Aula 02. Distribuições Amostrais e o Teorema Central do Limite

Estatística Inferencial

MBA CDIA

Prof. Dr. Allan Quadros
ENAP - Escola Nacional de Administração Pública

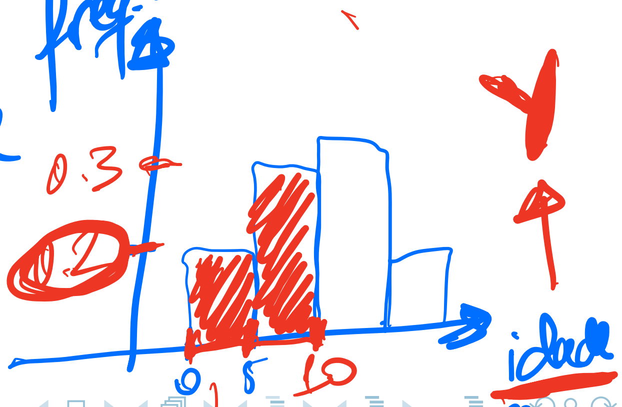
2025

pdf

X = renda

v.a. discreta

pmf



v.a. contínua

total de prods?

$P(X \le 120k)$

0 — 1

certo

$P(X \le 1 mi) = 1$

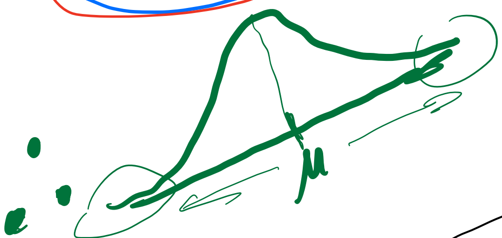
$= 0.2$

$= 0.3$

$P(Y \le 10)$

$= \sum_{i=1}^{z} p(w) \cdot w$

freq. rel.



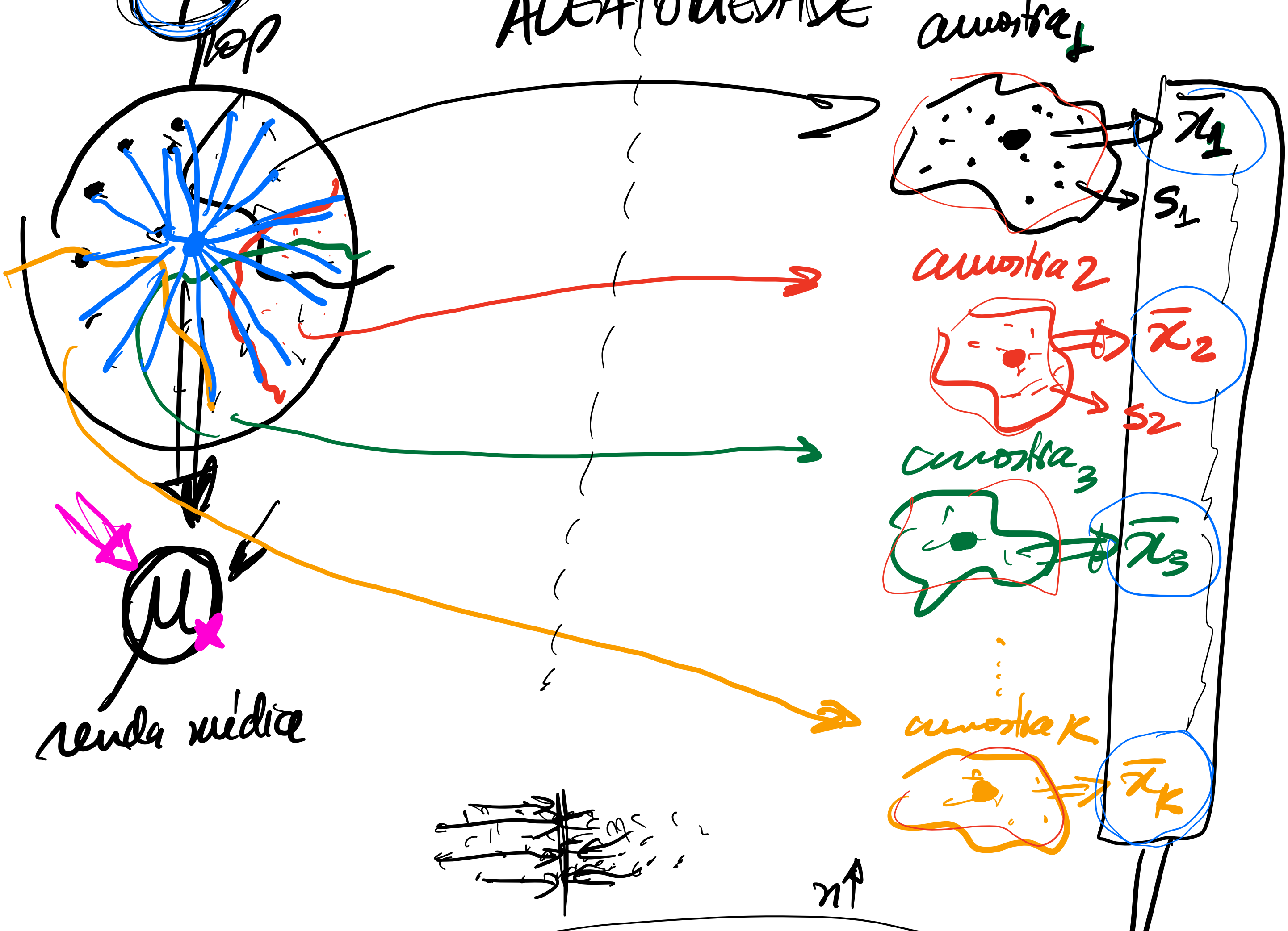
Renda $= X$

renda

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

todas c/ tamanho "n"

$n \uparrow$



$$\mu_{\bar{x}} = \frac{\bar{x}_1 + \bar{x}_2 + \bar{x}_3 + \dots + \bar{x}_{10000}}{10000} = \mu_x ?$$

V. a.

$$\sigma_{\bar{x}}^2 = \frac{\sum_{i=1}^{10000} (\bar{x}_i - \mu_{\bar{x}})^2}{n-1}$$

$$\frac{\sigma_x^2}{n}$$

$$\bar{X} \sim N\left(\mu_{\bar{x}} = \mu_x , \sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}}\right)$$

