INFERÊNCIA ESTATÍSTICA

RICARDO S. EHLERS

Primeira publicação em 2003
Segunda edição publicada em 2006
Terceira edição publicada em 2009

Sumário

Capítulo 1

Introdução

Inferência estatística é o processo pelo qual podemos tirar conclusões acerca de um conjunto maior (a *população*) usando informação de um conjunto menor (a *amostra*). Em Estatística, o termo população não se refere necessariamente a pessoas, plantas, animais, etc. Ele poderia também se referir, por exemplo, a fósseis, rochas e sedimentos num determinado local, itens produzidos em uma linha de montagem, etc.

A *população* se refere a todos os casos ou situações sobre as quais o pesquisador quer fazer inferências. Diferentes pesquisadores podem querer fazer inferências acerca da concentração de poluentes num determinado lençol freático; predizer a quantidade de petróleo num poço a ser perfurado e assim por diante.

Note que o investigador não está interessado em todos os aspectos da população. O pesquisador pode não estar interessado em estudar a concentração de todos os tipos de poluentes, somente alguns poluentes mais importantes para seu estudo.

Uma *amostra* é um subconjunto qualquer da população usado para obter informação acerca do todo. Algumas razões para se tomar uma amostra ao invés de usar a população toda são as seguintes,

- custo alto para obter informação da população toda,
- tempo muito longo para obter informação da população toda,
- algumas vezes impossível, por exemplo, estudo de poluição atmosférica
- algumas vezes logicamente impossível, por exemplo, em ensaios destrutivos.

Uma definição mais formal de amostra é dada a seguir.

Definição 1.1 *Sejam as variáveis aleatórias* $\boldsymbol{X} = (X_1, \ldots, X_n)$ *com função de (densidade) de probabilidade conjunta* $f(\boldsymbol{x})$ *fatorando nas densidades marginais como*

$$f(\boldsymbol{x}) = f(x_1)f(x_2)\ldots f(x_n)$$

sendo $f(\cdot)$ a densidade comum de todos $X_i's$. Então $X_1, \ldots, X_n$ é definida como uma amostra aleatória de tamanho n da população com densidade $f(\cdot)$.

Características de uma população que diferem de um indivíduo para outro e aquelas que temos interesse em estudar são chamadas *variáveis*. Alguns exemplos são comprimento, massa, idade, temperatura, número de ocorrências, etc. Cada membro da população que é escolhido como parte de uma amostra fornece uma medida de uma ou mais variáveis, chamadas *observações*.

1.1 Princípios de estimação

Suponha que estamos interessados em um parâmetro populacional (desconhecido) θ. O conjunto Θ aonde θ assume valores é denominado espaço paramétrico.

Exemplo 1.1 : Se $X \sim Poisson(\theta)$, então $\Theta = \{\theta : \theta > 0\}$.

Exemplo 1.2 : Se $X \sim N(\mu, 1)$, então $\Theta = \{\mu : -\infty < \mu < \infty\}$.

Exemplo 1.3 : Se $X \sim N(\mu, \sigma^2)$, então $\Theta = \{(\mu, \sigma^2) : -\infty < \mu < \infty,\ \sigma^2 > 0\}$.

Podemos *estimar* o parâmetro θ usando a informação de nossa amostra. Chamamos este único número que representa o valor mais plausível do parâmetro (baseado nos dados amostrais) de uma *estimativa pontual* de θ. Alguns exemplos são a média amostral, o desvio padrão amostral, a mediana amostral, os quais estimam a verdadeira média, desvio padrão e mediana da população (que são desconhecidos).

Definição 1.2 *Uma estatística é uma função qualquer das variáveis aleatórias observáveis $X_1, \ldots, X_n$ que não depende do parâmetro desconhecido.*

Note que por esta definição, uma estatística é também uma variável aleatória observável. Estatísticas são usualmente representadas por letras latinas, (por exemplo, $\overline{X}$ para a média amostral, S para o desvio padrão amostral), enquanto que parâmetros são usualmente representados por letras gregas (por exemplo, μ para a média populacional, σ para o desvio padrão populacional). É claro que à medida que a amostra aumenta, mais informação nós teremos acerca da população de interesse, e portanto mais precisas serão as estimativas dos parâmetros de interesse.

Definição 1.3 *Qualquer estatística que assume valores em Θ é denominada um estimador para θ.*

Das definições acima segue então que qualquer estimador é uma estatística mas nem toda estatística define um estimador.

Definição 1.4 *Momentos amostrais: Para uma amostra aleatória $X_1, \ldots, X_n$ o k-ésimo momento amostral é definido como*

$$M_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$$

e o k-ésimo momento amostral em torno de $\overline{X}$ é definido como

$$M_k' = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^k$$

1.2 Função de Verossimilhança

Seja uma única variável aleatória X cuja distribuição depende de um único parâmetro θ. Para um valor x fixo e variando θ, $p(x|\theta) = l(\theta; x)$ é a *plausibilidade* ou *verossimilhança* de cada um dos valores de θ. Assim, a função de verossimilhança de θ é uma função que associa o valor de $p(x|\theta)$ a cada um dos possíveis valores de θ. Vale notar que $l(\theta; x)$ não é uma função de densidade de probabilidade, i.e. em geral

$$\int l(\theta; x)d\theta \neq 1.$$

Exemplo 1.4 : Se $X \sim$ Binomial $(2, \theta)$ então

$$p(x|\theta) = l(\theta; x) = \binom{2}{x}\theta^x(1-\theta)^{2-x}, \quad x = 0, 1, 2 \quad \theta \in (0, 1)$$

e a integral da função de verossimilhança em relação a θ é dada por

$$\int l(\theta; x)d\theta = \binom{2}{x}\int_0^1 \theta^x(1-\theta)^{2-x}d\theta.$$

Mas o integrando é o núcleo da função de densidade de uma distribuição Beta (ver Apêndice A) com parâmetros $x + 1$ e $3 - x$, portanto

$$\int_0^1 \theta^x(1-\theta)^{2-x}d\theta = \frac{\Gamma(x+1)\Gamma(3-x)}{\Gamma(x+1+3-x)} = \frac{x!(2-x)!}{3!}.$$

Esta última igualdade vem do fato que sendo x um número inteiro positivo então $\Gamma(x) = (x-1)!$. Após algumas simplificações segue que

$$\int l(\theta; x) d\theta = \frac{1}{3}.$$

Além disso, para cada possível valor observado de X temos um valor mais plausível para θ,

(i) $l(\theta; x = 1) = 2\theta(1-\theta)$ e o valor mais provável de θ é $1/2$.

(ii) $l(\theta; x = 2) = \theta^2$ e o valor mais provável é 1.

(iii) $l(\theta; x = 0) = (1-\theta)^2$ e o valor mais provável é 0.

Claro que na prática um único valor de X será observado. Na Figura 1.1 estão representadas as funções de verossimilhança para uma única variável aleatória X com distribuições Binomial(2,θ), Poisson(θ) e Exponencial(θ).

Se $\boldsymbol{x} = (x_1, \ldots, x_n)$ são os valores observados das variáveis aleatórias $X_1, \ldots, X_n$ cuja função de (densidade) de probabilidade conjunta é $p(\boldsymbol{x}|\theta)$ então a função de verossimilhança de θ é $l(\theta; \boldsymbol{x}) = p(\boldsymbol{x}|\theta)$. No caso particular em que $X_1, \ldots, X_n$ são variáveis aleatórias independentes e identicamente distribuidas, a função de verossimilhança de θ correpondente à amostra observada $x_1, \ldots, x_n$ é dada por

$$l(\theta; \boldsymbol{x}) = \prod_{i=1}^{n} p(x_i|\theta).$$

Note porém que a definição de verossimilhança não requer que os dados sejam observações de variáveis aleatórias independentes ou identicamente distribuidas. Além disso, fatores que dependem somente de $\boldsymbol{x}$ e não dependem de θ podem ser ignorados quando se escreve a função de verossimilhança já que eles não fornecem informação sobre a plausibilidade relativa de diferentes valores de θ.

No caso geral θ pode ser um escalar, um vetor ou mesmo uma matriz de parâmetros.

Informação de Fisher

O conceito visto a seguir será útil no cálculo da variância de estimadores, estudo do comportamento assintótico de estimadores de máxima verossimilhança e em inferência Bayesiana.

Definição 1.5 *Considere uma única observação X com função de (densidade) de probabilidade $p(x|\theta)$. A medida de informação esperada de Fisher de θ através*

de X é definida como

$$I(\theta) = E\left[-\frac{\partial^2 \log p(x|\theta)}{\partial \theta^2}\right].$$

No caso de um vetor paramétrico $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$ define-se a matriz de informação esperada de Fisher de $\boldsymbol{\theta}$ através de X como

$$\boldsymbol{I}(\boldsymbol{\theta}) = E\left[-\frac{\partial^2 \log p(x|\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}\right].$$

Note que o conceito de informação aqui está sendo associado a uma espécie de curvatura média da função de verossimilhança no sentido de que quanto maior a curvatura mais precisa é a informação contida na verossimilhança, ou equivalentemente maior o valor de $I(\theta)$. Em geral espera-se que a curvatura seja negativa e por isso seu valor é tomado com sinal trocado. Note também que a esperança matemática é tomada em relação à distribuição amostral $p(x|\theta)$.

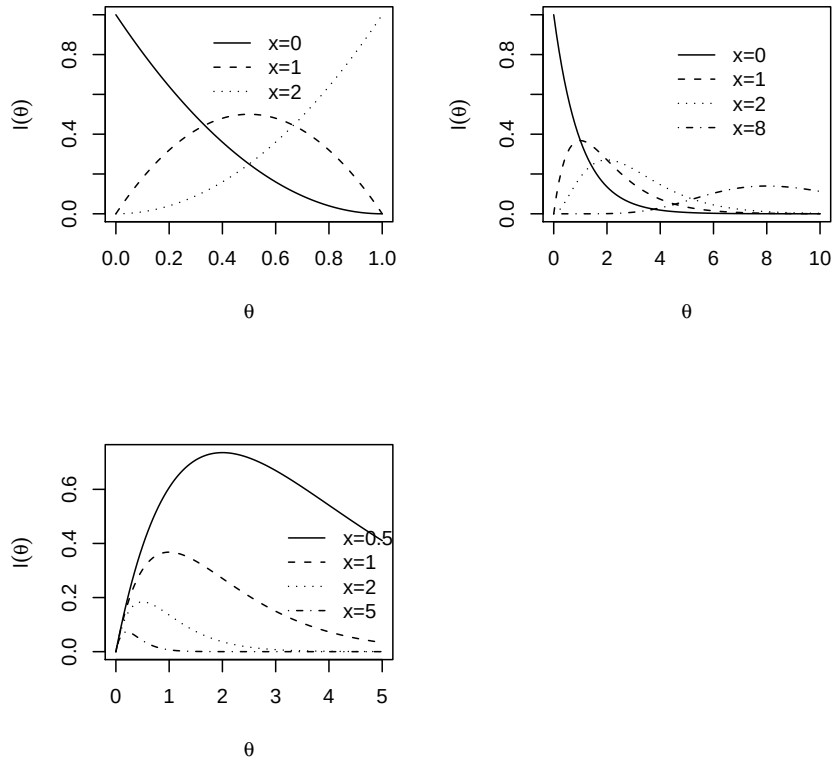

Figura 1.1: Funções de verossimilhança para uma única variável aleatória X com distribuições Binomial(2,θ), Poisson(θ) e Exponencial(θ).

Podemos considerar então $I(\theta)$ uma medida de informação global enquanto

que uma medida de informação local é obtida quando não se toma o valor esperado na definição acima. A medida de informação observada de Fisher $J(\theta)$ fica então definida como

$$J(\theta) = -\frac{\partial^2 \log p(x|\theta)}{\partial \theta^2}.$$

Lema 1.1 *Seja $\boldsymbol{X} = (X_1, \ldots, X_n)$ uma coleção de variáveis aleatórias independentes com distribuições $p_i(x|\theta)$, $i = 1, \ldots, n$ e sejam $I(\theta)$, $J(\theta)$, $J_i(\theta)$ e $I_i(\theta)$ as medidas de informação de θ obtidas através de $\boldsymbol{X}$ e de X_i, respectivamente. Então,*

$$I(\theta) = \sum_{i=1}^{n} I_i(\theta) \quad e \quad J(\theta) = \sum_{i=1}^{n} J_i(\theta).$$

Prova. A prova é simples e será deixada como exercício.

O lema nos diz então que a informação total contida em observações independentes é igual a soma das informações individuais. Um caso particular importante é quando as observações são também identicamente distribuidas já que neste caso $I_i(\theta)$ é constante e assim a informação total é simplesmente $nI(\theta)$.

Outra estatística muito importante no estudo da função de verossimilhança e que será útil é a função escore definida a seguir.

Definição 1.6 *A função escore de X denotada por $U(X;\theta)$ é dada por*

$$U(X;\theta) = \frac{\partial \log p(X|\theta)}{\partial \theta}.$$

No caso de um vetor paramétrico $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$ a função escore será um vetor $\boldsymbol{U}(X;\boldsymbol{\theta})$ com componentes $U_i(X;\boldsymbol{\theta}) = \partial \log p(X|\boldsymbol{\theta})/\partial \theta_i$.

Além disso, pode-se mostrar que sob certas condições de regularidade o valor esperado da função escore é zero e sua variância é dada por $I(\theta)$[1] (a prova será deixada como exercício). Segue então que uma forma alternativa de cálculo da informação de Fisher é obtida a partir da função escore como

$$I(\theta) = E[U^2(\boldsymbol{X};\theta)]$$

onde a esperança é tomada em relação à distribuição de $\boldsymbol{X}|\theta$. No caso de um vetor paramétrico o resultado fica

$$I(\boldsymbol{\theta}) = E[U(\boldsymbol{X};\boldsymbol{\theta})U(\boldsymbol{X};\boldsymbol{\theta})'].$$

[1] As condições de regularidade referem-se à verossimilhança ser derivável em todo o espaço paramétrico e à troca dos sinais de derivação e integração.

Exemplo 1.5 : Seja $X_1, \ldots, X_n$ uma amostra aleatória da distribuição exponencial com parâmetro θ. A função de densidade de cada X_i é dada por

$$p(x_i|\theta) = \theta e^{-\theta x_i}, \quad \theta > 0,$$

e portanto a função de densidade conjunta é dada por

$$p(\boldsymbol{x}|\theta) = \theta^n e^{-\theta t}, \quad \theta > 0, \quad \text{sendo } t = \sum_{i=1}^{n} x_i.$$

Tomando-se o logaritmo obtém-se

$$\log p(\boldsymbol{x}|\theta) = n \log(\theta) - \theta t$$

de modo que as derivadas de primeira e segunda ordem são

$$\frac{\partial \log p(\boldsymbol{x}|\theta)}{\partial \theta} = \frac{n}{\theta} - t \quad \text{e} \quad \frac{\partial^2 \log p(\boldsymbol{x}|\theta)}{\partial \theta^2} = -\frac{n}{\theta^2}$$

e a informação esperada de Fisher baseada na amostra é $I(\theta) = n/\theta^2$. Além disso, a função escore é dada por

$$U(\boldsymbol{X}, \theta) = \frac{n}{\theta} - \sum_{i=1}^{n} X_i.$$

Exemplo 1.6 : Seja $X_1, \ldots, X_n$ uma amostra aleatória da distribuição de Poisson com parâmetro θ. A função de densidade conjunta é dada por

$$p(\boldsymbol{x}|\theta) = \prod_{i=1}^{n} \frac{\exp(-\theta)\ \theta^{x_i}}{x_i!} = \exp(-n\theta)\ \theta^t \prod_{i=1}^{n} \frac{1}{x_i!}, \quad \theta > 0, \quad \text{sendo } t = \sum_{i=1}^{n} x_i.$$

As derivadas de primeira e segunda ordem do logaritmo da verossimilhança são

$$\frac{\partial \log p(\boldsymbol{x}|\theta)}{\partial \theta} = -n + \frac{t}{\theta} \quad \text{e} \quad \frac{\partial^2 \log p(\boldsymbol{x}|\theta)}{\partial \theta^2} = -\frac{t}{\theta^2}$$

e portanto a informação esperada de Fisher é

$$I(\theta) = \frac{1}{\theta^2} E\left[\sum_{i=1}^{n} X_i\right] = \frac{1}{\theta^2} \sum_{i=1}^{n} E(X_i) = \frac{n\theta}{\theta^2} = \frac{n}{\theta}.$$

1.3 Suficiência e família exponencial

Dado um conjunto de observações $\boldsymbol{X} = (X_1, \cdots, X_n)$, será que existe alguma função $\boldsymbol{T}(\boldsymbol{X})$ que resume toda a informação contida em $\boldsymbol{X}$? Esta idéia dá origem ao conceito de estatística suficiente definido a seguir.

Definição 1.7 *$\boldsymbol{T}(\boldsymbol{X})$ é uma estatística suficiente para o parâmetro θ se $p(\boldsymbol{x}|\boldsymbol{t}, \theta) = p(\boldsymbol{x}|\boldsymbol{t})$. Assim, dado $\boldsymbol{T}$, $\boldsymbol{X}$ não traz nenhuma informação adicional sobre o parâmetro θ.*

Ou seja por esta definição, ao invés de observar $\boldsymbol{X}$ basta observar $\boldsymbol{T}$ que pode ter dimensão muito menor. Na prática esta definição é difícil de ser aplicada e precisamos de uma ferramenta adicional.

Teorema 1.1 *(Critério de fatoração de Neyman) $\boldsymbol{T}(\boldsymbol{X})$ é suficiente para θ se somente se*

$$p(\boldsymbol{x}|\theta) = f(\boldsymbol{t}, \theta) g(\boldsymbol{x})$$

com f e g não negativas.

Exemplo 1.7: Sejam $\boldsymbol{X} = (X_1, \cdots, X_n)$ observações tipo 0-1 com $P(X_i = 1|\theta) = \theta$. Então para r sucessos e s falhas a função de densidade conjunta é

$$p(\boldsymbol{x}|\theta) = \theta^t (1-\theta)^{n-t}, \quad \text{onde} \quad t = \sum_{i=1}^{n} x_i$$

e portanto $\boldsymbol{T}(\boldsymbol{X}) = \sum_{i=1}^{n} X_i$ é uma estatística suficiente para θ.

Exemplo 1.8: Dado θ, $X_1, \cdots, X_n$ são independentes e identicamente distribuidos com funções de densidade $p(x_i|\theta)$. Então a função de densidade conjunta é

$$p(\boldsymbol{x}|\theta) = p(x_1, \cdots, x_n|\theta) = \prod_{i=1}^{n} p(x_i|\theta).$$

Definindo as estatísticas de ordem

$$Y_1 = X_{(1)} = \min_i X_i \leq \cdots \leq Y_n = X_{(n)} = \max_i X_i$$

e como a cada x_i corresponde um único y_i então

$$p(\boldsymbol{x}|\theta) = \prod_{i=1}^{n} p(x_i|\theta) = \prod_{i=1}^{n} p(y_i|\theta) = g(\boldsymbol{x}) f(\boldsymbol{t}, \theta)$$

com $g(\boldsymbol{x}) = 1$, $f(\boldsymbol{t}, \theta) = \prod_{i=1}^{n} p(y_i|\theta)$ e $\boldsymbol{t} = (y_1, \cdots, y_n)$.
Conclusão: $\boldsymbol{T}(\boldsymbol{X}) = (Y_1, \cdots, Y_n)$ é estatística suficiente para θ e a dimensão de $\boldsymbol{T}$ depende do tamanho amostral.

O que se pode notar deste último exemplo é que o conceito de suficiência não é necessariamente útil. Na prática estamos interessados em uma redução significativa em relação ao tamanho amostral. Um questão que se coloca é como obter estatísticas suficientes que gerem a maior redução possível nos dados.

Definição 1.8 *$\boldsymbol{T}(\boldsymbol{X})$ é estatística suficiente minimal para θ se for suficiente e se for função de qualquer outra estatística suficiente para θ.*

Além disso pode-se mostrar que,

- Se $\boldsymbol{S}(\boldsymbol{X})$ é função bijetiva de $\boldsymbol{T}(\boldsymbol{X})$ então $\boldsymbol{S}$ também é suficiente.
- Estatísticas suficientes minimais são únicas.

Existem distribuições com estatísticas suficientes cuja dimensão é igual ao número de parâmetros para qualquer tamanho n da amostra. Isto nos remete às definições da próxima seção.

1.3.1 Família Exponencial

A família exponencial inclui muitas das distribuições de probabilidade mais comumente utilizadas em Estatística, tanto contínuas quanto discretas. Uma característica essencial desta família é que existe uma estatística suficiente com dimensão fixa.

Definição 1.9 *A família de distribuições com função de (densidade) de probabilidade $p(x|\theta)$ pertence à família exponencial a um parâmetro se podemos escrever*

$$p(x|\theta) = a(x) \ \exp\{u(x)\phi(\theta) + b(\theta)\}.$$

Note que pelo critério de fatoração de Neyman $U(X)$ é uma estatística suficiente para θ.

A definição de família exponencial pode ser estendida ao caso multiparamétrico com $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_r)$,

$$p(x|\boldsymbol{\theta}) = a(x) \ \exp\left\{\sum_{j=1}^{r} u_j(x)\phi_j(\boldsymbol{\theta}) + b(\boldsymbol{\theta})\right\},$$

e quando se tem uma amostra aleatória $X_1, \ldots, X_n$, i.e.

$$p(\boldsymbol{x}|\boldsymbol{\theta}) = \left[\prod_{i=1}^{n} a(x_i)\right] \exp\left\{\sum_{j=1}^{r}\left[\sum_{i=1}^{n} u_j(x_i)\right]\phi_j(\boldsymbol{\theta}) + nb(\boldsymbol{\theta})\right\}$$

Neste caso, definindo $U_j(\boldsymbol{X}) = \sum_{i=1}^{n} U_j(x_i)$, $i = 1, \ldots, n$, então pelo critério de fatoração, $\boldsymbol{T}(\boldsymbol{X}) = (U_1(\boldsymbol{X}), \ldots, U_r(\boldsymbol{X}))$ é uma estatística *conjuntamente suficiente* para o vetor de parâmetros $(\theta_1, \ldots, \theta_r$.

Exemplo 1.9 : $(X_1, \cdots, X_n) \sim$ Bernoulli(θ)

$$\begin{aligned} p(x|\theta) &= \theta^x(1-\theta)^{1-x} I_x(\{0,1\}) \\ &= \exp\left\{x \log\left(\frac{\theta}{1-\theta}\right) + \log(1-\theta)\right\} I_x(\{0,1\}) \\ \Rightarrow p(\boldsymbol{x}|\theta) &= \exp\left\{\left(\sum_{i=1}^{n} x_i\right) \log\left(\frac{\theta}{1-\theta}\right) + n\log(1-\theta)\right\} I_{\boldsymbol{x}}(\{0,1\}^n) \end{aligned}$$

Conclusão: A Bernoulli pertence à família exponencial e $U = \sum_{i=1}^{n} X_i$ é estatística suficiente para θ.

Exemplo 1.10 : Sejam $X_1, \cdots, X_n \sim$ Poisson(λ). Então,

$$\begin{aligned} p(x|\lambda) &= \frac{e^{-\lambda}\lambda^x}{x!} I_x(\{0,1,\cdots\}) = \frac{1}{x!}\exp\{-\lambda + x\log\lambda\} I_x(\{0,1,\cdots\}) \\ \Rightarrow p(\boldsymbol{x}|\lambda) &= \frac{1}{\prod x_i!}\exp\{-n\lambda + \sum x_i \log\lambda\} I_{\boldsymbol{x}}(\{0,1,\cdots\}^n) \end{aligned}$$

Conclusão: A Poisson pertence à família exponencial e $U = \sum_{i=1}^{n} X_i$ é estatística suficiente para λ.

Exemplo 1.11 : Sejam $X_1, \cdots, X_n \sim$ Normal(θ, σ^2). Então,

$$\begin{aligned} p(x_i|\theta,\sigma^2) &= (2\pi\sigma^2)^{-1/2}\exp\{-(x_i-\theta)^2/2\sigma^2\} \\ &= (2\pi)^{-1/2}\exp\left\{\frac{\theta}{\sigma^2}x_i - \frac{1}{2\sigma^2}x_i^2 - \frac{\theta^2}{2\sigma^2} - \frac{1}{2}\log\sigma^2\right\} \\ \Rightarrow p(\boldsymbol{x}|\theta,\sigma^2) &= (2\pi)^{-n/2}\exp\left\{\frac{\theta}{\sigma^2}\sum_{i=1}^{n} x_i - \frac{1}{2\sigma^2}\sum_{i=1}^{n} x_i^2 - \frac{n}{2}\left(\frac{\theta^2}{\sigma^2} + \log\sigma^2\right)\right\} \end{aligned}$$

Conclusão: A Normal pertence à família exponencial e $\boldsymbol{U} = (\sum_{i=1}^{n} X_i, \sum_{i=1}^{n} X_i^2)$ é estatística conjuntamente suficiente para (θ, σ^2).

1.4 Problemas

1. Uma única variável aleatória X tem distribuição de Bernoulli com parâmetro θ desconhecido mas sabe-se que $\theta = 0,25$ ou $\theta = 0,75$. A tabela abaixo descreve a distribuição de X para cada possível valor de θ.

	θ	
X	0,25	0,75
0	1/4	5/6
1	3/4	1/6

   (a) Explique por que a soma em cada coluna é igual a 1 mas a soma em cada linha não é.

   (b) Qual valor de θ você escolheria como o mais plausível se $X = 1$ for observado?

2. Explique as diferenças entre estatísticas, estimadores e estimativas.

3. Se $X_1, \ldots, X_n$ é uma amostra aleatória da $N(\mu, \sigma^2)$ prove que se $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$ então

$$U = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}.$$

4. Prove o Lema 1.1, i.e. que a informação total contida em observações independentes é igual a soma das informações individuais.

5. Prove que a média da função escore é zero e sua variância é igual a $I(\theta)$.

6. Se $X_1, \ldots, X_n \sim$ Bernoulli(p) obtenha a informação de Fisher para p.

7. Se $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ obtenha a matriz de informação de Fisher para (μ, σ^2).

8. Seja uma amostra aleatória $X_1, \ldots, X_n$ de cada uma das distribuições abaixo. Mostre que a estatística T especificada é uma estatística suficiente para o parâmetro.

   (a) Distribuição de Bernoulli com parâmetro p desconhecido, $T = \sum_{i=1}^{n} X_i$.

   (b) Distribuição geométrica com parâmetro p desconhecido, $T = \sum_{i=1}^{n} X_i$.

   (c) Distribuição binomial negativa com parâmetros r conhecido e p desconhecido, $T = \sum_{i=1}^{n} X_i$.

(d) Distribuição normal com média μ conhecida e variância σ^2 desconhecida, $T = \sum_{i=1}^{n}(X_i - \mu)^2$.

(e) Distribuição gama com parâmetros α conhecido e β desconhecido, $T = \overline{X}$.

(f) Distribuição gama com parâmetros α desconhecido e β conhecido, $T = \prod_{i=1}^{n} X_i$.

(g) Distribuição beta com parâmetros α desconhecido e β conhecido, $T = \prod_{i=1}^{n} X_i$.

(h) Distribuição uniforme nos inteiros $1, 2, \ldots, \theta$ para θ desconhecido ($\theta = 1, 2, \ldots$), $T = \max(X_1, \ldots, X_n)$.

(i) Distribuição uniforme no intervalo (a, b) com a conhecido e b desconhecido ($b > a$), $T = \max(X_1, \ldots, X_n)$.

(j) Distribuição uniforme no intervalo (a, b) com a desconhecido e b conhecido ($a < b$), $T = \min(X_1, \ldots, X_n)$.

9. Verifique que cada uma das famílias de distribuições abaixo é uma família exponencial e obtenha as estatísticas suficientes de dimensão mínima.

(a) A família de distribuições de Bernoulli com parâmetro p desconhecido.

(b) A família de distribuições de Poisson com média desconhecida.

(c) A família de distribuições Normais com média desconhecida e variância conhecida.

(d) A família de distribuições Normais com média conhecida e variância desconhecida.

(e) A família de distribuições $Gama(\alpha, \beta)$ com α desconhecido e β conhecido.

(f) A família de distribuições $Gama(\alpha, \beta)$ com α conhecido e β desconhecido.

(g) A família de distribuições $Beta(\alpha, \beta)$ com α desconhecido e β conhecido.

(h) A família de distribuições $Beta(\alpha, \beta)$ com α conhecido e β desconhecido.

1.5 Teorema Central do Limite

Um resultado que nos permite conduzir alguns procedimentos de inferência sem qualquer conhecimento da distribuição da população é apresentado a seguir.

Teorema 1.2 *Se $X_1, X_2, \ldots$ são variáveis aleatórias independentes e identicamente distribuidas com média μ e variância $\sigma^2 < \infty$ e $\overline{X}_n = \sum_{i=1}^{n} X_i/n$ então*

$$\sqrt{n}(\overline{X}_n - \mu) \xrightarrow{D} Y, \text{ quando } n \to \infty \tag{1.1}$$

com $Y \sim N(0, \sigma^2)$.

Para simplificar a notação usualmente escreve-se (1.1) como

$$\sqrt{n}(\overline{X}_n - \mu) \xrightarrow{D} N(0, \sigma^2), \text{ quando } n \to \infty.$$

Assim, o Teorema 1.2 nos diz que qualquer que seja a distribuição da variável de interesse, a distribuição das médias amostrais tenderá a uma distribuição normal à medida que o tamanho de amostra cresce.

Capítulo 2

Propriedades dos Estimadores

Na inferência estatística clássica não existe um critério único para escolha de estimadores em um dado problema, mas sim um conjunto de critérios que podem ser utilizados para seleção e comparação. Estes critérios ou propriedades são descritos a seguir.

Definição 2.1 *Seja uma amostra aleatória $X_1, \ldots, X_n$ tomada de uma distribuição parametrizada por θ. O erro quadrático médio de um estimador $\hat{\theta}$ de θ é definido como*

$$EQM(\hat{\theta}) = E(\hat{\theta} - \theta)^2.$$

Podemos reescrever esta última expressão como

$$\begin{aligned} EQM(\hat{\theta}) &= E[(\hat{\theta} - E(\hat{\theta})) + (E(\hat{\theta}) - \theta)]^2 \\ &= E[\hat{\theta} - E(\hat{\theta})]^2 + [E(\hat{\theta}) - \theta]^2 \\ &= Var(\hat{\theta}) + [E(\hat{\theta}) - \theta]^2. \end{aligned}$$

onde o termo $E(\hat{\theta}) - \theta$ é chamado *vício* ou *viés* do estimador e denotado por $B(\hat{\theta})$.

Assim, o erro quadrático médio é definido como a variância do estimador mais o quadrado do seu viés. Um caso particular ocorre quando $B(\hat{\theta}) = 0$, ou equivalentemente $E(\hat{\theta}) = \theta$, i.e. o vício do estimador é nulo. Neste caso diz-se que $\hat{\theta}$ é um estimador não viesado (ENV) para θ e da Definição 2.1 segue que $EQM(\hat{\theta}) = Var(\hat{\theta})$. A interpretação clássica desta definição é que, após observar todas as possíveis amostras de tamanho n desta distribuição a média dos valores calculados de $\hat{\theta}$ será θ.

Se $E(\hat{\theta}) \neq \theta$ então o estimador $\hat{\theta}$ é dito ser viesado ou viciado. No entanto pode ocorrer que a esperança do estimador se aproxima do verdadeiro valor de θ à medida que aumenta o tamanho da amostra, i.e. $\lim_{n\to\infty} E(\hat{\theta}) = \theta$. Neste caso, $\hat{\theta}$ é dito ser uma estimador assintoticamente não viesado para θ.

Exemplo 2.1 : Sejam as variáveis aleatórias $X_1, \ldots, X_n$ independentes e identi-

camente distribuidas com $E(X_i) = \mu$ e $Var(X_i) = \sigma^2$. Então,

(i) $E(\overline{X}) = \dfrac{1}{n}\sum_{i=1}^{n} E(X_i) = \dfrac{1}{n}\sum_{i=1}^{n} \mu = \mu$

(i) $Var(\overline{X}) = \dfrac{1}{n^2}\sum_{i=1}^{n} Var(X_i) = \dfrac{1}{n^2}\sum_{i=1}^{n} \sigma^2 = \dfrac{\sigma^2}{n}.$

Portanto a média amostral $\overline{X}$ é um ENV da média populacional μ e sua variância dada por σ^2/n diminui com o tamanho da amostra.

Exemplo 2.2 : (continuação) Suponha agora que o seguinte estimador $\hat{\sigma}^2 = (1/n)\sum_{i=1}^{n}(X_i - \overline{X})^2$ é proposto para σ^2. Então

$$E(\hat{\sigma}^2) = \frac{1}{n} E\left[\sum_{i=1}^{n}(X_i - \overline{X})^2\right].$$

Mas a soma dos quadrados em torno da média amostral pode ser reescrita como

$$\begin{aligned}
\sum_{i=1}^{n}(X_i - \overline{X})^2 &= \sum_{i=1}^{n}[(X_i - \mu) - (\overline{X} - \mu)]^2 \\
&= \sum_{i=1}^{n}(X_i - \mu)^2 - 2(\overline{X} - \mu)\sum_{i=1}^{n}(X_i - \mu) + n(\overline{X} - \mu)^2 \\
&= \sum_{i=1}^{n}(X_i - \mu)^2 - n(\overline{X} - \mu)^2.
\end{aligned}$$

Assim, a esperança do estimador é dada por

$$E(\hat{\sigma}^2) = \frac{1}{n}\left[\sum_{i=1}^{n} E(X_i - \mu)^2 - nE(\overline{X} - \mu)^2\right] = \sigma^2 - \frac{\sigma^2}{n} = \left(\frac{n-1}{n}\right)\sigma^2$$

e conclui-se que $\hat{\sigma}^2$ não é um ENV para σ^2. Porém,

$$\lim_{n\to\infty}\left(\frac{n-1}{n}\right)\sigma^2 = \sigma^2$$

e portanto $\hat{\sigma}^2$ é assintoticamente não viesado para σ^2.

No exemplo acima note que nenhuma distribuição de probabilidades foi atribuida aos $X_i's$. Assim, as propriedades obtidas são válidas qualquer que seja a distribuição dos dados. Além disso, fica fácil obter um ENV para σ^2 notando-se que

$$E\left[\left(\frac{n}{n-1}\right)\hat{\sigma}^2\right] = \left(\frac{n}{n-1}\right)E(\hat{\sigma}^2) = \sigma^2.$$

Portanto, o estimador

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

é um ENV para a variância populacional σ^2.

Em geral o processo de estimação consiste em escolher o estimador que apresenta o menor erro quadrático médio. No caso de estimadores não viesados isto equivale a escolher aquele com a menor variância.

Exemplo 2.3 : (continuação) Seja o estimador $\hat{\mu} = X_1$ para a média populacional μ. Como $E(\hat{\mu}) = E(X_1) = \mu$ segue que $\hat{\mu} = X_1$ é também um ENV para μ. Portanto

$$EQM(\overline{X}) = \frac{\sigma^2}{n} < EQM(\hat{\mu}) = \sigma^2, \quad \text{para} \quad n > 1 \quad \text{e} \quad \forall \mu$$

e assim o estimador $\overline{X}$ deve ser escolhido.

O simples fato de um estimador ser não viesado não significa que ele seja bom, mas se a sua variância for pequena então necessariamente sua distribuição estará concentrada em torno da média e com alta probabilidade $\hat{\theta}$ estará próximo de θ.

Exemplo 2.4 : Seja $X_1, \ldots, X_n$ uma amostra aleatória da distribuição de Poisson com parâmetro λ. Como $E(X_i) = Var(X_i) = \lambda$ segue dos resultados nos Exemplos 2.1 e 2.2 que $\overline{X}$ e S^2 são ENV para λ. Além disso,

$$\hat{\theta} = \alpha\overline{X} + (1-\alpha)S^2$$

também é um ENV para λ já que

$$E(\hat{\theta}) = \alpha E(\overline{X}) + (1-\alpha)E(S^2) = \alpha\lambda + (1-\alpha)\lambda = \lambda.$$

Exemplo 2.5 : Seja $X_1, \ldots, X_n$ uma amostra aleatória da distribuição $N(\mu, \sigma^2)$ e seja o estimador $T^2 = c\sum_{i=1}^{n}(X_i - \overline{X})^2$. Nesta classe de estimadores vamos obter o de menor erro quadrático médio. Como

$$\frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{\sigma^2} \sim \chi^2_{n-1}$$

então

$$E(T^2) = c(n-1)\sigma^2 \quad \text{e} \quad Var(T^2) = c^2 2(n-1)\sigma^4$$

e portanto

$$EQM(T^2) = 2c^2(n-1)\sigma^4 + [c(n-1)\sigma^2 - \sigma^2]^2.$$

Para obter o valor de c tal que T^2 tem o menor erro quadrático médio vamos derivar a expressão acima em relação a c e igualar a zero, i.e.

$$\frac{d}{dc}EQM(T^2) = 4c(n-1)\sigma^4 + 2[c(n-1)\sigma^2 - \sigma^2](n-1)\sigma^2 = 0$$

ou equivalentemente

$$-4c(n-1)\sigma^4 = 2(n-1)\sigma^2[c(n-1)\sigma^2 - \sigma^2]$$

e finalmente

$$c = \frac{1}{n+1}.$$

Não é difícil mostrar que a segunda derivada em relação a c é maior do que zero para $n > 1$ de modo que o estimador

$$T_0^2 = \frac{1}{n+1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

tem o menor EQM nesta classe de estimadores, para todos os possíveis valores de μ e σ^2.

Vimos então que o erro quadrático médio é a ferramenta usualmente utilizada para comparar estimadores. Dizemos que $\hat{\theta}_1$ é *melhor* do que $\hat{\theta}_2$ se

$$EQM(\hat{\theta}_1) \leq EQM(\hat{\theta}_2)$$

com $\leq$ substituido por $<$ para ao menos um valor de θ. Neste caso o estimador $\hat{\theta}_2$ é dito ser *inadmissível*. Um estimador é dito ser *ótimo* (ou *admissível*) para θ se não existir nenhum outro estimador melhor do que ele. Assim, $\hat{\theta}^*$ é um estimador ótimo para θ se

$$EQM(\hat{\theta}^*) \leq EQM(\hat{\theta})$$

com $\leq$ substituido por $<$ para ao menos um valor de θ. No Exemplo 2.5 o estimador T_0^2 é ótimo naquela classe de estimadores.

No caso de estimadores não viesados a comparação é feita em termos de variâncias. Em particular, se $\hat{\theta}^*$ for um ENV para θ e

$$Var(\hat{\theta}^*) \leq Var(\hat{\theta}), \quad \forall\theta$$

com $\leq$ substituido por $<$ para ao menos um valor de θ então $\hat{\theta}^*$ é dito ser *não*

viesado de variância uniformemente mínima (UMVU). A seguir serão apresentados conceitos que possibilitarão a obtenção de estimadores não viesados ótimos.

2.1 Estimadores baseados em estatisticas suficientes

O teorema a seguir, conhecido como teorema de Rao-Blackwell mostra que é possível melhorar estimadores não viesados via estatísticas suficientes.

Teorema 2.1 *(Rao-Blackwell) Para uma amostra aleatória $X_1, \ldots, X_n$ sejam $T(X_1, \ldots, X_n)$ uma estatística suficiente para θ e $S(X_1, \ldots, X_n)$ um estimador não viesado de θ que não seja função de T. Então*

$$\hat{\theta} = E[S(\boldsymbol{X})|T(\boldsymbol{X})]$$

é um ENV de θ com $Var(\hat{\theta}) \leq Var[S(\boldsymbol{X})]$.

Basicamente, o teorema de Rao-Balckwell nos diz que é sempre possível melhorar um estimador não viesado condicionando em uma estatística suficiente. A pergunta que se faz aqui é como obter a menor redução possível na variância e para isto precisamos do conceito de estatística completa.

Definição 2.2 *Uma estatística $T(X_1, \ldots, X_n)$ é dita ser completa em relação à família $p(x|\theta)$ se a única função real g definida no domínio de T tal que $E[g(T)] = 0$, $\forall\theta$ é a função nula, i.e. $g(T) = 0$.*

Teorema 2.2 *(Lehmann-Scheffé) Se T é uma estatística suficiente e completa e S é um ENV para θ então $\hat{\theta}$ é o único ENV para θ baseado em T e tem variância uniformemente mínima (UMVU).*

2.2 Eficiência

Um resultado importante que será visto a seguir é que, na classe de estimadores não viesados para um parâmetro θ existe um limite inferior para sua variância. Veremos que isto está associado ao conceito de *eficiência* do estimador.

Teorema 2.3 *Sejam $X_1, \ldots, X_n$ uma amostra aleatória de $p(x|\theta)$ e $T(\boldsymbol{X})$ um estimador não viesado de θ. Sob condições de regularidade,*

$$Var[T(\boldsymbol{X})] \geq \frac{1}{I(\theta)}.$$

Este resultado é conhecido como desigualdade de Cramer-Rao e nos diz então que a variância mínima de um ENV para θ é dada pelo inverso da informação de Fisher.

Definição 2.3 *Um estimador de θ é dito ser eficiente se for não viesado e sua variância atingir o limite inferior da desigualdede de Cramer-Rao para todos os possíveis valores de θ.*

Com esta definição podemos calcular a *eficiência* do estimador como a razão entre o limite inferior da desigualdade e sua variância, i.e. para um estimador $\hat{\theta}$ de θ

$$\text{eficiência}(\hat{\theta}) = \frac{1/I(\theta)}{Var(\hat{\theta})} \leq 1.$$

Vale notar que a variância de um estimador UMVU não necessariamente atinge o limite inferior de Cramer-Rao e sua eficiência pode ser menor do que 1. Porém o contrário é sempre verdade, i.e. estimadores eficientes são necessariamente UMVU.

O Teorema 2.3 pode ser generalizado para o caso de $T(\boldsymbol{X})$ ser um ENV para uma função $h(\theta)$, i.e. $E[T(\boldsymbol{X})] = h(\theta)$. Neste caso, a desigualdade de Cramer-Rao é dada por

$$Var[T(\boldsymbol{X})] \geq \frac{[h'(\theta)]^2}{I(\theta)}$$

sendo $h'(\theta) = dh(\theta)/d\theta$.

Esta forma geral da desigualdade pode ser usada para calcular o limite inferior da variância de um estimador viesado. Seja $\hat{\theta}$ um estimador de θ com viés $b(\theta) = E(\hat{\theta}) - \theta$. Portanto $\hat{\theta}$ é um ENV para $b(\theta) + \theta$. Fazendo $h(\theta) = b(\theta) + \theta$ segue então que

$$Var[\hat{\theta}] \geq \frac{[b'(\theta) + 1]^2}{I(\theta)}.$$

2.3 Consistência

É bem intuitivo pensar que a informação a respeito de um parâmetro contida em uma amostra aumenta conforme o tamanho da amostra aumenta. Assim, é razoável esperar que bons estimadores assumam valores cada vez mais próximos do verdadeiro valor do parâmetro. A seguir serão discutidas propriedades teóricas dos estimadores quando o tamanho amostral torna-se cada vez maior.

Definição 2.4 *Seja $X_1, \ldots, X_n$ uma amostra aleatória de $p(x|\theta)$ e $T(\boldsymbol{X})$ um estimador de $h(\theta)$. Variando o tamanho amostral n obtém-se uma sequência de estimadores $T_n(\boldsymbol{X})$ de $h(\theta)$. Esta sequência é dita ser (fracamente) consistente para $h(\theta)$ se $T_n(\boldsymbol{X}) \to h(\theta)$, em probabilidade quando $n \to \infty$.*

Na prática tem-se uma única amostra de tamanho n e a definição é simplificada dizendo-se que o estimador é ou não consistente, ao invés de uma sequência consistente. A convergência da Definição 2.4 é em probabilidade e pode ser reescrita como

$$P(|T_n(\boldsymbol{X}) - h(\theta)| > \epsilon) \to 0, \ \forall \epsilon > 0, \text{ quando } n \to \infty.$$

Este resultado também é usualmente denotado por $plim\ T_n(\boldsymbol{X}) = h(\theta)$.

É importante também enfatizar a diferença de interpretação entre os conceitos de consistência e viés. Basicamente, consistência refere-se a um único experimento com um número infinitamente grande de replicações enquanto viés refere-se a um número infinitamente grande de experimentos, cada um deles com um número finito de replicações. Ou seja, um estimador consistente pode ser viesado no entanto ele será sempre assintoticamente não viesado.

