Planos amostrais complexos na estimação da média das notas de português

Pedro Henrique Corrêa de Almeida[1] and Gustavo Almeida Silva[1]

[1*]Estatística, UFJF, Juiz de Fora, Brasil, MG.

Abstract

O estudo de simulação realizado nesta pesquisa envolve a geração de 1.000 replicações de designs de amostragem estratificada e de múltiplos conglomerados. Cada replicação considera diferentes cenários, como tamanhos variados de cluster e níveis de correlação dentro do cluster. Os dados simulados são então analisados usando técnicas estatísticas apropriadas para avaliar o viés, erros padrão e outras medidas relevantes para cada projeto de amostragem. Os resultados desta pesquisa contribuem para a compreensão dos pontos fortes e limitações das pesquisas estratificadas e de múltiplos conglomerados. O estudo destaca a importância de considerar desenhos de amostragem complexos e suas métricas associadas para obter estimativas confiáveis e robustas. O conhecimento obtido com este trabalho pode ajudar pesquisadores e profissionais na seleção de estratégias de amostragem apropriadas para seus contextos de pesquisa específicos.

Keywords: Amostragem, Plano Amostral Complexo, Amostragem estratificada, Amostragem por conglomerado, Simulação Monte Carlo

1 Resumo

Este trabalho apresenta um estudo de simulação sobre métodos de amostragem complexa para a estimação da média amostral. Os métodos analisados são baseados em amostragem conglomerada em 1, 2 e 3 estágios. O objetivo é comparar o desempenho desses métodos em termos de eficiência e precisão da estimativa.

A amostragem complexa é amplamente utilizada em pesquisas em que a população de interesse possui uma estrutura hierárquica ou está dividida em subpopulações distintas. A amostragem conglomerada é uma técnica comumente aplicada nesse contexto, em que a população é dividida em conglomerados e, em seguida, uma amostra é selecionada em cada conglomerado.

Neste estudo, são simulados diferentes cenários com base em parâmetros de amostragem realistas. São considerados os métodos de amostragem conglomerada em 1, 2 e 3 estágios, nos quais a seleção dos conglomerados e das unidades amostrais é realizada de forma sequencial.

Através das simulações, são comparados os estimadores das médias amostrais obtidos pelos diferentes métodos, levando em consideração a variância da estimativa e a eficiência em relação ao tamanho da amostra. Além disso, são avaliados possíveis Viéses de estimadores e a precisão das estimativas em cada estágio da amostragem conglomerada.

Os resultados das simulações fornecem insights valiosos sobre a adequação e o desempenho dos métodos de amostragem conglomerada em diferentes estágios. Espera-se que este estudo contribua para a compreensão das complexidades da amostragem em pesquisas com estrutura hierárquica e auxilie pesquisadores na escolha do método de amostragem mais apropriado para suas necessidades.

2 Introdução

A amostragem desempenha um papel fundamental na estatística, permitindo aos pesquisadores obterem informações sobre uma população a partir de uma amostra representativa. Através de métodos estatísticos robustos, é possível extrapolar conclusões precisas e confiáveis sobre a população em geral. No entanto, a amostragem muitas vezes enfrenta desafios práticos, como a seleção adequada das unidades amostrais e a consideração de complexidades inerentes a certos planos de amostragem.

De forma geral, é amplamente reconhecido na teoria da amostragem que, embora o esquema de amostragem aleatória simples (AAS) seja teoricamente simples, na prática, é pouco utilizado devido às restrições orçamentárias e à busca por métodos probabilísticos que forneçam informações mais precisas. Além disso, é comum encontrar dificuldades na obtenção de cadastros adequados para o AAS, bem como lidar com situações de não resposta, o que requer considerar observações com pesos desiguais (**?**). A especificação inadequada na análise do plano amostral selecionado também pode resultar em estimativas enViésadas, destacando a importância de estudar metodologias que levem em conta o esquema de amostragem adotado.

Este artigo tem como objetivo explorar a interseção entre a amostragem em estatística e a simulação computacional, destacando como essa abordagem combinada

pode contribuir para aprimorar a qualidade das inferências estatísticas. Serão apresentados conceitos fundamentais da amostragem, incluindo diferentes métodos de seleção amostral e as respectivas propriedades, e, em seguida, será discutido como a simulação computacional pode ser aplicada para investigar essas técnicas em contextos específicos.

Ao integrar a simulação computacional à amostragem estatística, os pesquisadores podem explorar virtualmente uma ampla gama de cenários de amostragem, considerando diferentes planos amostrais, tamanhos de amostra e distribuições populacionais. Além disso, a simulação permite a avaliação de métricas de desempenho, como viés e erro padrão, fornecendo insights valiosos sobre a precisão e a eficiência dos métodos de amostragem em diferentes contextos.

3 Metodologia

O objetivo deste trabalho é comparar diferentes planos de amostragem em estágios complexos, como a amostragem estratificada e a amostragem conglomerada. Para realizar essa comparação, foi conduzido um estudo de simulação. O estudo tem como propósito investigar e avaliar o desempenho desses diferentes planos amostrais em termos de eficiência, precisão e viés. Através da simulação, é possível criar cenários controlados que permitem analisar o impacto de cada plano amostral em diferentes características da população. Com base nos resultados obtidos na simulação, será possível identificar quais planos de amostragem são mais adequados para determinados contextos e auxiliar na tomada de decisões estatísticas mais embasadas.

