

**Tera**

# AULA 13

Design de Experimentos  
& Teste A/B

Instrutor: [Raphael Ballet](#)

---

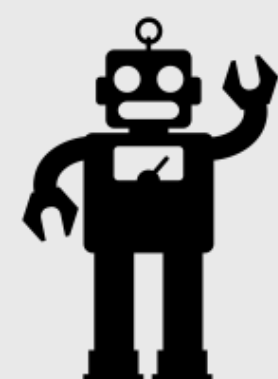
## Background:

- Engenheiro de Controle e Automação (IMT)
  - Mestre em Sistemas Aeroespaciais e Mecatrônica (ITA)
  - Data Scientist – Elo7
- 

## Interesses:



[Drones](#)



[Robótica](#)



[Aprendizado de Máquina](#)



[Visão Computacional](#)



[Processamento de Linguagem Natural](#)



[Sistemas de recomendação](#)

# Planejamento:

1. Apresentação Case
2. Introdução à estatística
3. Teste A/B
4. Testes de hipóteses
5. Case final e discussões

T

# CASE:

Empresa: TeraBuy

Objetivo: Venda online de equipamentos educacionais



# Objetivo:

➤ Vender mais!

Sim, claro, mas como?



- Quem são nossos clientes?
- Como utilizam nosso site?
- Quantas pessoas acessam meu site?
- E quantas clicam no botão “compre já”?
- O layout do meu site é bom?

# Objetivo:

➤ Vender mais!

Sim, claro, mas como?



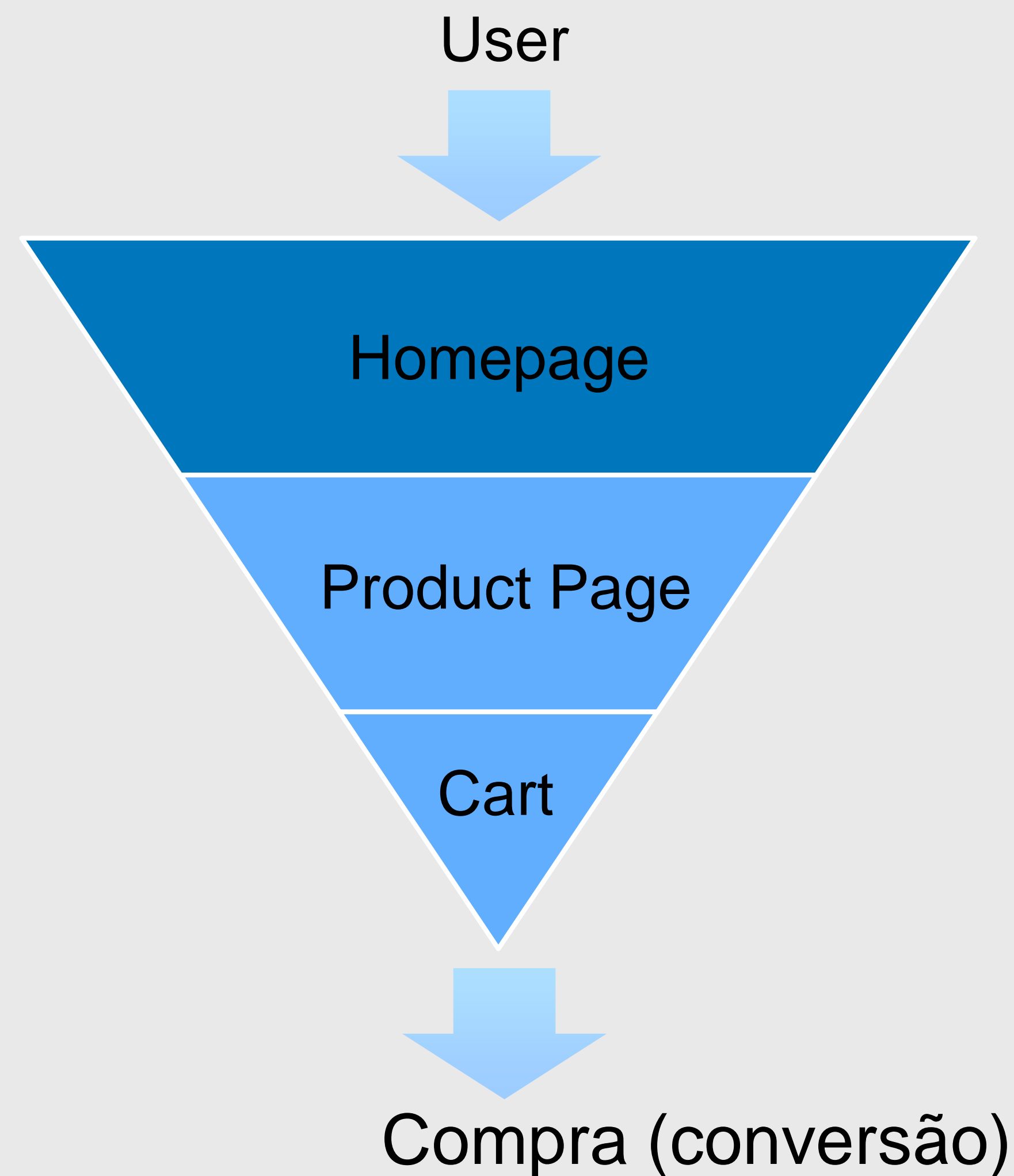
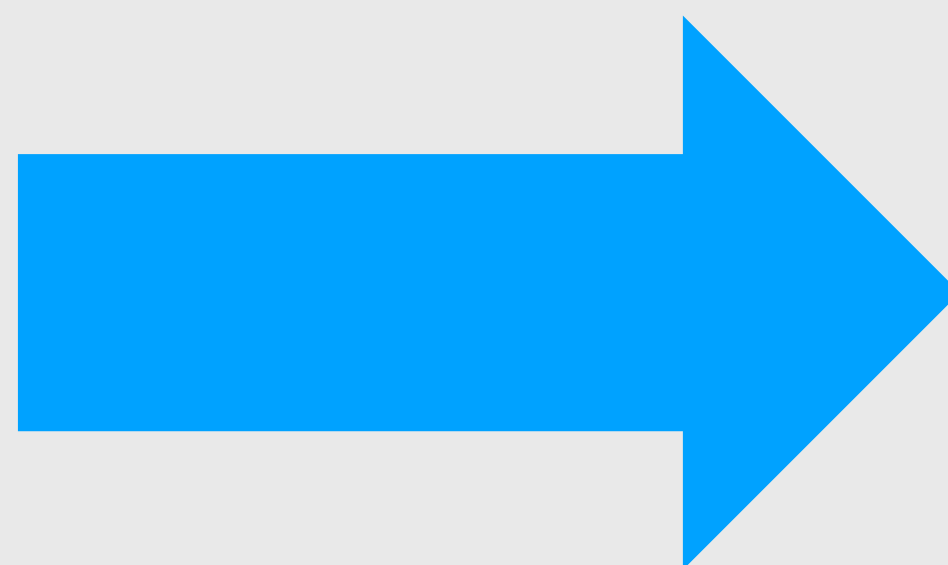
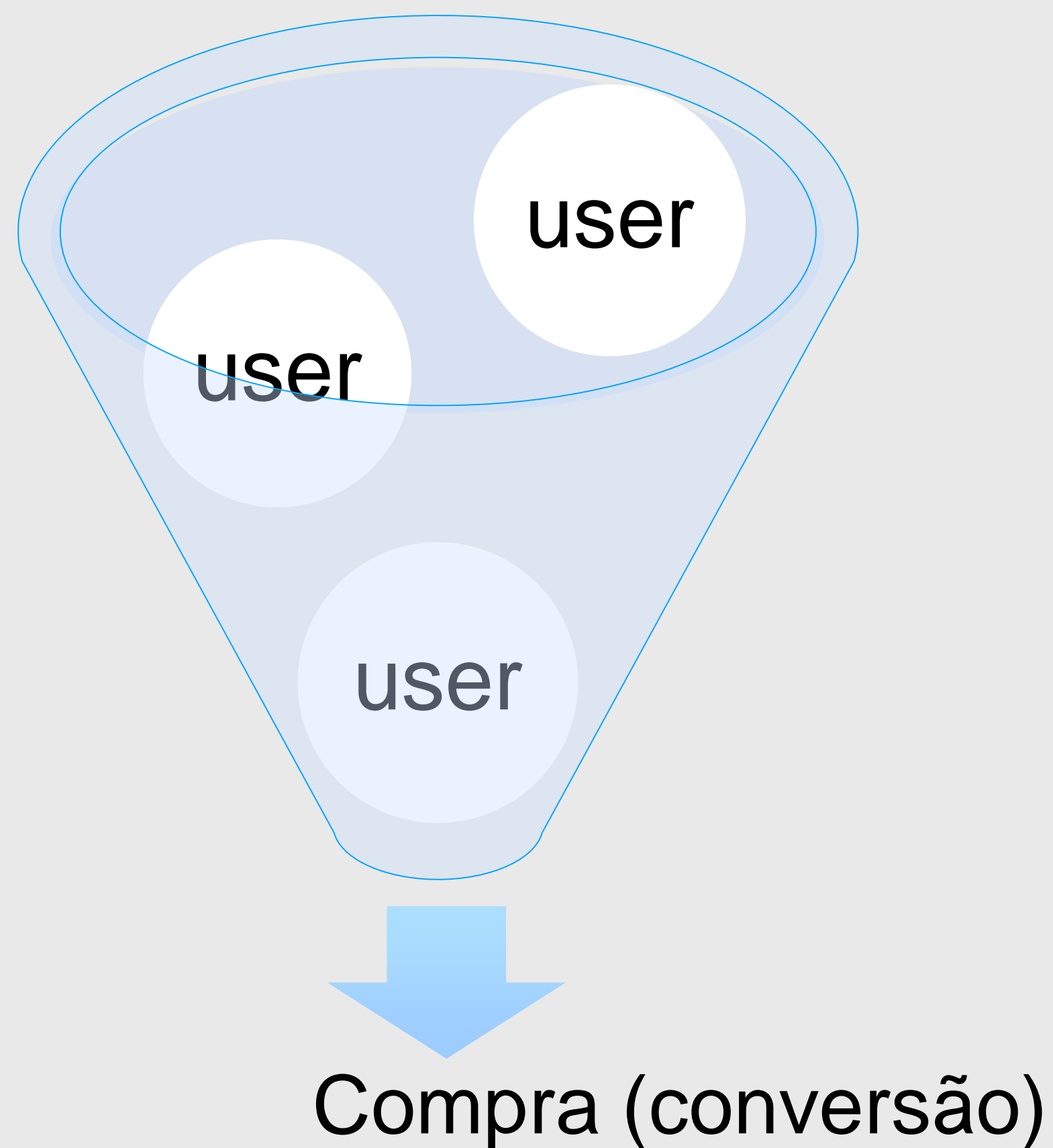
- Quem são nossos clientes?
- Como utilizam nosso site?
- Quantas pessoas acessam meu site?
- E quantas clicam no botão “compre já”?
- O layout do meu site é bom?





