:7], use = "complete.obs")
```

```
round(100*R)
```

Você deve gastar um tempo olhando a matriz de correlação R , a menos que ela seja muito grande. Você está planejando usar PCA ou FA para capturar a essência das correlações nesta matriz. Observe que há muitas correlações grandes e médias em R . Todas as variáveis tem algumas correlações grandes, com a exceção de `reputation` que é moderadamente (e negativamente) correlacionada com todo o resto. É óbvio que existe uma estrutura de correlação entre as variáveis.

O pacote `corrplot` permite visualizar a matriz de correlação R de um jeito muito legal. Veja em <http://cran.r-project.org/web/packages/corrplot/vignettes/corrplot-intro.html>. Instale e carregue este pacote. Faça um gráfico na forma da matriz R em que cada célula possui uma elipse representando grau de correlação entre as duas variáveis. Quanto mais achatada e parecida com uma linha reta, mais correlacionadas são as duas variáveis. Se a elipse for parecida com um círculo, é sinal de que a correlação é próxima de zero. Neste caso, a imagem estará quase transparente. Correlações positivas são azuis, negativas são vermelhas.

```
library(corrplot)
corrplot(R, method = "ellipse")
# plotando as elipses e os valores das correlacoes
corrplot.mixed(R, upper = "ellipse")
# rearranjando as linhas e colunas para agrupar variaveis com correlacoes parecidas
corrplot.mixed(R, order = "AOE", upper = "ellipse", cl.align = "r")
```

Parece haver dois grupos de variáveis, um formado por `COST`, `ALCOHOL`, `SIZE` e outro formado por `COLOR`, `AROMA`, `TASTE`. Elas são bem positivamente correlacionadas dentro de cada grupo e, ao mesmo tempo, pouco correlacionadas com as variáveis do outro grupo.

Uma variável, `REPUTATION`, forma um grupo à parte, sendo fracamente e negativamente correlacionada com todas as outras seis.

Gere uma nova matriz eliminando as poucas linhas em que existem NAs. A seguir, obtenha os autovetores e autovalores da matriz de covariância S com a função `eigen`. Vamos trabalhar com matriz de covariância porque os desvios-padrão das setes variáveis são parecidos.

```
newbeer = na.omit(beer)
S = cov(newbeer[,1:7])
fit = eigen(S) # usa o algoritmo QR em cima da matriz S
# autovalores
fit$values
# autovetores
fit$vectors
```

Na análise acima, tivemos de gerar a matriz de covariância e, a seguir, passá-la à função `eigen`. A função `eigen` não enxerga mais os dados originais, somente a matriz R . Outra maneira é fornecer diretamente a matriz X de dados $n \times p$ e pedir que os componentes principais da matriz de covariância induzida (ou da matriz de correlação) seja calculada. A função `prcomp` faz isto através da decomposição SVD de X .

```
pca.beer = prcomp(newbeer[,1:7])

# Se quiser obter PCA da matriz de correlação, use
# pca.beer = prcomp(newbeer[,1:7], scale. = TRUE)

# Os 7 autovetores
pca.beer$rot

# Os 7 autovalores
(pca.beer$sdev)^2

# verifique que os autovetores acima são os mesmos daqueles retornados por eigen.
# verifique que os autovetores tem norma euclidiana = 1.
# Por exemplo, o 1o PCA:
sum(pca.beer$rot[,1]^2)

# Grafico scree com os 7 autovalores (ou variancias de cada PCA)
plot(pca.beer)

# Barplot das variancias acumuladas indicando a escolha de 2 PCAs
barplot(cumsum(pca.beer$sdev^2))

# Resumo
```

```
summary(pca.beer)

# Note que o quadrado da linha Standard deviation acima eh igual aos autovalores
# obtidos com fit$values

# Vamos usar apenas os dois 1os PCs para representar R com dois fatores
# Carga do Fator = sqrt(LAMBDA) * EIGENVECTOR

cargafat1 = pca.beer$sdev[1] * pca.beer$rot[,1]
cargafat2 = pca.beer$sdev[2] * pca.beer$rot[,2]

# matriz de cargas
L = cbind(cargafat1, cargafat2)

rownames(L) = rownames(R)[1:7]

round(L, 2)

plot(L, type="n",xlim=c(-40, 20), ylim=c(-10, 25))
text(L, rownames(L))
abline(h=0)
abline(v=0)
```

A interpretação dos resultados obtidos não é simples. Com os eixos rotacionados conseguiremos um resultado bem mais interpretável. Não existe uma rotina nativa em R para obter a rotação ótima dos fatores no caso da estimação pelo método de componentes principais. Em R, a rotação ótima está implementada apenas para a estimação das cargas L por meio do método de máxima verossimilhança, um método que veremos em breve. Para o caso do método de componentes principais, o último gráfico mostra que uma rotação horária de aproximadamente $90^\circ + 15^\circ$ ou $\pi/2 + 15(\pi/180)$ deve colocar a maioria dos pontos em apenas um dos dois eixos ortogonais:

```
# Fazendo manualmente uma rotacao horaria de pi/2+15*pi/180
phi = pi/2 + 15*(pi/180)
T = matrix(c(cos(phi), -sin(phi), sin(phi), cos(phi)), ncol=2, byrow=T)

Lstar = L %*% T # usando a multiplicacao por linha da matriz L

plot(Lstar, type="n", xlim=c(-20, 30), ylim=c(-15, 35))
text(Lstar, rownames(L))
abline(h=0); abline(v=0)

round(Lstar,2)
```

A interpretação dos fatores é bem mais simples agora. O primeiro fator tem cargas positivas e grandes em COLOR, AROMA, TASTE e uma carga negativa moderada em REPUTATION.

Algém com uma nota (ou escore) muito elevado neste fator é alguém que preza e diferencia qualidades ligadas ao paladar da cerveja e também seus aspectos estéticos. Este componente poderia ser chamado de *Degustador*.

Os indivíduos que tiverem seu segundo fator muito positivo terão dado notas altas para os aspectos de **COST**, **ALCOHOL**, **SIZE** e, ao mesmo tempo, dado uma nota moderadamente baixa para **REPUTATION**. Alguém que possui uma nota muito alta neste segundo fator é alguém que gosta de muita cerveja barata e com muito álcool, e não se importa muito com a reputação da cerveja. Este componente poderia ser chamado de *Bebum Barato*.

Para obter uma estimativa das variâncias dos fatores específicos (isto é, da matriz Ψ), usamos o código abaixo:

```
matpsi = diag(diag(S - Lstar %*% t(Lstar)))
round(matpsi, 2)

sum( (S - Lstar %*% t(Lstar) - matpsi)^2 )/sum(S^2)
```

O último comando mostra que a matriz residual $\Sigma - \mathbf{L}^*(\mathbf{L}^*)' - \Psi$ tem uma soma de suas entradas (ao quadrado) muito pequena em comparação com a soma das entradas na matriz de covariância Σ (apenas 0.005303857 ou 0.5%).

No nosso modelo, os indivíduos recebem escores *independentes* destes dois fatores. Eles não são fatores competidores, um indivíduo pode receber altas doses dos dois fatores. Ele pode gostar de tomar muita cerveja com muito álcool e que seja barata. Isto é, ter um escore alto no fator 2. Ao mesmo tempo, este mesmo indivíduo pode apreciar também as cervejas mais refinadas, mais caras e com mais sabor e aroma. Isto é, ter um escore alto também no fator 1. Em suma, ele pode ser um esteta que adora se embriagar.

Para encontrar uma estimativa dos escores dos dois fatores para cada um dos 220 indivíduos que restaram na matriz **newbeer** após eliminar as 11 linhas com NAs, usamos o procedimento de regressão linear. Lembre-se que o modelo de análise fatorial estabelece que as 7 notas do indivíduo i é representada por

$$\underset{(7 \times 1)}{\mathbf{X}_i} = \underset{(7 \times 1)}{\boldsymbol{\mu}} + \underset{(7 \times 2)}{\mathbf{L}^*} \underset{(2 \times 1)}{\mathbf{F}_i} + \underset{(7 \times 1)}{\boldsymbol{\epsilon}_i}$$

onde $\mathbf{F}_i' = (F_{1i}, F_{2i})$ são os escores (ou as doses) que o indivíduo i possui dos fatores 1 e 2. Como observamos diretamente \mathbf{X}_i e como estimamos a média populacional $\boldsymbol{\mu}$ e a matriz de cargas rotacionadas \mathbf{L}^* , podemos usar mínimos quadrados ou regressão linear para estimar os escores F_{1i} e F_{2i} .

Por exemplo, o primeiro indivíduo na matriz **newbeer** tem a sua representação fatorial

$$\mathbf{X}_1 = \begin{bmatrix} 90 \\ 80 \\ 70 \\ 20 \\ 50 \\ 70 \\ 60 \end{bmatrix} = \hat{\boldsymbol{\mu}} + \hat{\mathbf{L}}^* \hat{\mathbf{F}}_1 + \hat{\boldsymbol{\epsilon}}_1 = \begin{bmatrix} 47.25 \\ 43.50 \\ 46.50 \\ 48.25 \\ 51.00 \\ 44.75 \\ 67.25 \end{bmatrix} + \begin{bmatrix} 0.11 & 31.74 \\ 4.84 & 32.07 \\ 3.09 & 30.19 \\ -12.95 & -10.83 \\ 25.81 & 0.05 \\ 24.79 & -1.37 \\ 22.94 & -2.28 \end{bmatrix} \begin{bmatrix} F_{1i} \\ F_{2i} \end{bmatrix} + \begin{bmatrix} \hat{\epsilon}_{11} \\ \hat{\epsilon}_{21} \\ \hat{\epsilon}_{31} \\ \hat{\epsilon}_{41} \\ \hat{\epsilon}_{51} \\ \hat{\epsilon}_{61} \\ \hat{\epsilon}_{71} \end{bmatrix}$$

A matriz $\hat{\mathbf{L}}^*$ está na matriz **Lstar** no final do script R e é a mesma para todos os indivíduos. O vetor $\hat{\boldsymbol{\mu}}$ também é o mesmo para todos os indivíduos e é obtido simplesmente tomando a média aritmética de cada uma das sete qualidades de modo que $\boldsymbol{\mu}$ é aproximadamente igual ao resultado do comando `mu = apply(newbeer[,1:7], 2, mean)`.

```
> apply(newbeer[,1:7], 2, mean)
      COST      SIZE ALCOHOL REPUTAT      COLOR      AROMA      TASTE
47.25  43.50  46.50  48.25  51.00  44.75  67.25
```

Assim, para o indivíduo i podemos estimar seus escores F_{1i} e F_{2i} pelos valores \hat{F}_{1i} e \hat{F}_{2i} que minimizam o comprimento (ao quadrado) da diferença entre \mathbf{X}_i e $\hat{\boldsymbol{\mu}} + \hat{\mathbf{L}}^* \mathbf{F}_i$:

$$\underset{\mathbf{F}_i}{\operatorname{argmin}} \|\mathbf{X}_i - \hat{\boldsymbol{\mu}} - \hat{\mathbf{L}}^* \mathbf{F}_i\|^2$$

Ou seja, para o indivíduo $i = 1$, queremos o vetor \mathbf{F}_1 que minimize a norma euclidiana (ao quadrado) do vetor

$$\mathbf{X}_1 - \hat{\boldsymbol{\mu}} - \hat{\mathbf{L}}^* \mathbf{F}_1 = \begin{bmatrix} 90 - 47.25 \\ 80 - 43.50 \\ 70 - 46.50 \\ 20 - 48.25 \\ 50 - 51.00 \\ 70 - 44.75 \\ 60 - 67.25 \end{bmatrix} - \begin{bmatrix} 0.11 & 31.74 \\ 4.84 & 32.07 \\ 3.09 & 30.19 \\ -12.95 & -10.83 \\ 25.81 & 0.05 \\ 24.79 & -1.37 \\ 22.94 & -2.28 \end{bmatrix} \begin{bmatrix} F_{11} \\ F_{21} \end{bmatrix}$$

O seguinte código em R faz isto através de loop sobre as linhas da matriz **newbeer**:

```
## Factor scores dos n=220 individuos
factors = matrix(0, nrow=nrow(beer), ncol=2)
mu = apply(newbeer[,1:7], 2, mean)
for(i in 1:nrow(newbeer)){
  y = newbeer[i, 1:7] - mu
  factors[i,] = lm(y ~ 0 + Lstar)$coef
}
```


Podemos visualizar os fatores de cada um dos 220 indivíduos pedindo um plot da matriz `factors`:

```
plot(factors, xlab="fator 1", ylab="fator2")

# mas... onde estao os 220 individuos?
# Varios individuos poduziram o MESMO valor x --> estimamos com os mesmos fatores

plot(jitter(factors, amount=0.05), xlab="fator 1", ylab="fator2")
```

Como vários indivíduos produziram o mesmo vetor \mathbf{X}_i seus fatores \mathbf{F}_i também coincidem. Assim, o comando `jitter` foi usado. Ele perturba as coordenadas de cada ponto aleatoriamente com um ruído uniforme entre `-amount` e `+amount`. No novo plot, podemos enxergar todos os indivíduos.

Respostas

- Para o problema do vinho:

```
x = c(13.95, 3.65, 2.25, 18.4, 90.18, 1.55, 0.48, 0.5, 1.34, 10.2, 0.71, 1.48, 58)
z = (x - apply(wine[,2:14], 2, mean))/apply(wine[,2:14], 2, sd)
y1 = sum( wine.pca$rot[,1] * z)
y2 = sum( wine.pca$rot[,2] * z)
plot(fscore1, fscore2, pch="*", col=wine[,1]+8)
points(y1, y2, pch="*", cex=4)
```

Capítulo 7

Classificação

1. Replicar a análise de classificação usando a função LDA de Fisher em duas páginas da web (uma sendo a sequência da seguinte): <http://www.aaronschlegel.com/discriminant-analysis/> e <https://www.r-bloggers.com/classification-with-linear-discriminant-analysis/>. Os dados não estão imediatamente visíveis apontado pelas páginas mas eu os coloquei na página da nossa disciplina.

-
2. Existem duas classes ou populações, 1 e 2, presentes nas proporções positivas π_1 e π_2 com $\pi_1 + \pi_2 = 1$. Suponha que o vetor aleatório contínuo $\mathbf{X} = (X_1, \dots, X_p)$ com p variáveis possua as densidades $f_1(\mathbf{x})$ e $f_2(\mathbf{x})$ quando o indivíduo pertence à população 1 ou 2, respectivamente. Sejam $c(1|2)$ o custo do erro de classificar erradamente no grupo 1 um indivíduo que seja do grupo 2. Analogamente, defina o custo do outro erro $c(2|1)$. A região ótima R_1 de classificação no grupo 1 é dada pela seguinte região do espaço \mathbb{R}^p :

$$R_1 = \left\{ \mathbf{x} \in \mathbb{R}^p \text{ tais que } \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} > \frac{c(1|2) \pi_2}{c(2|1) \pi_1} \right\}$$

Note que a ordem das populações na última fração é oposta à ordem na razão das densidades. Isto é, comparamos f_1/f_2 com π_2/π_1 .

- Suponha que $c(1|2) = c(2|1)$ e que $\pi_1 = \pi_2$. Neste caso, a regra ótima fica reduzida a uma simples comparação. Qual é esta regra de classificação?
- Imagine agora que $\pi_1 = 0.01$ e que $c(1|2) = c(2|1)$. Para tornar as coisas mais concretas, suponha que a população 1 sejam portadores de certo vírus e a população 2, os demais. A regra simples do item acima fica modificada. Agora não basta que $f_1(\mathbf{x})$ seja maior que $f_2(\mathbf{x})$. Ela precisa ser bem maior que $f_2(\mathbf{x})$. Quantas vezes maior $f_1(\mathbf{x})$ deve ser para que classifiquemos o item com característica \mathbf{x} em 1?
- Suponha que os custos de má-classificação sejam muito diferentes. O custo de classificar o portador do vírus como não pode custar-lhe a vida ou a vida de outras pessoas. Por outro lado, o indivíduo saudável ser classificado como infectado custa mais exames confirmatórios, algumas medidas de isolamento e outras coisas que são

-
3. Um programa é usado para classificar fotos de gatos (população 1) versus fotos de não-gatos (população 2). As fotos da população 1 (fotos de gatos) são chamadas de *relevantes*. O classificador seleciona algumas fotos para classificar no grupo 1 baseado em features aleatórias no vetor \mathbf{X} . A regra de classificação é representada pela função binária $D(\mathbf{X})$ que assume os valores 1 ou 2 dependendo do vetor aleatório \mathbf{X} cair ou não na região R_1 de classificação no grupo 1.

Haverá erros nesta classificação e queremos torná-los pequenos. Duas métricas muito populares para avaliar a qualidade de um classificador são: *precisão* (precision, em inglês) e *revocação* (recall, em inglês). A palavra revocação não é muito usada na linguagem diária. Ela significa “fazer voltar, retornar, chamar novamente”. Pode significar também revogação, anulamento de um contrato mas não é este o significado relevante para nosso contexto.

- Precisão: $\mathbb{P}(\text{foto} \in \text{gatos} \mid \text{classificado como gato}) = \mathbb{P}(\mathbf{X} \in 1 \mid D(\mathbf{x}) = 1)$
- Revocação: $\mathbb{P}(\text{classificado como gato} \mid \text{foto} \in \text{gatos}) = \mathbb{P}(D(\mathbf{x}) = 1 \mid \mathbf{X} \in 1)$

É claro que, tanto para precisão quanto para revocação, quanto maior, melhor. Precisão e revocação são probabilidades condicionais usando os mesmos eventos A e B mas um deles é $\mathbb{P}(A|B)$ enquanto o outro é simplesmente $\mathbb{P}(B|A)$. Sabemos que estas probabilidades podem ser muito diferentes. A Figura 7.1, retirada da página **Precision and recall** na Wikipedia, mostra itens nas suas classes reais: relevante (pop 1) ou não (pop 2). Mostra também a sua classificação na classe 1 (os itens dentro da elipse central) ou na classe 2 (os restantes). A Figura ainda mostra as probabilidades precisão e revocação como diagramas de Venn dos eventos envolvidos.

Marque V ou F nas afirmativas a seguir:

- A precisão mede o quanto os resultados da classificação são úteis.
- A revocação mede o quanto os resultados da aplicação da regra de classificação são completos.
- A soma de precisão e revocação é igual a 1.
- $\text{Precisão} = \text{Revocação} \times \frac{\mathbb{P}(\mathbf{X} \in 1)}{\mathbb{P}(D(\mathbf{x}) = 1)}$.
- Existe um trade-off entre precisão e revocação: se aumentarmos uma métrica, a outra tem de diminuir.

Solução: VVFVF

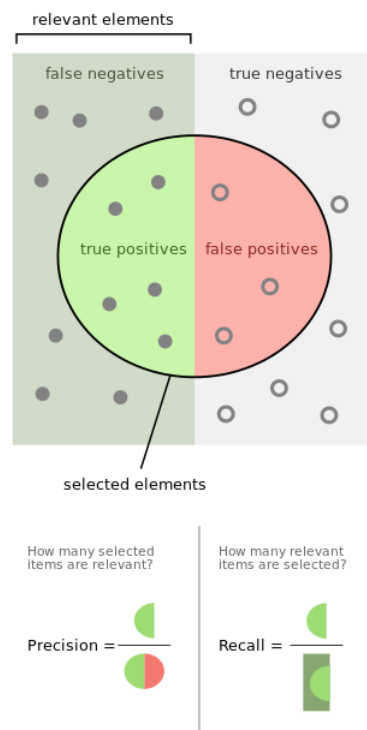


Figura 7.1: Retirado da Wikipedia.

- 74 4. Existem duas classes ou populações, 1 e 2, presentes nas proporções positivas π_1 e π_2 com $\pi_1 + \pi_2 = 1$. Suponha que $p_{i1} \approx 0$. O vetor aleatório contínuo $\mathbf{X} = (X_1, \dots, X_p)$ com p variáveis possua as densidades $f_1(\mathbf{x})$ e $f_2(\mathbf{x})$ quando o indivíduo pertence à população 1 ou 2, respectivamente. Sejam $c(1|2)$ o custo do erro de classificar erradamente no grupo 1 um indivíduo que seja do grupo 2. Analogamente, defina o custo do outro erro $c(2|1)$. A regra de classificação é representada pela função binária $D(\mathbf{X})$ que assume os valores 1 ou 2 dependendo do vetor aleatório \mathbf{X} cair ou não na região R_1 de classificação no grupo 1.

- Uma regra de decisão que vai errar pouco será atribuir a classe 2 a todo e qualquer item: $D(\mathbf{X}) \equiv 2$ para todo valor de \mathbf{X} . Obtenha a probabilidade de classificação errada. A probabilidade é próxima de zero?
- Se o custo de má-classificação for também desbalanceado, com $c(2|1) \gg c(1|2)$, a estratégia anterior pode ser muito ruim. Obtenha o custo esperado de má-classificação (ECM) da regra anterior.

Solução: π_1 e $ECM = c(2|1)\pi_1$

5. Você quer classificar objetos em duas classes, 1 ou 2, com base numa única variável X , um vetor de dimensão 1, usando a regra de classificação ótima. Seja $f_1(x) = (1 - |x|)/2$ para $x \in (-1, 1)$ a densidade de X na população 1 e $f_2(x) = (1 - |x - 0.5|)/2$ para $-0.5 < x < 1.5$ a densidade de X na população 2. Suponha que os custos de classificação errada são $c(2|1) = a$ e $c(1|2) = 2a$. Além disso, assuma que $p_1 = 0.3$ e $p_2 = 1 - p_1 = 0.7$.

- Esboce as duas densidades de probabilidade num gráfico.
 - Identifique as regiões de classificação ótima R_1 e R_2 .
 - Assumindo que $p_1 = p_2$, identifique as regiões.
 - Assuma agora que, além das probabilidades a priori, os custos também são iguais.
-

6. Na população 1, o vetor \mathbf{X} possui distribuição $N_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ e distribuição $N_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$ na população 2. Seja $d_k^2(\mathbf{x}, \boldsymbol{\mu}_k)$ a distância de Mahalanobis avaliada com os parâmetros $\boldsymbol{\mu}_k$ e $\boldsymbol{\Sigma}_k$ da população k . Mostre que a regra de classificação ótima implica que um novo objeto com medições \mathbf{x} é alocado a população 1 se

$$d_1^2(\mathbf{x}, \boldsymbol{\mu}_1) - d_2^2(\mathbf{x}, \boldsymbol{\mu}_2) + \log \left(\frac{|\boldsymbol{\Sigma}_1|}{|\boldsymbol{\Sigma}_2|} \right) \leq k$$

onde $|A| = \det(A)$. Encontre a constante k em função dos custos e probabilidades a priori p_1 e $p_2 = 1 - p_1$. Obtenha esta constante no caso de custos iguais e $p_1 = p_2$.

7. Quando temos $g > 2$ populações, a regra de classificação ótima aloca \mathbf{x} à população j para a qual 75

$$\sum_{\substack{i=1 \\ i \neq j}}^g p_i f_i(\mathbf{x}) c(j| \in i)$$

é mínimo. As probabilidades *a priori* p_1, \dots, p_g somam 1 e $c(j| \in i)$ é o custo de classificar em j um indivíduo da população i . Mostre que a regra vista em sala de aula é equivalente a esta no caso de $g = 2$ populações.

8. Suponha que $f(\mathbf{x})$ é a densidade de uma gaussiana $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ para $\mathbf{x} \in \mathbb{R}^p$. Sejam $\lambda_1 < \lambda_2 < \dots < \lambda_p$ os autovalores de $\boldsymbol{\Sigma}$ com os correspondentes autovetores $\mathbf{e}_1, \dots, \mathbf{e}_p$. Responda V ou F:

- A chance de observar um valor \mathbf{x} distante do perfil médio $\boldsymbol{\mu}$ decresce mais rápido se nos afastarmos de $\boldsymbol{\mu}$ ao longo da direção \mathbf{e}_1 .
- Se os autovalores λ_i forem todos iguais, os pontos \mathbf{x} que tem a mesma chance de serem selecionados ficam localizados em esferas centradas em $\boldsymbol{\mu}$.
- A regra de classificação ótima para duas populações com $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$ projeta ortogonalmente cada dado \mathbf{x} ao longo do autovetor com menor autovalor.

9. Thomson e Randall-Maciver (1905) escavaram e obtiveram crânios que, de acordo com o local em que foram encontrados, puderam ser datados em cinco períodos distintos da história do império egípcio.

- the early predynastic period (circa 4000 BC)
- the late predynastic period (circa 3300 BC)
- the 12th and 13th dynasties (circa 1850 BC)
- the Ptolemaic period (circa 200 BC)
- the Roman period (circa 150 BC)

Measurements in mm on of 30 male Egyptian skulls from each period were taken. The variables are:

- MB: Maximal Breadth of Skull
- BH: Basibregmatic Height of Skull
- BL: Basialveolar Length of Skull
- NH: Nasal Height of Skull

Queremos analisar os dados para determinar se existem diferenças na distribuição de probabilidade dos tamanhos dos crânio ao longo do tempo. Isto é, a distribuição estatística

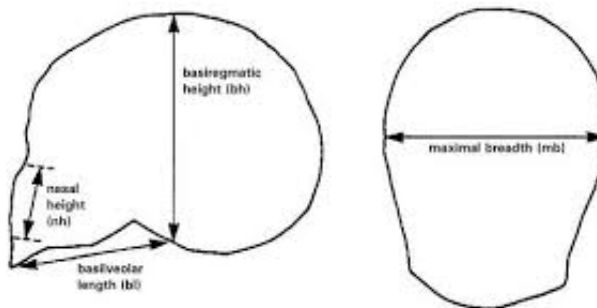


Figura 7.2: Medições em crânios encontrados em sítios arqueológicos.

dos tamanhos de crânios variou ao longo do tempo? Antropólogos teorizam que uma mudança no tamanho do crânio ao longo do tempo é uma evidência da miscigenação dos egípcios com populações de imigrantes ao longo dos séculos.

REF: Thomson, A. and Randall-Maciver, R. (1905) *Ancient Races of the Thebaid*, Oxford: Oxford University Press. Data also found in: Manly, B.F.J. (1986) *Multivariate Statistical Methods*, New York: Chapman & Hall.

Leia os dados do arquivo `EgyptianSkull.txt`, crie duas classes (ou populações) agregando as duas primeiras e as duas últimas e deletando os crânios do período intermediário, das 12a. e 13a. dinastias (cerca de 1850 a.C.).

Separe 4 dos crânios para alocar a uma das duas populações. Crie a regra de classificação com os crânios restantes em duas situações: com $\Sigma_1 = \Sigma_2$ e com $\Sigma_1 \neq \Sigma_2$.

10. Examples of the character images generated by these procedures are presented in Figure 7.3. Each character image was then scanned, pixel by pixel, to extract 16 numerical attributes. These attributes represent primitive statistical features of the pixel distribution. To achieve compactness, each attribute was then scaled linearly to a range of integer values from 0 to 15. This final set of values was adequate to provide a perfect separation of the 26 classes. That is, no feature vector mapped to more than one class. The attributes (before scaling to 0-15 range) are:

- The horizontal position, counting pixels from the left edge of the image, of the center of the smallest rectangular box that can be drawn with all on pixels inside the box.
- The vertical position, counting pixels from the bottom, of the above box.
- The width, in pixels, of the box.
- The height, in pixels, of the box.
- The total number of “on” pixels in the character image.
- The mean horizontal position of all “on” pixels relative to the center of the box and divided by the width of the box. This feature has a negative value if the image is “leftheavy” as would be the case for the letter L.



Figura 7.3: Examples of the character images from which features were extracted.

- (g) The mean vertical position of all “on” pixels relative to the center of the box and divided by the height of the box.
- (h) The mean squared value of the horizontal pixel distances as measured in 6 above. This attribute will have a higher value for images whose pixels are more widely separated in the horizontal direction as would be the case for the letters W or M.
- (i) The mean squared value of the vertical pixel distances as measured in 7 above.
- (j) The mean product of the horizontal and vertical distances for each “on” pixel as measured in 6 and 7 above. This attribute has a positive value for diagonal lines that run from bottom left to top right and a negative value for diagonal lines from top left to bottom right.
- (k) The mean value of the squared horizontal distance times the vertical distance for each “on” pixel. This measures the correlation of the horizontal variance with the vertical position.
- (l) . The mean value of the squared vertical distance times the horizontal distance for each “on” pixel. This measures the correlation of the vertical variance with the horizontal position.
- (m) The mean number of edges (an “on” pixel immediately to the right of either an “on” pixel or the image boundary) encountered when making systematic scans

from left to right at all vertical positions within the box. This measure distinguishes between letters like “W” or “M” and letters like “T” or “L”.

- (n) The sum of the vertical positions of edges encountered as measured in 13 above. This feature will give a higher value if there are more edges at the top of the box, as in the letter “Y”.
- (o) The mean number of edges (an “on” pixel immediately above either an “off” pixel or the image boundary) encountered when making systematic scans of the image from bottom to top over all horizontal positions within the box.
- (p) The sum of horizontal positions of edges encountered as measured in 15 above. A data file of the 16 attribute values and outcome category for each of the 20,000 stimulus items is on file with David Aha (aha@ics.uci.edu). The set of 20,000 unique letter images was organized into two files. Sixteen thousand items were used as a learning set and the remaining 4000 items were used for testing the accuracy of the rules.

11. Este exercício foi extraído da página web do livro *The Elements of Statistical Learning* de Hastie, Tibshirani e Friedman (2009), editado pela Springer-Verlag: <http://statweb.stanford.edu/~tibs/ElemStatLearn/> Este é um dos melhores livros de Machine Learning no momento. Clique no link *Data* e no link *ZIP code* para encontrar os dados e a sua descrição .

Normalized handwritten digits, automatically scanned from envelopes by the U.S. Postal Service. The original scanned digits are binary and of different sizes and orientations; the images here have been deslanted and size normalized, resulting in 16 x 16 grayscale images (Le Cun et al., 1990).

The data are in two gzipped files, and each line consists of the digit id (0-9) followed by the 256 grayscale values.

There are 7291 training observations and 2007 test observations, distributed as follows: 0 1 2 3 4 5 6 7 8 9 Total Train 1194 1005 731 658 652 556 664 645 542 644 7291 Test 359 264 198 166 200 160 170 147 166 177 2007

or as proportions: 0 1 2 3 4 5 6 7 8 9 Train 0.16 0.14 0.1 0.09 0.09 0.08 0.09 0.09 0.07 0.09 Test 0.18 0.13 0.1 0.08 0.10 0.08 0.08 0.07 0.08 0.09

Encare cada Assim, as porporções Alternatively, the training data are available as separate files per digit (and hence without the digit identifier in each row)

The test set is notoriously “difficult”, and a 2.5% error rate is excellent. These data were kindly made available by the neural network group at AT&T research labs (thanks to Yann Le Cun).

Problema dos crânios:

```
# R script for some basic classification and
# lda = linear discriminant analysis

skull <- read.table(file="EgyptianSkull.txt",header=T)

dim(skull)
head(skull)

period = skull[,5]
period[skull[,5] < -3000] = 1
period[skull[,5] < -1000 & skull[,5] > -3000] = 2
period[skull[,5] > -1000] = 3

colsk = c("red","green","blue")[period]
pairs(skull[,1:4], main="Egyptian skull", pch=21, bg=colsk)

# pch=21 specifies the marker type. See help(pch)
# bg = background colour of the marker
# In our example we want a different colour for each period.

# Os pontos das tres classes parecem bem misturados.
# Parece dificil ser capaz de classifica-los sem muito erro.
# Vamos fixar a atencao nos dois periodos mais extremos para
# trabalhar apenas com duas classes

skull = skull[period == 1 | period == 3, 1:4]
row.names(skull) = 1:120
period = period[period == 1 | period == 3]
period[period==3] = 2

colsk = c("red","blue")[period]
pairs(skull[,1:4], main="Egyptian skull, 2 periods", pch=21, bg=colsk)

# separando alguns dados, 2 de cada periodo,
# para classificar posteriormente:

teste = skull[c(23, 52, 88, 111), ]
treino = skull[-c(23, 52, 88, 111), ]
period.treino = rep(1:2, c(58,58))

# Visualizando os conjuntos de teste e treino
```

```

colsk = c("red","blue")[period]
colsk[c(23, 52, 88, 111)] = "green"
mark = rep(21, nrow(skull))
mark[c(23, 52, 88, 111)] = 22
pairs(skull[,1:4], pch=mark, bg=colsk)

# Regra de classificacao otima supondo Sigma_1 = Sigma_2
# vetor de medias das 4 variaveis
mu1 = apply(treino[period.treino ==1, ], 2, mean)
mu2 = apply(treino[period.treino ==2, ], 2, mean)
matcov1 = cov(treino[period.treino ==1, ])
matcov2 = cov(treino[period.treino ==2, ])
matcov = (matcov1 + matcov2)/2
maha1 = mahalanobis(teste, mu1, matcov)
maha2 = mahalanobis(teste, mu2, matcov)

maha1 - maha2
# O segundo ponto eh alocado a pop1, os demais a pop2
# Assim, cometemos um erro com o primeiro ponto, que
# deveria ser alocado a pop1

# Agora, vamos refazer os calculos supondo Sigma_1 != Sigma_2
mu1 = apply(treino[period.treino ==1, ], 2, mean)
mu2 = apply(treino[period.treino ==2, ], 2, mean)
matcov1 = cov(treino[period.treino ==1, ])
matcov2 = cov(treino[period.treino ==2, ])
det1 = log(det(matcov1))
# este termo adicional, log da matriz de covariancia,
# precisa ser subtraido da distancia de Mahalanobis
det2 = log(det(matcov2))

d1 = mahalanobis(teste, mu1, matcov) - det1
d2 = mahalanobis(teste, mu2, matcov) - det2

d1-d2

# Como a amostra e' muito pequena, vamos avaliar a classificacao
# omitindo um ponto x de cada vez da base, ajustando os parametros
# mu e Sigma SEM ESTE ponto x e calculando
# a distancia de Mahalanobis entre x e mu
# Esta seja a distancia que usariamos caso quisessemos alocar
# o novo ponto x usando os outros dados.
# Vamos avaliar as taxas de erro cometidos.
# Assumimos custo iguais e proporcoes iguais nas
# duas populacoes

```

```

maha = matrix(0, nrow=nrow(skull), ncol=2)

pop1 = skull[period == 1, ]
pop2 = skull[period == 2, ]
mu1 = apply(pop1, 2, mean); mu2 = apply(pop2, 2, mean)
matcov1 = cov(pop1); matcov2 = cov(pop2)
det1 = log(det(matcov1)); det2 = log(det(matcov2))

for(i in 1:60){
  aux1 = pop1[-i,]
  aux2 = pop2[-i,]
  mu1i = apply(aux1, 2, mean)
  mu2i = apply(aux2, 2, mean)
  matcov1i = cov(pop1[-i,])
  matcov2i = cov(pop2[-i,])
  det1i = log(det(matcov1i))
  det2i = log(det(matcov2i))
  maha[i,1] = mahalanobis(pop1[i,], mu1i, matcov1i) - det1i
  maha[i,2] = mahalanobis(pop1[i,], mu2, matcov2) - det2
  maha[i+60,1] = mahalanobis(pop2[i,], mu1, matcov1) - det1
  maha[i+60,2] = mahalanobis(pop2[i,], mu2i, matcov2i) - det2i
}

difmaha = maha[,1] - maha[,2]
boxplot(difmaha ~ period)

## Vemos que a maioria dos pontos da pop1 possuem distancia de
## Mahalanobis a pop1 menor que a distancia a pop 2 (isto eh,
## difmaha < 0 quando pop1 ==1), enquanto o oposto ocorre
## com os pontos da pop2.

# binaria indicando quem seria classificado em pop2
class2 = difmaha > 0

tabmaha = table(class2, period)
tabmaha

# proporcao de acerto global
sum(diag(tabmaha)) / sum(tabmaha)
# acerta 65.8% dos dados, independentemente de onde venham

# propocao de acerto dentro de cada populacao, estimativa de
# P(classif em k | pertence a k)
prop.table(table(class2, period), margin = 2)

```

82 Acerta 61.7% para x vindo da pop1 e acerta 70% para x vindo de pop2

```
# A probab reversa: P(pertence a k | classif em k)
prop.table(table(class2, period), margin = 1)
# 67.3% dos classificados na pop 1 sao, de fato, da pop1
# 64.6% dos classificados na pop 2 sao, de fato, da pop2
```

Capítulo 8

Teoremas Limite: LGN e TCL

Os exercícios abaixo são do curso de Patrick Breheny na Univ de Kentucky, o autor do material que usei em sala de aula: <http://web.as.uky.edu/statistics/users/pbreheny/index.html>. Por favor, leia o material para entender algumas das questões. São exercícios básicos, que exigem simples manipulação da distribuição normal. O fato fundamental que precisa ser usado várias vezes é o seguinte: se $\bar{X} \sim N(\mu, \sigma^2/n)$ então $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0, 1)$. Portanto, para qualquer valor a , temos

$$\mathbb{P}(\bar{X} > a) = \mathbb{P}\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} > \frac{a - \mu}{\sigma/\sqrt{n}}\right) \approx \mathbb{P}\left(N(0, 1) > \frac{a - \mu}{\sigma/\sqrt{n}}\right) = 1 - \text{pnorm}(b)$$

onde $b = \sqrt{n}(a - \mu)/\sigma$.

-
1. Você quer selecionar uma amostra para estimar a porcentagem θ de pessoas que vai votar num candidato X . Imagine que a resposta é uma v.a. X de Bernoulli com valores 1 e 0 (vai e não vai votar, respectivamente) e a probabilidade de sucesso é θ . As respostas de n indivíduos serão X_1, X_2, \dots, X_n e você vai estimar θ usando $\hat{\theta} = (X_1 + \dots + X_n)/n$, a proporção amostral. Se você assumir que as respostas são variáveis aleatórias i.i.d., determine o tamanho n da amostra necessário para que o erro de estimação $|\hat{\theta} - \theta|$ seja menor que 0.02 com probabilidade 0.99. Para isto, assumo que você sabe que seu candidato está estacionado entre 15% e 35% dos eleitores (baseado em outras pesquisas mais antigas).

-
2. No problema acima, determine um intervalo da forma $I = (\hat{\theta} - c, \hat{\theta} + c)$ tal que a probabilidade $\mathbb{P}(\hat{\theta} - c \leq \theta \leq \hat{\theta} + c)$ seja aproximadamente igual ou maior que 0.95.

-
3. A researcher wants her sample mean to be twice as accurate; how much does she have to increase her sample size by?
-

- 84 4. An article in the New England Journal of Medicine reported that among adults living in the United States, the average level of albumin in cerebrospinal fluid is 29.5 mg/dl, with a standard deviation of 9.25 mg/dl. We are going to select a sample of size 20 from this population.

- How does the variability of our sample mean compare with the variability of albumin levels in the population?
 - What is the probability that our sample mean will be greater than 33 mg/dl?
 - What is the probability that our sample mean will lie between 29 and 31 mg/dl?
 - What two values will contain the middle 50
-

5. The unemployment rate θ is the proportion of people actively looking for jobs and not finding them. Assume that it is known for sure that this rate is some number between 0.03 and 0.08 (or between 3% and 8%). Find the sample size need to estimate this rate θ in such a way that the estimation error is below 0.005 with probability 0.95. That is, we want an estimate $\hat{\theta}$ such that $|\hat{\theta} - \theta| < 0.005$ with probability 0.95.
-

6. According to an article in the American Journal of Public Health, the distribution of birth weights in a certain population is approximately normal with mean 3500 grams and standard deviation 430 grams.

- What is the probability that a newborn's weight will be less than 3200 grams?
 - Suppose we take a sample of 9 newborns. What is the probability that their average weight will be less than 3200 grams?
 - In the aforementioned sample of 9 newborns, how many newborns would you expect to weigh under 3200 grams?
 - What is the probability that our sample of 9 newborns will contain exactly 3 newborns who weigh less than 3200 grams?
 - Suppose we take 5 samples of 9 newborns. What is the probability that at least one of the sample averages will be less than 3200 grams?
 - How large must our sample be in order to ensure a 95% probability that the sample mean will be within 50 grams of the population mean?
-

7. In a 2006 study published in The New England Journal of Medicine, 78 pairs of patients with Parkinson's disease were randomly assigned to receive *treatment* (which consisted of deep-brain stimulation of a region of the brain affected by the disease) or *control* (which consisted of taking a prescription drug). The pairs were composed by individuals similar with respect to several risk factors such as sex, age, occupation, etc. This ensured that, within each pair, we could considered the individuals more or less coming from the same population except by the possible effect of the treatment.

The researchers found that in 50 of 78 pairs, the patients who received deep-brain stimulation⁸⁵ had improved more than their partner in the control group. We are interested in conducting a hypothesis test of these findings.

For each pair, define the random variable

$$X_i = \begin{cases} 1 & \text{if treatment improves more} \\ 0 & \text{if control improves more} \end{cases}$$

The key rationale is: IF INDEED THE TREATMENT HAS NO EFFECT AT ALL, the probability that the treatment individual is 1/2. Let us call this hypothesis or model the *null hypothesis*, represented by H_0 .

- Conduct a z-test of the null hypothesis that deep-brain stimulation has no effect on the disease by calculating the probability that you can observe something as large as 50 in 78 successes when indeed the “coin” has probability 1/2. That is, use the TCL to calculate approximately

$$\mathbb{P}(X_1 + \dots + X_n \geq 50 \mid H_0 \text{ is true})$$

- Construct a 95% confidence level for the proportion of patients who would do better on deep-brain stimulation than control (see the slides).

8. An irregularly shaped object of unknown area A is located in the unit square $0 \leq x \leq 1$ and $0 \leq y \leq 1$. Consider a random point distributed uniformly over the square. Let $Z = 1$ if the point lies inside the object and $Z = 0$ otherwise. Show that $E[Z] = A$. How could A be estimated from a sequence of n independent points uniformly distributed on the square?

Hint 1: Imagine this is actually a coin tossing experiment with unknown probability of getting Head, that is, the coin land on H if the point is inside the object and on T otherwise. How will you estimate the probability of getting H ?

Hint 2: Solution in <http://bit.ly/1T206Rf>

9. Suppose that a basketball player can score on a particular shot with probability $p = 0.3$. Use the central limit theorem to find the approximate distribution of S , the number of successes out of 25 independent shots. Find the approximate probabilities that S is less than or equal to 5, 7, 9, and 11 and compare these to the exact probabilities.

Hint 1 : Let X_1, X_2, \dots, X_{25} be the indicator random variables of the 25 shots, that is, $X_i = 1$ if the player scores on the i -th shot and $X_i = 0$, otherwise.

Hint 2: Solution in <http://bit.ly/1T206Rf>

8610. The amount of mineral water consumed by a person per day on the job is normally distributed with mean 19 ounces and standard deviation 5 ounces. A company supplies its employees with 2000 ounces of mineral water daily. The company has 100 employees.

- Find the probability that the mineral water supplied by the company will not satisfy the water demanded by its employees.
- Find the probability that in the next 4 days the company will not satisfy the water demanded by its employees on at least 1 of these 4 days. Assume that the amount of mineral water consumed by the employees of the company is independent from day to day.
- Find the probability that during the next year (365 days) the company will not satisfy the water demanded by its employees on more than 15 days.

11. Supply responses true or false with an explanation to each of the following:

- The probability that the average of 20 values will be within 0.4 standard deviations of the population mean exceeds the probability that the average of 40 values will be within 0.4 standard deviations of the population mean.
- $\mathbb{P}(\bar{X} \geq 4)$ is larger than $\mathbb{P}(X \geq 4)$ if $X \sim N(8, \sigma)$ and \bar{X} is the sample mean of $n > 1$ instances of X .
- If \bar{X} is the average of n values sampled from a normal distribution with mean μ and if c is any positive number, then $\mathbb{P}(\mu - c \leq \bar{X} \leq \mu + c)$ decreases as n gets large.

Capítulo 9

Regressão Linear

Vamos trabalhar com o modelo de regressão linear supondo $p - 1$ atributos e n observações. A matriz de desenho \mathbf{X} tem dimensão $n \times p$ e a sua primeira coluna é composta pelo vetor n -dim de 1's representado por $\mathbf{1} = (1, \dots, 1)'$. Vamos representar a matriz de desenho \mathbf{X} ora por meio de suas colunas, ora por meio de suas linhas.

$$\begin{aligned}\mathbf{Y} &= \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \\ &= \left(\mathbf{1}, \mathbf{x}^{(1)}, \dots, \mathbf{x}^{(p-1)} \right) \boldsymbol{\beta} + \boldsymbol{\varepsilon} \\ &= \begin{bmatrix} \mathbf{x}'_1 \\ \mathbf{x}'_2 \\ \vdots \\ \mathbf{x}'_n \end{bmatrix} \boldsymbol{\beta} + \boldsymbol{\varepsilon}\end{aligned}$$

A j -ésima coluna $\mathbf{x}^{(j)}$ de dimensão $n \times 1$ representa o conjunto de todos os valores do atributo j medidos na amostra. A i -ésima linha \mathbf{x}'_i , de dimensão $p \times 1$ representa a observação ou instância da amostra. Note que \mathbf{x}_i é um vetor coluna. A i -ésima linha da matriz de desenho \mathbf{X} é escrita como \mathbf{x}'_i .

Seja $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ a matriz de projeção ortogonal no espaço $\mathcal{C}(\mathbf{X})$ das combinações lineares das colunas de \mathbf{X} . O vetor resposta \mathbf{Y} pode ser decomposto em

$$\mathbf{Y} = \mathbf{H}\mathbf{Y} + (\mathbf{I} - \mathbf{H})\mathbf{Y} = \hat{\mathbf{Y}} + \mathbf{r}$$

onde $\hat{\mathbf{Y}}$ é o vetor de valores preditos (ou ajustados) pelo modelo para a variável resposta e \mathbf{r} é o vetor de resíduos.

1. O arquivo aptos.txt possui dados de apartamentos vendidos no bairro Sion em BH em 2011 obtidos com um coletor em páginas web de uma imobiliária. A primeira coluna é um identificador do anúncio. Use o script R abaixo para fazer uma regressão linear simples de preço versus area. As seguir, faça uma regressão múltipla de preço versus todas as covariáveis.

Solução:

```

setwd("u:/ESTATICS") # set the working directory
# Lendo os dados como um dataframe
aptos = read.table("aptosBH.txt", header = T)
attach(aptos)          # attach o dataframe
aptos[1:5,]            # visualizando as 5 1as linhas
par(mfrow=c(2,2))      # tela grafica dividida em 2 x 2
plot(area, preco)      # scatter plot de (area_i, preco_i)
plot(quartos, preco)
plot(suites, preco)
plot(vagas, preco)

# Para fazer uma regressao em R, use o comando lm (linear model)
# Mas vamos antes obter a regressao fazendo as operacoes matriciais
# que vimos em sala

# Ajustando um modelo de regressao linear SIMPLES apenas com area
x = as.matrix(cbind(1, aptos[,2])) # matriz de desenho n x 2
b.simples = (solve(t(x) %*% x)) %*% (t(x) %*% preco)
b.simples

# Ajustando um modelo de regressao linear MULTIPLA com 4 covariaveis
x = as.matrix(cbind(1, aptos[,2:5])) # matriz de desenho n x 2
b.all = (solve(t(x) %*% x)) %*% (t(x) %*% preco)
b.all

```

Veja a mudança do valor do coeficiente de área nas duas regressões: o efeito de área em preço depende de quais outras covariáveis estão no modelo. A razão é que as covariáveis são correlacionadas entre si: aptos grandes tendem a ter mais quartos, por exemplo. Parte do efeito de área medido no coeficiente da regressão simples está capturando também o efeito do número de quartos. Quando quartos entra na regressão, o efeito puro de área diminui.

-
2. Regressão no R é feita usando-se o comando `lm` que implementa uma série de algoritmos numéricos eficientes para lidar com matrizes, incluindo a decomposição QR, a principal técnica para obter o vetor estimado de coeficientes. A sintaxe básica é a seguinte: `lm(y ~ x1 + x2 + x3, data=aptos)` onde `y` é a variável que queremos modelar (a variável resposta ou variável dependente) e `x1`, `x2` etc. são as covariáveis do modelo. O argumento `data` fornece o nome do data.frame ou matriz que contém TODOS os dados (`y` e `x`).

