Estatística para Ciência de Dados

Profa. Rebeca Valgueiro

Quem sou eu?

- Graduada em Engenharia Civil
- MBA em Gestão Empresarial
- Trabalho a mais de 4 anos no mercado de tecnologia atuando em projetos de:
    - Desenvolvimento web
    - Desenvolvimento desktop- windows
    - Data Science
    - IA e visão computacional

https://www.linkedin.com/in/rebecavalgueiro/

Imagine que você acredita em uma teoria razoável

>>Surge uma nova teoria

>> Contraria o que você acredita

Exemplo

Sabe-se que a média dos homens brasileiros é de 1,71 m de altura.

TEORIA INICIAL > A média de altura dos homens holandeses é próximo desse valor

NOVA TEORIA > A média de altura é maior do que 1,71m (COMO PROVAR?)

Hipótese

Uma teoria ou suposição que é testada usando dados e análise estatística e que pode explicar um comportamento determinado do interesse de pesquisa.
É uma afirmação ou proposição testável sobre um fenômeno ou a relação entre variáveis

Tipos de Hipóteses

1. Hipótese Nula (H_0)
2. Hipótese Alternativa (H_1 ou H_a)

Hipótese Nula (Ho)

1. É a hipótese de "não efeito", "não diferença" ou "não mudança".
2. Representa o status quo, o que é geralmente aceito como verdadeiro ou o que se assume que é verdade antes de coletar os dados.
3. O objetivo do teste é tentar encontrar evidências para rejeitar a hipótese nula.

Hipótese Alternativa (H_1 ou H_a)

1. Indica a presença de um efeito, uma diferença, ou uma relação entre variáveis
2. Representa o "efeito", "diferença" ou "mudança" que você suspeita que exista.

Testes de hipóteses

1 - Você presume que a hipótese nula é verdadeira
2 - Avalia se os dados da sua amostra são improváveis o suficiente sob essa suposição
3 - Se os dados forem muito improváveis, você rejeita a hipótese nula em favor da alternativa.
4 - Se não forem, você não rejeita a hipótese nula (o que não significa que ela seja verdadeira, apenas que não há evidências suficientes para refutá-la).

Erro em teste de Hipóteses

- Erro Tipo I (α): Rejeitar H0 quando ela é verdadeira. O nível de significância (α) é a probabilidade máxima de cometer esse erro.
- Erro Tipo II (β): Não rejeitar H0 quando ela é falsa. É como libertar um culpado. A potência do teste (1 - β) é a probabilidade de rejeitar corretamente uma H0 falsa.

O α é definido pelo pesquisador para controlar o risco do Erro Tipo I

Passo a passo

1 - Formular as Hipóteses: Definir claramente o que você está testando.
2 - Escolher o Nível de Significância (α): Este é o risco máximo que você está disposto a correr de cometer um Erro Tipo I (rejeitar H0 quando ela é verdadeira). Os valores comuns são 0.05 (5%), 0.01 (1%) ou 0.10 (10%).
3 - Coletar os Dados Amostrais: Obter as informações necessárias da sua amostra.
4 - Calcular a Estatística de Teste: É um valor calculado a partir dos dados da amostra que nos permite avaliar a evidência contra H0
5 - Tomar uma Decisão (Usando p-valor ou Valor Crítico)
6 - Formular a Conclusão: Declarar a decisão em termos do problema original, de forma clara e não técnica.

Exemplo

TEORIA INICIAL (H_0): O tamanho médio dos crânios dos seres humanos não mudou significativamente ao longo dos anos.

