INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DE SÃO PAULO

Danilo Bento Lessa	SP3071715
Felipe Domingues Bonfim	SP3071227
Lucas Nogueira De Souza	SP3072703

Projeto de Estatística e Probabilidade - Análise exploratória de dados

SÃO PAULO

2023

INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DE SÃO PAULO

Danilo Bento Lessa	SP3071715
Felipe Domingues Bonfim	SP3071227
Lucas Nogueira De Souza	SP3072703

Projeto de Estatística e Probabilidade - Análise exploratória de dados

Atividade apresentada ao curso de Análise e Desenvolvimento de Sistemas do Instituto Federal de Educação, Ciência e Tecnologia de São Paulo.

Orientadora Prof.ª Dra. Josceli M. Tenorio

SÃO PAULO

2023

Sumário

1 Introdução

Este projeto tem como objetivo realizar uma análise exploratória dos dados de homicídio no Brasil, comparando o período pré-pandemia de COVID-19 (2018-2019) com o período durante a pandemia (2020-2021). Os dados de mortalidade utilizados foram extraídos das bases de dados do Sistema de Informação sobre Mortalidade(SIM) disponibilizadas pelo Ministério da Saúde através do portal openDataSUS.

O Sistema de Informação sobre Mortalidade (SIM) foi desenvolvido pelo Ministério da Saúde em 1975 com o objetivo de unificar mais de quarenta modelos de Declaração de Óbito utilizados anteriormente. Essa unificação permitiu a coleta de dados sobre mortalidade de forma padronizada em todo o país. O SIM desempenha um papel fundamental na obtenção de informações precisas e consistentes sobre as causas de morte no Brasil, auxiliando na formulação de políticas públicas de saúde e no monitoramento epidemiológico.

A análise foi conduzida com o objetivo de verificar se houve uma diminuição estatisticamente significativa nos casos de homicídios durante a pandemia (2020-2021) em comparação com os anos anteriores (2018-2019). Além disso, buscou-se correlacionar os resultados obtidos com fontes externas de informação confiáveis, como notícias baseadas em dados e bases complementares, a fim de verificar a consistência das conclusões obtidas. Essa abordagem permite uma análise mais precisa e embasada sobre o impacto da pandemia nos índices de homicídio no Brasil.

1.1 Descrição do problema

Durante os anos de 2019 e 2020, o Brasil enfrentou um período de transição complexo em relação à segurança pública, com um impacto significativo da pandemia de COVID-19 na criminalidade e nos índices de homicídios. Essa crise sanitária teve efeitos diversos nas diferentes regiões do país, resultando em variações na quantidade de homicídios e em suas taxas.

Em 2019, o Brasil apresentava uma taxa de homicídios de aproximadamente 27,8 por 100.000 habitantes, de acordo com dados do Atlas da Violência 2021,

produzido pelo Instituto de Pesquisa Econômica Aplicada (IPEA) em parceria com o Fórum Brasileiro de Segurança Pública. Com estes números, o Brasil ocupava a posição de número 10 no ranking mundial e a segunda colocação na América do Sul no ranking de países com maior taxa de homicídio em 2019, segundo o Relatório Global de Homicídios 2019 do Escritório das Nações Unidas sobre Drogas e Crime (UNODC).

Entretanto, apesar dos altos números apresentados, dados coletados pelo projeto Monitor da Violência, realizado pelo Núcleo de Estudos da Violência da USP, o G1 e o Fórum Brasileiro de Segurança Pública, indica que a taxa de homicídios no Brasil vem decrescendo de forma consistente desde 2017, como aponta a Figura 1 abaixo:

Figura 1 – Homicídios no Brasil (Fonte: Monitor da Violência, 2021)

1.2 Proposta da análise

Este projeto tem como objetivo realizar uma análise exploratória dos dados coletados em relação à taxa de homicídio no Brasil, em cada uma das cinco regiões: Norte, Nordeste, Centro-Oeste, Sudeste e Sul. A proposta deste estudo é confirmar se há redução efetiva na taxa que não tenha sido causada por coincidência,

realizando a verificação através da análise das bases de dados públicas do DataSUS do período entre 2018 e 2021.

