

# Introdução a vies

---

O **viés em estatística** é um conceito fundamental que se refere a desvios sistemáticos entre estimativas obtidas a partir de amostras e os valores reais dos parâmetros populacionais. Esse desvio pode ocorrer devido a erros no processo de coleta, seleção ou análise dos dados, levando a conclusões incorretas sobre a população estudada.

---

## Definições Fundamentais

No contexto estatístico, viés é definido como

“Um erro sistemático, ou desvio da verdade, nos resultados ou inferências.

Essa definição enfatiza que o viés não é resultado do acaso, mas sim de falhas sistemáticas que afetam a precisão das estimativas

---

## Tipos Comuns de Viés

### 1. Viés de Seleção

**Definição:** Ocorre quando a amostra escolhida **não representa adequadamente a população**, geralmente por **critério de inclusão/exclusão** inadequado.

#### Exemplo:

Uma universidade deseja avaliar a satisfação dos alunos com os serviços acadêmicos. No entanto, ela aplica o questionário **apenas em alunos que frequentam a biblioteca**.

- ❌ Problema: estudantes que **não frequentam a biblioteca** (por falta de tempo, por trabalharem, etc.) **ficam de fora** da amostra.
  - ✅ Resultado: a amostra tende a ser composta por alunos mais engajados ou satisfeitos, distorcendo os resultados gerais.
- 

### 2. Viés de Não Resposta

**Definição:** Ocorre quando **uma parte significativa da amostra não responde** à pesquisa, e esses não-respondentes têm **características diferentes** dos que responderam.

#### Exemplo:

Um instituto envia um questionário sobre **nível de estresse no trabalho** para 1.000 funcionários. Apenas 200 respondem.

- ❌ Problema: os que estão mais estressados podem **não ter tempo ou disposição** para responder, ou podem **evitar se expor**.
  - ✅ Resultado: a média de estresse será **subestimada**, pois os casos mais graves estão ausentes.
-



### 3. Viés de Medição

**Definição:** Ocorre quando o **instrumento ou método de coleta de dados é impreciso**, causando erro sistemático.



#### Exemplo:

Um estudo quer medir o peso médio de recém-nascidos. A balança usada está **descalibrada e adiciona sempre 150g a mais**.

- Problema: todos os dados coletados estão **sistematicamente incorretos**.
- Resultado: a média será **superestimada**, mesmo com uma amostra aleatória bem feita.



### 4. Viés de Confirmação

**Definição:** Acontece quando os pesquisadores **buscam ou interpretam evidências** que **confirmem hipóteses pré-existent**s, ignorando evidências contrárias.



#### Exemplo:

Um pesquisador acredita que **alunos que usam aplicativos educacionais têm melhor desempenho**. Ele coleta dados e dá mais atenção aos casos que confirmam isso, ignorando ou explicando com desculpas os casos que contradizem.

- Problema: a interpretação se torna **parcial**, e os resultados **não são objetivos**.
- Resultado: o estudo **reflete crenças do pesquisador** mais do que a realidade.



### Viés em Modelos Estatístico

David A. Freedman destacou que, em modelos de regressão, especialmente quando o número de variáveis explicativas é grande em relação ao número de observações, é comum identificar variáveis não relacionadas como estatisticamente significativas. Esse fenômeno, conhecido como "paradoxo de Freedman", ilustra como a seleção de modelos pode introduzir viés nas análises.



### Viés em Revisões Sistemáticas

A Cochrane Collaboration, reconhecida por suas revisões sistemáticas rigorosas, utiliza ferramentas específicas para avaliar o risco de viés em estudos clínicos. O viés é categorizado em diferentes domínios, como viés de seleção, desempenho e relato, permitindo uma avaliação abrangente da qualidade dos estudos incluídos [\[cite\]turn0search4](#).



### Importância de Identificar e Corrigir o Viés

A presença de viés pode comprometer a validade das conclusões estatísticas, levando a decisões baseadas em informações distorcidas. Portanto, é crucial:

- Utilizar métodos de amostragem adequados para garantir representatividade.

- Empregar instrumentos de medição calibrados e procedimentos padronizados.
- Aplicar técnicas estatísticas apropriadas que considerem possíveis fontes de vies.

## Definição Formal do Vies


O **vies de um estimador**  $\hat{\theta}$  em relação ao parâmetro verdadeiro  $\theta$  é dado por:

$$\text{Vies}(\hat{\theta}) = E[\hat{\theta}] - \theta$$

Onde:

- $E[\hat{\theta}]$ : Valor esperado (médio) do estimador  $\hat{\theta}$
- $\theta$ : Valor real do parâmetro da população

$$\text{Vies}(\hat{\theta}) = \mathbb{E}[\hat{\theta}] - \theta$$

 Leitura em português:

**"Vies de theta chapéu é igual à esperança de theta chapéu menos theta."**

 Explicando os termos:

Símbolo	Leitura	Significado
$\hat{\theta}$	<b>theta chapéu</b>	Um estimador (ex: média amostral) do parâmetro verdadeiro $\theta$
$\theta$	<b>theta</b>	Parâmetro verdadeiro da população (ex: média populacional)
$\mathbb{E}[\hat{\theta}]$	<b>esperança de theta chapéu</b>	Valor esperado do estimador, ou seja, média de seus resultados ao repetir a amostragem infinitamente

## Exemplo Prático para Excel

 Cenário

Vamos estimar a **média da população** com três amostragens diferentes. A população real tem:

- População: [5, 10, 15, 20, 25]
- Média populacional verdadeira  $\theta = 15$

Você coleta três amostras simples com reposição:

- Amostra 1: [10, 15, 20] → Média = 15

- Amostra 2: [5, 10, 15] → Média = 10
- Amostra 3: [15, 20, 25] → Média = 20

Agora, vamos calcular o viés do estimador da média.



## Como montar no Excel

Amostra	Valores	Média da Amostra ( $\hat{\theta}$ )
Amostra 1	10, 15, 20	=MÉDIA(10;15;20) → 15
Amostra 2	5, 10, 15	=MÉDIA(5;10;15) → 10
Amostra 3	15, 20, 25	=MÉDIA(15;20;25) → 20
Média das Médias ( $E[\hat{\theta}]$ )	=MÉDIA(15;10;20) → 15	
Valor Real da População ( $\theta$ )	15	
<b>Viés</b>	=15 - 15	0



## Interpretação

- Como o **valor esperado do estimador** (média das médias amostrais) é **igual ao valor verdadeiro da população**, o **viés é zero**.
- Isso mostra que, nesse caso, o estimador da média **não é enviesado**.



## O que é o Teorema do Erro Quadrático Médio (MSE)?

Ele mostra que o erro total de um estimador pode ser decomposto em **três componentes**:

$$\text{MSE}(\hat{\theta}) = \underbrace{(\mathbb{E}[\hat{\theta}] - \theta)^2}_{\text{Viés}^2} + \underbrace{\text{Var}(\hat{\theta})}_{\text{Variância}} + \underbrace{\text{Erro irreduzível}}_{\text{(às vezes negligenciado)}}$$



Termos explicados:

Componente	Significado
<b>Viés</b>	Diferença sistemática entre o valor esperado do estimador e o valor real
<b>Variância</b>	Quão sensível o estimador é a mudanças na amostra



Componente	Significado
<b>Erro irreduzível</b>	Ruído ou aleatoriedade natural no processo, que nenhum modelo consegue capturar

## O trade-off: Viés vs. Variância

- **Modelos com alto viés** tendem a **subestimar** a complexidade dos dados. São simples demais (ex: média constante). **Erros sistemáticos**.
- **Modelos com alta variância** são **muito sensíveis aos dados de entrada**. Geralmente se ajustam demais ao acaso (overfitting).

A **melhor solução** geralmente não é o modelo com o menor viés nem o de menor variância, mas **aquele com o menor MSE**.

 Visualização intuitiva (analogia):

	Baixo Viés + Alta Variância	Alto Viés + Baixa Variância
 Tiros no alvo	Espalhados, mas em média no centro	Agrupados, mas longe do centro
 Interpretação	Modelo complexo e instável	Modelo simples, mas incorreto

## Exemplo numérico simples:

Suponha que você quer estimar a média verdadeira de uma população:  $\theta = 10$

Você testa dois estimadores:

Estimador A:

- Média esperada:  $\mathbb{E}[\hat{\theta}] = 9$
- Variância:  $\text{Var}(\hat{\theta}) = 1$

$$\text{MSE}_A = (9 - 10)^2 + 1 = 1 + 1 = 2$$

Estimador B:

- Média esperada:  $\mathbb{E}[\hat{\theta}] = 10$
- Variância:  $\text{Var}(\hat{\theta}) = 4$

$$\text{MSE}_B = (10 - 10)^2 + 4 = 0 + 4 = 4$$

**Mesmo com viés**, o estimador A tem **menor erro quadrático médio** e é preferível nesse caso!

## Referências confiáveis

- HASTIE, Trevor; TIBSHIRANI, Robert; FRIEDMAN, Jerome. *The Elements of Statistical Learning*. Springer, 2009. (Cap. 2)
- GÉRON, Aurélien. *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*. O'Reilly Media, 2019. (Cap. 4)

# Introdução a amostragem

---

## O que é Teoria da Amostragem?

A **Teoria da Amostragem** é um ramo da estatística que estuda métodos e princípios para selecionar subconjuntos (amostras) de uma população com o objetivo de fazer inferências sobre toda a população. Ao invés de analisar todos os elementos de um universo (o que muitas vezes é inviável por razões econômicas, logísticas ou temporais), a teoria da amostragem busca formas eficientes de **representar a totalidade por meio de uma parte**.

Segundo **Triola (2015)**:

*"Amostragem é o processo de selecionar membros de uma população de forma que as inferências sobre a população possam ser feitas com base nas informações obtidas da amostra."*  
— Mario Triola, *Introdução à Estatística*.

O objetivo principal é garantir que a amostra seja **representativa** — ou seja, que reflita de maneira fiel as características da população de interesse. Isso permite a utilização de técnicas estatísticas para **estimar parâmetros populacionais** com base nas estatísticas amostrais.

De acordo com **Barbetta (2010)**:

*"A teoria da amostragem preocupa-se com a forma de se obter uma amostra representativa, de modo a possibilitar generalizações confiáveis para a população."*  
— Pedro Barbetta, *Estatística Aplicada às Ciências Sociais*.

Além disso, a teoria também lida com a **mensuração e controle dos erros**, especialmente o **erro amostral**, que é a diferença entre o valor estimado com base na amostra e o valor real do parâmetro populacional.

Segundo **Wonnacott & Wonnacott (1990)**:

*"A principal preocupação da teoria da amostragem é avaliar com que grau de confiança e precisão podemos estender conclusões obtidas a partir de uma amostra para a população como um todo."*  
— Wonnacott & Wonnacott, *Estatística*.

