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/

Estatística Inferencial

é o ramo da estatística que se dedica a fazer inferências ou generalizações sobre uma população inteira a partir da análise de uma amostra dessa população.

Com base no que observei nesta pequena parte (amostra), o que posso dizer sobre o todo (população)?

Objetivos

- **Estimar parâmetros populacionais:**: Utilizar estatísticas amostrais para obter uma estimativa do valor de um parâmetro da população
- **Testar hipóteses sobre a população:** Formular hipóteses sobre características da população e usar os dados da amostra para determinar se há evidências suficientes para rejeitar ou não essas hipóteses.
- **Prever tendências e comportamentos:** fazer previsões sobre valores futuros ou comportamentos da população com base nos padrões identificados na amostra.

IMPORTANTE!

A validade dessas conclusões depende crucialmente da qualidade da amostra e da adequação dos métodos estatísticos utilizados.

Amostragem e Distribuições Amostrais

Amostragem

Processo de selecionar uma amostra (um subconjunto) de indivíduos, objetos ou dados de uma população maior para inferir características sobre toda a população.

Ao invés de analisar cada elemento da população (o que muitas vezes é impraticável ou impossível), a amostragem permite coletar informações de uma porção menor e, com os métodos estatísticos adequados, generalizar esses achados para o grupo maior.

Tipos de Amostragem

1) **Amostragem Probabilística**: Cada membro da população tem uma probabilidade conhecida de ser selecionado para a amostra. Os principais tipos de amostragem probabilística incluem:

    a) **Amostragem Aleatória Simples (AAS)**: Cada indivíduo da população tem a mesma chance de ser selecionado, e cada possível amostra do mesmo tamanho tem a mesma probabilidade de ser escolhida. É como um sorteio.

    b) **Amostragem Sistemática**: Os indivíduos são selecionados em intervalos regulares a partir de um ponto de partida aleatório.

Tipos de Amostragem

c) Amostragem Estratificada: A população é dividida em subgrupos homogêneos com base em alguma característica (idade, sexo, escolaridade, etc.), e então uma amostra aleatória simples é retirada de cada estrato.

d) Amostragem por Conglomerados: A população é dividida em grupos (conglomerados), como bairros ou escolas, e alguns desses conglomerados são selecionados aleatoriamente. Todos os indivíduos dentro dos conglomerados selecionados são incluídos na amostra ou uma amostra aleatória é retirada deles.

Tipos de Amostragem

2) Amostragem Não Probabilística: A seleção dos membros da amostra não é baseada em probabilidades conhecidas.
Possuem um teor subjetivo na escolha dos elementos amostrais
Inferências sobre a população devem ser feitas com cautela. Os principais tipos incluem amostragem por conveniência, por julgamento, por cotas e bola de neve.

Viés de amostragem

ocorre quando a amostra selecionada de uma população não reflete adequadamente as características dessa população

O que pode causar?

- **Estimativas Incorretas:** Uma amostra enviesada pode levar a estimativas de parâmetros populacionais de forma errada
- **Conclusões Errôneas**: Testes de hipóteses realizados com dados de uma amostra enviesada podem levar a conclusões falsas sobre a população.
- **Generalizações Inválidas:** Os resultados obtidos da amostra não podem ser generalizados com confiança para a população de interesse.

Tipos

- **Viés de Seleção:** Quando favorece certos grupos ou indivíduos em detrimento de outros.
  - **Viés de Cobertura Insuficiente:** Alguns membros da população têm uma probabilidade nula ou baixa de serem selecionados. Exemplo: Realizar uma pesquisa por telefone em uma população com muitos moradores sem telefone fixo.
  - **Viés de Sobrecobrimento:** Membros da população são incluídos na amostra mais de uma vez. Exemplo: Listas de e-mail duplicadas sendo usadas para uma pesquisa online.

Tipos

- **Viés de Seleção**
  - **Viés de Autoseleção:** Os participantes da amostra se oferecem para participar, e aqueles que se voluntariam podem ter características diferentes daqueles que não o fazem. Exemplo: Pesquisas online não controladas ou enquetes em programas de rádio.
  - **Viés de Sobrevivência:** A amostra é composta apenas por indivíduos ou itens que "sobreviveram" a algum processo, ignorando aqueles que não sobreviveram e que poderiam ter características diferentes. Exemplo: Estudar o sucesso de empresas que ainda estão ativas, ignorando as que faliram.

