
Universidade Federal de Pernambuco
Centro de Ciências Exatas e da Natureza
Programa de Pós-Graduação em Estatística

PAULO RICARDO PEIXOTO DE ALENCAR FILHO

# AMOSTRAGEM INVERSA DE BERNOULLI E APLICAÇÕES

Recife

2022

PAULO RICARDO PEIXOTO DE ALENCAR FILHO

# AMOSTRAGEM INVERSA DE BERNOULLI E APLICAÇÕES

Dissertação apresentada ao Programa de Pós-Graduação em Estatística do Centro de Ciências Exatas e da Natureza da Universidade Federal de Pernambuco, como requisito parcial à obtenção do título de mestre em Estatística.

**Área de Concentração**: Estatística Aplicada

**Orientador (a)**: Cristiano Ferraz

Recife

2022

Catalogação na fonte
Bibliotecária Monick Raquel Silvestre da S. Portes, CRB4-1217

A368a Alencar Filho, Paulo Ricardo Peixoto de
Amostragem inversa de Bernoulli e aplicações / Paulo Ricardo Peixoto de Alencar Filho. – 2022.
42 f.: il., fig., tab.

Orientador: Cristiano Ferraz.
Dissertação (Mestrado) – Universidade Federal de Pernambuco. CCEN, Estatística, Recife, 2022.
.
Inclui referências.

1. Estatística aplicada. 2. Amostragem. I. Ferraz, Cristiano (orientador). II. Título.

310 CDD (23. ed.) UFPE - CCEN 2022-69

**PAULO RICARDO PEIXOTO DE ALENCAR FILHO**

**AMOSTRAGEM INVERSA DE BERNOULLI E APLICAÇÕES**

Dissertação apresentada ao Programa de Pós-Graduação em Estatística da Universidade Federal de Pernambuco, como requisito parcial para a obtenção do título de Mestre em Estatística.

Aprovada em: 18 de fevereiro de 2022.

**BANCA EXAMINADORA**

Prof. Dr. Cristiano Ferraz

DE/UFPE

Profª. Drª Fernanda De Bastiani

DE/UFPE

Prof.Dr. Pedro Luis do Nascimento Silva

ENCE/IBGE

## AGRADECIMENTOS

Agradeço primeiramente ao Grande Arquiteto do Universo por ter me concedido o dom da vida e sabedoria para encarar as dificuldades dia após dia, que por vezes nos fazem querer sucumbir, mas que sempre ao supera-las nos mostramos ser mais forte do que podíamos imaginar.

A minha família, sobretudo a minha mãe Maria Jucileide, meu pai Ricardo Peixoto e meus irmãos Alexandre Gernol e Danielle Lopes, por todo apoio neste momento de pandemia que não me permitiu estar ao lado de todos eles.

Ao meu professor Cristiano Ferraz por toda ajuda e orientação, mesmo não podendo conhece-lo devido as aulas e orientação remota, esteve sempre presente seja para ensinar e indicar os caminhos, mas também como ser um amigo ouvinte para toda dificuldade. Ao professor André Leite por sua ajuda na parte computacional utilizando o *R*.

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), pelo apoio financeiro durante o Mestrado.

## RESUMO

A ausência de uma listagem, ou cadastro, que identifique e dê acesso aos elementos da população-alvo é uma das adversidades mais recorrentes enfrentadas em levantamentos amostrais. Quando cadastros estão disponíveis, não raro necessitam ser atualizados para serem utilizados. Quando os elementos da população-alvo estão agrupados em conglomerados, o desafio recai com frequência na ausência ou desatualização de listagens de elementos dentro de cada conglomerado. Investir tempo para atualização de tais listagens, embora necessário, pode não ser uma opção, como pôde ser sentido, durante a pandemia de COVID-19 no Brasil, em que pesquisas sorológicas precisavam ser realizadas com a maior brevidade possível, dada a velocidade de contágio do vírus SARS-CoV-2. Consequentemente, estudar planos amostrais que permitam selecionar amostras ao mesmo tempo em que um cadastro seja atualizado, é de grande utilidade. Nesta dissertação, o plano de amostragem inversa de Bernoulli é apresentado, suas propriedades estatísticas discutidas, e o potencial de seu uso no segundo estágio de planos amostrais de dois estágios, para selecionar a amostra durante o processo de atualização do cadastro, investigado. O desempenho de planos em dois estágios combinando o uso de Amostragem de Pareto ou Amostragem Sequencial de Poisson no primeiro estágio, com Amostragem Inversa de Bernoulli ou Amostragem Sistemática no segundo estágio, é estudado por meio de um experimento computacional de Monte Carlo utilizando dados da Pesquisa Sorológica Continuar Cuidando, realizada no Estado da Paraíba, para monitoramento da epidemia de COVID-19.



**Palavras-chaves**: amostragem em dois estágios; amostragem de Pareto; amostragem sequencial de Poisson; amostragem sistemática; covid-19.

## ABSTRACT

The absence of a listing frame that identifies and provides access to the elements of a target-population is one of the most recurrent adversities faced by sampling surveys. When sampling frames are available, not seldom they need to be updated to be used in practice. When the elements of a target-population are grouped in clusters, the challenge rely very often on the non-existence or the outdating of existing listing frames of elements within clusters. Hence, to study sampling designs that allow for selecting the sample at the same time a frame is been updated is of great usefulness. Investing time to update such listing frames, although necessary, may not be an option, as felt during the COVID-19 pandemic in Brazil, where serological surveys needed to be carried out in the shortest time possible, given the quick contagiuos rate of SARS-CoV-2 virus. In this thesis the Inverse Bernoulli Sampling design is presented, its statistical properties discussed and its potential use in the second stage of two-stage sampling designs, to select a sample at the same time an updating screening process is carried out, is investigated. The performance of two-stage designs combining Pareto Sampling or Sequential Poisson sampling in the first stage, with Inverse Bernoulli Sampling or Systematic Sampling in the second stage, is studied by a computational Monte Carlo experiment using data from the serological Survey Sample Continuar Cuidando, carried out in the Brazilian state of Paraiba, to monitor the COVID-19 epidemics.



**Keywords**: COVID-19; Pareto Sampling; Sequential Poisson Sampling; Systematic Sampling; Two-stage Sampling.

## LISTA DE FIGURAS

## LISTA DE TABELAS

## SUMÁRIO

# 1 INTRODUÇÃO

Um dos grandes desafios encontrados em levantamentos amostrais é dispor de um cadastro, ou uma listagem de elementos de uma população-alvo, que esteja atualizado. Em diversas situações que envolvem o agrupamento de elementos de uma população-alvo em conglomerados, o cadastro que lista tais elementos está desatualizado ou mesmo, não existe. Em países com censos populacionais decenais, por exemplo, o nível de desatualização pode ser agravado quando a pesquisa precisa ser realizada pouco tempo antes de um censo. É o caso, no momento em que essa dissertação é escrita, de levantamentos de domicílios permanentes no Brasil. Na expectativa de ter um censo populacional que vem sendo adiado desde 2020, pesquisas com base domiciliar valem-se do Cadastro Nacional de Endereços para Fins Estatísticos (CNEFE), criado e disponibilizado pela Fundação Instituto Brasileiro de Geografia e Estatística (IBGE), por ocasião do censo de 2010. Consequentemente, a listagem de domicílios permanentes dentro de cada setor, necessita de atualização. Exemplos de mudanças que ocorrem naturalmente com o tempo, dentro dos setores censitários, incluem a formação de novos, o fechamento de antigos, ou até mesmo a mudança de estrutura de domicílios, dentre outros.

Processos de atualização de cadastros, embora necessários, levam tempo e usualmente acontecem em duas etapas. Numa primeira visita a campo, é realizada uma varredura para identificação, registro e atualização das unidades amostrais. Esses dados são analisados, uma amostra é selecionada a partir deles, e uma nova visita a campo é realizada para colher as informações das unidades amostrais que foram selecionadas para compor a amostra. Realizar a pesquisa de campo sem a primeira etapa de varredura pode significar perdas consideráveis, incidindo sobre o tamanho efetivo da amostra, já que uma unidade amostral eventualmente selecionada para a amostra pode não mais existir. Por outro lado, podem também existir unidades amostrais novas que não terão chance de fazerem parte da amostra, introduzindo viés não-amostral por falha de cobertura de cadastro. Como consequência, identificar e fazer uso de planos amostrais que possam reduzir o tempo investido no processo de atualização e seleção da amostra é de grande importância e utilidade. Uma classe de planos amostrais com potencial para atingir esse objetivo é a classe de planos de amostragem inversa. A essência de tais planos está em prosseguir o processo de seleção até que o tamanho desejado de amostra seja atingido. A amostragem inversa de

Bernoulli é um plano da classe de amostragem inversa cuja implementação não precisa do conhecimento prévio do tamanho da população. Sua implementação pode ser realizada em campo, em paralelo a um processo de atualização cadastral. Por essa razão, o foco dessa dissertação é o estudo desse plano e como ele pode ser empregado, através da amostragem de conglomerados em dois estágios, em situações que requerem um processo de atualização de cadastro simultaneamente ao processo de seleção de amostra. Uma dessas situações ocorre em pesquisas epidemiológicas que apresentem necessidade de gerar estimativas em tempo muito curto, como é o caso de pesquisas sorológicas para estimar prevalências da COVID-19. A Pesquisa Sorológica Continuar Cuidando, realizada pelo Governo do Estado da Paraíba para monitoramento da epidemia da COVID-19 na população do estado, é um exemplo de tais pesquisas, cujo plano amostral utiliza a amostragem inversa de Bernoulli no segundo estágio. Motivado por sua realização, esta dissertação tem como objetivos:

- apresentar o plano de amostragem inversa de Bernoulli, descrevendo suas propriedades e destacando seu potencial de utilização prática em situações nas quais é necessário atualizar cadastro e selecionar amostra dentro de um mesmo processo de visitação a campo; e

- comparar o desempenho de quatro planos amostrais em dois estágios, a saber: amostragem sequencial de Poisson e amostragem de Pareto em primeiro estágio, e para segundo estágio, amostragem inversa de Bernoulli e amostragem sistemática. Os planos foram escolhidos por serem alternativas de uso prático. Os planos com amostragem inversa de Bernoulli permitem seleção de amostra à medida que o cadastro é atualizado.

