Amostragem e Tipos de Amostras

Compreendendo a Importância da Amostragem e os Diferentes Métodos de Seleção de Amostras

Márcio Nicolau

2025-09-17

Table of contents

List of Figures

Introdução e Objetivos

Nas aulas anteriores, exploramos a estatística descritiva e a análise exploratória de dados, aprendendo a resumir e visualizar informações de um conjunto de dados. No entanto, muitas vezes, é inviável ou impossível coletar dados de uma população inteira (por exemplo, todos os cidadãos de um país, todos os clientes de uma empresa global, ou todas as transações financeiras). É aqui que entra a **amostragem**.

A amostragem é o processo de selecionar um subconjunto de indivíduos ou itens de uma população maior para estimar características dessa população. Um bom plano de amostragem é essencial para garantir que as conclusões tiradas da amostra sejam válidas e representativas da população, evitando vieses e economizando tempo e recursos.

Nesta aula, desvendaremos a importância da amostragem, definiremos seus termos chave e exploraremos os diversos métodos de seleção de amostras, classificando-os em probabilísticos e não probabilísticos.

Objetivos de Aprendizagem

Ao final desta aula, você será capaz de:

- Distinguir entre população, amostra, censo, parâmetro e estatística.
- Compreender a necessidade e os benefícios da amostragem.
- Diferenciar entre métodos de amostragem probabilística e não probabilística.
- Identificar e aplicar os principais tipos de amostragem probabilística (aleatória simples, estratificada, sistemática, por conglomerados).
- Identificar os principais tipos de amostragem não probabilística (por conveniência, por cotas, intencional).
- Reconhecer as vantagens e desvantagens de cada método de amostragem.
- Utilizar Python e R para implementar alguns métodos de amostragem probabilística.

Conceitos Fundamentais em Amostragem

Para compreender a amostragem, é fundamental estabelecer uma terminologia clara.

População (Universo)

A **população** (ou universo) é o conjunto completo de todos os elementos (indivíduos, objetos, eventos, etc.) que possuem uma ou mais características em comum e sobre os quais se deseja obter informação. É o grupo total de interesse para o estudo.

Exemplos:

- Todos os eleitores aptos em um país.
- Todas as lâmpadas produzidas por uma fábrica em um mês.
- Todos os clientes de uma operadora de telefonia.

Amostra

A **amostra** é um subconjunto (parte) da população que é selecionado para ser estudado. O objetivo é que a amostra seja representativa da população para que as conclusões obtidas a partir dela possam ser generalizadas para toda a população.

Exemplos:

- 1.000 eleitores entrevistados de um país.
- 100 lâmpadas testadas de um lote de produção.
- Um grupo de 500 clientes contatados para uma pesquisa de satisfação.

Censo

Um **censo** ocorre quando dados são coletados de *todos* os elementos da população. É um levantamento completo.

Características:

- Oferece informações completas sobre a população.
- Geralmente é caro, demorado e, em muitos casos, impraticável ou impossível.
- Pode ser destrutivo (por exemplo, testar a resistência de todos os parafusos de um lote).

Parâmetro vs. Estatística

- **Parâmetro:** É uma medida numérica que descreve uma característica da **população**. Geralmente é desconhecido e é o que desejamos estimar. (Ex: Média populacional μ, Desvio padrão populacional σ, Proporção populacional P).
- **Estatística:** É uma medida numérica que descreve uma característica da **amostra**. É calculada a partir dos dados da amostra e usada para estimar o parâmetro populacional. (Ex: Média amostral $\bar{x}$, Desvio padrão amostral s, Proporção amostral $\hat{p}$).

Por Que Amostrar?

A amostragem é preferível a um censo na maioria das situações devido a:

- **Custo:** Coletar dados de uma amostra é geralmente muito mais barato.
- **Tempo:** A coleta e análise de dados de uma amostra são mais rápidas.
- **Praticidade:** Muitas populações são muito grandes ou inacessíveis.
- **Precisão:** Em alguns casos, uma amostra bem planejada pode, paradoxalmente, fornecer resultados mais precisos que um censo mal executado, devido à menor complexidade e maior controle sobre a coleta de dados.
- **Natureza Destrutiva:** Se o processo de coleta destrói o item (e.g., teste de vida útil de produtos), a amostragem é a única opção.