NOVA TEORIA (H_1): O tamanho médio dos crânios aumenta ao longo dos anos

Exemplo

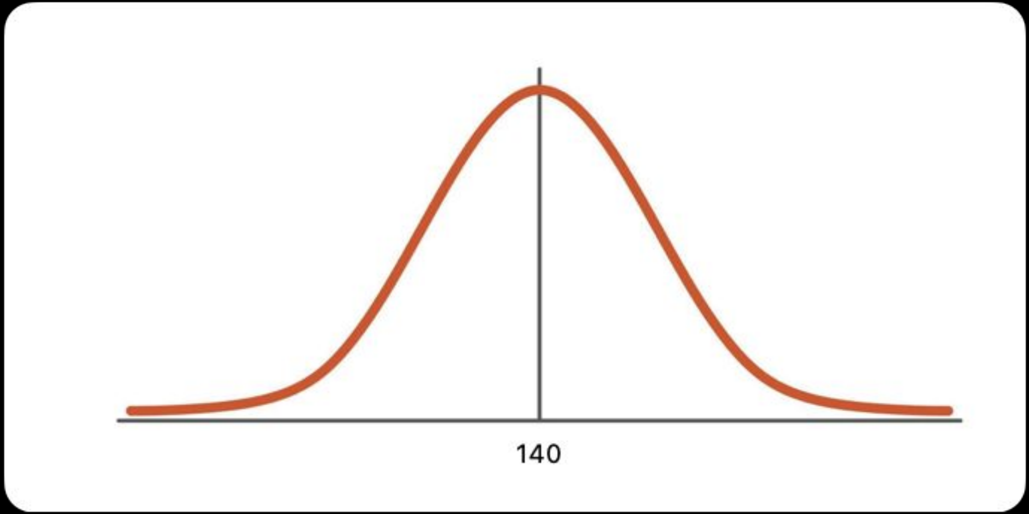
DADOS:

- Foi feita uma pesquisa sobre o tamanho dos crânios atuais e chegamos nas estatísticas (distribuição normal):
    - média = 140mm
    - desvio padrão = 26mm

- Foram encontrados 30 crânios de 600 anos atrás, foi feito as medidas e chegamos a seguinte estatística:
    - média: 131,37 mm

O QUE ISSO SIGNIFICA?
ISSO PROVA MINHA HIPÓTESE?
O VALOR É SIGNIFICATIVAMENTE MENOR?