Os planos amostrais em dois estágios considerados nesta dissertação são comparados através de um estudo de simulação de Monte Carlo baseado em informações da Pesquisa Sorológica Continuar Cuidando.

Esta dissertação está organizada com a seguinte estrutura: o capítulo 1 apresenta uma introdução e justificativa para estudar o plano de amostragem inversa de Bernoulli, bem como estudar o desempenho dos quatros planos em dois estágios considerados. O capítulo 2 introduz o plano de amostragem inversa de Bernoulli e discute seus principais aspectos teóricos. No capítulo 3 é apresentado o plano de amostragem sistemática, por se tratar de um plano amostral concorrente ao plano de amostragem inversa de Bernoulli.

No capítulo 4 são apresentados os planos amostrais com probabilidades proporcionais ao tamanho (PPT) que serão considerados para o primeiro estágio dos planos amostrais investigados pelo estudo de simulação: o Sequencial de Poisson (OHLSSON, 1998) e a amostragem de Pareto (ROSÉN, 2000). O capítulo 5 apresenta a teoria de planos amostrais de conglomerados em dois estágios, detalhando os planos que foram investigados no estudo de simulação. O capítulo 6 é dedicado a apresentar informações sobre a Pesquisa Sorológica Continuar Cuidando. Informações a respeito do plano amostral utilizado pela Pesquisa para a coleta de dados são detalhadas. No capítulo 7 é descrito o estudo de simulação de Monte Carlo e as análises realizadas para comparar o desempenho estatístico dos diversos planos amostrais considerados nesta dissertação. Finalmente, no capítulo 8 são realizadas as considerações finais deste trabalho.

# 2 AMOSTRAGEM INVERSA DE BERNOULLI

A amostragem inversa de Bernoulli é uma alternativa ao plano de amostragem de Bernoulli para selecionar elementos populacionais de modo a ter uma amostra de tamanho fixo. Ambos pertencem a classe de planos amostrais que tem a propriedade de fornecer probabilidades iguais de inclusão para todos os elementos de uma população.

Considere $U = \{1, \ldots, k, \ldots, N\}$ como o conjunto de elementos que compõem uma população-alvo de interesse, de tamanho finito $N$. Considere ainda que uma amostra $S$ seja selecionada de $U$ utilizando um plano amostral probabílistico, e defina a variável $I_k$ como indicadora de inclusão de um elemento $k \in U$ na amostra:

$$I_k = \begin{cases} 1, k \in S; \\ 0, k \notin S. \end{cases}$$

Cada $I_k$ corresponde a uma variável aleatória que segue uma distribuição de Bernoulli de parâmetro $\pi$, tal que $\pi = P(I_k = 1)$ é chamado de probabilidade de inclusão de primeira ordem na amostra. O plano de amostragem de Bernoulli (AB) consiste em realizar experimentos aleatórios independentes de Bernoulli sequencialmente, um para cada $k \in U$. Tais experimentos consideram como *sucesso* a inclusão de um elemento da população na amostra, e como *insucesso*, a não inclusão do mesmo na amostra. Quando a $k$-ésima repetição do experimento resulta em sucesso, inclui-se o elemento $k$ na amostra, de tal forma que $(I_k = 1)$. Do contrário, o elemento $k$ não participa da amostra $(I_k = 0)$. O procedimento continua até que experimentos aleatórios tenham sido realizados para todos os elementos da população, registrando ao final o número $n_S$ de elementos incluídos na amostra. Como consequência do plano AB, $n_S$ é uma variável aleatória com distribuição Binomial$(N, \pi)$, que pode ser descrita como:

$$n_S = \sum_{k \in U} I_k.$$

Dessa forma, têm-se:

$$\mathrm{P}(n_S = k) = \binom{N}{k} \pi^k (1 - \pi)^{N-k},$$

para $k = 0, 1, 2, \ldots$

O valor de esperado de $n_S$ é dado por:

$$\mathrm{E}(n_S) = N\pi,$$

e sua variância,

$$\mathrm{Var}(n_S) = N\pi(1-\pi).$$

A amostragem inversa de Bernoulli (AIB) corresponde a execução de experimentos aleatórios de Bernoulli sequencialmente para cada elemento $k \in U$ até que se observe $n$ elementos incluídos na amostra $S$. Este processo de amostragem transfere a aleatoriedade do tamanho da amostra, que sob o plano AB está em $n_S$, para o número $M$ de experimentos de Bernoulli necessários até que a amostra final tenha tamanho fixo $n$. Para tanto, uma regra de parada é adicionada ao algoritmo de seleção de uma amostra de Bernoulli.

É possível descrever um algoritmo de seleção de uma amostra de acordo com o plano AIB como segue:

- Considerar um valor para $\pi(0 < \pi < 1)$ e definir $n = 0$;
- Identificar o primeiro elemento da população $(k = 1)$ e observar a realização $\epsilon_1$ de uma Uniforme $(0, 1)$;
- Se $\epsilon_1 < \pi$, então, o elemento $k = 1$ é selecionado para compor a amostra e $n = n+1$. Do contrário, $k = 1$ não compõe a amostra e $n$ permanece inalterado;
- Para cada $k = 2, 3, \ldots$ observa-se a realização $\epsilon_k$ de uma Uniforme $(0, 1)$, de forma independente, e inclui-se o elemento $k$ na amostra quando $\epsilon_k < \pi$. Nesse caso, atualiza-se o tamanho da amostra fazendo $n = n + 1$; Caso $\epsilon_k > \pi$, o elemtno $k$ não entra na amostra e o valor de $n$ permanece inalterado;
- Continuar o processo de seleção até que o tamanho de amostra $n$ seja igual ao valor pretendido.

O número $M$ de tentativas necessárias para a seleção da amostra de tamanho fixo $n$ utilizando o algoritmo acima, com probabilidade de inclusão de primeira ordem $\pi$, é uma variável aleatória que segue distribuição Binomial Negativa com parâmetros $n$ e $\pi$, de maneira que:

$$P(M = m|n, \pi) = \binom{m-1}{n-1}\pi^n(1-\pi)^{m-n},$$

para $m = n, n+1, n+2, \ldots$

O valor esperado de $M$ é dado por:

$$\mathrm{E}(M) = \frac{n}{\pi},$$

e sua variância, dada por:

$$\mathrm{Var}(M) = \frac{n(1-\pi)}{\pi^2}.$$

O tamanho $M$ pode ser considerado como uma medida de esforço para que se alcance a amostra desejada. Dependendo dos valores de $N$ e $\pi$, no entanto, é possível que o algoritmo de uma AIB chegue até o último elemento de uma população sem atingir o critério desejado de observar uma amostra de tamanho fixo $n$. Por essa razão, é desejável que a condição $M \leq N$ seja observada de maneira a evitar situações críticas tais como $n = 0$ ou $n > N$.

Uma vantagem importante dos planos AIB e AB é que em ambos, o processo de seleção da amostra pode ser empregado no campo, sem necessidade de um cadastro a priori para seleção da amostra. Tanto a AIB quanto a AB podem ser utilizadas para seleção da amostra à medida que um cadastro vai sendo atualizado, mas apenas a AIB tem a possibilidade de garantir um tamanho de amostra fixo igual a $n$. Dessa forma, utilizar o plano AIB, com valor de $\pi$ cuidadosamente escolhido, pode implicar em economia de custos com respeito a tempo e orçamento, quando comparado a planos amostrais de tamanho de amostra fixo que precisam de informação prévia do tamanho da população $N$.

## 2.1 ESTIMAÇÃO DE TOTAIS POPULACIONAIS

Defina $y_k$ como o valor de uma variável de interesse associado ao elemento $k \in U$. Considere $t = \sum_{k \in U} y_k$, o total populacional, como parâmetro de interesse. Em situações nas quais $C$ representa um conjunto de elementos que possuem certa característica de interesse, e $y_k$ é definido como:

$$y_k = \begin{cases} 1, k \in C; \\ 0, k \notin C, \end{cases}$$

$t$ é interpretado como o número total de elementos da população que possuem a característica de interesse que define os elementos de $C$. Nesta dissertação, exemplos de tais características incluem testar positivo num determinado teste de diagnóstico de COVID-19, ou apresentar algum comportamento de risco, como não utilizar máscara durante o período de pandemia. Estimativas de tais totais são obtidas utilizando o estimador de Horvitz e Thompson (1952). Tal estimador pode ser escrito como:

$$\hat{t} = \sum_{k \in S} d_k y_k = \sum_{k \in S} \frac{y_k}{\pi_k}. \tag{2.1}$$

Na Equação (2.1) os $d_k$ são chamados de pesos amostrais básicos e correspondem ao inverso da probabilidade de inclusão do elemento $k$, ou seja $d_k = 1/\pi_k$. Sabe-se que o estimador (2.1) é centrado, e que sua variância pode ser escrita como:

$$\mathrm{Var}(\widehat{t}) = \sum_{k \in U} \sum_{l \in U} (\pi_{kl} - \pi_k \pi_l) \frac{y_k}{\pi_k} \frac{y_l}{\pi_l} = \sum_{k \in U} \sum_{l \in U} \Delta_{kl} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l}. \tag{2.2}$$

