ROSIELLE DA COSTA FARIAS

UM NOVO MÉTODO PARA ALOCAÇÃO DE UNIDADES EM SUBAMOSTRAS REPRESENTATIVAS BASEADO EM COVARIÁVEIS DISCRETAS

Dissertação apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Estatística Aplicada e Biometria, para obtenção do título de Magister Scientiae.

VIÇOSA
MINAS GERAIS - BRASIL
2018

Ficha catalográfica preparada pela Biblioteca Central da Universidade Federal de Viçosa - Câmpus Viçosa

T
F224n
2018

Farias, Rosielle, 1990-
Um novo método para alocação de unidades em subamostras representativas baseado em covariáveis discretas / Rosielle Farias. – Viçosa, MG, 2018.
xii, 55f. : il. (algumas color.) ; 29 cm.

Orientador: Fernando Luiz Pereira de Oliveira.
Dissertação (mestrado) - Universidade Federal de Viçosa.
Referências bibliográficas: f.52-55.

1. Amostragem (Estatística). I. Universidade Federal de Viçosa. Departamento de Estatística. Programa de Pós-Graduação em Estatística Aplicada e Biometria. II. Título.

CDD 22. ed. 519.52

ROSIELLE DA COSTAS FARIAS

UM NOVO MÉTODO PARA ALOCAÇÃO DE UNIDADES EM SUBAMOSTRAS REPRESENTATIVAS BASEADO EM COVARIÁVEIS DISCRETAS

Dissertação apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Estatística Aplicada e Biometria, para obtenção do título de *Magister Scientiae*.

APROVADA: 23 de março de 2018.

Graziela Dutra Rocha Gouvea

Adriano Marçal Pimenta

Fernando Luiz Pereira de Oliveira
(Orientador)

Dedicatória

Dedico este trabalho à todos os meus familiares e principalmente ao meu namorado Helgem.

Agradecimentos

Em primeiro lugar agradeço a toda minha família, por todo apoio recebido durante toda minha formação, em especial ao meu namorado Helgem, que foi o meu professor particular em tempo integral, e também, por ter suportado meu mal humor e *stress* durante todo o meu mestrado.

Ao Prof. Fernando, por todo apoio e conhecimento compartilhado comigo.

Á Érica por sempre ter sido uma ótima amiga e companheira nós caminhos da volta pra casa na UFV.

Á Juliana por ser uma amiga tão boa que além de ter cedido um lugarzinho no coração me ofereceu um cantinho em sua casa.

Brincar é condição fundamental para ser sério.

Aristóteles

Lista de Figuras

Lista de Tabelas

Resumo

FARIAS, Rosielle da Costa, M.Sc., Universidade Federal de Viçosa, março de 2018. **Um novo método para alocação de unidades em subamostras representativas baseado em covariáveis discretas**. Orientador: Fernando Luiz Pereira de Oliveira. Coorientador: Ivair Ramos Silva.

Em estudos experimentais, ensaios clínicos por exemplo, nos quais se deseja verificar a eficácia de alguma intervenção, é fundamental a presença de diferentes grupos que sofrerão ou não as intervenções para que futuras comparações possam ser realizadas. Para garantir que tais comparações sejam válidas, é necessário que os grupos apresentem características o mais semelhantes possíveis entre si e a amostra original. Este trabalho apresenta uma nova metodologia de divisão de uma amostra original em k subamostras representativas em relação à amostra original, com base em covariáveis que definem as características da amostra. Os resultados obtidos demonstram que a metodologia proposta apresenta resultados bastante satisfatórios, principalmente se comparados com a técnica tradicional de seleção de subamostras, o sorteio aleatório (amostragem aleatória simples). As subamostras delineadas pelo método apresentam altíssimo grau de similaridade com a amostra original, o que possibilitará estudos experimentais com viés de seleção bastante reduzido e resultados confiáveis.

Abstract

FARIAS, Rosielle da Costa, M.Sc., Universidade Federal de Viçosa, March, 2018. **A new method for unit allocation in representative subsamples based on discrete covariables**. Adviser: Fernando Luiz Pereira de Oliveira. Co-adviser: Ivair Ramos Silva.

In experimental studies, like clinic trials, where one wants to verify the eficacy of some intervention, the presence of different groups that will suffer the or not the interventions, so one can make future comparisons. To waranty that the comparisons will be valid, it's necessary that the groups shows the most similar characteristics among them and the original sample. This study brings a new methodology of division of an original sample in k representative sub-samples about the original sample, based in the covariates that defines the original sample characteristics. The results demonstrate that the proposed methodology shows very satisfatory results, mainly if compared to the traditional method, the random sampling. The sub-samples defined by the new method shows a high similarity with the original sample, which will made possible experimental studies with low selection bias and reliable results.

Índice

Referências Bibliográficas p. 52

1 Introdução

1.1 Motivação

O *Ensaio Clínico* é uma importante ferramenta para a avaliação de intervenções para saúde ou em qualquer área onde se deseja verificar a eficacia de algum tratamento. Entende-se por ensaio clínico um estudo planejado cuja a finalidade primária seria a avaliação da eficácia e da segurança de intervenções sanitárias médicas ou cirúrgicas.

Um pressuposto indispensável na realização de um ensaio clínico é a aleatorização de seus tratamentos entre os participantes do estudo. O ensaio é dito um *ensaio clínico aleatorizado* quando atende a este pressuposto, ou seja, quando os indivíduos elegíveis ao estudo são alocados nos diferentes grupos de tratamento de maneira casual.

Sempre que se propõe uma metodologia que culminará em algum teste estatístico, deseja-se realizar uma amostragem de modo que os grupos de tratamentos sejam heterogêneos dentre si, ou seja, possuam a maior representatividade possível entre seus elementos amostrais. Ademais busca-se homogêneos entre os grupos, de maneira que eles possuam características amostrais semelhantes. Além destas relações, espera-se que todos os grupos apresentem as mesmas características da população, de modo a tornar cada parcela do experimento representativa.

Usualmente, os grupos de tratamento em ensaios clínicos e demais experimentos são definidos por meio de amostragem aleatória simples (sorteio), o que nem

sempre garante as características desejadas nos grupos. Caso hajam características que possam influenciar o problema em estudo e estas não estejam distribuídas de forma homogênea entre os grupos, pode ocorrer um viés causado pela seleção.

Exemplo 1.1 *Suponhamos que se deseja testar a eficácia de um novo medicamento para a hipertensão arterial. Para tal, foi proposto um ensaio clínico aleatorizado composto por dois grupos: um grupo que receberá o medicamento de referência e um segundo grupo que receberá a nova droga. Para que a eficácia da droga seja comprovada a mesma deve funcionar de forma similar para em todos os indivíduos. Entretanto, suponha que existam indivíduos acima do peso, sedentários, sob estresse e que apresentam uma alimentação inadequada. Uma má distribuição destes indivíduos dentre cada grupo poderá influenciar diretamente o resultado do ensaio, pois sabidamente tais características impactam diretamente sobre a pressão arterial. Diante disso, idealmente, devemos obter grupos heterogêneos em relação à esta característica para que se possa garantir a fiabilidade do estudo.*

Exemplo 1.2 *Determinada empresa deseja testar se um novo método de produção é mais rápido e eficiente do que o antigo. Entretanto, diversas variáveis relacionadas aos operadores podem influenciar na execução dos procedimentos, tais como: turno, idade, sexo, tempo de serviço na empresa e tempo de experiência no processo antigo, dentre outros. Para isso deve-se selecionar grupos o mais homogêneos possíveis que contemplem todos os perfis de profissionais, pois caso hajam profissionais mais experientes e capacitados, em um determinado grupo, por exemplo, a dinâmica do experimento poderá ser comprometida.*

Em ambos os casos supracitados, uma seleção de grupos desbalanceada em relação às características populacionais poderá acarretar problemas de viés de seleção,

o que possivelmente acarretaria queda na qualidade dos resultados e possivelmente influenciaria nas conclusões dos respectivos testes.

Em problemas desta natureza, se faz necessária uma metologia capaz de produzir grupos amostrais que representem de maneira fidedigna a população, não só em termos da variável explicativa mas também em termos das covariáveis que influenciam no resultado do estudo. Considerando-se experimentos com grande número de covariáveis, fica inviável realizar a seleção dos grupos com base em técnicas de planejamento de experimentos, tais como blocagem, e a simples aleatorização poderá não ser capaz de garantir a homogeneidade necessária entre os grupos. Com base neste tipo de problema, bastante frequente na literatura, com destaque para pesquisas na área de saúde, foi desenvolvido um método computacional que seleciona os elementos dos grupos de acordo com o comportamento da população previamente estudada pelo pesquisador.

1.2 Objetivos

O objetivo geral deste trabalho é apresentar e validar um método de aleatorização que respeite a estrutura populacional na definição dos grupos em estudos experimentais. Os objetivos específicos deste trabalho são os seguintes:

- Descrever um método de aleatorização de grupos em ensaios clínicos;
- Validar metodologia apresentada baseada na análise de um conjunto de dados reais;

1.3 Organização

No Capítulo 2, será realizada uma revisão bibliográfica sobre metodologias de pesquisa na área da saúde e técnicas de aleatorização.

No Capítulo 3, o método proposto será discutido, bem como o conjunto de dados utilizado na validação da metodologia proposta e as técnicas empregadas na análise da eficiência do algorítimo proposto.

O Capítulo 4, será destinado à análise e validação do método de aleatorização, sendo seu desempenho comparado com a metodologia clássica de aleatorização via sorteio.

Para concluir, no Capítulo 5, serão apresentadas as conclusões e observações finais, além das propostas de continuidade do estudo.

2 Metodologias de Pesquisa na Área da Saúde

2.1 Introdução

Segundo Fontelles (2012), a Bioestatística é definida como a aplicação de métodos estatísticos em pesquisas relacionadas às áreas das ciências da vida e da saúde. A rigor, o que diferencia a Bioestatística da Estatística convencional é a utilização de conceitos próprios e metodologias consagradas em estudos relacionados às áreas supracitadas. O jargão diferenciado da Bioestatística induz os estudiosos da área à intuição de que se trata de áreas distintas, entretanto, os modelos, equações, teoremas e demais entes matemáticos utilizados nesta área de pesquisa são exatamente os mesmos utilizados nas demais aplicações de estatística, e portanto, o mesmo rigor matemático-científico deve ser exigido para que se garanta a validade e confiabilidade dos resultados de pesquisas na área da saúde.

A garantia da validade de uma análise estatística está vinculada a uma série de premissas que devem ser verificadas pelo pesquisador, afim de garantir que seus resultados apresentam as características inerentes de um estudo científico e corroborem com as metodologias estatísticas utilizadas no decorrer do estudo. De acordo com Montgomery e Runger (2010) a condução adequada de um experimento estatístico deve compreender as seguintes etapas:

- descrição detalhada do problema,

- identificação de fatores preponderantes que o afetam,
- proposição de um modelo estatístico adequado,
- coleta e processamento de dados,
- aplicação e validação do modelo proposto e
- a tomada de decisão com base nos resultados obtidos.

Naturalmente, estes elementos também devem estar presentes em estudos de problemas na área da saúde. A distinção da Bioestatística para os demais estudos estatísticos está principalmente na proposição de modelos adequados e sua coleta de dados, pois as aplicações estatísticas desta área, tanto por motivos técnicos quanto pode motivos éticos, são peculiares o suficiente para exigir metodologias únicas de seleção e coleta de dados.

O objetivo deste capítulo é realizar uma revisão bibliográfica acerca dos principais modelos e métodos estatísticos em pesquisas na área das ciências da saúde e suas aplicações, além de explorar um dos requisitos fundamentais à boa condução de um experimento estatístico, a aleatorização.

2.2 Estudos Observacionais

Os estudos observacionais, conforme o próprio nome indica, são aqueles nos quais o investigador analisa os eventos mas não intervém nos acontecimentos relacionados aos grupos observados. Eles podem ser descritivos, que se limitam a descrever a ocorrência de eventos de interesse e os analíticos, nos quais os pesquisadores abordam de forma mais detalhada as relações existentes entre os eventos de interesse e as possíveis variáveis relacionadas.

A técnicas estatísticas são amplamente utilizadas nos estudos observacionais analíticos. Nesta seção serão descritas as principais metodologias aplicadas nos estudos observacionais analíticos.

2.2.1 Estudos Transversais

De acordo com Bonita, Beaglehole e Kjellström (2008) os estudos transversais medem a prevalência da doença e, por essa razão, são frequentemente chamados de estudos de prevalência. Em um estudo transversal, as medidas de exposição e efeito (doença) são realizadas ao mesmo tempo. Por esse razão, não é fácil avaliar as associações encontradas nesses estudos. A questão-chave nesse tipo de delineamento é saber se a exposição precede ou é consequência do efeito. Se os dados coletados representam a exposição antes da ocorrência de qualquer efeito, a análise pode ser feita de modo semelhante à utilizada nos estudos de coorte.

A expressiva popularidade deste tipo de delineamento pode ser atribuída a diversos fatores, entre eles o baixo custo, a facilidade de realização, a rapidez com que é empregado e a objetividade na coleta de dados (BASTOS; DUQUIA, 2007)

Exemplos desses estudos podem ser encontrados em Chrestani et al. (2008) onde a ideia era avaliar a saúde infantil nos municípios estudados a partir do monitoramento de alguns indicadores básicos de saúde, dentre os quais aqueles relacionados à assistência à gestação e ao parto entre crianças menores de cinco anos. No artigo de Sitta et al. (2010) foi realizada uma analise de estudos epidemiológicos de caráter transversal que focam alterações em pacientes afásicos adultos para investigação das suas principais manifestações. No trabalho de Mendes et al. (2006) foi verificada a agregação familiar de fatores de risco para doenças cardiovasculares, observando frequência de excesso de peso e obesidade, sedentarismo, tabagismo e hipertensão arterial

2.2.2 Estudo de Caso-Controle

De acordo com Fontelles (2012), o estudo de caso-controle é um tipo de estudo observacional no qual o pesquisador seleciona, a partir de uma população, dois grupos de indivíduos. O primeiro grupo, definido como *caso*, é composto de indivíduos portadores de uma condição específica, como uma doença ou desfecho

clínico. O segundo grupo, definido como *controle*, é constituído por indivíduos que não apresentam tal condição específica. Os dois grupos são comparados em relação ao seu histórico com o objetivo de detectar possíveis eventos causadores da doença ou fatores de risco, tornado-o um estudo retrospectivo. Um desenho esquemático do método pode ser observado na Figura 1.