---

Resumo com palavras suas para aula

A **teoria da amostragem** é como o guia que nos ensina a escolher "um pedacinho" do todo de forma inteligente e criteriosa, para que possamos estudar esse pedacinho e aprender sobre o todo. Usamos isso quando não dá para medir tudo — como em uma eleição, onde ouvimos milhares de eleitores para tentar entender milhões.

## Objetivos da Amostragem

A **amostragem** é uma técnica estatística que visa estudar uma **parte representativa** de uma população para fazer **inferências sobre o todo**. Em vez de analisar todos os elementos de um grupo (o que, na prática, muitas vezes é inviável), os estatísticos trabalham com amostras, desde que essas sejam cuidadosamente selecionadas. A seguir, destacam-se os principais **objetivos da amostragem** na teoria estatística:

---

### 1. Economia de Tempo e Recursos

Um dos principais motivos para se utilizar amostragem é a **redução de custos, tempo e esforço** envolvidos na coleta e análise de dados.

- Realizar um **censo completo** (isto é, estudar todos os elementos da população) pode ser extremamente caro, demorado e, em muitos casos, inviável.
- A amostragem permite obter **resultados mais rapidamente**, possibilitando que decisões sejam tomadas com agilidade, o que é essencial em áreas como saúde, marketing, economia e políticas públicas.

*Exemplo real:* Uma empresa quer saber a satisfação dos seus 500 mil clientes. Aplicar uma pesquisa com todos seria lento e caro. Com uma amostra bem planejada, pode-se ter uma estimativa confiável da opinião geral com muito menos recursos.

"A amostragem permite uma redução significativa nos custos e no tempo de execução de pesquisas, sem comprometer a qualidade dos resultados, desde que a amostra seja representativa."

— Barbetta, 2010

---

### 2. Estimativa de Parâmetros com Precisão

Outro objetivo essencial da amostragem é a **estimativa de parâmetros populacionais** com um bom grau de **precisão e confiabilidade**.

- Parâmetros populacionais são valores verdadeiros da população, como média, proporção, desvio padrão etc.
- Como esses valores são, muitas vezes, desconhecidos, usamos **estatísticas amostrais** para **estimar** os parâmetros com **margem de erro e intervalo de confiança**.

A amostragem bem conduzida permite inferir, por exemplo:

- A média de renda de uma população.
- A taxa de aprovação de um produto ou político.

- A proporção de indivíduos com uma determinada doença.

O importante é que essas estimativas sejam feitas com **riscos controlados** de erro, especialmente o **erro amostral**, que decorre da variabilidade natural entre diferentes amostras.

“A teoria da amostragem nos dá ferramentas para calcular o erro associado às estimativas, garantindo que os resultados possam ser generalizados para a população com um grau conhecido de precisão.”

— Wonnacott & Wonnacott, 1990

Claro! Vamos entender o **viés estatístico** a partir de sua **formulação matemática**, com um **exemplo passo a passo**, e em seguida explicar **quando usá-la**.

## Fórmula Matemática do Viés

O **viés de um estimador**  $\hat{\theta}$  em relação a um parâmetro populacional  $\theta$  é dado por:

$$\text{Viés}(\hat{\theta}) = \mathbb{E}[\hat{\theta}] - \theta$$

- $\mathbb{E}[\hat{\theta}]$ : valor esperado (ou média) do estimador.
- $\theta$ : valor real do parâmetro da população.

## Exemplo Passo a Passo

**Contexto:** Suponha que queremos estimar a **média** da população  $\mu = 5$ , e usamos um **estimador enviesado**, como a **média truncada** (que ignora o maior valor).

### Passo 1: População e parâmetro

Considere a população:

$$\{3, 5, 7\}$$

$$\Rightarrow \mu = \frac{3 + 5 + 7}{3} = 5$$

### Passo 2: Todas as amostras possíveis (de tamanho 2, sem reposição)

- (3,5)
- (3,7)
- (5,7)

### Passo 3: Calcular o estimador enviesado — média truncada (descartando o maior valor da amostra)

Amostra	Maior valor	Média truncada
---------	-------------	----------------



Amostra	Maior valor	Média truncada
(3,5)	5	3
(3,7)	7	3
(5,7)	7	5

#### Passo 4: Valor esperado do estimador

$$\mathbb{E}[\hat{\mu}] = \frac{3 + 3 + 5}{3} = \frac{11}{3} \approx 3,67$$

#### Passo 5: Cálculo do viés

$$\text{Viés}(\hat{\mu}) = \mathbb{E}[\hat{\mu}] - \mu = 3,67 - 5 = -1,33$$

 **Interpretação:** O estimador subestima o valor real da média da população.

## Quando usar essa fórmula?

Você deve usar a fórmula do viés quando:

- Está comparando diferentes estimadores de um parâmetro (por exemplo, média amostral vs. média truncada).
- Quer verificar se um **estimador é insesgado** (bias = 0).
- Precisa analisar o **trade-off entre viés e variância** (ex: Teorema de decomposição de erro quadrático médio).

## Observações importantes

- Um estimador **insesgado** tem  $\mathbb{E}[\hat{\theta}] = \theta$ .
- Um estimador **enviesado** pode ser útil se reduzir a variância drasticamente (como na **regularização**).

Se quiser, posso fazer essa simulação em Python também ou aplicar ao caso de **estimativa de proporções**. Deseja que eu continue com isso?

A amostragem é uma solução poderosa para problemas práticos e teóricos na estatística. Seus principais objetivos — **economizar tempo e recursos** e **estimar parâmetros com precisão** — tornam-na indispensável em estudos que exigem agilidade e confiabilidade. Quando bem aplicada, a amostragem permite transformar dados parciais em conhecimento confiável sobre o todo.

# Tipos de Amostragem

---

Na estatística, existem diferentes **tipos de amostragem**, e a escolha do método mais adequado depende dos objetivos da pesquisa, das características da população e dos recursos disponíveis. Os métodos podem ser classificados, principalmente, em dois grandes grupos: **amostragem probabilística** e **amostragem não probabilística**.

---

## 1. Amostragem Probabilística

A amostragem probabilística é um conjunto de métodos de seleção de amostras no qual todos os elementos da população têm uma chance conhecida e diferente de zero de serem incluídos na amostra. Ou seja, a escolha dos elementos é feita com base em princípios de aleatoriedade, o que permite garantir a representatividade da amostra e a possibilidade de inferir estatisticamente os resultados para toda a população.

Esse tipo de amostragem é considerado o mais rigoroso do ponto de vista estatístico, pois permite calcular a margem de erro e os intervalos de confiança, fundamentais para validar cientificamente os resultados obtidos.

“A amostragem probabilística é a base para generalizações confiáveis, pois oferece garantias estatísticas sobre a precisão das estimativas.”

— Triola, 2015

### Por exemplo

#### a) Amostragem Aleatória Simples

Consiste em selecionar elementos de forma completamente aleatória, garantindo que cada membro da população tenha a **mesma chance de ser escolhido**. Pode ser feita com sorteio ou por meio de software.

*Exemplo:* Sortear 10 nomes entre 100 alunos usando uma tabela de números aleatórios.

#### b) Amostragem Sistemática

Nesse método, os elementos são escolhidos com base em um **intervalo fixo (k)** a partir de uma lista ordenada.

*Exemplo:* Em uma lista de 1.000 clientes, escolher 1 a cada 50 nomes (começando por um aleatório entre os 50 primeiros).

#### c) Amostragem Estratificada

A população é dividida em **estratos homogêneos** (grupos com características semelhantes), e uma amostra é tirada de cada estrato, proporcional ou igualmente.

*Exemplo:* Dividir alunos por curso (Engenharia, Administração, Direito) e sortear proporcionalmente em cada grupo.

#### d) Amostragem por Conglomerados (ou Clusters)

Em vez de selecionar indivíduos, são escolhidos **grupos inteiros** (conglomerados) de forma aleatória.  
*Exemplo:* Sortear 5 escolas e entrevistar todos os alunos dessas escolas.

---

## 2. Amostragem Não Probabilística

Seleção dos elementos **não segue critérios aleatórios**, e os elementos da população **não têm chance conhecida** de serem escolhidos. Por isso, os resultados obtidos não podem ser generalizados com o mesmo rigor estatístico.

#### a) Amostragem por Conveniência

Os elementos são escolhidos por serem **de fácil acesso** ao pesquisador.  
*Exemplo:* Entrevistar pessoas que estão passando na frente da faculdade.

#### b) Amostragem por Julgamento (ou Intencional)

O pesquisador escolhe os elementos com base em seu **conhecimento e critérios subjetivos** sobre a população.  
*Exemplo:* Selecionar apenas especialistas para responderem a um questionário técnico.

#### c) Amostragem por Cotas

Seleciona-se uma amostra que representa **proporcionalmente** certas características da população, mas a escolha dos elementos dentro de cada cota não é aleatória.  
*Exemplo:* Garantir que 60% da amostra sejam mulheres e 40% homens, mas escolhendo os participantes por conveniência.

#### d) Amostragem Bola de Neve

Usada em populações de difícil acesso, onde **os primeiros participantes indicam outros**.  
*Exemplo:* Pesquisas com usuários de drogas ou populações marginalizadas.

---

Compreender os diferentes tipos de amostragem é fundamental para garantir a **qualidade e a confiabilidade** de uma pesquisa. Os métodos probabilísticos são ideais quando se deseja fazer inferências estatísticas com maior precisão, enquanto os métodos não probabilísticos são úteis em contextos exploratórios ou quando não há acesso a uma lista completa da população. A escolha do método deve ser feita com base em critérios técnicos, mas também considerando as **limitações práticas** da pesquisa. Abaixo vamos passar um por um com mais detalhes

---

## Tipo de amostragem probabilística

---

### 1. Amostragem Aleatória Simples (AAS)

**Amostragem Aleatória Simples** é o tipo mais básico e fundamental de amostragem probabilística. Consiste em selecionar **elementos de uma população de forma totalmente aleatória**, garantindo que **cada indivíduo tenha exatamente a mesma probabilidade de ser escolhido**.

---

## Conceito

Na AAS, a seleção é feita **sem substituição**, o que significa que, uma vez escolhido, um elemento **não pode ser selecionado novamente** (em geral). A escolha pode ser feita por meio de sorteio manual, tabela de números aleatórios ou softwares estatísticos.

"Em uma amostragem aleatória simples, todos os subconjuntos possíveis de tamanho  $n$  têm a mesma probabilidade de serem selecionados."

— Wonnacott & Wonnacott, 1990

---

## Pré-requisitos

- Uma **lista completa da população** (também chamada de *frame amostral*).
  - Um método para garantir **aleatoriedade na seleção** (sorteio, números aleatórios, etc.).
  - Um **tamanho de amostra definido** ( $n$ ).
- 

## Exemplo Didático

Imagine que você é professor e tem uma **turma com 30 alunos**. Você deseja aplicar uma entrevista com **5 alunos**, escolhidos de forma **justa e aleatória**, para avaliar a opinião da turma sobre o uso de novas tecnologias em sala.