Na expressão (2.2) $\pi_{kl}$ é a probabilidade de inclusão de segunda ordem dos elementos $k$ e $l$ na amostra, para $k \neq l$, já $\Delta_{kl}$ é a expressão para $\Delta_{kl} = \mathrm{Cov}(I_k, I_l) = \pi_{kl} - \pi_k \pi_l$. Um estimador centrado para a variância é dado por

$$\widehat{\mathrm{Var}}(\widehat{t}) = \sum_{k \in S} \sum_{l \in S} \frac{\Delta_{kl}}{\pi_{kl}} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l} \tag{2.3}$$

## 2.2 ESTIMAÇÃO DE TOTAIS POPULACIONAIS SOB AIB

A estimção de totais populacionais sob um plano AIB depende de um valor-chave para $\pi$. Tal valor é função das quantidades referentes ao tamanho desejável $n$ da amostra, e das $M$ tentativas necessárias para selecioná-la. Para efeitos comparativos, no plano AB, o valor de $\pi$ é usualmente obtido pela razão entre o tamanho esperado $n_0$ de amostra, e o tamanho $N$ da população, $\pi = n_0/N$. Já no plano AIB, o valor de $\pi$ é determinado pela razão entre o tamanho de amostra desejável $n$ e pelo valor esperado do número de tentativas, $m_0$, isto é, $\pi = n/m_0$. Consequentemente, considerar um valor alto para $\pi$ sob um plano AIB, implica planejar um trabalho de coleta de dados em campo com menor esforço, no sentido de se esperar atingir o tamanho desejado de amostra $n$ num menor número de tentativas.

Em um plano AIB, as probabilidades de inclusão de primeira ordem, são tais que:

$$\begin{aligned}
\pi_k &= \mathrm{P}(k \in S) \\
&= \mathrm{P}(\epsilon_k \leq \pi) \\
&= \mathrm{P}(U(0,1) \leq \pi) \\
&= \int_{-\infty}^{\pi} f(t)dt \\
&= \int_{0}^{\pi} 1dt \\
&= \pi.
\end{aligned}$$

Por se tratarem de realizações independentes da distribuição Uniforme-padrão no algoritmo de seleção, as probabilidades de inclusão de segunda ordem dos elementos $k$ e $l$, são dadas por:

$$\begin{aligned}
\pi_{kl} &= \mathrm{P}(k, l \in S) \\
&= \mathrm{P}(\epsilon_k \leq \pi)\,\mathrm{P}(\epsilon_l \leq \pi) \\
&= \mathrm{P}(U(0,1) \leq \pi)\,\mathrm{P}(U(0,1) \leq \pi) \\
&= \left(\int_{-\infty}^{\pi} f(t)dt\right) \cdot \left(\int_{-\infty}^{\pi} f(w)dw\right) \\
&= \left(\int_{0}^{\pi} 1dt\right) \cdot \left(\int_{0}^{\pi} 1dw\right) \\
&= \pi^2.
\end{aligned}$$

Logo, sob o plano AIB, o estimador de Horvitz-Thompson para o total populacional $t$ assume a forma:

$$\widehat{t} = \frac{m_0}{n} \sum_{k \in S} y_k.$$

Sua variância é dada por:

$$\begin{aligned}
\mathrm{Var}(\widehat{t}) &= \mathrm{Var}\left(\frac{m_0}{n}\sum_{k\in S} y_k.\right) \\
&= \left(\frac{m_0}{n}\right)^2 \mathrm{Var}(\sum_{k\in S} y_k) \\
&= \left(\frac{m_0}{n}\right)^2 \mathrm{Var}(\sum_{k\in U} I_k y_k) \\
&= \left(\frac{m_0}{n}\right)^2 \sum_{k\in U} \mathrm{Var}(I_k) y_k^2 \\
&= \left(\frac{m_0}{n}\right)^2 \pi(1-\pi) \sum_{k\in U} y_k^2 \\
&= \left(\frac{m_0}{n}\right)^2 \frac{n}{m_0}\left(1-\frac{n}{m_0}\right) \sum_{k\in U} y_k^2 \\
&= \left(\frac{m_0}{n}-1\right) \sum_{k\in U} y_k^2,
\end{aligned}$$

onde $I_k$ é a variável indicadora de inclusão de um elemento $k \in U$. Um estimador centrado da variância de $\hat{t}$ é dado por:

$$\widehat{\mathrm{Var}}(\widehat{t}) = \left(\frac{m_0}{n}-1\right)\left(\frac{m_0}{n}\right) \sum_{k\in S} y_k^2.$$

# 3 AMOSTRAGEM SISTEMÁTICA

A amostragem sistemática é um plano amostral que seleciona com equiprobabilidade os elementos de uma população-alvo para compor a amostra. Trata-se de um plano de amostragem de fácil aplicação, uma vez que este método pode ser utilizado mesmo quando não se tenha cadastro prévio dos elementos da população, necessitando apenas de informação sobre o tamanho $N$ da população. Por sua natureza simples, é um plano concorrente a amostragem inversa de Bernoulli, mas que exige uma atualização do cadastro previamente a seleção da amostra.

No método da amostragem sistemática, na sua forma básica tem-se que um primeiro elemento é sorteado aleatoriamente entre os $a$ primeiros elementos da população. Nenhum outro sorteio aleatório será necessário devido ao fato de que os demais elementos da amostra são determinados sistematicamente a cada $a$-ésimo elemento, até o final da listagem. Logo, existem $a$ possíveis amostras, cada uma com a mesma probabilidade $1/a$ de seleção. O valor de $a$ é fixado como o valor inteiro da razão $N/n$, onde $N$ é o tamanho da população e $n$ é o tamanho da amostra. Este valor é denominado de intervalo de amostragem.

Numa amostra sistemática de intervalo amostral $a$ tem-se $N = na + c$, onde $c$ é uma constante que satisfaz $c \in [0, a)$. Para o caso de $c = 0$, uma amostra de tamanho $n$ será sorteada mediante o esquema amostral a seguir. Do contrário, se $c > 0$, o tamanho da amostra poderá ser $n$ ou $n + 1$.

- Selecione com igual probabilidade $1/a$ um inteiro aleatório, digamos $r$, de tal forma $1 \leq r \leq a$;
- A amostra selecionada será composta da seguinte forma

$$S_r = \{k : k = r + (j-1)a; j = 1, 2, \ldots, n_s\}$$

  com $n_s = n$ para o caso $r \leq c$ e $n_s = n + 1$ quando $c \leq r \leq a$. Assim, o conjunto de amostras possíveis, denotado por $\mathcal{S}$, consiste em $\mathcal{S} = \{S_1, S_2, \ldots, S_a\}$

Existe um outro método de seleção de amostra sistemática que evita o problema de ter um tamanho de amostra variável, denominado de método do intervalo fracionado. Esse método seleciona uma amostra sistemática através do uso de um valor fracionado de $a$, resultando em um tamanho de amostra fixo $n$. O esquema amostral é descrito como segue:

- Defina um valor inteiro ou fracionado de $a$ tal que $a = N/n$;
- Considere $\xi$, uma uma realização de uma variável aleatória com distribuição uniforme em $U(0, a)$;
- A amostra selecionada será composta da seguinte forma:

$$S_r = \{k : k - 1 < \xi + (j - 1)a \leq k; j = 1, 2, \ldots, n\}$$

Para as probabilidades de inclusão de um elemento $k$ na amostra tem-se que cada elemento $k$ pertence a uma, e somente uma, das possíveis amostras sistemáticas. Logo sob um plano de amostragem sistemática tem-se:

$$\pi_k = 1/a,$$

Considerando o caso da probabilidade de inclusão de segunda ordem para $k$ e $l$, onde $k \neq l$, tem-se:

$$\pi_{kl} = \begin{cases} 1/a, (k, l) \in S; \\ 0, (k, l) \notin S. \end{cases}$$

Considere a estimação do total populacional de uma população finita sob um plano amostral por amostragem sistemática. O estimador de Horvitz-Thompson assume a forma

$$\widehat{t} = a \sum_{k \in S} y_k,$$

com sua variância dada por

$$\text{Var}(\widehat{t}) = a(a - 1)S^2_{yt},$$

onde tem-se que $S^2_{yt} = \frac{1}{a-1} \sum_{r=1}^{a} (t_{s_r} - \overline{t})^2$ com $\overline{t} = \frac{1}{a} \sum_{r=1}^{a} t_{s_r}$. A respeito do estimador da variância, como tem-se que a condição $\pi_{kl} > 0, \forall k \neq l$ não é verificada, não se consegue utilizar a expressão (2.3) para estimar variância. Detalhes para estimação da variância sob um plano de amostragem sistemática podem ser consultados em Särndal, Swensson e Wretman (1992, pág. 76).

# 4 AMOSTRAGEM COM PROBABILIDADE PROPORCIONAL AO TAMANHO

Planos de amostragem com probabilidades proporcionais ao tamanho (PPT) selecionam elementos da população com probabilidades distintas de inclusão, sendo estas proporcionais a uma medida de tamanho associada a cada elemento.

Em situações práticas é possível contar com cadastros que contenham informações de uma ou mais variáveis auxiliares que sejam indicativas da importância dos elementos populacionais. Tal importância é medida por uma variável chamada de tamanho. De um modo geral, quando essas medidas de tamanho são correlacionadas positivamente com as variáveis de interesse, é possível utilizá-las para determinar valores de $\pi$ que aumentem a precisão de estimativas da seguinte forma. Defina $x_k$ como o valor da variável auxiliar $x$ que mede o tamanho (importância) do elemento $k \in U$. Defina $X = \sum_{k \in U} x_k$, e $n$ como o tamanho fixo de amostra desejado. Então, um plano de amostragem PPT usa probabilidades de inclusão $\pi_k = nx_k/X$.

