ISSN: 2317-0840

Um novo método para alocação de unidade em subamostras representativas baseado em covariáveis discretas

Rosielle C. Farias[1†], Helgem S. R. Martins[2], Ivair R. Silva[3], Graziela D. R. Gouvêa[4], Raimundo M. Nascimento Neto[5], Fernando L. P. Oliveira[6]

[1] *Universidade Federal de Viçosa.*
[2] *Universidade Federal de Viçosa. E-mail:* helgem.souza@gmail.com.
[3] *Universidade Federal de Ouro Preto. E-mail:* ivairest@gmail.com.
[4] *Universidade Federal de Ouro Preto. E-mail:* gragouvea@gmail.com.
[5] *Universidade Federal de Ouro Preto. E-mail:* raimundo.neto@ufop.edu.br.
[6] *Universidade Federal de Ouro Preto. E-mail:* fernandoluizest@gmail.com.

Resumo: *Em estudos experimentais nos quais se deseja verificar a eficácia de alguma intervenção, é usual a presença de grupos que sofrerão ou não estas intervenções para que comparações a respeito de fatores relacionados a estas intervenções possam ser medidos. Para garantir que tais comparações sejam válidas, é necessário que os grupos apresentem características o mais semelhantes possíveis entre si, definidas no início do estudo. Este trabalho apresenta uma nova metodologia de divisão, dado um conjunto de dados inicial, em* k *subamostras representativas em relação aos dados iniciais, com base em covariáveis que definem as características desta. Os resultados obtidos constatam que a metodologia de aleatorização proposta apresenta resultados satisfatórios, principalmente se comparados com a técnica tradicional de amostragem aleatória simples. As subamostras delineadas pelo método apresentam um alto grau de similaridade com a amostra original, o que possibilitará aos estudos experimentais deste trabalho uma redução no viés de seleção, proporcionando resultados mais satisfatórios.*

Palavras-chave: Estudos experimentais; alocação de subamostras; amostragem; aleatorização.

Abstract: *In experimental studies, like clinic trials, where one wants to verify the efficacy 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.*

Keywords: Experimental studies; sub-sample allocation; sampling; randomization.

[†]Autora correspondente: rosi-elle@hotmail.com.

Introdução

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.

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.

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 difícil realizar a seleção dos grupos com base em técnicas de planejamento de experimentos. Com base neste tipo de situação, 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.

Organização do trabalho

Na seção , será realizada uma revisão bibliográfica sobre metodologias de pesquisa na área da saúde e técnicas de aleatorização. Na seção , 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, enquanto a seção , será destinada à análise e validação do método de aleatorização, sendo seu desempenho comparado com a metodologia clássica de aleatorização via sorteio. Para concluir, na seção , serão apresentadas as conclusões e observações finais.

Metodologias de Pesquisa na Área da Saúde

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 desta seção é realizar uma breve revisão sobre estudos experimentais, com ênfase nos ensais clínicos aleatorizados, principal foco da técnica desenvolvida neste trabalho, além de explorar um dos requisitos fundamentais à boa condução de um experimento estatístico, a aleatorização.

Estudos Experimentais

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.

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 determina-

dos 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.

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.

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 introduzir 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.

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.

Material e Métodos

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.

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:

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

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:

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

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 1 apresenta o pseudo-código do método. 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.

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

algoritmo
leia o banco de dados
/ escolha de subgrupo inicial*/*
escolha S_0
/ cálculo da distância entre vetores de proporções */*
calcule $D_0 = \sum_{i=1}^{k} \sqrt{(p_{0i} - p_{obs})'(p_{0i} - p_{obs})}$
$\mathbf{D} \leftarrow \mathbf{D_0}$
$\mathbf{S_{opt}} \leftarrow \mathbf{S_0}$
para $j = 1$ faça
/ escolha de novo subgrupo */*
escolha S_j
/ cálculo da distância entre vetores de proporções */*
calcule $D_j = \sum_{i=1}^{k} \sqrt{(p_{ji} - p_{obs})'(p_{ji} - p_{obs})}$
se $D_j < D$ então
$S_{opt} \leftarrow S_j$
senão
mantenha S_{opt}
fim se
enquanto $j <= nsim$ ou $D = 0$
fim para
imprima S_{opt}
fim algoritmo

Figura 1: Algoritmo de seleção de subgrupos

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, a métrica de ajuste utilizada foi o RMSE (*root mean squared error*), que possui a 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.

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^o do parecer 596.741-0/2013), bem como o estudo de validação (n^o 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, (n^o do parecer CAAE: 39682014.7.0000.5150).

Resultados e Discussão

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.