Para isso, foi utilizado o conjunto de dados: **Alunos.txt**, que se trata de dados sobre notas de alunos na prova de portugues. Os dados são populacionais. Assim, diferentes métodos de amostragem complexa foram avaliados utilizando esse conjunto de dados.

O conjunto de dados possui 6 variáveis, são elas:

Variável	Descrição
Aluno	ID do aluno
Rede	Rede de ensino
Escola	ID da escola
Turma	ID da turma
Port	Nota no teste de portugues de cada aluno, é tambem a variável de interesse desse trabalho

Os metodos estudados foram:

- Amostragem Estratificada
  - Foram testados estratificação por Rede e estratificação por Escola
- Amostragem Conglomerada
  - 1 estágio por Escolas
  - 1 estágio por Turmas

  - 2 estágios: UPA-Escolas, USA-Turmas
  - 3 estágios: UPA-Escolas, USA-Turmas, UTA-Alunos
- Amostragem Conglomerada com PPT Poisson
  - 1 estágio por Escolas, tamanho via número de turmas
  - 1 estágio por Escolas, tamanho via número de alunos

4 Estudo de Simulação

Considerou-se como variavel de interesse a média da variavel Port com transformação logaritmo natural, ou seja, a variavel estimada via diferentes metodos de amostragem complexa foi: $ln(Port)$

Para a cada plano amostral, foram replicadas 1000 vezes amostras de tamanho 500 e 750, para cada replicação foram calculadas as estimativas pontuais, o erro padrão e o intervalo de confiança de 95%

Para avaliar o desempenho de cada plano amostral. foram consideradas metricas como **Víes, Erro-padrão e Erro Quadrático Médio**.

O verdadeiro valor da variavel estimada é de:

$$\frac{\sum_{i=1}^{n} ln(Port_i)}{n} = 6.218181$$

O conhecimento de tal valor é importantíssimo para o calculo do viés e consequente a decisão sobre o plano maostral mais adequado para o problema.

4.1 Amostragem Estratificada

A amostragem estratificada é uma técnica valiosa que permite uma seleção mais precisa e representativa da amostra, considerando as heterogeneidades presentes na população. Ao estratificar a população em subgrupos e selecionar uma amostra de cada estrato, é possível obter estimativas mais confiáveis e insights mais detalhados sobre os diferentes grupos presentes na população de interesse.

Nesse contexto, trabalhou-se com duas divisões de estratos: Rede e Escola. Primeiramente foi realizada 1000 replicações com um tamanho amostral igual a 500, após isso realizou-se novamente 1000 replicações com 750. Vemos um ótimo dos 3 tipos de alocação, onde todos apresentam um EQM extremamente baixo. O intervalo de confiança considerado foi de 95, assim vemos que o metodo de alocação Proporcional foi aquele que mais se aproximou desse valor. Os demais metodos apresentaram taxa de rejeição inferior ao nivel de significancia definida.

4.1.1 Estratificada por Rede

Após aplicar as 1000 replicações utilizando o método de Monte Carlo. Seguem as médias das estimativas produzidas utilizando cada uma das alocações.

n = 500

Alocação	$\hat{\theta}$	$E\hat{P}(\theta)$	$IC(95\%) \subset \theta$	Viés	EQM
Uniforme	6.218338	0.0104576	0.966	0.0001568	0.0001094
Proporcional	6.218293	0.0090885	0.951	0.0001121	0.0000826
Neyman	6.218179	0.0090841	0.964	-0.0000020	0.0000825

n = 750

Alocação	Estimativas	ErroPadrão	$IC(95\%) \subset \theta$	Viés	EQM
Uniforme	6.217622	0.0085606	0.968	-0.0005597	7.36e-05
Proporcional	6.218224	0.0074278	0.960	0.0000427	5.52e-05
Neyman	6.217922	0.0074324	0.966	-0.0002592	5.53e-05

Com base no resultados obtidos a partir do estudo de simulação, podemos verificar que os intervalos de confiança de 95% conteram o valor real, aproximadamente, 95% das vezes, como era o esperado. Além disso, para todos os métodos, tivemos um erro quadrático médio baixo,

4.1.2 Estratificada por Escola

Alocação	Estimativas	ErroPadrão	$IC(95\%) \subset \theta$	Viés	EQM
Uniforme	6.218014	0.0094822	0.981	-0.0001673	8.99e-05
Proporcional	6.217734	0.0084041	0.983	-0.0004475	7.08e-05
Neyman	6.217978	0.0086743	0.986	-0.0002035	7.53e-05

Alocação	Estimativas	ErroPadrão	$IC(95\%) \subset \theta$	Viés	EQM
Uniforme	6.218394	0.0081765	0.982	0.0002125	6.69e-05
Proporcional	6.218155	0.0070644	0.991	-0.0000268	4.99e-05
Neyman	6.218010	0.0072053	0.984	-0.0001719	5.19e-05

Os 3 metodos apresentaram um erro quadrático médio baixo, porém o intervalo de confiança não obteve desempenho desejado. Definido um nivel de confiança de 5, viu-se que os três métodos apresentaram um nível de rejeição menor que o esperado. Buscando uma melhor convergência, ambos os metodos foram replicados 1000 vezes novamente, porém fixando um tamanho amostral maior, igual a 750