Diagrama Conceitual de Amostragem

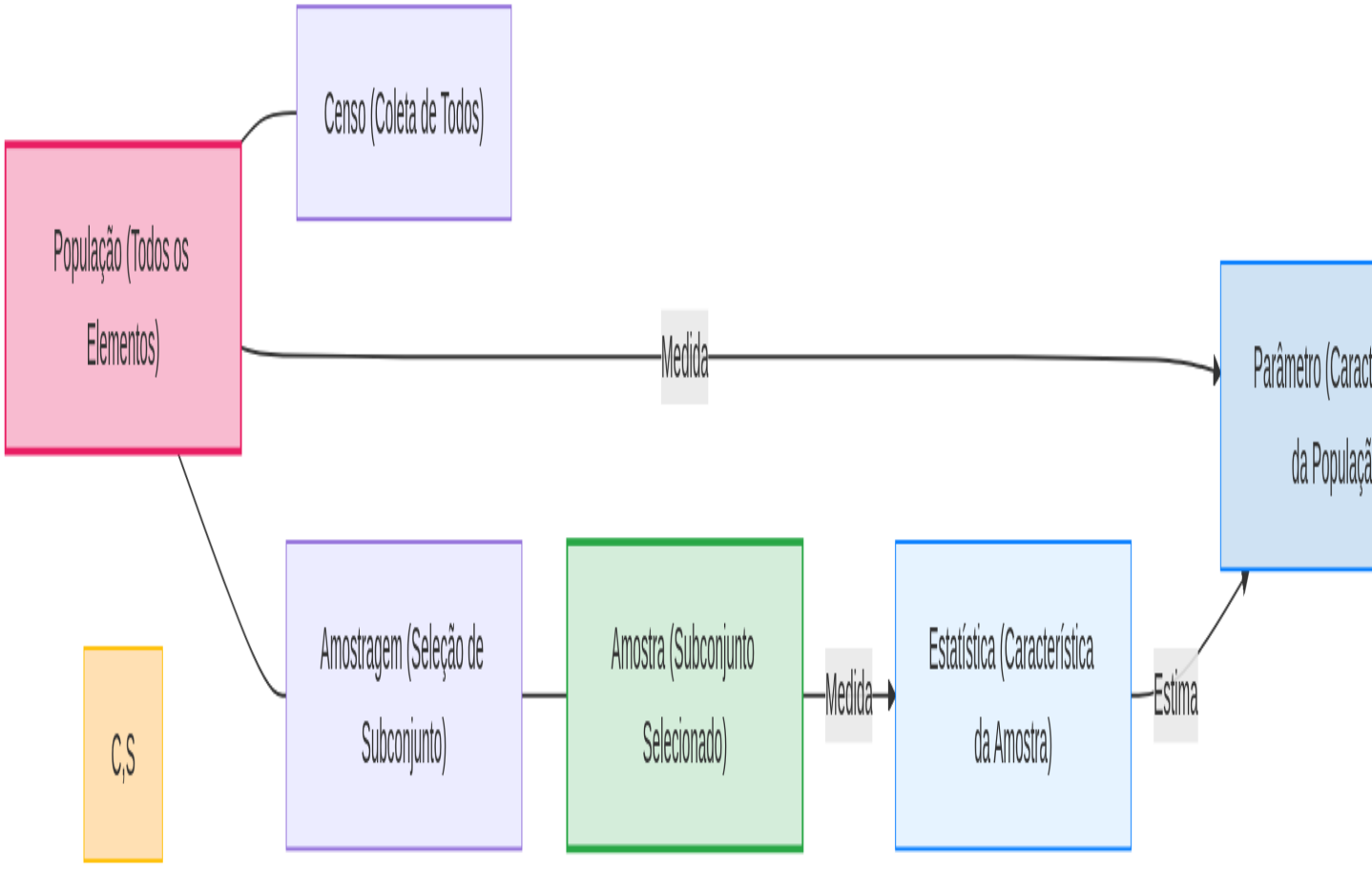

Figure 1: Conceitos Fundamentais em Amostragem

Métodos de Amostragem Probabilística

A **amostragem probabilística** (ou aleatória) é aquela em que todos os elementos da população têm uma chance conhecida (e diferente de zero) de serem selecionados para a amostra. Isso permite calcular a probabilidade de erro amostral e, portanto, generalizar os resultados para a população com um determinado nível de confiança. É a base da inferência estatística (Bussab; Morettin, 2017).