Desempenho computacional

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 , 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 , 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.

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.

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

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

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

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

Número	Tamanho	Nº de	Melhoria percentual da distância euclidiana inicial							
de grupos	da amostra	variáveis	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	*	*	*	*

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

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

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

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

no processo.

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.

A seguir será apresentado um dos cenários estudados, composto por uma amostra de tamanho 150 e 5 variáveis em 3 subgrupos. As análises realizadas neste foram replicadas nos demais cenários, cuja análise geral dos resultados será apresentada na sequência. A composição de cada cenário foi descrita na seção 4.2.

Na Figura 2 , observa-se que 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.

Neste caso, porém, as subamostras se apresentam mais homogêneas para a amostragem

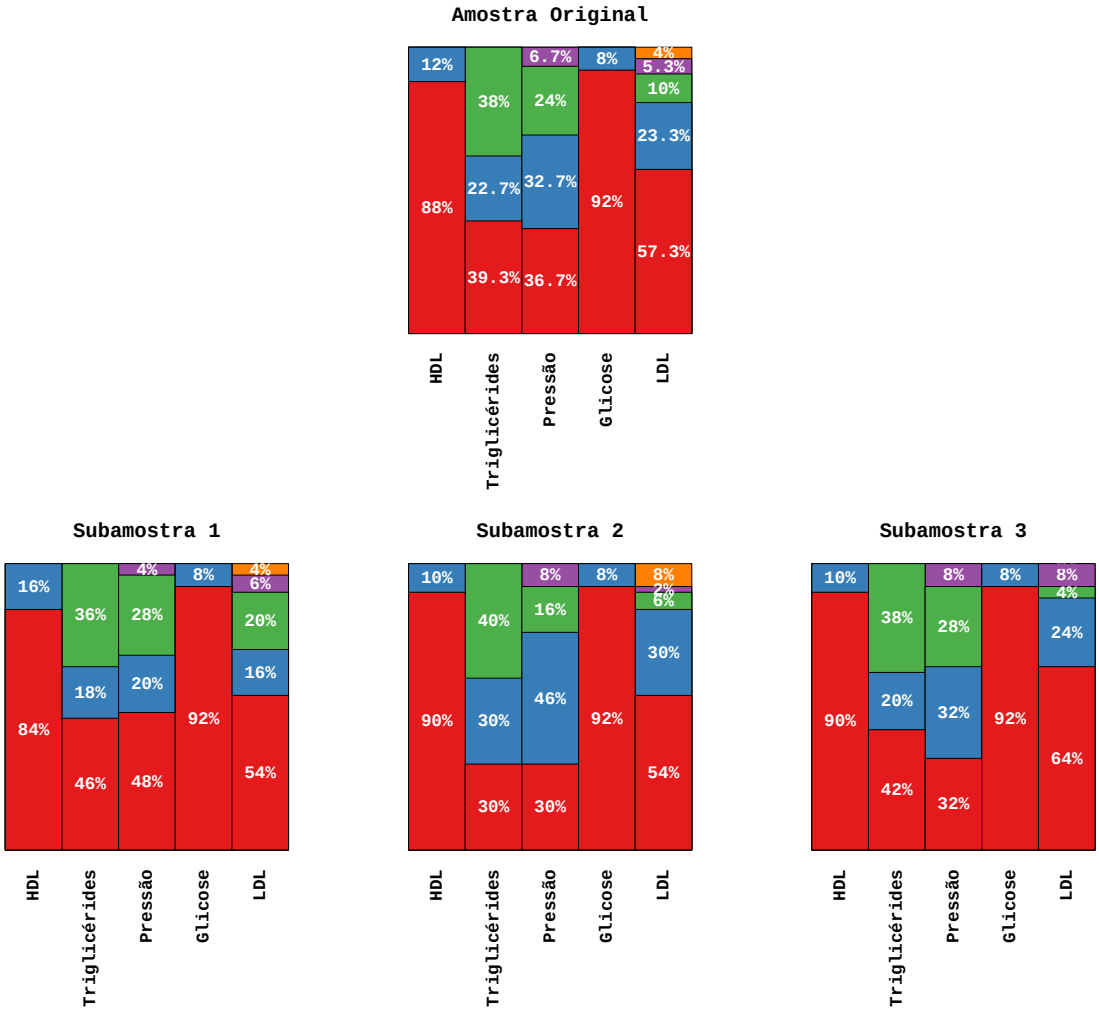

Figura 2: Distribuição dos subgrupos por AAS para amostra de tamanho 150 e 5 variáveis em 3 subgrupos

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 3) apresentou um resultado bastante satisfatório, pois todos os níveis de variáveis foram representados nas subamostras extraídas.