Exemplo

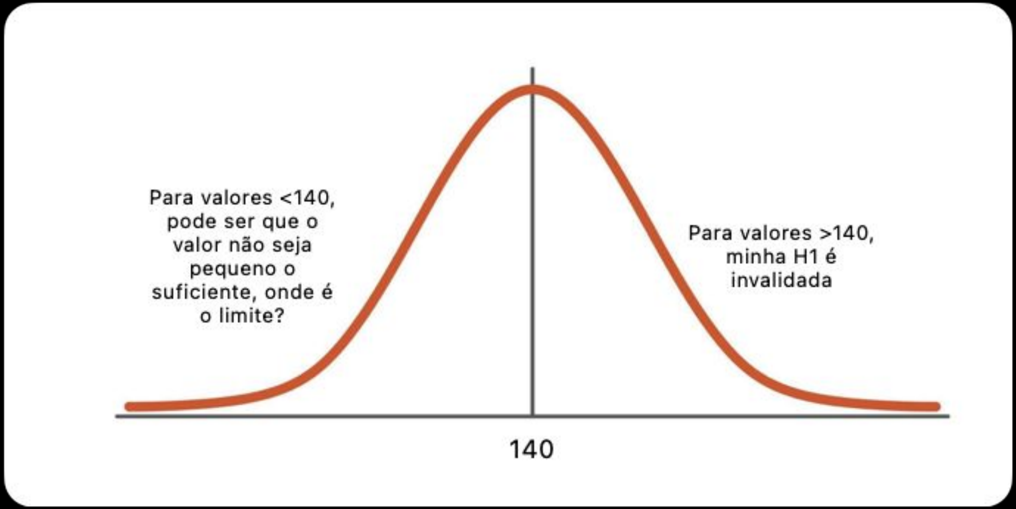


H0: média = 140
H1: média < 140

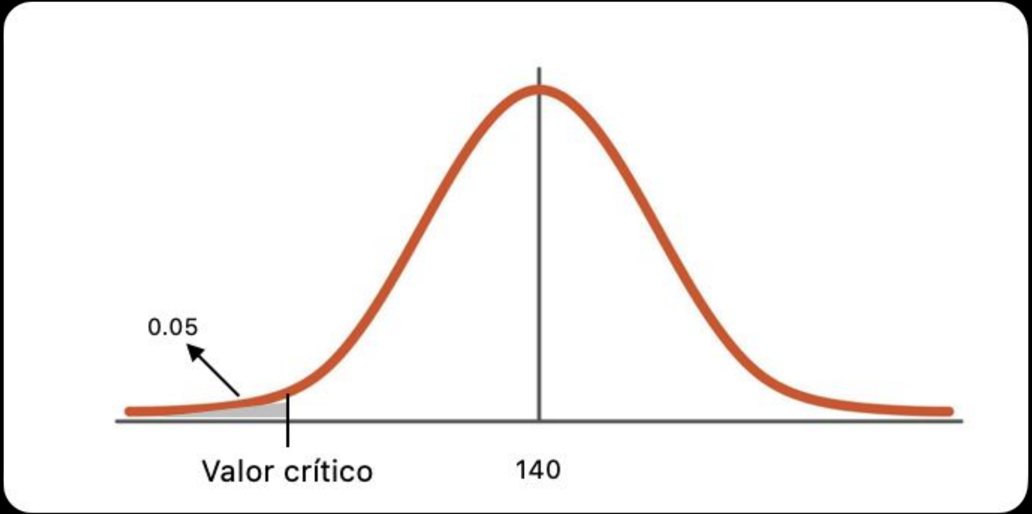
Já que da amostra deu 131,37 mm, já podemos validar H1?
Precisamos saber se a média é significativamente menor

Exemplo

Precisamos saber qual o nível de significância que buscamos no teste (α). Quanto você está disposto a errar?



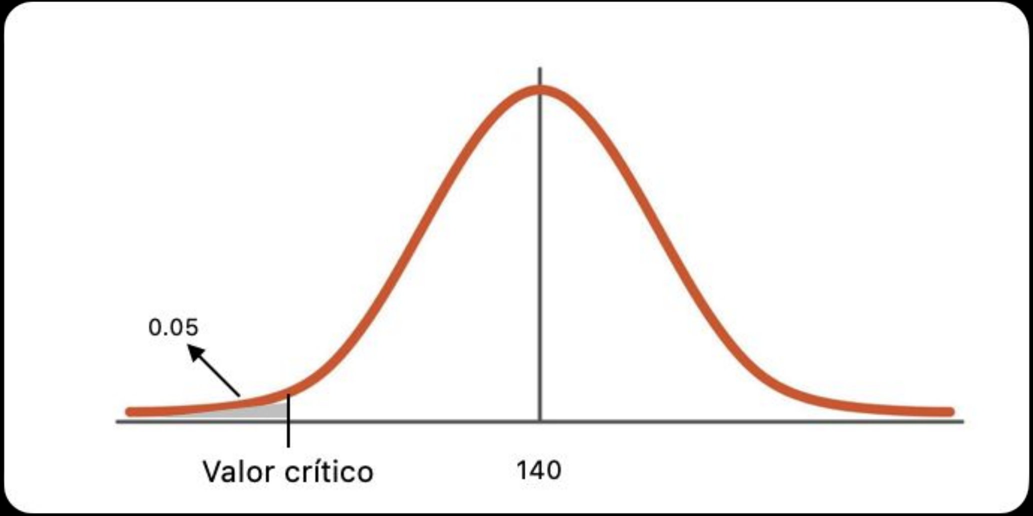
Exemplo



Todos os valores abaixo do valor crítico fazem parte da **ZONA CRÍTICA**

Se a média da amostra estiver na Zona crítica > H0 negada

Exemplo



P(Z ≤ Zcritico)=α

Z para α = 0.05 >> -1.645

$$IC = \bar{x} \pm Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

Xc = 140 - 1.645*26/sqrt(30)

