Desafio 2: Estudo do Estimador de Horvitz-Thompson em Amostragem Aleatória Simples com e sem Reposição

Amostragem

Data máxima de entrega: 17/05/2025

Objetivo

Este projeto tem como objetivo investigar, via simulação computacional no R, as propriedades do estimador de Horvitz-Thompson (HT) em populações finitas sob o plano de amostragem aleatória simples com e sem reposição. Espera-se que o aluno compreenda o comportamento do estimador, avalie sua distribuição empírica, estime sua variância, e interprete resultados à luz da teoria de amostragem.

1 Contexto Teórico

Seja uma população finita $U = \{1, 2, \dots, N\}$ com valores associados $y_1, y_2, \dots, y_N$. Deseja-se estimar o total populacional:

$$t = \sum_{k=1}^{N} y_k$$

O estimador de Horvitz-Thompson é dado por:

$$\hat{t}_\pi = \sum_{k \in s} \frac{y_k}{\pi_k}$$

onde π_k é a probabilidade de inclusão do elemento k na amostra s.

Algumas propriedades importantes:

- $\mathbb{E}[\hat{t}_\pi] = t$ (não-viesado);
- A variância do estimador pode ser estimada com fórmulas baseadas na estrutura do plano amostral;
- O estimador continua centrado mesmo que a distribuição de y_k na população não seja normal.

2 Etapas do Projeto

1. Geração da População

- Gerar uma população de tamanho N (ex: $N = 1000$) com valores y_k simulados segundo diferentes distribuições:
  - Normal padrão: $y_k \sim N(10, 4)$
  - Exponencial: $y_k \sim \text{Exp}(1)$
  - Log-normal: $y_k \sim \text{Lognormal}(0, 1)$

2. Amostragem Aleatória Simples com e sem Reposição

- Selecionar amostras de tamanhos variados, por exemplo: $n \in \{10, 50, 100, 200\}$.
- Estimar o total populacional usando o estimador de Horvitz-Thompson.

3. Repetição e Avaliação

- Repetir o processo de amostragem $R = 1000$ vezes para cada combinação de população e n.
- Avaliar:
  - Viés médio
  - Distribuição empírica do estimador
  - Comparar estimativa da variância empírica com variância teórica (quando possível)

3 Sugestões de Implementação no R

Use os pacotes `sampling` ou `survey` para criar os planos amostrais e calcular os estimadores. Por exemplo:

```
# # sem uso dos pacotes abaixo
# #library(sampling)
# #library(survey)

# # Criar população
# N <- 1000
# y <- rlnorm(N, meanlog = 0, sdlog = 1)
# pop <- data.frame(id = 1:N, y = y)

# # Amostragem com reposição
# n <- 100
# amostra <- sample(1:N, n, replace = TRUE)
# pik <- rep(1 - (1 - 1/N)^n, n) # Nota: Este pik é para cada unidade k ser incluída.
                                  # Se y_sample tem repetições, e você divide cada y_k repetido por este p
                                  # não é o estimador HT usual para SRSWR que soma sobre unidades distint
                                  # O estimador HT para SRSWR geralmente é sum_{k in s_distinct} y_k / (1
# y_sample <- y[amostra]

# # Estimador HT
# Y_HT <- sum(y_sample / pik)
```

4 Tópicos de Investigação

- A distribuição empírica de $\hat{t}_\pi$ se aproxima da normalidade? O que ocorre para n pequeno?
- Como o tamanho da amostra influencia o erro padrão?
- Como o viés e a variância mudam para diferentes distribuições da população?
- Compare a estimativa da variância fornecida pelo `svytotal()` com a variância empírica das 1000 repetições.
- Que acontece se no caso da amostragem aleatória simples sem reposição é usada a função `srswor1()` que usa o algoritmo Selection-rejection method do pacote `sampling` para tirar a amostra?

5 Entrega

- Relatório em PDF contendo:
  - Introdução teórica muito breve
  - Descrição da população e amostragens simuladas

  - Gráficos da distribuição do estimador
  - Tabelas com viés, variância empírica e teórica
  - **Discussão dos resultados**
- Código-fonte implementado em R bem comentado

Referências Sugeridas

- Särndal, C. E., Swensson, B., & Wretman, J. (2003). *Model Assisted Survey Sampling.* Springer.
- Tillé, Y. (2011). *Sampling and Estimation from Finite Populations.* Wiley.