Universidade de São Paulo
Escola Superior de Agricultura "Luiz de Queiroz"

Estudo do método de estimação do tipo razão multivariada para média populacional

Welinton Yoshio Hirai

Tese apresentada para obtenção do título de Doutor (ou Doutora) em Ciências.
Área de concentração: Estatística e Experimentação Agronômica

Piracicaba
2023

Welinton Yoshio Hirai
Licenciado em matemática

Estudo do método de estimação do tipo razão multivariada para média populacional

Orientadora:
Profa. Dra. **SÔNIA MARIA DE STEFANO PIEDADE**

Tese apresentada para obtenção do título de Doutor (ou Doutora) em Ciências.
Área de concentração: Estatística e Experimentação Agronômica

Piracicaba
2023

Dados Internacionais de Catalogação na Publicação
DIVISÃO DE BIBLIOTECA - DIBD/ESALQ/USP

Hirai, Welinton Yoshio

Estudo do método de estimação do tipo razão multivariada para média populacional / Welinton Yoshio Hirai. – – Piracicaba, 2023 .

72 p.

Tese (Doutorado) – – USP / Escola Superior de Agricultura "Luiz de Queiroz".

1. Amostras 2. Estratos 3. Variável interesse 4. Variáveis auxiliares . I. Título.

DEDICATÓRIA

Ao grande amor da minha vida, Roberta Mitie Matsunaka.
Aos meus pais Shindi e Adriana Hirai.

AGRADECIMENTOS

A minha orientadora professora Sônia Maria De Stefano Piedade, pela paciência, companheirismo e admiração profissional e pessoal.

Aos colegas e amigos que fiz durante o período do Doutorado que ajudaram no meu aprimoramento pessoal e profissional. Em especial a Vivian que sempre me auxiliou e auxilia na minha trajetória.

A todos os professores do Programa de Pós-graduação em Estatística e Experimentação Agronômica (PPGEEA) que construíram na minha formação acadêmica.

A todos os funcionários da PPGEEA, que sempre estão a disposição para ajudar e auxiliar todos os alunos.

Agradecer aos dois pesquisadores que forneceram os dados para as análises. O Dr. João Gabriel Ribeiro do Departamento de Matemática da Universidade do Estado de Mato Grosso (UNEMAT), e ao Dr. Fernando Antonio Souza de Aragão da EMBRAPA em Fortaleza.

E a CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico) que concedeu abolsa, sem a qual seria inviável o desenvolvimento de pesquisas no Brasil

SUMÁRIO

RESUMO

Estudo do método de estimação do tipo razão multivariada para média populacional

Nesse trabalho foram produzidos estudos para um método de estimação da média baseado em mais que uma variável auxiliar, o método da razão multivariada. A abordagem utiliza-se da razão entre a variável de interesse em conjunto com demais variáveis coletadas durante o processo de amostragem para aprimorar o valor preditivo da estimativa. No primeiro capítulo, foi realizada uma comparação entre o método usual, método da razão univariada, e o método da razão multivariada para cenários simulados de alta e baixa correlação entre grupos. No segundo capítulo, foi proposto um novo método de imputação utilização a razão multivariada, estudo das expressões matemáticas para boas propriedades e aplicados em dados simulados e reais. O método se apresenta com bons resultados para estruturas altamente correlacionadas. Os códigos foram implementados na linguagem R e anexadas no trabalho.

Palavras-chave: Estimador pela razão; Amostragem; Imputação; Correlação

ABSTRACT

Study of the multivariate ratio estimation method for population mean

In this work, studies were conducted to develop a mean estimation method based on more than one auxiliary variable, known as the multivariate ratio method. The approach utilizes the ratio between the variable of interest and other variables collected during the sampling process to enhance the predictive value of the estimate.

In the first chapter, a comparison was made between the usual method, the univariate ratio method, and the multivariate ratio method in simulated scenarios with high and low correlation between groups. In the second chapter, a new imputation method using the multivariate ratio was proposed, with a study of mathematical expressions to ensure good properties and application to simulated and real data. The method showed good results, especially in highly correlated structures.

The codes were implemented in the R language and are attached to the work.

Keywords: Ratio estimation; Sampling; Imputation; Correlation

1 INTRODUÇÃO

Machine learning, *deep learning*, *neural network* são palavras-chave que estão crescendo progressivamente em trabalhos acadêmicos e setores empresariais. Esses estudos estão gerando cada vez mais resultados, alavancando várias áreas. Porém, como qualquer novo campo de estudo, ainda existem dificuldades devem ser consideradas para as implementações e aplicações tecnológicas: podemos levantar uma delas a necessidade de um conjunto/banco de dados com grande volume de informações para a concepção de resultados verossímeis. Porém a obtenção do censo, não é realidade em todas as áreas do conhecimento devido às questões estruturais e financeiras que são impossibilitadas de obter informações de todos os elementos constituintes da população de interesse. Um exemplo é no campo de estudo das agrárias na qual muitas pesquisas são realizadas com mão-de-obra limitada, espaço físico definido.

Em virtude disto, faz-se necessário utilizar de uma das áreas do conhecimento estatísticos a Amostragem. Em que parte da população é obtida por meio de amostras, e depois são calculados estimadores gerando resultados verossímeis da população.

Segundo os autores Kruskal and Mosteller (1980), Rao and Bellhouse (1990) e Rao (2005) a primeira vez que utilizou-se o termo "amostra representativa" foi em meados de 1895/96 nos trabalhos do estatístico Kiear, que era diretor da agência central de estatística na Noruega. Kiear foi o primeiro em propor a coleta de amostras em vez de todos os elementos da população, para fins de realização de análises. Mas como relatado por Bodin (2020), Kiear foi criticado na época pelo seus colegas, devido aos receios de que os cálculos não representassem a verdadeira informação da população.

Mesmo assim, Kiear continuou aprimorando suas ideias, principalmente a de melhorar os procedimentos de coleta dessas unidades amostrais. Inicialmente aperfeiçoou o método partindo de características que agrupavam os elementos dentro de uma mesma população, e assim particionando-a em estratos. E desta forma, introduziu a ideia de amostragem estratificada que levava às ponderações, baseado nos tamanhos dos grupos e da população, e atribuiu pesos para os estimadores (Kruskal and Mosteller, 1980).

Com o passar dos anos, as técnicas de amostragem foram se tornando objeto de estudo de muitos outros autores, como Cochran que aprimorou teorias na amostragem como: a demonstração e aplicação do uso da análise de variância (ANOVA) para estimar o ganho de eficiência nas estratificação; o uso de componentes de variância em amostragem de dois estágios; estimação por regressão sobre amostras de dois estágios; e o estudo do efeito do erro em tamanhos de estratos (Rao and Bellhouse, 1990; Rao, 2005). Segundo, Rao and Bellhouse (1990) apresentou o conceito de super população, em que a própria população finita deve ser considerada como um plano amostral vinda de uma população infinita, criticando assim o conceito de população "fixa".

Em 1940, Cochran ainda formalizou o conceito de estimação por razão, porém segundo Rao (2005) o primeiro trabalho que apresentou esta ideia foi de Laplace em 1820. Após este trabalho, Cochran em 1942, desenvolveu também o método de estimação por regressão.

Neyman (1992) publicou um trabalho que validou os estimadores de totais, médias e proporções para a população que são obtidas sem depender de uma amostra representativa, por meio de um plano amostral que atribuía seleções com probabilidades equiprováveis. E além disto, introduziu o conceito de função custo para alocação de amostras em duas fases, minimizando assim variabilidades. Como conta Rao and Bellhouse (1990), inspirados por Neyman, posteriormente Horvitz e Thompson em 1952 e Narain em 1951 estenderam o conceito para amostras com probabilidades desiguais e sem reposição.

Na Índia, houve um grande avanço no estudo de amostragem, muito devido ao estatístico Mahalanobis, que em 1937 implementou o método de amostragem por múltiplos estágios para melhorar a produtividade agrícola de diversas culturas em pequenas comunidades. O objetivo foi delinear *grides* dentro de aldeias, depois parcelas dentro destes grides, e consecutivamente definir diferentes tamanhos

para as unidades amostrais, resultando em uma amostragem de 4 estágios (Kruskal and Mosteller, 1980). Em 1944, Mahalanobis destacou-se como o pioneiro no aprimoramento das funções de custo e de variância no planejamento de planos amostral para pesquisas. E assim foi uns dos nomes imprescindíveis para o desenvolvimento do centros de pesquisas estatística na Índia, a *National Sample Survey* (NSS) em conjunto à pesquisadores como D. B. Lahiri e M. N. Murthy.

Muitas pesquisas amostrais são realizadas diversas vezes, gerando coletas durante vários períodos durante um tempo. Pensando nisso autores como Yates et al. (1949), Patterson (1950) e Jessen (1943) introduziram o método de amostragem longitudinal em dois estágios, tendo parcial reposição das unidades, gerando as teorias para o plano amostral e estimação de medida repetidas, corroborando no ganho de eficiência quando se utiliza de informações anteriores, isto é, no tempo passado (Rao, 2005).

Deming (1990) começou a utilizar métodos computacionais para estimativas de variância por meio de réplicas simuladas, como jackknife e bootstrap. O autor aplicava os métodos de amostragem para procedimentos de controle de qualidade nas indústrias.

Nota-se então, que a amostragem abrange desde o planejamento dos processos de coleta dos dados, até os cálculos matemáticos para as análises estatísticas. Esses dois processos levam em consideração, possíveis fatores não controlados (erro), isto é, comportamentos não previstos pelo pesquisador.

Segundo a Figura 1.1 o erro pode ocorrer durante o processo de planejamento amostral desconsiderando possíveis características da população, o que pode acarretar numa coleta de informação viesada, ou seja, pouco representativa. E também no processo de cálculo das expressões (implementações), pois se não considerar todos as pressuposições do planejamento os resultados ocasionam em inferências errôneas.

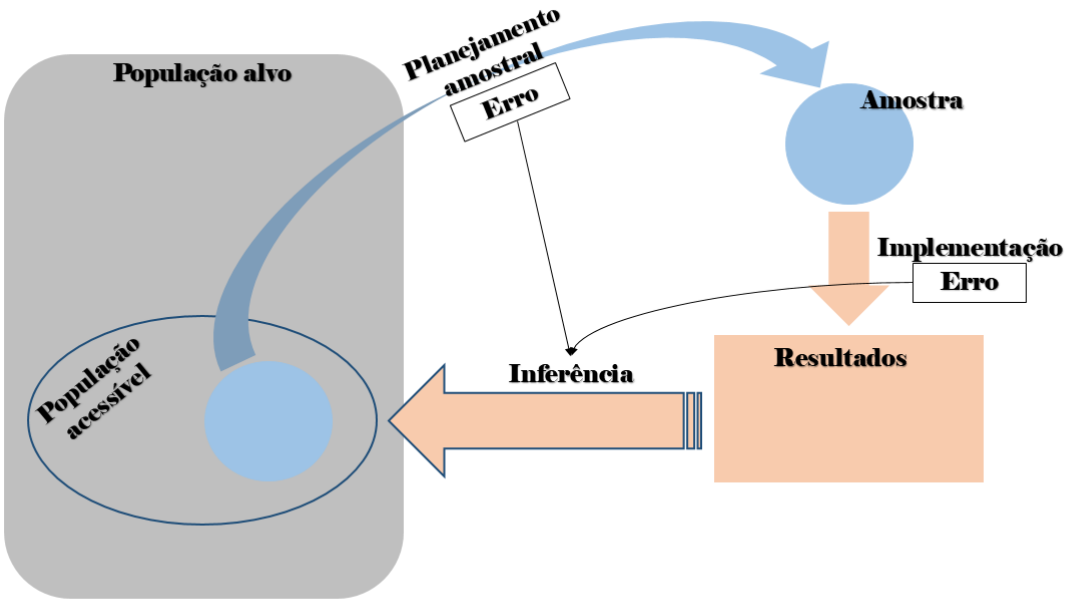


Figura 1.1. Representação do processo de amostragem
Fonte: Próprio autor

Há diferentes tipos de desenhos amostrais para realizar a coleta das informações que possam representar a população. Os métodos mais conhecidos são: a amostragem aleatória simples em que todos os indivíduos têm probabilidades equiprováveis de seleção; amostragem estratificada em que a população é particionada em grupos chamados estratos, fazendo com que cada indivíduo pertencente a seu respectivo estrato possa ser selecionado de forma independente; amostragem por conglomerados quando a população é dividida em grupos coletando todos os indivíduos pertencentes aos grupos; a amostragem sistemática a população é ordenada a partir de uma característica das unidades amostrais e quando selecionado um indivíduo, as próximas unidades são coletados por meio de comprimento.

As teorias de amostragem são fundamentadas a partir dos conceitos de estatística matemática, em que encontra-se expressões/funções que têm como objetivo estimar os parâmetros populacionais utilizando apenas as amostras. Estes cálculos são chamados de estimadores e são avaliado-s por meio propriedades de vieses e variâncias. Durante a inferência, o pesquisador tem o interesse de construir estimadores pontuais e/ou intervalares, levando em consideração a magnitude das fontes de variação não controladas e apresentem inferências verossímeis da população.

Então um dos interesses para a amostragem é conseguir estimadores que apresentem boas características. Somado à isto, existe uma área da estatística que proporciona ao pesquisador coletar resultados por meio de metodologias que utilizam informações conjuntas de mais de uma variável. Esta área é chamada de estatística multivariada.

Existem diversos métodos que utilizam a informação de uma variável auxiliar para aprimorar as estimativas de variáveis de interesse. Como o estimador tipo razão, tipo regressão e tipo diferença. Entretanto para estes estimadores é utilizado apenas uma variável para auxiliar nos cálculos.

Devido as evoluções tecnológicas, às diversas áreas de pesquisas conseguem coletar múltiplas informação de uma única unidade amostral. Como por exemplos nas agrárias: em estudo de solo que uma amostra gera informações físicas e químicas; na pecuária com características físicas do bovino; na climatologia com diversas informações meteorológicas em tempo real; e entre outros.

A partir das metodologias citadas este, trabalho tem como objetivo estudar o método de estimação por razão multivariada proposto por Olkin (1958). No qual, utiliza-se múltiplas variáveis auxiliares para o cálculo do estimador da média populacional. No primeiro capítulo desta tese, estudou-se 3 tipos de estimadores para um plano de amostragem estratificada: o estimador usual; por razão; e razão multivariada. Para isso, simulou-se 4 cenários diferentes de alta e baixa correlação entre as variáveis, e com 5 e 10 estratos.

1.1 Referências

Arnab, R., 2017 Survey Sampling: Theory and Applications. Academic Press, first edition.

Beunckens, C., C. Sotto, G. Molenberghs, and G. Verbeke, 2009 A multifaceted sensitivity analysis of the Slovenian public opinion survey data. JOURNAL OF THE ROYAL STATISTICAL SOCIETY SERIES C-APPLIED STATISTICS **58**: 171–196.

Bodin, J.-L., 2020 A view on 50 years of life of the ISI: With a focus on ISI relations with official statistics. Statistical Journal of the IAOS **36**: 303–308.

Bolfarine, H. and W. O. Bussab, 2004 Elementos de amostragem. Editora Blucher.

Brand, M., 2002 Incremental singular value decomposition of uncertain data with missing values. In COMPUTER VISON - ECCV 2002, PT 1, edited by A. Heyden, G. Sparr, M. Nielsen, and P. Johansen, volume 2350 of Lecture Notes in Computer Science, pp. 707–720, IT Univ Copenhagen; Univ Copenhagen; Lund Univ.

Cavalcanti, P. P. and C. T. d. S. Dias, 2021 Archetypal analysis as an imputation method and multivariate data augmentation .

Cebrián, A. A. and M. R. García, 1997 Variance Estimation Using Auxiliary Information: An Almost Unbiased Multivariate Ratio Estimator. Metrika **45**: 171–178.

Cochran, W. G., 1977 Sampling techniques. John Wiley & Sons, third edition.

Deming, W. E., 1990 Sample design in business research, volume 23. John Wiley & Sons.

Departamento de Astronomia Instituto de Astronomia, G. e. C. A., 2020 Início das estações do ano (2005–2020). Accessed: 2022-03-18.

Diana, G. and P. Francesco Perri, 2010 Improved estimators of the population mean for missing data. Communications in Statistics—Theory and Methods **39**: 3245–3251.

Ellington, E. H., G. Bastille-Rousseau, C. Austin, K. N. Landolt, B. A. Pond, E. E. Rees, N. Robar, and D. L. Murray, 2015 Using multiple imputation to estimate missing data in meta-regression. METHODS IN ECOLOGY AND EVOLUTION **6**: 153–163.

Epifanio, I., M. V. Ibanez, and A. Simo, 2020 Archetypal Analysis With Missing Data: See All Samples by Looking at a Few Based on Extreme Profiles. AMERICAN STATISTICIAN **74**: 169–183.

Gasparetto, S. C., S. M. D. S. Piedade, L. R. Angelocci, and V. A. Ozaki, 2021 COMPARAÇÃO ENTRE MÉTODOS DE IMPUTAÇÃO DE DADOS EM DIFERENTES INTENSIDADES AMOSTRAIS NA SÉRIE DE PRECIPITAÇÃO PLUVIAL DA ESALQ. Revista Brasileira de Climatologia **29**: 464–489.

Goodman, L. A. and H. O. Hartley, 1958 The Precision of Unbiased Ratio-Type Estimators. Journal of the American Statistical Association **53**: 491–508.

Hanif, M., Z. Ahmed, and M. Ahmad, 2009 Generalized multivariate ratio estimators using multi-auxiliary variables for multi-phase sampling. Pakistan Journal of Statistics **25**: 615–629.

He, Y., R. Yucel, and T. E. Raghunathan, 2011 A functional multiple imputation approach to incomplete longitudinal data. STATISTICS IN MEDICINE **30**: 1137–1156.

Howe, L. D., K. Tilling, A. Matijasevich, E. S. Petherick, A. C. Santos, L. Fairley, J. Wright, I. S. Santos, A. J. D. Barros, R. M. Martin, M. S. Kramer, N. Bogdanovich, L. Matush, H. Barros, and D. A. Lawlor, 2016 Linear spline multilevel models for summarising childhood growth trajectories: A guide to their application using examples from five birth cohorts. STATISTICAL METHODS IN MEDICAL RESEARCH **25**: 1854–1874.

Hulley, S. B., S. R. Cummings, W. S. Browner, D. G. Grady, and T. B. Newman, 2013 Designing clinical research. Lippincott Williams & Wilkins, fourth edi edition.

Jessen, R. J., 1943 Statistical investigation of a sample survey for obtaining farm facts. Iowa State University.

Johnson, R. A. and D. W. Wichern, 2007 Applied Multivariate Statistical Analysis. Prentice Hall, 6th edition.

Kenward, M. G. and J. Carpenter, 2007 Multiple imputation: current perspectives. Statistical Methods in Medical Research **16**: 199–218.

Kim, J. and M.-J. Park, 2019 Multiple imputation and synthetic data. KOREAN JOURNAL OF APPLIED STATISTICS **32**: 83–97.

Kruskal, W. and F. Mosteller, 1980 Representative Sampling, IV: The History of the Concept in Statistics, 1895-1939. International Statistical Review / Revue Internationale de Statistique **48**: 169.

Lohr, S. L., 2019 Sampling: Design and Analysis. Advanced (Cengage Learning), Cengage Learning.

Lohr, S. L., 2021 Sampling: design and analysis. CRC press.

Maitra, R., V. Melnykov, and S. N. Lahiri, 2012 Bootstrapping for Significance of Compact Clusters in Multidimensional Datasets. Journal of the American Statistical Association **107**: 378–392.

Melnykov, V., W.-C. Chen, and R. Maitra, 2012 MixSim: An R Package for Simulating Data to Study Performance of Clustering Algorithms. Journal of Statistical Software **51**: 1–25.

Myers, W. R., 2000 Handling Missing Data in Clinical Trials: An Overview. Drug information journal : DIJ / Drug Information Association **34**: 525–533.

Nath, K. and B. K. Sungh, 2018 Population Mean Estimation Using Ratio-cum Product Compromised-method of Imputation in Two-phase Sampling Scheme. Asian J. Math. Stat **11**: 27–39.

Neyman, J., 1992 pp. 123–150 in On the Two Different Aspects of the Representative Method: the Method of Stratified Sampling and the Method of Purposive Selection, edited by Kotz, S. and N. L. Johnson, Springer New York.

Olkin, I., 1958 Multivariate Ratio Estimation for Finite Populations. Biometrika **45**: 154.

Olufadi, Y. and C. Kadilar, 2014 A study on the chain ratio-type estimator of finite population variance. Journal of Probability and Statistics **2014**.

Patterson, H., 1950 Sampling on successive occasions with partial replacement of units. Journal of the Royal Statistical Society. Series B (Methodological) **12**: 241–255.

Peel, D., R. S. Waples, G. M. Macbeth, C. Do, and J. R. Ovenden, 2013 Accounting for missing data in the estimation of contemporary genetic effective population size (N-e). MOLECULAR ECOLOGY RESOURCES **13**: 243–253.

Qiu, W. and H. Joe., 2020 clusterGeneration: Random Cluster Generation (with Specified Degree of Separation). R package version 1.3.7.

Rao, J., 2005 Interplay Between Sample Survey Theory and Practice: An Appraisal. Survey Methodology **31**.

Rao, J. N. K. and D. R. Bellhouse, 1990 History and development of the theoretical foundations of survey based estimation and analysis. Survey Methodology **16**: 3–29.

Rao, J. N. K. and R. R. Sitter, 1995 Variance estimation under two-phase sampling with application to imputation for missing data. Biometrika **82**: 453–460.

Rubin, D. B., H. S. Stern, and V. Vehovar, 1995 Handling "Don't Know" Survey Responses: The Case of the Slovenian Plebiscite. Journal of the American Statistical Association **90**: 822–828.

Sano, N., 2020 Synthetic Data by Principal Component Analysis. In 20TH IEEE INTERNATIONAL CONFERENCE ON DATA MINING WORKSHOPS (ICDMW 2020), edited by G. DiFatta, V. Sheng, A. Cuzzocrea, C. Zaniolo, and X. Wu, International Conference on Data Mining Workshops, pp. 101–105, IEEE; IEEE Comp Soc; Univ Calabria; Mininglamp Technol.

Scheaffer, R. L., W. III Mendenhall, R. L. Ott, and K. Gerow, 2012 Elementary Survey Sampling. Cengage Learning, 7th edition.

Schneeberger, H. and K. Fleischer, 1993 The Multivariate Ratio Estimation: A Simulation Study. Jahrbücher für Nationalökonomie und Statistik **211**: 524–538.

Shukla, G. K., 1966 An Alternative Multivariate Ratio Estimate for Finite Population. Calcutta Statistical Association Bulletin **15**: 127–134.

Singh, G. N., S. Maurya, M. Khetan, and C. Kadilar, 2016 Some imputation methods for missing data in sample surveys. Hacettepe Journal of Mathematics and Statistics **45**: 1865–1880.

Singh, G. N. and S. Suman, 2019 Estimation of population mean using imputation methods for missing data under two-phase sampling design. Journal of Statistical Theory and Practice **13**: 1–24.

Singh, H. P., A. Gupta, and R. Tailor, 2021 Estimation of population mean using a difference-type exponential imputation method. Journal of Statistical Theory and Practice **15**: 1–43.

Singh, H. P., S. Kumar, and S. Bhougal, 2011 Multivariate ratio estimation in presence of non-response in successive sampling. Journal of Statistical Theory and Practice **5**: 591–611.

Singh, S., 2009 A new method of imputation in survey sampling. Statistics **43**: 499–511.

Tariq, M. U., M. N. Qureshi, and M. Hanif, 2021 Variance Estimators in the Presence of Measurement Errors Using Auxiliary Information **19**: 606–616.

Wissler, A., K. E. Blevins, and J. E. Buikstra, 2022 Missing data in bioarchaeology I: A review of the literature. AMERICAN JOURNAL OF BIOLOGICAL ANTHROPOLOGY **179**: 339–348.

Xu, Y. and R. Goodacre, 2018 On Splitting Training and Validation Set: A Comparative Study of Cross-Validation, Bootstrap and Systematic Sampling for Estimating the Generalization Performance of Supervised Learning. Journal of Analysis and Testing **2**: 249–262.

Yates, F. et al., 1949 Sampling methods for censuses and surveys. Sampling methods for censuses and surveys. .

Yuan, Y., 2011 Multiple Imputation Using SAS Software. Journal of Statistical Software **45**: 1–25.

Zhang, P., 2003 Multiple imputation: Theory and method. INTERNATIONAL STATISTICAL REVIEW **71**: 581–592.

2 ESTIMADOR TIPO RAZÃO MULTIVARIADA DA MÉDIA POPULACIONAL PARA PLANO DE AMOSTRAGEM ESTRATIFICADA COM VARIÁVEIS AUXILIARES DE BAIXA E ALTA CORRELAÇÃO

Resumo

Em estudos de amostragem uma necessidade é se estimar a média de uma variável de interesse, e uma das técnicas utilizadas variáveis auxiliares para ajudar na precisão. O presente trabalho compara 3 técnicas de estimar a média populacional: o estimador usual, por razão e razão multivariado. Foi realizado um estudo de simulação em que os dados possuíam: baixa e alta correção; 5 e 10 estratos; e tamanhos amostrais de 100, 250 e 500 observações. O método da razão apresentou uma melhor erro padrão para todos os cenários, enquanto o método da razão multivariada uma melhor correlação entre as médias estimadas em relação as médias verdadeiras. Foi aplicado também em um conjuntos de dados de usina hidrelétrica de Santo Antônio localizado no estado de Rondônia (Brasil). A estimativa da média para os diferentes métodos apresentaram valores próximos.

Palavras-chave: variáveis conjuntas; estratos; estimador; erro padrão.

2.1 Introdução

Na área de amostragem uma das principais preocupações vem do cálculo de estimadores que apresente boas propriedades na inferência populacional. Para isto desenvolve-se metodologias desde o processo de planejamento: caracterização da população, processo de coleta das amostras, identificação dos possíveis vieses, chegando enfim nas escolhas de estimadores que geram resultados e discussões representativos para a população.

Um método utilizado é o do estimador tipo razão, em que durante a coleta dos dados são coletadas informações tanto da variável de interesse do pesquisador, quanto uma variável auxiliar. Esta variável auxiliar pode ser em alguns casos informativa, isto é, se correlaciona com a variável de interesse, e desta forma ajuda nos cálculos dos estimadores. Olkin (1958) apresentou o cálculo de um estimador, que se baseia no método de razão, porém considerando mais de uma variável auxiliar.

Schneeberger and Fleischer (1993) comparam o estimador por razão e o estimador razão multivariada com seis variáveis auxiliares, levando em consideração a relação entre a variável de interesse e as variáveis auxiliares. No primeiro cenário foi utilizado apenas uma variável auxiliar (seis combinações) o segundo leva em consideração duas variáveis auxiliares (15 combinações), e no último 3 variáveis auxiliares (20 combinações).

Shukla (1966) utilizou o estimador tipo razão multivariada em um problema de estimação da média da variável de produção, em relação a duas variáveis auxiliares, o diâmetro da base e a largura da planta sendo altamente correlacionadas. Hanif et al. (2009) desenvolveram uma classe de estimadores para amostragem em dois estágios e múltiplos estágios, a partir do método proposto por Olkin. Singh et al. (2011) utilizaram a teoria desenvolvida por Olkin para propor estimadores que ajudam em ocasiões em que acontecem ausência da resposta.