Xc = 132,22

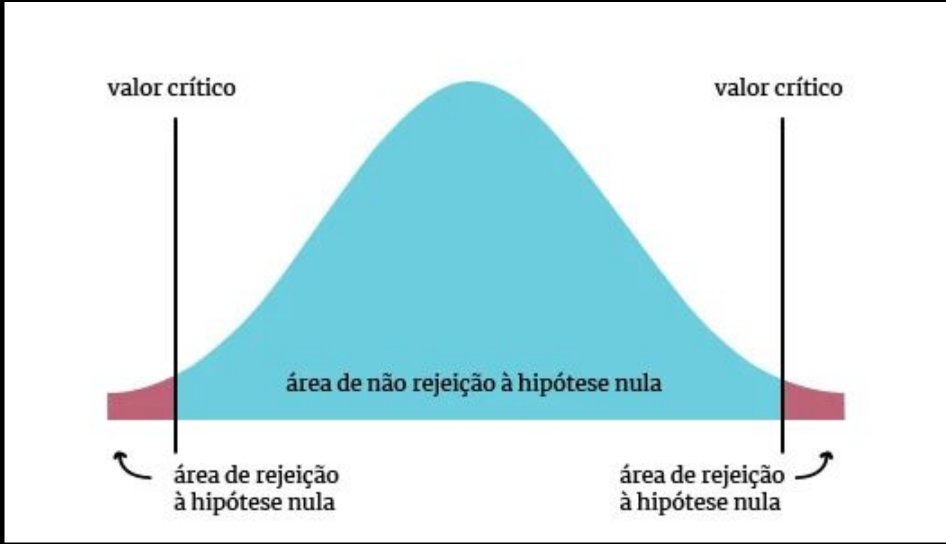
Exemplo

Logo se a média da amostra for menor que o valor crítico, minha hipótese Ho está Rejeitada
Xc = 132,22
média da amostra = 131,37mm

>> Concluímos que o tamanho dos crânios está aumentando com o tempo.

Observação

No nosso exemplo, vimos o caso de querermos validar um valor menor do que o proposto por H0, mas podemos validar também números maiores ou diferentes.



Valor Crítico

É o valor que serve como um limite de corte que separa a região de rejeição da região de não rejeição em uma distribuição de probabilidade. Ele é usado para decidir se a estatística de teste calculada a partir da amostra é "extrema" o suficiente para rejeitar a hipótese nula (H0).

p-valor (valor de probabilidade)

O p-valor é a probabilidade de se observar um resultado de amostra tão extremo ou mais extremo do que o resultado que você realmente obteve, assumindo que a hipótese nula (H_0) é verdadeira.

Se o p-valor é pequeno, significa que os seus dados observados seriam muito improváveis de acontecer se a H_0 fosse de fato verdadeira. Isso nos leva a duvidar da H_0.

Se o p-valor é grande, significa que os seus dados observados (ou dados ainda mais extremos) seriam razoavelmente prováveis de acontecer se a H_0 fosse verdadeira. Isso sugere que não há evidência forte o suficiente para rejeitar a H_0.

p-valor (valor de probabilidade)

A decisão em um teste de hipóteses usando o p-valor é baseada na comparação do p-valor com o **nível de significância (α)**, que é predefinido pelo pesquisador.

- **Se p-valor ≤ α:**
  - **Decisão:** Rejeitar a Hipótese Nula (H_0).
  - **Interpretação:** Há evidências estatísticas suficientes nos dados para concluir que a Hipótese Alternativa (H_1) é verdadeira (ou que a H_0 é falsa).

p-valor (valor de probabilidade)

- **Se p-valor >α:**
  - **Decisão:** Não rejeitar a Hipótese Nula (H_0).
  - **Interpretação:** Não há evidências estatísticas suficientes nos dados para rejeitar a Hipótese Nula (H_0). O resultado **não** é considerado "estatisticamente significativo". **Importante**: "Não rejeitar H_0" não significa que H_0 é verdadeira, apenas que não há evidências fortes o suficiente contra ela com base na amostra e no nível de significância escolhido.

Exemplo

Uma empresa desenvolveu um novo fertilizante e afirma que ele fornece uma produção média de cana-de-açúcar em até 80 toneladas/hectare. Um agricultor faz um teste com uma amostra de 35 hectares usando o novo fertilizante, obtendo uma produção média de 83 toneladas/hectare. O desvio padrão populacional histórico para a produção é conhecido como 10 toneladas/hectare. Queremos testar a 5% de significância.

