# 6 Amostragem e estimação pontual

## 6.1 Amostragem aleatória

### Estatística

Ciência que estuda a

- recolha
- organização e apresentação
- análise e interpretação

de **dados**.

#### Que dados são estes?

**Definição**

Uma **população** é um conjunto de entidades que têm pelo menos uma característica (*variável estatística*) em comum.

**Definição**

Uma **amostra** é um conjunto retirado de uma população com o objetivo de a representar.

#### O que fazer com os dados?

**Estatística descritiva**

Descrição dos dados recorrendo a:

- gráficos
- tabelas
- medidas numéricas

**Estatística inferencial**

Uso de dados para aprender algo sobre a população que se admite que representam.

Baseia-se na **Teoria da Probabilidade**.

### Objectivo

Estudo de uma variável estatística numa população

⇕ **Estatística paramétrica** ⇕

Estudo de uma variável aleatória $X$ com distribuição conhecida, $f_X(x;\theta)$, mas em que $\theta \in \Theta \subset \mathbb{R}^k$ tem um valor desconhecido.

### Nota

$\theta$ é um vector de **parâmetros** e $\Theta$ é o **espaço paramétrico** (conjunto de valores possíveis para $\theta$).

### Definição

Uma **amostra aleatória simples** é uma sucessão de variáveis aleatórias $\mathbf{X} = (X_1, \ldots, X_n)$ independentes e identicamente distribuídas de acordo com a distribuição que se admite para a variável estatística na população.

### Nota

Cada amostra observada é uma concretização particular da amostra aleatória!

$x_1, \ldots, x_n$ – amostra observada de dimensão $n$

$x_{(1)}, \ldots, x_{(n)}$ – amostra observada ordenada por ordem crescente

### A informação amostral

1. os valores da amostra observada;
2. a **distribuição da amostra aleatória**:

$$f_{\mathbf{X}}(x_1, \ldots, x_n \mid \theta) = \prod_{i=1}^{n} f_{X_i}(x_i \mid \theta) = \prod_{i=1}^{n} f_X(x_i \mid \theta)$$

## 6.2 Estimação pontual

A informação amostral é resumida através do uso de estatísticas.

### Definição

Uma **estatística** é uma função da amostra aleatória

$$T = T(X_1, \ldots, X_n) = T(\mathbf{X})$$

### Nota

Se a distribuição de $T(\mathbf{X})$ depende de $\theta$ então os valores observados da estatística devem conter informação sobre o valor de $\theta$ na população.

### Exemplo – Algumas estatísticas

1. $\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$ – média amostral;
2. $S_n^2 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n}$ – variância amostral;
3. $S^2 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n-1}$ – variância amostral (corrigida);
4. $X_{(1)} = \min\{X_1, \ldots, X_n\}$ – mínimo amostral;
5. $X_{(n)} = \max\{X_1, \ldots, X_n\}$ – máximo amostral.

**Estimação pontual** $\equiv$ cálculo, a partir de uma amostra observada, de valores plausíveis para os parâmetros da distribuição da variável estatística de interesse numa população.

### Definição

Um **estimador pontual** de um parâmetro $\theta$ é uma qualquer função da amostra aleatória que não dependa de parâmetros cujo valor seja desconhecido (estatística).

### Definição

Uma **estimativa pontual** de um parâmetro $\theta$ é um valor observado de um estimador pontual desse parâmetro.

### Exemplo – Estimação numa população Normal

Seja $X_1, \ldots, X_n$ uma amostra aleatória de uma população com distribuição $N(\mu, \sigma^2)$.

Alguns estimadores possíveis para $\mu$ são:

1. $\bar{X}$ – média amostral
2. mediana amostral
3. moda amostral
4. $\frac{X_{(1)}+X_{(n)}}{2}$ – centro do intervalo de variação amostral
5. . . .

## Como avaliar estimadores?

Há um grande número de critérios para avaliar e comparar estimadores de acordo com diferentes propriedades desejáveis.