Outros autores utilizaram a abordagem de utilizar várias variáveis auxiliares para melhorar a precisão do estimador porém com o objetivo de fazer inferência da variância populacional (Cebrián and García, 1997; Olufadi and Kadilar, 2014; Tariq et al., 2021).

A partir desses estudos, este trabalho tem por objetivo comparar os estimadores usuais, por razão e razão multivariado por meio de um conjunto de dados simulados que se aproximam de uma população estratificada quando as variáveis de estudo possuem baixa correlação ou alta correlação.

2.2 Materiais e Métodos

2.2.1 Notações e formalidades matemáticas

Nesta seção serão apresentadas as notações e índices matemáticos para os cálculos ao decorrer do trabalho.

Segundo HULLEY ET AL. (2013) a população possui 2 níveis de hierarquia, primeiramente tem-se a população alvo (Ω) que representa todos os elementos que contém a informação em estudo visualizada de forma abstrata pois se apresenta, em sua maioria, com uma quantidade incontáveis de elementos; imprescindivelmente indicada durante o planejamento amostral de uma pesquisa. E depois, tem-se a população referenciada (BOLFARINE and BUSSAB, 2004) ou acessível como citado por (HULLEY ET AL., 2013), aquela que contém as informações que o pesquisador possui interesse de estudar, e corroborando naquele que será estruturada as técnicas amostrais.

As informações da população acessível são chamadas como parâmetros e comumente são escritas por letras gregas. Na amostragem são coletados indivíduos que mais se assemelham com as características da população, a partir disso é construído um delineamento amostral. As representações das informações destas amostras são denotadas com letras do alfabeto latim.

Além disso, mesmo dentro de uma única população acessível, em determinadas áreas do conhecimento, podem conter perfis de indivíduos que se agrupam, isto é, apresentam características homogêneas entres si, e heterogêneas com indivíduos fora desse grupo, sendo necessário particionar a população acessível em estratos.

O tamanho populacional é indicado por N e cada estratos possuem tamanho N_h com $h = 1, 2, ..., L$ possíveis estratos (L é o número de estratos), sendo que $\sum_{h=1}^{L} N_h = N$. O tamanho amostral é definido por n, e é o número total de amostras que são coletadas, em cada um dos estrato tem-se $n_1, n_2, ..., n_L$, de forma que $\sum_{h=1}^{L} n_h = n$. As alocações podem ser do tipo proporcional quando são distribuídas a partir do tamanho populacional total e de cada estrato ($n_h = n\frac{N_h}{N}$), alocação uniforme quando define-se um tamanho único para cada estrato m ($n_h = \frac{n}{L}$).

2.2.2 Simulação

Foi realizado um estudo de simulação para comparar os três métodos de estimação da média populacional: a padrão, por razão e razão multivariada. Definiu-se 4 cenários de dados estruturados da seguinte forma: primeiro cenário com 5 grupos (estratos), alta correlação entre as variáveis auxiliares e 5000 observações; segundo cenário com 10 grupos (estratos), alta correlação entre as variáveis auxiliares e com 10000 observações; terceiro cenário com 5 grupos (estratos), baixa correlação entre as variáveis v e com 5000 observações; e o quarto cenário com 10 grupos (estratos), baixa correlação entre as variáveis auxiliares e com 10000 observações.

Para gerar esses cenários foi utilizada biblioteca `MixSim` (MELNYKOV ET AL., 2012) que permite simular misturas de uma distribuição Gaussiana multivariada com diferentes níveis de sobreposições entre agrupamentos de indivíduos. Outros autores utilizaram essa biblioteca para estudos com estrutura de agrupamentos, como XU and GOODACRE (2018) fizeram um estudo comparativo dos métodos de validação cruzada, *bootstrap*, e amostragem sistemática para performance de modelos de *machine learning*. MAITRA ET AL. (2012) também aplicaram as funções para o estudo do método *bootstrap* para teste de significância para quantificar agrupamentos em conjuntos de dados multidimensionais.

Para gerar as matrizes de covariâncias utilizou-se a função `genPositiveDefMat` da biblioteca `clusterGeneration` QIU and JOE. (2020). A função permite definir valores para os autovalores $(\lambda_1, \lambda_2, ..., \lambda_p)$ que são utilizado para gerar aleatoriamente os autovetores $\mathbf{Q} = [\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, ..., \boldsymbol{\alpha}_p]$, a assim construir as matrizes de covariância $\boldsymbol{\Sigma} = \mathbf{Q} \cdot diag(\lambda_1, ..., \lambda_p) \cdot \mathbf{Q}^\top$. Segundo JOHNSON and WICHERN (2007)

os primeiros autovalores possuem alta proporção explicativa das estrutura da covariância das variáveis originais, com essa propriedade foi possível gerar matrizes com alta e baixa correlação. Para a determinação dos vetores de médias da forma aleatória gerando valores de uma distribuição $X_i \sim Gaussian(70, 20^2)$.

Com isso gerou-se 1000 simulações para o estudo de amostragem e selecionadas observações de forma aleatória pelo método de alocação proporcional. Para os cenários 1 e 3, com 5 grupos, e obtidos tamanhos amostrais de 100, 250 e 500; agora para os cenários 2 e 4, com 10 grupos foram selecionados tamanhos amostrais de 200, 500 e 1000.

2.2.3 Aplicação em dado reais

Os dados utilizados para a análise são provenientes da usina hidrelétrica de Santo Antônio localizado no Rio Madeira, com sede no Município de Porto Velho, capital do estado de Rondônia, Brasil. As coordenadas geográficas 8°48'06"S e 63°57'03"W, sendo a quarta maior geradora de energia hídrica em operação do pais e considerada a primeira em sustentabilidade, conforme avaliação da International Hydropower Association (IHA). A empresa responsável pela implantação e funcionamento é a Santo Antônio Energia.

Para as análise foram consideradas 3 variáveis sendo elas: geração de energia (ge) em MWh; vazão afluente média (vam) em m^3/s; e nível horário do reservatório (nrm) em m, e coletadas periodicamente pelo período de uma hora diariamente no decorrer do ano de 2016. A ge foi variável de interesse e as demais foram consideradas como variáveis auxiliares. Para um plano de amostragem considerando estratificação, levou-se em consideração o perfil característico das estações definidas do ano em 2016, segundo o Departamento de Astronomia Instituto de Astronomia (2020) sendo: de janeiro até 19 de março e de 22 a 31 de dezembro; outono de 20 de março à 19 de junho; inverno de 20 de junho à 21 de setembro; e por fim a primavera entre os dias 22 de setembro e 21 de dezembro.

Para o método de amostragem, as amostras foram selecionadas de forma aleatória considerando os tamanho amostral pelo método de alocação proporcional para cada estação. Para o verão foram obtidos 487 amostras ($N = 2130$), outono com 505 amostras ($N = 2208$), inverno com 516 amostras ($N = 2256$), e primavera com 494 amostras ($N = 2159$).

2.2.4 Métodos de estimação para amostragem aleatória estratificada

2.2.5 Estimador da média populacional (sem variável auxiliar)

O estimador não viesado para o valor da média populacional μ, considerando o planejamento de amostragem estratificada com índice de $h = 1, ..., L$, é dada pela expressão:

$$\hat{\mu} = \sum_{h=1}^{L} W_h \bar{y}_h$$

em que: $W_h = \frac{N_h}{N}$ são os pesos de cada estrato (Cochran, 1977), ou constastes que satisfaçam a condição de não viés (Arnab, 2017), e $\bar{y}_h$ é a média amostral respectiva do h-ésimo estrato. O estimador da média populacional é o cálculo da média ponderada pelos estratos em que multiplica-se pela proporção de unidades populacionais em cada estrato (Lohr, 2019). Segundo Cochran (1977) cada estrato apresenta um estimador não viesado ($E[\hat{\mu}_h] = \mu_h$) corroborando assim que o estimador populacional ($E[\hat{\mu}] = \mu$) também não é viesado.

A variância do estimador da média é dado pela expressão:

$$Var[\hat{\mu}] = \sum_{h=1}^{L} W_h^2 (1 - f_h) \frac{S^2}{n_h}.$$

em que: $f_h = \dfrac{n_h}{N_h}$ é a razão que os estratos amostrais representam da população, chamando-se de fração de amostragem. Segundo Olkin (1958) para se obter o estimador da variância não viesado para a média, é realizada a substituição da variância populacional S_h^2 de cada estrato pela variância amostral s_h^2, obtendo $\hat{Var}[\hat{\mu}] = \sum_{h=1}^{L} W_h^2(1-f_h)\frac{s_h^2}{n_h}$.

2.2.6 Estimador tipo razão da média populacional (única variável auxiliar)

A estimação da média do tipo razão pressupõe que o pesquisador, além da variável de interesse (Y), colete outra variável auxiliar (X) da mesma unidade amostral. Abordado assim um critério univariado pois leva-se em consideração apenas uma única variável para aprimorar as inferências para a variável de interesse, e segundo Lohr (2019), o método oferece uma acréscimo na precisão dos estimadores da média e do total.

O estimador da média populacional para a variável de interesse (Y), considerando amostragem estratificada é expresso por:

$$\hat{\mu}_{(Y)}^{[R]} = \sum_{h=1}^{L} W_h \hat{\mu}_{h(Y)}^{[R]}$$

em que, $\hat{\mu}_{h(Y)}^{[R]} = r_h \mu_{h(X)}$ é o estimador tipo razão da média populacional da variável de interesse dentro de cada $h-$ésimo estrato, $\mu_{h(X)}$ é a média populacional da variável auxiliar, $r_h = \dfrac{\bar{y}_h}{\bar{x}_h}$ é a razão entre a média amostral da variável de interesse pela média amostral da variável auxiliar.

Segundo Cochran (1977) o estimador tipo razão é viesado, porém esse viés não prejudica as estimativas a medida que o tamanho amostral cresce ($n \Rightarrow \infty$). Para o plano de amostragem estratificado o viés deve ser considerado, quando o tamanho amostral é pequeno, e o número de estratos é grande (Goodman and Hartley, 1958). O estimador da variância da média tipo razão foi calculado separadamente para cada estrato, e depois combinado de forma ponderada pelos pesos de cada estrato.

$$Var[\hat{\mu}_Y^{[R]}] = \sum_{h=1}^{L} \left(\frac{N_h}{N}\right)^2 \left(1 - \frac{n_h}{N_h}\right) \frac{s_h^{2[R]}}{n_h}$$

em que $s_h^{2[R]} = \frac{\sum_{j=1}^{n_h}(y_{hj} - r_h x_{hj})^2}{n_h - 1}$ é uma expressão similar da variância amostral, porém utiliza o estimador razão para cálculo dos desvios das j-ésimas observações.

2.2.7 Estimador tipo razão multivariada da média populacional (diversas variáveis auxiliares)

Introduzido por Olkin (1958) este estimador é uma extensão do estimador por razão univariada, visto que incorpora mais que uma variável auxiliar conjuntamente $(X_1, X_2, ..., X_p)$. A expressão para o estimador da média populacional, do tipo razão multivariada, para a variável de interesse (Y) para amostragem estratificada é dada por:

$$\hat{\mu}_{(Y)}^{[\mathrm{MR}]} = \sum_{h=1}^{L} W_h \hat{\mu}_{h(Y)}^{[\mathrm{MR}]}$$

em que: $\hat{\mu}_{h(Y)}^{[\mathrm{MR}]} = \sum_{j=1}^{p} w_{hj} r_{hj} \mu_{h(X_j)}$ é estimador da média populacional, do tipo razão multivariada, para o h-ésimo estrato; $r_{hj} = \frac{\bar{y}}{\bar{x}_j}$ é a razão entre a média amostral a variável de interesse, sobre a média amostral da j-ésima variável auxiliar; $\mu_{h(X_j)}$ é a as média populacional da j-ésima variável auxiliar, dentro do h-ésimo estrato; w_{hj} são coeficientes que seguem a restrição $\sum_{j=1}^{p} w_{hj} = 1$. Segundo Olkin (1958), os coeficientes são estimados utilizando-se da generalização da inequação de Cauchy para a minimização da variância de um estimador.

Dentro de um h-ésimo estrato, o cálculo dos vetores de coeficientes $\hat{\boldsymbol{w}}_h = [w_1, ..., w_p]$ é dado por:

$$\hat{\boldsymbol{w}}_h = \frac{\boldsymbol{e}\mathbb{A}^{-1}}{\boldsymbol{e}\mathbb{A}^{-1}\boldsymbol{e}^\top}$$

sendo que, $\boldsymbol{e} = [1, ..., 1]^\top$, $\mathbb{A} = \mathbb{T}\mathbb{C}\mathbb{T}^\top$, com $\mathbb{C} \Rightarrow c_{ij} = \frac{S_{kj}}{\bar{X}_k \bar{X}_j}$ e

$$\underset{p\times(p+1)}{\mathbb{T}} = \begin{bmatrix} 1 & -1 & 0 & \dots & 0 \\ 1 & 0 & -1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 1 & 0 & 0 & \dots & -1 \end{bmatrix}$$

Segundo **?** esse estimador também apresenta viés, entretanto, para grandes amostras torna-se pequeno. A variância estimada para a estimativa da média, para cada h-ésimo estrato é dada pela expressão:

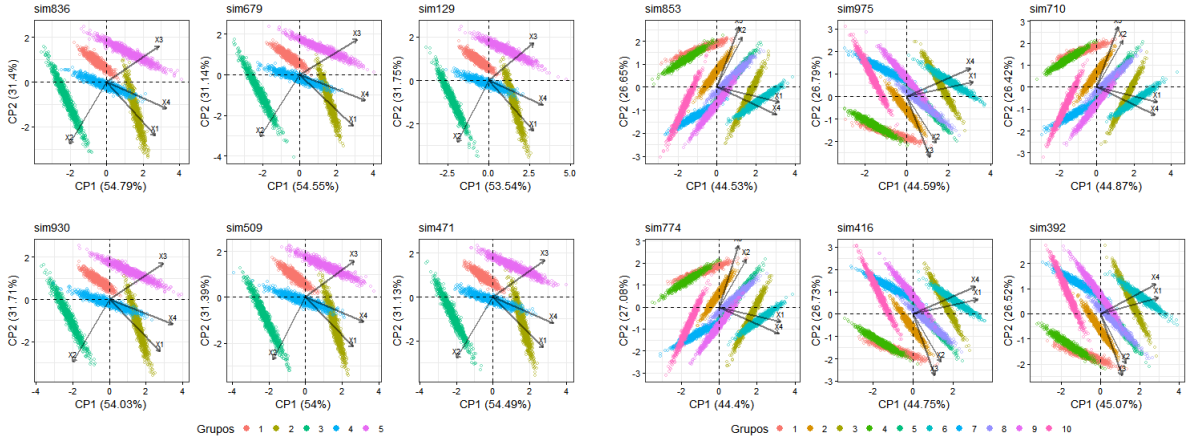
$$\hat{\text{Var}}\left[\hat{\mu}^{[\text{MR}]}_{h(Y)}\right] = \frac{1}{n(\boldsymbol{e}\mathbb{A}^{-1}\boldsymbol{e}^T)}$$

e como resultado a variância do estimador é obtido pela combinação das variâncias calculadas dentro de cada estrato sem a necessidade das ponderação dos tamanhos populacionais $\hat{\text{Var}}\left[\hat{\mu}^{[\text{MR}]}_{(Y)}\right] = \sum_{h=1}^{L} \hat{\text{Var}}\left[\hat{\mu}^{[\text{MR}]}_{h(Y)}\right]$.

2.3 Resultados e Discussões

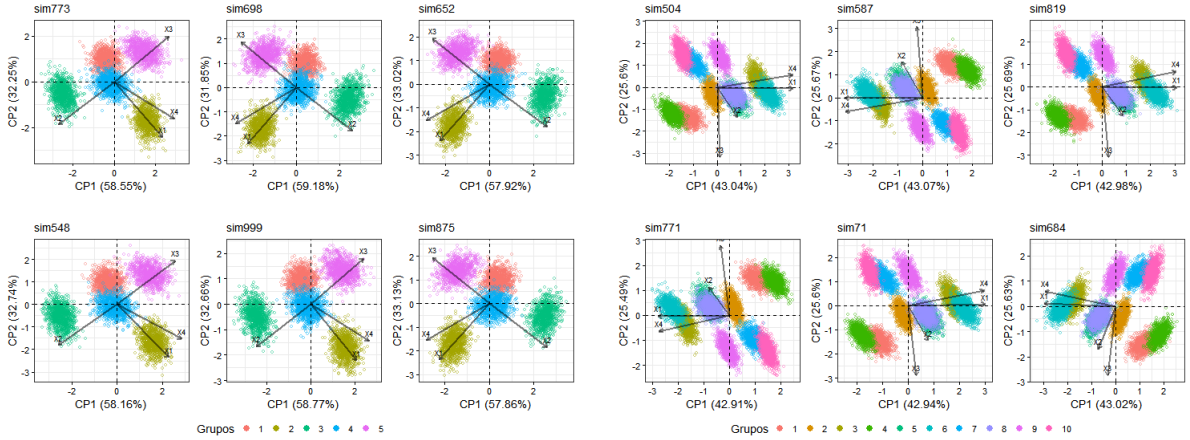
Na Figura Figura 2.1 inicialmente foi realizado uma análise exploratória das 1000 simulações para os 4 cenários gerados. Foram selecionados aleatoriamente 6 conjuntos simulados, e realizado uma análise de componentes principais (PCA) para visualizar alta e baixa correlação entre as variáveis dentro dos grupos que na simulação tem o intuito de representar os estratos. Para os 4 cenários, em média, as duas primeiras componentes principais possuem uma proporção acumulada explicativa de 70 a 80 % dos dados. Caracterizando uma boa proporção para visualizar a estrutura conjunta das variáveis.

Nas Figura 2.1a e Figura 2.1b observa-se que os pontos, dentro de cada grupo, formam uma elipse cuja a área está achatada. Isto devido a característica de alta correlação das variáveis de interesse e as auxiliares. Nas Figura 2.1c e Figura 2.1d, as simulações foram geradas com baixa correlação entre as variáveis, com isso as observações dentro do grupos se apresentaram mais espalhadas tendendo a formar um circunferência.



(a) *Biplot* para 6 simulações escolhidas aleatoriamente pertencentes ao cenário 1

(b) *Biplot* para 6 simulações escolhidas aleatoriamente pertencentes ao cenário 2

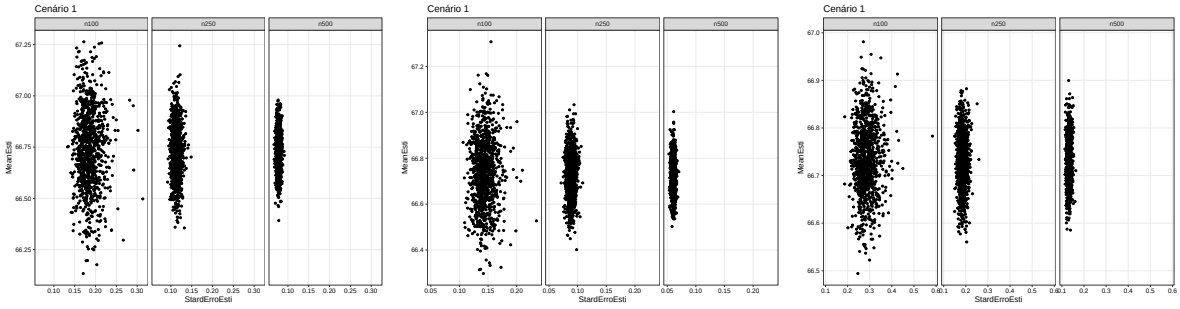


(c) *Biplot* para 6 simulações escolhidas aleatoriamente pertencentes ao cenário 3

(d) *Biplot* para 6 simulações escolhidas aleatoriamente pertencentes ao cenário 4

Figura 2.1. Análise de Componentes Principais para os 4 cenários de diferentes: intensidade de correlações alta (a e b) e baixa (c e d); e estratos de tamanho 5 (a e c) e 10 (b e d)

Para cada tamanho amostral, em cada cenário, foram calculados os valores das estimativa da média e estimativa do erro padrão (EP) do estimador. A partir dessas duas medidas gerou-se gráficos de dispersão (Figuras 2.2, 2.5, 2.3 e 2.6), a fim de compará-las entre cada tamanho amostral (n), cada método de estimação (ME) e principalmente cada cenário. Os gráficos dentro da cada cenário foram apresentados com os eixos fixos para cada tamanho amostral, porém nos métodos tem-se uma diferença entre as escalas. Isso devido essas comparações apresentarem valores sub e super estimados.

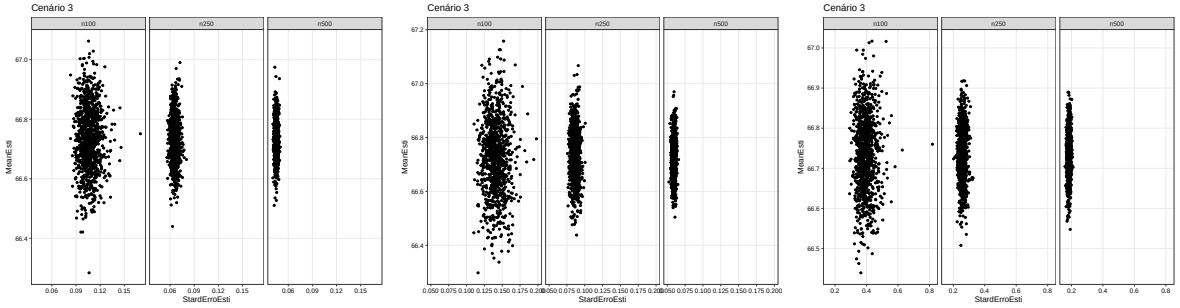


(a) ME padrão **(b)** ME tipo razão **(c)** ME tipo razão multivariada

Figura 2.2. Gráfico de dispersão para os valores de estimação (eixo y) em relação aos valores de erro padrão do estimador (eixo x) para o plano de amostragem estratificado utilizado dentro do cenário 1

Na Figura 2.2 as estimativas para a média populacional são mais dispersas quando o tamanho amostral é menor. Evidenciando que o aumento de amostras aprimora as estimativas e aproxima da média "verdadeira", independente do métodos es estimação. Com o erro padrão do estimador o método de estimação tipo razão (Figura 2.2b) apresentou menores valores, quando comparado com os demais (Figura 2.2a e Figura 2.2c). Como observa-se para o tamanho amostral de 100 observações, os erros padrão do método da razão ficaram entre 0,06 a 0,10 aproximadamente; para o estimador padrão estavam entre 0,15 a 0,23; e para o método da razão multivariada entre 0,2 à 0,35. O estimador tipo razão apresenta melhores resultados devido ao acréscimo de uma variável auxiliar nas estimativa. Todavia, quando aumenta-se a quantidade de variáveis auxiliares, o erro padrão do estimador pode ser inflacionados pela estrutura multivariada mais complexa.

Esse comportamento se repetem para o cenário 2 (Figura 2.5), dessa forma o gráfico foi apresentado no Apêndice A (Subseção 2.5.1).



(a) ME padrão **(b)** ME tipo razão **(c)** ME tipo razão multivariada

Figura 2.3. Gráfico de dispersão para os valores de estimação (eixo y) em relação aos valores de erro padrão do estimador (eixo x) para o plano de amostragem estratificado utilizado dentro do cenário 3

Entretanto no cenário 3 (Figura 2.3), para qual a variável de interesse não tem correlação com as variáveis auxiliares e considera-se o tamanho amostral de 100 observações, os resultados foram: para o erro padrão do estimador sem variável auxiliar, os valores ficaram entre 0,09 a 0,12; enquanto que o ME tipo razão os erros padrão ficaram com valores entre 0,125 a 0,150; e ME tipo razão multivariada os valores ficaram em torno de 0,3 a 0,5. Isso pode indicar que, quando as variáveis possuem uma alta correlação, para ME tipo razão a variabilidade do estimador é aprimorado, resultado que não é observado quando as variáveis não possuem alta correlação. Destacando-se o ME tipo razão multivariada, que acaba superestimando os resultados de variância do estimador.

Para observar o quanto as estimativa, se aproximam do valor verdadeiro da média populacional, foram gerados gráficos de dispersão entre os valores de médias estimadas com relação aos valores de

médias verdadeiras (Apêndice A 2.5.2). Na Tabela 2.1 foram apresentadas as medidas de correlação de Spearman entre as médias estimadas, em relação as médias populacionais de cada simulação. Desta forma, foi possível verificar quais métodos e tamanhos amostrais apresentam resultados que se aproximam do valor populacional esperado.

Tabela 2.1. Tabela com medidas de correlação de Spearman para cada cenário simulado, em relação aos métodos de estimação (padrão, razão e razão multivariada) e em cada tamanho amostral.

Cenários	Métodos de estimação								
	Padrão			Razão			Razão multivariada		
	Tamanho amostral			Tamanho amostral			Tamanho amostral		
	1	2	3	1	2	3	1	2	3
C1	0,207	0,339	0,484	0,258	0,392	0,575	0,491	0,707	0,826
C2	0,410	0,568	0,726	0,452	0,601	0,772	0,772	0,901	0,949
C3	0,278	0,444	0,597	0,226	0,364	0,491	0,346	0,524	0,676
C4	0,557	0,729	0,848	0,426	0,588	0,738	0,603	0,769	0,885

Foi observado no método da razão multivariada, que a correlação cresce para todos os cenários e tamanhos amostrais. Os métodos padrão e razão para os cenários 1 e 2 apresentam valores similares. Para o cenário 1, houve uma correlação forte de aproximadamente 0,826 para o tamanho amostral 3, enquanto que os métodos padrões e razão, as correlações foram de aproximadamente 0,484 e 0,575 respectivamente. No cenário 2, houve correlação forte para os tamanhos amostrais 2 e 3, evidenciando que a maior quantidade de grupos ajudou na precisão dos estimador.

No cenário 4, os métodos padrão e razão multivariada foram parecidos. Mostrando que, apesar da estrutura de baixa correlação entre as variáveis de interesse e auxiliares, a precisão do estimador não decaiu. O método da razão com apenas uma variável auxilar, quando comparado com o método padrão, foi mais preciso somente para os cenários que possuem alta correlação entre as variáveis.

Para os dados reais, inicialmente foi realizado uma análise exploratória com medidas descritivas para as 3 variáveis em estudo, em relação a cada estação do ano.

