UNIVERSIDADE FEDERAL DE PERNAMBUCO
CENTRO DE CIÊNCIAS EXATAS E DA NATUREZA
PROGRAMA DE PÓS-GRADUAÇÃO EM ESTATÍSTICA

JOSÉ CARLOS DA SILVA MELO

AMOSTRAGEM DE BERNOULLI

Recife
2020

JOSÉ CARLOS DA SILVA MELO

AMOSTRAGEM DE BERNOULLI

Este trabalho foi apresentado à Pós-graduação em Estatística do Centro de Ciências Exatas e da Natureza da Universidade Federal de Pernambuco como requisito parcial para obtenção do grau de Mestre em Estatística.

Área de Concentração: Estatística Aplicada

Orientador: Prof. Dr. Cristiano Ferraz

Recife
2020

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

M528a Melo, José Carlos da Silva
Amostragem de Bernoulli / José Carlos da Silva Melo. – 2020.
54 f.: il., tab.

Orientador: Cristiano Ferraz.
Dissertação (Mestrado) – Universidade Federal de Pernambuco. CCEN, Estatística, Recife, 2020.
Inclui referências e apêndices.

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

310 CDD (23. ed.) UFPE- CCEN 2021 - 31

JOSE CARLOS DA SILVA MELO

AMOSTRAGEM DE BERNOULLI

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: 27 DE FEVEREIRO DE 2020

BANCA EXAMINADORA

Prof.(º) Cristiano Ferraz
UFPE

Prof.(º) Raydonal Ospina Martínez
UFPE

Prof.(º) Hemílio Fernandes Campos Coêlho
UFPB

À minha amada, Aline Romão, e aos meus pais, Anaíde e Carlos.

Agradecimentos

Agradeço a Deus, por tudo que Ele fez e faz por mim.

À minha namorada, Aline Romão, por me acompanhar durante todo esse tempo, me apoiando, me incentivando, querendo sempre o meu melhor. Se não fosse pela força que ela me dá, eu não estaria onde estou hoje. Por sempre estar presente quando as coisas não vão bem, pelo afago, carinho, atenção e paciência. Por todos os anos que já vivi ao seu lado, sempre me dando amor e me mostrando como é ser verdadeiramente feliz.

Aos meus pais, Anaíde e Carlos, pela compreensão e pelo apoio, que mesmo sem entender muita coisa do que faço, sempre estão lá para me incentivar.

Ao meu orientador, Cristiano Ferraz, pela mão estendida e pela oportunidade dada num momento tão crítico da minha vida. Pelo seu caráter, seu incentivo, sua paciência e sua competência como professor.

Aos meus quase irmãos, Adenice, Diogo e Lukas, que sempre estão presentes na minha vida e me dando apoio e muitos momentos de alegria.

Aos meus amigos, André, Pedro e Jonas, pelo ano que passamos dividindo o apartamento e compartilhando os sofrimentos e os momentos de alegria vividos na pós-graduação.

Aos amigos que fiz no mestrado, pelo curto tempo que passamos juntos. Pela convivência, dores de cabeça e paciência. Em especial Cesar, que me ajudou em um dos momentos mais desesperadores.

A Valeria Bittencourt, por ser a melhor secretária que já conheci. Pelo carinho e pela sua competência.

Aos professores do Programa de Pós-Graduação em Estatística da Universidade Federal de Pernambuco, pelo ótimo trabalho prestado e dedicação para formar ótimos profissionais.

Por todos que aqui não mencionei, mas marcaram minha vida de algum modo.

À banca examinadora pelas sugestões relevantes que engradeceram e enriqueceram este trabalho.

À CAPES, pelo apoio financeiro.

"How many kinds are there?"
(GOODMAN, 1949)

Resumo

Comumente utilizada na literatura, a amostragem de Bernoulli consiste em um plano amostral para selecionar elementos de uma população finita, de tamanho N, por meio de experimentos de Bernoulli com probabilidade conhecida de sucesso, caracterizado usualmente como inclusão do elemento na amostra. Experimentos independentes são realizados, um para cada elemento, todos com mesma probabilidade de seleção, π. O tamanho de amostra de um plano amostral de Bernoulli (BE) é uma variável aleatória com distribuição Binomial de parâmetros N e π. Proposto originalmente por Goodman (1949) e com várias possibilidades de aplicações, este plano é pouco explorado em cursos de amostragem, em diferentes níveis, de graduação e pós-graduação, por limitação de tempo. Essa dissertação apresenta uma revisão de literatura sobre este plano amostral, ressaltando a sua importância na área de amostragem, com ênfase em seus aspectos históricos e leque de aplicações. Estudos de simulação de Monte Carlo foram realizados para comparar a eficiência estatística de estimadores de total populacional encontrados na literatura, com base em um plano de BE, sob cenários práticos, incluindo estratificação.

Palavras-chave: Amostragem Binomial. Estratificação. Estimador de Horvitz-Thompson.

Abstract

Commonly used in the literature, Bernoulli sampling is a sampling design to select elements from a finite population of size N, using Bernoulli experiments with known probability of success related to the inclusion of the element in the sample. Independent experiments are run, one for each element, all with the same selection probability, π. The sample size of a Bernoulli design (BE) is a random variable with Binomial distribution of parameters N and π. Originally proposed by Goodman (1949), and with many possibilities of application, this design is underexplored in sampling courses, in undergraduate and graduate levels, for time constraints. This dissertation presents a literature review about Bernoulli sampling design, highlighting its importance in sampling, with emphasis on its historical aspects and range of applications. Monte Carlo simulation studies were performed in order to compare the statistical efficiency of estimators of total population found in the literature, based on the BE design, in many practical scenarios, including stratification.

Keywords: Binomial Sampling. Stratification. Horvitz-Thompson Estimator.

Lista de Tabelas

Tabela 1 — Erro quadrático médio dos estimadores $\hat{t}_\pi$ e $\hat{Y}_1$ utilizando um plano de BE 37
Tabela 2 — Viés relativo dos estimadores $\hat{t}_\pi$ e $\hat{Y}_1$ utilizando um plano de BE 38
Tabela 3 — Coeficiente de variação dos estimadores $\hat{t}_\pi$ e $\hat{Y}_1$ utilizando um plano de BE 39
Tabela 4 — Erro quadrático médio para os estimadores $\hat{t}_\pi$ e $\hat{Y}_1$ aplicados a uma amostragem estratificada de BE 41
Tabela 5 — Viés relativo dos estimadores $\hat{t}_\pi$ e $\hat{Y}_1$ aplicados a uma amostragem estratificada de BE 41
Tabela 6 — Coeficiente de variação dos estimadores $\hat{t}_\pi$ e $\hat{Y}_1$ aplicados a uma amostragem estratificada de BE 42

Sumário

1 Introdução 11
2 História e Aplicações 13
2.1 Revisão histórica 13
2.2 Aplicações envolvendo o plano de BE 19
3 Amostragem de Bernoulli 25
3.1 Plano amostral 25
3.2 Estimação pontual 26
3.3 Estimação intervalar 28
3.4 Amostragem estratificada 29
4 Estudo Computacional 31
4.1 Estimadores 31
4.2 Cenários de simulação 33
4.3 Resultados 34
4.3.1 Resultados para a Amostragem de Bernoulli 36
4.3.2 Resultados para a Amostragem Estratificada de Bernoulli 40
5 Conclusão Geral 43
Referências 45
Apêndice A — Script para o plano BE 48
Apêndice B — Script para o plano BE estratificado 51

1 Introdução

A amostragem de Bernoulli consiste em associar cada elemento da população a um experimento probabilístico independente, cujo resultado define a inclusão ou não na amostra. O elemento é selecionado para a amostra se o resultado do experimento for considerado sucesso, o que ocorre com probabilidade fixa, conhecida e determinada previamente em acordo com o planejado. O tamanho amostral pode ser expresso como uma soma de variáveis aleatórias independentes, que seguem uma distribuição de Bernoulli de parâmetro π, que remete a probabilidade de inclusão amostral.

Apesar do sobrenome de referência em sua nomenclatura, o plano amostral de Bernoulli (BE) foi originalmente apresentado por Goodman (1949). Em seu estudo, Goodman demonstra como obter uma amostra a partir do plano e apresenta um estimador, que se adaptado para o estimador do total populacional, assume a forma do estimador de Horvitz-Thompson para amostragem de BE.

Com base nas referências consultadas nesta dissertação, o termo "Amostragem de Bernoulli" só foi efetivamente utilizado pelo pesquisador Beckwith (1973), quando este autor abordou um estudo do limite do tamanho amostral em um plano de BE e em um plano de BE modificado. O plano modificado retratado pelo autor equivale a amostragem de Poisson, vindo a ser um plano de BE com probabilidade de inclusão variável.

Com o passar dos anos o plano ganhou mais relevância, como pode ser visto em Strand (1979), que abordou o estudo sobre a amostragem de BE no processo de estimação do total populacional, investigando dois estimadores, que são comparados à amostragem aleatória simples por meio do erro quadrático médio. O autor apresenta situações em que os estimadores do plano de BE têm quase o mesmo erro quadrático médio que o estimador da amostragem aleatória simples.

Vários pesquisadores desenvolveram aplicações diretas (BE eleita como instrumento

da amostragem) ou indiretas (BE eleita para avaliar computacionalmente algum método proposto) envolvendo o método de BE. Cohen (1975) e Haas e König (2004) estão entre os autores cujos trabalhos são exemplos de aplicação direta do plano de BE. Cohen (1975) utilizou o plano de BE para testar e compreender modelos quantitativos e a composição demográfica de grupos sociais de espécies de orangotangos selvagens. Haas e König (2004) utilizaram um esquema de amostragem de dois níveis de BE, com o intuito de aprimorar o processo de consultas a banco de dados, devido à dimensão das informações.

Em outros casos, o plano de BE foi usado indiretamente com a finalidade de avaliar algum método computacionalmente proposto. Muitos são os exemplos nessa categoria. Dois deles, para exemplificar, podem ser conferidos em Duchesne (2003) e Rondon *et al.* (2012). Duchesne (2003) fez uso do plano de BE para avaliar estimadores para a proporção, utilizando estimadores de regressão. Rondon *et al.* (2012) realizaram uma aplicação envolvendo estimadores de regressão e o plano de BE, sendo este último usado para auxiliar na avaliação dos estimadores.

Essa dissertação tem por objetivos: (i) apresentar uma revisão de literatura que resgate a origem histórica do plano amostral de Bernoulli; e (ii) avaliar estatisticamente o desempenho de estimadores para o total populacional, sob esse plano, a partir de uma atualização e ampliação do estudo realizado por Strand (1979). Tal avaliação, leva em consideração a realização de experimentos de Monte Carlo, incluindo cenários de estratificação.

A dissertação está dividida em cinco seções, incluindo essa introdução. A segunda seção apresenta uma revisão histórica acerca da origem do plano de BE e de suas aplicações. A terceira seção engloba uma revisão de conteúdo acerca da amostragem de BE, elencando pontos relacionados ao plano amostral, à estimação pontual, à estimação intervalar e a amostragem estratificada.

A quarta seção realiza a comparação de desempenho entre os estimadores de Horvitz-Thompson e de Strand, para o total populacional, utilizando simulações de Monte Carlo. Por fim, a quinta seção apresenta a conclusão geral do trabalho.

2 História e Aplicações

2.1 Revisão histórica

A amostragem de Bernoulli foi originalmente apresentada por Goodman (1949), que em seu artigo a utilizou para lidar com a estimação do número de classes de uma população. O referido autor apresenta o conceito base de um plano de BE quando se refere a seleção de cada elemento da população como um evento aleatório e independente com mesma probabilidade de ocorrer, assim como quando discorre sobre o tamanho amostral, fazendo uso de uma variável aleatória indicadora que representa a inclusão ou não do elemento populacional na amostra. Dessa forma, a seleção de um elemento para a amostra é o resultado de um experimento aleatório com distribuição de Bernoulli de parâmetro igual a probabilidade de inclusão amostral π. Como consequência, o tamanho da amostra tem distribuição Binomial de parâmetros N e π, onde N é o tamanho da população. Por essa razão, Goodman (1949) adotou o nome de amostragem Binomial. No entanto, por conta da formulação do esquema amostral ele é considerado o precursor do plano amostral de BE.

O artigo de Goodman (1949) lida com a problemática de estimar o número total de classes que subdividem uma população finita. Imaginando uma urna com bolas coloridas, cada uma de uma cor, o problema equivaleria a, com base numa amostra de bolas tiradas dessa urna, estimar quantas cores distintas há na população de bolas. Goodman cita exemplos de situações corriqueiras para a época, que também podem ser aplicados nos dias atuais, como por exemplo: "Uma empresa recebe um grande número de solicitações de amostra grátis de seu produto. Sabe-se que as mesmas pessoas costumam enviar mais de uma solicitação. A partir de uma amostra das solicitações, desejamos estimar quantas pessoas diferentes enviaram solicitações".

Goodman leva em consideração quatro estimadores para a estimação das classes, sendo um deles

$$S'' = \frac{N}{n}\sum_{i=1}^{n} x_i, \tag{2.1}$$

onde x_i representa o número de classes que contém i elementos da amostra e n é tamanho amostral esperado. Ao aplicar o estimador S'' ao problema da estimação do total populacional, ou seja, saindo da estimação do número de classes e tendo como objetivo a estimação de uma característica de interesse y_j, que representa a j-ésima observação da variável de interesse y na amostra, obtém-se o estimador do total populacional do plano de BE que se dá por

$$t = \frac{N}{n}\sum_{j=1}^{n} y_j \tag{2.2}$$

Ao final do referido artigo, o pesquisador realiza uma aplicação prática comparando os resultados obtidos por meio dos estimadores. Dois deles apresentaram resultados mais preferíveis, um que aponta viés nulo e o outro que possui o menor erro quadrático médio (EQM). Todavia, o autor enfatiza que estes dois podem apresentar estimativas sem sentido. Logo, o pesquisador levanta modificações nos dois estimadores para sempre se obter estimativas razoáveis. Apesar de as estatísticas modificadas serem viesadas, elas possuem um EQM menor ou igual àquelas que não possuem modificação.