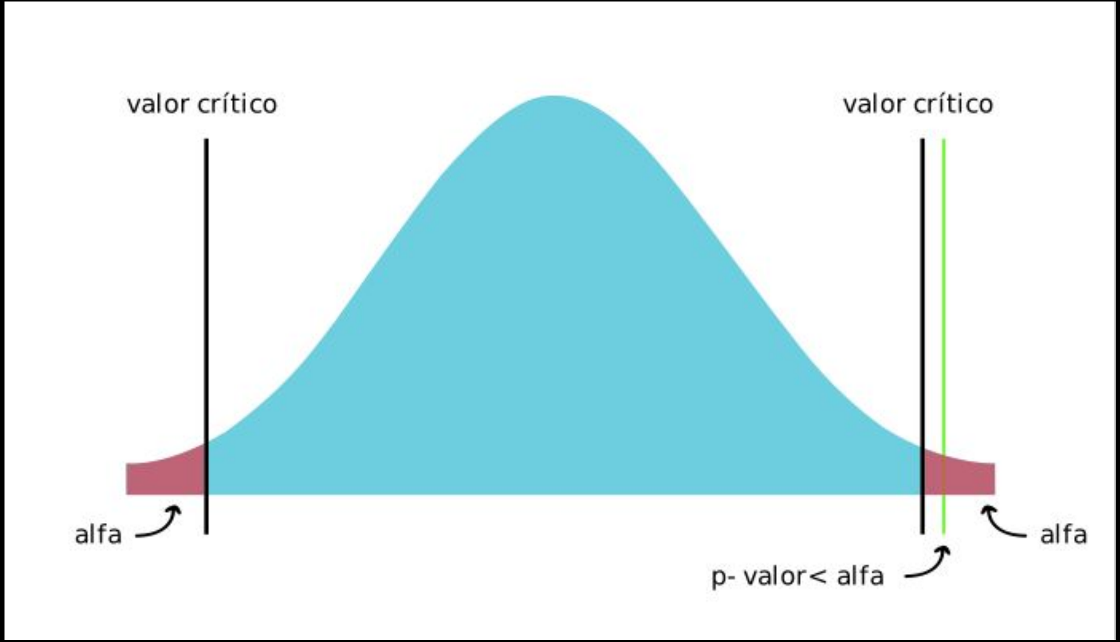
H0 : A produção média com o novo fertilizante é menor que 80 ton/ha. ($\mu \leq 80$)
H1: A produção média com o novo fertilizante é maior que 80 ton/ha. ($\mu > 80$)
$\alpha = 0.05$ (ou 5%)

Exemplo

DADOS:
Média amostral (μ) = 83 t/ha
Desvio padrão populacional (σ) = 10 t/ha
Tamanho da amostra (n) = 35 hectares
Média populacional sob H0 (μ0) = 80 t/ha
α=0.05 (ou 5%)

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{10}{\sqrt{35}} \approx \frac{10}{5.916} \approx 1.6904$$

Exemplo



Precisamos descobrir qual o Z para o valor de média amostral = 83:

$$IC = \bar{x} \pm Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

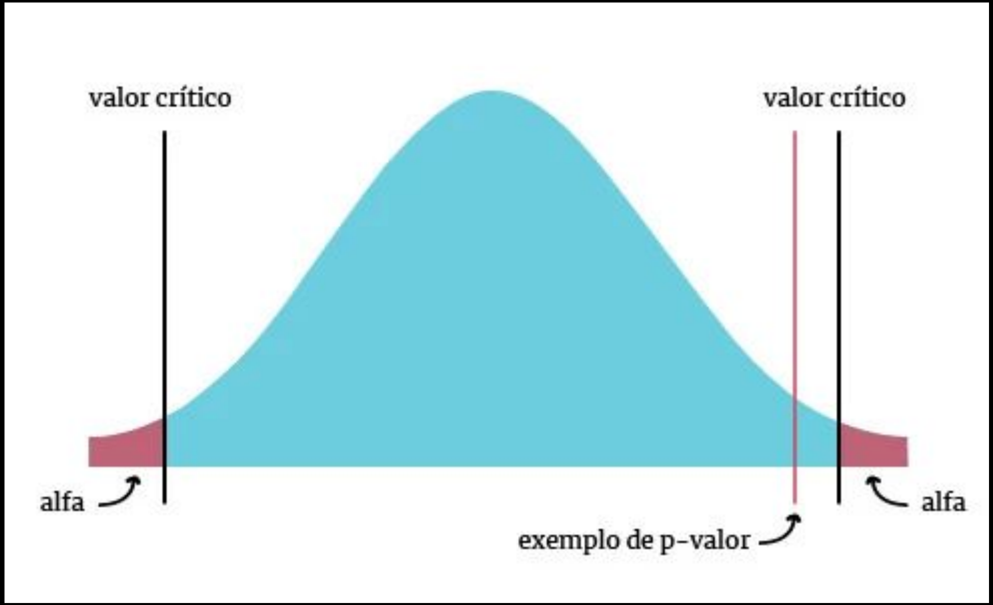
83 = 80 + Z * 1.6904

Z = 1,77 >> α = 0.0384

0.0384 < 0.05

REJEITA-SE Ho

Exemplo



p-valor X valor crítico

Método do Valor Crítico:
"Minha estatística de teste calculada é extrema o suficiente para cair na região definida pelo meu α?"

Método do p-valor:
"Qual é a probabilidade de obter meus resultados se H0 fosse verdadeira? É essa probabilidade menor que o meu α?"

Como calcular o Z ou t usando python

1. Para a Distribuição Normal Padrão (Z-scores) - scipy.stats.norm

stats.norm.cdf(z_value): Função de Distribuição Acumulada (Cumulative Distribution Function). Retorna a probabilidade acumulada à esquerda de um determinado z_value.

```
from scipy import stats
z_score = -1.96
p_left_tail = stats.norm.cdf(z_score)
print(f"P(Z <= {z_score}): {p_left_tail:.4f}") # Saída: 0.0250 (para um teste unilateral à esquerda)
```

Para testes unilaterais à esquerda usar: 1 - stats.norm.cdf(z_value)

Como calcular o Z ou t usando python

1. Para a Distribuição Normal Padrão (Z-scores) - scipy.stats.norm

stats.norm.ppf(probability): É a inversa da CDF. Retorna o valor Z (valor crítico) correspondente a uma dada probability acumulada à esquerda ou à direita.

```
from scipy import stats
alpha = 0.05
# Valor crítico para teste unilateral à esquerda (alpha na cauda esquerda)
z_critico_esq = stats.norm.ppf(alpha)
# Valor crítico para teste unilateral à direita (alpha na cauda direita, então 1-alpha à esquerda)
z_critico_dir = stats.norm.ppf(1 - alpha)
# Valores críticos para teste bilateral (alpha/2 em cada cauda)
z_critico_bilateral_inf = stats.norm.ppf(alpha / 2)
z_critico_bilateral_sup = stats.norm.ppf(1 - alpha / 2)
```

Como calcular o Z ou t usando python

2. Para a Distribuição t de Student (t-scores) - scipy.stats.t

stats.t.cdf(t_value, df): Retorna a probabilidade acumulada à esquerda de um determinado t_value para um dado número de df (graus de liberdade)

```
from scipy import stats
t_score = -2.10
df = 19 # Para n=20
p_left_tail = stats.t.cdf(t_score, df)
print(f"P(T <= {t_score}, df={df}): {p_left_tail:.4f}") # Exemplo: 0.0238
```

Para testes unilaterais à esquerda usar: 1 - stats.t.cdf(t_value, df)

Como calcular o Z ou t usando python

2. Para a Distribuição t de Student (t-scores) - scipy.stats.t

stats.t.ppf(probability, df): É a inversa da CDF. Retorna o valor t (valor crítico) correspondente a uma dada probability acumulada à esquerda, para um dado número de df.

```
from scipy import stats
alpha = 0.05
df = 15 # Exemplo para n=16
# Valor crítico para teste unilateral à esquerda
t_critico_esq = stats.t.ppf(alpha, df)
# Valor crítico para teste unilateral à direita
t_critico_dir = stats.t.ppf(1 - alpha, df)
# Valores críticos para teste bilateral
t_critico_bilateral_inf = stats.t.ppf(alpha / 2, df)
t_critico_bilateral_sup = stats.t.ppf(1 - alpha / 2, df)
```