Tabela 2.2. Medidas descritiva para as variáveis, em relação as estações

Variáveis	Estação	Mínimo	Média	Mediana	Máximo	Desvio Padrão	Curtose	Assimetria
[1]ge	Inverno	312,00	1052,89	899,00	2200,00	352,99	2,53	0,78
ge	Outono	214,00	1633,45	1693,00	2215,00	240,76	4,00	-0,75
ge	Primavera	7,00	1090,10	1020,00	1910,00	332,98	2,05	0,36
ge	Verão	446,00	1642,28	1671,50	2208,00	258,06	5,43	-1,13
[2]nrm	Inverno	70,42	70,49	70,49	70,56	0,02	4,67	-0,47
nrm	Outono	70,35	70,50	70,50	7,62	0,02	8,14	-0,58
nrm	Primavera	70,44	70,49	70,49	70,58	0,02	3,59	0,50
nrm	Verão	69,94	70,44	70,45	70,59	0,10	13,18	-3,00
[3]vam	Inverno	1806,00	5303,52	4296,00	12604,00	2188,00	2,79	0,91
vam	Outono	10861,00	23105,05	24893,50	33769,00	5837,19	1,84	-0,38
vam	Primavera	2090,00	5355,96	4788,00	12919,00	1965,75	2,25	0,53
vam	Verão	9137,00	22965,98	21895,00	42138,00	7911,03	1,97	0,22

Rodapé: [1]ge - geração de energia em MWh; [2]nrm nível horário do reservatória em m; e [3]vam - vazão afluente média em m^3/s

Para a variável ge, observa-se para inverno uma produção mínima de 312 MWh e máxima de 2200 MWh, na primavera a produção fica entre 7 MWh e 1910 MWh, no outono entre 214 MWh a 2215 MWh, e para o verão entre 446 MWh a 2208 MWh. Com as maiores valores de média e mediana nas estações do outono com 1633.45 e 1693, e de verão com 1642,28 e 1671,50.

Na variável de nrm, houve a menor variabilidade de observações, acarretando em desvios padrão baixos para todas as estações. Foram calculados para todas as estações valores alto para o coeficiente de curtose, principalmente para a estação do verão (13,18), indicando uma distribuição bem concentrada. No verão também houve um coeficiente de assimetria negativo (-3,00) indicando em cauda mais alongadas a direita da média.

Analisando as medidas obtidas para a variável vam, as estações que apresentaram menores valores foram o inverno com mínima de 1806 e máxima de 12604, e primavera entre 2090 a 12919. O verão apresentou maior variabilidade entre os valores com um desvio padrão de 7911,03, com mínima de 9137 e máxima de 42138.

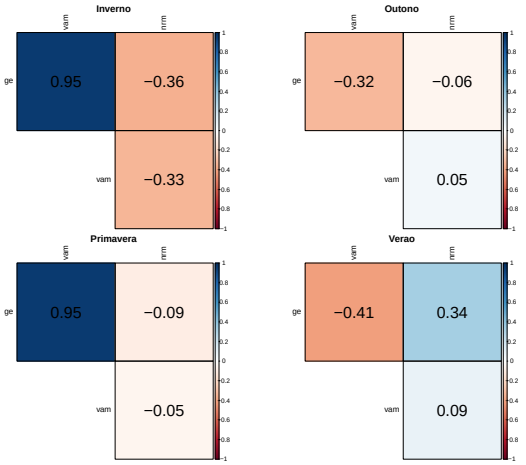


Figura 2.4. Gráfico de correlação entre as variáveis de estudo, para cada estação do ano (estratos)

Na Figura 2.4, foi gerado um gráfico de correlação de Spearman entre as variáveis para cada estação em virtude da não normalidade que as variáveis apresentam. Observa-se que, para as estações de inverno e primavera, a correlação entre a ge e vam foram fortes e positivas.

Considerando o conjunto de dados completo, a média populacional para a variável ge (μ_Y) foi de 1351,94. Para comparar os métodos foi realizado os cálculos de estimador da média e erro padrão do estimador da média, para os dados amostrais gerado por alocação proporcional. Os métodos comparados: padrão (Subseção 2.2.5 sem o uso de variável auxiliar; o método da razão para a variável vam e nrm (Subseção 2.2.6); e por fim o método da razão multivariada que considera as 2 variáveis auxiliares conjuntamente (Subseção 2.2.7).

Tabela 2.3. Estimativa da média e EP (erro padrão) para os métodos do tipo padrão, razão com uma variável auxiliar e razão multivariada

	Métodos			
Estimativa	Padrão	Tipo razão (vam)	Tipo razão (nrm)	Razão multivariada
Média	1350,448	1345,388	1350,452	1350,846
EP da estimativa da média	34,533	77,922	34,504	64,993

Com o método padrão a estimativa da média foi de 1350,448 MWh, com um erro padrão de 34,533. Para o método da razão utilizando a variável *vam* a estimativa foi de 1350,452 e o erro padrão de 77,922. Com o mesmo método, porém agora com a variável nrm a estimativa foi de 1350,452 e o erro padrão de 34,504. E finalmente para o método da razão multivariada, a estimativa foi de 1350,846 com erro padrão de 64,993.

Pelo método da razão com a variável vam houve uma subestimação da média verdadeira, além de um erro padrão maior quando comparado com o padrão e tipo razão (nrm). Indicando que a relação

de alta e baixa correlação entre as variáveis, dentro das estações pode influenciar na estimativa da média.

O método da razão multivariada apresentou segundo o maior erro padrão, todavia com mais precisão do estimador, indicando que a influência das duas variáveis auxiliares podem acarretar na melhora da precisão do estimador para a variável de interesse.

2.4 Conclusões

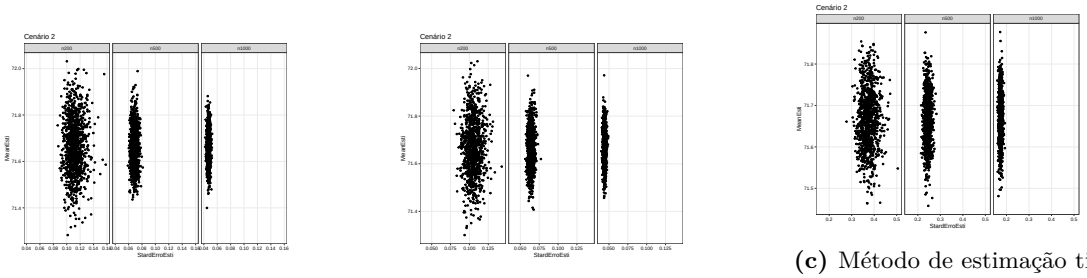
O método da razão multivariada apresentou maior precisão tanto para os dados simulado, quanto para os dados reais. Porém esse método possui um erro padrão inflacionado devido a estrutura de variância e covariância das variáveis.

O método da razão não apresenta uma melhora, mesmo com a característica que os grupos apresentam uma alta correlação. Para os dados reais, mesmo com a alta correlação para as estações do inverno e primavera, o estimador não se aproximou do valor verdadeiro.

Para um futuro estudo, tem-se a necessidade de ponderar a inflação do estimador de erro padrão para o método razão multivariada. Abordagem não paramétrica como, *bootstrap* e *jackknife* podem apresentar boas caracterizações para os dados multivariados.

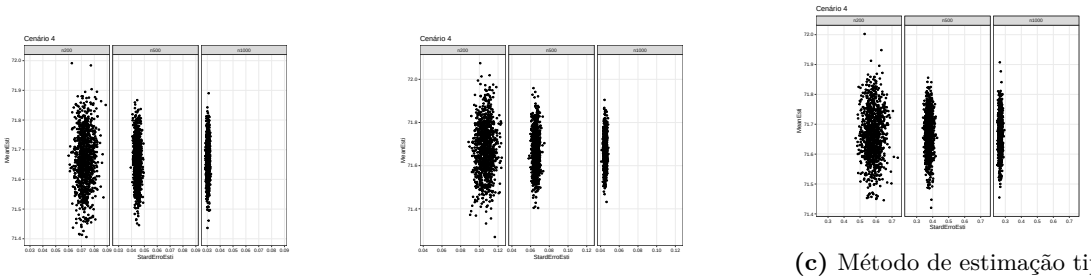
2.5 Apêndices

2.5.1 Gráfico de dispersão para os valores de estimação em relação aos valores de erro padrão do estimador para o plano de amostragem estratificado



(a) Método de estimação **(b)** Método de estimação tipo razão **(c)** Método de estimação tipo razão multivariada

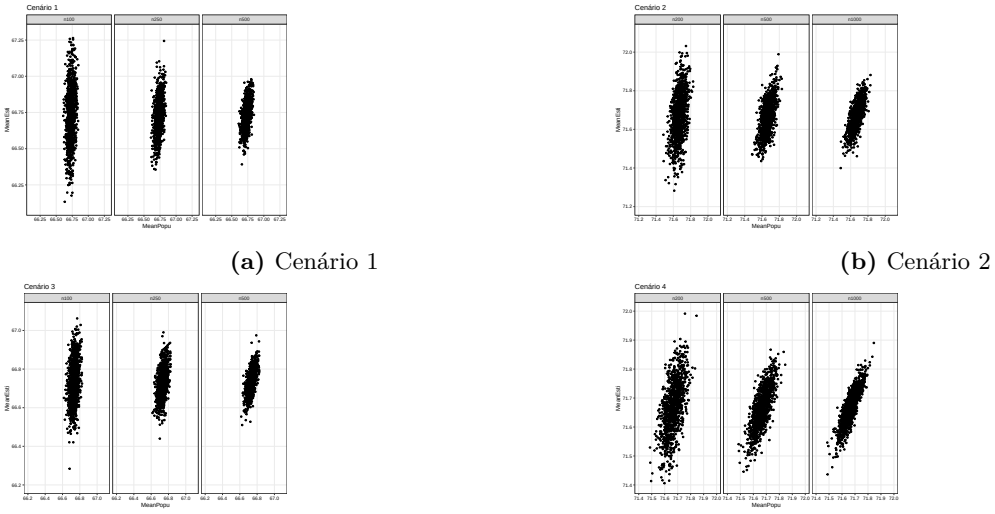
Figura 2.5. Gráfico de dispersão para os valores de estimação (eixo y) em relação aos valores de erro padrão do estimador (eixo x) para o plano de amostragem estratificado utilizado dentro do cenário 2



(a) Método de estimação **(b)** Método de estimação tipo razão multivariada **(c)** Método de estimação tipo razão

Figura 2.6. Gráfico de dispersão para os valores de estimação (eixo y) em relação aos valores de erro padrão do estimador (eixo x) para o plano de amostragem estratificado utilizado dentro do cenário 4

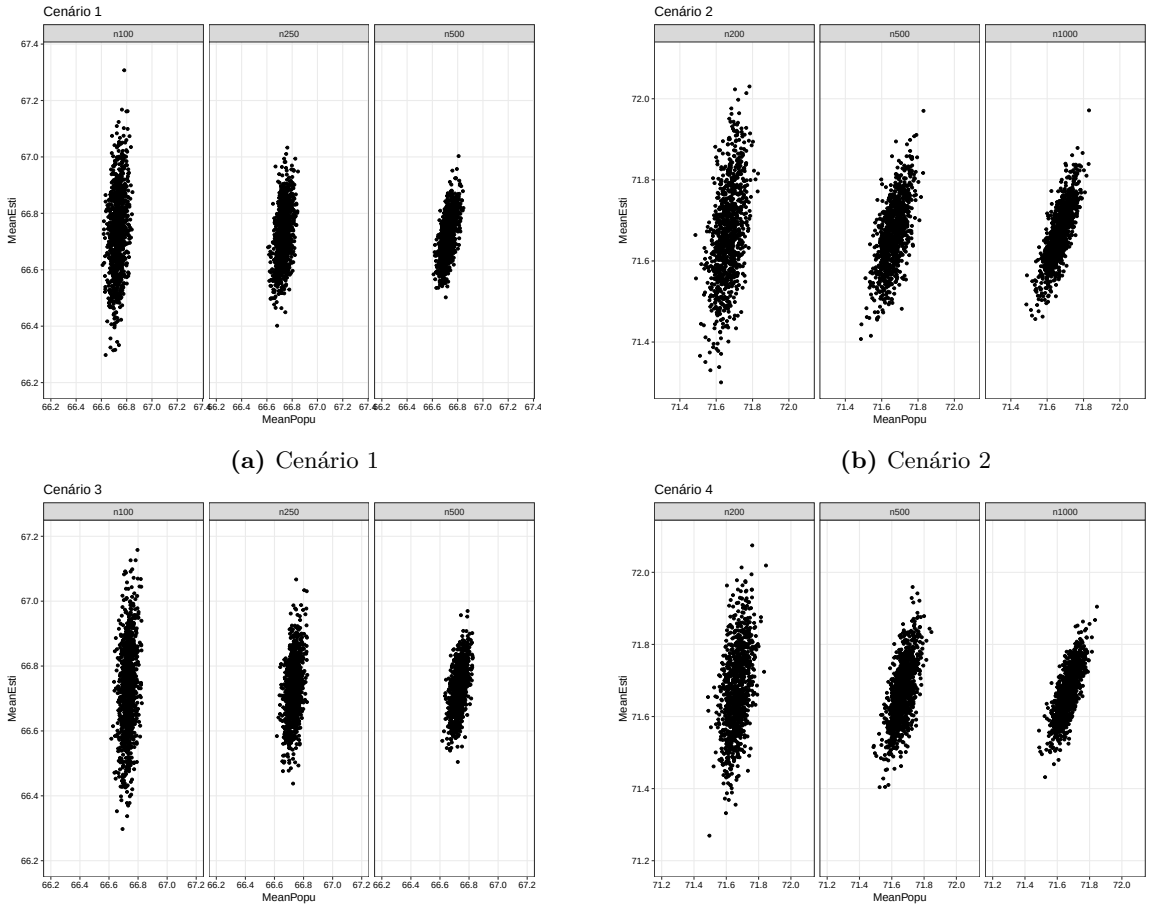
2.5.2 Gráficos de dispersão para os valores da média da média estimada orientada no eixo y, e os valores da média populacional orientada no eixo x



(a) Cenário 1 **(b)** Cenário 2

(c) Cenário 3 **(d)** Cenário 4

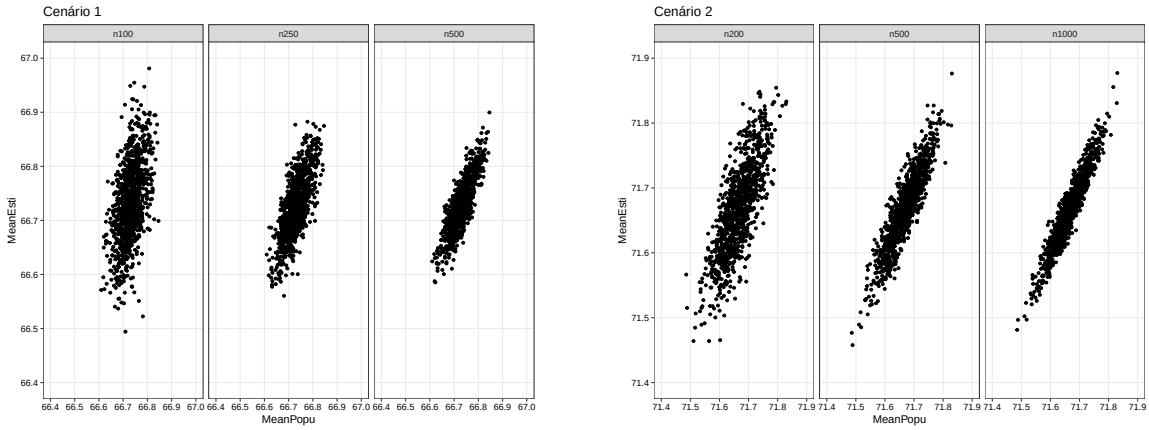
Figura 2.7. Gráficos de dispersão entre os valores estimados dados seus respectivos métodos (eixo y), pelo valor verdadeiro da média populacional (eixo x), para o método de estimação para o plano de amostragem estratificada.



(a) Cenário 1 **(b)** Cenário 2

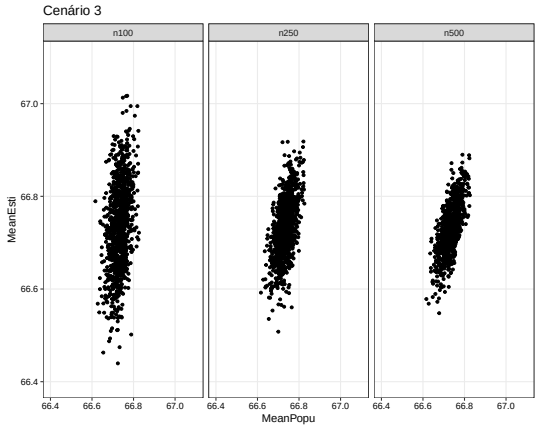
(c) Cenário 3 **(d)** Cenário 4

Figura 2.8. Gráficos de dispersão entre os valores estimados dados seus respectivos métodos (eixo y), pelo valor verdadeiro da média populacional (eixo x), para o método de estimação do tipo razão para o plano de amostragem estratificada.

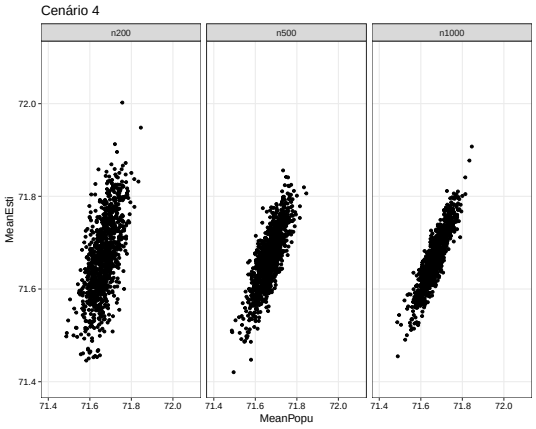


(a) Cenário 1

(b) Cenário 2



(c) Cenário 3



(d) Cenário 4

Figura 2.9. Gráficos de dispersão entre os valores estimados dados seus respectivos métodos (eixo y), pelo valor verdadeiro da média populacional (eixo x), para o método de estimação do tipo razão multivariada para o plano de amostragem estratificada.

2.5.3 Programação

2.5.4 Estimador da média pelo método usual

```
# :::::::::::::::::::: Parametros da funcao ::::::::::::::::::::

# model = ... (ex. X1 ~ strata) #modelo
# data = ... # dados amostrais
# N_strata = ... #tamanho populacional de cada estrato
# v_alpha = ... #nivel de significancia

default_meanestimation <- function(model, data, N_strata, v_alpha){

    # :::::::::::::::::::: Informacões ::::::::::::::::::::

    #estrutura das variáveis
    vars_ain <- model %>% as.formula() %>% all.vars() %>% {.[1]}
    vars_strata <- model %>% as.formula() %>% all.vars() %>% {.[2]}

    #id. strata
```

```
id_strata <- data %>%
magrittr::extract2(vars_strata) %>%
levels() %>%
paste('strata', ., sep = '_')

# :::::::::::::::::::: Informacões populacionais ::::::::::::::::::::

#tamanho da populacao para cada estrato
N_strata <- N_strata %>%
dplyr::group_split(strata) %>%
purrr::map(.f = ~ .x %>%
dplyr::select(-all_of(vars_strata)))

#tamanho total da populacao
N_all <- N_strata %>%
purrr::map(.f = ~ .x %>% dplyr::select(n)) %>%
unlist() %>% sum()

# :::::::::::::::::::: Cálculos ::::::::::::::::::::

#tamanho da amostra para cada estrato
n_strata <-
data %>%
dplyr::group_by_at(vars_strata) %>%
dplyr::summarise(n = n()) %>%
dplyr::group_split(strata) %>%
purrr::map(.f = ~ .x %>% dplyr::select(-all_of(vars_strata)))

#média amostral para cada estrato
mean_strata <-
data %>%
dplyr::group_by_at(vars_strata) %>%
dplyr::summarise_all(.funs = list(~ .x %>% mean())) %>%
dplyr::group_split(strata) %>%
purrr::map(.f = ~ .x %>% dplyr::select(-all_of(vars_strata)))

#desvio padrao amostral para cada estrato
sd_strata <-
data %>%
dplyr::group_by_at(vars_strata) %>%
dplyr::summarise_all(.funs = list(~ .x %>% sd())) %>%
dplyr::group_split(strata) %>%
purrr::map(.f = ~ .x %>% dplyr::select(-all_of(vars_strata)))

# :::::::::::::::::::: Estimacao para a média ::::::::::::::::::::

#cálculo dos pesos
```

```
        weigth <- purrr::map(.x = N_strata, .f = ~ .x/N_all)

        #média estimada
        estimated_mu <-
        purrr::pmap(.l = list('W' = weigth,
        'mean_s' = mean_strata),
        .f = function(W, mean_s) W * mean_s) %>%
        unlist() %>% sum()

        #cálculo da variancia do estimador da média
        estimated_varmu <-
        purrr::pmap(.l = list('W' = weigth,
        'sd_strata' = sd_strata,
        'n' = n_strata,
        'N' = N_strata),
        .f = function(W, sd_strata, n, N){
                W^2*(sd_strata^2/n)*(1 - n/N)}) %>%
        unlist() %>% sum()

        #cálculo do intervalo de confiancia para média
        propz <-  qnorm((1 - v_alpha)/2, lower.tail = F)
        CI <- c('Limite␣Inferior' = estimated_mu - propz*sqrt(estimated_varmu
        'Limite␣Superior' = estimated_mu + propz*sqrt(estimated_varmu))

        #saida
        output <- list(
        'Estimativa_da_média_popu._(Aleatória_com_estratos)' = estimated_mu,
        'Erro_padrao_do_estimador_da_média' = estimated_varmu %>% sqrt(),
        'Intervalo_de_confianca' = CI)

        return(output)
}
```

2.5.5 Estimador da média pelo método da razão univariado

```
# :::::::::::::::::::: Parametros da função ::::::::::::::::::::

# model = ... (e.g. X1/X2 ~ strata) #modelo agora considerando a razão
# data = ... #dados amostrais
# mean_popuAux = ... #média populacional para var. auxiliar
# N_strata = ... #tamanho populcional de cada estrato
# v_alpha = ... #nível de significância

ratio_meanestimation <- function(model, data, mean_popuAux,
N_strata, v_alpha){

        # :::::::::::::::::::: Informações ::::::::::::::::::::
```

```
#estrutura das variáveis
vars_ain <- model %>% as.formula() %>% all.vars() %>% {.[1]}
vars_aux <- model %>% as.formula() %>% all.vars() %>% {.[2]}
vars_strata <- model %>% as.formula() %>% all.vars() %>% {.[3]}

#id. strata
id_strata <- data %>%
magrittr::extract2(vars_strata) %>%
levels() %>%
paste('strata', ., sep = '_')

# :::::::::::::::::::: Informações populacionais ::::::::::::::::::::

#média populacional da variável auxiliar
mean_popuAux <- mean_popuAux %>%
dplyr::group_split(strata) %>%
purrr::map(.f = ~ .x %>% dplyr::select(-all_of(vars_strata)))

#tamanho total da população
N_strata <- N_strata %>%
dplyr::group_split(strata) %>%
purrr::map(.f = ~ .x %>% dplyr::select(-all_of(vars_strata)))

N_all <- N_strata %>%
purrr::map(.f = ~ .x %>% dplyr::select(n)) %>%
unlist() %>% sum()

# :::::::::::::::::::: Cálculos de amostragens ::::::::::::::::::::

#tamanho da amostra para cada estrato
n_strata <- data %>%
dplyr::group_by_at(vars_strata) %>%
dplyr::summarise(n = n()) %>%
dplyr::group_split(strata) %>%
purrr::map(.f = ~ .x %>% dplyr::select(-all_of(vars_strata)))

#média amostral para cada estrato
mean_strata <- data %>%
dplyr::group_by_at(vars_strata) %>%
dplyr::summarise_all(.funs = list(~ .x %>% mean())) %>%
dplyr::group_split(strata) %>%
purrr::map(.f = ~ .x %>% dplyr::select(-all_of(vars_strata)))

#desvio padrão amostral para cada estrato
sd_strata <- data %>%
dplyr::group_by_at(vars_strata) %>%
```

```
dplyr::summarise_all(.funs = list(~ .x %>% sd())) %>%
dplyr::group_split(strata) %>%
purrr::map(.f = ~ .x %>% dplyr::select(-all_of(vars_strata)))

# :::::::::::::::::::: Estimação ::::::::::::::::::::

#cálculo dos pesos
weigth <- N_strata %>%
purrr::map(.f = ~ .x/N_all)

#razão estimada para cada estrato
ration_esti <-
purrr::map(.x = mean_strata,
.f = ~ as.numeric(.x[vars_ain]/.x[vars_aux]))

#média estimada
estimated_mu <-
purrr::pmap(.l = list('W' = weigth,
'r' = ration_esti,
'MU' = mean_popuAux),
.f = function(W, r, MU) W * r * MU) %>%
unlist() %>% sum()

#cálculo da variância do estimador da média
VarRest <-
purrr::map2(.x = data %>% dplyr::group_split(strata),
.y = ration_esti,
.f = ~ sum((.x[vars_ain] -
.y*.x[vars_aux])^2)*(1/(nrow(.x) - 1))) %>%
magrittr::set_names(id_strata)

estimated_varmu <-
purrr::pmap(.l = list('Nstrata' = N_strata,
'VarRest' = VarRest,
'data' = data %>% dplyr::group_split(strata)),
.f = function(Nstrata, VarRest, data){
        (Nstrata/N_all)^2 * (1 - nrow(data)/Nstrata) *
        VarRest/nrow(data)
}) %>%
unlist() %>% sum()

#cálculo do intervalo de confiância para média
propz <- qnorm((1 - v_alpha)/2, lower.tail = F)
CI <- c('Limite␣Inferior' =
estimated_mu - propz*sqrt(estimated_varmu),
'Limite␣Superior' =
estimated_mu + propz*sqrt(estimated_varmu))
```

```
        # :::::::::::::::::::: Saída ::::::::::::::::::::

        output <- list(
        'Estimativa_da_média_popu._(Aleatória_com_estratos)' = estimated_mu,
        'Erro_padrão_do_estimador_da_média' = estimated_varmu %>% sqrt(),
        'Intervalo_de_confiança' = CI)

        return(output)
}
```

