CIÊNCIA DE DADOS (BIG DATA)

ANÁLISE ESTATÍSTICA

Professor curador: Mário Olímpio de Menezes

EaD Mackenzie

TRILHA 2

PARTE B – INTRODUÇÃO À PROBABILIDADE E INFERÊNCIA ESTATÍSTICA

PARTE B – INTRODUÇÃO À PROBABILIDADE E INFERÊNCIA ESTATÍSTICA

VARIÁVEIS ALEATÓRIAS E DISTRIBUIÇÕES DE PROBABILIDADE

VARIÁVEIS ALEATÓRIAS

- Uma variável numérica cujo valor depende do resultado de um experimento aleatório.
- Uma variável aleatória associa um valor numérico com cada resultado de um experimento aleatório.
- Pode ser **discreta** ou **contínua**.

DISTRIBUIÇÕES DE PROBABILIDADES

- A **distribuição de probabilidade de uma variável aleatória discreta** *x* dá a probabilidade associada com cada possível valor *x.*
- Distribuições mais conhecidas para Variáveis Aleatórias Discretas são:
  - Distribuição Binomial
  - Distribuição de Poisson

DISTRIBUIÇÃO BINOMIAL

- **Exemplo**: número de **caras (heads – H)** em quatro lançamentos de uma moeda.

P(X) = n(X)/n(Ω)

Ω = 16 resultados possíveis

		HTT		
		HTHT		
	HTTT	THTH	HHHT	
	THTT	HHTT	HHTH	
	TTHT	THHT	HTHH	
TTTT	TTTH	TTHH	THHH	HHHH
X = 0	X = 1	X = 2	X = 3	X = 4

Cada um dos 16 resultados possíveis tem a mesma probabilidade: 1/16

DISTRIBUIÇÕES CONTÍNUAS

- Alguns dados vêm de medidas em escalas essencialmente contínuas, tais como: temperatura, concentrações, distâncias etc.
- Para variáveis aleatórias contínuas, utilizamos o conceito de **densidade de probabilidade**.

INFERÊNCIA ESTATÍSTICA

- Inferência estatística tem como objetivo fazer afirmações sobre uma característica de uma **população** a partir do conhecimento de dados de uma parte desta população – uma **amostra**.
- A população é **representada** por uma distribuição de probabilidade com **parâmetros** de valores desconhecidos.

FAZENDO INFERÊNCIA ESTATÍSTICA

Quando abordamos um problema de estatística, algumas etapas possíveis para sua solução são:

- Estimação Pontual do parâmetro da população.
- Teste de Hipóteses.
- Estimação Intervalar.

ESTIMAÇÃO PONTUAL

- O objetivo é apresentar **um valor** para um parâmetro da população.
- Dentre os parâmetros que desejamos inferir para a população (com distribuição Normal), estão:
  - Média.
  - Desvio padrão.

ESTIMAÇÃO PONTUAL

DISTRIBUIÇÃO NORMAL

$$X \sim N(\mu, \sigma^2)$$

temos que $E(X) = \mu$ e $Var(X) = \sigma^2$

Um estimador para μ : $\bar{X}$

Um estimador para σ^2 : $s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$

ESTIMAÇÃO INTERVALAR (CONJUNTO DE VALORES)

- Objetivo é apresentar **um intervalo de possíveis valores** para o parâmetro da população, chamado de *intervalo de confiança*.
- Os limites do intervalo são funções da amostra (X1...Xn).
- A probabilidade de que o intervalo contenha o parâmetro deve ser **alta**.
- A *amplitude* do intervalo deve ser tão pequena quanto possível.

TESTE DE HIPÓTESES

- Uma hipótese estatística (H) é uma afirmação sobre o valor do parâmetro da população que estamos estimando.
- Pode ser verdadeira ou falsa.
- Utiliza-se duas hipóteses:
  - Hipótese Nula (H_0)
  - Hipótese Alternativa (H_1)

TIPOS DE ERROS

Quando fazemos um teste de hipótese, estamos sujeitos a dois tipos de erros:

- Erro tipo I: **rejeitar** H_0 quando H_0 é **verdadeira**.
- Erro tipo II: **não rejeitar (aceitar)** H_0 quando H_0 é **falsa**.