1.3 Descrição da base de dados

Os arquivos utilizados como base de dados para este projeto foram disponibilizados pelo Ministério da Saúde através do openDataSUS. As bases de dados utilizadas são denominadas Mortalidade Geral 2018, Mortalidade Geral 2019, Mortalidade Geral 2020 e Mortalidade 2021. Cada arquivo contém informações sobre os óbitos registrados nos respectivos anos. O arquivo Mortalidade Geral 2018 possui aproximadamente 438MB de dados, o Mortalidade Geral 2019 possui aproximadamente 449MB, o Mortalidade Geral 2020 possui aproximadamente 519MB e o Mortalidade 2021 possui 617MB de dados.

Os dados brutos estão estruturados em tabelas com 1.048.575 registros e 87 colunas cada. Essas tabelas fornecem informações detalhadas sobre cada óbito registrado, incluindo características como tipo de morte, contagem de homicídios, contagem total de mortes no Brasil, data do óbito, nome da região e idade da pessoa.

Para o propósito desta análise, os dados foram tratados e separados em grupos de amostras contendo 200.000 registros por ano. Essas amostras foram agrupadas em duas categorias: "Pré-pandemia de COVID-19" e "Durante a pandemia de COVID-19". O grupo "Pré-pandemia" consiste nas amostras coletadas nos anos de 2018 e 2019, enquanto as amostras coletadas nos dois anos subsequentes foram alocadas no grupo que representa o período durante a pandemia.

As variáveis selecionadas para a análise incluem o tipo de morte, a contagem de homicídios, a contagem total de mortes no Brasil, a data do óbito, o nome da região(Norte, Nordeste, Centro-oeste, Sudeste e Sul) e a idade da pessoa.

2 Metodologia

Inicialmente, os dados foram agrupados em duas amostras distintas divididos em pré-pandemia e durante a pandemia, período considerado pré-pandêmico engloba os anos 2018 e 2019, já o período durante a pandemia considerou os dados dos anos 2020 e 2021. Essa divisão tem o objetivo de analisar possíveis efeitos causados pela pandemia de COVID-19 na taxa de homicídio.

Utilizando a linguagem R, foram criadas funções para transformar e adequar os dados às necessidades da análise. As informações relacionadas à região geográfica, mês de óbito e outras variáveis relevantes foram tratadas e organizadas para facilitar a visualização e compreensão dos resultados.

Foram geradas tabelas agregando o número total de mortes e homicídios em cada região, permitindo uma visão geral dos dados. Além disso, gráficos foram criados para ilustrar os homicídios totais ao longo do período analisado, evidenciando possíveis tendências ou variações significativas.

A análise também inclui a geração de intervalos de confiança para cada região, o que possibilita avaliar a precisão das estimativas. Além disso, foram realizados testes de hipótese para verificar se há diferenças estatisticamente significantes na taxa de homicídio entre as regiões e no Brasil como um todo.

Todo o processo de análise, incluindo tabelas e gráficos gerados, é demonstrado com detalhes na seção 3 deste documento.

3 Análises

A análise exploratória de dados desempenha um papel crucial na compreensão e interpretação dos dados coletados em um estudo ou pesquisa. Nesta seção, é realizado um estudo dos dados obtidos, visando identificar padrões, tendências e informações relevantes que contribuam para a compreensão do fenômeno em estudo. Nesse contexto, identificamos as variáveis utilizadas, fizemos um levantamento de informações relevantes, realizamos uma análise descritiva e exploramos proporções de mortes por homicídio, além de conduzir testes de hipótese para investigar possíveis relações ou diferenças significativas.

3.1 Variáveis utilizadas

Conforme citado na Introdução, cada arquivo no conjunto de dados contém 87 variáveis. Para a análise realizada, foram utilizados os seguintes campos do conjunto de dados:

- *CODMUNRES:* Código do município de residência. Em caso de óbito fetal, considerar o município de residência da mãe. (Números).
- *IDADE:* Idade do falecido em minutos, horas, dias, meses ou anos. (Idade: composto de dois subcampos. - O primeiro, de 1 dígito, indica a unidade da idade (se 1 = minuto, se 2 = hora, se 3 = mês, se 4 = ano, se = 5 idade maior que 100 anos). - O segundo, de dois dígitos, indica a quantidade de unidades: Idade menor de 1 hora: subcampo varia de 01 e 59 (minutos); De 1 a 23 Horas: subcampo varia de 01 a 23 (horas); De 24 horas e 29 dias: subcampo varia de 01 a 29 (dias); De 1 a menos de 12 meses completos: subcampo varia de 01 a 11 (meses); Anos - subcampo varia de 00 a 99; - 9 - ignorado).
- *DTOBITO*: Data em que occoreu o óbito.(Data no padrão ddmmaaaa).
- *CIRCOBITO*: Tipo de morte violenta ou circunstâncias em que se deu a morte não natural. (1 – acidente; 2 – suicídio; 3 – homicídio; 4 – outros; 9 – ignorado).