### Passo a passo:

1. **Numerar os alunos** de 1 a 30.
2. Utilizar um método de seleção aleatória. Exemplo: usar uma tabela de números aleatórios ou uma função em Python, Excel, etc.
3. Sortear **5 números distintos entre 1 e 30**.
4. Os alunos correspondentes a esses números formarão a **amostra aleatória simples**.

Se os números sorteados forem 4, 11, 17, 22 e 28, então os alunos com essas numerações serão entrevistados.

---

## Benefícios da AAS

- ✓ **Fácil compreensão e aplicação.**
  - ✓ **Evita viés de seleção**, pois todos têm a mesma chance.
  - ✓ Permite **aplicação direta de fórmulas estatísticas**.
- 

## ⚠ Limitações

- ✗ Exige uma **lista completa e atualizada da população**.
  - ✗ Pode ser **logisticamente difícil em populações grandes e dispersas**.
  - ✗ Não garante representatividade de subgrupos (como homens/mulheres, faixas etárias, regiões, etc.).
- 

## Quando usar?

A Amostragem Aleatória Simples é indicada quando:

- A população é relativamente **pequena e homogênea**.
  - Há **acesso fácil a todos os elementos**.
  - Busca-se **imparcialidade e simplicidade**.
- 

## Ferramentas para selecionar AAS

- **Planilhas eletrônicas** (como Excel → `=ALEATÓRIOENTRE(1;30)`).
  - **Linguagens de programação** (ex: Python `random.sample()`).
  - **Softwares estatísticos** (SPSS, R, SAS, etc.).
  - **Tabelas de números aleatórios** impressas (método tradicional).
- 

A **Amostragem Aleatória Simples** é o alicerce da teoria da amostragem. Sua simplicidade e rigor teórico tornam-na um modelo ideal para estudos iniciais e para situações em que a população é pequena e acessível. No entanto, em cenários mais complexos, pode ser necessário recorrer a métodos mais avançados para garantir representatividade.

---

## **Formulação Matemática da Amostragem Aleatória Simples (AAS)**

A **formulação matemática da AAS** se baseia no conceito de **combinatória**, pois trata da seleção de subconjuntos da população sem reposição e sem importar a ordem.

---

## Definição dos Termos

- $N$ : Tamanho da população (número total de elementos).
  - $n$ : Tamanho da amostra (quantos elementos queremos selecionar).
  - $\binom{N}{n}$ : Número de maneiras de escolher  $n$  elementos de uma população de  $N$ , **sem considerar a ordem**.
  - $P$ : Probabilidade de seleção de uma amostra específica.
- 

## Número Total de Amostras Possíveis

A quantidade de diferentes amostras possíveis de tamanho  $n$  que podem ser retiradas da população de tamanho  $N$  é dada pelo **coeficiente binomial**:

**coeficiente binomial**, também chamado de número binomial, de um número  $N$ , na classe  $k$ , consiste no número de combinações de  $N$  termos,  $k$  a  $k$

$$\binom{N}{n} = \frac{N!}{n! \cdot (N - n)!}$$

Esse valor representa **todas as combinações possíveis** de  $N$  elementos retirados de  $N$ .

Exemplo: Se temos  $N = 5$  elementos e queremos uma amostra de  $n = 2$ , temos:

$$\binom{5}{2} = \frac{5!}{2! \cdot (5-2)!} = \frac{120}{2 \cdot 6} = 10$$

Ou seja, existem **10 amostras possíveis** com 2 elementos retirados de 5.

---

## Probabilidade de Seleção de uma Amostra

Na AAS, cada uma dessas combinações possíveis tem **a mesma chance de ser escolhida**. Então, a probabilidade de uma amostra específica ser selecionada é:

$$P(\text{amostra específica}) = \frac{1}{\binom{N}{n}}$$

**Exemplo:** Com  $N = 5$  e  $n = 2$ , cada amostra de dois elementos tem:

$$P = \frac{1}{10} = 0,1 \quad \text{ou} \quad 10\%$$

---

## Interpretação Prática

Essa igualdade de probabilidade é o que torna a AAS tão importante na inferência estatística. Como **todas as amostras possíveis são igualmente prováveis**, qualquer **estatística amostral (como a média)** tende a ser **não tendenciosa (não enviesada)** como estimador do parâmetro populacional.

---

## Exemplo Completo

**População:**

$N = 4$  elementos:  $\{A, B, C, D\}$

**Amostras possíveis com  $n = 2$ :**

1.  $\{A, B\}$
2.  $\{A, C\}$
3.  $\{A, D\}$
4.  $\{B, C\}$
5.  $\{B, D\}$
6.  $\{C, D\}$

\$

$\binom{4}{2} = 6 \quad \rightarrow \quad P = \frac{1}{6} \approx 16,67\%$

\$

Se uma dessas for sorteada ao acaso, **por exemplo, {A, C}**, a chance era a mesma de qualquer outra.

---

## Complemento: Estimativa da Média Amostral

Se estamos estimando a **média de uma variável  $XX$**  usando AAS, a **média amostral  $\bar{x}$**  é um **estimador não tendencioso** da média populacional  $\mu$ :

\$

$E(\bar{x}) = \mu$

\$

E a **variância da média amostral** (sem reposição) é:

\$

$\text{Var}(\bar{x}) = \frac{\sigma^2}{n} \cdot \frac{N - n}{N - 1}$

\$

**Onde:**

$\sigma^2$ : variância populacional

$n$ : tamanho da amostra

$N$ : tamanho da população

Esse fator adicional  $\frac{N - n}{N - 1}$  é chamado de **fator de correção para população finita (fpc)**.

Por que isso importa?

- Se você **não usa reposição** e tem uma **população pequena**, esse fator evita **superestimar a variância**.
- Se  $N \rightarrow \infty$ , o fator se aproxima de 1, e a fórmula se torna a da variância da média em amostragem com reposição.

Exemplo em python

---

## Cenário Didático:

Imagine que temos uma turma com 30 alunos, e queremos **selecionar aleatoriamente 5 alunos** para responder a um questionário. Todos os alunos têm **a mesma chance de serem escolhidos**.

---

## Código Completo em Python:

```
import random
import pandas as pd
```

```
# 1. Lista de 30 alunos fictícios
alunos = [f'Aluno_{i+1}' for i in range(30)]

# 2. Transformar em DataFrame para visualização
df_turma = pd.DataFrame({'Nome': alunos})
print("Turma completa:\n")
print(df_turma)

# 3. Amostragem aleatória simples de 5 alunos
amostra = random.sample(alunos, 5)

# 4. Exibir a amostra
print("\nAmostra aleatória simples (5 alunos):\n")
for i, nome in enumerate(amostra, 1):
    print(f"{i}. {nome}")
```

### 🔍 Explicação:

Etapa	O que foi feito?
1	Criamos uma lista com 30 nomes fictícios.
2	Usamos <b>pandas</b> para visualizar como se fosse uma tabela.
3	Usamos <b>random.sample()</b> para selecionar <b>5 elementos únicos sem reposição</b> , simulando a <b>amostragem aleatória simples</b> .
4	Imprimimos o resultado final.

### Características da Amostragem Aleatória Simples

- **Probabilidade igual** para todos os elementos da população.
- **Independente**: a seleção de um aluno **não afeta** a chance de outro ser escolhido.
- **Sem reposição**: um aluno sorteado **não é sorteado novamente**.

### Complemento com Probabilidades

Se quiser deixar mais avançado, dá pra calcular a **probabilidade de um aluno específico ser sorteado**:

$$P(\text{ser sorteado}) = \frac{\text{tamanho da amostra}}{\text{tamanho da população}} = \frac{5}{30} \approx 16,67\%$$

### Introdução a Amostragem Sistemática



A **Amostragem Sistemática** é um **tipo de amostragem probabilística** em que os elementos da amostra são selecionados a partir de **intervalos regulares** de uma lista ordenada da população.

Em vez de escolher os elementos completamente ao acaso (como na Amostragem Aleatória Simples), você escolhe **um ponto de início aleatório** e depois segue uma **regra fixa de espaçamento**.

---

## Exemplo Conceitual

Imagine uma população com 1000 pessoas e você deseja selecionar uma amostra de 100 pessoas. Na Amostragem Sistemática, você:

1. **Ordena** a população (por nome, matrícula, etc.).
2. **Calcula o intervalo de seleção** (também chamado de "salto" ou  $k$ ).
3. **Sorteia aleatoriamente** um número entre 1 e  $k$  como ponto de partida.
4. Seleciona os próximos elementos usando esse intervalo.

---

## Formulação Matemática

### Fórmula do Intervalo

A fórmula para definir o intervalo  $k$  é:

$$k = \left\lfloor \frac{N}{n} \right\rfloor$$

Onde:

- $N$  = tamanho da população
- $n$  = tamanho da amostra
- $k$  = intervalo de seleção (salto)

---

## 4. Ponto de partida:

Escolhe-se aleatoriamente um número inteiro  $r$  entre 1 e  $k$ :

$$r \in \{1, 2, \dots, k\}$$

---

## 5. Elementos selecionados:

A amostra será formada pelos elementos nas posições:

$$r, r + k, r + 2k, r + 3k, \dots, r + (n-1)k$$

\$

---

## Exemplo passo a passo

---



### Situação:

Temos uma **lista com 20 pessoas** numeradas de 1 a 20:

1. Ana
2. Joaquim
3. Carla
4. Daniel
5. Eduardo
6. Fernanda
7. Gabriela
8. Henrique
9. Isabel
10. João
11. Karina
12. Luis
13. Mariana
14. Lais
15. Olivia
16. Paulo
17. Quezia
18. Rodrigo
19. Sabrina
20. Thiago

Queremos selecionar uma **amostra de 5 pessoas** usando **amostragem sistemática**.

---

✅ Passo 1: Identificar o tamanho da população (N)

\$

$N = 20$

\$

---

✅ Passo 2: Definir o tamanho da amostra desejada (n)

\$

$n = 5$

\$

---

✅ Passo 3: Calcular o intervalo de seleção (k)

\$

$$k = \left\lfloor \frac{N}{n} \right\rfloor = \left\lfloor \frac{20}{5} \right\rfloor = 4$$

\$

👉 Vamos escolher **1 pessoa a cada 4 posições**.

✅ Passo 4: Escolher um número aleatório entre 1 e k

Vamos supor que o número sorteado foi:

\$

$\text{Início aleatório} = 3$

\$

✅ Passo 5: Selecionar os elementos da amostra

Começando da posição 3 (Carla), e pulando de 4 em 4:

Ordem na amostra	Índice na lista	Nome
1º	3	Carla
2º	7	Gabriela
3º	11	Karina
4º	15	Olivia
5º	19	Sabrina

Resultado da amostra sistemática:

1. Carla
2. Gabriela
3. Karina
4. Olivia
5. Sabrina

Observações importantes:

- O **intervalo k = 4** foi calculado dividindo o total da população pelo tamanho da amostra.
- O **início aleatório** é crucial para manter o caráter **probabilístico**.
- Esse método **espalha bem os elementos** ao longo da lista.