$\therefore \sigma_{\bar{x}}^2 = \frac{\sigma_x^2}{n}$

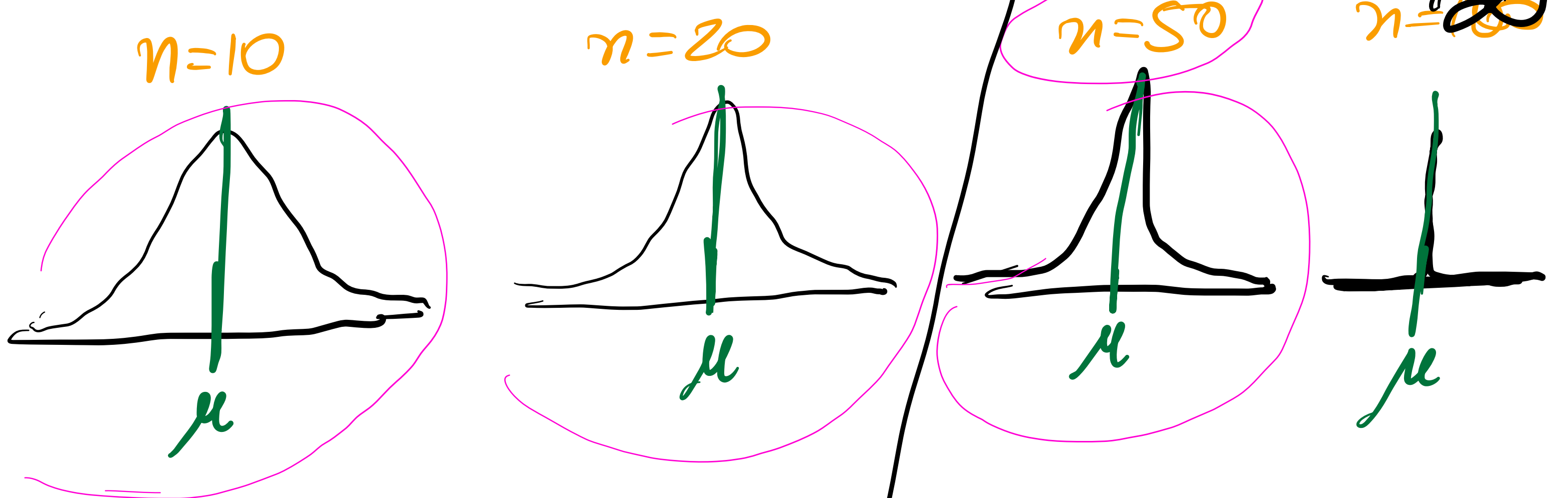
$\therefore$ se $X \sim N(\mu_x, \sigma_x)$

e se $X \sim (?)$?

↳ pelo TCL, se "n" for suficientemente grande ($n > 30$), então:

pop pop

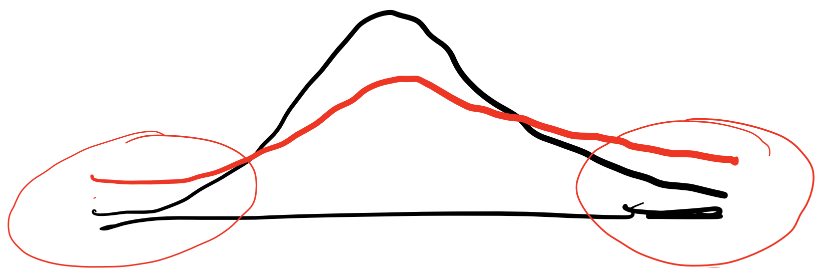
$$\bar{X} \sim N\left(\mu_{\bar{x}} = \mu_x , \boxed{\sigma_{\bar{x}}} = \frac{\sigma_x}{\sqrt{n}}\right)$$



∴ quando $\sigma_{\bar{x}}$ é desconhecido:

(TCL)

$$\bar{X} \sim \cancel{N}\left(\mu_{\bar{x}} = \mu_X \, , \, \boxed{\sigma_{\bar{x}}} = \frac{s}{\sqrt{n}}\right)$$



$$\bar{X} \sim t_{g.l.}$$

$g.l. = n-1$

$\Updownarrow$

$$\boxed{s^2 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}}$$

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i-\mu)^2}{N}$$

$\bar{x} = 3$

5

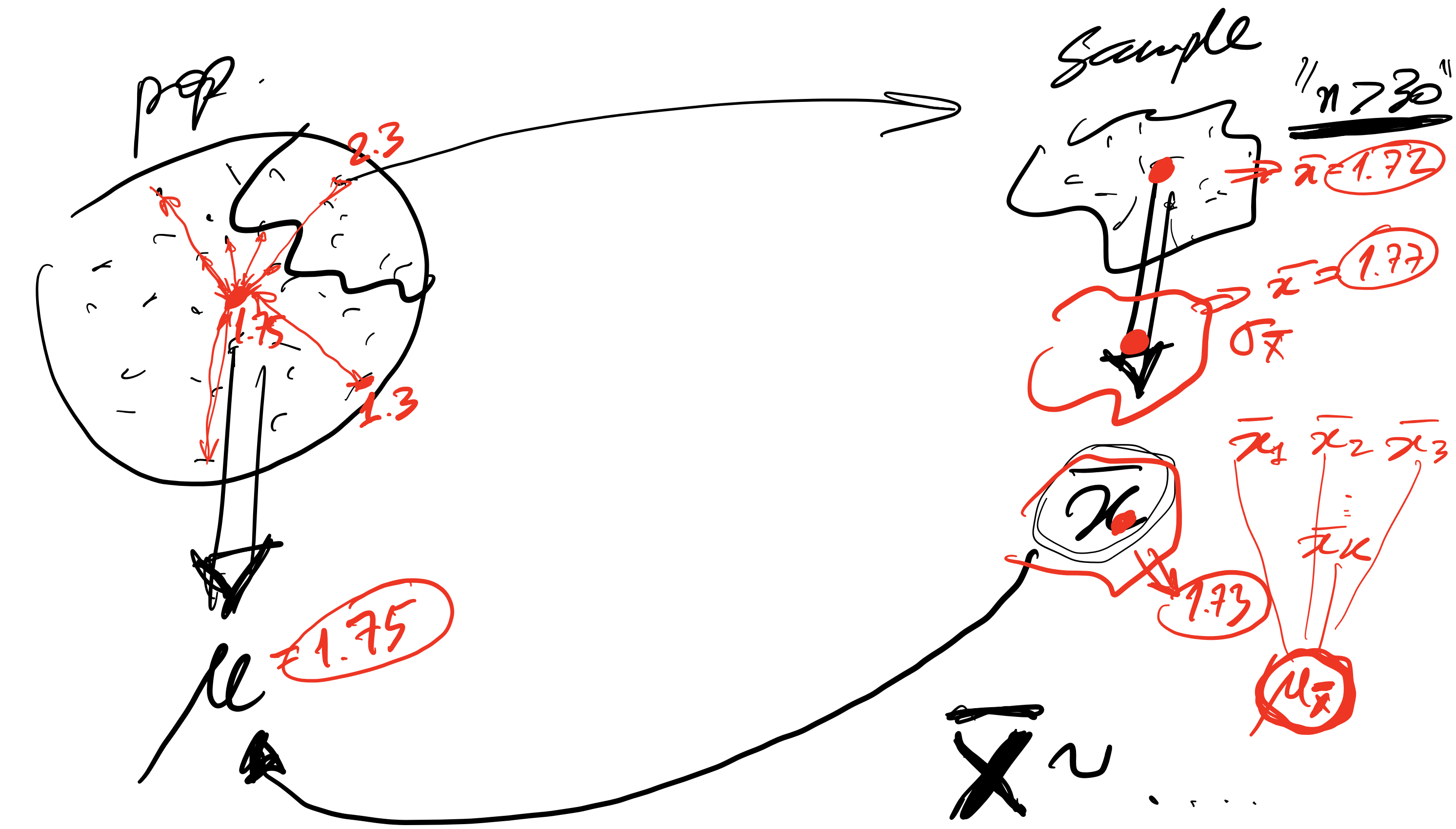
1 2 3 4 $\boxed{?}$

$s^2 \to \sigma^2$

$$\bar{x} = \frac{1+2+3+4+?}{5} \Rightarrow \frac{10+x}{5} = 3 \Rightarrow 10+x = 15$$