2.5.6 Estimador da média pelo método da razão multivariado

```
# :::::::::::::::::::: Parâmetros da função ::::::::::::::::::::

# model = ... # (e.g.X1/(X2 + X3 + X4) ~ strata) modelo considerando a razão
# data = ... #dados amostrais
# mean_popuAux = ... #média populacional para as var.'s auxiliares
# N_strata = ... #tamanho populcional de cada estrato
# v_alpha = ... #nível de significância

multiration_meanestimation <- function(model, data, mean_popuAux,
N_strata, v_alpha){

        # :::::::::::::::::::: Informações ::::::::::::::::::::

        #estrutura das variáveis pela formula
        vars_names <- model %>% as.formula() %>% all.vars()
        vars_ain <- vars_names[1]
        vars_aux <- vars_names[-c(1, length(vars_names))]
        vars_strata <- vars_names[length(vars_names)]

        #id. strata
        id_strata <- data %>%
        magrittr::extract2(vars_strata) %>%
        levels() %>%
        paste('strata', ., sep = '_')

        # :::::::::::::::::::: Informações populacionais ::::::::::::::::::::

        #média populacional da variável auxiliar
        mean_popuAux <- mean_popuAux %>%
        dplyr::group_split(strata) %>%
        purrr::map(.f = ~ .x %>% dplyr::select(-all_of(vars_strata)))

        #tamanho total da população
        N_strata <-
```

```
N_strata %>%
dplyr::group_split(strata) %>%
purrr::map(.f = ~ .x %>% dplyr::select(-all_of(vars_strata)))

N_all <-
N_strata %>%
purrr::map(.f = ~ .x %>% dplyr::select(n)) %>%
unlist() %>% sum()

# :::::::::::::::::::: Cálculos ::::::::::::::::::::

#tamanho da amostra para cada estrato
n_strata <-
data %>%
dplyr::group_by_at(vars_strata) %>%
dplyr::summarise(n = n()) %>%
dplyr::group_split(strata) %>%
purrr::map(.f = ~ .x %>% dplyr::select(-all_of(vars_strata)))

#média amostral para cada estrato
mean_strata <-
data %>%
dplyr::group_by_at(vars_strata) %>%
dplyr::summarise_all(.funs = list(~ .x %>% mean())) %>%
dplyr::group_split(strata) %>%
purrr::map(.f = ~ .x %>% dplyr::select(-all_of(vars_strata)))

#desvio padrão amostral para cada estrato
sd_strata <-
data %>%
dplyr::group_by_at(vars_strata) %>%
dplyr::summarise_all(.funs = list(~ .x %>% var())) %>%
dplyr::group_split(strata) %>%
purrr::map(.f = ~ .x %>% dplyr::select(-all_of(vars_strata)))

#cálculo dos pesos
weigth <- N_strata %>%
purrr::map(.f = ~ .x/N_all)

# :::::::::::::::::::: Estimação ::::::::::::::::::::

#matriz de médias amostrais
meanmatrix <-
purrr::map(.x = mean_strata,
.f = ~ .x %>%
as.numeric() %>%
matrix(nrow = c(vars_ain, vars_aux) %>% length(),
```

```
ncol = c(vars_ain, vars_aux) %>% length()))

#matriz de covariância amostral
Covariancematrix <-
data %>%
dplyr::group_by_at(vars_strata) %>%
dplyr::group_split() %>%
purrr::map(.f = ~ .x %>%
dplyr::select(-all_of(vars_strata)) %>%
cov())

#cálculo da matriz C (razão entre covariância e produto de médias)
Cmatrix <-
purrr::map2(.x = meanmatrix,
.y = Covariancematrix,
.f = ~ (1/(.x*t(.x)))*.y)

#vetor unitário
e <- rep_len(x = 1,
length.out = length(vars_aux)) %>%
as.matrix() %>% t()

Tmatrix <- cbind(rep_len(x = 1, length.out = length(vars_aux)),
diag(x = -1, nrow = length(vars_aux)))

Amatrix <- purrr::map(.x = Cmatrix,
.f = ~ Tmatrix%*%.x%*%t(Tmatrix))

num_expression <- purrr::map(.x = Amatrix,
.f = ~ e%*%solve(.x))

den_expression <- purrr::map(.x = Amatrix,
.f = ~ e%*%solve(.x)%*%t(e))

weight_esti <- purrr::map2(.x = num_expression ,
.y = den_expression,
.f = ~ .x/.y %>% as.numeric())

#razão estimada para cada estrato
ration_esti <-
purrr::map(.x = mean_strata,
.f = ~ as.numeric(.x[vars_ain])/as.numeric(.x[vars_aux]))

#média estimada por razão multivariada em cada estrato
estimated_mu_ration <-
purrr::pmap(.l = list('W' = weight_esti,
'r' = ration_esti,
```

```
        'MU' = mean_popuAux),
        .f = function(W, r, MU) sum(W * r * MU))

        #média estimada com ponderação por estratos
        estimated_mu <-
        purrr::map2(.x = weigth,
        .y = estimated_mu_ration,
        .f = ~ .x*.y) %>%
        unlist() %>% sum()

        #estimativa da variância do estimador da média
        estimated_varmu <-
        purrr::pmap(.l = list('MeanY' = estimated_mu_ration,
        'deno' = den_expression,
        'n' = n_strata),
        .f = function(MeanY, deno, n) ((MeanY^2)/n)*(1/deno)) %>%
        unlist() %>% sum()

        #cálculo do intervalo de confiância para média
        propz <- qnorm((1 - v_alpha)/2, lower.tail = F)
        CI <- c('Limite Inferior' = estimated_mu - propz*sqrt(estimated_varmu
        'Limite Superior' = estimated_mu + propz*sqrt(estimated_varmu))

        # :::::::::::::::::::: Saída ::::::::::::::::::::

        output <- list(
        'Estimativa_da_média_popu._(Aleatória_com_estratos)' = estimated_mu,
        'Erro_padrão_do_estimador_da_média' = estimated_varmu %>% sqrt(),
        'Intervalo_de_onfiança' = CI)

        return(output)

}
```

2.6 Referências

Arnab, R., 2017 Survey Sampling: Theory and Applications. Academic Press, first edition.

Beunckens, C., C. Sotto, G. Molenberghs, and G. Verbeke, 2009 A multifaceted sensitivity analysis of the Slovenian public opinion survey data. JOURNAL OF THE ROYAL STATISTICAL SOCIETY SERIES C-APPLIED STATISTICS **58**: 171–196.

Bodin, J.-L., 2020 A view on 50 years of life of the ISI: With a focus on ISI relations with official statistics. Statistical Journal of the IAOS **36**: 303–308.

Bolfarine, H. and W. O. Bussab, 2004 Elementos de amostragem. Editora Blucher.

Brand, M., 2002 Incremental singular value decomposition of uncertain data with missing values. In COMPUTER VISON - ECCV 2002, PT 1, edited by A. Heyden, G. Sparr, M. Nielsen, and P. Johan-

sen, volume 2350 of Lecture Notes in Computer Science, pp. 707–720, IT Univ Copenhagen; Univ Copenhagen; Lund Univ.

Cavalcanti, P. P. and C. T. d. S. Dias, 2021 Archetypal analysis as an imputation method and multivariate data augmentation .

Cebrián, A. A. and M. R. García, 1997 Variance Estimation Using Auxiliary Information: An Almost Unbiased Multivariate Ratio Estimator. Metrika **45**: 171–178.

Cochran, W. G., 1977 Sampling techniques. John Wiley & Sons, third edition.

Deming, W. E., 1990 Sample design in business research, volume 23. John Wiley & Sons.

Departamento de Astronomia Instituto de Astronomia, G. e. C. A., 2020 Início das estações do ano (2005–2020). Accessed: 2022-03-18.

Diana, G. and P. Francesco Perri, 2010 Improved estimators of the population mean for missing data. Communications in Statistics—Theory and Methods **39**: 3245–3251.

Ellington, E. H., G. Bastille-Rousseau, C. Austin, K. N. Landolt, B. A. Pond, E. E. Rees, N. Robar, and D. L. Murray, 2015 Using multiple imputation to estimate missing data in meta-regression. METHODS IN ECOLOGY AND EVOLUTION **6**: 153–163.

Epifanio, I., M. V. Ibanez, and A. Simo, 2020 Archetypal Analysis With Missing Data: See All Samples by Looking at a Few Based on Extreme Profiles. AMERICAN STATISTICIAN **74**: 169–183.

Gasparetto, S. C., S. M. D. S. Piedade, L. R. Angelocci, and V. A. Ozaki, 2021 COMPARAÇÃO ENTRE MÉTODOS DE IMPUTAÇÃO DE DADOS EM DIFERENTES INTENSIDADES AMOSTRAIS NA SÉRIE DE PRECIPITAÇÃO PLUVIAL DA ESALQ. Revista Brasileira de Climatologia **29**: 464–489.

Goodman, L. A. and H. O. Hartley, 1958 The Precision of Unbiased Ratio-Type Estimators. Journal of the American Statistical Association **53**: 491–508.

Hanif, M., Z. Ahmed, and M. Ahmad, 2009 Generalized multivariate ratio estimators using multi-auxiliary variables for multi-phase sampling. Pakistan Journal of Statistics **25**: 615–629.

He, Y., R. Yucel, and T. E. Raghunathan, 2011 A functional multiple imputation approach to incomplete longitudinal data. STATISTICS IN MEDICINE **30**: 1137–1156.

Howe, L. D., K. Tilling, A. Matijasevich, E. S. Petherick, A. C. Santos, L. Fairley, J. Wright, I. S. Santos, A. J. D. Barros, R. M. Martin, M. S. Kramer, N. Bogdanovich, L. Matush, H. Barros, and D. A. Lawlor, 2016 Linear spline multilevel models for summarising childhood growth trajectories: A guide to their application using examples from five birth cohorts. STATISTICAL METHODS IN MEDICAL RESEARCH **25**: 1854–1874.

Hulley, S. B., S. R. Cummings, W. S. Browner, D. G. Grady, and T. B. Newman, 2013 Designing clinical research. Lippincott Williams & Wilkins, fourth edi edition.

Jessen, R. J., 1943 Statistical investigation of a sample survey for obtaining farm facts. Iowa State University.

Johnson, R. A. and D. W. Wichern, 2007 Applied Multivariate Statistical Analysis. Prentice Hall, 6th edition.

Kenward, M. G. and J. Carpenter, 2007 Multiple imputation: current perspectives. Statistical Methods in Medical Research **16**: 199–218.

Kim, J. and M.-J. Park, 2019 Multiple imputation and synthetic data. KOREAN JOURNAL OF APPLIED STATISTICS **32**: 83–97.

Kruskal, W. and F. Mosteller, 1980 Representative Sampling, IV: The History of the Concept in Statistics, 1895-1939. International Statistical Review / Revue Internationale de Statistique **48**: 169.

Lohr, S. L., 2019 Sampling: Design and Analysis. Advanced (Cengage Learning), Cengage Learning.

Lohr, S. L., 2021 Sampling: design and analysis. CRC press.

Maitra, R., V. Melnykov, and S. N. Lahiri, 2012 Bootstrapping for Significance of Compact Clusters in Multidimensional Datasets. Journal of the American Statistical Association **107**: 378–392.

Melnykov, V., W.-C. Chen, and R. Maitra, 2012 MixSim: An R Package for Simulating Data to Study Performance of Clustering Algorithms. Journal of Statistical Software **51**: 1–25.

Myers, W. R., 2000 Handling Missing Data in Clinical Trials: An Overview. Drug information journal : DIJ / Drug Information Association **34**: 525–533.

Nath, K. and B. K. Sungh, 2018 Population Mean Estimation Using Ratio-cum Product Compromised-method of Imputation in Two-phase Sampling Scheme. Asian J. Math. Stat **11**: 27–39.

Neyman, J., 1992 pp. 123–150 in On the Two Different Aspects of the Representative Method: the Method of Stratified Sampling and the Method of Purposive Selection, edited by Kotz, S. and N. L. Johnson, Springer New York.

Olkin, I., 1958 Multivariate Ratio Estimation for Finite Populations. Biometrika **45**: 154.

Olufadi, Y. and C. Kadilar, 2014 A study on the chain ratio-type estimator of finite population variance. Journal of Probability and Statistics **2014**.

Patterson, H., 1950 Sampling on successive occasions with partial replacement of units. Journal of the Royal Statistical Society. Series B (Methodological) **12**: 241–255.

Peel, D., R. S. Waples, G. M. Macbeth, C. Do, and J. R. Ovenden, 2013 Accounting for missing data in the estimation of contemporary genetic effective population size (N-e). MOLECULAR ECOLOGY RESOURCES **13**: 243–253.

Qiu, W. and H. Joe., 2020 clusterGeneration: Random Cluster Generation (with Specified Degree of Separation). R package version 1.3.7.

Rao, J., 2005 Interplay Between Sample Survey Theory and Practice: An Appraisal. Survey Methodology **31**.

Rao, J. N. K. and D. R. Bellhouse, 1990 History and development of the theoretical foundations of survey based estimation and analysis. Survey Methodology **16**: 3–29.

Rao, J. N. K. and R. R. Sitter, 1995 Variance estimation under two-phase sampling with application to imputation for missing data. Biometrika **82**: 453–460.

Rubin, D. B., H. S. Stern, and V. Vehovar, 1995 Handling "Don't Know" Survey Responses: The Case of the Slovenian Plebiscite. Journal of the American Statistical Association **90**: 822–828.

Sano, N., 2020 Synthetic Data by Principal Component Analysis. In 20TH IEEE INTERNATIONAL CONFERENCE ON DATA MINING WORKSHOPS (ICDMW 2020), edited by G. DiFatta, V. Sheng, A. Cuzzocrea, C. Zaniolo, and X. Wu, International Conference on Data Mining Workshops, pp. 101–105, IEEE; IEEE Comp Soc; Univ Calabria; Mininglamp Technol.

Scheaffer, R. L., W. III Mendenhall, R. L. Ott, and K. Gerow, 2012 Elementary Survey Sampling. Cengage Learning, 7th edition.

Schneeberger, H. and K. Fleischer, 1993 The Multivariate Ratio Estimation: A Simulation Study. Jahrbücher für Nationalökonomie und Statistik **211**: 524–538.

Shukla, G. K., 1966 An Alternative Multivariate Ratio Estimate for Finite Population. Calcutta Statistical Association Bulletin **15**: 127–134.

Singh, G. N., S. Maurya, M. Khetan, and C. Kadilar, 2016 Some imputation methods for missing data in sample surveys. Hacettepe Journal of Mathematics and Statistics **45**: 1865–1880.

Singh, G. N. and S. Suman, 2019 Estimation of population mean using imputation methods for missing data under two-phase sampling design. Journal of Statistical Theory and Practice **13**: 1–24.

Singh, H. P., A. Gupta, and R. Tailor, 2021 Estimation of population mean using a difference-type exponential imputation method. Journal of Statistical Theory and Practice **15**: 1–43.

Singh, H. P., S. Kumar, and S. Bhougal, 2011 Multivariate ratio estimation in presence of non-response in successive sampling. Journal of Statistical Theory and Practice **5**: 591–611.

Singh, S., 2009 A new method of imputation in survey sampling. Statistics **43**: 499–511.

Tariq, M. U., M. N. Qureshi, and M. Hanif, 2021 Variance Estimators in the Presence of Measurement Errors Using Auxiliary Information **19**: 606–616.

Wissler, A., K. E. Blevins, and J. E. Buikstra, 2022 Missing data in bioarchaeology I: A review of the literature. AMERICAN JOURNAL OF BIOLOGICAL ANTHROPOLOGY **179**: 339–348.

Xu, Y. and R. Goodacre, 2018 On Splitting Training and Validation Set: A Comparative Study of Cross-Validation, Bootstrap and Systematic Sampling for Estimating the Generalization Performance of Supervised Learning. Journal of Analysis and Testing **2**: 249–262.

Yates, F. et al., 1949 Sampling methods for censuses and surveys. Sampling methods for censuses and surveys. .

Yuan, Y., 2011 Multiple Imputation Using SAS Software. Journal of Statistical Software **45**: 1–25.

Zhang, P., 2003 Multiple imputation: Theory and method. INTERNATIONAL STATISTICAL REVIEW **71**: 581–592.

3 UM NOVO MÉTODO DE IMPUTAÇÃO PARA DADOS AMOSTRAIS UTILIZANDO O MÉTODO DA RAZÃO MULTIVARIADA

Resumo

Existe uma necessidade de imputar dados ausentes para melhorar um perfil de resposta de modelos preditivos. Visto isso, o nosso trabalho propõe um novo método de imputação baseado no método da razão multivariada, baseado no método da razão multivariado de Olkin (1958). Foi realizado um estudo algébrico das expressões de viés e erro quadrático média do método da razão univariada para e da razão multivariada para imputação. Ambos os métodos foram comparados a dados simulados de uma distribuição normal multivariada, em que foi definido critério de: alta e baixa correlação; quantidades de 3, 5 e 10 variáveis auxiliares; e com proporções de 10%, 15%, 20% e 25% de valores ausentes. Observou-se que o método proposto apresentou melhores resultados de viés e erro quadrático médio para cenário com alta correlação. Os mesmos métodos também foi aplicado a um conjunto de dados da Embrapa afim de estima a média populacional do peso de melões. Observou-se que ambos apresentaram resultados similares.
Palavras-chave: valores ausentes; correlação; simulação multivariada.

3.1 Introdução

Inúmeras áreas necessitam do estudo de amostragem começando com a seleção de amostras representativas visando a características de interesse da populacional Lohr (2021) chegando em resultados inferenciais com menor erro e mair precisão. Entretanto, mesmo em delineamentos planejados, pode acontecer a ausência/perda de informações dessas características referentes a algumas unidades amostrais.

Essa falta de informação, é o que costuma-se chamar de valor ausente. Podendo ser ocasionada por diversos fatores, como a não resposta que unidades observacionais podem apresentar em pesquisas de opiniões Beunckens et al. (2009) Rubin et al. (1995), perda da unidade de estudo nas áreas de estudos biológicos e médicas Myers (2000)Peel et al. (2013)Ellington et al. (2015)Wissler et al. (2022), até por falha de equipamentos no momentos da coletadas informações na agrometeorologia Gasparetto et al. (2021). Em virtude disso, tem-se o estudo dos métodos de imputação, em que consiste no preenchimento do valor ausente por um novo valor mais verossímil, calculado por meio de métodos matemáticos e/ou probabilísticos.

Existem diferentes técnicas para gerar esses novos valores: como em estudos de modelagens supondo uma relação funcional entre as informações existentes das variáveis He et al. (2011); Howe et al. (2016). Método que envolvem técnicas da álgebra linear, como decomposição de valores singulares ou estudos de arquétipos Cavalcanti and Dias (2021); Epifanio et al. (2020); Brand (2002); **?**); **?**. Métodos que envolvem a aleatoriedades dos novos valores gerados como imputações múltiplas Zhang (2003); Yuan (2011); Kenward and Carpenter (2007). E mais atualmente, devido ao crescente do poder computacional, estudos que envolve a simulação extensiva de valores aleatórios Sano (2020); Kim and Park (2019).

Em estudo amostrais, em que o interesse é estimar uma característica populacional, como a média e o total, utiliza-se a relação entre as variáveis reposta e uma auxiliar para aumento da precisão dos estimadores, mesmo que em muitos casos sejam estimadores viesados, como os métodos da razão, regressão e da diferença Lohr (2021); Scheaffer et al. (2012). Para a imputação, diversos trabalhos utilizaram desses métodos para também estimar o valor ausente, e corrigindo esses viéses por meio de ponderações com medida estatísticas como a correlação, coeficiente de curtose e assimetria a variável auxiliar para gerar novos valores. Singh et al. (2021) sugeriu o uso de dois diferentes estimadores exponenciais que utilizam estatística amostrais da variável auxiliar para a imputação, e por meio disso

estimar a média populacional da variável de interesse. Esse estimador exponencial apresenta a combinação das estatística da correlação, coeficiente de variação, desvio-padrão, coeficientes de assimetria e curtose da variável auxiliar. SINGH (2009) também propôs o método de imputação que utiliza uma contante que possuem valores entre 0 e 1, como combinação das informações da variável auxiliares a fim de minimizar a variância do estimador da média populacional.

Esses artigos consideram apenas uma variável auxiliar, entretanto OLKIN (1958) realizou o estudo de um método multivariado que considerou mais de uma variável auxiliar para aprimorar o estimador da média populacional da variável resposta, porém não utilizou-se desse método para a imputação. Desta forma, o presente trabalho tem como propor a utilização do método da razão múltiplas para imputação com o objetivo de estimar a média.

Para isso realizou-se estudo algébricos das expressões matemáticas de viés e erro quadrático médio desse novo estimador, esses resultados foram aplicados a um conjunto de dados simulados com diferentes cenários de proporção de ausências de dados, e correlação entre as variáveis. E por fim, aplicar em um conjunto de dados reais que procura estimar o peso médio de melão que possui diferentes genótipos e ambientes.

3.2 Materiais

3.2.1 Conjunto de dados simulados

As simulações foram geradas por meio da biblioteca *MASS* com a função *mvrnorm* (**?**) oriundas de uma distribuição gaussiana (normal) multivariada $\mathbf{X} \sim Gaussian(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Os parâmetros foram fixados para o vetor de média ($\boldsymbol{\mu}$) de dimensão $p \times 1$, e matriz de variância e covariância ($\boldsymbol{\Sigma}$) com dimensão $p \times p$. Considerou-se 2 casos, variáveis tinham alta e baixa correlação, e com diferentes quantidades de variáveis conjuntas ($p = 3$, 5 e 10). O vetor de médias foram fixado para cada quantidade de variáveis auxiliares: para $p = 3$ com $\boldsymbol{\mu}^\top = [20, 40, 60]$; para $p = 5$ com $\boldsymbol{\mu}^\top = [20, 40, 60, 80, 100]$, e por fim com $p = 10$ com $\boldsymbol{\mu}^\top = [20, 40, 60, 80, 100, 30, 50, 70, 90, 110]$.

As matrizes de variância e covariância da distribuição normal multivariada foram geradas por meio da biblioteca *clusterGeneration* utilizando a função *genPositiveDefMat*, em que a partir de valores fixado de autovetores, são geradas aleatoriamente novamente autovalores $(\lambda_1, ..., \lambda_p)$ que são utilizados para a construção de matrizes ortogonais de autovetores ($\mathbf{Q} = [\boldsymbol{\alpha}_1, ..., \boldsymbol{\alpha}_p]^\top$), para que assim, seja possível calcular as matrizes de variância e covariância $= \mathbf{Q} \times diag(\lambda_1, \lambda_2, ..., \lambda_p) \times \mathbf{Q}^\top$. Para $p = 3$ foi considerado alta correlação $r > 0,9$ e baixa correlação $r < 0,2$; com $p = 5$ foi considerado alta correlação $r > 0,8$ e baixa correlação $r < 0,3$; e Para $p = 10$ foi considerado alta correlação $r > 0,75$ e baixa correlação $r < 0,4$. A seleção de matrizes que tivessem em média valores altos/baixos das correlação 2 - 2 das variáveis foi realizada de forma iterativa, isto é, cada iteração gerava uma matriz com as propriedades descritiva, e era realizada uma avaliação da correlação, caso essa matriz não apresentasse uma alta correlação, a matriz era descartada e gerada novamente.

Em seguida, com todos os cenários de quantidades de variáveis auxiliares (3 casos), e estrutura de alta e baixa correlação, realizou-se a exclusão de amostras, de forma aleatória considerando 4 proporções de ausência: com 10%; 15%; 20%; e 25% de valores.

A simulação procurou se aproximar de dados que tivessem características reais que necessitasse de imputação da variável de interesse.

3.2.2 Descrição da 2º análise - Dados EMBRAPA

O conjunto de dados foi fornecido pela EMBRAPA - Fortaleza, que tem como objetivo a caracterização a estrutura do fruto melão com 92 famílias (genótipos), realizado em 3 diferentes ambientes.

Por conta de fatores externos não controlados, ocorreu a ausência da germinação de 3 famílias: 52, 55 e 79. Em razão das condições locais serem homogêneas o experimento foi estruturado por meio de um DIC com 3 repetições. Certas atenções devem ser tomada pois em determinados ambientes, são apresentados quantidades menores de repetições ou até valores ausentes. Para visualizar esta estrutura, foi apresentado na Figura 3.1 os números de repetições em cada família no diferentes ambientes, no qual observa-se que as famílias 86 (ambiente 1) 5, 70 (ambiente 2), 31, 47, 92 e 95 (ambiente 3) apresentam valores ausentes, representados por quadrados em brancos.

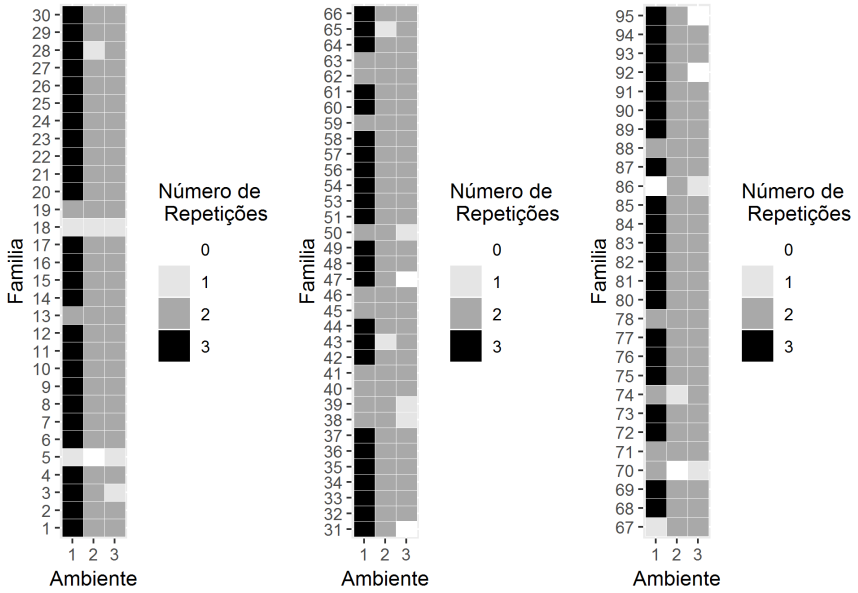


Figura 3.1. Heatmap para contagens de números de repetições de cada família em cada ambiente

3.3 Métodos

3.3.1 Método de imputação por razão para uma variável auxiliar

Considerando um população $\Omega = 1, 2, ..., N$, e nesta população tem-se como variáveis de interesse do pesquisador (Y) e uma variável auxiliar (X). Desta população foi realizado uma amostragem estratificada sem reposição $s = 1, 2, ..., i, ..., n$. No estudo houve observações que não foram possíveis de se obter a variável resposta y, apenas a auxiliar x.

Com isso, seja $r \subset n$ o número de observações que informaram a variável de interesse, vamos denotar o conjunto desses elementos como R. Consequentemente, o conjunto de elementos que são ausentes da variável de interesse é denotado por R^c.