```

setwd("u:/ESTATICS") # set the working directory

```

```
# Lendo os dados como um dataframe
aptos = read.table("aptosBH.txt", header = T)
lm(preco ~ area+quartos+suites+vagas, data=aptos)$coef
lm(preco ~ area, data=aptos)$coef
```

A saída de `lm` é uma lista (um objeto tipo `list`) que pertence à classe `lm`. A lista possui muitas informações para a análise dos dados e sobre as quais vamos aprender ao longo da disciplina. Podemos extrair informação da lista de diversas formas. Por exemplo:

```
aptos = read.table("aptosBH.txt", header = T)
reg.all = lm(preco ~ area+quartos+suites+vagas, data=aptos) # guardo a lista
class(reg.all) # classe do objeto reg.all
names(reg.all) # nomes dos elementos da lista. Eles sao vetores, matrizes, strings, etc.
reg.all$coef # extraindo o elemento da lista de nome coefficients
summary(reg.all) # saida padronizada de regressao quando o objeto e' da classe lm
reg.all$fitted # vetor dos precos preditos pelo modelo de regressao linear
reg.all$res # vetor dos residuos = Y - Yhat = preco observado - preco predito
```

3. Vamos começar simulando um modelo de regressão linear com UM atributo apenas e estimando o vetor de coeficientes $\beta = (\beta_0, \beta_1)'$. A seguir, vamos verificar que o comportamento estatístico do estimador $\hat{\beta}$ está de acordo com o comportamento estocástico deduzido teoricamente. Como este primeiro exercício envolve apenas um atributo, será simples visualizar os vários resultados.

Vamos fixar um modelo de regressão em que CONHECEMOS o vetor de coeficientes

$$\beta = (\beta_0, \beta_1)' = (1, 1.5)'.$$

Vamos usar $n = 25$ observações com um único atributo, x_1 . A matriz \mathbf{X} é de dimensão $n \times p = 25 \times 2$. A linha i da matriz \mathbf{X} é igual a $\mathbf{x}'_i = (1, x_{i1})$. O modelo será:

$$Y_i = \mathbf{x}'_i \beta + \varepsilon_i = \beta_0 + \beta_1 x_{i1} + \varepsilon_i = 1 + 1.5 x_{i1} + \varepsilon_i$$

Os erros ε_i serão i.i.d. $N(0, \sigma^2 = 25)$.

Rode o script abaixo no R:

```
set.seed(0) # fixando a semente do gerador aleatorio
x1 = 1:24 # coluna com 1o atributo
x1 # visualizando x1
beta = c(4, 1.5) # vetor beta
X = cbind(1, x1) # matriz de desenho X
mu = X %*% beta # vetor com E(Y)= X*beta
sigma = 5
epsilon = rnorm(24, 0, sigma) # vetor epsilon de "erros"
```

```

y = mu + epsilon                                # resposta y = X*beta + erros N(0,1)

plot(x1,y)                                       # scatterplot dos dados
abline(4,1.5, col="blue")                      # reta "verdadeira" beta_0 + beta_1 * x

plot(x1,y, xlim=c(0, max(x1)), ylim=c(0, max(y))) # redesenhando para ver o inter
abline(4,1.5, col="blue")                      # reta "verdadeira" beta_0 + beta_1 * x
cor(y, x1)                                     # correlacao entre y e x1

sim1 = lm(y ~ x1 )                             # sim1 e' objeto da classe lm com resultado
is.list(sim1)                                  # sim1 e' uma lista
names(sim1)                                    # nomes dos objetos que compoem a lista sim1
summary( sim1 )                               # funcao summary em sim1: info sobre ajuste

plot(x1,y, xlim=c(0, max(x1)), ylim=c(0, max(y))) # redesenhando para ver o inter
abline(4,1.5, col="blue")                      # reta "verdadeira" beta_0 + beta_1 * x
abline(sim1$coef, col="red")                   # reta ajustada beta_0HAT + beta_1HAT * x

```

O último gráfico mostra uma diferença FUNDAMENTAL entre $\hat{\beta}' = (5.7981, 1.3737)$ e $\beta' = (4, 1.5)$: eles não são iguais. $\hat{\beta}'$ é um vetor que está usando os 24 dados para ESTIMAR o valor verdadeiro de β' . Na prática, não saberemos o valor de β' e é por isto que estamos usando os dados da amostra para inferir sobre seu valor. Olhando a saída de `summary`, veja que

$$\hat{\beta}' = (5.7981, 1.3737) \neq (4, 1.5) = \beta'.$$

O erro de estimação NESTA AMOSTRA PARTICULAR é igual a

$$\hat{\beta}' - \beta' = (5.7981, 1.3737) - (4, 1.5) = (1.7981 - 0.1263)$$

Vamos gerar uma segunda amostra de 24 valores y com os mesmos x . Apenas os “erros” ϵ_{i} vão variar. Vamos estimar β novamente com esta segunda amostra.

```

set.seed(1)
epsilon2 = rnorm(24, 0, sigma)                 # NOVO vetor epsilon de "erros"
y2 = mu + epsilon2                             # NOVA resposta y = X*beta + NOVOS erros
sim2 = lm(y2 ~ x1 )                           # sim2 e' o ajuste dos NOVOS dados.
summary(sim2)

```

Note que a reta estimada com esta segunda amostra é $4.7559 + 1.4995x$, diferente da reta original e também diferente da reta estimada com a primeira amostra. Vamos visualizar estas diferentes retas e amostras.

```

plot(x1,y2)                                     # scatterplot dos NOVOS dados
abline(4, 1.5, col="blue")                     # reta "verdadeira" beta_0 + beta_1 * x

```

```

abline(sim2$coef, col="red")          # NOVA reta ajustada

# plotando os dois conjuntos de pontos
par(mfrow=c(1,2))
plot(x1,y2, main="Dados novos")      # scatterplot dos NOVOS dados
abline(4, 1.5, col="blue")           # reta "verdadeira"  $\beta_0 + \beta_1 * x$ 
abline(sim2$coef, col="red")         # NOVA reta ajustada

plot(x1,y,main="Dados antigos")      # scatterplot dos dados ANTIGOS
abline(4, 1.5, col="blue")           # reta "verdadeira" E' A MESMA
abline(sim1$coef, col="red")         # reta ajustada com os dados ANTIGOS

# os dois conjuntos de dados num unico plot
par(mfrow=c(1,1))
plot(x1,y2, col="black")             # scatterplot dos NOVOS dados
points(x1, y, col="red")             # dados antigos
abline(4, 1.5, col="blue")           # reta "verdadeira"  $\beta_0 + \beta_1 * x$ 
abline(sim2$coef, col="black")       # NOVA reta ajustada
abline(sim1$coef, col="red")         # reta ajustada com os dados ANTIGOS

```

Agora temos TRÊS retas distintas: a reta verdadeira que queremos estimar $\beta_0 + \beta_1 x = 4 + 1.5x$, a reta estimada com a primeira amostra de 24 dados $\hat{\beta}_0^{(1)} + \hat{\beta}_1^{(1)} x = 5.798 + 1.374x$ e a reta estimada com a segunda amostra de 24 dados $\hat{\beta}_0^{(2)} + \hat{\beta}_1^{(2)} x = 4.756 + 1.500x$. O erro de estimação com esta segunda amostra foi igual a

$$\hat{\beta}' - \beta' = (4.7559, 1.4995) - (4, 1.5) = (0.7559, -0.0005)$$

diferente do erro de estimação com a primeira amostra, que foi igual a $(1.7981 - 0.1263)$.

Como o vetor estimado $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ é uma função dos dados aleatórios \mathbf{Y} , ele próprio é um vetor aleatório. Para cada amostra, gerada sob o mesmo modelo probabilístico, temos diferentes valores para $\hat{\beta}$. O erro de estimação $\hat{\beta}' - \beta' = \hat{\beta}' - (4, 1.5)$ também é uma quantidade aleatória. Algumas vezes, este vetor será pequeno, algumas vezes será grande. Queremos saber o que seria um erro grande na estimação do vetor β e com que frequência ele vai ocorrer. Em suma, queremos conhecer a *distribuição de probabilidade* do VETOR erro de estimação

$$\hat{\beta}' - \beta' = (\hat{\beta}_0 - 4, \hat{\beta}_1 - 1.5)$$

Veja que a única parte aleatório nesta expressão é o estimador $\hat{\beta}$ de mínimos quadrados já que β é um vetor fixo. Para estudar o comportamento probabilístico do estimador $\hat{\beta}$ (ou do erro de estimação), podemos usar simulação Monte Carlo. Vamos gerar centenas de vetores \mathbf{Y} , sempre nas mesmas condições, e verificar como o estimador $\hat{\beta}$ e o erro de estimação

$$\hat{\beta}' - \beta' = (\hat{\beta}_0 - \hat{\beta}_1) - (4, 1.5)$$

se comportam estatisticamente.

Queremos calcular o R^2 em cada simulação. Uma maneira simples de extrair seu valor a partir do objeto retornado pelo comando `lm` é acessá-lo a partir do objeto `summary`. Digite `str(summary(sim1))` para ver o que pode ser extraído. No caso do R^2 basta usar `summary(sim1)$r.squared`. Outra estatística que vamos precisar é uma estimativa de σ , explicada mais abaixo e obtida com `summary(sim1)$sigma`. Com isto, vamos às simulações:

```
set.seed(1)
nsim = 1000                                # numero de simulacoes
betasim = matrix(0, ncol=nsim, nrow=2) # matriz para guardar as nsim estimativas
R2 = rep(0, nsim)
S2 = rep(0, nsim)
for(j in 1:nsim){
  y = mu + rnorm(24, 0, sigma)              # gera novo vetor y
  simj = lm(y ~ x1)
  betasim[, j] = simj$coef                  # estima beta e salva
  R2[j] = summary(simj)$r.squared
  S[j] = summary(simj)$sigma
}

# visualizando os resultados
par(mfrow=c(2,2))                          # particiona a janela grafica em 2 x 2
hist(betasim[1,], prob=T, main="beta0")    # histograma dos 1000 interceptos estimados
abline(v=beta[1], lwd=2, col="blue")       # verdadeiro beta_0

hist(betasim[2,], prob=T, main="beta1")    # histograma dos 1000 interceptos estimados
abline(v=beta[2], lwd=2, col="blue")       # verdadeiro beta_1

plot(t(betasim), xlab="beta0", ylab="beta1") # correlacao entre beta_0_hat e beta_1_hat
abline(v=beta[1], h=beta[2])               # valores verdadeiros beta_0 e beta_1

hist(R2, main="R2")
```

Alguns comentários *muito importantes* : observe que o valor verdadeiro dos parâmetros nunca mudou ao longo das simulações. Sempre tivemos $\beta_0 = 4$ e $\beta_1 = 1.5$. Os 24 valores do atributo x_1 também não mudaram. Apenas \mathbf{Y} variou e isto ocorreu por causa dos erros ϵ_i que não possuem nenhuma conexão com X_1 ou com o β_0 e β_1 . A análise que você vai fazer na prática é uma dessas 1000 simulações. Todas elas foram geradas da mesma forma e poderiam legitimamente ser qualquer uma delas, a única instância específica de dados que você terá em mãos na prática de análise de dados. Você então observar nos gráficos o que poderia acontecer com sua análise. Até onde você pode errar? E com que frequência erros grandes podem ocorrer?

Veja o histograma dos 1000 valores estimados da inclinação, $\beta_1^{(1)}, \dots, \beta_1^{(1000)}$. O valor verdadeiro usado para gerar os dados foi $\beta_1 = 1.5$. Os valores estimados estão centrados

aproximadamente em torno do valor verdadeiro $\beta_1 = 1.5$. Além disso, eles variaram entre 1.0 e 2.0 causando então um erro máximo de aproximadamente 0.5. Os valores acima de 1.8 ou abaixo 1.2 (e portanto, com um erro de estimação maior que 0.3) aconteceram com baixa frequência. De fato, a proporção de vezes em que isto ocorreu nas 1000 simulações foi igual a `sum(abs(betasim[2,] - beta[2]) > 0.3)/nsim`, que resultou em 0.034, ou apenas 3.4%. Assim, podemos concluir que ao estimar $\beta_1 = 1.5$ com o estimador de mínimos quadrados neste problema podemos ter uma boa confiança de que erraremos o seu valor verdadeiro por máximo 0.3 (isto vai ocorrer aproximadamente 98.6% das vezes). Além disso, os valores estimados estarão oscilando em torno do valor verdadeiro, ora um pouco mais, ora um pouco menos que β_1 .

Chegamos a uma conclusão semelhante olhando para o histograma de $\hat{\beta}_0$, que parece estar centrado em torno do valor verdadeiro $\beta_0 = 4$ e com um erro que, na maioria das vezes, não ultrapassa 4 (tivemos `sum(abs(betasim[1,] - beta[1]) > 4)/nsim` igual a 0.052 ou 5.2%).

Outro aspecto claro nos histogramas é que as distribuições de probabilidade de $\hat{\beta}_0$ e $\hat{\beta}_1$ se parecem com distribuições normais. De fato, vamos repetir os últimos gráficos ajustando uma densidade normal a cada um dos histogramas usando a média aritmética e o desvio-padrão amostra das 1000 simulações de cada um dos estimadores. Temos `mean(betasim[1,])` igual a 3.97 e `sd(betasim[1,])` igual a 2.08, enquanto para $\hat{\beta}_1$ temos `mean(betasim[2,])` igual a 1.50 e `sd(betasim[2,])` igual a 0.15. Usando estes valores com o comando `curve`:

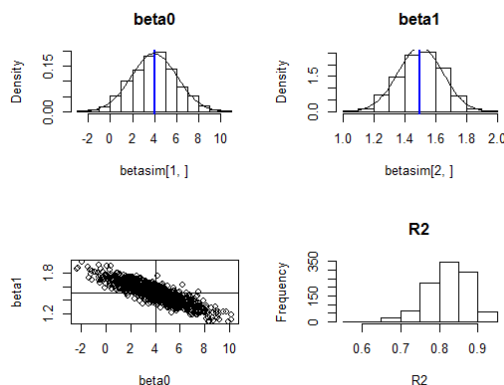
```
# visualizando os resultados
par(mfrow=c(2,2))                                # particiona a janela grafica em 2 x 2
hist(betasim[1,], prob=T, main="beta0")           # histograma dos 1000 interceptos estimados beta_0
abline(v=beta[1], lwd=2, col="blue")              # verdadeiro beta_0
# ajustando uma densidade normal aos dados de beta_0_hat
curve(dnorm(x, mean=mean(betasim[1,]), sd=sd(betasim[1,])), add=TRUE)

hist(betasim[2,], prob=T, main="beta1")           # histograma dos 1000 interceptos estimados beta_1
abline(v=beta[2], lwd=2, col="blue")              # verdadeiro beta_1
# ajustando uma densidade normal aos dados de beta_1_hat
curve(dnorm(x, mean=mean(betasim[2,]), sd=sd(betasim[2,])), add=TRUE)

plot(t(betasim), xlab="beta0", ylab="beta1")      # correlacao entre beta_0_hat e beta_1_hat
abline(v=beta[1], h=beta[2])                     # valores verdadeiros beta_0 e beta_1

hist(R2, main="R2")
```

Não somente a distribuição marginal de cada componente de $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1)$ parece seguir uma distribuição gaussiana mas, observando que o plot dos 1000 pares de pontos $(\hat{\beta}_0, \hat{\beta}_1)$ tem a forma de uma elipse, parece que a distribuição conjunta dos componentes do vetor segue uma distribuição gaussiana bivariada. A partir deste plot vemos que a correlação entre $\hat{\beta}_0$ e $\hat{\beta}_1$ é negativa. Quando a inclinação estimada $\hat{\beta}_1$ fica muito acima de sua média



(que é aproximadamente 1.5), o intercepto estimado $\hat{\beta}_0$ tende a ficar abaixo de sua média (que é aproximadamente 4).

Mais uma observação referente ao último gráfico: a estatística R^2 também é uma variável aleatória! Ao longo das 1000 simulações, R^2 teve um valor médio igual a 0.8234 e 50% de seus valores estão entre 0.79 e 0.86. Apenas 2.3% dos 1000 valores calculados foram menores que 0.7. Uma das amostras chegou a gerar um R^2 igual a 0.60, o mínimo nas 1000 simulações.

O drama do analista de dados: O estudo de simulação que fizemos mostra claramente as propriedades estatísticas do estimador de mínimos quadrados $\hat{\beta}$. Obtivemos o comportamento estatístico do erro de estimação descobrindo o valor esperado de cada componente de $\hat{\beta}$, incluindo um valor máximo que, com alta probabilidade, este erro pode atingir. Acontece que, na prática da análise de dados, este estudo é impossível de ser realizado. Para estudar o comportamento estatístico de $\hat{\beta}$, tivemos de gerar os dados \mathbf{Y} uma grande quantidade de vezes. Para esta geração, precisamos conhecer o verdadeiro valor do parâmetro β . Entretanto, o objetivo da estimação na prática é inferir um valor aproximado para β , supostamente desconhecido. Se conhecermos o valor verdadeiro de β , não há necessidade de estimá-lo, muito menos de conhecer as propriedades estatísticas do estimador $\hat{\beta}$. Parece que estamos num beco sem saída: para conhecer as propriedades estatísticas de $\hat{\beta}$ por simulação precisamos conhecer o verdadeiro valor de β . Mas, na prática, não saberemos este valor já que o objetivo de calcular $\hat{\beta}$ é exatamente obter um valor aproximado para o vetor β .

Todo este suspense é para fornecer um resultado surpreendente. Para conhecer como o estimador $\hat{\beta}$ vai variar considerando as diferentes amostras que poderiam ser geradas pelo modelo não vamos precisar simular o modelo milhares de vezes. Usando apenas a única amostra de dados que temos em mãos e o cálculo de probabilidades somos capazes de obter todo o comportamento estatístico de $\hat{\beta}$! Com uma única amostra de dados conseguimos descobrir todos os resultados obtidos nas simulações acima.

De fato, já obtivemos todo o comportamento estatístico de $\hat{\beta}$ nas notas de aula. Nós já aprendemos que, como $\hat{\beta}$ é uma matriz de constantes ($\mathbf{A} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}$) multiplicando o

vetor gaussiano multivariado \mathbf{Y} , podemos deduzir que $\hat{\boldsymbol{\beta}}$ é normal multivariado:

$$\hat{\boldsymbol{\beta}} \sim N_2(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}) .$$

Portanto, olhando para o componente 1 deste vetor, temos

$$\hat{\beta}_0 \sim N(\beta_0, \sigma^2(\mathbf{X}'\mathbf{X})_{11}^{-1})$$

enquanto que, para o segundo componente,

$$\hat{\beta}_1 \sim N(\beta_1, \sigma^2(\mathbf{X}'\mathbf{X})_{22}^{-1})$$

A matriz $(\mathbf{X}'\mathbf{X})^{-1}$ é facilmente obtida em R:

```
> solve(t(X) %*% X)
               x1
0.17753623 -0.0108695652
x1 -0.01086957  0.0008695652
```

Assim, sem nenhuma simulação nós podemos conhecer de forma quase completa todo o comportamento estatístico de $\hat{\boldsymbol{\beta}}$:

$$\hat{\beta}_0 \sim N(\beta_0, \sigma^2 0.1775)$$

enquanto que, para o segundo componente,

$$\hat{\beta}_1 \sim N(\beta_1, \sigma^2 0.0009)$$

Fica faltando apenas conhecer σ^2 , a variância dos erros ϵ_i no modelo de regressão linear $Y_i = \mathbf{x}_i'\boldsymbol{\beta} + \epsilon_i$. É claro que sabemos que $\sigma = 5$ pois nós mesmos geramos os dados. Mas também é claro que, numa análise de dados reais, temos apenas os dados \mathbf{Y} e a matriz \mathbf{X} . Não conhecemos o modelo que gerou os dados e portanto não conhecemos σ^2 .

Acontece que, embora não possamos conhecer σ^2 com exatidão, podemos estimá-lo com razoável precisão. Para isto usamos os resíduos $r_i = y_i - \hat{y}_i$. Pode-se mostrar que a variável aleatória

$$SSE = \mathbf{r}'\mathbf{r} = \sum_{i=1}^n r_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

possui uma distribuição qui-quadrado com $n - p = n - 2$ graus de liberdade multiplicada por σ^2 . Portanto $\mathbb{E}(SSE) = \sigma^2(n - p)$ e, como consequência, $S^2 = \mathbb{E}(SSE/(n - p)) = \sigma^2$. Isto é, a soma do quadrado dos resíduos dividida por $n - 2$ (quase a sua média, portanto) tem o valor esperado igual a σ^2 . Algumas vezes, um pouco mais que σ^2 , algumas vezes, um pouco menos.

Vamos ver como foi a estimativa S^2 do parâmetro σ^2 em cada uma das 1000 simulações. Com o comando `S[j] = summary(simj)$sigma`, nós guardamos os valores de $\sqrt{S^2}$ e portanto de uma estimativa de σ . Vamos elevá-lo ao quadrado e transformá-lo para verificar duas coisas:

- O ajuste da distribuição qui-quadrado com $df = n - 2$ graus de liberdade aos valores simulados de $SSE/\sigma^2 = SSE/25$. Este é um resultado teórico que estamos verificando empiricamente.
- Como $S = \sqrt{S^2} = \sqrt{SSE/(24 - 2)}$ comporta-se como estimador de σ^2 ao longo das 1000 simulações.

