Estudo Comparativo de Planos Amostrais Complexos na Estimação da Média de Notas Escolares

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

[1*]Dep. Estatística, Instituto de Ciências Exatas, Universidade Federal de Juiz de Fora, Juiz de Fora, MG, Brasil.

Abstract

Buscando comparar o desempenho de diferentes planos amostrais em um mesmo conjunto de dados, este trabalho utiliza o método de simulação Monte Carlo para geração de amostras. Os dados simulados são então analisados usando técnicas estatísticas apropriadas para avaliar o viés, erros padrão e outras medidas relevantes para cada plano amostral. Os resultados desta pesquisa contribuem para a compreensão dos pontos fortes e limitações das pesquisas estratificadas e em múltiplos conglomerados. O estudo destaca a importância de considerar desenhos de amostragem complexos e suas métricas associadas para obter estimativas confiáveis e robustas. 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.

Keywords: Amostragem, Plano Amostral Complexo, Amostragem Estratificada, Amostragem por Conglomerado, Simulação Monte Carlo

.

1 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 (Skinner and Vieira, 2005). A especificação inadequada na análise do plano amostral selecionado também pode resultar em estimativas enviesadas, destacando a importância de estudar metodologias que levem em conta o esquema de amostragem adotado.

A simulação de Monte Carlo é uma técnica estatística que envolve a geração de múltiplas amostras aleatórias com base em modelos probabilísticos(Kleijnen, 1995). É amplamente utilizada para avaliar incertezas, estimar parâmetros e estudar o desempenho de métodos estatísticos em uma variedade de áreas (Kroese and Rubinstein, 2012).

Nesse contexto, o trabalho tem como objetivo explorar a interseção entre a amostragem e a simulação computacional via método Monte Carlo, 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.

2 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 via método de Monte Carlo. 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, foi 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 em um determinado teste de português. 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

Nenhuma observação possuia valores faltantes, e portanto não foi necessário técnicas de imputação de dados

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

Para a seleção amostral utilizou-se o pacote do software R **sampling**, já para para os cálculos das estimativas foi utilizado o pacote **survey**. Ambas documentações detalhadas são encontradas em:

- Doc. Sampling
- Doc. Survey

3 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 e seu o erro padrão

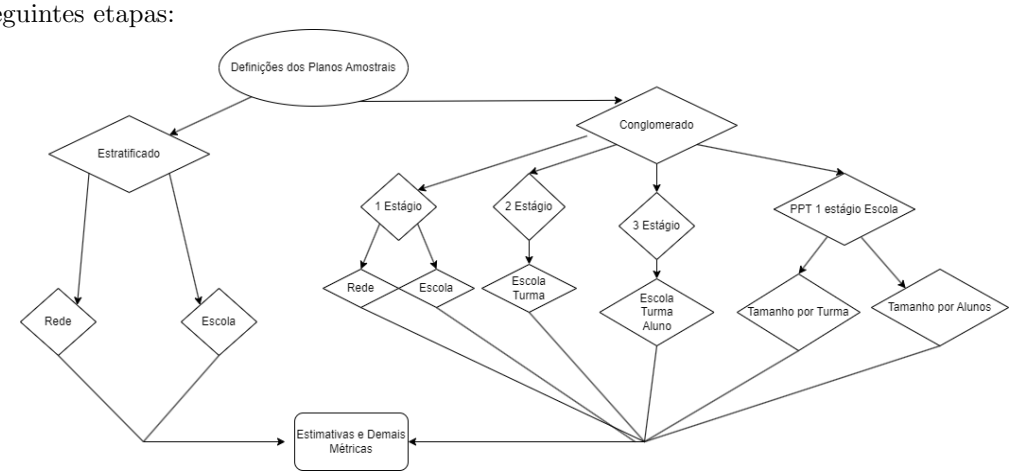
Para avaliar o desempenho de cada plano amostral. foram consideradas metricas como **Víes, erro padrão, Erro Quadrático Médio Relativo e a Proporção de captação do valor paramétrico no intervalo de confiança de** 95% **da estimativa**.

O verdadeiro valor da variável 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.

Construi-se o seguinte diagrama para facilitar passo a passo aplicado durante as seguintes etapas:



3.1 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.