Finalmente, segue da desigualdade de Chebychev que uma condição suficiente para um ENV ser consistente é que sua variância tenda a zero quando $n \to \infty$. Assim, as condições gerais para a consistência de um estimador $T(\boldsymbol{X})$ de $h(\theta)$ são

$$\lim_{n\to\infty} E[T(\boldsymbol{X})] = h(\theta) \quad \text{e} \quad \lim_{n\to\infty} Var[T(\boldsymbol{X})] = 0.$$

Exemplo 2.6 : Sejam as variáveis aleatórias $X_1, \ldots, X_n$ independentes e identicamente distribuidas com $E(X_i) = \mu$ e $Var(X_i) = \sigma^2$. Vimos no Exemplo 2.1 que $E(\overline{X}) = \mu$ e $Var(\overline{X}) = \sigma^2/n$, portanto $\overline{X}$ é um estimador consistente para a média populacional μ. Além disso,

$$E(\hat{\sigma}^2) = \left(\frac{n-1}{n}\right)\sigma^2 \to \sigma^2, \text{ quando } n \to \infty.$$

e a variância de $\hat{\sigma}^2$ é obtida usando o fato de que

$$Y = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{\sigma^2} \sim \chi^2_{n-1}$$

e $Var(Y) = 2(n-1)$. Assim,

$$Var(\hat{\sigma}^2) = Var\left(\frac{\sigma^2}{n}Y\right) = \frac{\sigma^4}{n^2}Var(Y) = \frac{2\sigma^4(n-1)}{n^2} \to 0, \text{ quando } n \to \infty$$

e segue que $\hat{\sigma}^2$ é um estimador consistente para σ^2.

2.4 Problemas

1. Para uma amostra aleatória $X_1, \ldots, X_n$ tomada de uma distribuição parametrizada por θ mostre que $E(\hat{\theta} - \theta)^2 = Var(\hat{\theta}) + [E(\hat{\theta}) - \theta]^2$

2. Um variável aleatória X tem distribuição desconhecida mas sabe-se que todos os momentos $E(X^k)$, $k = 1, 2, \ldots$ são finitos. Para uma amostra aleatória $X_1, \ldots, X_n$ desta distribuição mostre que o k-ésimo momento amostral $\sum_{i=1}^n X_i^k/n$ é um ENV para $E(X^k)$. Mostre também que este estimador é consistente.

3. Nas condições do exercício 2 encontre um estimador não viesado de $[E(X)]^2$. (Sugestão: $[E(X)]^2 = E(X^2) - Var(X)$)

4. Uma droga será administrada em 2 tipos diferentes A e B de animais. Sabe-se que a resposta média θ é a mesma nos dois tipos de animais mas seu valor é desconhecido e deve ser estimado. Além disso, a variância da resposta é 4 vezes maior em animais do tipo A. Sejam $X_1, \ldots, X_m$ e $Y_1, \ldots, Y_n$ amostras aleatórias independentes de respostas dos animais dos tipos A e B respectivamente.

   (a) Mostre que $\hat{\theta} = \alpha\overline{X} + (1 - \alpha)\overline{Y}$ é um ENV para θ.

   (b) Para valores fixos de m e n obtenha o valor de α que gera um ENV de variância mínima.

5. Seja $X_1, \ldots, X_n$ uma amostra aleatória da distribuição de Poisson com média $\theta > 0$ e $Y = \sum_{i=1}^n X_i$.

   (a) Determine a constante c tal que $\exp(-cY)$ seja um ENV para $\exp(-\theta)$.

   (b) Obtenha o limite inferior para a variância deste estimador.

   (c) Discuta a eficiência deste estimador.

6. Seja $X_1, \ldots, X_n$ uma amostra aleatória da distribuição de Bernoulli com parâmetro $\theta > 0$. Mostre que a variância de qualquer estimador não viesado de $(1 - \theta)^2$ deve ser pelo menos $4\theta(1 - \theta)^3/n$.

7. Descreva as seguintes propriedades fundamentais dos estimadores: consistência, não-tendenciosidade (ou não-viés) e eficiência.

8. Sejam $X_1, \ldots, X_n$ variáveis aleatórias independentes com $X_i \sim Exp(1/\theta)$. Mostre que a média amostral é um estimador eficiente para θ.

9. Sejam $X_1, \ldots, X_n$ variáveis aleatórias independentes com $X_i \sim N(\mu, \sigma^2)$, sendo μ conhecido e σ^2 desconhecido. Verifique se $T(\boldsymbol{X}) = \sum_{i=1}^n (X_i - \mu)^2/n$ é um estimador eficiente para σ^2. (Dica: $E(X - \mu)^4 = 3(\sigma^2)^2$).

10. Seja $X_1, \ldots, X_n$ uma amostra aleatória da distribuição $N(\mu, \sigma^2)$. Mostre que a estatística $T = \sum_{i=1}^{n} a_i X_i$ com $\sum_{i=1}^{n} a_i = 1$ é não viciada. Obtenha valores de a_i para os quais T seja consistente.

Capítulo 3

Métodos de Estimação

3.1 Estimadores de Máxima Verossimilhança

No Capítulo 1 foi introduzido o conceito de verossimilhança ou plausibilidade. Foi visto que esta medida está associada aos possíveis valores de um ou mais parâmetros e a função de verossimilhança define a plausibilidade de cada um destes possíveis valores. Em termos de estimação parece razoável selecionar o valor do parâmetro que recebe a maior verossimilhança, dada uma amostra da população de interesse. Estes conceitos são formalizados a seguir.

Definição 3.1 *Seja $X_1, \ldots, X_n$ uma amostra aleatória de $p(x|\theta)$, $\theta \in \Theta$. A função de verossimilhança de θ correspondente a esta amostra aleatória é dada por*

$$l(\theta; \boldsymbol{x}) = \prod_{i=1}^{n} p(x_i|\theta).$$

Definição 3.2 *O estimador de máxima verossimilhança (EMV) de θ é o valor $\hat{\theta} \in \Theta$ que maximiza $l(\theta; \boldsymbol{x})$. Seu valor observado é a estimativa de máxima verossimilhança.*

No caso uniparamétrico, i.e. θ é um escalar, temos que $\Theta \subseteq \mathbb{R}$ e o EMV pode ser obtido como solução da chamada *equação de verossimilhança*

$$\frac{\partial l(\theta; \boldsymbol{x})}{\partial \theta} = 0. \tag{3.1}$$

É claro que é sempre necessário verificar que a segunda derivada é negativa para garantir que a solução de (3.1) é um ponto de máximo. Ou seja, devemos ter

$$\left.\frac{\partial^2 l(\theta; \boldsymbol{x})}{\partial \theta^2}\right|_{\theta=\hat{\theta}} < 0.$$

Em muitas aplicações é mais simples algebricamente (e muitas vezes computacionalmente) trabalhar na escala dos logaritmos. Do ponto de vista da maximização não fará diferença já que a função logaritmo é estritamente crescente e o valor de θ que maximiza $l(\theta; \boldsymbol{x})$ é o mesmo que que maximiza $\log l(\theta; \boldsymbol{x})$. Portanto, a equação (3.1) pode ser reescrita em termos de logaritmo da verossimilhança e fica

$$\frac{\partial \log l(\theta; \boldsymbol{x})}{\partial \theta} = U(\boldsymbol{X}; \theta) = 0.$$

Trata-se portanto de um problema de otimização e a equação de verossimilhança pode não ter solução analítica.

A Definição 3.2 pode ser generalizada para o caso multiparamétrico, i.e. $\boldsymbol{\theta}$ pode ser um vetor de parâmetros de dimensão k, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$, ou mesmo uma matriz de parâmetros. Se $\boldsymbol{\theta}$ for um vetor de parâmetros as equações de verossimilhança são

$$\frac{\partial l(\boldsymbol{\theta}; \boldsymbol{x})}{\partial \theta_i} = \mathbf{0}, \quad i = 1, \ldots, k. \tag{3.2}$$

Neste caso as condições de segunda ordem para garantir que a solução de (3.2) seja um ponto de máximo referem-se à matriz de segundas derivadas (ou matriz Hessiana) da função de verossimilhança. A condição é de que a matriz

$$H = \left. \frac{\partial^2 l(\boldsymbol{\theta}; \boldsymbol{x})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \right|_{\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}}$$

seja negativa definida, i.e. $\boldsymbol{z}' H \boldsymbol{z} < 0$, $\forall \boldsymbol{z} \neq \mathbf{0}$ sendo cada elemento de H dado por

$$h_{ij} = \frac{\partial^2 l(\boldsymbol{\theta}; \boldsymbol{x})}{\partial \theta_i \partial \theta_j}.$$

Exemplo 3.1 : Seja $X_1, \ldots, X_n$ uma amostra aleatória da distribuição de Bernoulli com parâmetro θ. Para quaisquer valores observados cada x_i é igual a 0 ou 1 e a função de verossimilhança é dada por

$$l(\theta; \boldsymbol{x}) = p(\boldsymbol{x}|\theta) = \prod_{i=1}^{n} \theta^{x_i} (1 - \theta)^{1 - x_i}.$$

Como o valor de θ que maximiza $l(\theta; \boldsymbol{x})$ é o mesmo que maximiza $\log l(\theta; \boldsymbol{x})$ neste caso é mais conveniente algebricamente determinar o EMV obtendo o valor de θ

que maximiza

$$\begin{aligned}\log l(\theta;\boldsymbol{x}) &= \sum_{i=1}^{n}[x_i\log\theta+(1-x_i)\log(1-\theta)]\\ &= \left(\sum_{i=1}^{n}x_i\right)\log\theta+\left(n-\sum_{i=1}^{n}x_i\right)\log(1-\theta)\\ &= n[\bar{x}\log\theta+(1-\bar{x})\log(1-\theta)].\end{aligned}$$

Assim, a primeira derivada é dada por

$$n\left[\frac{\bar{x}}{\theta}-\frac{(1-\bar{x})}{(1-\theta)}\right]$$

e igualando a zero obtém-se que $\theta=\bar{x}$. A segunda derivada é dada por

$$-n\left[\frac{\bar{x}}{\theta^2}+\frac{(1-\bar{x})}{(1-\theta)^2}\right]<0$$

de modo que o EMV de θ é $\hat{\theta}=\bar{X}$, i.e. a proporção amostral de sucessos. Como $E(X)=\theta$ segue que este estimador é também não viesado. Note que esta solução só vale se $0<\hat{\theta}<1$ pois assumimos que $0<\theta<1$. No entanto, quando $\bar{x}=0$ temos que $\log l(\theta;\boldsymbol{x})=n\log(1-\theta)$ que é uma função decrescente de θ e portanto é maximizada em $\theta=0$. Analogamente, se $\bar{x}=1$ temos que $\log l(\theta;\boldsymbol{x})=n\log(\theta)$ que é maximizada em $\theta=1$. Assim, $\bar{X}$ é o EMV de θ mesmo que a proporção amostral de sucessos seja 0 ou 1.

Exemplo 3.2 : Seja $X_1,\ldots,X_n$ uma amostra aleatória da distribuição $N(\theta,1)$. A função de verossimilhança é dada por

$$\begin{aligned}l(\theta;\boldsymbol{x}) &= p(\boldsymbol{x}|\theta)=\prod_{i=1}^{n}(2\pi)^{-1/2}\exp(-(x_i-\theta)^2/2)\\ &= (2\pi)^{-n/2}\exp\left\{-\sum_{i=1}^{n}(x_i-\theta)^2/2\right\}\end{aligned}$$

e o logaritmo da verossimilhança é dado por

$$\log l(\theta;\boldsymbol{x})=-\frac{n}{2}\log(2\pi)-\sum_{i=1}^{n}(x_i-\theta)^2/2.$$

Tomando a primeira derivada e igualando a zero obtém-se a equação de verossim-

ilhança

$$\sum_{i=1}^{n}(x_i - \theta) = 0$$

cuja solução é $\theta = \sum_{i=1}^{n} x_i/n$. A segunda derivada é $-n < 0$ de modo que o EMV de θ é $\hat{\theta} = \bar{X}$. Além disso o estimador é não viesado para θ. Note que aqui não precisamos nos preocupar com valores extremos (como no exemplo anterior) pois o espaço paramétrico é ilimitado.

Exemplo 3.3 : Seja $X_1, \ldots, X_n$ uma amostra aleatória da distribuição $U(0, \theta)$, $\theta > 0$. A função de densidade é dada por

$$p(\boldsymbol{x}|\theta) = \begin{cases} 1/\theta^n, & 0 \leq x_i \leq \theta,\ i = 1, \ldots, n \\ 0, \text{caso contrário.} \end{cases}$$

Assim, a verossimilhança é uma função estritamente decrescente de θ e portanto seu máximo é atingido quando θ assume o menor dos seus possíveis valores. Esta condição é satisfeita quando $\theta = \max(x_1, \ldots, x_n)$, i.e. o EMV é $\hat{\theta} = \max(X_1, \ldots, X_n)$. Por outro lado a função de densidade poderia ser definida como

$$p(x|\theta) = \begin{cases} 1/\theta^n, & 0 < x_i < \theta,\ i = 1, \ldots, n \\ 0, \text{caso contrário.} \end{cases}$$

Neste caso, $\max(X_1, \ldots, X_n)$ não é um dos possíveis valores de θ já que $\theta > x_i$, $i = 1, \ldots, n$, i.e. $\theta > \max(X_1, \ldots, X_n)$. Portanto, o EMV não existe.

Exemplo 3.4 : Seja $X_1, \ldots, X_n$ uma amostra aleatória da distribuição $U(\theta, \theta + 1)$, $-\infty < \theta < \infty$. A função de densidade é dada por

$$p(\boldsymbol{x}|\theta) = \begin{cases} 1, & \theta \leq x_i \leq \theta + 1,\ i = 1, \ldots, n \\ 0, \text{caso contrário.} \end{cases}$$

A condição $\theta \leq x_i$ para $i = 1, \ldots, n$ é equivalente a $\theta \leq \min(x_1, \ldots, x_n)$ e a condição $x_i \leq \theta + 1$ para $i = 1, \ldots, n$ é equivalente a $\max(x_1, \ldots, x_n) \leq \theta + 1$. Assim, a função de densidade pode ser reescrita como

$$p(\boldsymbol{x}|\theta) = \begin{cases} 1,\ \max(x_1, \ldots, x_n) - 1 \leq \theta \leq \min(x_1, \ldots, x_n) \\ 0, \text{caso contrário.} \end{cases}$$

e qualquer valor de θ no intervalo $[\max(x_1, \ldots, x_n) - 1, \min(x_1, \ldots, x_n)]$ maximiza a função de verossimilhança. Em outras palavras, o EMV não é único.

Exemplo 3.5 : Seja $X_1, \ldots, X_n$ uma amostra aleatória da distribuição $N(\mu, \sigma^2)$.

A função de verossimilhança é dada por

$$\begin{aligned} l(\mu, \sigma^2; \boldsymbol{x}) &= p(\boldsymbol{x}|\mu, \sigma^2) = \prod_{i=1}^{n} (2\pi\sigma^2)^{-1/2} \exp(-(x_i - \mu)^2/2\sigma^2) \\ &= (2\pi\sigma^2)^{-n/2} \exp\left\{ -\sum_{i=1}^{n} (x_i - \mu)^2/2\sigma^2 \right\} \end{aligned}$$

e o logaritmo da verossimilhança é dado por

$$L(\mu, \sigma^2; \boldsymbol{x}) = \log l(\mu, \sigma^2; \boldsymbol{x}) = -\frac{n}{2} \log(2\pi\sigma^2) - \sum_{i=1}^{n} (x_i - \mu)^2/2\sigma^2.$$

Tomando a primeira derivada e igualando a zero obtém-se as seguintes equações de verossimilhança

$$\frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \mu) = \frac{n}{\sigma^2} (\bar{x} - \mu) = 0$$

$$-\frac{n}{\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^{n} (x_i - \mu)^2 = 0.$$

A solução da primeira equação é $\hat{\mu} = \bar{x}$ e a solução da segunda equação avaliada em $\hat{\mu} = \bar{x}$ é $\hat{\sigma}^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2/n$. As segundas derivadas avaliadas em $\hat{\mu}$ e $\hat{\sigma}^2$ são dadas por

$$\frac{-n}{\hat{\sigma}^2} < 0, \quad -\frac{n(\bar{x} - \hat{\mu})}{\hat{\sigma}^4} = 0 \quad \text{e} \quad \frac{n}{2\hat{\sigma}^4} - \frac{\sum_{i=1}^{n} (x_i - \hat{\mu})^2}{\hat{\sigma}^6} = -\frac{n}{\hat{\sigma}^4} < 0.$$

Conclui-se então que $\bar{X}$ e $\sum_{i=1}^{n} (X_i - \bar{X})^2/n$ são os EMV de μ e σ^2 respectivamente.

EMV e estatísticas suficientes

Se $X_1, \ldots, X_n$ é uma amostra aleatória de $p(x|\theta)$ e $T(\boldsymbol{X})$ é uma estatística suficiente para θ então, pelo critério de fatoração, a função de verossimilhança é dada por

$$l(\theta; \boldsymbol{x}) = f(t, \theta) g(\boldsymbol{x}).$$

Como $g(\boldsymbol{x})$ é constante em relação a θ então o valor $\hat{\theta}$ que maximiza $l(\theta; \boldsymbol{x})$ é o mesmo que maximiza $f(t, \theta)$, que depende de $\boldsymbol{x}$ somente através de $t(\boldsymbol{x})$. Assim $\hat{\theta}$ será necessariamente uma função de t e concluimos que o EMV é sempre função de uma estatística suficiente.

Invariância

Seja $X_1, \ldots, X_n$ uma amostra aleatória de $p(x|\theta)$ e $\hat{\theta}$ é o EMV de θ. Suponha que queremos inferir o valor de $\phi = g(\theta)$ onde g é uma função 1 a 1 (ou bijetora) de θ. Se $\theta = h(\phi)$ é a função inversa e $\hat{\phi}$ é o EMV de ϕ então $h(\hat{\phi})$ maximiza $p(x|h(\phi))$. Por outro lado $\hat{\theta}$ também maximiza $p(x|h(\phi))$, i.e. $h(\phi) = \hat{\theta}$ e portanto $h(\hat{\phi}) = \hat{\theta}$ ou equivalentemente $\hat{\phi} = g(\hat{\theta})$.

Conclui-se então que $g(\hat{\theta})$ é o EMV de $g(\theta)$. Esta propriedade é chamada *princípio da invariância*.

Exemplo 3.6 : No Exemplo 3.5, pelo princípio da invariância segue que o EMV de σ é $\hat{\sigma} = \sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2/n}$.

Exemplo 3.7 : Seja $X_1, \ldots, X_n \sim N(\theta, 1)$ e queremos estimar a probabilidade $g(\theta) = P(X < 0)$. Como $\hat{\theta} = \bar{X}$ é o EMV de θ e $P(X < 0) = P(X - \theta < -\theta) = \Phi(-\theta)$ então pelo princípio da invariância o EMV de $P(X < 0)$ é $\Phi(-\bar{X})$.

Exemplo 3.8 : Seja $X_1, \ldots, X_n$ uma amostra aleatória da distribuição exponencial com parâmetro β e queremos estimar a probabilidade $g(\beta) = P(X > 1)$. O EMV de β é $\hat{\beta} = 1/\bar{X}$ e a função de distribuição de X é $P(X < x) = 1 - e^{-\beta x}$, portanto $P(X > 1) = 1 - P(X < 1) = e^{-\beta}$. Assim, pelo princípio da invariância o EMV de $P(X > 1)$ é

$$g(\hat{\beta}) = e^{-\hat{\beta}} = e^{-1/\bar{X}}.$$

O EMV não depende do plano amostral

Se dois experimentos dão origem a funções de verossimilhança $l_1(\theta)$ e $l_2(\theta)$ que são proporcionais, i.e. $l_1(\theta) = k\ l_2(\theta)$, $k > 0$ e k não depende de θ, então o EMV de θ é o mesmo.

Exemplo 3.9 : O tempo (em minutos) entre chegadas de clientes em um banco é denotado pela variável aleatória $X \sim Exp(\beta)$. Deseja-se estimar o tempo médio entre chegadas a partir de uma amostra aleatória $X_1, \ldots, X_n$. O EMV de β é $\hat{\beta} = 1/\bar{X}$ e pela propriedade de invariância segue que o EMV de $\mu = E(X) = 1/\beta$ é $\hat{\mu} = 1/\hat{\beta} = \bar{X}$. Para uma amostra de tamanho $n = 20$ dois planos amostrais poderiam ter sido utilizados,

(i) Fixar $n = 20$ a priori.

(ii) Observar $X_1, X_2, \ldots$ até obter um tempo superior a 10 minutos.

Suponha que no segundo experimento observou-se $x_i < 10$, $i = 1, \ldots, 19$ e $x_{20} > 10$ e em ambos a média amostral foi igual 6 minutos. Então a estimativa de

máxima verossimilhança do tempo médio entre chegadas é $\bar{x} = 6$ não importando como a amostra foi obtida.

Diz-se que o método satisfaz ao chamado *princípio da verossimilhança*. Este princípio postula que, para fazer inferências sobre uma quantidade desconhecida θ só importa aquilo que foi realmente observado e não aquilo que "poderia" ter ocorrido mas efetivamente não ocorreu.

Observações incompletas

Em muitas situações práticas os dados fornecem informações incompletas sobre determinado fenômeno. Isto ocorre em geral quando o experimento precisa ser terminado por algum motivo de ordem prática e que pode ou não estar sob controle do pesquisador. Esta observação parcial dos dados é chamada de *censura* e os métodos para descrição e modelagem deste tipo de dados é chamada de *análise de sobrevivência* ou *análise de confiabilidade*. Esta informação parcial deve ser levada em conta ao se tentar estimar os parâmetros de interesse.

Exemplo 3.10 : No Exemplo 3.9, o tempo até a chegada do próximo cliente será observado até que: o cliente chegue ou o expediente se encerre, o que ocorrer primeiro. Suponha que esperou-se 15 minutos e o expediente se encerrou sem que ninguém tenha aparecido. Ou seja, X_{21} não foi observado mas sabe-se que $X_{21} > 15$. A média amostral baseada em 21 observações é maior do que 6 e a estimativa de máxima verossimilhança é obtida maximizando-se

$$p(x_1|\beta)\dots p(x_n|\beta)P(X_{21} > 15) = \beta^{20}\exp(-\beta\sum_{i=1}^{20} x_i)\exp(-15\beta).$$

Do Exemplo 3.9 temos que $\bar{x} = 6$ então o tempo total de espera dos 20 primeiros clientes foi $\sum_{i=1}^{20} x_i = 120$ e a função de verossimilhança fica $\beta^{20}e^{-135\beta}$.

Solução numérica

Em muitas situações práticas a função de verossimilhança está associada a modelos complexos e a equação de verossimilhança não apresenta solução analítica explicita.

Exemplo 3.11 : Suponha que uma variável aleatória X tem função de densidade $f(x) = \sum_{j=1}^{k} p_j f_j(x)$, sendo $p_j > 0$ e $\sum_{j=1}^{k} p_j = 1$. Para uma amostra aleatória $X_1, \dots, X_n$ a função de verossimilhança fica

$$f(\boldsymbol{x}) = \prod_{i=1}^{n}\left(\sum_{j=1}^{k} p_j f_j(x_i)\right).$$

Mesmo que as funções $f_j(x)$ sejam completamente conhecidas não há solução de máxima verossimilhança para os pesos p_j.

Exemplo 3.12 : Suponha que $X \sim Gama(\alpha, \beta)$. Para uma amostra aleatória $X_1, \ldots, X_n$ o logaritmo da função de verossimilhança fica

$$L(\alpha, \beta; \boldsymbol{x}) = \log\left(\prod_{i=1}^{n} \frac{\beta^\alpha}{\Gamma(\alpha)} x_i^\alpha \exp(-\beta x_i)\right) = n \log\left[\frac{\beta^\alpha}{\Gamma(\alpha)}\right] + \alpha \sum_{i=1}^{n} log(x_i) - \beta \sum_{i=1}^{n} x_i$$

e $\partial L(\alpha, \beta; \boldsymbol{x})/\partial\alpha = 0$ não tem solução analítica explícita.

Nestes casos pode-se recorrer a métodos numéricos para obter o EMV de um parâmetro θ. Lembrando que a função escore é definida como

$$U(\boldsymbol{X}; \theta) = \frac{\partial \log l(\theta; \boldsymbol{x})}{\partial \theta}$$

então, se $\hat{\theta}$ é o EMV de θ segue que $U(\boldsymbol{X}; \hat{\theta}) = 0$. Expandindo $U(\boldsymbol{X}; \hat{\theta})$ em série de Taylor em torno de θ_0 obtemos que

$$0 = U(\boldsymbol{X}; \hat{\theta}) = U(\boldsymbol{X}; \theta_0) + (\hat{\theta} - \theta_0) U'(\boldsymbol{X}; \theta_0) + \ldots$$

e desprezando os termos de ordem mais alta então para valores de $\hat{\theta}$ e θ_0 próximos segue que

$$0 = U(\boldsymbol{X}; \hat{\theta}) \approx U(\boldsymbol{X}; \theta_0) + (\hat{\theta} - \theta_0) U'(\boldsymbol{X}; \theta_0).$$

Resolvendo para $\hat{\theta}$ segue que

$$\hat{\theta} \approx \theta_0 - \frac{U(\boldsymbol{X}; \theta_0)}{U'(\boldsymbol{X}; \theta_0)} = \theta_0 + \frac{U(\boldsymbol{X}; \theta_0)}{J(\theta_0)}$$

onde $J(\cdot)$ é a informação observada de Fisher.

Assim, a partir de um valor inicial $\theta^{(0)}$ um procedimento iterativo para busca de máximo é dado por

$$\theta^{(j+1)} = \theta^{(j)} - \frac{U(\boldsymbol{X}; \theta^{(j)})}{U'(\boldsymbol{X}; \theta^{(j)})} = \theta^{(j)} + \frac{U(\boldsymbol{X}; \theta^{(j)})}{J(\theta^{(j)})}$$

que deve ser repetido até que o processo se estabilize segundo algum critério de convergência. Um critério tipicamente utilizado é $|\theta^{(j+1)} - \theta^{(j)}| < \epsilon$ onde ϵ é especificado arbitrariamente. Este é o conhecido algoritmo de Newton-Raphson e o ponto $\hat{\theta}$ aonde o algoritmo se estabiliza é tomado como a estimativa de máxima verossimilhança.

Uma modificação do algoritmo acima é obtida substituindo-se a informação

observada, $J(\theta)$, pela informação esperada de Fisher, $I(\theta)$. Sob algumas condições de regularidade, tipicamente verificadas na prática, este método modificado converge para o estimador de máxima verossimilhança.

Distribuição assintótica

Em muitas situações a equação de verossimilhança tem solução analítica explícita porém o EMV é uma função complicada da amostra. Neste caso, pode não ser uma tarefa fácil obter a distribuição do estimador ou verificar sua eficiência. Uma alternativa é estudar o comportamento do estimador quando o tamanho da amostra n tende a infinito (comportamento assintótico). Como na prática o tamanho amostral é finito os resultados obtidos são aproximadamente corretos para n suficientemente grande.

Pode-se mostrar que, sob condições de regularidade

$$\hat{\theta} \sim N(\theta, I^{-1}(\theta)), \quad \text{quando } n \to \infty.$$

A prova deste resultado está além do escopo destas notas e será omitida (ver Migon and Gamerman 1999). Na prática, i.e. para n finito, dizemos que para n suficientemente grande, o estimador de máxima verossimilhança $\hat{\theta}$ tem distribuição aproximadamente $N(\theta, I^{-1}(\theta))$. Ou seja, o EMV é sempre assintoticamente não viesado e eficiente já que sua esperança tende para θ e sua variância tende para o limite inferior da desigualdade de Cramer-Rao. Além disso, ele é consistente já que $Var(\hat{\theta}) \to 0$ quando $n \to \infty$.

O resultado pode ser generalizado para uma função $g(\theta)$, i.e.

$$g(\hat{\theta}) \sim N\left(g(\theta), \frac{[g'(\theta)]^2}{I(\theta)}\right), \quad \text{quando } n \to \infty.$$

Exemplo 3.13 : Suponha uma única observação X da distribuição binomial com parâmetros n e θ desconhecido. O EMV de θ é $\hat{\theta} = X/n$ e a informação de Fisher é $n/[\theta(1-\theta)]$ (verifique). Portanto, para n grande a distribuição aproximada da variável aleatória

$$\frac{\sqrt{n}(\hat{\theta} - \theta)}{\sqrt{\theta(1-\theta)}}$$

é $N(0, 1)$.

3.1.1 Comentários

Em muitas situações a função de verossimilhança pode ser muito difícil ou mesmo impossível de ser calculada. Assim, obter estimativas de máxima verossimilhança

e principalmente quantificar a incerteza associada pode ser uma tarefa complexa. Por outro lado a tendência atual é de propor modelos cada vez mais complexos para analisar conjuntos dados em quase todas as áreas da ciência (e.g. dados espacialmente distribuidos).

Alguns fatores que podem levar a dificuldades práticas no processo de estimação são,

- dados faltantes ou incompletos;
- função de verossimilhança complexa, com um número grande de parâmetros ou uma forma funcional computacionalmente intratável (e.g. modelos probito multinomiais, modelos de séries temporais para dados qualitativos);
- maximização pode ser extremamente lenta;
- não existência de um máximo único, ou máximo localizado no extremo do espaço dos parâmetros (e.g. modelos de misturas finitas).

Felizmente vários métodos computacionalmente intensivos (Bootstrap, algoritmo EM, métodos de Monte Carlo, algoritmos genéticos, etc) foram e continuam sendo desenvolvidos ou adaptados para tratar de situações cada vez mais complexas (e portanto mais realistas). Os recursos computacionais atualmente disponíveis vem contribuindo muito para disseminar o uso destas técnicas.

3.1.2 Problemas

1. Deseja-se estimar a proporção θ de mulheres em cursos de graduação em Estatística no Brasil. Uma amostra aleatória de 90 alunos matriculados foi selecionada e obteve-se que 58 eram mulheres e 32 eram homens. Encontre a estimativa de máxima verossimilhança de θ.

2. No exercício anterior sabe-se que $1/2 < \theta < 3/5$. Qual a estimativa de máxima verossimilhança de θ para aquela amostra.

3. Seja $X_1, \ldots, X_n$ uma amostra aleatória da distribuição de Bernoulli com parâmetro θ $(0 < \theta < 1)$. Mostre que o EMV de θ não existe se os valores observados forem todos iguais a 1 ou todos iguais a 0.

4. Seja $X_1, \ldots, X_n$ uma amostra aleatória da distribuição de Poisson com parâmetro λ desconhecido $(\lambda > 0)$.

   (a) Obtenha o EMV de λ assumindo que pelo menos um valor observado é diferente de zero.

   (b) Mostre que o EMV de λ não existe se todos os valores observados forem nulos.

5. Seja $X_1, \ldots, X_n$ uma amostra aleatória da distribuição $N(\mu, \sigma^2)$, com média conhecida e variância desconhecida. Obtenha o EMV de σ^2 e verifique se ele é não viesado.

6. Seja $X_1, \ldots, X_n$ uma amostra aleatória da distribuição exponencial com parâmetro β desconhecido ($\beta > 0$). Obtenha o EMV de β.

7. Seja $X_1, \ldots, X_n$ uma amostra aleatória da distribuição cuja função de densidade é dada por

$$p(x|\theta) = \begin{cases} \theta x^{\theta-1}, \ 0 < x < 1, \ \theta > 0 \\ 0, \ \text{caso contrário.} \end{cases}$$

   (a) Obtenha os EMV de θ e $g(\theta) = \theta/(1+\theta)$.

   (b) Obtenha as distribuições aproximadas destes estimadores para n grande.

8. Seja uma amostra aleatória $X_1, \ldots, X_n$ da distribuição $N(\theta, 1)$. Obtenha o EMV de $g(\theta) = P(X > 0)$ e sua distribuição aproximada quando n é grande.

9. Seja $X_1, \ldots, X_n$ uma amostra aleatória da distribuição de Poisson com média desconhecida. Obtenha o EMV do desvio padrão da distribuição.

10. O tempo de vida de um tipo de lâmpada tem distribuição exponencial com parâmetro β desconhecido. Uma amostra aleatória de n lâmpadas foi testada durante T horas e observou-se o número X de lâmpadas que falharam. Obtenha o EMV de β baseado em X.

11. Suponha que 21 observações são tomadas ao acaso de uma distribuição exponencial com média μ desconhecida. A média amostral de 20 observações foi igual a 6 e o valor da outra observação é desconhecido mas sabe-se que é maior do que 15. Calcule o EMV de μ.

12. Dois estatísticos precisam estimar uma quantidade desconhecida $\theta > 0$. O estatístico A observa uma variável aleatória $X \sim Gama(3, \theta)$ e o estatístico B observa uma variável aleatória Y com distribuição de Poisson e média 2θ. Se os valores observados foram $X = 2$ e $Y = 3$ mostre que as funções de verossimilhança são proporcionais e obtenha o EMV de θ.

3.2 Método dos Momentos

O método dos momentos para estimação de parâmetros é bastante simples e intuitivo. Basicamente, ele preconiza a estimação de momentos populacionais

(não observáveis) por seus equivalentes momentos amostrais. Assim, para uma variável aleatória X cuja distribuição depende de um parâmetro θ com momentos de ordem k dados por

$$\mu_k = E(X^k|\theta)$$

e uma amostra aleatória $X_1, \ldots, X_n$ desta distribuição, o método preconiza a estimação de μ_k por

$$\hat{\mu}_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k.$$

Qualquer outra função de θ é estimada a partir de sua relação com os momentos. Para um vetor de parâmetros $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_r)'$ os estimadores são obtidos como solução do sistema de equações criado igualando-se os r primeiros momentos amostrais e populacionais,

$$\hat{\mu}_k = \mu_k, \quad k = 1, \ldots, r.$$

Não é difícil verificar que o método sempre produz estimadores não viesados para os momentos populacionais, i.e.

$$E(\hat{\mu}_k) = \frac{1}{n}\sum_{i=1}^{n} E(X_i^k) = \mu_k.$$

com variância dada por

$$\begin{aligned} Var(\hat{\mu}_k) &= Var\left(\frac{1}{n}\sum_{i=1}^{n} X_i^k\right) \\ &= \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i^k) \\ &= \frac{1}{n^2}\sum_{i=1}^{n} \left[E(X_i^{2k}) - E^2(X_i^k)\right] \\ &= \frac{\mu_{2k} - \mu_k^2}{n}. \end{aligned}$$

O método também tem boas propriedades assintóticas já que as leis dos grandes números garantem que $\hat{\mu}_k \to \mu_k$ com probabilidade 1 quando $n \to \infty$.

Exemplo 3.14 : Seja uma amostra aleatória $X_1, \ldots, X_n$ tomada de uma distribuição com $E(X) = \mu_1$ e $Var(X) = \sigma^2$. Pelo método dos momentos, a média

populacional é estimada por $\bar{X}$ e o segundo momento é estimado por

$$\hat{\mu}_2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2.$$

Como $\sigma^2 = \mu_2 - \mu_1^2$ segue que a variância populacional é estimada por

$$\hat{\sigma}^2 = \hat{\mu_2} - \hat{\mu_1}^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2 = \frac{1}{n}\left(\sum_{i=1}^{n} X_i^2 - n\bar{X}^2\right) = \frac{1}{n}\sum_{i=1}^{n}(X_i^2 - \bar{X})^2.$$

Assim, os estimadores da média e da variância coincidem com os EMV no caso normal.

Exemplo 3.15 : Seja uma amostra aleatória $X_1, \ldots, X_n$ tomada de uma distribuição Gama com parâmetros α e β. A média e a variância populacionais são dados por

$$E(X) = \alpha/\beta \quad \text{e} \quad Var(X) = \alpha/\beta^2.$$

Portanto, pelo método dos momentos os estimadores para α e β são obtidos como solução das equações

$$\begin{aligned} \hat{\alpha}/\hat{\beta} &= \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X} \\ \hat{\alpha}/\hat{\beta}^2 + \hat{\alpha}^2/\hat{\beta}^2 &= \frac{1}{n}\sum_{i=1}^{n} X_i^2 \end{aligned}$$

A segunda equação pode ser reescrita como

$$\frac{\hat{\alpha}}{\hat{\beta}}\left(\frac{1}{\hat{\beta}} + \frac{\hat{\alpha}}{\hat{\beta}}\right) = \bar{X}\left(\frac{1}{\hat{\beta}} + \bar{X}\right) = \frac{1}{n}\sum_{i=1}^{n} X_i^2$$

obtendo-se

$$\frac{1}{\hat{\beta}} = \frac{\sum_{i=1}^{n} X_i^2/n}{\bar{X}} - \bar{X} = \frac{\sum_{i=1}^{n}(X_i^2 - \bar{X})^2/n}{\bar{X}} \Rightarrow \hat{\beta} = \frac{\bar{X}}{\sum_{i=1}^{n}(X_i^2 - \bar{X})^2/n}.$$

Substituindo na primeira equação obtém-se que

$$\hat{\alpha} = \frac{\bar{X}^2}{\sum_{i=1}^{n}(X_i^2 - \bar{X})^2/n}.$$

Neste exemplo, estimadores de máxima verossimilhança não podem ser obtidos explicitamente e métodos computacionais devem ser utilizados. Assim, uma possível aplicação do métodos dos momentos é utilizar este resultado para obter

valores iniciais em algoritmos de busca pelo máximo da função de verossimilhança.

3.3 Estimadores de Mínimos Quadrados

Seja agora uma amostra aleatória $Y_1, \ldots, Y_n$ tomada de uma distribuição tal que $E(Y_i|\theta) = f_i(\theta)$ e $Var(Y_i|\theta) = \sigma^2$. Ou seja, a média de cada Y_i assume uma forma específica, que pode depender de outras variáveis, e as variâncias são as mesmas. Uma forma equivente é

$$Y_i = f_i(\theta) + \epsilon_i$$

com $E(\epsilon_i) = 0$ e $Var(\epsilon_i) = \sigma^2$ para $i = 1, \ldots, n$.

O critério adotado aqui consiste em estimar θ de modo a minimizar os erros cometidos, $Y_i - f_i(\theta)$, minimizando uma função destes erros. Uma função que penaliza igualmente erros positivos e negativos e é comumente utilizada é a função quadrática. Assim, o critério pode ser expresso como, obter θ que minimiza

$$S(\theta) = \sum_{i=1}^{n} (Y_i - f_i(\theta))^2.$$

O valor $\hat{\theta}$ obtido é chamado de *estimador de mínimos quadrados* (EMQ) de θ.

Exemplo 3.16 : Regressão linear simples. Suponha que os valores da variável de interesse Y são afetados linearmente pelos valores de uma outra variável conhecida X. Dados n valores de X e Y um possível modelo para este problema é $E(Y_i) = \beta X_i$ e o EMQ do parâmetro β é obtido minimizando-se

$$S(\beta) = \sum_{i=1}^{n} (Y_i - \beta X_i)^2.$$

Derivando e igualando a zero esta soma de quadrados obtém-se que

$$-2 \sum_{i=1}^{n} (Y_i - \beta X_i)(X_i) = 0 \Leftrightarrow \beta = \frac{\sum_{i=1}^{n} Y_i X_i}{\sum_{i=1}^{n} X_i^2}$$

e como a segunda derivada é dada por $2 \sum_{i=1}^{n} X_i^2 > 0$ segue que o EMQ de β é

$$\hat{\beta} = \frac{\sum_{i=1}^{n} Y_i X_i}{\sum_{i=1}^{n} X_i^2}.$$

Note como nenhuma distribuição de probabilidades foi assumida para que o

método pudesse ser aplicado. Este é um dos motivos para sua grande utilização. Por outro lado, se os dados tiverem distribuição normal o procedimento coincide com a estimação de máxima verossimilhança, i.e. pode-se mostrar que minimizar a soma de quadrados dos erros é equivalente a maximizar a função de verossimilhança.

Outro fato importante é que o peso atribuido a cada observação na soma de quadrados foi o mesmo já que todas têm a mesma variância. O método pode ser estendido ao caso de variâncias desiguais e conhecidas a menos de uma constante, i.e. $Var(Y_i|\theta) = \sigma^2/w_i$. Neste caso a soma de quadrados a ser minimizada é

$$S(\theta) = \sum_{i=1}^{n} w_i(Y_i - f_i(\theta))^2$$

e observações com maior variância (menor w_i) terão um peso menor na estimação. Este procedimento é chamada de *estimação por mínimos quadrados ponderados*. O método anterior (sem ponderação) é então chamado de *estimação por mínimos quandrados ordinários* e é um caso particular onde todos os pesos são iguais a 1.

Exemplo 3.17 : No Exemplo 3.16 o estimador de mínimos quadrados ponderados de β é dado por

$$\hat{\beta} = \frac{\sum_{i=1}^{n} w_i Y_i X_i}{\sum_{i=1}^{n} w_i X_i^2}.$$

Finalmente, vale notar que a função $f_i(\theta)$ pode assumir várias formas distintas. Por exemplo, se f_i for um polinômio de ordem k em uma variável X conhecida, i.e. $\beta_0 + \beta_1 X + \beta_2 X^2 + \cdots + \beta_k X^k$ então os EMQ de $\beta_0, \beta_1, \ldots, \beta_k$ são obtidos minizando-se

$$S(\beta) = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i - \beta_2 X_i^2 - \cdots - \beta_k X_i^k)^2.$$

Por outro lado, se f_i define uma dependência linear em k variáveis conhecidas $X_1, \ldots, X_k$, i.e. $\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$ então os EMQ de $\beta_0, \beta_1, \ldots, \beta_k$ são obtidos minizando-se

$$S(\beta) = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_{i1} - \beta_2 X_{i2} - \cdots - \beta_k X_{ik})^2.$$

Em ambos os casos teremos um vetor de parâmetros $\beta_0, \beta_1, \ldots, \beta_k$ a serem estimados (além da variância σ^2) o que equivale a resolver um sistema de $k+1$ equações do tipo $\partial S/\partial \beta_j = 0$ para $j = 0, \ldots, k$.

3.4 Problemas

1. Seja $X_1, \ldots, X_n$ uma amostra aleatória tomada da distribuição Gama(θ,2). Obtenha um estimador para θ usando o método dos momentos.

2. Seja $X_1, \ldots, X_n$ uma amostra aleatória tomada da distribuição Exponencial(β). Obtenha um estimador para β usando o método dos momentos.

3. Seja $X_1, \ldots, X_n$ uma amostra aleatória tomada da distribuição Geométrica(p). Obtenha um estimador para p usando o método dos momentos.

4. Seja $X_1, \ldots, X_n$ uma amostra aleatória tomada da distribuição $N(\mu, \sigma^2)$. Obtenha estimadores de μ e σ^2 usando o método dos momentos. Obtenha o viés do estimador de σ^2.

5. Seja $X_1, \ldots, X_n$ uma amostra aleatória tomada da distribuição Gama(α, β). Obtenha estimadores de α e β usando o método dos momentos.

6. No Exemplo 3.16 mostre que o EMQ obtido é não viesado com variância $\sigma^2 / \sum_{i=1}^{n} X_i^2$.

7. No Exemplo 3.16 obtenha os EMQ de β_0 e β_1 supondo que $E(Y_i) = \beta_0 + \beta_1 X_i$ com variância constante.

8. Se $Y_i|\theta \sim N(f_i(\theta), \sigma^2)$ mostre que o EMV e o EMQ de θ coincidem.

Capítulo 4

Estimação Bayesiana

Considere uma amostra aleatória $X_1, \ldots, X_n$ tomada de uma distribuição de probabilidades com parâmetro θ desconhecido, $p(x|\theta)$. Em muitas situações, antes de observar a amostra o pesquisador tem condições de resumir sua informação e experiência anteriores sobre as chances de θ pertencer a determinadas regiões do espaço paramétrico. Este conhecimento pode ser quantificado construindo-se uma distribuição de probabilidades para θ, chamada *distribuição a priori*.

Exemplo 4.1 : Seja θ a probabilidade de obter cara quando uma moeda é lançada. Sabe-se que a moeda é honesta ou tem duas caras, i.e. os dois possíveis valores de θ são $1/2$ e 1. Se a probabilidade a priori de que a moeda seja honesta é p então a distribuição a priori de θ é $p(\theta = 1/2) = p$ e $p(\theta = 1) = 1 - p$.