```
par(mfrow=c(1,2))
S2 = S^2 # obtendo estimativa de sigma^2 em cada simulacao
SSE = (24-2)*S2 # obtendo a soma dos residuos ao quadrado em cada simulacao
hist(SSE/sigma^2, prob=T)
curve(dchisq(x, df=24-2), add=T) # ajuste de qui-quadrado
hist(S, prob=T)
sum(S > 7 | S < 3)/1000
```

A maioria dos valores estimados de σ estão entre 3 e 7. Já que $\sigma = 5$, o último comando acima produz 0.008, ou 0.8% das vezes, para a proporção das simulações em que S teve um erro de estimacão ao $S - \sigma$ maior que 2. Na prática, apenas um desses valores será usado, aquele associado com o conjunto particular de dados que está em sua mãos. Por exemplo, se tivermos em mãos apenas o primeiro conjunto de dados (dentre os 1000 simulados), teremos a estimativa $S[1] = 4.937436$, um valor bem próximo de σ . Nós fomos felizes neste primeiro conjunto de dados mas que valores mais altos ou mais baixo poderiam ter sido obtidos como estimativa de S , embora estas estimativas dificilmente ultrapassem 7 ou sejam menores que 3.

O ponto relevante é que, com esta aproximação de σ baseada em UMA ÚNICA amostra de dados, podemos agora conhecer o comportamento estatístico do estimador $\hat{\beta}$ COMO SE FIZÉSSSEMOS centenas de simulações. De fato, baseado na teoria que já deduzimos, como sabemos que

$$\hat{\beta} \sim N_2(\beta, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}) ,$$

podemos concluir que

$$\hat{\beta} \approx N_2(\beta, S^2(\mathbf{X}'\mathbf{X})^{-1}) ,$$

e que portanto, como $S = 4.937436$ na primeira amostra (supondo que é a única amostra de dados que temos), concluímos que

$$\hat{\beta}_0 \approx N(\beta_0, (4.937)^2 \cdot 0.1775)$$

enquanto que, para o segundo componente,

$$\hat{\beta}_1 \approx N(\beta_1, (4.937)^2 \cdot 0.0009)$$

O erro de estimacão: Com isto, podemos ter uma boa idéia do erro máximo que podemos estar cometendo ao estimar β pelo método de mínimos quadrados. De fato, como 95% da área de uma gaussiana fica entre dois desvios-padrão de sua esperança, temos para β_1 :

$$\mathbb{P}(|\hat{\beta}_1 - \beta_1| < 2\sqrt{(4.937)^2 \cdot 0.0009}) \approx 0.95$$

Isto é,

$$\mathbb{P}(|\hat{\beta}_1 - \beta_1| < 0.29622) \approx 0.95$$

Assim, com alta probabilidade, a diferença entre a estimativa $\hat{\beta}_1$ (uma variável aleatória calculada a partir dos dados) e β_1 (um valor fixo mas desconhecido) não deve ultrapassar 0.30. Manipulação simples do operador valor absoluto permite escrever o evento $|\hat{\beta}_1 - \beta_1| < 0.30$ de outra forma equivalente:

$$|\hat{\beta}_1 - \beta_1| < 0.30 \Leftrightarrow -0.30 < \hat{\beta}_1 - \beta_1 < 0.30 \Leftrightarrow \hat{\beta}_1 - 0.30 < \beta_1 < \hat{\beta}_1 + 0.30$$

Retornando ao cálculo de probabilidade, temos então

$$\mathbb{P}(\hat{\beta}_1 - 0.30 < \beta_1 < \hat{\beta}_1 + 0.30) \approx 0.95$$

Ou seja, com probabilidade 95% o intervalo (aleatório) $(\hat{\beta}_1 - 0.30, \hat{\beta}_1 + 0.30)$ vai cobrir o verdadeiro valor do parâmetro aproximadamente 95% das vezes que o procedimento de estimação de mínimos quadrados for adotado.

Lição a levar para casa: O ponto mais importante de todo este exercício é que: não apenas obtemos uma estimativa do vetor β mas conseguimos também ter uma boa aproximação para os outros possíveis valores que poderíamos ter se a amostra fosse um pouco diferente (mas gerada sob o mesmo modelo). Com isto, temos uma boa idéia do tamanho máximo do erro que podemos estar cometendo ao estimar β_i : dificilmente vamos ultrapassar $2S\sqrt{(\mathbf{X}'\mathbf{X})_{ii}^{-1}}$.

Outra maneira de expressar este resultado é apresentar o chamado *intervalo de confiança* de 95%:

$$(\hat{\beta}_i - 2S\sqrt{(\mathbf{X}'\mathbf{X})_{ii}^{-1}}, \hat{\beta}_i + 2S\sqrt{(\mathbf{X}'\mathbf{X})_{ii}^{-1}})$$

Este intervalo aleatório cobre o verdadeiro (e desconhecido) valor de β_i 95% das vezes, aproximadamente.

4. Repita o exercício anterior simulando um modelo de regressão linear com DOIS atributos. Este exercício tem o objetivo de forçá-lo a refletir sobre o significado da expressão “ $\hat{\beta}$ é um vetor aleatório”. Seu entendimento é crucial na teoria da aprendizagem estatística.

Vamos simular um modelo de regressão linear com dois atributos usando o R e estimar o vetor de coeficientes β em cada caso. A seguir, vamos verificar que o comportamento estatístico do estimador $\hat{\beta}$ está de acordo com o comportamento estocástico deduzido teoricamente.

Vamos fixar um modelo de regressão em que CONHECEMOS o vetor de coeficientes

$$\beta = (\beta_0, \beta_1, \beta_2)' = (1, 1.5, 0.7)'.$$

Vamos usar $n = 25$ observações com dois atributos. O modelo será:

$$Y_i = \mathbf{x}_i' \beta + \varepsilon_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i = 1 + 1.5x_{i1} + 0.7x_{i2} + \varepsilon_i$$

Os erros ε_i serão i.i.d. $N(0, \sigma^2 = 25)$.

Rode o script abaixo no R:

```

98  set.seed(0)                                # fixando a semente do gerador aleatorio
    x1 = 1:24                                  # coluna com 1o atributo
    x1                                     # visualizando x1
    x2 = round(25*runif(24),1)                # coluna com 2o atributo
    x2                                     # visualizando x2
    beta = c(1, 1.5, 0.7)                    # vetor beta
    mu = cbind(1, x1, x2) %*% beta            # vetor com E(Y)= X*b
    sigma = 5
    y = mu + rnorm(24, 0, sigma)              # resposta y = X*b + erros N(0,1)
    pairs(cbind(y, x1, x2))                  # scatterplots bivariado

```

Nos gráficos de dispersão que resultaram do comando `pairs` você deve ter notado como, visualmente, o atributo `x2` parece ter pouco efeito para explicar a variação de `y` quando comparado com a associação forte e óbvia entre `y` e `x1`. No entanto, você sabe que `x2` tem algum efeito sobre `y` pois seu coeficiente não é exatamente igual a zero. Na verdade, podemos comparar os dois coeficientes pois os dois atributos possuem aproximadamente a mesma escala (variando entre 0 e 24). O efeito de `x1` em `y` é medido pelo seu coeficiente (igual a 1.5) e o de `x2` pelo seu coeficiente (igual a 0.7). Assim, o efeito de `x1` parece ser $1.5/0.7 \approx 2.0$, ou duas vezes maior, que o de `x2`. No entanto, o gráfico de $y \times x_2$ dá a impressão visual de que `x2` tem muito menos efeito em `y`. Na verdade, o que afeta a nossa avaliação do efeito de `x2` neste gráfico é o efeito simultâneo de `x1`. Voltaremos a isto mais abaixo. Ainda nestes gráficos, note como os preditores `x1` e `x2` são muito pouco correlacionados.

```

x = cbind(x1, x2)                            # matriz de desenho (sem a constante 1)
sim1 = lm(y ~ x )                            # sim1 e' objeto da classe lm com results c
is.list(sim1)                                # sim1 e' uma lista
names(sim1)                                  # nomes dos objetos que compoem a lista sim1
summary( sim1 )                              # funcao summary em sim1: retorna info sobre

```

Veremos agora apenas alguns dos itens listados na saída de `summary`. Veja que

$$\hat{\beta}' = (0.46, 1.64, 0.58) \neq (1, 1.5, 0.7) = \beta',$$

considerando as estimativas com duas casas decimais. O erro de estimação NESTA AMOSTRA PARTICULAR é igual a

$$\hat{\beta}' - \beta' = (0.46, 1.64, 0.58) - (1, 1.5, 0.7) = (-0.54, 0.14, 0.12)$$

O valor ALEATÓRIO $\hat{\beta}'$ possui distribuição gaussiana p -variada (3-variada aqui) com

$$\hat{\beta}' \sim N_3(\beta, \sigma^2(\mathbf{X}'\mathbf{X})^{-1})$$

O que isto significa? O resto do exercício procura te dar uma idéia da resposta.

O vetor $\hat{\mathbf{Y}} = \mathbf{X}\hat{\beta}$ com os valores ajustados é obtido manipulando matrizes e vetores: ou diretamente a partir do objeto `sim1`:

```

betahat = sim1$coef          # extraindo o vetor beta-hat.
yhat = sim1$fitted          # Fitted values. Alternativa: yhat0 = cbind(1, x) %*% betahat
plot(yhat, y)               # R2 e' o (coeficiente de correlacao linear)^2 deste grafico
cor(yhat, y)^2              # compare com o valor de Multiple R-squared em summary(sim1)

r = y - yhat                # Vetor de residuos. Alternativa: sim1$res
S2 = sum(r*r)/(length(y) - 3) # estimativa nao-viciada de sigma^2

sqrt(S2)                    # 0 mesmo que ''Residual standard error...'' em summary(sim1)
                             # veja que o verdadeiro valor e' sigma=5

vcov(sim1)                  # extrai a estimativa da matriz de covariancia de
                             # beta-hat = S2 * (X'X)^{-1}

sdhat = sqrt(diag(vcov(sim1))) # sqrt dos elementos da diagonal: Estimativa dos DPs
                                # dos beta-hats.
                                # Mesmos valores que a coluna Std. Error em summary(sim1)

```

Vamos agora simular este processo 1000 vezes, sempre gerando um novo vetor resposta y , ajustando o modelo (com a matriz de desenho \mathbf{X} fixa) e calculando as estimativas. Teremos sempre um vetor $\hat{\beta}$ diferente do verdadeiro valor de β . Vamos avaliar empiricamente o comportamento deste estimador e comparar com o que a teoria diz.

```

nsim = 1000      # numero de simulacoes
betasim = matrix(0, ncol=nsim, nrow=3) # matriz para guardar as nsim estimativas de beta
betasim[,1] = betahat # primeira coluna = estimativa com 1a amostra
for(j in 2:nsim){
  y = mu + rnorm(24, 0, sigma) # gera y
  betasim[, j] = lm(y ~ x)$coef # estima e salva betahat na simulacao j
}

sdbeta = sigma * sqrt(diag(solve(t(cbind(1,x)) %*% cbind(1,x))))

par(mfrow=c(2,2)) # particiona a janela grafica em 2 x 2
hist(betasim[1,], prob=T) # histograma dos 1000 interceptos
abline(v=betasim[1,1], lwd=2, col="blue") # verdadeiro beta_0
aux = seq(min(betasim[1,]), max(betasim[1,]), len=100) # beta_0_hat e' gaussiano?
lines(aux, dnorm(aux, beta[1], sdbeta[1]))

hist(betasim[2,], prob=T); abline(v=beta[2], lwd=2, col="red")
abline(v=betasim[2,1], lwd=2, col="blue")
aux = seq(min(betasim[2,]), max(betasim[2,]), len=100)
lines(aux, dnorm(aux, beta[2], sdbeta[2]))

hist(betasim[3,], prob=T); abline(v=beta[3], lwd=2, col="red")

```

100 `abline(v=betasim[3,1], lwd=2, col="blue")` *CAPÍTULO 9. REGRESSÃO LINEAR*
`aux = seq(min(betasim[3,]), max(betasim[3,]), len=100)`
`lines(aux, dnorm(aux, beta[3], sdbeta[3]))`

`plot(betasim[2,], betasim[3,])`

5. The vector \mathbf{HY} is the orthogonal projection of \mathbf{Y} into the linear subspace of the linear combinations of the p columns of \mathbf{X} . Show that indeed \mathbf{HY} can be written as a linear combination of columns of \mathbf{X} .

Solution: A linear combination of the p columns of \mathbf{X} is a vector written as \mathbf{Xb} where \mathbf{b} is any p -dimensional vector. Using the definition of \mathbf{H} , we have

$$\mathbf{HY} = \mathbf{X} (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{Y} = \mathbf{X}\hat{\boldsymbol{\beta}}$$

where $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{Y}$ is a p -dimensional vector.

6. Numa análise de dados com o modelo de regressão linear comum (isto é, com a primeira coluna da matriz \mathbf{X} sendo a colunas de 1's), checar numericamente que o vetor de resíduos $\mathbf{r} = \mathbf{Y} - \hat{\mathbf{Y}}$ é ortogonal ao vetor projetado $\hat{\mathbf{Y}}$. Faça isto com os dados de apartamentos em BH.

Solution: Obtemos um valor aproximadamente zero, mas não exato, devido a erros de arredondamento numérico:

```
aptos = read.table("aptosBH.txt", header = T)
reg.all = lm(preco ~ area+quartos+suites+vagas, data=aptos)
sum(reg.all$fitted * reg.all$res) # produto interno de residuos x preditos
# resultado eh -0.003036737
```

7. Responda V ou F às questões abaixo:

- O vetor de resíduos $\mathbf{r} = \mathbf{Y} - \hat{\mathbf{Y}}$ é ortogonal ao vetor \mathbf{Y} .
- The orthogonal projection $\hat{\mathbf{Y}}$ is orthogonal to the data \mathbf{Y} .
- The inner product between \mathbf{r} and $\hat{\mathbf{Y}}$ is zero. That is,

$$\langle \mathbf{r}, \hat{\mathbf{Y}} \rangle = \underbrace{\mathbf{r}'}_{(1 \times n)} \underbrace{\hat{\mathbf{Y}}}_{(n \times 1)} = 0$$

- $\hat{\mathbf{Y}}$ pertence ao espaço das combinações lineares das colunas de \mathbf{X} .
- O vetor de resíduos $\mathbf{r} = \mathbf{Y} - \hat{\mathbf{Y}}$ pertence ao espaço das combinações lineares das colunas de \mathbf{X} .

Solução: $\mathbf{F}(\mathbf{r})$ é ortogonal a $\hat{\mathbf{Y}}$, $\mathbf{F}(\hat{\mathbf{Y}})$ is orthogonal to $\mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{r}$, T, T (pois $\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$), $\mathbf{F}(\mathbf{r})$ pertence ao espaço ortogonal ao espaço das combinações lineares das colunas de \mathbf{X} ; o vetor \mathbf{r} é ortogonal a cada coluna da matriz \mathbf{X} .

8. Em um modelo de regressão linear, a variável resposta é o rendimento de uma reação química em duas situações diferentes, 0 e 1. São feitas n_0 e n_1 repetições independentes da reação em cada um dos dois casos gerando os $n_0 + n_1$ valores da resposta Y_{ij} onde $i = 1, \dots, n_j$ e $j = 0, 1$. Suponha $\mathbf{Y} = (Y_{10}, Y_{20}, \dots, Y_{n_0 0}, Y_{11}, \dots, Y_{n_1 1})$ e que \mathbf{X} é a matriz de desenho com a primeira colunas de 1's e a segunda coluna com uma variável indicadora com valores 0 (se estado é 0) e 1 (se estado é 1). Isto é,

$$\mathbf{Y} = \begin{pmatrix} y_{10} \\ y_{20} \\ \vdots \\ y_{n_0 0} \\ y_{11} \\ y_{21} \\ \vdots \\ y_{n_1 1} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 1 & 1 \\ 1 & 1 \\ \vdots & \vdots \\ 1 & 1 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \boldsymbol{\epsilon} = \mathbf{X} \boldsymbol{\beta} + \boldsymbol{\epsilon}$$

Mostre que o estimador de mínimos quadrados é dado por

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \begin{bmatrix} \bar{Y}_0 \\ \bar{Y}_1 - \bar{Y}_0 \end{bmatrix}$$

onde \bar{Y}_j é a média aritmética das n_j observações no estado j .

OBS:

$$\mathbf{A}^{-1} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$$

9. Considere um modelo de regressão linear onde a matriz de desenho \mathbf{X} possui uma única coluna formada pelo atributo j de forma que \mathbf{X} é uma matriz $n \times 1$. Digamos que o atributo j seja o número total de linhas de código de um software e a resposta seja o tempo até a obtenção de uma primeira versão estável do software. Obtemos dados relacionados a n distintos software. Nosso modelo de regressão linear SEM INTERCEPTO é dado por:

$$\begin{aligned}
\mathbf{Y} &= \mathbf{x}^{(j)}\beta_j + \boldsymbol{\varepsilon} \\
\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} &= \begin{bmatrix} x_{1j} \\ x_{2j} \\ \vdots \\ x_{nj} \end{bmatrix} \beta_j + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} \\
&= \begin{bmatrix} x_{1j}\beta_j + \varepsilon_1 \\ x_{2j}\beta_j + \varepsilon_2 \\ \vdots \\ x_{nj}\beta_j + \varepsilon_n \end{bmatrix}
\end{aligned}$$

Note que o vetor-coeficiente neste caso é simplesmente o escalar β_j , um vetor de dimensão 1.

Ao contrário do que fazemos quase sempre por *default*, no modelo acima, nós não estamos usando a coluna de vetor n -dim de 1's representado por $\mathbf{1} = (1, \dots, 1)'$. Isto significa que o modelo assume que a resposta Y está relacionada ao atributo através de uma relação linear que passa pela origem da forma:

$$y \approx x\beta$$

Este modelo simplificado não é apropriado na maioria das situações práticas pois quase sempre podemos esperar um intercepto não-nulo. A utilidade deste modelo simplificado vai ficar clara no exercício 10.

- Seja $\langle \mathbf{x}, \mathbf{y} \rangle = \sum_i x_i y_i$ o produto interno de dois vetores \mathbf{x} e \mathbf{y} . Mostre que o estimador de mínimos quadrados de β_j neste modelo com um único atributo é dados por

$$\begin{aligned}
\beta_j &= \frac{\mathbf{x}^{(j)'} \mathbf{Y}}{\mathbf{x}^{(j)'} \mathbf{x}^{(j)}} \\
&= \frac{\sum_{i=1}^n x_{ij} y_i}{\sum_{i=1}^n x_{ij}^2} \\
&= \frac{\langle \mathbf{x}^{(j)}, \mathbf{y} \rangle}{\langle \mathbf{x}^{(j)}, \mathbf{x}^{(j)} \rangle} \\
&= \frac{\langle \mathbf{x}^{(j)}, \mathbf{y} \rangle}{\|\mathbf{x}^{(j)}\|^2}
\end{aligned}$$

- Suponha que o único atributo no modelo seja a coluna de 1's representada por $\mathbf{1} = (1, \dots, 1)'$. Mostre que o (único) coeficiente neste caso é dado pela média aritmética das observações

$$\beta_0 = \frac{\mathbf{1}' \mathbf{Y}}{\mathbf{1}' \mathbf{1}} = \frac{1}{n} \sum_i y_i = \bar{Y}$$

- Considerando o item anterior, conclua que o modelo no caso em que o único atributo é a coluna de 1's é da forma:

$$\mathbf{Y} = \mathbf{1} \beta_0 + \boldsymbol{\varepsilon}$$

o que significa que Y_1, Y_2, \dots, Y_n são i.i.d. $N(\beta_0, \sigma^2)$.

O modelo estimado por mínimos quadrados produz a seguinte decomposição do vetor \mathbf{Y} :

$$\mathbf{Y} = \bar{Y} \mathbf{1} + (\mathbf{Y} - \bar{Y} \mathbf{1})$$

Neste modelo, $\hat{\mathbf{Y}} = \bar{Y} \mathbf{1}$ e o vetor de resíduos é $\mathbf{r} = \mathbf{Y} - \bar{Y} \mathbf{1}$.

10. Em estudos experimentais, como nos testes AB feitos pelo Google, as colunas da matriz \mathbf{X} podem ser escolhidos de antemão pelo usuário. Um desenho experimental muito usado é o chamado *full factorial design*. Vou considerar um desses desenhos (se estiver interessado, estou tomando um desenho com dois fatores, dois níveis em cada um deles e apenas com duas replicações). Para este desenho, temos o seguinte modelo de regressão linear para uma resposta experimental.

$$\mathbf{Y} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \\ y_7 \\ y_8 \end{bmatrix} = \begin{bmatrix} 1 & -1 & -1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & 1 & 1 \\ 1 & -1 & -1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \varepsilon_6 \\ \varepsilon_7 \\ \varepsilon_8 \end{bmatrix}$$

Não se preocupe com significado exato dos atributos neste momento. Mas, para não ficar muito abstrato, você pode imaginar que, neste problema, a resposta Y é a produção por minuto de um processo químico numa indústria. Os dois primeiros atributos não constantes representam os efeitos individuais de dois fatores influentes na produção. O último atributo representa a interação ou efeito sinérgico entre estes dois fatores.

Ignorando a semântica do modelo e concentrando apenas na sua sintaxe, verifique o seguinte:

- As colunas da matriz de desenho são todas ortogonais entre si. Isto é, $\mathbf{x}^{(j)'} \mathbf{x}^{(k)} = \sum_i x_{ij} x_{ik} = 0$ se $j \neq k$.
- Conclua que a matriz $\mathbf{X}'\mathbf{X}$ é diagonal.
- Conclua que a matriz $(\mathbf{X}'\mathbf{X})^{-1}$ também é diagonal.
- Conclua que o estimador de mínimos quadrados de $\boldsymbol{\beta}$ é dado por

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \end{bmatrix} = \begin{bmatrix} \langle \mathbf{1}, \mathbf{y} \rangle / \|\mathbf{1}\|^2 \\ \langle \mathbf{x}^{(1)}, \mathbf{y} \rangle / \|\mathbf{x}^{(1)}\|^2 \\ \langle \mathbf{x}^{(2)}, \mathbf{y} \rangle / \|\mathbf{x}^{(2)}\|^2 \\ \langle \mathbf{x}^{(3)}, \mathbf{y} \rangle / \|\mathbf{x}^{(3)}\|^2 \end{bmatrix} = \begin{bmatrix} \sum_i y_i / 8 \\ \langle \mathbf{x}^{(1)}, \mathbf{y} \rangle / 8 \\ \langle \mathbf{x}^{(2)}, \mathbf{y} \rangle / 8 \\ \langle \mathbf{x}^{(3)}, \mathbf{y} \rangle / 8 \end{bmatrix}$$

- Conclua que o estimador $\hat{\beta}_j$ do atributo j é igual a aquele que seria obtido caso tivéssemos rodado uma regressão usando APENAS o atributo j .
- Verifique que esta conclusão é geral: caso as colunas de \mathbf{X} sejam ortogonais entre si, o estimador $\hat{\beta}_j$ do atributo j não é afetado pela presença ou ausência dos demais atributos na regressão.
- Verifique também que a matriz de covariância

$$\mathbb{V}(\hat{\beta}) = \sigma^2 (\mathbf{X}'\mathbf{X})^{-1}$$

do estimador $\hat{\beta}$ é uma matriz diagonal. Conclua que as estimativas de diferentes atributos são v.a.'s independentes pois $\text{Corr}(\hat{\beta}_j, \hat{\beta}_k) = 0$ se $j \neq k$.

- Apenas para seu conhecimento, é possível mostrar que a eficiência máxima na estimação dos coeficientes é alcançada quando as colunas em \mathbf{X} são ortogonais entre si. Mais especificamente, é possível mostrar que dada qualquer matriz de desenho \mathbf{X} , tal que $\|\mathbf{x}^{(j)}\|^2 = c_j^2 > 0$, então

$$\mathbb{V}(\hat{\beta}_j) \geq \frac{\sigma^2}{c_j^2}$$

e o mínimo é atingido quando $\mathbf{x}^{(j)'}\mathbf{x}^{(k)} = 0$ para todo par $j \neq k$ (isto é, quando as colunas são ortogonais entre si, o erro esperado de estimação é minimizado). Para mais informações, ver Rao (1973, pag. 236).

11. Suponha que a matriz de desenho possui apenas a coluna $\mathbf{1}$ e seja \mathbf{H}_1 a matriz de projeção no espaço $\mathcal{C}(\mathbf{1})$ dos múltiplos do vetor $\mathbf{1}$. Mostre que esta matriz é dada por

$$\mathbf{H}_1 = \frac{1}{n} \mathbf{1}\mathbf{1}' = \frac{1}{n} \begin{bmatrix} 1 & 1 & \dots & 1 \\ 1 & 1 & \dots & 1 \\ \vdots & \vdots & \dots & \vdots \\ 1 & 1 & \dots & 1 \end{bmatrix}$$

12. Verifique que, qualquer que seja o vetor resposta \mathbf{Y} , ele pode ser decomposto como

$$\mathbf{Y} = \mathbf{H}_1 \mathbf{Y} + (\mathbf{I} - \mathbf{H}_1) \mathbf{Y} = \bar{y}\mathbf{1} + (\mathbf{Y} - \bar{y}\mathbf{1})$$

e que os dois vetores do lado direito da equação são ortogonais.

13. Conclua que $\mathbf{Y} - \bar{y}\mathbf{1}$ pertence ao espaço ortogonal $\mathcal{C}(\mathbf{1})^\perp$ e que o comprimento (ao quadrado) de \mathbf{Y} pode ser decomposto da seguinte maneira

$$\|\mathbf{Y}\|^2 = \sum_{i=1}^n y_i^2 = n\bar{y}^2 + \sum_{i=1}^n (y_i - \bar{y})^2 = \|\bar{y}\mathbf{1}\|^2 + \|\mathbf{Y} - \bar{y}\mathbf{1}\|^2$$

14. Seja $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ onde \mathbf{X} é a matriz de desenho com a primeira coluna sendo o vetor de 1's. Seja \mathbf{H}_1 a matriz do exercício anterior. Mostre que as três matrizes \mathbf{H}_1 , $\mathbf{H} - \mathbf{H}_1$, e $\mathbf{I} - \mathbf{H}$ são matrizes de projeções ortogonais. Isto é, elas são simétricas e idempotentes.

15. Outra decomposição mais relevante é a seguinte:

$$\mathbf{I} = \mathbf{H}_1 + (\mathbf{H} - \mathbf{H}_1) + (\mathbf{I} - \mathbf{H}) .$$

Use esta decomposição matricial para decompor o vetor \mathbf{Y} em três outros vetores ortogonais entre si.

DICA: No produto matricial $\mathbf{AB} = \mathbf{C}$, a coluna j de \mathbf{C} é o resultado de multiplicar a matriz \mathbf{A} pela coluna j de \mathbf{B} .

Solução:

$$\mathbf{Y} = (\mathbf{H}_1 + (\mathbf{H} - \mathbf{H}_1) + (\mathbf{I} - \mathbf{H})) \mathbf{Y} = \bar{Y}\mathbf{1} + (\mathbf{H} - \mathbf{H}_1) \mathbf{Y} + (\mathbf{I} - \mathbf{H}) \mathbf{Y}$$

Basta fazer o produto interno desses vetores do lado direito para ver que são ortogonais entre si. Por exemplo,

$$\langle (\mathbf{H} - \mathbf{H}_1) \mathbf{Y}, (\mathbf{I} - \mathbf{H}) \mathbf{Y} \rangle = 0$$

usando que as matrizes são idempotentes e simétricas.

16. Como consequência, mostre que

$$\|\mathbf{Y}\|^2 = \sum_{i=1}^n y_i^2 = n\bar{y}^2 + \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n (y_i - \bar{y})^2$$

e também que

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

17. O índice de correlação múltipla R^2 é uma medida global do grau de proximidade do vetor ajustado ou predito pelo modelo $\hat{\mathbf{Y}}$ ao vetor resposta \mathbf{Y} e ele é definido como

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2} .$$

Quanto maior o R^2 , melhor o ajuste.

Mostre que sempre temos $R^2 \in [0, 1]$ e que o R^2 também pode ser escrito como

$$R^2 = \frac{\sum_i (\hat{y}_i - \bar{y})^2}{\sum_i (y_i - \bar{y})^2}$$

106. Mostre que a média aritmética dos valores no vetor ajustado $\hat{\mathbf{Y}}$ é igual a média dos valores observados sempre que $\mathbf{1}$ estiver na matriz \mathbf{X} . Isto é, mostre que $\sum_{i=1}^n \hat{y}_i/n = \bar{y}$. DICA: represente a soma como o produto interno de dois vetores.

19. Considere o índice empírico de correlação linear de Pearson entre os vetores \mathbf{Y} e $\hat{\mathbf{Y}}$ dado por

$$r = \frac{\sum_i (y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\{\sum_i (y_i - \bar{y})^2 \sum_i (\hat{y}_i - \bar{\hat{y}})^2\}^{1/2}}$$

Você vai mostrar que $R^2 = r^2$ trabalhando primeiro o numerador de r :

$$\begin{aligned} \langle \mathbf{Y} - \bar{y}\mathbf{1}, \hat{\mathbf{Y}} - \bar{\hat{y}}\mathbf{1} \rangle &= \langle \mathbf{Y} - \bar{y}\mathbf{1}, \hat{\mathbf{Y}} - \bar{y}\mathbf{1} \rangle \quad \text{pois } \sum_{i=1}^n \hat{y}_i/n = \bar{y}. \\ &= \langle \mathbf{Y} - \hat{\mathbf{Y}} + \hat{\mathbf{Y}} - \bar{y}\mathbf{1}, \hat{\mathbf{Y}} - \bar{y}\mathbf{1} \rangle \end{aligned}$$

Conclua que o numerador de r é igual a $\|\hat{\mathbf{Y}} - \bar{y}\mathbf{1}\|^2$ e que $r^2 = R^2$. Assim, o índice R^2 mede o quadrado da correlação entre os vetores \mathbf{Y} e $\hat{\mathbf{Y}}$.

20. Responda V ou F para as seguintes afirmativas:

- R^2 mede a proporção da variabilidade (ou variação) total da resposta que é explicada pelo modelo.
- R^2 mede a proporção da variação da resposta que o modelo não consegue explicar.
- R^2 é igual a zero se a matriz de desenho tiver apenas a coluna $\mathbf{1}$.

21. Seja X uma matriz de números reais de dimensão $n \times (p+1)$, seja $\beta = (\beta_0, \beta_1, \dots, \beta_p)$ um vetor $(p+1) \times 1$ e \mathbf{y} um vetor $n \times 1$.

- Sejam v_1, \dots, v_k vetores do \mathbb{R}^n . Verifique que o conjunto das combinações lineares desses vetores forma um sub-espço vetorial do \mathbb{R}^n .
- Verifique que $X\beta = \beta_0 X_0 + \beta_1 X_1 + \dots + \beta_p X_p$ onde X_0, X_1, \dots, X_p são os vetores colunas de X . Assim, o conjunto $\mathfrak{M}(X)$ das combinações lineares das colunas de X é igual a $\mathfrak{M}(X) = \{X\beta \mid \beta \in \mathbb{R}^{p+1}\}$.
- Seja W um subespaço do espaço vetorial V . Definimos o espaço ortogonal de W como sendo

$$W^\perp = \{u \in V \mid \langle u, w \rangle = 0 \quad \forall w \in W\}$$

Mostre que W^\perp é um subespaço vetorial de V .

Solução: $\vec{0} \in W^\perp$ pois $\langle \vec{0}, w \rangle = 0$ para todo $w \in W$. E também $\langle a_1 u_1 + a_2 u_2, w \rangle = 0$ se $\langle u_1, w \rangle = 0$ e $\langle u_2, w \rangle = 0$

- Seja $H = X(X'X)^{-1}X'$ de dimensão $n \times n$. Verifique que H é idempotente ($H^2 = H$) e simétrica ($H' = H$).

- Seja $y \in \mathbb{R}^n$. A matriz P de dimensão $n \times n$ é dita de projeção ortogonal num certo subespaço vetorial se $y - Py \perp Py$ para todo $y \in \mathbb{R}^n$. Mostre que $H = X(X'X)^{-1}X'$ é uma matriz de projeção ortogonal usando que H é idempotente e simétrica.

- Como $H = X(X'X)^{-1}X'$ é uma matriz de projeção ortogonal, resta saber em que sub-espaço vetorial W a matriz H projeta os vetores $y \in \mathbb{R}^n$. Mostre que H projeta ortogonalmente em $\mathfrak{M}(X)$ (isto é, mostre que $W = \mathfrak{M}(X)$.)

Solução: Para todo $y \in \mathbb{R}^n$, temos $Hy = X(X'X)^{-1}X'y = Xb$ onde $b = (X'X)^{-1}X'y$. Assim, $Hy \in \mathfrak{M}(X)$ para todo y e portanto $W \subset \mathfrak{M}(X)$. Por outro lado, tome um elemento Xb qualquer de $\mathfrak{M}(X)$. Por definição, $Hy \in W$ para todo y . Em particular, tomando $y = Xb$, temos então $HXb \in W$. Mas $HXb = X(X'X)^{-1}X'Xb = Xb$. Isto é, $Xb \in W$ e portanto $\mathfrak{M}(X) \subset W$. Concluimos então que $W = \mathfrak{M}(X)$.

- Seja $H = X(X'X)^{-1}X'$ a matriz de projeção ortogonal no espaço $\mathfrak{M}(X)$ das combinações lineares das colunas de X . Mostre que ao escolher β tal que $X\beta = Hy$ estamos minimizando a distância $\|y - X\beta\|^2$. DICA: Escreva some e subtraia Hy em $\|y - X\beta\|^2$ e use que $\|v\|^2 = \langle v, v \rangle$

Solução: $\|y - X\beta\|^2 = \langle y - X\beta, y - X\beta \rangle$. Somando e subtraindo Hy obtemos

$$\begin{aligned} \|y - X\beta\|^2 &= \langle y - Hy + Hy - X\beta, y - Hy + Hy - X\beta \rangle \\ &= \langle y - Hy, y - Hy \rangle + \langle Hy - X\beta, Hy - X\beta \rangle - 2 \langle y - Hy, Hy - X\beta \rangle \\ &= \|y - Hy\|^2 + \|Hy - X\beta\|^2 + 0. \end{aligned}$$

O último termo acima é zero pois $Hy - X\beta \in \mathfrak{M}(X)$ já que $Hy \in \mathfrak{M}(X)$ e $X\beta \in \mathfrak{M}(X)$ e o conjunto $\mathfrak{M}(X)$ é um sub-espaço vetorial (e portanto contém a diferença dos vetores). Além disso, $y - Hy \in \mathfrak{M}(X)^\perp$. Portanto, o produto interno $\langle y - Hy, Hy - X\beta \rangle$ é nulo.

Assim, $\|y - X\beta\|^2 = \|y - Hy\|^2 + \|Hy - X\beta\|^2$. O primeiro termo do lado direito não depende de β e o segundo é não-negativo. Ele será minimizado se for igual a zero, o que ocorre se tomamos $X\beta = Hy$.

22. *Regressão linear e distribuição condicional:* Vamos considerar um modelo (na verdade, mais uma caricatura) de como a renda do trabalho Y de um indivíduo qualquer está associada com o número de anos de estudo X desse mesmo indivíduo. Vamos supor que, para um indivíduo com $X = x$ anos de estudo teremos a renda Y como uma variável aleatória com distribuição normal com esperança $\mathbb{E}(Y|X = x) = g(x) = 300 + 100 * x$ e variância $\sigma^2 = 50^2$.

Responda V ou F às afirmações abaixo:

- Se $X = 10$ para um indivíduo (isto é, se ele possui 10 anos de estudo), então a sua renda é uma variável aleatória com distribuição $N(1300, 50^2)$.
- $\mathbb{E}(Y) = 300 + 100 * x$.
- $\mathbb{E}(Y|X = x) = 300 + 100 * x$.
- $\mathbb{V}(Y) = 50^2$.

- $\mathbb{V}(Y|X = x) = 50^2$.
- A distribuição de Y é normal (ou gaussiana).
- Dado o valor de x , a distribuição de Y é normal (ou gaussiana).

23. Numa lista anterior, você usou os dados do arquivo `aptos.txt` com preços de apartamento no bairro Sion em Belo Horizonte para criar um modelo de regressão. Usando álgebra matricial no \mathbb{R} , obtenha uma estimativa da matriz de covariância do estimador de mínimos quadrados $\hat{\beta}$ dada por

$$\mathbb{V}(\hat{\beta}) = \widehat{\sigma^2} (\mathbf{X}'\mathbf{X})^{-1}$$

onde

$$\widehat{\sigma^2} = \frac{1}{n-p} (\mathbf{Y} - \hat{\mathbf{Y}})' (\mathbf{Y} - \hat{\mathbf{Y}}) = \frac{1}{n-p} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

24. Considere um vetor $\mathbf{X} = (X_1, X_2)$ com distribuição normal bivariada com vetor esperado $\boldsymbol{\mu} = (\mu_1, \mu_2)$ e matriz de covariância

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{11} & \rho\sqrt{\sigma_{11}\sigma_{22}} \\ \rho\sqrt{\sigma_{11}\sigma_{22}} & \sigma_{22} \end{bmatrix}$$

Usando o resultado dos slides, mostre que a distribuição condicional de $(X_2|X_1 = x_1)$ é $N(\mu_c, \sigma_c^2)$ onde

$$\mu_c = \mu_2 + \rho\sqrt{\frac{\sigma_{22}}{\sigma_{11}}}(x_1 - \mu_1) = \mu_2 + \rho\sqrt{\sigma_{22}}\frac{x_1 - \mu_1}{\sqrt{\sigma_{11}}}$$

e

$$\sigma_c^2 = \sigma_{22}(1 - \rho^2)$$

A partir desses resultados, verifique se as afirmações abaixo são V ou F:

- Saber que o valor $X_1 = x_1$ está dois desvios-padrão acima de seu valor esperado (isto é, $(x_1 - \mu_1)/\sqrt{\sigma_{11}} = 2$) implica que devemos esperar que X_2 também fique dois desvios-padrão acima de seu valor esperado.
- Dado que $X_1 = x_1$, a variabilidade de X_2 em torno de seu valor esperado é maior se $x_1 < \mu_1$ do que se $x_1 > \mu_1$.
- Conhecer o valor de X_1 (e assim eliminar parte da incerteza existente) sempre diminui a incerteza da parte aleatória permanece desconhecida (isto é, compare a variabilidade de X_2 condicionada e não-condicionada no valor de X_1).
- μ_c é uma função linear de x_1 .

25. Considere um modelo de regressão linear onde a matriz de desenho \mathbf{X} possui uma única ¹⁰⁹ coluna formada pelo atributo j de forma que \mathbf{X} é uma matriz $n \times 1$. Digamos que o atributo j seja o número total de linhas de código de um software e a resposta seja o tempo até a obtenção de uma primeira versão estável do software. Obtemos dados relacionados a n distintos software. Nosso modelo de regressão linear SEM INTERCEPTO é dado por:

$$\begin{aligned}\mathbf{Y} &= \mathbf{x}^{(j)}\beta_j + \boldsymbol{\varepsilon} \\ \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} &= \begin{bmatrix} x_{1j} \\ x_{2j} \\ \vdots \\ x_{nj} \end{bmatrix} \beta_j + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} \\ &= \begin{bmatrix} x_{1j}\beta_j + \varepsilon_1 \\ x_{2j}\beta_j + \varepsilon_2 \\ \vdots \\ x_{nj}\beta_j + \varepsilon_n \end{bmatrix}\end{aligned}$$

Note que o vetor-coeficiente neste caso é simplesmente o escalar β_j , um vetor de dimensão 1.

Ao contrário do que fazemos quase sempre por *default*, no modelo acima, nós não estamos usando a coluna de vetor n -dim de 1's representado por $\mathbf{1} = (1, \dots, 1)'$. Isto significa que o modelo assume que a resposta Y está relacionada ao atributo através de uma relação linear que passa pela origem da forma:

$$y \approx x\beta$$

É claro que a maioria das situações práticas terá um intercepto não-nulo. A utilidade deste modelo vai ficar clara no exercício 10.

- Seja $\langle \mathbf{x}, \mathbf{y} \rangle = \sum_i x_i y_i$ o produto interno de dois vetores \mathbf{x} e \mathbf{y} . Mostre que o estimador de mínimos quadrados de β_j neste modelo com um único atributo é dados por

$$\begin{aligned}\beta_j &= \frac{\mathbf{x}^{(j)'}\mathbf{Y}}{\mathbf{x}^{(j)'}\mathbf{x}^{(j)}} \\ &= \frac{\sum_{i=1}^n x_{ij}y_i}{\sum_{i=1}^n x_{ij}^2} \\ &= \frac{\langle \mathbf{x}^{(j)}, \mathbf{y} \rangle}{\langle \mathbf{x}^{(j)}, \mathbf{x}^{(j)} \rangle} \\ &= \frac{\langle \mathbf{x}^{(j)}, \mathbf{y} \rangle}{\|\mathbf{x}^{(j)}\|^2}\end{aligned}$$

- Suponha que o único atributo no modelo seja a coluna de 1's representada por $\mathbf{1} = (1, \dots, 1)'$. Mostre que o (único) coeficiente neste caso é dado pela média aritmética das observações

$$\beta_0 = \frac{\mathbf{1}'\mathbf{Y}}{\mathbf{1}'\mathbf{1}} = \frac{1}{n} \sum_i y_i = \bar{Y}$$

- Considerando o item anterior, conclua que o modelo no caso em que o único atributo é a coluna de 1's é da forma:

$$\mathbf{Y} = \mathbf{1} \beta_0 + \boldsymbol{\varepsilon}$$

o que significa que Y_1, Y_2, \dots, Y_n são i.i.d. $N(\beta_0, \sigma^2)$.

O modelo estimado por mínimos quadrados produz a seguinte decomposição do vetor \mathbf{Y} :

$$\mathbf{Y} = \bar{Y} \mathbf{1} + (\mathbf{Y} - \bar{Y} \mathbf{1})$$

Neste modelo, $\hat{\mathbf{Y}} = \bar{Y} \mathbf{1}$ e o vetor de resíduos é $\mathbf{r} = \mathbf{Y} - \bar{Y} \mathbf{1}$.

26. Suponha que o modelo de regressão correto envolve dois conjuntos de atributos. O primeiro deles é de fato medido empiricamente e está armazenado na matriz \mathbf{X} de dimensão $n \times p$ onde a primeira coluna é o vetor $\mathbf{1}$ de 1's. O segundo conjunto de atributos também é importante para explicar a variação de Y mas esses atributos não são medidos porque são desconhecidos ou porque sua medição é impossível ou inviável em termos práticos. Vamos supor que existam k atributos nesse segundo conjunto. Caso eles fossem medidos, estariam numa matriz \mathbf{Z} de dimensão $n \times k$ (sem a coluna $\mathbf{1}$). Assim, o modelo correto é dado por

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}. \quad (9.1)$$

onde $\boldsymbol{\gamma}$ é um vetor $k \times 1$ dos coeficientes associados com as colunas-atributos em \mathbf{Z} .

O analista de dados possui apenas a matriz \mathbf{X} para fazer a regressão e ele obtém a estimativa de mínimos quadrados para o coeficiente $\boldsymbol{\beta}$ da maneira usual:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{Y}$$

- Substitua a expressão (9.1) de \mathbf{Y} na fórmula de $\hat{\boldsymbol{\beta}}$ e mostre que $\mathbb{E}(\hat{\boldsymbol{\beta}})$ é viciado para estimar $\boldsymbol{\beta}$ sendo que o vício de estimação é dado por

$$\mathbb{E}(\hat{\boldsymbol{\beta}}) - \boldsymbol{\beta} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{Z}\boldsymbol{\gamma}$$

- Para entender melhor este vício, imagine que \mathbf{Z} contenha um único atributo. Dessa forma, \mathbf{Z} é uma matriz $n \times 1$, um vetor-coluna. Imagine que vamos explicar a variação deste atributo \mathbf{Z} usando os p atributos da matriz \mathbf{X} . Isto é, vamos imaginar um modelo de regressão linear da seguinte forma:

$$\mathbf{Z} = \mathbf{X}\boldsymbol{\alpha} + \boldsymbol{\varepsilon}^*$$

onde $\boldsymbol{\alpha}$ é um vetor $p \times 1$ de coeficientes. Verifique que o estimador de mínimos quadrados de $\boldsymbol{\alpha}$ é dado por

$$\hat{\boldsymbol{\alpha}} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{Z}$$

Interprete agora o vício associado com o estimador $\hat{\boldsymbol{\beta}}$.

Solução: O estimador de regressão de $\boldsymbol{\beta}$ no modelo reduzido é viciado por uma quantidade correspondente a $\boldsymbol{\gamma}$ (coef do modelo full) vezes o coef da regressão de \mathbf{Z} em \mathbf{X} .

- Generalize para \mathbf{Z} com varios atributos usando o fato de que um produto matricial \mathbf{AB} é igual a uma matriz cuja j -ésima coluna é a aplicação da matriz \mathbf{A} à j -ésima coluna de \mathbf{B} . 111

27. Considere agora o problema inverso, em que colocamos no modelo de regressão mais atributos do que o necessário. O modelo correto é dado por

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}.$$

mas, sem saber disso, assumimos um modelo da seguinte forma

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}^*.$$

onde $\boldsymbol{\gamma}$ é um vetor $k \times 1$ de coeficientes associados com as colunas-atributos em \mathbf{Z} que não necessárias pois a distribuição de \mathbf{Y} não depende desses atributos. Vamos estimar os coeficientes dos atributos com uma matriz de desenho da forma $\mathbf{W} = [\mathbf{X}|\mathbf{Z}]$. Note que o verdadeiro valor do parâmetro $\boldsymbol{\gamma}$ é o vetor $\mathbf{0} = (0, \dots, 0)'$.

Usando apenas um simples argumento em palavras, mostre que o estimador usual de mínimos quadrados estima $\boldsymbol{\beta}$ sem vício.

28. O arquivo `exames.txt` mostra os escores F em um exame final e os escores em dois exames preliminares P_1 e P_2 de 22 estudantes. Use o comando `pairs` e `cor` para visualizar e estimar a correlação entre os dados.

```
dados = read.table("exames.txt", header=T)
head(dados)
attach(dados)
pairs(dados)
cor(dados)
```

Veja que F é bastante correlacionada com as notas prévias, sendo um pouco mais correlacionada com a nota mais recente P_2 . Rode um modelo de regressão linear para explicar F usando apenas P_1 como variável preditora e depois usando ambas, P_1 e P_2 . Veja que o coeficiente linear de P_1 é bem diferente nos dois casos.

```
summary(lm(F ~ P1))
summary(lm(F ~ P1 + P2))
```

A relação entre os coeficientes de regressão simples (com um único atributo) e os coeficientes de regressão múltipla podem ser vistos quando comparamos as seguintes equações de regressão:

$$\begin{aligned}\hat{F} &= \hat{\beta}_0 + \hat{\beta}_1 P_1 + \hat{\beta}_2 P_2 \\ \hat{F} &= \hat{\beta}'_0 + \hat{\beta}'_1 P_1 \\ \hat{F} &= \hat{\beta}''_0 + \hat{\beta}''_2 P_2 \\ \hat{P}_1 &= \hat{\alpha}_0 + \hat{\alpha}_2 P_2 \\ \hat{P}_2 &= \hat{\alpha}'_0 + \hat{\alpha}'_1 P_1\end{aligned}$$

$$\hat{\beta}'_1 = \hat{\beta}_1 + \hat{\beta}_2 \hat{\alpha}'_1$$

Isto é, o coeficiente da regressão linear simples de F em P_1 é o coeficiente de regressão múltipla de P_1 mais coeficiente da regressão múltipla de P_2 vezes o coeficiente da regressão do atributo P_2 regredido em P_1 . Compare com o resultado teórico que voce encontrou no exercício 4 da Lista 08. É possível fazer uma prova matemática e geral deste fato mas ela exige muita manipulação matricial.

29. Muitas vezes, o modelo de regressão linear ajusta-se perfeitamente a dados não-lineares após fazermos uma transformação nos dados. Este é o caso dos dados no arquivo `mortalidade.txt` com informações obtidas junto a uma seguradora brasileira referentes ao público consumidor de planos de seguro de vida num certo ano. Neste arquivo consideramos apenas a população masculina com 21 anos ou mais. Ele possui três colunas. A primeira apresenta x , a idade em anos. A segunda apresenta o número n_x de participantes do fundo que possuíam a idade x no dia 01 de Janeiro. A terceira coluna apresenta o número d_x desses indivíduos da segunda coluna que não chegaram vivos ao dia 01 de janeiro do ano seguinte.

- Leia os dados no R e crie um vetor $\mathbf{mx} = \mathbf{dx}/\mathbf{nx}$ com a razão entre a terceira e a segunda colunas. Dado que um indivíduo da população chega a fazer x anos de idade, o valor $m_x = d_x/n_x$ estima a probabilidade dele falecer antes de completar $x + 1$ anos.
- Queremos um modelo para o aumento de m_x com x . Faça um gráfico de dispersão de m_x versus x . Existem dois problemas aqui. O primeiro é que o aumento de m_x é claro mas não a forma exata pela qual este aumento ocorre. O segundo é que as últimas idades tem um número muito pequeno de indivíduos e assim m_x não é uma estimativa razoável. Vamos eliminar os dados de idades superiores a 77 anos. Isto ajuda com o segundo problema.
- Para lidar com o primeiro problema, faça um gráfico de dispersão da variável $\log(m_x)$ versus x . Você deve observar uma nítida relação linear entre os dois. Isto é, temos

$$\log(m_x) \approx \beta_0 + \beta_1 x$$

o que implica, tomando exponencial dos dois lados, em

$$m_x \approx e^{\beta_0} e^{\beta_1 x} = b_0 \left(e^{\beta_1} \right)^x = b_0 b_1^x$$

- Interprete o parâmetro $b_1 = e^{\beta_1}$ em termos do aumento da mortalidade com o aumento da idade. Para isto, calcule aproximadamente m_{x+1}/m_x e conclua que a cada ano adicional de vida a chance de falecer antes de completar o próximo aniversário aumenta em aproximadamente e^{β_1} .
- Ajuste um modelo de regressão linear simples tomando $\log(m_x)$ como resposta e x como atributo (além da coluna **1**).

- Conclua que a cada ano adicional de vida a chance de falecer antes de completar o próximo aniversário aumenta em aproximadamente 5%. O mais relevante: este aumento não depende de x : seja um jovem ou um idoso, um ano a mais de vida faz seu risco de morte anual aumentar em 5% do era antes.

Solução:

```
# Read the data
dados = read.table("mortalidade.txt", header=T); attach(dados)
mx = dados$mortes/dados$pop ; plot(x, mx)
# eliminando as ultimas faixas etarias, ficando apenas com idades < 78 anos
par(mfrow=c(1,2))
plot(x[x < 78], mx[x < 78]) ; plot(x[x < 78], mx[x < 78], type="l")
plot(x[x < 78], log(mx[x < 78])); plot(x[x < 78], log(mx[x < 78]), type="l")
res = lm( log(mx[x < 78]) ~ x[x < 78]); summary(res)
```

30. O arquivo `TurtleEggs.txt` contém dados de Ashton *et al.* (2007). Eles mediram o comprimento da carapaça (em mm), de 18 Tartarugas Gopher fêmeas (*Gopherus Polifemo*) do Okechee County Park, Florida. Tomaram também um raio-X para contar o número de ovos em cada uma delas. Faça um gráfico de **eggs** versus *length* para verificar que um modelo linear não é apropriado. Uma regressão com um modelo polinomial de segundo grau parece razoável:

$$\mathbb{E}(Y_i|x) = \beta_0 + \beta_1 x + \beta_2 x^2$$

onde Y é o número de ovos e x é o comprimento da carapaça.

Ajuste uma regressão de segundo grau. Superponha a parábola de melhor ajuste aos dados no gráfico de dispersão de **eggs** versus *length*. De acordo com <http://udel.edu/~mcdonald/statcurvreg.html>, “a primeira parte do gráfico não É surpreendente, É fácil imaginar por que as tartarugas maiores teriam mais ovos. o declínio no número de ovos acima 310 milímetros comprimento da carapaça É o interessante Este resultado sugere que a produção de ovos diminui nestes tartarugas a medida em que envelhecem e ficam grandes”.

Solução:

```
dados = read.table("TurtleEggs.txt", header=T)
attach(dados); plot(length, eggs)
x = length; x2 = length^2
res = lm(eggs ~ x + x2) ; summary(res)
xx = seq(280, 340, by=1)
yy = -8.999e+02 + 5.857*xx -9.425e-03*xx^2
lines(xx, yy)
```

-
31. Movimentos planetários em torno das suas estrelas podem causar variações na velocidade radial da estrela. Os dados do arquivo `starvelocity.txt` foram obtidos por Geoff Marcy,

no Observatório Lick, e referem-se à estrela Pegasus 51, uma estrela similar ao nosso sol e localizada na constelação de Pégaso a 50 anos-luz da Terra.

- Faça um gráfico da velocidade (Y) versus o tempo t , medido em dias. A velocidade da estrela 51 Pegasi varia de maneira cíclica com um período de aproximadamente 4.231 dias e com uma amplitude de 56 metros por segundo. Isto sugere que a estrela está cercada por um corpo celeste com uma massa aproximadamente igual a metade daquela de Júpiter.
- Use regressão linear para estimar o modelo

$$\begin{aligned} Y_t &= a + b \cos(2\pi wt + \phi) + \varepsilon \\ &= a + b \cos(2\pi wt) \sin(\phi) + b \sin(2\pi wt) \cos(\phi) + \varepsilon \\ &= \beta_0 + \beta_1 \cos(2\pi wt) + \beta_2 \sin(2\pi wt) \\ &= \beta_0 + \beta_1 x_{t1} + \beta_2 x_{t2} \end{aligned}$$

Se a frequência w tiver de ser estimada teremos um problema de mínimos quadrados não-lineares. Para evitar este problema nesta altura do curso, assuma que a frequência w é conhecida igual a $w = 1/4.231$ (isto é, o período orbital é de $41/w = 4.231$ dias). Com w fixado, podemos obter as colunas X_1 e X_2 para cada instante t e ajustar um modelo de regressão usual.

- Trace a curva encontrada no gráfico com os pontos.

Solução:

```
dias = c(2.65, 2.80, 2.96, 3.80, 3.90, 4.60, 4.80, 4.90, 5.65, 5.92)
vel = c(-45.5, -48.8, -54.0, -13.5, -7.0, 42.0, 50.5, 54.0, 36.0, 14.5)
plot(dias, vel); % w = 1/4.231; x1 = sin(2*pi*w*dias); x2 = cos(2*pi*w*dias)
res = lm(vel ~ x1 + x2); % summary(res); dd = seq(2.60, 6, by=0.01)
vv = 1.3471 + 49.8894*sin(2*pi*w*dd) + 17.9099*cos(2*pi*w*dd); lines(dd,vv)
```

32. *O efeito de centrar os atributos.* O objetivo deste exercício é mostrar que, ao centrar os atributos, temos coeficientes relacionados de forma simples aos coeficientes obtidos com coeficientes não-centrados. Seja $\hat{\beta}^* = (\hat{\beta}_0^*, \hat{\beta}_1^*, \hat{\beta}_2^*)$ o vetor que minimiza

$$\sum_i (y_i - (\beta_0^* + \beta_1^*(x_{i1} - \bar{x}_1) + \beta_2^*(x_{i2} - \bar{x}_2)))^2$$

Isto é, a matriz de desenho tem suas colunas com média zero (exceto a primeira coluna). Seja $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2)$ o coeficiente que minimiza a distância entre \mathbf{Y} e as combinações lineares das colunas *não-centradas*:

$$\sum_i (y_i - (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}))^2$$

Mostre que as soluções dos dois problemas estão relacionadas da seguinte forma:

115

$$\begin{aligned}\hat{\beta}_0 &= \hat{\beta}_0^* - \hat{\beta}_1^* \bar{x}_1 - \hat{\beta}_2^* \bar{x}_2 \\ \hat{\beta}_1 &= \hat{\beta}_1^* \\ \hat{\beta}_2 &= \hat{\beta}_2^*\end{aligned}$$

-
33. Prove que $m = \mathbb{E}(Y)$ minimiza $\mathbb{E}(Y - m)^2$.

Solução: some e subtraia $\mathbb{E}(Y)$ dentro do parênteses, expanda o quadrado e tome esperança de cada termo. A seguir, derive com relação a m .

34. INCOMPLETO AINDA: Estudar a matriz $\mathbf{X}'\mathbf{X}^{-1}$: Se \mathbf{X} tiver colunas linearmente independentes então $\mathbf{X}'\mathbf{X}$ é inversível e definida positiva. Ajudar na interpretação do elemento (i, j) de $\mathbf{X}'\mathbf{X}^{-1}$ como correlação parcial.
-

35. Obtenha o arquivo `miete03.asc` no site http://www.stat.uni-muenchen.de/service/datenarchiv/miete/miete03_e.html. Ele contém os dados de aluguel de 2053 apartamentos em Munique em 2002 com vários atributos/preditores potenciais, todos descritos na página. Use estes dados no restante desta lista.

Existem duas possíveis variáveis resposta: `nm`, o aluguel líquido em euros, e `nmqm`, este aluguel dividido pela área do apartamento. Vamos usar a primeira delas como resposta y . Elimine a segunda variável do restante da análise (não a coloque entre os preditores!!).

- (a) Separe o conjunto de 2053 exemplos em dois conjuntos, um com 600 exemplos escolhidos ao acaso para avaliar a qualidade do ajuste de vários modelos (amostra de validação) e outro com os $2053 - 600 = 1453$ restantes para ser a amostra de treinamento. Use o comando `sample` para selecionar os dados.
- (b) Usando a amostra de treinamento e as 11 variáveis preditoras, ajuste um modelo de regressão linear.
- (c) Obtenha o R^2 deste modelo completo.
- (d) Obtenha o valor da estatística F (e seu p-valor) para a hipótese nula de que todos os coeficientes são zero.
- (e) Verifique que todos os preditores é significativo num teste de $H_0 : \beta_j = 0$ versus $H_A : \beta_j \neq 0$ observando o valor da estatística t e o seu p-valor.
- (f) Obtenha o intervalo de confiança de 95% de cada um dos coeficientes.
- (g) As últimas 7 variáveis são qualitativas com duas categorias apenas. Considerando cada uma delas individualmente, qual o efeito esperado em preço que cada uma delas produz? Qual tem o maior efeito? Note que esta variável de efeito máximo não é aquela com o menor p-valor (ou o maior valor absoluto t^* da estatística t ? Isto é, grande significância estatística não implica maior efeito na resposta.

- 116 (h) Considere um modelo alternativo sem os preditores x_2 e x_3 . Sejam M_f o modelo completo e M_r o modelo reduzido. Avalie qual dos dois é melhor calculando o $SPSE$ de cada modelo onde

$$SPSE_M = \frac{1}{n} \sum_{i=1}^n (y_i - \mathbf{x}_i' \boldsymbol{\beta}^{(-i)})^2 = \frac{1}{n} \sum_{i=1}^n \left(\frac{y_i - \hat{y}_i}{1 - h_{ii}} \right)^2$$

Isto pode ser obtido com os seguintes comandos em R:

```
# ajustando o modelo completo de regressao linear
modg = lm(nm ~ ., dados)

# obtendo vetor com os valores de H[i,i]
hii = influence(modg)$hat

# calculando o leave-one-out cross-validation measure (loocv)
loocv = sum( ( modg$res/(1-hii) )^2 )/length(hii)
```

Bibliografia

- Rao, C.R. (1973) *Linear Statistical Inference and Its Applications*, 2nd Edition. Wiley-Interscience, New York.
- Seber, G. A. F., Lee, A. J. (2003) *Linear Regression Analysis*, 2nd Edition. Wiley-Interscience, New York.

Capítulo 10

Teoria da Estimação Pontual

1. Imagine uma fabricação de peças cujo diâmetro é uma variável aleatória com distribuição normal com parâmetro μ . Duas amostras de variáveis aleatórias i.i.d., ambas de tamanho 5, foram retiradas da população e os valores observados são os seguintes:

Amostra 1	5.8	9.8	12.2	14.4	14.5	média = 11.3
Amostra 2	9.7	10.7	12.6	10.7	4.6	média = 9.7

- Qual é a melhor estimativa de μ , a primeira média ou a segunda? Isto é, qual delas possui menor erro de estimação?
 - A estimativa $(11.3 + 9.7)/2$ possui erro de estimação menor que a primeira ou segunda média calculada na tabela?
 - E quanto ao estimador $T = (\bar{X}_1 + \bar{X}_2)/2$, a média aritmética das amostras 1 e 2? Ele é melhor que o estimador \bar{X}_1 ? Qual o critério que você usou para decidir sobre essa pergunta?
 - Considerando a primeira amostra acima, calcule a mediana amostral. Ela possui erro de estimação menor que a média amostral 11.3? Qual das duas estimativas você usaria? Por quê?
2. Suponha que X_1, X_2, \dots, X_n sejam v.a.'s i.i.d. com distribuição Poisson com parâmetro θ . Obtenha o MLE de θ . Supondo que $n = 4$ e que as observações tenham sido 3, 0, 1, 1 diga qual o valor do MLE.
 3. No Brasil, o Ministério da Previdência Social assegura um benefício de auxílio-doença ao trabalhador que paga pelos dias em que ele foi impedido de trabalhar por doença ou acidente. Como parte dos pagamentos é efetuado pelas empresas, um sindicato patronal de empresas de ônibus urbano de certa cidade contratou uma consultoria atuarial para analisar o número de faltas ao trabalho por doença. Dados de 20 empresas de ônibus urbano foram coletados.

O número de acidentes na empresa i é denotado por Y_i e depende do número de homem-hora trabalhado. Este número de homem-hora é representado por h_i e é medido em unidades de 10 mil horas. Assim, $h_i = 3.5$ representa 35 mil homens-horas. Quanto maior

118 h_i , maior tende a ser o número de dias parados pois a exposição ao risco é maior na empresa com h_i maior.

Para cada empresa defina o parâmetro λ_i que representa o número médio ou esperado de acidentes na empresa i por cada 10 mil homens-hora. Veja que 12 homens trabalhando 40 horas por semana ao longo de $21 = 30 - 9$ dias por mês dá um total de $12 \times 40 \times 21 = 10080$. Assim, grosseiramente, 10 mil homens/horas representam 12 homens trabalhando em tempo integral num mês.

O modelo estatístico para os dados é que Y_1, \dots, Y_{20} são independentes. Entretanto, essas variáveis aleatórias não são identicamente distribuídas pois os valores de h_1, \dots, h_{20} são diferentes. Assim, supõe-se que

$$Y_i \sim \text{Poisson}(\lambda_i) \text{ ou } \text{Poisson}(h_i\theta)$$

para $i = 1, \dots, 20$. Os dados são do seguinte tipo:

Empresa	h_i	y_i
1	94.5	5
2	15.7	1
3	62.9	5
...

O interesse é em fazer inferência sobre o valor desconhecido θ , a taxa de ausência por doença nas empresas. Em particular, para uma nova empresa que vai fazer seguro-doença para cobrir faltas de empregados, é preciso saber qual a expectativa de faltas durante um mês se ela tem um certo número h de homens-hora de trabalho.

Calcule:

- o EMV de θ .
- O EMV é não-viciado?
- Calcule o número de informação de Fisher $I(\theta)$.
- É verdade que a informação $I_n(\theta)$ com n observações é igual a n vezes a informação com uma única informação? Como a informação aumenta com n ?
- Suponha que os dados sejam esses abaixo:

```
h <- c(118.3, 13.4, 68.3, 141.6, 113.1, 63.6, 135.5, 107.5, 35.2,
      28.1, 34.7, 42.5, 139.7, 140.4, 79.2, 148.2, 28, 72.4, 50.9, 89.1)
y <- c(8, 0, 6, 18, 7, 6, 13, 4, 7, 4, 0, 3, 26, 10, 4, 15, 7, 10,
      9, 14)
```

Usando um diagrama de dispersão, verifique se existe associação entre h e y . Estime θ usando o MLE.

4. No problema acima, as empresas podem ser classificadas em quatro grupos distintos e esperamos que o valor de θ seja diferente nesses quatro grupos. As empresas são classificadas em 4 grupos de acordo com as seguintes categorias:

- $Z_1 = 0$ se a empresa faz cursos de treinamento e regulares sobre cuidados ao dirigir, e $Z_1 = 1$, caso contrário.
- $Z_2 = 0$ se a empresa cobra de cada motorista responsável por um acidente parte das perdas causadas (o valor cobrado é limitado a um máximo) e $Z_2 = 1$, caso contrário

No caso de Z_2 , é óbvio que se o motorista envolvido num acidente não é considerado culpado do mesmo, a cobrança não é feita. A cobrança também é limitada a um máximo de 2 salário mensais. Em cada mês, até um máximo de 15% do salário é descontado da folha de pagamentos.

Espera-se que empresas com $Z_1 = 1$ e $Z_2 = 1$ tenham θ maiores que aquelas com $Z_1 = 0$ e $Z_2 = 0$. Empresas com $Z_1 = 0$ e $Z_2 = 1$ ou $Z_1 = 1$ e $Z_2 = 0$ seriam casos intermediários.

É claro que, mesmo dentro de um dos grupos ($Z_1 = 1$ e $Z_2 = 1$, por exmplo) ainda esperamos que as empresas não sejam idênticas e portanto que seus θ 's possam diferir um pouco. Esperamos que, dentre as 20 empresas, boa parte das diferenças entre seus θ 's possa ser devido às políticas de prevenção e de punição. Ainda sobriaria um resíduo de causas não controladas tornando os θ 's dentro de cada um dos 4 grupos ligeiramente diferentes. No entnato, nós vamos ignorar estas diferenças residuais e supor que, empresa numa mesma categoria, tenham θ 's idênticos.

Queremos agora estimar o impacto dessas políticas de prevenção/punição na redução do número de faltas. Existe algum impacto? Qual política é mais eficaz? Como estudar isto? Uma possibilidade é fazer uma análise separada para cada um dos 4 grupos possíveis. Uma saída *muito melhor* é fazer um único modelo em que todos os quatro grupos estejam presentes e com parâmetros medindo o impacto dos programas de prevenção/punição. Para isto, precisamos alterar o modelo do exercício anterior. Vamos assumir que

$$Y_i \sim \text{Poisson}(h_i \theta_i)$$

para $i = 1, \dots, 20$ que elas sejam v.a.'s independentes.

Existem quatro valores possíveis para θ_i , dependendo da categoria à qual a empresa i pertence. Vamos supor que, ao passar de $Z_1 = 0$ para $Z_1 = 1$ o valor de θ seja aumentado por um fator multiplicativo. O mesmo ao passar de $Z_2 = 0$ para $Z_2 = 1$. Isto é, supomos que

$$\theta_i = e^{\beta_0} e^{\beta_1 Z_{1i}} e^{\beta_2 Z_{2i}}$$

onde Z_{1i} e Z_{2i} são os valores das variáveis independentes para a empresa i . Os dados de h_i e y_i são os mesmos de antes e os de Z_1 e Z_2 são os seguintes:

```
Z1 <- c(0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1),
Z2 <- c(0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 1, 1, 1),
```

- Descreva o modelo estatístico.
- Interprete os parâmetros do modelo. Para isto, verifique qual o efeito em $E(Y_i)$ de uma empresa mudar de uma categoria para outra categoria. Compare o valor de θ_i e

de $\log(\theta_i)$ para uma empresa que tenha $Z_1 = 0$ e $Z_2 = 0$ e outra empresa com $Z_1 = 1$ e $Z_2 = 0$. E com uma empresa com $Z_1 = 1$ e $Z_2 = 1$. Faça comparações entre todos os 6 tipos de *pares* de empresas possíveis.

- Obtenha o EMV e a MATRIZ 3×3 de informação de Fisher para os parâmetros do modelo.
- Obtenha o EMV e a MATRIZ 3×3 de informação de Fisher para os parâmetros $e^{\beta_0}, e^{\beta_1}, e^{\beta_2}$.

5. Leia no livro texto a demonstração e os exemplos associados com as propriedades básicas de esperança e variância:
6. Considere um modelo estatístico para dados de sobrevivência que supõe que X_1, X_2, \dots, X_n são i.i.d. com distribuição exponencial parametrizada por λ . Suponha que, com $n = 10$, os dados observados sejam os seguintes: 1.21, 0.15, 2.02, 0.37, 2.55, 2.20, 1.06, 0.10, 0.35, e 0.15. Use o conjunto de comandos R dados abaixo para desenhar a função de verossimilhança para o parâmetro λ . Procure entender o que cada comando está fazendo.

```
dados <- c(1.21,0.15,2.02,0.37,2.55,2.20,1.06,0.10,0.35,0.15)
lambda <- seq(0,3,length=300)
veross <- (lambda^10) * exp(-lambda * sum(dados))
plot(lambda,veross,type="l")
```

Os comandos abaixo sobrepõem uma segunda curva de verossimilhança baseada numa possível segunda amostra independente da primeira e gerada pelo mesmo mecanismo.

```
dados2 <- c(0.08,2.72,0.04 0.27 0.99 0.25 0.39 1.84 1.60 3.51
veross2 <- (lambda^10) * exp(-lambda * sum(dados2))
lines(lambda,veross2,lty=2)
```

Construa agora uma função que recebe os dados da amostra como entrada e que retorna o gráfico da verossimilhança para $\lambda \in (a, b)$.