O estimador não viciado do parâmetro de média é dado por:

$$\bar{y_h} = \sum_{i=1}^{n_h} \frac{y_{hi}}{n_h}$$

Enquanto o estimador da variância do estimador de média é dado por:

$$\hat{V}_{AES}(\bar{y}_{AES}) = \sum_{h=1}^{H} W_h^2 (\frac{1}{n_h} - \frac{1}{N_h}) s_h^2$$

Tais estimadores foram utilizados para os cálculos das estimativas (via pacote **survey**)

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 tamanho amostral igual a 750.

Foram consideradas 3 tipos de alocação amostral

- Uniforme
- Proporcional ao Tamanho
- Ótima de Neyman

3.1.1 Estratificada por Rede

Após aplicar as 1000 replicações utilizando o método de Monte Carlo, obteu-se as seguintes estimativas

n = 500

Alocação	$\hat{\theta}$	$\hat{EP(\theta)}$	$IC(95\%) \subset \theta$	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

Observou-se um ótimo desempenho dos 3 tipos de alocação, onde todos apresentaram um EQM baixo. A proporção de vezes em que o intervalo de confiança de 95% incluiu o valor paramétrico se mostrou próximo de 95% para os 3 métodos, onde a Alocação Ótima de Neyman foi aquele a apresentar a melhor métrica, apresentando a maior proporção de inclusão do valor paramétrico tendo o menor erro padrão entre os 3 métodos, indicando uma boa precisão das estimativas

n = 750

Alocação	$\hat{\theta}$	$\hat{EP(\theta)}$	$IC(95\%) \subset \theta$	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

Aumentando o tamanho amostral, observou-se que os 3 metodos tiveram uma melhora em seus desempenhos. O EQM se mostrou ainda menor. Além disso, a proporção de vezes de inclusão do valor paramétrico no intervalo de confiança de 95% mostrou um aumento em relação ao tamanho amostral 500, onde todos os métodos apresentaram uma taxa de inclusão maior com um menor erro padrão, indicando uma boa precisão das estimativas

3.1.2 Estratificada por Escola

Após aplicar as 1000 replicações utilizando o método de Monte Carlo, obteu-se as seguintes estimativas

n = 500

Alocação	$\hat{\theta}$	$\hat{EP}(\theta)$	$IC(95\%) \subset \theta$	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

Os 3 metodos apresentaram um erro quadrático médio baixo, sendo menor que aquele obtido via estratificação por Rede. Além disso, a proporção de vezes de inclusão do valor paramétrico no intervalo de confiança de 95% mostrou-se alto, com todos os métodos apresentando taxa de inclusão maior que 0.980. O método de Alocação Ótima de Neyman e Alocação Proporcional obtiveram as melhores métricas, possuindo as maiores taxas de inclusão do valor paramétrico possuindo os menores erros padrões

n = 750

Alocação	$\hat{\theta}$	$\hat{EP}(\theta)$	$IC(95\%) \subset \theta$	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

Aumentando o tamanho amostral, não observou-se uma melhora no EQM dos 3 métodos (valor esse que ja era extremamente baixo). Apesar disso, a proporção de vezes de inclusão do valor paramétrico no intervalo de confiança de 95% apresntou melhora em todos os métodos. O método de Alocação Proporcional foi aquele a apresentar a melhor métrica, tendo incluindo o valor paramétrico em seu intervalo em 99.1% das replicações, e apresentando o menor erro padrão entre os métodos, indicando uma ótima precisão das estimativas

3.2 Amostragem Conglomerada

A principal motivação por trás da amostragem conglomerada é simplificar o processo de amostragem quando a população é muito grande ou geograficamente dispersa. Em vez de lidar com cada elemento individualmente, os conglomerados podem ser selecionados de forma mais eficiente, reduzindo os custos e o tempo necessários para a coleta de dados. Esse método apresenta custos de operações menores quando comparados a outro métodos amostrais