✅ Características

Característica	Explicação
<b>Probabilística</b>	Sim, desde que a lista seja ordenada aleatoriamente.
<b>Simplicidade operacional</b>	Muito fácil de aplicar, especialmente com grandes populações.
<b>Requer ordenação?</b>	Sim — a população precisa estar organizada em uma sequência.
<b>Risco de viés</b>	Sim — se a lista tiver um <b>padrão cíclico</b> , a amostragem sistemática pode capturar esse padrão e <b>introduzir viés</b> .
<b>Rapidez</b>	Mais rápida do que a aleatória simples, porque não exige sorteio de todos os elementos.

Quando usar?

- Quando a população está **fisicamente ou logicamente ordenada**.
- Quando você quer um método de amostragem **simples e rápido**.
- Quando a população **não tem padrões cíclicos** que possam interferir.

Exemplo Didático com python

#### Cenário:

Você tem uma lista com 20 funcionários e quer selecionar **5** para uma pesquisa.

#### ✅ Passo a passo com Python

```
import pandas as pd
import random

# 1. Criar lista de 20 funcionários
funcionarios = [f'Funcionario_{i+1}' for i in range(20)]
df = pd.DataFrame({'ID': range(1, 21), 'Nome': funcionarios})

# 2. Parâmetros
N = len(df)      # Tamanho da população
n = 5            # Tamanho da amostra desejada
k = N // n       # Intervalo sistemático

print(f"Tamanho da população: {N}")
print(f"Tamanho da amostra: {n}")
print(f"Intervalo k: {k}")

# 3. Escolher ponto de partida aleatório entre 1 e k
ponto_inicial = random.randint(1, k)
print(f"Ponto de partida aleatório: {ponto_inicial}")
```

```
# 4. Selecionar os índices da amostra
indices_amostra = list(range(ponto_inicial - 1, N, k))
amostra_sistemica = df.iloc[indices_amostra]

# 5. Mostrar resultado
print("\nAmostra Sistemática:\n")
print(amostra_sistemica)
```

---

## Interpretação

- Se  $N = 20$  e  $n = 5$ , então  $k = 4$ .
- Suponha que o número aleatório inicial seja 2.
- A amostra será composta pelos elementos nas posições: 2, 6, 10, 14 e 18.

---

## ✓ Vantagens da Amostragem Sistemática

- Simples de aplicar.
- Boa distribuição da amostra ao longo da população.
- Útil quando os dados estão organizados em uma lista (alfabética, por data, etc).

---

## ⚠ Cuidados

- Evitar **padrões cíclicos** nos dados que possam coincidir com o intervalo  $k$ , pois isso pode introduzir **viés**.
- A lista precisa estar **bem organizada e representativa** da população.

A amostragem sistemática é como "contar de forma regular" dentro de uma população ordenada, começando de um ponto aleatório e saltando de forma fixa. É eficiente, fácil de aplicar e útil em pesquisas populacionais ou listas grandes, **mas exige cuidado com a ordenação da população**, para evitar padrões que distorçam os resultados.

---

## Introdução amostragem estratificada

A **amostragem estratificada** é uma técnica de amostragem **probabilística** na qual a população é dividida em **subgrupos homogêneos** chamados de *estratos*. Em seguida, uma amostra é extraída de cada estrato de forma separada.

Segundo **Cochran (1977)**, essa técnica é especialmente útil quando se sabe de antemão que a população pode ser dividida em **subpopulações com características distintas**, pois isso **melhora a precisão** das estimativas estatísticas.

*"Stratified sampling provides more precise estimates of population parameters than simple random sampling when the strata are internally homogeneous."*

— **W.G. Cochran**, *Sampling Techniques*, 3rd ed., 1977.

---

## Objetivo da Amostragem Estratificada

- **Garantir representatividade** de todos os subgrupos relevantes.
- **Reduzir a variância** das estimativas.
- **Melhorar a eficiência** estatística sem necessariamente aumentar o tamanho da amostra.

Como destaca **Silva et al. (2010)**:

*"A estratificação é indicada sempre que for possível dividir a população em grupos internamente homogêneos e externamente heterogêneos."*

— Silva, M. A. F., **Estatística: Fundamentos e Aplicações**, 2010.

---

## Tipos de Alocação

### 1. Alocação Proporcional (Neyman Simples)

Distribui amostras com base no tamanho de cada estrato.

Usada quando os estratos têm variâncias similares.

### 2. Alocação Iguatária

Todos os estratos recebem o mesmo número de elementos, independentemente de seu tamanho.

### 3. Alocação Ótima (Neyman Alocação)

Considera a variância dentro de cada estrato  $S_h^2$ , buscando minimizar o erro amostral:

$$n_h = \frac{N_h S_h}{\sum_{h=1}^L N_h S_h} \cdot n$$

Essa abordagem foi proposta por **Jerzy Neyman** em 1934, sendo ideal quando se sabe que os estratos têm **diferente variabilidade interna**.

*"In optimal allocation, more units are sampled from strata with greater variability to reduce overall sampling error."*

— **J. Neyman**, *On the two different aspects of the representative method*, 1934.

---

## Quando usar Amostragem Estratificada?

- Quando há **conhecimento prévio da população** e sua segmentação lógica.
  - Quando se deseja **controlar o erro amostral**.
  - Quando se precisa **garantir a presença de grupos pequenos** mas importantes na amostra (como minorias).
- 

## Alocação Proporcional

A **alocação proporcional** é um método de distribuição do tamanho da amostra entre os estratos de forma **proporcional ao tamanho de cada estrato na população**.

Esse é o **tipo mais comum** de alocação usado na amostragem estratificada e é **indicado quando a variabilidade dentro de cada estrato é semelhante** (ou seja, os estratos têm **variâncias parecidas**).

---

## Fórmula da Alocação Proporcional

Seja:

- $N$ : total da população
- $N_h$ : número de elementos no estrato  $h$
- $n$ : tamanho total da amostra
- $n_h$ : número de elementos a serem sorteados no estrato  $h$

A **fórmula** da alocação proporcional é:

$$n_h = \frac{N_h}{N} \cdot n$$

Isso garante que o **percentual do estrato na população** será **mantido igual** na amostra.

---

## Exemplo Didático Completo (Passo a Passo)

Imagine a seguinte população:

Estrato (Setor da Empresa)	Número de Funcionários ( $N_h$ )
Administrativo	100
Produção	300
Logística	200
<b>Total <math>N</math></b>	<b>600</b>

Você quer **entrevistar 60 funcionários** para uma pesquisa de clima organizacional.

---

### Passo 1: Aplicar a fórmula para cada estrato

Sabemos que  $N = 600$  e  $n = 60$

#### **a) Administrativo:**

$$n_1 = \frac{100}{600} \cdot 60 = \frac{1}{6} \cdot 60 = 10$$

#### **b) Produção:**

$$n_2 = \frac{300}{600} \cdot 60 = \frac{1}{2} \cdot 60 = 30$$

### c) Logística:

$$n_3 = \frac{200}{600} \cdot 60 = \frac{1}{3} \cdot 60 = 20$$

### ✓ Resultado Final da Amostra

Estrato	$N_h$	$n_h$
Administrativo	100	10
Produção	300	30
Logística	200	20
<b>Total</b>	<b>600</b>	<b>60</b>

## Quando Usar a Alocação Proporcional?

- Quando os **estratos têm tamanhos diferentes**.
- Quando **não há grandes diferenças na variância** entre os estratos.
- Quando você **não tem dados prévios sobre a variabilidade** dos estratos (então usa a proporcional por segurança).

## Citação Acadêmica

"A alocação proporcional é recomendada quando os estratos são homogêneos internamente, mas variam em tamanho. Ela assegura a representatividade proporcional dos estratos na amostra."

— **Cochran, W.G.** (1977). *Sampling Techniques*.

## Resumo Visual

$$n_h = \frac{N_h}{N} \cdot n$$

- $N_h$ : tamanho do estrato
- $N$ : população total
- $n$ : tamanho da amostra
- $n_h$ : amostra para o estrato



## Introdução Alocação Igualitária

A **amostragem estratificada** é uma técnica em que a população é dividida em **estratos homogêneos** (grupos com características similares). Em seguida, uma amostra é retirada de **cada estrato**.

A **alocação igualitária** é uma das formas de definir quantos elementos devem ser amostrados em cada estrato. Neste tipo de alocação:

**Todos os estratos contribuem com o mesmo número de elementos para a amostra, independentemente do tamanho real que têm na população.**

“Na alocação igualitária, o mesmo número de unidades é selecionado de cada estrato. Este método é simples, mas pode ser ineficiente se os estratos forem de tamanhos muito diferentes.”  
— **Cochran, W.G.** (1977). *Sampling Techniques*.

---

### Por que usar Alocação Igualitária?

A alocação igualitária é especialmente útil quando:

- O interesse está em **comparar os grupos entre si** com o **mesmo peso estatístico**, e não necessariamente refletir a distribuição da população.
- Os estratos têm **variâncias muito diferentes**, e o pesquisador quer garantir **representação adequada mesmo de estratos pequenos**.
- Quando há **dificuldade de acesso a informações sobre o tamanho real de cada estrato** (ou seja, você sabe que há grupos, mas não quantos elementos em cada).

---

### Fórmula da Alocação Igualitária

Se temos:

- $H$ : número total de estratos
- $n$ : tamanho total da amostra
- $n_h$ : número de elementos na amostra de cada estrato

A fórmula é:

$$n_h = \frac{n}{H}$$

Ou seja, **divide-se igualmente o tamanho da amostra pelo número de estratos**.

---

### Exemplo Didático Passo a Passo

Imagine a seguinte população de uma universidade:

Curso (Estrato)	Número de Alunos ( $n_h$ )
-----------------	----------------------------

---

Curso (Estrato)	Número de Alunos (\$N_h\$)
Administração	200
Engenharia	500
Psicologia	300
<b>Total</b> \$N\$	<b>1000</b>

Você deseja entrevistar **60 alunos**, aplicando a **alocação igualitária**.

Passo 1: Calcular quantos estratos existem

Temos **\$H = 3\$** (Administração, Engenharia e Psicologia)

Passo 2: Aplicar a fórmula da alocação igualitária

$$n_h = \frac{n}{H} = \frac{60}{3} = 20$$

Ou seja, você irá sortear **20 alunos de cada curso, mesmo que os cursos tenham tamanhos diferentes**.

✅ Resultado da Amostragem

Curso (Estrato)	\$N_h\$	\$n_h\$
Administração	200	20
Engenharia	500	20
Psicologia	300	20
<b>Total</b>	<b>1000</b>	<b>60</b>

⚠ Limitação

Esse tipo de alocação **pode gerar viés** se os estratos forem muito diferentes em tamanho, porque estratos menores terão um peso **maior** na amostra do que na população real.

