Aula 01. Introdução à Estatística Inferencial e Conceitos de Amostragem

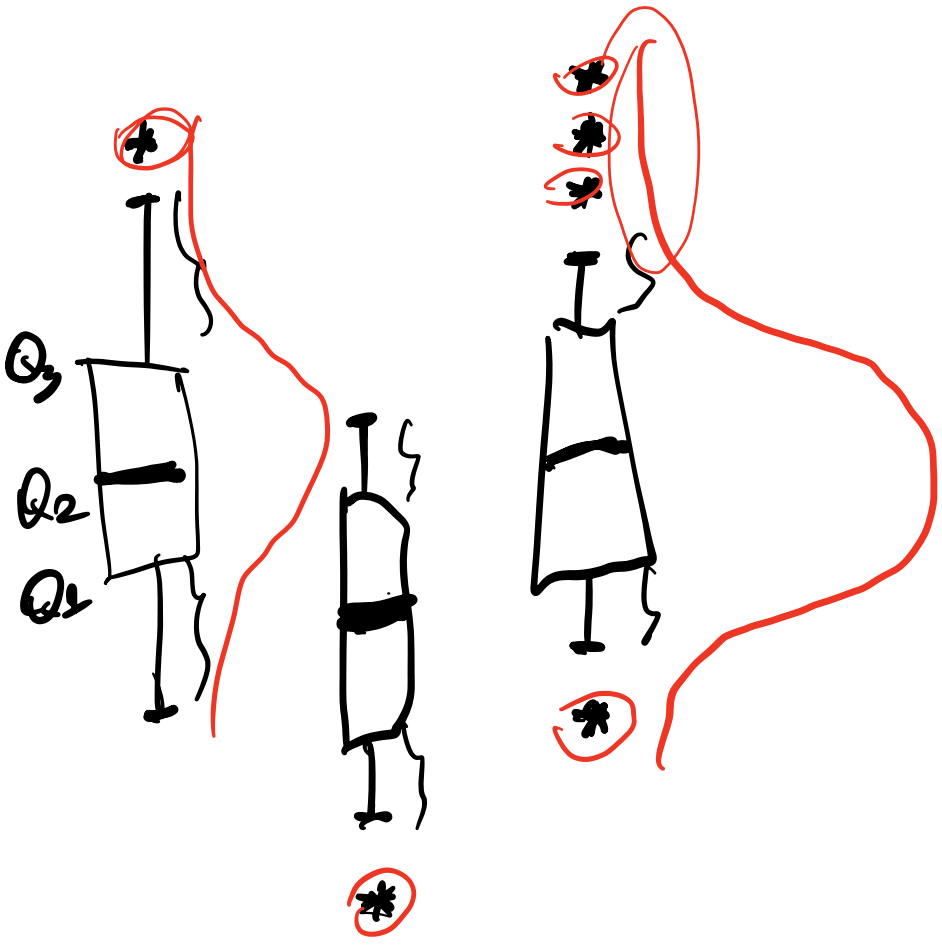
Estatística Inferencial

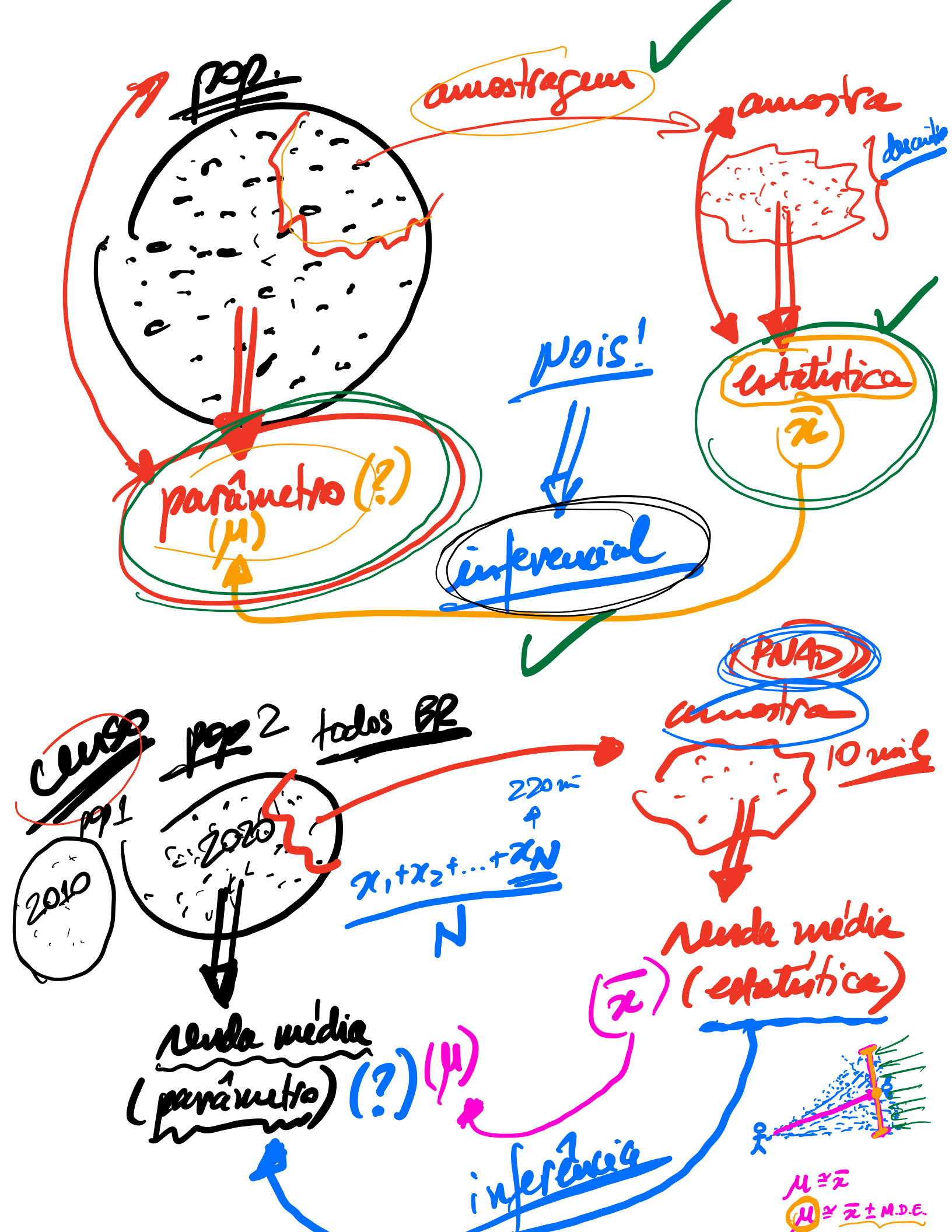
MBA CDIA
ENAP - Escola Nacional de Administração Pública
2025

id	...	...	renda 1	renda 2
1			80.000,00	50-100k
2			65.530,00	50-100k
3			130.422,00	100-150k
4			40.507,28	<50k
5			260.600,00	>150k
.			.	.
.			.	.
.			.	.
n			79.483,00	50-100k