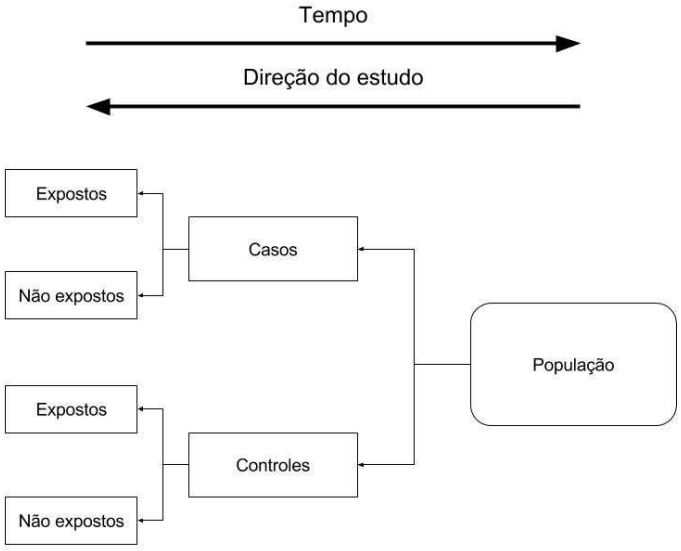

Figura 1: Delineamento de um estudo de caso-controle

Deste modo, a proporção de indivíduos expostos a determinados fatores é medida e estudada em ambos os grupos. Caso a proporção de expostos ao fator de risco no grupo *caso* seja superior à do grupo *controle*, existem evidências que tal fator de risco seja preponderante na manifestação da condição em estudo. Por outro lado, se a proporção for superior no grupo controle, tal fator pode ser um potencial fator de proteção, ou seja, a exposição ao fator em estudo reduz a incidência da condição em estudo.

A seleção de indivíduos que irão compor os grupos caso e controle deve ser realizada com base em critérios rígidos para evitar possíveis vieses que possam comprometer a qualidade do estudo. A composição dos grupos deve ser o mais

próxima possível entre si em relação às características sociodemográficas dos indivíduos, tais como localização geográfica, proporções de sexo, faixa etária, etnia, etc. O tamanho dos grupos não necessita ser necessariamente igual, pois determinadas condições podem apresentar baixa prevalência, como por exemplo doenças raras.

As grandes vantagens do estudo de caso-controle são o baixo custo, a rápida execução e a possibilidade de verificação de vários fatores de risco associados a um determinado desfecho. Uma das desvantagens dos estudos de caso-controle está associada ao rigor na seleção dos grupos, citado anteriormente, o que pode tornar difícil a obtenção dos indivíduos que atendam aos critérios de seleção.

No artigo de Loffredo et al. (1994) foi feito um estudo de caso-controle onde o grupo de casos foi composto pelos portadores de fissura labial ou labio-palatina e de fissura palatina, sem malformações associadas, enquanto o grupo controle foi formado pelos não-portadores de quaisquer anomalias, sendo casos e controles crianças menores de um ano. Já em Silva et al. (2006) foi feito um estudo para verificar a existência de relação entre os acidentes ocupacionais e os riscos ergonômicos no âmbito da organização do processo de trabalho de Enfermagem. Outros exemplos de estudos de caso-controle podem ser vistos em Niobey et al. (1992), Baughman et al. (2001) e Vanderlei, Silva e Braga (2003).

2.2.3 Estudo de Coorte

Os estudos de coorte, também conhecidos como estudos longitudinais ou de incidência, são iniciados com um grupo de indivíduos livres da enfermidade em estudo, que são classificadas em subgrupos de acordo com a exposição às potenciais causas da doença. Definidas as variáveis de interesse, estas variáveis são medidas e a coorte inteira é acompanhada com o objetivo de identificar o surgimento de novos casos da doença e verificar se a incidência difere entre os grupos formados conforme a exposição aos fatores de risco. (BONITA; BEAGLEHOLE; KJELLSTRÖM, 2008). Um fluxograma do método é apresentado na Figura 2.

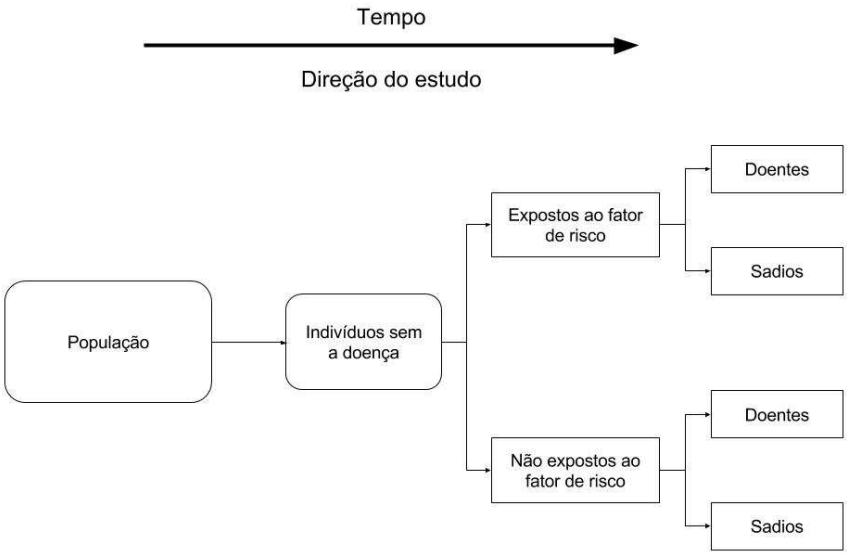

Figura 2: Delineamento de um estudo de coorte

Estudos de coorte são excelentes formas de avaliar associação de fatores e são os mais relevantes dos estudos observacionais frequentemente realizados. Sua grande desvantagem é o custo, visto que muitos indivíduos devem ser monitorados e analisados (ARAGÃO, 2013). Além da necessidade de acompanhamento de um grande número de indivíduos, os estudos de coorte podem requerer acompanhamento por grandes períodos de tempo, visto que diversas doenças apenas se manifestam após exposição prolongada aos fatores de risco, como por exemplo os canceres induzidos por exposição à radiação, tabagismo devido à exposição ao tabaco na infância, entre outros. Em alguns casos a informação sobre a exposição no passado pode ser coletada no momento da definição da coorte.

Os estudos de coorte também podem ser utilizados na investigação de efeitos crônicos ou tardios. Um outro percalço que pode surgir neste tipo de metodologia é o estudo de doenças raras, dificultando a garantia de que o grupo em estudo é grande o suficiente para ser significativo. Em Bloch, Melo e Nogueira (2008), uma coorte é analisada com o intuito de verificar a adesão ao tratamento de pres-

são arterial. Já em Moreira et al. (2013), o estudo de uma coorte composta por adolescentes indicou que a obesidade é o principal agente causador de hipertensão arterial sistêmicas em jovens desta faixa etária.

Uma alternativa bastante utilizada na redução dos custos é a análise de coortes históricas, baseadas em registros médicos ou dados prontuários. Em Rodrigues et al. (2006), uma coorte histórica foi construída com base em dados de prontuários para o estudo de pacientes infectados com leishmaniose cutânea tratados com antimoniato de meglumina, buscando compreender os motivos para o insucesso do tratamento. Palacio, Candeloro e Lopes (2009) investigaram, com base em dados registrados no departamento médico de um clube de futebol, os fatores preponderantes à ocorrência de lesões em jogadores de futebol e seu tempo de recuperação. Outras aplicações de coortes históricas podem ser encontradas em Dias, Araújo e Laranjeira (2011) e Maciel et al. (2009).

2.3 Estudos Experimentais

2.3.1 Ensaios Clínicos Aleatorizados

Entende-se por ensaio clínico um estudo planejado cuja finalidade primária seria a avaliação da eficácia e da segurança de intervenções sanitárias, médicas ou cirúrgicas. A organização Mundial da Saúde assim define ensaios clínicos:"...um experimento planejado ética e cuidadosamente com o propósito de responder a algumas perguntas precisas e bem delineadas"(OLIVEIRA, 2006).

Os ensaios clínicos constituem-se numa poderosa ferramenta para a avaliação de intervenções para a saúde, sejam elas medicamentosas ou não. O primeiro ensaio clínico, nos moldes que hoje conhecemos, foi publicado no final da década de 40, quando o estatístico Sir Austin Bradford Hill alocou aleatoriamente pacientes com tuberculose pulmonar em dois grupos: os que receberiam estreptomicina e os que não receberiam o medicamento. Desta forma, ele pode avaliar, de maneira não

viesada, a eficácia deste medicamento (COUTINHO; CUNHA, 2005).

Segundo Lachin (1988) o objetivo de qualquer atividade científica é a aquisição de novos conhecimentos. Na investigação científica empírica, novos conhecimentos ou resultados científicos são gerados por uma investigação ou estudo. A validade de quaisquer resultados científicos depende da forma como os dados ou observações são coletados, ou seja, no projeto e na condução do estudo, bem como a forma como os dados são analisados. Tais considerações são muitas vezes as áreas de especialização do estatístico. A análise estatística por si só não é suficiente para fornecer validade científica, porque a qualidade de qualquer informação derivada de uma análise de dados é determinada principalmente pela qualidade dos próprios dados. Portanto, no esforço para adquirir informações cientificamente válidas, é preciso considerar todos os aspectos de um estudo: desenho, execução e análise

Estudos que utilizam ensaios clínicos são amplamente utilizados, sobretudo em estudos relacionados às ciências da saúde, por exemplo em estudos de bioequivalência, teste de novas drogas e novas terapias. Em Pereira, Mesquita e Gomes (2014) ensaios clínicos foram utilizados para comparação entre métodos minimamente invasivos no tratamento da doença venosa crônica dos membros inferiores; no estudo de Fukuda et al. (2011) testou a eficácia a curto prazo do laser de baixa intensidade em pacientes com osteoartrite do joelho; no experimento de Amorim e Santos (2003) buscavam testar a eficácia e a tolerância do gel de aroeira (Schinus terebinthifolius Raddi) para tratamento da vaginose bacteriana.

Para que os ensaios clínicos apesentem validade científica, é necessária a observação de aspectos relacionados ao planejamento estatístico do estudo. Um ensaio clínico que atende a todos os requisitos metodológicos referentes a tal planejamento é denominado ensaio clínico aleatorizado.

Diz-se que um ensaio clínico é aleatorizado quando os indivíduos elegíveis ao estudo são alocados nos diferentes grupos de tratamentos de maneira casual, segundo, por exemplo, a geração de uma sequência de números aleatórios em um programa de computador. Em um ensaio aleatorizado, portanto, não há qualquer

controle do pesquisador sobre a decisão de destinar um paciente a um ou outro grupo; e nem os pacientes participam desta escolha. Os primeiros experimentos aleatorizados foram realizados na agricultura, e suas ideias foram posteriormente adaptadas aos outras áreas da pesquisa científica. Os propósitos da aleatorização são: (a) evitar vieses e (b) garantir que os pressupostos exigidos pelos métodos tradicionais de análise estatística sejam respeitados. (MARTINEZ, 2007)

O princípio da aleatorização é agora uma característica fundamental do método científico e é empregado em muitos campos de pesquisa empírica. A aleatorização é um problema em cada um dos três componentes de um ensaio clínico: planejamento, conduta e análise. Ensaios clínicos aleatorizados utilizam a probabilidade como um método de atribuição de tratamentos aos pacientes (ROSENBERGER; LACHIN, 2015).

Diversos são os estudos baseados em ensaios clínicos aleatorizados. Em Marinho et al. (2007) foi realizado um ensaio clinico aleatorizado para verificar se a prática do Tai Chi Chuan na população idosa apresenta efeitos positivos no controle do equilíbrio, na incidência de quedas e no medo de cair. Na pesquisa de Marinho, Chaves e Tarabal (2014) foi realizado uma revisão sistemática de ensaios clínicos aleatorizados do efeito da intervenção da dupla-tarefa na marcha em portadores da doença de Parkinson. No trabalho Lustosa et al. (2011) foi feito um ensaio para verificar o efeito do treinamento de força muscular com carga na capacidade funcional e força muscular dos extensores do joelho e sua associação, após treinamento, em idosas pré-frágeis da comunidade. Diante dos estudos citados, percebe-se a ampla utilização de ensaios clínicos aleatorizados em pesquisas na área da saúde.

2.4 Métodos de Aleatorização

A aleatorização é o principal fundamento na utilização dos métodos estatísticos na experimentação. Por aleatorização, entende-se que tanto a alocação do material

experimental quanto a ordem na qual os ensaios individuais do experimento serão executados são determinados de maneira aleatória (MONTGOMERY, 2001).

De acordo com Altman (1990), existem dois motivos principais para utilização da aleatorização. O primeiro motivo é a prevenção de vícios. A aleatorização visa garantir que os grupos que receberão determinados tipos de intervenção sejam o mais homogêneos possível. Em ensaios clínicos aleatorizados, por exemplo, se a seleção dos sujeitos que receberão tratamento foi de responsabilidade do pesquisador, existe uma grande chance de que a seleção seja viciada, seja de forma inconsciente ou mesmo conscientemente. Nestes estudos, a aleatorização também garante a distribuição ética dos tratamentos, sem privilégios a nenhum dos sujeitos.

O segundo motivo para utilização de técnicas de aleatorização está balizado na metodologia de modelagem estatística. Toda a teoria estatística está baseada na ideia de amostras aleatórias. De modo geral, a modelagem estatística utilizada em experimentação tem como pressuposto que os erros experimentais sejam variáveis aleatórias independentemente distribuídas. A aleatorização geralmente garante que esse pressuposto seja válido. Além disto, uma amostragem realizada adequadamente garante uma distribuição igualitária dos fatores externos não controláveis que podem estar presentes na experimentação

Existe uma infinidade de técnicas de aleatorização, cada qual aplicada a determinados tipos de experimentos. A seguir serão apresentadas as metodologias mais utilizadas na aleatorização em experimentação, em particular na área da saúde: as aleatorizações simples, em bloco e estratificada.

2.4.1 Aleatorização Simples

De acordo com Vaz et al. (2004) é a forma mais básica de aleatorização, também pode ser denominada aleatorização completa e é melhor explicada através de exemplo: ao lançar uma moeda não viciada, cada vez que um participante é apresentado para aleatorização; se o lançamento resultar em cara, o participante

é integrado no controle. Se o resultado do lançamento for coroa o participante é alocado no grupo de intervenção (por exemplo).

A sua vantagem é o fácil entendimento do método. De modo geral, este método gera grupos com número similar de participantes, entretanto, desequilíbrios podem ocorrer em qualquer estágio do processo, particularmente em pequenos grupos. Por exemplo, em um grupo de 20 indivíduos, a probabilidade de ocorrer uma divisão de grupos de do tipo 12/8 ou mais desbalanceados é de 50%. Tais desequilíbrios tendem a reduzir a habilidade de detecção de diferenças entre grupos e por este motivo, este método de aleatorização não é recomendado para grupos pequenos.