$x = 15-10$

$x = 5$



① se distrib. pop é Normal
$(X \sim N(\mu_X, \sigma_X))$

$$\sigma_{\bar{X}}^2 = \frac{\sigma_X^2}{n}$$

então $\bar{X} \sim N\left(\mu_{\bar{X}} = \mu_X, \sigma_{\bar{X}} = \frac{\sigma_X}{\sqrt{n}}\right)$

② se distrib. pop é desconhecida $(X \sim ?)$ ou distrib. pop. é sabidamente não-Normal $(X \not\sim N)$, mas "n" é suficientemente grande $(n > 30)$, então, pelo TCL, podemos assumir

$$\bar{X} \sim N\left(\mu_{\bar{X}} = \mu_X, \sigma_{\bar{X}} = \frac{\sigma_X}{\sqrt{n}}\right)$$

$n > 30$

$\bar{X}$

μ

$\bar{x}_1 = 1.7$

68%

~95%

$\mu-2\sigma$ $\mu-\sigma$ μ $\mu+\sigma$ $\mu+2\sigma$

obrigado

5% 95%

- Um dos objetivos da estatística é estimar o valor verdadeiro de um parâmetro populacional desconhecido sem examinar toda a população.

~99%

2σ -2σ

- Lembre-se que um **parâmetro** é um número que descreve uma população.

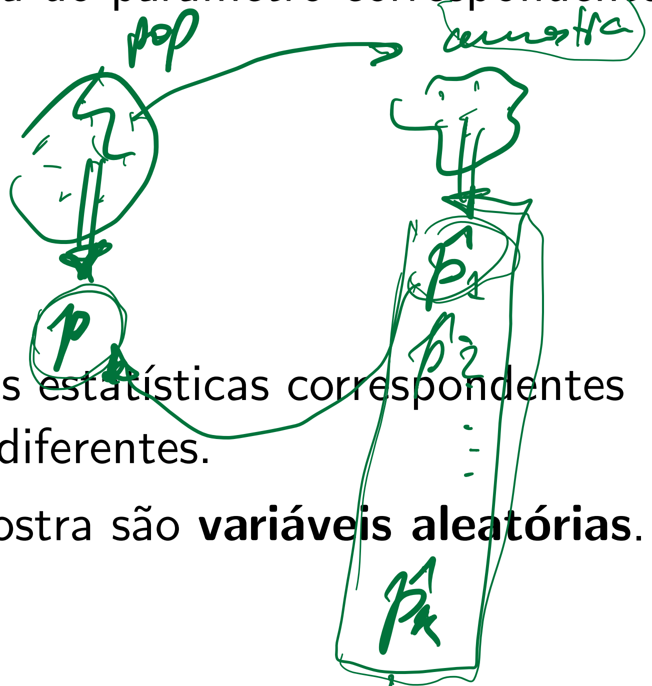
Exemplos de parâmetros:

- $\mu =$ média populacional – o ponto de equilíbrio ou média de longo prazo da população
- $\sigma^2 =$ variância populacional – a distância quadrática média da população em relação a μ
- $p =$ proporção populacional – a proporção de unidades que possuem certa propriedade

- Para gerar informação sobre um parâmetro, extraímos uma amostra da população e calculamos uma **estatística** para usar como estimativa do parâmetro correspondente.

Exemplos de estatísticas:

- $\bar{x}$ = média amostral – uma estimativa de μ
- s^2 = variância amostral – uma estimativa de σ^2
- $\hat{p}$ = proporção amostral – uma estimativa de p
- Se várias amostras fossem extraídas da população e as estatísticas correspondentes fossem calculadas, elas provavelmente teriam valores diferentes.
- Portanto, todas as estatísticas calculadas de uma amostra são **variáveis aleatórias**.



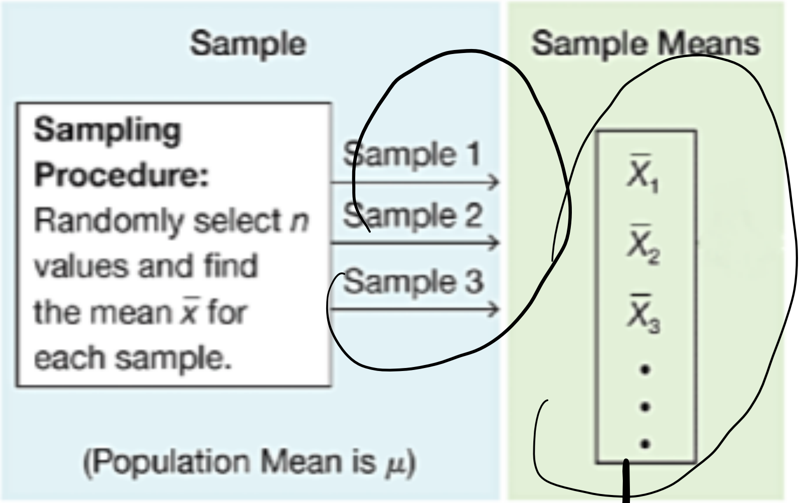
Distribuições Amostrais

v.a.: $\boxed{\hat{p}} \sim ?$

- Lembre-se que toda variável aleatória tem uma distribuição de probabilidade associada.
- A distribuição de probabilidade de uma estatística, em particular, é chamada de **distribuição amostral**.
- Estudaremos algumas propriedades das distribuições amostrais, como a média, desvio padrão e forma da distribuição.
- Vamos focar primeiro na distribuição amostral da média amostral.

Distribuição de médias amostrais

- O valor da média amostral varia cada vez que uma amostra é extraída, então é uma variável aleatória.
- Quando nos referimos a uma média amostral como variável aleatória, escrevemos $\bar{X}$.
- A distribuição de probabilidade de $\bar{X}$ é chamada de **distribuição amostral de** $\bar{X}$.



v.a. $\to \bar{X} \sim ?$

Seja $\bar{X}$ a média de uma amostra aleatória de tamanho n, extraída de uma população com média μ e desvio padrão σ.

Propriedades

- A **média** de $\bar{X}$ é μ. Ou seja:

$$\boxed{\mu_{\bar{X}} = \mu = \text{média populacional}}$$

- O **desvio padrão** de $\bar{X}$ é $\sigma/\sqrt{n}$. Ou seja:

$$\boxed{\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{\text{desvio padrão populacional}}{\sqrt{\text{tamanho da amostra}}}}$$

O desvio padrão $\sigma_{\bar{X}}$ também é chamado de **erro padrão da média**.
Note que o erro padrão $\sigma_{\bar{X}}$ **diminui** quando o tamanho da amostra n aumenta.