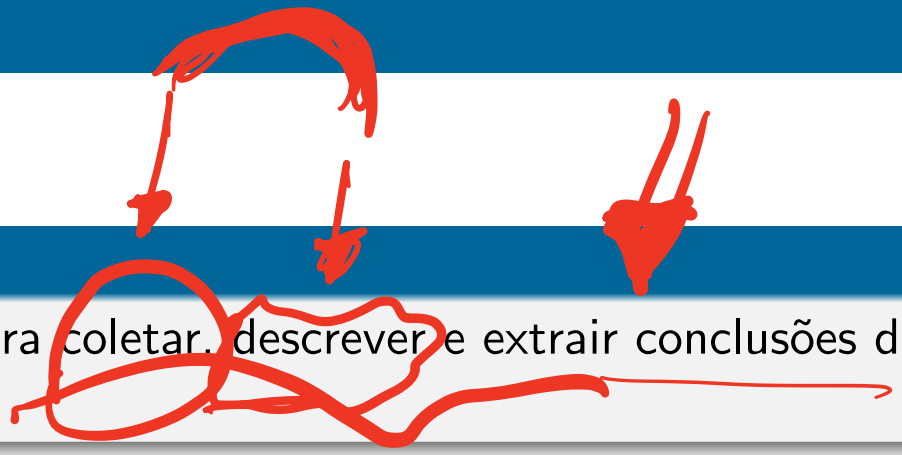
67

78




pop.
amostragem
amostra
descarte
estatística
$\bar{x}$
parâmetro (?)
(μ)
dois!
inferencial
censo
pop 1
2010
pop 2
todos BR
2020
$\frac{x_1 + x_2 + \ldots + x_N}{N}$
220 mi
PNAD
amostra
10 mil
renda média
(parâmetro)
(?) (μ)
renda média
(estatística)
$(\bar{x})$
inferência
$\mu \pm \bar{x}$
$\mu = \bar{x} \pm$ M.D.E.

O que é estatística?



Definição

Estatística é o estudo de procedimentos para coletar, descrever e extrair conclusões de informações.

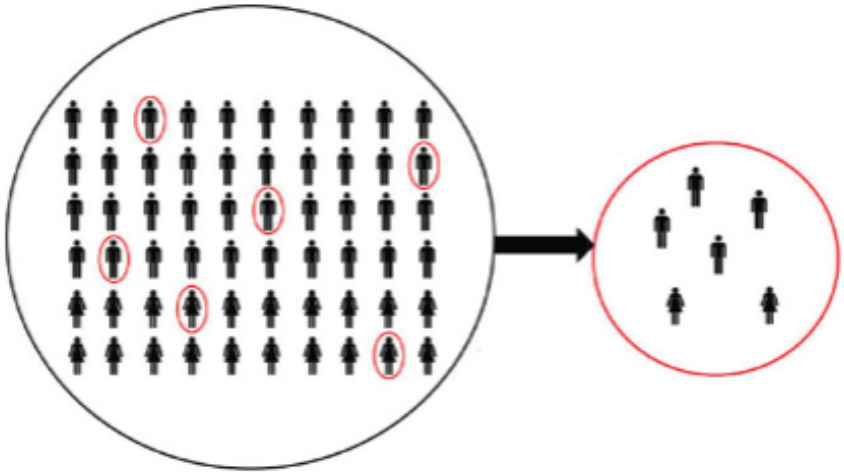
$\Rightarrow$ Estatística é um processo investigativo que envolve

- formular questões
- coletar dados
- descrever e resumir dados (estatística descritiva)
- fazer generalizações a partir dos dados (estatística inferencial)

Ex: Processo de pesquisas eleitorais, previsão do tempo, ensaios clínicos, etc.

População vs. amostra

- Uma **população** é toda a coleção de indivíduos sobre os quais se busca informação.
- Uma **amostra** é um subconjunto da população, contendo os indivíduos que são realmente observados.



https://nces.ed.gov

População vs. amostra

Exemplos de população e amostra:

- **população:** todos os servidores públicos federais atualmente ativos
  amostra: 50 servidores federais selecionados aleatoriamente
- **população:** todos os acidentes de trânsito no Brasil no ano passado
  amostra: acidentes de trânsito no Distrito Federal no ano passado
- **população:** todas as vagas de emprego para analista de políticas públicas
  amostra: os 30 primeiros resultados de busca por vagas de analista de políticas públicas em 1° de janeiro de 2025

Exemplo 1

No artigo científico "Taxas de Falha de Detectores Residenciais de Monóxido de Carbono nos Estados Unidos", foi declarado que existem 38 milhões de detectores de monóxido de carbono instalados nos Estados Unidos. Quando 30 deles foram selecionados aleatoriamente e testados, descobriu-se que 12 deles falharam em fornecer um alarme em condições perigosas de monóxido de carbono.