Exemplo 4.2 : A proporção θ de itens defeituosos em um grande lote é desconhecida e supõe-se que os possíveis valores de θ se distribuem uniformemente no intervalo (0,1). A distribuição a priori é então dada por $\theta \sim U(0, 1)$ ou

$$p(\theta) = \begin{cases} 1, & 0 < \theta < 1 \\ 0, & \text{caso contrário.} \end{cases}$$

Exemplo 4.3 : O tempo de vida de um certo tipo de lâmpada tem distribuição exponencial com parâmetro θ. Com base em experiências anteriores assume-se que a distribuição a priori de θ é Gama com média 0,0002 e desvio padrão 0,0001. Assim, a distribuição a priori é dada por $\theta \sim Gama(\alpha, \beta)$ onde os parâmetros α e β são tais que

$$\frac{\alpha}{\beta} = 0,0002 \quad \text{e} \quad \frac{\alpha}{\beta^2} = 0,0001^2$$

de onde se obtém que

$$\beta = \frac{0,0002}{0,0001^2} = 20\,000 \quad \text{e} \quad \alpha = 0,0002\beta = 4.$$

Portanto, a distribuição a priori de θ é dada por $\theta \sim Gama(4, 20\,000)$ ou equivalentemente,

$$p(\theta) = \begin{cases} \dfrac{20\,000^4}{3!}\theta^3 e^{-20\,000\theta}, \ \theta > 0 \\ \\ 0, \ \theta \leq 0. \end{cases}$$

4.1 Distribuição a Posteriori

Por simplicidade vamos assumir que todas as quantidades envolvidas são contínuas de modo que $p(\boldsymbol{x}|\theta)$ e $p(\theta)$ são funções de densidade de probabilidade. Multiplicando estas duas densidades obtém-se a densidade conjunta de $X_1, \ldots, X_n$ e θ, i.e.

$$p(\boldsymbol{x}, \theta) = p(\boldsymbol{x}|\theta)p(\theta).$$

A função de densidade conjunta marginal de $X_1, \ldots, X_n$ pode ser obtida por integração como

$$p(\boldsymbol{x}) = \int p(\boldsymbol{x}|\theta)p(\theta)d\theta.$$

Além disso, do cálculo das probabilidades, a função de densidade condicional de θ dados $x_1, \ldots, x_n$ é dada por

$$p(\theta|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\theta)p(\theta)}{p(\boldsymbol{x})} = \frac{1}{p(\boldsymbol{x})} \times p(\boldsymbol{x}|\theta)p(\theta). \tag{4.1}$$

A função de densidade (4.1) está representando a distribuição de θ após os dados serem observados, e portanto é chamada *distribuição a posteriori* de θ. Todos estes resultados valem também para distribuições discretas de probabilidade.

Note que $1/p(\boldsymbol{x})$ em (4.1) não depende de θ e funciona como uma constante normalizadora de $p(\theta|\boldsymbol{x})$. Assim, a forma usual do teorema de Bayes é

$$p(\theta|x) \propto p(\boldsymbol{x}|\theta)p(\theta). \tag{4.2}$$

Em palavras temos que

$$\text{distribuição a posteriori} \propto \text{verossimilhança} \times \text{distribuição a priori.}$$

Ou seja, ao omitir o termo $p(\boldsymbol{x})$, a igualdade em (4.1) foi substituída por uma proporcionalidade. Esta forma simplificada do teorema de Bayes será útil em

problemas que envolvam estimação de parâmetros já que o denominador é apenas uma constante normalizadora.

É intuitivo também que a probabilidade a posteriori de um particular conjunto de valores de θ será pequena se $p(\theta)$ ou $p(\boldsymbol{x}|\theta)$ for pequena para este conjunto. Em particular, se atribuirmos probabilidade a priori igual a zero para um conjunto de valores de θ então a probabilidade a posteriori será zero qualquer que seja a amostra observada.

Exemplo 4.4 : No Exemplo 4.2 suponha que uma amostra aleatória $X_1, \ldots, X_n$ é tomada do lote, onde $X_i = 1$ se o item i for defeituoso e $X_i = 0$ caso contrário para $i = 1, \ldots, n$. Assim,

$$p(\boldsymbol{x}|\theta) = \begin{cases} \theta^y(1-\theta)^{n-y}, \ x_i = 0, 1, \ i = 1, \ldots, n \\ 0, \ \text{caso contrário} \end{cases}$$

onde $y = \sum_{i=1}^{n} x_i$. Como a distribuição a priori é uniforme no intervalo (0,1) segue que

$$p(\theta|\boldsymbol{x})p(\theta) = \begin{cases} \theta^y(1-\theta)^{n-y}, \ y \geq 0, \ 0 < \theta < 1 \\ 0, \ \text{caso contrário} \end{cases}$$

Por comparação pode-se notar que, a menos de uma constante (que não depende de θ), o lado direito desta expressão tem a forma da função de densidade de uma distribuição Beta com parâmetros $\alpha = y + 1$ e $\beta = n - y + 1$. Assim, como a distribuição a posteriori de θ é proporcional ao lado direito desta expressão conclui-se que

$$\theta|\boldsymbol{x} \sim Beta(y + 1, n - y + 1).$$

Exemplo 4.5 : No Exemplo 4.3 suponha que uma amostra aleatória $X_1, \ldots, X_n$ com os tempos de vida de n lâmpadas é tomada. Neste caso, definindo $y = \sum_{i=1}^{n} x_i$, a densidade conjunta para $x_i > 0$, $i = 1, \ldots, n$ é

$$p(\boldsymbol{x}|\theta) = \theta^n e^{-\theta y}.$$

Usando o teorema de Bayes na forma (4.2) segue que

$$\begin{aligned} p(\theta|\boldsymbol{x}) &\propto \theta^n e^{-\theta y}\, \theta^3 e^{-20\,000\theta} \\ &\propto \theta^{3+n} e^{-\theta(20\,000+y)} \end{aligned}$$

onde todos os termos que não dependem de θ foram omitidos. Por comparação, o lado direito desta expressão tem a mesma forma da função de densidade de uma distribuição Gama com parâmetros $\alpha = n + 4$ e $\beta = 20\,000 + y$. Assim, para

$\theta > 0$ conclui-se que a distribuição a posteriori de θ é dada por

$$\theta|\boldsymbol{x} \sim Gama(n + 4, 20\,000 + y).$$

4.1.1 Observações Sequenciais

Uma questão que se coloca aqui é se a distribuição a posteriori depende da ordem em que as observações foram processadas. Observando-se as variáveis aleatórias $X_1, \ldots, X_n$, que são independentes dado θ e relacionadas a θ através de $p_i(x_i|\theta)$ segue que

$$\begin{aligned}
p(\theta|x_1) &\propto p_1(x_1|\theta)p(\theta) \\
p(\theta|x_2, x_1) &\propto p_2(x_2|\theta)p(\theta|x_1) \\
&\propto p_2(x_2|\theta)p_1(x_1|\theta)p(\theta) \\
&\vdots \quad \vdots \\
p(\theta|x_n, x_{n-1}, \cdots, x_1) &\propto \left[\prod_{i=1}^{n} p_i(x_i|\theta)\right] p(\theta) \\
&\propto p_n(x_n|\theta)\, p(\theta|x_{n-1}, \cdots, x_1).
\end{aligned}$$

Ou seja, a ordem em que as observações são processadas pelo teorema de Bayes é irrelevante. Na verdade, elas podem até ser processadas em subgrupos.

4.2 Problemas

1. Suponha que a proporção θ de itens defeituosos em um grande lote é igual a 0,1 ou 0,2 e que a função de probabilidade a priori de θ é $p(0,1) = 0,7$ e $p(0,2) = 0,3$. Se 8 itens foram selecionados ao acaso do lote e observou-se exatamente 2 defeituosos obtenha a distribuição a posteriori de θ.

2. Suponha que o número de defeitos em um tipo de fita magnética tem distribuição de Poisson com parâmetro λ cujo valor é igual a 1 ou 1,5. A distribuição a priori de λ é $p(1) = 0,4$ e $p(1,5) = 0,6$. Se uma fita selecionada ao acaso apresentou 3 defeitos obtenha a distribuição a posteriori de λ.

3. Suponha que a distribuição a priori de um parâmetros $\theta > 0$ é Gama com média 10 e variância 5. Determine os parâmetros desta distribuição a priori.

4. Suponha que a distribuição a priori de um parâmetros $\theta \in (0,1)$ é Beta com média 1/3 e variância 1/45. Determine os parâmetros desta distribuição a priori.

5. Suponha que a proporção θ de itens defeituosos em um grande lote é desconhecida e que sua distribuição a priori é uniforme no intervalo (0,1). Se 8 itens foram selecionados ao acaso do lote e observou-se exatamente 3 defeituosos obtenha a distribuição a posteriori de θ.

6. Considere novamente as condições do Problema 5 mas suponha que a função de densidade a priori de θ é

$$p(\theta) = \begin{cases} 2(1-\theta), & 0 < \theta < 1 \\ 0, & \text{caso contrário.} \end{cases}$$

   Determine a distribuição a posteriori de θ.

7. Suponha que uma única observação X é tomada da distribuição uniforme no intervalo $(\theta - 1/2, \theta + 1/2)$ e o valor de θ é desconhecido. Supondo que a distribuição a priori de θ é uniforme no intervalo (10,20) e observou-se $X = 12$ obtenha a distribuição a posteriori de θ.

4.3 Distribuições a Priori Conjugadas

A partir do conhecimento que se tem sobre θ, pode-se definir uma família paramétrica de distribuições. Neste caso, a distribuição a priori é representada por uma forma funcional, cujos parâmetros devem ser especificados de acordo com este conhecimento. Estes parâmetros indexadores da família de distribuições a priori são chamados de *hiperparâmetros* para distingui-los dos parâmetros de interesse θ.

Esta abordagem em geral facilita a análise e o caso mais importante é o de prioris conjugadas. A idéia é que as distribuições a priori e a posteriori pertençam a mesma classe de distribuições e assim a atualização do conhecimento que se tem de θ envolve apenas uma mudança nos hiperparâmetros. Neste caso, o aspecto sequencial do método Bayesiano pode ser explorado definindo-se apenas a regra de atualização dos hiperparâmetros já que as distribuições permanecem as mesmas.

A forma da distribuição conjugada depende da distribuição dos dados através da função de verossimilhança e alguns casos são listados a seguir.

4.3.1 Amostrando de um Distribuição de Bernoulli

Seja $X_1, \ldots, X_n$ uma amostra aleatória da distribuição de Bernoulli com parâmetro θ. Definindo $y = \sum_{i=1}^{n} x_i$ sua função de probabilidade conjunta para $x_i = 0, 1$,

$i = 1, \ldots, n$ é dada por

$$p(\boldsymbol{x}|\theta) = \theta^y(1-\theta)^{n-y}$$

e assumindo que a distribuição a priori é Beta com parâmetros $\alpha > 0$ e $\beta > 0$ então

$$p(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}.$$

Usando o teorema de Bayes, a distribuição a posteriori é dada por

$$\begin{aligned} p(\theta|\boldsymbol{x}) &\propto \theta^y(1-\theta)^{n-y}\theta^{\alpha-1}(1-\theta)^{\beta-1} \\ &\propto \theta^{\alpha+y-1}(1-\theta)^{\beta+n-y-1},\ 0 < \theta < 1. \end{aligned}$$

Exceto por uma constante que não depende de θ o lado direito desta expressão pode ser reconhecido como a função de densidade de uma distribuição Beta com parâmetros $\alpha + y$ e $\beta + n - y$. Portanto esta é a distribuição a posteriori de θ, i.e.

$$\theta|\boldsymbol{x} \sim Beta(\alpha + y, \beta + n - y).$$

Uma extensão direta é o modelo binomial, i.e. se $Y|\theta \sim$ Binomial(n, θ) então

$$p(y|\theta) \propto \theta^y(1-\theta)^{n-y}$$

e portanto a priori conjugada é Beta(α, β).

4.3.2 Amostrando de uma Distribuição de Poisson

Seja $X_1, \ldots, X_n$ uma amostra aleatória da distribuição de Poisson com parâmetro θ. Sua função de probabilidade conjunta é dada por

$$p(x|\theta) = \frac{e^{-n\theta}\theta^t}{\prod x_i!} \propto e^{-n\theta}\theta^t, \quad \theta > 0, \quad t = \sum_{i=1}^{n} x_i.$$

O núcleo da verossimilhança é da forma $\theta^a e^{-b\theta}$ que caracteriza a família de distribuições Gama. Assim, vamos assumir que a distribuição a priori é Gama com parâmetros positivos $\alpha > 0$ e $\beta > 0$, i.e.

$$p(\theta) \propto \theta^{\alpha-1}e^{-\beta\theta}, \quad \alpha, \beta > 0 \quad \theta > 0.$$

A densidade a posteriori fica

$$p(\theta|x) \propto \theta^{\alpha+t-1} \exp\{-(\beta+n)\theta\}$$

que corresponde à densidade Gama$(\alpha + t, \beta + n)$. Ou seja, a distribuição Gama é a priori conjugada para o modelo de Poisson.

4.3.3 Amostrando de uma Distribuição Exponencial

Seja $X_1, \ldots, X_n$ uma amostra aleatória da distribuição Exponencial com parâmetro θ. Sua função de densidade de probabilidade conjunta é dada por

$$p(\boldsymbol{x}|\theta) = e^{-\theta t}\theta^n, \quad \theta > 0, \quad t = \sum_{i=1}^{n} x_i.$$

O núcleo da verossimilhança é novamente da forma $\theta^a e^{-b\theta}$ e assim vamos assumir que a distribuição a priori é Gama com parâmetros positivos $\alpha > 0$ e $\beta > 0$. Neste caso a densidade a posteriori fica

$$p(\theta|\boldsymbol{x}) \propto \theta^{\alpha+n-1} \exp\{-(\beta + t)\theta\}$$

que corresponde à densidade Gama$(\alpha + n, \beta + t)$. Ou seja, a distribuição Gama é a priori conjugada para o modelo exponencial.

4.3.4 Amostrando de uma Distribuição Multinomial

Denotando por $\boldsymbol{X} = (X_1, \ldots, X_p)$ o número de ocorrências em cada uma de p categorias em n ensaios independentes, e por $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$ as probabilidades associadas deseja-se fazer inferência sobre estes p parâmetros. No entanto, note que existem efetivamente $p - 1$ parâmetros já que temos a seguinte restrição $\sum_{i=1}^{p} \theta_i = 1$. Além disso, a restrição $\sum_{i=1}^{p} X_i = n$ obviamente também se aplica. Dizemos que $\boldsymbol{X}$ tem distribuição multinomial com parâmetros n e $\boldsymbol{\theta}$ e a função de probabilidade conjunta das p contagens $\boldsymbol{X}$ é dada por

$$p(\boldsymbol{x}|\boldsymbol{\theta}) = \frac{n!}{\prod_{i=1}^{p} x_i!} \prod_{i=1}^{p} \theta_i^{x_i}.$$

Note que esta é uma generalização da distribuição binomial que tem apenas duas categorias. A função de verossimilhança para $\boldsymbol{\theta}$ é

$$l(\boldsymbol{\theta}; \boldsymbol{x}) \propto \prod_{i=1}^{p} \theta_i^{x_i}$$

que tem o mesmo núcleo da função de densidade de uma distribuição de Dirichlet (ver Apêndice A). Esta é uma generalização da distribuição Beta para um vetor aleatório com elementos definidos no intervalo (0,1). Usando esta distribuição

como priori para o vetor $\boldsymbol{\theta}$ então a função de densidade a priori é dada por

$$p(\boldsymbol{\theta}) \propto \prod_{i=1}^{p} \theta_i^{a_i-1}, \quad a_i > 0, \quad i = 1, \ldots p$$

sendo $a_1, \ldots, a_p$ os parâmetros da distribuição a priori Dirichlet. A distribuição a posteriori é dada por

$$p(\boldsymbol{\theta}|\boldsymbol{x}) \propto \prod_{i=1}^{p} \theta_i^{x_i} \prod_{i=1}^{p} \theta_i^{a_i-1} = \prod_{i=1}^{p} \theta_i^{x_i+a_i-1}.$$

ou seja, a posteriori é também Dirichlet com parâmetros $a_1 + x_1, \ldots, a_p + x_p$. Assim temos uma priori conjugada ao modelo multinomial. Note que estamos generalizando a análise conjugada para amostras Binomiais com priori Beta.

4.3.5 Amostrando de uma Distribuição Normal

Um outro resultado importante ocorre quando se tem uma única observação da distribuição normal com média desconhecida. Se a média tiver priori normal então os parâmetros da posteriori são obtidos de uma forma bastante intuitiva.

Teorema 4.1 *Se* $X|\theta \sim N(\theta, \sigma^2)$ *com* σ^2 *conhecido e* $\theta \sim N(\mu_0, \tau_0^2)$ *então* $\theta|x \sim N(\mu_1, \tau_1^2)$ *sendo*

$$\mu_1 = \frac{\tau_0^{-2}\mu_0 + \sigma^{-2}x}{\tau_0^{-2} + \sigma^{-2}} \quad e \quad \tau_1^{-2} = \tau_0^{-2} + \sigma^{-2}.$$

Note que, definindo *precisão* como o inverso da variância, segue do teorema que a precisão a posteriori é a soma das precisões a priori e da verossimilhança e não depende de x. Interpretando precisão como uma medida de informação e definindo $w = \tau_0^{-2}/(\tau_0^{-2} + \sigma^{-2}) \in (0, 1)$ então w mede a informação relativa contida na priori com respeito à informação total. Podemos escrever então que

$$\mu_1 = w\mu_0 + (1 - w)x$$

ou seja, μ_1 é uma *combinação linear convexa* de μ_0 e x e portanto $\min\{\mu_0, x\} \leq \mu_1 \leq \max\{\mu_0, x\}$.

Exemplo 4.6 : (Box & Tiao, 1992) Os físicos A e B desejam determinar uma constante física θ. O físico A tem mais experiência nesta área e especifica sua priori como $\theta \sim N(900, 20^2)$. O físico B tem pouca experiência e especifica uma priori muito mais incerta em relação à posição de θ, $\theta \sim N(800, 80^2)$. Assim, não

é difícil verificar que

$$\text{para o físico } A\text{:} \quad P(860 < \theta < 940) \approx 0,95$$

$$\text{para o físico } B\text{:} \quad P(640 < \theta < 960) \approx 0,95.$$

Faz-se então uma medição X de θ em laboratório com um aparelho calibrado com distribuição amostral $X|\theta \sim N(\theta, 40^2)$ e observou-se $X = 850$. Aplicando o teorema 1.1 segue que

$$(\theta|X = 850) \sim N(890, 17, 9^2) \quad \text{para o físico } A$$

$$(\theta|X = 850) \sim N(840, 35, 7^2) \quad \text{para o físico } B.$$

Note também que os aumentos nas precisões a posteriori em relação às precisões a priori foram,

- para o físico A: precisão(θ) passou de $\tau_0^{-2} = 0,0025$ para $\tau_1^{-2} = 0,00312$ (aumento de 25%).
- para o físico B: precisão(θ) passou de $\tau_0^{-2} = 0,000156$ para $\tau_1^{-2} = 0,000781$ (aumento de 400%).

A situação está representada graficamente na Figura 4.1 a seguir. Note como a distribuição a posteriori representa um compromisso entre a distribuição a priori e a verossimilhança. Além disso, como as incertezas iniciais são bem diferentes o mesmo experimento fornece muito pouca informação adicional para o físico A enquanto que a incerteza do físico B foi bastante reduzida.

Para uma única observação vimos pelo Teorema 4.1 que a família de distribuições normais é conjugada ao modelo normal. Para uma amostra de tamanho n, a função de verssimilhança pode ser escrita como

$$\begin{aligned} l(\theta; x) &= (2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \theta)^2\right\} \\ &\propto \exp\left\{-\frac{n}{2\sigma^2}(\overline{x} - \theta)^2\right\} \end{aligned}$$

onde os termos que não dependem de θ foram incorporados à constante de proporcionalidade. Portanto, a verossimilhança tem a mesma forma daquela baseada em uma única observação bastando substituir x por $\overline{x}$ e σ^2 por σ^2/n. Logo vale o Teorema 4.1 com as devidas substituições, i.e. a distribuição a posteriori de θ dado $\boldsymbol{x}$ é $N(\mu_1, \tau_1^2)$ onde

$$\mu_1 = \frac{\tau_0^{-2}\mu_0 + n\sigma^{-2}\overline{x}}{\tau_0^{-2} + n\sigma^{-2}} \quad \text{e} \quad \tau_1^{-2} = \tau_0^{-2} + n\sigma^{-2}.$$

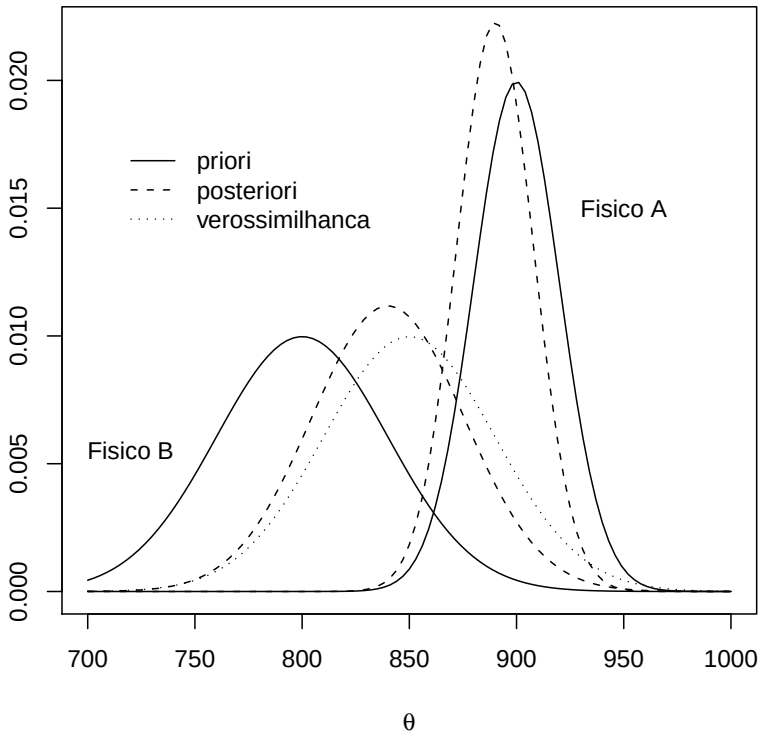

Figura 4.1: Densidades a priori e a posteriori e função de verossimilhança para o Exemplo 4.6.

4.4 Problemas

1. A proporção θ de itens defeituosos em um grande lote é desconhecida e deve ser estimada. Assume-se que a distribuição a priori de θ é uniforme no intervalo (0,1). Itens são selecionados ao acaso e inspecionados até que a variância a posteriori de θ seja menor ou igual a 0,01. Determine o número total de itens que devem ser selecionados.

2. No problema anterior suponha que a priori é Beta com parâmetros $\alpha = 2$ e $\beta = 200$. Se 100 itens foram selecionados ao acaso e 3 eram defeituosos obtenha a distribuição a posteriori de θ.

3. Mostre que a família de distribuições Beta é conjugada em relação às distribuições amostrais binomial, geométrica e binomial negativa.

4. Suponha que o tempo, em minutos, para atendimento a clientes segue uma distribuição exponencial com parâmetro θ desconhecido. Com base na experiência anterior assume-se uma distribuição a priori Gama com média 0,2

e desvio-padrão 1 para θ.

(a) Se o tempo médio para atender uma amostra aleatória de 20 clientes foi de 3,8 minutos, qual a distribuição a posteriori de θ.

(b) Qual o menor número de clientes que precisam ser observados para que o coeficiente de variação a posteriori se reduza para 0,1?

5. Seja $X_1, \ldots, X_n$ uma amostra aleatória da distribuição de Poisson com parâmetro θ.

(a) Determine os parâmetros da priori conjugada de θ sabendo que $E(\theta) = 4$ e o coeficiente de variação a priori é 0,5.

(b) Quantas observações devem ser tomadas até que a variância a posteriori se reduza para 0,01 ou menos?

(c) Mostre que a média a posteriori é da forma $\gamma_n \overline{x} + (1 - \gamma_n)\mu_0$, onde $\mu_0 = E(\theta)$ e $\gamma_n \to 1$ quando $n \to \infty$. Interprete este resultado.

6. O número médio de defeitos por 100 metros de uma fita magnética é desconhecido e denotado por θ. Atribui-se uma distribuição a priori Gama(2,10) para θ. Se um rolo de 1200 metros desta fita foi inspecionado e encontrou-se 4 defeitos qual a distribuição a posteriori de θ?

7. Seja $X_1, \ldots, X_n$ uma amostra aleatória da distribuição Bernoulli com parâmetro θ e usamos a priori conjugada $Beta(a, b)$. Mostre que a média a posteriori é da forma $\gamma_n \overline{x} + (1 - \gamma_n)\mu_0$, onde $\mu_0 = E(\theta)$ e $\gamma_n \to 1$ quando $n \to \infty$. Interprete este resultado.

8. Para uma amostra aleatória $X_1, \ldots, X_n$ tomada da distribuição $U(0, \theta)$, mostre que a família de distribuições de Pareto com parâmetros a e b, cuja função de densidade é $p(\theta) = ab^a/\theta^{a+1}$, é conjugada à uniforme.

9. Para uma amostra aleatória de 100 observações da distribuição normal com média θ e desvio-padrão 2 foi especificada uma priori normal para θ. Mostre que o desvio-padrão a posteriori será sempre menor do que 1/5 (Interprete este resultado).

10. Para uma amostra aleatória da distribuição normal com média θ e desvio-padrão 2 foi especificada uma priori normal para θ com variância igual a 1. Qual deve ser o menor número de observações para que o desvio-padrão a posteriori seja 0,1?

11. Para uma variável aleatória $\theta > 0$ a família de distribuições Gama-invertida tem função de densidade de probabilidade dada por

$$p(\theta) = \frac{\beta^\alpha}{\Gamma(\alpha)}\theta^{-(\alpha+1)}e^{-\beta/\theta}, \quad \alpha, \beta > 0.$$

Mostre que esta família é conjugada ao modelo normal com média μ conhecida e variância θ desconhecida.

4.5 Estimadores de Bayes

A distribuição a posteriori de um parâmetro θ contém toda a informação probabilística a respeito deste parâmetro e um gráfico da sua função de densidade a posteriori é a melhor descrição do processo de inferência. No entanto, algumas vezes é necessário resumir a informação contida na posteriori através de alguns poucos valores numéricos. O caso mais simples é a estimação pontual de θ onde se resume a distribuição a posteriori através de um único número, $\hat{\theta}$. Como veremos a seguir, será mais fácil entender a escolha de $\hat{\theta}$ no contexto de teoria da decisão.

4.5.1 Introdução à Teoria da Decisão

Um problema de decisão fica completamente especificado pela descrição dos seguintes espaços:

(i) Espaço do parâmetro ou estados da natureza, Θ.

(ii) Espaço dos resultados possíveis de um experimento, Ω.

(iii) Espaço de possíveis ações, A.

Uma regra de decisão δ é uma função definida em Ω que assume valores em A, i.e. $\delta : \Omega \to A$. A cada decisão δ e a cada possível valor do parâmetro θ podemos associar uma perda $L(\delta, \theta)$ assumindo valores positivos. Definimos assim uma função de perda $L(\delta, \theta) : \Theta \times A \to \mathbb{R}^+$. Algumas funções de perda comumente utilizadas em problemas de estimação serão vistas na próxima seção.

Intuitivamente, gostariamos de obter uma regra de decisão que minimiza a função de perda, no entanto isto não é possível já que esta depende do valor desconhecido de θ. Uma forma de contornar este problema é especificar uma regra de decisão que minimiza a perda média, o que nos leva a definição a seguir.

Definição 4.1 *O risco de uma regra de decisão, denotado por $R(\delta)$, é a perda esperada a posteriori, i.e. $R(\delta) = E_{\theta|\boldsymbol{x}}[L(\delta, \theta)]$.*

Definição 4.2 *Uma regra de decisão δ^* é ótima se tem risco mínimo, i.e. $R(\delta^*) < R(\delta)$, $\forall \delta$. Esta regra será denominada regra de Bayes e seu risco, risco de Bayes.*

Exemplo 4.7 : Um laboratório farmaceutico deve decidir pelo lançamento ou não de uma nova droga no mercado. É claro que o laboratório só lançará a droga se achar que ela é eficiente mas isto é exatamente o que é desconhecido. Podemos associar um parâmetro θ aos estados da natureza: droga é eficiente ($\theta = 1$), droga não é eficiente ($\theta = 0$) e as possíveis ações como lança a droga ($\delta = 1$), não lança a droga ($\delta = 0$). Suponha que foi possível construir a seguinte tabela de perdas levando em conta a eficiência da droga,

	eficiente	não eficiente
lança	-500	600
não lança	1500	100

Vale notar que estas perdas traduzem uma avaliação subjetiva em relação à gravidade dos erros cometidos. Suponha agora que a incerteza sobre os estados da natureza é descrita por $P(\theta = 1) = \pi$, $0 < \pi < 1$ avaliada na distribuição atualizada de θ (seja a priori ou a posteriori). Note que, para δ fixo, $L(\delta, \theta)$ é uma variável aleatória discreta assumindo apenas dois valores com probabilidades π e $1 - \pi$. Assim, usando a definição de risco obtemos que

$$\begin{aligned} R(\delta = 0) &= E(L(0, \theta)) = \pi 1500 + (1 - \pi)100 = 1400\pi + 100 \\ R(\delta = 1) &= E(L(1, \theta)) = \pi(-500) + (1 - \pi)600 = -1100\pi + 600 \end{aligned}$$

Uma questão que se coloca aqui é, para que valores de π a regra de Bayes será de lançar a droga. Não é difícil verificar que as duas ações levarão ao mesmo risco, i.e. $R(\delta = 0) = R(\delta = 1)$ se somente se $\pi = 0,20$. Além disso, para $\pi < 0,20$ temos que $R(\delta = 0) < R(\delta = 1)$ e a regra de Bayes consiste em não lançar a droga enquanto que $\pi > 0,20$ implica em $R(\delta = 1) < R(\delta = 0)$ e a regra de Bayes deve ser de lançar a droga.

4.5.2 Estimadores de Bayes

Seja agora uma amostra aleatória $X_1, \ldots, X_n$ tomada de uma distribuição com função de (densidade) de probabilidade $p(x|\theta)$ aonde o valor do parâmetro θ é desconhecido. Em um problema de inferência o valor de θ deve ser estimado a partir dos valores observados na amostra.

Se $\theta \in \Theta$ então é razoável que os possíveis valores de um estimador $\delta(\boldsymbol{X})$ também devam pertencer ao espaço Θ. Além disso, um bom estimador é aquele

para o qual, com alta probabilidade, o erro $\delta(\boldsymbol{X}) - \theta$ estará próximo de zero. Para cada possível valor de θ e cada possível estimativa $a \in \Theta$ vamos associar uma perda $L(a, \theta)$ de modo que quanto maior a distância entre a e θ maior o valor da perda. Neste caso, a perda esperada a posteriori é dada por

$$E[L(a,\theta)|\boldsymbol{x}] = \int_{\Theta} L(a,\theta)p(\theta|\boldsymbol{x})d\theta$$

e a regra de Bayes consiste em escolher a estimativa que minimiza esta perda esperada. Assim, a forma do estimador de Bayes vai depender tanto da função de perda quanto da distribuição a priori.

Aqui vamos discutir apenas funções de perda simétricas, já que estas são mais comumente utilizadas. Dentre estas a mais utilizada em problemas de estimação é certamente a função de perda quadrática, definida como $L(a, \theta) = (a - \theta)^2$. Neste caso, pode-se mostrar que o estimador de Bayes para o parâmetro θ será a média de sua distribuição atualizada. Note também que neste caso o risco de Bayes é simplesmente $E(E(\theta|\boldsymbol{x}) - \theta)^2 = Var(\theta|\boldsymbol{x})$.

Exemplo 4.8 : Suponha que queremos estimar a proporção θ de itens defeituosos em um grande lote. Para isto será tomada uma amostra aleatória $X_1, \ldots, X_n$ de uma distribuição de Bernoulli com parâmetro θ. Usando uma priori conjugada Beta(α, β) sabemos que após observar a amostra a distribuição a posteriori é Beta$(\alpha + t, \beta + n - t)$ onde $t = \sum_{i=1}^{n} x_i$. A média desta distribuição Beta é dada por $(\alpha + t)/(\alpha + \beta + n)$ e portanto o estimador de Bayes de θ usando perda quadrática é

$$\delta(\boldsymbol{X}) = \frac{\alpha + \sum_{i=1}^{n} X_i}{\alpha + \beta + n}.$$

Note também que fazendo $\alpha \to 0$ e $\beta \to 0$ segue que o estimador de Bayes coincide com o estimador de máxima verossimilhança $\hat{\theta} = \sum_{i=1}^{n} X_i/n$. Esta priori é chamada de *priori não informativa*.

Exemplo 4.9 : No Exemplo 4.8 suponha que foi especificada uma priori Beta(1,1) (ou equivalentemente U(0,1)) para θ e 10 itens foram inspecionados dos quais 8 eram defeituosos. A estimativa de Bayes de θ é $(1+8)/(2+10) = 0,75$ enquanto $\hat{\theta} = 0,80$.

A perda quadrática é as vezes criticada por penalizar demais o erro de estimação. A função de perda absoluta, definida como $L(a, \theta) = |a - \theta|$, introduz punições que crescem linearmente com o erro de estimação e pode-se mostrar que o estimador de Bayes associado é a mediana da distribuição atualizada de θ.

Para reduzir ainda mais o efeito de erros de estimação grandes podemos considerar funções que associam uma perda fixa a um erro cometido, não importando

sua magnitude. Uma tal função de perda, denominada perda 0-1, é definida como

$$L(a,\theta) = \begin{cases} 1 & \text{se} \quad |a-\theta| > \epsilon \\ 0 & \text{se} \quad |a-\theta| < \epsilon \end{cases}$$

para todo $\epsilon > 0$. Neste caso pode-se mostrar que o estimador de Bayes é a moda da distribuição atualizada de θ. A moda da posteriori de θ também é chamado de estimador de máxima verossimilhança generalizado (EMVG) e é o mais fácil de ser obtido dentre os estimadores vistos até agora. No caso contínuo devemos obter a solução da equação

$$\frac{\partial p(\theta|\boldsymbol{x})}{\partial \theta} = 0.$$

Um caso particular interessante é quando $p(\theta)$ é proporcional a uma constante (como no Exemplo 4.9). Pelo teorema de Bayes segue que $p(\theta|\boldsymbol{x}) \propto p(\boldsymbol{x}|\theta)$ e o estimador de Bayes coincide com o estimador de máxima verossimilhança.

Exemplo 4.10 : Se $X_1, \ldots, X_n$ é uma amostra aleatória da $N(\theta, \sigma^2)$ com σ^2 conhecido e usarmos a priori conjugada, i.e. $\theta \sim N(\mu_0, \tau_0^2)$ então a posteriori também será normal e neste caso média, mediana e moda coincidem. Portanto, o estimador de Bayes de θ é dado por

$$\delta(\boldsymbol{X}) = \frac{\tau_0^{-2}\mu_0 + n\sigma^{-2}\overline{\boldsymbol{X}}}{\tau_0^{-2} + n\sigma^{-2}}.$$

Note que se $\tau_0^{-2} \to 0$ segue que $\delta(\boldsymbol{X}) \to \overline{\boldsymbol{X}}$. Ou seja, na prática se atribuirmos uma variância a priori muito grande para θ a estimativa de Bayes deverá ser similar à média amostral.

Exemplo 4.11 : No Exemplo 4.8 suponha que foram observados 100 itens dos quais 10 eram defeituosos. Usando perda quadrática a estimativa de Bayes de θ é

$$\delta(\boldsymbol{x}) = \frac{\alpha + 10}{\alpha + \beta + 100}.$$

Assim, se a priori for Beta(1,1), ou equivalentemente $U(0,1)$, então $\delta(\boldsymbol{x}) = 0,108$. Por outro lado se especificarmos uma priori Beta(1,2), que é bem diferente da anterior, então $\delta(\boldsymbol{x}) = 0,107$. Ou seja, as estimativas de Bayes são bastante próximas, e isto é uma consequência do tamanho amostral ser grande. Note também que ambas as estimativas são próximas da proporção amostral de defeituosos 0,1, que é a estimativa de máxima verossimilhança.

4.6 Problemas

1. Sabendo que um paciente pode ter a doença A ou a doença B um médico deve decidir pelo diagnóstico de uma das duas doenças. Associando um parâmetro θ aos estados da natureza: paciente tem a doença A ($\theta = 1$), paciente tem a doença B ($\theta = 0$), e as possíveis ações do médico como diagnosticar a doença A ($\delta = 1$) ou diagnosticar a doença B ($\delta = 0$) foi possível construir a seguinte tabela de perdas,

	diagnóstico	
θ	doença A	doença B
1	0	5
0	10	0

   Pela experiência do médico com estas doenças ele atribui a probabilidade $P(\theta = 1) = \rho$, $0 < \rho < 1$. Calcule os riscos associados como função de ρ, esboce estes riscos graficamente e deduza a decisão de menor risco.

2. Em que condições o estimador de Bayes usando perda 0-1 coincide com o estimador de máxima verossimilhança?

3. A proporção θ de itens defeituosos em um grande lote é desconhecida e deve ser estimada. Assume-se que a distribuição a priori de θ é Beta(5,10). Suponha que 20 itens foram selecionados ao acaso e inspecionados e encontrou-se exatamente um defeituoso.

   (a) Obtenha a estimativa de Bayes de θ usando perda quadrática.

   (b) Repita a estimação usando perda 0-1.

   (c) Comente os resultados e compare com a estimativa de máxima verossimilhança.

4. O número de defeitos em rolos de 100 metros de uma fita magnética tem distribuição de Poisson com média θ desconhecida. A distribuição a priori de θ é Gama(3,1). Se cinco rolos são selecionados ao acaso e observa-se 2, 2, 6, 0 e 3 defeitos obtenha a estimativa Bayesiana de θ usando perda quadrática.

5. Suponha que as alturas (em cm) de indivíduos de uma população seguem uma distribuição normal cuja média θ é desconhecida e o desvio-padrão é 5 cm. A distribuição a priori de θ é normal com média 173 cm e desvio-padrão 2,5 cm. Uma amostra aleatória de 10 indivíduos foi selecionada e sua altura média foi de 177 cm. Calcule a estimativa de Bayes de θ.

6. Suponha que o tempo em minutos para atender um cliente tem distribuição exponencial com parâmetro θ desconhecido. A distribuição a priori de θ é Gama com média 0,2 e desvio-padrão 1. Se o tempo médio para atender uma amostra aleatória de 20 clientes foi 3,8 minutos calcule a estimativa de Bayes de θ usando função de perda quadrática.

Capítulo 5

Estimação por Intervalos

A principal restrição da estimação pontual é que quando estimamos um parâmetro através de um único valor numérico toda a informação presente nos dados é resumida através deste número. É importante encontrar também um intervalo de valores plausíveis para o parâmetro.

A idéia é construir um intervalo em torno da estimativa pontual de modo que ele tenha uma probabilidade conhecida de conter o verdadeiro valor do parâmetro. Tipicamente as distribuições amostrais de estimadores dos parâmetros desconhecidos serão utilizadas. Antes de descrever o procedimento geral veremos um exemplo simples de construção do intervalo de confiança.

Exemplo 5.1 : Seja $X_1, \ldots, X_n$ uma amostra aleatória da distribuição $N(\theta, \sigma^2)$, com σ^2 conhecido. Para fazer inferências sobre θ nos baseamos na média amostral $\overline{\boldsymbol{X}}$ e sabemos que

$$U = \frac{\sqrt{n}\ (\overline{\boldsymbol{X}} - \theta)}{\sigma} \sim N(0,1).$$

Note que a estatística U é uma função da amostra e também de θ, o parâmetro de interesse, mas sua distribuição de probabilidades não depende de θ. Usando uma tabela da distribuição normal padronizada podemos obter o valor do percentil $z_{\alpha/2}$ tal que

$$P(-z_{\alpha/2} \leq U \leq z_{\alpha/2}) = 1 - \alpha$$

e assim, após isolar θ, obtemos que

$$P\left(\overline{\boldsymbol{X}} - z_{\alpha/2}\ \frac{\sigma}{\sqrt{n}} \leq \theta \leq \overline{\boldsymbol{X}} + z_{\alpha/2}\ \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha.$$

Esta última igualdade pode dar margem a interpretações errôneas, o que aliás acontece com bastante frequência. Na inferência clássica, o parâmetro θ é desconhecido mas fixo e portanto não é passível de descrição probabilística, ou seja não se trata de um intervalo de probabilidade para θ. Na verdade os limites do

intervalo é que são variáveis aleatórias. Após a amostra ser observada teremos um valor numérico para a média amostral, i.e. $\overline{X} = \overline{x}$ e dizemos que

$$\overline{\boldsymbol{x}} - z_{\alpha/2}\ \frac{\sigma}{\sqrt{n}} \leq \theta \leq \overline{\boldsymbol{x}} + z_{\alpha/2}\ \frac{\sigma}{\sqrt{n}}$$

com confiança $100(1-\alpha)\%$. Não se pode fazer afirmações do tipo "o verdadeiro valor de θ tem 95% de chances de estar no intervalo $\overline{\boldsymbol{x}} \pm 1,96\ \sigma/\sqrt{n}$".

Vale notar também que, para um dado valor de $1-\alpha$, é possível construir muitos intervalos de confiança diferentes para θ. Na verdade, quaisquer constantes c_1 e c_2 tais que $P(c_1 \leq U \leq c_2) = 1-\alpha$ podem ser usadas para construir um intervalo com limites

$$\overline{\boldsymbol{x}} - c_2\ \frac{\sigma}{\sqrt{n}} \quad \text{e} \quad \overline{\boldsymbol{x}} - c_1\ \frac{\sigma}{\sqrt{n}}.$$

No entanto, pode-se mostrar que dentre todos os intervalos de confiança com esta característica, aquele definido acima que é simétrico em torno do média amostral $\overline{\boldsymbol{x}}$ é o de menor comprimento.

Exemplo 5.2 : No Exemplo 5.1 suponha que foram observados os dados abaixo,

```
-3.83 -1.88 -1.55 -12.64 -0.4 -18.87 4.98 -9.52 -14.06 13.56
```

e queremos construir um intervalo de 95% para a média θ com $\sigma = 10$. A média amostral é $\overline{x} = -4.421$. Na Tabela 5.1 abaixo encontram-se os valores de c_1 e c_2 obtidos para diferentes probabilidades nas caudas da distribuição normal padrão. Na última coluna estão os comprimentos $\sigma(c_2 - c_1)/\sqrt{n}$ dos intervalos.

	P(Z<c_1)	P(Z>c_2)	c_1	c_2	comp
1	0.010	0.040	-2.326	1.751	12.890
2	0.020	0.030	-2.054	1.881	12.440
3	0.025	0.025	-1.960	1.960	12.400
4	0.045	0.005	-1.695	2.576	13.510

Tabela 5.1: Valores de c_1 e c_2 para diferentes probabilidades nas caudas e os comprimentos dos intervalos.

5.1 Procedimento Geral

O procedimento geral para construção de intervalos de confiança para um parâmetro θ consiste nos seguintes passos,

1. Obter uma estatística que depende de θ, $U = G(\boldsymbol{X}, \theta)$, mas cuja distribuição não depende de θ.

2. Usando a distribuição de U, encontrar as constantes a e b tais que $P(a \leq U \leq b) \geq 1 - \alpha$.

3. Definir $\{\theta : a \leq G(\boldsymbol{x}, \theta) \leq b\}$ como o intervalo (ou região) de confiança 100(1-α)% para θ.

A exigência de que a probabilidade no item 2 acima possa ser maior do que o nível de confiança é essencialmente técnica pois queremos que o intervalo seja o menor possível, o que em geral implica em usar uma igualdade. A desigualdade será útil principalmente no caso de distribuições discretas onde nem sempre é possível satisfazer a igualdade.

Note que a variável aleatória U, comumente denominada quantidade pivotal ou *pivot*, é fundamental para o funcionamento do método. Idealmente ela deve depender da amostra através de estatísticas suficientes minimais e ter distribuição conhecida.

É importante notar também que este intervalo não pode ser interpretado como um intervalo de probabilidade para θ já que a aleatoriedade presente é devida à amostra $X_1, \ldots, X_n$. Ou seja, o procedimento leva à construção de um intervalo probabilístico para a variável aleatória U e não para θ.