2.4.2 Aleatorização em Bloco

De acordo com Suresh (2011), o método de aleatorização em blocos tem como objetivo aleatorizar sujeitos em grupos de tamanhos iguais, garantindo o equilíbrio entre os tamanhos de amostra entre os grupos durante todo o período de estudo. Blocos são pequenos e equilibrados grupos de sujeitos com determinadas características pré-estabelecidas, que manterão o mesmo número de indivíduos em todos os tempos. O tamanho do bloco é determinado pelo pesquisador e deve ser um múltiplo do número de grupos. Por exemplo, com dois grupos de tratamentos, o tamanho do bloco deve ser de 4, 6, 8 ou outro múltiplo de 2. Após a determinação do tamanho do bloco, todas as possíveis combinações de tratamentos dentro dos blocos deverão ser calculadas. Então, os blocos são sorteados aleatoriamente para determinar quais pacientes serão alocados em que grupos.

Embora o equilíbrio no tamanho da amostra possa ser obtido com esta metodologia, é possível que os grupos gerados apresentem incomparabilidade em termos de suas covariáveis. Pode ocorrer por exemplo que um dos grupos atribuídos possua uma maior incidência de doenças secundárias ao tratamento. Estas podem atuar como variáveis de confundimento e influenciar negativamente os resultados do experimento. Como medida preventiva, é de suma importância o controle de possíveis covariáveis inerentes às entidades em pesquisa, pois tais desequilíbrios podem in-

troduzir vício nas análises estatísticas e reduzir o poder do estudo. Deste modo, métodos de aleatorização de grupos que sejam capazes de levar em consideração as características individuais com base nas covariáveis existentes na composição dos blocos são altamente desejáveis.

2.4.3 Aleatorização Estratificada

O método de aleatorização estratificado surge da necessidade de controlar e equilibrar a influência das covariaveis (SURESH, 2011). Este método pode ser usado para garantir equilíbrio entre os grupos em termos de suas características relacionadas às covariáveis. Covariaveis específicas podem ser identificadas pelo pesquisador que compreende a influência de cada covariável tem na variável resposta. A ideia geral do método consiste na definição de um bloco para cada combinação de covariáveis possível e assim, designar os indivíduos no bloco apropriado em função de suas características. Após a destinação dos indivíduos para os respectivos blocos, uma aleatorização simples é realizada para designar os indivíduos que receberão cada tipo de tratamento.

A aleatorização estratificada controla as possíveis influências de covariáveis que podem viciar as conclusões de uma pesquisa. Por exemplo, em um estudo de reabilitação motora, sabe-se que a idade dos indivíduos afeta diretamente a capacidade de recuperação da mobilidade. Deste modo, deve-se estratificar primeiramente a amostra por idade e só depois desta estratificação deve-se proceder à distribuição dos indivíduos dentre os tratamentos. E a principal vantagem da estratificação, além do controle dos efeitos das covariáveis é a simplicidade de sua aplicação e sua aplicabilidade em pequenos estudos. Entretanto, para grupos grandes de indivíduos e covariáveis, a estratificação se torna bastante complexa. Desta complexidade emerge a necessidade do desenvolvimento de métodos que tornem a tarefa de estratificar grupos de indivíduos com diversas covariaveis envolvidas mais simples e eficiente.

3 Material e Métodos

3.1 Introdução

Nesta seção será apresentada uma metodologia de aleatorização de tratamentos dentro de uma amostra que se submeterá a um ensaio clínico aleatorizado, bem como os procedimentos utilizados na aferição de sua precisão em garantir a formação de subgrupos homogêneos dentro da amostra inicial.

3.2 Método Simulado de Alocação de Grupos

Conforme descrito no capítulo anterior, um dos problemas encontrados durante a realização de um ensaio clínico experimental é a definição dos grupos de tratamento e controle. Por motivos éticos e de representatividade estatística da amostra, tal definição deve ser realizada completamente ao acaso e garantir que cada grupo, enquanto subamostra, apresente uma estrutura de dados que seja representativa em relação à amostra total, que por sua vez deve ser representativa em termos da população de interesse.

Para garantir os objetivos mencionados, este trabalho propõe a criação de um algoritmo baseado em simulação computacional que realizará a separação dos grupos contemplando as características de interesse de um estudo experimental. O algoritmo tem como objetivo delimitar em grupos uma amostra que se sujeitará um ensaio clínico, garantindo a similaridade dos grupos em relação à amostra total

e à população. O método de divisão dos grupos dos ensaios clínicos aleatorizados foi desenvolvido em linguagem R e, para uma melhor compreensão da metodologia, os passos utilizados no processo de alocações dos indivíduos nos grupos dos ensaios clínicos são apresentados a seguir, considerando uma amostra de p variáveis com um total de m categorias.

1º Passo. Inicialmente, o algoritmo efetua a leitura dos dados, registra o número de categorias presente em cada variável e os armazena em um vetor ordenado.

2º Passo. O algoritmo recebe as proporções das categorias de cada variável e cria um vetor ordenado contendo as proporções das categorias de cada variável na amostra original completa $p_{obs} = (p_1,\ p_2,\ \cdots,\ p_m)$.

3º Passo. Realiza, de forma aleatória, a divisão da amostra inicial em k grupos de tamanho igual n/k, em que n é o tamanho da amostra. Se o resultado obtido de n/k não for um algarismo inteiro, o tamanho de cada grupo será arredondado para o próximo inteiro. Assim, fica definido o número de elementos nos grupos.

4º Passo. Após a divisão dos k grupos, são calculadas as proporções das categorias de cada variável em cada um dos grupos, de acordo com os passos 1-2 e é criado o vetor de proporções $p_{1k} = (p_{1k},\ p_{2k},\ \cdots,\ p_{mk})$.

5º Passo. Calcula-se a somas das distâncias euclidianas entre os vetores de proporções das k subamostras e as proporções originais da amostra total, definida como D_0. Ou seja:

$$D_0 = \sum_{i=1}^{k} \sqrt{(p_{1i} - p_{obs})'(p_{1i} - p_{obs})} \tag{3.1}$$

Essa será a métrica otimizada na simulação.

6º Passo. Geram-se novas subamostras de acordo com o 4º passo e calcula-se uma nova métrica para as novas subamostras, em conformidade com o passo 6, a saber:

$$D_1 = \sum_{i=1}^{k} \sqrt{(p_{2i} - p_{obs})'(p_{2i} - p_{obs})} \tag{3.2}$$

7º Passo. As métricas das subamostras são comparadas. A amostra cuja divisão de subgrupos apresentar o menor valor de D é mantida como amostra mais verossímil à original e será a base da próxima simulação. A amostra que apresenta o maior valor de D será descartada.

8º Passo. Repete-se os passos 6 e 7 até que a métrica D atinja um critério de parada pré-estabelecido ou até que determinado número máximo de simulações ocorra. A amostra que apresentar o menor valor de D será aquela que possui maior similaridade em termos de proporção das variáveis categóricas entre os grupos se comparado às proporções originais.

A Figura 3 apresenta o pseudo-código do método.

```
algoritmo
   leia o banco de dados
      /* escolha de subgrupo inicial*/
   escolha S_0
      /* cálculo da distância entre vetores de proporções */
   calcule D_0 = Σ_{i=1}^{k} sqrt((p_0i − p_obs)'(p_0i − p_obs))
   D ← D_0
   S_opt ← S_0
      para j = 1 faça
         /* escolha de novo subgrupo */
         escolha S_j
         /* cálculo da distância entre vetores de proporções */
         calcule D_j = Σ_{i=1}^{k} sqrt((p_ji − p_obs)'(p_ji − p_obs))
         se D_j < D então
            S_opt ← S_j
         senão
            mantenha S_opt
         fim se
      enquanto j <= nsim ou D = 0
       fim para
   imprima S_opt
fim algoritmo
```

Figura 3: Algoritmo de seleção de subgrupos

De forma resumida, o método verifica a diferença existente entre as proporções das categorias de cada variável entre os subgrupos simulados e a amostra original. Uma divisão bem definida dos subgrupos acarretará em um valor de D próximo de zero, que em última análise significa que a divisão manteve a estrutura presente nos dados originais em relação à proporção de casos ocorrido em cada variável que o compõe.

3.3 Métodos de verificação da qualidade da divisão

O algorítimo proposto tem como objetivo a distribuição dos elementos da população em estudo em k subgrupos amostrais, de modo que as proporções originais de cada variável esteja representada da forma mais fidedigna possível em cada um dos grupos experimentais. Em suma, é necessário que as proporções das variáveis em cada subgrupo seja o mais próxima possível do valor amostral. Além disto, espera-se que o método apresente resultados superiores à amostragem aleatória simples, metodologia tradicionalmente utilizada na composição de grupos amostrais.

Para verificar tais suposições, serão utilizados métodos de estatística clássica univariada, paramétricos e não paramétricos. As técnicas a seguir irão compor os descritores que indicaram a qualidade da distribuição dos elementos nos subgrupos amostrais.

3.3.1 Raiz do Erro Quadrático Médio - RMSE

A *raiz do erro quadrático médio*, ou *RMSE* (do inglês *root mean squared error*), como o próprio nome indica, é uma medida quadrática dos erros de estimação (HAIR et al., 2009). Ele calcula a magnitude média do erro de estimação. O cálculo do *RMSE* pode ser realizado a partir da seguinte fórmula:

$$RMSE = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}$$

em que

- y_j: Valor observado nos dados;
- $\hat{y}_j$: valor estimado com base na modelagem de interesse;
- n: tamanho da amostra em estudo.

O RMSE é uma medida bastante empregada na avaliação de ajuste de modelos preditivos e sua qualidade de ajuste. Uma importante característica desta métrica é que esta atribuí maior peso a erros de grande magnitude. Esta característica decorre do fato de que os erros são elevados ao quadrado anteriormente ao cálculo da média. Devido a esta característica, o RMSE é bastante útil quando grandes erros são particularmente indesejáveis.

Neste estudo, o *RMSE* será bastante empregado na validação do modelo, pois busca-se uma distribuição de proporções simuladas tão próximas das proporções observadas quanto se consiga e portanto, grandes erros devem ser evitados, pois um erro grande em uma proporção, digamos, p implica diretamente em seu complementar *1-p*.

3.3.2 Scatterplot

Segundo Montgomery e Runger (2010), o *Scatterplot* é uma das nossas ferramentas mais poderosas para a análise de dados. O gráfico de dispersão é uma representação gráfica da relação entre duas ou mais variáveis. Num gráfico de duas variáveis x e y, cada ponto no gráfico é um par x-y. A Figura 4 apresenta um exemplo de *Scatterplot*, também conhecido como diagrama de dispersão.

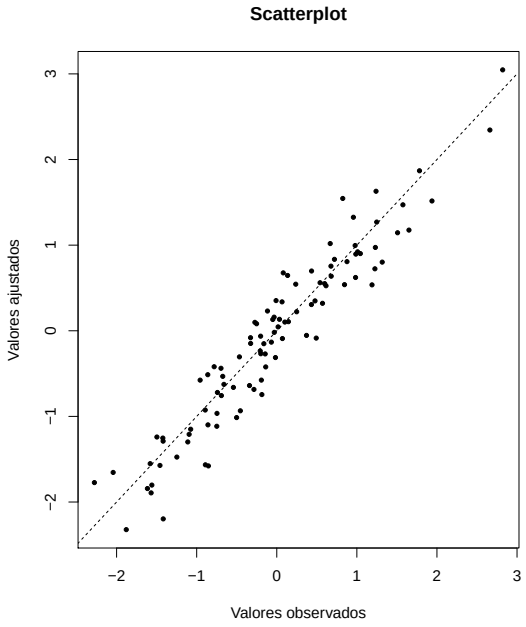

Figura 4: Scatterplot relacionando valores observados e esperados

3.4 Banco de dados

O emprego da técnica proposta, bem como sua capacidade de criar grupos de indivíduos homogêneos em relação as variáveis de interesse da população alvo, foi realizado em um banco de dados proveniente de um projeto de pesquisa denominado Coorte de Universidades Mineiras - CUME que foi aprovado pelos Comitês de Ética em Pesquisa com Seres Humanos da UFV e da UFMG (n° do parecer 596.741-0/2013), bem como o estudo de validação (n° do parecer 1.588.799/2016).

O desenvolvimento desta técnica também foi motivado para suprir uma demanda de uma aleatorização para formação de grupos de indivíduos homogêneos, considerando determinadas variáveis populacionais de interesse, que participaram de um ensaio clínico randomizado do projeto nomidado de Prevenção da Fadiga, aprovado pelo Comitê de Ética em Pesquisa da UFOP, (nº do parecer CAAE: 39682014.7.0000.5150).

4 Resultados e Discussão

4.1 Introdução

Definido o algorítimo de distribuição dos grupos amostrais, neta seção serão realizados experimentos computacionais para primeiramente estudar a convergência do método, de forma a se estabelecer o número de simulações que propicie uma divisão de grupos que seja suficientemente representativa. Definido tal número mínimo de simulações que reduz consideravelmente a distância euclidiana em comparação com a amostragem aleatória simples, serão analisados alguns casos selecionados para exemplificar a eficácia do método proposto.

.

4.2 Desempenho computacional

4.2.1 Determinação do número de simulações

Para estabelecer um número de simulações que seja razoável e atenda a critérios de qualidade em relação à resposta inicial, oferecida pela amostragem aleatória simples (sorteio), considerando o banco de dados apresentado na seção 3.4, foram estabelecidos 36 cenários hipotéticos, onde se deseja dividir amostras de tamanho 50, 100 e 150 em grupos de tamanho *2, 3* e *4* considerando-se um total de 3, 4, 5 e 6 variáveis. A combinação destas três característica gera os 36 cenários mencionados.

Em cada um dos cenários apresentados foram realizadas 30 mil simulações. Os resultados foram gerados a partir de um computador dotado de processador *Intel core i5 octa-core 2.3 Ghz* e *8gb* de memória *ram*.

Segundo a metodologia descrita no capítulo 3, todas as variáveis se apresentam na escala nominal. Em casos em que as variáveis explicativas se apresentem na escala ordinal ou contínua, os dados deverão ser categorizados a critério do pesquisador. O critério de otimização definido é a soma das distâncias euclidianas entre o vetor ordenado de proporções de cada nível das variáveis explicativas da população e os vetores de proporções cada subgrupo amostral definido pelo método.

Os resultados apresentados na Tabela 1 se referem ao número de simulações necessários para se atingir uma melhoria percentual de $k\%$ da distância euclidiana do método em relação à amostragem aleatória simples, sendo k obtido a partir da seguinte relação:

$$k = \frac{d_{AAS} - d_{NM}}{d_{AAS}}$$

em que:

- d : distância euclidiana entre os vetores de proporções;
- AAS : Amostragem aleatória simples;
- NM : Nova metodologia apresentada neste trabalho.

Mesmo que a distância entre os vetores de proporções, exclusivamente, não represente indícios suficientes sobre a eficácia do método, ela servirá como uma base inicial para definição de um número aproximado de simulações que apresenta resultados consistentes para diversos cenários. Pode-se entender que o algorítimo atingiu um patamar estável nos casos em que após, um dado número de simulações, o método não consegue produzir redução percentual, independente do número de simulações, indicando que foi atingido um mínimo local.