Beckwith (1973) parece ser o primeiro pesquisador a utilizar o termo "Amostragem de Bernoulli", em sua pesquisa, ele aborda o problema de avaliar o limite do tamanho de amostral em planos BE e BE modificado, tendo aplicações em pesquisas de opinião. A contribuição do plano para a época se deu pelo aperfeiçoamento de coletas de dados com base em questionários que se limitavam a apresentar perguntas dicotômicas, geralmente com respostas do tipo sim ou não. Beckwith demonstra que ao substituir as respostas dicotômicas por categorias $k \geqslant 3$ faz com que seja possível aumentar a precisão do estimador da característica da população em tamanhos amostrais pequenos.

A diferença entre o plano de BE e o modificado, proposto por Beckwith é que, na amostragem de BE, tem-se a suposição de que cada indivíduo da população possui igual probabilidade de inclusão π, $0 < \pi < 1$, enquanto que na modificada, a probabilidade de inclusão varia para cada indivíduo. Logo, a amostragem de BE modificada nada mais é do que a amostragem de Poisson.

Para delimitar os limites do intervalo de confiança para o tamanho amostral, o referido autor utilizou os limites de Chebyshev e da Normal. Com isso, ele comparou os dois tipos de amostragem, baseando-se nas estatísticas apresentadas, e chegou a conclusão de que a amostragem modificada, é superior à amostragem de BE. Por fim, o autor realizou duas aplicações para ilustrar o método. Na primeira, utilizando-se de um caso hipotético,

ilustrou o uso dos limites do tamanho amostral, em que um distribuidor de eletrodoméstico deseja avaliar a probabilidade subjetiva de que uma família selecionada aleatoriamente em uma cidade compraria algum dos seus produtos no próximo ano. Já na segunda, embora não aborde o limite do tamanho da amostra, fornece informações sobre a forma de se trabalhar com a não resposta, no qual o plano de BE se destaca. Nesta aplicação, deseja-se estimar o comparecimento dos membros de uma sociedade profissional cerca de seis meses antes do evento.

Com isso, uma amostra aleatória estratificada dos membros da associação foi gerada, sendo selecionada uma pessoa por página do diretório de filiação. Os membros que fazem parte da amostra foram solicitados a indicar sua intenção de comparecer ou não a reunião em uma escala de 0 à 10, em que 0 indica o não comparecimento e 10 indica a presença do membro na reunião. A estimação da previsão pontual da presença foi de 585 membros, tendo limites inferior e superior a 95% de confiança de 276 e 895, respectivamente. Como era almejado maior participação na reunião, esforços foram feitos para atrair não membros da sociedade, de modo que a participação efetiva foi de 1052, dos quais 587 eram membros da sociedade.

Já em Strand (1979), o pesquisador trabalha com o estudo sobre o plano de BE no processo de estimação do total populacional. No artigo, o referido autor investiga dois estimadores para o plano de BE. Estes são comparados ao estimador usual da amostragem aleatória simples (AAS), apresentando as situações em que os estimadores do plano de BE têm quase o mesmo erro quadrático médio que o estimador da AAS. Para isso, ele utiliza dois programas computacionais, o STRATA e o SPSS. Os estimadores para o plano de BE utilizados pelo referido autor se dão por $\hat{Y}_1$ e $\hat{Y}_2$, como descritos a seguir.

$$\hat{Y}_1 = \begin{cases} \left(\frac{N}{n_s}\right) \sum_{i=1}^{n_s} y_i, & \text{se} \quad n_s = 1, 2, \ldots, N \\ 0, & \text{se} \quad n_s = 0 \end{cases} \tag{2.3}$$

onde n_s é o tamanho amostral observado, com média $N\pi$ e variância $N\pi(1-\pi)$ e y_i representa o i-ésimo elemento selecionado para a amostra. A variância do estimador é dada por

$$V_{BE}(\hat{Y}_1) = N^2[HS^2 + Q^N(1-Q^N)\bar{Y}^2], \tag{2.4}$$

onde

$$\begin{aligned} H &= \sum_{k=1}^{N} (1/k) \binom{N}{k} P^k Q^{N-k} - 1(1-Q^N)/N, \\ S^2 &= \frac{\sum_{k=1}^{N} (y_k - \bar{Y})^2}{(N-1)}, \\ \bar{Y} &= \frac{\sum_{k=1}^{N} y_k}{N}, \end{aligned}$$

e $Q = 1 - P$, sendo $P = \pi$.

Já o segundo estimador é dado por

$$\hat{Y}_2 = \left(\frac{N}{n}\right) \sum_{i=1}^{n_s} y_i, \tag{2.5}$$

em que n é o tamanho amostral esperado e y_i é definido da mesma forma que em $\hat{Y}_1$. Tomando $n = N\pi$, este estimador é análogo ao estimador de Horvitz-Thompson. A variância para este estimador é dada por

$$V_{BE}(\bar{Y}_2) = (NQ/P)[(N-1)S^2/N + \bar{Y}^2]. \tag{2.6}$$

Strand (1979) realiza uma simulação comparando ambos os estimadores com o estimador da AAS, abordando diferentes tamanhos amostrais, populacionais e probabilidades de seleção. O pesquisador toma como medida de desempenho o EQM, chegando a conclusão de que o plano de BE é um procedimento alternativo vantajoso quando utilizado para amostrar uma população finita.

Baseado no artigo anterior, Wright (1982) propõe uma amostragem de BE inversa, com o intuito de estimar a proporção populacional e com aplicação em materiais nucleares, com o intuito de investigar o movimento de pequenas quantidades de materiais radioativos por indivíduo. A proposta do autor se dá pela combinação da amostragem Binomial inversa e a amostragem de BE, assumindo que N é muito grande e que fatores de correção para população finita podem ser ignorados. Levando em consideração a aplicação desta método, o processo se dá assumindo que todos os elementos da população são dispostos um a um, como em uma esteira transportadora, sendo selecionada e inspecionada apenas uma fração (π) de elementos, que é especificada pelo inspetor. Dividindo a população em duas categorias C e $\bar{C}$, o interesse do autor é estimar a proporção de elementos na população da categoria C.

Para tal, o referido autor assume que a probabilidade de um elemento i ser selecionado é π e que a probabilidade de um elemento i estar na categoria C é dado por p. Assumindo que a produção dos elementos é estatisticamente controlada e que a probabilidade de produção na categoria C é constante e igual a p. Com isso, existem quatro possibilidades para o elemento i, são elas

e_1 : o elemento i é selecionado e está na categoria C;

e_2 : o elemento i é selecionado e não está na categoria C;

e_3 : o elemento i não é selecionado e está na categoria C;

e_4 : o elemento i não é selecionado e não está na categoria C.

Já as probabilidades para estas quatro possibilidades são dadas por

$$\begin{aligned} P(e_1) =& \pi p; \\ P(e_2) =& \pi(1-p); \\ P(e_3) =& (1-\pi)p; \\ P(e_4) =& (1-\pi)(1-p). \end{aligned}$$

Tendo como interesse e_1, ensaios independentes de Bernoulli são aplicados para cada elemento da população. Logo, tem-se que

$$\begin{aligned} P(sucesso) &= P(e_1) = \pi p; \\ P(falha) &= P(e_2) + P(e_3) + P(e_4) = 1 - \pi p. \end{aligned}$$

Este procedimento é repetido até se obter k sucessos. Tomando $\tilde{n}$ como sendo o número de elementos (tentativas) que devem ser disponibilizados para se obter k sucessos, o design da amostragem de BE inversa é dado por

$$p(\tilde{n} = m) = \binom{m-1}{k-1}(\pi p)^k(1-\pi p)^{m-k} \quad para \quad m = k, k+1, k+2, \ldots \quad . \tag{2.7}$$

O mesmo autor também compara o plano da BE inversa com o plano da Binomial inversa quanto a variância, afirmando que quando π tende a 1 elas são aproximadas. Estes resultados são exibidos em uma tabela, para vários valores de π, p e k. Ao final é apresentado um exemplo para ilustrar o método proposto, sendo este aplicado a uma instalação de pesquisa nuclear.

Com base na proposta desenvolvida por Kingman (1963), Deshmukh (1991) introduziu os conceitos de processo de contagem de BE em um processo pontual e a amostragem de BE de um processo estocástico de parâmetros discretos. O artigo tem por objetivo demonstrar que o processo criado, ao desconstruir um processo estocástico de parâmetros discretos pelo plano de BE, satisfaz a mesma propriedade que o processo de contagem de BE. Também é constatado que a estacionariedade e a propriedade de Markov são invariantes no plano de BE.

Bunge e Fitzpatrick (1993) apresentam uma amostragem múltipla de BE que possui a característica de lidar com um tamanho populacional infinito. Com isso, dada uma população infinita particionada em C classes, a amostra pode ser representada pela matriz $C \times n$, em que n representa o número de ocasiões/observações realizadas e, para cada ocasião, cada classe pode ser ou não observada. Assim, a diferença para o plano usual de BE é que para o plano múltiplo de BE se pode ter mais de uma observação do plano amostral, no qual o resultado é posto em uma matriz. Sua origem se deu em experimentos de captura-recaptura, porém foi utilizado para a estimativa de C. Além disso, os autores

apresentam uma revisão sobre a estimação do número de espécies/classes, expondo alguns planos amostrais para população finita e infinita. Por fim, no que consideram como "o estado da arte", eles discutem sobre o uso dos estimadores e apresentam recomendações de trabalhos futuros.

Särndal (1996) aborda, em sua pesquisa, a utilização de um estimador de regressão generalizado nos planos amostrais em que a estimativa da variância requer probabilidade de inclusão de segunda ordem. O autor busca a diminuição do peso computacional, uma vez que, para a maioria dos esquemas de probabilidade proporcional ao tamanho da amostra $N\pi$, as probabilidades de segunda ordem possuem cálculos tediosos computacionalmente. Fazendo uso de um estimador de regressão, a estimativa da variância envolve apenas o cálculo de uma soma residual quadrada ponderada simples, substituindo o cenário $N\pi$ tradicional por uma alternativa mais simples, preservando a característica da alta eficiência. Uma amostragem de BE é utilizada, bem como uma amostragem de BE estratificada, para se avaliar o estimador desenvolvido.

Já Ghosh e Vogt (2002) estudam o tamanho amostral nas distribuições de BE e de Poisson, que é considerado uma variável aleatória que varia de 0 ao tamanho da população N. Os autores supracitados discutem formas de contornar esta limitação, como o uso de uma amostragem por rejeição, que se dá pela realização de uma amostragem em que a amostra obtida pode ser rejeitada a menos que seja alcançado um tamanho amostral almejado.

Outro método exposto pelos mesmos pesquisadores é o ajuste da probabilidade de inclusão por um fator de escala, para aumentar ou diminuir o tamanho de uma amostra previamente determinada, tendo por objetivo regularizar o tamanho da amostra. Todavia, ao realizar a regularização, se perde a independência conjunta das variáveis de inclusão e, com isso, não se pode calcular a variância diretamente. Dessa forma, os pesquisadores propõem que as variáveis de inclusão só precisam ser independentes em pares e que as probabilidades de inclusão tenham valores estipulados. Com isso, desenvolvem os métodos de amostragem generalizados de BE e Poisson.

A amostragem generalizada de BE apresentada pelos referidos autores, pode ser exemplificada da seguinte forma: atribui-se probabilidades, segundo uma variável aleatória Binomial generalizada, para cada tamanho de amostra de 0 a N - desde que $|2\pi - 1| \leqslant [1 - (4/N)]^{1/2}$ são permitidos tamanhos amostrais tão pequenos quanto 2 ou 3 - e depois é escolhido o tamanho de amostra de acordo com essas probabilidades. Logo após, dado o tamanho de amostra n, é selecionada uma amostra aleatória de tamanho n da população.

2.2 Aplicações envolvendo o plano de BE

O plano amostral de BE tem sido aplicado em diversos problemas práticos, como no estudo realizado por Cohen (1975), em que o pesquisador objetivou testar e compreender modelos quantitativos e a composição demográfica de grupos sociais de espécies de orangotangos selvagens. Para atingir seu objetivo, o mesmo autor fez uso do plano de Bernoulli, assumindo que os grupos fossem formados por amostragem independente de BE, de forma a estimar o número de indivíduos pertencentes aos grupos sociais de orangotangos, onde cada membro possui probabilidade idêntica "p" de inclusão em um grupo.

O autor utiliza uma base de dados referente a observação de orangotangos selvagens obtidas por MacKinnon (1974) para realização da análise, apresentando os resultados por meio de tabelas. Com isso, Cohen conclui que assumir uma probabilidade de inclusão independente e idêntica para todo orangotango, desconsiderando as classes de idade e sexo, não é viável, pois os resultados obtidos podem destoar da realidade. Por fim, o autor discute sobre a escolha de planos amostrais simples que, em alguns casos, fornecem uma descrição completa dos dados e, em outros, falham em descrever partes igualmente importantes desses, cabendo ao pesquisador identificar e avaliar o cerne do problema.