- *ASSISTMED*: Se refere ao atendimento médico continuado que o paciente recebeu, ou não, durante a enfermidade que ocasionou o óbito. (1 – sim; 2 – não; 9 – ignorado).

3.2 Levantamento de informações relevantes

Nesta etapa, foram realizados procedimentos de pré-processamento e análise de dados para extrair informações relevantes a partir do conjunto de dados fornecido.

Em primeiro lugar, foram selecionadas amostras aleatórias de 200.000 observações para cada ano (2018, 2019, 2020 e 2021) a partir do conjunto de dados completo. Essa amostragem facilitou a análise, reduzindo o tamanho dos conjuntos de dados.

Em seguida, foram aplicados tratamentos nos dados das amostras. Uma transformação foi feita na coluna que representa o código do município, atribuindo a cada código a sua região geográfica correspondente. Além disso, a coluna que indica a idade foi processada para calcular a idade em anos completos.

Posteriormente, as datas de óbito foram formatadas corretamente, convertendo-as para o formato de data adequado. Adicionalmente, o mês de cada óbito foi extraído a partir das datas convertidas e armazenado em uma coluna separada.

Por fim, foram gerados novos quadros de dados que contém as mesmas informações, uma no período pré-pandemia e outro no período durante a pandemia. Esses quadros (*data frames)* contém, para cada região do Brasil, seu nome, seu total de mortes, seu total de mortes não tratadas e o total de mortes por homicídio.

Essas etapas de seleção de amostragens, tratamento de dados e separação de novos quadros de dados agrupados por região permitiram extrair informações relevantes da base de dados de mortalidade geral. Essas informações são fundamentais para análises posteriores e para entender características e padrões nos períodos pré-pandemia e durante a pandemia.

3.3 Análise Descritiva

Realizamos uma análise descritiva dos dados de mortalidade geral, com foco na contagem de homicídios nos períodos pré-pandemia e durante a pandemia. A análise descritiva nos permite ter uma visão geral das ocorrências de homicídios nas diferentes regiões do Brasil.

A Figura 2 apresenta o gráfico de barras que representa a contagem mediana de homicídios nas regiões do Brasil durante os anos de 2018 e 2019, ou seja, o período pré-pandemia.