Tabela 1: Número de iterações necessárias para redução da distância inicial até determinados limiares percentuais

Número de grupos	Tamanho da amostra	Nº de variáveis	Melhoria percentual da distância euclidiana inicial							
			5%	10%	20%	30%	40%	50%	60%	70%
2	50	3	2	2	2	3	3	3	15	15
		4	4	4	10	10	10	54	2916	3640
		5	4	4	4	10	10	72	775	3640
		6	4	4	9	13	44	44	2361	6266
	100	3	6	19	19	33	166	166	2860	*
		4	9	9	37	343	404	404	5856	*
		5	2	4	7	9	9	37	1811	8034
		6	4	4	4	38	38	1214	5990	*
	150	3	4	4	4	13	13	151	5761	*
		4	2	2	2	12	12	13	13	1316
		5	2	2	2	13	13	13	13	20502
		6	2	2	13	13	419	767	*	*
3	50	3	4	14	14	14	26	97	3357	*
		4	25	33	34	54	2014	26597	*	*
		5	25	25	50	253	681	28440	*	*
		6	4	4	9	52	253	253	28440	*
	100	3	2	2	11	50	69	418	*	*
		4	2	4	4	61	61	628	*	*
		5	2	2	2	4	61	61	*	*
		6	2	2	4	61	3663	*	*	*
	150	3	4	14	20	623	1203	5643	*	*
		4	2	2	32	47	253	299	299	*
		5	2	2	2	32	299	6519	*	*
		6	2	2	2	47	299	*	*	*
4	50	3	7	7	27	164	164	*	*	*
		4	4	4	17	17	297	1121	*	*
		5	4	4	17	17	401	*	*	*
		6	4	4	17	37	297	19262	*	*
	100	3	19	19	55	72	966	*	*	*
		4	37	40	2372	*	*	*	*	*
		5	2	2	2	387	2372	*	*	*
		6	2	2	2081	*	*	*	*	*
	150	3	2	2	121	597	623	*	*	*
		4	2	2	2	2	299	4713	*	*
		5	2	2	2	2	335	*	*	*
		6	2	2	335	335	*	*	*	*

Entre os cenários estudados, observou-se uma melhoria percentual máxima de 70%, sobretudo nos cenários em que a amostra foi dividida em dois subgrupos. Casos nos quais a amostra foi separada em três subgrupos apresentaram melhoria máxima de 60% enquanto a divisão em quatro grupos gera uma redução percentual máxima de 50%.

Entre todos os cenários observados, um número de 7 mil simulações garantiu a estabilização de aproximadamente 83,33% dos casos. Dentre os casos que não atingiram estabilidade, o mesmo número de simulações garantiu uma redução percentual da distância no máximo 10% menor que a redução máxima alcançada, indicando que apesar de não obter um valor ótimo em todos os casos, tal número de simulações apresenta resultados satisfatórios em todos os cenários estudados. Por se tratar de um número razoável de simulações, que garante uma redução considerável na métrica de otimização, nos casos que serão estudados na sequência serão utilizadas 7 mil simulações.

4.2.2 Tempo médio de simulação

O tempo médio de duração do procedimento de simulação foi avaliado nos 36 cenários, considerando o número estabelecido de 7 mil simulações, com o intuito de se verificar a eficiência do método em termos do tempo computacional. Para tal, foram realizadas análises descritivas para verificar o comportamento do tempo diante das diferentes situações simuladas com a utilização de tabelas e boxplots.

Diante dos resultados obtidos na Figura 5 e na Tabela 2 percebe-se que o aumento do número de subgrupos impacta diretamente no tempo de simulação, sendo que quanto menor o número de subgrupos em que se dividirá a amostra total, menor será o tempo de convergência do método.

Em relação ao tamanho da amostra (Tabela 3, Figura 6), não existem evidências que indiquem um impacto no tempo de convergência com o seu incremento, pois observa-se que o tempo médio e mediano é praticamente o mesmo pra qual-

Tabela 2: Estatísticas descritivas - Tempo de simulação em segundos vs Número de Subgrupos

Número de Subgrupos	Média	Mediana	Desvio Padrão	Distância Interquartílica
2	6,88	7,15	1,33	1,68
3	9,87	9,54	2,2	3,46
4	12,19	11,87	2,54	3,71

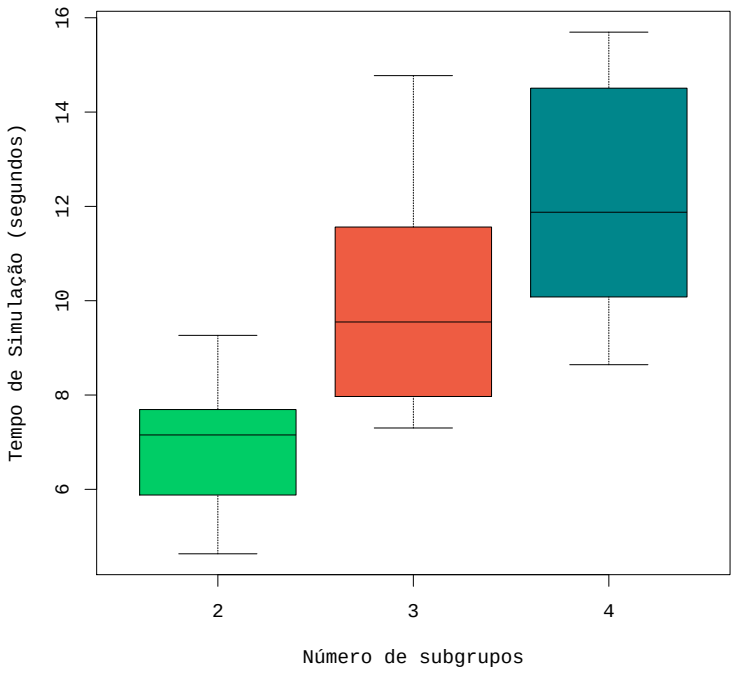

Figura 5: Boxplot: Tempo de simulação vs Número de Subgrupos

quer tamanho. Tal resultado implica que o método é eficiente em termos de tempo de simulação, independente do tamanho da amostra original.

Tabela 3: Estatísticas descritivas - Tempo de simulação em segundos vs Tamanho da amostra

Tamanho da Amostra	Média	Mediana	Desvio Padrão	Distância Interquartílica
50	9,45	8,86	2,72	3,2
100	9,88	9,03	3,43	4,5
150	9,61	9,39	3,04	3,6

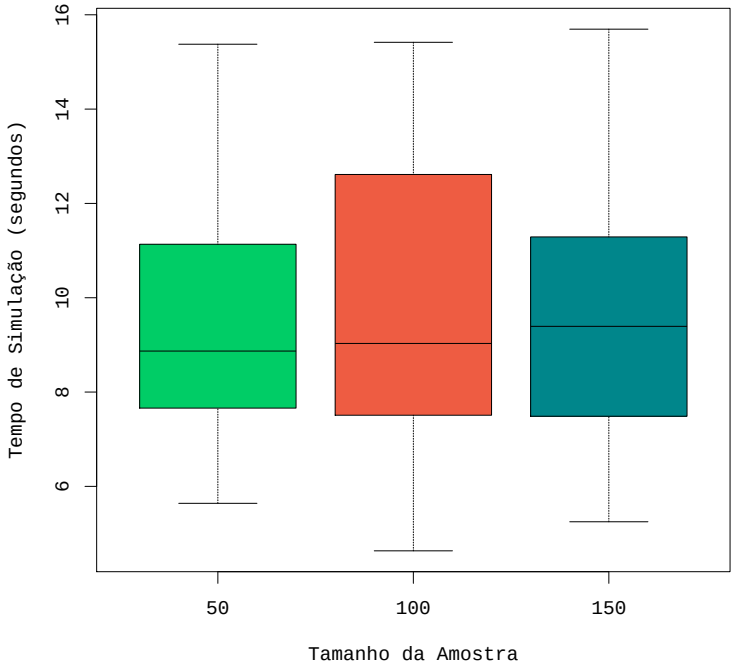

Figura 6: Boxlot: Tempo de simulação vs Tamanho da amostra

Já o número de variáveis do banco de dados tem impacto direto no tempo de simulação, de acordo com os resultados apresentados na Tabela 7 e na Figura 4, indicando que quanto maior o número de variáveis maior será o tempo de convergência.

Tabela 4: Estatísticas descritivas - Tempo de simulação em segundos vs Número de variáveis

Número de Variáveis	Média	Mediana	Desvio Padrão	Distância Interquartílica
3	7,29	7,76	1,74	3
4	8,8	8,67	1,92	2,93
5	10,37	10,19	3,64	4,94
6	12,15	11,63	3,29	6,11

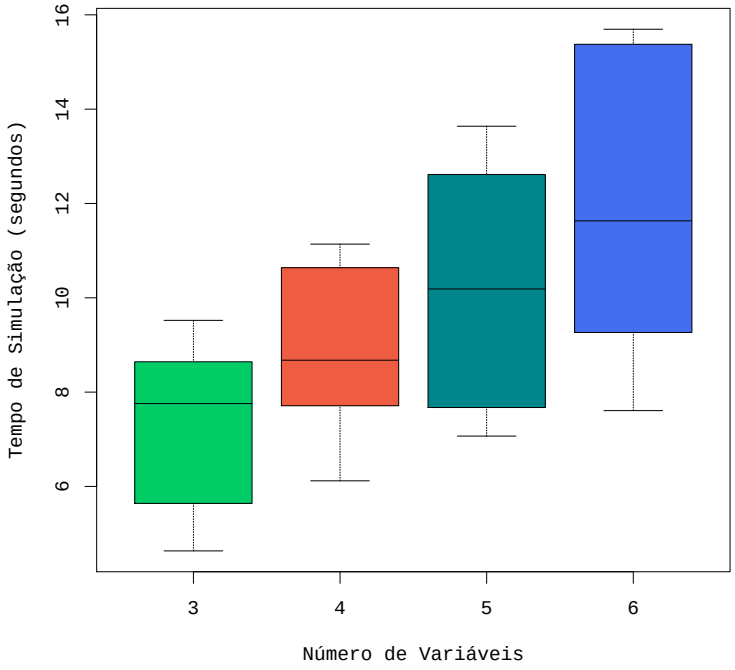

Figura 7: Boxplot: Tempo de simulação vs Número de variáveis

Sobre o tempo de simulação, observa-se que apenas as o número de variáveis e o número de subgrupos impacta diretamente no tempo de simulação, ou seja, apenas estes fatores acrescentam complexidade ao problema, impactando em alterações significativas no tempo computacional. Entretanto, independente das características da simulação, em todos os cenários o algorítimo convergiu rapidamente, com tempo médio de convergência inferior a 16 segundos para 7 mil simulações. Tal eficácia permite que, a critério do pesquisador, possam ser realizados maiores números de simulações sem que haja lentidão no processo.

4.3 Análise de eficiência do método

Verificado o desempenho computacional do algorítimo proposto, foi realizada a análise da eficiência do método em termos da qualidade da distribuição dos indivíduos em cada subgrupo em relação à sua capacidade de gerar tais grupos de forma homogênea e verossímil com a população alvo. Para tal, comparamos os resultados obtidos com a nova metodologia com as proporções da população original, bem como os resultados obtidos pelo método tradicional de distribuição, a amostragem aleatória simples (sorteio aleatório). Foram utilizadas ferramentas gráficas e como medida de desempenho a raiz do erro quadrático médio, que a partir deste ponto, por simplicidade foi representada pela sigla RMSE.

Alguns dos cenários simulados serão apresentados individualmente e, na sequência, uma análise geral dos resultados será realizada.

4.3.1 Distribuição de amostra de tamanho 50 e 6 variáveis em 2 subgrupos

Inicialmente foi analisado o método de amostragem aleatória simples. Na Figura 8, observa-se que método tradicional apresentou subamostras com certa similaridade em relação à amostra original. Porém, na subamostra 2, percebe-se que o método falhou em distribuir os casos na variável LDL, pois tal subgrupo não apre-

sentou os três níveis presentes na amostra original. Além desta ausência, observa-se desequilíbrio entre as subamostras, sendo que se espera uma homogeneidade maior entre as mesmas.

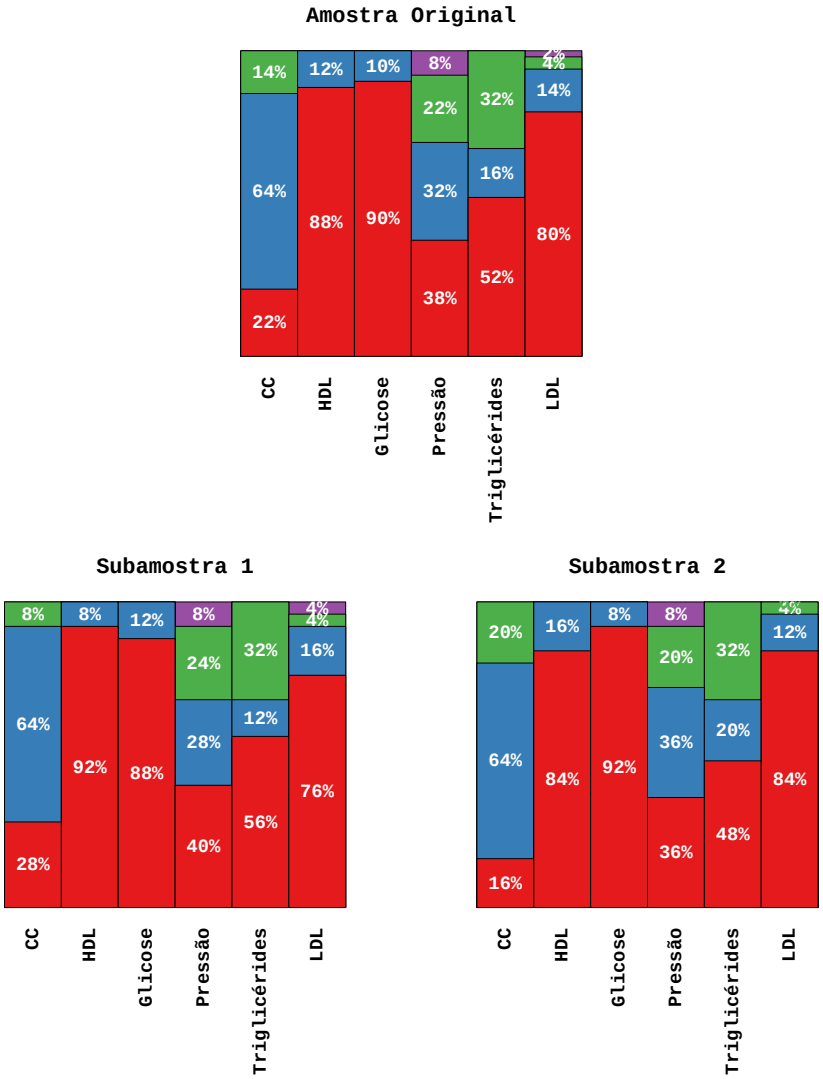

Figura 8: Distribuição dos subgrupos por AAS para amostra de tamanho 50 e 6 variáveis em 2 subgrupos

Já o método proposto neste trabalho apresentou resultados mais consistentes (Figura 9) para o cenário. Apesar de também pecar na ausência de todos os níveis na segunda subamostra, o método apresentou resultados mais compatíveis com a

amostra original e homogêneos entre si.