T

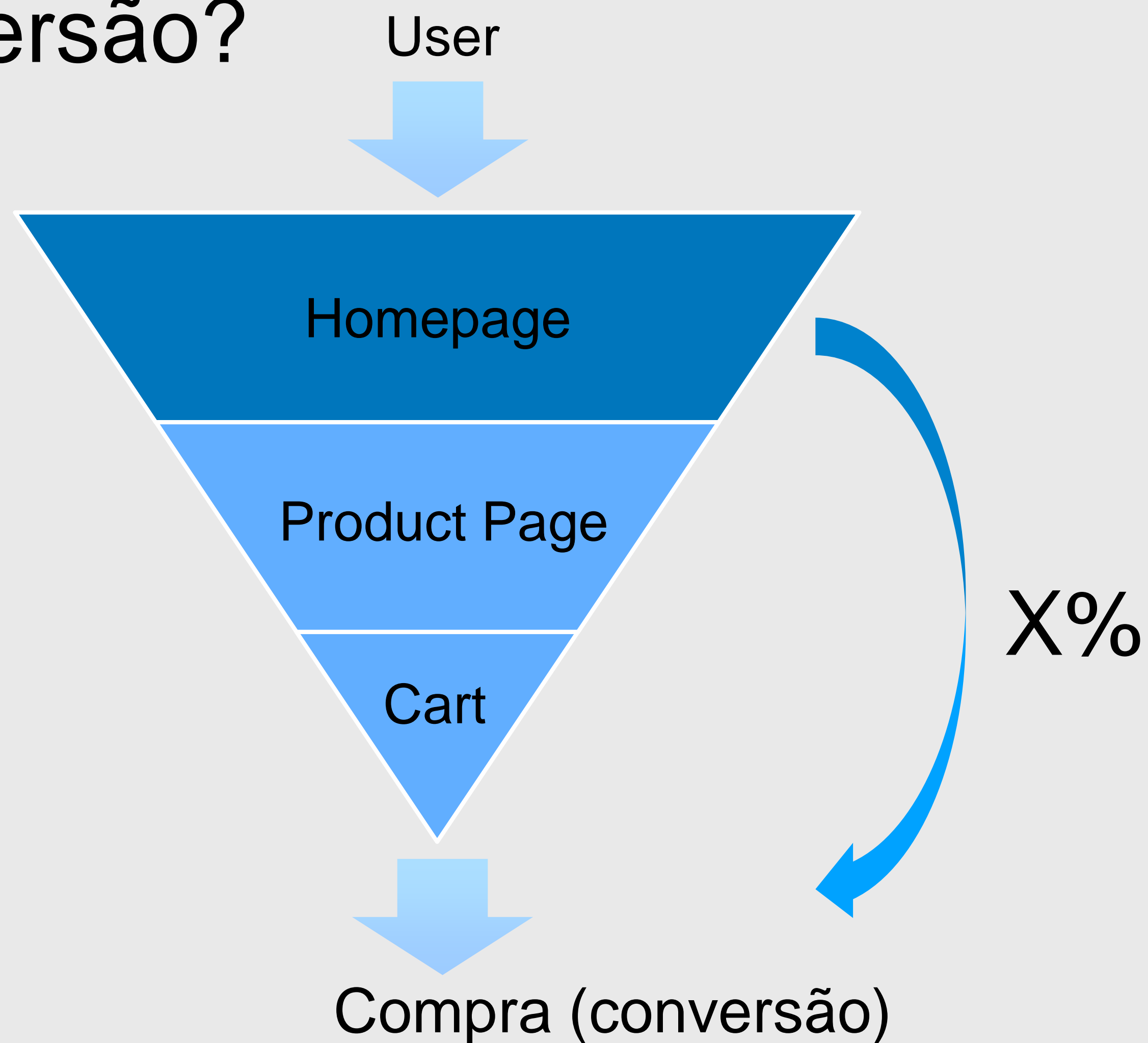
## Modelo funil:





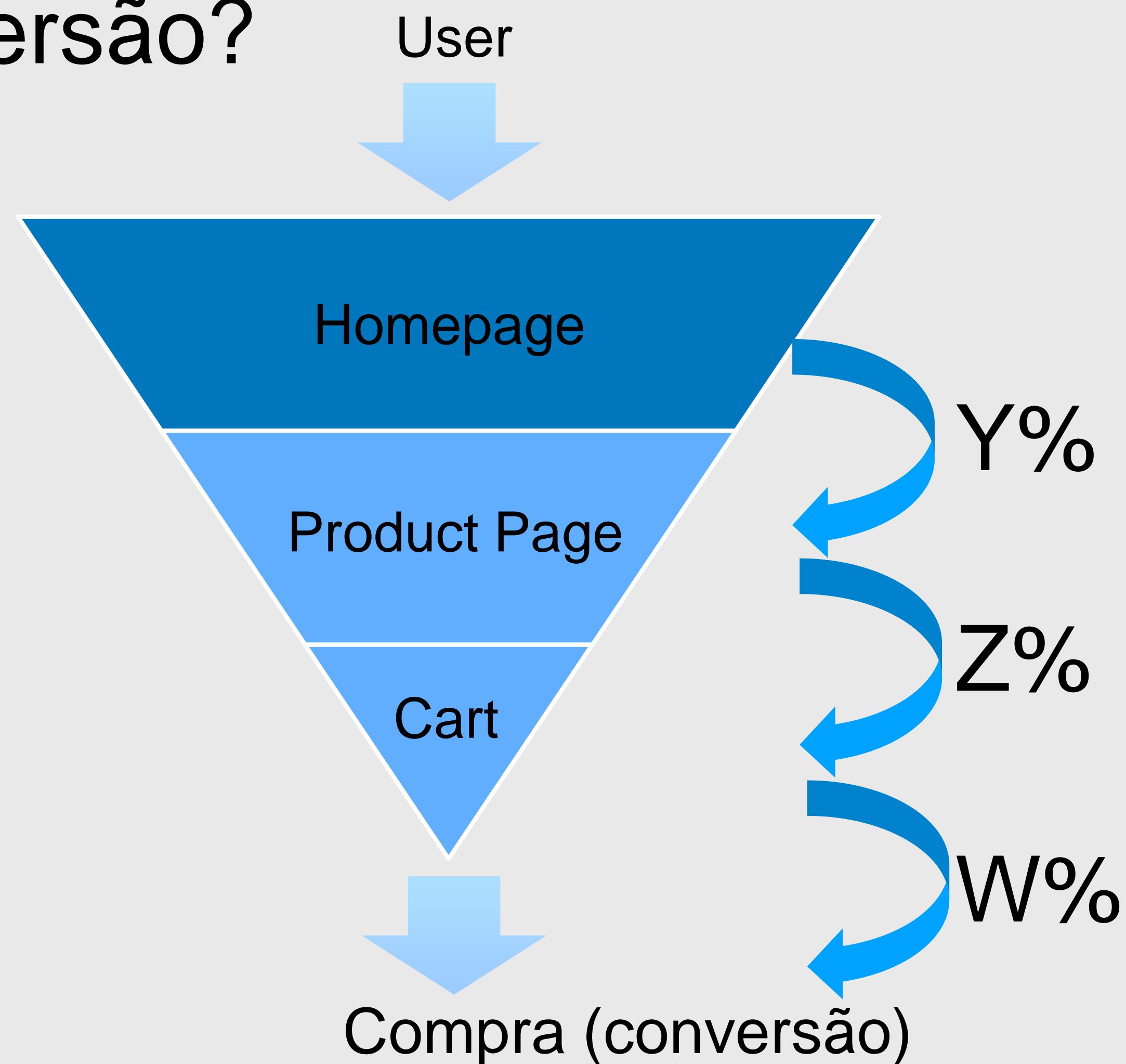
## Modelo funil:

Taxa de conversão?



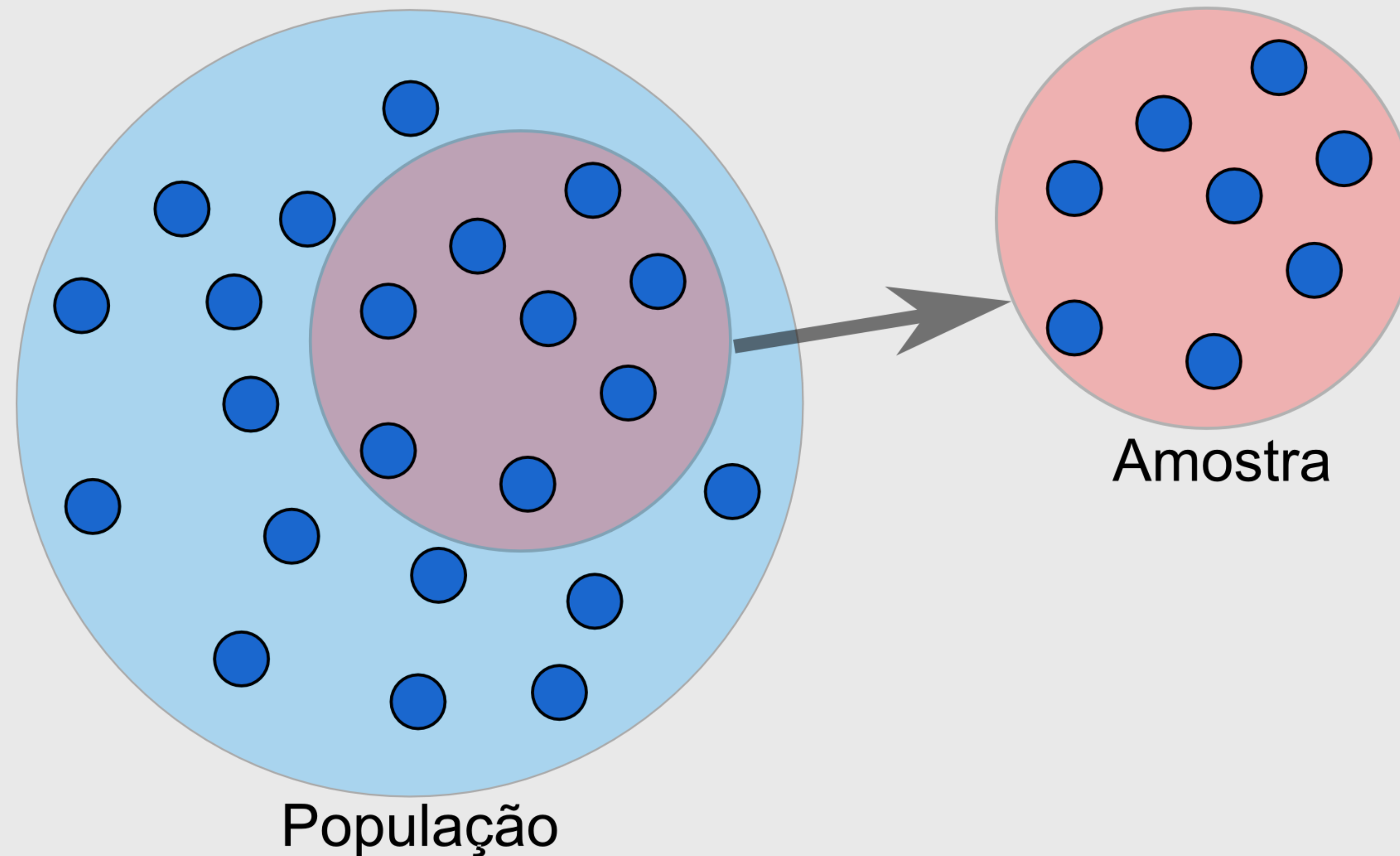
## Modelo funil:

Taxa de conversão?