Tecnicamente, dizemos que $100(1-\alpha)\%$ de todos os intervalos de confiança que construirmos conterão o verdadeiro valor do parâmetro (dado que todas as suposições envolvidas estejam corretas). Por exemplo se $1 - \alpha = 0,95$ então, em média, somente 5 a cada 100 intervalos não conterão θ. A probabilidade $1 - \alpha$ é denominada *nível de confiança* e sua escolha depende da precisão com que queremos estimar o parâmetro, sendo em geral $1 - \alpha \geq 0,90$ os valores mais utilizados na prática. Esta idéia está representada na Figura 5.1.

Exemplo 5.3 : Seja $X_1, \ldots, X_n \sim U[0, \theta]$, para $\theta > 0$ desconhecido. A função de distribuição acumulada de $\max\{X_i\}$ é dada por

$$F(x) = P(\max\{X_i\} < x) = P(X_1 < x, \ldots, X_n < x) = \prod_{i=1}^{n} P(X_i < x),$$

e como $P(X_i < x) = x/\theta$ segue que

$$P(\max\{X_i\} < x) = (x/\theta)^n, \ 0 \leq x \leq \theta.$$

Consequentemente a distribuição de $\max\{X_i\}/\theta$ também pode ser facilmente obtida como

$$P(\max\{X_i\}/\theta < x) = P(\max\{X_i\} < x\theta) = x^n, \ 0 \leq x \leq \theta.$$

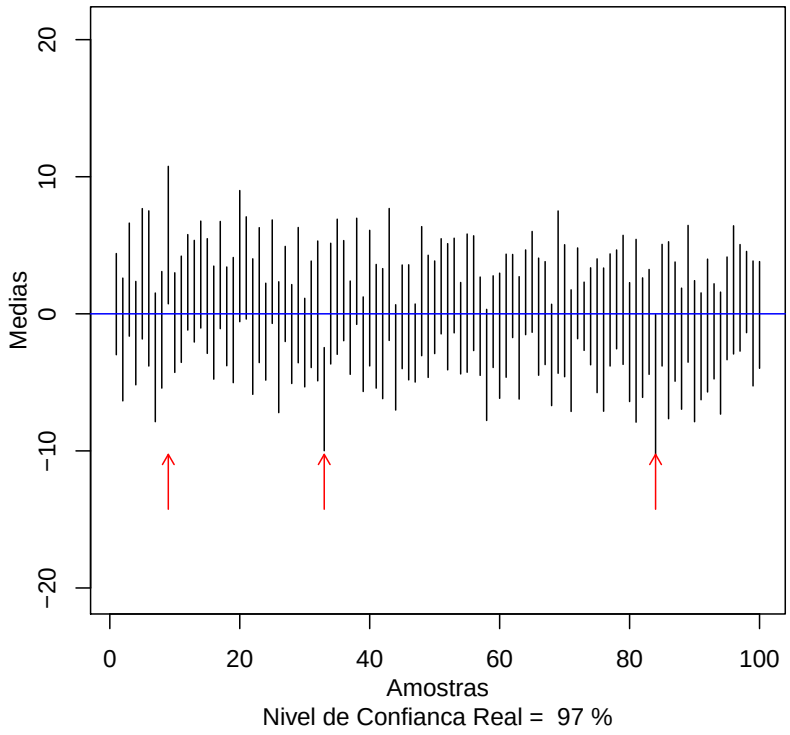

Figura 5.1: Intervalos de 95% de confiança para média de 100 amostras de tamanho $n = 20$ simuladas de uma distribuição $N(0, 100)$. As setas indicam os intervalos que não contém o valor verdadeiro (zero).

Assim, $\max\{X_i\}/\theta$ é uma estatística que depende da amostra através de $\max\{X_i\}$ e do parâmetro desconhecido mas cuja distribuição não depende de θ. Para um coeficiente de confiança $1 - \alpha = 0,90$ podemos obter os limites c_1 e c_2 tais que

$$P(c_1 < \max\{X_i\}/\theta < c_2) = 0,90$$

e se as áreas à esquerda de c_1 e à direita de c_2 forem iguais então

$$\begin{aligned} P(\max\{X_i\}/\theta < c_2) &= 0,95 = c_2^n \Rightarrow c_2 = 0,95^{1/n} \\ P(\max\{X_i\}/\theta < c_1) &= 0,05 = c_1^n \Rightarrow c_1 = 0,05^{1/n} \end{aligned}$$

Agora, isolando θ obtemos o I.C. de 90%

$$\frac{\max\{X_i\}}{0,95^{1/n}} < \theta < \frac{\max\{X_i\}}{0,05^{1/n}}.$$

Os dados abaixo foram simulados de uma distribuição uniforme no intervalo (0;10)

```
0.87 7.94 3.16 9.85 3.39 1.53 5.15 4.38 8.5 7.02
```

Usando a expressão acima então segue que $9.9 < \theta < 13.29$ com confiança 0,90.

5.2 Estimação no Modelo Normal

Nesta seção serão discutidos os casos em que os dados provém de uma distribuição normal. Inicialmente veremos o caso em que temos uma única amostra de uma distribuição normal e queremos estimar sua média e sua variância. Na Seção 5.2.2 estudaremos o caso de duas amostras tomadas de distribuições normais independentes.

5.2.1 O caso de uma amostra

No exemplo 5.1, se σ^2 for desconhecido não podemos usar a mesma quantidade pivotal já que ela depende de σ. Ou seja, precisamos obter uma outra quantidade pivotal que depende apenas de $\boldsymbol{X}$ e de θ e com uma distribuição que seja conhecida e não dependa de nenhum parâmetro desconhecido. No modelo normal isto será possível usando os resultados a seguir.

Teorema 5.1 *Seja $X_1, \ldots, X_n$ uma amostra aleatória da distribuição $N(\theta, \sigma^2)$ e sejam $\overline{X}$ e S^2 a média e a variância amostrais. Então, condicionado em θ e σ^2, $\overline{X}$ e S^2 são independentes com distribuições amostrais*

$$\frac{\sqrt{n}(\overline{\boldsymbol{X}} - \theta)}{\sigma} \sim N(0,1) \quad e \quad \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}.$$

Lema 5.1 *Se $U \sim N(0,1)$ e $W \sim \chi^2_\nu$ e se U e W são independentes então*

$$\frac{U}{\sqrt{\frac{W}{\nu}}} \sim t_\nu(0,1).$$

Prova. A prova é deixada como exercício.

A notação $t_\nu(0,1)$ denota a distribuição t de Student com ν graus de liberdade centrada em zero e com variância 1 (ver Apêndice A).

Corolário 5.1 *Seja $X_1, \ldots, X_n$ uma amostra aleatória da distribuição $N(\theta, \sigma^2)$ e sejam $\overline{X}$ e S^2 a média e a variância amostrais. Então, condicionado em θ e*

σ^2, $\overline{X}$ tem distribuição amostral

$$\frac{\sqrt{n}(\overline{X}-\theta)}{S} \sim t_{n-1}(0,1)$$

Prova. Aplicação direta do Lema 5.1 acima com $U = \sqrt{n}(\overline{X}-\theta)/\sigma$, $W = (n-1)S^2/\sigma^2$ e $\nu = n-1$.

Estes resultados nos permitem definir quantidades pivotais para construção de intervalos de confiança para θ e σ^2. No caso da média θ, o valor desconhecido de σ é substituido pelo seu estimador S levando a uma quantidade pivotal com distribuição t com $n-1$ graus de liberdade. Assim, podemos obter o percentil $t_{\alpha/2,n-1}$ tal que

$$P\left(-t_{\alpha/2,n-1} \leq \frac{\sqrt{n}\ (\overline{X}-\theta)}{S} \leq t_{\alpha/2,n-1}\right) = 1-\alpha$$

e, após isolar θ, obtemos que

$$P\left(\overline{\boldsymbol{X}} - t_{\alpha/2,n-1}\frac{S}{\sqrt{n}} \leq \theta \leq \overline{\boldsymbol{X}} + t_{\alpha/2,n-1}\frac{S}{\sqrt{n}}\right) = 1-\alpha.$$

Note que, mesmo se S pudesse estimar σ sem erro, esta substituição implica em um aumento da amplitude do intervalo de confiança pois $t_{\alpha,n} > z_\alpha$ para n pequeno.

Finalmente, após observar a amostra substituimos as estimativas e dizemos que

$$\overline{\boldsymbol{x}} - t_{\alpha/2,n-1}\ \frac{s}{\sqrt{n}} \leq \theta \leq \overline{\boldsymbol{x}} + t_{\alpha/2,n-1}\ \frac{s}{\sqrt{n}}$$

com confiança $100(1-\alpha)\%$.

Estimando a Variância

Para obter estimativas da variância populacional σ^2 usamos a seguinte quantidade pivotal

$$Q = \frac{(n-1)S^2}{\sigma^2}$$

que tem distribuição qui-quadrado com $n-1$ graus de liberdade. Devemos então obter os percentis $\underline{\chi}^2_{\alpha/2,n-1}$ e $\overline{\chi}^2_{\alpha/2,n-1}$ desta distribuição tais que

$$P\left(\underline{\chi}^2_{\alpha/2,n-1} \leq \frac{(n-1)S^2}{\sigma^2} \leq \overline{\chi}^2_{\alpha/2,n-1}\right) = 1-\alpha.$$

Após observar a amostra teremos o valor numérico s^2 de S^2 e o intervalo de confiança de $100(1-\alpha)\%$ para σ^2 é dado por

$$\left(\frac{(n-1)s^2}{\overline{\chi}^2_{\alpha/2,n-1}}, \frac{(n-1)s^2}{\underline{\chi}^2_{\alpha/2,n-1}} \right).$$

Note que o intervalo não depende da média amostral $\bar{x}$ mas somente do grau de dispersão dos dados, medido através do desvio padrão amostral s^2. Além disso, este intervalo não é simétrico em torno da estimativa pontual e por isso temos que obter 2 valores da distribuição qui-quadrado cujos valores absolutos são diferentes, um a ser utilizado no limite inferior e outro a ser utilizado no limite superior do intervalo.

Exemplo 5.4 : Considere novamente os dados do Exemplo 5.2 com média e variância desconhecidas e construa um I.C. de 90% para estes parâmetros.

A média amostral é $\overline{x} = -4.421$ e a variância amostral é $s^2 = 93.128$. Da tabela da distribuição t com $n-1 = 9$ graus de liberdade obtemos que $P(T > 1.833) = 0,05$. Portanto,

$$\overline{x} - 1.833\ s/\sqrt{n} \leq \theta \leq \overline{x} + 1.833\ s/\sqrt{n}$$

é um I.C. de 90% para θ. Substituindo os valores de $\overline{x}$ e s obtemos que $-10.015 \leq \theta \leq 1.173$.

Da tabela da distribuição qui-quadrado com $n-1 = 9$ graus de liberdade obtemos que $P(Q > 3.325) = 0,95$ e $P(Q > 16.919) = 0,05$. Portanto,

$$\left(\frac{(n-1)s^2}{16.919}, \frac{(n-1)s^2}{3.325} \right)$$

é um I.C. de 90% para σ^2. Substituindo os valores numéricos obtemos que $5.133 \leq \sigma^2 \leq 26.121$.

5.2.2 O caso de duas amostras

Nesta seção vamos assumir que $X_{11}, \ldots, X_{1n_1}$ e $X_{21}, \ldots, X_{2n_2}$ são amostras aleatórias das distribuições $N(\theta_1, \sigma_1^2)$ e $N(\theta_2, \sigma_2^2)$ respectivamente e que as amostras são independentes.

Podemos comparar as médias populacionais estimando a diferença $\beta = \theta_1 - \theta_2$. A estimação é baseada na diferença entre médias amostrais, i.e. $\overline{X}_1 - \overline{X}_2$ que é o estimador de máxima verossimilhança de β. Se as variâncias

populacionais forem conhecidas então a distribuição amostral é dada por

$$\overline{X}_1 - \overline{X}_2 \sim N(\theta_1 - \theta_2, \sigma^2)$$

onde

$$\sigma^2 = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}.$$

já que as médias amostrais são independentes. Assim, o intervalo de confiança de $100(1-\alpha)\%$ para a diferença entre médias é dado por

$$\left(\overline{x}_1 - \overline{x}_2 - z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \; ; \; \overline{x}_1 - \overline{x}_2 + z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \right).$$

No caso de variâncias populacionais desconhecidas porém iguais, i.e. $\sigma_1^2 = \sigma_2^2 = \sigma^2$ podemos combinar os duas variâncias amostrais para formar uma estimativa combinada da variância. Atribuímos mais peso às amostras maiores e esta variância combinada é dada por

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2},$$

isto é, a média ponderada das variâncias amostrais com pesos dados por $n_1 - 1$ e $n_2 - 1$. Agora podemos calcular o erro padrão das diferenças nas médias como

$$EP(\overline{X}_1 - \overline{X}_2) = S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}.$$

Note que $\min\{s_1^2, s_2^2\} \leq s_p^2 \leq \max\{s_1^2, s_2^2\}$ sempre já que a soma dos coeficientes é igual a 1. Se isto não ocorrer seus cálculos estão errados.

Note também que

$$\frac{(n_1 - 1)S_1^2}{\sigma^2} \sim \chi^2_{n_1 - 1} \quad \text{e} \quad \frac{(n_2 - 1)S_2^2}{\sigma^2} \sim \chi^2_{n_2 - 1}$$

e como S_1^2 e S_2^2 são independentes segue que

$$\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_1^2}{\sigma^2} \sim \chi^2_{n_1 + n_2 - 2}.$$

Agora fica fácil verificar que

$$\frac{(n_1 + n_2 - 2)S_p^2}{\sigma^2} \sim \chi^2_{n_1 + n_2 - 2}$$

Do exposto acima, um intervalo de confiança para a diferença entre médias $\theta_1 -$

θ_2 assumindo desvios padrão iguais pode ser construído usando-se a quantidade pivotal

$$\frac{\hat{\beta} - \beta}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_\nu(0,1)$$

onde $\nu = n_1 + n_2 - 2$ e $\hat{\beta} = \overline{X}_1 - \overline{X}_2$. Assim, o intervalo de confiança de $100(1-\alpha)\%$ para a diferença fica,

$$\left(\overline{x}_1 - \overline{x}_2 - t_{\alpha/2,\nu}\ s_p\ \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\ ;\ \overline{x}_1 - \overline{x}_2 + t_{\alpha/2,\nu}\ s_p\ \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right).$$

Analogamente ao caso de uma amostra, o intervalo de confiança para σ^2 é construído usando-se a quantidade pivotal

$$\frac{\nu S_p^2}{\sigma^2} \sim \chi^2_\nu.$$

Então devemos obter os quantis $\alpha/2$ inferior e superior desta distribuição qui-quadrado e o intervalo de confiança de $100(1-\alpha)\%$ para a variância populacional fica

$$\left(\frac{\nu s_p^2}{\overline{\chi}^2_{\alpha/2,\nu}}\ ;\ \frac{\nu s_p^2}{\underline{\chi}^2_{\alpha/2,\nu}}\right)$$

5.2.3 Variâncias desiguais

Até agora assumimos que as variâncias populacionais desconhecidas eram iguais (ou pelo menos aproximadamente iguais). A violação desta suposição leva a problemas teóricos e práticos uma vez que não é trivial encontrar uma quantidade pivotal para β com distribuição conhecida. Na verdade, se existem grandes diferenças de variabilidade entre as duas populações pode ser mais apropriado analisar conjuntamente as consequências das diferenças entre as médias e as variâncias. Assim, caso o pesquisador tenha interesse no parâmetro β deve levar em conta os problemas de ordem teóricas introduzidos por uma diferença substancial entre σ_1^2 e σ_2^2.

A literatura estatística apresenta vários métodos para resolver este problema mas nenhum deles é completamente satisfatório. Um procedimento possível (e aproximado) consiste em utilizar a estatística

$$T = \frac{\hat{\beta} - \beta}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}.$$

No entanto, a distribuição exata de T depende da razão σ_1^2/σ_2^2, que é desconhecida. Se n_1 e n_2 forem grandes T tem distribuição aproximadamente normal padrão, mas quando eles são ambos pequenos uma solução simples é utilizar uma distribuição t de Student com $k-1$ graus de liberdade onde $k = \min(n_1, n_2)$. Outra solução aproximada (método aproximado de Aspin-Welch) consiste em utilizar a estatística acima com distribuição t de Student e número de graus de liberdade dado por

$$\nu = \frac{(w_1 + w_2)^2}{\dfrac{w_1^2}{n_1 - 1} + \dfrac{w_2^2}{n_2 - 1}}$$

onde

$$w_1 = \frac{s_1^2}{n_1} \quad \text{e} \quad w_2 = \frac{s_2^2}{n_2}.$$

No caso de estar utilizando valores tabelados então ν deve ser arrendondado para o inteiro mais próximo.

Novamente pode-se construir intervalos de confiança para a diferença entre as médias populacionais usando esta estatística.

5.2.4 Comparação de variâncias

Outra situação de interesse é a comparação das duas variâncias populacionais. Neste caso, faz mais sentido utilizar a razão de variâncias ao invés da diferença já que elas medem a escala de uma distribuição e são sempre positivas. Ou seja estamos interessados em estimar a quantidade σ_1^2/σ_2^2 construindo intervalos de confiança em torno da estimativa pontual s_1^2/s_2^2. Para obter a distribuição amostral apropriada usaremos o teorema a seguir.

Teorema 5.2 *Sejam as variáveis aleatórias U e W independentes com distribuições qui-quadrado com ν_1 e ν_2 graus de liberdade respectivamente. Então a variável aleatória dada por*

$$X = \frac{U/\nu_1}{W/\nu_2}$$

tem distribuição F com ν_1 e ν_2 graus de liberdade. Usaremos a notação $X \sim F(\nu_1, \nu_2)$.

Do Teorema 5.1 temos que

$$\frac{(n_i - 1)S_i^2}{\sigma_i^2} \sim \chi^2_{n_i=1}, \ i = 1, 2$$

sendo que estas duas variáveis aleatórias são independentes. Então pelo Teorema

5.2 não é difícil mostrar que

$$\frac{S_1^2}{S_2^2}\frac{\sigma_2^2}{\sigma_1^2} \sim F(n_1 - 1, n_2 - 1).$$

Embora sua função de distribuição não possa ser obtida analiticamente os valores estão tabelados em muitos livros de estatística e também podem ser obtidos na maioria dos pacotes computacionais. Os percentis podem então ser utilizados na construção de intervalos de confiança para a razão de variâncias.

Uma propriedade bastante útil para calcular probabilidade com a distribuição F vem do fato de que se $X \sim F(\nu_2, \nu_1)$ então $X^{-1} \sim F(\nu_1, \nu_2)$ por simples inversão na razão de distribuições qui-quadrado independentes. Assim, denotando os quantis α e $1 - \alpha$ da distribuição $F(\nu_1, \nu_2)$ por $\underline{F}_\alpha(\nu_1, \nu_2)$ e $\overline{F}_\alpha(\nu_1, \nu_2)$ respectivamente segue que

$$\underline{F}_\alpha(\nu_1, \nu_2) = \frac{1}{\overline{F}_\alpha(\nu_2, \nu_1)}.$$

Note que é usual que os livros forneçam tabelas com os percentis superiores da distribuição F para várias combinações de valores de ν_1 e ν_2 devido à propriedade acima. Por exemplo, se temos os valores tabelados dos quantis 0,95 podemos obter também um quantil 0,05. Basta procurar o quantil 0,95 inverterndo os graus de liberdade.

Exemplo 5.5 : Suponha que $X \sim F(4, 6)$ e queremos obter o valor x tal que $P(X < x) = 0,05$. Neste caso devemos obter primeiro o valor y tal que $P(X^{-1} > y) = 0,05$ sendo que $X^{-1} \sim F(6, 4)$. Este valor é dado por $y = 6,16$. Podemos agora calcular $x = 1/y \approx 0,16$.

5.2.5 Amostras pareadas

Nas seções anteriores fizemos a suposição de que as amostras eram independentes, mas esta nem sempre é uma suposição razoável. Em estudos chamados *pareados* ou *emparelhados*, temos duas amostras mas cada observação na primeira amostra é pareada (ou emparelhada) com uma observação da segunda amostra. Tal situação ocorre por exemplo em um estudo de medidas feitas antes e depois no mesmo indivíduo (ou mesma máquina, ou mesmo processo de produção, etc). Como esperado, as duas observações do mesmo indivíduo são mais prováveis de serem similares, e portanto não podem ser consideradas estatisticamente independentes.

Analogamente ao caso anterior, as observações pareadas são representadas pelas variáveis aleatórias,

$$X_{11}, \ldots, X_{1n} : \text{medida 1}$$

$$X_{21}, \ldots, X_{2n} : \text{medida 2}$$

e então escrevemos as diferenças nas medidas de cada par como $D_i = X_{2i} - X_{1i}$, $i = 1, \ldots, n$. Temos agora uma amostra de diferenças e assumindo que

$$D_1, \ldots, D_n \sim N(\mu_D, \sigma_D^2)$$

podemos usar os métodos com os quais já estamos familiares. Ou seja, podemos calcular um intervalo de confiança para a diferença média e testar se a diferença média é igual a um particular valor (usualmente zero) ou não. Nos referimos a tal teste como um *teste t pareado*.

A estatística (pivot) utilizada então é

$$\frac{\overline{D} - \mu_D}{S_D/\sqrt{n}} \sim t_{n-1}$$

e o intervalo de confiança de 100(1-α)% para μ_D é

$$\overline{d} - t_{\alpha/2,n-1} s_D/\sqrt{n} \leq \mu_D \leq \overline{d} + t_{\alpha/2,n-1} s_D/\sqrt{n}.$$

Note que neste caso estamos interessados na diferença média enquanto que para duas amostras independentes, estamos interessados na diferença das médias. Ainda que numericamente estas quantidades possam ser as mesmas, conceitualmente elas são diferentes. Outra observação importante é que estamos assumindo normalidade para as diferenças e não para os dados originais. Lembre-se, mesmo que X_{1i} e X_{2i} sejam normais isto não implica que os D_i sejam normais já que aqui não há independência. Portanto a hipótese de normalidade deve ser feita nas diferenças.

Exemplo 5.6 : A mudança nos níveis de um contaminante numa certa área do início ao final de seis meses de observação foram (em μ/l):

$$-1,5 \quad -0,6 \quad -0,3 \quad 0,2 \quad -2,0 \quad -1,2$$

Aqui não estamos interessados nos níveis de contaminação mas sim em sua variação. A média e o desvio padrão amostrais são $\bar{d} = -0,9$ e $s = 0,81$ respectivamente. Então o erro padrão é $0,81/\sqrt{6} = 0,33$. Podemos agora construir um intervalo de confiança para verificar se a perda na concentração média é nula. Para α = 0,05 e 5 graus de liberdade obtém-se $t_{0,025} = 2.45$ e o I.C. de 95% para μ_D fica

$$-0,9 - 2.45 \times 0,81/\sqrt{6} \leq \mu_D \leq -0,9 + 2.45 \times 0,81/\sqrt{6}$$

ou seja com 95% de confiança $\mu \in$ [-1.75; -0.05]. Neste caso há indícios nos dados

de que a perda na concentração média não é nula, ao contrário é negativa.

5.2.6 Comentário

Os intervalos de confiança obtidos nesta seção dependem fortemente da suposição de independência e normalidade dos dados (ou das diferenças). Na prática dificilmente poderemos garantir que os dados seguem um modelo teórico simples e que estas suposições estão corretas.

5.3 Intervalos de confiança para uma proporção

Seja $X_1, \ldots, X_n$ uma amostra aleatória de uma distribuição de Bernoulli com parâmetro θ. Assim,

$$Y = \frac{1}{n}\sum_{i=1}^{n} X_i$$

é a proporção amostral de sucessos e será o nosso estimador pontual da verdadeira probabilidade de sucesso θ. Vamos considerar agora a construção de um intervalo de confiança para θ.

Pelo Teorema Central do Limite, para n grande e θ não muito próximo de 0 ou 1, a distribuição de Y será aproximadamente normal com média θ e um desvio padrão dado por

$$\sqrt{\frac{\theta(1-\theta)}{n}}.$$

já que $E(X_i) = \theta$ e $V(X_i) = \theta(1-\theta)$. Ou seja, a quantidade pivotal será dada por

$$\frac{Y-\theta}{\sqrt{\frac{\theta(1-\theta)}{n}}} \sim N(0,1).$$

Assim, após observar a amostra o intervalo de confiança de 100(1-α)% para θ fica

$$\left(y - z_{\alpha/2}\sqrt{\frac{\theta(1-\theta)}{n}}, y + z_{\alpha/2}\sqrt{\frac{\theta(1-\theta)}{n}}\right).$$

Note que os limites do intervalo dependem do valor desconhecido de θ e aqui duas abordagens são possíveis. Podemos usar o fato de que o valor máximo de $\theta(1-\theta)$ é atingido para $\theta = 1/2$, logo $\theta(1-\theta) \leq 1/4$, ou equivalentemente $\sqrt{\theta(1-\theta)/n} \leq 1/\sqrt{4n}$. Neste caso, um intervalo de confiança *conservativo* é dado por

$$\left(y - z_{\alpha/2}\sqrt{\frac{1}{4n}}, y + z_{\alpha/2}\sqrt{\frac{1}{4n}}\right).$$

No entanto, se o verdadeiro valor de θ estiver afastado do seu valor máximo e estiver próximo de 0 ou de 1 então este intervalo tem amplitude desnecessariamente grande porque substituimos $\theta(1-\theta)$ pelo seu valor máximo. Um enfoque mais otimista consiste em substituir θ pela sua estimativa de máxima verossimilhança, i.e. a proporção amostral de sucessos y e utilizar o intervalo

$$\left(y - z_{\alpha/2}\sqrt{\frac{y(1-y)}{n}}, y + z_{\alpha/2}\sqrt{\frac{y(1-y)}{n}} \right).$$

Note que, para n e $1-\alpha$ fixos a amplitude do intervalo conservativo será a mesma para todas as possíveis amostras de tamanho n. Por outro lado, usando-se esta última expressão o intervalo terá amplitude $2z_{\alpha/2}\sqrt{y(1-y)/n}$ que varia de amostra para amostra.

Uma função geral pode ser escrita no `R` para se obter o intervalo de confiança.

```
> ic.binom = function(dados, nivel = 0.95) {
+     x = sum(dados)
+     n = length(dados)
+     alpha = 1 - nivel
+     xbar = x/n
+     EP = sqrt(xbar * (1 - xbar)/n)
+     q = qnorm(c(alpha/2, 1 - (alpha/2)))
+     IC = xbar + q * EP
+     return(IC)
+ }
```

5.4 Intervalos de Confiança Assintóticos

Utilizando os conceitos do método da quantidade pivotal e a propriedade de normalidade assintótica dos estimadores de máxima verossimilhança podemos construir intervalos de confiança para θ. Para isto serão usadas as definições de medida de informação de Fisher e função escore vistas no Capítulo 1.

Vimos em estimação pontual que, para grandes amostras, o estimador de máxima verossimilhança $\hat{\theta}_n$ para um parâmetro θ tem distribuição aproximadamente normal com média θ sob condições de regularidade gerais. Assim, mesmo que $\hat{\theta}_n$ seja viesado para n fixo ele será assintoticamente não viesado. A variância assintótica é dada pelo inverso da informação esperada de Fisher $1/I(\theta)$. Ou seja, para n grande $\hat{\theta}_n$ tem distribuição aproximadamente $N(\theta, I^{-1}(\theta))$ e podemos construir intervalos de confiança aproximados para θ. Neste caso,

$$(\hat{\theta}_n - \theta)\sqrt{I(\theta)} \sim N(0,1)$$

pode ser tratado como uma quantidade pivotal aproximada e se for possível isolar θ na desigualdade

$$-z_{\alpha/2} < (\hat{\theta}_n - \theta)\sqrt{I(\theta)} < z_{\alpha/2}$$

teremos um intervalo de confiança com coeficiente de confiança aproximado igual a $1 - \alpha$.

Exemplo 5.7 : Seja $X_1, \ldots, X_n$ uma amostra aleatória da distribuição exponencial com parâmetro θ. A função de densidade conjunta é dada por

$$p(\boldsymbol{x}|\theta) = \theta^n e^{-\theta t},\ \theta > 0,\ t = \sum_{i=1}^{n} x_i.$$

Tomando-se o logaritmo obtém-se

$$\log p(\boldsymbol{x}|\theta) = n \log(\theta) - \theta t$$

de modo que as derivadas de primeira e segunda ordem são

$$\frac{\partial \log p(|\theta)}{\partial \theta} = \frac{n}{\theta} - t \quad \text{e} \quad \frac{\partial^2 \log p(|\theta)}{\partial \theta^2} = -\frac{n}{\theta^2}$$

e a informação esperada de Fisher baseada na amostra é $I(\theta) = n/\theta^2$. Sabemos também que o estimador de máxima verossimilhança de θ é $1/\overline{\boldsymbol{X}}$ e portanto, para n grande, $1/\overline{\boldsymbol{X}}$ tem distribuição aproximadamente normal com média θ e variância θ^2/n. Assim, o intervalo de confiança aproximado é obtido fazendo-se

$$P\left(-z_{\alpha/2} < \frac{1/\overline{\boldsymbol{X}} - \theta}{\sqrt{\theta^2/n}} < z_{\alpha/2}\right) \approx 1 - \alpha.$$

Isolando θ obtemos que

$$P\left(\frac{\sqrt{n}/\overline{X}}{\sqrt{n} + z_{\alpha/2}} < \theta < \frac{\sqrt{n}/\overline{X}}{\sqrt{n} - z_{\alpha/2}}\right) \approx 1 - \alpha.$$

Exemplo 5.8 : Os dados abaixo (simulados) formam uma amostra aleatória de $X \sim Exp(\theta)$, com $\theta = 0,5$. Deseja-se construir um intervalo de confiança de 95% para θ.

```
5.02 1.11 0.04 0.48 1.59 0.75 5.1 0.38 2.33 0.68
```

Aplicando o resultado do exemplo anterior devemos obter $z_{\alpha/2}$ tal que

$$P\left(-z_{\alpha/2} < \frac{1/\overline{\boldsymbol{X}} - \theta}{\sqrt{\theta^2/n}} < z_{\alpha/2}\right) = 0,95$$

isto é, $z_{\alpha/2} = 1,96$. Da amostra obtemos que $\overline{x} = 1.7$ e isolando θ na desigualdade acima segue que

$$\frac{3.16/1.7}{3.16 + 1.96} < \theta < \frac{3.16/1.7}{3.16 - 1.96}$$

e o I.C. de 95% é $[0.36; 1.55]$.

Um fato importante é que, em geral, na distribuição assintótica normal do estimador de máxima verossimilhança a sua variância $I^{-1}(\theta)$ pode ser substituida pelo seu estimador $I^{-1}(\hat{\theta})$ sem afetar muito a acurácia da aproximação. Este fato, que não será provado aqui, simplifica bastante a conversão das desigualdades para obtenção de intervalos de confiança aproximados. Assim,

$$P\left(-z_{\alpha/2} < (\hat{\theta} - \theta)\sqrt{I(\hat{\theta})} < z_{\alpha/2}\right) \approx 1 - \alpha$$

é facilmente convertido para

$$P\left(\hat{\theta} - z_{\alpha/2}\sqrt{I^{-1}(\hat{\theta})} < \theta < \hat{\theta} + z_{\alpha/2}\sqrt{I^{-1}(\hat{\theta})}\right) \approx 1 - \alpha.$$

Note que este resultado foi utilizado na Seção 5.3 para construir um intervalo de confiança aproximado para uma proporção. Naquele caso, $\theta(1-\theta)/n$ era a variância de $\overline{\boldsymbol{X}}$ que foi substituida pelo seu estimador de máxima verossimilhança.

5.4.1 Usando a Função Escore

Em algumas situações não se tem uma forma explícita para o estimador de máxima verossimilhança e neste caso a função escore será particularmente útil. Lembrando que a função escore de X tem média zero e variância igual a $I(\theta)$ então temos pelo teorema central do limite que $\sum_{i=1}^{n} U(X_i; \theta)$ converge em distribuição para uma $N(0, I(\theta))$. Podemos usar este resultado para fazer inferência aproximada sobre θ e assim o intervalo de confiança aproximado de $100(1-\alpha)\%$ é obtido fazendo-se

$$P\left(\left|\frac{1}{\sqrt{I(\theta)}}\sum_{i=1}^{n} U(X_i; \theta)\right| < z_{\alpha/2}\right) \approx 1 - \alpha.$$

Finalmente, vale ressaltar que todos os resultados desta seção podem ser es-

tendidos para o caso de um vetor paramétrico $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$. Neste caso, a distribuição assintótica do estimador de máxima verossimilhança será normal multivariada com vetor de médias $\boldsymbol{\theta}$ e matriz de variância-covariância igual a $\boldsymbol{I}^{-1}(\boldsymbol{\theta})$ onde $\boldsymbol{I}(\boldsymbol{\theta})$ é a matriz de informação de Fisher.

5.5 Problemas

1. Seja X uma única observação de uma distribuição com densidade

$$p(x|\theta) = \theta x^{\theta-1}, 0 < x < 1, \theta > 0.$$

   (a) Mostre que $-\theta \log X$ é uma quantidade pivotal.

   (b) Use este pivot para construir um intervalo de confiança para θ com coeficiente de confiança 0,90.

2. No problema anterior, se $Y = (-\log X)^{-1}$ e $(Y/2, Y)$ é o intervalo de confiança para θ, calcule o coeficiente de confiança.

3. Seja $X_1, \ldots, X_n$ uma amostra aleatória da distribuição Exponencial(θ). Obtenha uma quantidade pivotal e mostre como construir um I.C. para θ. (Dica: mostre que $\min\{X_i\} \sim$ Exponencial($n\theta$).

4. Seja $X_1, \ldots, X_n$ uma amostra aleatória da distribuição $N(\theta, \theta)$. Obtenha uma quantidade pivotal para construir um intervalo de confiança para θ.

5. Se $X_{11}, \ldots, X_{1n_1}$ e $X_{21}, \ldots, X_{2n_2}$ são amostras aleatórias independentes das distribuições $N(\theta_1, \sigma_1^2)$ e $N(\theta_2, \sigma_2^2)$ mostre que

$$\frac{(\overline{X}_1 - \overline{X}_2) - (\theta_1 - \theta_2)}{S_p\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$$

   tem distribuição t-Student com $n_1 + n_2 - 2$ graus de liberdade.

6. Os pulsos em repouso de 920 pessoas sadias foram tomados, e uma média de 72,9 batidas por minuto (bpm) e um desvio padrão de 11,0 bpm foram obtidos. Construa um intervalo de confiança de 95% para a pulsação média em repouso de pessoas sadias com base nesses dados.

7. Tendo sido medido o eixo maior de 9 grãos de quartzo de um corpo arenoso em uma lâmina de arenito, obteve-se um comprimento amostral médio de 1,5mm e um desvio padrão de 0,3mm. Deseja-se construir um intervalo de confiança para o comprimento médio dos grãos de quartzo do corpo arenoso.

8. O tempo médio, por operário, para executar uma tarefa, tem sido de 100 minutos com desvio padrão de 15 minutos. Foi introduzida uma modificação para reduzir este tempo e após alguns meses foi selecionada uma amostra de 16 operários medindo-se o tempo de execução de cada um. Obteve-se um tempo médio amostral de 90 minutos e um desvio padrão de 16 minutos.

   (a) Estime o novo tempo médio de execução por um intervalo com 95% de confiança.

   (b) Inteprete o I.C. obtido no item anterior. Você diria que a modificação surtiu efeito? (Justifique).

   (c) Estime a nova variância populacional por um intervalo com 98% de confiança. É razoável concluir que a variância populacional se alterou?

9. Os QIs de 181 meninos com idades entre 6-7 anos de Curitiba foram medidos. O QI médio foi 108,08, e o desvio padrão foi 14,38.

   (a) Calcule um intervalo de confiança de 95% para o QI médio populacional dos meninos entre 6-7 anos de idade em Curitiba usando estes dados.

   (b) Interprete o intervalo de confiança com palavras.

   (c) Foi necessário assumir que os QIs têm distribuição normal neste caso? Por quê?

10. Em um experimento sobre o efeito do álcool na habilidade perceptual, 10 indivíduos são testados duas vezes, uma depois de ter tomado dois drinks e uma depois de ter tomado dois copos de água. Os dois testes foram realizados em dois dias diferentes para evitar influência do efeito do álcool. Metade dos indivíduos tomou a bebida alcoólica primeiro e a outra metade água. Os escores dos 10 indivíduos são mostrados abaixo. Escores mais altos refletem uma melhor performance. Verifique se a bebida alcoólica teve um efeito significativo com 99% de confiança.

	1	2	3	4	5	6	7	8	9	10
agua	16	15	11	20	19	14	13	15	14	16
alcool	13	13	12	16	16	11	10	15	9	16

11. Em um estudo de captura e recaptura a massa de 10 pássaros migratórios foi medida em duas ocasiões distintas. Os dados obtidos estão na tabela abaixo. Construa um intervalo de confiança para a diferença média de massas e verifique se houve ganho, redução ou manutenção de massa.

	1	2	3	4	5	6	7	8	9	10
medicao 1	10.3	11.4	10.9	12.0	10.0	11.9	12.2	12.3	11.7	12.0
medicao 2	12.2	12.1	13.1	11.9	12.0	12.9	11.4	12.1	13.5	12.3

12. Uma indústria compra componentes eletrônicos dos fornecedores A e B, mas o fornecedor A garante que o tempo médio de vida (em horas) do seu produto supera o da marca B em 300 horas. Para testar esta afirmação foram selecionadas duas amostras de 5 e 4 componentes, das marcas A e B respectivamente. As médias amostrais foram 1492 e 1182 e as variâncias amostrais foram 770 e 990.

    (a) Compare as variâncias dos tempos de vida através de um intervalo de confiança de 98%. É razoável assumir igualdade de variâncias?

    (b) Construa um intervalo de confiança de 95% para a diferença entre os tempos médios de vida.

    (c) Este intervalo dá alguma indicação sobre a afirmação do fornecedor A? Explique.

13. Os dados abaixo são uma amostra aleatória da distribuição de Bernoulli com $P(\text{sucesso})=p$. Construa os intervalos de confiança de 90% e 99% para p.

    0 0 0 1 1 0 1 1 1 1 0 1 1 0 1 1 1 1 0 1 1 1 1 1 1

14. Para decidir se uma moeda é balanceada (honesta) ela é lançada 40 vezes e cara aparece 13 vezes. Construa um intervalo de 95% de confiança para a verdadeira proporção de caras p. O que você conclui?

15. Numa pesquisa eleitoral, 57 dentre 150 entrevistados afirmaram que votariam no candidato X. Com uma confiança de 90%, o que você pode dizer acerca da proporção real de votos aquele candidato terá?

16. Dentre 100 peixes capturados num certo lago, 18 não estavam apropriados para consumo devido aos níveis de poluição do ambiente. Construa um intervalo de confiança de 99% para a verdadeira proporção de peixes contaminados.

17. Seja $X_1, \ldots, X_n$ uma amostra aleatória da distribuição normal com média μ desconhecida e variância σ^2 conhecida. Qual deve ser o tamanho da amostra tal que exista um intervalo de confiança para μ com coeficiente de confiança 0,95 e comprimento menor do que $0{,}01\sigma$?

18. Seja $X_1, \ldots, X_n$ uma amostra aleatória da distribuição exponencial com média θ desconhecida. Descreva um método para construir um intervalo de confiança de $100(1-\alpha)\%$ para θ. (Sugestão: Determine as constantes c_1 e c_2 tais que $P(c_1 < (1/\theta)\sum_{i=1}^{n} X_i < c_2) = 1-\alpha$).

19. Seja $X_1, \ldots, X_n$ uma amostra aleatória da distribuição $Beta(\theta, 1)$. Obtenha o intervalo de confiança aproximado de $100(1-\alpha)\%$ baseando-se na distribuição assintótica da função escore.

20. Se uma variável aleatória X tem distribuição de Poisson com média θ obtenha a informação esperada de Fisher $I(\theta)$ através de X.

21. Suponha que uma variável aleatória X tem distribuição normal com média zero e desvio-padrão desconhecido σ. Obtenha a informação esperada de Fisher $I(\sigma)$ através de X. Suponha agora que a variância seja o parâmetro de interesse e obtenha a informação de Fisher de σ^2 através de X.

22. Seja $X_1, \ldots, X_n$ uma amostra aleatória da distribuição $N(0, \sigma^2)$. Construa um intervalo de confiança aproximado para o desvio-padrão σ baseado no seu estimador de máxima verossimilhança.

23. Seja $X_1, \ldots, X_n$ uma amostra aleatória da distribuição Exponencial com parâmetro θ. Construa um intervalo de confiança aproximado para $E(X)$ e $Var(X)$.

5.6 Intervalos Bayesianos

Do ponto de Bayesiano, todas as quantidades desconhecidas (parâmetros, dados omissos, etc) são variáveis aleatórias. Em princípio, a forma mais adequada de expressar a informação que se tem sobre um parâmetro é através de sua distribuição a posteriori. Nesta seção vamos introduzir um compromisso entre o uso da própria distribuição a posteriori e uma estimativa pontual. Será discutido o conceito de intervalo de credibilidade (ou intervalo de confiança Bayesiano) baseado no distribuição a posteriori.

Definição 5.1 *C é um intervalo de credibilidade de 100(1-α)%, ou nível de credibilidade (ou confiança) $1-\alpha$, para θ se $P(\theta \in C) \geq 1-\alpha$.*

Note que a definição expressa de forma probabilística a pertinência ou não de θ ao intervalo. Assim, quanto menor for o tamanho do intervalo mais concentrada é a distribuição do parâmetro, ou seja o tamanho do intervalo informa sobre a dispersão de θ. Além disso, a exigência de que a probabilidade acima possa ser maior do que o nível de confiança é essencialmente técnica pois queremos que o

intervalo seja o menor possível, o que em geral implica em usar uma igualdade. Ou seja, queremos obter c_1 e c_2 tais que,

$$\int_{c_1}^{c_2} p(\theta|\boldsymbol{x})d\theta = 1 - \alpha.$$

No entanto, a desigualdade será útil se θ tiver uma distribuição discreta onde nem sempre é possível satisfazer a igualdade.

Outro fato importante é que os intervalos de credibilidade são invariantes a transformações 1 a 1, $\phi(\theta)$. Ou seja, se $C = [a, b]$ é um intervalo de credibilidade 100(1-α)% para θ então $[\phi(a), \phi(b)]$ é um intervalo de credibilidade 100(1-α)% para $\phi(\theta)$. Note que esta propriedade também vale para intervalos de confiança na inferência clássica.

É possível construir uma infinidade de intervalos usando a definição acima mas estamos interessados apenas naquele com o menor comprimento possível. Pode-se mostrar que intervalos de comprimento mínimo são obtidos tomando-se os valores de θ com maior densidade a posteriori, e esta idéia é expressa matematicamente na definição abaixo.

Definição 5.2 *Um intervalo de credibilidade C de 100(1-α)% para θ é de máxima densidade a posteriori (MDP) se $C = \{\theta \in \Theta : p(\theta|\boldsymbol{x}) \geq k(\alpha)\}$ onde $k(\alpha)$ é a maior constante tal que $P(\theta \in C) \geq 1 - \alpha$.*

Usando esta definição, todos os pontos dentro do intervalo MDP terão densidade maior do que qualquer ponto fora do intervalo. Além disso, no caso de distribuições com duas caudas, e.g. normal, t de Student, o intervalo MDP é obtido de modo que as caudas tenham a mesma probabilidade.

Um problema com os intervalos MDP é que eles não são invariantes a transformações 1 a 1, a não ser para transformações lineares. O mesmo problema ocorre com intervalos de comprimento mínimo na inferência clássica.

5.7 Estimação no Modelo Normal

Os resultados desenvolvidos nos capítulos anteriores serão aplicados ao modelo normal para estimação da média e variância em problemas de uma ou mais amostras e em modelos de regressão linear. A análise será feita com priori conjugada e priori não informativa quando serão apontadas as semelhanças com a análise clássica. A abordagem aqui é introdutória, um tratamento mais completo do enfoque Bayesiano em modelos lineares pode ser encontrado em Broemeling (1985) e Box e Tiao (1992).

Nesta seção considere uma amostra aleatória $X_1, \cdots, X_n$ tomada da distribuição $N(\theta, \sigma^2)$.