O estimador não viciado do parâmetro de média por conglomerado é dado por:

$$\bar{y_N} = \frac{\hat{Y}}{M_0}$$

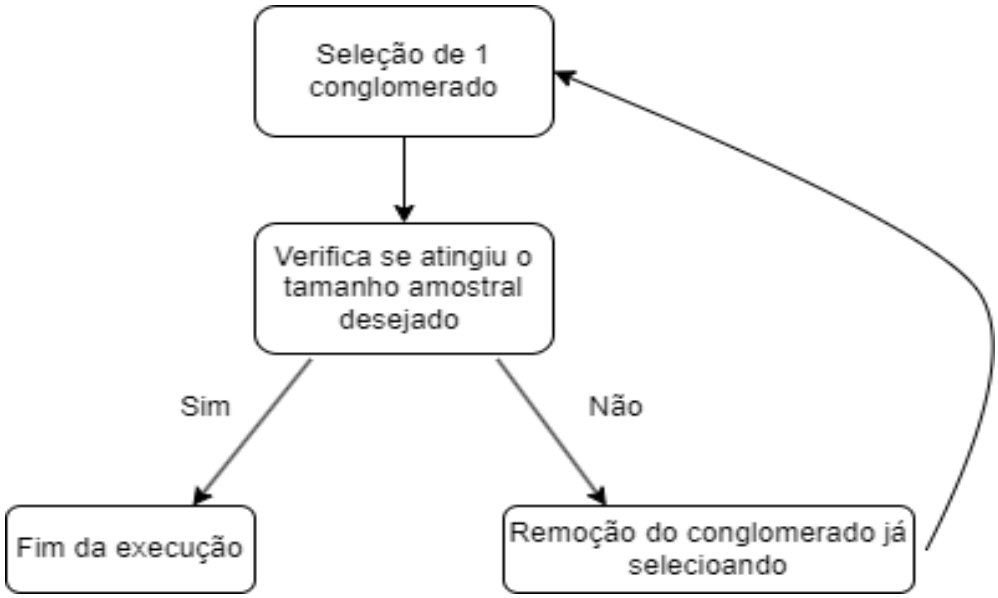
Enquanto o estimador da variância do estimador natural é dado por:

$$\hat{V}_{ACS}(\bar{y}_N) = \frac{1}{\bar{M}^2}\frac{1-f}{n}s_e^2$$

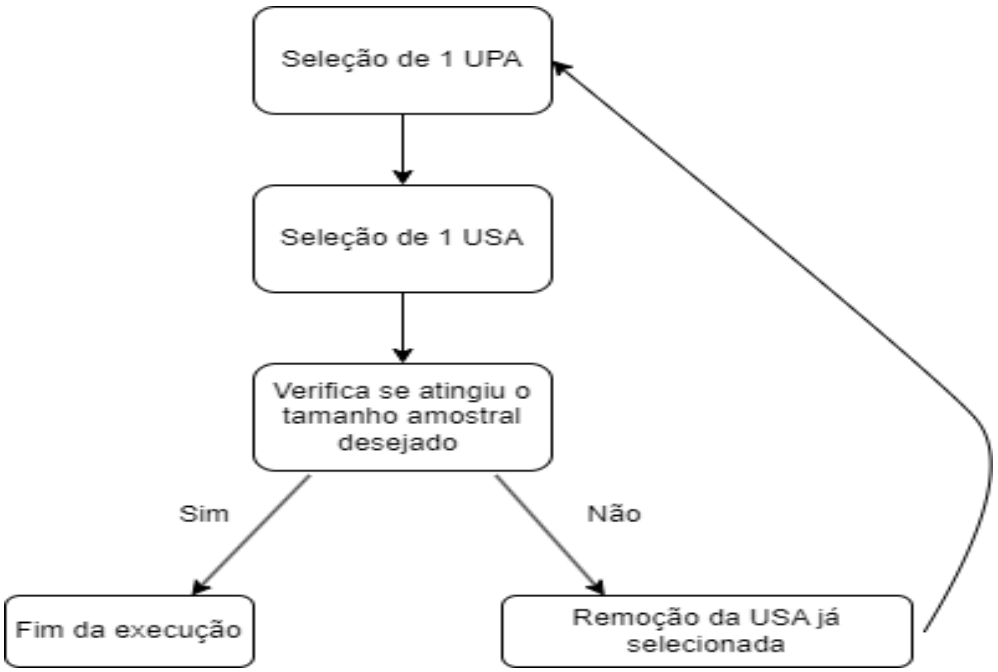
Tais estimadores foram utilizados para os cálculos das estimativas (via pacote **survey**)

As funções computacionais utilizadas nesse trabalho do pacote **sampling** não permitem a seleção do tamanho amostral final em uma amostragem conglomerada, é possível apenas definir o tamanho de cada conglomerado. Assim, buscando um tamanho amostral próximo a 500 e 750, foi construído o seguinte algoritmo de seleção amostral.