Exemplo 1

~~$\bar{X} \sim N$~~

a. Uma população tem média $\mu = 6$ e desvio padrão $\sigma = 4$. Encontre $\mu_{\bar{X}}$ e $\sigma_{\bar{X}}$ para um tamanho de amostra de $n = 25$.

$\mu_{\bar{x}} = \mu = 6$

$\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}} = \frac{4}{\sqrt{25}} = \frac{4}{5} = 0{,}8$

b. Uma população tem média $\mu = 17$ e desvio padrão $\sigma = 20$. Encontre $\mu_{\bar{X}}$ e $\sigma_{\bar{X}}$ para um tamanho de amostra de $n = 100$.

Solução - Exemplo 1

a. Para $\mu = 6$, $\sigma = 4$ e $n = 25$:

$$\mu_{\bar{X}} = \mu = 6$$

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{4}{\sqrt{25}} = \frac{4}{5} = 0.8$$

b. Para $\mu = 17$, $\sigma = 20$ e $n = 100$:

$$\mu_{\bar{X}} = \mu = 17$$

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{20}{\sqrt{100}} = \frac{20}{10} = 2$$

Resumo da distribuição amostral de $\bar{X}$

Média e desvio padrão da média amostral $\bar{X}$:

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

$$\boxed{\sigma_{\bar{X}}} = \frac{\sigma}{\boxed{\sqrt{n}}}$$

Observações importantes:

- O resultado acima é verdadeiro mesmo quando os valores verdadeiros de μ e σ são desconhecidos
- Este resultado é útil para fazer inferências sobre parâmetros populacionais desconhecidos
- O desvio padrão de $\bar{X}$, $\sigma_{\bar{X}}$, é sempre menor que o desvio padrão populacional, σ:

$$\boxed{\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \leq \sigma} \longleftrightarrow \sigma_{\bar{x}} \leq \sigma_x$$

- Isso implica que $\bar{X}$ é sempre menos dispersa que a distribuição populacional
- Na próxima aula, estudaremos a "forma" da distribuição amostral de $\bar{X}$

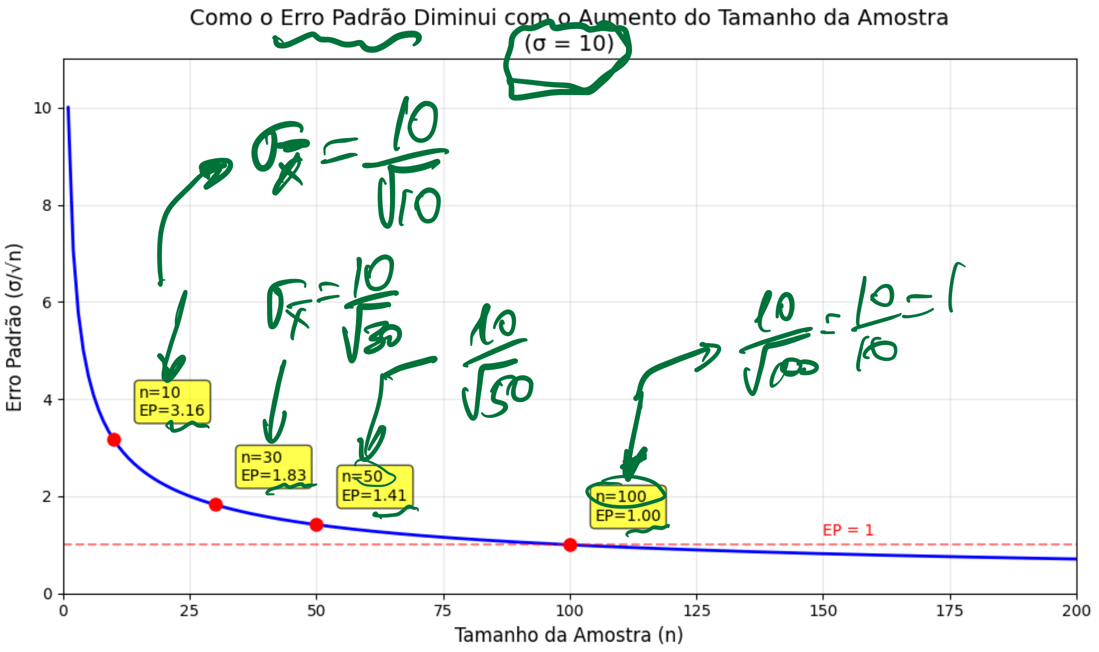
Aplicações Práticas no Setor Público

Por que isso é importante para a gestão pública?

- **Pesquisas de satisfação:** Ao avaliar a satisfação com serviços públicos, usamos amostras para estimar a satisfação média populacional
- **Controle de qualidade:** Para monitorar o tempo médio de atendimento em órgãos públicos
- **Planejamento orçamentário:** Estimar gastos médios por departamento usando amostras
- **Avaliação de políticas:** Medir o impacto médio de programas governamentais

Quanto maior a amostra, mais precisa é nossa estimativa!

Como o erro padrão $\boxed{\sigma_{\bar{X}}} = \frac{\sigma}{\sqrt{n}}$ diminui com o aumento de n:



Interpretação: Amostras maiores produzem estimativas mais precisas da média populacional.

1. Uma população de processos administrativos tem tempo médio de tramitação $\mu = 15$ dias e desvio padrão $\sigma = 6$ dias. Se extrairmos uma amostra de 36 processos:

a. Qual é a média esperada do tempo de tramitação na amostra? $\mu_{\bar{x}} = \mu_x = 15$ dias

b. Qual é o erro padrão dessa estimativa? $\boxed{\sigma_{\bar{x}}} = \frac{\sigma_x}{\sqrt{n}} = \frac{6}{\sqrt{36}} = \frac{6}{6} = 1 \rightarrow 0.5$

2. Para reduzir o erro padrão pela metade, por quanto devemos multiplicar o tamanho da amostra?

$0.5 = \frac{\sigma_x}{\sqrt{m \cdot n}} \rightarrow 0.5 = \frac{6}{\sqrt{m \cdot 36}} \rightarrow 0.5 = \frac{6}{\sqrt{m} \cdot \sqrt{36}} \rightarrow 0.5 = \frac{6}{\sqrt{m} \cdot 6} \rightarrow 0.5 = \frac{1}{\sqrt{m}} \rightarrow \frac{1}{2} = \frac{1}{\sqrt{m}}$

$\rightarrow (\sqrt{m})^2 = (2)^2 \rightarrow m = 4$

3. Por que a média amostral é menos variável que os valores individuais da população?

Pq $\sigma_{\bar{x}} < \sigma_x$?

Simulação de Distribuições Amostrais

Google Colab

Exercícios para fixação dos conceitos

- Simulação de médias amostrais
- Verificação do erro padrão
- Visualização de distribuições amostrais
- Efeito do tamanho da amostra

Exercício 1: Simulando Médias Amostrais

Enunciado: Crie uma população normal com $\mu = 100$ e $\sigma = 15$. Extraia 1000 amostras de tamanho $n = 30$ e calcule a média de cada amostra. Verifique se a média e desvio padrão das médias amostrais correspondem aos valores teóricos.