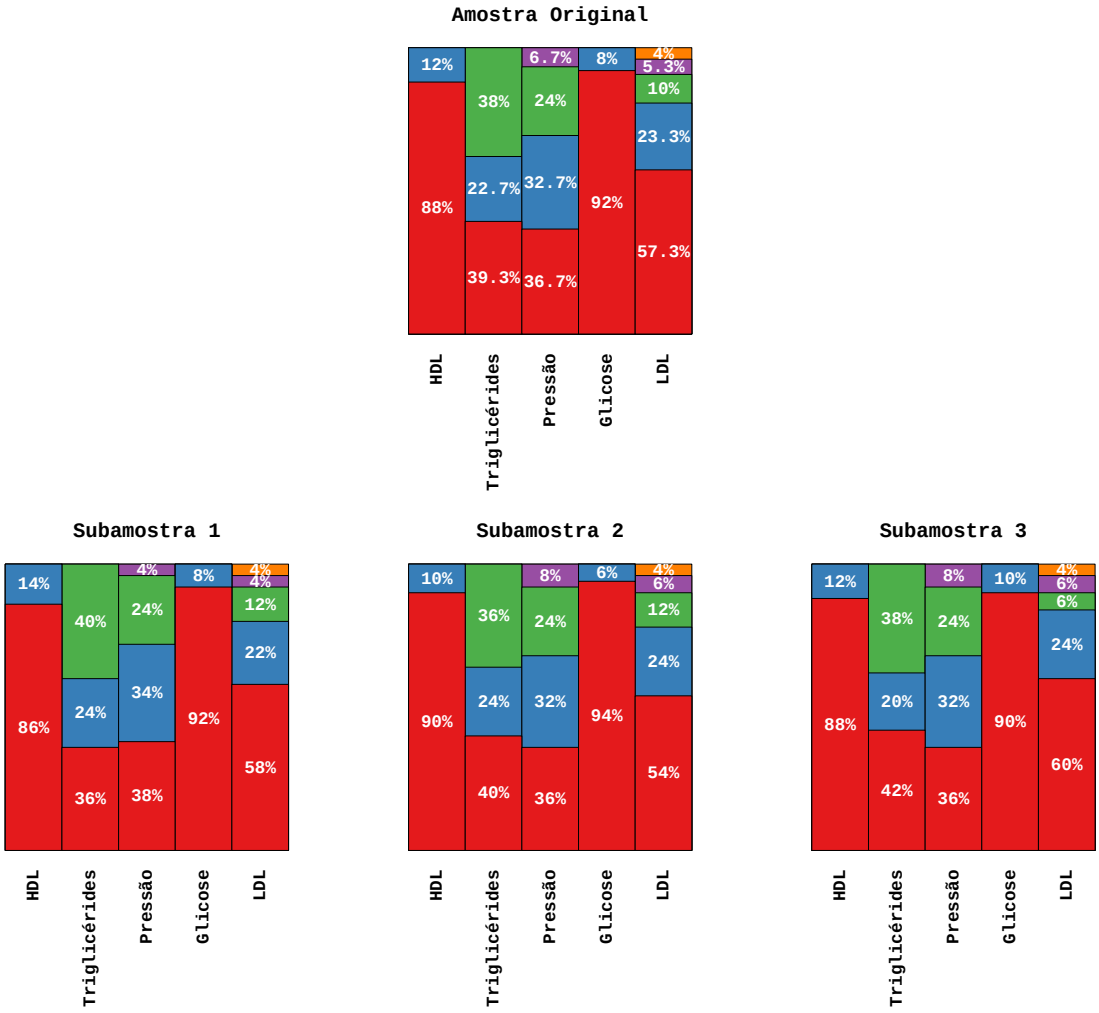

Figura 3: Distribuição dos subgrupos pelo novo método para amostra de tamanho 150 e 5 variáveis em 3 subgrupos

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 das amostras obtidas via AAS, se comparadas ao novo método com 7 mil simulações.

