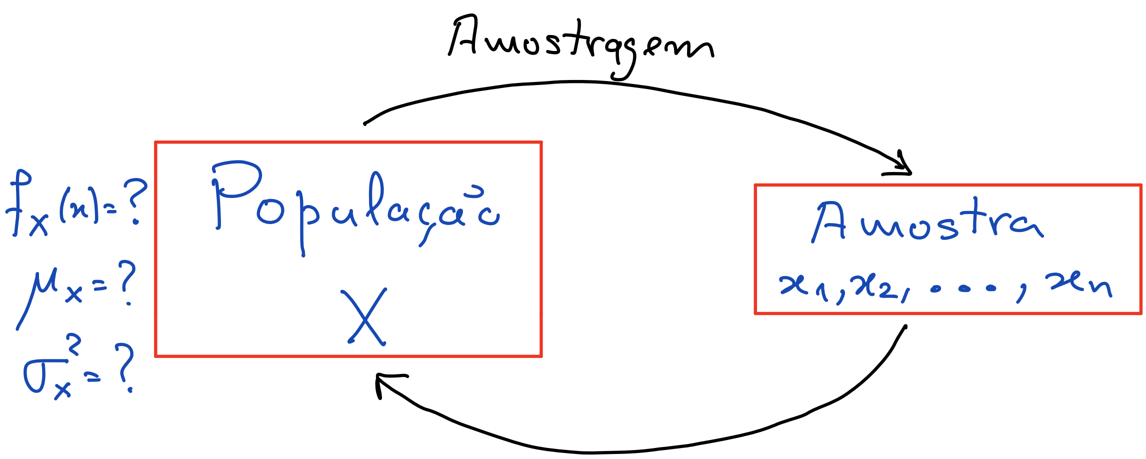
• Inferência Estatística



Estatística Descritiva

$\bar{x}, s^2, \ldots$

Inferência estatística

(ou estatística indutiva)

Dois níveis de "ignorância":

1) F_X / f_X são desconhecidas, sabendo-se apenas que X é do tipo discreto ou contínuo.

2) Sabe-se que F_X / f_X pertencem a det. família (e.g. Poisson ou normal) mas com parâmetros desconhecidos (e.g. λ ou (μ, σ^2)).

Objetivos da Inferência Estatística

- 1) Estimar F_X / f_X ou
  2) estimar parâmetros

- Fazer testes relativos ao tipo de F_X / f_X ou ao valor dos parâmetros.

Estimação de parâmetros:

- pontual → Cap. 6
- intervalar → Cap. 7

Testes de hipóteses:

- sobre parâmetros
- sobre forma de F_X / f_X

→ Cap. 8

- Amostragem aleatória: cada elemento da amostra é obtido totalmente ao acaso na população X e de forma independente dos outros elementos.

Def.: As v.a.'s $(X_1, \ldots, X_n)$ constituem uma amostra aleatória (a.a.) de dimensão n da população X se forem independentes e identicamente distribuídas a X (i.i.d.).

$$\begin{array}{ccccc} X \longrightarrow & X_1 & X_2 & \cdots & X_n \\ & \downarrow & \downarrow & & \downarrow \\ & x_1 & x_2 & & x_n \end{array}$$

n v.a.'s i.i.d. a X (amostra aleatória)

n observações (amostra casual)

Caracterização da a.a.:

$$f_{X_1, \ldots, X_n}(x_1, \ldots, x_n) = \prod_{i=1}^{n} f_{X_i}(x_i)$$

- Estatísticas

Def.: Uma estatística é uma v.a. que é função da a.a. e que não depende de parâmetros desconhecidos. Denota-se usualm/ por $T_n = T(X_1, \ldots, X_n)$.

Exemplos

1) Média amostral: $\bar{X} = \dfrac{X_1 + \cdots + X_n}{n} = \dfrac{\sum_{i=1}^{n} X_i}{n}$

$\bar{X}$ é uma v.a.. Podemos por exemplo calcular $E(\bar{X})$ e $V(\bar{X})$:

$$E(\bar{X}) = \frac{\sum_{i=1}^{n} E(X_i)}{n} \underset{\uparrow \atop \text{id. dist.}}{=} \frac{\cancel{n} \cdot \mu}{\cancel{n}} = \mu = E(X)$$

$$V(\bar{X}) = V\left(\frac{\sum_{i=1}^{n} X_i}{n}\right) \underset{\uparrow \atop \text{ind.}}{=} \frac{1}{n^2} \sum_{i=1}^{n} V(X_i) \underset{\uparrow \atop \text{id. dist.}}{=} \frac{1}{n^2} \cdot n\sigma^2 = \frac{\sigma^2}{n}$$

ind. + id. dist.

Dada uma amostra $(x_1, \ldots, x_n)$, o valor de $\bar{X}$ é $\bar{x} = \dfrac{x_1 + \cdots + x_n}{n}$.

2) Variância amostral: $S^2 = \dfrac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n-1} =$

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

S^2 é uma v.a..

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

$$= \frac{1}{n-1}\left(\sum_{i=1}^{n}\left(V(X_i) + E^2(X_i)\right) - n\left(V(\overline{X}) + E^2(\overline{X})\right)\right)$$

$$= \frac{1}{n-1}\left(n\sigma^2 + \cancel{n\mu^2} - \cancel{n}\frac{\sigma^2}{\cancel{n}} - \cancel{n\mu^2}\right)$$

$$= \frac{1}{n-1}\left((n-1)\sigma^2\right) = \sigma^2 = V(X)$$

3) Mínimo da a.a.: $X_{(1)} = \min\{X_1, \ldots, X_n\}$.