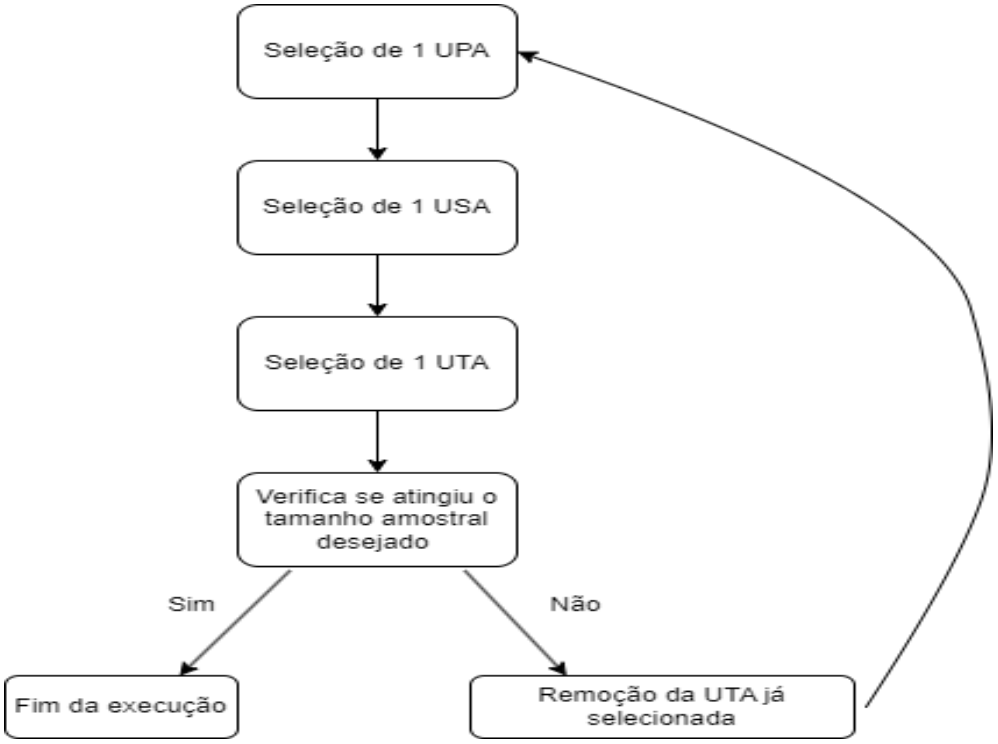
- 1 estágio



- 2 estágios


Seleção de 1 UPA
Seleção de 1 USA
Verifica se atingiu o tamanho amostral desejado
Sim
Não
Fim da execução
Remoção da USA já selecionada

- 3 estágios


Seleção de 1 UPA
Seleção de 1 USA
Seleção de 1 UTA
Verifica se atingiu o tamanho amostral desejado
Sim
Não
Fim da execução
Remoção da UTA já selecionada

3.2.1 Conglomerados

Após aplicar as 1000 replicações utilizando o método de Monte Carlo, obteu-se as seguintes estimativas

n = 500

Conglomeração	$\hat{\theta}$	$E\hat{P}(\theta)$	$IC(95\%) \subset \theta$	Viés	EQM
Escola	6.220454	0.0288410	0.904	0.0022722	0.0009215
Turma	6.219460	0.0244669	0.944	0.0012786	0.0005880
Escola e Turma	6.220118	0.0262029	0.917	0.0019367	0.0008067
Escola, Turma e Aluno	6.220172	0.0138810	0.987	0.0019904	0.0001177