Já o autor Frank (1977b), com base no estudo da teoria de grafos, aplicou o plano de BE em conjunto com o estimador de Horvitz-Thompson para estimar o total $T = \sum_{i \in V} \sum_{j \in V} y_{ij}$ utilizando observações y_{ij}, para $(i,j) \in S^2$, no qual S é uma amostra aleatória, selecionada a partir do plano de BE, de V, sendo $V = 1, \ldots, N$. Neste artigo, o mesmo autor indaga tecnicamente sobre o estimador de Horvitz-Thompson aplicado a problemática dos grafos, com o intuito de avaliar, estimar e testar vários parâmetros dos grafos, apresentando o estimador para o total, sua variância e um estimador não viesado para a variância. O documento é criado a fim de se referir a um problema de estimativa em associação com a amostragem de pesquisa em grafos, abordada pelo autor em outros estudos, vide Frank (1971, 1977a, 1977c, 1978), e abordado mais criteriosamente neste artigo.

Com base em Goodman (1949), o autor Berg (1977) lida com a problemática de duplicidade e/ou sobreposição em algumas unidades da amostragem populacional. O autor debate sobre a amostragem de multiplicidade de Sirken, ao longo do que foi trabalhado por Goodman, e pondera que a referida pesquisa pode ser considerada uma adaptação, e ligeira extensão, do trabalho desenvolvido por Goodman. Neste artigo, o plano de BE é utilizado como uma medida de análise, aplicado a situação em que é realizada uma pesquisa na qual não existe uma correspondência individual entre as unidades de amostragem e as unidades de interesse.

Ao longo da referida pesquisa são mencionados e utilizados três métodos de amostragem, a saber: amostragem por conglomerado, amostragem de multiplicidade de Sirken e

amostragem a partir de uma lista na qual algumas unidades estão duplicadas. Por fim, o autor argumenta sobre três exemplos de aplicações, ilustrando-os. No primeiro, objetiva-se examinar o caso de amostragem de várias listas, cada uma contendo duplicações. No segundo exemplo, é abordado o problema de estimar o número de peixes em um lago. Já no terceiro, a partir dos registros da população contidos no Escritório Central de Estatística de Estocolmo, o autor discute sobre possíveis problemáticas que se enquadrariam a sua pesquisa, como, por exemplo, agrupar indivíduos em estratos de acordo com algum critério, tendo em vista que os limites do estrato podem cortar as unidades de interesse, que são representadas pela família do indivíduo, por exemplo, apresentando situações em que os estratos se sobrepõem e as unidades de interesse são duplicadas.

Laborando com a problemática da estimação do número de classes em uma população finita, tem-se o artigo de Haas e Stokes (1998). Nele, os autores esclarecem alguns pontos sobre a estimação, apresentando situações em que ela é empregada, como exemplo, em um concurso patrocinado por uma empresa, em que muitas inscrições são recebidas. Tendo em vista que algumas pessoas enviaram a inscrição mais de uma vez, o objetivo é estimar o número de pessoas diferentes que entraram na amostra.

Tomando como base o plano de BE, e realizando aproximações a partir dele, o trabalho visa a avaliação de oito estimadores, que incorporam diferentes abordagens para lidar com as dificuldades geradas pela variação nos tamanhos das classes. O desempenho é analisado por meio da variância assintótica em um estudo de simulação de Monte Carlo.

Com base nos resultados obtidos no estudo, os referidos autores chegaram a conclusão de que, dos estimadores considerados, o melhor é o estimador de ramificação, no qual um estimador de Shlosser modificado é utilizado quando o coeficiente de variação dos tamanhos das classes é estimado como sendo muito grande e sem suavização. Já os estimadores de HT, segundo os mesmos autores, apresentaram desempenho inferior, indicando que é improvável que as abordagens baseadas na estimativa direta dos tamanhos das classes sejam bem-sucedidas.

Já os autores Skinner e Elliot (2002), propuseram uma nova medida de risco de divulgação de arquivos de microdados de pesquisas, uma vez que a avaliação do risco geralmente envolve julgamentos difíceis e formas sistemáticas são necessárias para dar apoio a esses julgamentos. Para tal, adotou-se uma estrutura de amostragem de pesquisa, na qual as quantidades finitas de população são fixas e a única fonte de aleatoriedade está na seleção da amostra. Com o intuito de demonstrar que a medida de risco proposta por eles não precisa de fortes suposições de modelagem, como geralmente ocorre nas avaliações do risco existentes, e por sua simplicidade, os pesquisadores utilizaram o plano de Bernoulli, no qual todas as unidades populacionais foram amostradas independentemente, com uma probabilidade comum.

O artigo supracitado foca em três tipos de medidas de risco de divulgação. As duas primeiras abordam 100% do uso de dados do censo, sofrendo dificuldades nas suas inferências, devido à extensão dos dados. Enquanto a terceira, proposta pelos autores, além de ser potencialmente uma medida de risco mais realista, também fornece um meio de superar a dificuldade inferencial das duas primeiras medidas. Esta medida é definida como a proporção de correspondências corretas entre as unidades populacionais que diz respeito a um registro de microdados exclusivo da amostra.

Além do plano de BE, os mesmos autores também trabalham com a AAS e a amostragem estratificada. Ao final é feita uma aplicação para os dados do censo de 1991 da Grã Bretanha com $N = 450000$, para diferentes valores de π, chegando à conclusão de que a proposta de medida de risco pode ser utilizada para decidir entre formas alternativas de liberar microdados de uma pesquisa amostral. Com isso, o risco de divulgação pode ser avaliado de maneira relativa, comparando estratégias alternativas de liberação ou exigindo que não exceda alguma probabilidade específica.

O artigo de Duchesne (2003) teve por objetivo avaliar estimadores para a proporção, fazendo uso de estimadores de regressão. O autor faz uso do plano de BE e do plano de AAS, como também de um cenário estratificado de ambos para realizar uma análise entre os estimadores de HT, o estimador de regressão logística (RL) e o estimador de regressão clássico.

O referido autor afirma que o uso de informações auxiliares melhora a precisão das estimativas da média ou do total quando um estimador de regressão é utilizado. No estudo, é realizada uma aplicação a dados reais referente a 7193 escolas que apresentaram um índice de performance acadêmico igual ou superior a 2000. Os resultados são dispostos em tabelas e demonstram que os estimadores de proporções podem ser mais eficientes devido à utilização de conhecimentos auxiliares, bem como que uma variável de estratificação eficiente pode dar apoio para uma estimativa mais precisa. O estimador de RL foi o que apresentou menor variação, sendo o cenário estratificado o que resultou em um melhor desfecho.

Para Duchesne, o plano de BE pode ter um importante valor pedagógico, porque os alunos geralmente têm problemas com o conceito de probabilidade de inclusão. Deste modo, o plano de BE pode ser uma forma de ajudá-los a ver quais são as probabilidades de inclusão, tendo em vista que é muito fácil de implementar e que pode lançar luz sobre a parte aleatória do experimento de amostragem.

No mesmo ano, os autores De e Sengupta (2003) objetivaram, na pesquisa por eles realizada, estudar a amostragem CMLR, com base na estrutura padrão de uma amostragem sequencial de Bernoulli. A amostragem CMLR consiste em, inicialmente, "Capturar", "Marcar" e "Liberar" j unidades populacionais na população-alvo e, em se-

guida, “Recuperar” aleatoriamente unidades da população em uma ou mais amostras, até que a amostragem seja interrompida de acordo com uma regra de parada. Os autores utilizaram noções da amostragem sequencial de BE (GUPTA, 1975)(SINHA; SINHA, 1975)(SINHA; BOSE, 1985) e forneceram uma abordagem do problema da estimação imparcial de N e, em particular, demonstraram, por meio de exemplos, estimação sob uma regra de parada arbitrária.

Com a evolução da tecnologia, surgiram novas aplicações envolvendo o uso de uma amostragem de BE, como é o caso da pesquisa desenvolvida Haas e König (2004). Nesse artigo, os autores abordaram a problemática da quantidade massiva de informações armazenadas na *Web* para possíveis minerações e explorações. Para tanto, utilizam um esquema de amostragem de dois níveis de Bernoulli, com o intuito de aprimorar o processo de consultas de dados, tendo em vista que o uso de algoritmos de análise de dados têm se tornado menos vantajosos, devido à dimensão das informações e ao acúmulo de *terabytes* em repositórios, bem como devido à perceptível exigência dos usuários por sistemas de processamento e análise de dados mais rápidos.

Baseando-se na amostragem de BE, os referidos autores desenvolveram a amostragem de dois níveis de BE, que combina os métodos de amostragem de nível de linha e de página, utilizados pela maioria dos sistemas comerciais da época, em um esquema que permite uma troca sistemática entre velocidade de processamento e precisão estatística, ao ajustar os parâmetros do método. Os pesquisadores também propõem um método heurístico prático com o intuito de aprimorar as configurações de parâmetros ideais, abdicando de uma amostra piloto. Por fim, são realizados mais de 1100 experimentos em 372 conjuntos de dados reais e sintéticos, que resultam em erros amostrais pequenos e comprovam a eficiência do método até mesmo em casos mais complexos.

No estudo desenvolvido por Rondon *et al.* (2012) foi realizada uma aplicação envolvendo estimadores de regressão e o plano de Bernoulli, este último, utilizado para auxiliar na avaliação dos estimadores. No referido artigo, os autores utilizam cinco planos amostrais, dentre os quais está o plano de BE, com o objetivo de desenvolver um estimador do tipo regressão.

A distribuição populacional finita da variável de interesse é vista como se fosse gerada por um membro da família exponencial, que inclui distribuições normais (mesmo com heterogeneidade de variância), Bernoulli, Poisson, gama e gaussiana inversa. Os autores exploram três aplicações utilizando bancos de dados reais para aferirem sobre os estimadores, resultando que o estimador de regressão de modelo linear generalizado, desenvolvido pelos autores, apresenta melhor desempenho que o estimador de regressão clássico.

Li *et al.* (2015) desenvolveram um artigo sobre a operação de consulta de frequências. Esta operação se dá quando se obtém uma frequência de ocorrência de cada valor em um

conjunto de dados sensoriais, sendo popular em Redes Móveis *Ad hoc* (do inglês, MANET).

O objetivo do mencionado artigo foi propor um algoritmo aproximado que se baseia em um plano de BE para processar consultas de frequência em MANETs, isto é, redes continuamente autoconfiguráveis e sem estrutura, que consistem em nós móveis, sendo bastante utilizadas. São exemplos de sua utilização, dentre outros, em monitoramento de poluição e vigilância de animais e, até mesmo, na estimação do número de pessoas feridas e sobreviventes de um desastre.

Para o cálculo do algoritmo da frequência aproximada, foram necessários os seguintes passos. Em primeiro lugar, determinar a probabilidade de amostragem "q" de acordo com um ϵ ($\epsilon > 0$) e δ ($0 \leqslant \delta \leqslant 1$) especificado. Em segundo lugar, usar um algoritmo de amostragem Bernoulli para amostragem de dados sensoriais. Finalmente, baseado nos dados amostrados, calcular frequência aproximada. Conforme resultados da pesquisa, o algoritmo desenvolvido foi utilizado em simulações e observou-se que ele possuiu alto desempenho quanto aos aspectos de eficiência energética e precisão.

O trabalho proposto por Kawakami *et al.* (2017) trouxe a temática da segurança das transferências de dados e das dificuldades de distribuir a chave dos dados criptografados entre pessoas de uma maneira segura. Os autores supramencionados objetivaram propor um método de análise de chaves finitas baseado no plano de BE, ao invés de amostragem aleatória simples, comumente utilizada em pesquisas anteriores.

A distribuição de chaves quânticas se baseia nos fundamentos da física quântica, tornando-se invulnerável às inseguranças nas transferências de dados. Um passo essencial na distribuição de chaves quânticas é a estimativa de parâmetros relacionados à quantidade vazada de informações, o que geralmente é feito por amostragem dos dados de comunicação. Quando o tamanho dos dados é finito, a taxa final depende de como o processo de estimativa lida com as flutuações estatísticas. Muitas das análises de segurança atuais são baseadas no método com amostragem aleatória simples, onde a distribuição hipergeométrica ou seus limites conhecidos são usados para a estimativa.

Para o protocolo BB84, utilizado na referida pesquisa como uma escolha de base de monitoramento tendenciosa, os dados coletados em uma das bases são utilizados exclusivamente para monitorar informações vazadas na base principal. Nesse caso, todos os dados da base de monitoramento são considerados uma amostra, em que cada rodada é selecionada com uma probabilidade constante, ditada no protocolo como aquela da escolha da base. Isso sugere que os dados da base de monitoramento estão relacionados ao plano de BE.

Com base nos resultados do mencionado estudo, concluiu-se que o novo método, baseado no plano de BE, é especialmente adequado para o protocolo BB84, fornecendo uma análise mais simples com menos processos de estimativa, além de alcançar uma taxa-chave mais

alta em comparação com a análise com amostragem aleatória simples.

Já Wiuf (2018), aplica o plano de BE na área da genética, apresentando um processo reconstruído condicionado a amostragem de BE para descrever a evolução de uma população de indivíduos originados de um único indivíduo. Este tipo de processo é utilizado, por exemplo, em casos de infecção por vírus, que pode ser causada pela transmissão de uma ou poucas partículas de vírus do ambiente para o hospedeiro. Acredita-se que uma mutação pontual de DNA no passado pode fundar uma população de cromossomos portadores do tipo mutante, produzida por um único cromossomo.

As auditorias de limitação de risco foram introduzidas como um mecanismo para detectar e corrigir erros de mudança de resultado na tabulação de votos, qualquer que seja a sua causa, incluindo hackers, configuração incorreta e erro humano. Na pesquisa realizada por Ottoboni *et al.* (2019), os autores propuseram apresentar um método e um software para auditorias de limitação de risco de votação com base no plano de BE: as cédulas são incluídas na amostra com probabilidade π.

Os autores argumentaram que o plano de BE tem várias vantagens, dentre as quais: a de que pode ser conduzida independentemente em diferentes locais, em vez de exigir que uma autoridade central selecione a amostra de toda a população de cédulas de votação ou exigir amostragem estratificada e a de que pode começar nos locais de votação na noite da eleição, antes que as margens sejam conhecidas.