Amostragem Aleatória Simples (AAS) - Simple Random Sampling (SRS)

É o método mais básico. Cada elemento da população tem a mesma probabilidade de ser selecionado.

Como funciona:

1. Obter uma lista completa de todos os elementos da população (frame amostral).
2. Atribuir um número único a cada elemento.
3. Utilizar um gerador de números aleatórios para selecionar os elementos.

Vantagens:

- Simples de entender e implementar.
- Livre de vieses do pesquisador.
- Base para outros métodos.

Desvantagens:

- Exige uma lista completa da população (nem sempre disponível).
- Pode ser impraticável para populações muito grandes.
- Pode não garantir a representatividade de subgrupos importantes (chance de selecionar apenas um tipo de elemento por acaso).

Exemplo: Amostragem Aleatória Simples

Python

```python
import pandas as pd
import numpy as np

# População de exemplo: 1000 IDs
populacao_ids = list(range(1, 1001))
print(f"Tamanho da população: {len(populacao_ids)}")

# Definir o tamanho da amostra
tamanho_amostra = 50

# Amostragem Aleatória Simples
# np.random.choice é excelente para isso, com replace=False para não repetir
amostra_aas = np.random.choice(populacao_ids, size=tamanho_amostra, replace=False)

print(f"\nAmostra Aleatória Simples (primeiros 10 IDs): {amostra_aas[:10]}...")
print(f"Tamanho da amostra AAS: {len(amostra_aas)}")
```

R

```r
# População de exemplo: 1000 IDs
populacao_ids <- 1:1000
cat(sprintf("Tamanho da população: %d\n", length(populacao_ids)))

# Definir o tamanho da amostra
tamanho_amostra <- 50
```

```r
# Amostragem Aleatória Simples
# sample() é a função básica para amostragem em R, com replace=FALSE por padrão
amostra_aas <- sample(populacao_ids, size = tamanho_amostra, replace = FALSE)

cat(sprintf("\nAmostra Aleatória Simples (primeiros 10 IDs): %s...\n", paste(head(amostra_aas, 10), col
cat(sprintf("Tamanho da amostra AAS: %d\n", length(amostra_aas)))
```

Amostragem Estratificada

A população é dividida em subgrupos (estratos) homogêneos e não sobrepostos. Em seguida, uma amostra aleatória simples é retirada de *cada* estrato. Isso garante que subgrupos importantes sejam representados na amostra na proporção correta.

Quando usar: Quando a população é heterogênea, mas pode ser dividida em estratos homogêneos (ex: renda, sexo, região geográfica).

Vantagens:

- Garante representatividade de subgrupos.
- Reduz o erro amostral se os estratos forem bem definidos.
- Permite análises separadas para cada estrato.

Desvantagens:

- Exige conhecimento prévio da população para definir os estratos.
- O frame amostral precisa conter informações sobre o estrato de cada elemento.

Exemplo: Amostragem Estratificada

Python

```python
import pandas as pd
import numpy as np

# População de exemplo com "gênero" como estrato
# Criando um DataFrame simulado
np.random.seed(42)
populacao_df = pd.DataFrame({
    'ID': range(1, 1001),
    'Genero': np.random.choice(['Masculino', 'Feminino', 'Outro'], size=1000, p=[0.48, 0.50, 0.02]),
    'Idade': np.random.randint(18, 65, size=1000)
})
print(f"Distribuição de gênero na população:\n{populacao_df['Genero'].value_counts(normalize=True)}")

# Definir o tamanho total da amostra
tamanho_amostra_total = 100
```