Observou-se que os 4 métodos apresentaram um EQM baixo, onde o melhor foi aquele dado pelo metodo divido em 3 estágios. Observando a proporção de vezes que o intervalo de 95% conteve o valor paramétrico, vemos que novamente o método com 3 estágios apresentou melhor desempenho, apresentando uma alta taxa de englobamento do valor paramétrico e um erro padrão baixo indicando que as estimativas estão precisas. O método a apresentar pior desempenho foi aquele com apenas 1 estágio de conglomeração por escola, apresentando um EQM alto e uma baixo inclusão do valor parametrico em seu intervalo de confiança de 95%

n = 750

Conglomeração	$\hat{\theta}$	$E\hat{P}(\theta)$	$IC(95\%) \subset \theta$	Viés	EQM
Escola	6.218883	0.0239705	0.916	0.0007016	0.0006286
Turma	6.219579	0.0198613	0.934	0.0013977	0.0004302
Escola e Turma	6.219389	0.0220708	0.926	0.0012072	0.0005188
Escola, Turma e Aluno	6.219319	0.0125530	0.992	0.0011375	0.0000814

Aumentando o tamanho amostral, observou-se que os 4 metodos tiveram uma melhora em seus desempenhos. O melhor método continou sendo aquele com 3 estágios de conglomeração e o pior método foi aquele com apenas 1 estágio de conglomeração por escola

3.2.2 Conglomerados com PPT Poisson

Realizou-se amostragem conglomerada em 1 estágio por escola com PPT Poisson das variáveis Turma e Escola. Foram obtidas as seguintes estimativas após 1000 replicações

n = 500

Proporcional	$\hat{\theta}$	$E\hat{P}(\theta)$	$IC(95\%) \subset \theta$	Viés	EQM
Turma	6.218823	0.0299626	0.912	0.0006420	0.0009433
Aluno	6.221696	0.0305077	0.918	0.0035142	0.0009983

Observou-se que ambos os métodos apresnetaram um EQM baixo, porém maior que todos os métodos conglomerados sem PPT Poisson. Além disso, vemos que ambos os intervalos de confiança de 95% tiverem uma baixa inclusão do valor parametrico, em torno de 0.92, indicando um pior desempenho quando comparados aos demais métodos ja vistos

n = 750

$\hat{\theta}$	$E\hat{P}(\theta)$	$IC(95\%) \subset \theta$	Viés	EQM	
Turma	6.219528	0.0253059	0.937	0.0013468	0.0006852
Aluno	6.219404	0.0253226	0.940	0.0012230	0.0005924

Aumentando o tamanho amostral, observou-se que os 2 metodos tiveram uma melhora em seus desempenhos. O EQM apresentou valores menores e a proporção que o intervalo de confiança de 95% incluiu o valor paramétrico se aproximou do valor teórico fixado de 95%, porém ainda se mostrando abaixo desse nível

4 Recomendações

Através do estudo de simulação realizado, padrões e insights valiosos de cada plano amostral puderam ser vistos.

A amostragem via estratificação foi aquela a apresentar melhor desempenho (assim como era esperado por conta de resultados teóricos). Ambas as variáveis utilizadas para estratificação apresentaram um baixo EQM, assim como uma boa taxa de inclusão do valor paramétrico em seus intervalos de confiança de 95%. Realizando um estudo da taxa de inclusão por erro padrão, observou-se que a variável Escola, se mostrou como uma variável de estratificação superior que Rede, com taxas maiores e erros padrões menores.

Dentre os métodos de alocações utilizados, a Alocação Proporcional ao Tamanho e a Alocação Ótima de Neyman foram aquelas a apresentarem melhor desempenho. Utilizando a variável de estratificação como Escola, a Alocação Proporcional ao Tamanho apresentou métricas superiores a Alocação Ótima de Neyman, com um EQM menor, erro padrão menor e uma taxa de inclusão do valor paramétrico maior

A seleção amostral estratificada apresntou um tempo razoável de execução, não passando de 17 minutos para realizar 1000 replicações, indicando um tempo menor que 1.02 segundos por replicação

A amostragem em conglomerados apresentou um desempenho pior que a amostragem estratificada. Os EQM entre ambos planos amostrais foram semelhantes, porém a taxa de inclusão do valor paramétrico nos intervalos de confiança de 95% para

cada replicação foram inferiores, apresentando um erro padrão maior com uma menor taxa de inclusão, ou seja, apesar dos intervalos de confianças serem maiores e portanto serem capazes de englobarem uma maior gama de valores, o valor paramétrico não foi incluido em maiores proporções

Uma tipo de conglomeração que fugiu da observação acima foi aquela por 3 estágios, onde ela apresentou o menor erro padrão entre os métodos de conglomeração (ainda maior que todos os planos estratificados) sendo capaz de apresentar um alta taxa de inclusão do valor paramétrico, 0.987 para tamanho amostral igual a 500 e 0.992 para tamanho amostral igual a 750.