Desta forma um método para imputação pelo método da razão é expressa:

$$y_{\cdot i} = \begin{cases} y_i, & \text{se } i \in R \\ \hat{b}x_i, & \text{se } i \in R^c \end{cases}$$

em que $\hat{b} = \dfrac{\sum_{i \in R} y_i}{\sum_{i \in R} x_i} = \dfrac{\bar{y}_r}{\bar{x}_r}$. Desta forma, tem-se que a média amostral é dado por:

$$\bar{y}_s = \frac{1}{n} \sum_{i=1}^{n} y_{\cdot i} \tag{3.1}$$

A partir da Equação 3.5 pode-se reescrever-la da forma:

$$
\begin{aligned}
\bar{y}_s &= \frac{1}{n}\sum_{i=1}^{n} y_{\cdot i} = \\
&= \frac{1}{n}(\sum_{i=1}^{r} y_i + \sum_{i=r+1}^{n-r} \hat{b}x_i) = \\
&= \frac{1}{n}(\sum_{i=1}^{r} y_i + \frac{\bar{y}_r}{\bar{x}_r}\sum_{i=r+1}^{n-r} x_i) = \\
&= \frac{1}{n}(\sum_{i=1}^{r} y_i + \frac{\bar{y}_r}{\bar{x}_r}\sum_{i=r+1}^{n-r} x_i)\frac{r}{r} = \\
&= \frac{1}{n}(\sum_{i=1}^{r} y_i\frac{r}{r} + \frac{\bar{y}_r}{\bar{x}_r}\sum_{i=r+1}^{n-r} x_i\frac{r}{r}) = \\
&= \frac{1}{n}(\bar{y}_r r + \frac{\bar{y}_r}{\bar{x}_r}\sum_{i=r+1}^{n-r} x_i\frac{r}{r}) = \\
&= \frac{1}{n}\bar{y}_r(r + \frac{1}{\bar{x}_r}\sum_{i=r+1}^{n-r} x_i\frac{r}{r}) = \\
&= \frac{1}{n}\bar{y}_r r(1 + \frac{1}{r\bar{x}_r}\sum_{i=r+1}^{n-r} x_i) = \\
&= \frac{1}{n}\bar{y}_r r(\frac{r\bar{x}_r + \sum_{i=r+1}^{n-r} x_i}{r\bar{x}_r}) = \\
&= \bar{y}_r\frac{1}{\bar{x}_r}\frac{1}{n}(r\bar{x}_r + \sum_{i=r+1}^{n-r} x_i) = \\
&= \bar{y}_r\frac{1}{\bar{x}_r}\frac{1}{n}(r\frac{\sum_{i=1}^{r} x_i}{r} + \sum_{i=r+1}^{n-r} x_i) = \\
&= \bar{y}_r\frac{1}{\bar{x}_r}\frac{1}{n}(\sum_{i=1}^{r} x_i + \sum_{i=r+1}^{n-r} x_i) = \\
&= \bar{y}_r\frac{1}{\bar{x}_r}\frac{(\sum_{i=1}^{r} x_i + \sum_{i=r+1}^{n-r} x_i)}{n} = \\
&= \bar{y}_r\frac{1}{\bar{x}_r}\frac{\sum_{i=1}^{n} x_i}{n} = \\
&= \bar{y}_r\frac{\bar{x}_n}{\bar{x}_r}
\end{aligned}
$$

Para o cálculo das expressões aproximadas do viés e erro quadrático médio, utiliza-se axiomas discutidas nos trabalhos Singh (2009); Singh et al. (2016); Nath and Sungh (2018); Singh and Suman (2019).

$$\epsilon = \frac{\bar{y}_r - \mu_Y}{\mu_Y}, \qquad \delta = \frac{\bar{x}_r - \mu_X}{\mu_X} \text{ e} \qquad \eta = \frac{\bar{x}_n - \mu_X}{\mu_X} \tag{3.2}$$

utilizando conceitos de amostragem em dois estágios (Rao and Sitter, 1995)

$$E(\epsilon) = E(\delta) = E(\eta) = 0$$

e

$$E(\epsilon^2) = (\tfrac{1}{r} - \tfrac{1}{N})C_Y^2 \quad E(\delta^2) = (\tfrac{1}{r} - \tfrac{1}{N})C_X^2 \quad E(\epsilon\delta) = (\tfrac{1}{r} - \tfrac{1}{N})\rho C_Y C_X$$

$$E(\eta^2) = (\tfrac{1}{n} - \tfrac{1}{N})C_X^2 \quad E(\delta\eta) = (\tfrac{1}{n} - \tfrac{1}{N})C_X^2 \quad E(\epsilon\eta) = (\tfrac{1}{n} - \tfrac{1}{N})\rho C_Y C_X$$

em que $C_Y^2 = \dfrac{S_Y^2}{\mu_Y^2}$, $C_X^2 = \dfrac{S_X^2}{\mu_X^2}$ e $\rho = \dfrac{S_{XY}}{S_X S_Y}$

Reescrevendo as expressões Equação 3.7 temos:

$$\bar{y}_r = (1+\epsilon)\mu_Y, \quad \bar{x}_r = (1+\delta)\mu_X \text{ e } \quad \bar{x}_n = (1+\eta)\mu_X \tag{3.3}$$

substituindo as expressões na equação, tem-se-que.

$$\begin{aligned}\bar{y}_r \frac{\bar{x}_n}{\bar{x}_r} &= (1+\epsilon)\mu_Y \frac{(1+\eta)\mu_X}{(1+\delta)\mu_X} = \\ &= \mu_Y(1+\epsilon)(1+\eta)\frac{1}{(1+\delta)} =\end{aligned}$$

por expansão de séries de Taylor. Seja $|\delta| < 1$, tem-se-que $\dfrac{1}{(1+\delta)} = 1 - \delta + \delta^2 + O(\delta^3)$.

$$\begin{aligned}&= \mu_Y(1+\epsilon)(1+\eta)(1-\delta+\delta^2+O(\delta^3)) = \\ &= \mu_Y(1+\eta+\epsilon+\epsilon\eta)(1-\delta+\delta^2+O(\delta^3)) = \\ &= \mu_Y(1+\eta+\epsilon+\epsilon\eta-\delta-\eta\delta-\epsilon\delta+\delta^2+O(\delta^3))\end{aligned}$$

aplicando o valor esperado, apenas nas expressões até segunda ordem, tem-se

$$\begin{aligned}Bias[\bar{y}_s] &= E[\bar{y}_s] - \mu_Y \approx \\ &\approx E[\mu_Y(1+\eta+\epsilon+\epsilon\eta-\delta-\eta\delta-\epsilon\delta+\delta^2)] - \mu_Y = \\ &= \mu_Y(E[1]+E[\eta]+E[\epsilon]+E[\epsilon\eta]-E[\delta]-E[\eta\delta]-E[\epsilon\delta]+E[\delta^2]) - \mu_Y = \\ &= \mu_Y\left[1+0+0+(\frac{1}{n}-\frac{1}{N})\rho C_Y C_X - 0 - (\frac{1}{n}-\frac{1}{N})C_X^2 - (\frac{1}{r}-\frac{1}{N})\rho C_Y C_X + (\frac{1}{r}-\frac{1}{N})C_X^2\right] - \mu_Y = \\ &= \mu_Y\left[1+(\frac{1}{n}-\frac{1}{N})\rho C_Y C_X - (\frac{1}{n}-\frac{1}{N})C_X^2 - (\frac{1}{r}-\frac{1}{N})\rho C_Y C_X + (\frac{1}{r}-\frac{1}{N})C_X^2 - 1\right] = \\ &= \mu_Y\left[(\frac{1}{n}-\frac{1}{N}-\frac{1}{r}+\frac{1}{N})\rho C_Y C_X - (\frac{1}{n}-\frac{1}{N}+\frac{1}{r}-\frac{1}{N})C_X^2\right] = \\ &= \mu_Y\left[(\frac{1}{n}-\frac{1}{r})\rho C_Y C_X - (\frac{1}{n}-\frac{1}{r})C_X^2\right] = \\ &= (\frac{1}{n}-\frac{1}{r})(\rho C_Y C_X - C_X^2)\mu_Y\end{aligned}$$

Agora o cálculo do erro quadrático médio para o estimador vamos utilizar a as aproximações de primeira ordem:

$$
\begin{aligned}
MSE[\bar{y}_s] &= E[\bar{y}_s - \mu_Y]^2 \approx \\
&\approx E[\mu_Y(1+\eta+\epsilon-\delta) - \mu_Y]^2 = \\
&= E[\mu_Y(1+\eta+\epsilon-\delta-1)]^2 = \\
&= E[\mu_Y^2(\eta+\epsilon-\delta)^2] = \\
&= \mu_Y^2 E[(\eta+\epsilon-\delta)^2] = \\
&= \mu_Y^2 E[(\eta^2+\epsilon^2+\delta^2+2\eta\epsilon-2\eta\delta-2\epsilon\delta)] = \\
&= \mu_Y^2 \Big\{ E[\eta^2] + E[\epsilon^2] + E[\delta^2] + 2E[\eta\epsilon] - 2E[\eta\delta] - 2E[\epsilon\delta] \Big\} = \\
&= \mu_Y^2 \Big\{ (\frac{1}{n} - \frac{1}{N})C_X^2 + (\frac{1}{r} - \frac{1}{N})C_Y^2 + (\frac{1}{r} - \frac{1}{N})C_X^2 + \\
&\qquad + 2[(\frac{1}{n} - \frac{1}{N})\rho C_Y C_X] - 2[(\frac{1}{n} - \frac{1}{N})C_X^2] - 2[(\frac{1}{r} - \frac{1}{N})\rho C_Y C_X] \Big\} = \\
&= \mu_Y^2 \Big\{ (\frac{1}{n} - \frac{1}{N} + \frac{1}{r} - \frac{1}{N} - \frac{2}{n} + \frac{2}{N})C_X^2 + (\frac{1}{r} - \frac{1}{N})C_Y^2 + (\frac{1}{n} - \frac{1}{N} - \frac{1}{r} + \frac{1}{N})2\rho C_Y C_X \Big\} = \\
&= \mu_Y^2 \Big\{ (\frac{1}{r} - \frac{1}{n})C_X^2 + (\frac{1}{r} - \frac{1}{N})C_Y^2 + (\frac{1}{n} - \frac{1}{r})2\rho C_Y C_X \Big\} = \\
&= \mu_Y^2 \Big\{ (\frac{1}{r} - \frac{1}{N})C_Y^2 + (\frac{1}{r} - \frac{1}{n})(C_X^2 - 2\rho C_Y C_X) \Big\}
\end{aligned}
\tag{3.4}
$$

3.3.2 Método de imputação por razão para diversas variáveis auxiliares

Considerando um população $\Omega = 1, 2, ..., N$, e nesta população tem-se como variáveis de interesse do pesquisador (Y) e diversas variáveis auxiliares $(X_1, X_2, ..., X_j, ..., X_p)$. Desta população foi realizado uma amostragem estratificada sem reposição $s = 1, 2, ..., i, ..., n$. No estudo houve observações que não foram possíveis de se obter a variável resposta y, apenas as auxiliares $x_1, x_2, ..., x_j, ..., x_p$.

Com isso, seja $r \subset n$ o número de observações que informaram a variável de interesse, vamos denotar o conjunto desses elementos como R. Consequentemente, o conjunto de elementos que são ausentes da variável de interesse é denotado por R^c

Desta forma um método para imputação, é pelo método da razão multivariada.

As observações imputada tem como expressão matemática:

$$
y_i^* = \begin{cases} y_i, & \text{se } i \in R \\ \frac{1}{p}(\hat{b}_1 x_{i1} + \hat{b}_2 x_{i2} + ... + \hat{b}_p x_{ip}) = \frac{1}{p}\sum_{j=1}^{p} \hat{b}_j x_{ij}, & \text{se } i \in R^c \end{cases}
$$

em que $\hat{b}_j = \dfrac{\sum_{i \in R} y_i}{\sum_{i \in R} x_{ij}} = \dfrac{\bar{y}_r}{\bar{x}_{rj}}$. Desta forma, tem-se que a média amostral é dado por:

$$
\bar{y}_s = \frac{1}{n}\sum_{i=1}^{n} y_i^* \tag{3.5}
$$

A partir da Equação 3.5 pode-se reescrever-la da forma:

$$
\begin{aligned}
\bar{y}_s &= \frac{1}{n}\sum_{i=1}^{n} y_i^* = \\
&= \frac{1}{n}\left[\sum_{i=1}^{r} y_i + \frac{1}{p}\sum_{i=r+1}^{n-r}(\hat{b}_1 x_{i1} + \hat{b}_2 x_{i2} + ... + \hat{b}_p x_{ip})\right] = \\
&= \frac{1}{n}\left[\sum_{i=1}^{r} y_i + \frac{1}{p}(\sum_{i=r+1}^{n-r}\hat{b}_1 x_{i1} + \sum_{i=r+1}^{n-r}\hat{b}_2 x_{i2} + ... + \sum_{i=r+1}^{n-r}\hat{b}_p x_{ip})\right] = \\
&= \frac{1}{n}\left[\sum_{i=1}^{r} y_i + \left(\frac{1}{p}\hat{b}_1\sum_{i=r+1}^{n-r} x_{i1} + \frac{1}{p}\hat{b}_2\sum_{i=r+1}^{n-r} x_{i2} + ... + \frac{1}{p}\hat{b}_p\sum_{i=r+1}^{n-r} x_{ip}\right)\right] = \\
&= \frac{1}{n}\frac{p}{p}\left[\sum_{i=1}^{r} y_i + \left(\frac{1}{p}\hat{b}_1\sum_{i=r+1}^{n-r} x_{i1} + \frac{1}{p}\hat{b}_2\sum_{i=r+1}^{n-r} x_{i2} + ... + \frac{1}{p}\hat{b}_p\sum_{i=r+1}^{n-r} x_{ip}\right)\right] = \\
&= \frac{1}{n}\frac{1}{p}\left[p\sum_{i=1}^{r} y_i + \left(\hat{b}_1\sum_{i=r+1}^{n-r} x_{i1} + \hat{b}_2\sum_{i=r+1}^{n-r} x_{i2} + ... + \hat{b}_p\sum_{i=r+1}^{n-r} x_{ip}\right)\right] = \\
&= \frac{1}{n}\frac{1}{p}\left[(\sum_{i=1}^{r} y_i + \hat{b}_1\sum_{i=r+1}^{n-r} x_{i1}) + (\sum_{i=1}^{r} y_i + \hat{b}_2\sum_{i=r+1}^{n-r} x_{i2}) + ... + (\sum_{i=1}^{r} y_i + \hat{b}_p\sum_{i=r+1}^{n-r} x_{ip})\right] = \\
&= \frac{1}{n}\frac{1}{p}\left[(\sum_{i=1}^{r} y_i + \frac{\sum_{i=1}^{r} y_i}{\sum_{i=1}^{r} x_{i1}}\sum_{i=r+1}^{n-r} x_{i1}) + (\sum_{i=1}^{r} y_i + \frac{\sum_{i=1}^{r} y_i}{\sum_{i=1}^{r} x_{i2}}\sum_{i=r+1}^{n-r} x_{i2}) + ... + (\sum_{i=1}^{r} y_i + \frac{\sum_{i=1}^{r} y_i}{\sum_{i=1}^{r} x_{ip}}\sum_{i=r+1}^{n-r} x_{ip})\right] = \\
&= \frac{1}{n}\frac{1}{p}\left[\frac{\sum_{i=1}^{r} y_i}{\sum_{i=1}^{r} x_{i1}}(\sum_{i=1}^{r} x_{i1} + \sum_{i=r+1}^{n-r} x_{i1}) + \frac{\sum_{i=1}^{r} y_i}{\sum_{i=1}^{r} x_{i2}}(\sum_{i=1}^{r} x_{i2} + \sum_{i=r+1}^{n-r} x_{i2}) + ... + \frac{\sum_{i=1}^{r} y_i}{\sum_{i=1}^{r} x_{ip}}(\sum_{i=1}^{r} x_{ip} + \sum_{i=r+1}^{n-r} x_{ip})\right] = \\
&= \frac{1}{n}\frac{1}{p}\left[\frac{\sum_{i=1}^{r} y_i}{\sum_{i=1}^{r} x_{i1}}(\sum_{i=1}^{n} x_{i1}) + \frac{\sum_{i=1}^{r} y_i}{\sum_{i=1}^{r} x_{i2}}(\sum_{i=1}^{n} x_{i2}) + ... + \frac{\sum_{i=1}^{r} y_i}{\sum_{i=1}^{r} x_{ip}}(\sum_{i=1}^{n} x_{ip})\right] = \\
&= \frac{1}{p}\left[\frac{\sum_{i=1}^{r} y_i}{\sum_{i=1}^{r} x_{i1}}\frac{\sum_{i=1}^{n} x_{i1}}{n} + \frac{\sum_{i=1}^{r} y_i}{\sum_{i=1}^{r} x_{i2}}\frac{\sum_{i=1}^{n} x_{i2}}{n} + ... + \frac{\sum_{i=1}^{r} y_i}{\sum_{i=1}^{r} x_{ip}}\frac{\sum_{i=1}^{n} x_{ip}}{n}\right] = \\
&= \frac{1}{p}\left[\frac{\bar{y}_r}{\bar{x}_{r1}}\bar{x}_{n1} + \frac{\bar{y}_r}{\bar{x}_{r2}}\bar{x}_{n2} + ... + \frac{\bar{y}_r}{\bar{x}_{rp}}\bar{x}_{np}\right] = \\
&= \frac{1}{p}\bar{y}_r\left(\frac{\bar{x}_{n1}}{\bar{x}_{r1}} + \frac{\bar{x}_{n2}}{\bar{x}_{r2}} + ... + \frac{\bar{x}_{np}}{\bar{x}_{rp}}\right)
\end{aligned}
\tag{3.6}
$$

Para o cálculo das expressões aproximadas do viés e erro quadrático médio, utiliza-se axiomas discutidas nos trabalhos Singh (2009); Singh et al. (2016); Nath and Sungh (2018); Singh and Suman (2019) estendido para o caso de múltiplas variáveis auxiliares ($j = 1, ..., p$).

$$
\epsilon_Y = \frac{\bar{y}_r - \mu_Y}{\mu_Y}, \qquad \epsilon^r_{X_j} = \frac{\bar{x}^r_j - \mu_{X_j}}{\mu_{X_j}} \text{ e} \qquad \epsilon^n_{X_j} = \frac{\bar{x}^n_j - \mu_{X_j}}{\mu_{X_j}} \tag{3.7}
$$

utilizando conceitos de amostragem em dois estágios (Rao and Sitter, 1995)

$$
E[\epsilon_Y] = E[\epsilon^r_{X_j}] = E[\epsilon^n_{X_j}] = 0
$$

e

$$
E[\epsilon^2_Y] = (\tfrac{1}{r} - \tfrac{1}{N})C^2_Y \qquad E[\epsilon^{r2}_{X_j}] = (\tfrac{1}{r} - \tfrac{1}{N})C^2_{X_j} \qquad E[\epsilon^{n2}_{X_j}] = (\tfrac{1}{n} - \tfrac{1}{N})C^2_{X_j}
$$

$$
E[\epsilon_Y\epsilon^r_{X_j}] = (\tfrac{1}{r} - \tfrac{1}{N})\rho C_Y C_{X_j} \quad E[\epsilon_Y\epsilon^n_{X_j}] = (\tfrac{1}{n} - \tfrac{1}{N})\rho C_Y C_{X_j} \quad E[\epsilon^r_{X_j}\epsilon^n_{X_j}] = (\tfrac{1}{n} - \tfrac{1}{N})C^2_{X_j}
$$

em que $C^2_Y = \dfrac{S^2_Y}{\mu^2_Y}$, $C^2_{X_j} = \dfrac{S^2_{X_j}}{\mu^2_{X_j}}$ e $\rho = \dfrac{S_{X_jY}}{S_{X_j}S_Y}$

Reescrevendo as expressões Equação 3.7 temos:

$$\bar{y}_r = (1+\epsilon_Y)\mu_Y, \qquad \bar{x}_r = (1+\epsilon^r_{X_j})\mu_{X_j} \text{ e} \qquad \bar{x}_n = (1+\epsilon^n_{X_j})\mu_{X_j} \tag{3.8}$$

Agora, reescrevendo a Equação 3.6 com os termos definidos pela Equação 3.8.

$$\begin{aligned}
\bar{y}_s &= \frac{1}{p}\bar{y}_r\left(\frac{\bar{x}_{n1}}{\bar{x}_{r1}} + \frac{\bar{x}_{n2}}{\bar{x}_{r2}} + ... + \frac{\bar{x}_{np}}{\bar{x}_{rp}}\right) = \\
&= \frac{1}{p}(1+\epsilon_Y)\mu_Y\left(\frac{(1+\epsilon^n_{X_1})\mu_{X_1}}{(1+\epsilon^r_{X_1})\mu_{X_1}} + \frac{(1+\epsilon^n_{X_2})\mu_{X_2}}{(1+\epsilon^r_{X_2})\mu_{X_2}} + ... + \frac{(1+\epsilon^n_{X_p})\mu_{X_p}}{(1+\epsilon^r_{X_p})\mu_{X_p}}\right) = \\
&= \frac{1}{p}(1+\epsilon_Y)\mu_Y\left((1+\epsilon^n_{X_1})\frac{1}{(1+\epsilon^r_{X_1})} + (1+\epsilon^n_{X_2})\frac{1}{(1+\epsilon^r_{X_2})} + ... + (1+\epsilon^n_{X_p})\frac{1}{(1+\epsilon^r_{X_p})}\right)
\end{aligned}$$

com $|\epsilon^r_{X_j}| < 1$ por expansão de séries de Taylor de segunda ordem, tem-se-que $\frac{1}{(1+\epsilon^r_{X_j})} = 1 - \epsilon^r_{X_j} + \epsilon^{r2}_{X_j} + O(\cdot^3)$

$$\begin{aligned}
&= \frac{1}{p}(1+\epsilon_Y)\mu_Y\Big((1+\epsilon^n_{X_1})(1-\epsilon^r_{X_1}+\epsilon^{r2}_{X_1}) + (1+\epsilon^n_{X_2})(1-\epsilon^r_{X_2}+\epsilon^{r2}_{X_2}) + ...+ \\
&\qquad + (1+\epsilon^n_{X_p})(1-\epsilon^r_{X_p}+\epsilon^{r2}_{X_p}) + O(\cdot^3)\Big) = \\
&= \frac{\mu_Y}{p}(1+\epsilon_Y)\Big((1-\epsilon^r_{X_1}+\epsilon^{r2}_{X_1}+\epsilon^n_{X_1}-\epsilon^n_{X_1}\epsilon^r_{X_1}) + (1-\epsilon^r_{X_2}+\epsilon^{r2}_{X_2}+\epsilon^n_{X_2}-\epsilon^n_{X_2}\epsilon^r_{X_2}) + ...+ \\
&\qquad + (1-\epsilon^r_{X_p}+\epsilon^{r2}_{X_p}+\epsilon^n_{X_p}-\epsilon^n_{X_p}\epsilon^r_{X_p}) + O(\cdot^3)\Big) = \\
&= \frac{\mu_Y}{p}\Big[(1-\epsilon^r_{X_1}+\epsilon^{r2}_{X_1}+\epsilon^n_{X_1}-\epsilon^n_{X_1}\epsilon^r_{X_1}) + ... + (1-\epsilon^r_{X_p}+\epsilon^{r2}_{X_p}+\epsilon^n_{X_p}-\epsilon^n_{X_p}\epsilon^r_{X_p}) + \\
&\qquad + (\epsilon_Y - \epsilon_Y\epsilon^r_{X_1} + \epsilon_Y\epsilon^n_{X_1}) + ... + (\epsilon_Y - \epsilon_Y\epsilon^r_{X_p} + \epsilon_Y\epsilon^n_{X_p}) + O(\cdot^3)\Big] = \\
&= \frac{\mu_Y}{p}\Big[\sum_{j=1}^{p}(1-\epsilon^r_{X_j}+\epsilon^{r2}_{X_j}+\epsilon^n_{X_j}-\epsilon^n_{X_j}\epsilon^r_{X_j}) + \sum_{j=1}^{p}(\epsilon_Y - \epsilon_Y\epsilon^r_{X_j} + \epsilon_Y\epsilon^n_{X_j}) + O(\cdot^3)\Big]
\end{aligned}$$

Desta forma, para o calculo do viés, calcula-se

$$
\begin{aligned}
Bias[\bar{y}_s] &= E\left[\bar{y}_s\right] - \mu_Y \approx \\
&\approx E\left[\frac{\mu_Y}{p}\left(\sum_{j=1}^{p}(1-\epsilon^r_{X_j}+\epsilon^{r2}_{X_j}+\epsilon^n_{X_j}-\epsilon^n_{X_j}\epsilon^r_{X_j})+\sum_{j=1}^{p}(\epsilon_Y-\epsilon_Y\epsilon^r_{X_j}+\epsilon_Y\epsilon^n_{X_j})\right)\right]-\mu_Y = \\
&= \frac{\mu_Y}{p}E\left[\sum_{j=1}^{p}(1-\epsilon^r_{X_j}+\epsilon^{r2}_{X_j}+\epsilon^n_{X_j}-\epsilon^n_{X_j}\epsilon^r_{X_j})+\sum_{j=1}^{p}(\epsilon_Y-\epsilon_Y\epsilon^r_{X_j}+\epsilon_Y\epsilon^n_{X_j})\right]-\mu_Y = \\
&= \frac{\mu_Y}{p}\left[\sum_{j=1}^{p}(E[1]-E[\epsilon^r_{X_j}]+E[\epsilon^{r2}_{X_j}]+E[\epsilon^n_{X_j}]-E[\epsilon^n_{X_j}\epsilon^r_{X_j}])+\right. \\
&\qquad\left.+\sum_{j=1}^{p}(E[\epsilon_Y]-E[\epsilon_Y\epsilon^r_{X_j}]+E[\epsilon_Y\epsilon^n_{X_j}])\right]-\mu_Y = \\
&= \frac{\mu_Y}{p}\left[\sum_{j=1}^{p}\left(1-0+(\frac{1}{r}-\frac{1}{N})C^2_{X_j}+0-(\frac{1}{n}-\frac{1}{N})C^2_{X_j}+\right.\right. \\
&\qquad\left.\left.+0-(\frac{1}{r}-\frac{1}{N})\rho C_Y C_{X_j}+(\frac{1}{n}-\frac{1}{N})\rho C_Y C_{X_j}\right)\right]-\mu_Y = \\
&= \frac{\mu_Y}{p}\left[p+\sum_{j=1}^{p}\left((\frac{1}{r}-\frac{1}{N})C^2_{X_j}-(\frac{1}{n}-\frac{1}{N})C^2_{X_j}-\right.\right. \\
&\qquad\left.\left.-(\frac{1}{r}-\frac{1}{N})\rho C_Y C_{X_j}+(\frac{1}{n}-\frac{1}{N})\rho C_Y C_{X_j}\right)\right]-\mu_Y = \\
&= \mu_Y\left[1+\frac{1}{p}\sum_{j=1}^{p}\left((\frac{1}{r}-\frac{1}{N})C^2_{X_j}-(\frac{1}{n}-\frac{1}{N})C^2_{X_j}-\right.\right. \\
&\qquad\left.\left.-(\frac{1}{r}-\frac{1}{N})\rho C_Y C_{X_j}+(\frac{1}{n}-\frac{1}{N})\rho C_Y C_{X_j}\right)-1\right] = \\
&= \mu_Y\left[\frac{1}{p}\sum_{j=1}^{p}\left((\frac{1}{r}-\frac{1}{N}-\frac{1}{n}+\frac{1}{N})C^2_{X_j}+(\frac{1}{n}-\frac{1}{N}-\frac{1}{r}+\frac{1}{N})\rho C_Y C_{X_j}\right)\right] = \\
&= \mu_Y\left[\frac{1}{p}\sum_{j=1}^{p}\left((\frac{1}{r}-\frac{1}{n})C^2_{X_j}+(\frac{1}{n}-\frac{1}{r})\rho C_Y C_{X_j}\right)\right] = \\
&= \mu_Y\left[\frac{1}{p}\sum_{j=1}^{p}\left((\frac{1}{r}-\frac{1}{n})(C^2_{X_j}-\rho C_Y C_{X_j})\right)\right]
\end{aligned}
$$

Agora o cálculo do erro quadrático médio para o estimador vamos utilizar a as aproximações de primeira ordem:

$$
\begin{aligned}
MSE[\bar{y}_s] &= E[\bar{y}_s - \mu_Y]^2 \approx \\
&\approx E\left[\frac{\mu_Y}{p}\left(\sum_{j=1}^{p}(1+\epsilon_Y+\epsilon^n_{X_j}-\epsilon^r_{X_j})\right)-\mu_Y\right]^2 = \\
&= E\left[\mu_Y^2\left(\frac{1}{p}\sum_{j=1}^{p}(1+\epsilon_Y+\epsilon^n_{X_j}-\epsilon^r_{X_j})-1\right)^2\right] = \\
&= \mu_Y^2 E\left[\left(\frac{1}{p}p+\frac{1}{p}\sum_{j=1}^{p}(\epsilon_Y+\epsilon^n_{X_j}-\epsilon^r_{X_j})-1\right)^2\right] = \\
&= \mu_Y^2 E\left[\left(1+\frac{1}{p}\sum_{j=1}^{p}(\epsilon_Y+\epsilon^n_{X_j}-\epsilon^r_{X_j})-1\right)^2\right] = \\
&= \mu_Y^2 E\left[\left(\frac{1}{p}\sum_{j=1}^{p}(\epsilon_Y+\epsilon^n_{X_j}-\epsilon^r_{X_j})\right)^2\right] = \\
&= \frac{\mu_Y^2}{p^2} E\left[\left(\sum_{j=1}^{p}(\epsilon_Y+\epsilon^n_{X_j}-\epsilon^r_{X_j})\right)^2\right] = \\
&= \frac{\mu_Y^2}{p^2} E\left[\left(p\epsilon_Y+\sum_{j=1}^{p}(\epsilon^n_{X_j}-\epsilon^r_{X_j})\right)^2\right] = \\
&= \frac{\mu_Y^2}{p^2} E\left[(p\epsilon_Y)^2+2p\epsilon_Y\sum_{j=1}^{p}(\epsilon^n_{X_j}-\epsilon^r_{X_j})+\left(\sum_{j=1}^{p}(\epsilon^n_{X_j}-\epsilon^r_{X_j})\right)^2\right] = \\
&= \frac{\mu_Y^2}{p^2}\left\{E\left[(p\epsilon_Y)^2\right]+E\left[2p\epsilon_Y\sum_{j=1}^{p}(\epsilon^n_{X_j}-\epsilon^r_{X_j})\right]+E\left[\sum_{j=1}^{p}(\epsilon^n_{X_j}-\epsilon^r_{X_j})\right]^2\right\} = \\
&= \frac{\mu_Y^2}{p^2}\left\{p^2E\left[\epsilon_Y^2\right]+2pE\left[\sum_{j=1}^{p}(\epsilon_Y\epsilon^n_{X_j}-\epsilon_Y\epsilon^r_{X_j})\right]+E\left[\sum_{j=1}^{p}(\epsilon^n_{X_j}-\epsilon^r_{X_j})\right]^2\right\} = \\
&= \frac{\mu_Y^2}{p^2}\left\{p^2E\left[\epsilon_Y^2\right]+2p\sum_{j=1}^{p}E[\epsilon_Y\epsilon^n_{X_j}-\epsilon_Y\epsilon^r_{X_j}]+E\left[\sum_{j=1}^{p}(\epsilon^n_{X_j}-\epsilon^r_{X_j})\right]^2\right\}
\end{aligned}
$$

Supondo que $E\left[\sum_{j=1}^{p}(\epsilon^n_{X_j}-\epsilon^r_{X_j})^2\right]$ é finito, pela desigualdade de Jensen para funções convexa (ex. $f(x)=x^2$) tem-se que $E\left[\sum_{j=1}^{p}(\epsilon^n_{X_j}-\epsilon^r_{X_j})\right]^2 \leq E\left[\sum_{j=1}^{p}(\epsilon^n_{X_j}-\epsilon^r_{X_j})^2\right]$.