```
veross.exp <-function(x, a, b)
{
# funcao para desenhar a funcao de verossimilhanca
# para dados exponenciais
# INPUT: x = dados de entrada
ene <- length(x)
lambda <- seq(a,b,length=300)
veross <- (lambda^ene) * exp(-lambda * sum(x))
plot(lambda, veross, type= "l",ylab= "verossimilhanca")
title("Funcao de verossimilhanca para dados exponenciais")
return()
}
```

Gere um conjunto de dados de uma exponencial com um parâmetro λ num intervalo (a, b) a sua escolha. Chame a função acima com os dados que você gerou e verifique se o máximo da função de verossimilhança fica próximo do valor λ que você escolheu. Repita o exercício escolhendo diferentes tamanhos n de amostra.

7. Algoritmo EM: Leia o exemplo de misturas gaussianas (MoG ou Mixture of Gaussians, em machine learning) no final da página da Wikipedia: http://en.wikipedia.org/wiki/Expectation-maximization_algorithm. Usando os dados do Old Faithful dataset (pegue os dados, por exemplo, em <http://tinyurl.com/cndlsp3>), estime os parâmetros de uma mistura de duas gaussianas bivariadas usando o algoritmo EM.
-

8. Suponha que X_1, \dots, X_n forme uma amostra aleatória de v.a.'s i.i.d. com uma das seguintes densidades (caso contínuo) ou função de probabilidade (caso discreto). Encontre o EMV de θ e verifique se ele é função da estatística suficiente para estimar θ em cada caso.

- $f(x; \theta) = \theta^x e^{-\theta} / x!$ para $x = 0, 1, 2, \dots$ e com $\theta \in (0, \infty)$ (função de probabilidade Poisson).
 - $f(x; \theta) = \theta e^{-\theta x}$ para $x > 0$ com $\theta \in (0, \infty)$ (densidade exponencial).
 - $f(x; \theta) = \theta c^\theta x^{-(\theta+1)}$ para $x \geq c$ com $\theta \in (0, \infty)$ (densidade Pareto).
 - $f(x; \theta) = \sqrt{\theta} x^{-\sqrt{\theta}-1}$ para $0 \leq x \leq 1$ com $\theta \in (0, \infty)$ (densidade beta($\sqrt{\theta}, 1$)).
 - $f(x; \theta) = (x/\theta)^2 \exp(-x^2/(2\theta^2))$ para $x > 0$ com $\theta \in (0, \infty)$ (densidade Rayleigh).
 - $f(x; \theta) = \theta c x^{c-1} \exp(-\theta x^c)$ para $x \geq 0$ com $\theta \in (0, \infty)$ e $c > 0$ sendo uma constante conhecida (densidade Weibull com c fixo).
 - $f(x; \theta) = \theta(1 - \theta)^{x-1}$ para $x = 1, 2, \dots$ e com $\theta \in (0, \infty)$ (função de probabilidade geométrica).
 - $f(x; \theta) = \theta^2 x e^{-\theta x}$ para $x > 0$ com $\theta \in (0, \infty)$ (densidade Gama(2, θ)).
-

9. Sejam X_1 e X_2 duas v.a.'s independentes com esperança comum $\theta \in \mathbb{R}$ e com $\text{Var}(X_1) = \sigma^2$ e $\text{Var}(X_2) = \sigma^2/4$. Isto é, X_1 e X_2 tendem a oscilar em torno de θ mas X_2 possui um desvio-padrão duas vezes menor que X_1 . Podemos usar estes dois pedaços de informação para estimar θ . Podemos, por exemplo, formar uma combinação linear de X_1 e X_2 propondo o estimador $\hat{\theta} = c_1 X_1 + c_2 X_2$ onde $c_1 - 1$ e c_2 são constantes conhecidas. Por exemplo, podemos pensar em usar $\hat{\theta} = (X_1 + X_2)/2$ ou então usar $\hat{\theta} = \frac{2X_1}{3} + \frac{X_2}{3}$ ou até mesmo $\hat{\theta} = 4X_1 - 2X_2$.

Mostre que $\hat{\theta} = c_1 X_1 + c_2 X_2$ é não-viciado para estimar θ (qualquer que seja o valor de $\theta \in \mathbb{R}$) se, e somente se, $c_1 + c_2 = 1$. Dentre os estimadores da forma $\hat{\theta} = c_1 X_1 + c_2 X_2$ e que são não-viciados para estimar θ , encontre aquele que minimiza o MSE, dado por $\mathbb{E}(\hat{\theta} - \theta)^2$.

10. Generalize o problema anterior para n v.a.s! Sejam X_1, \dots, X_n v.a.s independentes com esperança comum θ e com a variância de X_i igual a σ^2/a_i , sendo os $a_i > 0$ conhecidos e com $\sigma^2 > 0$ desconhecido. Considere a classe de todos os estimadores lineares de θ . Isto é, considere a classe de todos os estimadores que podem ser escritos como $\hat{\theta} = \sum_i c_i X_i$ onde c_i são contantes.

Mostre que na classe dos estimadores lineares, $\hat{\theta}$, é não-viciado para estimar θ se, e somente se $\sum_i c_i = 1$. Dentre todos os estimadores lineares $\sum_i \alpha_i X_i$ de θ que são não-viciados (isto é, satisfazendo $\sum_i \alpha_i = 1$), encontre aquele que minimiza o MSE.

11. Suponha que X_1, X_2, \dots, X_n sejam v.a.'s i.i.d. com distribuição Poisson com parâmetro comum θ . É possível mostrar matematicamente que

$$\hat{\theta}_1 = \bar{X} = \frac{X_1 + \dots + X_n}{n}$$

e

$$\hat{\theta}_2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

são ambos estimadores não viciados para estimar θ . Considere adicionalmente um terceiro estimador não-viciado para θ :

$$\hat{\theta}_3 = (\hat{\theta}_1 + \hat{\theta}_2)/2$$

Faça um pequeno estudo de simulação para identificar qual dos três possui um erro de estimação MSE menor. Para isto, fixe o valor de $\theta = 3$. Gere um grande número de amostras (digamos, 10 mil), cada uma delas de tamanho $n = 10$. Para cada amostra calcule os valores dos três estimadores de θ . Estime o MSE $\mathbb{E}(\hat{\theta}_j - \theta)^2$ de cada estimador usando a média das diferenças ao quadrado entre os 10 mil valores de $\hat{\theta}_j$ e θ . Qual dos estimadores produz um erro MSE menor? Isto significa que o melhor estimador teve SEMPRE o seu valor mais próximo do verdadeiro valor do parâmetro θ ? Estime a probabilidade de que, baseados numa mesma amostra, $\hat{\theta}_2$ esteja mais próximo de θ que $\hat{\theta}_1$.

A conclusão muda se você tomar $n = 20$ e $\theta = 10$?

12. Responda V ou F para as afirmações abaixo.

- Como o parâmetro θ não pode ser predito antes do experimento, ele é uma variável aleatória.
- Num problema de estimação de uma população com distribuição normal $N(\mu, \sigma^2)$ encontrou-se $\bar{x} = 11.3$ numa amostra de tamanho $n = 10$. A distribuição de probabilidade desse valor 11.3 é também uma normal com média μ e variância $\sigma^2/10$.
- Suponha que \bar{X} esteja sendo usado como estimador da média populacional μ . Como a variância de \bar{X} decresce com o tamanho da amostra, então toda estimativa obtida a partir de uma amostra de tamanho 15 possui erro de estimação menor que qualquer outra estimativa obtida a partir de uma amostra de tamanho 10.

- Um estimador não viciado é sempre melhor que um estimador viciado.
- Considere uma estimativa da média populacional μ baseada na média aritmética de uma amostra de tamanho 10 e outra estimativa com uma amostra de tamanho 15. Nunca devemos preferir a estimativa baseada na amostra de 15 pois a estimativa baseada na amostra de tamanho 10 tem alguma chance de estar mais perto do verdadeiro valor desconhecido de μ .

13. Sejam X_1, \dots, X_n i.i.d.'s com distribuição exponencial com parâmetro λ . O interesse é estimar $E(X_i) = 1/\lambda$. Suponha que apenas as variáveis X_i 's que ficarem maiores ou iguais a $x = 10$ sejam observadas. Todas as observações menores que $x = 10$ são perdidas. Assim, a amostra final é possui um número $0 < k \leq n$ de observações.

O estimador baseado na média amostral da amostra de k variáveis é viciado. Ele subestima ou superestima sistematicamente $E(X_i)$? Não precisa calcular o vício.

A distribuição de X_i DADO QUE $X_{>10}$ tem a densidade dada por

$$f(x; \lambda) = \begin{cases} 0, & \text{se } x < 10 \\ \lambda \exp(-\lambda(x - 10)), & \text{se } x \geq 10 \end{cases}$$

Se X_1, \dots, X_k é uma amostra desta distribuição TRUNCADA (em que só observamos os X_i 's maiores que 10), encontre o MLE de λ .

RESP: O MLE é $k / \sum_i (x_i - 10)$.

14. Suponha que será coletada uma amostra de observações independentes Y com distribuição normal. Elas *não são* identicamente distribuídas. A média de Y varia de acordo com o valor de uma covariável x de forma que $Y = \alpha + \beta x + \epsilon$, onde ϵ possui distribuição normal com média 0 e variância σ^2 . Os valores possíveis de x são três: baixo ($x = -1$), médio ($x = 0$) e alto ($x = 1$). Os valores de x são fixos e conhecidos. Eles não são variáveis aleatórias.

São feitas três observações em cada nível de x . Podemos representar os dados na tabela e no gráfico dos valores observados de Y versus x que Figura 11.1.

$x_j = -1$	$x_j = 0$	$x_j = 1$
Y_{11}	Y_{21}	Y_{31}
Y_{12}	Y_{22}	Y_{32}
Y_{13}	Y_{23}	Y_{33}

Vamos representar as observações como $Y_{ij} = \alpha + \beta x_j + \epsilon_{ij}$ onde $x_j = -1, 0$ ou 1 , e os ϵ_{ij} são i.i.d. com distribuição $N(0, \sigma^2)$.

- É correto dizer que $(Y_{ij}|x_j) \sim N(\alpha + \beta x_j, \sigma^2)$ e que as variáveis Y_{ij} são independentes? DICA: não existe pegadinha aqui.
- Calcule $E(Y_{ij}|x_j)$ e $\text{Var}(Y_{ij}|x_j)$ nos três casos: $x_j = -1$, $x_j = 0$ e $x_j = 1$. A variância depende do valor de x_j ? E o valor esperado?

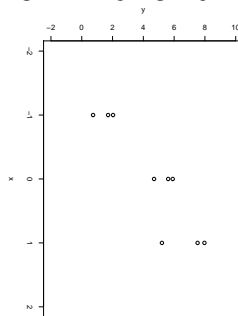


Figura 10.1: Gráfico dos valores observados y_{ij} versus x_j .

- Deseja-se um estimador para $E(Y|x_j = 0) = \alpha$ quando $x = 0$. Um primeiro estimador bem simples é proposto:

$$\frac{Y_{21} + Y_{22} + Y_{23}}{3}$$

Ele simplesmente toma a média das três observações quando $x = 0$. Mostre que este estimador é não viciado para α e encontre sua variância. Qual o risco quadrático desse estimador? OBS: Risco quadrático de um estimador é o seu MSE.

- Um segundo estimador é proposto:

$$\frac{Y_{11} + Y_{12} + Y_{13} + Y_{21} + Y_{22} + Y_{23} + Y_{31} + Y_{32} + Y_{33}}{9}$$

Ele toma a média aritmética simples de todas as 9 observações disponíveis. Mostre que este estimador também é não viciado para α e encontre sua variância.

- Qual dos dois estimadores é preferível?
- O interesse agora é em estimar β , o quanto Y aumenta em média quando passamos de um nível de x para o nível seguinte. Um primeiro estimador é o valor médio de Y quando $x = 0$ menos o valor médio de Y quando $x = -1$. Isto é,

$$T_1 = \bar{Y}_0 - \bar{Y}_{-1} = \frac{Y_{21} + Y_{22} + Y_{23}}{3} - \frac{Y_{11} + Y_{12} + Y_{13}}{3}$$

Mostre que T_1 é uma combinação linear $\sum_{ij} a_{ij} Y_{ij}$ dos Y 's e identifique os valores de a_{ij} .

- Mostre que $E(T_1)$ é não-viciado para β e ache sua variância.
- De maneira análoga, defina

$$T_1 = \bar{Y}_1 - \bar{Y}_0$$

e ache sua média e variância.

- Um terceiro estimador, melhor que os dois anteriores, leva em conta apenas as observações nos dois extremos, quando $x = -1$ e $x = 1$.

$$T_3 = \frac{1}{2} \bar{Y}_1 - \bar{Y}_{-1}$$

Mostre que T_3 também é uma combinação linear dos Y 's, que é não-viciado e ¹²⁵que possui risco quadrático (ou MSE) menor que T_1 e T_2 .

15. Uma operadora de planos de saúde sabe que o custo médio das internações varia muito de acordo com a idade do cliente. Aqueles com mais de 70 anos de idade acarretam a maior parte dos custos embora eles tenham uma participação pequena no portfolio de clientes.

A operadora decidiu investigar um pouco mais a incidência de internações entre seus clientes idosos. Para isto, escolheu uma amostra de clientes com idade acima de 70 anos e obteve o número de internações que cada um teve nos últimos dois anos. Decidiu-se adotar um modelo de Poisson para as contagens do número de internações.

Nem todos os selecionados foram clientes por todo o período de dois anos. Aqueles que estão na operadora há pouco tempo devem apresentar, em média, menos internações do que aqueles que estão na operadora durante os últimos dois anos. Por isto, a média da Poisson deveria refletir o tempo de permanência no plano de cada cliente. Dessa forma chegou-se ao seguinte modelo estatístico.

Sejam Y_1, \dots, Y_n a amostra de clientes. Suponha que essas sejam variáveis aleatórias independentes e que $Y_i \sim \text{Poisson}(\lambda t_i)$ onde t_i é o tempo de permanência do i -ésimo cliente na empresa (em meses) e $\lambda > 0$ é desconhecido e representa o número esperado de internações *por mês*. O interesse é estimar λ a partir dos dados que podem ser representados como na tabela abaixo:

i	t_i	y_i
1	24	4
2	12	1
3	3	0
4	24	1
...

- Pensou-se inicialmente em estimar λ simplesmente tomando o número médio de internações e dividir pelo tempo de observação de 24 meses. Isto é, $T_1 = \bar{Y}/24$. Mostre que este estimador é viciado para estimar λ a menos que $\sum_i t_i = 24n$. Por exemplo, se todos os clientes tiverem $t_i = 24$ esta condição seria válida.
- Tentando corrigir o vício do estimador T_1 , pensou-se então em adotar

$$T_2 = \frac{\bar{Y}}{\bar{t}} = \frac{Y_1 + \dots + Y_n}{t_1 + \dots + t_n}$$

Mostre que T_2 é não-viciado para estimar λ e encontre seu risco quadrático de estimação.

- Mais tarde, outro analista resolveu considerar o estimador

$$T_3 = \frac{1}{n} \left(\frac{Y_1}{t_1} + \dots + \frac{Y_n}{t_n} \right)$$

Mostre que T_3 é não-viciado para estimar λ e encontre seu risco quadrático de estimação.

- É possível dizer que T_2 é sempre melhor ou igual a T_3 considerando-se os testes quadráticos dos dois. Prove isto usando a desigualdade entre a média aritmética e a média harmônica que diz que

$$\frac{x_1 + \cdots + x_n}{n} \geq \frac{n}{\frac{1}{x_1} + \cdots + \frac{1}{x_n}}$$

para quaisquer números reais positivos x_1, \dots, x_n .

16. Suponha que X_1, \dots, X_n sejam n variáveis aleatórias com distribuição de Rayleigh com parâmetro $\theta > 0$ com densidade dada por

$$f(x; \theta) = \begin{cases} 0, & \text{se } x < 0 \\ x/\theta^2 \exp(-x^2/(2\theta^2)), & \text{se } x \geq 0 \end{cases}$$

Encontre o estimador de máxima verossimilhança de θ .

17. A distribuição logarítmica serve para modelar contagens em ecologia. Essa distribuição tem função de probabilidade dada por

$$P(X = x; \theta) = \frac{-\theta^x}{x \log(1 - \theta)}$$

para $x = 1, 2, \dots$ onde θ é um parâmetro desconhecido no intervalo $(0, 1)$. Mostre que, se X_1, X_2, X_3, X_4 é uma amostra aleatória da distribuição acima, a estimativa de máxima verossimilhança $\hat{\theta}$ satisfaz a equação

$$\hat{\theta} = \bar{X}(1 - \hat{\theta}) \log(1 - \hat{\theta})$$

onde \bar{X} é a média aritmética dos dados. Se $x_1 = 1$, $x_2 = 2$, $x_3 = 3$ e $x_4 = 2$, e se você tiver o valor inicial $\theta_0 = 0.6$, encontre o valor θ_1 do processo iterativo de Newton-Raphson (faça as contas).

OBS: Como $\theta \in (0, 1)$, temos $-\log(1 - \theta) > 0$, mas não faz sentido tomar $\log(-\theta)$.

18. Num estudo de seguro agrícola, acredita-se que a produção de trigo X_i da área i é normalmente distribuída com média θz_i , onde z_i é a quantidade *CONHECIDA* de fertilizante utilizado na área. Assumindo que as produções em diferentes áreas são independentes, e que a variância é conhecida e igual a 1 (ou seja, que $X_i \sim N(\theta z_i, 1)$, para $i = 1, \dots, n$:

- Encontre o EMV de θ
- Mostre que o EMV é não viciado para θ . Lembre-se que os valores de z_i são constantes.

19. Uma função de interesse quando trabalha-se com resseguros de perdas X com uma certa ¹²⁷distribuição é a função média do excesso definida como $e(a) = E(X - a | X > a)$. Isto é, $e(a)$ é o valor médio do excesso de X acima da constante a . Para cada valor a existe um valor de $e(a)$. Com base numa amostra X_1, \dots, X_n de v.a.'s i.i.d., sugerem-se dois possíveis estimadores para $e(a)$:

$$\hat{e}(a) = \frac{\sum_{i=1}^n X_i I(X_i > a)}{n} - a$$

e

$$\hat{e}(a) = \frac{\sum_{i=1}^n X_i I(X_i > a)}{\sum_{i=1}^n I(X_i > a)} - a$$

Um deles é muito ruim e foi proposto por alguém que não entendeu a definição de $e(a)$. Diga qual é e explique por quê.

20. Dados de perda geralmente tem caudas pesadas. Suponha que x_0 é um valor de franquia *conhecido* (isto é, são observadas apenas as perdas que têm valores monetários acima de x_0). Uma possibilidade sempre pensada é usar a distribuição de Pareto com função distribuição acumulada dada por:

$$F_X(x) = P(X \leq x) = \begin{cases} 0, & \text{se } x \leq x_0 \\ \int_0^x \alpha x_0^\alpha / y^{\alpha+1} dy = 1 - (x_0/x)^\alpha, & \text{se } x > x_0 \end{cases}$$

onde $\alpha > 0$.

- Mostre que a densidade de probabilidade de X é igual a $f(x) = \alpha x_0^\alpha / x^{\alpha+1}$.
- Se X_1, \dots, X_n é uma amostra de v.a.'s i.i.d., obtenha o EMV de α dado por

$$\hat{\alpha} = \frac{n}{\sum_i \log(x_i/x_0)} = \frac{1}{\log(\sqrt[n]{x_1 x_2 \dots x_n} / x_0)}$$

21. A distribuição de Gumbel é uma escolha popular para modelar dados de catástrofes naturais tais como enchentes. Temos dados da maior precipitação pluvial diária durante um ano para o período de 1956 a 2001. Suponha que os valores aleatórios X_1, \dots, X_{46} das maiores precipitações diárias anuais sigam uma distribuição de Gumbel com parâmetros μ e σ e com densidade dada por

$$f(x; \mu, \sigma) = \frac{1}{\sigma} \exp\left(-\frac{x - \mu}{\sigma}\right) \exp\left(-e^{\frac{x - \mu}{\sigma}}\right)$$

- Mostre que o EMV de μ e de σ são as soluções do sistema de equações simultâneas não-lineares

$$\begin{aligned} \hat{\mu} &= -\hat{\sigma} \log\left(\frac{1}{n} \sum_i e^{-x_i/\hat{\sigma}}\right) \\ \hat{\sigma} &= \frac{1}{n} \sum_i x_i - \frac{\sum_i x_i e^{-x_i/\hat{\sigma}}}{\sum_i e^{-x_i/\hat{\sigma}}} \end{aligned}$$

- Estimativas para iniciar o procedimento numérico e obter o EMV são as estimativas de momentos (não vimos isto no nosso curso) e que são dadas por

$$\hat{\mu} = \bar{x} - 0.45s$$

$$\hat{\sigma} = s/1.283$$

onde s é o desvio-padrão amostral. Explique como você faria para obter as estimativas de máxima verossimilhança usando estas estimativas iniciais. Monte a equação recursiva necessária para o procedimento numérico.

22. Suponha que X_1, \dots, X_n forme uma amostra aleatória de v.a.'s i.i.d. com distribuição Poisson com esperança θ desconhecida.

- Encontre a função escore $\frac{\partial l}{\partial \theta}$.
- Considerando a função escore como uma variável aleatória, calcule o seu valor esperado $\mathbb{E}(\frac{\partial l}{\partial \theta})$.
- Calcule também a sua variância.
- Calcule a informação de Fisher $I(\theta)$ de duas formas distintas: $I(\theta) = \mathbb{E}(\frac{\partial l}{\partial \theta})^2$ e como $I(\theta) = -\mathbb{E}(\frac{\partial^2 l}{\partial \theta^2})$.

23. Repita o exercício acima supondo que X_1, \dots, X_n sejam iid com distribuição $\exp(\lambda)$.

24. Repita o exercício acima supondo que X_1, \dots, X_n sejam iid com distribuição Pareto, com densidade $f(x; \theta) = \theta c^\theta x^{-(\theta+1)}$ para $x \geq c$ com $\theta \in (0, \infty)$. Esta distribuição é muito usada para modelar dados de resseguro, quando as perdas podem chegar a valores muito grandes.

25. Suponha que X_1, \dots, X_5 forme uma amostra aleatória de 23 v.a.'s i.i.d. com distribuição Poisson com esperança $\theta = 1.2$. No R, o comando `rpois(5, 1.2)` gerou a amostra $\mathbf{x} = (1, 2, 1, 2, 1)$. Suponha que você não conhecesse este valor verdadeiro de θ .

- Faça um gráfico da função de log-verossimilhança $\log f(\mathbf{x}, \theta) = \log f((1, 2, 1, 2, 1), \theta)$ versus θ . Use um intervalo para θ grande o suficiente para cobrir o verdadeiro valor do parâmetro.
- Faça também um gráfico da função escore $\frac{\partial}{\partial \theta} \log f(\mathbf{x}, \theta) = \frac{\partial}{\partial \theta} \log f((1, 2, 1, 2, 1), \theta)$ versus θ .
- O EMV neste caso é a média aritmética. Portanto, o valor do EMV observado nesta amostra particular é igual a $\hat{\theta} = 7/5$. Verifique graficamente que $\hat{\theta} = 7/5$ é ponto de máximo da função de log-verossimilhança.

- Verifique graficamente que o verdadeiro valor do parâmetro $\theta = 1.2$ **não** maximiza a função de log-verossimilhança.
 - Verifique graficamente que $\hat{\theta} = 7/5$ é o zero da função escore.
 - Verifique também que a função escore não se anula no ponto $\theta = 1.2$.
-

Capítulo 11

Método de Máxima Verossimilhança

1. Suponha que X_1, X_2, \dots, X_n sejam v.a.'s i.i.d. com distribuição Poisson com parâmetro θ . Obtenha o MLE de θ . Supondo que $n = 4$ e que as observações tenham sido 3, 0, 1, 1 diga qual o valor do MLE.
2. No Brasil, o Ministério da Previdência Social assegura um benefício de auxílio-doença ao trabalhador que paga pelos dias em que ele foi impedido de trabalhar por doença ou acidente. Como parte dos pagamentos é efetuado pelas empresas, um sindicato patronal de empresas de ônibus urbano de certa cidade contratou uma consultoria atuarial para analisar o número de faltas ao trabalho por doença. Dados de 20 empresas de ônibus urbano foram coletados.

O número de acidentes na empresa i é denotado por Y_i e depende do número de homem-hora trabalhado. Este número de homem-hora é representado por h_i e é medido em unidades de 10 mil horas. Assim, $h_i = 3.5$ representa 35 mil homens-horas. Quanto maior h_i , maior tende a ser o número de dias parados pois a exposição ao risco é maior na empresa com h_i maior.

Para cada empresa defina o parâmetro λ_i que representa o número médio ou esperado de acidentes na empresa i por cada 10 mil homens-hora. Veja que 12 homens trabalhando 40 horas por semana ao longo de $21 = 30 - 9$ dias por mês dá um total de $12 \times 40 \times 21 = 10080$. Assim, grosseiramente, 10 mil homens/horas representam 12 homens trabalhando em tempo integral num mês.

O modelo estatístico para os dados é que Y_1, \dots, Y_{20} são independentes. Entretanto, essas variáveis aleatórias não são identicamente distribuídas pois os valores de h_1, \dots, h_{20} são diferentes. Assim, supõe-se que

$$Y_i \sim \text{Poisson}(\lambda_i) \text{ ou } \text{Poisson}(h_i\theta)$$

para $i = 1, \dots, 20$. Os dados são do seguinte tipo:

Empresa	h_i	y_i
1	94.5	5
2	15.7	1
3	62.9	5
...

O interesse é em fazer inferência sobre o valor desconhecido θ , a taxa de ausência por doença nas empresas. Em particular, para uma nova empresa que vai fazer seguro-doença para cobrir faltas de empregados, é preciso saber qual a expectativa de faltas durante um mês se ela tem um certo número h de homens-hora de trabalho.

Calcule:

- o EMV de θ .
- O EMV é não-viciado?
- Calcule o número de informação de Fisher $I(\theta)$.
- É verdade que a informação $I_n(\theta)$ com n observações é igual a n vezes a informação com uma única informação? Como a informação aumenta com n ?
- Suponha que os dados sejam esses abaixo:

```
h <- c(118.3, 13.4, 68.3, 141.6, 113.1, 63.6, 135.5, 107.5, 35.2,
      28.1, 34.7, 42.5, 139.7, 140.4, 79.2, 148.2, 28, 72.4, 50.9, 89.1)
y <- c(8, 0, 6, 18, 7, 6, 13, 4, 7, 4, 0, 3, 26, 10, 4, 15, 7, 10,
      9, 14)
```

Usando um diagrama de dispersão, verifique se existe associação entre h e y . Estime θ usando o MLE.

3. No problema acima, as empresas podem ser classificadas em quatro grupos distintos e esperamos que o valor de θ seja diferente nesses quatro grupos. As empresas são classificadas em 4 grupos de acordo com as seguintes categorias:

- $Z_1 = 0$ se a empresa faz cursos de treinamento e regulares sobre cuidados ao dirigir, e $Z_1 = 1$, caso contrário.
- $Z_2 = 0$ se a empresa cobra de cada motorista responsável por um acidente parte das perdas causadas (o valor cobrado é limitado a um máximo) e $Z_2 = 1$, caso contrário

No caso de Z_2 , é óbvio que se o motorista envolvido num acidente não é considerado culpado do mesmo, a cobrança não é feita. A cobrança também é limitada a um máximo de 2 salário mensais. Em cada mês, até um máximo de 15% do salário é descontado da folha de pagamentos.

Espera-se que empresas com $Z_1 = 1$ e $Z_2 = 1$ tenham θ maiores que aquelas com $Z_1 = 0$ e $Z_2 = 0$. Empresas com $Z_1 = 0$ e $Z_2 = 1$ ou $Z_1 = 1$ e $Z_2 = 0$ seriam casos intermediários.

é claro que, mesmo dentro de um dos grupos ($Z_1 = 1$ e $Z_2 = 1$, por exmplo) ainda esperamos que as empresas não sejam idênticas e portanto que seus θ 's possam diferir um

pouco. Esperamos que, dentre as 20 empresas, boa parte das diferenças entre seus θ 's possa ser devido às políticas de prevenção e de punição. Ainda sobraria um resíduo de causas não controladas tornando os θ 's dentro de cada um dos 4 grupos ligeiramente diferentes. No entanto, nós vamos ignorar estas diferenças residuais e supor que, empresa numa mesma categoria, tenham θ 's idênticos.

Queremos agora estimar o impacto dessas políticas de prevenção/punição na redução do número de faltas. Existe algum impacto? Qual política é mais eficaz? Como estudar isto? Uma possibilidade é fazer uma análise separada para cada um dos 4 grupos possíveis. Uma *saída muito melhor* é fazer um único modelo em que todos os quatro grupos estejam presentes e com parâmetros medindo o impacto dos programas de prevenção/punição. Para isto, precisamos alterar o modelo do exercício anterior. Vamos assumir que

$$Y_i \sim \text{Poisson}(h_i \theta_i)$$

para $i = 1, \dots, 20$ que elas sejam v.a.'s independentes.

Existem quatro valores possíveis para θ_i , dependendo da categoria à qual a empresa i pertence. Vamos supor que, ao passar de $Z_1 = 0$ para $Z_1 = 1$ o valor de θ seja aumentado por um fator multiplicativo. O mesmo ao passar de $Z_2 = 0$ para $Z_2 = 1$. Isto é, supomos que

$$\theta_i = e^{\beta_0} e^{\beta_1 Z_{1i}} e^{\beta_2 Z_{2i}}$$

onde Z_{1i} e Z_{2i} são os valores das variáveis independentes para a empresa i . Os dados de h_i e y_i são os mesmos de antes e os de Z_1 e Z_2 são os seguintes:

```
Z1 <- c(0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1),
Z2 <- c(0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1),
```

- Descreva o modelo estatístico.
- Interprete os parâmetros do modelo. Para isto, verifique qual o efeito em $E(Y_i)$ de uma empresa mudar de uma categoria para outra categoria. Compare o valor de θ_i e de $\log(\theta_i)$ para uma empresa que tenha $Z_1 = 0$ e $Z_2 = 0$ e outra empresa com $Z_1 = 1$ e $Z_2 = 0$. E com uma empresa com $Z_1 = 1$ e $Z_2 = 1$. Faça comparações entre todos os 6 tipos de *pares* de empresas possíveis.
- Obtenha o EMV e a MATRIZ 3×3 de informação de Fisher para os parâmetros do modelo.
- Obtenha o EMV e a MATRIZ 3×3 de informação de Fisher para os parâmetros $e^{\beta_0}, e^{\beta_1}, e^{\beta_2}$.

-
4. Considere um modelo estatístico para dados de sobrevivência que supõe que X_1, X_2, \dots, X_n são i.i.d. com distribuição exponencial parametrizada por λ . Suponha que, com $n = 10$, os dados observados sejam os seguintes: 1.21, 0.15, 2.02, 0.37, 2.55, 2.20, 1.06, 0.10, 0.35, e 0.15. Use o conjunto de comandos R dados abaixo para desenhar a função de verossimilhança para o parâmetro λ . Procure entender o que cada comando está fazendo.

134 `dados <- c(1.21,0.15,2.62,0.37,2.55,2.20,1.06,0.40,0.35,0.18)`
`lambda <- seq(0,3,length=300)`
`veross <- (lambda^10) * exp(-lambda * sum(dados))`
`plot(lambda,veross,type="l")`

Os comandos abaixo sobrepõem uma segunda curva de verossimilhança baseada numa possível segunda amostra independente da primeira e gerada pelo mesmo mecanismo.