É comum a aplicação de planos amostrais PPT em pesquisas domiciliares, já que as variáveis de interesse são usualmente correlacionadas com medidas de tamanho que estão disponíveis nos cadastros utilizados para a seleção da amostra. A Pesquisa Industrial Anual - Produção Física, do IBGE, de 1981 é um exemplo de pesquisa que usa amostragem PPT, como destaca Silva, Bianchini e Dias (2021, seção 10.4).

A seguir, apresentamos os dois planos amostrais PPT utilizados para o estudo desta dissertação. Os planos amostrais PPT adotados foram a amostragem sequencial de Poisson e a amostragem de Pareto. A escolha por tais planos PPT ocorre por serem simples e práticos para selecionar unidades primárias de amostragem em planos por conglomerados de dois estágios. A amostragem sequencial de Poisson, por exemplo, é empregada no Sistema de Avaliação da Educação Básica - SAEB, onde a mesma preserva a simplicidade do processo de seleção da amostra. Tal plano também foi utilizado para a seleção de escolas para aplicação de testes no âmbito do Sistema de Avaliação da Educação Básica - SAEB, como descrito em Bussab, Andrade e Silva (1999).A amostragem de Pareto ganha destaque e atenção ao ser utilizada na nova pesquisa domiciliar contínua, que integra a PME e a PNAD. O plano tem sido aplicado no sorteio de unidades primárias de amostragem da PNAD Contínua, como descrito em Freitas e Antonaci (2014). Costa (2007) estudou este plano amostral para a coordenação de amostras PPT em pesquisas repetidas.

## 4.1 AMOSTRAGEM SEQUENCIAL DE POISSON

Proposto por Ohlsson (1998), este plano amostral PPT é um caso particular dos planos de amostragem de Poisson, diferenciando-se por eliminar o tamanho variável da amostra, isto é, o tamanho da amostra será igual ao requerido antes da retirada final. O seguinte esquema amostral seleciona uma amostra sequencial de Poisson:

1. Gere números aleatórios $\epsilon_k$ de uma distribuição Uniforme (0,1) para cada elemento $k$ da população;
2. Calcule a medida de tamanho relativo $p_k = x_k / \sum_{k \in U} x_k$;
3. Calcule o número aleatório modificado $C_k = \frac{\epsilon_k}{p_k}$;
4. Ordene os elementos de forma crescente de acordo com os valores de $C_k$;
5. Selecione para a amostra os $n$ primeiros elementos com os menores valores de $C_k$.

A forma do estimador de Horvitz-Thompson para o total populacional sob o plano descrito é dada por:

$$\widehat{t} = \sum_{k \in s} \frac{y_k}{\pi_k} = \frac{1}{n} \sum_{k \in s} \frac{y_k}{p_k}.$$

Sua variância é dada por:

$$\mathrm{Var}(\widehat{t}) = \frac{1}{n} \frac{N}{N-1} \sum_{k \in U} (1 - np_k) \left( \frac{y_k}{p_k} - t \right)^2 p_k,$$

e um estimador centrado para a variância é escrito como:

$$\widehat{\mathrm{Var}}(\widehat{t}) = \frac{1}{n(n-1)} \sum_{k \in s} (1 - np_k) \left( \frac{y_k}{p_k} - \widehat{t} \right)^2.$$

## 4.2 AMOSTRAGEM DE PARETO

Proposto por Rosén (2000), trata-se de um plano amostral PPT semelhante a de amostragem sequencial de Poisson, por carregar consigo a vantagem de eliminar o efeito do tamanho variável da amostra de uma amostragem de Poisson. Sob tamanhos de amostras iguais, a amostragem de Pareto tende a ser mais eficiente do que a amostragem sequencial de Poisson. O seguinte esquema amostral seleciona uma amostra de Pareto:

1. Gere números aleatórios $\epsilon_k$ de uma distribuição Uniforme (0,1) para cada elemento $k$ da população;

2. Calcule a probabilidade de inclusão desejável da unidade $k$: $\lambda_k = nx_k / \sum_{k \in U} x_k$;

3. Calcule o número aleatório modificado $C_k = \frac{\epsilon_k(1-\lambda_k)}{\lambda_k(1-\epsilon_k)}$;

4. Ordene os elementos de forma crescente de acordo com os valores de $C_k$;

5. Selecione para a amostra os $n$ primeiros elementos com os menores valores de $C_k$.

Como descreve Silva, Bianchini e Dias (2021, seção 10.7), as probabilidades exatas de inclusão sob o plano de amostragem de Pareto não são estritamente proporcionais ao tamanho, além de serem difíceis de se calcular. Entretanto, estudos mostram que o valor $\lambda_k$, dado no passo 2 do esquema amostral, é suficientemente próximo do valor desejado de $\pi_k$ para grande parte dos levantamentos amostrais. Informações sobre resultados para essas aproximações podem ser consultadas em Aires e Rosén (2005).

A forma do estimador de Horvitz-Thompson para o total populacional sob um plano de Pareto é dada por:

$$\widehat{t} = \sum_{k \in s} \frac{y_k}{\lambda_k} = \frac{1}{n} \sum_{k \in s} \frac{y_k}{p_k}.$$

Sua variância é expressa por:

$$\mathrm{Var}(\widehat{t}) = \frac{N}{N-1} \sum_{k=1}^{N} \left( \frac{y_k}{\lambda_k} - \frac{\sum_{l=1}^{N} y_l(1-\lambda_l)}{\sum_{l=1}^{N} \lambda_l(1-\lambda_l)} \right)^2 \lambda_k(1-\lambda_k),$$

e um estimador centrado da sua variância é dado por:

$$\widehat{\mathrm{Var}}(\widehat{t}) = \frac{n}{n-1} \sum_{k=1}^{n} \left( \frac{y_k}{\lambda_k} - \frac{\sum_{l=1}^{n} y_l(1-\lambda_l)/\lambda_l}{\sum_{l=1}^{n}(1-\lambda_l)} \right)^2 (1-\lambda_k).$$

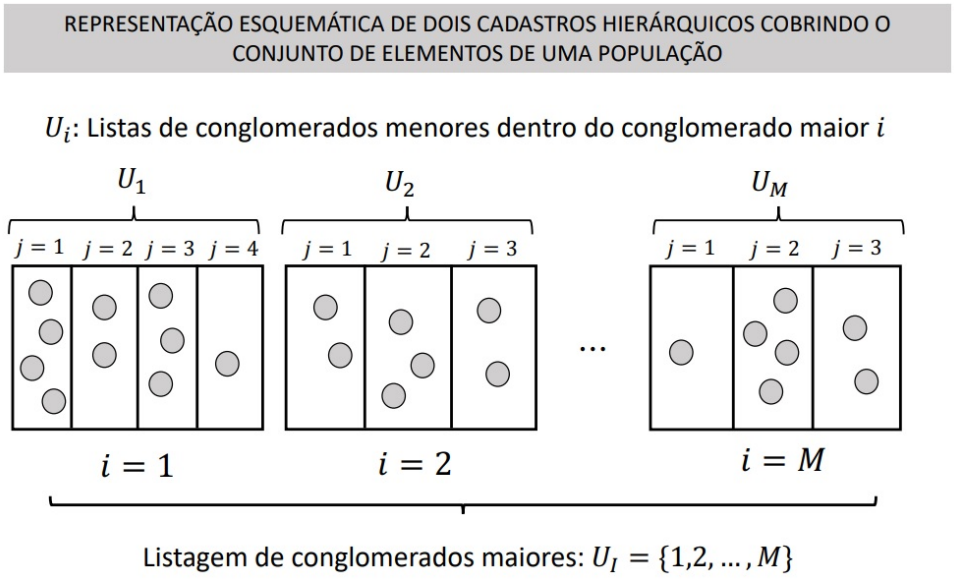
# 5 AMOSTRAGEM POR CONGLOMERADOS EM DOIS ESTÁGIOS

Considere uma população-alvo $U$, formada por $N$ elementos, identificada a partir de dois conjuntos de cadastros de conglomerados hierárquicos, $U_I$ e $U_i$, para $i \in U_I$, de tal forma que:

$$U_I = \bigcup_{i=1}^{M} U_i.$$

O conjunto $U_I$ identifica a população de $M$ conglomerados maiores, e o conjunto $U_i$ identifica $N_i$ conglomerados menores dentro do conglomerado $i \in U_I$. Logo, $U_I = \{1, 2, \ldots, M\}$, e $U_i = \{1, 2, \ldots, N_i\}$, para $i \in U_I$. É possível definir ainda $U_{ij} = \{1, 2, \ldots, N_j\}$ como o conjunto de elementos do conglomerado $j$ dentro do conglomerado $i$. A Figura 1 ilustra a situação de hierarquia.

Figura 1 – Identificação dos cadastros hierárquicos na população $U$.




Fonte: O Autor.

Levantamentos amostrais domiciliares ocorrem em populações nas quais a estrutura acima pode ser identificada de diversas formas. No caso da Pesquisa Sorológica Continuar Cuidando, por exemplo, a população-alvo é de moradores do estado da Paraíba. Os conglomerados maiores são setores censitários, e os conglomerados menores são os domicílios dentro de cada setor censitário.

Nesse contexto, um plano de amostragem de conglomerados em dois estágios é selecionado da seguinte forma:

- **Estágio 1:** selecione uma amostra $S_I \subset U_I$ de conglomerados maiores de acordo com o plano amostral probabilístico $p_I(\cdot)$ que assinala probabilidade de inclusão de primeira ordem $\pi_i$ para o conglomerado $i \in U_I$;

- **Estágio 2:** para cada conglomerado $i \in S_I$, selecione uma amostra $S_i$ de conglomerados menores de acordo com o plano amostral probabilístico $p_{II}(\cdot)$ que assinala probabilidade de inclusão de primeira ordem $\pi_{j|i}$ para o conglomerado menor $j \in S_i$.