Tipos

- **<u>Viés de Não Resposta:</u>** Ocorre quando indivíduos selecionados para a amostra não participam da pesquisa, e aqueles que não respondem podem diferir significativamente daqueles que respondem em relação às variáveis de interesse.

- **<u>Viés de Mensuração (ou Resposta):</u>** Envolve problemas na forma como as perguntas são feitas, a influência do entrevistador ou a tendência dos participantes a fornecerem respostas socialmente desejáveis ou imprecisas.

Como Evitar o Viés de Amostragem?

1. **Definir Claramente a População Alvo:** Certifique-se de ter uma compreensão precisa de quem ou o quê constitui a população de interesse para o seu estudo.
2. **Criar um Quadro de Amostragem Abrangente:** Desenvolva uma lista o mais completa e atualizada possível de todos os membros da população.
3. **Utilizar Amostragem Aleatória Simples (AAS):** Se possível e prático, use a AAS para garantir que cada membro tenha a mesma chance de ser selecionado. Implemente o método adequado!

Como Evitar o Viés de Amostragem?

4. **Maximizar a Taxa de Resposta:** Questionários claros e concisos, oferecer incentivos, utilizar diferentes métodos de coleta de dados para alcançar um público mais amplo.
5. **Realizar Análises de Viés:** Se houver suspeita de viés, tente analisar se existem diferenças significativas entre a amostra e a população em características conhecidas.

Distribuição Amostral de uma Estatística

Distribuição de probabilidade de uma estatística (como a média amostral, a proporção amostral, a variância amostral, etc.) que é calculada a partir de todas as possíveis amostras aleatórias do mesmo tamanho retiradas de uma determinada população.

Distribuição Amostral de uma Estatística

1. **Seleção de Amostras:** Retiramos um grande número de amostras independentes e aleatórias, todas com o mesmo tamanho (n), de uma população específica.
2. **Cálculo da Estatística:** Para cada uma dessas amostras, calculamos a estatística de interesse (por exemplo, a média aritmética da amostra).
3. **Distribuição dos Valores:** Se plotarmos a frequência desses valores da estatística amostral em um histograma ou construirmos sua distribuição de probabilidade teórica, obteremos a distribuição amostral dessa estatística.

Distribuição Amostral de uma Estatística

- **Dependência da População e do Tamanho da Amostra:** A forma, a média e o desvio padrão da distribuição amostral de uma estatística dependem das **características da população original** (sua distribuição, média e variância) e do **tamanho da amostra** (n).
- **Erro Padrão:** O desvio padrão da distribuição amostral de uma estatística é conhecido como erro padrão dessa estatística. Ele mede a variabilidade da estatística amostral de amostra para amostra. Um erro padrão menor indica que as estatísticas amostrais tendem a estar mais próximas do parâmetro populacional.

Teorema do Limite Central

A distribuição das médias amostrais se aproxima de uma distribuição normal à medida que o tamanho da amostra aumenta, **independentemente da forma da distribuição da população original.**

Esse teorema é fundamental para a inferência estatística.

Distribuição Amostral da Média Amostral

➔ **Média: A média da distribuição amostral da média é igual à média da população.**

$$\mu_{\bar{x}} = \mu$$

➔ **Desvio Padrão (Erro Padrão da Média):**

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

onde σ é o desvio padrão da população e n é o tamanho da amostra.

Distribuição Amostral da Média Amostral

Imagine que a altura de todos os estudantes adultos da UFPE segue uma distribuição normal com uma média populacional (μ) de 170 cm e um desvio padrão populacional (σ) de 10 cm. Essa é a nossa população.

Vamos supor que um pesquisador decide retirar muitas amostras aleatórias independentes de um tamanho fixo de n=25 estudantes dessa população. Para cada amostra, ele calcula a média da altura amostral.

1. Média da Distribuição Amostral: A média de todas as possíveis médias amostrais será igual à média da população:

$$\mu_{\bar{x}} = \mu = 170 \text{ cm}$$

Distribuição Amostral da Média Amostral

2. **Desvio Padrão da Distribuição Amostral (Erro Padrão da Média): O desvio padrão das médias amostrais (o erro padrão) será o desvio padrão populacional dividido pela raiz quadrada do tamanho da amostra:**

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{10 \text{ cm}}{\sqrt{25}} = \frac{10}{5} = 2 \text{ cm}$$

Estimativa de parâmetros

Consiste no processo de usar dados de uma amostra para inferir os valores de características desconhecidas de uma população.