Com base nos resultados, os autores identificaram que se as margens informadas para a eleição presidencial dos EUA em 2016 estiverem corretas, ao realizar uma auditoria de Bernoulli com um limite de risco de 5%, mesmo utilizando uma probabilidade de inclusão de $\pi = 0.01$ teria, pelo menos, uma probabilidade de 99% de confirmar o resultado em 42 estados. Os outros estados tinham maior chance de ser necessário examinar cédulas adicionais. Ademais destacaram que a simplicidade logística do plano de BE pode torná-la útil para auditorias eleitorais, tendo em vista que a auditoria no local da votação pode ser mais interessante do que o custo de examinar mais cédulas que alguns outros métodos podem exigir.

3 Amostragem de Bernoulli

3.1 Plano amostral

A amostragem de Bernoulli (BE) se destaca pela sua simplicidade e aplicabilidade. Imagine um conjunto de índices $U = \{1, 2, \ldots, k, \ldots, N\}$, que identificam os elementos pertencentes a uma população finita de tamanho N. Para extrair uma amostra s, $(s \subseteq U)$, pelo plano amostral de BE, realiza-se um experimento aleatório para cada elemento de U. Um exemplo simples consiste em jogar uma moeda e nas vezes em que cair coroa, incluir o elemento k na amostra. Defina I_k como uma variável aleatória que designa a inclusão do elemento k na amostra, tal que

$$I_k = \begin{cases} 1, & \text{se} \quad k \in s \\ 0, & \text{c.c.} \end{cases}$$

I_k segue uma distribuição de Bernoulli em que $\Pr(I_k = 1) = \pi$ e $\Pr(I_k = 0) = 1 - \pi$. O tamanho da amostra n_s é obtido ao somar os valores de I_k:

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

Deste modo, n_s é uma variável aleatória com distribuição Binomial de parâmetros (N, π). Por conseguinte, a probabilidade de selecionar uma amostra s de tamanho n_s, com o plano de BE é dado por

$$p(s) = \pi^{n_s}(1 - \pi)^{N - n_s}. \tag{3.1}$$

Além do mais, a probabilidade de que o tamanho da amostra aleatória seja exatamente n segue distribuição Binomial, em que

$$Pr(n_s = n) = \binom{N}{n} \pi^n (1 - \pi)^{N-n}.$$

Consequentemente, a esperança e variância de n_s são dadas por

$$E(n_s) = N\pi \quad \text{e}$$
$$Var(n_s) = N\pi(1-\pi)$$

A probabilidade de inclusão π do plano de BE é igual para cada elemento populacional. Exemplificando, toda vez que se joga uma moeda se tem 50% de chance de sucesso. Portanto, para todo $k \in U$, a probabilidade de inclusão de primeira ordem, que é a probabilidade de uma elemento fazer parte da amostra, se dá por $\pi_k = \pi$ e a de segunda ordem, que é a probabilidade de incluir um par de elementos, é dada por $\pi_{kl} = \pi^2$ para todo $k \neq l$.

Um algoritmo comumente utilizado para seleção amostral deste plano é o de fixar um valor para π, $0 < \pi < 1$; obter N valores ω_k $(k \in U)$ que representam realizações independentes de uma distribuição Uniforme no intervalo [0, 1]; e selecionar o k-ésimo elemento para fazer parte da amostra se $\omega_k < \pi$. Como n_s é uma variável aleatória, existe a possibilidade de não se obter uma amostra ($n_s = 0$), bem como de se realizar um censo ($n_s = N$).

3.2 Estimação pontual

Objetivando estimar o total populacional $Y = \sum_{k \in U} y_k$, em que y_k representa uma variável populacional de interesse, dois estimadores se destacaram dentre as referências trabalhadas, a saber: o estimador de Horvitz-Thompson (HT) (HORVITZ; THOMPSON, 1952) e um estimador alternativo, que nada mais é que o estimador proposto por Strand (1979) que é denominado como $\hat{Y}_1$, sendo este último uma espécie de correção para o estimador de HT. Tais estimadores são uma função dos valores amostrais que é utilizada para estimar o valor de um parâmetro desconhecido da população.

Começando pelo estimador de HT, ele é um estimador não viesado que foi desenvolvido para lidar com amostras retiradas sem reposição de um universo finito com probabilidades de seleção distintas, sendo considerado um caso geral da amostragem sem reposição. Este estimador se dá pela soma ponderada dos valores amostrais, em que o peso de cada elemento amostral é o inverso de sua probabilidade de inclusão. Para o caso da BE, o estimador para o total populacional segundo HT é dado por

$$\hat{t}_\pi = \frac{1}{\pi} \sum_{k \in s} y_k. \tag{3.2}$$

Já a sua variância é dada por

$$V_{BE}(\hat{Y}_{HT}) = \sum_{k=1}^{U} \frac{1-\pi_k}{\pi_k} y_k^2 + \sum_{k=1}^{U}\sum_{k\neq l}^{U} \frac{\pi_{kl}-\pi_k\pi_l}{\pi_k\pi_l} y_k y_l. \tag{3.3}$$

Aplicando ao plano de BE, a variância do estimador HT se reduz a

$$V_{BE}(\hat{t}_\pi) = \left(\frac{1}{\pi} - 1\right) \sum_{k\in U} y_k^2. \tag{3.4}$$

Um estimador não viesado para a variância do estimador do total populacional é dado por

$$\widehat{V_{BE}}(\hat{Y}_{HT}) = \sum_{k=1}^{s} \frac{1-\pi_k}{\pi_k^2} y_k^2 + \sum_{k=1}^{s}\sum_{k\neq l}^{s} \frac{\pi_{kl}-\pi_k\pi_l}{\pi_k\pi_l\pi_{kl}} y_k y_l. \tag{3.5}$$

Aplicando ao plano de BE, se reduz a

$$\widehat{V_{BE}}(\hat{t}_\pi) = \frac{1}{\pi}\left(\frac{1}{\pi} - 1\right) \sum_{k\in s} y_k^2. \tag{3.6}$$

é possível reescrever a variância do estimador (3.4), dado que o valor esperado do tamanho amostral é $n = N\pi$, de forma a obter

$$V_{BE}(\hat{t}_\pi) = N^2 \left(\frac{1}{n} - \frac{1}{N}\right) S_{yU}^2 \left[1 - \frac{1}{N} + (cv_{yU})^{-2}\right], \tag{3.7}$$

no qual

$$\begin{aligned} cv_{yU} &= \frac{S_{yU}}{\bar{y}_U}, \\ S_{yU}^2 &= \frac{1}{N-1}\sum_{k\in U}(y_k - \bar{y}_U)^2 \quad \text{e} \\ \bar{y}_U &= \frac{\sum_{k=1}^{N} y_k}{N}, \end{aligned}$$

representam o coeficiente de variação de y na população U, a variância populacional e a média populacional, respectivamente.

Um estimador alternativo ao de HT é descrito por Strand (1979) como estimador $\hat{Y}_1$, apresentado em (2.3), e sendo denominado assim daqui em diante. O estimador $\hat{Y}_1$ é expresso por

$$\hat{Y}_1 = \frac{N}{n_s} \sum_{k\in s} y_k. \tag{3.8}$$

Tomando a média amostral como $\sum_{k\in s} = y_k/n_s$, tem-se que

$$\hat{Y}_1 = N\bar{y}_s = \frac{n}{n_s}\hat{t}_\pi, \tag{3.9}$$

onde $n = N\pi = E_{BE}(n_s)$ é o tamanho amostral esperado. Tomando por base a equação (3.2), nota-se que $\hat{t}_\pi$ tem uma alta variabilidade, devido a variação do tamanho amostral

esperado, ao colocar o tamanho aleatório n_s no denominador, espera-se que o estimador $\hat{Y}_1$ esteja reduzindo parte dessa variabilidade observada no estimador de HT. Todavia, segundo Strand (1979) a vantagem do estimador $\hat{Y}_1$ se dá apenas quando se sabe o tamanho populacional N. A variância deste estimador pode ser aproximada por

$$V_{BE}(\hat{Y}_1) \cong N\left(\frac{1}{\pi} - 1\right) S_{yU}^2 = N^2(1-f)\frac{S_{yU}^2}{n}, \tag{3.10}$$

onde $f = n/N = \pi$.

Utilizando (3.7), nota-se que

$$\begin{aligned}
\frac{V_{BE}(\hat{t}_\pi)}{V_{BE}(\hat{Y}_1)} =& \frac{N^2\left(\frac{1}{n} - \frac{1}{N}\right) S_{yU}^2\left[1 - \frac{1}{N} + (cv_{yU})^{-2}\right]}{N^2(1-f)\frac{S_{yU}^2}{n}} \\
=& \frac{N^2\left(\frac{1}{n} - \frac{1}{N}\right) S_{yU}^2\left[1 - \frac{1}{N} + (cv_{yU})^{-2}\right]}{N^2(1-\frac{n}{N})\frac{S_{yU}^2}{n}} \\
=& \frac{N^2\left(\frac{1}{n} - \frac{1}{N}\right) S_{yU}^2\left[1 - \frac{1}{N} + (cv_{yU})^{-2}\right]}{N^2\left(\frac{1}{n} - \frac{1}{N}\right) S_{yU}^2} \\
=& 1 - \frac{1}{N} + (cv_{yU})^{-2} \cong 1 + (cv_{yU})^{-2}.
\end{aligned} \tag{3.11}$$

Särndal *et al.* (2003, p. 62–65) afirmam que, mesmo sendo viesado, o estimador $\hat{Y}_1$ possui variância menor que o estimador de HT. Quando são trabalhadas amostras muito grandes, o viés do estimador acaba tendendo a zero. O estimador $\hat{Y}_1$ passa a ser uma correção do estimador de HT, em que quanto mais próximo n_s for de seu valor esperado, menor é a correção.

3.3 Estimação intervalar

Devido a variabilidade amostral, é relevante incluir uma estimativa intervalar, de forma a estabelecer limites com uma probabilidade de incluir o verdadeiro valor do parâmetro de interesse. Diante disso, é possível calcular um intervalo de confiança para os possíveis valores de n_s. Uma vez que as condições que atendem a um Teorema Central do Limite sejam satisfeitas, é possível construir o seguinte intervalo de confiança assintótico para n_s

$$N\pi \pm z_{1-\alpha/2}\sqrt{N\pi(1-\pi)}, \tag{3.12}$$

onde $z_{1-\alpha/2}$ representa uma constante de modo que $P(Z > z_{1-\alpha/2}) = \alpha/2$, no qual $Z \sim N(0,1)$. Logo, com $100(1-\alpha)\%$ de confiança, o verdadeiro valor de n_s está contido dentro do intervalo.

3.4 Amostragem estratificada

Ao estratificar a população, obtém-se mais precisão na estimação de parâmetros, pois uma vez que se tenha homogeneidade dentro dos estratos, as estimativas calculadas mediante a amostra de um estrato qualquer são mais precisas em razão da pouca variabilidade dos elementos nele contidos. O objetivo da amostragem estratificada é dar um tratamento específico a cada estrato populacional, seja por razões econômicas, como aplicar esquemas operacionais distintos para vários estratos; administrativas, quando a população já é dividida em subgrupos formados naturalmente; ou logísticas, quando se quer diminuir a variância da estimação.

Para compreender a funcionalidade desta técnica, imagine uma população U que pode ser dividida em partições $U_1, \ldots, U_H$ de U, que representem subpopulações bem definidas (estratos), de forma que $U = \bigcup_{h=1}^{H} U_h$ e $U_h \bigcap U_{h'} = \varnothing$, para $h \neq h' = 1, \ldots, H$. Cada elemento da população pertence a somente um estrato e cada subpopulação U_h possui tamanho N_h e, portanto, $N = \sum_{h=1}^{H} N_h$. Quanto maior for a capacidade em produzir estratos homogêneos, mais eficaz é o resultado obtido por meio da estratificação (BOLFARINE; BUSSAB, 2005, p. 93–99).

Para garantir a precisão do procedimento amostral se faz necessário a alocação da amostra, que é a distribuição das n unidades da amostra pelos estratos. Neste trabalho foi utilizada a alocação proporcional com o intuito de se obter camadas que tenham as mesmas proporções observadas na população, que se dá ao distribuir a amostra de tamanho n proporcionalmente ao tamanho dos estratos (BOLFARINE; BUSSAB, 2005, p. 102–103). A alocação proporcional é escrita da forma

$$n_{s_h} = \frac{nN_h}{N}. \tag{3.13}$$

Diante do exposto, aplicando a estratificação ao plano de BE, para cada estrato da população a amostragem é realizada a partir de ensaios independentes de BE, em que $\pi_h = n_{s_h}/N_h$. Com isso, o tamanho amostral esperado no estrato U_h é n_{s_h} e uma amostra s_h pode ser selecionada de U_h a partir do plano. Deste modo, em cada amostra s_h usam-se os estimadores apresentados em (3.2) e (3.8) e, por conseguinte, monta-se um estimador por meio da combinação dos estimadores de cada estrato. Logo, o estimador de HT para a amostragem estratificada de Bernoulli (AEBE) é dado por

$$\hat{t}_\pi = \sum_{h=1}^{H} \sum_{k \in s_h} \hat{t}_{\pi_h}. \tag{3.14}$$

A variância desse estimador se dá por

$$V_{AEBE}(\hat{t}_\pi) = \sum_{h=1}^{H} \left(\frac{1}{\pi_h} - 1 \right) \sum_{k \in U_h} y_{k_h}^2. \tag{3.15}$$

Já o estimador não viesado da variância fica expresso como

$$\widehat{V_{AEBE}}(\hat{t}_\pi) = \sum_{h=1}^{H} \frac{1}{\pi_h}\left(\frac{1}{\pi_h} - 1\right) \sum_{k \in s_h} y_{k_h}^2. \tag{3.16}$$

De mesmo modo, para o estimador $\hat{Y}_1$ de Strand (1979), tem-se que o total populacional é dado por

$$\hat{Y}_1 = \sum_{h=1}^{H} \frac{N_h}{n_{s_h}} \sum_{k \in s_h} y_{k_h}. \tag{3.17}$$

Sua variância pode ser escrita como

$$V_{AEBE}(\hat{Y}_1) \cong \sum_{h=1}^{H} N_h^2 (1 - \pi_h) \frac{S_{yU_h}^2}{n_h}, \tag{3.18}$$

onde $S_{yU_h}^2$ é escrito da forma

$$S_{yU_h}^2 = \frac{1}{N_h - 1} \sum_{k \in U_h} (y_{k_h} - \bar{y}_{U_h})^2.$$

4 Estudo Computacional

4.1 Estimadores

Para o estudo de simulação foi considerado o método de Monte Carlo (MC) com o intuito de se obter estimativas mais precisas e concisas dos estimadores (3.2) e (3.8), para o total populacional. Isto é, dado o interesse em estimar um parâmetro θ através de um estimador $\hat{\theta}$ de uma população finita U, ao utilizar o método de MC é possível obter uma aproximação das caracterÃsticas do estimador, uma vez que não podem ser descritas analiticamente devido à sua complexidade.