Defina $t_{ij}$ como sendo o total da variável de interesse $y$ observada no conglomerado $j$ dentro do conglomerado $i$. Logo,

$$t_{ij} = \sum_{k \in U_{ij}} y_{ijk}$$

é um valor conhecido para todos os conglomerados da amostra do segundo estágio. O total populacional $t$ pode então ser escrito como:

$$t = \sum_{i=1}^{M} \sum_{j=1}^{N_i} t_{ij}.$$

O estimador de Horvitz-Thompson para o total populacional é dado por:

$$\widehat{t} = \sum_{i \in S_I} \sum_{j \in S_i} \frac{t_{ij}}{\pi_i \pi_{j|i}}.$$

Sua variância pode ser escrita como:

$$\begin{aligned} \mathrm{Var}(\widehat{t}) &= \sum_{i \in U_I} \sum_{i' \in U_I} \Delta_{ii'} \frac{t_i}{\pi_i} \frac{t_{i'}}{\pi_{i'}} + \sum_{i \in U_I} \frac{\mathrm{V}_i}{\pi_i} \\ &= \sum_{i \in U_I} \sum_{i' \in U_I} \Delta_{ii'} \frac{t_i}{\pi_i} \frac{t_{i'}}{\pi_{i'}} + \sum_{i \in U_I} \left(\frac{1}{\pi_i}\right) \sum_{j \in U_i} \sum_{j' \in U_i} \Delta_{jj'|i} \frac{t_{ij}}{\pi_{j|i}} \frac{t_{ij'}}{\pi_{j'|i}}, \end{aligned}$$

onde $t_i = \sum_{j \in U_i} y_{ij}$, $\Delta_{ii'} = \pi_{ii'} - \pi_i \pi_{i'}$ e $\Delta_{jj'|i} = \pi_{jj'|i} - \pi_{j|i}\pi_{j'|i}$. A variância do estimador do total populacional sob amostragem por conglomerados em dois estágios é composta por duas componentes: uma devido ao processo de aleatorização do primeiro estágio e outra devido ao processo de aleatorização empregada no segundo estágio.

O estimador centrado para a variância do total populacional é dado por

$$\widehat{\mathrm{Var}}(\widehat{t}) = \sum_{i \in S_I} \sum_{i' \in S_I} \frac{\Delta_{ii'}}{\pi_{ii'}} \frac{\widehat{t}_i}{\pi_i} \frac{\widehat{t}_{i'}}{\pi_{i'}} + \sum_{i \in S_I} \left(\frac{1}{\pi_i}\right) \sum_{j \in S_i} \sum_{j' \in S_i} \frac{\Delta_{jj'|i}}{\pi_{jj'|i}} \frac{t_{ij}}{\pi_{j|i}} \frac{t_{ij'}}{\pi_{j'|i}},$$

onde $\widehat{t}_i = \sum_{j \in S_i} \frac{y_{ij}}{\pi_{j|i}}$.

Gerar estimativas através deste estimador pode requer muito trabalho dependendo da complexidade do plano amostral. Uma alternativa nesses casos é a utilização de estimadores simplificados. Uma proposta de estimador simplificado para a variância do total populacional de um plano amostral por conglomerados em dois estágios é apresentada por Särndal, Swensson e Wretman (1992, pág. 422) e dada pela expressão

$$\mathrm{Var}(\widehat{t}) = \frac{1}{m(m-1)} \sum_{i=1}^{m} \left( \frac{\widehat{t}_i}{p_i} - \widehat{t} \right)^2 ,$$

onde $m$ é o tamanho de amostra do primeiro estágio, e $p_i = m \frac{x_i}{\sum_{i=1}^{M} x_i}$ é a probabilidade de inclusão do primeiro estágio do conglomerado $i \in U_I$.

# 6 APLICAÇÃO EM MONITORAMENTO EPIDEMIOLÓGICO

A Pesquisa Sorológica Continuar Cuidando foi uma ação do Governo do Estado da Paraíba para retratar os cenários da pandemia do novo coronavírus (COVID-19) nos municípios do Estado. Para esta pesquisa foram firmadas parcerias com a Universidade Federal da Paraíba (UFPB), a Fundação de Apoio à Pesquisa do Estado da Paraíba (FAPESQ-PB) e com a Sociedade para o Desenvolvimento da Pesquisa Cientifica (SCIENCE) que definiram as estratégias para o levantamento amostral, coleta e análise dos dados.

A equipe de coleta de dados contou com Agentes Comunitários de Saúde (ACS) que realizaram punção digital e coleta de material orofaringe para realizar os testes rápidos e o Teste RT-PCR no período de 03/11/2020 a 22/12/2020. Os testes rápidos visam averiguar os anticorpos IgM, que apontam se a infecção está ativa e, com isso, se o individuo está infectado e transmitindo o vírus. teste rápidos também verificam os anticorpos IgG para ter uma indicação sobre se o individuo já contraiu o vírus mas se encontra curado; o Teste RT-PCR investiga a detecção do vírus causador, acusando o resultado nas categorias detectável ou não-detectável através do material orofaringe.

No contexto desta dissertação, a Pesquisa Sorológica Continuar Cuidando é um exemplo de pesquisa em que os conceitos aqui estudados se aplicam, isto é, temos um plano amostral por conglomerados em dois estágios para situações que requerem um processo de atualização de cadastro simultaneamente ao processo de seleção de amostra, para gerar estimativas em tempo compatível com o necessário ao lidar com casos de COVID-19. O plano amostral da Pesquisa Sorológica Continuar Cuidando é detalhado a seguir.

## 6.1 O PLANO AMOSTRAL DA PESQUISA

O plano amostral da Pesquisa Sorológica Continuar Cuidando está descrito em Silva et al. (2020). Ele foi definido como estratificado e conglomerado em dois estágios. O método empregado para a seleção das unidades primárias de amostragem em cada estrato, isto é, selecionar os setores censitários, foi a amostragem de Pareto. Para o sorteio de domicílios em cada setor censitário selecionado, foi utilizada a amostragem inversa de Bernoulli. A utilização da amostragem inversa de Bernoulli fez com que após atingir 25 aplicações de testes por setores censitários, a coleta se encerre e com isto ocorra a economia do trabalho

de campo, ao evitar o processo de listagem completa de todos os domicílios do setor em uma etapa prévia.

Para o primeiro nível da estratificação foi levado em consideração o interesse em fornecer resultados para os quatros domínios de estudo definidos como recortes do território do estado da Paraíba. Estes recortes foram definidos reconhecendo que a epidemia tem se espalhado pelo interior e que a gestão da saúde leva em consideração a divisão regional considerada. A Tabela 1 apresenta os domínios correspondentes aos níveis de estratificação territorial.

Tabela 1 – Domínios de estudo considerados para estratificação da PB-COVID19.

| Estratos naturais | Setores | Domicílios | Pessoas | Setores na amostra por mês | Fração amostral setores (%) | Pessoas na amostra por mês | Fração amostral pessoas (por mil) |
|---|---|---|---|---|---|---|---|
| João Pessoa | 945 | 213.230 | 718.744 | 40 | 4,2 | 1.000 | 1,4 |
| Macro 1 | 1.566 | 298.353 | 1.054.552 | 44 | 2,8 | 1.100 | 1,0 |
| Macro 2 | 1.576 | 309.309 | 1.060.620 | 48 | 3,0 | 1.200 | 1,1 |
| Macro 3 | 1.384 | 254.318 | 897.480 | 60 | 4,3 | 1.500 | 1,7 |
| Total Geral | 5.471 | 1.075.210 | 3.731.396 | 192 | 3,5 | 4.800 | 1,3 |

Fonte: Silva et al. (2020).

Para o tamanho total da amostra, estipulou-se um valor fixo de 4.800 indivíduos a serem testados no período de quatro semanas. De posse da amostra de indivíduos no período mensal, é esperado que estime prevalências com margem de erro máximo de 2%. Para subconjuntos populacionais, as margens de erro podem ser maiores ao levar em conta o tamanho da amostra no respectivo subconjunto, assim como os valores para prevalências e efeito do plano amostral.

Além dessa estratificação geográfica mais agregada, foram formados estratos geográficos mais finos, buscando assegurar um espalhamento da amostra no território do estado. A estratificação da amostra foi concluída subdividindo os estratos geográficos da Capital (João Pessoa) em estratos de renda, usando pontos de corte para delimitar os quartis da variável “Valor do rendimento nominal médio mensal das pessoas de 10 anos ou mais de idade (com rendimento)” para cada setor. Com isso, a amostra em cada um dos estratos de seleção, foi formada pela combinação do estrato natural, o estrato geográfico e o estrato de renda. Todas as estratificações estão descritas na Tabela 2.