```
dados2 <- c(0.08,2.72,0.04 0.27 0.99 0.25 0.39 1.84 1.60 3.51
veross2 <- (lambda^10) * exp(-lambda * sum(dados2))
lines(lambda,veross2,lty=2)
```

Construa agora uma função que recebe os dados da amostra como entrada e que retorna o gráfico da verossimilhança para $\lambda \in (a, b)$.

```
veross.exp <-function(x, a, b)
{
# funcao para desenhar a funcao de verossimilhanca
# para dados exponenciais
# INPUT: x = dados de entrada
ene <- length(x)
lambda <- seq(a,b,length=300)
veross <- (lambda^ene) * exp(-lambda * sum(x))
plot(lambda, veross, type= "l",ylab= "verossimilhanca")
title("Funcao de verossimilhanca para dados exponenciais")
return()
}
```

Gere um conjunto de dados de uma exponencial com um parâmetro λ num intervalo (a, b) a sua escolha. Chame a função acima com os dados que você gerou e verifique se o máximo da função de verossimilhança fica próximo do valor λ que você escolheu. Repita o exercício escolhendo diferentes tamanhos n de amostra.

-
5. Algoritmo EM: Leia o exemplo de misturas gaussianas (MoG ou Mixture of Gaussians, em machine learning) no final da página da Wikipedia: http://en.wikipedia.org/wiki/Expectation-maximization_algorithm. Usando os dados do Old Faithful dataset (pegue os dados, por exemplo, em <http://tinyurl.com/cndlsp3>), estime os parâmetros de uma mistura de duas gaussianas bivariadas usando o algoritmo EM.
-

6. Suponha que X_1, \dots, X_n forme uma amostra aleatória de v.a.'s i.i.d. com uma das seguintes densidades (caso contínuo) ou função de probabilidade (caso discreto). Encontre o EMV de θ e verifique se ele é função da estatística suficiente para estimar θ em cada caso.

- $f(x; \theta) = \theta^x e^{-\theta} / x!$ para $x = 0, 1, 2, \dots$ e com $\theta \in (0, \infty)$ (função de probabilidade Poisson).
- $f(x; \theta) = \theta e^{-\theta x}$ para $x > 0$ com $\theta \in (0, \infty)$ (densidade exponencial).
- $f(x; \theta) = \theta c^\theta x^{-(\theta+1)}$ para $x \geq c$ com $\theta \in (0, \infty)$ (densidade Pareto).
- $f(x; \theta) = \sqrt{\theta} x^{-\sqrt{\theta}-1}$ para $0 \leq x \leq 1$ com $\theta \in (0, \infty)$ (densidade beta($\sqrt{\theta}, 1$)).
- $f(x; \theta) = (x/\theta)^2 \exp(-x^2/(2\theta^2))$ para $x > 0$ com $\theta \in (0, \infty)$ (densidade Rayleigh).
- $f(x; \theta) = \theta c x^{c-1} \exp(-\theta x^c)$ para $x \geq 0$ com $\theta \in (0, \infty)$ e $c > 0$ sendo uma constante conhecida (densidade Weibull com c fixo).
- $f(x; \theta) = \theta(1 - \theta)^{x-1}$ para $x = 1, 2, \dots$ e com $\theta \in (0, \infty)$ (função de probabilidade geométrica).
- $f(x; \theta) = \theta^2 x e^{-\theta x}$ para $x > 0$ com $\theta \in (0, \infty)$ (densidade Gama(2, θ)).

7. Sejam X_1 e X_2 duas v.a.'s independentes com esperança comum $\theta \in \mathbb{R}$ e com $\text{Var}(X_1) = \sigma^2$ e $\text{Var}(X_2) = \sigma^2/4$. Isto é, X_1 e X_2 tendem a oscilar em torno de θ mas X_2 possui um desvio-padrão duas vezes menor que X_1 . Podemos usar estes dois pedaços de informação para estimar θ . Podemos, por exemplo, formar uma combinação linear de X_1 e X_2 propondo o estimador $\hat{\theta} = c_1 X_1 + c_2 X_2$ onde c_1 e c_2 são constantes conhecidas. Por exemplo, podemos pensar em usar $\hat{\theta} = (X_1 + X_2)/2$ ou então usar $\hat{\theta} = \frac{2X_1}{3} + \frac{X_2}{3}$ ou até mesmo $\hat{\theta} = 4X_1 - 2X_2$.

Mostre que $\hat{\theta} = c_1 X_1 + c_2 X_2$ é não-viciado para estimar θ (qualquer que seja o valor de $\theta \in \mathbb{R}$) se, e somente se, $c_1 + c_2 = 1$. Dentre os estimadores da forma $\hat{\theta} = c_1 X_1 + c_2 X_2$ e que são não-viciados para estimar θ , encontre aquele que minimiza o MSE, dado por $\mathbb{E}(\hat{\theta} - \theta)^2$.

8. Generalize o problema anterior para n v.a.'s: Sejam X_1, \dots, X_n v.a.'s independentes com esperança comum θ e com a variância de X_i igual a σ^2/a_i , sendo os $a_i > 0$ conhecidos e com $\sigma^2 > 0$ desconhecido. Considere a classe de todos os estimadores lineares de θ . Isto é, considere a classe de todos os estimadores que podem ser escritos como $\hat{\theta} = \sum_i c_i X_i$ onde c_i são constantes.

Mostre que na classe dos estimadores lineares, $\hat{\theta}$, é não-viciado para estimar θ se, e somente se $\sum_i c_i = 1$. Dentre todos os estimadores lineares $\sum_i c_i X_i$ de θ que são não-viciados (isto é, satisfazendo $\sum_i c_i = 1$), encontre aquele que minimiza o MSE.

9. Suponha que X_1, X_2, \dots, X_n sejam v.a.'s i.i.d. com distribuição Poisson com parâmetro comum θ . É possível mostrar matematicamente que

$$\hat{\theta}_1 = \bar{X} = \frac{X_1 + \dots + X_n}{n}$$

$$\hat{\theta}_2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

são ambos estimadores não viciados para estimar θ . Considere adicionalmente um terceiro estimador não-viciado para θ :

$$\hat{\theta}_3 = (\hat{\theta}_1 + \hat{\theta}_2)/2$$

Faça um pequeno estudo de simulação para identificar qual dos três possui um erro de estimação MSE menor. Para isto, fixe o valor de $\theta = 3$. Gere um grande número de amostras (digamos, 10 mil), cada uma delas de tamanho $n = 10$. Para cada amostra calcule os valores dos três estimadores de θ . Estime o MSE $\mathbb{E}(\hat{\theta}_j - \theta)^2$ de cada estimador usando a média das diferenças ao quadrado entre os 10 mil valores de $\hat{\theta}_j$ e θ . Qual dos estimadores produz um erro MSE menor? Isto significa que o melhor estimador teve SEMPRE o seu valor mais próximo do verdadeiro valor do parâmetro θ ? Estime a probabilidade de que, baseados numa mesma amostra, $\hat{\theta}_2$ esteja mais próximo de θ que $\hat{\theta}_1$.

A conclusão muda se você tomar $n = 20$ e $\theta = 10$?

10. Responda V ou F para as afirmações abaixo.

- Como o parâmetro θ não pode ser predito antes do experimento, ele é uma variável aleatória.
- Num problema de estimação de uma população com distribuição normal $N(\mu, \sigma^2)$ encontrou-se $\bar{x} = 11.3$ numa amostra de tamanho $n = 10$. A distribuição de probabilidade desse valor 11.3 é também uma normal com média μ e variância $\sigma^2/10$.
- Suponha que \bar{X} esteja sendo usado como estimador da média populacional μ . Como a variância de \bar{X} decresce com o tamanho da amostra, então toda estimativa obtida a partir de uma amostra de tamanho 15 possui erro de estimação menor que qualquer outra estimativa obtida a partir de uma amostra de tamanho 10.
- Um estimador não viciado é sempre melhor que um estimador viciado.
- Considere uma estimativa da média populacional μ baseada na média aritmética de uma amostra de tamanho 10 e outra estimativa com uma amostra de tamanho 15. Nunca devemos preferir a estimativa baseada na amostra de 15 pois a estimativa baseada na amostra de tamanho 10 tem alguma chance de estar mais perto do verdadeiro valor desconhecido de μ .

11. Sejam X_1, \dots, X_n i.i.d.'s com distribuição exponencial com parâmetro λ . O interesse é estimar $E(X_i) = 1/\lambda$. Suponha que apenas as variáveis X_i 's que ficarem maiores ou iguais a $x = 10$ sejam observadas. Todas as observações menores que $x = 10$ são perdidas. Assim, a amostra final é possui um número $0 < k \leq n$ de observações.

O estimador baseado na média amostral da amostra de k variáveis é viciado. Ele subestima ou superestima sistematicamente $E(X_i)$? Não precisa calcular o vício.

A distribuição de X_i DADO QUE $X_i > 10$ tem a densidade dada por

$$f(x; \lambda) = \begin{cases} 0, & \text{se } x < 10 \\ \lambda \exp(-\lambda(x - 10)), & \text{se } x \geq 10 \end{cases}$$

Se X_1, \dots, X_k é uma amostra desta distribuição TRUNCADA (em que só observamos os X_i 's maiores que 10), encontre o MLE de λ .

RESP: O MLE é $k / \sum_i (x_i - 10)$.

12. Suponha que será coletada uma amostra de observações independentes Y com distribuição normal. Elas *não são* identicamente distribuídas. A média de Y varia de acordo com o valor de uma covariável x de forma que $Y = \alpha + \beta x + \epsilon$, onde ϵ possui distribuição normal com média 0 e variância σ^2 . Os valores possíveis de x são três: baixo ($x = -1$), médio ($x = 0$) e alto ($x = 1$). Os valores de x são fixos e conhecidos. Eles não são variáveis aleatórias.

São feitas três observações em cada nível de x . Podemos representar os dados na tabela e no gráfico dos valores observados de Y versus x que Figura 11.1.

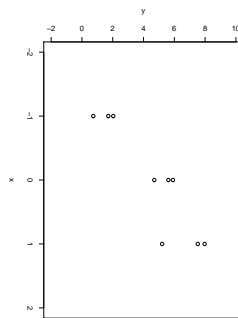


Figura 11.1: Gráfico dos valores observados y_{ij} versus x_j .

$x_j = -1$	$x_j = 0$	$x_j = 1$
Y_{11}	Y_{21}	Y_{31}
Y_{12}	Y_{22}	Y_{32}
Y_{13}	Y_{23}	Y_{33}

Vamos representar as observações como $Y_{ij} = \alpha + \beta x_j + \epsilon_{ij}$ onde $x_j = -1, 0$ ou 1 , e os ϵ_{ij} são i.i.d. com distribuição $N(0, \sigma^2)$.

- é correto dizer que $(Y_{ij}|x_j) \sim N(\alpha + \beta x_j, \sigma^2)$ e que as variáveis Y_{ij} são independentes? DICA: não existe pegadinha aqui.
- Calcule $E(Y_{ij}|x_j)$ e $\text{Var}(Y_{ij}|x_j)$ nos três casos: $x_j = -1$, $x_j = 0$ e $x_j = 1$. A variância depende do valor de x_j ? E o valor esperado?

- Deseja-se um estimador para $E(Y|x) = \alpha$ quando $x = 0$. Um primeiro estimador bem simples é proposto:

$$\frac{Y_{21} + Y_{22} + Y_{23}}{3}$$

Ele simplesmente toma a média das três observações quando $x = 0$. Mostre que este estimador é não viciado para α e encontre sua variância. Qual o risco quadrático desse estimador? OBS: Risco quadrático de um estimador é o seu MSE.

- Um segundo estimador é proposto:

$$\frac{Y_{11} + Y_{12} + Y_{13} + Y_{21} + Y_{22} + Y_{23} + Y_{31} + Y_{32} + Y_{33}}{9}$$

Ele toma a média aritmética simples de todas as 9 observações disponíveis. Mostre que este estimador também é não viciado para α e encontre sua variância.

- Qual dos dois estimadores é preferível?
- O interesse agora é em estimar β , o quanto Y aumenta em média quando passamos de um nível de x para o nível seguinte. Um primeiro estimador é o valor médio de Y quando $x = 0$ menos o valor médio de Y quando $x = -1$. Isto é,

$$T_1 = \bar{Y}_0 - \bar{Y}_{-1} = \frac{Y_{21} + Y_{22} + Y_{23}}{3} - \frac{Y_{11} + Y_{12} + Y_{13}}{3}$$

Mostre que T_1 é uma combinação linear $\sum_{ij} a_{ij} Y_{ij}$ dos Y 's e identifique os valores de a_{ij} .

- Mostre que $E(T_1)$ é não-viciado para β e ache sua variância.
- De maneira análoga, defina

$$T_1 = \bar{Y}_1 - \bar{Y}_0$$

e ache sua média e variância.

- Um terceiro estimador, melhor que os dois anteriores, leva em conta apenas as observações nos dois extremos, quando $x = -1$ e $x = 1$.

$$T_3 = \frac{1}{2} \bar{Y}_1 - \bar{Y}_{-1}$$

Mostre que T_3 também é uma combinação linear dos Y 's, que é não-viciado e que possui risco quadrático (ou MSE) menor que T_1 e T_2 .

13. Uma operadora de planos de saúde sabe que o custo médio das internações varia muito de acordo com a idade do cliente. Aqueles com mais de 70 anos de idade acarretam a maior parte dos custos embora eles tenham uma participação pequena no portfolio de clientes. A operadora decidiu investigar um pouco mais a incidência de internações entre seus clientes idosos. Para isto, escolheu uma amostra de clientes com idade acima de 70 anos e obteve o número de internações que cada um teve nos últimos dois anos. Decidiu-se adotar um modelo de Poisson para as contagens do número de internações.

Nem todos os selecionados foram clientes por todo o período de dois anos. Aqueles que estão na operadora há pouco tempo devem apresentar, em média, menos interações do que aqueles que estão na operadora durante os últimos dois anos. Por isto, a média da Poisson deveria refletir o tempo de permanência no plano de cada cliente. Dessa forma chegou-se ao seguinte modelo estatístico.

Sejam Y_1, \dots, Y_n a amostra de clientes. Suponha que essas sejam variáveis aleatórias independentes e que $Y_i \sim \text{Poisson}(\lambda t_i)$ onde t_i é o tempo de permanência do i -ésimo cliente na empresa (em meses) e $\lambda > 0$ é desconhecido e representa o número esperado de interações por mês. O interesse é estimar λ a partir dos dados que podem ser representados como na tabela abaixo:

i	t_i	y_i
1	24	4
2	12	1
3	3	0
4	24	1
...

- Pensou-se inicialmente em estimar λ simplesmente tomando o número médio de interações e dividir pelo tempo de observação de 24 meses. Isto é, $T_1 = \bar{Y}/24$. Mostre que este estimador é viciado para estimar λ a menos que $\sum_i t_i = 24n$. Por exemplo, se todos os clientes tiverem $t_i = 24$ esta condição seria válida.
- Tentando corrigir o vício do estimador T_1 , pensou-se então em adotar

$$T_2 = \frac{\bar{Y}}{\bar{t}} = \frac{Y_1 + \dots + Y_n}{t_1 + \dots + t_n}$$

Mostre que T_2 é não-viciado para estimar λ e encontre seu risco quadrático de estimação.

- Mais tarde, outro analista resolveu considerar o estimador

$$T_3 = \frac{1}{n} \left(\frac{Y_1}{t_1} + \dots + \frac{Y_n}{t_n} \right)$$

Mostre que T_3 é não-viciado para estimar λ e encontre seu risco quadrático de estimação.

- É possível dizer que T_2 é sempre melhor ou igual a T_3 considerando-se os riscos quadráticos dos dois. Prove isto usando a desigualdade entre a média aritmética e a média harmônica que diz que

$$\frac{x_1 + \dots + x_n}{n} \geq \frac{n}{\frac{1}{x_1} + \dots + \frac{1}{x_n}}$$

para quaisquer números reais positivos x_1, \dots, x_n .

14. Suponha que X_1, \dots, X_n sejam n variáveis aleatórias com distribuição de Rayleigh com parâmetro $\theta > 0$ com densidade dada por

$$f(x; \theta) = \begin{cases} 0, & \text{se } x < 0 \\ x/\theta^2 \exp(-x^2/(2\theta^2)), & \text{se } x \geq 0 \end{cases}$$

Encontre o estimador de máxima verossimilhança de θ .

15. A distribuição logarítmica serve para modelar contagens em ecologia. Essa distribuição tem função de probabilidade dada por

$$P(X = x; \theta) = \frac{-\theta^x}{x \log(1 - \theta)}$$

para $x = 1, 2, \dots$ onde θ é um parâmetro desconhecido no intervalo $(0, 1)$. Mostre que, se X_1, X_2, X_3, X_4 é uma amostra aleatória da distribuição acima, a estimativa de máxima verossimilhança $\hat{\theta}$ satisfaz a equação

$$\hat{\theta} = \bar{X}(1 - \hat{\theta}) \log(1 - \hat{\theta})$$

onde \bar{X} é a média aritmética dos dados. Se $x_1 = 1, x_2 = 2, x_3 = 3$ e $x_4 = 2$, e se você tiver o valor inicial $\theta_0 = 0.6$, encontre o valor θ_1 do processo iterativo de Newton-Raphson (faça as contas).

OBS: Como $\theta \in (0, 1)$, temos $-\log(1 - \theta) > 0$, mas não faz sentido tomar $\log(-\theta)$.

16. Num estudo de seguro agrícola, acredita-se que a produção de trigo X_i da área i é normalmente distribuída com média θz_i , onde z_i é a quantidade *CONHECIDA* de fertilizante utilizado na área. Assumindo que as produções em diferentes áreas são independentes, e que a variância é conhecida e igual a 1 (ou seja, que $X_i \sim N(\theta z_i, 1)$, para $i = 1, \dots, n$:
- Encontre o EMV de θ
 - Mostre que o EMV é não viciado para θ . Lembre-se que os valores de z_i são constantes.
-

17. Uma função de interesse quando trabalha-se com resseguros de perdas X com uma certa distribuição é a função média do excesso definida como $e(a) = E(X - a | X > a)$. Isto é, $e(a)$ é o valor médio do excesso de X acima da constante a . Para cada valor a existe um valor de $e(a)$. Com base numa amostra X_1, \dots, X_n de v.a.'s i.i.d., sugerem-se dois possíveis estimadores para $e(a)$:

$$\hat{e}(a) = \frac{\sum_{i=1}^n X_i I(X_i > a)}{n} - a$$

e

$$\hat{e}(a) = \frac{\sum_{i=1}^n X_i I(X_i > a)}{\sum_{i=1}^n I(X_i > a)} - a$$

Um deles é muito ruim e foi proposto por alguém que não entendeu a definição de $e^{141/d}$. Diga qual é e explique por quê.

18. Dados de perda geralmente tem caudas pesadas. Suponha que x_0 é um valor de franquia conhecido (isto é, são observadas apenas as perdas que têm valores monetários acima de x_0). Uma possibilidade sempre pensada é usar a distribuição de Pareto com função distribuição acumulada dada por:

$$F_X(x) = P(X \leq x) = \begin{cases} 0, & \text{se } x \leq x_0 \\ \int_0^x \alpha x_0^\alpha / y^{\alpha+1} dy = 1 - (x_0/x)^\alpha, & \text{se } x > x_0 \end{cases}$$

onde $\alpha > 0$.

- Mostre que a densidade de probabilidade de X é igual a $f(x) = \alpha x_0^\alpha / x^{\alpha+1}$.
- Se X_1, \dots, X_n é uma amostra de v.a.'s i.i.d., obtenha o EMV de α dado por

$$\hat{\alpha} = \frac{n}{\sum_i \log(x_i/x_0)} = \frac{1}{\log(\sqrt[n]{x_1 x_2 \dots x_n} / x_0)}$$

19. A distribuição de Gumbel é uma escolha popular para modelar dados de catástrofes naturais tais como enchentes. Temos dados da maior precipitação pluvial diária durante um ano para o período de 1956 a 2001. Suponha que os valores aleatórios X_1, \dots, X_{46} das maiores precipitações diárias anuais sigam uma distribuição de Gumbel com parâmetros μ e σ e com densidade dada por

$$f(x; \mu, \sigma) = \frac{1}{\sigma} \exp\left(-\frac{x-\mu}{\sigma}\right) \exp\left(-e^{\frac{x-\mu}{\sigma}}\right)$$

- Mostre que o EMV de μ e de σ são as soluções do sistema de equações simultâneas não-lineares

$$\begin{aligned} \hat{\mu} &= -\hat{\sigma} \log\left(\frac{1}{n} \sum_i e^{-x_i/\hat{\sigma}}\right) \\ \hat{\sigma} &= \frac{1}{n} \sum_i x_i - \frac{\sum_i x_i e^{-x_i/\hat{\sigma}}}{\sum_i e^{-x_i/\hat{\sigma}}} \end{aligned}$$

- Estimativas para iniciar o procedimento numérico e obter o EMV são as estimativas de momentos (não vimos isto no nosso curso) e que são dadas por

$$\begin{aligned} \hat{\mu} &= \bar{x} - 0.45s \\ \hat{\sigma} &= s/1.283 \end{aligned}$$

onde s é o desvio-padrão amostral. Explique como você faria para obter as estimativas de máxima verossimilhança usando estas estimativas iniciais. Monte a equação recursiva necessária para o procedimento numérico.

20. Suponha que X_1, \dots, X_n forme uma amostra aleatória de v.a.'s i.i.d. com distribuição Poisson com esperança θ desconhecida.

- Encontre a função escore $\frac{\partial l}{\partial \theta}$.
 - Considerando a função escore como uma variável aleatória, calcule o seu valor esperado $\mathbb{E}(\frac{\partial l}{\partial \theta})$.
 - Calcule também a sua variância.
 - Calcule a informação de Fisher $I(\theta)$ de duas formas distintas: $I(\theta) = \mathbb{E}(\frac{\partial l}{\partial \theta})^2$ e como $I(\theta) = -\mathbb{E}(\frac{\partial^2 l}{\partial \theta^2})$.
-

21. Repita o exercício acima supondo que X_1, \dots, X_n sejam iid com distribuição $\exp(\lambda)$.

22. Repita o exercício acima supondo que X_1, \dots, X_n sejam iid com distribuição Pareto, com densidade $f(x; \theta) = \theta c^\theta x^{-(\theta+1)}$ para $x \geq c$ com $\theta \in (0, \infty)$. Esta distribuição é muito usada para modelar dados de resseguro, quando as perdas podem chegar a valores muito grandes.

23. Suponha que X_1, \dots, X_5 forme uma amostra aleatória de 23 v.a.'s i.i.d. com distribuição Poisson com esperança $\theta = 1.2$. No R, o comando `rpois(5, 1.2)` gerou a amostra $\mathbf{x} = (1, 2, 1, 2, 1)$. Suponha que você não conhecesse este valor verdadeiro de θ .

- Faça um gráfico da função de log-verossimilhança $\log f(\mathbf{x}, \theta) = \log f((1, 2, 1, 2, 1), \theta)$ versus θ . Use um intervalo para θ grande o suficiente para cobrir o verdadeiro valor do parâmetro.
 - Faça também um gráfico da função escore $\frac{\partial}{\partial \theta} \log f(\mathbf{x}, \theta) = \frac{\partial}{\partial \theta} \log f((1, 2, 1, 2, 1), \theta)$ versus θ .
 - O EMV neste caso é a média aritmética. Portanto, o valor do EMV observado nesta amostra particular é igual a $\hat{\theta} = 7/5$. Verifique graficamente que $\hat{\theta} = 7/5$ é ponto de máximo da função de log-verossimilhança.
 - Verifique graficamente que o verdadeiro valor do parâmetro $\theta = 1.2$ **não** maximiza a função de log-verossimilhança.
 - Verifique graficamente que $\hat{\theta} = 7/5$ é o zero da função escore.
 - Verifique também que a função escore não se anula no ponto $\theta = 1.2$.
-