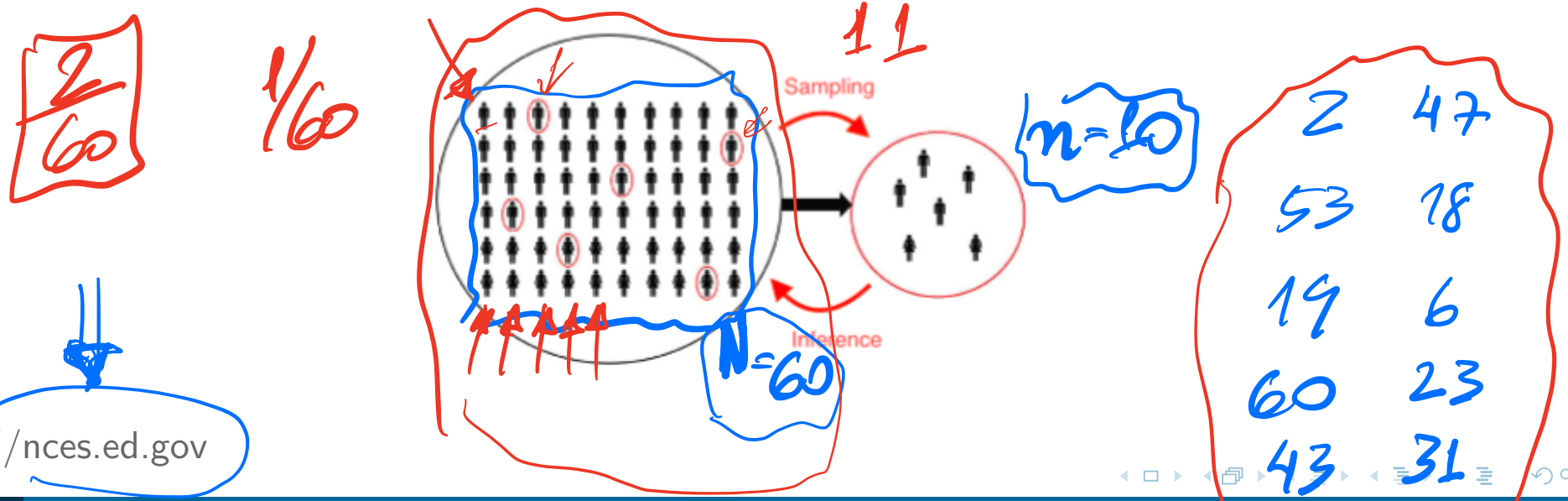
a. No estudo estatístico descrito acima, identifique a população e a amostra.

b. O que o estudo sugere sobre os detectores de monóxido de carbono nos Estados Unidos?

$$\hat{p} = \frac{12}{30} = 0.4 \Rightarrow 40\%$$

Amostragem

- Coletar dados de toda a população é frequentemente impraticável, ou até impossível.
- Em estatística, fazemos inferências sobre a população baseadas em observações feitas em uma amostra.
- Uma boa amostra deve representar a população o mais próximo possível.
- Discutiremos métodos básicos de amostragem usados para obter uma boa amostra.



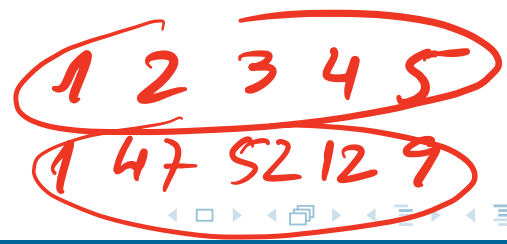
Amostragem aleatória simples

O método de amostragem mais básico (na maioria dos casos o melhor) é a amostragem aleatória simples.

Definição

Uma amostra aleatória simples (AAS) de tamanho n é uma amostra escolhida por um método no qual toda coleção de n itens da população tem igual probabilidade de compor a amostra.

Ex: Selecionar um bilhete premiado em uma loteria.