Todavia, como já mencionado para o plano de BE, existe a possibilidade de não se obter uma amostra. Neste caso, foram omitidas nos cálculos dos estimadores as réplicas geradas em que nenhuma amostra foi selecionada. Dentre as 10000 réplicas geradas de MC, foi constatado nas simulações que menos de 1% das réplicas apresentaram $n_s = 0$, ocorrendo mais frequentemente quando a probabilidade de inclusão π é muito pequena.

A avaliação desses estimadores foi feita por meio das seguintes medidas:

1. Viés relativo do estimador

$$VR_{\hat{\theta}} = \frac{|\bar{\hat{\theta}} - Y|}{Y}, \tag{4.1}$$

onde $Y = \sum_{k=1}^{U} y_k$ e $\bar{\hat{\theta}}$ se refere à média do estimador mediante as réplicas de MC, isto é,

$$\bar{\hat{\theta}} = \sum_{t=1}^{M} \frac{\hat{\theta}(s_t)}{M}, \tag{4.2}$$

onde $t = 1, \ldots, M$ e M é o número de réplicas de MC.

2. Erro quadrático médio

$$EQM(\hat{\theta}) = S^2_{(\hat{\theta})} + (\bar{\hat{\theta}} - Y)^2 \tag{4.3}$$

onde $S^2_{\hat{\theta}}$ representa a variância do estimador segundo o método de MC.

3. Coeficiente de variação do estimador

$$CV_{(\hat{\theta})} = 100\frac{\sqrt{S^2_{(\hat{\theta})}}}{\bar{\hat{\theta}}} \tag{4.4}$$

O desempenho dos seguintes estimadores foi avaliado:

1. Estimador de Horvitz-Thompson para o total populacional (3.2) e sua respectiva variância (3.6):

$$\hat{t}_\pi = \frac{1}{\pi}\sum_{k\in s} y_k,$$

$$\widehat{V_{BE}}(\hat{t}_\pi) = \frac{1}{\pi}\left(\frac{1}{\pi} - 1\right)\sum_{k\in s} y_k^2.$$

2. Estimador $\hat{Y}_1$ de Strand (1979) para o total populacional (3.6) e sua variância (3.8):

$$\hat{Y}_1 = \frac{n}{n_s}\hat{t}_\pi,$$

$$V_{BE}(\hat{Y}_1) \cong N^2(1-\pi)\frac{S^2_{yU}}{n}.$$

3. Estimador de Horvitz-Thompson para o total populacional utilizando uma amostragem estratificada de BE (3.14) e sua respectiva variância (3.16):

$$\hat{t}_\pi = \sum_{h=1}^{H}\sum_{k\in s_h}\hat{t}_{\pi_h},$$

$$\widehat{V_{AEBE}}(\hat{t}_\pi) = \sum_{h=1}^{H}\frac{1}{\pi_h}\left(\frac{1}{\pi_h} - 1\right)\sum_{k\in s_h} y_{k_h}^2.$$

4. Estimador $\hat{Y}_1$ para o total populacional segundo uma amostragem estratificada de BE (3.17) e sua variância (3.18):

$$\hat{Y}_1 = \sum_{h=1}^{H}\frac{N_h}{n_{s_h}}\sum_{k\in s_h} y_{k_h},$$

$$V_{AEBE}(\hat{Y}_1) \cong \sum_{h=1}^{H} N_h^2(1-\pi_h)\frac{S^2_{yU_h}}{n_h}.$$

4.2 Cenários de simulação

Para o primeiro cenário foi considerado o plano amostral padrão de Bernoulli, apresentado na sessão 3.2. Para tal, foram trabalhados cinco tamanhos amostrais esperados, $n = 5$, 50, 100, 500 e 1000, adotando uma probabilidades de inclusão de 0.5, 0.2, 0.05, 0.01 e 0.001, estes valores foram abordados com o intuito de avaliar os estimadores em meio a n e π de tamanhos distintos, como também pode ser observado em Strand (1979). Este cenário é comumente utilizado em análises amostrais e avaliações de estimadores devido a simplicidade de sua aplicação. A exemplo da simulação de MC, o procedimento realizado para avaliar os estimadores segundo o plano de BE se deu por:

Passo 1. Obter um vetor τ_k, $\tau_k = (\tau_1, \tau_2, \ldots, \tau_N)$, que contenha os valores de N uniformes no intervalo (0,1), no qual N representa o tamanho populacional;

Passo 2. Associar, a cada elemento k da população, um valor referente a uma variável de interesse y;

Passo 3. Adotar uma probabilidade de inclusão π de forma a selecionar um elemento da população para fazer parte da amostra s toda vez que $\tau_k \leqslant \pi$;

Passo 4. Aplicar os estimadores na amostra obtida.

Este procedimento é repetido 10000 vezes e as medidas de avaliação dos estimadores para o total populacional, anteriormente discutidas, são aplicadas e feitas as devidas análises.

Para o segundo cenário, levou-se em consideração uma amostragem estratificada de Bernoulli, discutida na sessão 3.4. Para isso, foi adotado um tamanho populacional de $N = 10000$, particionado em três estratos, a saber: $N_1 = 7000$, $N_2 = 2000$ e $N_3 = 1000$. O tamanho amostral adotado foi de $n = 50, 100, 500, 1000$ e 5000. Já o tamanho amostral aplicado a cada estrato é proporcional ao tamanho do estrato, uma vez que foi utilizada uma alocação proporcional no qual $n_{s_h} = nN_s/N$.

Deste modo, ao aplicar a alocação, tem-se que as quantidades para o primeiro estrato são $n_{s_1} = 35, 70, 350, 700$ e 3500, para o segundo estrato se tem $n_{s_2} = 10, 20, 100, 200$ e 1000 e para o terceiro estrato, $n_{s_3} = 5, 10, 50, 100$ e 500, tendo em vista que as amostras selecionadas em cada estrato seguem o plano de BE.

A simulação de MC para avaliar os estimadores, segundo a amostragem estratificada de BE, se dá da seguinte forma:

Passo 1. Dividir a população N em H estratos, de forma a obter N_h, no qual $\sum_{h=1}^{H} N_h = N$.

Passo 2. Obter, um vetor τ_{k_h}, que contenha os valores de N_h uniformes no intervalo (0,1);

Passo 3. Associar, a cada elemento k_h da população, um valor referente a uma variável de interesse y_h;

Passo 4. Adotar uma probabilidade de inclusão π_h de forma a selecionar um elemento da população para fazer parte da amostra s_h toda vez que $\tau_{k_h} \leqslant \pi_h$;

Passo 5. Aplicar os estimadores para os diferentes estratos a partir das amostragens dos mesmos.

Este processo é repetido 10000 vezes de forma a se obter as medidas de avaliação dos estimadores para cada estrato e para o total. As análises referentes a estes cenários são abordadas a seguir e os scripts utilizados estão dispostos nos apêndices A e B.

Tendo por objetivo estimar o número total de habitantes em uma cidade por meio do plano de BE, a variável de interesse gerada representa o número de residentes em uma moradia. Já para o plano de BE estratificado, a variável de interesse é subdividida de modo que cada estrato contenha apenas as moradias com uma quantidade X de residentes. A utilização da alocação proporcional faz com que, pela lógica, moradias de até quatro pessoas sejam mais frequentes do que aquelas com até oito, sendo, respectivamente, o maior e menor estrato.

4.3 Resultados

Nesta seção foram expostos os resultados obtidos por meio da simulação de MC, como desenvolvido nas seções 4.1 e 4.2 Para tal, foi utilizado o *software* R, versão 3.4.2. Um adendo a simulação desenvolvida, é que existe um pacote no R denominado "TeachingSampling" que possibilita aplicar o plano de BE para o estimador de HT, fornecendo como variáveis de entrada o vetor da variável de interesse e a probabilidade de inclusão. Este pacote, além de permitir selecionar amostras probabilísticas, possibilita inferências de uma população finita com base em vários planos de amostragem. Para mais detalhes ver (ROJAS, 2014).

Para dar início ao estudo dos resultados, como forma de análise descritiva, a seguir encontram-se figuras referentes ao histograma da densidade dos estimadores utilizados na pesquisa. Estes histogramas foram gerados segundo 10000 réplicas de MC.

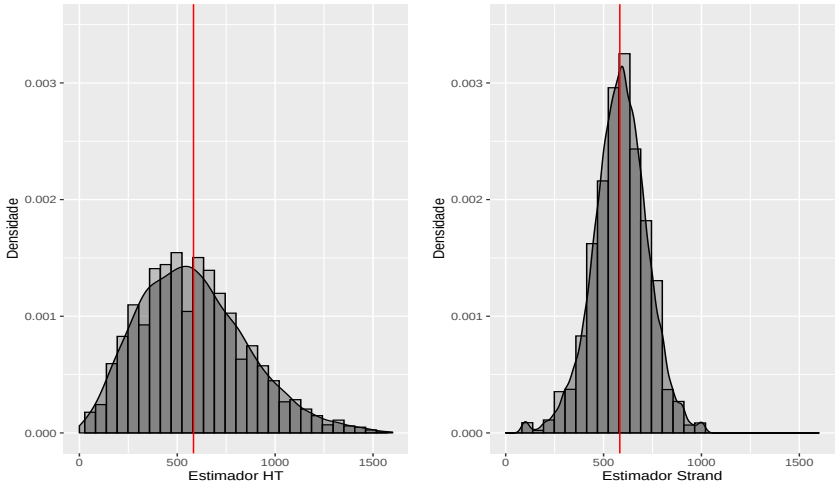

Figura 1: Valores dos estimadores de HT ($\hat{t}_\pi$) e de Strand ($\hat{Y}_1$) para $n = 5$, $N = 100$ e $\pi = 0.05$.

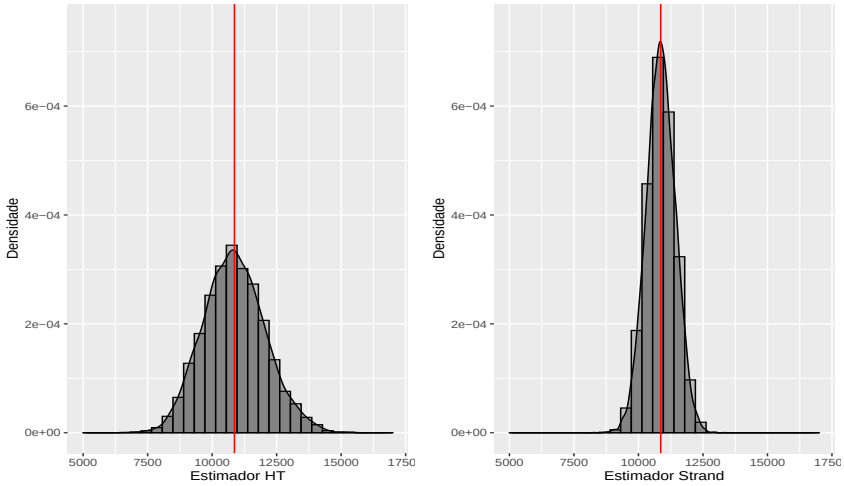

Figura 2: Valores dos estimadores de HT ($\hat{t}_\pi$) e de Strand ($\hat{Y}_1$) para $n = 100$, $N = 2000$ e $\pi = 0.05$.

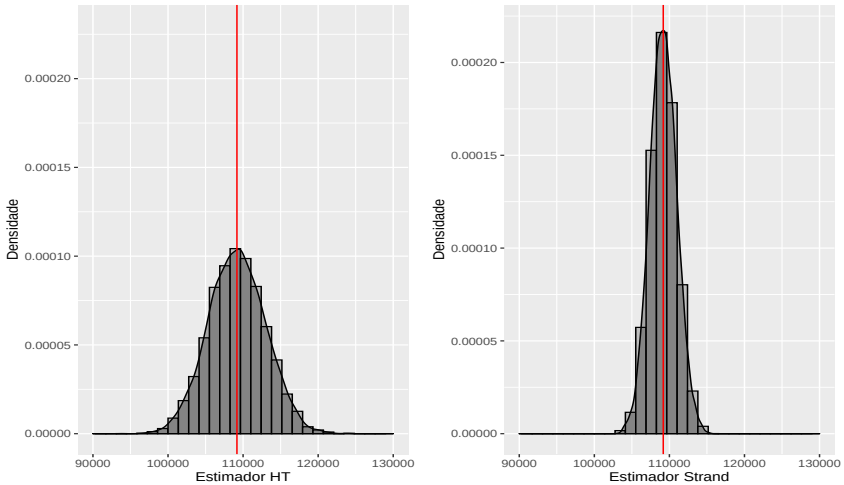

Figura 3: Valores dos estimadores de HT ($\hat{t}_\pi$) e de Strand ($\hat{Y}_1$) para $n = 1000$, $N = 20000$ e $\pi = 0.05$.