Dicas:

- Use `numpy.random.normal()` para gerar a população
- Use um loop para extrair as amostras
- Compare os resultados empíricos com $\mu_{\bar{X}} = \mu$ e $\sigma_{\bar{X}} = \sigma/\sqrt{n}$

Bibliotecas necessárias:

```
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
```

Gabarito - Exercício 1 - Parte 1

```
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd

# Parametros populacionais
mu = 100
sigma = 15
n = 30  # tamanho da amostra
num_amostras = 1000

# Simular distribuicao amostral
np.random.seed(42)
medias_amostrais = []

for i in range(num_amostras):
    amostra = np.random.normal(mu, sigma, n)
    media_amostra = np.mean(amostra)
    medias_amostrais.append(media_amostra)

medias_amostrais = np.array(medias_amostrais)

# Valores teoricos
mu_teorico = mu
sigma_teorico = sigma / np.sqrt(n)

# Valores empiricos
mu_empirico = np.mean(medias_amostrais)
sigma_empirico = np.std(medias_amostrais, ddof=0)
```

Gabarito - Exercício 1 - Parte 2

```
print(f"Valores Teoricos:")
print(f"Media das medias amostrais: {mu_teorico}")
print(f"Erro padrao: {sigma_teorico:.4f}")
print(f"\nValores Empiricos:")
print(f"Media das medias amostrais: {mu_empirico:.4f}")
print(f"Erro padrao: {sigma_empirico:.4f}")
print(f"\nDiferenca percentual do erro padrao: {abs(sigma_empirico - sigma_teorico)/sigma_teorico * 100:.2f}%")
```

Exercício 2: Visualização da Distribuição Amostral

Enunciado: Usando os dados do exercício anterior, crie um histograma das médias amostrais e sobreponha a curva normal teórica. Compare visualmente a distribuição empírica com a teórica.

Dicas:

- Use `plt.hist()` com `density=True`
- Use `scipy.stats.norm` para a curva teórica
- Adicione linhas verticais para as médias

Objetivo: Verificar visualmente que a distribuição das médias amostrais segue uma distribuição normal.

Gabarito - Exercício 2

```
from scipy import stats

# Criar histograma
plt.figure(figsize=(10, 6))
plt.hist(medias_amostrais, bins=30, density=True, alpha=0.7,
         color='skyblue', edgecolor='black', label='Distribuicao empirica')

# Curva normal teorica
x = np.linspace(medias_amostrais.min(), medias_amostrais.max(), 100)
y = stats.norm.pdf(x, mu_teorico, sigma_teorico)
plt.plot(x, y, 'r-', linewidth=2, label='Distribuicao teorica N({}, {:.2f})'.format(mu_teorico, sigma_teorico))

# Adicionar linhas verticais
plt.axvline(mu_empirico, color='blue', linestyle='--',
            label=f'Media empirica = {mu_empirico:.2f}')
plt.axvline(mu_teorico, color='red', linestyle='--',
            label=f'Media teorica = {mu_teorico}')

plt.xlabel('Media amostral')
plt.ylabel('Densidade')
plt.title('Distribuicao Amostral da Media (n=30)')
plt.legend()
plt.grid(True, alpha=0.3)
plt.show()
```

Exercício 3: Efeito do Tamanho da Amostra

Enunciado: Demonstre como o erro padrão diminui com o aumento do tamanho da amostra. Para uma população com $\mu = 50$ e $\sigma = 10$, simule distribuições amostrais para $n = 5, 10, 25, 50, 100$ e compare os erros padrão.

Dicas:

- Crie uma função para simular cada cenário
- Plote os resultados em subgráficos
- Compare com os valores teóricos

Desafio: Crie um gráfico mostrando a relação entre n e $\sigma_{\bar{X}}$.

Gabarito - Exercício 3 - Parte 1

```
# Parametros
mu = 50
sigma = 10
tamanhos_amostra = [5, 10, 25, 50, 100]
num_simulacoes = 1000

# Funcao para simular
def simular_distribuicao_amostral(mu, sigma, n, num_sim):
    medias = []
    for _ in range(num_sim):
        amostra = np.random.normal(mu, sigma, n)
        medias.append(np.mean(amostra))
    return np.array(medias)

# Simular para diferentes tamanhos
resultados = {}
for n in tamanhos_amostra:
    medias = simular_distribuicao_amostral(mu, sigma, n, num_simulacoes)
    resultados[n] = {
        'medias': medias,
        'erro_empirico': np.std(medias, ddof=0),
        'erro_teorico': sigma / np.sqrt(n)
    }
```

Gabarito - Exercício 3 - Parte 2

```
# Criar visualizacao
fig, axes = plt.subplots(2, 3, figsize=(15, 10))
axes = axes.flatten()

for i, n in enumerate(tamanhos_amostra):
    ax = axes[i]
    ax.hist(resultados[n]['medias'], bins=30, density=True, alpha=0.7)
    ax.set_title(f'n = {n}\nErro padrao = {resultados[n]["erro_empirico"]:.3f}')
    ax.set_xlim(35, 65)
    ax.axvline(mu, color='red', linestyle='--')

# Remover eixo extra
axes[-1].remove()
plt.tight_layout()
plt.show()
```

Exercício 4: Aplicação Prática - Tempo de Atendimento

Enunciado: Um órgão público registrou tempos de atendimento (em minutos) com média histórica $\mu = 25$ e desvio padrão $\sigma = 8$.

1. Simule 100 dias de trabalho, cada um com uma amostra de 20 atendimentos
2. Calcule a média diária e construa um gráfico de controle
3. Identifique dias com médias fora do intervalo $\mu \pm 2\sigma_{\bar{X}}$

Objetivo: Aplicar conceitos de distribuição amostral em controle de qualidade.

Gabarito - Exercício 4 - Parte 1

```
# Parametros do problema
mu_atendimento = 25  # minutos
sigma_atendimento = 8  # minutos
n_atendimentos_dia = 20
num_dias = 100

# Calcular limites de controle
erro_padrao = sigma_atendimento / np.sqrt(n_atendimentos_dia)
limite_superior = mu_atendimento + 2 * erro_padrao
limite_inferior = mu_atendimento - 2 * erro_padrao

# Simular 100 dias
np.random.seed(42)
medias_diarias = []
dias_fora_controle = []

for dia in range(num_dias):
    # Simular atendimentos do dia
    tempos_dia = np.random.normal(mu_atendimento, sigma_atendimento, n_atendimentos_dia)
    media_dia = np.mean(tempos_dia)
    medias_diarias.append(media_dia)

    # Verificar se esta fora de controle
    if media_dia < limite_inferior or media_dia > limite_superior:
        dias_fora_controle.append(dia + 1)

medias_diarias = np.array(medias_diarias)
```