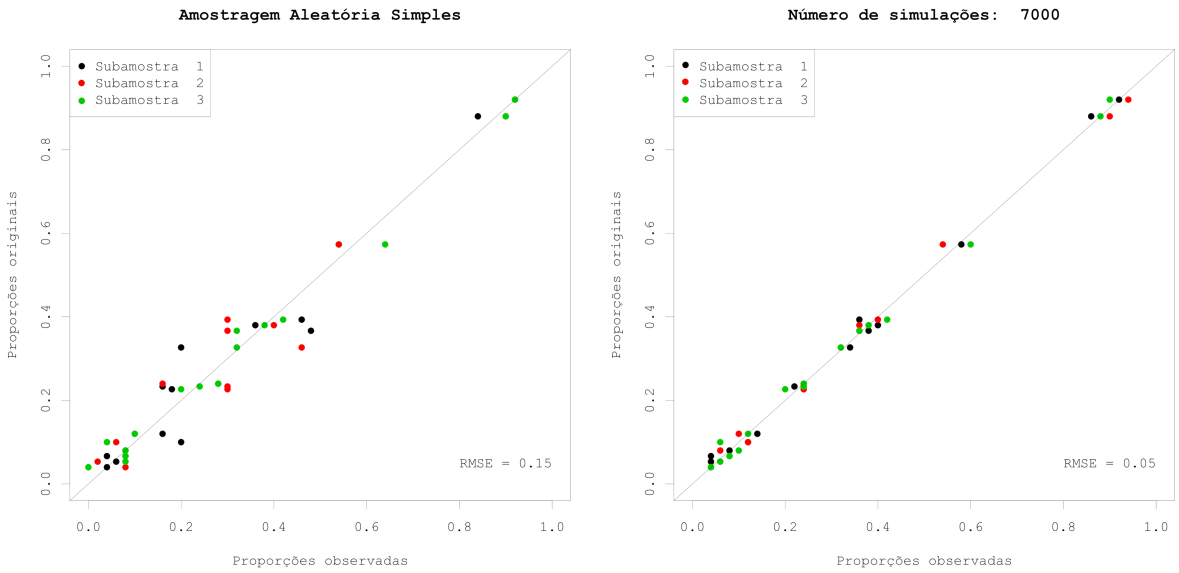

Figura 4: Diagrama de dispersão - AAS vs Proporções reais - amostra de tamanho 150 e 5 variáveis em 3 subgrupos

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 do cenário apresentado, bem como as demais não apresentadas neste texto, 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.

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

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

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%

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 a tabela 5, 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. Observa-se também que o novo método apresenta melhores resultados que a amostragem aleatória simples, 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, 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 metodologia 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.

Conclusões e Observações Finais

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.

Realizou-se 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 foram necessários em média, tempos inferiores a 15 segundos para a execução de 7 mil simulações, número estipulado para estabilização dos resultados.

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, de acordo com este estudo, se apresenta como alternativa viável e eficaz para uma divisão de amostras em subgrupos semelhantes. 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.

Agradecimentos

Esta pesquisa foi parcialmente financiada pelas agências brasileiras CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico), grant 300825/2016-1, FAPEMIG (Fundação de Amparo à Pesquisa do Estado de Minas Gerais), grant PPM-00321-18 e CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior). Agradecemos aos Projetos Prevenção da Fadiga e CUME.

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*, v.25, n.2, 2003.

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

FUKUDA, V. O.; FUKUDA, T.Y.; GUIMARÃES, M.; SHIWA, S.; DE LIMA, B.D.C.; LOPES-MARTINS, A.B.; CASAROTTO, R.A.; ALFREDO, P.P.; BJORDAL, J.M.; FUCS, P.M.M.B. Eficácia a curto prazo do laser de baixa intensidade em pacientes com osteoartrite do joelho: ensaio clínico aleatório, placebo-controlado e duplo-cego. *Revista Brasileira de Ortopedia*, v.46, n.5, p.526-33, 2011.

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

LUSTOSA, L.P.; SILVA, J.P.; COELHO, F.M.; PEREIRA, D.S.; PARENTONI, A.N.; PEREIRA, L.S.M. Efeito de um programa de resistência muscular na capacidade funcional e na força muscular dos extensores do joelho em idosas pré-fráges da comunidade: ensaio clínico aleatorizado do tipo crossover. *Revista Brasileira de Fisioterapia*, v.15, n.4, p.318-24, 2011.

MARINHO, M.S.; CHAVES, P.d.M.; TARABAL, T.d.O. Dupla-tarefa na doença de parkinson: uma revisão sistemática de ensaios clínicos aleatorizados. *Revista Brasileira de Geriatria e Gerontologia*, v.17, n.1, p.191-199, 2014.

MARINHO, M. S.; DA SILVA, J.F.; PEREIRA, L.S.M.; SALMELA, L.F.T. 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. *Revista Brasileira de Geriatria e Gerontologia*, v.10, n.2, p.243-256, 2007.

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

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.

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

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

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

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.

VAZ, D.; SANTOS, L.; MACHADO, M.; VAZ, A.C. Métodos de aleatorização em ensaios clínicos. *Revista portuguesa de cardiologia*, Sociedade Portuguesa de Cardiologia, v. 23, n.5, p.741-755, 2004.