Tabela 2 – Estratos geográficos detalhados considerados para PB-COVID19.

| Estrato natural | Região de saúde | Estrato geográfico | Setores | Domicílios | Pessoas | Setores na amostra por mês | Fração amostral setores (%) | Pessoas na amostra por mês | Fração amostral pessoas (por mil) |
|---|---|---|---|---|---|---|---|---|---|
| João Pessoa | 1 | João Pessoa.1 | 166 | 35.234 | 123.553 | 8 | 4,8 | 200 | 1,6 |
| | 1 | João Pessoa.2 | 159 | 39.343 | 137.610 | 8 | 5,0 | 200 | 1,5 |
| | 1 | João Pessoa.3 | 182 | 43.464 | 148.357 | 8 | 4,4 | 200 | 1,3 |
| | 1 | João Pessoa.4 | 155 | 33.048 | 112.713 | 8 | 5,2 | 200 | 1,8 |
| | 1 | João Pessoa.5 | 283 | 62.141 | 196.511 | 8 | 2,8 | 200 | 1,0 |
| Macro 1 sem João Pessoa | 1 | Macro 1 sem João Pessoa.1 | 647 | 128.537 | 456.512 | 16 | 2,5 | 400 | 0,9 |
| | 2 | Macro 1 sem João Pessoa.2 | 471 | 83.651 | 294.706 | 12 | 2,5 | 300 | 1,0 |
| | 12 | Macro 1 sem João Pessoa.12 | 229 | 48.485 | 168.910 | 8 | 3,5 | 200 | 1,2 |
| | 14 | Macro 1 sem João Pessoa.14 | 219 | 37.680 | 134.424 | 8 | 3,7 | 200 | 1,5 |
| Macro 2 | 3 | Macro 2.3 | 286 | 53.125 | 187.313 | 8 | 2,8 | 200 | 1,1 |
| | 4 | Macro 2.4 | 177 | 31.175 | 105.882 | 8 | 4,5 | 200 | 1,9 |
| | 5 | Macro 2.5 | 189 | 33.978 | 106.675 | 8 | 4,2 | 200 | 1,9 |
| | 15 | Macro 2.15 | 239 | 41 | 144.511 | 8 | 3,3 | 200 | 1,4 |
| | 16 | Macro 2.16 | 685 | 149.616 | 516.239 | 16 | 2,3 | 400 | 0,8 |
| Macro 3 | 6 | Macro 3.6 | 363 | 63.895 | 223.416 | 12 | 3,3 | 300 | 1,3 |
| | 7 | Macro 3.7 | 216 | 40.740 | 146.150 | 8 | 3,7 | 200 | 1,4 |
| | 8 | Macro 3.8 | 172 | 30.999 | 110.811 | 8 | 4,7 | 200 | 1,8 |
| | 9 | Macro 3.9 | 250 | 48.206 | 167.424 | 8 | 3,2 | 200 | 1,2 |
| | 10 | Macro 3.10 | 143 | 31.724 | 110.381 | 8 | 5,6 | 200 | 1,8 |
| | 11 | Macro 3.11 | 141 | 22.019 | 80.866 | 8 | 5,7 | 200 | 2,5 |
| | 13 | Macro 3.13 | 99 | 16.735 | 58.432 | 8 | 8,1 | 200 | 3,4 |
| Total Geral | | | | | | 192 | 3,5 | 4.800 | 1,3 |

Fonte: Silva et al. (2020).

Com base nas informações do Censo Demográfico de 2010, os pesquisadores foram orientados a percorrer todo o setor censitário de acordo com o percurso de mapas fornecidos pelo Instituto Brasileiro de Geografia e Estatística (IBGE) em sua Base Operacional Geográfica. No caso do endereço sendo considerado residencial, isto é, um domicilio particular permanente, o dispositivo móvel efetuou um sorteio de um número aleatório com o intuito de fazer com que o domicilio deva ou não ser incluído na amostra, em caso de ser incluído os pesquisadores abordam visando obtenção da entrevista e aplicação dos testes nos indivíduos participantes. Para mais detalhes sobre a Pesquisa Sorológica Continuar Cuidando, deve-se consultar Silva et al. (2020).

## 6.2 ESTIMAÇÕES DE TOTAIS

O processo de calibração dos pesos amostrais relacionados a pesquisa foram: calcular os pesos básicos para cada domicilio e os seus moradores mediante o plano amostral considerado. Os pesos básicos foram calibrados segundo informações referentes ao número de habitantes do Estado da Paraíba e seus respectivos estratos conforme descrição da

pesquisa.

Desta forma, o estimador de Horvitz-Thompson para o total populacional para os casos de COVID-19 para o Estado da Paraíba é dado por:

$$\widehat{t} = \sum_{h \in H} \sum_{i \in S_I} \sum_{j \in S_i} d_{hij} y_{hij},$$

sendo $y_{hij}$ o número de pessoas que testaram positivo para a presença do anticorpo IgG no domicílio $j$ selecionado no setor censitário $i$ do estrato $h$. O peso básico $d_{hij}$ que refere-se a um domicilio $j$ selecionado no setor censitário $i$ do estrato $h$ é:

$$d_{hij} = \left( m_h \frac{t_{hi}}{T_h} \right)^{-1} \times (\pi_{hij})^{-1} \times \left( \frac{n_{hij}}{n_{hi}} \right), \tag{6.1}$$

da equação (6.1), tem-se que

- $m_h$ : Identifica o número de setores censitários selecionados no estrato $h$;
- $t_{hi}$ : Identifica a medida de tamanho do setor censitário $i$ pertencente ao estrato $h$;
- $T_{hi}$ : Identifica o somatório das medidas de tamanho dos setores censitários pertencentes ao estrato $h$;
- $\pi_{hij}$ : Identifica a probabilidade de inclusão dos domicílios no setor censitário conforme empregada na amostragem de Bernoulli para cada setor;
- $n_{hi}$ : Identifica o número de domicílios que são elegíveis dentre os domicílios que foram selecionados no setor censitário $i$ do estrato $h$;
- $n_{hij}$ : Identifica o número $j$ de domicílios elegíveis que foram entrevistados no setor censitário $i$ do estrato $h$.

A terceira parcela da expressão corresponde a um fator de correção atribuída aos domicílios que foram selecionados podem ser inelegíveis, assim sendo o produto das probabilidades traria como resultado a probabilidade de inclusão na amostra do domicilio.

# 7 SIMULAÇÃO

Neste capítulo é descrito como foi feita a reamostragem dos dados para a obtenção da população artificial e cálculo das estimativas através de simulação de Monte Carlo. A população artificial foi criada a partir dos dados amostrais da Pesquisa Sorológica Continuar Cuidando (PSCC). O estudo de simulação comparou as estimativas de prevalências da COVID-19 por meio de quatro planos amostrais de dois estágios, sendo eles: Pareto com inversa de Bernoulli, Pareto com sistemática, Sequencial Poisson com inversa de Bernoulli e Sequencial Poisson com sistemática.

## 7.1 GERAÇÃO DE POPULAÇÃO ARTIFICIAL

Criada a partir do conjunto de dados amostrais da PSCC, a população artificial, denominada de POP19, foi elaborada com a finalidade de dar suporte ao estudo de simulação de Monte Carlo. O nome POP19 foi atribuido por remeter a COVID-19.

A estrutura da POP19 manteve os moldes da população real de moradores do estado da Paraíba, garantindo a hierarquia sucessiva de codificação de municípios, setores censitários, domicílios e seus respectivos moradores. A POP19 possui 344.161 linhas onde cada uma é uma identificação de um residente, e 12 variáveis descritas na Tabela 3.

A POP19 possui um total de 384 setores censitários que correspondem aos setores selecionados na PSCC, e 89.578 domicílios que correspondem ao total de domicílios nesses setores de acordo com o Censo Demográfico de 2010.

Tabela 3 – Descrição das variáveis de POP19.

| Variáveis de POP19 | |
|---|---|
| Variável | Descrição |
| cod_municipio | Identificação para o código do município. |
| cod_setor | Identificação para o código do setor censitário. |
| a06_domicílio | Identificação do domicílio dentro do setor censitário. |
| NumberHits | Número de vezes que o domicílio foi selecionado em reposição. |
| b02_sexo | 1 (Masculino) e 2 (Feminino). |
| b04_idade | Identificação para a idade do morador entrevistado. |
| resp_igg_tr | 1 (Positivo), 2 (Negativo), 3 (Inconclusivo) e 4 (Não testado) |
| resp_igm_tr | 1 (Positivo), 2 (Negativo), 3 (Inconclusivo) e 4 (Não testado) |
| rep | Idetificação da réplica. |
| elegível | 1 (elegível) e 0 (não elegível). |
| domicílio | Identificador único do domicílio. |

Fonte: Silva et al. (2020).

Sobre a POP19, segue os seguintes comentários:

- Para identificação única de um determinado domicilio da POP19, a mesma deve ser feita com base na variável **domicílio**, a mesma é a correspondência da hierarquia de setor censitário (cod-setor), domicilio selecionado (a-06-domicilio) e réplica(rep);
- A codificação para a variável **sexo** (b02-sexo) atribui dois valores, sendo: 1 (Masculino) e 2 (Feminino). Esta codificação é útil para situação de perca da formatação da população real;
- A variável **elegível** trata-se de realizações independentes da distribuição de probabilidade Bernoulli, atribuindo sucesso (ser elegível) igual a 0.8. A sua codificação é então: 1 (elegível) e 0 (não elegível). Esta variável foi criada para simular uma situação de varredura de campo, uma vez que o conceito de elegibilidade é atribuído conforme o contexto da pesquisa;
- As variáveis **resp_igg_tr** e **resp_igm_tr** são a identificação para os resultados dos testes rápidos do IgG e IgM, respectivamente.

## Construção da POP19

Para o processo de construção da população artificial POP19, utilizou-se a base de dados da PSCC, que possui um total de 10.872 observações (pessoas) e 190 variáveis.

Os 389 setores censitários listados foram identificados e seus respectivos tamanhos populacionais (número de domicílios) resgatados a partir do Censo Demográfico de 2010. Verificou-se que tais setores possuem um total de 89.834 domicílios.

Os 3.192 domicílios listados da PSCC foram identificados e separados em um arquivo para formar a base de geração dos dados da população a partir de um processo de reamostragem com reposição, considerando os setores censitários como os estratos.

Dentro de cada um dos 389 setores censitários dos dados amostrais, uma amostra aleatória de domicílios foi extraída com reposição e de tamanho igual ao número de domicílios daquele setor na população.

A título ilustrativo, considere um setor censitário S1, que tenha listado na amostra 10 domicílios, e que com base no censo de 2010, S1 tenha 100 domicílios na população. Então, dentro de S1 uma amostra aleatória de tamanho 100 é retirada com reposição do conjunto de 10 domicílios listados na amostra.