Gabarito - Exercício 4 - Parte 2

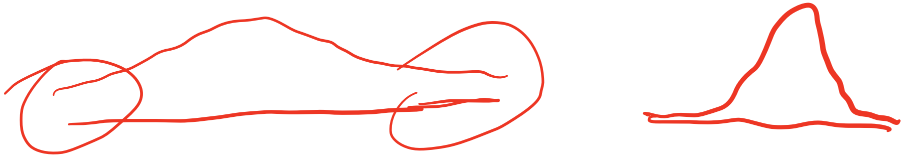
```
# Criar grafico de controle
plt.figure(figsize=(14, 8))

# Plotar medias diarias
plt.plot(range(1, num_dias + 1), medias_diarias, 'b-', marker='o',
         markersize=4, label='Media diaria')

# Linhas de controle
plt.axhline(mu_atendimento, color='green', linestyle='-', linewidth=2,
            label=f'Media populacional = {mu_atendimento}')
plt.axhline(limite_superior, color='red', linestyle='--', linewidth=2,
            label=f'Limite superior (2) = {limite_superior:.2f}')
plt.axhline(limite_inferior, color='red', linestyle='--', linewidth=2,
            label=f'Limite inferior (2) = {limite_inferior:.2f}')

# Destacar pontos fora de controle
for dia in dias_fora_controle:
    plt.scatter(dia, medias_diarias[dia-1], color='red', s=100, zorder=5)

plt.xlabel('Dia')
plt.ylabel('Tempo medio de atendimento (minutos)')
plt.title('Grafico de Controle - Tempo de Atendimento')
plt.legend()
plt.grid(True, alpha=0.3)
plt.show()

print(f"Dias fora de controle: {dias_fora_controle}")
print(f"Percentual de dias fora de controle: {len(dias_fora_controle)}%")
print(f"Esperado teoricamente: ~5% (regra 95% para 2 desvios)")
```

Desafio Extra: Comparando Distribuições

Enunciado: Compare o comportamento da distribuição amostral para diferentes distribuições populacionais:

- Normal
- Uniforme
- Exponencial

Cauchy



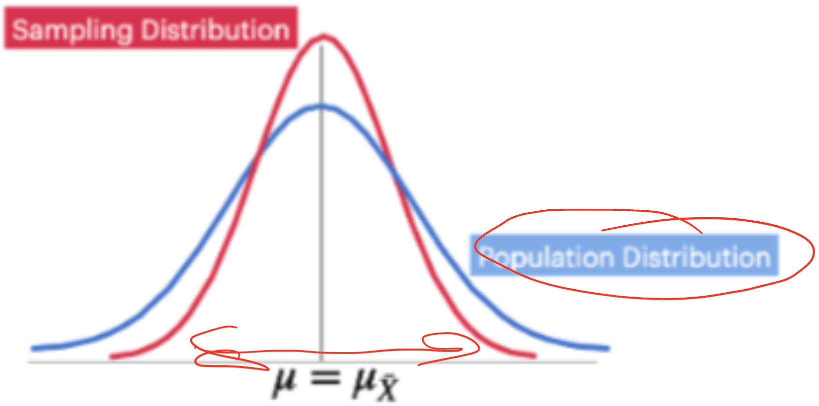
Verifique que, independentemente da distribuição original, a distribuição das médias tende a ser normal (prévia do Teorema Central do Limite).

Sugestão: Use $n = 30$ e visualize as distribuições lado a lado.

Resultado esperado: https://allanvc.shinyapps.io/CLT-app/

Média amostral para população normal

- Suponha que você tem uma população com média μ e desvio padrão σ.
- Suponha também que a população é **normalmente distribuída**.
- Então a distribuição amostral de $\bar{X}$ é **também normal** com: $\mu_{\bar{X}} = \mu$ e $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$



- Podemos tratar a média amostral de uma população normal como qualquer outra variável aleatória normal.

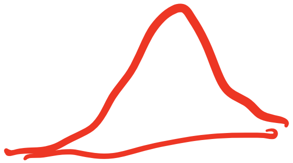
Exemplo 2

Suponha que extraímos uma amostra aleatória simples de tamanho 25 de uma população normal com média 20 e desvio padrão 4.

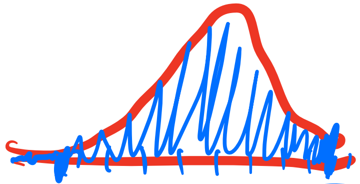
a. Qual é a distribuição de $\bar{X}$?

b. Encontre a probabilidade de observarmos uma média amostral acima de 22.

c. Encontre o percentil 95 de $\bar{X}$.



$N(\mu, \sigma) \Longleftrightarrow$



$N(0,1) \sim Z$

-3.49 3.49

$1 - 0.0002$

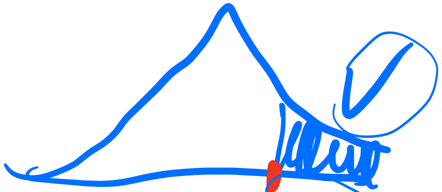
Dados: $n = 25$, $\mu = 20$, $\sigma = 4$, população normal

a. Como a população é normal, $\bar{X}$ também é normal: $\bar{X} \sim N\left(20, \frac{4^2}{25}\right) = N(20, 0.64)$ onde $\sigma_{\bar{X}} = \frac{4}{\sqrt{25}} = 0.8$

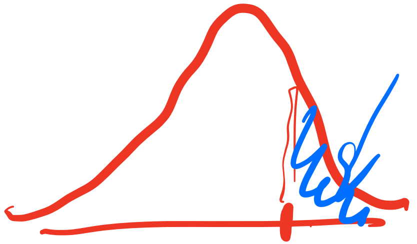
0.9998

b. $P(\bar{X} > 22) = P\left(Z > \frac{22-20}{0.8}\right) = P(Z > 2.5) = 0.0062$

c. Para o percentil 95: $z_{0.95} = 1.645$ $\bar{x}_{0.95} = \mu + z_{0.95} \cdot \sigma_{\bar{X}} = 20 + 1.645(0.8) = 21.316$



$z = ?$



?

z ✓

Média amostral para população não-normal

- E se a população da qual estamos amostrando **NÃO for normal**?
- Seria bom se ainda pudéssemos assumir que $\bar{X}$ é uma variável aleatória normal.
- O próximo teorema nos permite fazer isso, desde que o tamanho da amostra seja grande o suficiente.

Teorema Central do Limite

Teorema Central do Limite (TCL) para médias

Teorema Central do Limite

Seja $\bar{X}$ a média de uma amostra aleatória grande de qualquer população com média μ e desvio padrão σ.
Então a distribuição de $\bar{X}$ é **aproximadamente normal**, com média $\mu_{\bar{X}} = \mu$ e desvio padrão $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$.

P: Por que isso é importante?
R: Se n for grande o suficiente, temos $\bar{X} \sim N(\mu, \sigma^2/n)$ mesmo quando a população original não é normal.

P: Quão grande n precisa ser?
R: Depende da forma da população original – quanto mais assimétrica, maior n precisa ser.
Para a maioria das distribuições, $n > 30$ é suficiente.

