#867 #DesenvolveDados Seg • Qua • Sex

Estatística I

Conteúdo

6 Inferência estatística

Inferência estatística

Seja X uma variável aleatória com função densidade (ou de probabilidade) denotada por $f(x, \theta)$, em que θ é um parâmetro desconhecido e que se deseja conhecer. Chamamos de inferência estatística o problema que consiste em especificar um ou mais valores para θ, baseado em um conjunto de valores X. Há duas formas de estimar θ estimativas pontuais ou intervalares.

1.1 Estimador Pontual

É um estimador pontual para

$$\theta$$

é qualquer estatística que possa ser usada para estimar esse parâmetro, ou seja,

$$\hat{\theta} = T(X)$$

Isto é, o estimador é uma função da amostra, e a estimativa é o valor observado de um estimador (um número) de uma amostra particular.

Um estimador paramétrico deve apresentar as seguintes propriedades:

- **Suficiente**: um estimador $T(X)$ é suficiente para θ se e somente se $T(X)$ não depender de θ. Por exemplo, o estimador $\overline{X}$ é suficiente para a média populacional, pois a função paramétrica não depende da média populacional μ:

$$\overline{X} = T(X) = \frac{\sum_i^n x_i}{n}$$

- **Consistência**: um estimador é dito consistente para parâmetro

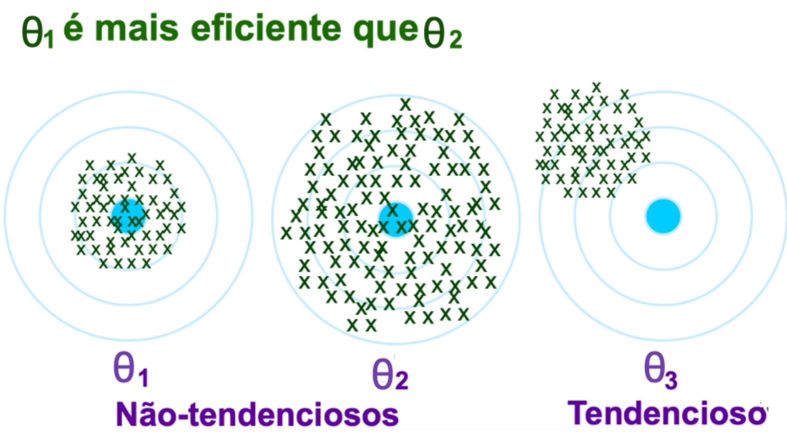
$$\theta$$

  quando, a medida que se aumenta o tamanho **n** da amostra, também aumenta a precisão na estimativa ou seja, considerando $T_n(X)$ o estimador de θ para uma amostra de tamanho n, temos:

$$P(\mid T_n(X) - \theta \mid > \epsilon) \rightarrow 0, n \rightarrow \infty$$

- **Não enviesado**: um estimador de θ é não enviesado (ou não tendencioso) para o parâmetro, se $E(T(X)) = \theta$, para qualquer valor de θ.

- **Eficiência**: Sejam $\hat{\theta}_1$ e $\hat{\theta}_2$ dois estimadores não viciados de θ, dizemos que $\hat{\theta}_1$ é mais eficiente que $\hat{\theta}_2$ se $Var(\hat{\theta}_1) < Var(\hat{\theta}_2)$.



São exemplos de estimadores que satisfazem as quatro propriedades acima:

- Para a média populacional μ

$$\widehat{\mu} = \overline{X} = \frac{\sum_i^n x_i}{n}$$

- para a proporção populacional p (em que $x_i = 0, 1$)

$$\hat{p} = \overline{p} = \frac{\sum_i^n x_i}{n}$$

- para a variância populacional σ^2

$$\widehat{\sigma^2} = s^2 = \frac{\sum_i^n (x_i - \overline{X})^2}{n-1}$$

1.2 Erro Quadrático Médio

Vale ressaltar que nem sempre um estimador eficiente produz um bom resultado. Considere a expressão do Erro Quadrático Médio (EQM) do estimador $\hat{\theta}$:

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

Sendo assim:

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

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

$$= Var(\hat{\theta}) + (E(\hat{\theta}))^2 + E(\theta^2) - 2\theta E(\hat{\theta})$$