5.7.1 Variância Conhecida

Se σ^2 é conhecido e a distribuição a priori de θ é $N(\mu_0, \tau_0^2)$ então, do Teorema 4.1, obtém-se que distribuição a posteriori de θ também é normal com média μ_1 e variância τ_1^2 dados por

$$\mu_1 = \frac{\tau_0^{-2}\mu_0 + n\sigma^{-2}\overline{x}}{\tau_0^{-2} + n\sigma^{-2}} \quad \text{e} \quad \tau_1^{-2} = \tau_0^{-2} + n\sigma^{-2}.$$

Assim temos que,

$$\begin{aligned} X_1, \cdots, X_n &\sim N(\theta, \sigma^2) \\ \theta &\sim N(\mu_0, \tau_0^2) \\ \theta|\boldsymbol{x} &\sim N(\mu_1, \tau_1^2) \end{aligned}$$

Portanto, intervalos de confiança Bayesianos para θ podem então ser construídos usando o fato de que

$$\frac{\theta - \mu_1}{\tau_1}|\boldsymbol{x} \sim N(0,1).$$

Assim, usando uma tabela da distribuição normal padronizada podemos obter o valor do percentil $z_{\alpha/2}$ tal que

$$P\left(-z_{\alpha/2} \leq \frac{\theta - \mu_1}{\tau_1} \leq z_{\alpha/2}\right) = 1 - \alpha$$

e após isolar θ, obtemos que

$$P\left(\mu_1 - z_{\alpha/2}\ \tau_1 \leq \theta \leq \mu_1 + z_{\alpha/2}\ \tau_1\right) = 1 - \alpha.$$

Portanto $\left(\mu_1 - z_{\alpha/2}\ \tau_1; \mu_1 + z_{\alpha/2}\ \tau_1\right)$ é o intervalo de credibilidade 100(1-α)% MDP para θ, devido à simetria da normal.

A priori não informativa pode ser obtida fazendo-se a variância da priori tender a infinito, i.e. $\tau_0^2 \to \infty$. Neste caso, é fácil verificar que

$$\tau_1^{-2} \to n\sigma^{-2} \quad \text{e} \quad \mu_1 \to \overline{\boldsymbol{x}},$$

ou seja a média e a precisão da posteriori convergem para a média e a precisão amostrais. Média, moda e mediana a posteriori coincidem então com a estimativa clássica de máxima verossimilhança, $\overline{\boldsymbol{x}}$. O intervalo de confiança Bayesiano de 100(1-α)% para θ é dado por

$$\left(\overline{\boldsymbol{x}} - z_{\alpha/2}\ \frac{\sigma}{\sqrt{n}};\ \overline{\boldsymbol{x}} + z_{\alpha/2}\ \frac{\sigma}{\sqrt{n}}\right)$$

e também coincide numericamente com o intervalo de confiança clássico.

É importante notar que esta coincidência é apenas numérica uma vez que a interpretação do intervalo Bayesiano é como uma afirmação probabilística sobre θ.

Exemplo 5.9 : Sejam $X_1, \ldots, X_n$ os tempos (em minutos) de execução de uma tarefa medidos para 16 operários selecionados ao acaso. Sabe-se que o desvio padrão populacional destes tempos é igual a 15 minutos e obteve-se um tempo médio amostral de 90 minutos.

Assumindo que $X \sim N(\theta, \sigma^2)$ com $\sigma = 15$ e usando uma distribuição a priori não informativa para θ segue que a sua distribuição a posteriori é

$$\theta | x_1, \ldots, x_n \sim N(\overline{x}, \sigma^2/n)$$

e para construir o I.C. Bayesiano de 95% para θ obtemos de uma tabela da normal padrão que $z_{0,025} = 1.96$. Assim, o intervalo fica

$$\left[90 - 1.96 \times \frac{15}{\sqrt{16}}; 90 + 1.96 \times \frac{15}{\sqrt{16}}\right] = [82.65; 97.35].$$

Ou seja, após observar os dados a probabilidade do tempo médio de execução estar neste intervalo é 0,95, i.e.

$$P(82.65 < \theta < 97.35) = 0.95.$$

Uma função geral pode ser escrita no `R` para se obter o intervalo MDP e opcionalmente fazer os gráficos das densidades.

```
> ic.mdp = function(x, sigma, mu0, tau0, plot = F, conf = 0.95) {
+     n = length(x)
+     xbar = mean(x)
+     ep = sigma/sqrt(n)
+     sigma2 = sigma^2
+     precisao = n * (1/sigma2) + (1/tau0)
+     mu1 = (n * (1/sigma2) * xbar + (1/tau0) * mu0)/precisao
+     tau1 = 1/precisao
+     if (plot) {
+         curve(dnorm(x, xbar, ep), xbar - 3 * ep, xbar + 3 * ep)
+         curve(dnorm(x, mu0, sqrt(tau0)), add = T, col = 2)
+         curve(dnorm(x, mu1, 1/sqrt(precisao)), add = T, col = 3)
+     }
+     z = qnorm((1 - conf)/2, lower = F)
+     c(mu1 - z * sqrt(tau1), mu1 + z * sqrt(tau1))
```

```
+ }
```

Exemplo 5.10 : No Exemplo 5.9 sabe-se que o tempo médio de execução tem sido de 100 minutos com desvio padrão igual a 10 minutos. Podemos usar esta informação como priori para o tempo médio ou seja $\theta \sim N(\mu_0, \tau_0^2)$ com $\mu_0 = 100$ e $\tau_0 = 10$. Assim, segue que

$$\begin{aligned} \theta|x_1, \ldots, x_n &\sim N(\mu, \tau_1^2) \\ \tau_1^{-2} &= \frac{16}{15^2} + \frac{1}{10^2} = 0.0811 \\ \mu_1 &= \frac{(16/15^2)(90) + (1/10^2)(100)}{0.0811} = 91.245 \end{aligned}$$

e o I.C. Bayesiano de 95% fica

$$\left[91.245 - 1.96\sqrt{\frac{1}{0.0811}};\ 91.245 + 1.96\sqrt{\frac{1}{0.0811}}\right] = [84.36;\ 98.13].$$

Usando a função "ic.mdp" obtemos

```
ic.mdp(x=rep(90,16),sigma=15,mu0=100,tau0=100,plot=F,conf=0.95)
[1] 84.35098 98.11477
```

5.7.2 Média e Variância desconhecidas

Neste caso deve-se obter uma distribuição a posteriori para os 2 parâmetros (θ, σ^2) via teorema de Bayes, i.e.

$$p(\theta, \sigma^2|\boldsymbol{x}) \propto p(\boldsymbol{x}|\theta, \sigma^2)\ p(\theta, \sigma^2).$$

Começaremos especificando uma priori não informativa e uma forma de fazer isto é assummir que θ e σ são a priori independentes e que $(\theta, \log(\sigma))$ tem distribuição uniforme. Isto equivale a dizer que

$$p(\theta, \sigma^2) \propto 1/\sigma^2.$$

A função de verossimilhança é dada por

$$\begin{aligned} p(\boldsymbol{x}|\theta,\sigma^2) &= (2\pi\sigma^2)^{-n/2}\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\theta)^2\right\} \\ &= (2\pi\sigma^2)^{-n/2}\exp\left\{-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{n}(x_i-\overline{x})^2+n(\overline{x}-\theta)^2\right)\right\} \\ &\propto \sigma^{-n}\exp\left\{-\frac{1}{2\sigma^2}\left((n-1)s^2+n(\overline{x}-\theta)^2\right)\right\} \end{aligned}$$

sendo s^2 a variância amostral. Aplicando o teorema de Bayes obtemos então que

$$p(\theta,\sigma^2|\boldsymbol{x}) \propto \sigma^{-n-2}\exp\left\{-\frac{1}{2\sigma^2}\left((n-1)s^2+n(\overline{x}-\theta)^2\right)\right\}. \tag{5.1}$$

Da expressão (5.1) e usando novamente o Teorema 4.1 não é difícil verificar que a densidade a posteriori de θ condicionada em σ^2 fica

$$p(\theta|\boldsymbol{x},\sigma^2) \propto p(\theta,\sigma^2|\boldsymbol{x}) \propto \exp\left\{-\frac{n}{2\sigma^2}(\overline{x}-\theta)^2\right\}$$

ou seja, $\theta|\boldsymbol{x},\sigma^2 \sim N(\overline{x},\sigma^2/n)$.

Distribuição Marginal de σ^2

O próximo passo é obter a distribuição a posteriori marginal de σ^2 e para isto basta integrar a densidade a posteriori conjunta em relação a θ. Assim,

$$\begin{aligned} p(\sigma^2|\boldsymbol{x}) &= \int_{-\infty}^{\infty}\sigma^{-n-2}\exp\left\{-\frac{1}{2\sigma^2}[(n-1)s^2+n(\overline{x}-\theta)^2]\right\}d\theta \\ &= \sigma^{-n-2}\exp\left\{-\frac{1}{2\sigma^2}(n-1)s^2\right\}\int_{-\infty}^{\infty}\exp\left\{-\frac{n}{2\sigma^2}(\overline{x}-\theta)^2\right\}d\theta \end{aligned}$$

Nesta última integral temos o núcleo de uma função de densidade normal com média $\overline{x}$ e variância σ^2/n, portanto ela é dada simplesmente por

$$\int_{-\infty}^{\infty}\exp\left\{-\frac{n}{2\sigma^2}(\overline{x}-\theta)^2\right\}d\theta = \sqrt{2\pi\sigma^2/n}.$$

Conclui-se então que

$$p(\sigma^2|\boldsymbol{x}) \propto (\sigma^2)^{-(n+1)/2}\exp\left\{-\frac{(n-1)s^2}{2\sigma^2}\right\}.$$

e portanto (ver Apêndice A)

$$\sigma^2|\boldsymbol{x} \sim GI\left(\frac{n-1}{2}, \frac{(n-1)s^2}{2}\right).$$

Finalmente, pelo teorema de transformação de variáveis pode-se mostrar que

$$\sigma^{-2}|\boldsymbol{x} \sim Gama\left(\frac{n-1}{2}, \frac{(n-1)s^2}{2}\right)$$

ou equivalentemente,

$$\frac{(n-1)s^2}{\sigma^2}|\boldsymbol{x} \sim \chi^2_{n-1}.$$

Agora podemos então construir um intervalo de probabilidade para σ^2. Obtenha os percentis $\underline{\chi}^2_{\alpha/2,n-1}$ e $\overline{\chi}^2_{\alpha/2,n-1}$ desta distribuição qui-quadrado tais que

$$P\left(\underline{\chi}^2_{\alpha/2,n-1} \leq \frac{(n-1)s^2}{\sigma^2} \leq \overline{\chi}^2_{\alpha/2,n-1}\right) = 1-\alpha.$$

O intervalo de credibilidade de $100(1-\alpha)\%$ para σ^2 é dado então por

$$\left(\frac{(n-1)s^2}{\overline{\chi}^2_{\alpha/2,n-1}}, \frac{(n-1)s^2}{\underline{\chi}^2_{\alpha/2,n-1}}\right).$$

Exemplo 5.11 : No Exemplo 5.9 suponha agora que a variância populacional é desconhecida e sua estimativa amostral foi $s^2 = 16$. Neste caso a distribuição a posteriori de $(15)(16)/\sigma^2$ é qui-quadrado com 15 graus de liberdade e os percentis de ordem 0.025 inferior e superior são 6.2621 e 27.4884 respectivamente, i.e.

$$P(6.2621 < (15)(16)/\sigma^2 < 27.4884) = 0.95.$$

Assim, o intervalo de probabilidade fica

$$\left[\frac{(15)(16)}{27.4884};\ \frac{(15)(16)}{6.2621}\right] = [8.73;\ 38.33].$$

Note que este intervalo não é simétrico em torno de $s^2 = 16$,

$$\begin{aligned} P(8,73 < (15)(16)/\sigma^2 < 15) &= 0,4398 \\ P(15 < (15)(16)/\sigma^2 < 38,33) &= 0,4506. \end{aligned}$$

Exemplo 5.12 : Ainda no Exemplo 5.9, recebemos a informação de que em

outro setor da empresa o tempo de execução desta mesma tarefa tem variância igual a 10. Após introduzir algumas alterações foram observados 16 operários e seus tempos de execução em minutos resultaram em $s^2 = 16$. O intervalo, tanto clássico quanto Bayesiano, de 95% para σ^2 é exatamente [8,73; 38,33]. O estatístico clássico diria que não indicação nos dados de que a variância tenha se alterado (de 10 para outro valor). No entanto,

$$\begin{aligned} P(8.73 < \sigma^2 < 10) &= P\left(\frac{15\times 16}{10} < \frac{15\times 16}{\sigma^2} < \frac{15\times 16}{8.73}\right) \\ &= P\left(24 < \frac{15\times 16}{\sigma^2} < 27.49\right) = 0.04 \\ P(10 < \sigma^2 < 38.33) &= P\left(\frac{15\times 16}{38.33} < \frac{15\times 16}{\sigma^2} < \frac{15\times 16}{10}\right) \\ &= P\left(6.26 < \frac{15\times 16}{\sigma^2} < 24\right) = 0.91. \end{aligned}$$

A situação está descrita na Figura 5.2.

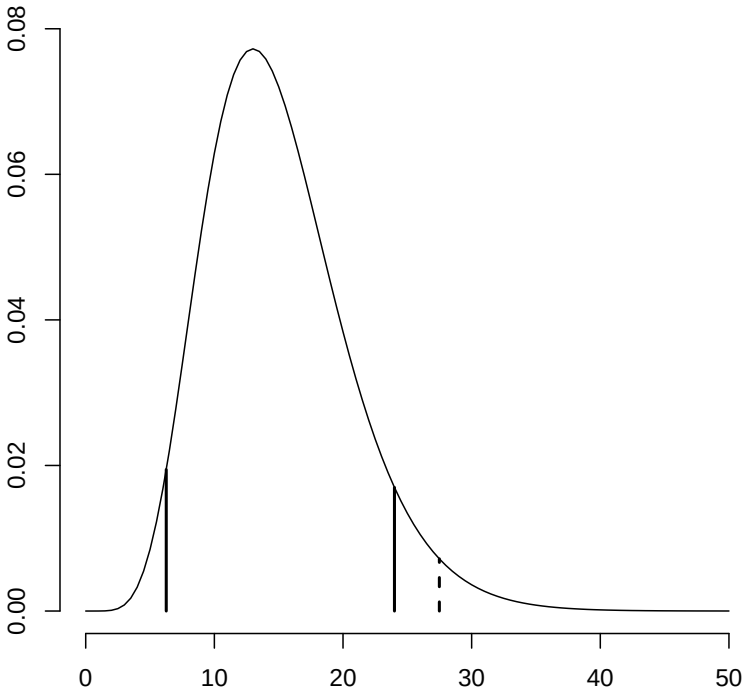

Figura 5.2: Intervalo de 95% de probabilidade para $(n-1)s^2/\sigma^2$.

Distribuição Marginal de θ

Tipicamente estaremos interessados em estimar a média do processo, i.e. o parâmetro θ. Do ponto de vista Bayesiano, todo a inferência é feita com base na distribuição a posteriori marginal de θ obtida como

$$p(\theta|\boldsymbol{x}) = \int_0^\infty p(\theta, \sigma^2|\boldsymbol{x})d\sigma^2 = \int_0^\infty p(\theta|\boldsymbol{x}, \sigma^2)p(\sigma^2|\boldsymbol{x})d\sigma^2.$$

Usando a expressão (5.1) segue que

$$p(\theta|\boldsymbol{x}) \propto \int_0^\infty (\sigma^2)^{-(n/2+1)} \exp\left\{-\frac{1}{2\sigma^2}\left((n-1)s^2 + n(\overline{x}-\theta)^2\right)\right\} d\sigma^2$$

e do Apêndice A pode-se notar que o integrando é o núcleo de uma densidade Gama Inversa com parâmetros $n/2$ e $(n-1)s^2 + n(\overline{x}-\theta)^2$. Portanto a integral é dada por

$$\begin{aligned} p(\theta|\boldsymbol{x}) &\propto \frac{\Gamma(n/2)}{[(n-1)s^2 + n(\overline{x}-\theta)^2]^{n/2}} \propto [(n-1)s^2 + n(\overline{x}-\theta)^2]^{-n/2} \\ &\propto \left[(n-1) + \frac{n(\overline{x}-\theta)^2}{s^2}\right]^{-n/2} \end{aligned}$$

que é o núcleo da distribuição t de Student com $n-1$ graus de liberdade, parâmetro de locação $\overline{x}$ e parâmetro de escala s^2/n (ver Apêndice A). Ou seja,

$$\theta|\boldsymbol{x} \sim t_{n-1}(\overline{x}, s^2/n).$$

ou equivalentemente,

$$\frac{\theta - \overline{x}}{s/\sqrt{n}}|\boldsymbol{x} \sim t_{n-1}(0,1).$$

A conclusão final é que mais uma vez um intervalo Bayesiano irá coincidir numericamente com um intervalo de confiança clássico. O intervalo de probabilidade 100(1-α)% de MDP é dado por

$$\left[\overline{x} - t_{\alpha/2,n-1}\frac{s}{\sqrt{n}};\ \overline{x} + t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}\right]$$

e a probabilidade de θ pertencer a este intervalo é $1-\alpha$.

Mais uma vez vale enfatizar que esta coincidência com as estimativas clássicas é apenas numérica uma vez que as interpretações dos intervalos diferem radicalmente.

Exemplo 5.13 : Voltando ao Exemplo 5.9, usando priori não informativa o

intervalo Bayesiano será exatamente o mesmo, i.e. [82,65; 97,35], porém com uma interpretação probabilística,

$$P(\theta \in [82,65; 97,35] \mid \boldsymbol{x}) = 0,95.$$

5.7.3 O Caso de duas Amostras

Nesta seção vamos assumir que $X_{11}, \ldots, X_{1n_1}$ e $X_{21}, \ldots, X_{2n_2}$ são amostras aleatórias das distribuições $N(\theta_1, \sigma_1^2)$ e $N(\theta_2, \sigma_2^2)$ respectivamente e que as amostras são independentes.

Para começar vamos assumir que as variâncias σ_1^2 e σ_2^2 são conhecidas. Neste caso, a função de verossimilhança é dada por

$$\begin{aligned} p(\boldsymbol{x}_1, \boldsymbol{x}_2|\theta_1, \theta_2) &= p(\boldsymbol{x}_1|\theta_1)\ p(\boldsymbol{x}_2|\theta_2) \\ &\propto \exp\left\{-\frac{n_1}{2\sigma_1^2}(\theta_1 - \overline{x}_1)^2\right\}\ \exp\left\{-\frac{n_2}{2\sigma_2^2}(\theta_2 - \overline{x}_2)^2\right\} \end{aligned}$$

isto é, o produto de verossimilhanças relativas a θ_1 e θ_2. Assim, se assumirmos que θ_1 e θ_2 são independentes a priori então eles também serão independentes a posteriori já que, pelo Teorema de Bayes

$$\begin{aligned} p(\theta_1, \theta_2|\boldsymbol{x}_1, \boldsymbol{x}_2) &= \frac{p(\boldsymbol{x}_1|\theta_1)\ p(\theta_1)}{p(\boldsymbol{x}_1)} \times \frac{p(\boldsymbol{x}_2|\theta_2)\ p(\theta_2)}{p(\boldsymbol{x}_2)} \\ &= p(\theta_1|\boldsymbol{x}_1)p(\theta_2|\boldsymbol{x}_2) \end{aligned}$$

Se usarmos a classe de prioris conjugadas da Seção 5.7.1 ou seja

$$\theta_i \sim N(\mu_i, \tau_i^2)$$

então as distribuições a posteriori independentes serão

$$\theta_i|\boldsymbol{x}_i \sim N(\mu_i^*, \tau_i^{*^2}),\ i = 1, 2$$

sendo a média e a variância dadas por

$$\mu_i^* = \frac{\tau_i^{-2}\mu_i + n_i\ \sigma_i^{-2}\ \overline{\boldsymbol{x}}_i}{\tau_i^{-2} + n_i\ \sigma_i^{-2}} \quad \text{e} \quad \tau_i^{*^2} = 1/(\tau_i^{-2} + n_i\ \sigma_i^{-2}), \quad i = 1, 2.$$

Em geral estaremos interessados em comparar as médias populacionais, i.e queremos estimar $\beta = \theta_1 - \theta_2$. Neste caso, a posteriori de β é facilmente obtida, devido à independência, como

$$\beta|\boldsymbol{x}_1, \boldsymbol{x}_2 \sim N(\mu_1^* - \mu_2^*, \tau_1^{*^2} + \tau_2^{*^2})$$

e podemos usar $\mu_1^* - \mu_2^*$ como estimativa pontual para a diferença e também construir um intervalo de credibilidade MDP para esta diferença. Note que se usarmos priori não informativa, i.e. fazendo $\tau_i^2 \to \infty$, $i = 1, 2$ então a posteriori fica

$$\beta|\boldsymbol{x}_1, \boldsymbol{x}_2 \sim N\left(\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$$

e o intervalo obtido coincidirá mais uma vez com o intervalo de confiança clássico. Podemos escrever então que o intervalo de credibilidade MDP digamos de 95% é

$$\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2 - 1,96\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}; \overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2 + 1,96\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}.$$

Variâncias Desconhecidas

No caso de variâncias populacionais desconhecidas porém iguais, temos que $\sigma_1^2 = \sigma_2^2 = \sigma^2$ e novamente podemos definir a variância amostral combinada

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}.$$

Vejamos agora como fica a análise usando priori não informativa. Neste caso, pode-se mostrar que a distribuição a priori é dada por

$$p(\theta_1, \theta_2, \sigma^2) \propto 1/\sigma^2$$

e as distribuições a posteriori marginais de $\theta_1 - \theta_2$ e σ^2 são

$$\theta_1 - \theta_2|\boldsymbol{x}_1, \boldsymbol{x}_2 \sim t_{n_1+n_2-2}\left(\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2, s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right)$$

ou equivalentemente,

$$\frac{\theta_1 - \theta_2 - (\boldsymbol{x}_1 - \boldsymbol{x}_2)}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2}$$

e

$$\sigma^{-2} \sim Gamma\left(\frac{n_1 + n_2 - 2}{2}, \frac{(n_1 + n_2 - 2)s_p^2}{2}\right)$$

ou equivalentemente,

$$\frac{(n_1 + n_2 - 2)s_p^2}{\sigma^2} \sim \chi^2_{n_1+n_2-2}.$$

O intervalo de $100(1 - \alpha)\%$ para σ^2 é obtido de maneira análoga ao caso de uma amostra usando a distribuição qui-quadrado, agora com $n_1 + n_2 - 2$ graus de

liberdade, i.e.

$$\left(\frac{(n_1+n_2-2)s_p^2}{\overline{\chi}^2_{\frac{\alpha}{2}}}, \frac{(n_1+n_2-2)s_p^2}{\underline{\chi}^2_{\frac{\alpha}{2}}} \right).$$

Variâncias desiguais

Até agora assumimos que as variâncias populacionais desconhecidas eram iguais (ou pelo menos aproximadamente iguais). Na inferência clássica a violação desta suposição leva a problemas teóricos e práticos uma vez que não é trivial encontrar uma quantidade pivotal para β com distribuição conhecida ou tabelada. Do ponto de vista Bayesiano o que precisamos fazer é combinar informação a priori com a verossimilhança e basear a estimação na distribuição a posteriori. A função de verossimilhança agora pode ser fatorada como

$$p(\boldsymbol{x}_1, \boldsymbol{x}_2|\theta_1, \theta_2, \sigma_1^2, \sigma_2^2) = p(\boldsymbol{x}_1|\theta_1, \sigma_1^2)p(\boldsymbol{x}_2|\theta_2, \sigma_2^2).$$

A análise usando priori não informativa pode ser feita assumindo que

$$p(\theta_1, \theta_2, \sigma_1^2, \sigma_2^2) \propto \sigma_1^{-2}\sigma_2^{-2}$$

e a obtenção das posterioris marginais de $(\theta_1 - \theta_2)$, σ_1^2 e σ_2^2 será deixada como exercício.

5.8 Problemas

1. Refaça o Exemplo 5.9 sabendo que o tempo médio de execução tem sido de 100 minutos com desvio padrão igual a 10 minutos. Use esta informação como priori para o tempo médio e construa um I.C. Bayesiano de 95%.

2. Seja $X_1, \ldots, X_n$ uma amostra aleatória da distribuição normal com média μ desconhecida e variância σ^2 conhecida. Usando uma priori não informativa para μ, obtenha o tamanho da amostra tal que $P(a < \mu < b|\boldsymbol{x}) = 0,95$ e o comprimento do intervalo (a, b) seja menor do que $0{,}01\sigma$.

3. Seja $X_1, \ldots, X_n$ uma amostra aleatória de tamanho 16 da distribuição $N(\mu, 1)$. Sabendo-se que foi observado $\sum_{i=1}^n x_i = 160$ e usando uma priori não informativa, obtenha um intervalo de credibilidade MDP de 95% para μ. Interprete este intervalo.

4. Repita o problema 3 supondo agora que a variância populacional (σ^2) também é desconhecida, assumindo uma priori não informativa e sabendo que foi observado $s^2 = 1$. Construa também um intervalo de credibilidade para σ^2.

5. Suponha que $X_1, \ldots, X_n \sim N(\theta, \phi)$ sendo θ conhecido e $\phi = \sigma^{-2}$ (o inverso da variância) desconhecido. Se a distribuição a priori de ϕ for $\phi \sim Gama(a, b)$ mostre que a sua distribuição a posteriori será

$$\phi|\boldsymbol{x} \sim Gama\left(a + \frac{n}{2}, b + \frac{1}{2}\sum_{i=1}^{n}(x_i - \theta)^2\right)$$

6. Seja $X_1, \ldots, X_n$ uma amostra aleatória de tamanho 10 da distribuição $Poisson(\theta)$ sendo cada X_i o número de defeitos por m^2 de uma placa metálica. Usando uma distribuição a priori não informativa para θ dada por $p(\theta) \propto \theta^{-1/2}$,

   (a) Verifique que a distribuição a posteriori de θ é dada por

   $$\theta|\boldsymbol{x} \sim Gama\left(\sum_{i=1}^{n} x_i + \frac{1}{2}, n\right).$$

   (b) Obtenha um intervalo de credibilidade de 95% para o número médio de defeitos sabendo que o total observado de defeitos foi 10.

   (c) Repita os itens anteriores usando uma priori $p(\theta) \propto \theta^{a-1}\exp(-b\theta)$ com $a > 0$ e $b > 0$.

7. Uma moeda cuja probabilidade de cara é desconhecida foi lançada 10 vezes e observou-se 7 caras e 3 coroas. Usando uma distribuição a priori uniforme no intervalo (0,1) calcule um intervalo Bayesiano de 90% para a probabilidade de cara.

8. O número de defeitos em um item manufaturado tem distribuição de Poisson com parâmetro λ. Atribui-se uma distribuição a priori exponencial com parâmetro 1 para λ. Se em 5 itens selecionados ao acaso observou-se um total de 8 defeitos calcule o intervalo Bayesiano de 90% para λ.

Capítulo 6

Testes de Hipóteses

6.1 Introdução e notação

Em geral, intervalos de confiança são a forma mais informativa de apresentar os achados principais de um estudo. Contudo, algumas vezes existe um particular interesse em verificar determinadas afirmações ou conjecturas. Por exemplo, podemos estar interessados em determinar se uma moeda é honesta, se certas quantidades são independentes, ou se populações distintas são similares do ponto de vista probabilístico. Cada uma destas afirmações constitui uma hipótese que pode ser associada a um modelo, i.e. pode ser parametrizada. O material deste capítulo é fortemente baseado em DeGroot (1989), Migon and Gamerman (1999) e DeGroot and Schervish (2002). A teoria clássica de testes de hipóteses é apresentada a um nível mais formal em Lehman and Romano (2005).

Chamamos de *hipótese estatística* qualquer afirmação que se faça sobre um parâmetro populacional desconhecido. A idéia básica é que a partir de uma amostra da população iremos estabelecer uma *regra de decisão* segundo a qual rejeitaremos ou não a hipótese proposta. Esta regra de decisão é chamada de *teste*. Normalmente existe uma hipótese que é mais importante para o pesquisador que será denotada por H_0 e chamada *hipótese nula*. Qualquer outra hipótese diferente de H_0 será chamada de *hipótese alternativa* e denotada por H_1.

Exemplo 6.1 : (Teste Binomial) Um professor aplica um teste do tipo certo-errado com 10 questões. Queremos testar a hipótese de que o aluno está advinhando.

Nossa hipótese nula é que o aluno acerta as questões ao acaso e a hipótese alternativa é que ele tem algum conhecimento da matéria. Denotando por p a probabilidade (desconhecida) do aluno acertar cada questão a hipótese estatística de interesse pode ser formulada como $H_0 : p = 1/2$. Neste caso, a hipótese alternativa mais adequada é $H_1 : p > 1/2$ indicando que o aluno tem algum conhecimento

sobre o assunto. Temos então 10 repetições do experimento com p constante e vamos assumir também que as questões são resolvidas de forma independente. Portanto a variável aleatória $X =$ "número de acertos" tem distribuição binomial com parâmetros $n = 10$ e p desconhecido. Suponha que adotamos a seguinte regra de decisão: o aluno não está advinhando se acertar 8 ou mais questões. Isto equivale a

rejeitar H_0 se $X \geq 8$ (*região de rejeição* ou *região crítica*) e
aceitar H_0 se $X < 8$ (*região de aceitação*).

No entanto, é possível que um aluno acerte 8 ou mais questões e esteja advinhando, ou seja podemos rejeitar H_0 quando ela é verdadeira. A probabilidade de que isto ocorra é

$$P(X \geq 8 \mid p = 1/2) = \sum_{k=8}^{10} 0.5^k (1 - 0.5)^{10-k} \approx 0.055.$$

Esta probabilidade é chamada *nível de significância* e será denotada por α. Fica claro então que o valor de α depende da regra de decisão, por exemplo se a região crítica fosse $X \geq 7$ teriamos $\alpha \approx 0,171$. No próximo exemplo veremos como usar o nível de significância para construir uma regra de decisão.

Exemplo 6.2 : Um fornecedor garante que 90% de sua produção não apresenta defeito. Para testar esta afirmação selecionamos ao acaso 10 itens de um lote e contamos o número de defeituosos. Com base nesta amostra tomaremos uma decisão: comprar ou não comprar o lote. É bem intuitivo que devemos decidir não comprar o lote se o número observado de não defeituosos for muito pequeno. O nosso problema é definir o quão pequeno.

Seja a variável aleatória $X =$ "número de não defeituosos na amostra de 10 itens". Temos então uma distribuição binomial com parâmetros $n = 10$ e p desconhecido, e queremos testar $H_0 : p = 0.9$. Aqui p é a proporção de itens não defeituosos no lote e portanto a hipótese alternativa deve ser $H_1 : p < 0.9$. Suponha que decidimos manter $\alpha \leq 0.025$ e a partir deste valor vamos estabelecer a nossa regra de decisão, ou seja obter o valor da constante c tal que H_0 é rejeitada

se $X \leq c$. Para isto vamos calcular α para diferentes regiões críticas,

$$\begin{aligned} P(X \leq 5 \mid p = 0.9) &= \sum_{k=0}^{5} 0.9^k (1-0.9)^{10-k} = 0.002 \\ P(X \leq 6 \mid p = 0.9) &= \sum_{k=0}^{6} 0.9^k (1-0.9)^{10-k} = 0.013 \\ P(X \leq 7 \mid p = 0.9) &= \sum_{k=0}^{7} 0.9^k (1-0.9)^{10-k} = 0.07. \end{aligned}$$

Portanto, devemos usar a região crítica $X \leq 6$. Isto é, vamos rejeitar o lote se o número de itens defeituosos na amostra for maior ou igual a 4.

Nestes dois exemplos os testes são chamados de *unilaterais* porque somente valores de um lado do espaço amostral foram utilizados para construir a região crítica. As regiões críticas são mostradas nos gráficos da Figura 6.1. Podemos ter também testes *bilaterais* aonde os dois extremos do espaço amostral são usados como região crítica. A variável aleatória X é chamada *estatística de teste*, sua distribuição deve ser conhecida e ela deve depender do parâmetro que está sendo testado.

No caso geral então temos uma amostra aleatória $\boldsymbol{X} = (X_1, \ldots, X_n)$ tomada de uma distribuição que envolve um parâmetro θ desconhecido, definido em um espaço paramétrico Θ. Assim, as hipóteses podem ser definidas como

$$\begin{aligned} H_0 &: \theta \in \Theta_0 \\ H_1 &: \theta \in \Theta_1 \end{aligned}$$

sendo que Θ_0 e Θ_1 são subconjuntos disjuntos de Θ. Um teste é especificado particiondo-se o espaço amostral em dois subconjuntos. Um sobconjunto contém os valores de $\boldsymbol{X}$ para os quais H_0 será rejeitada e é chamado região crítica do teste, e o outro contem os valores de $\boldsymbol{X}$ para os quais H_0 será aceita e é chamado região de aceitação do teste. Em resumo, um teste fica determinado quando especificamos sua região crítica.

Além disso, uma hipótese pode ser classificada da seguinte maneira. Se o subconjunto Θ_i, $i = 0$ ou $i = 1$ contém um único valor então H_i é uma hipótese simples. Caso contrário, se Θ_i contém mais de um valor então H_i é uma hipótese composta. Nos Exemplos 6.1 e 6.2 H_0 é uma hipótese simples enquanto H_1 é composta. Ou seja, se C e $\overline{C}$ denotam a região de rejeição e aceitação respectivamente então

$$P(\boldsymbol{X} \in C | \theta \in \Theta_0) = \alpha \quad \text{e} \quad P(\boldsymbol{X} \in \overline{C} | \theta \in \Theta_1) = \beta$$

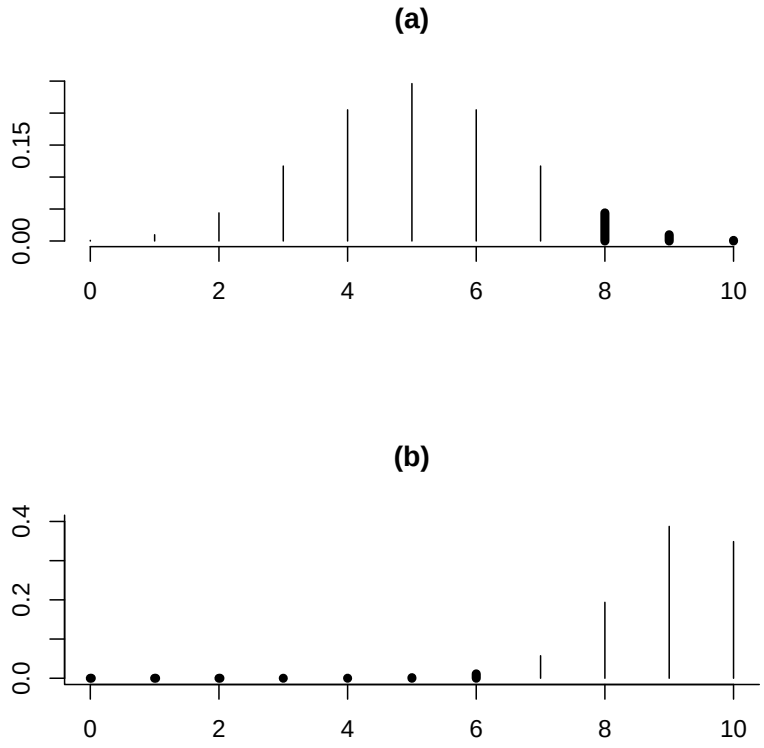

Figura 6.1: Probabilidades binomiais e regiões criticas para os Exemplos 6.1 e 6.2.

6.1.1 Tipos de Decisão

Ao tomar uma decisão a favor ou contra uma hipótese existem dois tipos de erros que podemos cometer. Podemos rejeitar a hipótese nula quando de fato ela é verdadeira (erro tipo I) ou podemos falhar em rejeitar H_0 quando de fato ela é falsa (erro tipo II). Frequentemente denotamos as probabilidades destes dois tipos de erro como α e β respectivamente.

Existe um balanço entre esses dois tipos de erros, no sentido de que ao tentar-se minimizar α, aumenta-se β. Isto é, não é possível minimizar estas duas probabilidades simultaneamente e na prática é costume fixar um valor (pequeno) para α. Na Tabela 6.1 estão descritos as decisões que podemos tomar e os tipos de erro associados.

Tabela 6.1: Tipos de decisão e tipos de erro associados a testes de hipóteses.

	Decisão	
Verdade	Aceitar H_0	Rejeitar H_0
H_0 verdadeira	Decisão correta (probabilidade $1-\alpha$)	Erro Tipo I (probabilidade α)
H_0 falsa	Erro Tipo II (probabilidade β)	Decisão correta (probabilidade $1-\beta$)

6.1.2 A Função Poder

As características probabilísticas de um teste podem ser descritas através de uma função que associa a cada valor de θ a probabilidade $\pi(\theta)$ de rejeitar H_0. A função $\pi(\theta)$ é chamada função de poder (ou potência) do teste. Assim, denotando por C a região crítica a função de poder é definida como

$$\pi(\theta) = P(\boldsymbol{X} \in C \mid \theta), \quad \forall\ \theta \in \Theta.$$

A função de poder é a ferramenta utilizada para verificar a adequação de um teste ou para comparar dois ou mais testes. É claro que uma função de poder ideal seria tal que $\pi(\theta) = 0$ para θ satisfazendo H_0 e $\pi(\theta) = 1$ para θ satisfazendo H_1. Em um problema prático no entanto raramente existirá um teste com estas características. Na Figura 6.2 abaixo está representada a função poder para o Exemplo 6.2, i.e. $P(X \leq 6 \mid p)$, para $0 < p < 1$ com $X \sim Binomial(10, p)$. Note que neste exemplo se p for maior do que digamos 0,8 então o teste quase certamente aceitará H_0, indicando que o teste é adequado. Por outro lado, para valores de p entre 0,7 e 0,8 o teste ainda rejeita H_0 com probabilidade baixa.

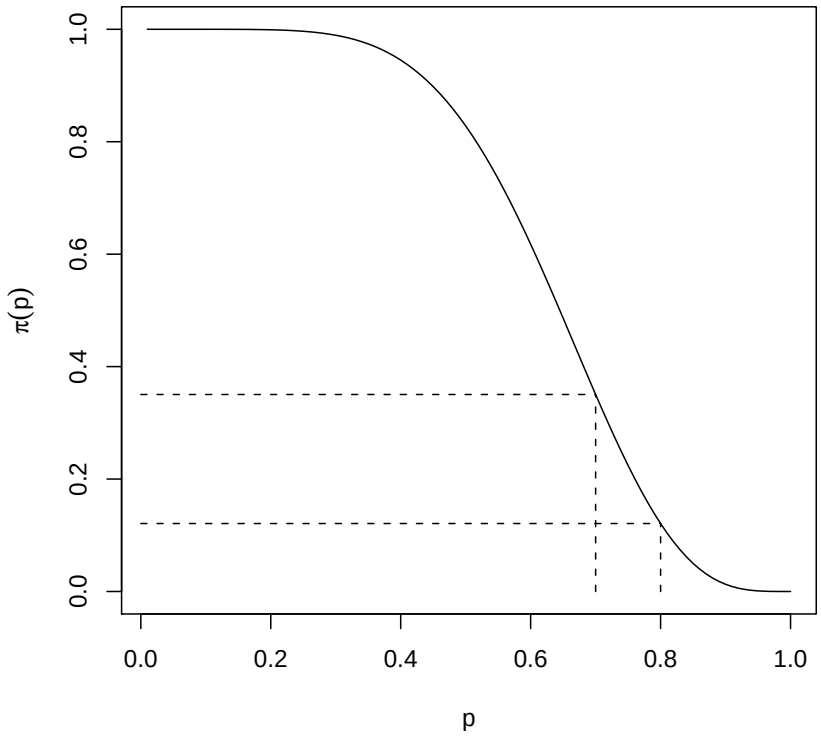

Figura 6.2: Gráfico da função de poder para o Exemplo 6.2.

O *tamanho* ou nível de significância α de um teste é definido como

$$\alpha \geq \sup_{\theta \in \Theta_0} \pi(\theta).$$

Assim como no caso de níveis de confiança na Seção 5.1, a desigualdade acima é essencialmente técnica já que estaremos interessados em valores de α tão pequenos quanto possível. Na prática isto implicará em usar uma igualdade e o tamanho do teste então será a probabilidade máxima, para $\theta \in \Theta_0$, de tomar uma decisão errada. A desigualdade será útil principalmente no caso de espaços amostrais discretos.

Exemplo 6.3 : Seja $X_1, \ldots, X_n$ uma amostra aleatória da distribuição $N(\theta, \sigma^2)$ com $\sigma^2 = 25$ e suponha que queremos testar $H_0 : \theta \leq 17$. Suponha que a regra de decisão consiste em rejeitar H_0 se somente se $\overline{X} > 17 + \sigma/\sqrt{n}$. Neste caso a função poder é dada por

$$\pi(\theta) = P(\text{rejeitar } H_0 \mid \theta) = P(\overline{X} > 17 + \sigma/\sqrt{n}) = P\left(Z > \frac{17 + \sigma/\sqrt{n} - \theta}{\sigma/\sqrt{n}}\right)$$

onde $Z \sim N(0,1)$. Para $n = 25$ segue que,

$$\pi(\theta) = P(Z > 18 - \theta)$$

e calculando esta probabilidade para vários valores de θ podemos construir o gráfico da Figura 6.3 para a função poder do teste. Note que o valor máximo da função quando H_0 é verdadeira ($\theta \leq 17$) é obtido para $\theta = 17$ e portanto o tamanho do teste é dado por

$$\sup_{\theta \leq 17} \left[P\left(Z > \frac{17 + \sigma/\sqrt{n} - \theta}{\sigma/\sqrt{n}} \right) \right] = \pi(17) = P(Z > 1) \approx 0,159.$$

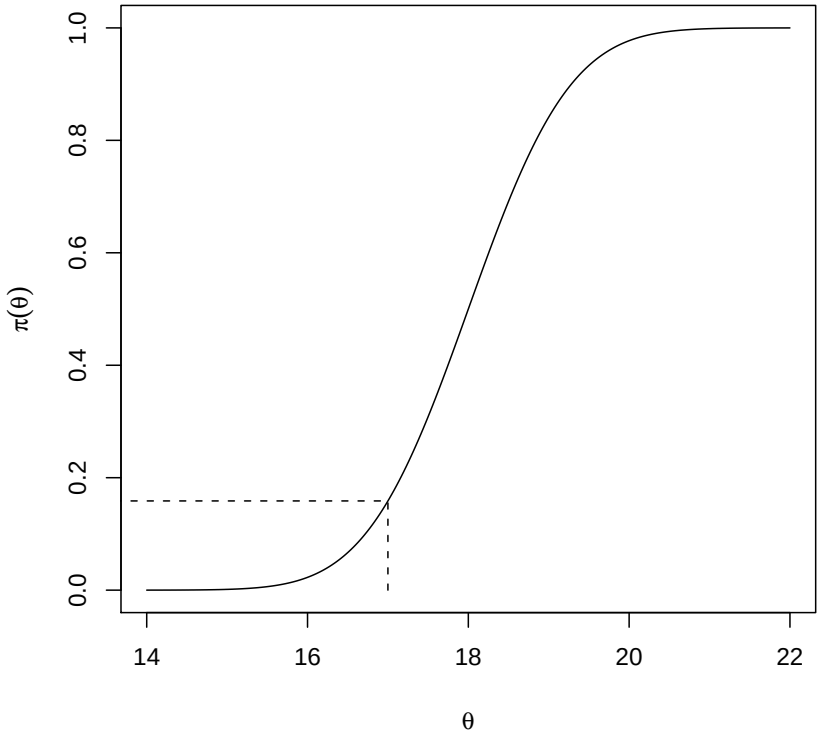

Figura 6.3: Gráfico da função de poder para o Exemplo 6.3.

Comentário

Fica claro que os testes de hipóteses clássicos dependem basicamente da probabilidade de $\boldsymbol{X}$ pertencer a uma determinada região do espaço amostral. Isto significa que os testes dependem da probabilidade de dados que "poderiam" ter

sido observados mas na realidade não foram. Portanto, estes testes violam o princípio da verossimilhança.

6.1.3 Problemas

1. Suponha que $X_1, \ldots, X_n$ é uma amostra aleatória da distribuição $U(0, \theta)$, $\theta > 0$ e queremos testar as hipóteses $H_0 : \theta \geq 2 \times H_1 : \theta < 2$. Seja $Y_n = \max(X_1, \ldots, X_n)$ e um teste que rejeita H_0 se $Y_n \leq 1$.

   (a) Determine a função poder do teste.