TESTE DE HIPÓTESES

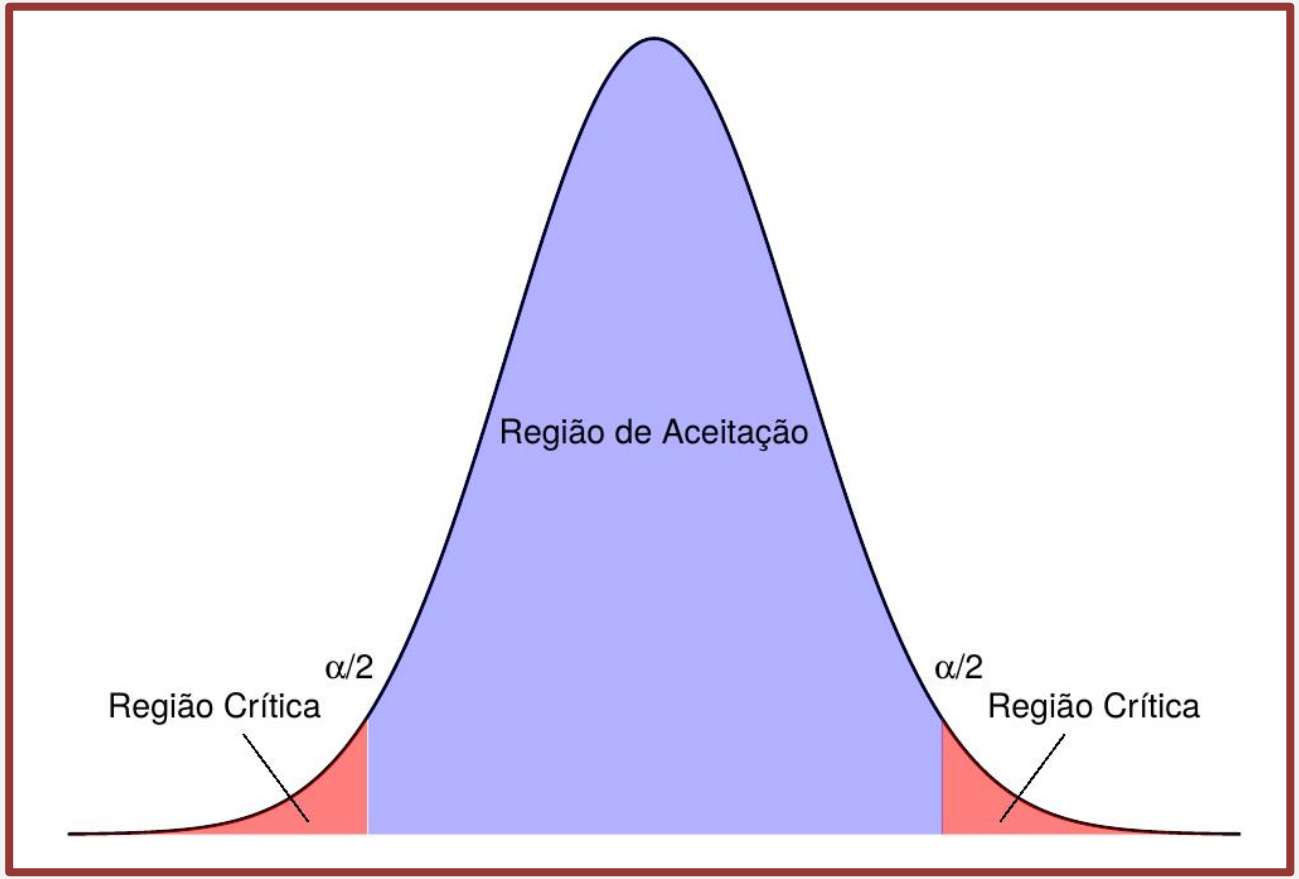
- Para efetuarmos um teste de hipóteses, definimos um **limite superior** para a probabilidade máxima de **Erro Tipo I** que deve ser tolerada.
- Este limite é o ***nível de significância*** do teste, α.

  - α = 0,05
  - α = 0,01

Valores típicos de níveis de significância

REGIÃO CRÍTICA E REGIÃO DE ACEITAÇÃO

- No Teste de Hipóteses, chamamos de **Região Crítica** (R_c) ou região **de rejeição** o conjunto de valores assumidos pela estatística de teste para os quais a hipótese nula é **rejeitada**.
  Seu complementar é a **região de aceitação** (R_a)


Região de Aceitação
α/2
Região Crítica
α/2
Região Crítica


M
Mackenzie