Podemos avaliar a **exatidão** de um estimador recorrendo ao seu valor esperado.

### Definição

Um estimador pontual $T$ de um parâmetro $\theta$ diz-se **centrado** se e só se

$$E[T] = \theta, \forall \theta \in \Theta.$$

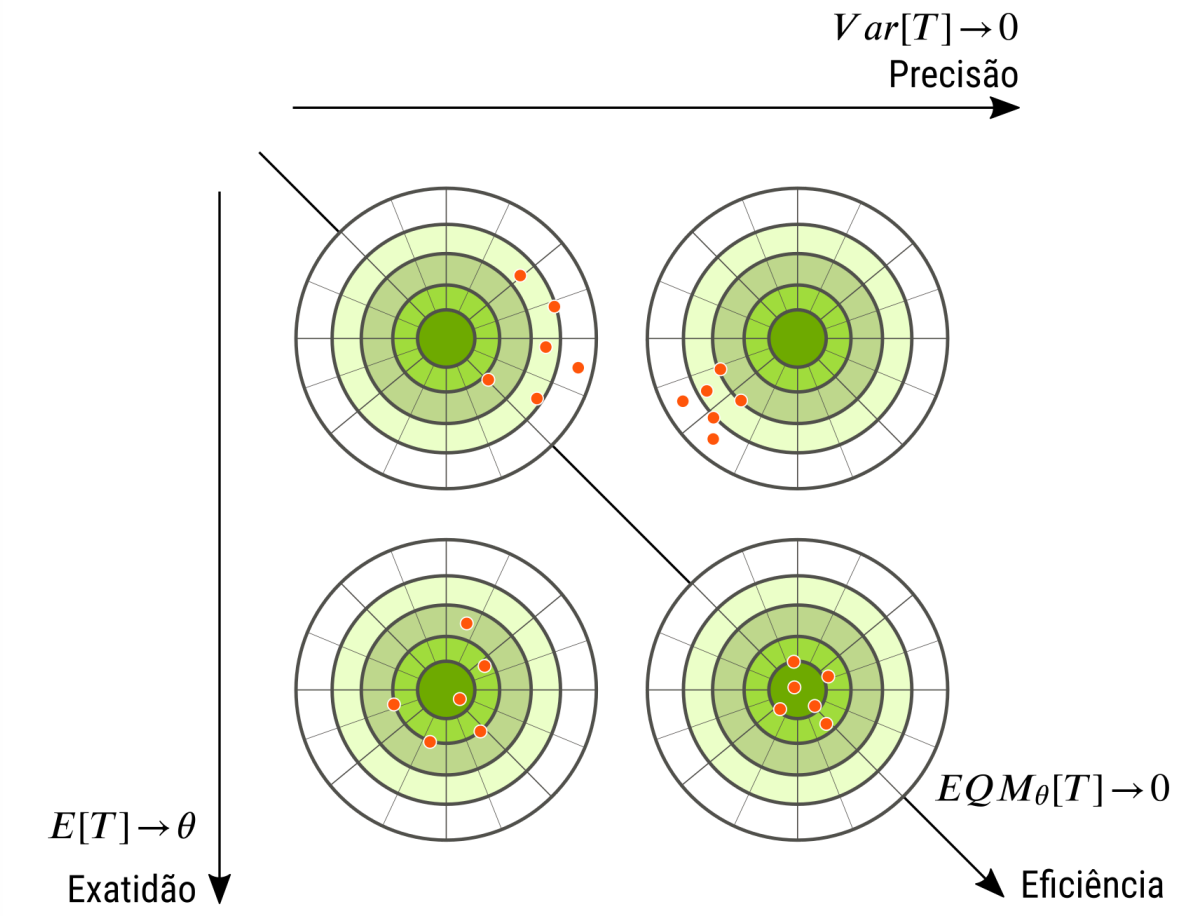
Podemos avaliar a **precisão** de um estimador recorrendo à sua variância.

E podemos avaliar em conjunto as duas caraterísticas usando o **erro quadrático médio**.