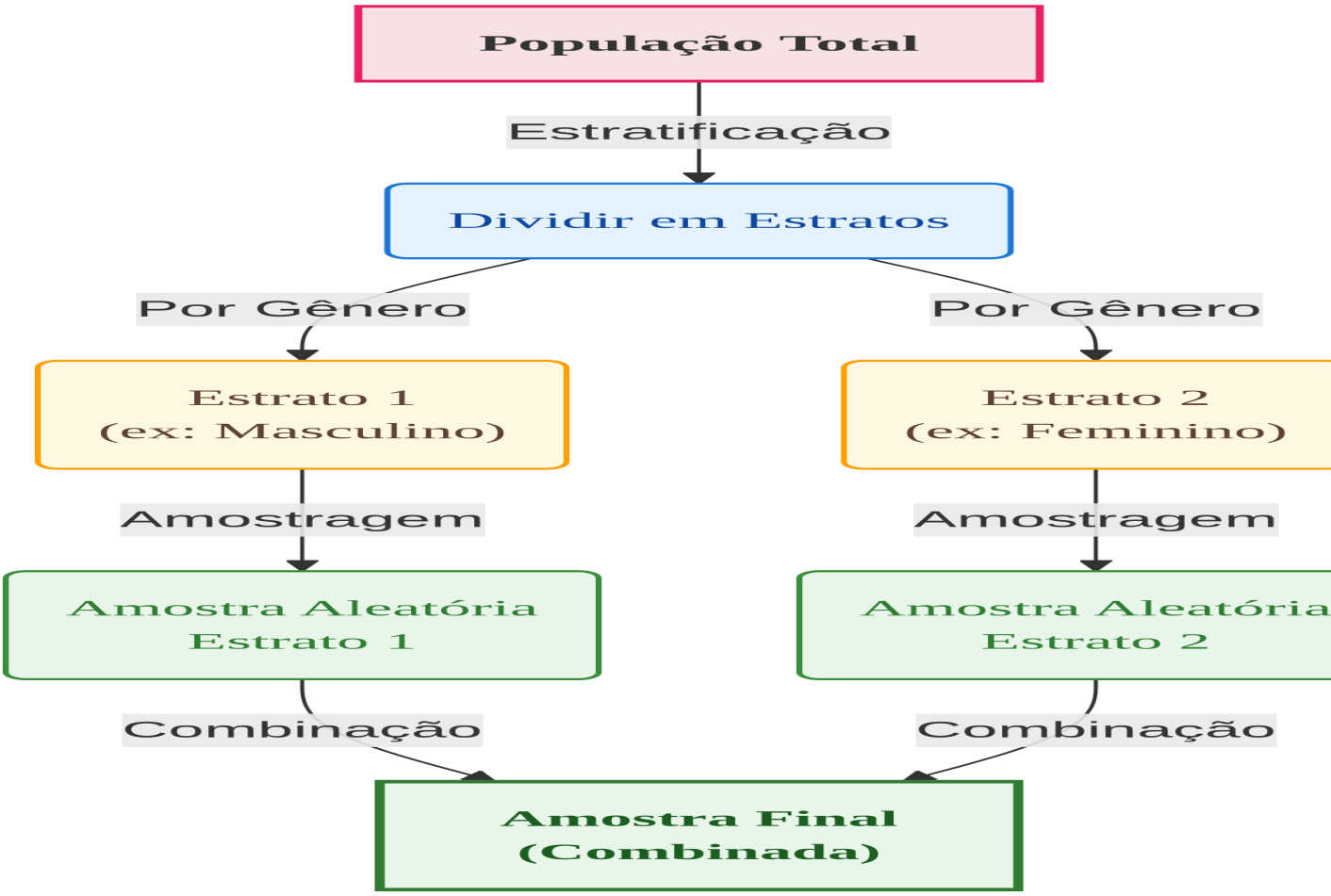

Figure 2: Diagrama de Amostragem Estratificada

```python
# Calcular as proporções dos estratos na população
proporcoes_estratos = populacao_df['Genero'].value_counts(normalize=True)

# Calcular o tamanho da amostra para cada estrato
tamanhos_amostra_estratos = (proporcoes_estratos * tamanho_amostra_total).round().astype(int)
print(f"\nTamanhos de amostra por estrato:\n{tamanhos_amostra_estratos}")

amostra_estratificada = pd.DataFrame()
for genero, tamanho in tamanhos_amostra_estratos.items():
    # Selecionar aleatoriamente dentro de cada estrato
    sub_amostra = populacao_df[populacao_df['Genero'] == genero].sample(n=tamanho, random_state=42)
    amostra_estratificada = pd.concat([amostra_estratificada, sub_amostra])

print(f"\nAmostra Estratificada (primeiras 5 linhas):\n{amostra_estratificada.head()}")
print(f"\nTamanho da amostra estratificada: {len(amostra_estratificada)}")
print(f"Distribuição de gênero na amostra estratificada:\n{amostra_estratificada['Genero'].value_counts
```