Podemos assumir que $\bar{X} \sim N(\mu, \sigma^2/n)$ se:

(i) A população original é normal (devido às propriedades da distribuição normal)

OU

(ii) O tamanho da amostra é grande ($n > 30$) (devido ao TCL)

Se pelo menos uma dessas condições for verdadeira, podemos tratar $\bar{X}$ como qualquer outra variável aleatória normal.

Exemplo 3

Dados recentes do Censo indicam que a idade média de estudantes universitários no Brasil é $\mu = 25$ anos, com desvio padrão $\sigma = 9.5$ anos. Uma amostra aleatória simples de 125 estudantes é extraída.

Se $\bar{X}$ = idade média amostral dos estudantes, qual é a distribuição de $\bar{X}$? (Justifique sua resposta.)

Solução - Exemplo 3

Dados: $\mu = 25$, $\sigma = 9.5$, $n = 125$

- Não sabemos se a população é normal
- Mas $n = 125 > 30$ (amostra grande)
- Pelo TCL, $\bar{X}$ é aproximadamente normal

Calculando os parâmetros:

$$\mu_{\bar{X}} = \mu = 25$$
$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{9.5}{\sqrt{125}} = \frac{9.5}{11.18} = 0.85$$

Portanto: $\bar{X} \sim N(25, 0.85^2)$ aproximadamente nº sucessos ≥ 10   nº fracassos ≥ 10

TCL ✓

$\boxed{n > 30}$

Distribuição Amostral de Proporções

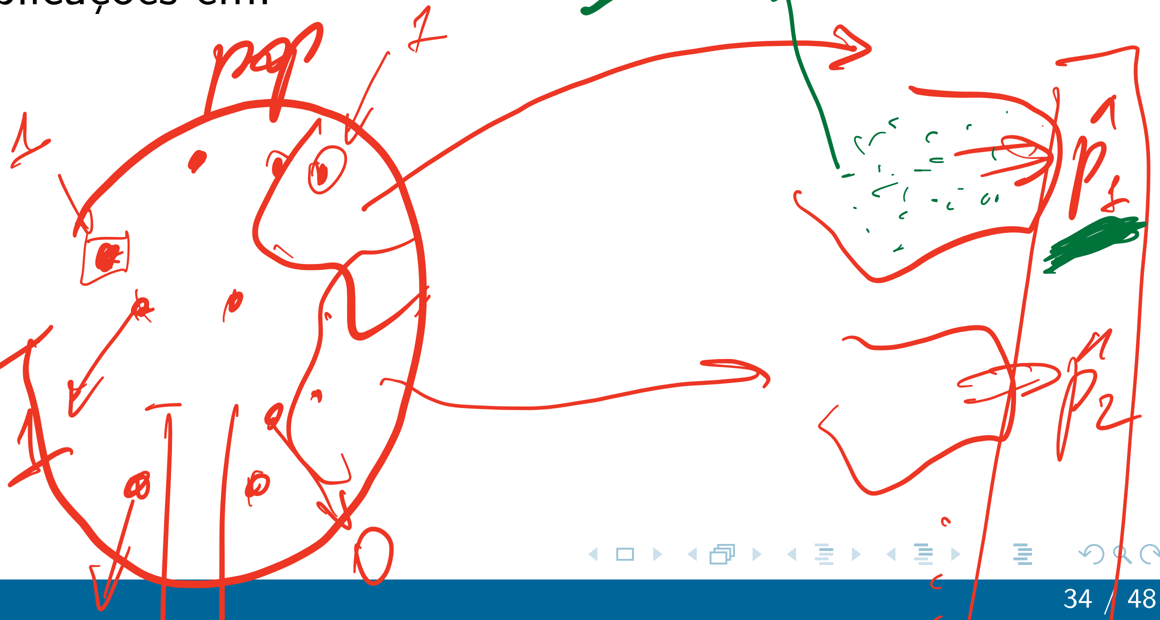
$n\hat{p} \geq 10$ $n(1-\hat{p}) \geq 10$

39 fracasso   1 sucesso

Aplicações em:

- Pesquisas de opinião
- Controle de qualidade
- Estudos epidemiológicos
- Avaliação de políticas públicas

$Y \sim Bin(N, p)$



$p = ?$ $\hat{p}_k$

- Frequentemente estamos interessados em estimar a proporção de indivíduos em uma população com certa propriedade.
- Podemos pensar em um indivíduo como sendo um "sucesso" ou "fracasso" dependendo se possui a propriedade.

v.a. $\hat{p} \sim ?$

Exemplos:

- sucesso = "aprova a gestão" vs. fracasso = "não aprova a gestão"
- sucesso = "utilizou o serviço digital" vs. fracasso = "não utilizou"
- sucesso = "possui ensino superior" vs. fracasso = "não possui"

$\hat{p} \sim N(\mu_{\hat{p}}, \sigma_{\hat{p}})$

$\mu_{\hat{p}} = \mu_Y$, $\sigma_{\hat{p}} = \frac{\sigma_Y}{\sqrt{n}}$

Proporção amostral

- Seja p a proporção de sucessos na população.
- Seja $X_1, \ldots, X_n$ uma amostra aleatória de tamanho n tal que:

$$X_i = \begin{cases} 1 & \text{se a i-ésima observação é sucesso} \\ 0 & \text{se a i-ésima observação é fracasso} \end{cases}$$

- A proporção populacional pode ser estimada pela proporção amostral: $\boxed{\hat{p} = \frac{X}{n}} = \frac{\sum_{i=1}^{n} X_i}{n}$ onde $X = \sum_{i=1}^{n} X_i$ representa o número de sucessos na amostra.

$$\left.\begin{aligned} E(Y) &= np \\ VAR(Y) &= np(1-p) \end{aligned}\right\}$$

Y é v.a. Binomial

L é nº menos na pop

Média e desvio padrão da proporção amostral

$$\hat{p} = \frac{\hat{x}}{n} \qquad E(\hat{p}) = \mu_{\hat{p}}$$

Propriedades de $\hat{p}$

- A **média** da proporção amostral $\hat{p}$ é: $\boxed{\mu_{\hat{p}} = p}$
- O **desvio padrão** da proporção amostral $\hat{p}$ é: $\boxed{\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}}$

- Novamente, o teorema central do limite nos dirá a "forma" da distribuição de $\hat{p}$.

$$\hat{p} \sim N\left(\mu_{\hat{p}} = p,\ \sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}\right)$$

$$\boxed{\hat{p}(1-\hat{p})} \qquad n$$

Teorema Central do Limite para proporções

$H_0: p = \boxed{0.27}$

$H_1: p > \boxed{0.27}$

$p > 0.3$

est. teste $= \dfrac{\text{algo}}{\text{E.P.}}$

Teorema Central do Limite (TCL)

Se $np \geq 10$ e $n(1-p) \geq 10$, então pelo TCL a distribuição de $\hat{p}$ é aproximadamente normal com média $\mu_{\hat{p}} = p$ e desvio padrão $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$.

Ou seja: $\hat{p} \sim N\left(p, \frac{p(1-p)}{n}\right)$ quando $np \geq 10$ e $n(1-p) \geq 10$.