   (b) Determine o tamanho do teste.

2. Um aluno faz um teste de múltipla escolha com 10 questões, cada uma com 5 alternativas (somente uma alternativa correta). O aluno acerta 4 questões. É possível deduzir (estatisticamente) que este aluno sabe alguma coisa da matéria?

3. Suponha que a proporção p de itens defeituosos em uma população de itens é desconhecida e queremos testar as hipóteses $H_0 : p = 0,2 \times H_1 : p \neq 0,2$. Uma amostra aleatória de 20 itens é tomada desta população e a regra de decisão consiste em rejeitar H_0 se o número amostral de defeituosos for menor ou igual a 1 ou maior ou igual a 7.

   (a) Faça um esboço do gráfico da função poder para $p = 0; 0,1; 0,2, \ldots, 1$

   (b) Determine o tamanho do teste.

6.2 Testando Hipóteses Simples

É mais útil começar o estuda da teoria de testes de hipóteses considerando apenas hipóteses simples. Isto equivale a dizer que uma amostra aleatória $X_1, \ldots, X_n$ foi tomada de uma dentre duas possíveis distribuições e queremos decidir de qual delas vem a amostra. Neste caso o espaço paramétrico Θ contém apenas dois pontos, digamos θ_0 e θ_1 e queremos testar

$$H_0 : \theta = \theta_0$$
$$H_1 : \theta = \theta_1.$$

Neste caso, as probabilidades dos dois tipo de erro são dadas por

$$\alpha = P(\text{rejeitar } H_0 \mid \theta = \theta_0)$$

$$\beta = P(\text{aceitar } H_0 \mid \theta = \theta_1)$$

e gostariamos de poder construir um teste para o qual estas probabilidades fossem as menores possíveis. Na prática é impossível encontrar um teste que minimize α e β simultaneamente mas pode-se construir testes que minimizam combinações lineares destas probabilidades. Assim, para constantes positivas a e b queremos encontrar um teste δ para o qual $a\alpha(\delta) + b\beta(\delta)$ seja mínima.

Teorema 6.1 (Teste Ótimo) *Seja $X_1, \ldots, X_n$ uma amostra aleatória de uma distribuição com função de (densidade) de probabilidade $p(x|\theta)$ e defina $p_i = p(x|\theta_i)$. Se um teste δ^* rejeita H_0 quando $p_0/p_1 < k$, aceita H_0 quando $p_0/p_1 > k$ e nada decide se $p_0/p_1 = k$, então qualquer outro teste δ é tal que*

$$a\alpha(\delta^*) + b\beta(\delta^*) \leq a\alpha(\delta) + b\beta(\delta).$$

A razão p_0/p_1 é chamada *razão de verossimilhanças* (RV). O teorema estabelece então que um teste ótimo, no sentido de minimizar $a\alpha(\delta) + b\beta(\delta)$, rejeita H_0 quando a razão de verossimilhanças é pequena e aceita H_0 quando esta razão é grande.

Outro resultado vem do fato de que a hipótese H_0 e o erro tipo I são em geral privilegiados em problemas práticos. Assim, é usual considerar testes tais que $\alpha(\delta)$ não seja maior do que um nível especificado, digamos α_0, e tentar minimizar $\beta(\alpha)$.

Lema 6.1 (Neyman-Pearson) *Seja $X_1, \ldots, X_n$ uma amostra aleatória de uma distribuição com função de (densidade) de probabilidade $p(x|\theta)$ e defina $p_i = p(x|\theta_i)$. Se um teste δ^* rejeita H_0 quando $p_0/p_1 < k$, aceita H_0 quando $p_0/p_1 > k$ e nada decide se $p_0/p_1 = k$, então para qualquer outro teste δ tal que $\alpha(\delta) \leq \alpha(\delta^*)$, $\beta(\delta) \geq \beta(\delta^*)$. E também, $\alpha(\delta) < \alpha(\delta^*)$ implica em $\beta(\delta) > \beta(\delta^*)$.*

Exemplo 6.4 : Seja $X_1, \ldots, X_n$ uma amostra aleatória da distribuição $N(\theta, 1)$ e queremos testar $H_0 : \theta = 0 \times H_1 : \theta = 1$. Neste caso a razão de verossimilhanças é dada por

$$\begin{aligned}
\frac{p_0}{p_1} &= \frac{(2\pi)^{-n/2}\exp(-(1/2)\sum_{i=1}^n x_i^2)}{(2\pi)^{-n/2}\exp(-(1/2)\sum_{i=1}^n (x_i-1)^2)} \\
&= \exp\left\{-\frac{1}{2}\left[\sum_{i=1}^n x_i^2 - \sum_{i=1}^n (x_i-1)^2\right]\right\} \\
&= \exp\left[-n\left(\overline{x} - \frac{1}{2}\right)\right].
\end{aligned}$$

Portanto rejeitar H_0 quando $p_0/p_1 < k$ é equivalente a rejeitar H_0 quando

$$\overline{x} > (1/2) - (1/n)\log k = c.$$

Não é difícil obter o valor da constante c tal que

$$P(\overline{X} > c \mid \theta = 0) = P(Z > c\sqrt{n}) = \alpha \quad \text{com} \quad Z \sim N(0,1).$$

Por exemplo para $\alpha = 0,05$ obtemos da tabela da normal padronizada que $c\sqrt{n} = 1,645$ e o teste ótimo (que minimiza β) consiste em rejeitar H_0 se $\overline{X} > 1,645/\sqrt{n}$.

Exemplo 6.5 : Seja $X_1, \ldots, X_n$ uma amostra aleatória da distribuição exponencial com parâmetro θ e queremos testar $H_0 : \theta = \theta_0 \ \times \ H_1 : \theta = \theta_1$, com $\theta_1 > \theta_0$. A razão de verossimilhanças é dada por

$$\frac{p_0}{p_1} = \left(\frac{\theta_0}{\theta_1}\right)^n \exp\left[-(\theta_0 - \theta_1)\sum_{i=1}^n x_i\right]$$

então, pelo lema de Neyman-Pearson, o teste mais poderoso (teste ótimo) rejeita H_0 se $p_0/p_1 < k$ ou equivalentemente se

$$\sum_{i=1}^n x_i < -\frac{1}{\theta_0 - \theta_1}\log\left[k\left(\frac{\theta_1}{\theta_0}\right)^n\right] = c$$

A constante c é obtida fixando-se o valor de α, ou seja calcule c tal que

$$\alpha = P\left(\sum_{i=1}^n X_i < c \mid \theta = \theta_0\right).$$

Note que se $X_i \sim Exp(\theta)$ então quando $\theta = \theta_0$ temos que $\sum_{i=1}^n X_i \sim Gama(n, \theta_0)$ e portanto $2\theta_0 \sum_{i=1}^n X_i$ tem distribuição χ^2_{2n}.

Exemplo 6.6 : Seja $X_1, \ldots, X_{10} \sim Exp(\theta)$ uma amostra aleatória de tempos (em horas) até a falha de equipamento eletrônicos. Suponha que queremos testar $H_0 : \theta = 1 \ \times \ H_1 : \theta = 2$ ao nível de 5%. Do exemplo anterior, devemos obter o valor de uma constante c tal que

$$P\left(2\sum_{i=1}^n X_i < 2c\right) = 0,05$$

sendo que $2\sum_{i=1}^n X_i \sim \chi^2_{20}$. Usando uma tabela da distribuição qui-quadrado com 20 graus de liberdade obtemos que $2c = 10.85$. Assim, a regra de decisão consiste em rejeitar H_0 se $\sum_{i=1}^n X_i < 5.425$, ou equivalentemente se $\overline{X} < 0.5425$.

6.2.1 Problemas

1. Sejam as hipóteses $H_0 : \theta = 1/2$ e $H_1 : \theta = 2/3$ sendo θ a probabilidade de sucesso em um experimento de Bernoulli. O experimento é repetido 2 vezes e aceita-se H_0 se forem obtidos 2 sucessos. Calcule as probabilidades de erro tipo I e II.

2. Sabe-se que uma caixa contém 3 bolas vermelhas e 5 pretas ou 5 vermelhas e 3 pretas. Um experimento consiste em retirar 3 bolas da caixa. Se menos do que 3 bolas retiradas forem vermelhas a decisão será que a caixa contém 3 bolas vermelhas e 5 pretas. Calcule as probabilidades de erro (tipo I e tipo II).

3. Com base em uma amostra de tamanho n da variável aleatória X sendo

$$f(x|\theta) = (\theta + 1)x^{\theta} I_{[0,1]}(x),\ \theta > 0,$$

deseja-se testar as hipóteses $H_0 : \theta = \theta_0$ contra $H_1 : \theta = \theta_1$ com $\theta_0 > \theta_1$. Construa um teste ótimo (use o Lema de Neyman-Pearson).

4. Deseja-se testar $H_0 : \theta = \theta_0$ contra $H_1 : \theta = \theta_1$ $(\theta_1 > \theta_0)$ com base em uma amostra de tamanho n da variável aleatória X sendo

$$f(x|\theta) = \theta \exp(-\theta x) I_{[0,\infty)}(x),\ \theta > 0.$$

Construa um teste ótimo usando o Lema de Neyman-Pearson.

5. Uma v.a. X é tal que $f(x|\theta) = (1-\theta)\theta^{x-1}$, para $x \in \{1, 2, \dots\}$ e $\theta \in (0, 1)$. Encontre uma região crítica para testar $H_0 : \theta = 3/4$ contra $H_1 : \theta = 2/3$ com base em um único valor de X e que satisfaça $\alpha \leq 0,5$.

6. Dispõe-se de uma amostra aleatória de tamanho 50 da v.a. $X \sim N(\mu, 25)$. Sabendo que a média amostral foi $\bar{x} = 28$ teste $H_0 : \mu = 30$ contra $H_1 : \mu = 29$ com $\alpha = 0,05$.

6.3 Probabilidade de significância (P-valor)

Vimos que a escolha do nível de significância do teste é completamente arbitrária. Além disso, quando a distribuição da estatística de teste é discreta, como no Exemplo 6.2 da binomial, o nível escolhido pode nem mesmo ser atingido. Por outro lado, a decisão de aceitar ou rejeitar H_0 claramente depende desta escolha. Na maioria das aplicações práticas o valor escolhido é 0,05 ou 0,01 mas não há nada que justifique formalmente o uso destes valores em particular.

Um enfoque alternativo consiste em calcular uma quantidade chamada *nível crítico*, *probabilidade de significância* ou *p-valor*. Em geral, se T é uma estatística de teste e H_0 é rejeitada por exemplo para $T > c$ então o p-valor é a probabilidade $P(T > t \mid H_0)$ onde t é o valor observado de T.

Exemplo 6.7 : No Exemplo 6.1 suponha que o número observado de questões certas foi $X = 9$. Então o p-valor será

$$P(X \geq 9 \mid p = 1/2) = \binom{10}{9} 0,5^{10} + \binom{10}{10} 0,5^{10} = 0,0107$$

e rejeitaremos H_0 para todo nível de significância maior do que este valor. Por exemplo, rejeitaremos H_0 para os valores usuais $\alpha = 0,025$ ou $\alpha = 0,05$. Por outro lado, H_0 seria aceita para $\alpha = 0,01$.

Exemplo 6.8 : No Exemplo 6.2 suponha que o número observado de não defeituosos foi $X = 4$. Neste caso o p-valor é dado por

$$P(X \leq 4 \mid p = 0,90) = 0,000146$$

ou seja, rejeitaremos H_0 para praticamente todos os níveis de significância usuais.

Portanto, o p-valor é a probabilidade de observar resultados tão extremos quanto aqueles que foram obtidos se a hipótese nula for verdadeira. A idéia é que se o p-valor for grande ele fornece evidência de que H_0 é verdadeira, enquanto que um p-valor pequeno indica que existe evidência nos dados contra H_0. As seguintes interpretações de p-valores (P) podem ser úteis,

$P \geq 0,10$	Não existe evidência contra H_0
$0,05 \leq P < 0,10$	Fraca evidência contra H_0
$0,01 \leq P < 0,05$	Evidência significativa ...
$0,001 \leq P < 0,01$	Evidência altamente significativa ...
$P < 0,001$	Evidência extremamente significativa ...

Comentários

Da forma como a metodologia clássica de testes de hipóteses foi desenvolvida podemos ter a impressão de que estamos calculando probabilidades a respeito de uma hipótese. De fato, algumas vezes é incorretamente afirmado que rejeitar H_0 ao nível α indica que a probabilidade de H_0 ser verdadeira é menor do que α.

Esta interpretação não é válida e o p-valor calculado em um teste não fornece nenhuma indicação sobre qualquer probabilidade a respeito de H_0.

Por exemplo, um p-valor próximo de zero nos fornece (do ponto de vista clássico) muita evidência contra H_0 porém isto não significa de maneira alguma que $P(H_0$ ser verdadeira) seja também próxima de zero. Esta última afirmação probabilística sequer faz sentido na inferência clássica, embora seja exatamente isto que gostariamos de calcular.

Para que esta interpretação fosse válida teriamos que usar a abordagem Bayesiana. Basicamente, teríamos que atribuir uma probabilidade *a priori*, i.e. antes de observar os dados, para a hipótese H_0. Após a observação dos dados amostrais esta probabilidade seria atualizada, segundo regras da inferência Bayesiana, e teríamos uma probabilidade *a posteriori* para a hipótese H_0. Para maiores detalhes ver por exemplo Migon and Gamerman (1999) ou DeGroot (1989).

6.4 Testes Uniformemente mais Poderosos

Na Seção 6.2 foram definidos testes ótimos para testar hipóteses simples. Nesta seção os resultados serão generalizados para hipóteses compostas. Considere então um teste em que H_0 pode ser uma hipótese simples ou composta e H_1 é sempre uma hipótese composta.

Definição 6.1 *Um teste δ de $H_0 : \theta \in \Theta_0 \ \times \ H_1 : \theta \in \Theta_1$ é dito ser uniformemente mais poderoso (UMP) de tamanho α se e somente se*

$$\sup_{\theta \in \Theta_0} \pi(\theta) = \alpha$$

e para qualquer outro teste δ^* que satisfaça esta igualdade

$$\pi(\theta|\delta) \geq \pi(\theta|\delta^*), \quad \forall\ \theta \in \Theta_1.$$

Assim, de acordo com esta definição, precisamos especificar um teste cuja probabilidade máxima de rejeitar H_0 quando ela é verdadeira seja α e que ao mesmo tempo maximize a probabilidade de rejeitar H_0 quando ela é falsa. Veremos a seguir que os testes UMP só existem em situações especiais, por exemplo quando a distribuição pertence à família exponencial vista na Seção 1.3.1.

Teorema 6.2 *Se $X_1, \ldots, X_n$ é uma amostra aleatória de um membro da família exponencial e ϕ for estritamente crescente em θ então o teste UMP de nível α para testar $H_0 : \theta \leq \theta_0 \times H_1 : \theta > \theta_0$ rejeita H_0 se $T(\boldsymbol{x}) > c$. Se as hipóteses forem invertidas ou ϕ for estritamente decrescente em θ então o teste UMP rejeita H_0 se $T(\boldsymbol{x}) < c$. Se ambas as condições ocorrerem o teste fica inalterado.*

Um fato importante é que, em qualquer condição estes testes têm função poder crescente em θ e portanto seu valor máximo sob H_0 é atingido em $\theta = \theta_0$. Assim a constante c acima é obtida de modo que $P(\text{rejeitar } H_0 \mid \theta = \theta_0) \leq \alpha$, com igualdade no caso contínuo.

Exemplo 6.9 : Seja $X_1, \ldots, X_n$ uma amostra aleatória da distribuição de Bernoulli com parâmetro θ. Suponha que queremos testar $H_0 : \theta \leq 0,1 \times H_1 : \theta > 0,1$ ao nível máximo de 5% com base em uma amostra de tamanho $n = 15$. Então, definindo $t(\boldsymbol{x}) = \sum_{i=1}^n x_i$

$$\begin{aligned} p(\boldsymbol{x}|\theta) &= \theta^{t(\boldsymbol{x})}(1-\theta)^{n-t(\boldsymbol{x})} = \exp[t(\boldsymbol{x})\log\theta + (n - t(\boldsymbol{x}))\log(1-\theta)] \\ &= \exp\left\{t(\boldsymbol{x})\log\left(\frac{\theta}{1-\theta}\right) + n\log(1-\theta)\right\}. \end{aligned}$$

Logo, a distribuição pertence à família exponencial e $\phi(\theta) = \log(\theta/(1-\theta))$ é uma função estritamente crescente de θ. Assim, um teste UMP deve rejeitar H_0 se $\sum_{i=1}^n X_i > c$ onde c é tal que $P(\sum_{i=1}^n X_i > c \mid \theta = 0,1) \leq \alpha$. Como $Y = \sum_{i=1}^n X_i \sim Binomial(n, \theta)$ segue que

$$\begin{aligned} P(Y > 3 \mid \theta = 0,1) &= 0,056 \\ P(Y > 4 \mid \theta = 0,1) &= 0,013 \\ P(Y > 5 \mid \theta = 0,1) &= 0,002 \\ P(Y > 6 \mid \theta = 0,1) &= 0,0003. \end{aligned}$$

e a regra de decisão consiste em rejeitar H_0 se $\sum_{i=1}^n X_i > 4$.

Exemplo 6.10 : Seja $X_1, \ldots, X_n$ uma amostra aleatória da distribuição exponencial com parâmetro θ e queremos testar $H_0 : \theta \leq \theta_0 \times H_1 : \theta > \theta_0$. Definindo $t(\boldsymbol{x}) = \sum_{i=1}^n x_i$ a função de densidade conjunta é

$$p(\boldsymbol{x}|\theta) = \theta^n e^{-\theta\, t(\boldsymbol{x})} = \exp(n\log\theta - \theta\, t(\boldsymbol{x})).$$

Portanto a distribuição pertence à família exponencial e $\phi(\theta) = -\theta$ é uma função estritamente decrescente de θ. Então pelo Teorema 6.2 o teste UMP deve rejeitar H_0 se $\sum_{i=1}^n X_i < c$. Fixando o valor de α a constante c é a solução da equação $P(\sum_{i=1}^n X_i < c \mid \theta = \theta_0) = \alpha$ com $\sum_{i=1}^n X_i \sim Gama(n, \theta_0)$ e portanto $2\theta_0 \sum_{i=1}^n X_i \sim \chi^2_{2n}$.

A propriedade que garante a existência de testes UMP na família exponencial pode ser estendida a famílias de distribuições com razão de verossimilhança monótona.

Definição 6.2 *A família de distribuições com função de (densidade) de probabilidade $p(\boldsymbol{x}|\theta)$ é dita ter razão de verossimilhança monótona se existe uma estatística $T(\boldsymbol{X})$ tal que $\forall\, \theta_1, \theta_2 \in \Theta$, com $\theta_1 < \theta_2$, a razão $p(\boldsymbol{x}|\theta_2)/p(\boldsymbol{x}|\theta_1)$ é uma função monótona em $t(\boldsymbol{x})$.*

Intuitivamente, quanto maior for a razão de verossimilhança mais plausível é o valor θ_2 em relação a θ_1. Assim, se queremos testar $H_0 : \theta \leq \theta_0 \times H_1 : \theta > \theta_0$ e se a RV for uma função crescente de $T(\boldsymbol{X})$ então é razoável rejeitar H_0 para valores grandes de $T(\boldsymbol{X})$. Pode-se mostrar que neste caso o teste UMP rejeita H_0 se $T(\boldsymbol{X}) > c$. Analogamente, se as hipóteses forem invertidas ou se a RV for uma função decrescente de $T(\boldsymbol{X})$ então o teste UMP rejeita H_0 se $T(\boldsymbol{X}) < c$. Se ambas as condições ocorrerem o teste fica inalterado.

Em qualquer destas condições o fato importante é que a função poder é sempre crescente em θ. Portanto, a constante c acima é obtida de modo que $P(\text{rejeitar } H_0 \mid \theta = \theta_0) \leq \alpha$, com igualdade no caso contínuo.

Exemplo 6.11 : Seja $X_1, \ldots, X_n$ uma amostra aleatória da distribuição de Bernoulli com parâmetro θ e queremos testar $H_0 : \theta \leq \theta_0 \times H_1 : \theta > \theta_0$. Então, definindo $t(\boldsymbol{x}) = \sum_{i=1}^{n} x_i$ temos que

$$p(\boldsymbol{x}|\theta) = \theta^{t(\boldsymbol{x})}(1-\theta)^{n-t(\boldsymbol{x})}$$

e para $\theta_1 < \theta_2$ a razão de verossimilhança fica

$$\frac{\theta_2^{t(\boldsymbol{x})}(1-\theta_2)^{n-t(\boldsymbol{x})}}{\theta_1^{t(\boldsymbol{x})}(1-\theta_1)^{n-t(\boldsymbol{x})}} = \left[\frac{\theta_2(1-\theta_1)}{\theta_1(1-\theta_2)}\right]^t \left(\frac{1-\theta_2}{1-\theta_1}\right)^n = \alpha^t \beta^n.$$

Como $\theta_2 > \theta_1$ e $1-\theta_1 > 1-\theta_2$ então $\alpha > 1$ e a RV é uma função crescente em t. Portanto, o teste UMP rejeita H_0 se $\sum_{i=1}^{n} X_i > c$ confirmando assim o resultado no Exemplo 6.9.

6.4.1 Problemas

1. Para cada uma das distribuições abaixo considere uma amostra aleatória $X_1, \ldots, X_n$ e obtenha o teste UMP para testar as hipóteses $H_0 : \theta \leq \theta_0 \times H_0 : \theta > \theta_0$.

   (a) Poisson com parâmetro θ.

   (b) Normal com média conhecida e variância desconhecida.

   (c) Gama com parâmetro α desconhecido e β conhecido.

   (d) Gama com parâmetro α conhecido e β desconhecido.

2. Seja $X_1, \ldots, X_n$ uma amostra aleatória da distribuição $N(0, \sigma^2)$ com σ^2 desconhecido. Obtenha o teste UMP para testar as hipóteses $H_0 : \sigma^2 \leq 2 \times H_0 : \sigma^2 > 2$ com $n = 10$ e $\alpha = 0,05$.

3. Suponha que $X_1, \ldots, X_n$ seja uma amostra aleatória da distribuição exponencial com parâmetro θ e queremos testar $H_0 : \theta \geq 1/2 \times H_0 : \theta < 1/2$. Obtenha o teste UMP para estas hipóteses com $n = 10$ e $\alpha = 0,05$.

4. Suponha que $X_1, \ldots, X_n$ seja uma amostra aleatória da distribuição de Poisson com parâmetro θ e queremos testar $H_0 : \theta \leq 1 \times H_0 : \theta > 1$. Obtenha o teste UMP para estas hipóteses com $n = 10$ e $\alpha = 0,05$.

5. Seja $X_1, \ldots, X_n$ uma amostra aleatória da distribuição com função de densidade $p(x|\theta) = \theta x^{\theta-1}$, para $x \in (0,1)$ e $\theta > 0$ desconhecido. Encontre o teste UMP para as hipóteses $H_0 : \theta \leq 1 \times H_1 : \theta > 1$ com nível de significância $\alpha = 0,05$.

6. A proporção p de itens defeituosos em um grande lote de manufaturados é desconhecida. Uma amostra aleatória de 20 itens foi selecionada e inspecionada, e queremos testar as hipóteses $H_0 : p \leq 0,1 \times H_1 : p > 0,1$. Obtenha o teste UMP.

7. Suponha que $X_1, \ldots, X_n$ seja uma amostra aleatória da distribuição de Poisson com média λ desconhecida e queremos testar $H_0 : \lambda \geq 1 \times H_1 : \lambda < 1$. Para $n = 10$, verifique para quais níveis de significância no intervalo $0 < \alpha < 0,03$ existem testes UMP.

8. Seja $X_1, \ldots, X_n$ uma amostra aleatória da distribuição $N(\mu, 1)$ com μ desconhecido e queremos testar as hipóteses $H_0 : \mu \leq 0 \times H_1 : \mu > 0$. Sejam δ^* o teste UMP ao nível $\alpha = 0,025$ e $\pi(\mu|\delta^*)$ função poder do teste.

   (a) Determine o menor valor de n para o qual $\pi(\mu|\delta^*) \geq 0,9$ para $\mu \geq 0,5$.

   (b) Determine o menor valor de n para o qual $\pi(\mu|\delta^*) \leq 0,001$ para $\mu \leq -0,1$.

9. Seja $X_1, \ldots, X_n$ uma amostra aleatória da distribuição χ^2 com número de graus de liberdade θ desconhecido, $\theta = 1, 2, \ldots$. Suponha que queremos testar as hipóteses $H_0 : \theta \leq 8 \times H_1 : \theta \geq 9$ ao nível de significância α. Mostre que existe um teste UMP que rejeita H_0 se $\sum_{i=1}^{n} \log X_i > k$ para uma constante k.

6.5 Testes Bilaterais

Suponha agora que queremos testar hipóteses do tipo

$$\begin{aligned} H_0 &: \theta = \theta_0 \\ H_1 &: \theta \neq \theta_0, \end{aligned} \tag{6.1}$$

ou seja H_0 é uma hipótese simples e H_1 é uma alternativa bilateral. Como veremos nas próximas seções este tipo de teste pode ser útil na comparação de tratamentos. O problema é que neste caso não existe um teste UMP para estas hipóteses, i.e. não é possível construir um teste cuja probabilidade de rejeitar H_0 seja maximizada quando ela é falsa.

Um procedimento alternativo seria construir testes tais que as chances de rejeitar H_0 sejam maiores quando ela é falsa do que quando ela é verdadeira. Isto nos leva à definição de testes não viesados a seguir.

Definição 6.3 *Um teste δ é dito ser não viesado para as hipóteses $H_0 : \theta \in \Theta_0$ $\times$ $H_1 : \theta \in \Theta_1$ se $\forall\ \theta \in \Theta_0$ e $\theta' \in \Theta_1$ então $\pi(\theta) \leq \pi(\theta')$. Caso contrário o teste é dito viesado.*

Ou seja, em testes não viesados a probabilidade de rejeitar H_0 quando ela é falsa é no mínimo tão grande quanto para H_0 verdadeira.

Podemos agora tentar construir testes para hipóteses bilaterais que sejam UMP dentro da classe de testes não viesados. Se a distribuição pertence à família exponencial, pode-se mostrar que se $\phi(\theta)$ for uma função estritamente crescente em θ então o teste UMP não viesado de nível α para as hipóteses (6.1) aceita H_0 quando $c_1 < T(\boldsymbol{X}) < c_2$. As constantes c_1 e c_2 são obtidas de modo que $P(c_1 < T(\boldsymbol{X}) < c_2 \mid \theta = \theta_0) = 1 - \alpha$.

Note que existe uma infinidade de valores de c_1 e c_2 satisfazendo a esta condição. Em muitas situações é conveniente tomar valores tais que

$$P(T(\boldsymbol{X}) < c_1 \mid \theta = \theta_0) = P(T(\boldsymbol{X}) > c_2 \mid \theta = \theta_0) = \alpha/2$$

e se $T(\boldsymbol{X})$ tem uma distribuição simétrica em torno de um ponto isto implica em escolher c_1 e c_2 simetricamente em relação a este ponto. No entanto, nada impede que outros valores possam ser considerados. Por exemplo, o pesquisador pode considerar mais grave aceitar H_0 quando $\theta < \theta_0$ do que quando $\theta > \theta_0$ e neste caso é melhor considerar testes com função poder assimétrica.

6.5.1 Testes Gerais

Em muitas situações não é possível obter nem mesmo um teste não viesado. Um procedimento geral para testar $H_0 : \theta \in \Theta_0 \times H_1 : \theta \in \Theta_1$ é baseado na estatística da razão de máxima verossimilhança (RMV) dada por

$$\lambda(\boldsymbol{X}) = \frac{\sup_{\theta\in\Theta_0} p(\boldsymbol{X}|\theta)}{\sup_{\theta\in\Theta_1} p(\boldsymbol{X}|\theta)}.$$

Deste modo estaremos comparando o valor máximo atingido pela função de verossimilhança quando $\theta \in \Theta_0$ com o valor máximo atingido quando $\theta \in \Theta_1$. Neste caso, é razoável decidir pela rejeição de H_0 se $\lambda(\boldsymbol{X}) < c$ onde a constante c é obtida de modo que

$$\sup_{\theta\in\Theta_0} P(\lambda(\boldsymbol{X}) < c \mid \theta) \leq \alpha.$$

Novamente, a igualdade será usada sempre que possível ficando a desigualdade para o caso de distribuições discretas.

Equivalentemente, podemos usar o logaritmo da verossimilhança

$$-2\log\lambda = 2(\ell_1^* - \ell_0^*)$$

e neste caso, a região de rejeição será $\{\boldsymbol{X} : -2\log\lambda(\boldsymbol{X}) > k\}$.

Existem duas dificuldades práticas associadas a estes testes:

- obter os valores $\hat{\theta}_0$ e $\hat{\theta}_1$ que maximizam a verossimilhança sob H_0 e H_1.
- determinar a distribuição amostral de $\lambda(\boldsymbol{X})$ (ou $-2\log\lambda(\boldsymbol{X})$).

Este segundo problema será discutido em mais detalhes quando falarmos de testes assintóticos na Seção 6.7.

6.6 Testes de Hipóteses no Modelo Normal

Os resultados desenvolvidos nas seções anteriores serão aplicados ao modelo normal para testes sobre média e variância em problemas de uma ou mais amostras e em modelos de regressão linear. Nesta seção considere uma amostra aleatória $X_1, \cdots, X_n$ tomada da distribuição $N(\theta, \sigma^2)$.

Suponha que queremos testar $H_0 : \theta = \theta_0 \times H_1 : \theta \neq \theta_0$ e inicialmente vamos

assumir que σ^2 é conhecida. Neste caso,

$$\begin{aligned} p(\boldsymbol{x}|\theta) &= (2\pi\sigma^2)^{-n/2}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\theta)^2\right) \\ &= (2\pi\sigma^2)^{-n/2}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}x_i^2\right)\exp\left(\frac{\overline{x}n\theta}{\sigma^2}-\frac{n\theta^2}{2\sigma^2}\right) \end{aligned}$$

e como $n\theta$ é uma função estritamente crescente de θ segue que o teste UMP não viesado rejeita H_0 se $\overline{X} < c_1$ ou $\overline{X} > c_2$. Ao nível de significância α podemos obter as constantes c_1 e c_2 tais que

$$P(\overline{X} < c_1 \mid \theta = \theta_0) + P(\overline{X} > c_2 \mid \theta = \theta_0) = \alpha.$$

Conforme discutido anteriormente, existe uma infinidade de valores que satisfazem esta condição. Na maioria dos experimentos envolvendo o modelo normal será conveniente tomar c_1 e c_2 simétricos em relação a $E(\overline{X})$. Assim, usando uma tabela da distribuição normal padronizada podemos obter o valor do percentil $z_{\alpha/2}$ tal que

$$P\left(-z_{\alpha/2} \leq \frac{\sqrt{n}(\overline{\boldsymbol{X}}-\theta_0)}{\sigma} \leq z_{\alpha/2}\right) = 1-\alpha$$

e o teste bilateral UMP não viesado rejeita H_0 se

$$\overline{X} < \theta_0 - z_{\alpha/2}\sigma/\sqrt{n} \quad \text{ou} \quad \overline{X} > \theta_0 + z_{\alpha/2}\sigma/\sqrt{n}.$$

No caso em que a variância populacional é também desconhecida o espaço dos parâmetros é $\Theta = \{(\theta, \sigma^2) : \theta \in \mathbb{R}, \sigma^2 > 0\}$ e vamos obter o teste da RMV. Note que, como H_0 é uma hipótese simples então $\Theta_0 = \{(\theta_0, \sigma^2) : \sigma^2 > 0\}$ e não é difícil verificar que o valor de σ^2 que maximiza a verossimilhança para θ_0 fixo é $\hat{\sigma}_0^2 = \sum_{i=1}^{n}(x_i-\theta_0)^2/n$ (faça as contas). Portanto,

$$\sup_{(\theta,\sigma^2)\in\Theta_0} p(\boldsymbol{X}|\theta,\sigma^2) = p(\boldsymbol{x}|\theta_0,\hat{\sigma}_0^2).$$

Para $\theta \neq \theta_0$ a função de verossimilhança é maximizada em $(\hat{\theta}, \hat{\sigma}^2)$ onde $\hat{\theta} = \overline{x}$ e $\hat{\sigma}^2 = \sum_{i=1}^{n}(x_i-\overline{x})^2/n$. Portanto

$$\sup_{(\theta,\sigma^2)\in\Theta_1} p(\boldsymbol{X}|\theta,\sigma^2) = p(\boldsymbol{x}|\hat{\theta},\hat{\sigma}^2).$$

Assim, a estatística da RMV é dada por

$$\lambda(\boldsymbol{X}) = \frac{(2\pi\hat{\sigma}_0^2)^{-n/2}\exp\{-\sum_{i=1}^{n}(X_i-\theta_0)^2/2\hat{\sigma}_0^2\}}{(2\pi\hat{\sigma}^2)^{-n/2}\exp\{-\sum_{i=1}^{n}(X_i-\overline{X})^2/2\hat{\sigma}^2\}}$$

e substituindo as somas de quadrados obtemos que $\lambda(\boldsymbol{X}) = (\hat{\sigma}_0^2/\hat{\sigma}^2)^{-n/2}$. Mas,

$$\frac{\hat{\sigma}_0^2}{\hat{\sigma}^2} = \frac{\sum_{i=1}^n (X_i - \overline{X})^2 + n(\overline{X} - \theta_0)^2}{\sum_{i=1}^n (X_i - \overline{X})^2} = 1 + \frac{n(\overline{X} - \theta_0)^2}{(n-1)S^2} = 1 + \frac{T^2}{n-1}$$

onde $T = \sqrt{n}(\overline{X} - \theta_0)/S$ e então podemos reescrever a RMV como

$$\lambda(\boldsymbol{X}) = \left(1 + \frac{T^2}{n-1}\right)^{-n/2}.$$

Finalmente, o teste da RMV rejeita H_0 se $\lambda(\boldsymbol{X}) < c^*$ ou equivalentemente se $T^2 > c$ ou $|T| > c$. Como $T \sim t_{n-1}$ a constante c é simplesmente o percentil $t_{\alpha/2,n-1}$ desta distribuição.

O teste desenvolvido acima é conhecido como teste t e talvez um dos mais utilizados em Estatística. Pode-se mostrar que o teste t é não viesado já que o valor mínimo da função poder ocorre em $\theta = \theta_0$. Além disso, as propriedades do teste não são afetadas pelo valor de σ^2 (parâmetro de distúrbio) já que σ^2 foi substituido pelo seu estimador S^2 e T é uma quantidade pivotal. O teste também é invariante a transformações lineares das observações.

Testes bilaterais do tipo $H_0 : \sigma^2 = \sigma_0^2 \times H_1 : \sigma^2 \neq \sigma_0^2$ para a variância podem ser construídos fazendo-se analogia com intervalos de confiança. Vimos na Seção 5.2.1 do Capítulo 5 que o intervalo de confiança de $100(1-\alpha)\%$ para σ^2 é dado por

$$\left(\frac{(n-1)s^2}{q_2}, \frac{(n-1)s^2}{q_1}\right)$$

sendo q_1 e q_2 são os quantis $\alpha/2$ e $1-\alpha/2$ da distribuição χ^2_{n-1}. Assim, o teste deve aceitar H_0 se e somente se σ_0^2 estiver contido neste intervalo. Será deixado como exercício mostrar que este é o teste da razão de máxima verossimilhança para as hipóteses acima.

6.6.1 Testes para Várias Médias

Para começar vamos assumir que temos duas amostras aleatórias $X_{11}, \ldots, X_{1n_1}$ e $X_{21}, \ldots, X_{2n_2}$ das distribuições $N(\theta_1, \sigma_1^2)$ e $N(\theta_2, \sigma_2^2)$ respectivamente e que as amostras são independentes. Neste caso o vetor de parâmetros é $(\theta_1, \theta_2, \sigma_1^2, \sigma_2^2)$ e em geral estaremos interessados em testar as hipóteses

$$\begin{aligned} H_0 &: \theta_1 = \theta_2, \sigma_1^2 > 0, \sigma_2^2 > 0 \\ H_1 &: \theta_1 \neq \theta_2, \sigma_1^2 > 0, \sigma_2^2 > 0 \end{aligned} \tag{6.2}$$

Se pudermos assumir que as variâncias populacionais são iguais, i.e. $\sigma_1^2 = \sigma_2^2 = \sigma^2$, o problema de construção do teste se torna relativamente simples usando a

estatística da razão de máxima verossimilhança. Neste caso, como as amostras são independentes, podemos escrever a função de verossimilhança como

$$p(\boldsymbol{x}_1, \boldsymbol{x}_2|\theta_1, \theta_2, \sigma^2) = p(\boldsymbol{x}_1|\theta_1, \sigma^2)p(\boldsymbol{x}_2|\theta_2, \sigma^2)$$

e após algum algebrismo segue que a verossimilhança de $(\theta_1, \theta_2, \sigma^2)$ é dada por

$$(2\pi\sigma^2)^{-(n_1+n_2)/2} \exp\left\{-\frac{1}{2\sigma^2}\left[(n_1-1)S_1^2 + n_1(\theta_1 - \overline{x}_1)^2 + (n_2-1)S_2^2 + n_2(\theta_2 - \overline{x}_2)^2\right]\right\}.$$

Quando $\theta_1 \neq \theta_2$ as estimativas de máxima verossimilhança de θ_1, θ_2 e σ^2 são respectivamente $\overline{x}_1$, $\overline{x}_2$ e

$$\hat{\sigma}^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}$$

onde S_1^2 e S_2^2 são as variâncias amostrais. Quando $\theta_1 = \theta_2 = \theta$ segue que as estimativas de máxima verossimilhança de θ e σ^2 são

$$\hat{\theta} = \frac{n_1\overline{x}_1 + n_2\overline{x}_2}{n_1+n_2} \quad \text{e} \quad \hat{\sigma}_0^2 = \hat{\sigma}^2 + \frac{n_1 n_2}{(n_1+n_2)^2}(\overline{x}_1 - \overline{x}_2)^2.$$

Substituindo estas expressões na razão de verossimilhanças pode-se mostrar que o teste da RMV rejeita H_0 se

$$|T| = \left|\frac{(\overline{X}_1 - \overline{X}_2)}{\hat{\sigma}\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}\right| > c.$$

Pode-se mostrar que T tem distribuição t de Student com $\nu = n_1 + n_2 - 2$ graus de liberdade de modo que a constante c é simplesmente o percentil $t_{\alpha/2,\nu}$ desta distribuição. Este teste é conhecido como teste t para duas amostras.

6.6.2 Variâncias Desconhecidas e Desiguais

O procedimento visto na seção anterior para variâncias iguais pode ser estendido facilmente para o caso de variâncias desconhecidas e desiguais, desde que a razão de variâncias σ_1^2/σ_2^2 seja conhecida. Suponha por exemplo que $\sigma_1^2 = k\sigma_2^2$ onde k é uma constante positiva conhecida. Definindo-se

$$\hat{\sigma}^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2/k}{n_1+n_2-2}$$

então pode-se mostrar que quando $\theta_1 = \theta_2$ a variável aleatória

$$U = \frac{(\overline{X}_1 - \overline{X}_2)}{\hat{\sigma}\sqrt{\frac{1}{n_1} + \frac{k}{n_2}}}$$

tem distribuição t de Student com $n_1 + n_2 - 2$ graus de liberdade.

Finalmente, se mesmo a razão de variâncias for desconhecida então o problema de testar as hipóteses 6.2 torna-se bastante complexo. Este problema é conhecido na literatura como o *problema de Behrens-Fisher*. Vários procedimentos de teste já foram propostos e a maioria foi objeto de controvérsia em relação a sua utilidade e correção.

6.6.3 Comparação de Variâncias

Em problemas com duas ou mais amostras de distribuições normais é natural que se tenha interesse em comparar as variâncias populacionais. Neste caso, a distribuição F é utilizada para testar as hipóteses associadas. No caso de duas amostras suponha que queremos testar

$$H_0 : \sigma_1^2 \leq \sigma_2^2$$
$$H_1 : \sigma_1^2 > \sigma_2^2$$

Pode-se mostrar que não existe teste UMP para estas hipóteses e é prática comum utilizar-se o chamado teste F. Este teste é não viesado e na verdade é UMP dentro da classe de testes não viesados. Usando a estatística da razão de máxima verossimilhança pode-se mostrar que o teste F rejeita H_0 se

$$\frac{\sum_{i=1}^{n_1}(x_{1i} - \overline{x}_1)^2/(n_1 - 1)}{\sum_{i=1}^{n_2}(x_{2i} - \overline{x}_2)^2/(n_2 - 1)} = \frac{s_1^2}{s_2^2} > c.$$

Vimos na Seção 5.2.4 que

$$\frac{S_1^2}{S_2^2}\frac{\sigma_2^2}{\sigma_1^2} \sim F(n_1 - 1, n_2 - 1).$$

e portanto a constante c pode ser obtida tal que

$$P\left(\frac{S_1^2}{S_2^2}\frac{\sigma_2^2}{\sigma_1^2} > c \mid \sigma_1^2 = \sigma_2^2\right) = P\left(\frac{S_1^2}{S_2^2} > c\right) = \alpha$$

usando os valores tabelados da distribuição F com $n_1 - 1$ e $n_2 - 1$ graus de liberdade.

No caso de testes bilaterais, i.e.

$$H_0 : \sigma_1^2 = \sigma_2^2$$
$$H_1 : \sigma_1^2 \neq \sigma_2^2$$

o teste F rejeita H_0 se $S_1^2/S_2^2 < c_1$ ou $S_1^2/S_2^2 > c_2$ onde as constantes c_1 e c_2 são mais uma vez obtidas como percentis da distribuição F com $n_1 - 1$ e $n_2 - 1$ graus de liberdade. Analogamente ao teste t, é prática comum escolher c_1 e c_2 tal que as probabilidades nas caudas sejam iguais, i.e. $\alpha/2$.

6.6.4 Problemas

1. Suponha que $X_1, \ldots, X_n$ é uma amostra aleatória da distribuição $N(\mu, 1)$ e queremos testar as hipóteses $H_0 : \mu = \mu_0 \times H_1 : \mu \neq \mu_0$. Considere um teste que rejeita H_0 se $\overline{X} \leq c_1$ ou $\overline{X} \geq c_2$.

   (a) Determine os valores de c_1 e c_2 tais que $\pi(\mu_0) = 0,10$ e $\pi(\mu)$ seja simétrica em torno de μ_0.

   (b) Determine os valores de c_1 e c_2 tais que $\pi(\mu_0) = 0,10$ e o teste seja não viesado.

   (c) Suponha que $c_1 = \mu_0 - 1,96/\sqrt{n}$. Determine c_2 tal que $\pi(\mu_0) = 0,10$.

   (d) Determine o menor valor de n para o qual $\pi(\mu_0) = 0,10$ e $\pi(\mu_0 + 1) = \pi(\mu_0 - 1) \geq 0,95$.

2. Suponha que $X_1, \ldots, X_n$ é uma amostra aleatória da distribuição $N(\mu, 1)$ e queremos testar as hipóteses

   $$H_0 : 0,1 \leq \mu \leq 0,2$$
   $$H_1 : \mu < 0,1 \text{ ou } \mu > 0,2.$$

   Considere um teste que rejeita H_0 se $\overline{X} \leq c_1$ ou $\overline{X} \geq c_2$.

   (a) Para $n = 25$ determine c_1 e c_2 tais que tais que $\pi(0,1) = \pi(0,2) = 0,07$.

   (b) Idem para $\pi(0,1) = 0,02$ e $\pi(0,2) = 0,05$.

3. Os comprimentos de fibras metálicas (em milímetros) produzidas por uma máquina têm distribuição normal com média μ e variância σ^2 desconhecidos. Suponha que queremos testar as seguintes hipóteses

   $$H_0 : \mu \leq 5,2$$
   $$H_1 : \mu > 5,2.$$

Os comprimentos de 15 fibras selecionadas ao acaso foram medidos e obteve-se a média amostral $\overline{x} = 5,4$ e $\sum_{i=1}^{n}(x_i - \overline{x})^2 = 2,5$.

(a) Construa um teste t ao nível de 0,05 baseado nestes resultados.

(b) Repita o item anterior para as hipóteses $H_0 : \mu = 5,2 \times H_1 : \mu \neq 5,2$. Qual a conclusão do exercício?