Implementação de AAS

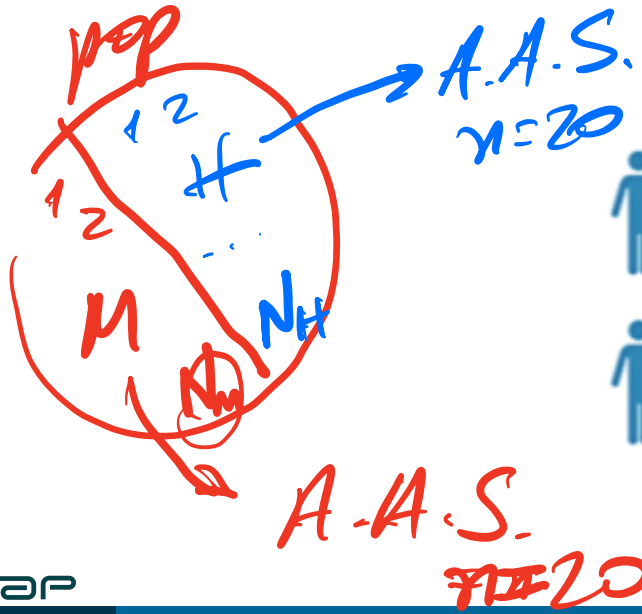
- Suponha que há 300 servidores em um determinado órgão público e que queremos extrair uma AAS de 20 servidores.
- AAS pode ser construída da seguinte forma:

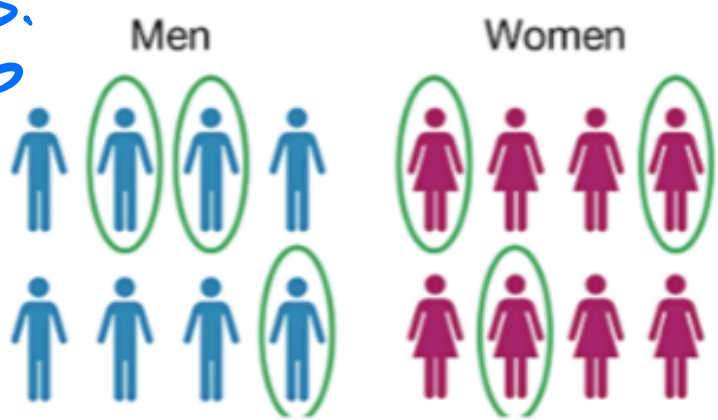
Passo 1: Fazer uma lista dos 300 servidores e numerá-los de 1 a 300.
Passo 2: Gerar 20 números aleatórios entre 1 e 300.
Passo 3: Selecionar os servidores que correspondem a esses números.

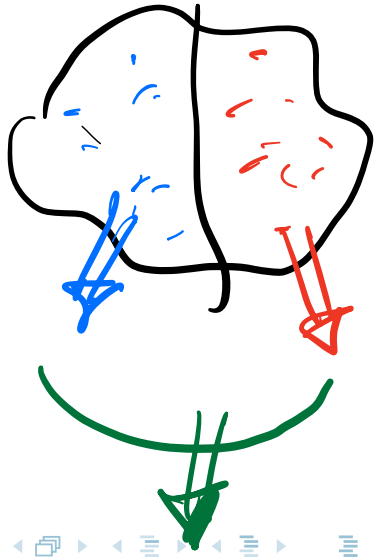
Nota: Qualquer grupo de $n = 20$ servidores tem igual probabilidade de ser selecionado no procedimento acima.

Amostragem estratificada

- Na **amostragem estratificada**, a população é dividida em grupos, chamados estratos, onde os membros de cada estrato são similares de alguma forma. Então uma AAS é extraída de cada estrato.
- Amostragem estratificada é útil quando os estratos diferem uns dos outros, mas indivíduos dentro de um estrato tendem a ser similares.







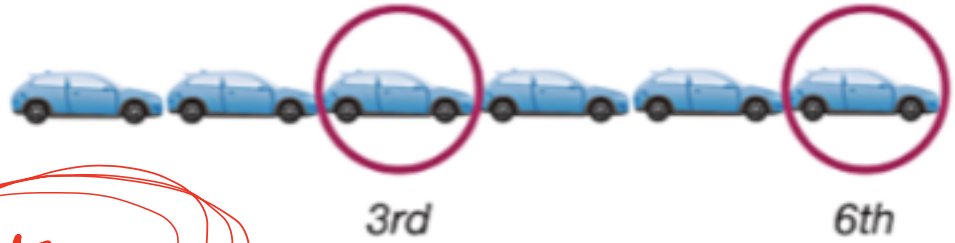
Amostragem por conglomerados

clusters

- Na **amostragem por conglomerados**, itens são extraídos da população em grupos, ou conglomerados.
- Amostragem por conglomerados é útil quando a população é muito grande e espalhada para amostragem aleatória simples.