Como em muitos casos é impraticável medir uma população inteira, dependemos de amostras para fazer essas "adivinhações".

Estimativa Pontual

- É a forma mais direta de estimar um parâmetro. Ela consiste em utilizar uma única estatística calculada a partir da amostra como o "melhor palpite" ou o valor mais provável para o parâmetro populacional correspondente.
- **Exemplos:**
  - A **média amostral** é a estimativa pontual mais comum para a média populacional (μ). Por exemplo, se a média salarial em uma amostra de trabalhadores em Recife é R$ 2.500, essa é a estimativa pontual para a média salarial de todos os trabalhadores na cidade.
  - A **proporção amostral** é a estimativa pontual para a proporção populacional (P). Por exemplo, se 60% dos turistas em uma amostra de Olinda se dizem satisfeitos, essa é a estimativa pontual para a proporção de todos os turistas satisfeitos.

Estimativa Pontual

- **Propriedades de um bom estimador pontual:**
  - **Não-viesado**: Significa que, em média, o estimador não superestima nem subestima o parâmetro verdadeiro.
  - **Eficiente**: Um estimador eficiente tem a menor variabilidade possível entre os estimadores não-viesados. Isso significa que suas estimativas tendem a estar mais próximas do valor real do parâmetro.
  - **Consistente**: À medida que o tamanho da amostra aumenta, a estimativa pontual tende a se aproximar cada vez mais do verdadeiro valor do parâmetro.

Estimativa Pontual

- **Limitação:**
  - A estimativa pontual fornece apenas um único valor.
  - Não nos diz nada sobre a precisão dessa estimativa ou o quão "próxima" ela provavelmente está do verdadeiro parâmetro populacional.

Intervalo de Confiança (IC)

- Aborda a limitação da estimativa pontual, fornecendo uma faixa de valores (um intervalo) dentro da qual o parâmetro populacional tem uma alta probabilidade de estar, juntamente com um nível de confiança associado a essa probabilidade.
- O IC é uma medida da incerteza em torno de uma estimativa de parâmetro.

- **Exemplos:**
  - Estimativa pontual - A média é 58.3 anos
  - Intervalo de confiança - A média está entre 56.5 e 60.1 com 95% de confiança

Intervalo de Confiança

- Componentes:
  - **Estimativa Pontual**: O centro do intervalo
  - **Margem de Erro**: É a distância do centro do intervalo até cada extremidade, que quantifica a incerteza da estimativa.
  - **Nível de Confiança**: Representa a probabilidade de que o intervalo construído contenha o verdadeiro parâmetro populacional.

$$\text{Intervalo de Confiança} = \text{Média Amostral} \pm \text{Margem de Erro}$$

Cálculo do Intervalo de Confiança para a média populacional

- Depende principalmente de dois fatores:
  - Se o desvio padrão populacional (σ) é conhecido ou desconhecido.
  - O tamanho da amostra (n).

Caso 1: Desvio Padrão Populacional (σ) Conhecido

- Quando o desvio padrão da população (σ) é conhecido, a distribuição amostral da média é considerada normal (graças ao Teorema do Limite Central para n≥30, ou se a população original já for normal). Usamos a **distribuição Z** para encontrar o valor crítico.

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

$\bar{x}$ = média amostral

$Z_{\alpha/2}$ = valor crítico da distribuição normal padrão para o nível de confiança desejado (α é o nível de significância, $1 - \alpha$ é o nível de confiança).

σ = desvio padrão populacional

n = tamanho da amostra

Caso 1: Desvio Padrão Populacional (σ) Conhecido

➔ Exemplo 1:

Em uma pesquisa sobre o tempo de espera (em minutos) em uma fila de um órgão público em Recife, sabe-se por estudos anteriores que o desvio padrão populacional (σ) é de 5 minutos. Uma amostra de 49 pessoas revelou um tempo médio de espera (x) de 12 minutos. Calcule o intervalo de confiança de 95% para o tempo médio de espera.