É uma v.a..

4) Máximo da a.a.: $X_{(n)} = \max\{X_1, \ldots, X_n\}$.

É uma v.a..

6 Estimação Pontual

A.a. $(X_1, \ldots, X_n)$ de uma população X cuja função massa/densidade prob. f_X pertence a uma família $\mathcal{F} = \{f(x;\theta), \theta \in \Theta\}$.

Desconhece-se apenas o valor do parâmetro θ que caracteriza X. Como usar a amostra para o estimar?

Def.: Uma estatística T_n diz-se um estimador de θ caso tome valores exclusivamente no espaço paramétrico Θ.
Uma estimativa para θ é o valor concreto do estimador numa amostra particular $(x_1, \dots, x_n)$, i.e. $t_n = T_n(x_1, \dots, x_n)$.

Como encontrar bons estimadores?

6.1 Método da máxima verosimilhança

Ideia: θ desconhecido, $(x_1, \dots, x_n)$ conhecidos.
Estimar θ de forma a maximizar a massa/densidade de probabilidade de ocorrência de $(x_1, \dots, x_n)$.

Exemplo: $X \sim \text{Bernoulli}(p)$, $P(X=1) = p$, $P(X=0) = 1-p$, $p = ?$

A. a.: $(X_1, \dots, X_{10})$, $X_i \sim Ber(p)$, i.i.d.

Amostra particular: $x_1 = x_4 = x_7 = x_8 = x_9 = 0$

$x_2 = x_3 = x_5 = x_6 = x_{10} = 1$

Função de verosimilhança desta amostra:

$$L(p \mid x_1, \dots, x_{10}) = f_X(x_1, p) \times \dots \times f_X(x_{10}, p)$$
$$= p^5 \times (1-p)^5, \quad 0 < p < 1$$

Qual o valor de p que maximiza L?

$$\frac{dL}{dp} = \frac{d\left(p^5(1-p)^5\right)}{dp} = 5p^4(1-p)^5 - 5p^5(1-p)^4$$

$$= 5p^4(1-p)^4(1-2p) = 0$$

$\Rightarrow p = 0 \ \vee \ p = 1 \ \vee \ p = 1/2$

p	0		$1/2$		1
L'	0	$+$	0	$-$	0
L	0	$\nearrow$	Máx	$\searrow$	0

$\Rightarrow$ estimativa de máxima verosimilhança (m.v.) de p com base nesta amostra é $\hat{p} = 1/2$.

Def.: X - v.a. com função massa/densidade de probabilidade $f(x, \theta)$, onde θ é um parâmetro desconhecido.

$(x_1, \ldots, x_n)$ - valores observados de uma a.a. de dimensão n.

A função de verosimilhança da amostra é

$$L(\theta | x_1, \ldots, x_n) = \prod_{i=1}^{n} f(x_i, \theta)$$

A estimativa de máxima verosimilhança de θ é o valor $\hat{\theta}$ que maximiza L, i.e. tal que

$$\boxed{L(\hat{\theta} | x_1, \ldots, x_n) = \max_{\theta} L(\theta | x_1, \ldots, x_n)}$$

Exemplo $X \sim Ber(p)$, $f_X(x) = P(X=x) = p^x (1-p)^{1-x}$, $0<p<1$

Dada uma amostra $(x_1, \dots, x_n)$, temos que

$$L(p) \equiv L(p|x_1,\dots,x_n) = f_X(x_1) \times \dots \times f_X(x_n) =$$

$$= \prod_{i=1}^{n} p^{x_i} (1-p)^{1-x_i} = p^{\sum_{i=1}^{n} x_i} (1-p)^{\sum_{i=1}^{n}(1-x_i)} =$$

$$= p^k (1-p)^{n-k}, \quad \text{com } k = \sum_{i=1}^{n} x_i = \text{nº de sucessos na amostra.}$$

Observação importante: para uma função $f > 0$ qualquer, f e $\log f$ têm máximo e mínimo nos mesmos pontos (pq log é estritam/ crescente). Assim, em vez de determinar p que maximiza L pode-se determinar p que maximiza $\log L$, o que envolve cálculos mais simples. A função $\log L$ é designada por log-verosimilhança.

- $\log L(p) = \log\left(p^k (1-p)^{n-k}\right) = k \log p + (n-k)\log(1-p)$

$$\Rightarrow \frac{d \log L}{dp}(p) = \frac{k}{p} - \frac{n-k}{1-p} = \frac{k(1-p)-(n-k)p}{p(1-p)} = 0$$

$$\overset{p\neq 0,1}{\Longrightarrow} \quad k - np = 0 \Rightarrow p = \frac{k}{n} = \frac{\sum_{i=1}^{n} x_i}{n} = \bar{x}$$

- $\dfrac{d^2 \log L}{dp^2}(p) = \dfrac{-k}{p^2} - \dfrac{n-k}{(1-p)^2} < 0, \quad \forall\, 0<p<1$

máximo

$$\Longrightarrow \boxed{\begin{array}{l} \hat{p} = \dfrac{\sum_{i=1}^{n} x_i}{n} = \bar{x} = \text{estimativa de m.v.} \\ \qquad\qquad\qquad\qquad\qquad\qquad \text{(valor numérico)} \\ EMV(p) = \dfrac{\sum_{i=1}^{n} X_i}{n} = \bar{X} = \text{estimador de m.v.} \\ \qquad\qquad\qquad\qquad\qquad\qquad \text{(variável aleatória)} \end{array}}$$