Amostragem sistemática

- Na **amostragem sistemática**, um ponto de partida é escolhido aleatoriamente e então cada k-ésimo item da população é selecionado.
- Amostragem sistemática é às vezes usada para amostrar produtos conforme saem de uma linha de montagem, para verificar se atendem aos padrões de qualidade.



Outros tipos de amostra

- Note que todos os métodos de amostragem que discutimos têm componente aleatório.
- Uma **amostra de conveniência** é uma amostra que não é extraída por um método aleatório bem definido.
- Frequentemente, itens são incluídos na amostra de conveniência porque são fáceis de localizar ou medir.

Exemplos:

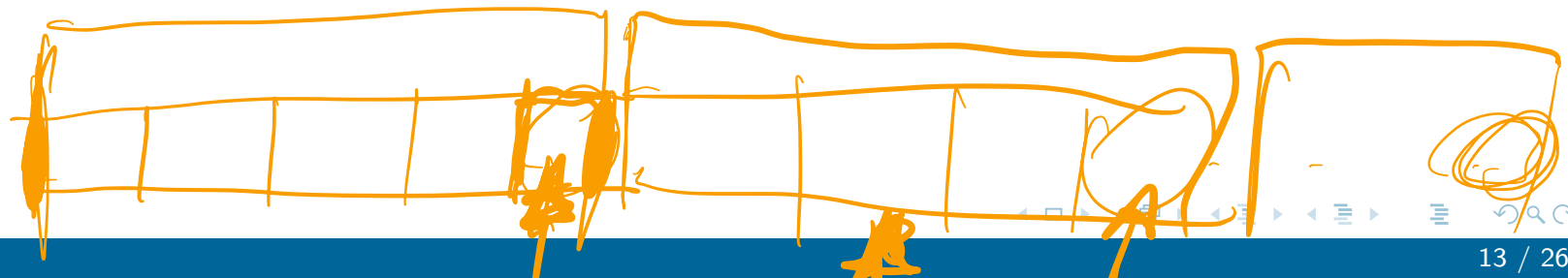
- usar os servidores de um único departamento para representar todos os servidores do órgão
- pesquisar as primeiras 10 pessoas que saem de um grande evento
- Amostras de conveniência não são prováveis de ser representativas da população.

Outros tipos de amostra

- **Amostras de resposta voluntária** (ou amostras auto-selecionadas) são compostas por pessoas que escolheram se incluir na amostra.

Exemplos:

- uma enquete postada em um website
- cartões de feedback de usuários de serviços públicos
- Amostras de resposta voluntária são sempre enviesadas porque os indivíduos que respondem voluntariamente provavelmente terão opiniões mais fortes (positivas ou negativas) que o resto da população.



Exemplo 2

Identifique o tipo de amostra que é descrito.

a. Um pesquisador caminha por um centro administrativo e aborda servidores que passam para perguntar sobre satisfação no trabalho.

b. Para coletar dados sobre qualidade do atendimento, um pesquisador coletou avaliações de atendimento a cada 5 usuários atendidos.

c. Cem servidores participam de um evento e cada um recebe um número. Os 100 números são colocados em uma caixa e completamente misturados, e 10 deles são sorteados para uma pesquisa.

d. Para estimar o número de processos em um arquivo, um analista selecionou aleatoriamente 20 gavetas diferentes e contou todos os processos em cada uma.

e. Em uma pesquisa sobre gestão pública, um órgão foi selecionado de cada página do diretório de órgãos federais.

Definições

5%. trimmed mean 5%.

- Frequentemente usamos números para descrever, ou resumir, uma amostra ou uma população.
- Um **parâmetro** é um número que descreve uma população inteira.
- Uma **estatística** é um número que descreve uma amostra.