R

```r
library(dplyr)

# População de exemplo com "gênero" como estrato
set.seed(42)
populacao_df <- data.frame(
  ID = 1:1000,
  Genero = sample(c('Masculino', 'Feminino', 'Outro'), size = 1000, replace = TRUE, prob = c(0.48, 0.50
  Idade = sample(18:65, size = 1000, replace = TRUE)
)
cat("Distribuição de gênero na população:\n")
print(prop.table(table(populacao_df$Genero)))

# Definir o tamanho total da amostra
tamanho_amostra_total <- 100

# Calcular as proporções dos estratos na população
proporcoes_estratos <- prop.table(table(populacao_df$Genero))

# Calcular o tamanho da amostra para cada estrato
tamanhos_amostra_estratos <- round(proporcoes_estratos * tamanho_amostra_total)
cat("\nTamanhos de amostra por estrato:\n")
print(tamanhos_amostra_estratos)

amostra_estratificada <- populacao_df %>%
  group_by(Genero) %>%
```

```
  sample_n(size = as.numeric(tamanhos_amostra_estratos[cur_group_id()]), replace = FALSE) %>%
  ungroup()

cat("\nAmostra Estratificada (primeiras 5 linhas):\n")
print(head(amostra_estratificada, 5))
cat(sprintf("\nTamanho da amostra estratificada: %d\n", nrow(amostra_estratificada)))
cat("Distribuição de gênero na amostra estratificada:\n")
print(prop.table(table(amostra_estratificada$Genero)))
```

Amostragem Sistemática

Seleciona elementos de uma lista ordenada em intervalos regulares.

Como funciona:

1. Obter uma lista ordenada da população.
2. Calcular o intervalo de amostragem $k = N/n$ (população/amostra).
3. Selecionar um ponto de partida aleatório entre 1 e k.
4. Selecionar cada k-ésimo elemento a partir desse ponto.

Vantagens:

- Mais fácil de implementar que a AAS para grandes populações.
- Pode ser mais representativa que a AAS se a lista tiver um padrão subjacente que não seja detectado pela aleatoriedade.

Desvantagens:

- Se houver um padrão oculto na lista que se alinha com o intervalo k, pode introduzir viés.
- Exige uma lista ordenada da população.

Amostragem por Conglomerados (Cluster Sampling)

A população é dividida em grupos (conglomerados) que são heterogêneos internamente, mas homogêneos entre si (miniaturas da população). Em vez de amostrar indivíduos de cada grupo, amostra-se *grupos inteiros*.

Como funciona:

1. Dividir a população em conglomerados (ex: bairros, escolas, cidades).
2. Selecionar aleatoriamente alguns conglomerados.
3. Coletar dados de *todos* os elementos dentro dos conglomerados selecionados.
    - (Pode haver estágios múltiplos, por exemplo, selecionar cidades e depois bairros dentro das cidades).

Quando usar: Quando a população está geograficamente dispersa ou quando é mais eficiente amostrar grupos do que indivíduos isolados.

Vantagens:

- Economia de custos e tempo, especialmente para populações dispersas.

- Não requer uma lista completa de todos os indivíduos, apenas dos conglomerados.

Desvantagens:

- Pode levar a um erro amostral maior do que AAS ou estratificada se os conglomerados não forem realmente representativos da população.
- Os resultados podem ser menos precisos.

Métodos de Amostragem Não Probabilística

A **amostragem não probabilística** é aquela em que a seleção dos elementos da amostra não envolve aleatoriedade, e a probabilidade de cada elemento ser selecionado é desconhecida. Isso significa que não se pode quantificar o erro amostral, e as conclusões não podem ser generalizadas estatisticamente para a população.

Quando usar: Para estudos exploratórios, pesquisas qualitativas, quando a amostragem probabilística é inviável, ou para obter insights rápidos e de baixo custo.

Amostragem por Conveniência

Seleciona os elementos mais acessíveis ou fáceis de alcançar.

Exemplo: Entrevistar pessoas que passam na rua ou alunos de uma sala de aula específica.

Vantagens:

- Muito rápida e barata.