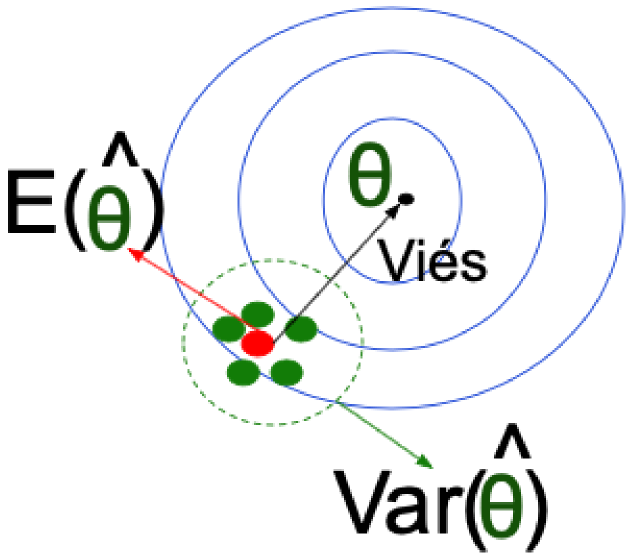
$$= Var(\hat{\theta}) + (E(\hat{\theta}))^2 + \theta^2 - 2\theta E(\hat{\theta})$$

$$= Var(\hat{\theta}) + [E(\hat{\theta}) - \theta]^2$$

Nesta expressão

$$E(\hat{\theta}) - \theta$$

indica o viés do estimador $\hat{\theta}$, sendo assim, perceba mesmo que a variância seja pequena, não necessariamente o erro produzido pelo estimador é baixo (exemplificado na figura abaixo).



1.3 Estimador Mínimos Quadrados

Considere duas medidas $(X_1, X_2, \cdots, X_n)$ e $(Y_1, Y_2, \cdots, Y_n)$, em que Y_i pode ser estimado da seguinte forma:

$$Y_i = \theta X_i + \epsilon$$

Sendo assim, é supomos que X é uma variável explicativa para Y, o que implica que Y seguem uma distribuição de probabilidade centrada em θX. Por consequência, $\epsilon = Y - \theta X$ segue uma distribuição de probabilidade centrada no zero.

O método de estimação por mínimos quadrados consiste em minimizar o quadrado das diferenças entre os valores observados de uma amostra e seus respectivos valores estimados. Sendo assim, este método consiste em minimizar:

$$\sum_i^n \epsilon^2 = \sum_i^n (Y_i - \theta X)^2$$

O que equivale a resolver a seguinte equação:

$$\frac{\partial}{\partial \theta} \sum_i^n (Y_i - \theta X)^2 = 0$$

Exemplo 1: Suponha que se queira estimar a média $T(X)$ de um conjunto de dados:

$$\frac{\partial}{\partial T(X)} \sum_i^n (X_i - T(X))^2 = 0$$

Abrindo o somatório:

$$\frac{\partial}{\partial T(X)} \left[\sum_i^n X_i^2 - 2T(X) + \sum_i^n T(X)^2 \right] = 0$$

Derivando:

$$-2\sum_{i}^{n} X_i + 2\sum_{i}^{n} T(X) = 0$$

$$-2\sum_{i}^{n} X_i + 2nT(X) = 0$$

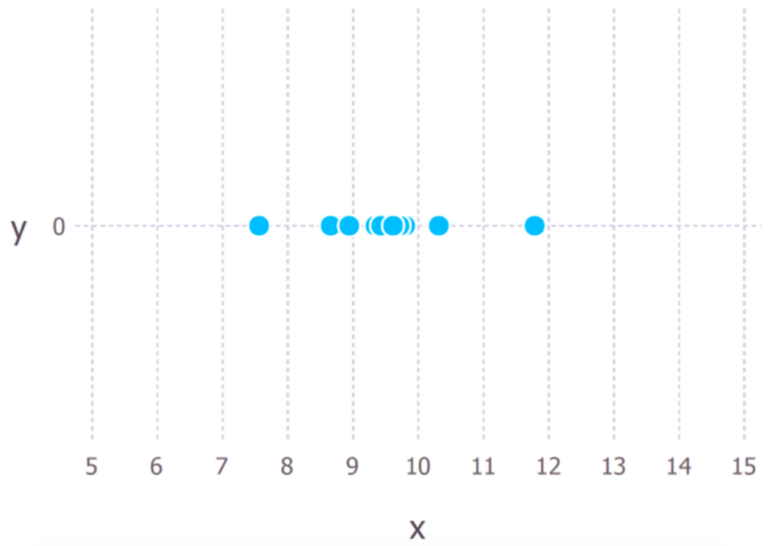
$$T(X) = \frac{-2\sum_{i}^{n} X_i}{2n} = \frac{\sum_{i}^{n} X_i}{n}$$

Note que a expressão já conhecida da média amostral coincide com o estimador de mínimos quadrados para a média.