Ex. Descreva se o número descrito é uma estatística ou um parâmetro.

a. Dos deputados federais brasileiros, 300 deles votaram a favor de uma medida específica.

b. Em uma amostra de 100 usuários de um serviço público que utilizaram um novo sistema digital, 78% deles relataram melhoria na experiência.

Implementação de Técnicas de Amostragem

Google Colab

Exercícios para fixação dos conceitos

- Amostragem Aleatória Simples
- Amostragem Estratificada
- Amostragem Sistemática
- Análise Comparativa

Exercício 1: Amostragem Aleatória Simples

Enunciado: Você possui uma lista com 1000 servidores públicos federais. Implemente uma função para extrair uma amostra aleatória simples de 50 servidores e calcule a idade média da amostra.

Dicas:

- Use `numpy.random.choice()` ou `random.sample()`
- Gere dados simulados de idades entre 25 e 65 anos
- Compare a média amostral com a média populacional

Bibliotecas necessárias:

```
import numpy as np
import pandas as pd
import random
```

Gabarito - Exercício 1

```
import numpy as np
import pandas as pd
import random

# Gerar populacao de 1000 servidores
np.random.seed(42)  # Para reprodutibilidade
populacao = {
    'servidor_id': range(1, 1001),
    'idade': np.random.normal(45, 10, 1000).round().astype(int)
}
populacao['idade'] = np.clip(populacao['idade'], 25, 65)
df_populacao = pd.DataFrame(populacao)

print(f"Populacao total: {len(df_populacao)} servidores")
print(f"Idade media populacional: {df_populacao['idade'].mean():.2f}")

# Amostragem Aleatoria Simples
def amostra_aleatoria_simples(df, tamanho_amostra):
    return df.sample(n=tamanho_amostra, random_state=42)

amostra = amostra_aleatoria_simples(df_populacao, 50)
print(f"\nAmostra: {len(amostra)} servidores")
print(f"Idade media amostral: {amostra['idade'].mean():.2f}")
print(f"Diferença: {abs(df_populacao['idade'].mean() - amostra['idade'].mean()):.2f}")
```

Exercício 2: Amostragem Estratificada

Enunciado: Considerando a mesma população de servidores, agora dividida em estratos por nível de escolaridade (Medio, Superior,
Pos$_g$*raduacao*), *implementeamostragemestratificadaproporcional*.

Dicas:

- Calcule a proporção de cada estrato na população
- Mantenha essas proporções na amostra
- Use `pandas.groupby()` para trabalhar com estratos

Objetivo: Extrair amostra de 60 servidores mantendo as proporções dos estratos.

Gabarito - Exercício 2

```
# Adicionar estratos a populacao
escolaridade = np.random.choice(['Medio', 'Superior', 'Pos_graduacao'],
                                1000, p=[0.3, 0.5, 0.2])
df_populacao['escolaridade'] = escolaridade

# Verificar distribuicao populacional
print("Distribuicao por escolaridade na populacao:")
prop_populacao = df_populacao['escolaridade'].value_counts(normalize=True)
print(prop_populacao)

# Amostragem estratificada proporcional
def amostra_estratificada(df, coluna_estrato, tamanho_total):
    amostras = []
    for estrato, grupo in df.groupby(coluna_estrato):
        prop = len(grupo) / len(df)
        tamanho_estrato = int(tamanho_total * prop)
        if tamanho_estrato > 0:
            amostra_estrato = grupo.sample(n=tamanho_estrato, random_state=42)
            amostras.append(amostra_estrato)
    return pd.concat(amostras, ignore_index=True)

amostra_estratificada_df = amostra_estratificada(df_populacao, 'escolaridade', 60)
print(f"\nAmostra estratificada: {len(amostra_estratificada_df)} servidores")
print("Distribuição por escolaridade na amostra:")
prop_amostra = amostra_estratificada_df['escolaridade'].value_counts(normalize=True)
print(prop_amostra)
```

Exercício 3: Amostragem Sistemática

Enunciado: Implemente amostragem sistemática para selecionar 40 servidores de uma lista ordenada por número de matrícula. Compare os resultados com amostragem aleatória simples.