Método de PPT de Poisson não apresentaram um bom desempenho, possuindo uma baixa taxa de inclusão do valor paramétrico em seus intervalos de confiança apesar do erro padrão alto. Além disso, apresentaram os piores valores do EQM entre todos os métodos testados

Um grande empecilho para seleção amostral via conglomeração foi o tempo de execução, onde certas simulações apresentaram tempo de execução maiores que 3 horas, ou seja mais que 108 segundos por replicação. Esse tempo pode ser explicado pelo algoritmo de seleção utilizado, que buscou otimizar a busca pelo tamanho amostral desejado. Outro método testado de se encontrar a combinação ideal entre número de conglomerados para a otimização do tamanho amostral desejado foi via grid de valores. Eram fixados valores inteiros e atráves de intensa simulação e replicação, anotava-se a combinação de número de conglomerados para cada estágio que otimizase o tamanho amostral desejado. O método se mostrou altamente ineficaz, dado o tamanho amostral final para cada combinação do grid era uma variável aleátoria. Observou-se que se tratava de uma distribuição Normal, onde desejou-se que sua média fosse o tamanho amostral desejado, apesar disso, a distribuição apresentava um desvio padrão altíssimo, onde portanto por mais que fossem encontrados os valores do grid que otimizassem o tamanho amostral desejado (parâmetroi de locação da distribuição), o tamanho amostral simulado em cada replicação era extramente volátil e com alta variabilidade

5 Conclusão

O advento da era computacional no século XX afetou a sociedade como um todo, em especial, os estatísticos, fazendo com que a profissão passasse a ser composta por uma tríplice combinação entre estatística, matemática e computação. Sob esta nova perspectiva profissional, para se conduzir um processo de modelagem estatística é necessário que as três componentes deste processo sejam exploradas (Pacheco, 2021).

Assim, foram realizadas comparações detalhadas entre diferentes planos amostrais complexos por meio do uso do método de simulação Monte Carlo. Através da geração de múltiplas amostras simuladas, pode-se analisar o desempenho de cada plano amostral em termos de precisão, viés e outros indicadores de qualidade das estimativas. Os resultados obtidos forneceram insights valiosos para auxiliar na seleção do plano amostral mais adequado, levando em consideração as características específicas da população em estudo.

Ao realizar comparações entre os planos amostrais, viu-se um desempenho superior dos métodos de estratificação, apresentando uma boa taxa de inclusão do valor paramétrico para os 3 métodos de alocações vistos. Além disso, eles apresentaram um baixo tempo de execução e portanto se caracterizaram como o melhor método estudado. Dentro deles, a estratificação por Escola se mostrou superior a estratificação por Rede, apresentando melhores EQMs e uma melhor taxa de inclusão do valor paramétrico por erro padrão

Dentro do amostragem conglomerada, o tempo de execução se mostrou como um empecilho, onde certas replicações apresentaram tempo de execução maiores que 3 horas. Entres esses planos amostrais, aqueles que se utilizaram PPT de Poisson apresentaram um desempenho ruim. Conglomeração em 3 estãgios se mostrou como o melhor método através do algoritmo de seleção amostral utilizado

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.

6 Código

O código utilizado nesse trabalho pode visto em : github.com/monte_carlo_complex_survey_rtargets

Ou em: github.com/monte_carlo_complex_survey_r

References

Kleijnen JP (1995) Verification and validation of simulation models. European journal of operational research 82(1):145–162

Kroese DP, Rubinstein RY (2012) Monte carlo methods. Wiley Interdisciplinary Reviews: Computational Statistics 4(1):48–58

Pacheco PM (2021) Modelagem de dados longitudinais complexos no r: desenvolvimento de um pacote estatístico. URL https://repositorio.ufjf.br/jspui/bitstream/ufjf/13437/1/pedrohenriquedemesquitapacheco.pdf

Skinner C, Vieira MdT (2005) Design effects in the analysis of longitudinal survey data. University of Southampton Institutional Repository