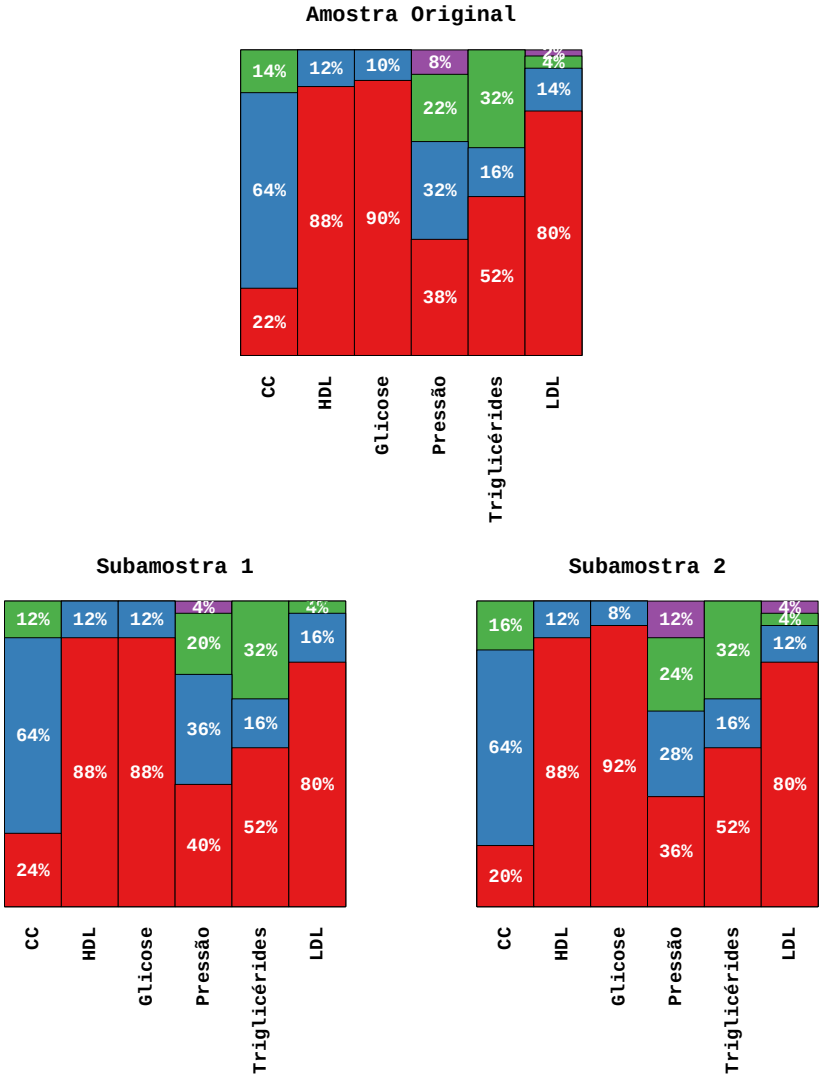

Figura 9: Distribuição dos subgrupos pelo novo método para amostra de tamanho 50 e 6 variáveis em 2 subgrupos

Verificando-se o diagrama de dispersão entre as proporções observadas e as proporções de cada subamostra, observa-se que apesar de apresentar certa proximidade da diagonal do primeiro quadrante, existe uma maior dispersão dos valores no método de amostragem aleatória simples (Figura 10). Para a distribuição realizada, foi observado um valor de RMSE de 0,065.

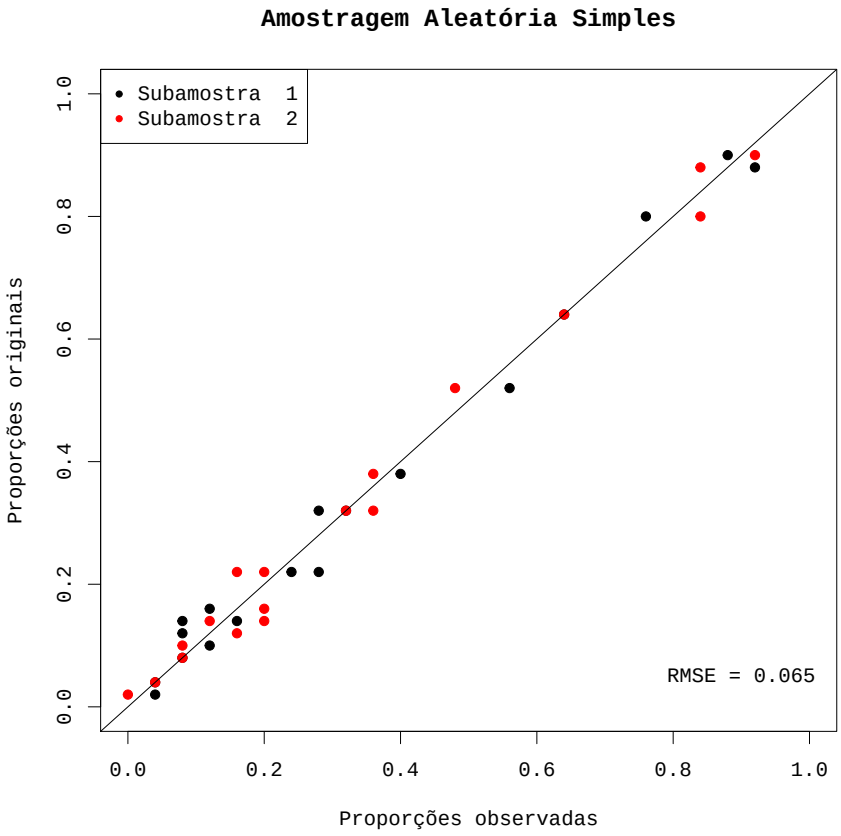

Figura 10: Diagrama de dispersão - AAS vs Proporções reais - amostra de tamanho 50 e 6 variáveis em 2 subgrupos

Para a mesma amostra, a metodologia proposta apresentou um diagrama de dispersão com maior aproximação dos valores observados à diagonal do primeiro quadrante (Figura 11), o que indica uma maior similaridade entre as proporções simuladas e aquelas observadas na amostra original.

Além desta percepção gráfica, o valor de RMSE observado foi de 0,038, o que representa uma redução de aproximadamente 50% no valor do RMSE observado na AAS, fato que corrobora com a conclusão anterior. Assim, observa-se que, para o cenário apresentado, a metodologia proposta apresenta resultados superiores se comparados com aqueles apresentados pela abordagem tradicional.

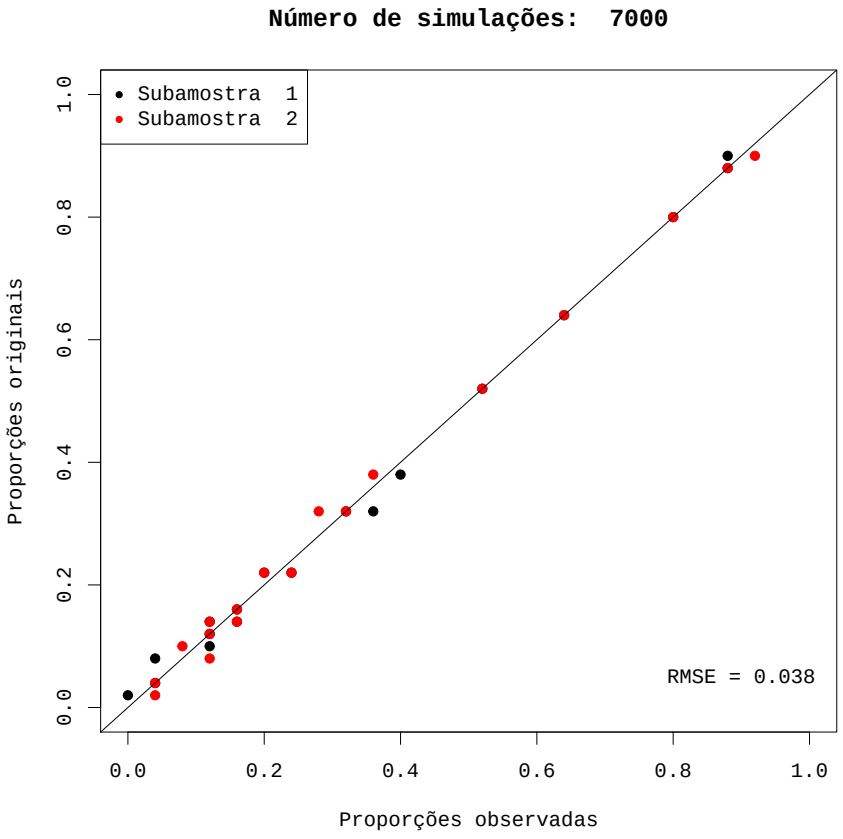

Figura 11: Diagrama de dispersão - Método simulado vs Proporções reais - amostra de tamanho 50 e 6 variáveis em 2 subgrupos

A seguir, apresentaremos os resultados para um segundo cenário, com mais

subamostras e menos variáveis.

4.3.2 Distribuição de amostra de tamanho 100 e 4 variáveis em 4 subgrupos

Os resultados obtidos pelo sorteio aleatório simples (Figura 12), assim como no cenário anterior, pecou em representar a totalidade dos níveis em duas das quatro subamostras, sobretudo na variável LDL. Tal fato pode ser explicado pela baixa proporção dos níveis da variável.

Além destas ausências, novamente observa-se uma heterogeneidade entre as quatro subamostras. Por exemplo, enquanto o nível mais frequente da variável LDL originalmente apresenta-se em 66% dos indivíduos, as subamostras apresentam proporções de 64%, 80%, 48% e 72%, respectivamente. Ou seja, além do afastamento das proporções em relação ao valor original, existe discrepância entre si, o que não é desejado pelos pesquisadores, sobretudo em estudos na área da saúde.

Os resultados obtidos pelo novo método, apresentados na Figura 13, apresentaram uma maior representatividade em termos dos níveis de cada variável, sendo que em apenas uma das subamostras os níveis com baixa proporção deixaram de ser representados. Outro ponto positivo foi a homogeneidade apresentada entre as subamostras, bastante superior se comparada àquelas geradas pela amostragem aleatória simples.

O diagrama de dispersão das proporções reais vs valores ajustados (Figura 14) apresenta um afastamento considerável da reta diagonal, indicando que as proporções geradas pela amostragem aleatória simples, conforme mostrou o gráfico de perfil, apresenta proporções de subamostras que não são similares às originais.

Em contrapartida, o diagrama apresentado na Figura 15, gerado pelas subamostras obtidas pelo novo método, apresentam-se mais próximos à diagonal do primeiro quadrante e com menor dispersão, indicando que as subamostras geradas

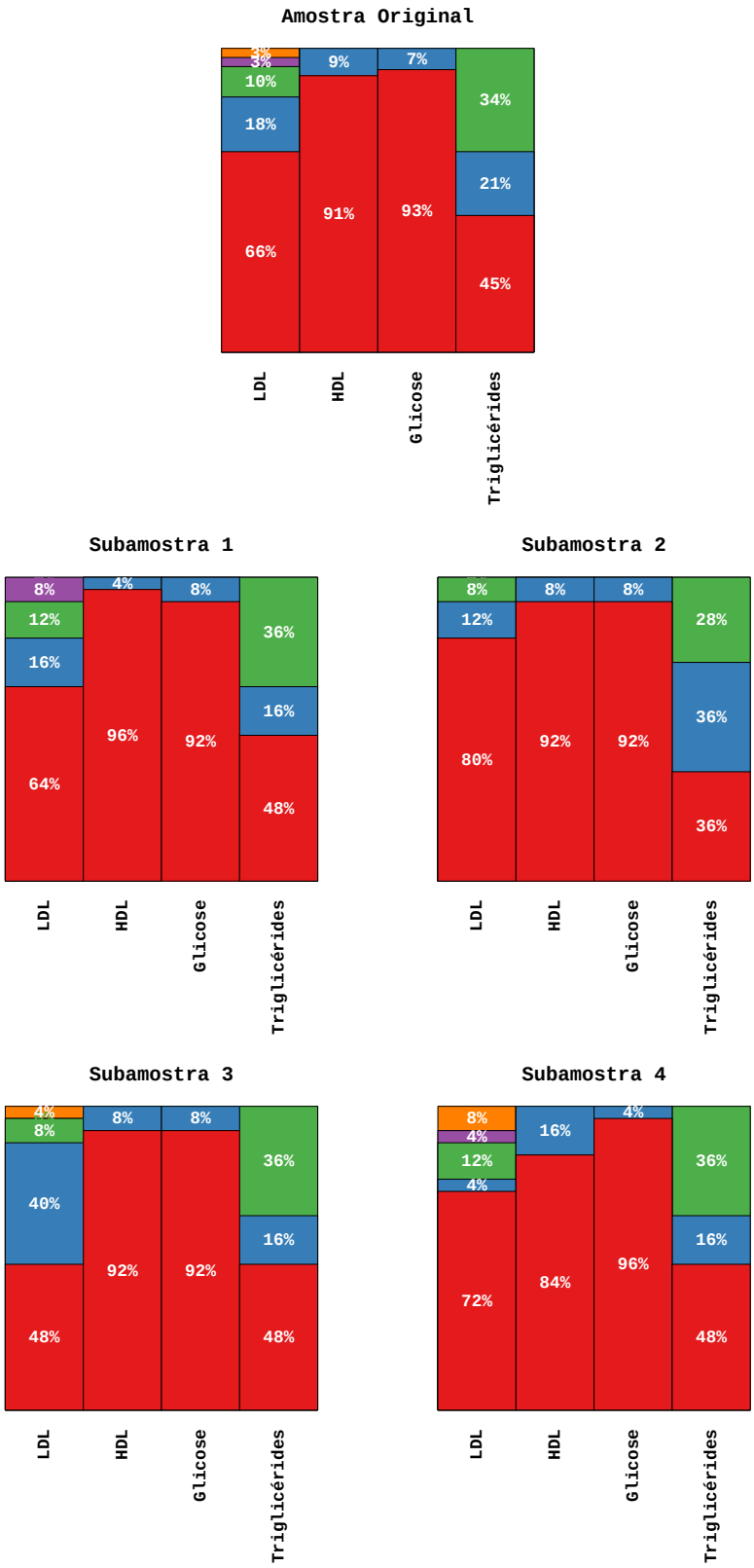

Figura 12: Distribuição dos subgrupos por AAS para amostra de tamanho 100 e 4 variáveis em 4 subgrupos