Verificando $E\left[\sum_{j=1}^{p}(\epsilon^n_{X_j}-\epsilon^r_{X_j})^2\right] < \infty$, temos

$$E\Big[\sum_{j=1}^{p}(\epsilon^n_{X_j}-\epsilon^r_{X_j})^2\Big]=$$

$$=E\Big[\sum_{j=1}^{p}(\epsilon^{n2}_{X_j}-2\epsilon^n_{X_j}\epsilon^r_{X_j}+\epsilon^{r2}_{X_j})\Big]=$$

$$=\sum_{j=1}^{p}E\Big[(\epsilon^{n2}_{X_j}-2\epsilon^n_{X_j}\epsilon^r_{X_j}+\epsilon^{r2}_{X_j})\Big]=$$

$$=\sum_{j=1}^{p}\Big\{E[\epsilon^{n2}_{X_j}]-2E[\epsilon^n_{X_j}\epsilon^r_{X_j}]+E[\epsilon^{r2}_{X_j}]\Big\}=$$

$$=\sum_{j=1}^{p}\Big\{(\frac{1}{n}-\frac{1}{N})C^2_{X_j}-2(\frac{1}{n}-\frac{1}{N})C^2_{X_j}+(\frac{1}{r}-\frac{1}{N})C^2_{X_j}\Big\}=$$

$$=\sum_{j=1}^{p}\Big\{C^2_{X_j}(\frac{1}{n}-\frac{1}{N}-\frac{2}{n}+\frac{2}{N}+\frac{1}{r}-\frac{1}{N})\Big\}=$$

$$=\sum_{j=1}^{p}\Big\{C^2_{X_j}(\frac{1}{r}-\frac{1}{n})\Big\}=$$

$$=(\frac{1}{r}-\frac{1}{n})\sum_{j=1}^{p}C^2_{X_j}$$

Desta forma novamente por meio de aproximação tem-se-que

$$\frac{\mu^2_Y}{p^2}\Big\{p^2E\Big[\epsilon^2_Y\Big]+2p\sum_{j=1}^{p}E[\epsilon_Y\epsilon^n_{X_j}-\epsilon_Y\epsilon^r_{X_j}]+E\Big[\sum_{j=1}^{p}(\epsilon^n_{X_j}-\epsilon^r_{X_j})\Big]^2\Big\}\leq$$

$$\leq\frac{\mu^2_Y}{p^2}\Big\{p^2E\Big[\epsilon^2_Y\Big]+2p\sum_{j=1}^{p}E[\epsilon_Y\epsilon^n_{X_j}-\epsilon_Y\epsilon^r_{X_j}]+E\Big[\sum_{j=1}^{p}(\epsilon^n_{X_j}-\epsilon^r_{X_j})^2\Big]\Big\}=$$

$$=\frac{\mu^2_Y}{p^2}\Big\{p^2E\Big[\epsilon^2_Y\Big]+2p\sum_{j=1}^{p}\Big(E[\epsilon_Y\epsilon^n_{X_j}]-E[\epsilon_Y\epsilon^r_{X_j}]\Big)+E\Big[\sum_{j=1}^{p}(\epsilon^n_{X_j}-\epsilon^r_{X_j})^2\Big]\Big\}=$$

$$=\frac{\mu^2_Y}{p^2}\Big\{p^2(\frac{1}{r}-\frac{1}{N})C^2_Y+2p\sum_{j=1}^{p}\Big((\frac{1}{n}-\frac{1}{N})\rho C_YC_{X_j}-(\frac{1}{r}-\frac{1}{N})\rho C_YC_{X_j}\Big)+(\frac{1}{r}-\frac{1}{n})\sum_{j=1}^{p}C^2_{X_j}\Big\}=$$

$$=\frac{\mu^2_Y}{p^2}\Big\{p^2(\frac{1}{r}-\frac{1}{N})C^2_Y+2p\sum_{j=1}^{p}\Big((\frac{1}{n}-\frac{1}{r})\rho C_YC_{X_j}\Big)+(\frac{1}{r}-\frac{1}{n})\sum_{j=1}^{p}C^2_{X_j}\Big\}=$$

$$=\frac{\mu^2_Y}{p^2}\Big\{p^2(\frac{1}{r}-\frac{1}{N})C^2_Y+2p(\frac{1}{n}-\frac{1}{r})\rho C_Y\sum_{j=1}^{p}C_{X_j}+(\frac{1}{r}-\frac{1}{n})\sum_{j=1}^{p}C^2_{X_j}\Big\}=$$

$$=\mu^2_Y\Big\{(\frac{1}{r}-\frac{1}{N})C^2_Y+2(\frac{1}{n}-\frac{1}{r})\rho C_Y\frac{1}{p}\sum_{j=1}^{p}C_{X_j}+(\frac{1}{r}-\frac{1}{n})\frac{1}{p^2}\sum_{j=1}^{p}C^2_{X_j}\Big\}=$$

$$=\mu^2_Y\Big\{(\frac{1}{r}-\frac{1}{N})C^2_Y+(\frac{1}{r}-\frac{1}{n})\Big(\frac{1}{p^2}\sum_{j=1}^{p}C^2_{X_j}-2\rho C_Y\frac{1}{p}\sum_{j=1}^{p}C_{X_j}\Big)\Big\}$$

3.3.3 Medidas de avaliação

Foi calculado a diferença entre os valores médios verdadeiros em comparação aos valores médios encontrados pelo método da imputação. Essa medida tem como finalidade a avaliação descritiva da predição quando a diferença se aproxima do valor nulo. Além disso, tem-se a proporção de médias sub e super estimados, por meio de valores positivos e negativos respectivamente.

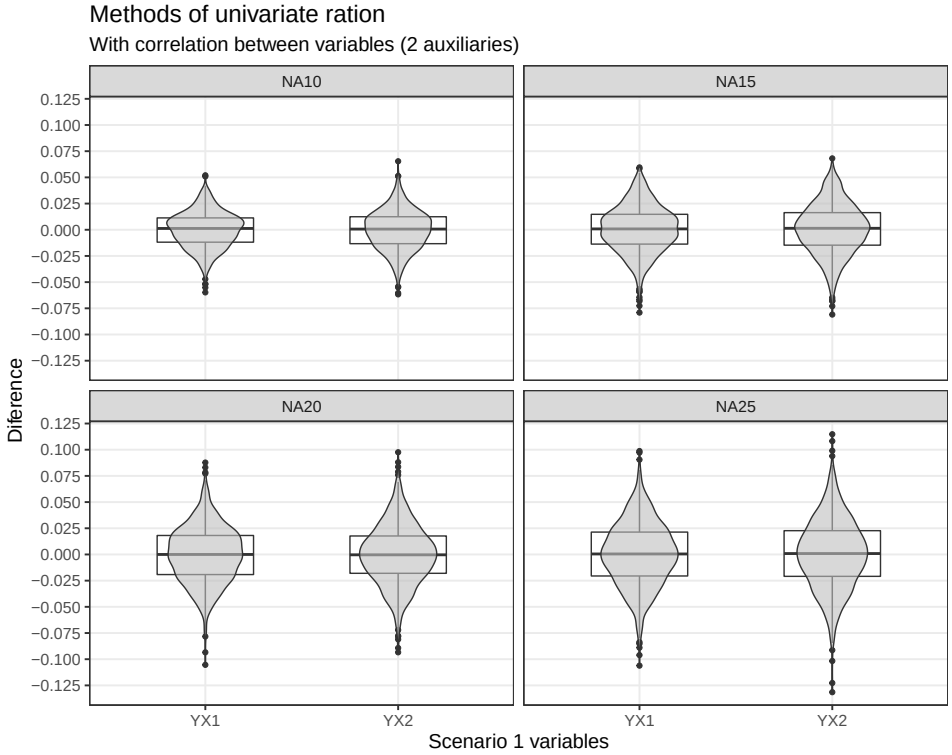
$$d_{p\%}^{[k]} = \mu_Y^{[k]} - \bar{Y}_{(Impu)p\%}^{[k]}, \quad \text{sendo } k = 1, 2, ..., 1000 \text{ (simulações)}$$

em que: $\mu_Y^{[k]}$ é a média populacional da variável de interesse da k-ésima simulação; $\bar{Y}_{(Impu)p\%}^{[k]}$ é a média calculada da variável de interesse com valores imputados com $p\%$ porcentagem de NA na k-ésima simulação.

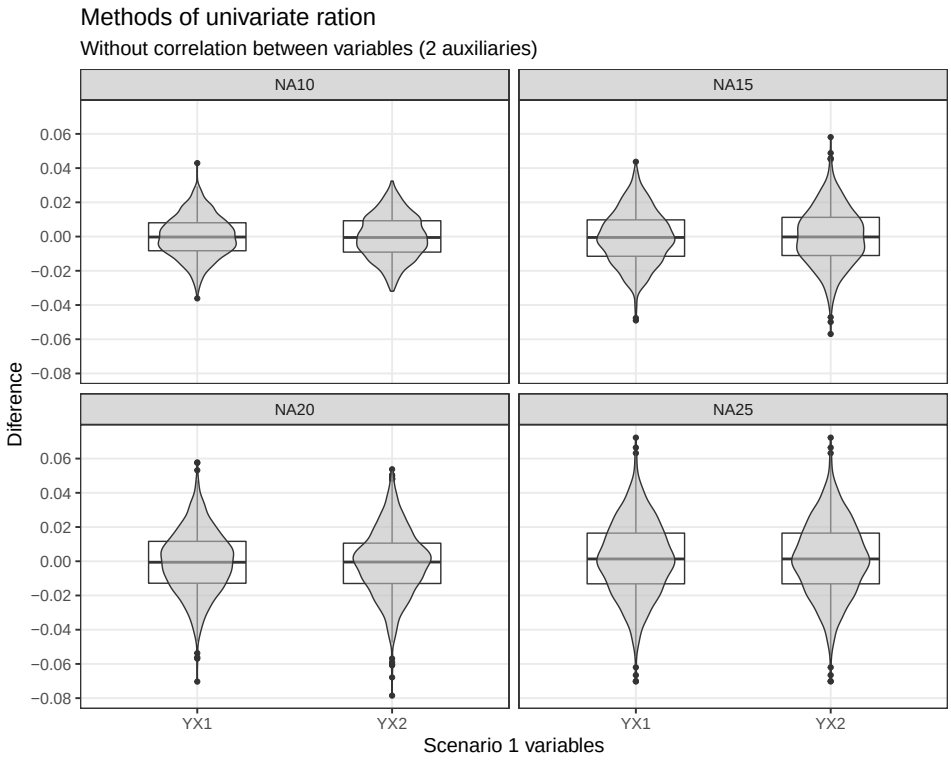
3.4 Resultados

Diana and Francesco Perri (2010) mostrou que os estimadores do tipo regressão, para a média populacional com ausência de dados, apresentaram maiores valores de eficiência percentual relativa (PRE) quando comparados aos outros métodos que utilizam componentes do estimador tipo razão e regressão conjuntamente. A quantidade menor de termos apresenta uma boa solução tanto em perspectiva teórica e prática.

Foram gerados boxplot para cada método de imputação e cada variável auxiliar que foi utilizadas, levando em consideração cada cenário causado pelos pesquisadores. Neste gráficos, tem-se em cada janela as proporções de valores ausentes (5%, 15%, 20% e 25%), e na abcissa a variável auxiliar que foi utilizada durante o processo de imputação por razão univariada (X's).

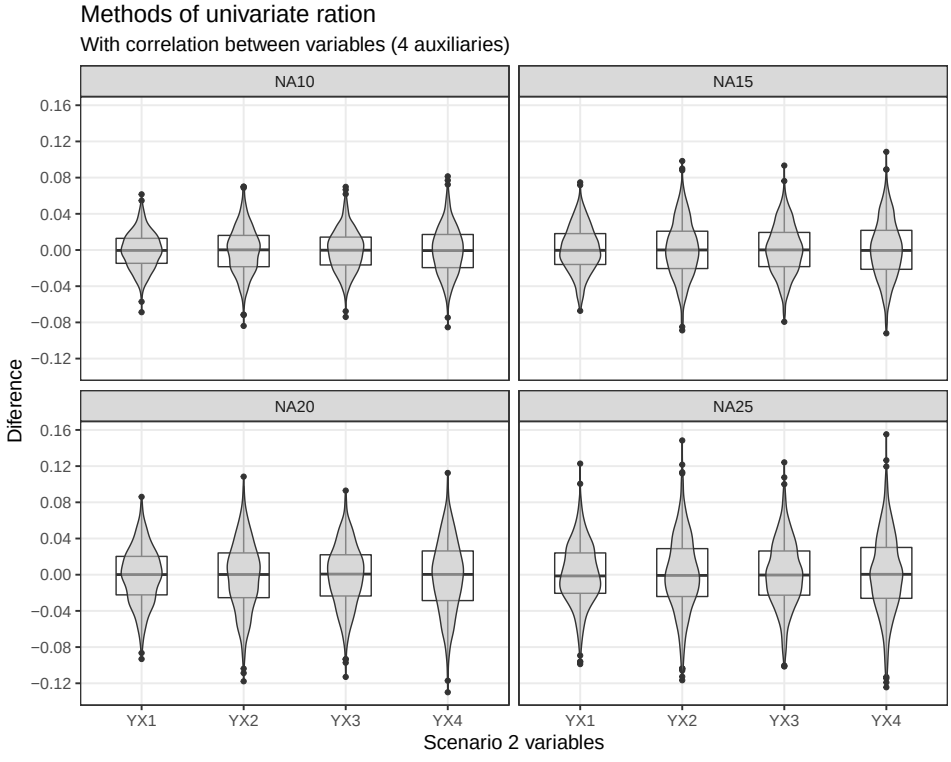


Observa-se para o método de imputação univariado para o cenário 1 com correlação que a diferença aumentou a medida de a proporção de valores ausente aumenta. Para a proporção de 10% a diferente ficou entre $(-0,075; 0,075)$. As 2 variáveis auxiliares tiveram dispersão similares. Na proporção de 25% houve uma maior variabilidade de diferença, para o método de imputação por razão univariada, quando aplicada-o a variável X2.



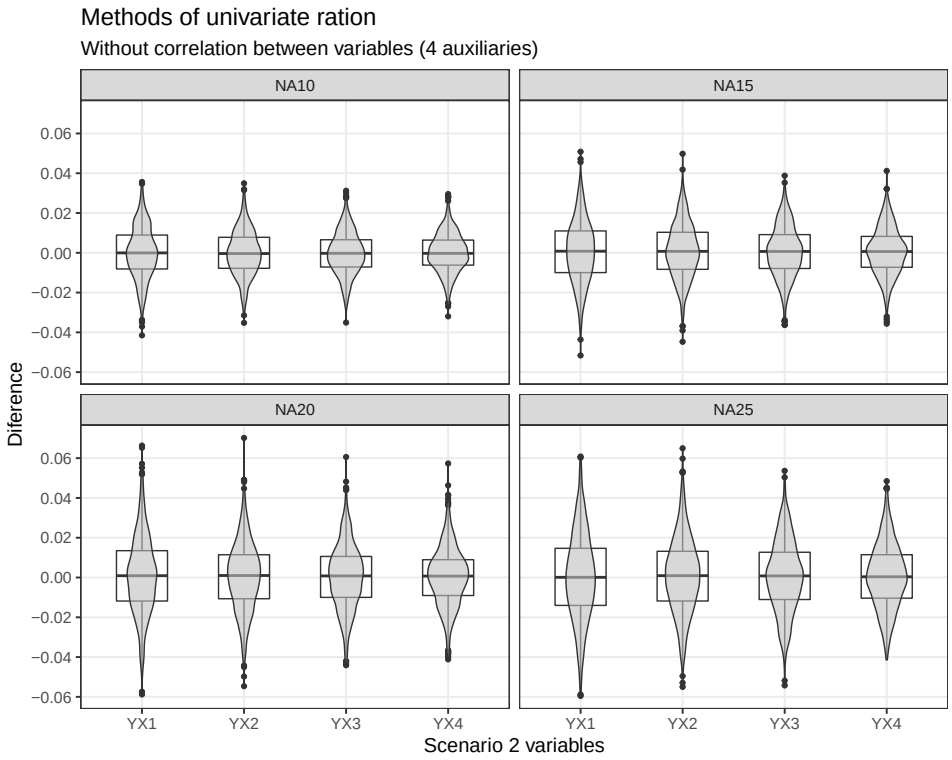
No método de imputação da razão univariado para o cenário 1 em que as variáveis não possuíam correlação, houve um aumento da diferença a medida que a proporção de valores ausentes aumentava. Para a proporção de 10%, os valores de diferença ficaram entre -0,04 a 0,05. E no maior nível de valores ausente (25%) a diferença esteve entre -0,08 a 0,07.

Entre os cenários de correlação houve maior variabilidade para os valores sem correlação.

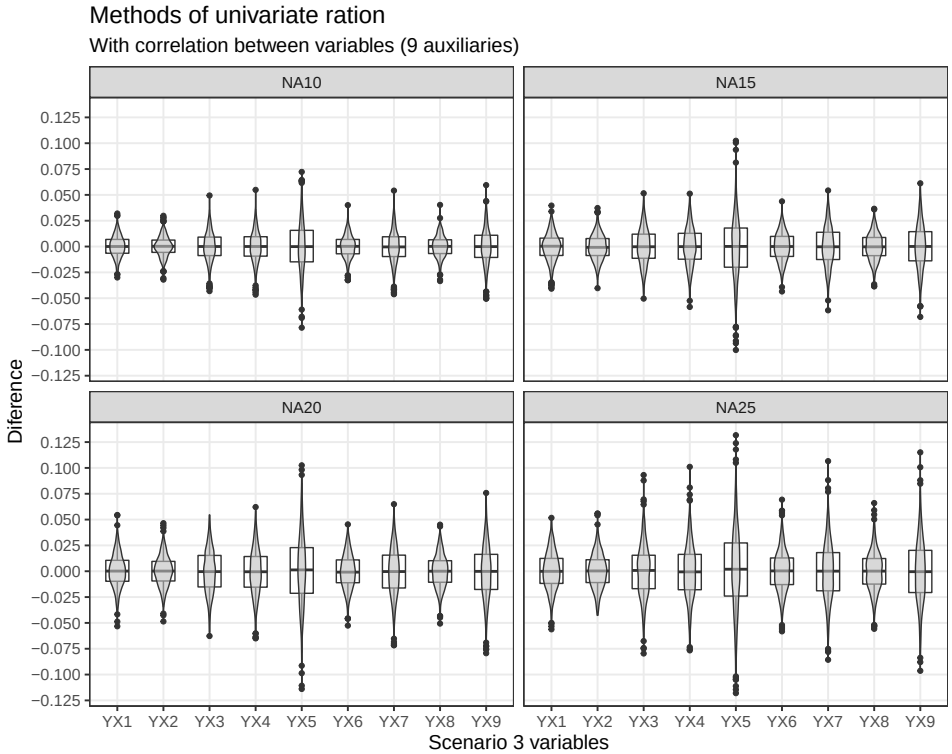


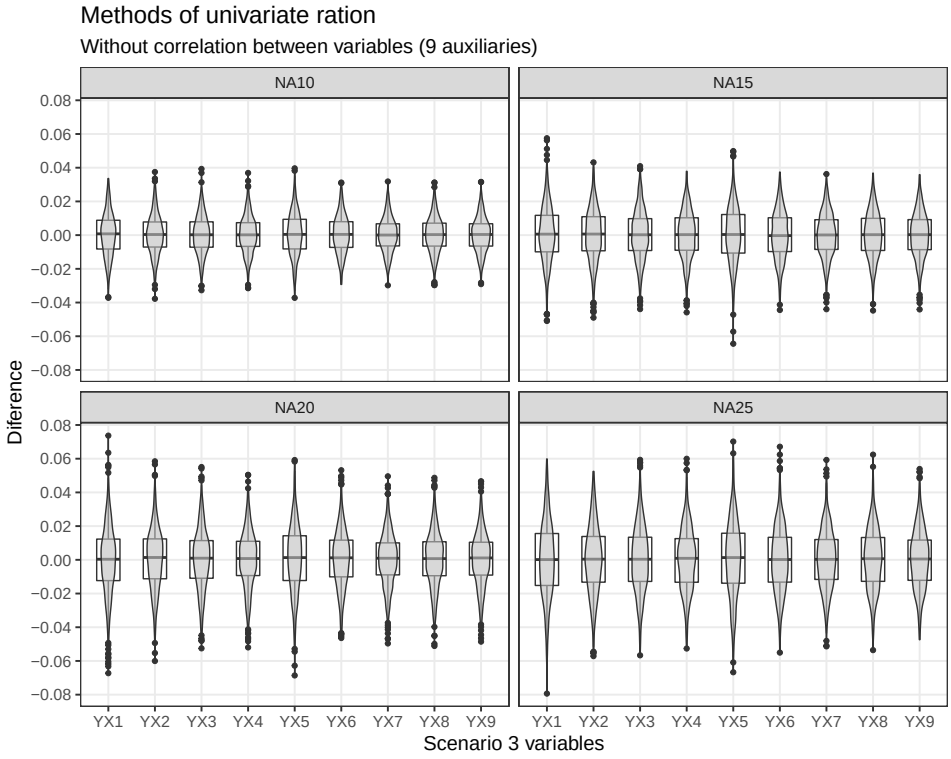
Para o cenário com 4 variáveis auxiliares altamente correlacionadas, pode-se observar que con-

forme a proporção de valores ausentes aumenta a diferença entre os valores verdadeiros com os imputados também aumenta.

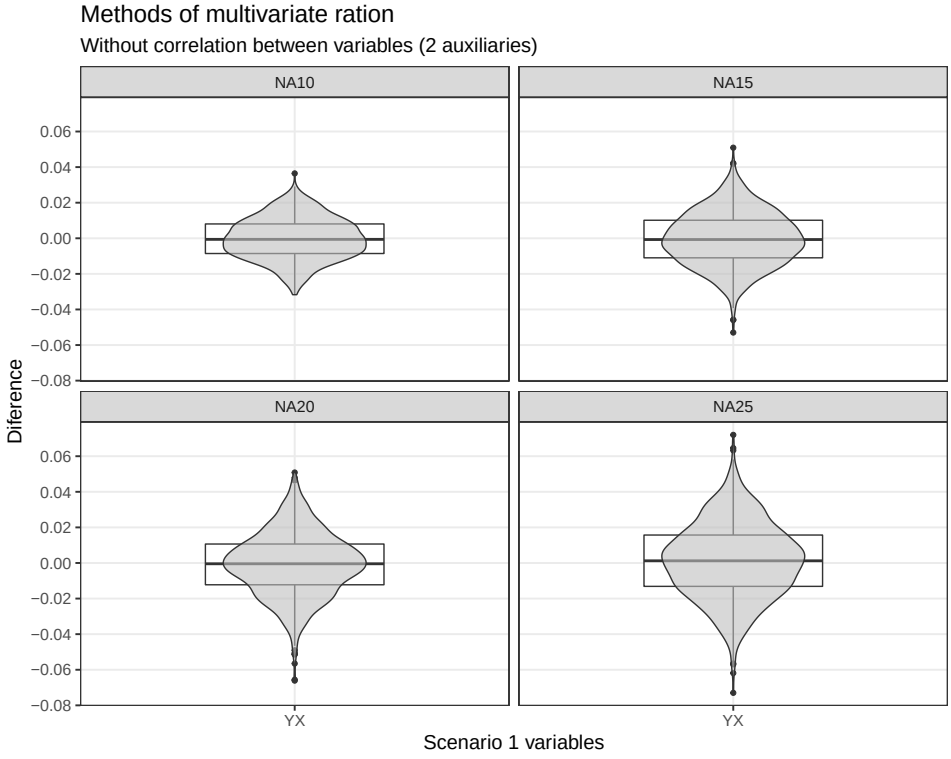


A imputação utilizando o método da razão univariada para o cenário com 4 variáveis auxiliares com correlação baixa, nota-se que os valores de diferença ficaram entre -0,06 a 0,06. Representando valores bem abaixo quando comparados aos mesmo cenário, porém com correlação alta. 7



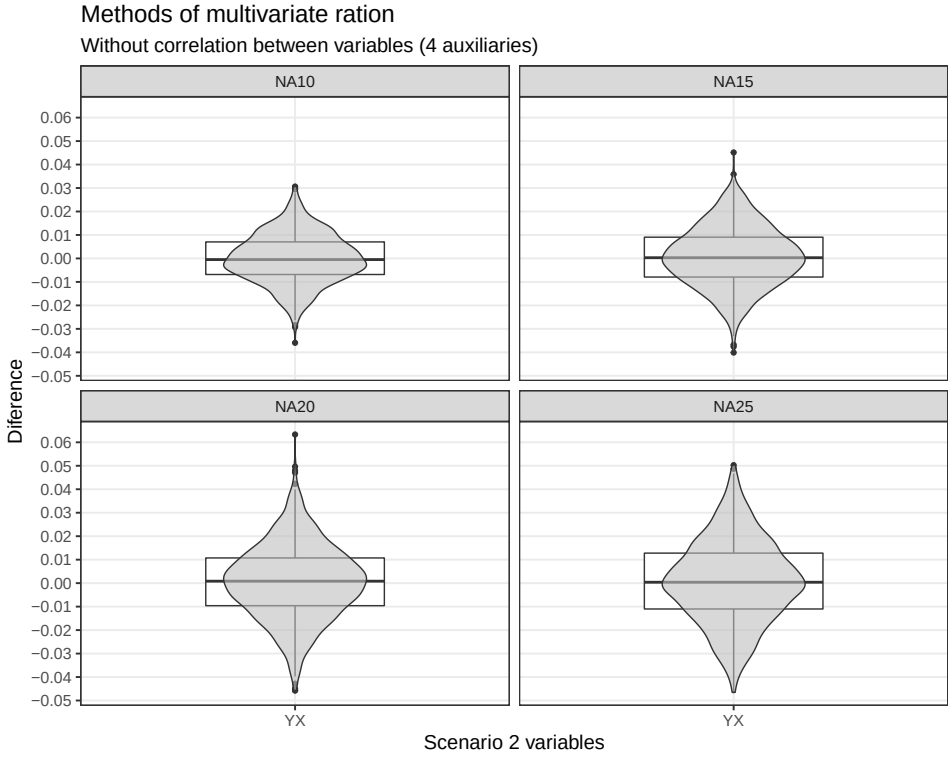


No 3º cenário com 9 variáveis auxiliares, a diferença quando para correlação alta ficaram entre -0,125 a 0,125; enquanto para correlação baixa foram de -0,08 a 0,08. Esses resultados para os casos de imputação univariada pelo metodo da razão, pode ser um indicativa que a influencia entres as variáveis (correlação) pode acarreta em valores imputados menos verossímeis.

