5 Estimação

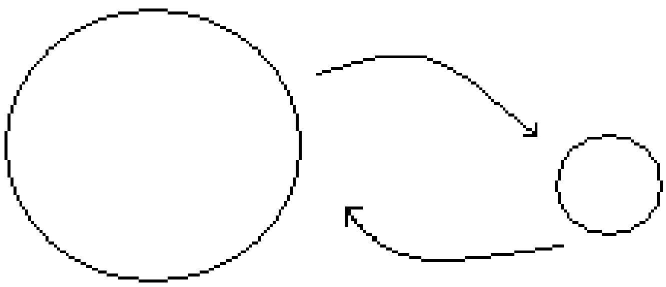
5.1 Introdução

Seja X a v.a. (população) em estudo, cuja função de probabilidade/densidade é representada por $f(x,\theta)$, onde θ é um (ou mais) parâmetro(s) desconhecido(s). Por exemplo (com leis nossas conhecidas):

- $X \sim \mathcal{P}(\lambda)$ $\quad f(x,\lambda) = e^{-\lambda}\dfrac{\lambda^x}{x!},\ \forall x \in \mathbb{N}_0$ $\quad \longrightarrow\ \theta = \lambda = E(X)$
- $X \sim \mathcal{E}(\lambda)$ $\quad f(x,\lambda) = \lambda\, e^{-\lambda x}\mathbb{I}_{[0,+\infty[}(x)$ $\quad \longrightarrow\ \theta = \lambda,\ g(\lambda) = \dfrac{1}{\lambda} = E(X)$
- $X \sim \mathcal{N}(\mu,\sigma)$ $\quad f(x,\mu,\sigma) = \dfrac{1}{\sqrt{2\pi}\sigma}\, e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2},\ \forall x \in \mathbb{R}$ $\quad \longrightarrow\ \overrightarrow{\theta} = (\mu,\sigma) = (E(X), \sqrt{V(X)})$

O nosso problema: Que valor(es) atribuir a θ, ou a uma função de θ, g(θ) ?

População X **Amostra Aleatória** $X_1, X_2, ..., X_n$



$\downarrow$

Inferência Estatística

Parâmetros Populacionais	**Características Amostrais/Estatísticas**
$E(X) = \mu$	$\overline{X} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} X_i$
$V(X) = \sigma^2,\ \sigma(X) = \sigma, ...$	$S_n^2 = \dfrac{1}{n-1}\displaystyle\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2,\ S_n, ...$

5.2 Estimação Pontual

Estimador é qualquer Estatística usada para estimar um parâmetro populacional (ou função desse parâmetro).
Estimativa é um valor do estimador para uma amostra em concreto.

Notação Estimador do parâmetro $\theta \;\rightarrow\; \widehat{\Theta}$

Exemplos

População $X \longrightarrow$	a.a. $X_1, X_2, ..., X_n \longrightarrow$	concretização da a.a. $x_1, x_2, ..., x_n$
$\downarrow$	$\downarrow$	$\downarrow$
(parâmetros desconhecidos)	(Estimador)	(Estimativa)
μ	$\widehat{\mu} = \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$	$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
σ^2	$\widehat{\sigma}^2 = S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$	$s_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$
$\vdots$		

Os valores possíveis de um Estimador (note-se v.a.) são as estimativas obtidas em cada concretização da amostra aleatória.

Algumas propriedades de um (*bom*) Estimador:

- Um estimador $\widehat{\Theta}$ do parâmetro θ diz-se **cêntrico** (ou **centrado** ou **não enviesado**) se $E(\widehat{\Theta}) = \theta$. Por exemplo, $\overline{X}$ é um estimador centrado de $E(X)$!
- Um estimador $\widehat{\Theta}$ do parâmetro θ diz-se **enviesado** se não for cêntrico e o seu **enviesamento** é dado por $viés(\widehat{\Theta}) = E(\widehat{\Theta}) - \theta$.
- A **eficiência** de um estimador cêntrico $\widehat{\Theta}$ é dada por $V(\widehat{\Theta})$.

  Assim, dados dois estimadores cêntricos do parâmetro θ, $\widehat{\Theta}_1$ e $\widehat{\Theta}_2$, $\widehat{\Theta}_1$ diz-se **mais eficiente** do que $\widehat{\Theta}_2$ se $V(\widehat{\Theta}_1) < V(\widehat{\Theta}_2)$.