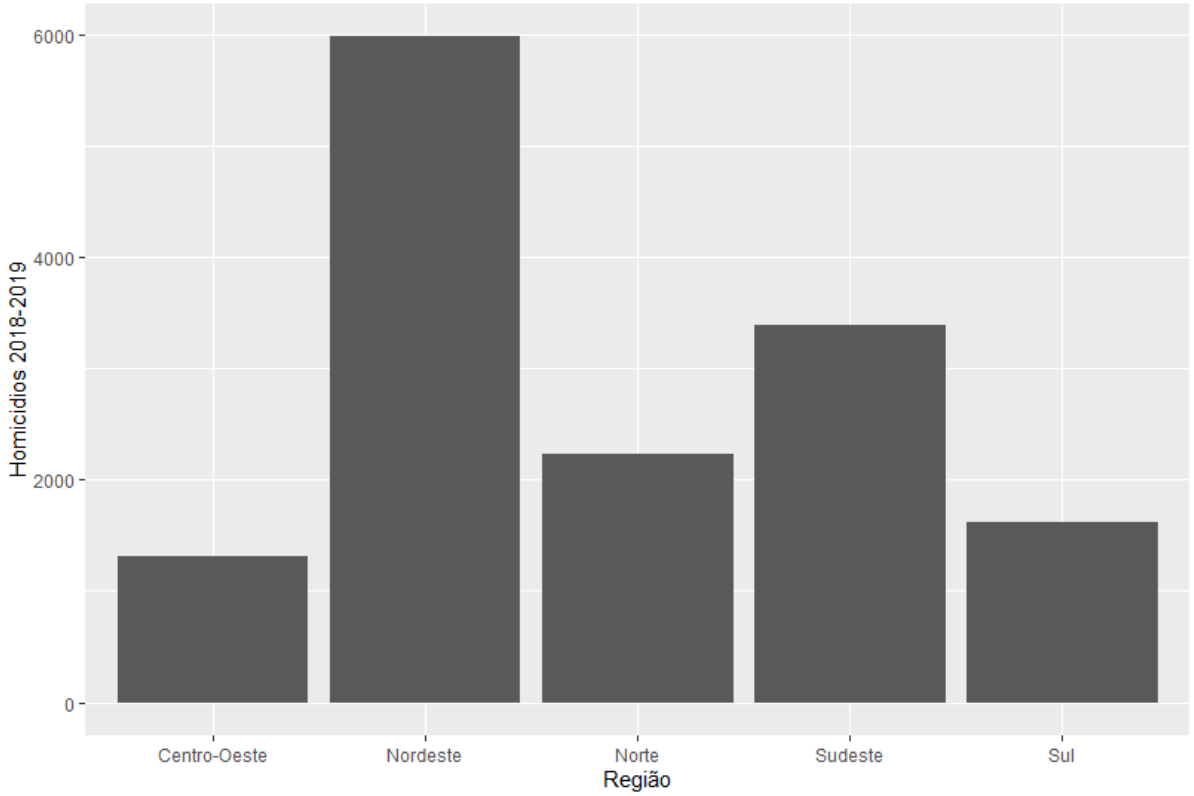

Figura 2 – Homicídios no período pré-pandemia (Fonte: Os autores)

Já a Figura 3 apresenta o gráfico de barras que representa a contagem mediana de homicídios nas regiões do Brasil durante os anos de 2020 e 2021, ou seja, o período durante a pandemia.

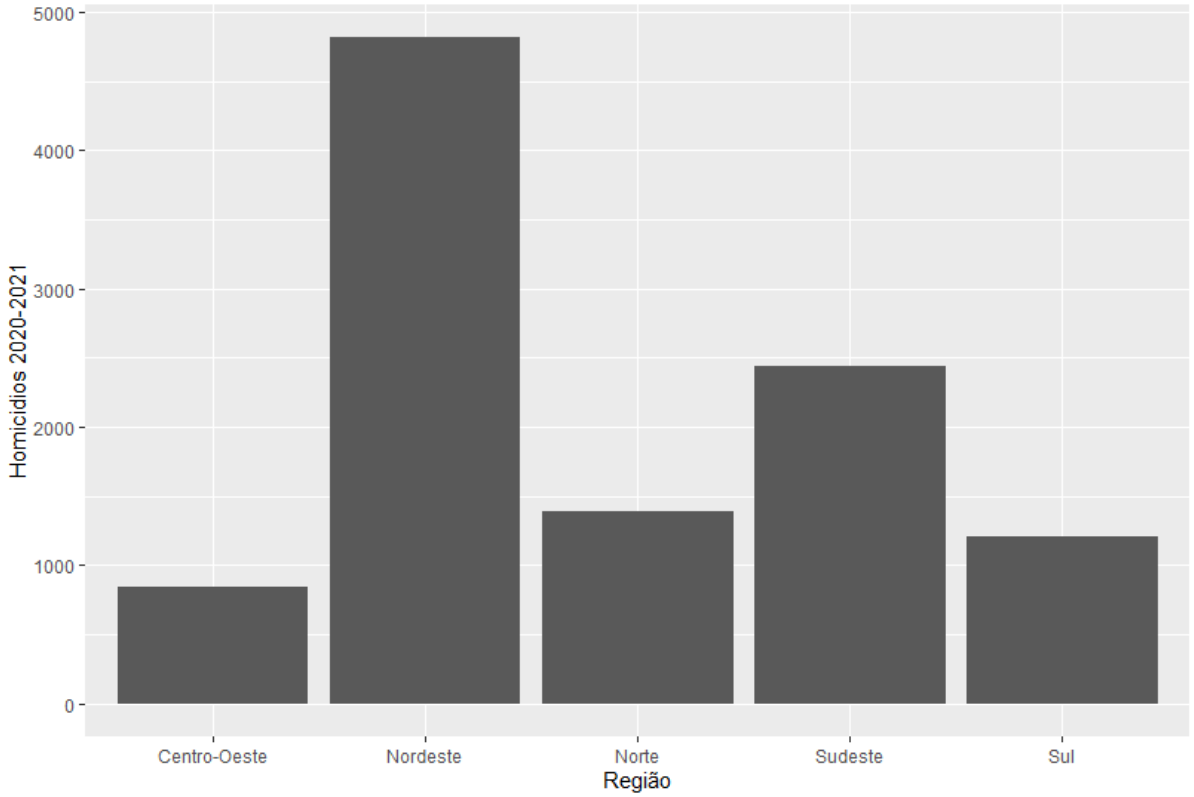

Figura 3 – Homicídios no período durante a pandemia (Fonte: Os autores)

Nas Figuras 4 e 5, é possível observar a distribuição dos homicídios por região nos períodos pré e durante a pandemia, respectivamente.