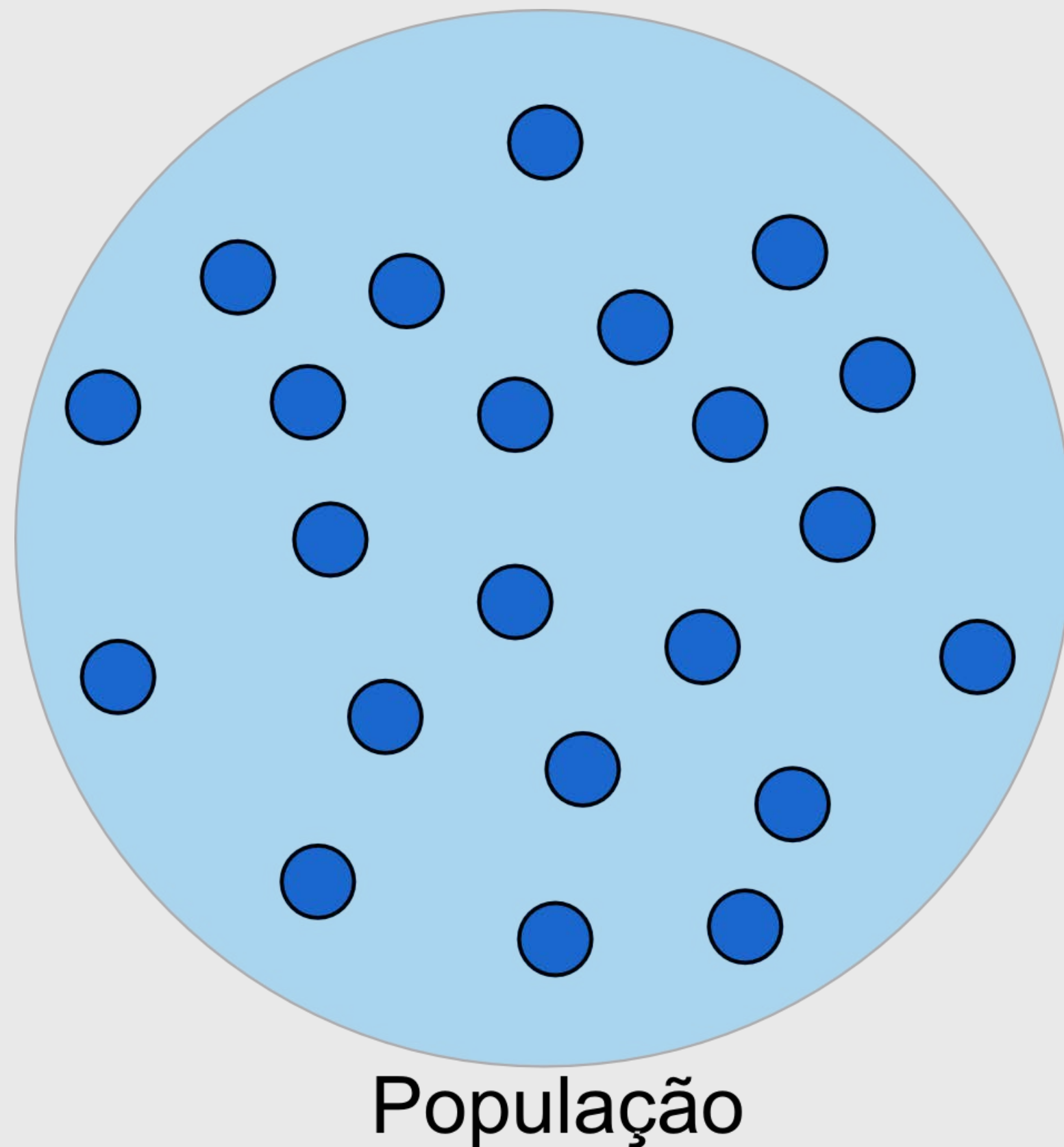
## 2. INTRODUÇÃO À ESTATÍSTICA

- População x Amostra



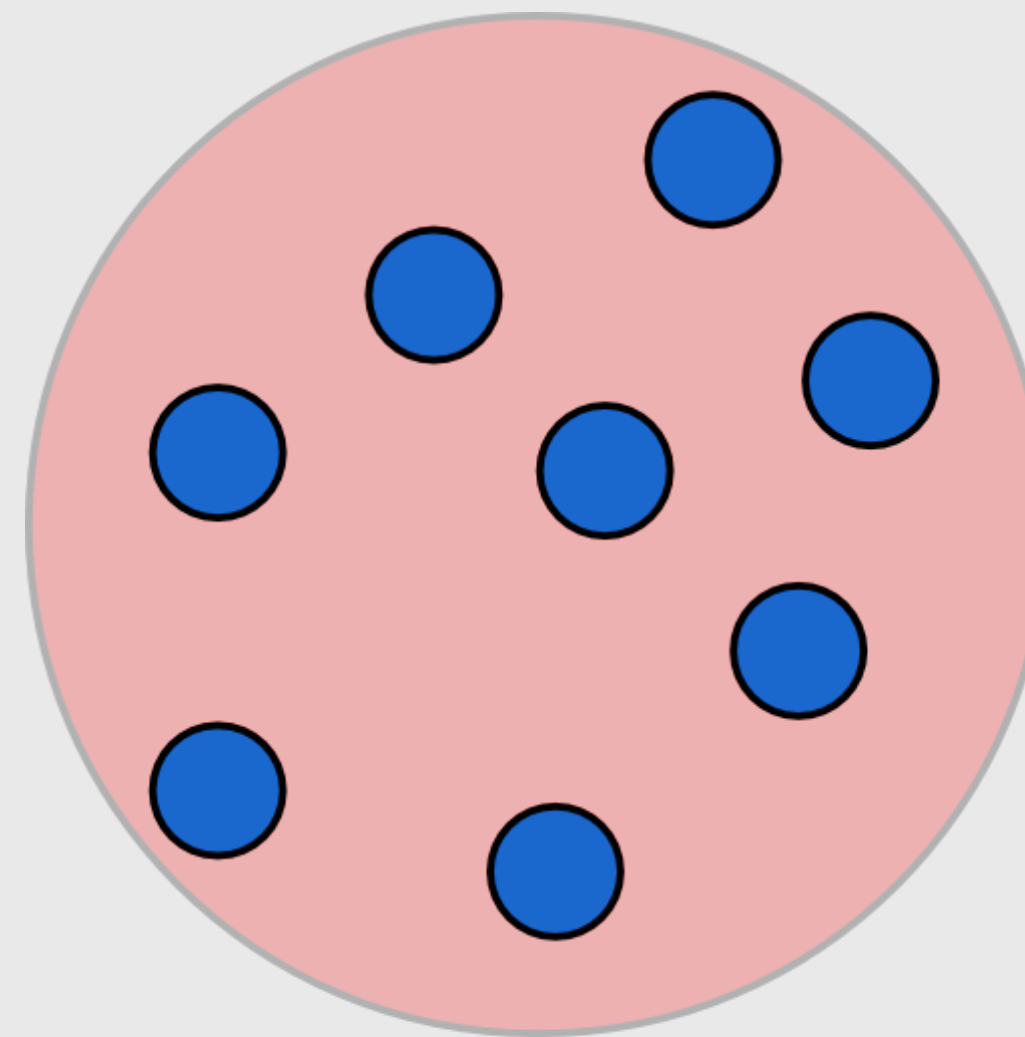
## 2. INTRODUÇÃO À ESTATÍSTICA

- Descrição da população:  
parâmetros: média, desvio padrão etc.



## 2. INTRODUÇÃO À ESTATÍSTICA

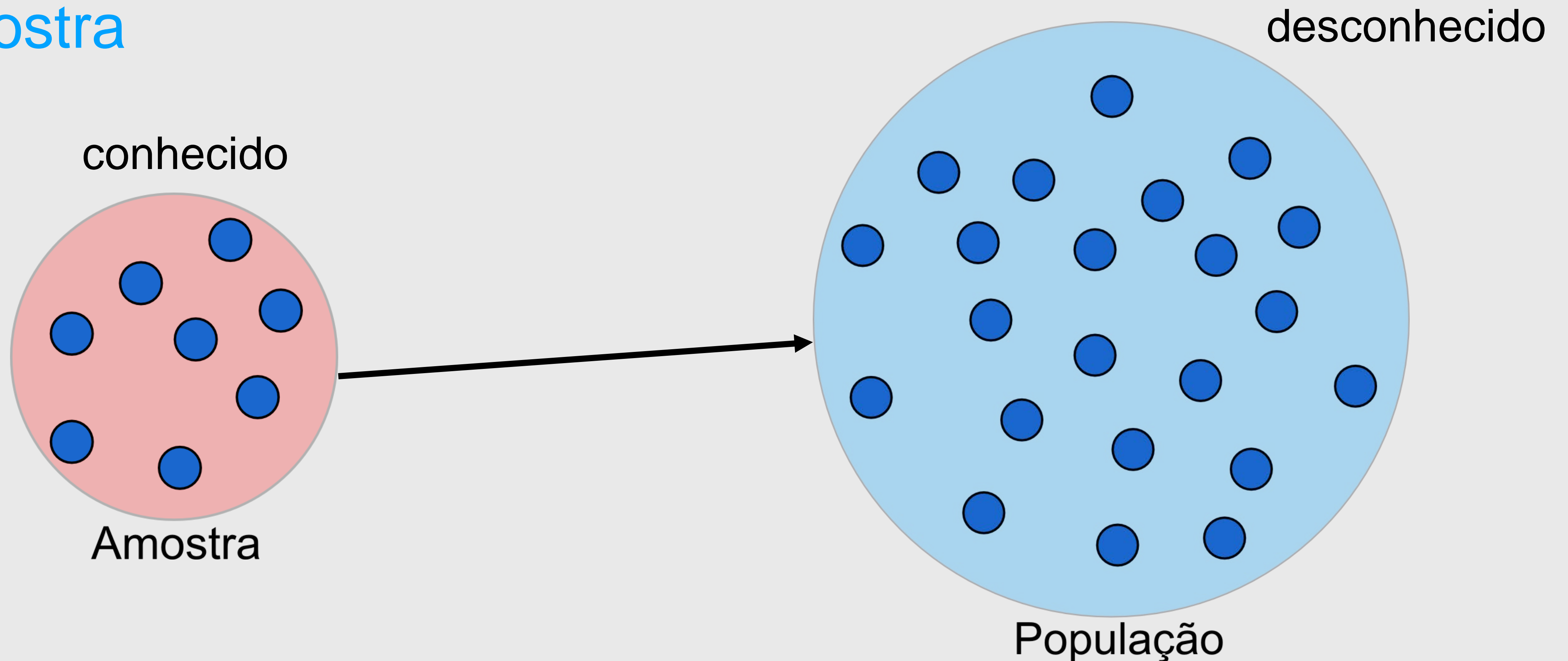
- Amostra: estatísticas de sumário  
estatísticas: média, desvio padrão (amostral) etc.