Entre estimadores não enviesados, é dada preferência ao estimador com menor variância, isto é, ao mais eficiente. Por exemplo, considerem-se duas a.a.'s de uma população X, de tamanhos n e m, respetivamente, com $n > m$. As médias amostrais $\overline{X}_n$ e $\overline{X}_m$ são estimadores cêntricos de $\mu = E(X)$; no entanto, $\overline{X}_n$ é mais eficiente do que $\overline{X}_m$. De facto,

$$V(\overline{X}_n) = \frac{V(X)}{n} \; < \; V(\overline{X}_m) = \frac{V(X)}{m}$$

- A **eficiência** de um estimador $\widehat{\Theta}$ do parâmetro θ é dada pelo seu **erro quadrático médio** (**EQM**), definido por

$$EQM(\,\widehat{\Theta}\,) = E[\,(\widehat{\Theta} - \theta)^2\,] = V(\,\widehat{\Theta}\,) + [\text{viés}(\,\widehat{\Theta}\,)\,]^2\,.$$

Entre estimadores enviesados e não enviesados, é dada preferência ao estimador com menor erro quadrático médio (o mais eficiente). Note-se que um estimador $\widehat{\Theta}_1$ enviesado pode ser mais eficiente do que um estimador cêntrico $\widehat{\Theta}_2$, bastando que $EQM(\,\widehat{\Theta}_1\,) \; < \; V(\widehat{\Theta}_2)$.

5.3 Estimação por Intervalos

Como determinar um intervalo no qual se espera encontrar, com uma certa *confiança*, o valor de um parâmetro populacional θ (ou de $g(\theta)$)?

Para $\alpha \in]0,1[$, um intervalo de confiança $(1-\alpha)\%$ para o parâmetro θ (ou $g(\theta)$) é um **intervalo aleatório** $]L_1, L_2[$, com L_1 e L_2 duas estatísticas amostrais (v.a.'s) tais que

$$P\,(L_1 < \theta < L_2) = 1 - \alpha.$$

(a probabilidade de $]L_1, L_2[$ conter θ é $1 - \alpha$)

- $1 - \alpha$ é a probabilidade do intervalo conter θ (o grau de confiança atribuído ao intervalo);
- $\alpha \in]0,1[$ é a probabilidade do intervalo não conter θ (valor preferencialmente *pequeno*).

Como calcular L_1 e L_2 (os limites do intervalo)?

Método da Variável Fulcral

Seja X uma v.a. cuja distribuição contém um parâmetro θ desconhecido. Pretende-se um intervalo de confiança $(1 - \alpha)\%$ para θ (ou para g(θ)). Considere-se uma amostra aleatória $X_1, X_2, ..., X_n$, de X. Uma Variável Fulcral é uma função da amostra aleatória $X_1, X_2, ..., X_n$ e do parâmetro θ, mas com distribuição (exata ou aproximada) independente de θ.

Passos do método:

1- Escolher a Variável Fulcral Z_n e sua lei (tabelas da disciplina)

2- Determinar $z_1,\ z_2\ \in \mathbb{R}$: $P\left(z_1 < Z_n < z_2\right) = 1-\alpha$

Nota: $z_1:\ P\left(Z_n < z_1\right) = \dfrac{\alpha}{2} \quad \wedge \quad z_2:\ \ P\left(Z_n < z_2\right) = 1 - \dfrac{\alpha}{2}$

3- Com z_1 e z_2 conhecidos, encontrar

$$L_1 \equiv L_1(X_1, X_2, ..., X_n) \ \text{ e } \ L_2 \equiv L_2(X_1, X_2, ..., X_n)$$

tais que

$$P\left(z_1 < Z_n < z_2\right) = 1-\alpha \Leftrightarrow P\left(L_1 < \theta < L_2\right) = 1-\alpha$$

$$(\text{ probabilidade } 1-\alpha \text{ de }]L_1, L_2[\text{ conter } \theta)$$

4- Para uma amostra particular $x_1, x_2, ..., x_n$, determinar estimativas para L_1 e L_2, respetivamente,

$$l_1 \equiv l_1(x_1, x_2, ..., x_n) \ \text{ e } \ l_2 \equiv l_2(x_1, x_2, ..., x_n)$$

Obtém-se, assim, uma estimativa para o intervalo de θ, calculado com confiança $(1-\alpha)\%$,

$$IC_\theta =]l_1, l_2[$$

Nota: O intervalo $]L_1, L_2[$ é aleatório, com probabilidade $1-\alpha$ de conter θ; isto é, se forem recolhidas amostras em número bastante elevado, espera-se que $(1-\alpha)\%$ dos intervalos estimados com base nessas amostras contenha θ; o intervalo $]l_1, l_2[$ é apenas uma estimativa, calculado com confiança $(1-\alpha)\%$, mas ao qual não se pode atribuir uma "probabilidade" de conter θ.