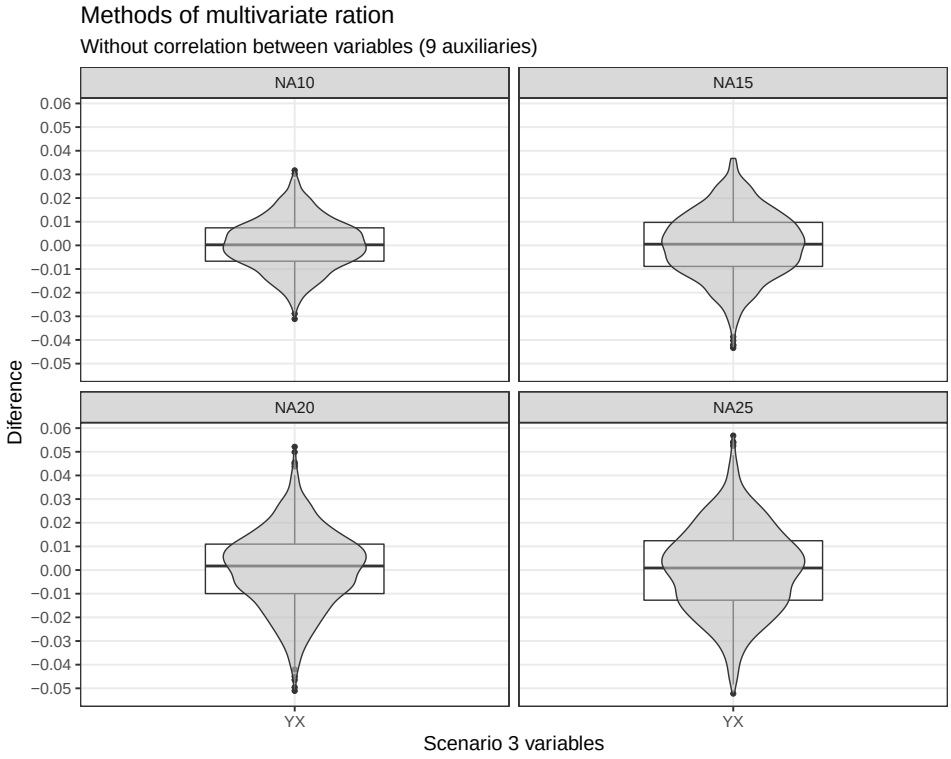


No método de imputação da razão multivariada, a medida que a proporção de valores ausentes da simulação aumentou as medidas diferença apresentaram maios amplitude. Em 10% os valores ficaram

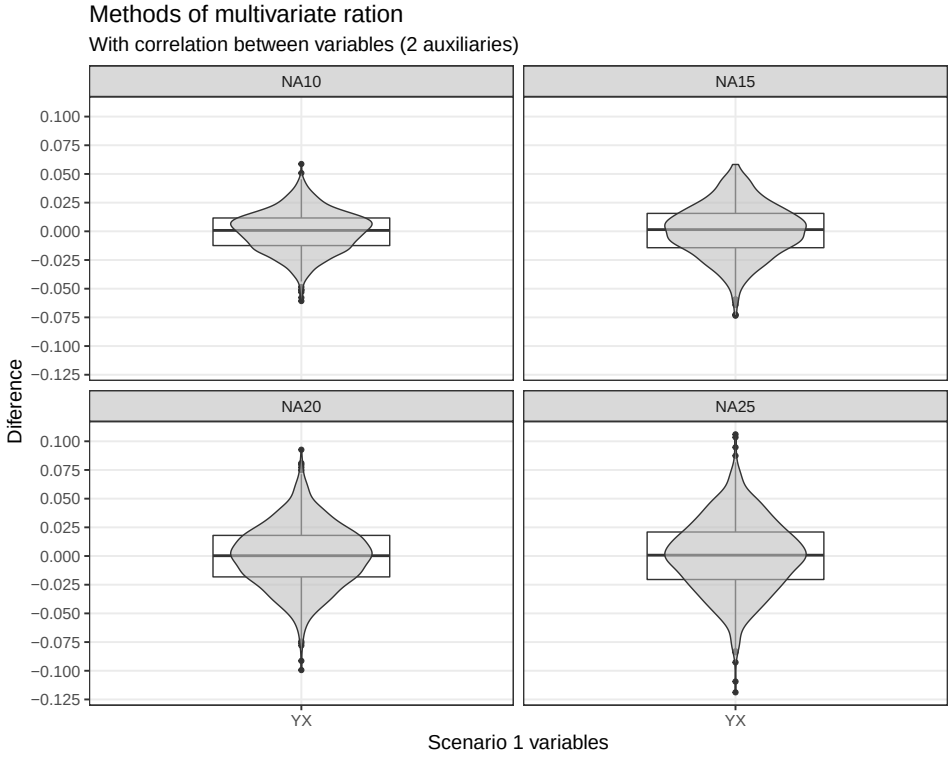
em torno de -0,04 a 0,04; para 15% entre -0,05 a 0,05; para 20% entre -0,07 a 0,05; e para 25% entre -0,08 a 0,07. Observa-se que, conforme a proporções aumentam, apresenta-se uma tendência de maior valores negativos; como descrito o cálculo da diferença, essa frequência maior corrobora que os valores de imputação foram maiores do que a média populacional.



No cenário 2, o método de imputação da razão multivariada apresenta resultados menores para as proporções de 10 % e 15%; e levemente maiores para as proporções de 20 % e 25%. Houve, para a proporção de 20% um valor discrepante que fica acima de 0,06.

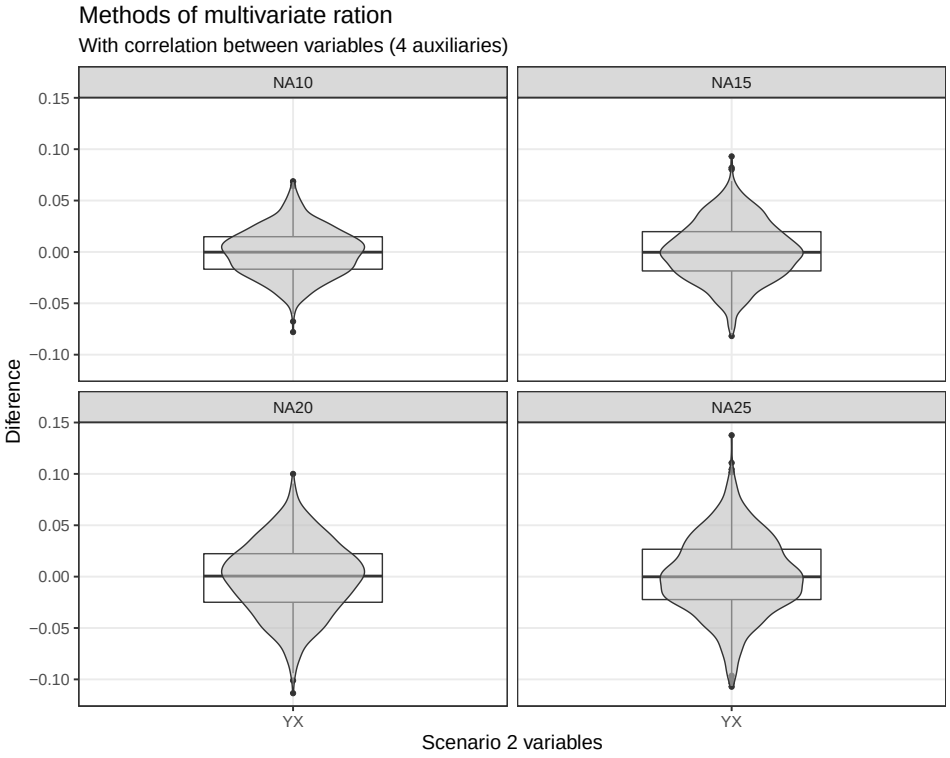


No cenário 3, houve um crescimento de amplitude de diferença conforme o aumento de proporções de valores ausentes. Em 10% ficaram entre -0,04 a 0,04; e a maior proporção entre -0,06 a 0,06. Todavia, mesmo assim, os resultados de diferença foram parecido para os demais cenários com menores quantidades de variáveis auxiliares.

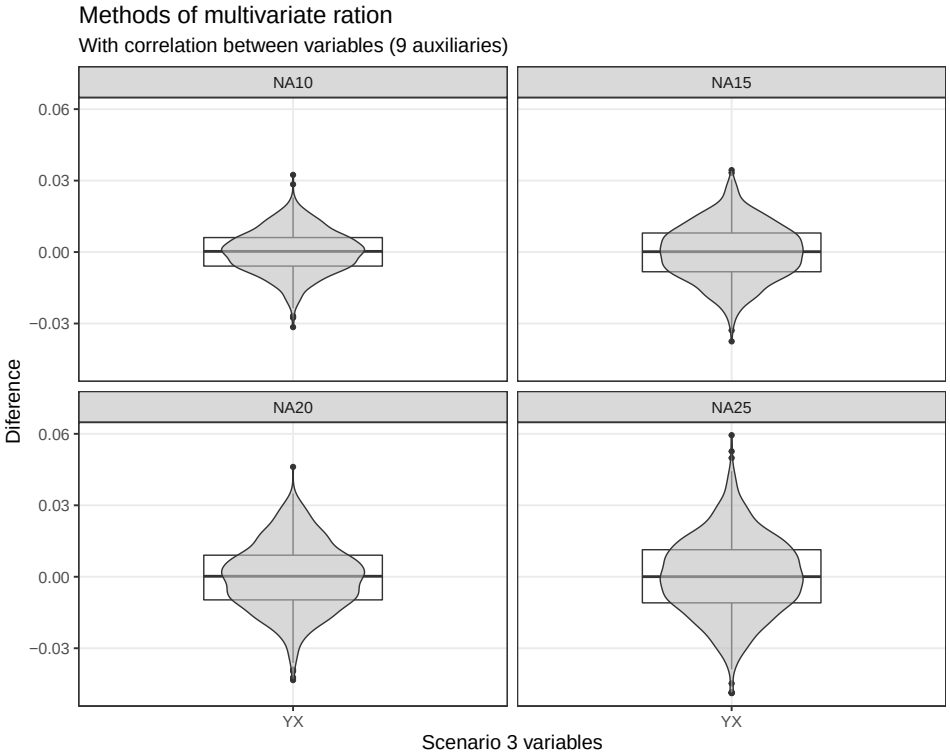


Os resultados das imputações do cenário 1, isto é 2 variáveis auxiliares com alta correlação houve um aumento na amplitude da diferença quando comparada-as quando não havia correlação entre

as variáveis. Para 10 % valores ausentes ficou-se em torno de -0,06 a 0,06; com 15 % entre -0,075 a 0,06; em 20 % -0,1 a 0,1; e por fim para 25 % entre -0,12 a 0,12. Essas diferenças foram semelhantes para a imputação pelo método da razão univariada, em que houve crescimento por conta do aumenta das proporções de valores ausentes.



Mesmo no cenário 2 com 4 variáveis auxiliares com correlação alta, os resultados foram similares ao do cenário 1.



Todavia no cenário 3, houve uma melhora nos valores de diferença entre os valores imputados em relação a média populacional. Para a proporção de 10 % os valores ficaram entre -0,04 a 0,04; similar para proporção de 15 %; em 20 % ficou entre -0,05 a 0,05; e finalmente para 25 % entre -0,06 a 0,06.

Cenários	Variável Auxiliar	Proporção de valores ausentes 10 %	15%	20%	25%
C1 Alta Correlação	X1	-0.00015	0.00055	-0.00021	0.00040
	X2	-0.00018	0.00067	-0.00022	0.00033
C1 Baixa Correlação	X1	-0.00019	-0.00046	-0.00050	0.00153
	X2	-0.00025	-0.00011	-0.00080	0.00153
C2 Alta Correlação	X1	-0.00056	0.00060	-0.00099	0.00068
	X2	-0.00035	0.00049	-0.00122	0.00100
	X3	-0.00052	0.00053	-0.00108	0.00094
	X4	-0.00059	0.00049	-0.00154	0.00107
C2 Baixa Correlação	X1	0.00013	0.00057	0.00112	0.00057
	X2	-0.00015	0.00105	0.00048	0.00101
	X3	-0.00020	0.00067	0.00033	0.00070
	X4	-0.00007	0.00065	0.00025	0.00060
C3 Alta Correlação	X1	0.00020	-0.00028	0.00045	0.00048
	X2	0.00020	-0.00050	0.00033	0.00017
	X3	0.00016	0.00005	-0.00056	0.00015
	X4	0.00005	0.00003	-0.00045	-0.00029
	X5	0.00051	-0.00092	0.00133	0.00117
	X6	0.00013	-0.00005	-0.00021	0.00004
	X7	-0.00009	0.00010	-0.00051	-0.00017
	X8	0.00005	-0.00007	-0.00009	-0.00004
	X9	-0.00001	0.00019	-0.00058	-0.00024
C3 Baixa Correlação	X1	0.00038	0.00083	-0.00005	0.00006
	X2	0.00034	0.00019	0.00039	0.00042
	X3	0.00035	0.00035	0.00022	0.00034
	X4	0.00042	0.00027	0.00034	0.00021
	X5	0.00058	0.00024	0.00075	0.00057
	X6	0.00048	-0.00007	0.00044	0.00018
	X7	0.00021	0.00020	0.00025	0.00034
	X8	0.00035	0.00019	0.00028	0.00034
	X9	0.00034	0.00011	0.00033	0.00032

Tabela 3.1. Viés para imputação univariada

Na Tabela 3.1 foi calculada a medida de víes para o método de imputação por razão univariado, aplicando para os cenários de alta e baixa correlação e utilizando cada variável auxiliar separadamente (coluna 1 e 2), para cada proporção de valores ausentes simulados (10%, 15%, 20% e 25%). Com 2 variáveis auxiliares (C1), para 10% houve uma similaridade entre os vieses, com 15% utilizando a variável X_1 o valores em valor absoluto foram próximo (0,00055 e 0,00046), porém para X_2 o cenário com alta correlação apresentou maior viés (0,00067), quando comparado-o ao com baixa correlação (0,00011). Na proporções de 20% e 25% houve um maior viés para os casos com baixa correlação. Para C2 houve um maior viés para quase todos os cenários com alta correlação, exceto para as característica da imputação realizada utilizando a variável auxiliar X_2 e com 15% (Alta Correlação - 0,00049 e Baixa Correlação - 0,00105).

Cenários	Variável Auxiliar	Proporção de valores ausentes 10 %	15%	20%	25%
C1 Alta Correlação	X1	0.00029	0.00049	0.00071	0.00092
	X2	0.00035	0.00058	0.00080	0.00110
C1 Baixa Correlação	X1	0.00014	0.00023	0.00033	0.00047
	X2	0.00016	0.00027	0.00036	0.00047
C2 Alta Correlação	X1	0.00039	0.00061	0.00094	0.00106
	X2	0.00062	0.00092	0.00138	0.00162
	X3	0.00048	0.00073	0.00111	0.00129
	X4	0.00066	0.00100	0.00153	0.00179
C2 Baixa Correlação	X1	0.00017	0.00024	0.00014	0.00047
	X2	0.00013	0.00019	0.00029	0.00035
	X3	0.00011	0.00016	0.00026	0.00031
	X4	0.00009	0.00014	0.00021	0.00026
C3 Alta Correlação	X1	0.00010	0.00017	0.00024	0.00031
	X2	0.00009	0.00014	0.00020	0.00027
	X3	0.00018	0.00029	0.00042	0.00059
	X4	0.00020	0.00032	0.00042	0.00063
	X5	0.00054	0.00081	0.00106	0.00152
	X6	0.00011	0.00018	0.00026	0.00036
	X7	0.00023	0.00036	0.00051	0.00072
	X8	0.00010	0.00016	0.00023	0.00033
	X9	0.00028	0.00044	0.00062	0.00088
C3 Baixa Correlação	X1	0.00016	0.00027	0.00039	0.00050
	X2	0.00013	0.00023	0.00031	0.00040
	X3	0.00012	0.00021	0.00030	0.00037
	X4	0.00012	0.00019	0.00027	0.00034
	X5	0.00016	0.00027	0.00039	0.00050
	X6	0.00013	0.00020	0.00028	0.00036
	X7	0.00010	0.00016	0.00023	0.00029
	X8	0.00011	0.00018	0.00025	0.00032
	X9	0.00010	0.00017	0.00024	0.00030

Tabela 3.2. MSE para imputação univariada

A medida de MSE frequentemente é utilizada para verificar a precisão do estimador, principalmente quando o estimador é viesado a fim de substituindo o cálculo da variância Lohr (2021). Para o C1 e C2, os maiores valores de MSE eram para as variáveis auxiliares que possuíam alta correlação corroborando na formula analística da Equação 3.4 por apresentar o ρ. Observou-se que para 9 variáveis auxiliares as medidas de MSE ficaram próximo. Mas independente dos cenários, o MSI cresceu a medida que a proporção de valores ausentes aumentaram.

Cenários		Proporção de valores ausentes 10 %	15%	20%	25%
C1	Viés	-0.00016	0.00061	-0.00022	0.00036
Alta Correlação	MSE	0.00031	0.00052	0.00073	0.00098
C1	Viés	-0.00022	-0.00029	-0.00065	0.00127
Baixa Correlação	MSE	0.00014	0.00023	0.00032	0.00045
C2	Viés	-0.00051	0.00053	-0.00121	0.00092
Alta Correlação	MSE	0.00052	0.00079	0.00121	0.00139
C2	Viés	-0.00007	0.00073	0.00055	0.00072
Baixa Correlação	MSE	0.00011	0.00016	0.00025	0.00031
C3	Viés	0.00013	-0.00016	-0.00003	0.00014
Alta Correlação	MSE	0.00008	0.00013	0.00019	0.00027
C3	Viés	0.00038	0.00026	0.00033	0.00031
Baixa Correlação	MSE	0.00011	0.00019	0.00026	0.00033

Tabela 3.3. Viés e MSE para imputação multivariada

3.4.1 Aplicação da imputação nos dados Embrapa

Para o conjunto de dados da Embrapa, a imputação foi realizada a fim de calcular o valor médio do peso para as famílias de melão que não apresentaram nenhuma observação. Na Figura 3.2 tem-se a identificação das famílias no eixo horizontal, a mensuração do valor média para a variável de peso no eixo vertical, e para os 3 ambientes um gráfico de pontos. As família que não apresentaram valores ausentes são representadas pela linha vertical vermelha e quanto os pontos são os valores médios, como exemplo a família 86 que no ambiente 1 não apresentou valores, porém no ambiente 2 e 3 possuem valores médios de aproximadamente de 1,4 kg e 0,6 kg, respectivamente.

Para a utilização dos métodos de imputação por razão e razão múltipla foi considerados para a implementação da análise os ambientes como as variáveis e famílias como observações, levando em conta estudo de modelos para interação genótipos $\times$ ambientes. Cada família que necessitou de imputação, levou-se em conta o ambiente ausente como sendo a variável de interesse, e os outros ambientes como as variáveis auxiliares, na imputação univariada obteve-se 2 valores imputados, em virtude de cada ambiente considerado.

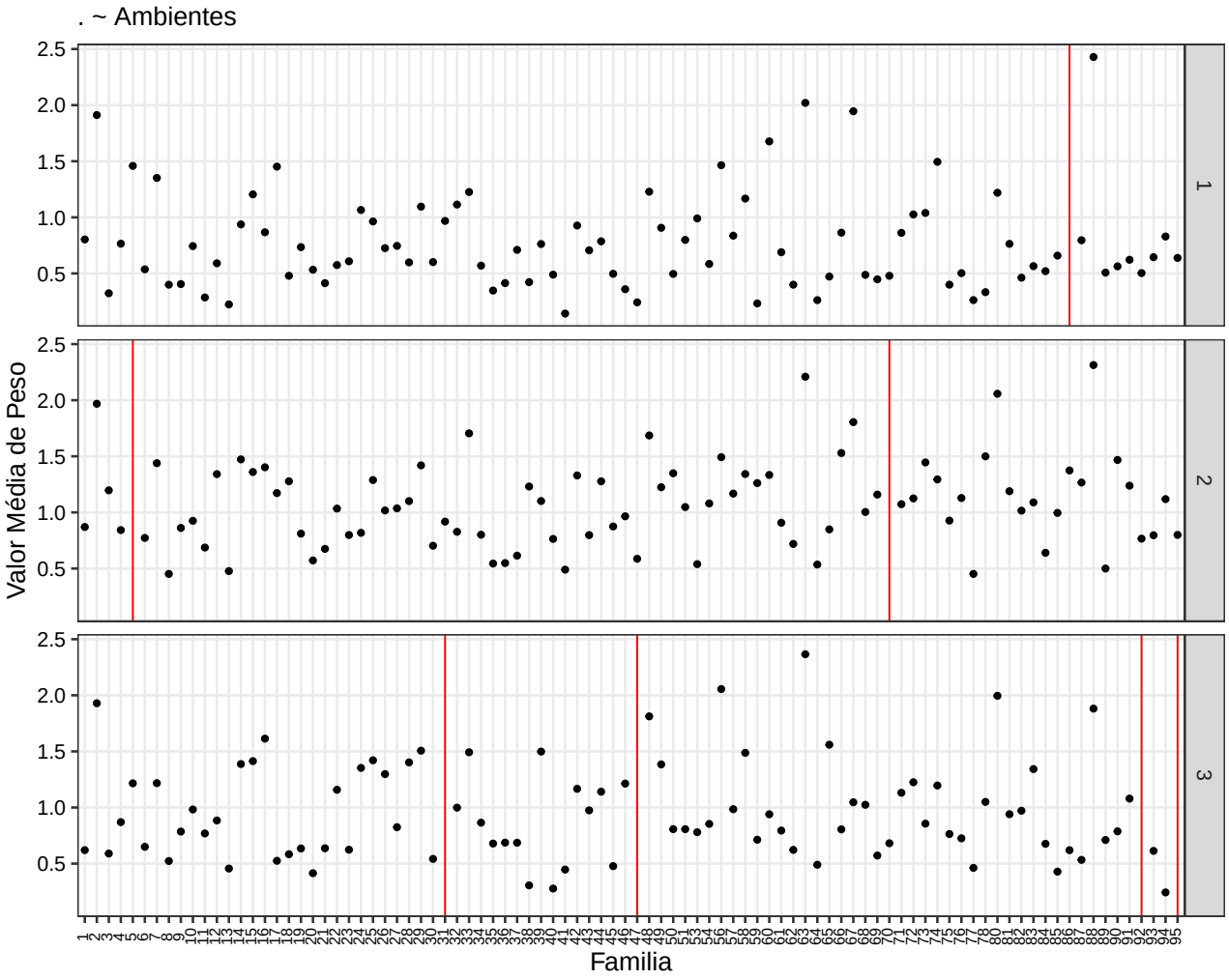


Figura 3.2. Gráfico de pontos dos valores médios para peso (eixo y), em relação as 92 famílias (eixo x), para os 3 ambientes (multi-gráficos) e com identificação de valores ausentes para as respectivas informações por meio de uma reta vertical

Na Figura 3.3 para a família 86, devido a falta de informação no ambiente 1 obteve-se 2 imputações, primeiro considerando o ambiente 2 como variável auxiliar, e em virtude dessa família apresentar um maior valor médio para o peso nesse ambiente, o valor foi próximo de 1 kg (ponto verde). Agora, considerando o ambiente 3 como de variável auxiliar, a imputação por razão do peso ficou próxima do valor 0,5 kg, devido ao peso da família no ambiente 3 ser menor.

Para o ambiente 2 as famílias 5 e 70 não tinha valores, e desta forma, utilizou-se as medidas dessas famílias no ambiente 1 e 3, identificadas com pontos de cor roxa e rosa, respectivamente (Figura 3.3). Vale notar que, para essa duas famílias, houve uma diferença entre seus resultados. Como a observação da família 5 apresentava maiores valores para o ambiente 1 e menores valores para o ambiente 3, gerando nas imputações uma diferença entres os valores imputados de 0,6 unidades de quilos. Todavia, para a família 70, devido a semelhança desses valores as imputações também resultaram em medidas próximas.

E finalmente, para o ambiente 3 foi onde houve mais valores ausentes, resultando assim em 4 imputações utilizando as informações do ambiente 1, caracterizadas pelos pontos em verde, e 4 imputações aplicando a variável auxiliar do ambiente 2 com pontos em amarelo (Figura 3.3).