$Z_{\alpha/2}$ para 95% de confiança é $Z_{0.025} = 1.96$.

$$IC = 12 \pm 1.96 \cdot \frac{5}{\sqrt{49}}$$

$$IC = 12 \pm 1.96 \cdot \frac{5}{7}$$

$$IC = 12 \pm 1.96 \cdot 0.7143$$

$$IC = 12 \pm 1.399$$

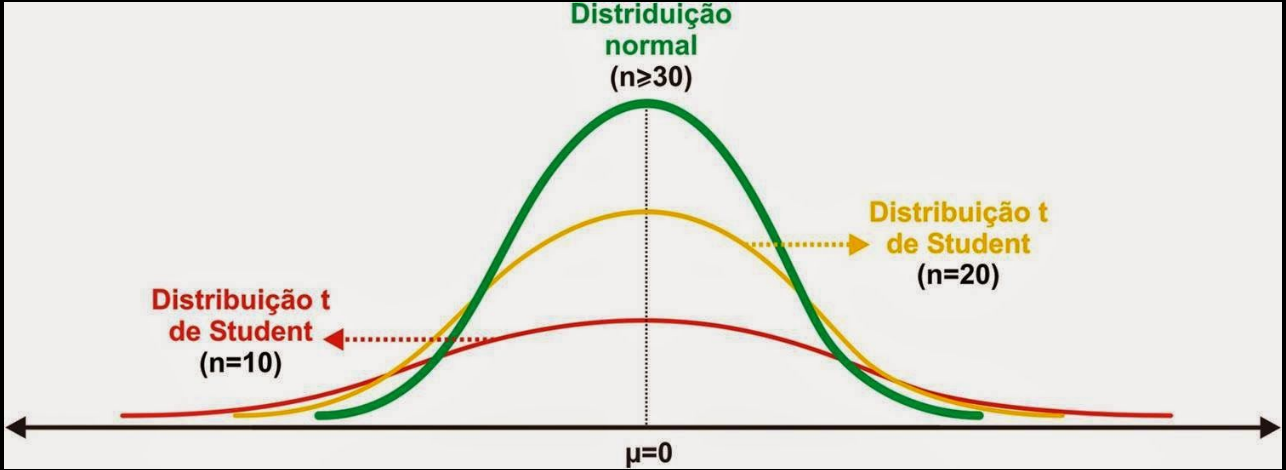
$$IC = [10.601, 13.399]$$

Caso 2: Desvio Padrão Populacional (σ) Desconhecido

- ➔ Este é o caso mais comum na prática, pois raramente conhecemos o desvio padrão de toda a população.
- ➔ Quando σ é desconhecido, nós o estimamos usando o desvio padrão amostral (s).
- ➔ No entanto, usar s em vez de σ introduz uma incerteza adicional, especialmente para amostras pequenas.
- ➔ Para levar em conta essa incerteza extra, utilizamos a distribuição t de Student em vez da distribuição Z.

Caso 2: Desvio Padrão Populacional (σ) Desconhecido

- Distribuição t de Student:
  - Distribuição de probabilidade simétrica, em forma de sino, semelhante à normal padrão, mas com maior dispersão. Sua forma varia com os graus de liberdade (gl = n-1). À medida que o tamanho da amostra (n) aumenta, a distribuição t se aproxima da distribuição normal padrão.



Caso 2: Desvio Padrão Populacional (σ) Desconhecido

$$IC = \bar{x} \pm t_{n-1,\alpha/2} \cdot \frac{s}{\sqrt{n}}$$

$\bar{x}$ = média amostral

$t_{n-1,\alpha/2}$ = valor crítico da distribuição t de Student com $n-1$ graus de liberdade e nível de confiança desejado.

s = desvio padrão amostral

n = tamanho da amostra

Caso 2: Desvio Padrão Populacional (σ) Desconhecido

➔ Exemplo 2:

Um restaurante em Boa Viagem quer estimar o valor médio que os clientes gastam por refeição. Uma amostra de 30 clientes (n=30) revela um gasto médio de R$ 65.00 e um desvio padrão amostral (s) de R$ 12.00. Calcule o intervalo de confiança de 90% para o gasto médio populacional.

Graus de liberdade (gl) = $n - 1 = 30 - 1 = 29$.