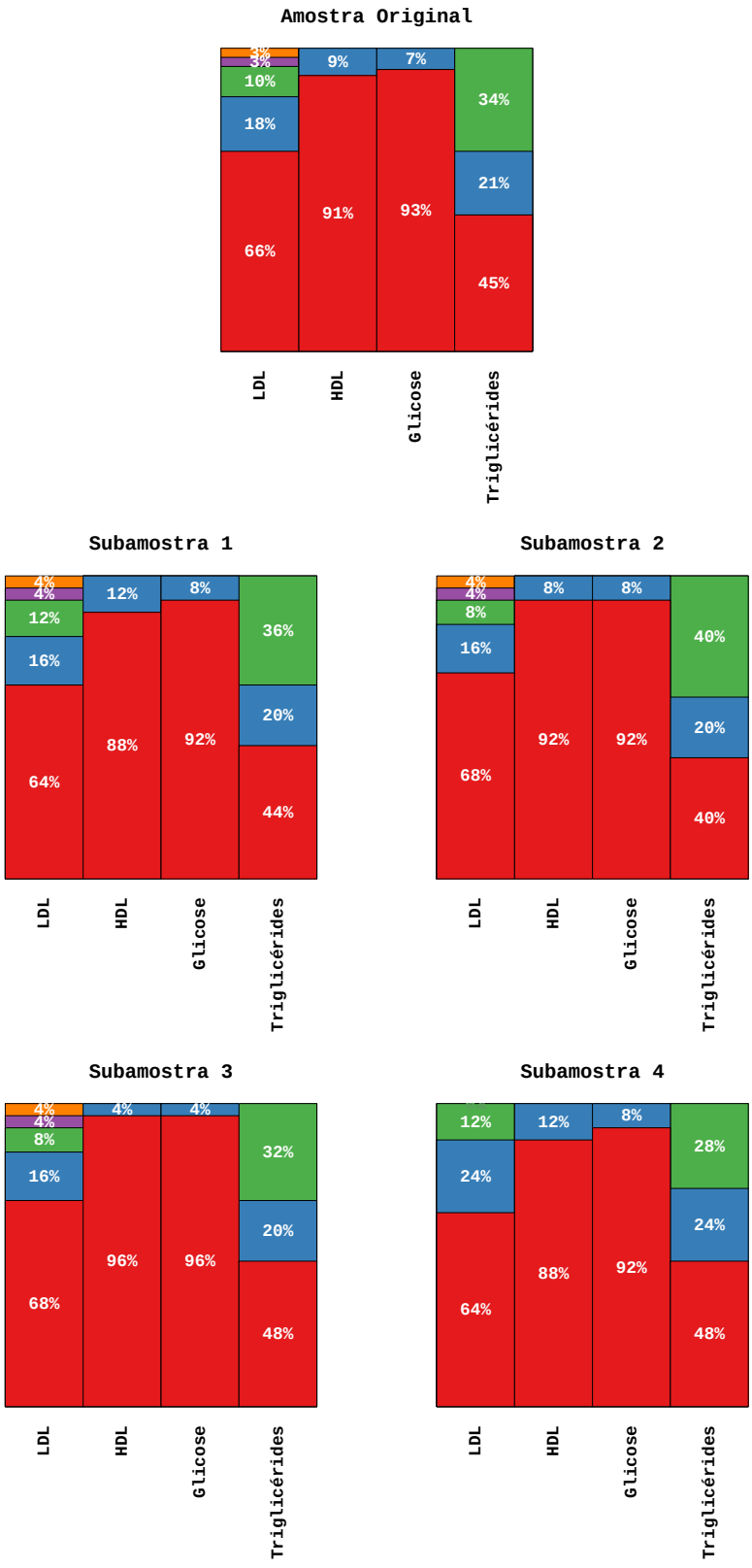

Figura 13: Distribuição dos subgrupos pelo novo método para amostra de tamanho 100 e 4 variáveis em 4 subgrupos

pelo novo método de fato são mais fidedignas à amostra original se comparadas àquelas geradas pela AAS. Em termos do RMSE, enquanto o método tradicional apresenta valores próximos de 0,25, o novo método apresentou RMSE de aproximadamente 0,11, ou seja, observa-se uma redução da métrica superior a 50%.

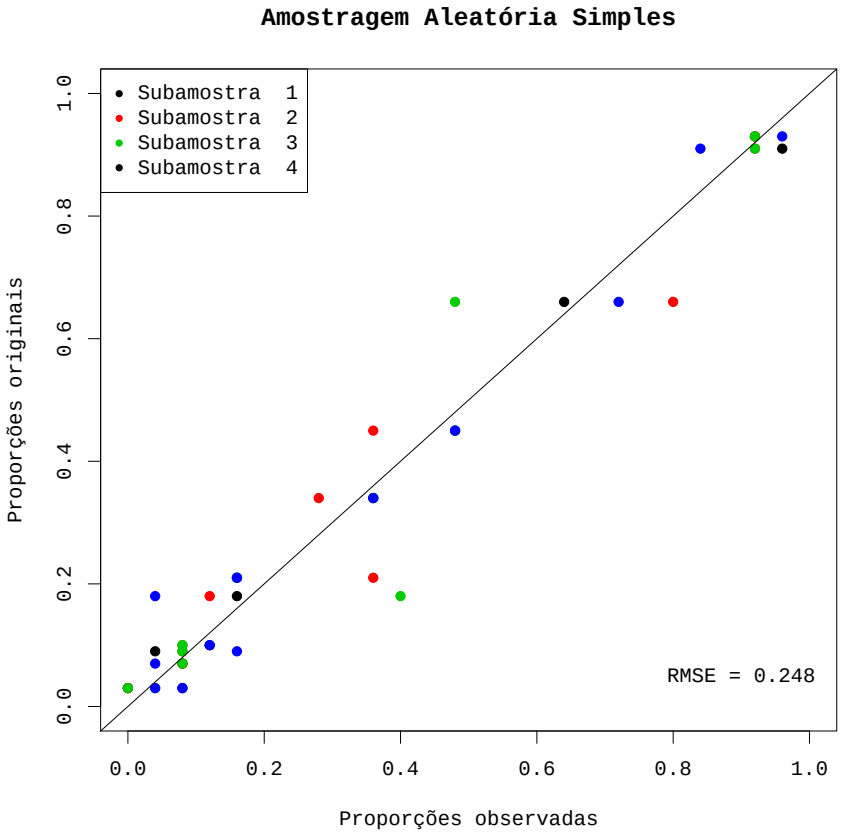

Figura 14: Diagrama de dispersão - AAS vs Proporções reais - amostra de tamanho 100 e 4 variáveis em 4 subgrupos

Assim como no caso explorado anteriormente, neste cenário, o método simulado também supera o tradicionalmente aplicado, gerando subamostras mais similares à população original que garantirão uma maior representatividade de cada uma delas.

O terceiro e último caso analisado individualmente apresenta um cenário com um número intermediário de grupos e de variáveis.

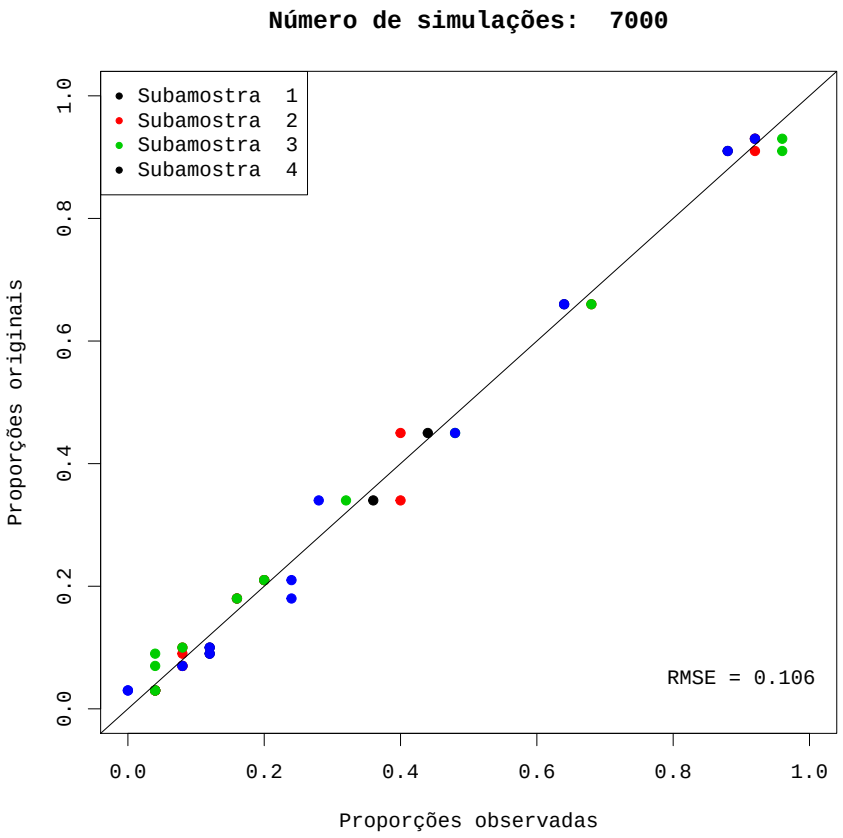

Figura 15: Diagrama de dispersão - Método simulado vs Proporções reais - amostra de tamanho 100 e 4 variáveis em 4 subgrupos

4.3.3 Distribuição de amostra de tamanho 150 e 5 variáveis em 3 subgrupos

O cenário a seguir apresenta uma distribuição da amostra original em três subgrupos. Na Figura 16 , observa-se que assim como os casos anteriores, a amostragem aleatória simples dos grupos peca na representatividade de níveis com baixa frequência, sendo que na subamostra 3, novamente na variável *LDL* não está presente um dos níveis.

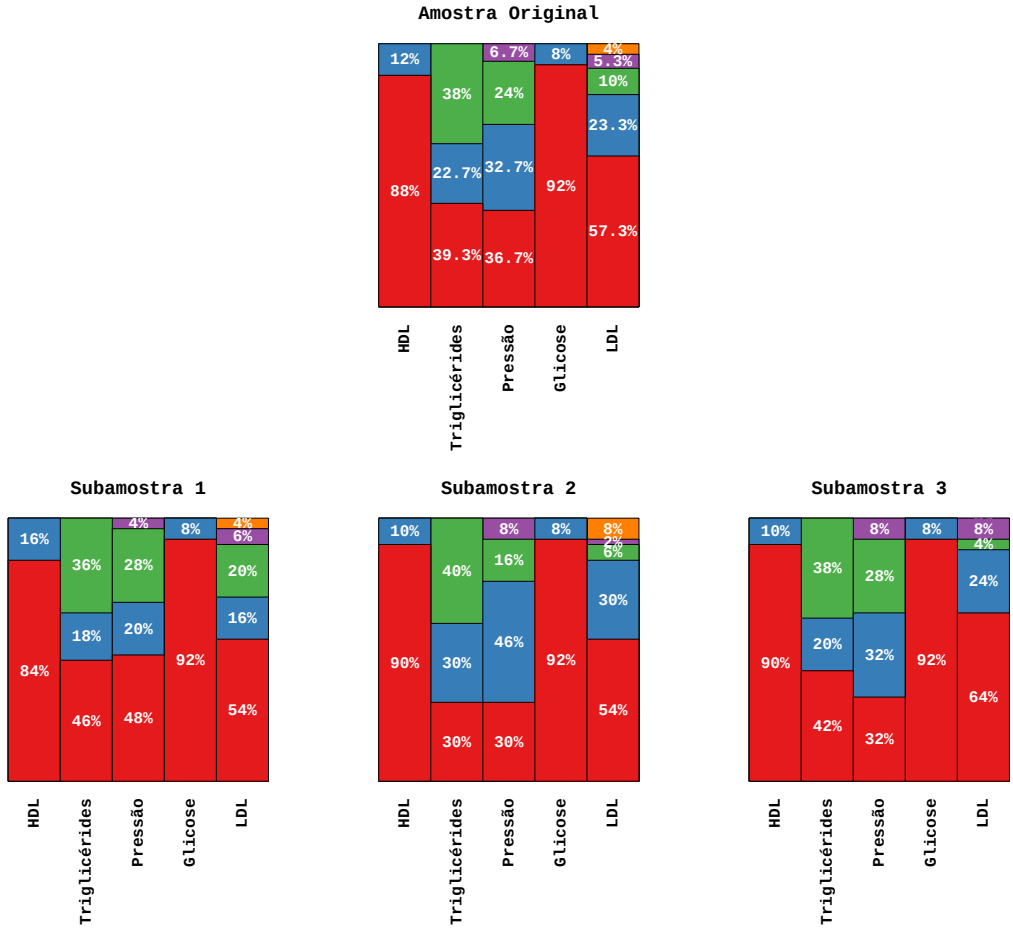

Figura 16: Distribuição dos subgrupos por AAS para amostra de tamanho 150 e 5 variáveis em 3 subgrupos

Neste caso, porém, as subamostras se apresentam mais homogêneas que as

apresentadas anteriormente para a amostragem aleatória simples. Apesar disso, esperava-se resultados superiores, dado que a amostra apresenta mais indivíduos que os casos anteriores (150).

Já a nova metodologia (Figura 17) apresentou um resultado bastante satisfatório, pois todos os níveis de variáveis foram representados nas subamostras extraídas.