Amostra

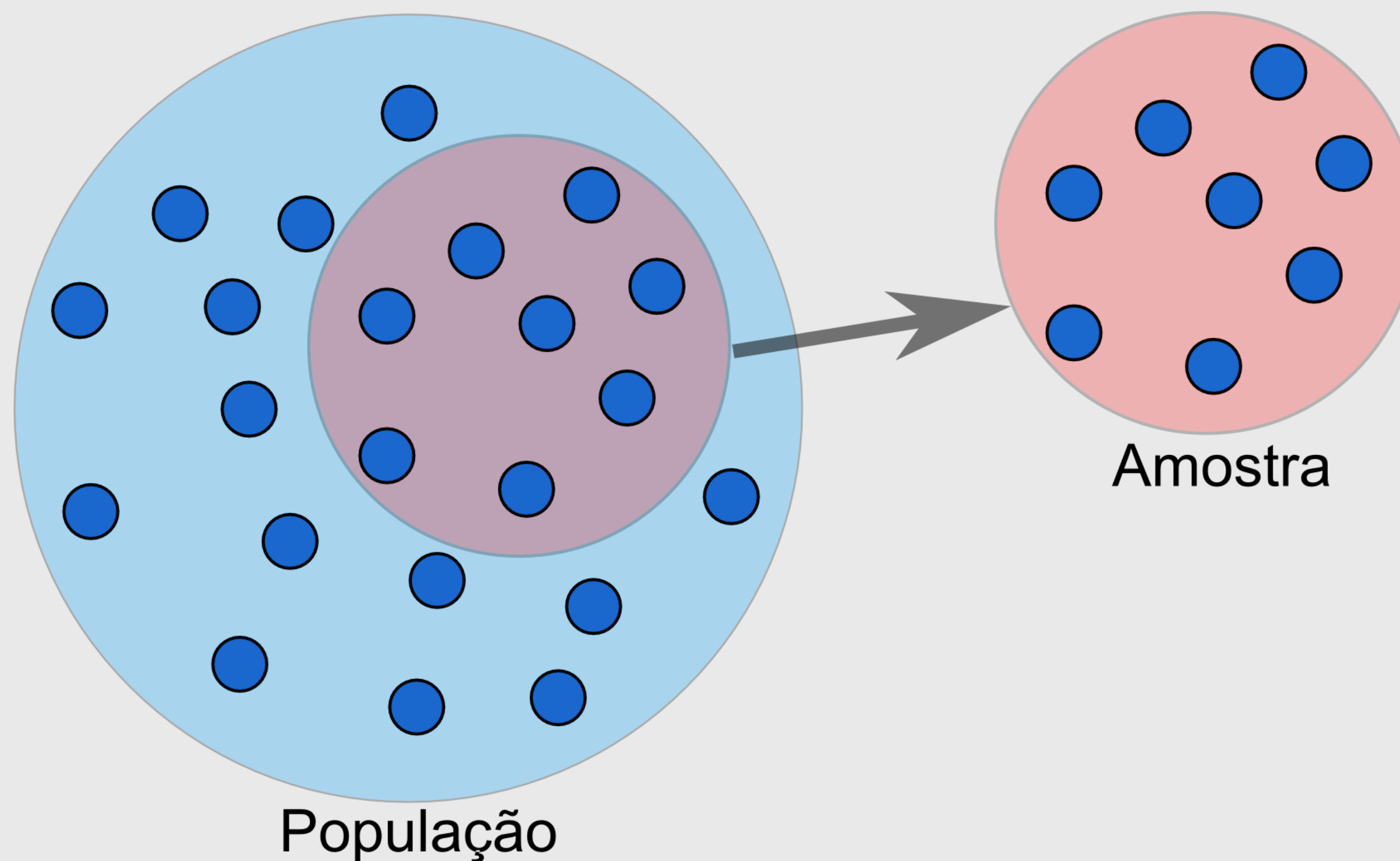
## 2. INTRODUÇÃO À ESTATÍSTICA

- Estatística inferencial:  
prever comportamento da população a partir da amostra



## 2. INTRODUÇÃO À ESTATÍSTICA

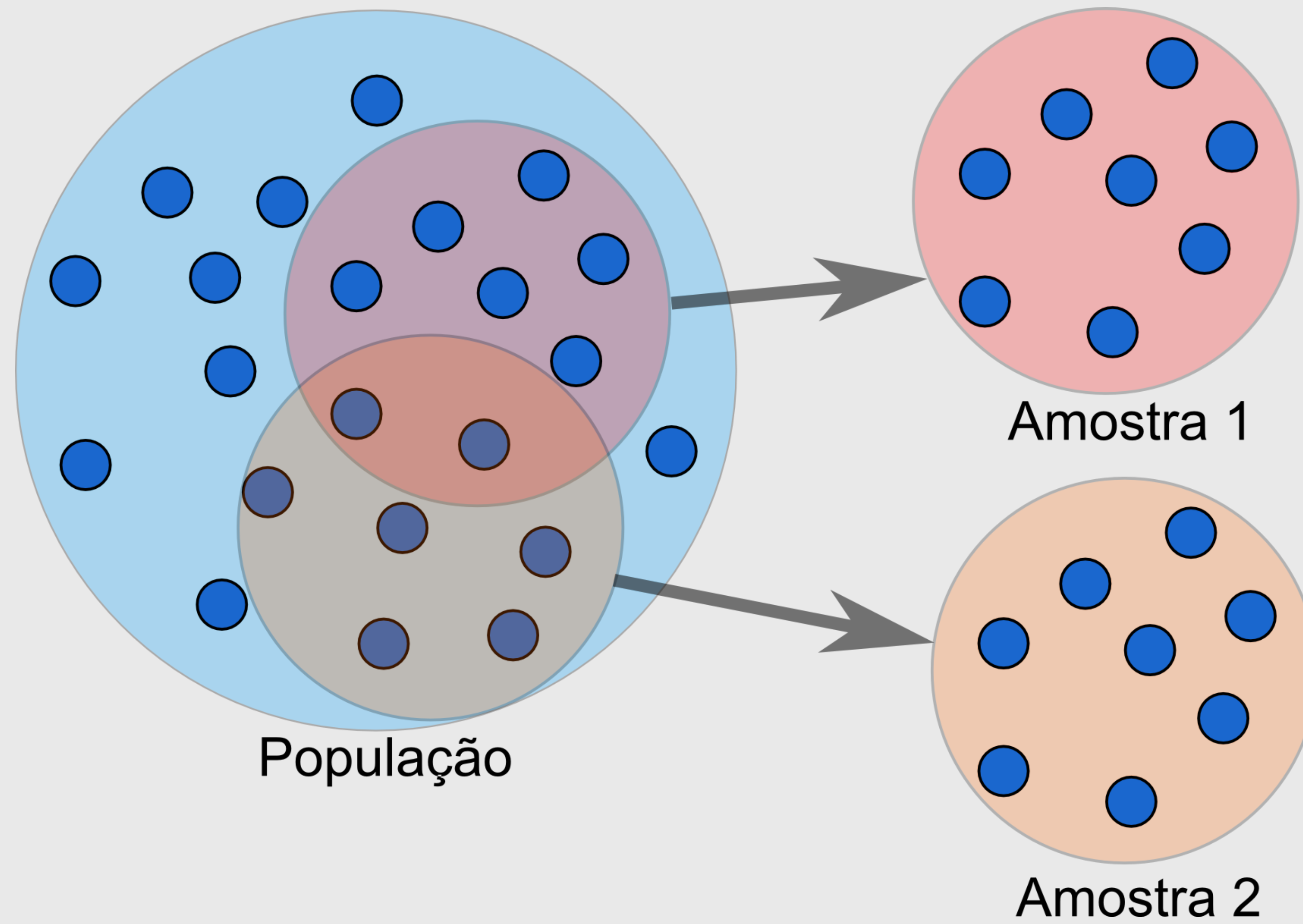
- Amostra representa bem a população?  
Teoria da amostragem: amostras aleatórias