Figura 4 – Distribuição de homicídios por região no período pré-pandemia (Fonte: Os autores)

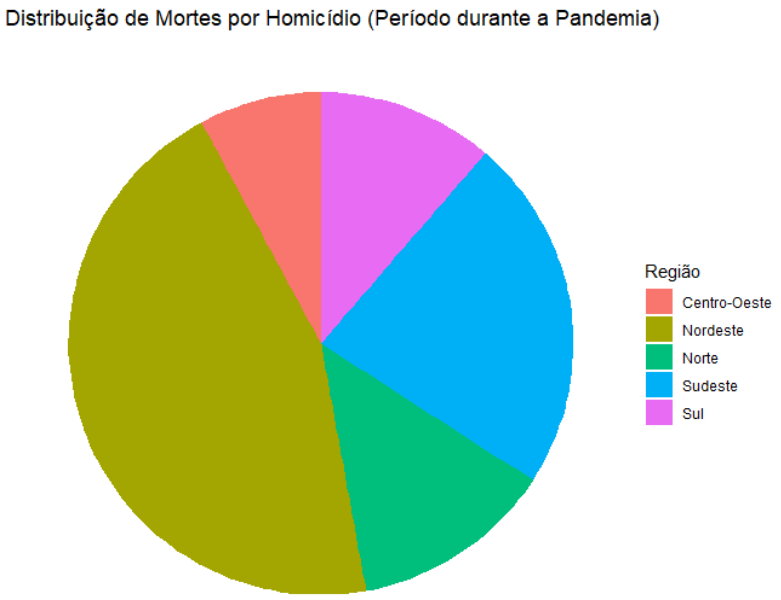

Figura 5 – Distribuição de homicídios por região no período da pandemia (Fonte: Os autores)

Na figura 6, é possível comparar os casos de homicídio antes e durante a pandemia por região através de um gráfico de barras empilhadas.

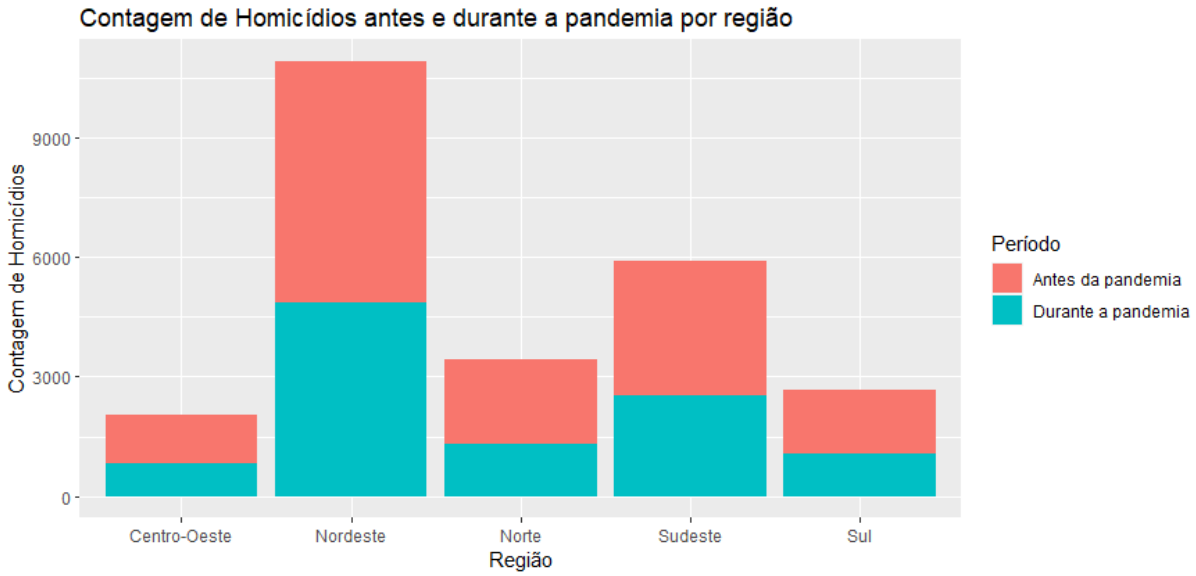

Figura 6 – Gráfico de barras empilhadas comparando os homicídios antes e durante a pandemia por região (Fonte: Os autores)