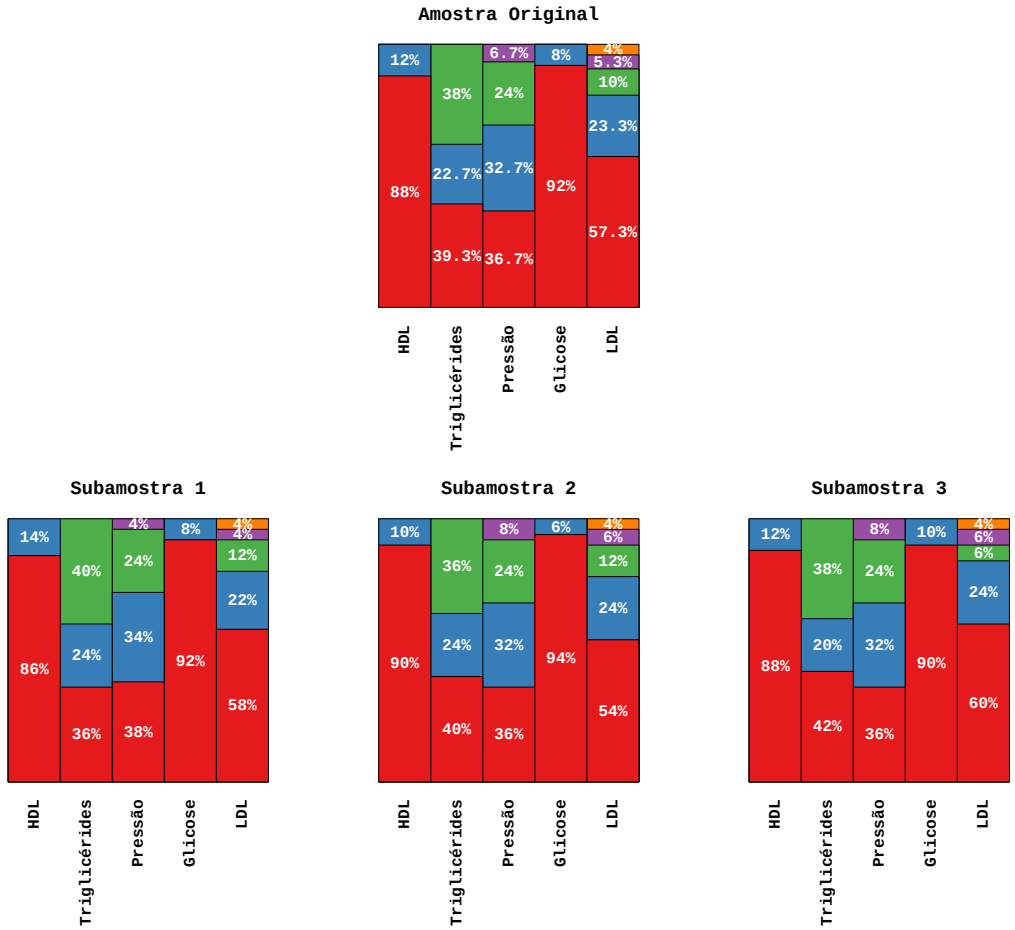

Figura 17: Distribuição dos subgrupos pelo novo método para amostra de tamanho 150 e 5 variáveis em 3 subgrupos

A homogeneidade entre as subamostras também deve ser destacada neste caso. Os perfis apresentados são bastante similares entre si e também se comparados à

amostra original. O maior desvio entre as proporções ficou na casa dos 4% na variável *LDL* da subamostra 4.

Se compararmos os diagramas de dispersão da amostragem aleatória simples com aquele obtido pela nova metodologia, é possível perceber uma maior dispersão dos pontos na Figura 18 se comparados aos obtidos na Figura 19

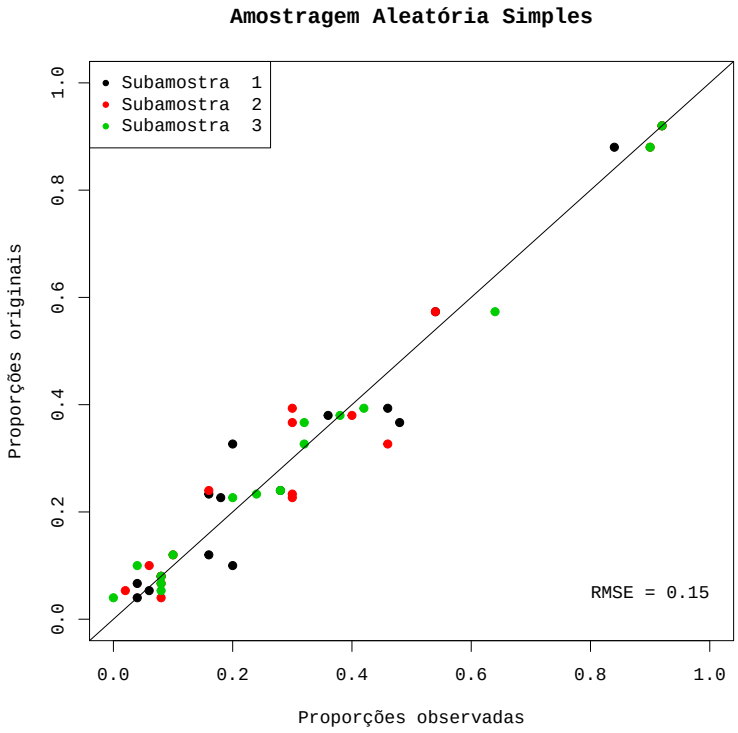

Figura 18: Diagrama de dispersão - AAS vs Proporções reais - amostra de tamanho 150 e 5 variáveis em 3 subgrupos

Em termos do RMSE, o sorteio simples apresentou valores da ordem de 0,15, sendo que a metodologia proposta neste trabalho apresentou valores três vezes inferiores, da ordem de 0,05. Em suma, para este cenário, o método proposto apresentou resultados bastante superiores se comparados com o método padrão.

A análise dos três cenários apresentados indicou uma superioridade da nova metodologia se comparada à amostragem aleatória simples, método tradicional de

definição de subgrupos, que não leva em consideração as variáveis que caracterizam os indivíduos. Entretanto, uma análise de poucos casos pode não ser conclusiva e nem permite a identificação de caraterísticas que possibilitem uma compreensão dos fatores que influenciam na precisão do método.

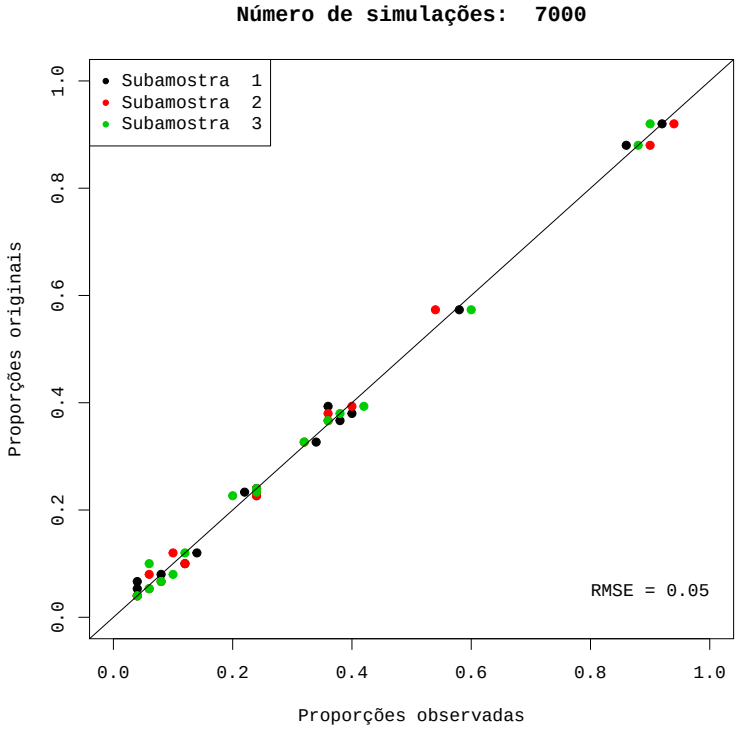

Figura 19: Diagrama de dispersão - Método simulado vs Proporções reais - amostra de tamanho 150 e 5 variáveis em 3 subgrupos

Em seguida, uma análise detalhada de todos os cenários estudados será realizada e analisados fatores que podem influenciar no desempenho do método.

4.4 Avaliação em função das características

Para uma melhor exploração dos resultados comparados, bem como dos fatores que podem afetar a precisão do método em estudo, foram analisados conjuntamente

todos os 36 casos estudados, considerando os fatores *número de subamostras*, *tamanho da amostra original* e *número de variáveis explicativas*. Foram observados os valores de RMSE obtidos em cada cenário pelos métodos concorrentes e sua eficiência comparada via métodos gráficos.

A Tabela 5 apresenta os valores de RMSE obtidos, organizados pelas características das simulações. Em todos os casos analisados, o novo método apresentou redução do RMSE se comparado aos obtidos na amostragem aleatória simples. A menor redução foi de 34,24% com o cenário de 4 grupos, 100 elementos e 6 variáveis. A maior redução foi de 83,78% com o cenário de de 2 grupos, 150 elementos e 3 variáveis. A redução média foi de 57,14%, indicando que, de fato, a nova metodologia apresenta melhoria significativa da homogeneidade da distribuição das subamostras em relação às proporções originais.

De acordo com o boxplot apresentado na Figura 20, o novo método possui menor RMSE para todos os tamanhos de amostra, ou seja, mais precisa é a distribuição das subamostras. Observa-se também que o tamanho da amostra influencia diretamente na precisão dos métodos, pois quanto maior a amostra original, menores os valores de RMSE observados.

Assim como a Figura 20, o novo método apresenta melhores resultados que a amostragem aleatória simples, conforme se verifica na Figura 21, se comparadas as distribuições com o mesmo número de subamostras. Observa-se que quanto menor o número de subamostras a ser gerada, melhor a precisão do método.

Já em termos do número de variáveis, de acordo com a Figura 22, na amostragem aleatória simples não há influência de tal fator na precisão. Entretanto, na nova metodologia, a precisão aparenta ser ligeiramente afetada pelo número de variáveis explicativas. Novamente observa-se que o patamar do RMSE é inferior no novo método, reforçando as conclusões obtidas até aqui.

Após todos os experimentos realizados, é possível afirmar que para todos os casos utilizados o método proposto constitui uma alternativa superior à metodo-

Tabela 5: Precisão dos métodos de distribuição de subamostras

Grupos	Tam. amostra	Variáveis	RMSE		
			Amostragem aleatória	Método Simulado	Redução
2	50	3	0,126	0,033	74,21%
		4	0,070	0,030	58,04%
		5	0,069	0,040	42,20%
		6	0,065	0,038	42,27%
	100	3	0,044	0,011	75,75%
		4	0,063	0,016	74,18%
		5	0,070	0,021	69,54%
		6	0,068	0,027	60,51%
	150	3	0,055	0,009	83,78%
		4	0,059	0,014	76,43%
		5	0,065	0,017	73,77%
		6	0,076	0,024	67,66%
3	50	3	0,250	0,061	75,68%
		4	0,190	0,094	50,76%
		5	0,184	0,098	46,63%
		6	0,171	0,106	38,11%
	100	3	0,138	0,039	71,52%
		4	0,103	0,053	48,67%
		5	0,127	0,063	50,51%
		6	0,126	0,079	37,51%
	150	3	0,101	0,034	65,85%
		4	0,148	0,048	67,63%
		5	0,150	0,050	66,80%
		6	0,144	0,058	59,36%
4	50	3	0,367	0,146	60,18%
		4	0,292	0,175	40,21%
		5	0,337	0,213	36,61%
		6	0,316	0,208	34,24%
	100	3	0,276	0,081	70,65%
		4	0,248	0,106	57,27%
		5	0,239	0,142	40,48%
		6	0,259	0,154	40,49%
	150	3	0,231	0,075	67,34%
		4	0,203	0,107	47,28%
		5	0,204	0,117	42,55%
		6	0,228	0,131	42,65%