Desvantagens:

- Alto risco de viés, pois a amostra pode não ser representativa.
- Resultados não generalizáveis.

Amostragem por Quotas

Semelhante à estratificada, mas a seleção dentro de cada "cota" não é aleatória. A população é dividida em subgrupos com base em características específicas (cotas), e o pesquisador preenche essas cotas com elementos acessíveis.

Exemplo: Entrevistar 50 homens e 50 mulheres, escolhendo as primeiras 50 pessoas de cada gênero que encontrar.

Vantagens:

- Garante representatividade de certos subgrupos (nas proporções definidas).
- Mais rápida e barata que a amostragem estratificada.

Desvantagens:

- Viés de seleção dentro das cotas.
- Não é probabilística, portanto, não generalizável estatisticamente.

Considerações Finais sobre Amostragem

A escolha do método de amostragem depende de diversos fatores, incluindo:

- **Objetivo da Pesquisa:** Se o objetivo é generalizar para a população, métodos probabilísticos são indispensáveis.
- **Recursos (Tempo, Custo):** Restrições de recursos podem levar à escolha de métodos não probabilísticos.
- **Conhecimento da População:** A disponibilidade de um frame amostral e informações sobre estratos.
- **Precisão Requerida:** O nível de precisão desejado para as estimativas.

É crucial sempre declarar o método de amostragem utilizado e suas implicações para a interpretabilidade e generalização dos resultados.

Verificação de Aprendizagem

Resolva os problemas abaixo, aplicando os conceitos de amostragem.

1. **Problema 1 (Terminologia):**

   Considere um estudo sobre a altura média dos estudantes de uma universidade.

   a) Defina a **população** para este estudo.
   b) Defina uma possível **amostra**.
   c) O que seria um **parâmetro** neste contexto?
   d) O que seria uma **estatística** neste contexto?

2. **Problema 2 (Escolha do Método de Amostragem):**

   Para cada cenário abaixo, indique qual método de amostragem probabilística seria mais apropriado (Amostragem Aleatória Simples, Estratificada, Sistemática, por Conglomerados) e justifique sua escolha:

   a) Um pesquisador deseja avaliar a satisfação dos funcionários de uma grande empresa (5.000 funcionários) com o novo plano de benefícios. Ele tem acesso a uma lista alfabética completa de todos os funcionários.
   b) Uma organização de saúde quer estimar a prevalência de uma doença em crianças em idade escolar em um grande estado, que possui muitas cidades e escolas.
   c) Uma empresa de pesquisa de mercado quer saber a opinião de jovens (18-24 anos), adultos (25-54 anos) e idosos (55+ anos) sobre um novo produto, garantindo que cada faixa etária esteja representada proporcionalmente na amostra.
   d) Um analista de qualidade precisa inspecionar 100 itens de um lote de 10.000 itens que saem de uma linha de produção. Os itens estão numerados sequencialmente.

3. **Problema 3 (Amostragem em Python/R):**

   Você possui um dataset simulado de clientes de uma loja online:

```
ID_Cliente | Idade | Renda | Regiao
-----------------------------------
1          | 25    | 3000  | Norte
2          | 30    | 5000  | Sul
```

```
3          | 45    | 7000  | Leste
... (total de 500 clientes)
```

Assuma que há 150 clientes na região Norte, 200 no Sul, 100 no Leste e 50 no Oeste.

a) Crie um `DataFrame` (Python) ou `data.frame` (R) com 500 clientes simulados, seguindo a distribuição de regiões e com `Idade` entre 18 e 70, `Renda` entre 1500 e 10000.
b) Utilizando **Amostragem Aleatória Simples**, selecione uma amostra de 50 clientes. Mostre as primeiras 5 linhas da amostra e sua distribuição de `Regiao`.
c) Utilizando **Amostragem Estratificada** pela coluna `Regiao`, selecione uma amostra de 50 clientes, garantindo que a proporção de clientes de cada região na amostra seja a mesma da população. Mostre as primeiras 5 linhas da amostra e sua distribuição de `Regiao`. Compare esta distribuição com a da amostra aleatória simples.

Referências Bibliográficas

BUSSAB, Luiz O. de M.; MORETTIN, Pedro A. **Estatística Básica**. 9. ed. São Paulo: Saraiva, 2017.