Com base na análise descritiva dos dados de homicídios nas diferentes regiões do Brasil, podemos concluir que, de forma geral, houve uma diminuição na

quantidade de homicídios durante o período da pandemia em comparação com o período pré-pandemia.

Além disso, as Figuras 4 e 5 mostram a distribuição dos homicídios por região nos períodos pré e durante a pandemia. É possível notar que, em ambos os períodos, a região com a maior contagem de homicídios é o Nordeste, seguido pelo Sudeste, Norte, Sul e Centro-Oeste. No entanto, durante a pandemia, ocorre uma redução na distribuição dos homicídios em todas as regiões.

O gráfico de barras empilhadas na Figura 6 também permite uma comparação visual entre os casos de homicídio antes e durante a pandemia por região. Ele reforça a tendência de redução nos homicídios durante o período da pandemia em todas as regiões do Brasil.

3.4 Proporções de Mortes por Homicídio e Intervalos de Confiança

Ao observar a Figura 7, podemos visualizar a diferença das proporções de mortes por homicídio antes e durante a pandemia, representada por um gráfico de linhas. Cada linha representa uma região específica, e o eixo x indica os períodos "Antes da pandemia" e "Durante a pandemia". As cores das linhas correspondem a cada região, facilitando a comparação visual.

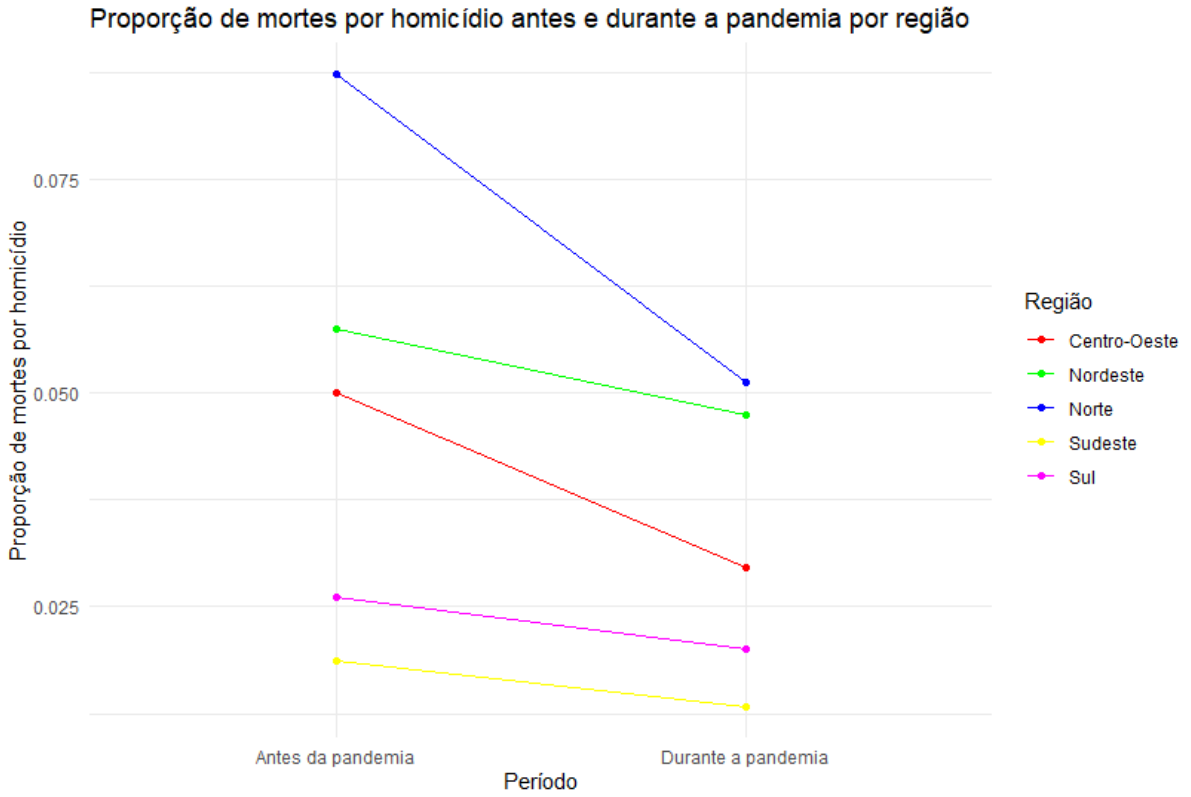

Figura 7 – Comparação de proporção de mortes por homicídio antes e durante a pandemia por região (Fonte: Os autores)

Durante o período pré-pandemia, observamos as seguintes proporções de mortes por homicídio nas regiões: Região Norte (0,084 a 0,090), Região Nordeste (0,056 a 0,059), Região Centro-Oeste (0,048 a 0,052), Região Sudeste (0,018 a 0,019) e Região Sul (0,025 a 0,027). Esses intervalos de confiança, com um nível de confiança de 90%, indicam a faixa em que a proporção real de homicídios provavelmente se encontra.

Por exemplo, para a Região Norte, a proporção de mortes por homicídio variou de 0,084 a 0,090, enquanto que para a Região Nordeste, a proporção ficou entre 0,056 a 0,059. Esses intervalos fornecem uma estimativa da incerteza associada às proporções.