📊 Comparação com Alocação Proporcional

Critério	Alocação Proporcional	Alocação Igualitária
Considera \$N_h\$	Sim	Não
Representatividade	Alta	Pode ser distorcida

Critério	Alocação Proporcional	Alocação Igualitária
Comparação entre grupos	Desbalanceada se $N_h$ for desigual	Equilibrada por construção
Recomendado para	Inferência populacional	Estudos comparativos entre estratos

## ✓ Exemplo 2

Imagine a seguinte situação: uma escola tem alunos divididos por turno:

Turno (Estrato)	Número de Alunos ( $N_h$ )
Manhã	400
Tarde	300
Noite	100
<b>Total</b> $N$	<b>800</b>

Você quer **entrevistar 30 alunos**, usando **alocação igualitária**.

Passo 1: Contar o número de estratos

Temos:

$H = 3$

Passo 2: Calcular o número de alunos a serem sorteados em cada estrato

$n_h = \frac{n}{H} = \frac{30}{3} = 10$

Resultado

Turno	$N_h$	$n_h$
Manhã	400	10
Tarde	300	10
Noite	100	10
<b>Total</b>	<b>800</b>	<b>30</b>

## Interpretação

- Mesmo o **turno da noite**, que tem apenas 12,5% da população (100 de 800), representa **33,3% da amostra**.
- Isso **distorce a representatividade da população**, mas **permite fazer comparações equilibradas entre os turnos**.
- Exemplo: comparar satisfação média dos turnos com o **mesmo peso estatístico** (cada um contribui igualmente para o resultado).



## Citações Relevantes

"A alocação igualitária pode ser preferida quando o interesse se volta para o estudo dos estratos individualmente, e não para a estimativa de parâmetros populacionais globais."

— Barbetta, P.A. (2010). *Estatística Aplicada às Ciências Sociais*.

"A amostragem com alocação uniforme é recomendada quando se deseja comparação direta entre os grupos, pois garante o mesmo número de observações por estrato, independentemente do seu tamanho."

— Bolfarine, H., Bussab, W.O. (2005). *Elementos de Amostragem*.



## Riscos e Cuidados

- Pode **super-representar grupos pequenos** (dando a eles mais importância do que têm na população).
- Pode **sub-representar grupos grandes**, levando a **erros de inferência se o objetivo for estimar valores populacionais**.



## Cenário:

Temos alunos divididos por **turno** (estratos): Manhã, Tarde e Noite.

Queremos sortear **10 alunos de cada turno**, totalizando 30 alunos na amostra.



## Passo a Passo em Python

```
import pandas as pd
import numpy as np

# Criar população fictícia
np.random.seed(42) # Para reprodutibilidade

# Quantidade de alunos por turno (população)
populacao = {
    'Manhã': 400,
    'Tarde': 300,
    'Noite': 100
}
```

```

# Gerar DataFrame com a população total
dados = []
for turno, quantidade in populacao.items():
    for i in range(quantidade):
        dados.append({'nome': f'Aluno_{turno}_{i+1}', 'turno': turno})

df_populacao = pd.DataFrame(dados)

# Verificar tamanho da população
print("População total por turno:")
print(df_populacao['turno'].value_counts())

# -----
# Amostragem Estratificada com Alocação Igualitária
# -----

n_por_estrato = 10 # Alocação igualitária

# Função para amostrar n alunos de cada estrato
def amostragem_igualitaria(df, coluna_estrato, n):
    return (
        df.groupby(coluna_estrato)
        .apply(lambda x: x.sample(n=n, random_state=42))
        .reset_index(drop=True)
    )

# Gerar amostra
df_amostra = amostragem_igualitaria(df_populacao, 'turno',
n_por_estrato)

print("\nAmostra obtida:")
print(df_amostra.head(10))
print("\nDistribuição na amostra:")
print(df_amostra['turno'].value_counts())

```

Saída esperada:

```

População total por turno:
Manhã    400
Tarde    300
Noite    100
Name: turno, dtype: int64

Amostra obtida:
      nome  turno
0  Aluno_Manhã_103  Manhã
1  Aluno_Manhã_279  Manhã
...

```

```
Distribuição na amostra:
Manhã      10
Tarde      10
Noite      10
Name: turno, dtype: int64
```

---

## ✓ Observações

- A **distribuição da amostra é perfeitamente uniforme**: 10 alunos de cada turno.
- Ideal para **comparar opiniões ou comportamentos por turno**, com o mesmo peso.
- Não ideal para inferência estatística geral, pois **não representa a proporção real da população**.

---

## Alocação Ótima (Neyman)

A **Alocação Ótima de Neyman** é uma técnica usada na **amostragem estratificada** para **minimizar o erro padrão da estimativa** de uma média ou proporção, ao mesmo tempo em que se respeita um **tamanho total fixo de amostra**.

Ela é mais **eficiente** que a alocação proporcional quando os **estratos têm diferentes variabilidades internas** (ou seja, diferentes desvios padrão). Em vez de alocar apenas proporcionalmente ao tamanho do estrato, ela considera também **a variabilidade dentro de cada estrato**.

## Por que é chamada de "ótima"?

Porque, entre todas as maneiras possíveis de distribuir a amostra entre os estratos (como alocação igualitária ou proporcional), a de Neyman é **a que resulta na menor variância** para a estimativa da média ou proporção populacional, **sob um custo total fixo ou tamanho de amostra fixo**.

---

## Objetivo

Minimizar a variância da estimativa da média da população, dada uma amostra total  $N$ , distribuída entre  $L$  estratos.

---

## Fórmula da Alocação de Neyman

Seja:

- $N$ : tamanho total da população
- $N_h$ : tamanho do estrato  $h$
- $S_h$ : desvio padrão da variável de interesse no estrato  $h$
- $n$ : tamanho total da amostra desejada
- $n_h$ : número de elementos da amostra no estrato  $h$  (o que queremos calcular)
- $L$ : número total de estratos

A fórmula para **calcular \$N\_h\$** é:

$$N_h = \frac{N_h S_h}{\sum_{i=1}^L N_i S_i} \cdot n$$

## Interpretação

- Quanto **maior o estrato \$N\_h\$**, **maior** deve ser **\$n\_h\$**
- Quanto **maior a variabilidade \$S\_h\$**, **mais elementos** da amostra devem ser coletados nesse estrato
- A soma dos **\$n\_h\$** é igual ao total da amostra:

$$\sum_{h=1}^L n_h = n$$

## Quando Usar?

- Quando os **estratos têm tamanhos e variabilidades diferentes**
- Quando se deseja obter **maior precisão** nas estimativas
- Quando se tem **acesso ao desvio padrão** (ou estimativa) dos estratos

## Exemplo Conceitual

Imagine que temos:

Estrato (h)	<b>\$N_h\$</b>	<b>\$S_h\$</b>
1	1000	10
2	500	30
3	1500	20

Queremos uma amostra de tamanho total **\$n = 300\$**.

Aplicando a fórmula:

1. Calcular **\$N\_h S\_h\$** para cada estrato:

- **\$1000 \times 10 = 10,000\$**
- **\$500 \times 30 = 15,000\$**
- **\$1500 \times 20 = 30,000\$**

2. Soma dos produtos:

$$\sum N_h S_h = 10,000 + 15,000 + 30,000 = 55,000$$

\$

3. Calcular  $n_h$  para cada estrato:

\$

$$n_1 = \frac{10,000}{55,000} \cdot 300 \approx 54.55$$

\$

\$

$$n_2 = \frac{15,000}{55,000} \cdot 300 \approx 81.82$$

\$

\$

$$n_3 = \frac{30,000}{55,000} \cdot 300 \approx 163.63$$

\$

Assim, a amostra alocada seria aproximadamente:

- **Estrato 1: 55 elementos**
- **Estrato 2: 82 elementos**
- **Estrato 3: 163 elementos**

---

A **alocação ótima de Neyman** direciona mais elementos da amostra para **estratos com maior variabilidade**, garantindo maior **eficiência estatística**.

Segundo Cochran (1977), essa alocação reduz a variância da média estratificada "ao alocar mais observações para estratos com maior contribuição à variância total".

---

## Intuição por trás da fórmula

A alocação ótima considera dois fatores:

1. **Tamanho do estrato ( $N_h$ )**: estratos maiores devem contribuir mais para a amostra.
2. **Variabilidade ( $S_h$ )**: estratos mais heterogêneos (maior desvio padrão) também devem receber mais elementos da amostra, pois a incerteza estatística é maior neles.

Logo, a amostra é alocada mais densamente onde:

- Há **mais indivíduos**, e
- Há **maior variação** nos dados (indicando maior incerteza a ser medida).

---

## Propriedades Estatísticas

A variância da estimativa da média da população sob amostragem estratificada com alocação de Neyman é:

\$

$$\operatorname{Var}(\bar{y}_{str}) = \sum_{h=1}^L \left( \left( \frac{N_h}{N} \right)^2 \cdot \frac{S_h^2}{n_h} \right)$$



$\{n_h\}$  \right)

\$

A alocação de Neyman **minimiza essa variância**, pois distribui  $n_h$  de modo que os termos  $\frac{S_h^2}{n_h}$  fiquem equilibrados com os pesos  $\left( \frac{N_h}{N} \right)^2$ .

## Observações importantes

- É **necessário conhecer (ou estimar)  $SS_h$**  para cada estrato antes da amostragem.
- A técnica assume **custos iguais** para coleta de dados em cada estrato.
  - Se os custos forem diferentes, uma generalização chamada **alocação ótima com custo variável** deve ser usada.
- É muito usada em pesquisas sociais, estudos de opinião, e pesquisas por amostragem em auditorias e estatísticas oficiais.

## Comparando com outras alocações

Tipo de Alocação	Leva em conta $N_h$ ?	Leva em conta $SS_h$ ?	Eficiência
Igualitária	✗	✗	Baixa
Proporcional	✓	✗	Média
Ótima (Neyman)	✓	✓	Alta

## Citação de Autor

"A allocation that minimizes the variance of the stratified mean estimator is called Neyman allocation. It gives more weight to strata with larger sizes and greater variability."

— William G. Cochran, **Sampling Techniques (1977)**

## Exemplo em python

### Exemplo (recapitulando):

Estrato	$N_h$ (Tamanho da População)	$SS_h$ (Desvio Padrão)
1	500	10
2	300	20
3	200	15

Total de população:  $N = 1000$

Tamanho da amostra total:  $n = 60$

## ✓ COMO FAZER NO EXCEL

1. Crie uma planilha com os seguintes cabeçalhos:

A: Estrato  
B: N\_h  
C: S\_h  
D: N\_h \* S\_h  
E: Proporção Neyman  
F: Alocação n\_h (amostra)

2. Insira os dados:

Linha 2: 1 | 500 | 10  
Linha 3: 2 | 300 | 20  
Linha 4: 3 | 200 | 15

3. Na coluna D (D2), calcule  $N_h \cdot S_h$ :

=D2 → =B2\*C2

Arraste até D4.