4. Suponha que foi selecionada uma amostra aleatória de 9 observações da distribuição $N(\mu, \sigma^2)$ com parâmetros desconhecidos. Obteve-se $\overline{X} = 22$ e $\sum_{i=1}^{n}(X_i - \overline{X})^2 = 72$.

(a) Teste as hipóteses $H_0 : \mu \leq 20 \times H_1 : \mu > 20$ ao nível de significância 0,05.

(b) Teste as hipóteses $H_0 : \mu = 20 \times H_1 : \mu \neq 20$ ao nível de significância 0,05. Use um teste simétrico com probabilidade 0,025 em cada cauda.

5. O tempo médio, por operário, para executar uma tarefa, tem sido de 100 minutos com desvio padrão de 15 minutos. Foi introduzida uma modificação para reduzir este tempo e após alguns meses foi selecionada uma amostra de 16 operários medindo-se o tempo de execução de cada um. Obteve-se um tempo médio amostral de 90 minutos e um desvio padrão amostral de 16 minutos. Estabeleça claramente as suposições que precisam ser feitas.

(a) Verifique se existem evidências, ao nível de significância 0,05, de que a modificação surtiu efeito?

(b) Verifique se há evidências, ao nível de significância 0,05, de que a modificação alterou a variância populacional.

6. Uma indústria compra componentes eletrônicos dos fornecedores A e B, mas o fornecedor A garante que o tempo médio de vida (em horas) do seu produto supera o da marca B em 300 horas. Para testar esta afirmação foram selecionadas duas amostras de componentes, uma de cada fornecedor, e obteve-se os seguintes tempos de vida:

marca A	1500 1450 1480 1520 1510
marca B	1100 1200 1180 1250

Após estabelecer claramente as suposições que precisam ser feitas,

(a) teste a hipótese de igualdade das variâncias dos tempos de vida, ao nível de significância 0,02;

(b) teste a afirmação do fornecedor A, ao nível de significância 0,05.

7. Uma droga A foi administrada em um grupo de 8 pacientes selecionados ao acaso. Após um período fixo de tempo a concentração da droga em certas células de cada paciente foi medida (em unidades apropriadas). O procedimento foi repetido em um outro grupo de 6 pacientes selecionados ao acaso usando uma droga B. As concentrações obtidas foram

droga A	1,23 1,42 1,41 1,62 1,55 1,51 1,60 1,76
droga B	1,76 1,41 1,87 1,49 1,67 1,81

Após estabelecer claramente as suposições que precisam ser feitas,

(a) teste a hipótese de que a concentração média de droga A entre todos os pacientes é pelo menos tão grande quanto da droga B;

(b) teste a hipótese de que as concentrações médias das duas drogas são iguais.

8. Mostre que o teste bilateral para a variância dado na Seção 6.6 é o teste da RMV.

6.7 Testes Assintóticos

Vimos que a construção de um teste envolve a obtenção de constantes através da distribuição de probabilidades de uma estatística. Em muitas situações, particularmente para a razão de máxima verossimilhança, estas distribuições não podem ser determinadas de forma exata e precisamos recorrer a resultados aproximados. Nesta seção serão desenvolvidos testes baseados em distribuições assintóticas das estatísticas de teste envolvidas. Iremos nos concentrar em testes baseados na distribuição assintótica da razão de máxima verossimilhança, do estimador de máxima verossimilhança e da função escore.

Suponha que uma amostra aleatória $X_1, \ldots, X_n$ é tomada de uma distribuição com parâmetro $\theta \in \Theta \subseteq \mathbb{R}$ desconhecido e queremos testar $H_0 : \theta = \theta_0$. Expandindo em série de Taylor a função $L(\theta_0) = \log p(\boldsymbol{x}|\theta_0)$ em torno do estimador de máxima verossimilhança $\hat{\theta}$ obtemos

$$L(\theta_0) \approx L(\hat{\theta}) + U(\boldsymbol{x}; \hat{\theta})(\theta_0 - \hat{\theta}) - \frac{1}{2} J(\hat{\theta})(\theta_0 - \hat{\theta})^2$$

sendo que J é a informação observada de Fisher e podemos desprezar os termos de ordem mais alta já que, sob H_0, θ_0 e $\hat{\theta}$ estão próximos para n grande.

Mas a função escore avaliada em $\hat{\theta}$ é igual a zero por definição. Além disso, a

razão de máxima verossimilhança neste caso é

$$\lambda(\boldsymbol{X}) = \frac{p(\boldsymbol{X}|\theta_0)}{p(\boldsymbol{X}|\hat{\theta})}$$

e podemos escrever então que

$$-2\log\lambda(\boldsymbol{X}) = -2\log\left(\frac{p(\boldsymbol{X}|\theta_0)}{p(\boldsymbol{X}|\hat{\theta})}\right) = -2[L(\theta_0) - L(\hat{\theta})] \approx J(\hat{\theta})(\theta_0 - \hat{\theta})^2.$$

Lembrando que $\hat{\theta}$ é assintóticamente normal com média θ e usando o fato de que $J(\hat{\theta})/n$ converge quase certamente para o seu valor esperado $I(\theta_0)/n$ quando H_0 é verdadeira então a distribuição assintótica de $-2\log\lambda(\boldsymbol{X})$ é χ^2_1. Assim, um teste com nível de significância assintótico α rejeita H_0 se $-2\log\lambda(\boldsymbol{X}) > c$ onde c é tal que $P(-2\log\lambda(\boldsymbol{X}) > c|\theta = \theta_0) = \alpha$.

Este resultado pode ser generalizado para o caso de um vetor de parâmetros $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$ de dimensão k. Neste caso, a estatística $-2\log\lambda(\boldsymbol{X})$ tem distribuição assintótica χ^2_k.

6.7.1 Teste Qui-quadrado

Um caso de particular interesse em Estatística é quando os dados são tais que cada observação pode ser classificada de acordo com um número finito de possíveis categorias. Por isso, observações deste tipo são chamadas *dados categóricos* e estaremos interessados em fazer inferência sobre as probabilidades de cada categoria.

Suponha que uma população consiste de itens que podem ser classificados em k diferentes categorias. Seja θ_i a probabilidade de que um item selecionado ao acaso pertença à categoria i, $i = 1, \ldots, k$. Assumimos também que $\theta_i \geq 0$, $i = 1, \ldots, k$ e $\sum_{i=1}^n \theta_i = 1$. Sejam agora os valores específicos $\theta_1^0, \ldots, \theta_k^0$ tais que $\theta_i^0 > 0$, $i = 1, \ldots, k$ e $\sum_{i=1}^n \theta_i^0 = 1$ e queremos testar as hipóteses

$$\begin{aligned} &H_0 : \theta_i = \theta_i^0, \quad i = 1, \ldots, k \\ &H_0 : \theta_i \neq \theta_i^0, \quad \text{para ao menor um valor de } i. \end{aligned} \tag{6.3}$$

Suponha agora que uma amostra aleatória de tamanho n é tomada desta população e as hipóteses (6.3) serão testadas com base nesta amostra. Para isto vamos denotar por N_i o número amostral de observações na categoria i, i.e. $N_1, \ldots, N_k$ são inteiros não negativos tais que $\sum_{i=1}^k N_i = n$. Quando H_0 é verdadeira, o número esperado de observações do tipo i é $n\theta_i^0$ e a diferença entre o número observado e o número esperado tende a ser menor quando H_0 é verdadeira do que quando ela é falsa. Parece razoável então basear o teste nas

magnitudes relativas destas diferenças. Neste caso, usando-se a função escore pode-se mostrar que o teste assintótico rejeita H_0 se

$$Q = \sum_{i=1}^{k} \frac{(N_i - n\theta_i^0)^2}{n\theta_i^0} > c$$

onde a estatística Q tem distribuição assintótica χ^2_{k-1}. Estes testes também são conhecidos na literatura como *testes de qualidade de ajuste* ou *testes de aderência* e estão entre os mais utilizados em Estatística.

Uma observação de ordem prática é que as frequências esperadas $n\theta_i^0$ não devem ser muito pequenas para que a distribuição χ^2 seja uma boa aproximação da distribuição de Q. Especificamente, pode-se mostrar que a aproximação será muito boa se $n\theta_i^0 \geq 5$ e apenas razoável $n\theta_i^0 \geq 1,5$.

Várias aplicações para dados categóricos e métodos não paramétricos que utilizam testes qui-quadrado podem ser vistas por exemplo em DeGroot (1989).

Testes de Aderência

Suponha agora que deseja-se testar a hipótese de que a amostra foi tomada de uma certa distribuição indexada por um vetor de parâmetros $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)$. Neste caso a hipótese alternativa é que a amostra foi tomada de alguma outra distribuição. Além disso, suponha que foram observados k valores de uma variável categórica ou os dados foram agrupados em k classes de valores.

Agora, para calcular as probabilidades de que um item pertença a cada uma das classes ou categorias precisamos dos valores estimados dos parâmetros $\theta_1, \ldots, \theta_m$. Se usarmos estimativas de máxima verossimilhança pode-se mostrar que a estatística Q tem distribuição assintótica qui-quadrado com $k-1-m$ graus de liberdade sendo m o número de parâmetros estimados no modelo teórico. Uma condição de validade desta distribuição é que $e_i \geq 5$, $i = 1, \ldots, k$.

Exemplo 6.12 : A proporção p de itens defeituosos em um grande lote é desconhecida e deseja-se testar as hipóteses

$$H_0 : p = 0,1$$
$$H_1 : p \neq 0,1$$

com base em uma amostra aleatória de 100 itens dos quais 16 são defeituosos. Podemos usar o teste qui-quadrado com duas categorias (defeituoso e não de-

feituoso) reformulando as hipóteses acima como

$$H_0 : p_1 = 0,1 \quad \text{e} \quad p_2 = 0,9$$
$$H_1 : H_0 \text{ é falsa}$$

sendo p_1 e p_2 as probabilidades de um item ser defeituoso ou não defeituoso respectivamente. As frequências observadas e esperadas sob H_0 são

$$N_1 = 16, N_2 = 84,\ np_1^0 = 10,\ np_2^0 = 90$$

e portanto o valor observado da estatística de teste é

$$Q = \frac{(16-10)^2}{10} + \frac{(84-90)^2}{90} = 4.$$

Usando uma tabela da distribuição qui-quadrado com 1 grau de liberdade obtém-se que 0,025 $<$ P-valor $<$ 0,05 e assim H_0 deve ser rejeitada ao nível de 5% e aceita ao nível de 2,5%.

Exemplo 6.13 : O teste χ^2 também pode ser aplicado no estudo da relação entre duas variaveis categóricas com p e k possíveis categorias. Neste caso queremos testar se as variáveis são independentes (hipótese nula). A estatística de teste é a mesma porém com número de graus de liberdade igual a $(p-1)(k-1)$ Considere por exemplo a Tabela 6.3 na qual estão apresentados os número de alunos matriculados nos colégios A e B, em relação à sua classe social. Se as

Tabela 6.2:

	Classe social			
Colégio	Alta	Media	Baixa	Total
A	20	40	40	100
B	50	40	30	120
Total	70	80	70	220

variáveis Colégio e Classe social forem independentes espera-se que as frequências de alunos das 3 classes sejam as mesmas nos 2 colégios, i.e. 70/220, 80/220 e 70/220. As frequências esperadas sob a hipótese de independência são então dadas por

$$\text{Colégio A:}\ \ 100\frac{70}{220} = 31,82\ \ 100\frac{80}{220} = 36,36\ \ 100\frac{70}{220} = 31,82$$

$$\text{Colégio B:}\ \ 120\frac{70}{220} = 38,18\ \ 120\frac{80}{220} = 43,64\ \ 120\frac{70}{220} = 38,18$$

e podemos construir a tabela abaixo.

Tabela 6.3: Frequências esperadas sob a hipótese de independência.

	Classe social		
Colégio	Alta	Media	Baixa
A	31,82	36,36	31,82
B	38,18	43,64	38,18

Podemos agora avaliar a estatística de teste

$$T = \frac{(20-31,82)^2}{31,82} + \frac{(40-36,36)^2}{36,36} + \frac{(40-31,82)^2}{31,82} + \frac{(50-38,18)^2}{38,18} + \frac{(40-43,64)^2}{43,64} + \frac{(30-38,18)^2}{38,18} = 12,57.$$

Ao nível de significância 0,05 obtemos da tabela χ^2 com $(p-1)(k-1) = 2$ graus de liberdade que $P(T > 5,99) = 0,05$ e como $12,57 > 5,99$ a hipótese de independência é rejeitada. Para calcular o P-valor, note que a tabela qui-quadrado com 2 graus de liberdade nos fornece,

$$P(T > 12,429) = 0,002$$

e portanto podemos concluir que P-valor $<$ 0,002. Ou seja, existe forte evidência contra a hipótese de independência entre as variáveis Colégio e Classe social.

6.8 Problemas

1. Em uma amostra de 100 lotes com 5 itens cada um, verificou-se que o número de itens defeituosos tem a distribuição de frequências abaixo. Teste a adequação do modelo binomial.

n^o de defeituosos	0	1	2	3	4	5	total
n^o de lotes	75	21	3	1	0	0	100

2. Em uma amostra de 300 itens, o número de defeitos observados em cada um deles tem a distribuição de frequências dada na tabela abaixo. Teste a adequação do modelo Poisson.

n^o de defeitos	0	1	2	3	4	total
n^o de itens	80	122	53	31	14	300

3. Em seus experimentos com ervilhas, Mendel ao cruzar plantas de sementes amarelas lisas com plantas de sementes verdes enrugadas observou a seguinte descendência na $2^{\underline{a}}$ geração: 315 plantas com sementes amarelas lisas, 108 com sementes amarelas enrugadas, 101 com sementes verdes lisas e 32 com sementes verdes enrugadas. De acordo com os postulados de Mendel a segregação esperada nesta geração deveria seguir a proporção de 9:3:3:1. Verifique se a teoria de Mendel explica a segregação observada.

4. Em uma amostra de 1800 valores no intervalo (0,1) obteve-se 391 valores entre 0 e 0,2, 490 valores entre 0,2 e 0,5, 580 entre 0,5 e 0,8; e 339 maiores do que 0,8. Teste a hipótese de que a amostra foi tomada de uma distribuição uniforme no intervalo (0,1) (neste caso a probabilidade de um valor cair no intervalo (a, b) é $b - a$).

6.9 Testes Bayesianos

Do ponto de vista Bayesiano, podemos atribuir probabilidades a priori $p(H_0)$ e $p(H_1)$ para um par de hipóteses estatísticas H_0 e H_1. Após observar uma amostra aleatória $X_1, \ldots, X_n$ e aplicando o teorema de Bayes obtemos as probabilidades a posteriori das hipóteses,

$$p(H_0|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|H_0)p(H_0)}{p(\boldsymbol{x})} \quad \text{e} \quad p(H_1|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|H_1)p(H_1)}{p(\boldsymbol{x})}.$$

Tomando-se a razão das probabilidades a posteriori (e notando que o termo $p(\boldsymbol{x})$ se cancela) obtemos

$$\underbrace{\frac{P(H_0|\boldsymbol{x})}{P(H_1|\boldsymbol{x})}}_{\substack{\text{razão de chances} \\ \text{a posteriori}}} = \underbrace{\frac{P(\boldsymbol{x}|H_0)}{P(\boldsymbol{x}|H_1)}}_{\text{fator de Bayes}} \underbrace{\frac{P(H_0)}{P(H_1)}}_{\substack{\text{razão de chances} \\ \text{a priori}}}.$$

O fator de Bayes (FB) será usado para testar as hipóteses e pode ser reescrito como

$$FB = \frac{P(\boldsymbol{x}|H_0)}{P(\boldsymbol{x}|H_1)} = \frac{\displaystyle\int p(\theta|H_0)p(\boldsymbol{x}|\theta, H_0)d\theta}{\displaystyle\int p(\theta|H_1)p(\boldsymbol{x}|\theta, H_1)d\theta}.$$

Note que o fator de Bayes é similar à razão de verossimilhanças porém ao invés de maximizar a verossimilhança toma-se uma média ponderada com pesos $p(\theta|H_i)$. Na escala logarítmica o fator de Bayes é algumas vezes chamado de *força (ou peso) da evidência* fornecido pelos dados para H_0 contra H_1.

Um fator de Bayes grande indica evidência a favor de H_0 e a seguinte escala pode ser usada,

FB	log FB	Força da evidência
< 1	< 0	negativa (suporta H_1)
$[1, 3]$	$[0, 5]$	fraca
$(3, 12]$	$(5, 11]$	positiva
$(12, 150]$	$(11, 22]$	forte
> 150	> 22	muito forte

Capítulo 7

Correlação e Regressão

Em diversas investigações deseja-se avaliar a relação entre duas medidas quantitativas. Por exemplo, as alturas dos filhos estão relacionadas com as alturas dos seus pais? O faturamento de uma empresa é afetado pelo número de funcionários? A produção de uma máquina depende do nível de treinamento do operador? Note que nestes casos não estamos mais interessados em amostras independentes como na seção anterior.

Em geral os principais objetivos de tais investigações são os seguintes.

- Verificar se as variáveis estão *associados*, isto é se os valores de uma variável tendem a crescer (ou decrescer) à medida que os valores da outra variável crescem.

- Predizer o valor de uma variável a partir de um valor conhecido da outra.

- Descrever a relação entre as variáveis, isto é dado um aumento específico numa variável, qual o crescimento médio esperado para a outra variável?

Uma primeira aproximação para o tipo de associação entre duas variáveis é através de funções lineares. O grau de associação linear entre duas variáveis é medido usando um parâmetro chamado *coeficiente de correlação*. Já para predizer o valor de uma variável contínua a partir de uma outra variável e para descrever a relação entre duas variáveis utiliza-se *métodos de regressão* que serão estudados no próximo capítulo.

O primeiro estágio em qualquer um dos casos é fazer um gráfico de pontos dos dados para ter alguma idéia da forma e grau de associação entre duas variáveis (como na Figura tipo de gráfico. Mesmo com apenas 18 observações, parece existir algum tipo de associação entre estas variáveis.

7.1 Definições

Seja $x_1, \ldots, x_n$ e $y_1, \ldots, y_n$ os valores amostrais de duas variáveis X e Y. Sejam $\bar{x}$, $\bar{y}$, s_x e s_y as médias e desvios padrão amostrais dos dois conjuntos de dados. A idéia aqui é tentar quantificar o grau de associação linear entre X e Y a partir dos desvios em torno das médias amostrais, definindo-se para cada par (x_i, y_i) o produto

$$c_i = (x_i - \bar{x}) \times (y_i - \bar{y}).$$

Intuitivamente, se valores altos de x tendem a acompanhar valores altos de y, e se valores baixos de x acompanham valores baixos de y então c_i tenderá a ser positivo em sua maioria (correlação positiva). Se valores altos de x acompanham valores baixos de y e vice-versa então a maioria dos valores c_i serão negativos (correlação negativa). Se não existir associação entre x e y então se tomarmos a média aritmética dos valores c_i, valores positivos e negativos tenderão a se cancelar e a média será próxima de zero.

A *covariância amostral* de x e y é definida como

$$s_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{n-1}.$$

sendo essencialmente a média dos valores de c_i acima.

Embora seja intuitiva esta medida é de difícil interpretação pois está definida na reta dos reais e depende das escalas dos dados. Por exemplo se multiplicarmos todos os valores de x por uma constante então a covariância também fica multiplicada por esta constante.

Dividindo-se a covariância amostral pelos desvios padrão amostrais obtemos uma medida do grau de associação linear entre duas variáveis que é adimensional e de mais fácil interpretação. Esta nova medida chama-se *coeficiente de correlação amostral* e é definida como

$$r = \frac{s_{xy}}{s_x s_y}.$$

Pode-se mostrar que $-1 \leq r \leq 1$. Quanto mais próximo de -1 ou 1 estiver o valor calculado de r maior é o grau de associação linear (negativa ou positiva) entre as variáveis e quanto mais próximo de zero menor é o grau de associação.

Exemplo 7.1 : Foram observados $n = 18$ valores de duas variáveis x e y e obteve-se $\bar{x} = 0,48$, $\bar{y} = 1,58$, $s_x = 0,18$, $s_y = 0,54$ e $\sum x_i y_i = 12,44$. A partir destes valores podemos calcular a covariância amostral $s_{xy} = -0,0712$ e portanto a correlação amostral é $r = -0,732$. Isto indica que possivelmente estas variáveis estão negativamente correlacionadas (ao menos linearmente).

O coeficiante de correlação populacional (que é um parâmetro desconhecido) é

denotado pela letra grega ρ e também está definido no intervalo [-1,1]. Os valores -1 e 1 representam correlação linear perfeita (negativa ou positiva) enquanto o valor zero representa ausência de correlação linear. Podemos considerar r como sendo uma estimativa de ρ. Na Figura com seus coeficientes de correlação amostrais calculados.

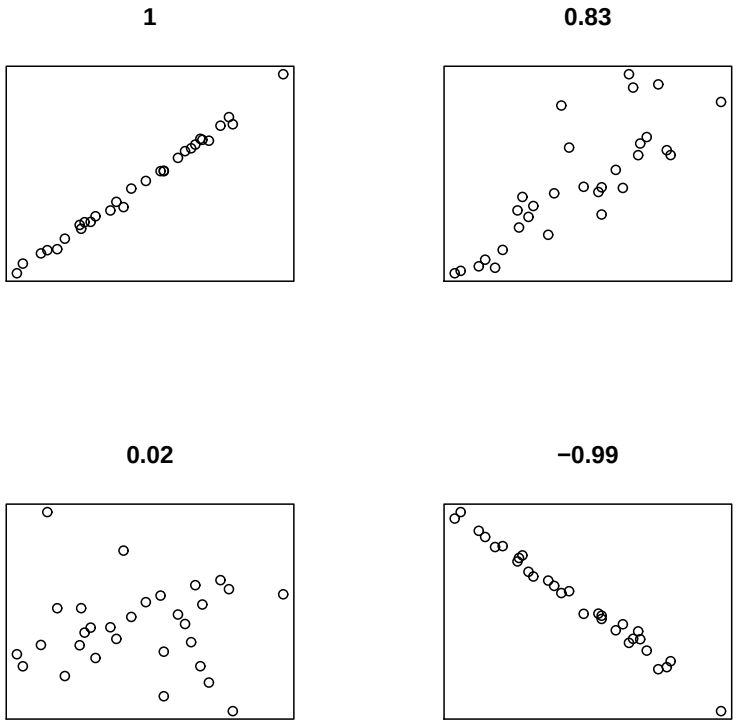

Figura 7.1: Exemplos de correlação entre variáveis.

7.2 Interpretação do coeficiente de correlação

O valor de r está sempre entre -1 e 1, com $r = 0$ correspondendo à não associação.

$$\text{Valores de } r \left\{ \begin{array}{c} \text{negativos} \\ \text{positivos} \end{array} \right\} \text{ indicam uma associação } \left\{ \begin{array}{c} \text{negativa} \\ \text{positiva} \end{array} \right\}$$

Usamos o termo correlação positiva quando $r > 0$, e nesse caso à medida que x cresce também cresce y, e correlação negativa quando $r < 0$, e nesse caso à medida que x cresce, y decresce (em média).

Quanto maior o valor de r (positivo ou negativo), mais forte a associação. Nos extremos, se $r = 1$ ou $r = -1$ então todos os pontos no gráfico de dispersão caem exatamente numa linha reta. No outro extremo, se $r = 0$ não existe nenhuma associação linear.

A seguinte quadro fornece um guia de como podemos descrever uma correlação em palavras dado o valor numérico. É claro que as interpretações dependem de cada contexto em particular.

Valor de ρ (+ ou −)	Interpretação
0,00 a 0,19	Uma correlação bem fraca
0,20 a 0,39	Uma correlação fraca
0,40 a 0,69	Uma correlação moderada
0,70 a 0,89	Uma correlação forte
0,90 a 1,00	Uma correlação muito forte

É importante notar que as correlações não dependem da escala de valores dos dados. Por exemplo, obteríamos o mesmo valor de r se medíssemos altura e peso em metros e quilogramas ou em pés e libras.

Se pudermos supor que as amostras são provenientes de distribuições normais então testes de hipóteses e intervalos de confiança podem ser construidos para o coeficiente de correlação teórico ρ. Neste caso a estatística a ser utilizada é

$$T = \frac{r - \rho}{\sqrt{\frac{1 - r^2}{n - 2}}}$$

que tem distribuição t com $n - 2$ graus de liberdade. Assim, um I.C. de $100(1 - \alpha)\%$ para ρ após as amostras serem observadas é dado por

$$r - t_{\alpha/2}\sqrt{\frac{1 - r^2}{n - 2}} < \rho < r + t_{\alpha/2}\sqrt{\frac{1 - r^2}{n - 2}}.$$

As hipóteses de interesse são em geral do tipo bilateral, ou seja

$$H_0 : \rho = 0$$
$$H_1 : \rho \neq 0.$$

Assim, após observar as amostras calcula-se o valor de r e o p-valor do teste é dado pela probabilidade

$$P\left(|T| > \frac{r}{\sqrt{\frac{1 - r^2}{n - 2}}} \right)$$

onde $T \sim t_{n-2}$. Note que a não rejeição de H_0 nos diz que não há evidência amostral de haver correlação linear entre as variáveis. Em outras palavras, a correlação entre as variáveis não é significativa.

Exemplo 7.2 : Na Figura 7.1 os dados foram simuladas de distribuições normais e cada amostra tem 30 observações. As correlações amostrais r, estatistica t observadas, I.C. de 95% e os p-valores de testes de hipóteses bilaterais estão na Tabela 7.1.

Tabela 7.1: Correlações amostrais, estatisticas observadas, I.C. de 95% e p-valores bilaterais.

r	t	g.l.	IC 95%	p-valor
0,9914	40,1368	28	0,9819 0,9960	$< 0{,}001$
0,7477	5,9590	28	0,5303 0,8729	$< 0{,}001$
0,0259	0,1372	28	-0,3375 0,3826	0,8918
-0,9981	-84,8822	28	-0,9991 -0,9959	$< 0{,}001$

Observações Discrepantes

A reta de regressão é estimada com base na soma de quadrados das distâncias dos pontos em relação à reta. Por isso, observações discrepantes ou *outliers* podem ter uma grande influência na estimativa da inclinação da reta e consequentemente no coeficiente de correlação amostral.

Linearidade e normalidade

É bom enfatizar que somente relações lineares são detectadas pelo coeficiente de correlação que acabamos de descrever (também chamado coeficiente de correlação de Pearson). Ou seja, aceitar a hipótese de que $\rho = 0$ não necessariamente implica que as variáveis não estejam de alguma forma associadas.

Por exemplo, nos gráficos da Figura 7.2, mesmo existindo uma clara relação (não-linear) entre as variáveis x e y, o coeficiente de correlação é estatisticamente zero (Verifique!).
A mensagem aqui é que deve-se sempre fazer o gráfico dos dados de modo que se possa tentar visualizar tais relações.

Transformações

Em alguns casos pode ser apropriado e mesmo justificável fazer transformações em x e/ou y induzindo uma relação linear na escala transformada. Por exemplo, na

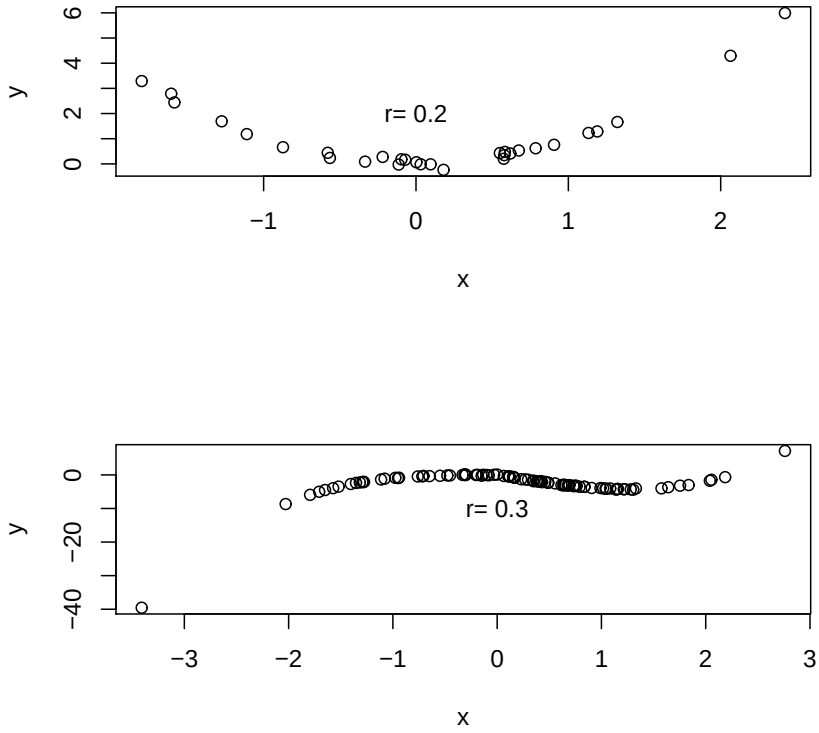

Figura 7.2: Exemplos de associação não linear entre duas variáveis simuladas.

Figura militares contra o produto interno bruto em 75 países. É difícil vizualizar uma relação linear entre estas variáveis especialmente para valores grandes. No gráfico da direita foi tomado o logaritmo natural das variáveis e a relação linear fica bem mais aparente.

Correlação não significa causalidade

Um dos erros de interpretação mais comuns é assumir que correlações significativas necessariamente implicam em uma relação de causa e efeito entre duas variáveis. Esta interpretação é incorreta. Na verdade é extremamente difícil estabelecer relações causais a partir de dados observados. Seria preciso realizar experimentos controlados para obter mais evidências de um relação causal.

Também é preciso ter cuidado ao assumir que existe correlação somente porque duas variáveis seguem o mesmo padrão de variabilidade. A correlação pode ser devida a uma terceira variável influenciando as duas primeiras.

Finalmente, vale notar que correlações estatisticamente significativas (i.e. quando se rejeita a hipótese de correlação nula) não necessariamente tem sig-

nificado prático. Por exemplo, que conclusões poderia-se tirar de uma correlação significativa positiva entre nascimento de bebês e número de cegonhas em determinada região?

Resumindo, se encontramos uma associação ou correlação entre duas variáveis X e Y podem existir diversas explicações do porque elas variam conjuntamente, incluindo:

- Mudanças em X causam mudanças em Y.
- Mudanças em Y causam mudanças em X.
- Mudanças em outras variáveis causam mudanças tanto em X quanto em Y.
- A relação observada é somente uma coincidência.

A terceira explicação é frequentemente a mais apropriada. Por exemplo, o número de pessoas usando óculos-de-sol e a quantidade de sorvete consumido num particular dia são altamente correlacionados. Isto não significa que usar óculos-de-sol causa a compra de sorvetes ou vice-versa, mas sim que existe uma outra variável, a temperatura, causando as duas primeiras.

Coeficiente de determinação

O quadrado do coeficiente de correlação de Pearson é chamado de *coeficiente de determinação* e costuma ser denotado por R^2. Esta é uma medida da proporção da variabilidade em uma variável que é explicada pela variabilidade da outra. Na prática não se espera encontrar uma correlação perfeita (i.e. $R^2 = 1$), porque existem muitos fatores que determinam as relações entre variáveis no mundo real.

Por exemplo, na Figura **??** se o coeficiente de correlação calculado para os logaritmos dos gastos militares e PIB dos países for $r = 0,80$, então $R^2 = 0,64$ ou 64%. Ou seja, cerca de 36% da variabilidade nos gastos militares não pode ser descrita ou explicada pela variabilidade nos PIB e portanto fica claro que existem outros fatores que poderiam ser importantes.

7.3 Problemas

1. Dados os valores x=(-2,-1,0,1,2) e y=(4,2,0,1,2) calcule o coeficiente de correlação amostral e teste a hipótese de correlação nula. Faça um gráfico de dispersão e comente os resultados.

2. Dados os valores x=(-2,-1,0,1,2) e y=(-8,-1,0,1,8) calcule o coeficiente de correlação amostral. Teste a hipótese de não haver correlação linear. Qual a sua conclusão notando que $y_i = x_i^3$?

3. Verifique o que ocorre com a covariância amostral se

   (a) uma constante k for somada a todos os valores de x;

   (b) todos os valores de x forem multiplicados por uma constante k;

   (c) idem para o coeficiente de correlação amostral

4. Comente os resultados da Tabela 7.1. Nesta tabela obtenha intervalos de confiança de 98% para o coeficiente de correlação linear populacional.

5. Os resultados abaixo foram obtidos em um pacote estatístico. Comente.

```
        correlação de Pearson

t = -2,0134, gl = 28, p-valor = 0,05378

Hipótese alternativa: correlação verdadeira diferente de 0.

Intervalo de confiança de 95%: [-0,6346; 0,0053]
estimativa amostral: -0,3556
```

6. No item anterior suponha que o gráfico das variáveis é similar ao da Figura 7.2. Qual a sua conclusão?

7. Um estudo geoquímico foi realizado utilizando amostras compostas de sedimentos de corrente com granulometria de 100-150 *mesh* e profundidade de $40cm$, provenientes de riachos correndo sobre granulitos e revelou os seguintes resultados em *ppm*:

Ni	Cr	Ni	Cr
5.2	16,8	4,5	15,5
5.0	20,0	5,4	13,0
6.8	14,2	8,8	12,5
7.5	17,5	18,0	20,2
2.5	10,1	6,2	12,5
5.0	15,5	20,5	13,5
7.5	13,8	10,0	17,8
7.0	18,2	4,0	12,8
8.0	13,0	4,4	12,2
4.0	15,0	15,9	13,0

   (a) Faça o gráfico destes dados com Ni no eixo x.

(b) Calcule o coeficiente de correlação amostral para estes dados e verifique se o valor obtido parece consistente com seu gráfico.

(c) Qual proporção da variabilidade na concentração de Cr pode ser explicada pela concentração de Ni?

8. Em um estudo da influência de processos praianos no condicionamento do ângulo de inclinação do fundo oceânico situado logo após a linha da maré baixa a estirâncio mediu-se a profundidade da lâmina d'água (em pés). Os dados coletados foram:

ângulo de	0.68	0.85	0.66	0.50	1.86	2.33	2.17	1.83	1.68
inclinação y	2.05	1.83	1.84	1.87	1.82	1.85	1.75	1.51	1.38
profundidade x	12.4	11.4	10.7	11.6	11.3	10.7	11.1	12.8	13.3
	13.3	14.1	13.4	13.5	13.3	14.4	14.1	15.3	14.0

(a) Faça o gráfico desses dados com profundidade da lâmina d'água no eixo x.

(b) Calcule o coeficiente de correlação, r e interprete o resultado obtido.

(c) Qual proporção da variabilidade em ângulo de inclinação pode ser explicada por profundidade da lâmina d'água?

7.4 Regressão

Em muitas situações o fenômeno a ser estudado envolve duas ou mais variáveis e para responder a certas questões científicas precisamos estabelecer uma relação funcional entre elas. Um problema de regressão consiste em determinar a função que descreve esta relação. Aqui estudaremos somente o caso em que esta relação é descrita por uma função linear. Veremos primeiro o caso particular de duas variáveis.

Por exemplo, se conhecemos a altura de um indivíduo, mas não o seu peso, qual seria um bom chute para o peso deste indivíduo? O coeficiente de correlação apenas indica a grau de associação como um único número. Suponha que dispomos de amostras de alturas $x_1, \ldots, x_n$ e pesos $y_1, \ldots, y_n$ de n indivíduos. Por enquanto vamos ignorar se eles são do sexo masculino ou feminino. Se estamos interessados em predizer o peso a partir da altura então não temos uma relação simétrica entre as duas variáveis. Chamamos peso de *variável resposta* ou *dependente*, e altura de *variável explicativa*, *preditora*, *regressora* ou *independente*.

Em um gráfico de pontos os valores da variável resposta (y) são em geral dispostos no eixo vertical, e da variável explicativa (x) no eixo horizontal. Por exemplo, na Figura 7.3 temos 30 observações de pesos e alturas de indivíduos selecionados aleatoriamente em uma população.

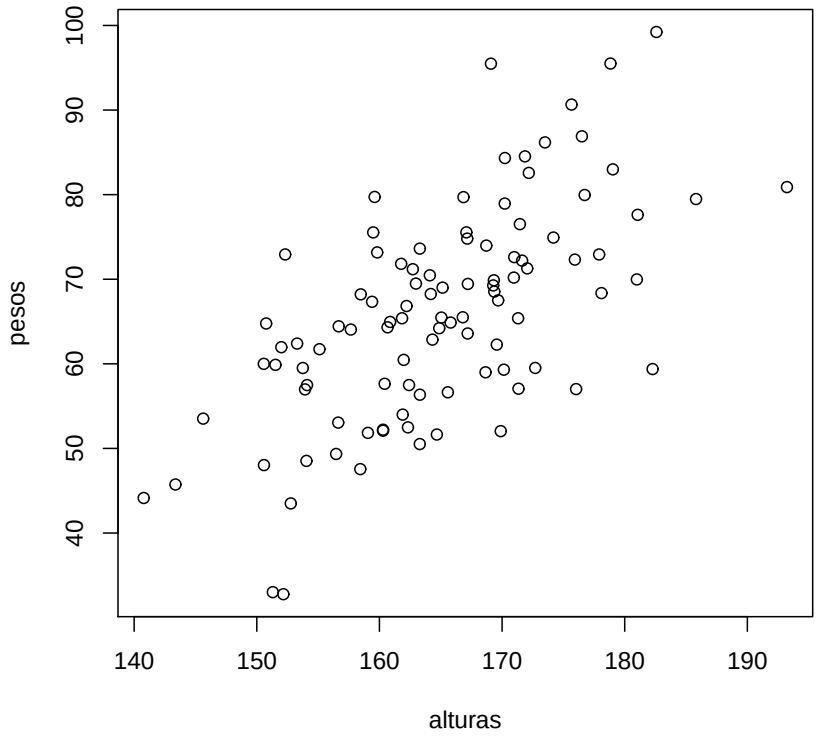

Figura 7.3: 30 observações de pesos e alturas de indivíduos em uma população.

Se a relação entre as duas variáveis é aproximadamente linear, então os dados podem ser resumidos através do ajuste de uma reta passando pelos pontos. A equação dessa reta é dada por

$$y = \alpha + \beta x$$

onde o intercepto α e a inclinação β são parâmetros desconhecidos. Existe uma infinidade de possíveis retas passando pelos pontos mas intuitivamente queremos aquela que forneça pequenas diferenças entre os pesos observados (y_i) e aqueles dados pela reta para as alturas correspondentes. Estas diferenças (ou erros) são então dadas por

$$y_i - \alpha - \beta x_i$$

e estão representadas pelas linhas verticais na Figura 7.4 para 11 pontos.

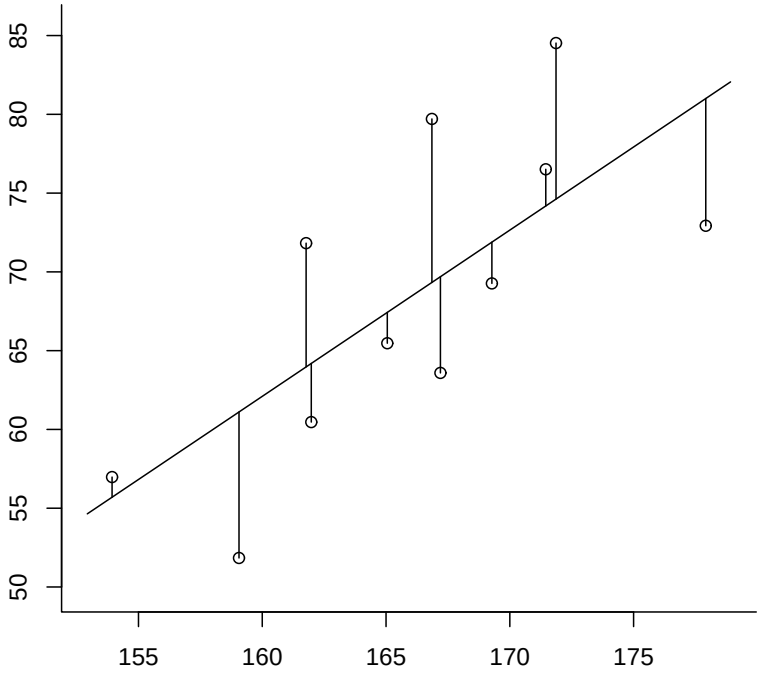

Figura 7.4: Diferenças entre valores de y e uma reta de regressão hipotética.

Parece razoável tentar minimizar alguma função destes erros. Em geral não importa se as diferenças são positivas ou negativas e todas elas tem o mesmo grau de importância. Assim, uma função que pode ser minimizada é

$$S(\alpha, \beta) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \quad = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2$$

onde $\hat{y}_i = \alpha + \beta x_i$ são chamados *valores ajustados*. O método que minimiza esta soma de quadrados dos erros para obter a melhor reta ajustada é chamado *método de mínimos quadrados* (MMQ) e as estimativas de α e β, denotadas por $\hat{\alpha}$ e $\hat{\beta}$, são então as *estimativas de mínimos quadrados*[1].

Igualando a zero a primeira derivada de $S(\alpha, \beta)$ em relação a α e β e resolvendo para $\hat{\alpha}$ e $\hat{\beta}$ não é difícil verificar que a melhor reta segundo este critério de

[1]Outras funções dos erros podem ser consideradas, e.g. soma dos erros absolutos, erro absoluto máximo, etc.

estimação é aquela tal que

$$\begin{aligned} \hat{\beta} &= \frac{\sum(y_i - \bar{y})(x_i - \bar{x})}{\sum(x_i - \bar{x})^2} = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x^2 - n\bar{x}^2} = \frac{s_{xy}}{s_x^2} \\ \hat{\alpha} &= \bar{y} - \hat{\beta}\bar{x}. \end{aligned} \tag{7.1}$$

As condições de segunda ordem também devem ser verificadas para garantir que este é um ponto de mínimo. Note que $\hat{\beta}$ pode reescrito como

$$\hat{\beta} = \frac{s_y}{s_y} \frac{s_{xy}}{s_x^2} = r \frac{s_y}{s_x}$$

e assim o coeficiente de correlação amostral pode ser obtido a partir da reta estimada.

Exemplo 7.3 : Suponha que para o exemplo das alturas e pesos de indivíduos obtivemos $\hat{\alpha} = -51,17$ kg e $\hat{\beta} = 0,68$ kg/cm. Então a reta de regressão estimada é dada por

$$y = -51,17 + 0,68x.$$

O valor estimado de β (0,68) pode ser interpretado como o aumento médio (ou aumento esperado) no peso quando a altura aumenta de 1cm. O valor estimado de α (-51,17) não possui qualquer significado já que neste caso não faz sentido incluir o ponto $x = 0$. Esta reta ajustada é uma estimativa da reta de regressão populacional (desconhecida), $y = \alpha + \beta x$.

O próximo passo é construir intervalos de confiança e testar hipóteses para α e β, mas para fazer isto precisamos pensar mais cuidadosamente sobre nossas suposições acerca da população.

7.4.1 Modelo de regressão linear simples

Este é o modelo mais simples para descrever a relação entre uma variável explicativa x e uma variável resposta y. O modelo faz a seguintes suposições, em ordem decrescente de importância:

1. o valor médio da variável resposta é uma função linear de x,

2. a variância de y é constante, ou seja é a mesma para todos os valores de x,

3. a variação aleatória de y para qualquer valor fixo de x segue uma distribuição normal, e estes termos de erro são independentes.

Em termos algébricos, dada uma amostra de pontos (x_i, y_i), $i = 1, \ldots, n$ o *modelo de regressão linear* é dado por

$$y_i = \alpha + \beta x_i + \varepsilon_i \tag{7.2}$$

onde ε_i representa desvios aleatórios (supostos independentes) da relação linear entre y e x. Para satisfazer às três suposições acima segue então que

$$\varepsilon_i \sim Normal(0, \sigma^2)$$

o que é equivalente a

$$y_i | x_i \sim Normal(\alpha + \beta x_i \ , \ \sigma^2).$$

Os parâmetros α e β são frequentemente chamados de coeficientes da regressão. Em particular, β é denominado *coeficiente* ou *efeito* de x sobre y já que representa o aumento (ou redução) esperado em y quando x aumenta de 1 unidade. O parâmetro α é a resposta média no ponto $x = 0$ e só tem interpretação prática se o modelo inclui este ponto.

Na Figura parecem satisfazer às três suposições, enquanto os dados representados no gráfico da direita não satisfazem à nenhuma das suposições.