Já durante a pandemia, as proporções de mortes por homicídio estimadas e os intervalos de confiança para cada região foram os seguintes: Região Norte (0,049 a 0,053), Região Nordeste (0,046 a 0,048), Região Centro-Oeste (0,028 a 0,031), Região Sudeste (0,013 a 0,014) e Região Sul (0,019 a 0,021). Com um nível de confiança de 90%, podemos afirmar que a proporção real de homicídios em cada região durante a pandemia está contida dentro dessas faixas. Por exemplo, para a Região Norte, a proporção estimada de homicídios durante a pandemia variou de 0,049 a 0,053, enquanto que para a Região Nordeste, a proporção ficou entre 0,046 a 0,048. Novamente, esses intervalos de confiança com um nível de confiança de 90% indicam a faixa em que a proporção real de homicídios durante a pandemia provavelmente se encontra.

Outra informação observada é que os intervalos de confiança das proporções de mortes por homicídio antes e durante a pandemia não se sobrepõem em nenhuma das regiões. Isso indica que as diferenças nas proporções de homicídios entre os períodos são estatisticamente significativas. Essa observação sugere que houve uma redução consistente nas proporções de homicídios em todas as regiões durante a pandemia, em comparação ao período pré-pandemia.

Antes da pandemia, as regiões apresentaram diferentes níveis de proporções de mortes por homicídio. A Região Norte teve as maiores taxas, seguida pela Região Nordeste, Região Centro-Oeste, Região Sudeste e, por fim, Região Sul. Essas diferenças regionais podem estar relacionadas a fatores socioeconômicos, culturais e estruturais específicos de cada região.

Durante a pandemia, a tendência de redução nas proporções de mortes por homicídio também foi observada em todas as regiões. No entanto, as diferenças entre as regiões persistiram. A Região Norte continuou apresentando as maiores taxas, seguida pela Região Nordeste, Região Centro-Oeste, Região Sudeste e, por fim, Região Sul.

3.5 Testes de Hipótese

Além dos intervalos de confiança, realizamos um teste de hipótese para investigar se há uma diferença significativa nas proporções de mortes por homicídio entre os dois períodos. Utilizamos um teste de proporções com a hipótese alternativa de que a proporção de mortes por homicídio durante a pandemia é maior do que antes da pandemia.

Na Figura 8, apresentamos os resultados do teste de hipótese de forma geral, ou seja, considerando todas as regiões. O valor-p indica a probabilidade de obter os resultados observados ou mais extremos, considerando-se a hipótese nula de que não há diferença nas proporções de mortes por homicídio entre os períodos. Quanto menor o valor-p, mais evidências temos contra a hipótese nula.