Um exemplo

Intervalo de confiança $(1-\alpha)\%$ para o valor médio de uma população normal com σ conhecido:

Seja $(X_1, X_2, \ldots, X_n)$ uma a.a. de $X \sim \mathcal{N}(\mu, \sigma)$, sendo σ conhecido. Vimos que um estimador para a média μ da população é dado pela estatística $\overline{X}$. Neste caso, segue-se que

$$Z_n = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

Fixado o valor de α, e designando por $z_{\alpha/2}$ o valor tal que $P(Z > z_{\alpha/2}) = \alpha/2$, tem-se

$$P\big(-z_{\alpha/2} < Z_n < z_{\alpha/2}\big) = 1-\alpha \quad \Leftrightarrow \quad P\Big(-z_{\alpha/2} < \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} < z_{\alpha/2}\Big) = 1-\alpha$$

$$-z_{\alpha/2} < \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} < z_{\alpha/2} \quad \Leftrightarrow \quad \underbrace{\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}}_{L_1} < \mu < \underbrace{\overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}}_{L_2}$$

Intervalo (aleatório) a $(1-\alpha)\%$ de confiança para μ:

$$IAC_{(1-\alpha)\%}(\mu) = \Big]\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\Big[$$

Para uma realização da a.a., ou seja, uma amostra concreta $(x_1, x_2, \ldots, x_n)$, e designando por $\overline{x}$ o valor concreto da estatística $\overline{X}$, obtém-se o intervalo (concreto) para o parâmetro μ:

$$IC(\mu) = \left]\overline{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\ ,\ \overline{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right[$$

Exercícios:

1- O conteúdo médio de nicotina de uma amostra de 10 cigarros de uma certa marca é de 1 mg (miligramas). O laboratório sabe, pela longa experiência neste tipo de análise, que o conteúdo de nicotina é uma v.a. com distribuição normal de desvio padrão 0.15 mg.

(a) Determine um intervalo de confiança 90% para o conteúdo médio de nicotina.

(b) Com a mesma confiança, quantos cigarros devem de ser analisados de modo a que o erro máximo cometido na estimação não ultrapasse 0.01?

População: X = conteúdo de nicotina (em mg) num cigarro de uma certa marca; $X \sim N(\mu, 0.15)$.

Amostra: $n = 10; \overline{x} = 1$.

(a) Variável Fulcral e Lei: $Z_n = \dfrac{\overline{X} - \mu}{0.15/\sqrt{n}} \sim N(0,1)$

$z : P\big(-z < Z_n < z\big) = 0.9 \Leftrightarrow P\big(Z_n < z\big) = 0.95 \Leftrightarrow z = 1.645$

$$P\big(-1.645 < Z_n < 1.645\big) = 0.9 \quad \Leftrightarrow \quad P\Big(-1.645 < \frac{\overline{X} - \mu}{0.15/\sqrt{n}} < 1.645\Big) = 0.9$$

$$-1.645 < \frac{\overline{X} - \mu}{0.15/\sqrt{n}} < 1.645 \quad \Leftrightarrow \quad \overline{X} - 1.645\frac{0.15}{\sqrt{n}} < \mu < \overline{X} + 1.645\frac{0.15}{\sqrt{n}}$$

Intervalo (aleatório) a 90% de confiança para μ:

$$IAC_{90\%}(\mu) = \left]\overline{X} - 1.645\frac{0.15}{\sqrt{n}}, \overline{X} + 1.645\frac{0.15}{\sqrt{n}}\right[$$

Concretização:

$$IC(\mu) = \left]1 - 1.645\frac{0.15}{\sqrt{10}},\ 1 + 1.645\frac{0.15}{\sqrt{10}}\right[=]0.922,\ 1.078[\ \ (mg)$$

(b) $1.645\frac{0.15}{\sqrt{n}} \leq 0.01 \Leftrightarrow n \geq 609$

2- A resistência das cordas produzidas por determinada fábrica tem distribuição normal. A fábrica testou uma amostra aleatória de 51 cordas, tendo-se obtido uma resistência média de 300 Kg e desvio padrão 24 Kg.

(a) Determine um intervalo de confiança 95% para a resistência média deste tipo de cordas.

(b) Determine um intervalo de confiança 90% para o desvio padrão.