Como calcular o Z ou t usando python

Distribuição	Valor Crítico	p-valor
Z	stats.norm.cdf(z_value)	stats.norm.ppf(probability)
t	stats.t.cdf(t_value, df)	stats.t.ppf(probability, df)

Exemplo 1

Valor Crítico

Uma empresa de construção civil em Goiana, Pernambuco, afirma que a resistência média à compressão de um novo tipo de concreto que ela produz é de 30 MPa (Megapascals). Um engenheiro de controle de qualidade da prefeitura suspeita que a resistência média real seja inferior a 30 MPa. Ele coleta uma amostra aleatória de 18 blocos de concreto desse novo tipo e mede sua resistência à compressão. Os resultados da amostra são:

Média Amostral: 28.5 MPa

Desvio Padrão Amostral: 3.5 MPa

Com base nesses dados, o engenheiro quer verificar sua suspeita utilizando um nível de significância de $\alpha=0.01$ (1%).

```
from scipy import stats
import numpy as np

m_ho = 30
m_amostra = 28.5
n=18
dp_amostra = 3.5
nivel_confianca = 0.01

grau_liberdade = n-1
t_critico = stats.t.ppf(nivel_confianca,grau_liberdade)
erro_padrao = dp_amostra/np.sqrt(n)
valor_critico = m_ho + t_critico*erro_padrao
print(valor_critico)
if valor_critico < m_amostra:
    print("H0 não foi rejeitada")
else:
    print("H0 foi rejeitada")
```

```python
from scipy import stats
import numpy as np

m_ho = 30
m_amostra = 28.5
n=18
dp_amostra = 3.5
nivel_confianca = 0.04
grau_liberdade = n-1
erro_padrao = dp_amostra/np.sqrt(n)
t_critico = (m_amostra - m_ho)/erro_padrao
print(t_critico)
alpha_critico = stats.t.cdf(t_critico, grau_liberdade)
if alpha_critico > nivel_confianca:
    print("H0 não foi rejeitada")
else:
    print("H0 foi rejeitada")
```

Inferência sobre a Diferença entre Duas Médias Populacionais

Utilizamos quando queremos saber se existe uma diferença significativa entre dois grupos.

Por que Comparar Duas Médias?

- ➔ Economia: O gasto médio de turistas nacionais é diferente do gasto médio de turistas estrangeiros em Pernambuco?
- ➔ Saúde: Um novo tratamento reduz o tempo de recuperação mais do que o tratamento padrão?
- ➔ Educação: A média de notas de alunos de escolas públicas é diferente da média de notas de alunos de escolas privadas?
- ➔ Engenharia: A resistência média de um novo material é maior que a de um material tradicional?

Distribuição Amostral da Diferença entre Duas Médias

Quando retiramos amostras de duas populações independentes e calculamos suas médias (x1 e x2), a diferença entre essas médias amostrais (x1 - x2) também forma uma distribuição.

Média da diferença

$$E(\bar{x}_1 - \bar{x}_2) = \mu_1 - \mu_2$$

Erro Padrão da Diferença

$$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Dessa forma conseguimos fazer cálculo de Intervalo de confiança e Teste de hipóteses para essa diferença

Exemplo

Imagine que uma agência de turismo em Recife quer saber se há diferença no gasto médio diário entre turistas nacionais e estrangeiros. Eles têm dados históricos que fornecem os desvios padrão populacionais.

Dados Históricos:

- Turistas Nacionais (População 1): $\sigma_1 = \text{R\$ } 80,00$
- Turistas Estrangeiros (População 2): $\sigma_2 = \text{R\$ } 100,00$

Amostras Coletadas:

- Amostra de $n_1 = 60$ turistas nacionais: $\bar{x}_1 = \text{R\$ } 520,00$
- Amostra de $n_2 = 45$ turistas estrangeiros: $\bar{x}_2 = \text{R\$ } 550,00$

Questão: Existe uma diferença estatisticamente significativa no gasto médio diário entre turistas nacionais e estrangeiros? Use α=0.05.