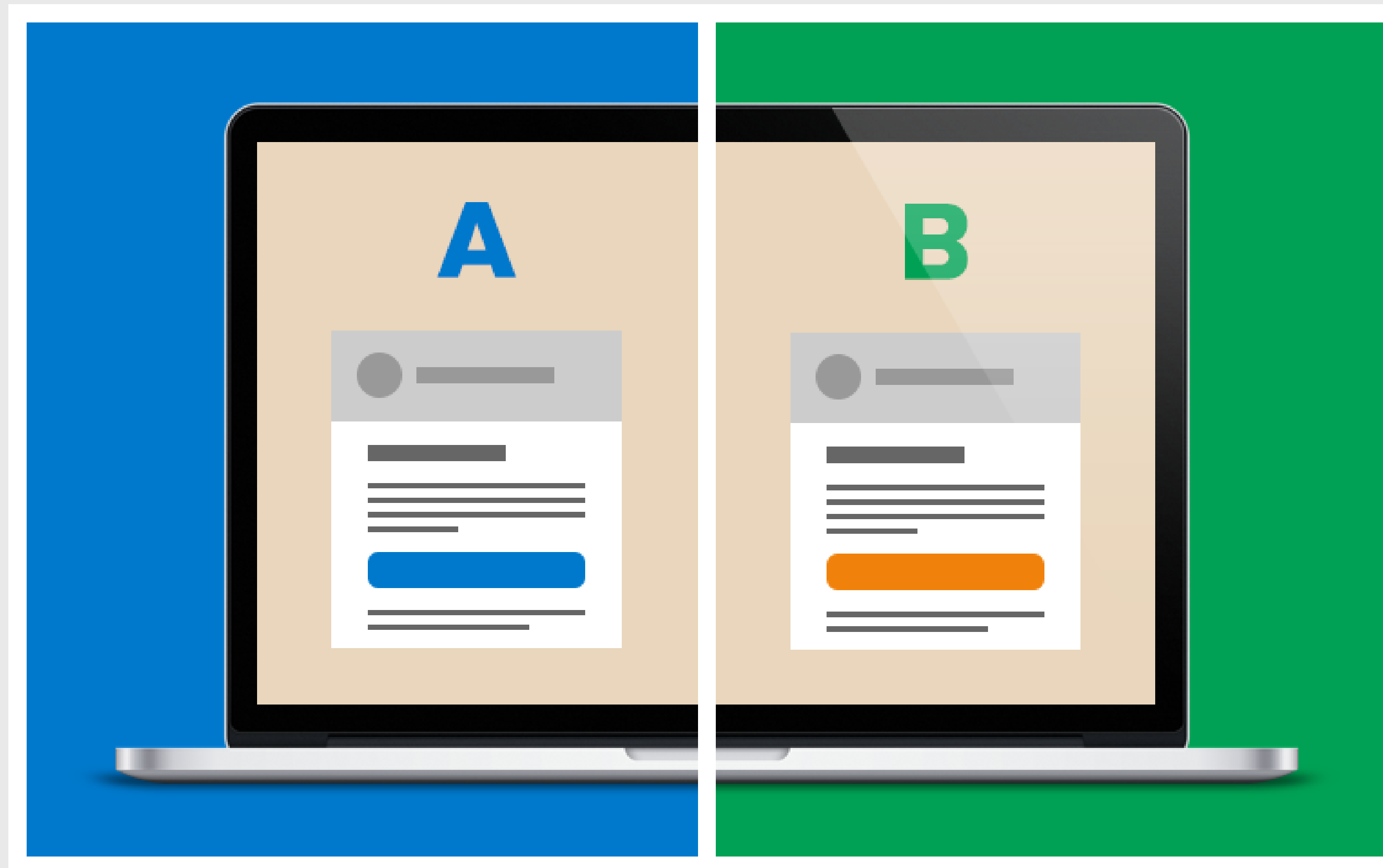
## 2. INTRODUÇÃO À ESTATÍSTICA

- Múltiplas amostragens: Lei dos Grandes Números



### 3. TESTE A/B

- Teste entre versões diferentes de uma mesma funcionalidade



### 3. TESTE A/B

- Teste de hipóteses: diferença entre populações



30%

vs

35%



## 3. TESTE A/B

- Quando usar?
  - Análises relativas
  - Testes de funcionalidade

### 3. TESTE A/B

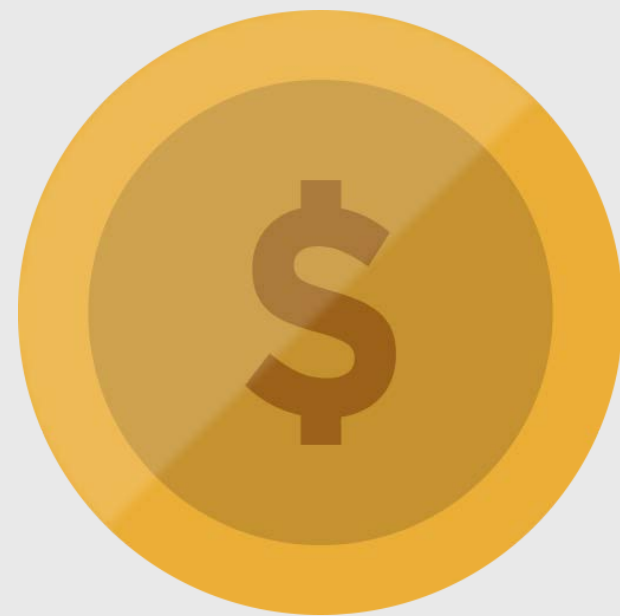
- Quando não usar?
  - Análise exploratória
  - Encontrar melhor funcionalidade
  - Testes de grande impacto:  
efeito de aprendizado (learning effect) e preconceito

## 4. TESTES DE HIPÓTESES

- Diferença entre populações/amostras

**Exemplo:** moedas honestas/viciadas

$P1(\text{cara}) = 50\%$



Conhecido

$P2(\text{cara}) = P1(\text{cara}) ??$



Desconhecido

## 4. TESTES DE HIPÓTESES

- Hipótese: As moedas são idênticas (honestas)

Hipótese nula ( $H_0$ )

- Hipótese 2: A moeda “2” é viciada

Hipótese alternativa ( $H_1$ )

- Teste de hipóteses: verificar se há evidências para rejeitar  $H_0$ .



- **Experimento:** Jogar a moeda “2” 10 vezes e contar quantas vezes saiu “cara”.

Caras	Coroas	A moeda é honesta ?	Probabilidade
0	10		
1	9		
2	8		
3	7		
4	6		
5	5		
6	4		
7	3		
8	2		
9	1		
10	0		



- **Experimento:** Jogar a moeda “2” 10 vezes e contar quantas vezes saiu “cara”.

Caras	Coroas	A moeda é honesta ?	Probabilidade
0	10	Não	
1	9	?	
2	8	?	
3	7	?	
4	6	?	
5	5	Sim	
6	4	?	
7	3	?	
8	2	?	
9	1	?	
10	0	Não	



- Probabilidade de tirar  $k$  caras: Distribuição binomial



Caras	Coroas	A moeda é honesta ?	Probabilidade
0	10	Não	
1	9	?	
2	8	?	
3	7	?	
4	6	?	
5	5	Sim	
6	4	?	
7	3	?	
8	2	?	
9	1	?	
10	0	Não	

- Probabilidade de tirar k caras: Distribuição binomial



Caras	Coroas	A moeda é honesta ?	Probabilidade
0	10	Não	0.0976%
1	9	Não	0.977%
2	8	?	4.4 %
3	7	Sim	11.7 %
4	6	Sim	20.5 %
5	5	Sim	24.6%
6	4	Sim	20.5 %
7	3	Sim	11.7 %
8	2	?	4.4 %
9	1	Não	0.977%
10	0	Não	0.0976%

T

- Intervalo de confiança:

Intervalo  
de  
confiança

Caras	Coroas	A moeda é honesta ?	Probabilidade
0	10	Não	0.0976%
1	9	Não	0.977%
2	8	?	4.4 %
3	7	Sim	11.7 %
4	6	Sim	20.5 %
5	5	Sim	24.6%
6	4	Sim	20.5 %
7	3	Sim	11.7 %
8	2	?	4.4 %
9	1	Não	0.977%
10	0	Não	0.0976%



- Realizamos o experimento e tiramos 10 caras. Podemos dizer que é apenas acaso?

Em 10.000 repetições do experimento, podemos obter esse resultado entre 9 e 10 vezes!