Observação: As condições $np \geq 10$ e $n(1-p) \geq 10$ garantem que temos sucessos e fracassos suficientes na amostra.

Exemplo 4

De acordo com uma pesquisa do Datafolha 27% dos brasileiros preferem chocolate como sabor de sorvete favorito. Se uma amostra de 100 brasileiros é selecionada, qual é a probabilidade de que a proporção amostral dos que preferem chocolate seja maior que 0.30?

v.a.

$$P(\hat{p} > 0.30) = ?$$

$\hat{p} \sim N$

$$P(Z > z) = ?$$

Solução - Exemplo 4

cdf → probs acumuladas

$\boxed{\text{PPF}}$ → valores no eixo x

Dados: $p = 0.27$, $n = 100$

Verificando as condições do TCL:

- $np = 100 \times 0.27 = 27 \geq 10$
- $n(1-p) = 100 \times 0.73 = 73 \geq 10$

Parâmetros da distribuição:

$$\mu_{\hat{p}} = p = 0.27$$

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.27 \times 0.73}{100}} = 0.0444$$

Calculando a probabilidade: $P(\hat{p} > 0.30) = P\left(Z > \frac{0.30-0.27}{0.0444}\right) = P(Z > 0.676) = 0.2495$

Aproximadamente 24.95% de chance.

Aplicações do Teorema Central do Limite

Google Colab

Exercícios complementares

- Verificação do TCL para diferentes distribuições
- TCL para proporções
- Aplicações em pesquisas de opinião
- Cálculo de probabilidades

Exercício 5: Verificando o TCL

Enunciado: Demonstre o TCL simulando médias amostrais de três distribuições diferentes:

- Uniforme(0, 1)
- Exponencial($\lambda = 2$)
- Binomial(n=10, p=0.3)

Use amostras de tamanho n=5, 30 e 100. Visualize como a distribuição das médias se aproxima da normal conforme n aumenta.

Objetivo: Verificar empiricamente o Teorema Central do Limite.

Gabarito - Exercício 5 - Parte 1

```
import numpy as np
import matplotlib.pyplot as plt
from scipy import stats
# Configuracao
tamanhos = [5, 30, 100]
num_simulacoes = 10000
# Funcao para plotar
def plotar_tcl(dist_func, dist_params, dist_nome):
    fig, axes = plt.subplots(1, 3, figsize=(15, 5))

    for i, n in enumerate(tamanhos):
        medias = []
        for _ in range(num_simulacoes):
            if dist_nome == 'Uniforme':
                amostra = np.random.uniform(*dist_params, n)
            elif dist_nome == 'Exponencial':
                amostra = np.random.exponential(1/dist_params[0], n)
            else:  # Binomial
                amostra = np.random.binomial(*dist_params, n)
            medias.append(np.mean(amostra))

        axes[i].hist(medias, bins=50, density=True, alpha=0.7, edgecolor='black')
        axes[i].set_title(f'{dist_nome} - n={n}')

        # Adicionar curva normal teorica
        media_teorica = np.mean(medias)
        std_teorico = np.std(medias)
        x = np.linspace(media_teorica - 4*std_teorico, media_teorica + 4*std_teorico, 100)
        axes[i].plot(x, stats.norm.pdf(x, media_teorica, std_teorico), 'r-', lw=2)
```

Gabarito - Exercício 5 - Parte 2

```
def plotar_tcl(dist_func, dist_params, dist_nome):
    fig, axes = plt.subplots(1, 3, figsize=(15, 5))

    for i, n in enumerate(tamanhos):
        medias = []
        for _ in range(num_simulacoes):

            [...]

    plt.tight_layout()
    plt.show()

# Testar distribuicoes
plotar_tcl(None, (0, 1), 'Uniforme')
plotar_tcl(None, (2,), 'Exponencial')
plotar_tcl(None, (10, 0.3), 'Binomial')
```

Exercício 6: TCL para Proporções

Enunciado: Uma pesquisa eleitoral indica que 45% dos eleitores apoiam o candidato A.

1. Simule 1000 pesquisas com n=200 eleitores cada
2. Calcule a proporção de apoio em cada pesquisa simulada
3. Compare a distribuição empírica com a distribuição teórica
4. Calcule a probabilidade de uma pesquisa mostrar mais de 50% de apoio

Aplicação: Entender margens de erro em pesquisas eleitorais.

Gabarito - Exercício 6 - Parte 1

```
# Parametros
p_real = 0.45  # Proporcao real de apoio
n_eleitores = 200
num_pesquisas = 1000

# Verificar condicoes do TCL
print(f"np = {n_eleitores * p_real} >= 10? {n_eleitores * p_real >= 10}")
print(f"n(1-p) = {n_eleitores * (1-p_real)} >= 10? {n_eleitores * (1-p_real) >= 10}")

# Simular pesquisas
np.random.seed(42)
proporcoes_simuladas = []

for _ in range(num_pesquisas):
    # Simular votos (1 = apoia, 0 = nao apoia)
    votos = np.random.binomial(1, p_real, n_eleitores)
    prop_amostra = np.mean(votos)
    proporcoes_simuladas.append(prop_amostra)

proporcoes_simuladas = np.array(proporcoes_simuladas)

# Parametros teoricos
mu_p_hat = p_real
sigma_p_hat = np.sqrt(p_real * (1 - p_real) / n_eleitores)
```

p

$n \cdot p \geq 10$

$n(1-p) \geq 10$

1 ou 0 Bernoulli

0 : 999

Gabarito - Exercício 6

```
# Visualizar
plt.figure(figsize=(10, 6))
plt.hist(proporcoes_simuladas, bins=30, density=True, alpha=0.7,
         edgecolor='black', label='Distribuicao empirica')

# Curva normal teorica
x = np.linspace(0.35, 0.55, 100)
y = stats.norm.pdf(x, mu_p_hat, sigma_p_hat)
plt.plot(x, y, 'r-', lw=2, label=f'Normal({mu_p_hat}, {sigma_p_hat:.4f})')

plt.axvline(0.5, color='green', linestyle='--', label='50% de apoio')
plt.xlabel('Proporcao de apoio ao candidato A')
plt.ylabel('Densidade')
plt.title('Distribuicao amostral da proporcao (n=200)')
plt.legend()
plt.grid(True, alpha=0.3)
plt.show()

# Calcular probabilidade
prob_maior_50 = 1 - stats.norm.cdf(0.5, mu_p_hat, sigma_p_hat)
print(f"\nProbabilidade de uma pesquisa mostrar > 50%: {prob_maior_50:.4f}")
print(f"Probabilidade empirica: {np.mean(proporcoes_simuladas > 0.5):.4f}")
```

Conclusões e Próximos Passos

O que aprendemos hoje:

- Distribuição amostral da média para populações normais
- Teorema Central do Limite para médias
- Teorema Central do Limite para proporções
- Aplicações práticas em pesquisas e controle de qualidade
- Quando podemos usar a aproximação normal

Próxima aula:

- Intervalos de Confiança
- Estimação por intervalo
- Margem de erro

Dúvidas?