### Definição

O **erro quadrático médio** de um estimador pontual $T$ de um parâmetro $\theta$ é definido por

$$EQM_\theta[T] = E\left[(T-\theta)^2\right] = Var[T] + (E[T]-\theta)^2.$$




## Como obter estimadores ou estimativas?

Diversos métodos de estimação estão disponíveis:

- Método da máxima verosimilhança
- Método dos mínimos quadrados
- ...

## Método da máxima verosimilhança

Retomemos a distribuição amostral

$$f_{\mathbf{X}}(x_1, \ldots, x_n \mid \theta) = \prod_{i=1}^{n} f_{X_i}(x_i \mid \theta) = \prod_{i=1}^{n} f_X(x_i \mid \theta)$$

No caso discreto, esta função dá-nos a probabilidade de se observar qualquer ponto amostral $(x_1, \ldots, x_n)$ dado um valor de $\theta$.

## Exemplo – Estimação numa população de Bernoulli

Numa amostra de dimensão 10 de uma população $X \sim \text{Ber}(p)$ observou-se $\sum_{i=1}^{10} x_i = 3$.

Tendo em conta que $p \in \{1/4, 1/2, 3/4\}$, como podemos estimar o valor desconhecido de $p$?

### Definição

A função $\mathcal{L}(\theta \mid \mathbf{x}) \equiv f_{\mathbf{X}}(\mathbf{x} \mid \theta)$ diz-se uma **função de verosimilhança**.

### Definição

O valor

$$\hat{\theta} = \arg\max_{\theta \in \Theta} \mathcal{L}(\theta \mid \mathbf{x})$$

diz-se a **estimativa de máxima verosimilhança** do parâmetro $\theta$.

## Notas

1. O estimador associado, quando é determinável analiticamente, diz-se o **estimador de máxima verosimilhança**;

2. Frequentemente é mais conveniente maximizar $\log \mathcal{L}(\theta \mid \mathbf{x})$.

## Exemplo – Estimação numa população Geométrica

Seja $X_1, \ldots, X_n$ uma amostra aleatória de uma população com distribuição $Geo(p)$, com $0 < p < 1$.

Determine o estimador de máxima verosimilhança do parâmetro $p$ e de $q = P(X > 1)$.

### Invariância dos estimadores de máxima verosimilhança

Seja $\hat{\theta}$ o estimador de máxima verosimilhança de um parâmetro $\theta \in \Theta \subset \mathbb{R}^k$ e $g$ uma função de $\mathbb{R}^k$ em $\mathbb{R}^p$ com $p \leq k$. Então, o estimador de máxima verosimilhança de $g(\theta)$ é $g(\hat{\theta})$.

Seja $X_1, \ldots, X_n$ uma amostra aleatória de $X$ com distribuição uniforme contínua num intervalo $(0, \theta)$ em que $\theta \in \mathbb{R}^+$.

Determine o estimador de máxima verosimilhança de $\theta$.

## 6.3 Distribuições amostrais de estimadores

Já vimos que

$$E[\bar{X}] = E[X]$$

Temos ainda que

$$Var[\bar{X}] = \frac{Var[X]}{n}$$

Melhor ainda é conhecer a **distribuição amostral** de um estimador ou estatística!

### Teorema (revisão)

Seja $X_1, \ldots, X_n$ uma amostra aleatória de uma população $N(\mu, \sigma^2)$. Então

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \iff \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0, 1)$$

Para populações não normais, se a amostra for suficientemente grande, podemos aplicar o Teorema do Limite Central, obtendo-se:

$$\frac{\bar{X} - E[X]}{\sqrt{\frac{Var[X]}{n}}} \overset{a}{\sim} N(0, 1) \iff \bar{X} \overset{a}{\sim} N\left(E[X], \frac{Var[X]}{n}\right)$$