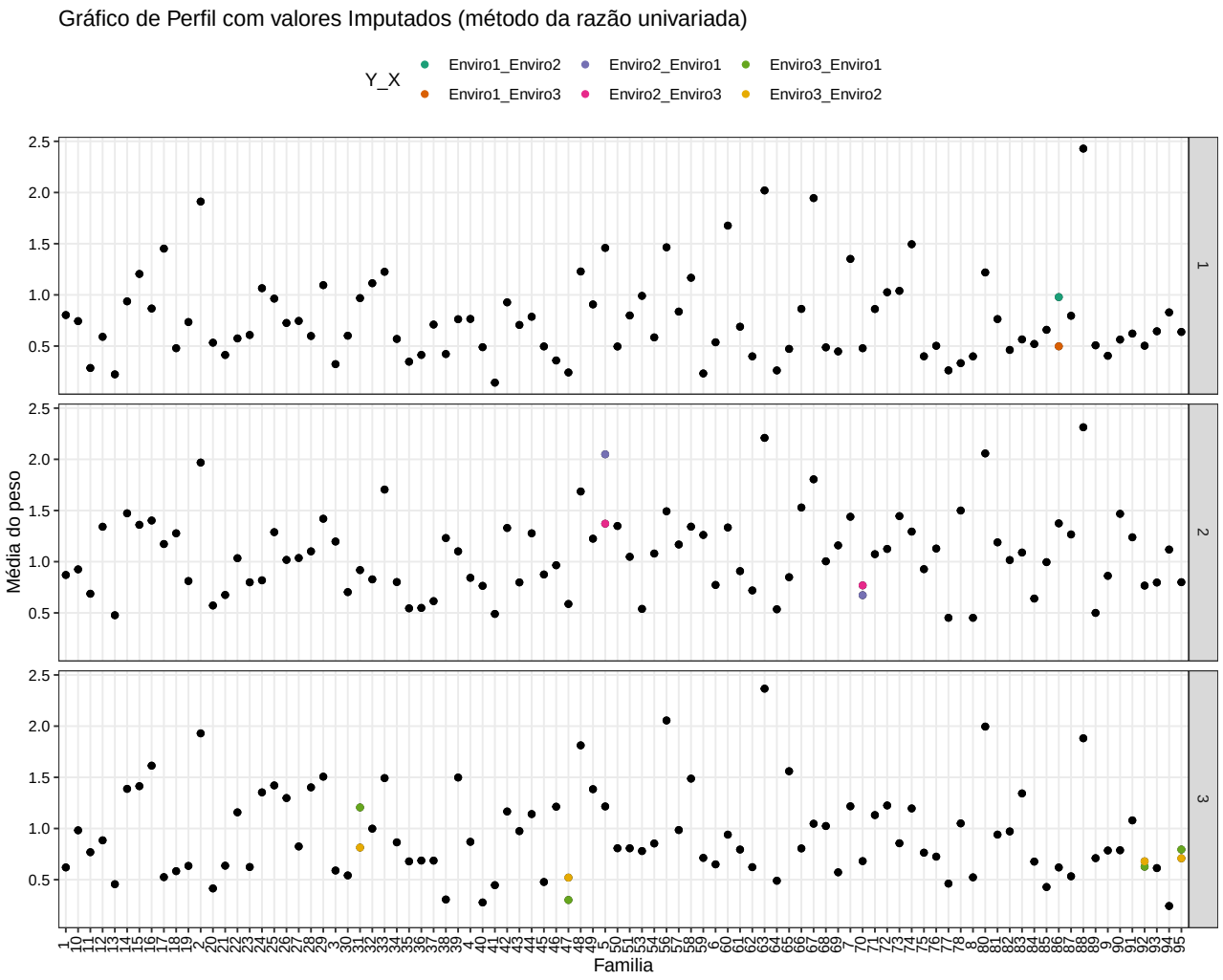


Figura 3.3. Gráfico de pontos dos valores médios para peso (eixo y), em relação as 92 famílias (eixo x), para os 3 ambientes (multi-gráficos) e com valores imputados pela técnica de razão, utilizando cada ambiente como variável auxiliar (cores dos pontos)

Para a imputação das famílias pelo método da razão múltipla considerou-se os demais ambientes com a família preenchida como auxiliares. Na Figura 3.4 essas imputações foram identificadas por meio de pontos vermelhos. Na família 86 do ambientes 1, o valor ficou próximo de 0,7 kg, para as famílias 5 e 70 ficaram com valores médios de 1,7 e 0,7, respectivamente. E finalmente, para o ambiente 3, as famílias 31 (1 kg), 47 (0,4 kg), 92 (0,6 kg) e 95 (0,7 kg).

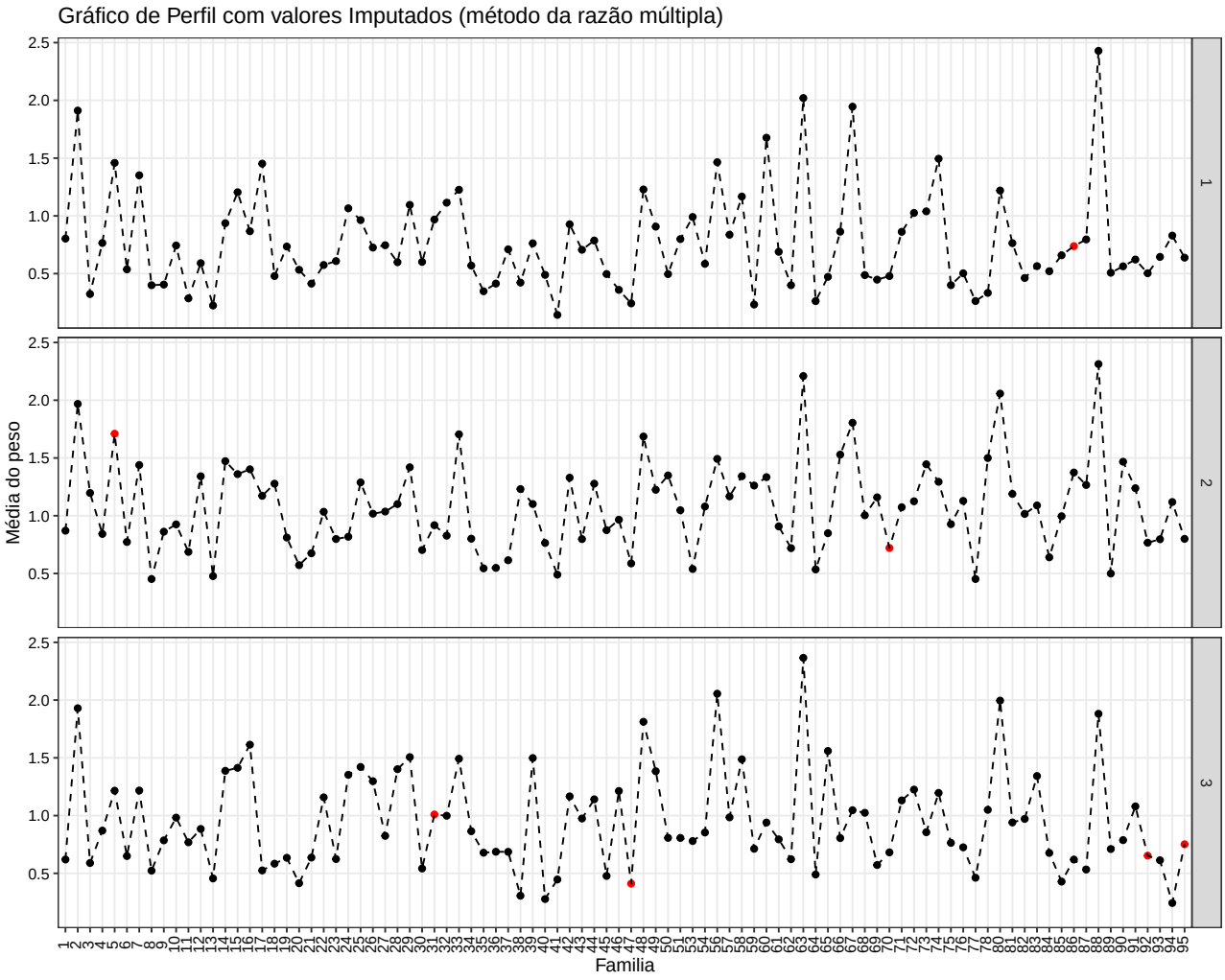


Figura 3.4. Gráfico de pontos dos valores médios para peso (eixo y), em relação as 92 famílias (eixo x), para os 3 ambientes (multi-gráficos) e com valores imputados pela técnica de razão múltipla (pontos vermelhos)

Na Tabela 3.4 calculou-se os valores médios do peso para cada ambiente: sem imputação, utilizando o método da imputação da razão univariada e da razão múltipla. Para o ambiente 1, os métodos de imputação por razão univariada utilizando o ambiente A2, A3 e pelo método da razão múltipla, apresentaram médias de 0,773, 0,768, 0,771 respectivamente, e superiores quando comparando com a média sem imputação 0,762. Para o ambiente 2, houve uma semelhança entre as médias calculadas sem imputação, com imputação univariada utilizando o ambiente A1, e com razão múltipla. E no ambiente A3, o método sem imputação apresentou uma média superior aos médias calculadas com os dados imputados.

Tabela 3.4. Valores da média amostral para cada ambiente considerando 3 casos: sem imputação retirando os valores ausentes; razão univariada utilizando a imputação com o respectivo ambiente (Variável Auxiliar) para o cenário de imputação; e razão múltipla considerando todos ambientes (as 2 variáveis auxiliares)

Métodos de Imputação		Médias do peso (kg)	
Sem Imputação			
	A1	0,762	
Ambientes	A2	1,080	
	A3	0,969	
Razão Univariada		Variável Auxiliar	
		A2	A3
	A1	0,773	0,768
Ambientes		A1	A3
	A2	1,084	1,077
		A1	A2
	A3	0,951	0,949
Razão Multipla			
	A1	0,771	
Ambientes	A2	1,080	
	A3	0,950	

3.5 Conclusões

Com esses resultados podemos verificar os seguintes aspectos: como o método de razão univariada se comportou para cada variável auxiliar gerada na simulação, principalmente devido a informação de correlação; como os 2 métodos se comportaram para cenários iguais, todavia são avaliações sem validação entre eles, pois o método da razão multivariada acaba tendo mais fontes de informação (variáveis auxiliares) na estrutura de cálculo de imputação.

3.5.1 Programação

3.5.1.1 Cenários de Simulação

```
# Informações iniciais ------------------------------------------------

n_simu = 1000 #quantidade de simulações
N_popu = 5000 #tamanho populacional

# Simulação 1 - Alta correlação ---------------------------------------

# S1.1 com 3 variáveis ------------------------------------------------

#quantidade de variáveis auxiliares
p_var = 3
#média populacional
mean_vector1 <- c(20, 40, 60)

#processo iterativo para gerar dados com alta correlação
mean_corr = 0
while (mean_corr <= 0.9) {
        #matriz de covariância populacional
        matrix_cov1 <-
        clusterGeneration::genPositiveDefMat(dim = p_var,
        covMethod = 'eigen',
        eigenvalue = c(100, 5, 1)) %>% #mudar algumento para diminuir a correlação
        magrittr::extract2('Sigma')

        #cálculo da matriz de correlação pela matriz de covariância
        Sd_Matrix <- diag(matrix_cov1) %>% sqrt() %>% diag()
        Corr_Matrix <- solve(Sd_Matrix) %*% matrix_cov1 %*% solve(Sd_Matrix)

        #cálculo médio das correlações
        mean_corr <- Corr_Matrix[upper.tri(Corr_Matrix)] %>% mean()
}

datalist_simu1 = list()
for(i in 1:n_simu){

        #simulação dos dados
        datalist_simu1[[i]] <- MASS::mvrnorm(n = N_popu,
        mu = mean_vector1,
        Sigma = matrix_cov1)
}

# S1.2 com 5 variáveis ------------------------------------------------

#quantidade de variáveis auxiliares
```

```
p_var = 5
#média populacional
mean_vector2 <- c(20, 40, 60, 80, 100)

#processo iterativo para gerar dados com alta correlação
mean_corr = 0
while (mean_corr <= 0.8) {
        #matriz de covariância populacional
        matrix_cov2 <-
        clusterGeneration::genPositiveDefMat(dim = p_var,
        covMethod = 'eigen',
        eigenvalue = c(100, 5, 5, 2, 1)) %>%
        magrittr::extract2('Sigma')

        #cálculo da matriz de correlação pela matriz de covariância
        Sd_Matrix <- diag(matrix_cov2) %>% sqrt() %>% diag()
        Corr_Matrix <- solve(Sd_Matrix) %*% matrix_cov2 %*% solve(Sd_Matrix)

        #cálculo médio das correlações
        mean_corr <- Corr_Matrix[upper.tri(Corr_Matrix)] %>% mean()
}

datalist_simu2 = list()
for(i in 1:n_simu){

        #simulação dos dados
        datalist_simu2[[i]] <- MASS::mvrnorm(n = N_popu,
        mu = mean_vector2,
        Sigma = matrix_cov2)
}

# S1.3 com 10 variáveis ----------------------------------------------------

#quantidade de variáveis auxiliares
p_var = 10
#média populacional
mean_vector3 <- c(20, 40, 60, 80, 100,
30, 50, 70, 90, 110)

#processo iterativo para gerar dados com alta correlação
mean_corr = 0
while (mean_corr <= 0.75) {
        #matriz de covariância populacional
        matrix_cov3 <-
        clusterGeneration::genPositiveDefMat(dim = p_var,
        covMethod = 'eigen',
        eigenvalue = c(500, 5, 5, 5, 5,
```

```
            1, 1, 1, 1, 1)) %>%
            magrittr::extract2('Sigma')

            #cálculo da matriz de correlação pela matriz de covariância
            Sd_Matrix <- diag(matrix_cov3) %>% sqrt() %>% diag()
            Corr_Matrix <- solve(Sd_Matrix) %*% matrix_cov3 %*% solve(Sd_Matrix)

            #cálculo médio das correlações
            mean_corr <- Corr_Matrix[upper.tri(Corr_Matrix)] %>% mean()
    }

    datalist_simu3 = list()
    for(i in 1:n_simu){

            #simulação dos dados
            datalist_simu3[[i]] <- MASS::mvrnorm(n = N_popu,
            mu = mean_vector3,
            Sigma = matrix_cov3)
    }
```

3.5.1.2 Cenários de Simulação - Com proporção de valores ausentes

```
# NA scenario ------------------------------------------------------------

# for scenario 1 ---------------------------------------------------------

# with 10%
dataNA_simu110 <- datalist_simu1
for(i in 1:1000){
        #NA individuals random
        id_na <- sample.int(n = 5000, size = 5000*.1)

        #selecting individuals in 'Y' for NA
        dataNA_simu110[[i]][id_na,'Y'] <- NA
}

# with 15%
dataNA_simu115 <- datalist_simu1
for(i in 1:1000){
        #NA individuals random
        id_na <- sample.int(n = 5000, size = 5000*.15)

        #selecting individuals in 'Y' for NA
        dataNA_simu115[[i]][id_na,'Y'] <- NA
}

# with 20%
```

```
dataNA_simu120 <- datalist_simu1
for(i in 1:1000){
        #NA individuals random
        id_na <- sample.int(n = 5000, size = 5000*.2)

        #selecting individuals in 'Y' for NA
        dataNA_simu120[[i]][id_na,'Y'] <- NA
}

# with 25%
dataNA_simu125 <- datalist_simu1
for(i in 1:1000){
        #NA individuals random
        id_na <- sample.int(n = 5000, size = 5000*.25)

        #selecting individuals in 'Y' for NA
        dataNA_simu125[[i]][id_na,'Y'] <- NA
}
```

3.5.1.3 Funções para cálculo de imputação pelo método da razão univariada e multivariada

```
ratio_imputation <- function(x, NA_names){
        # x = dataset
        # NA_names = name of variable contains NA

        #names variables without NA
        Var_names = setdiff(names(x), NA_names)

        #NA index
        id_NA <-
        x[NA_names] %>%
        is.na() %>%
        which()

        #partition data without NA
        data_R <- x[-id_NA,]

        #partition data without NA
        data_RC <- x[id_NA,]

        # imputation method (ration)
        mean_R <- apply(data_R, 2, mean)
        ration_R <- mean_R[NA_names]/mean_R[Var_names]

        # apply imputation
        Xm <- as.matrix(data_RC[Var_names])
        bs <- as.numeric(ration_R)
```

```
        # product between ratio estimation and value of auxiliar variable
        pv <- bs*Xm

        #new colunm for NA identificator
        x$var_NA <- 0
        x$var_NA[id_NA] <- '1'

        #subtitute
        x[id_NA, NA_names] <- pv

        return(x)
}

multiration_imputation <- function(x, NA_names){
        # x = dataset
        # NA_names = name of variable contains NA

        Var_names = setdiff(names(x), NA_names)

        #NA index
        id_NA <-
        x[NA_names] %>%
        is.na() %>%
        which()

        #partition data without NA
        data_R <- x[-id_NA,]

        #partition data without NA
        data_RC <- x[id_NA,]

        # imputation method (multivariate ration)
        mean_RC <- apply(data_R, 2, mean)
        ration_RC <- mean_RC[NA_names]/mean_RC[Var_names]

        p_scale = mean_RC[Var_names] %>% length()

        # apply imputation
        Xm <- as.matrix(data_RC[Var_names]) #matrix with auxiliar variable
        bv <- as.vector(ration_RC[Var_names]) #vector with ration estimation

        #sum of product between ration estimation and value of auxiliar variable
        sp = (1/p_scale)*(Xm%*%bv)

        #new colunm for NA identificator
        x$var_NA <- 0
```

```
        x$var_NA[id_NA] <- '1'

        #subtistuindo imputação
        x[id_NA, 1] <- sp

        return(x)
}
```

3.6 Referências

Arnab, R., 2017 Survey Sampling: Theory and Applications. Academic Press, first edition.

Beunckens, C., C. Sotto, G. Molenberghs, and G. Verbeke, 2009 A multifaceted sensitivity analysis of the Slovenian public opinion survey data. JOURNAL OF THE ROYAL STATISTICAL SOCIETY SERIES C-APPLIED STATISTICS **58**: 171–196.

Bodin, J.-L., 2020 A view on 50 years of life of the ISI: With a focus on ISI relations with official statistics. Statistical Journal of the IAOS **36**: 303–308.

Bolfarine, H. and W. O. Bussab, 2004 Elementos de amostragem. Editora Blucher.

Brand, M., 2002 Incremental singular value decomposition of uncertain data with missing values. In COMPUTER VISON - ECCV 2002, PT 1, edited by A. Heyden, G. Sparr, M. Nielsen, and P. Johansen, volume 2350 of Lecture Notes in Computer Science, pp. 707–720, IT Univ Copenhagen; Univ Copenhagen; Lund Univ.

Cavalcanti, P. P. and C. T. d. S. Dias, 2021 Archetypal analysis as an imputation method and multivariate data augmentation .

Cebrián, A. A. and M. R. García, 1997 Variance Estimation Using Auxiliary Information: An Almost Unbiased Multivariate Ratio Estimator. Metrika **45**: 171–178.

Cochran, W. G., 1977 Sampling techniques. John Wiley & Sons, third edition.

Deming, W. E., 1990 Sample design in business research, volume 23. John Wiley & Sons.

Departamento de Astronomia Instituto de Astronomia, G. e. C. A., 2020 Início das estações do ano (2005–2020). Accessed: 2022-03-18.

Diana, G. and P. Francesco Perri, 2010 Improved estimators of the population mean for missing data. Communications in Statistics—Theory and Methods **39**: 3245–3251.

Ellington, E. H., G. Bastille-Rousseau, C. Austin, K. N. Landolt, B. A. Pond, E. E. Rees, N. Robar, and D. L. Murray, 2015 Using multiple imputation to estimate missing data in meta-regression. METHODS IN ECOLOGY AND EVOLUTION **6**: 153–163.

Epifanio, I., M. V. Ibanez, and A. Simo, 2020 Archetypal Analysis With Missing Data: See All Samples by Looking at a Few Based on Extreme Profiles. AMERICAN STATISTICIAN **74**: 169–183.

Gasparetto, S. C., S. M. D. S. Piedade, L. R. Angelocci, and V. A. Ozaki, 2021 COMPARAÇÃO ENTRE MÉTODOS DE IMPUTAÇÃO DE DADOS EM DIFERENTES INTENSIDADES AMOSTRAIS NA SÉRIE DE PRECIPITAÇÃO PLUVIAL DA ESALQ. Revista Brasileira de Climatologia **29**: 464–489.

Goodman, L. A. and H. O. Hartley, 1958 The Precision of Unbiased Ratio-Type Estimators. Journal of the American Statistical Association **53**: 491–508.

Hanif, M., Z. Ahmed, and M. Ahmad, 2009 Generalized multivariate ratio estimators using multi-auxiliary variables for multi-phase sampling. Pakistan Journal of Statistics **25**: 615–629.

He, Y., R. Yucel, and T. E. Raghunathan, 2011 A functional multiple imputation approach to incomplete longitudinal data. STATISTICS IN MEDICINE **30**: 1137–1156.

Howe, L. D., K. Tilling, A. Matijasevich, E. S. Petherick, A. C. Santos, L. Fairley, J. Wright, I. S. Santos, A. J. D. Barros, R. M. Martin, M. S. Kramer, N. Bogdanovich, L. Matush, H. Barros, and D. A. Lawlor, 2016 Linear spline multilevel models for summarising childhood growth trajectories: A guide to their application using examples from five birth cohorts. STATISTICAL METHODS IN MEDICAL RESEARCH **25**: 1854–1874.

Hulley, S. B., S. R. Cummings, W. S. Browner, D. G. Grady, and T. B. Newman, 2013 Designing clinical research. Lippincott Williams & Wilkins, fourth edi edition.

Jessen, R. J., 1943 Statistical investigation of a sample survey for obtaining farm facts. Iowa State University.

Johnson, R. A. and D. W. Wichern, 2007 Applied Multivariate Statistical Analysis. Prentice Hall, 6th edition.

Kenward, M. G. and J. Carpenter, 2007 Multiple imputation: current perspectives. Statistical Methods in Medical Research **16**: 199–218.

Kim, J. and M.-J. Park, 2019 Multiple imputation and synthetic data. KOREAN JOURNAL OF APPLIED STATISTICS **32**: 83–97.

Kruskal, W. and F. Mosteller, 1980 Representative Sampling, IV: The History of the Concept in Statistics, 1895-1939. International Statistical Review / Revue Internationale de Statistique **48**: 169.

Lohr, S. L., 2019 Sampling: Design and Analysis. Advanced (Cengage Learning), Cengage Learning.

Lohr, S. L., 2021 Sampling: design and analysis. CRC press.

Maitra, R., V. Melnykov, and S. N. Lahiri, 2012 Bootstrapping for Significance of Compact Clusters in Multidimensional Datasets. Journal of the American Statistical Association **107**: 378–392.

Melnykov, V., W.-C. Chen, and R. Maitra, 2012 MixSim: An R Package for Simulating Data to Study Performance of Clustering Algorithms. Journal of Statistical Software **51**: 1–25.

Myers, W. R., 2000 Handling Missing Data in Clinical Trials: An Overview. Drug information journal : DIJ / Drug Information Association **34**: 525–533.

Nath, K. and B. K. Sungh, 2018 Population Mean Estimation Using Ratio-cum Product Compromised-method of Imputation in Two-phase Sampling Scheme. Asian J. Math. Stat **11**: 27–39.

Neyman, J., 1992 pp. 123–150 in On the Two Different Aspects of the Representative Method: the Method of Stratified Sampling and the Method of Purposive Selection, edited by Kotz, S. and N. L. Johnson, Springer New York.

Olkin, I., 1958 Multivariate Ratio Estimation for Finite Populations. Biometrika **45**: 154.

Olufadi, Y. and C. Kadilar, 2014 A study on the chain ratio-type estimator of finite population variance. Journal of Probability and Statistics **2014**.

Patterson, H., 1950 Sampling on successive occasions with partial replacement of units. Journal of the Royal Statistical Society. Series B (Methodological) **12**: 241–255.

Peel, D., R. S. Waples, G. M. Macbeth, C. Do, and J. R. Ovenden, 2013 Accounting for missing data in the estimation of contemporary genetic effective population size (N-e). MOLECULAR ECOLOGY RESOURCES **13**: 243–253.

Qiu, W. and H. Joe., 2020 clusterGeneration: Random Cluster Generation (with Specified Degree of Separation). R package version 1.3.7.

Rao, J., 2005 Interplay Between Sample Survey Theory and Practice: An Appraisal. Survey Methodology **31**.

Rao, J. N. K. and D. R. Bellhouse, 1990 History and development of the theoretical foundations of survey based estimation and analysis. Survey Methodology **16**: 3–29.

Rao, J. N. K. and R. R. Sitter, 1995 Variance estimation under two-phase sampling with application to imputation for missing data. Biometrika **82**: 453–460.

Rubin, D. B., H. S. Stern, and V. Vehovar, 1995 Handling "Don't Know" Survey Responses: The Case of the Slovenian Plebiscite. Journal of the American Statistical Association **90**: 822–828.

Sano, N., 2020 Synthetic Data by Principal Component Analysis. In 20TH IEEE INTERNATIONAL CONFERENCE ON DATA MINING WORKSHOPS (ICDMW 2020), edited by G. DiFatta, V. Sheng, A. Cuzzocrea, C. Zaniolo, and X. Wu, International Conference on Data Mining Workshops, pp. 101–105, IEEE; IEEE Comp Soc; Univ Calabria; Mininglamp Technol.

Scheaffer, R. L., W. III Mendenhall, R. L. Ott, and K. Gerow, 2012 Elementary Survey Sampling. Cengage Learning, 7th edition.

Schneeberger, H. and K. Fleischer, 1993 The Multivariate Ratio Estimation: A Simulation Study. Jahrbücher für Nationalökonomie und Statistik **211**: 524–538.

Shukla, G. K., 1966 An Alternative Multivariate Ratio Estimate for Finite Population. Calcutta Statistical Association Bulletin **15**: 127–134.

Singh, G. N., S. Maurya, M. Khetan, and C. Kadilar, 2016 Some imputation methods for missing data in sample surveys. Hacettepe Journal of Mathematics and Statistics **45**: 1865–1880.

Singh, G. N. and S. Suman, 2019 Estimation of population mean using imputation methods for missing data under two-phase sampling design. Journal of Statistical Theory and Practice **13**: 1–24.

Singh, H. P., A. Gupta, and R. Tailor, 2021 Estimation of population mean using a difference-type exponential imputation method. Journal of Statistical Theory and Practice **15**: 1–43.

Singh, H. P., S. Kumar, and S. Bhougal, 2011 Multivariate ratio estimation in presence of non-response in successive sampling. Journal of Statistical Theory and Practice **5**: 591–611.

Singh, S., 2009 A new method of imputation in survey sampling. Statistics **43**: 499–511.

Tariq, M. U., M. N. Qureshi, and M. Hanif, 2021 Variance Estimators in the Presence of Measurement Errors Using Auxiliary Information **19**: 606–616.

Wissler, A., K. E. Blevins, and J. E. Buikstra, 2022 Missing data in bioarchaeology I: A review of the literature. AMERICAN JOURNAL OF BIOLOGICAL ANTHROPOLOGY **179**: 339–348.

Xu, Y. and R. Goodacre, 2018 On Splitting Training and Validation Set: A Comparative Study of Cross-Validation, Bootstrap and Systematic Sampling for Estimating the Generalization Performance of Supervised Learning. Journal of Analysis and Testing **2**: 249–262.

Yates, F. et al., 1949 Sampling methods for censuses and surveys. Sampling methods for censuses and surveys. .

Yuan, Y., 2011 Multiple Imputation Using SAS Software. Journal of Statistical Software **45**: 1–25.

Zhang, P., 2003 Multiple imputation: Theory and method. INTERNATIONAL STATISTICAL REVIEW **71**: 581–592.