A partir dos histogramas, nota-se que há uma maior dispersão das estimativas para o estimador $\hat{t}_\pi$ em confronto ao estimador $\hat{Y}_1$. A linha que divide o gráfico representa o verdadeiro valor do parâmetro. Considerando um número razoável de réplicas de MC, é visível que os resultados dos estimadores para o total populacional se aproximam de uma distribuição simétrica como a Normal, pelo teorema central do limite.

4.3.1 Resultados para a Amostragem de Bernoulli

No cenário amostral de Bernoulli, o resultado das medidas de avaliação podem ser observados por meio das Tabelas 4.1 a 4.3. Com isso, é possível realizar a avaliação e comparação dos estimadores para o total populacional abordados neste trabalho.

Na Tabela 4.1, pode-se observar o EQM dos estimadores, segundo 4.3. Tomando por base essa medida de avaliação, é possível avaliar a eficácia dos estimadores do total populacional $\hat{t}_\pi$ e $\hat{Y}_1$. Com base nos resultados apresentados, pode-se averiguar que o estimador mais eficaz, no que diz respeito ao EQM, é o estimador $\hat{Y}_1$. Este, por sua vez, se sai melhor em todas as cinco conjunturas levantadas, sendo inferior ao $\hat{t}_\pi$ em todas essas. Isso se dá devido à baixa variabilidade do estimador $\hat{Y}_1$, como ficou evidenciado anteriormente por meio dos histogramas.

Também é possível averiguar, por meio desta tabela, que à medida que se aumenta o tamanho populacional há um aumento no erro quadrático de ambos estimadores. O mesmo ocorre para os tamanhos amostrais abordados.

Na Tabela 4.2, é possível observar o viés relativo dos estimadores para o total populacional, a partir do cenário abordado. Em um primeiro momento, é notável que os estimadores $\hat{t}_\pi$ e $\hat{Y}_1$ possuem o viés relativo abaixo de 1%, ratificando que são estimadores aproximadamente não viesados para o total populacional, como era de se esperar.

Apesar do estimador $\hat{Y}_1$ possuir o viés relativo menor que o estimador $\hat{t}_\pi$, em algumas conjunturas, como por exemplo, tomando $n = 100$ e $N = 100000$, bem como $n = 500$ e $N = 500000$, o estimador $\hat{t}_\pi$ se saiu melhor que o $\hat{Y}_1$. Isto também ocorre quando tomamos $n = 5$ e $N = 5000$, podendo indicar que, ao assumir probabilidades de inclusão π muito pequenas o estimador $\hat{t}_\pi$ poderá ter um viés relativo menor que o estimador $\hat{Y}_1$.

Quanto à Tabela 4.3, observam-se os valores resultantes do coeficiente de variação dos estimadores, que tem por objetivo a inferência sobre a estabilidade dos mesmos, pois quanto menor o coeficiente, mais estável é o estimador. Portanto, verifica-se que o estimador $\hat{Y}_1$ é o mais estável e que seu coeficiente diminui à medida que o tamanho amostral aumenta.

Também é possível identificar que os coeficientes do estimador $\hat{t}_\pi$ diminuem conforme se aumenta o tamanho amostral. Os estimadores apresentaram coeficientes menores quanto maior a probabilidade de inclusão. Devido à dispersão observada por meio do coeficiente, não é aconselhável utilizar estes estimadores para um tamanho amostral de $n = 5$.

Tabela 1: Erro quadrático médio dos estimadores $\hat{t}_\pi$ e $\hat{Y}_1$ utilizando um plano de BE.

n	N	π	Estimadores	
			$\hat{t}_\pi$	$\hat{Y}_1$
	10	0.500	568.743	135.7518
	25	0.200	4961.536	1170.247
5	100	0.050	79123.75	19166
	500	0.010	1857690	518620.7
	5000	0.001	180349287	51165822
	100	0.500	4241.575	782.6659
	250	0.200	40292.02	8283.285
50	1000	0.050	733281.2	163596.8
	5000	0.010	18623629	4227759
	50000	0.001	1883714328	422689374
	200	0.500	8296.222	1633.862
	500	0.200	77144.53	15864.87
100	2000	0.050	1436857	316671.9
	10000	0.010	37658613	8155426
	100000	0.001	3791913346	823533324
	1000	0.500	39358.39	8316.624
	2500	0.200	364857	84124.84
500	10000	0.050	7031588	1549361
	50000	0.010	191856670	40277728
	500000	0.001	19025236337	4054386147
	2000	0.500	76700.51	16501.26
	5000	0.200	754107.2	168049.1
1000	20000	0.050	14357975	3182644
	100000	0.010	374061764	81277862
	1000000	0.001	38649825602	8351084769

Tabela 2: Viés relativo dos estimadores $\hat{t}_\pi$ e $\hat{Y}_1$ utilizando um plano de BE

n	N	π	Estimadores	
			$\hat{t}_\pi$	$\hat{Y}_1$
	10	0.500	0.00771206	0.0006180496
	25	0.200	0.007560976	0.001838868
5	100	0.050	0.003880828	0.003775704
	500	0.010	0.00063023	0.003515832
	5000	0.001	0.001557554	0.007139337
	100	0.500	0.0005492281	0.0000075769
	250	0.200	0.0003239943	0.0003721825
50	1000	0.050	0.001450009	0.0004449264
	5000	0.010	0.001116939	0.0003443828
	50000	0.001	0.00117865	0.0008817949
	200	0.500	0.0001903593	0.0001124902
	500	0.200	0.0004712023	0.0002798792
100	2000	0.050	0.0003934094	0.00002355497
	10000	0.010	0.0006593085	0.00005158665
	100000	0.001	0.0003965271	0.0005326101
	1000	0.500	0.0005294074	0.0002739164
	2500	0.200	0.000200783	0.0002090275
500	10000	0.050	0.0002515696	0.0001816903
	50000	0.010	0.0006931966	0.0003940657
	500000	0.001	0.0000912358	0.0003905552
	2000	0.500	0.0002289396	0.0002449205
	5000	0.200	0.0004510416	0.0002975343
1000	20000	0.050	0.0001855102	0.00005075545
	100000	0.010	0.00007459745	0.00006044363
	1000000	0.001	0.0001223449	0.00003238309

Tabela 3: Coe iciente de variação dos estimadores $\hat{t}_\pi$ e $\hat{Y}_1$ utilizando um plano de BE

n	N	π	Estimadores	
			$\hat{t}_\pi$	$\hat{Y}_1$
	10	0,500	34.28975	16.89621
	25	0,200	43.41562	21.20795
5	100	0,050	48.06054	23.65405
	500	0,010	49.0321	26.01251
	5000	0,001	49.52526	26.60055
	100	0,500	11.16481	4.798693
	250	0,200	14.4248	6.540587
50	1000	0,050	15.73152	7.419793
	5000	0,010	15.92151	7.5971
	50000	0,001	15.87553	7.504442
	200	0,500	7.981245	3.542183
	500	0,200	9.993346	4.532692
100	2000	0,050	11.03787	5.179945
	10000	0,010	11.22492	5.226918
	100000	0,001	11.23227	5.233593
	1000	0,500	3.637196	1.672323
	2500	0,200	4.462947	2.142939
500	10000	0,050	4.8524	2.277868
	50000	0,010	5.0567	2.317497
	500000	0,001	5.019213	2.316019
	2000	0,500	2.548546	1.181869
	5000	0,200	3.208611	1.514297
1000	20000	0,050	3.469898	1.633905
	100000	0,010	3.528998	1.645215
	1000000	0,001	3.576459	1.662208

4.3.1 Resultados para a Amostragem Estratificada de Bernoulli

Com base no cenário amostral estratificado de Bernoulli, os resultados obtidos por meio das simulações estão dispostos nas Tabelas 4.4 a 4.6. Desse modo, por meio das medidas de avaliação é factível a análise e comparação dos estimadores $\hat{t}_\pi$ e $\hat{Y}_1$, quando se é utilizado um plano estratificado de BE na seleção amostral.

Na Tabela 4.4, observam-se os resultados referentes à eficácia dos estimadores quanto ao EQM. Nota-se que os estimadores $\hat{t}_\pi$ e $\hat{Y}_1$ apresentam resultados análogos ao cenário anterior, visto que o estimador $\hat{Y}_1$ é mais eficaz que o $\hat{t_\pi}$ em todos os estratos e tamanhos amostrais. Para ambos estimadores, quanto maior o tamanho amostral do estrato, menos eficazes são. E à medida que o tamanho amostral cresce a eficácia aumenta.

Na Tabela 4.5, é possível verificar que os estimadores são aproximadamente não viesados quanto ao total populacional. De mesmo modo que o cenário anterior, o estimador $\hat{Y}_1$ possui viés relativo inferior ao do estimador $\hat{t}_\pi$. Também é percebível que quanto maior o tamanho do estrato, menor é o viés dos estimadores.

Já na Tabela 4.6, em todas as conjunturas abordadas, o estimador $\hat{Y}_1$ apresentou o menor coeficiente, indicando uma homogeneidade no estimador. Também é notável que, conforme o tamanho amostral aumenta, menor é o coeficiente dos estimadores. De forma geral, os maiores estratos apresentam menores coeficientes, ou seja, nestes estratos os estimadores são mais homogêneos.

Os resultados obtidos a partir deste cenário, corroboraram com os obtidos no cenário anterior. Isso sugere que o estimador proposto por Strand (1979) é o estimador mais adequado para se trabalhar em um plano amostral de BE, com ou sem estratificação, em confronto com estimador de HT.

Tabela 4: Erro quadrático médio para os estimadores $\hat{t}_\pi$ e $\hat{Y}_1$ aplicados a uma amostragem estratificada de BE.

n	Estimadores	Estratos			Total
		h_1	h_2	h_3	
50	$\hat{t}_\pi$	10259177	6215439	4924256	21396734
	$\hat{Y}_1$	1794052	1291107	1316374	4401448
100	$\hat{t}_\pi$	5070494	3056682	2559471	10686134
	$\hat{Y}_1$	865200.8	595489.7	583412.5	2044046
500	$\hat{t}_\pi$	953521	569440.7	487520.8	2010410
	$\hat{Y}_1$	166056.2	107913.5	103589.4	377608.2
1000	$\hat{t}_\pi$	452436.8	278159.2	227919.6	958527.7
	$\hat{Y}_1$	77909.24	52576.22	48732.72	179217.1
5000	$\hat{t}_\pi$	51010.43	31493.33	25275.41	107783.3
	$\hat{Y}_1$	8773.709	5826.605	5349.037	19950.04

Tabela 5: Viés relativo dos estimadores $\hat{t}_\pi$ e $\hat{Y}_1$ aplicados a uma amostragem estrati icada de BE.

n	Estimadores	Estratos			Total
		h_1	h_2	h_3	
50	$\hat{t}_\pi$	0.000218826	0.002394604	0.01214832	0.001164754
	$\hat{Y}_1$	0.000003828	0.001153417	0.00117055	0.0001048577
100	$\hat{t}_\pi$	0.001238856	0.001991287	0.001448014	0.0004752009
	$\hat{Y}_1$	0.0002948755	0.0007984306	0.0002356093	0.0000569393
500	$\hat{t}_\pi$	0.0005058469	0.0008288364	0.001199018	0.0002844971
	$\hat{Y}_1$	0.0000563227	0.0004019725	0.001265506	0.0003292238
1000	$\hat{t}_\pi$	0.000178129	0.0005581788	0.0001970103	0.0002135633
	$\hat{Y}_1$	0.000041690	0.0003145046	0.0001576177	0.0000770399
5000	$\hat{t}_\pi$	0.000001447262	0.0001716133	0.0003950022	0.0001027501
	$\hat{Y}_1$	0.000002530763	0.0000935951	0.0001346814	0.0000424615

Tabela 6: Coe iciente de variação dos estimadores $\hat{t}_\pi$ e $\hat{Y}_1$ aplicados a uma amostragem estratificada de BE.

n	Estimadores	Estratos			Total
		h_1	h_2	h_3	
50	$\hat{t}_\pi$	18.54633	35.1181	48.90165	16.00221
	$\hat{Y}_1$	7.754009	15.98583	25.5685	7.267182
100	$\hat{t}_\pi$	13.01892	24.61731	35.64275	11.31679
	$\hat{Y}_1$	5.383082	10.85266	17.04582	4.95214
500	$\hat{t}_\pi$	5.649827	10.61293	15.55935	4.90947
	$\hat{Y}_1$	2.358897	4.614354	7.170825	2.127399
1000	$\hat{t}_\pi$	3.894564	7.415509	10.64959	3.391639
	$\hat{Y}_1$	1.615914	3.22108	4.924564	1.466133
5000	$\hat{t}_\pi$	1.307486	2.49338	3.545511	1.136938
	$\hat{Y}_1$	0.5422502	1.072543	1.631521	0.4891705

5 Conclusão Geral

Considerando a importância do entendimento do plano amostral de Bernoulli, tendo em vista as possibilidades de inúmeras aplicações, essa pesquisa teve por objetivos **apresentar uma revisão de literatura que resgate a origem histórica do plano amostral de Bernoulli** e **avaliar estatisticamente o desempenho de estimadores para o total populacional, sob esse plano, a partir de uma atualização e ampliação do estudo realizado por Strand (1979)**. A partir disso, serão apresentadas algumas considerações.