Por fim, este procedimento descrito de reamostragem gerou para a população artificial um total de 89.578 domicílios, uma quantidade ligeiramente menor do que os 89.834 domicílios identificados no censo.

## Parâmetros da população POP19

Verifica-se que a POP19 possui 57,7% de pessoas do sexo feminino e 42,3% de pessoas do sexo masculino, como é mostrado na Tabela 4.

Tabela 4 – Frequências de sexo em POP19.

| Sexo | Frequência | Porcentagem | Frequência acumulada | Porcentagem acumulada |
|---|---|---|---|---|
| Masculino | 145.559 | 42,30 | 145.559 | 42,30 |
| Feminino | 198.602 | 57,70 | 344.161 | 100,00 |

Fonte: O Autor.

A Tabela 5 resume informações de idade por sexo em POP19.

Tabela 5 – Estatísticas de idade por sexo em POP19.

| Idade | | | |
|---|---|---|---|
| Sexo | Observações | Média | Desvio padrão |
| Masculino | 145.559 | 37,70 | 22,43 |
| Feminino | 198.602 | 39,24 | 22,23 |

Fonte: O Autor.

As Tabelas 6 e 7 mostram parâmetros de resultados de testes rápidos de IgG e IgM. Tem-se 9,10% de casos positivos e 84,44% de casos negativos para as respostas do teste rápido do IgG. Para o IgM tem-se 3,40% de casos positivos e 90,10% de casos negativos.

Tabela 6 – Frequência de respostas ao teste rápido de IgG em POP19.

| Resposta ao teste rápido de IgG | | | | |
|---|---|---|---|---|
| Categoria | Frequência | Porcentagem | Frequência acumulada | Porcentagem acumulada |
| Positivo | 31.323 | 9,10 | 31.323 | 9,10 |
| Negativo | 290.611 | 84,44 | 321.934 | 93,54 |
| Inconclusivo | 51 | 0,01 | 321.985 | 93,56 |
| Não testado | 22.176 | 6,44 | 344.161 | 100,00 |

Fonte: O Autor.

Tabela 7 – Frequência de respostas ao teste rápido de IgM em POP19.

| Resposta ao teste rápido de IgM | | | | |
|---|---|---|---|---|
| Categoria | Frequência | Porcentagem | Frequência acumulada | Porcentagem acumulada |
| Positivo | 11.755 | 3,42 | 11.755 | 3,42 |
| Negativo | 310.102 | 90,10 | 321.857 | 93,52 |
| Inconclusivo | 128 | 0,04 | 321.985 | 93,56 |
| Não testado | 22.176 | 6,44 | 344.161 | 100,00 |

Fonte: O Autor.

## 7.2 SIMULAÇÃO DE MONTE CARLO

A simulação de Monte Carlo envolveu selecionar amostras da população POP19 de acordo com cada um dos planos amostrais em dois estágios considerados. Ao todo foram selecionadas $B = 5.000$ amostras (réplicas).

Os planos amostrais foram comparados a partir de medidas de viés, viés relativo, erro padrão e da raiz do erro quadrático médio relativo do estimador de HT. Considere $\hat{\theta}_i$ como a estimativa observada pelo estimador de HT para o parâmetro $\theta$ na $i$-ésima réplica de Monte Carlo. O viés de cada plano amostral $p(.)$ é calculado da seguinte forma:

$$\text{Viés}_p(\widehat{\theta}) = \sum_{i=1}^{B} \frac{\hat{\theta}_i}{B} - \theta.$$

A expressão para o viés relativo é dada por

$$\text{Viés Relativo}_p(\widehat{\theta}) = \text{Viés}_p(\widehat{\theta})/\theta.$$

A raiz do erro quadrático médio relativo (REQMR) foi calculada usando a seguinte expressão:

$$\text{REQMR}_p(\widehat{\theta}) = \sqrt{EQM_p(\widehat{\theta})}/\theta,$$

onde $EQM_p(\widehat{\theta}) = \frac{1}{B-1}\sum_{i=1}^{B}[\hat{\theta}_i - \theta]^2$.

Por fim, o erro padrão foi calculado como:

$$\text{Erro Padrão}_p(\widehat{\theta}) = \sqrt{\frac{1}{B-1}\sum_{i=1}^{B}[\widehat{\theta}_i - \overline{\theta}]^2},$$

onde $\overline{\theta} = \sum_{i=1}^{B}\hat{\theta}_i/B$.

## 7.3 RESULTADOS

Nesta seção serão apresentados os resultados das 5000 réplicas da simulação de Monte Carlo, para as estimativas de totais populacionais relativos aos resultados dos testes rápidos do IgG e IgM, seguindo os planos amostrais abordados. A Tabela 8 apresenta os seguintes resultados:

- Estimativas de totais populacionais para os testes rápidos do IgG e IgM obtidas da simulação de 5000 réplicas;
- Estimativas do viés para os testes rápidos do IgG e IgM obtidas da simulação de 5000 réplicas;
- Estimativas do viés relativo para os testes rápidos do IgG e IgM obtidas da simulação de 5000 réplicas.

Tabela 8 – Estimativas para os totais populacionais, para o viés e o viés relativo dos testes rápidos de IgG e IgM baseado nas 5000 réplicas da simulação de Monte Carlo.

| Teste | Resultado | Valor do Parâmetro | Valores Estimados | Pareto Inversa | Pareto Sistemática | Poisson Inversa | Poisson Sistemática |
|---|---|---|---|---|---|---|---|
| Igg | Positivo | 31.323 | Estimativa | 31.121,6 | 31.429.2 | 31.140,8 | 31.253,8 |
| | | | Viés | -201,4 | 106,2 | -182,2 | -69,2 |
| | | | Viés Relativo | -0,00643 | 0,00339 | -0,00582 | -0,00221 |
| | Negativo | 290.611 | Estimativa | 290.650,2 | 290.759,8 | 290.610,0 | 290.699 |
| | | | Viés | 39,2 | 148,8 | -1 | 88 |
| | | | Viés Relativo | 0,00013 | 0,00051 | 0 | 0,00003 |
| | Inconclusivo | 51 | Estimativa | 46,0 | 52,2 | 48,0 | 49,7 |
| | | | Viés | -5 | 1,2 | -3 | -1,3 |
| | | | Viés Relativo | -0,09804 | 0,02353 | -0,05882 | -0,02549 |
| | Não testado | 22.176 | Estimativa | 22.152,1 | 22.135,6 | 22.122,9 | 22.160,4 |
| | | | Viés | -23,9 | -40,4 | -53,1 | -15,6 |
| | | | Viés Relativo | -0,00108 | -0,00182 | -0,00239 | 0,00001 |
| Igm | Positivo | 11.755 | Estimativa | 11.756,2 | 11.831,4 | 11.734,7 | 11695.1 |
| | | | Viés | 1,2 | 76,4 | -20,3 | -59,9 |
| | | | Viés Relativo | 0,00001 | 0,0065 | -0,00173 | -0,0051 |
| | Negativo | 310.102 | Estimativa | 309.935,6 | 310.278,5 | 309.936,5 | 310.182,3 |
| | | | Viés | -166,4 | 176,5 | -165,5 | 80,3 |
| | | | Viés Relativo | -0,00054 | 0,00057 | -0,00053 | 0,00026 |
| | Inconclusivo | 128 | Estimativa | 125,9 | 131,3 | 127,5 | 125,1 |
| | | | Viés | -2,1 | 3,3 | -0,5 | -2,9 |
| | | | Viés Relativo | -0,01641 | 0,02578 | -0,00391 | -0,02266 |
| | Não testado | 22.176 | Estimativa | 22.152,1 | 22.135,6 | 22.122,9 | 22.160,4 |
| | | | Viés | -23,9 | -40,4 | -53,1 | -15,6 |
| | | | Viés Relativo | -0,00108 | -0,00182 | -0,00239 | 0,00001 |

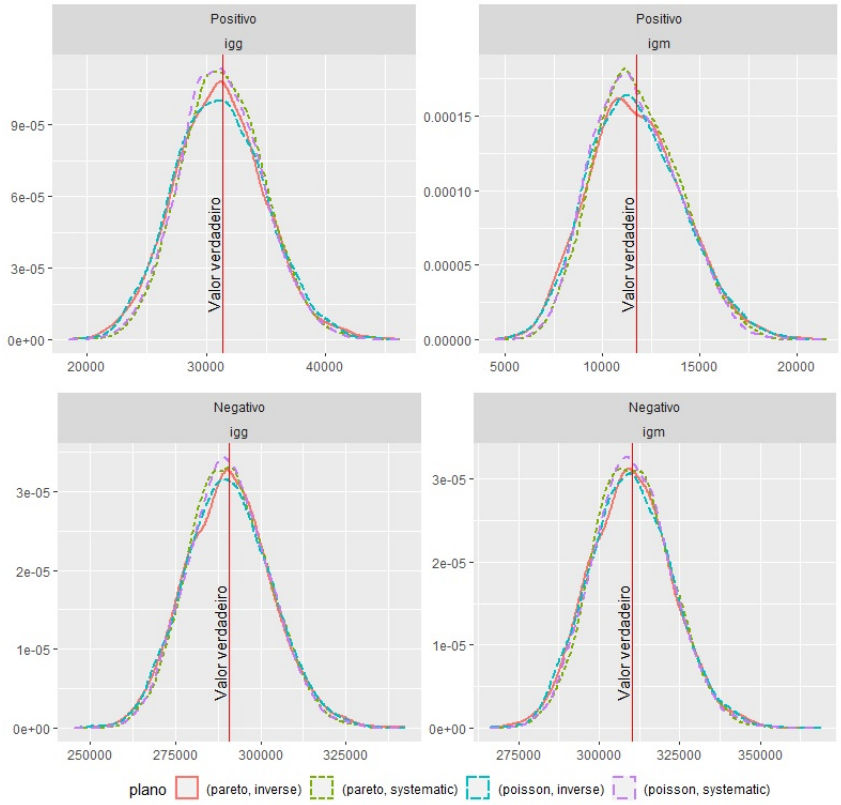
Fonte: O Autor.