4. Em alguma célula fora da tabela (ex: D6), some a coluna D:

=SUM(D2:D4) → isso será  $\sum N_h * S_h$

5. Coluna E: Proporção Neyman

E2: =D2/\$D\$6

Arraste até E4.

6. Coluna F: Alocação

F2: =E2\*60

Arraste até F4.

**Pronto!** A coluna F mostra quantos elementos sortear por estrato.

## ✓ COMO FAZER EM PYTHON

```
import pandas as pd

# Dados
estratos = ['Estrato 1', 'Estrato 2', 'Estrato 3']
N_h = [500, 300, 200]
S_h = [10, 20, 15]
n = 60 # Tamanho da amostra total

# DataFrame com os dados
df = pd.DataFrame({
    'Estrato': estratos,
    'N_h': N_h,
    'S_h': S_h
})

# Etapas de cálculo
df['N_h_S_h'] = df['N_h'] * df['S_h']
total_Nh_Sh = df['N_h_S_h'].sum()
df['Proporcao'] = df['N_h_S_h'] / total_Nh_Sh
df['n_h'] = (df['Proporcao'] * n).round().astype(int)

# Resultado
print(df[['Estrato', 'N_h', 'S_h', 'n_h']])
```

🖨 Saída esperada:

	Estrato	N_h	S_h	n_h
0	Estrato 1	500	10	21
1	Estrato 2	300	20	26
2	Estrato 3	200	15	13

## Tipo de amostragem por conglomerado

A **amostragem por conglomerados** consiste em dividir a população em grupos ou "**conglomerados**" – que podem ser geográficos, organizacionais ou baseados em outras características naturais – e, em seguida, selecionar aleatoriamente alguns desses conglomerados para compor a amostra. Diferentemente da amostragem estratificada, em que se deseja que cada estrato esteja representado proporcionalmente, na amostragem por conglomerados a unidade de seleção é o conglomerado e não o indivíduo.

**Exemplo Conceitual:** Em uma pesquisa sobre hábitos de consumo em uma cidade, os bairros (ou blocos residenciais) podem ser considerados conglomerados. Em vez de listar e sortear

indivíduos de toda a cidade, escolhe-se aleatoriamente alguns bairros e, em seguida, todos ou uma amostra dos residentes desses bairros são incluídos na pesquisa.

## Motivações para Utilizá-la

- **Custo e Logística:**

Quando a população é grande e dispersa geograficamente, realizar um censo ou uma amostragem aleatória simples pode ser impraticável. Selecionar conglomerados pode reduzir custos e facilitar a coleta de dados, concentrando os esforços em áreas específicas.

- **Estrutura Natural da População:**

Em muitas situações, a população já se organiza naturalmente em grupos (por exemplo, escolas, bairros, empresas). Essa divisão natural pode ser explorada para facilitar a amostragem.

- **Facilidade na Obtenção do Quadro Amostral:**

Em vez de se ter uma lista de todos os indivíduos, pode ser mais fácil obter uma lista dos conglomerados e, uma vez selecionados, realizar uma amostragem dentro deles.

## Processos de Amostragem por Conglomerados

A amostragem por conglomerados pode ser \*\*realizada em diferentes estágios:

### 1. Conglomerado em Uma Etapa (One-Stage Cluster Sampling)

- **Seleção dos Conglomerados:** São selecionados aleatoriamente alguns conglomerados.
- **Inclusão Total dos Elementos:** Todos os elementos dos conglomerados selecionados são incluídos na amostra.

*Exemplo:* Selecionar 5 escolas (conglomerados) aleatoriamente de uma cidade e entrevistar todos os alunos de cada escola escolhida.

### 2. Conglomerado em Duas Etapas (Two-Stage Cluster Sampling)

- **Primeira Etapa – Seleção dos Conglomerados:** Seleciona-se aleatoriamente alguns conglomerados.
- **Segunda Etapa – Amostragem Dentro dos Conglomerados:** Dentro de cada conglomerado selecionado, realiza-se uma amostragem aleatória (geralmente, amostragem aleatória simples) para selecionar um subconjunto dos elementos.

*Exemplo:* Selecionar 5 bairros aleatoriamente e, dentro de cada bairro, sortear 20 residências para a pesquisa.

---

## Formulação Matemática e Cálculo da Probabilidade

Embora a amostragem por conglomerados seja um método prático, a modelagem matemática pode se tornar mais complexa em razão da estrutura hierárquica dos dados. Aqui estão alguns pontos-chave:

### 1. Probabilidade de Seleção do Conglomerado:

Se existirem  $M$  conglomerados na população e  $m$  forem selecionados de forma aleatória, a probabilidade de um conglomerado específico ser selecionado é:

\$

$$P(\text{conglomerado selecionado}) = \frac{m}{M}$$

\$

### 2. Probabilidade de Seleção de um Elemento Dentro do Conglomerado:

Se, dentro de um conglomerado  $h$ , existem  $N_h$  elementos e uma amostragem de  $n_h$  elementos é realizada (por exemplo, por amostragem aleatória simples), a probabilidade de um determinado elemento ser selecionado dentro desse conglomerado é:

\$

$$P(\text{elemento selecionado no conglomerado } h) = \frac{n_h}{N_h}$$

\$

### 3. Probabilidade Global de Seleção de um Elemento:

Se um elemento pertence a um conglomerado que tem a probabilidade de ser selecionado  $\frac{m}{M}$  e, depois, o elemento tem uma probabilidade  $\frac{n_h}{N_h}$  de ser selecionado dentro desse conglomerado, a probabilidade total de seleção é o produto:

\$

$$P(\text{elemento}) = \frac{m}{M} \times \frac{n_h}{N_h}$$

\$

Essa formulação é crucial para o cálculo de estimadores e de suas variâncias, permitindo a aplicação de técnicas de inferência estatística.

## Exemplo prático

### ✓ 1. Probabilidade de Seleção do Conglomerado (Escola)

Na amostragem por conglomerados, **não sorteamos indivíduos diretamente**, mas **grupos (conglomerados)** que contêm esses indivíduos. Neste caso, **as escolas são os conglomerados**.

A fórmula é:

\$

$$P(\text{escola selecionada}) = \frac{m}{M}$$

\$

- $M$  = número total de conglomerados (escolas): 20
- $m$  = número de conglomerados sorteados: 5

\$

$$P(\text{escola}) = \frac{5}{20} = 0,25$$

\$

#### 📌 Interpretação:

Cada escola tem **25% de chance** de ser escolhida. Isso acontece porque estamos **sorteando 5 entre 20 escolas**, com **igual probabilidade** para todas.

## ✓ 2. Probabilidade de Seleção de um Elemento Dentro do Conglomerado

Após selecionar uma escola, fazemos **uma nova amostragem dentro dessa escola**. Nesse caso, é feita uma **amostragem aleatória simples** (AAS) com os alunos da escola.

A fórmula:

$$P(\text{aluno dentro da escola}) = \frac{n_h}{N_h}$$

Onde:

- $N_h$ : total de elementos no conglomerado  $h$  (número de alunos na escola): 100
- $n_h$ : número de elementos sorteados dentro do conglomerado: 10

$$P(\text{aluno dentro da escola}) = \frac{10}{100} = 0,10$$

### 📌 Interpretação:

Se a escola foi escolhida, **cada aluno tem 10% de chance de ser selecionado**.

---

## ✓ 3. Probabilidade Global de Seleção de um Elemento

Essa é a **probabilidade real de um aluno qualquer da rede ser escolhido**, considerando as duas etapas:

1. A escola dele ser sorteada.
2. Ele ser sorteado **dentro da escola**.

Como as etapas são independentes (primeiro sortearmos a escola, depois o aluno), usamos o produto:

$$P(\text{aluno}) = \frac{m}{M} \times \frac{n_h}{N_h} = 0,25 \times 0,10 = 0,025$$

### 📌 Interpretação:

No total, **qualquer aluno da rede tem 2,5% de chance** de ser selecionado para participar da pesquisa.

---

### 🔍 Importância disso

Esse tipo de cálculo é fundamental para:

- **Planejar amostras representativas.**
- **Avaliar viés de seleção.**
- **Calcular pesos amostrais** para inferência estatística.

---

## Dica Didática

Use esse raciocínio quando:

- Os grupos são bem definidos (turmas, bairros, escolas).
- Não é viável ou eficiente sortear diretamente todos os indivíduos.

---

## Vantagens e Desvantagens

Vantagens:

- **Custos Reduzidos:** Coletar dados de alguns conglomerados é menos oneroso do que coletar de indivíduos dispersos.
- **Facilidade Operacional:** Aproveita a estrutura natural da população.
- **Aplicabilidade em Grandes Populações:** É particularmente útil em pesquisas de grande escala, como censos e pesquisas domiciliares.

Desvantagens:

- **Aumento da Variância:** A amostragem por conglomerados, especialmente em uma etapa, tende a ter uma variância maior do que a amostragem aleatória simples, pois os elementos dentro de um mesmo conglomerado são muitas vezes mais similares entre si.
- **Efeito de Cluster:** Se os conglomerados são muito homogêneos, a variabilidade entre os elementos selecionados pode ser limitada, reduzindo a eficiência das estimativas.
- **Necessidade de Correções:** Para análises estatísticas, é muitas vezes necessário aplicar um **fator de desenho** (design effect) para ajustar os erros padrão e os intervalos de confiança.

---

## Citações Acadêmicas

"Cluster sampling is a cost-effective alternative to simple random sampling when the population is large and geographically widespread. However, the efficiency gained in terms of logistics might be offset by higher variances due to intracluster homogeneity."

— **Cochran, W.G.**, *Sampling Techniques*, 3rd ed. (1977).

"In cluster sampling, the key is to adequately account for the clustering in the analysis phase, often requiring the use of complex survey design techniques to obtain unbiased estimates."

— **Lohr, S.L.**, *Sampling: Design and Analysis*, 2nd ed. (2009).

---

## Resumo Didático Passo a Passo

### 1. Divisão da População:

Identifique os conglomerados (grupos naturais) na população.

### 2. Seleção dos Conglomerados:

Determine quantos conglomerados serão selecionados aleatoriamente da lista total de

conglomerados.

### 3. Amostragem Dentro dos Conglomerados:

Se optar por uma amostragem em duas etapas, realize uma amostragem adicional dentro de cada conglomerado selecionado para escolher os elementos finais.

### 4. Cálculo das Probabilidades:

Utilize as fórmulas apresentadas para determinar a probabilidade global de seleção de cada elemento (útil para ajustes e análise inferencial).

### 5. Avaliação da Variância:

Considere as implicações do efeito de cluster e, se necessário, calcule o design effect para ajustar os erros padrão.