### Exemplo – Estimação numa população de Bernoulli

Seja $X_1, \ldots, X_n$ uma amostra aleatória de uma população com distribuição $Ber(p)$. Para $n$ suficientemente grande tem-se

$$\frac{\bar{X} - p}{\sqrt{\frac{p(1-p)}{n}}} \overset{a}{\sim} N(0, 1) \iff \bar{X} \overset{a}{\sim} N\left(p, \frac{p(1-p)}{n}\right)$$

## Outros resultados para populações normais

### Teorema

1. Se $X \sim N(0,1)$ então $Y = X^2 \sim \chi^2_{(1)}$.

2. Se $X_1, \ldots, X_n$ são variáveis aleatórias independentes com $X_i \sim \chi^2_{(1)}$ então $Y = \sum_{i=1}^n X_i \sim \chi^2_{(n)}$.

### Definição

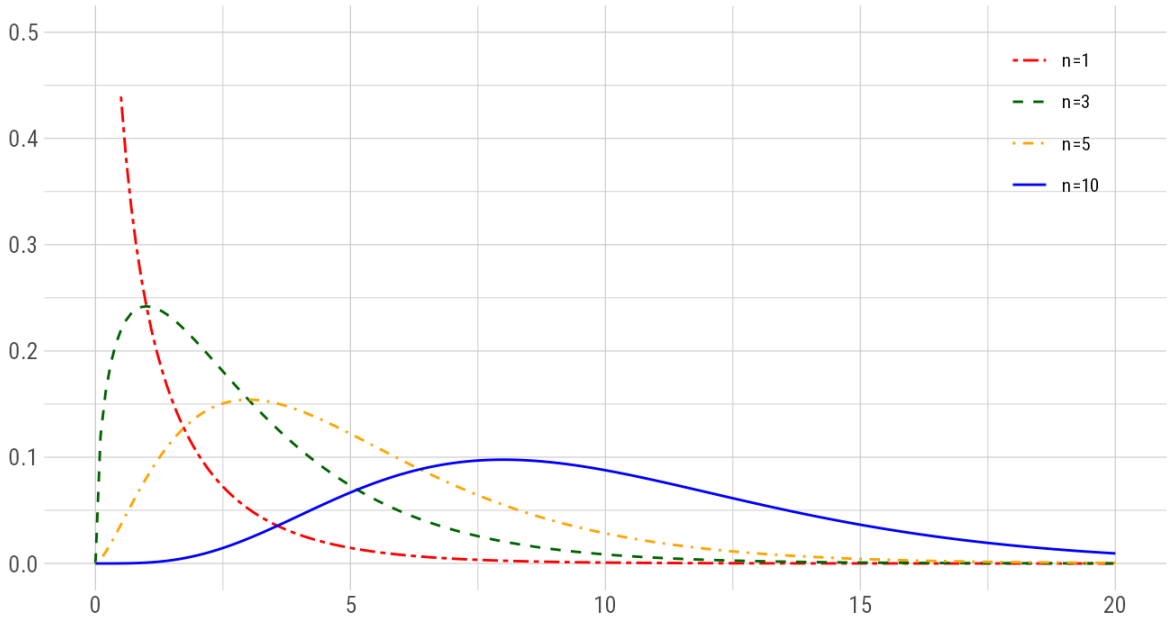
Uma variável aleatória $X$ com função densidade de probabilidade dada por

$$f_Y(y) = \frac{(1/2)^{n/2}}{\Gamma(n/2)} y^{n/2-1} e^{-y/2},\ y > 0,$$

diz-se que tem uma **distribuição do qui-quadrado** com $n$ graus de liberdade, $X \sim \chi^2_{(n)}, n \in \mathbb{N}$.

- $E[X] = n$ e $Var[X] = 2n$
- $\Gamma(x) = \int_0^{+\infty} t^{x-1} e^{-t}\, dt, \Gamma(n) = (n-1)!,\ \forall n \in \mathbb{N}$

### Exemplo – Distribuições do qui-quadrado



## Exemplo – Aplicações

Sejam $X_1, \ldots, X_n$ variáveis aleatórias independentes com $X_i \sim N\left(\mu, \sigma^2\right)$. Então:

1. $\dfrac{\sum_{i=1}^{n}(X_i-\mu)^2}{\sigma^2} \sim \chi^2_{(n)}$;
2. $\dfrac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{\sigma^2} = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{(n-1)}$.