Quanto ao primeiro objetivo, com base nos resultados identificados, verificou-se que o plano amostral de Bernoulli foi originalmente apresentado na pesquisa realizada por Goodman (1949). Nessa pesquisa, a inclusão de um elemento em uma amostra se deu por eventos aleatórios de Bernoulli. Além do que, mesmo o referido autor abordando uma problemática de estimação do número total de classes, ao substituirmos o interesse de estimar classes pelo de estimar o total populacional, obtemos o estimador do plano de BE, $\hat{t} = N/n \sum_{j \in U} y_j$. Com isso, este autor foi considerado o precursor deste plano amostral por diversos pesquisadores.

Dando continuidade, foram identificadas algumas pesquisas que fizeram uso do referido plano amostral. Importantes artigos foram encontrados na literatura, a exemplo da pesquisa de 1973, na qual Beckwith aborda um estudo do limite do tamanho amostral para um plano de BE, apresentando uma modificação para o mesmo plano, transformando-o no plano amostral de Poisson. Além da pesquisa realizada por Strand, em 1979, na qual o autor retratou o estudo do plano de BE quanto ao processo de estimação do total populacional, apresentando dois estimadores populacionais e realizando simulação segundo o EQM.

Ademais, vários pesquisadores desenvolveram aplicações envolvendo o método de

Bernoulli, seja tratando-o como instrumento da amostragem ou utilizando o plano de BE para avaliar computacionalmente algum método proposto. A exemplo, tem-se o pesquisa de Cohen (1975), que utilizou o plano para verificar o tamanho e composição demográfica de grupos sociais de orangotangos selvagens e de Haas e König (2004), que aplicaram o plano visando a mineração e exploração massiva de informações armazenadas na *Web*. Logo, com base na revisão bibliográfica realizada, pode-se observar que o plano de BE tem importante papel no campo amostral, tendo em vista que pode ser aplicado em inúmeras áreas da ciência.

Quantoa ao segundo objetivo, com base nos resultados obtidos por meio das análises da simulação, pode-se afirmar que o estimador $\hat{Y}_1$ é o mais adequado para se trabalhar segundo um plano de BE ou em um plano de BE estratificado, pois apresenta menor viés relativo, menor erro quadrático e maior eficiência, ao ser comparado com o estimador $\hat{t}_\pi$.

No primeiro cenário abordado, em que se utiliza o plano de BE padrão, verificou-se que o estimador $\hat{Y}_1$ é o mais eficaz no que se refere ao EQM. O viés relativo dos estimadores, no geral, foram inferiores a 1%, sendo todos não viesados. Já para o coeficiente de variação, o estimador $\hat{Y}_1$ é visto como o mais estável.

Quanto ao cenário estratificado de BE, de mesmo modo que foi visto no cenário anterior, apresentou como estimador mais eficaz o $\hat{Y}_1$. De forma geral, ambos os estimadores são não viesados. Quanto ao coeficiente de variação, à medida que o tamanho do estrato aumenta, menor é o coeficiente. Novamente o estimador $\hat{Y}_1$ apresenta o menor valor, sendo este o mais homogêneo.

Com base nos resultados obtidos, é possível afirmar que os estimadores apresentam comportamento similar nos dois cenários, cabendo ao pesquisador escolher o cenário que melhor se adeque a sua situação. Levando em conta a estimativa do total populacional, dentre os estimadores aqui abordados, o melhor estimador seria o $\hat{Y}_1$, que remete a uma correção do estimador de HT.

Este trabalho retrata a importância da amostragem de Bernoulli, apresentando sua evolução e sua relevância na área de amostragem, desde o seu desenvolvimento até os dias atuais. Sendo de fácil aplicação, este plano pôde fornecer subsídios a outros e fazer parte no desenvolvimento de estimadores, bem como suas modificações possibilitaram novos tipos de análise em diversas áreas da ciência.

Como trabalhos futuros, sugere-se o estudo de novos cenários e da utilização de outros estimadores para fins diferentes ao da finalidade adotada nesta pesquisa. Também pode-se sugerir a utilização de outras formas de alocação para a amostragem estratificada, bem como o cálculo da taxa de cobertura dos estimadores fazendo uso de um intervalo de confiança bootstrap. Como também, baseando-se nos trabalhos expostos, existem diversas possibilidades de pesquisas que poderiam ser implementadas e abordadas.

Referências

BECKWITH, R. E. Bounds on sample size in modified Bernoulli sampling, with applications in opinion surveys. *Decision Sciences*, Wiley Online Library, v. 4, n. 1, p. 31–43, 1973.

BERG, S. Sampling from a population in which some units are duplicated and/or overlap. *Scandinavian Actuarial Journal*, Taylor & Francis, v. 1977, n. 4, p. 188–202, 1977.

BOLFARINE, H.; BUSSAB, W. O. *Elementos de amostragem*. São Paulo: Ed. Edgard Blucher, 2005. ISBN 85-212-0367-5.

BUNGE, J.; FITZPATRICK, M. Estimating the number of species: a review. *Journal of the American Statistical Association*, Taylor & Francis, v. 88, n. 421, p. 364–373, 1993.

COHEN, J. E. The size and demographic composition of social groups of wild orangutans. *Animal behaviour*, Elsevier, v. 23, p. 543–550, 1975.

DE, M.; SENGUPTA, S. Bernoulli sequential estimation of the size of a finite population. *Sequential Analysis*, Taylor & Francis, v. 22, n. 1-2, p. 95–106, 2003.

DESHMUKH, S. R. Bernoulli sampling. *Australian Journal of Statistics*, Wiley Online Library, v. 33, n. 2, p. 167–176, 1991.

DUCHESNE, P. Estimation of a proportion with survey data. *Journal of Statistics Education*, Taylor & Francis, v. 11, n. 3, 2003.

FRANK, O. *Statistical inference in graphs*. [S.l.]: FOA Repro, 1971. 280–284 p.

FRANK, O. Estimation of graph totals. *Scandinavian Journal of Statistics*, JSTOR, v. 4, n. 2, p. 81–89, 1977.

FRANK, O. A note on Bernoulli sampling in graphs and Horvitz-Thompson estimation. *Scandinavian Journal of Statistics*, JSTOR, v. 4, n. 4, p. 178–180, 1977.

FRANK, O. Survey sampling in graphs. *Journal of Statistical Planning and Inference*, Elsevier, v. 1, n. 3, p. 235–264, 1977.

FRANK, O. Estimation of the number of connected components in a graph by using a sampled subgraph. *Scandinavian Journal of Statistics*, JSTOR, v. 5, n. 4, p. 177–188, 1978.

GHOSH, D.; VOGT, A. Sampling methods related to Bernolli and Poisson sampling. In: AMERICAN STATISTICAL ASSOCIATION ALEXANDRIA, VA. *Proceedings of the Joint Statistical Meetings*. [S.l.], 2002. p. 3569–3570.

GOODMAN, L. A. On the estimation of the number of classes in a population. *The Annals of Mathematical Statistics*, Institute of Mathematical Statistics, v. 20, n. 4, p. 572–579, 1949.

GUPTA, M. K. Unbiased estimate of l/p. *Annals of the Institute of Statistical Mathematics*, v. 27, n. 1, p. 245–258, 1975.

HAAS, P. J.; KÖNIG, C. A bi-level Bernoulli scheme for database sampling. In: ACM. *Proceedings of the 2004 ACM SIGMOD international conference on Management of data*. [S.l.], 2004. p. 275–286.

HAAS, P. J.; STOKES, L. Estimating the number of classes in a finite population. *Journal of the American Statistical Association*, Taylor & Francis, v. 93, n. 444, p. 1475–1487, 1998.

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.

KAWAKAMI, S.; SASAKI, T.; KOASHI, M. Finite-key analysis for quantum key distribution with weak coherent pulses based on Bernoulli sampling. *Physical Review A*, APS, v. 96, n. 1, p. 012305, 2017.

KINGMAN, J. F. C. Poisson counts for random sequences of events. *Ann. Math. Statist.*, The Institute of Mathematical Statistics, v. 34, n. 4, p. 1217–1232, 12 1963.

LI, J. *et al.* Bernoulli sampling based (epsilon, delta)-approximate frequency query in mobile ad hoc networks. In: SPRINGER. *International Conference on Wireless Algorithms, Systems, and Applications*. [S.l.], 2015. p. 315–324.

MACKINNON, J. The behaviour and ecology of wild orangutans (pongo pygmaeus). *Animal behaviour*, Elsevier, v. 22, n. 1, p. 3–74, 1974.

ROJAS, H. Teachingsampling: Selection of samples and parameter estimation in finite population. *R package version*, v. 3, n. 1, 2014.

RONDON, L. M.; VANEGAS, L. H.; FERRAZ, C. Finite population estimation under generalized linear model assistance. *Computational Statistics & Data Analysis*, Elsevier, v. 56, n. 3, p. 680–697, 2012.

SÄRNDAL, C.-E. Efficient estimators with simple variance in unequal probability sampling. *Journal of the American Statistical Association*, Taylor & Francis, v. 91, n. 435, p. 1289–1300, 1996.

SÄRNDAL, C.-E.; SWENSSON, B.; WRETMAN, J. *Model assisted survey sampling*. [S.l.]: Springer Science & Business Media, 2003.

SINHA, B. K.; BOSE, A. Unbiased sequential estimation of 1/p: settlement of a conjecture. *Annals of the Institute of Statistical Mathematics*, v. 37, n. 3, p. 455–460, 1985.

SINHA, B. K.; SINHA, B. K. Some problems of unbiased sequential binomial estimation. *Annals of the Institute of Statistical Mathematics*, Springer, v. 27, n. 1, p. 245–258, 1975.

SKINNER, C. J.; ELLIOT, M. A measure of disclosure risk for microdata. *Journal of the Royal Statistical Society: series B (statistical methodology)*, Wiley Online Library, v. 64, n. 4, p. 855–867, 2002.

STRAND, M. M. Estimation of a population total under a "Bernoulli sampling" procedure. *The American Statistician*, Taylor & Francis, v. 33, n. 2, p. 81–84, 1979.

WIUF, C. Some properties of the conditioned reconstructed process with Bernoulli sampling. *Theoretical population biology*, Elsevier, v. 122, p. 36–45, 2018.

WRIGHT, T. An inverse sampled Bernoulli (ISB) procedure for estimating a population proportion, with nuclear material applications. *American Journal of Mathematical and Management Sciences*, v. 2, n. 2, p. 123–134, 1982.

Apêndice A - Script para o plano BE

```
 # Programa de simulacao para a avaliacao dos estimadores sob BE#

library(ggplot2)

rm(list = ls())
set.seed(1994)

N  <- 10000  #Tamanho da populacao
n  <- 100    #Tamanho da amostral
r  <- 10000  #Numero de replicas

y     <- sample(1:10, N, replace=TRUE) #variavel de interesse
K     <- sum(y)                        #total populacional
ybaru <- mean(y)                       #media populacional
k     <- seq(1:N)                      #sequencia p/ calculo do H

# Matrizes que para armazenamento #

EHT  <- matrix(0,r,1)
EY1  <- matrix(0,r,1)

VEHT  <- matrix(0,r,1)
VEY1  <- matrix(0,r,1)

IC <- matrix(0,2,2)

TCEHT <- matrix(0,r,1)
TCEY1 <- matrix(0,r,1)

##### laco de monte carlo #####

for(i in 1:r)
{
P    <- n/N                    #pi
U    <- runif(N)               #geracao da uniforme para cada elemento da populacao
Ik   <- ifelse(U <= P, 1, 0)    #selecao da amostra
ns   <- sum(Ik)                #tamanho amostral
s1   <- (1:N)[Ik == 1]         #observacoes que fazem parte da amostra
y_k  <- y[s1]                  #valores de interesse referentes as observacoes

Q    <- 1-P                                                  #1-pi
S2yU <- (1/(N-1))*sum((y_k-ybaru)^2)  #calculo da variancia

###### Estimadores #######

## Estimador de Horvitz-Thompson

EHT[i,1] <- (1/P)*sum(y_k)
```

```
## Estimador Y1 Strand

EY1[i,1] <- N/ns*sum(y_k)

###### Estimadores da variancia #######

## Estimador da variancia de Horvitz-Thompson

VEHT[i,1] <- 1/P*(1/P-1)*sum(y_k^2)

## Estimador da variancia de Y1 Strand

VEY1[i,1] <- N^2*(1/n-1/N)*S2yU

####### Intervalo de confianca #########

EHT[EHT==0] <- NaN  #em casos que nao se obtem amostra este est. resultava 0.

                    #Agora o retorno dele e igual ao est. Y1.

var <- cbind(VEHT[i,1], VEY1[i,1])
est <- cbind(EHT[i,1], EY1[i,1])

IC[,1] <- t(est - 1.96*sqrt(var))
IC[,2] <- t(est + 1.96*sqrt(var))

## Taxa de cobertura

TCEHT[i,1] <- (K>=IC[1,1])*(K<=IC[1,2])
TCEY1[i,1] <- (K>=IC[2,1])*(K<=IC[2,2])
}

###### Medias a partir de MC ######

MEHT  <- mean(EHT[,1],  na.rm = TRUE)
MEY1  <- mean(EY1[,1],  na.rm = TRUE)

###### Variancias a partir de MC ######

VARHT  <- var(EHT[,1],  na.rm = TRUE)
VARY1  <- var(EY1[,1],  na.rm = TRUE)

###### Erro quadratico medio ######
EQMHT  <- VARHT + (MEHT-K)^2
EQMY1  <- VARY1 + (MEY1-K)^2

###### Vies Relativo dos estimadores ######

VIEHT  <- abs(MEHT-K)/K
VIEY1  <- abs(MEY1-K)/K

###### Coeficiente de variacao ######

CVHT  <- 100*(sqrt(VARHT)/MEHT)
CVY1  <- 100*(sqrt(VARY1)/MEY1)

###### Taxa de cobertura ######

TC <- cbind(mean(TCEHT[,1], na.rm = TRUE), mean(TCEY1[,1], na.rm = TRUE))

sink("C://Users//Pichau//Desktop//Dissertacao//Simulacao//C1_BE.txt")

name1   <- cbind('HT', 'Y1')
medida1 <- cbind(EQMHT, EQMY1)
try("ERRO QUADRaTICO MeDIO")
cbind(t(name1),t(medida1))

medida2 <- cbind(VIEHT, VIEY1)
try("VIeS RELATIVO DOS ESTIMADORES")
cbind(t(name1),t(medida2))
```

```
medida3 <- cbind(CVHT, CVY1)
try("COEFICIENTE DE VARIAcaO")
cbind(t(name1),t(medida3))

medida4 <- TC
try("TAXA DE COBERTURA")
cbind(t(name1),t(medida4))

sink()

#### taxa de amostras nao obtidas ####
colSums(VEHT=="NaN")/r*100
colSums(VEY1=="NaN")/r*100

#### geracao de histograma dos estimadores ####

# se faz necessario omitir os erros de estimacao mediante a ns=0
EHT <- na.omit(EHT)
EY1 <- na.omit(EY1)

# para gerar o histograma se faz necessario um data.frame
estm <- data.frame(EHT, EY1)

ggplot(estm, aes(x=EHT)) +
   geom_histogram(aes(x=EHT), color="black", fill="gray") +
     labs(x = "Estimador HT", y = "Frequencia") +
       geom_vline(xintercept=mean(EHT), color = "red")

 ggplot(estm, aes(x=EY1)) +
   geom_histogram(aes(x=EY1), color="black", fill="gray") +
   labs(x = "Estimador Strand", y = "Frequencia")+
      geom_vline(xintercept=mean(EY1), color = "red")
```