Dicas:

- Calcule o intervalo k = população / amostra
- Escolha um ponto de partida aleatório entre 1 e k
- Selecione cada k-ésimo elemento

Desafio: Analise se há diferença significativa entre as médias obtidas pelos dois métodos.

Gabarito - Exercício 3

```
# Amostragem sistematica
def amostra_sistematica(df, tamanho_amostra):
    n = len(df)
    k = n // tamanho_amostra  # Intervalo
    inicio = random.randint(0, k-1)  # Ponto de partida aleatório

    indices = [inicio + i*k for i in range(tamanho_amostra)]
    indices = [i for i in indices if i < n]  # Garantir que não exceda o tamanho

    return df.iloc[indices]

# Ordenar por servidor_id para simular lista ordenada
df_ordenado = df_populacao.sort_values('servidor_id').reset_index(drop=True)

amostra_sistematica_df = amostra_sistematica(df_ordenado, 40)
amostra_aleatoria_df = df_populacao.sample(n=40, random_state=42)

print("Comparação de metodos:")
print(f"Amostra sistemática - Idade media: {amostra_sistematica_df['idade'].mean():.2f}")
print(f"Amostra aleatória - Idade media: {amostra_aleatoria_df['idade'].mean():.2f}")
print(f"População - Idade media: {df_populacao['idade'].mean():.2f}")

diferenca = abs(amostra_sistematica_df['idade'].mean() - amostra_aleatoria_df['idade'].mean())
print(f"Diferença entre metodos: {diferenca:.2f} anos")
```

Exercício 4: Análise Comparativa Completa

Enunciado: Implemente todos os três métodos de amostragem estudados e crie uma análise comparativa visualizando:

- Histogramas das distribuições de idade para cada método
- Estatísticas descritivas comparativas
- Análise da representatividade de cada método

Dicas:

- Use `matplotlib` ou `seaborn` para visualizações
- Calcule média, mediana, desvio padrão para cada amostra
- Compare com os parâmetros populacionais

Gabarito - Exercício 4 - Parte 1

```
import matplotlib.pyplot as plt
import seaborn as sns

# Obter amostras pelos tres métodos
aas = amostra_aleatoria_simples(df_populacao, 50)
estratificada = amostra_estratificada(df_populacao, 'escolaridade', 50)
sistematica = amostra_sistematica(df_ordenado, 50)

# Análise estatistica
resultados = pd.DataFrame({
    'Metodo': ['População', 'AAS', 'Estratificada', 'Sistematica'],
    'Media': [df_populacao['idade'].mean(), aas['idade'].mean(),
              estratificada['idade'].mean(), sistematica['idade'].mean()],
    'Mediana': [df_populacao['idade'].median(), aas['idade'].median(),
                estratificada['idade'].median(), sistematica['idade'].median()],
    'Desvio Padrao': [df_populacao['idade'].std(), aas['idade'].std(),
                      estratificada['idade'].std(), sistematica['idade'].std()]
})

print("Analise Comparativa:")
print(resultados.round(2))
```

Gabarito - Exercício 4 - Parte 2

```
# Visualizacao
plt.figure(figsize=(12, 8))

plt.subplot(2, 2, 1)
plt.hist(df_populacao['idade'], bins=20, alpha=0.7)
plt.title('População')

plt.subplot(2, 2, 2)
plt.hist(aas['idade'], bins=15, alpha=0.7)
plt.title('AAS')

plt.subplot(2, 2, 3)
plt.hist(estratificada['idade'], bins=15, alpha=0.7)
plt.title('Estratificada')

plt.subplot(2, 2, 4)
plt.hist(sistematica['idade'], bins=15, alpha=0.7)
plt.title('Sistemática')

plt.tight_layout()
plt.show()
```

Conclusões e Próximos Passos

O que aprendemos hoje:

- Diferença entre população e amostra
- Métodos de amostragem probabilística
- Implementação prática em Python
- Comparação entre diferentes técnicas

Próxima aula:

- Distribuições Amostrais
- Teorema Central do Limite

Dúvidas?