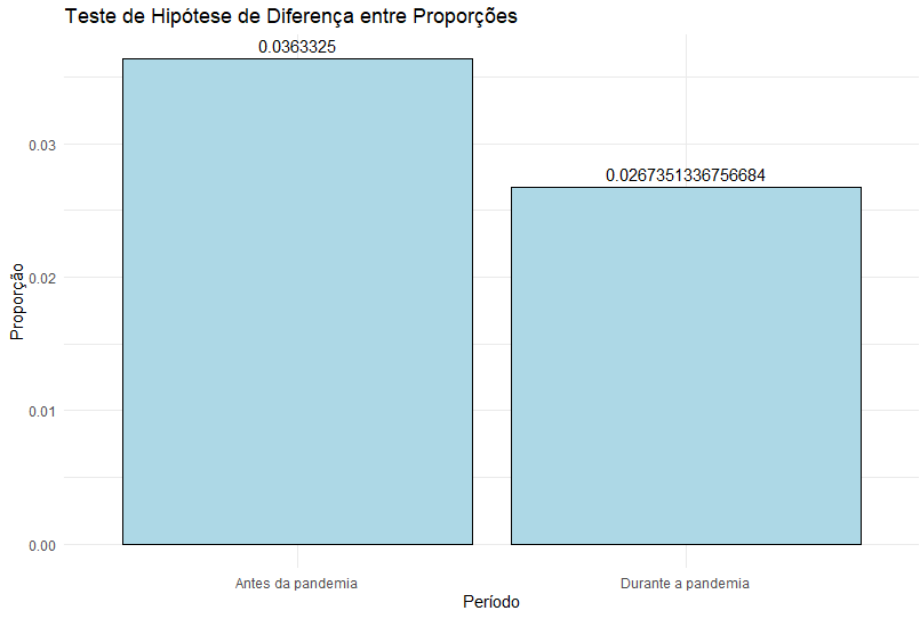

Figura 8 –Teste de hipótese de diferença entre proporções (Fonte: Os autores)

Com base nos resultados apresentados, podemos notar que há fortes evidências contra a hipótese nula de que não há diferença nas proporções de mortes por homicídio entre os dois períodos (antes da pandemia e durante a pandemia). O valor-p é extremamente baixo em todas as regiões, indicando que a probabilidade de obter os resultados observados ou mais extremos, considerando-se a hipótese nula, é muito pequena.

Portanto, com base nessas informações, é observado que há uma diferença significativa nas proporções de mortes por homicídio entre os dois períodos, com um aumento durante a pandemia em comparação com o período anterior.

4 Conclusão

Com base na análise realizada, foi possível confirmar uma redução estatisticamente significativa na taxa de homicídios em todas as cinco regiões do Brasil. Esses resultados respaldam a premissa inicial de que houve uma diminuição dos índices de violência letal no país. No entanto, é importante ressaltar que, com base nas informações analisadas, não é possível atribuir exclusivamente o impacto da pandemia de COVID-19 como o único responsável por essas reduções.

Diversos fatores devem ser considerados para uma compreensão mais abrangente dessas mudanças. Entre eles, destacam-se as políticas de segurança pública implementadas pelas autoridades em cada região, bem como a influência do crime organizado e do tráfico de drogas. Além disso, fatores socioeconômicos, como níveis de desigualdade, pobreza e acesso a oportunidades, também desempenham um papel importante na incidência de homicídios.

Embora a pandemia possa ter contribuído indiretamente para a redução da taxa de homicídios, através de medidas de isolamento social e restrições à circulação, não é possível afirmar categoricamente que esse foi o único fator determinante. A análise estatística indica uma tendência geral de queda nos índices de violência letal, mas é fundamental considerar a interação complexa entre esses diversos elementos e seu impacto na segurança pública.

5 Referência bibliográfica

United Nations Office on Drugs and Crime (UNODC). **Global Study on Homicide**. [Online]. Disponível em: <https://www.unodc.org/documents/data-and-analysis/gsh/Booklet1.pdf>. Acesso em: 17 jun. 2023.

Instituto de Pesquisa Econômica Aplicada (IPEA) e Fórum Brasileiro de Segurança Pública. (2021). **Atlas da Violência 2021**. [Online]. Disponível em: <https://www.ipea.gov.br/atlasviolencia/arquivos/artigos/5141-atlasdaviolencia2021completo.pdf>. Acesso em: 17 jun. 2023.

Monitor da Violência. **As mortes violentas mês a mês no país** [Online]. Disponível em: <http://especiais.g1.globo.com/monitor-da-violencia/2018/mortes-violentas-no-brasil/#/dados-mensais-2022>. Acesso em: 17 de junho de 2023.