$t_{n-1,\alpha/2}$ para 90% de confiança (e $gl = 29$) é $t_{29,0.05} = 1.699$. (Você buscaria esse valor em uma tabela t de Student).

$$IC = 65 \pm 1.699 \cdot \frac{12}{\sqrt{30}}$$

$$IC = 65 \pm 1.699 \cdot \frac{12}{5.477}$$

$$IC = 65 \pm 1.699 \cdot 2.191$$

$$IC = 65 \pm 3.722$$

$$IC = [61.278, 68.722]$$

Fatores que afetam o Intervalo de Confiança

- **Nível de Confiança**: Um nível de confiança maior (por exemplo, 99% em vez de 95%) resultará em um intervalo mais largo. Para ter mais certeza de que o intervalo contém o parâmetro, precisamos de uma faixa maior.
- **Tamanho da Amostra** (n): Um tamanho de amostra maior resultará em um erro padrão menor. Um erro padrão menor, por sua vez, leva a um intervalo mais estreito e, portanto, a uma estimativa mais precisa.
- **Variabilidade dos Dados**: Uma maior variabilidade na população (maior σ) resultará em um erro padrão maior e, consequentemente, em um intervalo mais largo

Exemplo 1

Intervalo de Confiança para a Média (Desvio Padrão Populacional σ Conhecido)

Imagine que, por experiência passada com o turismo em Recife, sabemos que o desvio padrão do valor gasto por dia por um turista é de R$ 50,00 (σ=50). Coletamos uma amostra de 100 turistas (n=100) e a média de gastos diários nesta amostra foi de R$ 320,00. Queremos construir um IC de 95% para o gasto médio diário de todos os turistas.

Exemplo 1

```python
import numpy as np
from scipy import stats # Para distribuição normal e t-student

# Dados
media_amostral = 320.0
desvio_padrao_populacional = 50.0
tamanho_amostra = 100
nivel_confianca = 0.95

# 1. Calcular o erro padrão da média
erro_padrao = desvio_padrao_populacional / np.sqrt(tamanho_amostra)

# Usando a função stats.norm.interval (mais direta)
# Essa função já faz o cálculo do erro padrão e usa o valor crítico Z
intervalo_z_direto = stats.norm.interval(nivel_confianca,
                                         loc=media_amostral,
                                         scale=erro_padrao)
print(f"Intervalo de Confiança (Direto com norm.interval): [R$ {intervalo_z_direto[0]:.2f}, R$ {intervalo_z_direto[1]:.2f}]")
```

Exemplo 1

Função interval

```
intervalo_z = stats.norm.interval(nivel_confianca, loc=media_amostral, scale=erro_padrao)
```

Exemplo 2

Intervalo de Confiança para a Média (Desvio Padrão Populacional σ Desconhecido)

Uma amostra de 30 apartamentos em Boa Viagem (Recife) foi selecionada para estimar o valor médio de aluguel. A média amostral foi de R$ 3500,00 e o desvio padrão amostral (s) foi de R$ 700,00. Queremos construir um IC de 90% para o valor médio de aluguel de todos os apartamentos em Boa Viagem.

Exemplo 2

```python
media_amostral_t = 3500.0
desvio_padrao_amostral = 700.0 # 's' - Sigma desconhecido
tamanho_amostra_t = 30
nivel_confianca_t = 0.90

# 1. Calcular os graus de liberdade
graus_liberdade = tamanho_amostra_t - 1

# 2. Calcular o erro padrão da média (usando s)
erro_padrao_t = desvio_padrao_amostral / np.sqrt(tamanho_amostra_t)

# Usando a função stats.t.interval
# Essa função já calcula graus de liberdade, erro padrão e usa o valor crítico t
intervalo_t_direto = stats.t.interval(nivel_confianca_t,
                                      df=graus_liberdade,
                                      loc=media_amostral_t,
                                      scale=erro_padrao_t)
print(f"Intervalo de Confiança (Direto com t.interval): [R$ {intervalo_t_direto[0]:.2f}, R$ {intervalo_t_direto[1]:.2f}]")
```

Exemplo 2

Função interval

```
intervalo_t = stats.t.interval(nivel_confianca_t, df=graus_liberdade, loc=media_amostral_t, scale=erro_padrao_t)
```

Cálculo do Intervalo de Confiança para a proporção populacional

- Técnica estatística que nos permite estimar a proporção real (P) de uma característica em toda uma população, com base nos dados de uma amostra
- Por Que Usar Intervalos de Confiança para Proporções?
  - A proporção de eleitores em Pernambuco que apoiam um determinado candidato.
  - A proporção de produtos defeituosos em um grande lote fabricado.
  - A proporção de clientes satisfeitos com um novo serviço.