Apêndice B - Script para o plano de BE estratificado

```
# Programa de simulacao para a avaliacao dos estimadores sob BE estratificada#

rm(list = ls())
set.seed(1994)

N1  <- 7000            #Tamanho do estrato 1
N2  <- 2000            #Tamanho do estrato 2
N3  <- 1000            #Tamanho do estrato 3
N   <- N1+N2+N3        #Tamanho da populacao
n   <- 100             #Tamanho da amostral
n1  <- ceiling(n*N1/N) #Tamanho da amostra no estrato 1
n2  <- ceiling(n*N2/N) #Tamanho da amostra no estrato 2
n3  <- ceiling(n*N3/N) #Tamanho da amostra no estrato 3
r   <- 10000           #Numero de replicas

h     <- sample(1:4, N1, replace=TRUE) #variavel de interesse estrato 1
j     <- sample(1:6, N2, replace=TRUE) #variavel de interesse estrato 2
l     <- sample(1:8, N3, replace=TRUE) #variavel de interesse estrato 3

h1    <- sum(h)                        #total populacional estrato 1
j1    <- sum(j)                        #total populacional estrato 2
l1    <- sum(l)                        #total populacional estrato 3
K     <- h1+j1+l1

ybaruh <- mean(h)                      #media populacional estrato 1
ybaruj <- mean(j)                      #media populacional estrato 2
ybarul <- mean(l)                      #media populacional estrato 3

k1     <- seq(1:N1)                    #sequencia p/ calculo do H estrato 1
k2     <- seq(1:N2)                    #sequencia p/ calculo do H estrato 2
k3     <- seq(1:N3)                    #sequencia p/ calculo do H estrato 3

# Matrizes que para armazenamento #

EHT  <- matrix(0,r,3)
EY1  <- matrix(0,r,3)

VEHT  <- matrix(0,r,3)
VEY1  <- matrix(0,r,3)

IC1 <- matrix(0,2,2)
IC2 <- matrix(0,2,2)
IC3 <- matrix(0,2,2)
IC  <- matrix(0,2,2)

TCEHT  <- matrix(0,r,3)
TCEY1  <- matrix(0,r,3)
TCEHT1 <- matrix(0,r,3)
```

```
TCEY11 <- matrix(0,r,3)
TCEHT2 <- matrix(0,r,3)
TCEY12 <- matrix(0,r,3)
TCEHT3 <- matrix(0,r,3)
TCEY13 <- matrix(0,r,3)

##### laco de monte carlo #####

for(i in 1:r)
{
  P1    <- n1/N1                  #pi p/ estrato 1
  P2    <- n2/N2
  P3    <- n3/N3

  U1    <- runif(N1)              #geracao da uniforme para cada elemento da populacao no estrato 1
  U2    <- runif(N2)
  U3    <- runif(N3)

  Ik1   <- ifelse(U1 <= P1, 1, 0) #selecao da amostra no estrato 1
  Ik2   <- ifelse(U2 <= P2, 1, 0)
  Ik3   <- ifelse(U3 <= P3, 1, 0)

  ns1   <- sum(Ik1)               #tamanho amostral do estrato 1
  ns2   <- sum(Ik2)
  ns3   <- sum(Ik3)

  s1   <- (1:N1)[Ik1 == 1]        #observacoes que fazem parte da amostra no estrato 1
  s2   <- (1:N2)[Ik2 == 1]
  s3   <- (1:N3)[Ik3 == 1]

  y_k1  <- h[s1]                  #valores de interesse referentes as observacoes do estrato 1
  y_k2  <- j[s2]
  y_k3  <- l[s3]

  S2yU1 <- 1/(N1-1)*sum((y_k1-mean(h, na.rm = TRUE))^2, na.rm = TRUE)  #calculo da variancia
  S2yU2 <- 1/(N2-1)*sum((y_k2-mean(j, na.rm = TRUE))^2, na.rm = TRUE)
  S2yU3 <- 1/(N3-1)*sum((y_k3-mean(l, na.rm = TRUE))^2, na.rm = TRUE)

  ###### Estimadores #######

  ## Estimador de Horvitz-Thompson

  EHT[i,1] <- (1/P1)*sum(y_k1)
  EHT[i,2] <- (1/P2)*sum(y_k2)
  EHT[i,3] <- (1/P3)*sum(y_k3)

  ## Estimador Y1 Strand

  EY1[i,1] <- N1/ns1*sum(y_k1)
  EY1[i,2] <- N2/ns2*sum(y_k2)
  EY1[i,3] <- N3/ns3*sum(y_k3)

  ###### Estimadores da variancia #######

  ## Estimador da variancia de Horvitz-Thompson

  VEHT[i,1] <- 1/P1*(1/P1-1)*sum(y_k1^2, na.rm = TRUE)
  VEHT[i,2] <- 1/P2*(1/P2-1)*sum(y_k2^2, na.rm = TRUE)
  VEHT[i,3] <- 1/P3*(1/P3-1)*sum(y_k3^2, na.rm = TRUE)

  ## Estimador da variancia Y1 Strand

  VEY1[i,1] <- N1^2*(1/n1-1/N1)*S2yU1
  VEY1[i,2] <- N2^2*(1/n2-1/N2)*S2yU2
  VEY1[i,3] <- N3^2*(1/n3-1/N3)*S2yU3

  ####### Intervalo de confianca #########

  EHT[EHT==0] <- NaN  #em casos que nao se obtem amostra este est. resultava 0.
                      #Agora o retorno dele e igual ao est. Y1.

  var1 <- cbind(VEHT[i,1], VEY1[i,1])
  var2 <- cbind(VEHT[i,2], VEY1[i,2])
```

```
  var3 <- cbind(VEHT[i,3], VEY1[i,3])
  var  <- cbind(sum(VEHT[i,1:3]), sum(VEY1[i,1:3]))

  est1 <- cbind(EHT[i,1], EY1[i,1])
  est2 <- cbind(EHT[i,2], EY1[i,2])
  est3 <- cbind(EHT[i,3], EY1[i,3])
  est  <- cbind(sum(EHT[i,1:3]), sum(EY1[i,1:3]))

  IC1[,1] <- t(est1 - 1.96*sqrt(var1))
  IC1[,2] <- t(est1 + 1.96*sqrt(var1))

  IC2[,1] <- t(est2 - 1.96*sqrt(var2))
  IC2[,2] <- t(est2 + 1.96*sqrt(var2))

  IC3[,1] <- t(est3 - 1.96*sqrt(var3))
  IC3[,2] <- t(est3 + 1.96*sqrt(var3))

  IC[,1] <- t(est - 1.96*sqrt(var))
  IC[,2] <- t(est + 1.96*sqrt(var))

  ## Taxa de cobertura

  TCEHT[i,1] <- (K>=IC[1,1])*(K<=IC[1,2])
  TCEY1[i,1] <- (K>=IC[2,1])*(K<=IC[2,2])

  TCEHT1[i,1] <- (h1>=IC1[1,1])*(h1<=IC1[1,2])
  TCEY11[i,1] <- (h1>=IC1[2,1])*(h1<=IC1[2,2])

  TCEHT2[i,1] <- (j1>=IC2[1,1])*(j1<=IC2[1,2])
  TCEY12[i,1] <- (j1>=IC2[2,1])*(j1<=IC2[2,2])

  TCEHT3[i,1] <- (l1>=IC3[1,1])*(l1<=IC3[1,2])
  TCEY13[i,1] <- (l1>=IC3[2,1])*(l1<=IC3[2,2])

}

###### Medias ######

MEHT  <- sum(colMeans(EHT,  na.rm = TRUE))
MEY1  <- sum(colMeans(EY1,  na.rm = TRUE))

###### Variancias ######

VARHT1  <- var(EHT[,1],  na.rm = TRUE)
VARHT2  <- var(EHT[,2],  na.rm = TRUE)
VARHT3  <- var(EHT[,3],  na.rm = TRUE)
VARHT   <- VARHT1+VARHT2+VARHT3

VARY11  <- var(EY1[,1],  na.rm = TRUE)
VARY12  <- var(EY1[,2],  na.rm = TRUE)
VARY13  <- var(EY1[,3],  na.rm = TRUE)
VARY1   <- VARY11+VARY12+VARY13

###### Erro quadratico medio ######

EQMHT1 <- VARHT1+(mean(EHT[,1], na.rm=TRUE)-h1)^2
EQMHT2 <- VARHT2+(mean(EHT[,2], na.rm=TRUE)-j1)^2
EQMHT3 <- VARHT3+(mean(EHT[,3], na.rm=TRUE)-l1)^2
EQMHT  <- VARHT + (MEHT-K)^2

EQMY11 <- VARY11+(mean(EY1[,1], na.rm=TRUE)-h1)^2
EQMY12 <- VARY12+(mean(EY1[,2], na.rm=TRUE)-j1)^2
EQMY13 <- VARY13+(mean(EY1[,3], na.rm=TRUE)-l1)^2
EQMY1  <- VARY1 + (MEY1-K)^2

###### Vies Relativo dos estimadores ######

VIEHT1 <- abs(mean(EHT[,1], na.rm=TRUE)-h1)/h1
VIEHT2 <- abs(mean(EHT[,2], na.rm=TRUE)-j1)/j1
VIEHT3 <- abs(mean(EHT[,3], na.rm=TRUE)-l1)/l1
VIEHT  <- abs(MEHT-K)/K

VIEY11 <- abs(mean(EY1[,1], na.rm=TRUE)-h1)/h1
```

```
VIEY12             <-             abs(mean(EY1[,2],             na.rm=TRUE)-j1)/j1
VIEY13             <-             abs(mean(EY1[,3],             na.rm=TRUE)-l1)/l1
VIEY1  <- abs(MEY1-K)/K

###### Coeficiente de variacao ######

CVHT1            <-           100*(sqrt(VARHT1)/mean(EHT[,1],           na.rm=TRUE))
CVHT2            <-           100*(sqrt(VARHT2)/mean(EHT[,2],           na.rm=TRUE))
CVHT3            <-           100*(sqrt(VARHT3)/mean(EHT[,3],           na.rm=TRUE))
CVHT   <- 100*(sqrt(VARHT)/MEHT)

CVY11           <-           100*(sqrt(VARY11)/mean(EY1[,1],           na.rm=TRUE))
CVY12           <-           100*(sqrt(VARY12)/mean(EY1[,2],           na.rm=TRUE))
CVY13           <-           100*(sqrt(VARY13)/mean(EY1[,3],           na.rm=TRUE))
CVY1  <- 100*(sqrt(VARY1)/MEY1)

###### Taxa de Cobertura ######

TC <-  cbind(mean(TCEHT[,1], na.rm = TRUE), mean(TCEY1[,1], na.rm = TRUE))
TC1 <- cbind(mean(TCEHT1[,1], na.rm = TRUE), mean(TCEY11[,1], na.rm = TRUE))
TC2 <- cbind(mean(TCEHT2[,1], na.rm = TRUE), mean(TCEY12[,1], na.rm = TRUE))
TC3 <- cbind(mean(TCEHT3[,1], na.rm = TRUE), mean(TCEY13[,1], na.rm = TRUE))

EQM1 <- cbind(EQMHT1, EQMHT2, EQMHT3, EQMHT)
EQM2 <- cbind(EQMY11, EQMY12, EQMY13, EQMY1)

VIES1 <- cbind(VIEHT1, VIEHT2, VIEHT3, VIEHT)
VIES2 <- cbind(VIEY11, VIEY12, VIEY13, VIEY1)

CV1 <- cbind(CVHT1,CVHT2, CVHT3, CVHT)
CV2 <- cbind(CVY11,CVY12,CVY13,CVY1)

TXHT <- cbind(TC1[1,1],TC2[1,1],TC3[1,1],TC[1,1])
TXY1 <- cbind(TC1[1,2],TC2[1,2],TC3[1,2],TC[1,2])
```