### Teorema

Se $X \sim N(0,1)$ e $Y \sim \chi^2_{(n)}$ são variáveis aleatórias independentes então

$$\frac{X}{\sqrt{Y/n}} \sim t_{(n)}.$$

### Definição

Uma variável aleatória com a função densidade de probabilidade

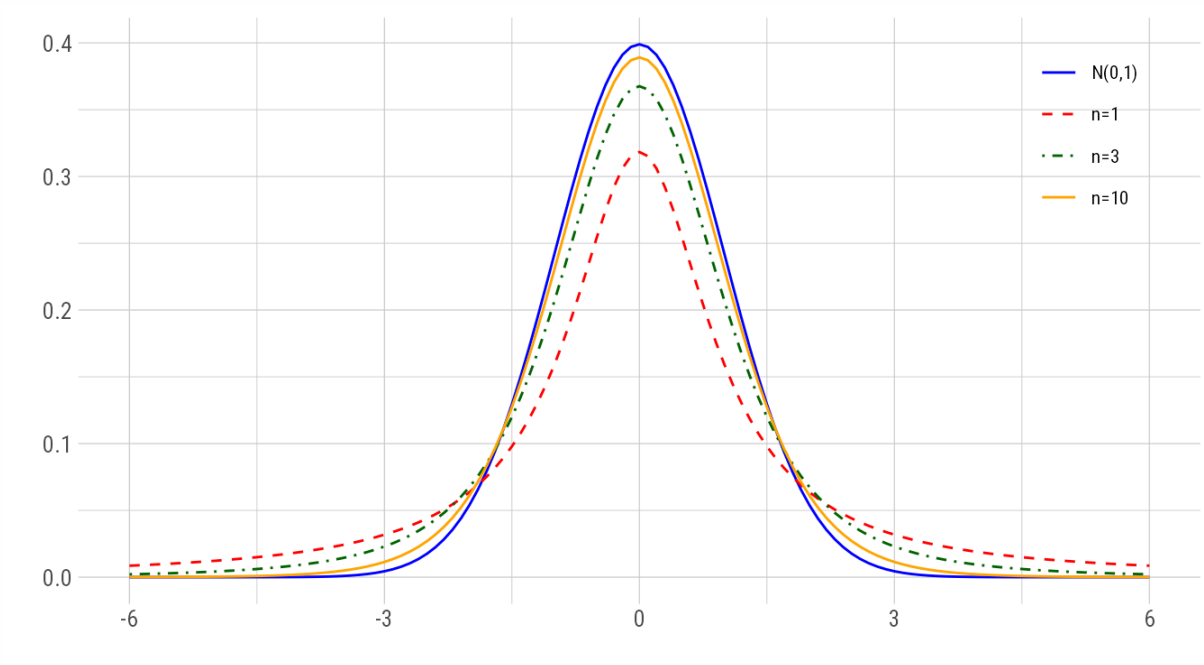
$$f_X(x) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\Gamma\left(\frac{n}{2}\right)}\left(1+\frac{x^2}{n}\right)^{-\frac{n+1}{2}}, \; x \in \mathbb{R}$$

diz-se que tem uma **distribuição t-Student** com $n$ graus de liberdade, $X \sim t_{(n)}$, $n \in \mathbb{N}$.

## Notas

- $E[X] = 0, n \geq 2$
- $Var[X] = \frac{n}{n-2}, n \geq 3$
- $X \xrightarrow{\mathcal{D}} N(0,1)$

## Exemplo – Distribuições t-Student




## Exemplo – Aplicação

Seja $X_1, \ldots, X_n$ uma amostra aleatória de uma população com distribuição $N(\mu, \sigma^2)$.

Então,

$$\frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}} \sim t_{(n-1)}.$$