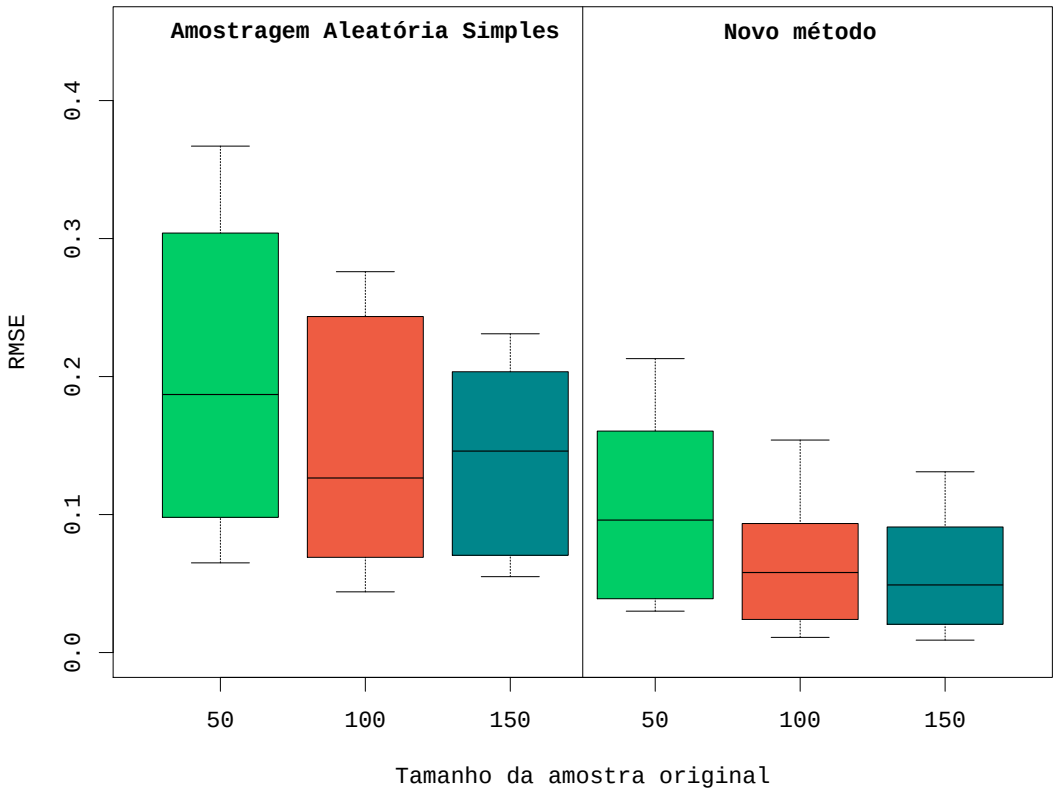

Figura 20: Boxplot: RMSE pelo tamanho da amostra

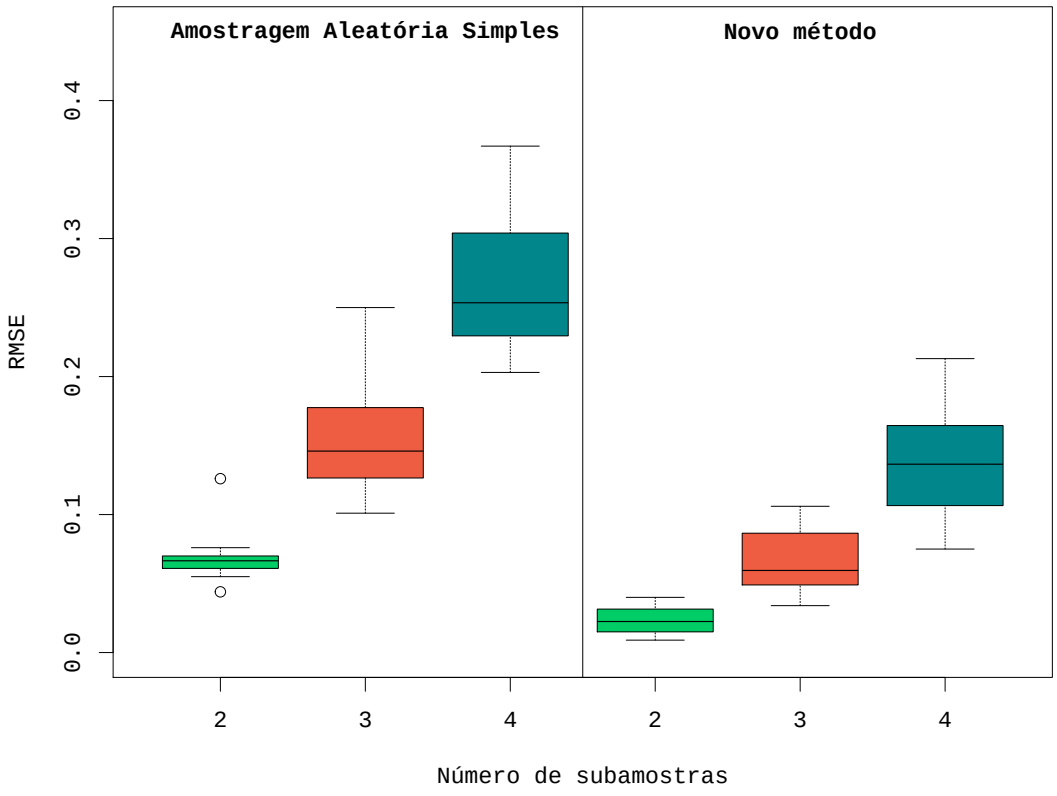

Figura 21: Boxplot: RMSE pelo número de subgrupos

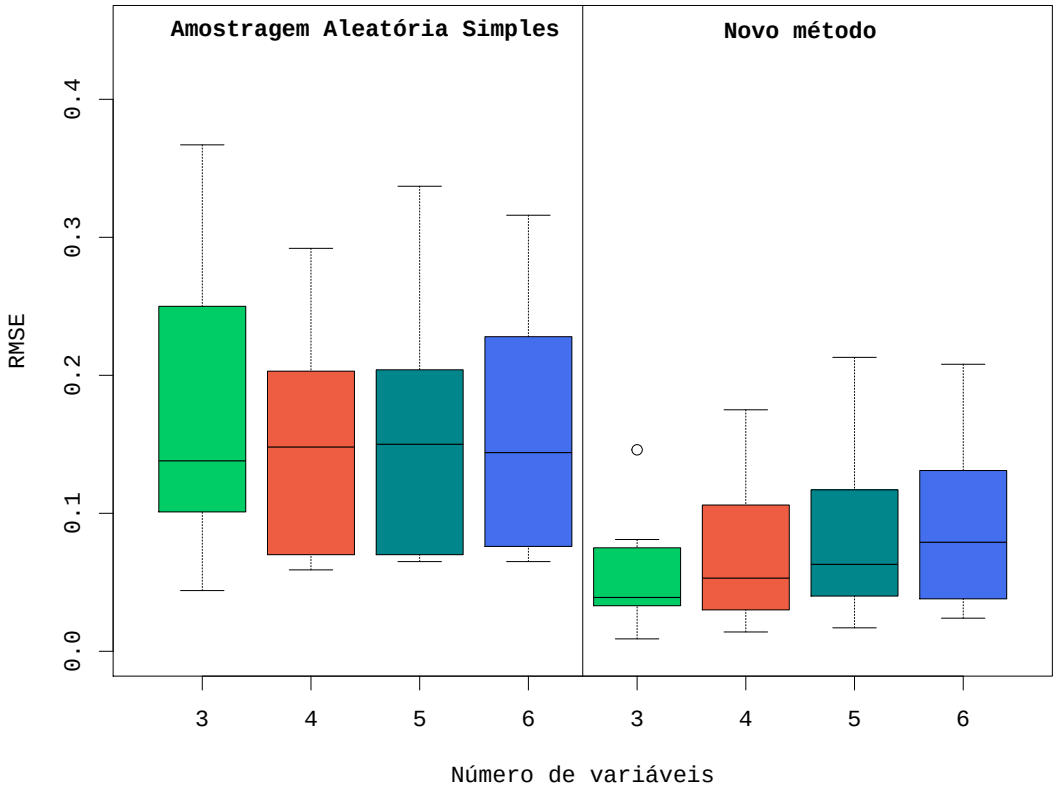

Figura 22: Boxplot: RMSE pelo número de variáveis

,

logia padrão que é aplicada em estudos experimentais no que se refere à divisão de subamostras. Tal método apresenta uma redução significativa da raiz do erro quadrático médio, redução esta que indica uma maior precisão entre as subamostras geradas e que pode garantir uma maior representatividade de cada uma delas em relação à amostra original completa, o que impacta diretamente na redução de viés causado pela ausência de controle sobre as variáveis explicativas.

5 Conclusões e Observações Finais

5.1 Sumário

Este trabalho apresentou inicialmente uma revisão dos principais estudos experimentais empregados na área da saúde, que são potenciais metodologias que poderão usufruir da metodologia desenvolvida. Após a revisão, foi proposta uma metodologia que permite a divisão de uma amostra completa em subamostras heterogêneas dentro e homogêneas entre si, com base em covariáveis discretas que descrevem características populacionais.

Proposto o método, foram realizadas uma série de experimentos computacionais que comprovam a eficácia do algoritmo proposto em produzir subamostras altamente similares à amostra principal, em termos das proporções de cada nível de covariável presentes na amostra original, sendo que o método apresentou reduções de até 83% da raiz do erro quadrático médio em comparação com a abordagem tradicional, amostragem aleatória simples.

Verificou-se que o algorítimo apresenta desempenho computacional bastante satisfatório em termos de tempo computacional, sendo que em todos os cenários estudados, na máquina utilizada, foram necessários em média, tempos inferiores a 15 segundos para a execução de 7 mil simulações, número estipulado para estabilização dos resultados.

Por fim, observou-se os fatores que influenciam a qualidade do agrupamento.

Verifica-se que o método se torna mais eficaz com o aumento do tamanho da amostra original. Este também apresenta resultados superiores quando se divide a amostra original em um menor número de subamostras, apresentando resultados excelentes na divisão em 2 grupos. Por fim, a qualidade das respostas é afetada negativamente pelo acréscimo do número de covariáveis do banco de dados. Entretanto, percebe-se que o impacto não é tão agressivo se comparado com os demais fatores.

Dito isto, a metodologia proposta se apresenta como alternativa viável e eficaz para uma divisão de amostras em subgrupos semelhantes, técnica esta que pode agregar bastante valor a estudos experimentais. A garantia de amostras semelhantes é fundamental para a redução de viés de seleção e eliminar fatores de confundimento e não controláveis, que podem comprometer totalmente a qualidade dos resultados obtidos por um estudo experimental.

5.2 Propostas de Continuidade

Como trabalhos futuros, podem-se citar os seguintes:

- Aplicação da metodologia de divisão de subamostras em bancos de dados contínuos com base em discretização dos dados;
- Utilização de novas técnicas de otimização para obtenção de melhores resultados;
- Utilização da técnica em estudos experimentais.
- Comparação de resultados de intervenções com subamostras geradas via sorteio aleatória e via nova metodologia com base em simulações.

Referências Bibliográficas

ALTMAN, D. G. *Practical statistics for medical research*. [S.l.]: CRC press, 1990.

AMORIM, M. d.; SANTOS, L. C. Tratamento da vaginose bacteriana com gel vaginal de aroeira (schinus terebinthifolius raddi): ensaio clínico randomizado. *RBGO*, SciELO Brasil, v. 25, n. 2, 2003.

ARAGÃO, J. Introdução aos estudos quantitativos utilizados em pesquisas científicas. *Revista práxis*, v. 3, n. 6, 2013.

BASTOS, J. L. D.; DUQUIA, R. P. Um dos delineamentos mais empregados em epidemiologia: estudo transversal. *Scientia Medica*, v. 17, n. 4, p. 229–232, 2007.

BAUGHMAN, R. P. et al. Clinical characteristics of patients in a case control study of sarcoidosis. *American journal of respiratory and critical care medicine*, Am Thoracic Soc, v. 164, n. 10, p. 1885–1889, 2001.

BLOCH, K. V.; MELO, A. N. d.; NOGUEIRA, A. R. Prevalence of anti-hypertensive treatment adherence in patients with resistant hypertension and validation of three indirect methods for assessing treatment adherence. *Cadernos de saude publica*, SciELO Brasil, v. 24, n. 12, p. 2979–2984, 2008.

BONITA, R.; BEAGLEHOLE, R.; KJELLSTRÖM, T. *Epidemiologia básica*. [S.l.]: OPS, 2008.

CHRESTANI, M. A. D. et al. Assistência à gestação e ao parto: resultados de dois estudos transversais em áreas pobres das regiões norte e nordeste do brasil. *Cadernos de Saúde Pública*, SciELO Public Health, v. 24, n. 7, p. 1609–1618, 2008.

COUTINHO, E. S. F.; CUNHA, G. M. da. Conceitos básicos de epidemiologia e estatística para a leitura de ensaios clínicos controlados basic concepts in epidemiology and statistics for reading controlled clinical trials. *Rev Brasileira de Psiquiatria*, SciELO Brasil, v. 27, n. 2, p. 146–151, 2005.

DIAS, A. C.; ARAÚJO, M. R.; LARANJEIRA, R. Evolução do consumo de crack em coorte com histórico de tratamento. *Revista de Saúde Pública*, SciELO Brasil, v. 45, n. 5, p. 938–948, 2011.

FONTELLES, M. *Bioestatística aplicada à pesquisa experimental*. [S.l.: s.n.], 2012.

FUKUDA, V. O. et al. Eficácia a curto prazo do laser de baixa intensidade em pacientes com osteoartrite do joelho: ensaio clínico aleatório, placebo-controlado e duplo-cego. *Rev Bras Ortop*, v. 46, n. 5, p. 526–33, 2011.

HAIR, J. F. et al. *Análise multivariada de dados*. [S.l.]: Bookman Editora, 2009.

LACHIN, J. M. Statistical properties of randomization in clinical trials. *Controlled clinical trials*, Elsevier, v. 9, n. 4, p. 289–311, 1988.

LOFFREDO, L. d. C. M. et al. Fissuras lábio-palatais: estudo caso-controle. *Revista de Saúde Pública*, v. 28, n. 3, p. 213–217, 1994.

LUSTOSA, L. P. et al. Efeito de um programa de resistência muscular na capacidade funcional e na força muscular dos extensores do joelho em idosas pré-frágeis da comunidade: ensaio clínico aleatorizado do tipo crossover. *Rev Bras Fisioter*, SciELO Brasil, v. 15, n. 4, p. 318–24, 2011.

MACIEL, E. L. N. et al. Fatores associados ao abandono da quimioprofilaxia de tb no município de vitória (es): um estudo de coorte histórica. *Jornal Brasileiro de Pneumologia*, Jornal Brasileiro de Pneumologia, v. 35, n. 9, p. 884–891, 2009.

MARINHO, M. S.; CHAVES, P. d. M.; TARABAL, T. d. O. Dupla-tarefa na doença de parkinson: uma revisão sistemática de ensaios clínicos aleatorizados. *Rev. bras. geriatr. gerontol*, v. 17, n. 1, p. 191–199, 2014.

MARINHO, M. S. et al. Efeitos do tai chi chuan na incidência de quedas, no medo de cair e no equilíbrio em idosos: uma revisão sistemática de ensaios clínicos aleatorizados. *Rev. bras. geriatr. gerontol*, v. 10, n. 2, p. 243–256, 2007.

MARTINEZ, E. Z. Metanálise de ensaios clínicos controlados aleatorizados: aspectos quantitativos. *Medicina (Ribeirao Preto. Online)*, v. 40, n. 2, p. 223–235, 2007.

MENDES, M. J. F. d. L. et al. Associação de fatores de risco para doenças cardiovasculares em adolescentes e seus pais. *Rev. bras. saúde matern. infant*, p. s49–s54, 2006.

MONTGOMERY, D. C. Analysis of experiments. *New York: Tohn Wiley and Sons*, v. 1, p. 976, 2001.

MONTGOMERY, D. C.; RUNGER, G. C. *Applied statistics and probability for engineers*. [S.l.]: John Wiley & Sons, 2010.

MOREIRA, N. F. et al. Obesidade: principal fator de risco para hipertensão arterial sistêmica em adolescentes brasileiros participantes de um estudo de coorte. *Arq Bras Endocrinol Metab*, v. 57, n. 7, p. 520–6, 2013.

NIOBEY, F. M. L. et al. Fatores de risco para morte por pneumonia em menores de um ano em uma região metropolitana do sudeste do brasil. um estudo tipo caso-controle. *Revista de Saúde Pública*, v. 26, n. 4, 1992.

OLIVEIRA, G. G. d. *Ensaios clínicos: princípios e prática*. [S.l.]: Anvisa; Sobravime, 2006.

PALACIO, E. P.; CANDELORO, B. M.; LOPES, A. d. A. Lesões nos jogadores de futebol profissional do marília atlético clube: estudo de coorte histórico do campeonato brasileiro de 2003 a 2005. *Rev. bras. med. esporte*, p. 31–35, 2009.

PEREIRA, A. F. A.; MESQUITA, A.; GOMES, C. Abordagens cirúrgicas no tratamento de varizes. *Angiologia e Cirurgia Vascular*, Elsevier, v. 10, n. 3, p. 132–140, 2014.

RODRIGUES, A. M. et al. Factors associated with treatmet failure of cutaneous leishmaniasis with meglumine antimoniate. *Revista da Sociedade Brasileira de Medicina Tropical*, SciELO Brasil, v. 39, n. 2, p. 139–145, 2006.

ROSENBERGER, W. F.; LACHIN, J. M. *Randomization in clinical trials: theory and practice*. [S.l.]: John Wiley & Sons, 2015.

SILVA, V. F. d. et al. Fatores associados à ideação suicida na comunidade: um estudo de caso-controle. *Cadernos de Saúde Pública*, Escola Nacional de Saúde Pública Sergio Arouca, Fundação Oswaldo Cruz, 2006.

SITTA, E. et al. A contribuição de estudos transversais na área da linguagem com enfoque em afasia. *Rev. CEFAC*, SciELO Brasil, v. 12, n. 6, p. 1059–66, 2010.

SURESH, K. An overview of randomization techniques: an unbiased assessment of outcome in clinical research. *Journal of human reproductive sciences*, Medknow Publications & Media Pvt. Ltd., v. 4, n. 1, p. 8, 2011.

VANDERLEI, L. C. d. M.; SILVA, G. A. P. d.; BRAGA, J. U. Fatores de risco para internamento por diarréia aguda em menores de dois anos: estudo de caso-controle. *Cad. saúde pública*, v. 19, n. 2, p. 455–463, 2003.

VAZ, D. et al. Métodos de aleatorização em ensaios clínicos. *Revista portuguesa de cardiologia*, Sociedade Portuguesa de Cardiologia, v. 23, n. 5, p. 741–755, 2004.