---

## Conglomerado em Uma Etapa (One-Stage Cluster Sampling)

Na **amostragem por conglomerado em uma etapa**, a seleção da amostra ocorre em **duas fases conceituais, mas apenas uma etapa operacional**:

1. A **unidade amostral primária** é o **conglomerado** (por exemplo, escolas, bairros, empresas, turmas, etc.).
2. Após a seleção aleatória de alguns conglomerados, **todos os elementos dentro de cada conglomerado escolhido são incluídos na amostra**, sem uma subamostragem posterior.

Essa técnica é útil quando:

- A população está naturalmente agrupada em unidades;
- O levantamento de todos os elementos dentro dos grupos selecionados é factível;
- Há restrições logísticas e de custos que dificultam o sorteio de elementos individuais espalhados na população.

---

## Exemplo Teórico

Situação:

Uma prefeitura deseja estimar a média de consumo de água por residência em uma cidade composta por 100 bairros. Cada bairro tem, em média, 1.000 residências.

Passo a passo:

1. A prefeitura **define os bairros como os conglomerados**.
2. Ela **sorteia aleatoriamente 10 bairros** dos 100 disponíveis (sem considerar características específicas).
3. Todos os moradores dos **10 bairros sorteados** são incluídos na amostra.
4. Os dados coletados dessas residências servirão para estimar o consumo médio da cidade inteira.



## Formulação Matemática



Seja:

- $N$  o número total de conglomerados na população.
- $n$  o número de conglomerados selecionados aleatoriamente.
- $M_i$  o número de elementos no conglomerado  $i$ .
- $y_{ij}$  o valor observado no  $j$ -ésimo elemento do  $i$ -ésimo conglomerado.
- $\bar{y}_i$  a média dos elementos do conglomerado  $i$ :

$$\bar{y}_i = \frac{1}{M_i} \sum_{j=1}^{M_i} y_{ij}$$

Estimador da média populacional:

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n \bar{y}_i$$

Ou seja, é a média das médias dos conglomerados escolhidos.

Estimador da variância:

Se os conglomerados forem de tamanhos semelhantes, a variância da média amostral pode ser estimada por:

$$\text{Var}(\hat{\mu}) = \frac{S^2_c}{n} \left(1 - \frac{n}{N}\right)$$

Onde:

- $S^2_c$  é a variância entre as médias dos conglomerados:

$$S^2_c = \frac{1}{n-1} \sum_{i=1}^n (\bar{y}_i - \hat{\mu})^2$$

---

## ✓ Vantagens

- **Economia de tempo e custo**, especialmente quando a população está dispersa.
- **Facilidade de execução**, pois é mais prático visitar todos os elementos de um grupo do que sortear indivíduos espalhados.
- **Utiliza agrupamentos naturais** já existentes (bairros, escolas, turmas...).

---

## ✗ Desvantagens

- **Alta variância intra-conglomerado:** Se os elementos dentro dos conglomerados forem muito semelhantes entre si (homogêneos), a variância entre os grupos será maior e a precisão da estimativa será menor.
  - **Menor eficiência estatística** em comparação com a amostragem aleatória simples ou estratificada, a menos que os conglomerados sejam bem diversificados internamente.
- 

Exemplo didático passo a passo:

Uma universidade quer estimar a **média de horas semanais de estudo** dos seus alunos. A universidade tem **6 cursos** (que usaremos como **conglomerados**). Cada curso tem **10 alunos**.

Como não é viável entrevistar todos os alunos, a universidade decide aplicar uma **amostragem por conglomerado em uma etapa**: vai sortear 2 cursos e **entrevistar todos os alunos desses cursos**.

---

✅ Passo 1: Listar os conglomerados (cursos)

Vamos nomear os cursos de A a F:

- Curso A
  - Curso B
  - Curso C
  - Curso D
  - Curso E
  - Curso F
- 

✅ Passo 2: Dados simulados (horas de estudo por aluno)

Curso	Aluno 1	Aluno 2	...	Aluno 10	Média
A	10	11	...	9	10.0
B	7	6	...	8	7.5
C	12	14	...	13	13.0
D	5	6	...	6	5.8
E	9	10	...	10	9.8
F	15	16	...	14	15.0

(Suponha que as médias de cada curso já foram calculadas.)

---

✅ Passo 3: Sorteio dos conglomerados

Vamos supor que sorteamos **Curso B** e **Curso E**.

---

✓ Passo 4: Calcular a média dos conglomerados escolhidos

- Média do Curso B = 7.5 horas
- Média do Curso E = 9.8 horas

---

✓ Passo 5: Estimar a média populacional

Pela fórmula da média na Amostragem por Conglomerado em Uma Etapa:

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n \bar{y}_i$$

Onde:

- $n = 2$  cursos sorteados
- $\bar{y}_1 = 7.5$ ,  $\bar{y}_2 = 9.8$

$$\hat{\mu} = \frac{7.5 + 9.8}{2} = \frac{17.3}{2} = 8.65$$

✓ **Estimativa da média de horas de estudo dos alunos da universidade: 8,65 horas semanais.**

---

✓ Passo 6: Estimar a variância (se quiser confiabilidade)

$$\begin{aligned} S_c^2 &= \frac{1}{n-1} \sum_{i=1}^n (\bar{y}_i - \hat{\mu})^2 \\ &= \frac{1}{1} [(7.5 - 8.65)^2 + (9.8 - 8.65)^2] \\ &= (1.3225 + 1.3225) = 2.645 \end{aligned}$$

$$\begin{aligned} \text{Var}(\hat{\mu}) &= \frac{S_c^2}{n} \left(1 - \frac{n}{N}\right) \\ &= \frac{2.645}{2} \left(1 - \frac{2}{6}\right) \\ &= 1.3225 \cdot \left(\frac{4}{6}\right) = 0.8816 \end{aligned}$$

Assim, o **erro padrão da média** estimada seria:

$$\text{Erro padrão} = \sqrt{\text{Var}(\hat{\mu})} = \sqrt{0.8816} \approx 0.94$$

---

Perfeito! Vamos fazer os **três** itens com calma e clareza. Começaremos com o gráfico didático e, depois, avançamos para o código e a comparação conceitual.


---

## 1. Gráfico Didático: Amostragem por Conglomerados em Uma Etapa

Cenário:

- Temos **20 escolas** (conglomerados).
- Cada escola tem **100 alunos**.
- Seleccionamos **5 escolas** aleatoriamente.
- Em cada escola sorteada, escolhemos **10 alunos**.

Vamos visualizar isso:

 Visualização esperada

- Blocos (escolas) com alunos representados por bolinhas.
- As escolas sorteadas destacadas.
- Alunos escolhidos em vermelho, por exemplo.

△ Por favor, me envie se você quer o gráfico como **imagem gerada agora** ou se deseja o **código em Python** para gerar no seu ambiente.

## 2. Simulação em Python: Amostragem por Conglomerado em Uma Etapa

Aqui está um exemplo completo com pandas:

```
import pandas as pd
import numpy as np

# Parâmetros
total_escolas = 20
alunos_por_escola = 100
escolas_sorteadas = 5
alunos_sorteados_por_escola = 10

# Gerar base de dados
data = []
for escola in range(1, total_escolas + 1):
    for aluno in range(1, alunos_por_escola + 1):
        data.append({'Escola': f'Escola_{escola}', 'Aluno':
f'Aluno_{escola}_{aluno}'})

df = pd.DataFrame(data)

# Sorteando escolas (conglomerados)
escolas_amostradas = np.random.choice(df['Escola'].unique(),
size=escolas_sorteadas, replace=False)

# Sorteando alunos dentro das escolas
```

```

amostra_final =
df[df['Escola'].isin(escolas_amostradas)].groupby('Escola').sample(n=alunos_sorteados_por_escola, random_state=42)

print("Amostra final:\n", amostra_final)

```

Você pode rodar esse código para visualizar os alunos sorteados.

### 3. Comparação com Amostragem Estratificada

Característica	Amostragem Estratificada	Amostragem por Conglomerados (1 etapa)
Divisão da população	Em <b>estratos homogêneos</b>	Em <b>conglomerados (grupos naturais)</b>
Sorteio	Amostragem feita <b>dentro de cada estrato</b>	Sorteia-se <b>conglomerados inteiros</b>
Custo	Mais alto (coleta em vários locais)	Mais baixo (coleta em poucos grupos)
Precisão estatística	Mais alta (menor variância amostral)	Menor (maior variância entre grupos)
Quando usar	Quando os <b>estratos são internamente homogêneos</b>	Quando <b>é difícil ou caro acessar todos os elementos</b>

## Conclusão

A **amostragem por conglomerados** é uma ferramenta valiosa para situações onde a população é extensa e a coleta de dados deve ser prática e econômica. Apesar de poder introduzir uma variância adicional devido à similaridade dos elementos dentro dos conglomerados, seu uso é amplamente justificado por questões logísticas e de custo, desde que os pesquisadores estejam atentos à necessidade de ajustar as análises para o desenho amostral.

Essa abordagem, quando bem aplicada, permite a realização de estudos significativos em grandes populações, mantendo a representatividade e viabilidade operacional da pesquisa.

Claro! A seguir, apresento uma explicação formal e detalhada sobre **Amostragem por Conglomerado em Duas Etapas (Two-Stage Cluster Sampling)**, como solicitado — sem exemplos computacionais.

## Amostragem por Conglomerado em Duas Etapas

### Definição Geral


A **Amostragem por Conglomerado em Duas Etapas** é um método probabilístico de seleção amostral que combina dois níveis de sorteio:

1. **Primeira etapa:** seleção de conglomerados (grupos naturais da população, como escolas, bairros, hospitais etc.).
2. **Segunda etapa:** seleção de elementos individuais **dentro dos conglomerados sorteados**.

Esse método é amplamente utilizado em estudos populacionais e pesquisas por amostragem quando é logisticamente difícil ou caro construir uma lista completa de todos os elementos da população.

### Diferença-chave em relação à amostragem por conglomerado em uma etapa

- **Uma etapa:** após a seleção dos conglomerados, **todos os elementos** dentro dos conglomerados sorteados são incluídos na amostra.
- **Duas etapas:** após a seleção dos conglomerados, **um subconjunto de elementos** é amostrado dentro de cada conglomerado.

 Como define Cochran (1977), "a amostragem em duas etapas permite maior flexibilidade e economia, uma vez que reduz o esforço de coleta mantendo boa representatividade".

### Formulação Matemática

Seja:

- $M$ : número total de conglomerados na população.
- $m$ : número de conglomerados selecionados na primeira etapa.
- $N_h$ : número de elementos no conglomerado  $h$ .
- $n_h$ : número de elementos amostrados no conglomerado  $h$ .
- $y_{hi}$ : valor da variável de interesse para o  $i$ -ésimo elemento do conglomerado  $h$ .