Intervalo de Confiança

- Componentes:
  - **Proporção Amostral**: É a estimativa pontual da proporção populacional
  - **Margem de Erro**: É a distância do centro do intervalo até cada extremidade, que quantifica a incerteza da estimativa.
  - **Nível de Confiança**: Representa a probabilidade de que o intervalo construído contenha o verdadeiro parâmetro populacional.

Intervalo de Confiança

$$IC = \hat{p} \pm Z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$\hat{p}$ = proporção amostral

$Z_{\alpha/2}$ = valor crítico da distribuição normal padrão (baseado no nível de confiança)

n = tamanho da amostra

$\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ = Erro Padrão da Proporção Amostral (estimado, já que P é desconhecido e é substituído por $\hat{p}$)

Intervalo de Confiança

Exemplo:

- Uma pesquisa de opinião em João Pessoa entrevistou 400 adultos selecionados aleatoriamente e descobriu que 220 deles aprovam a nova gestão municipal. Construa um intervalo de confiança de 95% para a proporção de todos os adultos na cidade que aprovam a gestão.

$$\hat{p} = 220/400 = 0.55$$

Para 95% de confiança, $\alpha = 0.05$, então $\alpha/2 = 0.025$.

. O valor crítico $Z_{0.025}$ é 1.96.

$$\sigma_{\hat{p}} = \sqrt{\frac{0.55(1-0.55)}{400}} \approx 0.02487$$

$$IC = 0.55 \pm 0.0487$$

$$IC = [0.55 - 0.0487, 0.55 + 0.0487]$$

$$IC = [0.5013, 0.5987]$$

Exemplo 3

Intervalo de Confiança para a Proporção

A Companhia de Transportes Urbanos do Grande Recife (CTTU) deseja estimar a proporção de passageiros que estão satisfeitos com os serviços de ônibus na cidade. Eles realizaram uma pesquisa com uma amostra aleatória de 500 passageiros. Desses, 320 passageiros declararam-se satisfeitos.

Construa um intervalo de confiança de 95% para a verdadeira proporção de passageiros satisfeitos com o transporte público em Recife.

Exemplo 3 - usando stats

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

# --- Dados do Exercício ---
num_satisfeitos = 320  # Número de "sucessos" na amostra
tamanho_amostra = 500  # Tamanho total da amostra
nivel_confianca = 0.95 # Nível de confiança desejado (95%)

# --- Calcular a Proporção Amostral (p_amostral) ---
p_amostral = num_satisfeitos / tamanho_amostra

# --- Calcular o Intervalo de Confiança ---
# 1.Encontrar o valor crítico Z
valor_critico_z = stats.norm.ppf(1 - (1 - nivel_confianca) / 2)
# 2. Calcular o Erro Padrão da Proporção (usando p_amostral como estimativa de P)
# A fórmula do erro padrão da proporção é sqrt(P*(1-P)/n).
erro_padrao_p = np.sqrt((p_amostral * (1 - p_amostral)) / tamanho_amostra)
# 3. Calcular a Margem de Erro
margem_erro_p = valor_critico_z * erro_padrao_p
# 4. Calcular os Limites do Intervalo de Confiança
limite_inferior = p_amostral - margem_erro_p
limite_superior = p_amostral + margem_erro_p


print(f"IC de {int(nivel_confianca*100)}% para a proporção populacional: [{limite_inferior:.4f}, {limite_superior:.4f}]")
print(f"Ou seja, entre {limite_inferior*100:.2f}% e {limite_superior*100:.2f}%")
```

Exemplo 3 - usando statsmodels

```
import numpy as np
from statsmodels.stats.proportion import proportion_confint

# --- Dados do Exercício ---
num_satisfeitos = 320  # Número de "sucessos" na amostra
tamanho_amostra = 500  # Tamanho total da amostra
nivel_confianca = 0.95 # Nível de confiança desejado (95%)

# --- Calcular o Intervalo de Confiança ---
limite_inferior_sm, limite_superior_sm = proportion_confint(
    count=num_satisfeitos,
    nobs=tamanho_amostra,
    alpha=1 - nivel_confianca, # alpha é o nível de significância (1 - nível de confiança)
    method="normal"
)

print(f"IC de {int(nivel_confianca*100)}% para a proporção populacional: [{limite_inferior_sm:.4f}, {limite_superior_sm:.4f}]")
print(f"Ou seja, entre {limite_inferior_sm*100:.2f}% e {limite_superior_sm*100:.2f}%")
```

Obrigada!
Bons estudos
💪