1.4 Máxima Verossimilhancia

A estimativa de máxima verossimilhança é um método que determina valores para os parâmetros, de modo a maximizar a probabilidade de que o processo descrito pelo modelo produza os dados que foram realmente observados.

Para **exemplificar**, suponha que se tenha observado 10 alunos e o tempo em segunds que cada um leva para responder a uma pergunta específica de um exame. Esses 10 pontos de dados são mostrados na figura abaixo:



Definição: Considere que uma amostra aleatória

$$(x_1, x_2, \cdots, x_n)$$

seja retirada de uma população, em que a função de densidade de probabilidade é

$$f(x, \theta)$$

, a qual depende do vetor de parâmetros

$$\theta$$

, então a função de verossimilhancia é definida pela equaçao abaixo. O estimador de máxima verossimilhancia é o valor $\hat{\theta}_{MV}$ que maximiza $L(\theta; x_1, x_2, \cdots, x_n)$.

$$L(\theta; x_1, x_2, \cdots, x_n) = \prod_{i=1}^{n} f(x_i, \theta) = f(x_1, \theta) f(x_2, \theta) \cdots f(x_n, \theta)$$

Este problema se traduz em encontrar as derivadas parciais da função de verossimilhancia, igualar a zero e encontrar o vetor $\hat{\theta}_{MV}$ que resolve as equações. é comum trabalhar com o logaritmo natural da função de verossimilhança (lnL), pois maximizar o logaritmo natural de uma função é em geral mais simples e produz os mesmos resultados da maximização da função original.

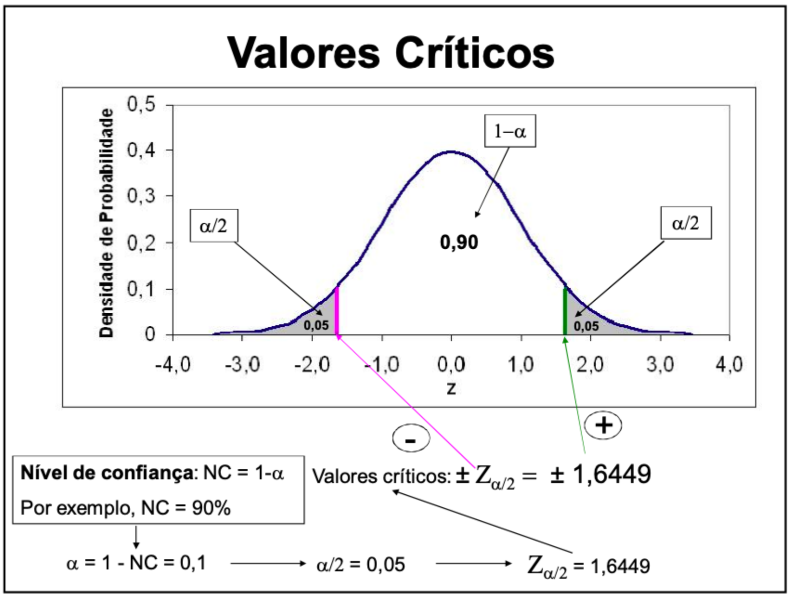
1.5 Intervalo de Confiança

A estimção pontual não permite julgar a magnetude do erro cometido na estimativa. Então surge a necessidade de um estimador intervalar (Intervalo de Confiança, IC), que permite construir um intervalo que é previsto para conter o parâmetro estimado, baseado na distribuição amostral do estimador pontual. A confiança que atribuimos ao intervalo é a probabilidade de que ele irá conter o parâmetro.

Seja $(1 - \alpha)$ uma probabilidade especificada e L e U funções dos valores amostrais X, de modo que

$$P(L < \theta < U) = 1 - \alpha$$

O intervalo (L, U) é chamado intervalo de confiança e $(1 - \alpha)$ é o nível deconfiança (NC) associado ao intervalo. Pode-se definir nível de confiança como a probabilidade de o intervalo conter θ, em outras palavras é a proporção de vezes que o intervalo de confiança realmente contém o parâmetro populacional, supondo que o processo seja repetido um número grande de vezes.



Tópico anterior

Próximo Tópico