Conforme a Tabela 8, verificou-se para os resultados do teste do IgG, que a magnitude do viés, foi menor nos planos com amostragem inversa de Bernoulli, e no segundo estágio como visto nos resultados Negativos, Inconclusivo, e para os casos não testados. Contudo, ao comparar o viés relativo, percebe-se, que não há grandes diferenças entre os planos amostrais no estudo comparativo para o teste do IgG. Análise similar ocorre no teste do IgM, onde verificam-se menores magnitudes de viés para os planos com amostragem inversa de Bernoulli nos resultados Positivos, inconclusivos e Não testados. Mas ao tomar

o viés relativo, tem-se planos que se assemelham aos resultados.

A Figura 2 apresenta a distribuição amostral dos estimadores de totais populacionais quando comparamos os quatros planos amostrais testados para os resultados positivo e negativos dos testes para o IgG e IgM.

Figura 2 – Distribuição das estimativas dos totais populacionais para os resultados positivo e negativo dos testes rápidos do IgG e IgM com base na 5000 réplicas de simulação de Monte Carlo.




Fonte: O Autor.

A Tabela 9 apresenta as seguintes estimativas:

- Estimativas para o desvios-padrão para os testes rápidos do IgG e IgM obtidos das 5000 réplicas da simulação de Monte Carlo;
- Estimativas para o REQMR para os testes rápidos do IgG e IgM obtidos das 5000 réplicas da simulação de Monte Carlo.

Tabela 9 – Estimativas dos erros-padrão e REQMR dos testes rápidos de IgG e IgM baseado nas 5000 réplicas da simulação de Monte Carlo.

| Teste | Resultado | Valor do Parâmetro | Valores Estimados | Pareto Inversa | Pareto Sistemática | Poisson Inversa | Poisson Sistemática |
|---|---|---|---|---|---|---|---|
| Igg | Positivo | 31.323 | Erro padrão | 4.419,7 | 4.076,06 | 4.418,46 | 4.042,76 |
| | | | REQMR | 0,1412 | 0,1301 | 0,1411 | 0,1290 |
| | Negativo | 290.611 | Erro padrão | 14.846,86 | 14.032,54 | 14.868,73 | 14.035,46 |
| | | | REQMR | 0,05109 | 0,01404 | 0,05116 | 0,04830 |
| | Inconclusivo | 51 | Erro padrão | 87,81 | 81,54 | 91,04 | 79,05 |
| | | | REQMR | 1,72471 | 1,59914 | 1,78614 | 1,55022 |
| | Não testado | 22176 | Erro padrão | 3.158,72 | 2.887,19 | 3.158,24 | 2.886,82 |
| | | | REQMR | 0,14244 | 0,13021 | 0,14244 | 0,13018 |
| Igm | Positivo | 11.755 | Erro padrão | 2.837,81 | 2.629,31 | 2.836,31 | 2.574,57 |
| | | | REQMR | 0,24141 | 0,22377 | 0,24129 | 0,21908 |
| | Negativo | 310.102 | Erro padrão | 15.324,7 | 14.502,97 | 15.351,74 | 14.510,75 |
| | | | REQMR | 0,04942 | 0,04677 | 0,04951 | 0,04679 |
| | Inconclusivo | 128 | Erro padrão | 192,05 | 178,24 | 194,36 | 173,65 |
| | | | REQMR | 1,50054 | 1,39279 | 1,51851 | 1,35685 |
| | Não testado | 22.176 | Erro padrão | 3.158,72 | 2.887,19 | 3.158,24 | 2.886,82 |
| | | | REQMR | 0,14244 | 0,13021 | 0,14244 | 0,13018 |

Fonte: O Autor.

Analisando a Tabela 9 verificam-se magnitudes maiores para o erro padrão nos planos com amostragem inversa de Bernoulli em segundo estágio, para todos os resultados dos testes do IgG e do IgM. Porém, considerando o REQMR, percebe-se que não há diferença de desempenho entre os planos amostrais considerados.

Para avaliar a eficiência da amostragem por conglomerados em dois estágios com amostragem inversa de Bernoulli em segundo estágio, comparado à amostragem sistemática em segundo estágio, foi utilizado o Efeito do Plano Amostral (EPA). Esta medida permite analisar a eficiência de uma estratégia de estimação, tomando como base outra estratégia, comparando as variâncias dos respectivos planos amostrais.

A Tabela (10) apresenta os valores calculados para o EPA tomando como base o plano de amostragem Sequencial de Poisson em primeiro estágio e amostragem inversa de Bernoulli em segundo estágio.

Tabela 10 – Comparação de eficiência dos resultados dos testes rápidos de IgG e IgM dos planos amostrais.

| Teste | Resultado | Pareto Inversa | Pareto Sistemática | Poisson Sistemática |
|---|---|---|---|---|
| IgG | Positivo | 1,0012 | 0,8344 | 0,8330 |
| | Negativo | 0,9804 | 0,8942 | 0,9011 |
| | Inconclusivo | 1,2364 | 0,6944 | 0,7178 |
| | Não testado | 1,0273 | 0,8254 | 0,8216 |
| IgM | Positivo | 1,0541 | 0,9035 | 0,8775 |
| | Negativo | 0,9824 | 0,8922 | 0,8992 |
| | Inconclusivo | 1,0028 | 0,9035 | 0,8982 |
| | Não testado | 1,0273 | 0,8254 | 0,8216 |

Fonte: O Autor.

Observa-se que boa parte dos valores do EPA são menores do que 1 para ambos os testes. Dos 24 valores calculados, verificou-se que 13 apresentaram um ganho de eficiência de pouco mais de 10%. Para este caso, temos indícios de que para uma parte das variáveis do estudo, o uso da amostragem inversa de Bernoulli acarreta uma pequena perda de eficiência em comparação com a amostragem sistemática.

# 8 CONCLUSÕES GERAIS

A falta de um cadastro que permita a identificação das unidades da população-alvo é uma recorrente adversidade enfrentada em levantamentos amostrais. Investir tempo para que se possa atualizar o cadastro, algo necessário, pode não ser uma opção viável quando pesquisas necessitam ser realizadas com urgência, como pôde ser visto na pandemia da COVID-19, em que dada a velocidade de contágio do vírus SARS-CoV-2, as pesquisas sorológicas precisavam dar respostas em um tempo bastante reduzido. Em razão disto, a aplicação de planos amostrais que permitam selecionar amostras e atualizar cadastros de forma simultânea é de grande utilidade.

Esta dissertação se propôs a analisar o plano de Amostragem inversa de Bernoulli, um plano com grande potencial para ser utilizado no segundo estágio de planos amostrais de conglomerados em dois estágios, devido a sua capacidade de selecionar uma amostra durante o processo de atualização do cadastro. O desempenho da amostragem inversa de Bernoulli em segundo estágio, combinado com o uso da Amostragem de Pareto ou Amostragem Sequencial de Poisson no primeiro estágio, foi estudado de maneira comparativa com a amostragem Sistemática em segundo estágio, através de simulação de Monte Carlo. O estudo comparativo utilizou uma população artificial (POP19) construída a partir dos dados da Pesquisa Sorológica Continuar Cuidando, realizada no Estado da Paraíba para monitorar o avanço da epidemia da COVID-19. Foi verificado que as estimativas pontuais para os resultados dos testes de contaminação do vírus SARS-CoV-2 através dos planos estudados por esta dissertação foram bem próximas dos valores da população artifical, como era de se epserar. A amostragem inversa de Bernoulli mostrou ter grande potencial de aplicação em segundo estágio por apresentar estimativas e medidas de desempenho tão boas quanto as dos planos concorrentes, sendo ela preferível por seu baixo custo de implementação e capacidade de realização de uma amostra simultaneamente ao trabalho de atualização do cadastro.

# REFERÊNCIAS

AIRES, N.; ROSÉN, B. On inclusion probabilities and relative estimator bias for pareto $\pi$ps sampling. *Journal of statistical Planning and inference*, Elsevier, v. 128, n. 2, p. 543–567, 2005.

BUSSAB, W.; ANDRADE, O.; SILVA, P. Plano amostral saeb-99: definição do plano amostral. *Relatório técnico, INEP/MEC*, 1999.

COSTA, G. T. L. da. Coordenação de amostras ppt em pesquisas repetidas, utilizando o método de amostragem de pareto. 2007.

FREITAS, M. P. S. de; ANTONACI, G. de A. Amostra mestra 2010 e amostra da pnad contínua. 2014.

HORVITZ, D. G.; THOMPSON, D. J. A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association*, Taylor & Francis, v. 47, n. 260, p. 663–685, 1952.

OHLSSON, E. Sequential poisson sampling. *Journal of official Statistics*, Statistics Sweden (SCB), v. 14, n. 2, p. 149, 1998.

ROSÉN, B. On inclusion probabilities for order $\pi$ps sampling. *Journal of statistical planning and inference*, Elsevier, v. 90, n. 1, p. 117–143, 2000.

SÄRNDAL, C.-E.; SWENSSON, B.; WRETMAN, J. *Model assisted survey sampling*. [S.l.]: Springer Science & Business Media, 1992.

SILVA, P.; BIANCHINI, Z.; DIAS, A. *Amostragem: Teoria e prática usando R*. [S.l.]: Github, 2021.

SILVA, P.; VASCONCELLOS, M.; COELHO, H.; FERRAZ, C. Plano amostral da pesquisa pb-covid19: definição do plano amostral. 2020.