#### ◆ Probabilidade de Seleção de um Conglomerado

\$

$$P(C_h) = \frac{m}{M}$$

\$

#### ◆ Probabilidade de Seleção de um Elemento dentro do Conglomerado $h$

\$

$$P(E_{hi} \mid C_h) = \frac{n_h}{N_h}$$

\$

#### ◆ Probabilidade Total de Seleção de um Elemento da População

\$

$$P(E_{hi}) = P(C_h) \times P(E_{hi} \mid C_h) = \frac{m}{M} \cdot \frac{n_h}{N_h}$$

\$

Essa fórmula permite o cálculo do peso amostral  $w_{hi}$  de cada elemento, usado posteriormente para estimativas estatísticas:

\$

$$w_{\{hi\}} = \frac{1}{P(E_{\{hi\}})} = \frac{M}{m} \cdot \frac{N_h}{n_h}$$

\$

---

## Estimativa do Total Populacional

O estimador do total populacional  $\hat{T}$  pode ser definido por:

\$

$$\hat{T} = \sum_{h \in S} \frac{M}{m} \cdot \frac{N_h}{n_h} \cdot \sum_{i \in s_h} y_{\{hi\}}$$

\$

Onde:

- $S$ : conjunto de conglomerados sorteados.
- $s_h$ : amostra de elementos dentro do conglomerado  $h$ .

---

## Vantagens

- Reduz custos de coleta de dados em grandes populações dispersas.
- Possibilita amostragem mesmo quando não se conhece a lista completa de todos os elementos da população.
- Flexível para ajustar o tamanho da amostra de acordo com a variação esperada.

---

## Desvantagens

- A variância da estimativa pode ser maior que em métodos como a amostragem estratificada ou aleatória simples.
- Exige cuidado na segunda etapa para garantir aleatoriedade dentro dos conglomerados.

---

## Exemplo aplicado: Amostragem por Conglomerado em Duas Etapas no IBGE

### Contexto do IBGE

O **Instituto Brasileiro de Geografia e Estatística (IBGE)** trabalha com milhões de domicílios espalhados pelo país. Seria inviável listar todos e aplicar uma amostragem aleatória simples. Assim, utiliza a amostragem por conglomerados em duas etapas.

---

### Primeira Etapa: Seleção de Setores Censitários

A **unidade primária de amostragem** (UPA) do IBGE são os **setores censitários**, que são pequenas áreas geográficas homogêneas com aproximadamente 300 domicílios urbanos ou 100 domicílios rurais.

- A seleção dos setores é feita **proporcional ao tamanho (PPT)**, ou seja, setores com mais domicílios têm maior chance de serem escolhidos.

\$

$$P(SC_i) = \frac{\text{tamanho do setor } i}{\text{soma dos tamanhos de todos os setores}}$$

\$

## Segunda Etapa: Seleção de Domicílios

Dentro dos setores selecionados, o IBGE sorteia um número fixo de domicílios para serem entrevistados — por exemplo, 10 ou 12 domicílios.

\$

$$P(D_{ij} \mid SC_i) = \frac{n_i}{N_i}$$

\$

- Onde:
  - $n_i$ : número de domicílios amostrados no setor  $i$ .
  - $N_i$ : número total de domicílios no setor  $i$ .

## Probabilidade Total de Seleção de um Domicílio

A probabilidade total de um domicílio ser selecionado é:

\$

$$P(D_{ij}) = P(SC_i) \cdot P(D_{ij} \mid SC_i)$$

\$

E o peso amostral:

\$

$$w_{ij} = \frac{1}{P(D_{ij})}$$

\$

## Estimativas com os Pesos

Os pesos amostrais são utilizados para ajustar as estimativas, tornando-as representativas da população total, compensando os diferentes tamanhos dos setores e o número de domicílios sorteados.

Por exemplo, o número total estimado de pessoas com determinada característica seria:

\$

$$\hat{T} = \sum_i \sum_{j \in s_i} w_{ij} \cdot y_{ij}$$

\$

## Resumo das Vantagens na Prática



- **Custos reduzidos:** evita a listagem nacional de domicílios.
- **Eficiência logística:** os entrevistadores atuam em áreas específicas.
- **Representatividade:** mantida através dos pesos amostrais.

```
import pandas as pd
import numpy as np

# Parâmetros da população
np.random.seed(42)
total_sectors = 6
households_per_sector = 6

# Gerar população completa
population = []
for sector_id in range(1, total_sectors + 1):
    for household_id in range(1, households_per_sector + 1):
        population.append({
            "sector": sector_id,
            "household_id": household_id,
            "income": np.random.randint(1000, 5000) # renda fictícia
        })

df_population = pd.DataFrame(population)

# 1ª Etapa: selecionar setores (conglomerados)
selected_sectors = np.random.choice(df_population["sector"].unique(),
size=2, replace=False)
df_stage1 =
df_population[df_population["sector"].isin(selected_sectors)]

# 2ª Etapa: selecionar domicílios dentro dos setores escolhidos
sample = df_stage1.groupby("sector").sample(n=3, random_state=42)

# Mostrar amostra selecionada
sample.reset_index(drop=True)
```

Aqui está uma tabela comparativa clara entre os principais **tipos de amostragem probabilística**, destacando a **Amostragem por Conglomerados** em relação às outras:

Critério	Aleatória Simples	Sistemática	Estratificada	Conglomerado
<b>Unidade básica de amostragem</b>	Indivíduo	Indivíduo	Indivíduo dentro de um estrato	<b>Grupo de indivíduos (conglomerado)</b>

Critério	Aleatória Simples	Sistemática	Estratificada	Conglomerado
Como seleciona a amostra	Sorteia indivíduos da população total	Escolhe 1 ponto inicial e salta em k	Divide a população em grupos homogêneos (estratos) e sorteia em cada grupo	Sorteia grupos inteiros e usa todos ou parte dos elementos desses grupos
Objetivo	Amostra representativa aleatória	Simplicidade e praticidade	Garantir representação proporcional de subgrupos	Reduzir custos logísticos em populações grandes e dispersas
Necessidade de lista completa?	Sim	Sim	Sim (com identificação de estratos)	Não necessariamente (basta listar os conglomerados)
Custo logístico	Alto (contato direto com todos)	Moderado	Alto (precisa identificar e dividir estratos)	Baixo (trabalha com grupos geográficos, escolas, etc.)
Precisão estatística	Alta	Moderada	Alta (quando estratos são bem definidos)	Menor precisão se os conglomerados forem internamente homogêneos
Exemplo típico	Sorteio de 100 pessoas de um cadastro	Seleção de cada 10º cliente	Sorteio de alunos por série escolar	Sorteio de escolas inteiras e depois alunos dentro delas (ou todos os alunos)

#### Resumo:

- A **amostragem por conglomerado** é muito útil quando é difícil ou caro acessar toda a população.
- Ela **reduz custos**, mas pode **sacrificar precisão** estatística se os grupos sorteados forem muito semelhantes internamente.
- Em contraste, métodos como a **estratificada** aumentam a precisão, mas exigem mais informações e estrutura.

#### Referências Acadêmicas

- Cochran, W. G. (1977). *Sampling Techniques*. 3rd edition. John Wiley & Sons.
- Lohr, S. L. (2010). *Sampling: Design and Analysis*. 2nd edition. Brooks/Cole.
- Thompson, S. K. (2012). *Sampling*. Wiley Series in Probability and Statistics.

## Amostragem Não Probabilística

A amostragem **não probabilística** é um método de seleção em que **nem todos os elementos da população têm chance conhecida e igual de serem incluídos na amostra**. A escolha dos elementos geralmente depende de critérios subjetivos, conveniência ou julgamento do pesquisador.

Segundo Babbie (2010), "na amostragem não probabilística, os casos não são selecionados aleatoriamente; em vez disso, a escolha dos elementos reflete decisões tomadas pelo pesquisador com base em sua própria avaliação".



#### Referência:

Babbie, E. (2010). *The Practice of Social Research*. Wadsworth.



## Tipos de Amostragem Não Probabilística

### 1. Amostragem por Conveniência

Seleção de elementos que estão mais facilmente disponíveis para o pesquisador.

- **Exemplo:** Entrevistar os primeiros 50 estudantes que saem de uma sala.
- **Vantagem:** Rápida e de baixo custo.
- **Limitação:** Alto risco de viés, pois a amostra pode não representar adequadamente a população.



Segundo Malhotra (2006), é "um método útil em estágios exploratórios, mas deve ser usado com cautela, pois não garante representatividade".

### 2. Amostragem por Julgamento (ou Intencional)

O pesquisador seleciona os elementos que, em sua opinião, são mais representativos.

- **Exemplo:** Escolher especialistas para opinar sobre uma tecnologia.
- **Vantagem:** Útil em estudos qualitativos ou pilotos.
- **Limitação:** Subjetividade elevada; depende da expertise do pesquisador.



Kotler e Keller (2012) destacam que essa abordagem é útil quando os participantes devem ter características específicas para fornecer informações relevantes.

### 3. Amostragem por Quotas

A população é dividida em subgrupos (como sexo, idade, escolaridade), e são selecionadas cotas de cada grupo com base em proporções conhecidas.

- **Exemplo:** Selecionar 40% mulheres e 60% homens, conforme distribuição da população.
- **Vantagem:** Tenta garantir proporcionalidade.
- **Limitação:** Seleção dentro dos subgrupos ainda pode ser enviesada (não aleatória).




Segundo Malhotra (2006), "a amostragem por cotas busca imitar a estrutura da população, mas sua eficácia depende da qualidade dos dados demográficos e do controle rigoroso da aplicação".

## 4. Amostragem Bola de Neve (Snowball)

Usada quando os elementos da população são difíceis de identificar. Um participante indica outro, formando uma "cadeia".

- **Exemplo:** Pesquisas com usuários de drogas, populações marginalizadas ou grupos profissionais raros.
- **Vantagem:** Permite acessar populações ocultas ou difíceis de rastrear.
- **Limitação:** Pode gerar amostras altamente correlacionadas, com baixo grau de diversidade.

 Biernacki e Waldorf (1981) observaram que "esse método é particularmente útil para estudar redes sociais ou populações ocultas".

---

## Considerações Críticas

Tipo	Vantagem Principal	Limitação Principal
Conveniência	Simple e econômica	Alto viés e baixa representatividade
Julgamento	Foco em casos representativos	Subjetividade do pesquisador
Quotas	Proporcionalidade controlada	Seleção dentro da cota é não aleatória
Bola de Neve	Útil para populações difíceis de acesso	Amostra tende a ser homogênea

---

## Referências Bibliográficas

- Babbie, E. (2010). *The Practice of Social Research*. Wadsworth.
- Malhotra, N. K. (2006). *Pesquisa de Marketing: Uma Orientação Aplicada*. Bookman.
- Kotler, P., & Keller, K. L. (2012). *Administração de Marketing*. Pearson.
- Biernacki, P., & Waldorf, D. (1981). *Snowball Sampling: Problems and Techniques of Chain Referral Sampling*. Sociological Methods & Research.