Exemplo

1 - Formular Hipóteses
H0: $\mu1 - \mu2 = 0$ (Não há diferença no gasto médio)
H1: $\mu1 - \mu2 != 0$ (Há uma diferença no gasto médio)

2 - Calcular Estatísticas

Erro padrão da diferença:

$$SE(\bar{x}_1 - \bar{x}_2) = \sqrt{\frac{80^2}{60} + \frac{100^2}{45}} = \sqrt{\frac{6400}{60} + \frac{10000}{45}}$$

$$= \sqrt{106.67 + 222.22} = \sqrt{328.89} \approx \text{R\$ } 18.135$$

Exemplo

2 - Calcular Estatísticas

Erro padrão da diferença = R$ 18.135

$$IC = \bar{x} \pm Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

550 = 520 + Z * 18.135 >> Z = 1.654 >> α = 0.049

0.049 < 0.05 (intervalo de confiança sugerido)

NÃO rejeitamos a H0, logo:

Com um nível de significância de 5%, não há evidências estatísticas suficientes para concluir que existe uma diferença significativa no gasto médio diário entre turistas nacionais e estrangeiros em Pernambuco, dado os desvios padrão populacionais conhecidos.

Exemplo 3

Uma pesquisa de mercado em Recife, Pernambuco, está investigando o tempo médio que consumidores gastam em dois grandes centros de compras.Estudos históricos de comportamento do consumidor indicam os seguintes desvios padrão populacionais para o tempo de permanência:

Shopping RioMar (População 1): Desvio padrão populacional ($\sigma1$) = 35 minutos

Shopping Recife (População 2): Desvio padrão populacional ($\sigma2$) = 40 minutos

A pesquisadora decide coletar amostras aleatórias de visitantes para estimar a diferença no tempo médio de permanência:

Amostra de n1=100 visitantes do Shopping RioMar, com tempo médio de permanência = 185 minutos.

Amostra de n2=80 visitantes do Shopping Recife, com tempo médio de permanência = 170 minutos.

A pesquisadora quer testar se o tempo médio de permanência no Shopping RioMar é maior do que no Shopping Recife. Ela vai usar um nível de significância de α=0.05 (5%).

Exemplo 3

Passo 1: Formular as Hipóteses (H0 e H1)

H0: O tempo médio de permanência no RioMar NÃO é maior que no Shopping Recife. Ou seja, a diferença (RioMar - Recife) é menor ou igual a zero.

H0: mu1 - mu2 <= 0

H1: O tempo médio de permanência no RioMar É maior que no Shopping Recife. Ou seja, a diferença (RioMar - Recife) é maior que zero.

H1: mu1 - mu2 > 0

```python
import numpy as np
from scipy import stats

# --- Dados do Problema ---
# População 1 (Shopping RioMar)
sigma1 = 35.0  # Desvio padrão populacional conhecido
n1 = 100       # Tamanho da amostra
mu_amostral1 = 185.0 # Média amostral

# População 2 (Shopping Recife)
sigma2 = 40.0  # Desvio padrão populacional conhecido
n2 = 80        # Tamanho da amostra
mu_amostral2 = 170.0 # Média amostral
alpha = 0.05   # Nível de significância

# Calcular a diferença observada entre as médias amostrais
diff_mu_amostral = mu_amostral1 - mu_amostral2
# Calcular o Erro Padrão da Diferença entre Duas Médias (sigma conhecido)
se_diff = np.sqrt((sigma1**2 / n1) + (sigma2**2 / n2))
# O valor hipotetizado da diferença sob H0 é D0 = 0
D0 = 0
# Calcular o Z-score
z_calculated = (diff_mu_amostral - D0) / se_diff
# Para um teste unilateral à direita, o p-valor é P(Z > z_calculated)
p_value = 1 - stats.norm.cdf(z_calculated)

if p_value <= alpha:
    print(f"Decisão: REJEITAR H0 (pois {p_value:.4f} <= {alpha})")
else:
    print(f"Decisão: NÃO REJEITAR H0 (pois {p_value:.4f} > {alpha})")
```

Obrigada!
Bons estudos

💪