7.4.2 Estimando os parâmetros do modelo

Aqui também os coefficientes da regressão (e agora σ^2) precisam ser estimados para obter a equação da reta ajustada. Um método de estimação muito utilizado em estatística é chamado *método de máxima verossimilhança*. No caso particular em que assumimos distribuição normal para os erros este método leva às mesmas estimativas de mínimos quadrados, i.e.

$$\hat{\beta} = s_{xy}/s_x^2 \qquad \text{e} \qquad \hat{\alpha} = \overline{y} - \hat{\beta}\overline{x}.$$

Em aplicações práticas, não existe garantia de que o modelo de regressão linear será razoável para nossos dados. Por isso devemos sempre sobrepor a reta ajustada $y = \hat{\alpha} + \hat{\beta}x$ sobre um diagrama de dispersão dos dados para checar se o modelo é razoável. Devemos procurar por evidências de uma relação não-linear, ou desvios muito extremos da reta ajustada.

Se julgamos que o modelo está razoável, podemos também estimar σ^2, a variância dos erros ε_i. Em geral a fórmula utilizada é

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^{n} (y_i - \hat{\alpha} - \hat{\beta}x_i)^2$$

e substituindo as expressões de $\hat{\alpha}$ e $\hat{\beta}$ obtém-se que

$$\hat{\sigma}^2 = \frac{n-1}{n-2}\left(S_y^2 - \frac{S_{xy}^2}{S_x^2}\right) \tag{7.3}$$

Exemplo 7.4 : Para os dados de alturas (x) e pesos (y) na Figura 7.3, sabendo-se que as médias amostrais são $\bar{x} = 164,3$ e $\bar{y} = 66,7$, as variâncias amostrais são $S_x^2 = 91$ e $S_y^2 = 81$ e a covariância amostral é $S_{xy} = 52,6$ segue que as estimativas dos coeficientes são

$$\hat{\beta} = 52,6/91 = 0,58$$

e

$$\hat{\alpha} = 66,7 - 0,58 \times 164,3 = -28,6.$$

Podemos agora obter uma estimativa da variância dos erros,

$$\hat{\sigma}^2 = \frac{29}{28}\left(81 - \frac{52,6^2}{91}\right) = 52,4.$$

Um gráfico dos dados com a reta ajustada é dado na Figura 7.5

O ajuste da reta não parece tão bom. Existem dois pontos bem distantes da reta ajustada, que parecem ter tido uma grande influência no ajuste. Na prática é aconselhavel investigar a acurácia destes valores e/ou verificar quanto muda a reta ajustada quando estes pontos são removidos.

7.4.3 Construindo intervalos e testando hipóteses

Usualmente é de interesse saber qual a precisão nas estimativas de α e principalmente de β. Para construir intervalos de confiança e testar hipóteses usaremos as seguintes estatísticas

$$\sqrt{\frac{n\sum(x_i-\bar{x})^2}{\sum x_i^2}}\left(\frac{\hat{\alpha}-\alpha}{\hat{\sigma}}\right) \quad \text{e} \quad \sqrt{\sum(x_i-\bar{x})^2}\left(\frac{\hat{\beta}-\beta}{\hat{\sigma}}\right).$$

Ambas tem distribuição t de Student com $n-2$ graus de liberdade e as demonstrações são omitidas. Assim, podemos construir intervalos de confiança obtendo o valor de t na tabela apropriada

$$\hat{\alpha} \pm t\hat{\sigma}\sqrt{\frac{\sum x_i^2}{n\sum(x_i-\bar{x})^2}} \quad \text{e} \quad \hat{\beta} \pm \frac{t\hat{\sigma}}{\sqrt{\sum(x_i-\bar{x})^2}}.$$

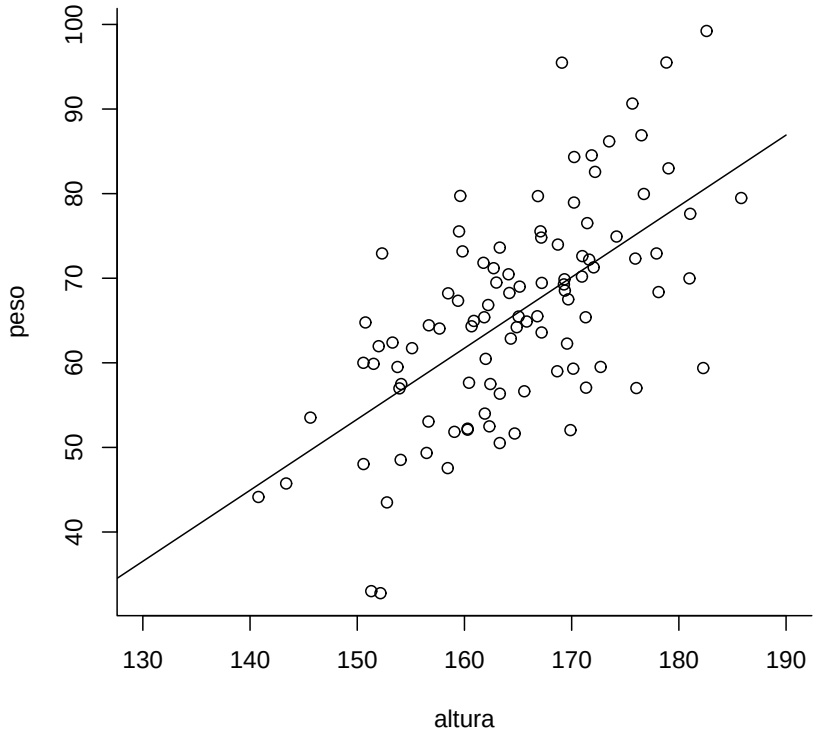

Figura 7.5: Dados de altura e peso com a reta de regressão ajustada.

Geralmente estamos interessados em testar as hipóteses

$$H_0 : \beta = 0$$
$$H_1 : \beta \neq 0$$

ou seja, de que não exista relação entre x e y. Nesse caso, após observar os dados calcula-se o valor da estatística de teste sob H_0

$$t = \sqrt{\sum (x_i - \bar{x})^2} \left(\frac{\hat{\beta}}{\hat{\sigma}} \right)$$

e na tabela t de Student com $n - 2$ graus de liberdade obtém-se o p-valor na forma usual.

Exemplo 7.5 : Para os dados da Figura 7.3, para testar a hipótese nula de não existência de relação entre altura e peso calculamos

$$\sqrt{\sum (x_i - \bar{x})^2} = \sqrt{(n-1)S_x^2} = \sqrt{29 * 91} = 51,37$$

e a estatística de teste fica

$$t = 51,37 \times 0,58 / \sqrt{52,4} = 4,12.$$

Na tabela t com 28 graus de liberdade obtém-se que o p-valor é menor do que 0,1% e portanto há evidência forte contra a hipótese $H_0 : \beta = 0$.

7.4.4 Transformações de dados

Uma forma de estender a aplicabilidade do modelo de regressão linear é aplicar uma transformação em x ou y, ou ambos, antes de ajustar o modelo. Se a relação entre duas variáveis é não-linear (uma curva pareceria ajustar melhor do que uma reta), então frequentemente a relação pode ser *feita* linear transformando uma ou ambas as variáveis.

No entanto deve-se tomar um certo cuidado com transformações. Elas podem ser muito úteis em algumas situações, mas só devem ser consideradas como um último recurso já que quando uma ou ambas as variáveis são transformadas, os coeficientes deixam de ter interpretações diretas.

Na prática precisamos então escolher uma transformação que faça a relação ser aproximadamente linear e que ainda permaneça interpretável. Por exemplo, frequentemente as relações são multiplicativas ao invés de aditivas e nestes casos transformações logarítmicas são particularmente úteis.

7.4.5 Representação Matricial

O modelo de regressão linear (7.2) pode ser representado em forma matricial. Empilhando todas as observações e definindo

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \quad \boldsymbol{X} = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \quad \boldsymbol{\theta} = \begin{bmatrix} \alpha \\ \beta \end{bmatrix} \quad \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

podemos reescrever o modelo como $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\theta} + \boldsymbol{\epsilon}$. Esta representação será útil quando mais variáveis explicativas forem introduzidas.

7.4.6 Problemas

1. No modelo de regressão linear simples (7.2),

   (a) mostre que as estimativas de mínimos quadrados são aquelas dadas pelas expressões (7.1);

   (b) verifique o que ocorre com as estimativas se uma constante k for somada a todos os valores de x;

   (c) verifique o que ocorre com as estimativas se todos os valores de x forem multiplicados por uma constante k;

   (d) derive a expressão (7.3) para estimativa de σ^2.

2. No modelo de regressão $\log(y_i) = \alpha + \beta x_i + \epsilon_i$, $i = 1, \ldots, n$, qual o efeito esperado sobre y quando x aumenta de 1 unidade.

3. Em um modelo de regressão linear as variáveis y e x são medidas em Kg e metros respectivamente. Se o modelo estimado foi $y = -2,3 + 8,2\ x$ qual o aumento esperado em y se x aumentar em 1cm?

4. Explique porque na Figura **??**(b) os dados não parecem satisfazer às suposições do modelo de regressão linear (7.2).

5. Comente os resultados na Figura 7.5.

6. Deseja-se verificar se uma determinada substância encontrada em pastos pode ser usada para melhorar o ganho de peso em bovinos. Foram selecionadas acaso 15 bois de mesma raça e idade e cada um recebeu uma concentração X da substância (em mg/l). Após 30 dias foram medidos os ganhos de peso Y (em Kg) para cada animal. Com os dados resultantes obteve-se: $\bar{x} = 2,7$, $\bar{y} = 16,14$, $\sum x_i y_i = 785,55$, $\sum x_i^2 = 163,39$ e $\sum y_i^2 = 4329,43$.

(a) Estime a reta de regressão e interprete os valores dos coeficientes.

(b) Teste a hipótese de que esta substância influencia no ganho de peso dos animais.

7. Na análise de um certo combustível obteve-se as observações abaixo das variáveis "poder calorífico" (y) e "percentual de cinzas" (x)

y	13100	11200	10200	9600	8800
x	18,3	27,5	36,4	48,5	57,8

(a) Obtenha a reta de regressão estimada e interprete os coeficientes estimados,

(b) estime o poder calorífico para 30% de cinzas,

(c) esboce o diagrama de pontos com a reta ajustada.

8. Comente os resultados abaixo que foram obtidos ao estimar um modelo de regressão linear em um pacote estatístico.

```
            Estimativas  EP        estatistica t   P-valor
Intercepto  3.7960       2.1616    1.756           0.09001
inclinação -0.7400       0.2417   -3.062           0.00482

variancia dos erros: 11.33 com 28 graus de liberdade
```

9. Os dados abaixo são referentes ao consumo per capita de vinho (x) e a taxa de mortalidade por infarto (y) observada em 9 países.

x	2,8	3,2	3,3	5,1	5,9	6,0	7,9	10,2	10,8
y	11,2	14,0	12,6	8,2	7,0	4,2	2,6	1,8	3,2

Sabe-se que $\bar{x} = 6,1$, $\bar{y} = 7,2$, $\sum_{i=1}^{9} x_i y_i = 299,5$, $\sum_{i=1}^{9} x_i^2 = 409$ e $\sum_{i=1}^{9} y_i^2 = 634$.

(a) Calcule o coeficiente de correlação amostral e comente.

(b) Teste a hipótese de que não existe correlação linear.

(c) Obtenha a reta de regressão estimada e interprete os coeficientes estimados.

(d) Estime a taxa de mortalidade se o consumo per capita for igual a 9.

(e) Teste a hipótese de que o consumo per capita de vinho não influencia a taxa de mortalidade por infarto.

(f) Como fica a reta de regressão se estes 9 países duplicarem o consumo per capita de vinho?

7.5 Regressão Linear Múltipla

Dada uma variável dependente y e k variáveis explicativas $x_1, \ldots, x_k$ e n observações destas variáveis o modelo de regressão linear múltipla é dado por

$$y_i = \alpha + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \epsilon_i, \quad i = 1, \ldots, n. \tag{7.4}$$

Em palavras,

$$y = \text{combinação linear dos } X's + \text{erro}.$$

Os erros ϵ_i representam desvios (supostos independentes) da relação linear entre y e $x_1, \ldots, x_k$ e assume-se que $\epsilon_i \sim N(0, \sigma^2)$. Equivalentemente,

$$y_i | x_{i1}, \ldots, x_{ik} \sim N(\alpha + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} \ , \ \sigma^2).$$

Aqui cada coeficiente β_j representa o efeito de x_j sobre y quando todas as outras variáveis são mantidas constantes. Neste caso temos $k + 2$ parâmetros a serem estimados.

Exemplo 7.6 : Em um problema de regressão com uma variável resposta y e 3 variáveis explicativas x_1, x_2, x_3 podemos investigar o grau de associação entre cada par de variáveis através de gráficos de dispersão como na Figura 7.6. Parece haver alguma associação linear entre y e cada uma das variáveis explicativas, e um modelo de regressão linear múltipla levará em conta todas estas correlações simultaneamente.

Exemplo 7.7 : Um fabricante de borrachas (de apagar lápis) tem interesse em determinar a perda de abrasividade após certo tempo de uso, porém esta variável é muito cara de ser medida diretamente. Uma saída é tentar medi-la indiretamente a partir de outras variáveis e para isto foi coletada uma amostra de 30 borrachas aonde foram medidas as variáveis Perda de abrasividade, Dureza e Resistência à tensão. Os dados estão disponíveis em `http://www.stats.bris.ac.uk/ peter/Teach/LM`. O grau de associação entre as variáveis pode ser investigado através das Figuras 7.7 e 7.8.

Para usar a representação matricial em regressão múltipla, i.e. $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\theta} + \boldsymbol{\epsilon}$ definimos

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \quad \boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1k} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{bmatrix} \quad \boldsymbol{\theta} = \begin{bmatrix} \alpha \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} \quad \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}.$$

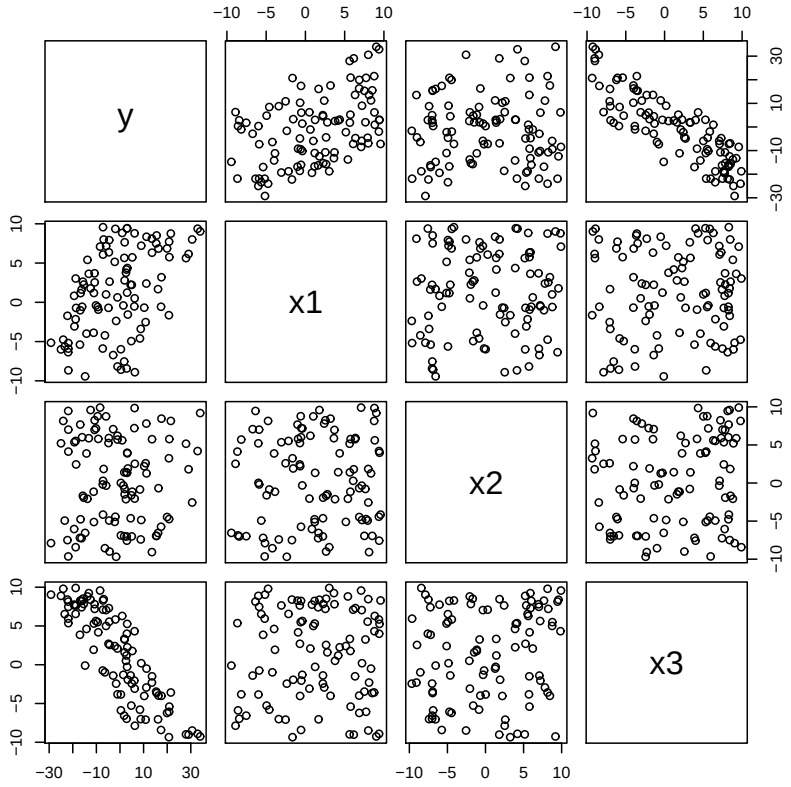

Figura 7.6: Investigando associação entre as 4 variáveis do Exemplo 7.6.

O elemento x_{ij} da matriz $\boldsymbol{X}$ representa a i-ésima observação da variável x_j e queremos estimar os elementos do vetor $\boldsymbol{\theta}$.

Pode-se mostrar que as estimativas dos coeficientes da regressão são dadas por

$$\hat{\boldsymbol{\theta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$$

sendo $\boldsymbol{X}'$ a transposta da matrix $\boldsymbol{X}$. Os valores ajustados da variável resposta são $\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\theta}}$ e uma estimativa de σ^2 é dada por

$$\hat{\sigma}^2 = \frac{1}{n-k-1} \sum_{i=1}^{n} (y_i - \hat{\alpha} - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_k x_{ik})^2.$$

Para construir intervalos de confiança e testar hipóteses sobre os coeficientes usa-se novamente a distribuição t. Neste caso pode-se mostrar que a estatística

$$\frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}\sqrt{h_j}},$$

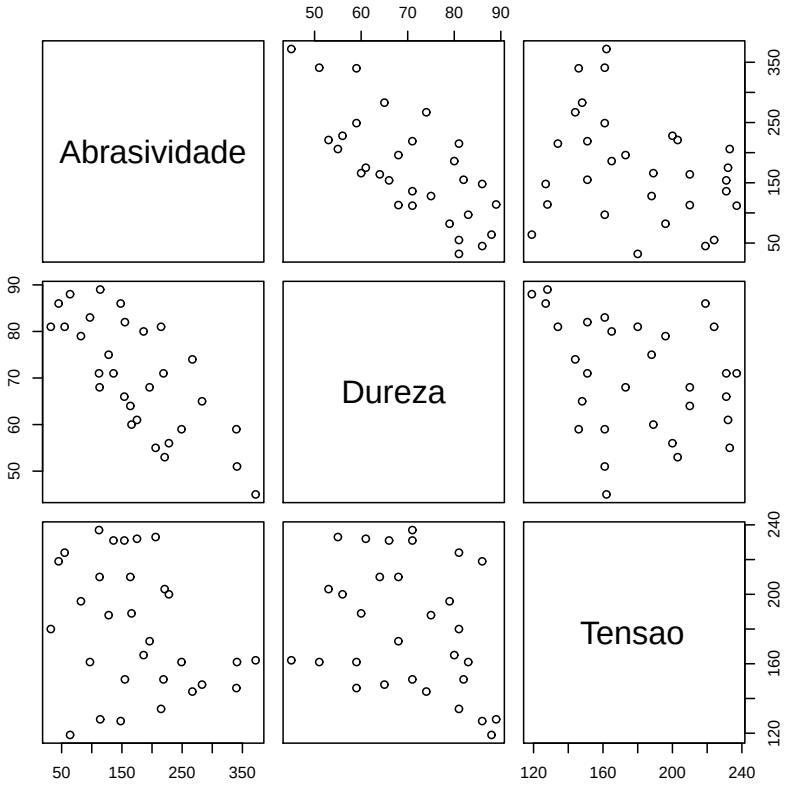

Figura 7.7: Investigando associação entre as 3 variáveis do Exemplo 7.7.

sendo que h_j é o elemento j na diagonal da matriz $(\boldsymbol{X}'\boldsymbol{X})^{-1}$, tem distribuição t com $n-k-1$ graus de liberdade. Assim, um I.C. para β_j fica

$$\hat{\beta}_j \pm t_{\frac{\alpha}{2}} \hat{\sigma} \sqrt{h_j}.$$

Em geral as hipóteses a serem testadas são do tipo

$$H_0 : \beta_j = 0$$
$$H_1 : \beta_j \neq 0$$

ou seja queremos testar se x_j não explica significativamente a variabilidade em y.

Após estimar o modelo gostariamos de ter uma idéia sobre qual proporção da variabilidade em y está sendo explicada pelas outras variáveis. Esta variabilidade

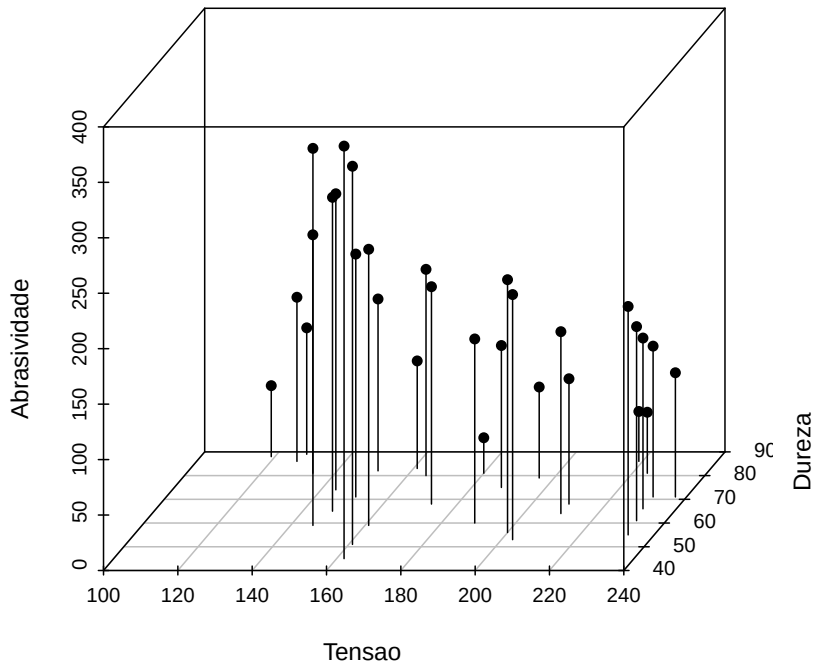

Figura 7.8: Associação entre as 3 variáveis do Exemplo 7.7 em perspectiva.

pode ser particionada da seguinte forma,

$$\underbrace{\sum_{i=1}^{n}(y_i-\bar{y})^2}_{SQT}=\underbrace{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}_{SQReg}+\underbrace{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}_{SQR}$$

sendo

- SQT: a soma de quadrados total (a variabilidade total em y).
- SQReg: a soma de quadrados da regressão (a variabilidade em y induzida pelas variáveis regressoras).
- SQR: a soma de quadrados residual (a variabilidade em y não induzida pelas variáveis regressoras).

O ajuste será tanto melhor quanto mais próximo a SQReg estiver da SQT, ou equivalentemente quanto menor for a SQR. Uma forma de medir isto é através

do chamado *coeficiente de correlação múltipla* denotado por R^2 e definido como

$$R^2 = \frac{SQReg}{SQT} = 1 - \frac{SQR}{SQT} = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2}$$

sendo que $0 \leq R^2 \leq 1$. Quanto mais próximo de 1 melhor é o ajuste do modelo.

Exemplo 7.8 : Para um conjunto de 100 observações foi ajustando um modelo de regressão usando um pacote estatístico e obteve-se os resultados abaixo.

```
           Estimativa EP   estatistica t p-valor
Intercepto 0.19       0.09  2.11         0.03746
x1        -1.51       0.51 -2.96         0.00387
x2         2.23       1.21  1.84         0.06842
x3        -1.25       1.01 -1.24         0.218
sigma: 0.9695 com 96 graus de liberdade
correlação multipla: 0.68
```

Note que os coeficientes de x_2 e x_3 não são significativos ao nível de 5% já que os p-valores correspondentes são maiores do que 0,05. Ou seja existe evidência nos dados de que $\beta_2 = 0$ e $\beta_3 = 0$. O coeficiente de correlação múltipla também é muito pequeno (0,68) indicando que em torno de 32% da variabilidade em y é explicada por outras variáveis que não entraram no modelo.

Deve-se ter um certo cuidado na interpretação do R^2 uma vez que é sempre possível aumentar o seu valor acrescentando-se mais variáveis regressoras ao modelo. Uma forma de corrigir isto é calcular o R^2 ajustado,

$$R^2\text{ajustado} = 1 - \frac{(1 - R^2)(n-1)}{n-k}.$$

Este valor não necessariamente aumentará com a inclusão de mais regressoras já que isto aumentará o valor de k.

O Teste F

Suponha agora que queremos testar a hipótese mais geral de que não existe qualquer relação linear entre a variável dependente e as regressoras no seu modelo. Este teste pode ser formulado como

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$$
$$H_1 : \text{ao menos um coeficiente é não nulo.}$$

ou seja os coeficientes de todas as regressoras são conjuntamente iguais a zero. A estatística de teste neste caso é

$$F = \frac{SQReg/k}{SQR/(n-k+1)}.$$

Quando H_0 é verdadeira esta estatística tem distribuição F com k e $n-k+1$ graus de liberdade no numerador e denominador. Se H_0 for falsa então espera-se que SQReg > SQR e portanto valores grandes de F indicam evidência contra H_0. Assim, o teste é do tipo unilateral.

Também não é difícil verificar a relação da estatística F com o R^2 já que

$$SQReg = R^2 SQT \quad \text{e} \quad SQR = (1-R^2)SQT.$$

Portanto,

$$F = \frac{n-k+1}{k} \frac{R^2}{1-R^2}.$$

Exemplo 7.9 : No Exemplo 7.8 temos que $n = 100$, $k = 3$ e $R^2 = 0,68$. A estatística F então fica

$$F = \frac{100-3+1}{3} \frac{0,68}{0,32} = 69,41667$$

e comparando com o valor tabelado para o

Efeito de Interação

Considere o seguinte modelo de regressão linear com duas variáveis regressoras

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \epsilon.$$

Neste modelo, $x_1 x_2$ representa a *interação* entre as variáveis independentes x_1 e x_2. Se a interação é significativa, i.e. rejeitamos a hipótese $\beta_3 = 0$, então, o efeito de x_1 na resposta média depende do nível de x_2 e, analogamente, o efeito de x_2 na resposta média depende do nível de x_1.

A interpretação dos coeficientes da regressão agora fica

- Quando x_2 é mantida constante, a cada mudança de uma unidade em x_1, a mudança na resposta média será $\beta_1 + \beta_3 x_2$.

- Quando x_1 é mantida constante, a cada mudança de uma unidade em x_2, a mudança na resposta média será $\beta_2 + \beta_3 x_1$.

7.6 Problemas

1. Comente os resultados na Figura 7.6.

2. Comente os resultados na Figura 7.7.

3. No Exemplo 7.7, foi estimado um modelo de regressão linear para a Abrasividade tendo Dureza e Resistência à tensão como regressoras. Comente os resultados obtidos abaixo.

```
           Estimativa    E.P. Estatistica t  p-valor
Intercepto   885.1611 61.7516        14.334 3.84e-14
Dureza        -6.5708  0.5832       -11.267 1.03e-11
Tensao        -1.3743  0.1943        -7.073 1.32e-07

E.P. residual: 36.49 com 27 g.l.
R-Quadrado 0.8402
Estatistica F: 71 com 2 e 27 g.l.,  p-valor: 1.767e-11
```

4. Em um conjunto de dados econômicos para 50 países temos os valores médios para o período 1960-1970 das seguintes variáveis: Renda per capita (Renda), Taxa de crescimento da renda per capita (Taxa), Poupança agregada dividida pela renda disponível (PoupR), percentual da população abaixo dos 15 (Pop15) e acima dos 75 anos (Pop75). Interprete o resultado abaixo de um modelo de regressão linear tendo a variável PoupR como resposta e as outras como regressoras. Estes dados estão disponiveis em `http://www.maths.bath.ac.uk/ jjf23/LMR`.

```
           Estimativa E.P. Estatistica t p-valor
Intercepto 28.566     7.35         3.884 0.000334
Pop15      -0.461     0.14        -3.189 0.002603
Pop75      -1.691     1.08        -1.561 0.125530
Renda      -0.000     0.00        -0.362 0.719173
Taxa        0.409     0.19         2.088 0.042471

E.P. residual: 3.803 com 45 g.l.
R-Quadrado: 0.3385
Estatistica F: 5.76 com 4 e 45 g.l.,  p-valor: 0.0007904
```

5. Escreva em notação matricial os seguintes modelos

   (a) $y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \epsilon_i, \;\; i = 1, \ldots, n.$

(b) $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i2} + \epsilon_i, \;\; i = 1, \ldots, n.$

(c) $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 \log(x_{i2}) + \beta_3 x_{i1} \text{sen} x_{i2} + \epsilon_i, \;\; i = 1, \ldots, n.$

(d) $y_{ij} = \alpha_i + \epsilon_{ij}, \;\; i = 1, 2$ e $j = 1, \ldots, n_i.$

6. No itens (b) e (c) do Exercício 5 qual o efeito de um aumento de 1 unidade em x_1 sobre a resposta média?

7. Explique intuitivamente por que a inclusão de variáveis regressoras no modelo aumenta o valor de R^2.

8. No modelo de regressão $\log(y_i) = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i$, $i = 1, \ldots, n$, qual o efeito esperado sobre y quando x_1 aumenta de 2 unidades.

Apêndice A

Lista de Distribuições

Neste apêndice são listadas as distribuições de probabilidade utilizadas no texto para facilidade de referência. São apresentadas suas funções de (densidade) de probabilidade além da média e variância. Uma revisão exaustiva de distribuições de probabilidades pode ser encontrada em Johnson et al. (1992, 1995) e Evans et al. (1993).

A.1 Distribuição Normal

X tem distribuição normal com parâmetros μ e σ^2, denotando-se $X \sim N(\mu, \sigma^2)$, se sua função de densidade é dada por

$$p(x|\mu, \sigma^2) = (2\pi\sigma^2)^{-1/2} \exp[-(x-\mu)^2/2\sigma^2], \quad -\infty < x < \infty,$$

para $-\infty < \mu < \infty$ e $\sigma^2 > 0$. Quando $\mu = 0$ e $\sigma^2 = 1$ a distribuição é chamada normal padrão. A distribuição log-normal é definida como a distribuição de e^X.

No caso vetorial, $\boldsymbol{X} = (X_1, \ldots, X_p)$ tem distribuição normal multivariada com vetor de médias $\boldsymbol{\mu}$ e matriz de variância-covariância Σ, denotando-se $\boldsymbol{X} \sim N(\boldsymbol{\mu}, \Sigma)$ se sua função de densidade é dada por

$$p(x|\boldsymbol{\mu}, \Sigma) = (2\pi)^{-p/2}|\Sigma|^{-1/2} \exp[-(\boldsymbol{x}-\boldsymbol{\mu})'\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})/2]$$

para $\boldsymbol{\mu} \in \mathbb{R}^p$ e Σ positiva-definida.

A.2 Distribuição Gama

X tem distribuição Gama com parâmetros α e β, denotando-se $X \sim Ga(\alpha, \beta)$, se sua função de densidade é dada por

$$p(x|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}, \quad x > 0,$$

para $\alpha, \beta > 0$.

$$E(X) = \alpha/\beta \quad \text{e} \quad V(X) = \alpha/\beta^2.$$

Casos particulares da distribuição Gama são a distribuição de Erlang, $Ga(\alpha, 1)$, a distribuição exponencial, $Ga(1, \beta)$, e a distribuição qui-quadrado com ν graus de liberdade, $Ga(\nu/2, 1/2)$.

A.3 Distribuição Gama Inversa

X tem distribuição Gama Inversa com parâmetros α e β, denotando-se $X \sim GI(\alpha, \beta)$, se sua função de densidade é dada por

$$p(x|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{-(\alpha+1)} e^{-\beta/x}, \quad x > 0,$$

para $\alpha, \beta > 0$.

$$E(X) = \frac{\beta}{\alpha - 1} \quad \text{e} \quad V(X) = \frac{\beta^2}{(\alpha-1)^2(\alpha-2)}.$$

Não é difícil verificar que esta é a distribuição de $1/X$ quando $X \sim Ga(\alpha, \beta)$.

A.4 Distribuição Beta

X tem distribuição Beta com parâmetros α e β, denotando-se $X \sim Be(\alpha, \beta)$, se sua função de densidade é dada por

$$p(x|\alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}, \quad 0 < x < 1,$$

para $\alpha, \beta > 0$.

$$E(X) = \frac{\alpha}{\alpha + \beta} \quad \text{e} \quad V(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}.$$

A.5 Distribuição de Dirichlet

O vetor aleatório $\boldsymbol{X} = (X_1, \ldots, X_k)$ tem distribuição de Dirichlet com parâmetros $\alpha_1, \ldots, \alpha_k$, denotada por $D_k(\alpha_1, \ldots, \alpha_k)$ se sua função de densidade conjunta é dada por

$$p(\boldsymbol{x}|\alpha_1, \ldots, \alpha_k) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1), \ldots, \Gamma(\alpha_k)} x_1^{\alpha_1 - 1} \ldots x_k^{\alpha_k - 1}, \quad \sum_{i=1}^{k} x_i = 1,$$

para $\alpha_1, \ldots, \alpha_k > 0$ e $\alpha_0 = \sum_{i=1}^{k} \alpha_i$.

$$E(X_i) = \frac{\alpha_i}{\alpha_0}, \quad V(X_i) = \frac{(\alpha_0 - \alpha_i)\alpha_i}{\alpha_0^2(\alpha_0 + 1)}, \quad \text{e} \quad Cov(X_i, X_j) = -\frac{\alpha_i \alpha_j}{\alpha_0^2(\alpha_0 + 1)}$$

Note que a distribuição Beta é obtida como caso particular para $k = 2$.

A.6 Distribuição t de Student

X tem distribuição t de Student (ou simplesmente t) com média μ, parâmetro de escala σ e ν graus de liberdade, denotando-se $X \sim t_\nu(\mu, \sigma^2)$, se sua função de densidade é dada por

$$p(x|\nu, \mu, \sigma^2) = \frac{\Gamma(\frac{\nu+1}{2})\nu^{\nu/2}}{\Gamma(\frac{\nu}{2})\sqrt{\pi}\ \sigma} \left[\nu + \frac{(x-\mu)^2}{\sigma^2}\right]^{-(\nu+1)/2}, \quad x \in \mathbb{R},$$

para $\nu > 0$, $\mu \in \mathbb{R}$ e $\sigma^2 > 0$.

$$E(X) = \mu, \quad \text{para } \nu > 1 \quad \text{e} \quad V(X) = \frac{\nu\sigma^2}{\nu - 2}, \quad \text{para } \nu > 2.$$

Um caso particular da distribuição t é a distribuição de Cauchy, denotada por $C(\mu, \sigma^2)$, que corresponde a $\nu = 1$.

A.7 Distribuição F de Fisher

X tem distribuição F com ν_1 e ν_2 graus de liberdade, denotando-se $X \sim F(\nu_1, \nu_2)$, se sua função de densidade é dada por

$$p(x|\nu_1, \nu_2) = \frac{\Gamma(\frac{\nu_1 + \nu_2}{2})}{\Gamma(\frac{\nu_1}{2})\Gamma(\frac{\nu_2}{2})}\ \nu_1^{\nu_1/2}\nu_2^{\nu_2/2}x^{\nu_1/2-1}(\nu_2 + \nu_1 x)^{-(\nu_1+\nu_2)/2}$$

$x > 0$, e para $\nu_1, \nu_2 > 0$.

$$E(X) = \frac{\nu_2}{\nu_2 - 2}, \quad \text{para } \nu_2 > 2 \quad \text{e} \quad V(X) = \frac{2\nu_2^2(\nu_1 + \nu_2 - 2)}{\nu_1(\nu_2 - 4)(\nu_2 - 2)^2}, \quad \text{para } \nu_2 > 4.$$

A.8 Distribuição Binomial

X tem distribuição binomial com parâmetros n e p, denotando-se $X \sim bin(n, p)$, se sua função de probabilidade é dada por

$$p(x|n, p) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0, \ldots, n$$

para $n \geq 1$ e $0 < p < 1$.

$$E(X) = np \quad \text{e} \quad V(X) = np(1-p)$$

e um caso particular é a distribuição de Bernoulli com $n = 1$.

A.9 Distribuição Multinomial

O vetor aleatório $\boldsymbol{X} = (X_1, \ldots, X_k)$ tem distribuição multinomial com parâmetros n e probabilidades $\theta_1, \ldots, \theta_k$, denotada por $M_k(n, \theta_1, \ldots, \theta_k)$ se sua função de probabilidade conjunta é dada por

$$p(\boldsymbol{x}|\theta_1, \ldots, \theta_k) = \frac{n!}{x_1!, \ldots, x_k!} \theta_1^{x_1}, \ldots, \theta_k^{x_k}, \quad x_i = 0, \ldots, n, \quad \sum_{i=1}^{k} x_i = n,$$

para $0 < \theta_i < 1$ e $\sum_{i=1}^{k} \theta_i = 1$. Note que a distribuição binomial é um caso especial da multinomial quando $k = 2$. Além disso, a distribuição marginal de cada X_i é binomial com parâmetros n e θ_i e

$$E(X_i) = n\theta_i, \quad V(X_i) = n\theta_i(1-\theta_i), \quad \text{e} \quad Cov(X_i, X_j) = -n\theta_i\theta_j.$$

A.10 Distribuição de Poisson

X tem distribuição de Poisson com parâmetro θ, denotando-se $X \sim Poisson(\theta)$, se sua função de probabilidade é dada por

$$p(x|\theta) = \frac{\theta^x e^{-\theta}}{x!}, \quad x = 0, 1, \ldots$$

para $\theta > 0$.

$$E(X) = V(X) = \theta.$$

A.11 Distribuição Binomial Negativa

X tem distribuição de binomial negativa com parâmetros r e p, denotando-se $X \sim BN(r, p)$, se sua função de probabilidade é dada por

$$p(x|r,p) = \binom{r+x-1}{x} p^r (1-p)^x, \quad x = r, r+1, \ldots$$

para $r \geq 1$ e $0 < p < 1$.

$$E(X) = r(1-p)/p \quad \text{e} \quad V(X) = r(1-p)/p^2.$$

Um caso particular é quando $r = 1$ e neste caso diz-se que X tem distribuição geométrica com parâmetro p.

Apêndice B

Propriedades de Algumas Distribuições de Probabilidade

Nos resultados a seguir assume-se que $X_1, \ldots, X_k$ são k variáveis aleatórias independentes.

1. Se $X_i \sim Binomial(n_i, p)$, $i = 1, \ldots, k$. Então

$$Y = \sum_{i=1}^{k} X_i \sim Binomial\left(\sum_{i=1}^{k} n_i, p\right).$$

2. Se $X_i \sim Poisson(\lambda_i)$, $i = 1, \ldots, k$. Então

$$Y = \sum_{i=1}^{k} X_i \sim Poisson\left(\sum_{i=1}^{k} \lambda_i\right).$$

3. Se $X_i \sim Geometrica(p)$, $i = 1, \ldots, k$. Então

$$Y = \sum_{i=1}^{k} X_i \sim Binomial - Negativa(k, p).$$

4. Se $X_i \sim Normal(\mu_i, \sigma_i^2)$, $i = 1, \ldots, k$. Então para constantes $a_1, \ldots, a_k$ e b diferentes de zero,

$$Y = b + \sum_{i=1}^{k} a_i X_i \sim Normal\left(b + \sum_{i=1}^{k} a_i \mu_i, \sum_{i=1}^{k} a_i^2 \sigma_i^2\right).$$

5. Se $X_i \sim Gama(\alpha_i, \beta)$, $i = 1, \ldots, k$. Então

$$Y = \sum_{i=1}^{k} X_i \sim Gama\left(\sum_{i=1}^{k} \alpha_i, \beta\right).$$

6. Se $X_i \sim Exponencial(\beta)$, $i = 1, \ldots, k$. Então

$$Y = \min\{X_i\} \sim Exponencial(k\beta).$$

Apêndice C

Soluções de Exercícios Selecionados

Capítulo 4

Seção 4.4

5. (a) $a = 4$ e $b = 1$

10. $n = 396$

Seção 4.6

3. (a) média $\approx 0,17$, (b) moda $\approx 0,15$

4. média $\approx 2,67$

4. média $\approx 175,86$

Capítulo 5

Seção 5.5

1. (a) Definindo $Y = -\theta \log X$ segue por transformação de variáveis que

$$p(y) = p(x(y))|\frac{dx}{dy}| = \theta[e^{-y/\theta}]^{\theta-1}\frac{e^{-y/\theta}}{\theta} = e^{-y},\ y > 0 \text{ (pois } \log x < 0).$$

Portanto, $Y \sim Exp(1)$ é um pivot.

(b) Se $Y \sim Exp(1)$ então $Y \sim Gama(1,1)$ e portanto $-2\theta \log X \sim \chi^2_2$. Para construir o intervalo pedido obter na tabela da distribuição qui-quadrado

com 2 graus de liberdade c_1 e c_2 tais que

$$P(c_1 < -2\theta \log X < c_2) = 0,90.$$

3. Se $X_1, \ldots, X_n \sim Exp(\theta)$ então $X_i \sim Gama(1, \theta)$, $i = 1, \ldots, n$ e como os X_i's são independentes segue que $\sum_{i=1}^n X_i \sim Gama(n, \theta)$ e portanto $2\theta \sum_{i=1}^n X_i \sim \chi^2_{2n}$. Assim, basta obter as constantes c_1 e c_2 tais que $P(c_1 < 2\theta \sum_{i=1}^n X_i < c_2) = 1 - \alpha$ em uma tabela qui-quadrado com $2n$ graus de liberdade. Isolando θ segue que

$$\frac{c_1}{2\sum_{i=1}^n X_i} < \theta < \frac{c_2}{2\sum_{i=1}^n X_i}$$

10. O I.C. de 99% para a diferença média é [0,2143177; 4,185682]. Com 99% de confiança podemos afirmar que a bebida teve efeito significativo pois em média houve aumento nos escores após ingestão de água.

11. O I.C. de 95% para a diferença média de massas é $[0.117847354321697, 1.6421526456783]$. Com 95% de confiança podemos afirmar que houve ganho de massa já que o intervalo contém somente valores positivos.

Seção 5.8

6. (b) Usando a distribuição a posteriori do item (a) com $\sum_{i=1}^n x_i = 10$ e $n = 10$, segue que $\theta|\boldsymbol{x} \sim Gama(10,5;10)$. Portanto, $20\theta|\boldsymbol{x} \sim \chi^2_{21}$. Da tabela qui-quadrado com 21 graus de liberdade obtemos que,

$$P(20\theta < 10.283) = 0.025 \quad \text{e} \quad P(20\theta > 35.479) = 0.025$$

e segue então que $10.283/20 < \theta < 35.479/20$ com probabilidade 0.95 e o intervalo de credibilidade é $0.51415 < \theta < 1.77395$.

Capítulo 6

Seção 6.1.3

1. (a) A função poder é dada por $\pi(\theta) = P(\max\{X_i\} \leq 1) = P(X_1 \leq 1, \ldots, X_n \leq 1) = \prod_{i=1}^n P(X_i \leq 1) = 1/\theta^n$.

   (b) O tamanho do teste é dado por $\sup_{\theta \geq 2} \pi(\theta) = \pi(2) = 1/2^n$.

Seção 6.4

1. (a) $\sum_{i=1}^{n} X_i > c$, (b) $\sum_{i=1}^{n}(X_i - \mu)^2 > c$, (c) $\prod_{i=1}^{n} X_i > c$, (d) $-\overline{X} > c$.

2. Rejeitar H_0 se $\sum_{i=1}^{n} X_i^2 > 36,62$.

3. Rejeitar H_0 se $\sum_{i=1}^{n} X_i > 31,41$.

5. Teste UMP rejeita H_0 se $\sum \log X_i > c$ ou equivalentemente se $-\sum \log X_i < -c$ sendo que $-\log X_i \sim Exp(\theta)$.

Seção 6.6.4

6 Da Tab. A.7 obtemos $P(Y > 28.71)$=0,01 sendo $Y \sim F(4,3)$ então F_{SUP}=28,71 Da Tab. A.7 obtemos $P(Z > 16.69)$=0,01 sendo $Z \sim F(3,4)$ então F_{INF}=1/16,69 S_A^2/S_B^2= 770/3892 = 0.1978417 e como $0,0599 < 0,197841 < 28,71$ aceita-se H_0 ao nível de 2%.

Referências

Broemeling, L. (1985). *Bayesian Analysis of Linear Models*. New York: Marcel Dekker.

DeGroot, M. H. (1989). *Probability and Statistics* (2nd ed.). Addison Wesley.

DeGroot, M. H. and M. J. Schervish (2002). *Probability and Statistics* (3rd ed.). Addison Wesley.

Evans, M., N. Hastings, and B. Peacock (1993). *Statistical Distributions, Second Edition* (Second ed.). Wiley Interscience.

Johnson, N. L., S. Kotz, and N. Balakrishnan (1995). *Continuous Univariate Distributions* (2nd ed.), Volume 2. John Wiley, New York.

Johnson, N. L., S. Kotz, and A. W. Kemp (1992). *Univariate Discrete Distributions* (2nd ed.). John Wiley, New York.

Lehman, E. and J. P. Romano (2005). *Testing Statistical Hypothesis* (Third ed.). Springer.

Migon, H. S. and D. Gamerman (1999). *Statistical Inference: An Integrated Approach*. Arnold.