Caras	Coroas	A moeda é honesta ?	Probabilidade
0	10	Não	0.0976%
1	9	Não	0.977%
2	8	?	4.4 %
3	7	Sim	11.7 %
4	6	Sim	20.5 %
5	5	Sim	24.6%
6	4	Sim	20.5 %
7	3	Sim	11.7 %
8	2	?	4.4 %
9	1	Não	0.977%
10	0	Não	0.0976%

- Possíveis resultados:

## Inferência

### Realidade

	Aceitamos $H_0$	Rejeitamos $H_0$
$H_0$ é verdadeira	OK	Erro Tipo 1 (Falso positivo)
$H_0$ é falsa	Erro Tipo 2 (Falso negativo)	OK

- Evento positivo: rejeitar  $H_0$



- Parâmetros do teste:

	Aceitamos H0	Rejeitamos H0
H0 é verdadeira	OK	Erro Tipo 1 (Falso positivo)
H0 é falsa	Erro Tipo 2 (Falso negativo)	OK

- Erro tipo 1 = Taxa de falsos positivos =  $\alpha$
- Erro tipo 2 = Taxa de falsos negativos =  $\beta$
- Nível de significância =  $1 - \alpha$
- Poder do teste =  $1 - \beta$

- Parâmetros do teste:

	Aceitamos H0	Rejeitamos H0
H0 é verdadeira	OK	Erro Tipo 1 (Falso positivo)
H0 é falsa	Erro Tipo 2 (Falso negativo)	OK

- $\alpha$ : Porcentagem das vezes que detectamos uma variação, mas, na realidade, não há. (Podemos encontrar a variação por acaso)
- $1-\beta$ : Porcentagem das vezes que detectamos uma variação, caso ela exista (Se houver mudança, vamos encontrar)

Se  $\alpha$  for alto: temos maior probabilidade de detectarmos falsas mudanças.

Se  $1-\beta$  for alto: temos maior probabilidade de encontrarmos mudanças, se elas, de fato, houverem.

- 5 parâmetros definem o teste de hipóteses:
- Tamanho da amostra
- Nível de significância ( $1-\alpha$ )
- Poder do teste ( $1-\beta$ )
- Efeito mínimo (significância prática)
- Número de direções do teste
- É necessário definir 4 deles e o outro é automático

<http://www.evanmiller.org/ab-testing/sample-size.html>

T

# Exercício Case:

## Teste A/B para cor do botão





T

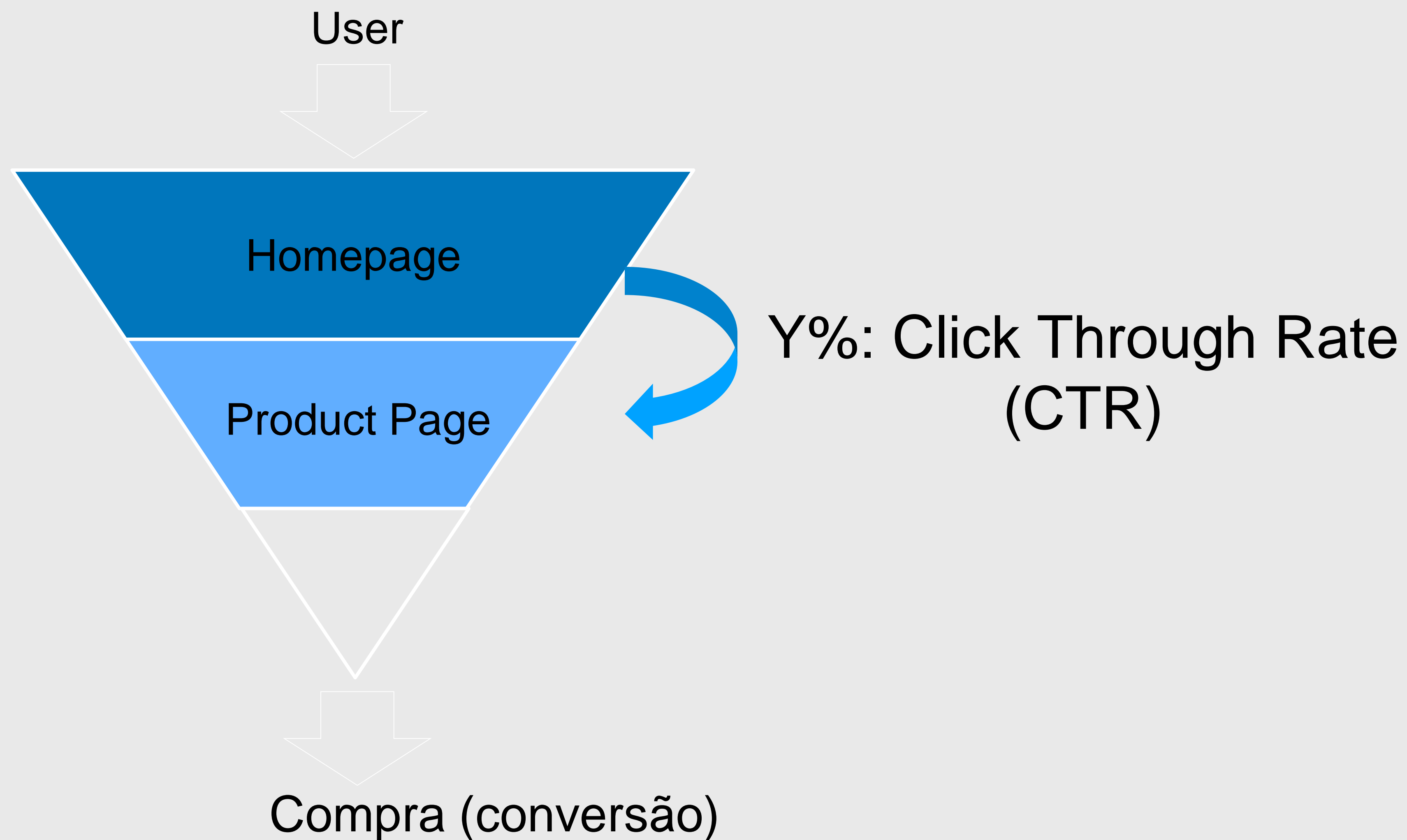
# Exercício Case:

## Teste A/B para cor do botão



# Exercício Case:

Foco: H – P



# Teste A/B: Desafios na prática

- Escolha de métricas:
  - Sensibilidade vs Robustez
  - Viés
- P-Hacking
- Datas festivas (inconsistência)
- Variáveis espúrias (Paradoxo de Simpson)



# Teste A/B: Paradoxo de Simpson

## Tratamentos para pedras no rim

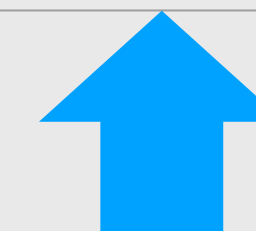
	Tratamento A	Tratamento B
Pedras pequenas	93% (81/87)	87% (234/270)
Pedras grandes	73% (192/263)	69% (55/80)
Total		



## Teste A/B: Paradoxo de Simpson

### Tratamentos para pedras no rim

	Tratamento A	Tratamento B
Pedras pequenas	<b>93%</b> (81/87)	87% (234/270)
Pedras grandes	<b>73%</b> (192/263)	69% (55/80)
Total	78% (273/350)	<b>83%</b> (289/350)



# Teste A/B: Paradoxo de Simpson

Desbalanceamento dos dados nos tratamentos  
Variáveis espúrias!

	Tratamento A	Tratamento B
Pedras pequenas	93% (81/87)	87% (234/270)
Pedras grandes	73% (192/263)	69% (55/80)
Total	78% (273/350)	83% (289/350)



## Case final: Desconto frete

**Hipótese:** Dar desconto no frete aumentará receita

**Perigo:** Podemos ter prejuízo

Como modelar o teste A/B?

T

OBRIGADO!