ANÁLISE EXPLORATÓRIA DOS REGISTROS TOTAIS MENSAIS DE HOMICÍDIO DOLOSO NO ESTADO DE SANTA CATARINA NO PERÍODO DE JANEIRO DE 2010 A DEZEMBRO DE 2018

Dewes, Vinícios Munari[1]

2613810

PADILHA, Eliandro José [2]

RESUMO

A quantidade de pessoas assassinadas sempre foi motivo preocupação por parte dos governantes. A definição formal de homicídio doloso engloba o ato de matar alguém e mais alguns desdobramentos, descritos no Código Penal, artigo 121. Conforme a definição legal, os registros do crime de homicídio doloso serão contabilizados aos meses e usados numa análise exploratória. Selecionados do site da Secretaria de Segurança Pública de Santa Catarina (SSP/SC, os registros foram avaliados quando as características básicas (medidas de posição, dispersão, assimetria e achatamento); verificação de valores atípicos; testados quando a presença de tendência e sazonalidade determinística e por fim verificada a aderência da amostra a distribuição normal de probabilidades. Foi verificado que a dispersão dos dados não é linear, não existe tendência significativa, mas foi confirmada a sazonalidade determinística. O box-plot evidenciou apenas um valor *atípico* que foi corrigido, sendo realizada então uma análise das estatísticas básicas e dos gráficos que indicaram aderência a distribuição normal de probabilidade, sendo confirmado pelos valores do teste de Lilliefors.

Palavras-chave: Homicídio Doloso. Testes de Hipótese. Análise de Dados.

1 INTRODUÇÃO

As situações de convivência não pacífica que resultam em mortes violentas recebem atenção especial do legislador brasileiro, que reserva na parte especial, título primeiro do Decreto-Lei 2848, um capítulo aos crimes contra a vida, sendo descrito no artigo 121 o crime de homicídio doloso que pode ser apresentado na forma dolosa ou culposa e já inclui o feminicídio.

---

1 Aluno do Centro Universitário Internacional UNINTER. Artigo apresentado como Trabalho de Conclusão de Curso. 02- 2020.

2 Professor Orientador no Centro Universitário Internacional UNINTER.

A atuação por oito anos como Policial Militar em Santa Catarina possibilitou atuar na análise de dados criminais, contabilizando os registros mensais dos principais crimes e verificando a correspondência dos resultados obtidos com atividades operacionais. Como o Estado de Santa Catarina não dispõe de grandes investimentos nessa área, foi necessário atuar com recursos computacionais limitados, e, através desta dificuldade, desenvolvida uma sequência de aplicação conteúdos que possibilitam a extração de informações, auxiliando na tomada de decisões relacionadas às atividades operacionais.

Acreditando que o crime de homicídio é o que possui a quantidade de registros efetuados como melhor representatividades entre os crimes contra a vida, e, acreditando no poder de extração de informações proporcionado pela ciência estatística, será realizado um estudo exploratório inicial dos dados que possui a finalidade de verificar as características básicas para obtenção de informações sobre o crime de homicídio.

Pode-se sintetizar o problema em estudo com a seguinte pergunta: é possível realizar essa análise exploratória dos registros totais mensais de homicídio doloso de Santa Catarina, obtendo-se informações que possam direcionar a análise conjunta com outras variáveis?

Esta análise tem por objetivo principal responder o questionamento supracitado, avaliando a consistência dos registros totais mensais de homicídio doloso no Estado de Santa Catarina, verificando a possibilidade extração de informações para melhor caracterização do fenômeno. Os objetivos complementares são listados:

- Realizar uma pesquisa bibliográfica elencando as principais técnicas utilizadas nesta análise preliminar de dados;
- Verificar se existem diferenças na média de registros entre os meses;
- Analisar se existe algum valor de variância entre os meses que pode ser considerado diferente dos demais;
- Mensurar o grau de aderência à distribuição normal de probabilidades.

Para assegurar a concretização dos objetivos, será inicialmente realizada uma pesquisa que inclui revisão bibliográfica com aplicação dos conceitos levantados nos registros totais mensais de homicídio doloso no Estado de Santa Catarina, coletados no site da Secretaria de Segurança Pública, que são de acesso público.

2 METODOLOGIA, REVISÃO DE CONCEITOS E APLICAÇÃO DO MÉTODO

Na primeira subseção será apresentada a metodologia. Na segunda serão apresentados os conceitos através da revisão bibliográfica, devidamente referenciada, incluindo comentários próprios do autor que relacionam o conteúdo pesquisado com o desenvolvimento do trabalho. Na terceira subseção serão discutidos os resultados obtidos.

2.1 METODOLOGIA

A consulta dos registros pode se feita através da referência Santa Catarina (2020), sendo a última atualização em 22 de março de 2019. Foram selecionados os registros totais mensais de homicídio doloso do Estado de Santa Catarina, no período de janeiro de 2010 até dezembro de 2018, totalizando 108 registros ordenados pelos meses, conforme observados no Apêndice A, embora o tempo não seja variável neste estudo.

A escolha dos valores totais foi devido à impossibilidade de quantificação exata da população residente, que aumenta muito nos meses de dezembro a março, devido ao intenso turismo na região. Dessa forma, qualquer provável influência que seja estatisticamente significante provinda do aumento populacional na variação deste crime será testada comparando as médias e variâncias entre os meses e os anos. Necessário ressaltar que estes dados constituem uma amostra, pois em algumas ocorrências o crime pode ser encoberto ou nem mesmo ser registrado. Apesar de ser uma variável discreta, a amostra é formada por material contínuo, tendo em vista que o banco de dados que a constitui é alimentado mensalmente.

A realização das análises será feita com uso do software Statistica versão 12.5 e com planilhas eletrônicas que servem de apoio. Inicialmente, serão verificados os valores que se destacam em relação aos demais, como *outliers (ou atípicos)* ou extremos, observando sua distribuição com gráfico de box-plot. A partir daí, através de análise do gráfico de linhas será observada a existência da

linearidade na dispersão dos valores amostrais, ordenados pelo tempo conforme coletados.

A análise de um fator relacionado à tendência e a sazonalidade determinística serão realizadas incialmente com visualização do gráfico de linha e posteriormente complementada com testes não paramétricos, já que nem sempre estas informações ficam bem definidas com a visualização gráfica.

Também será verificado se a amostra é proveniente da distribuição normal de probabilidades, através da análise visual do histograma de frequências, do gráfico pp-plot e dos valores das estatísticas básicas (medidas de posição, dispersão), da assimetria e curtose. Como última análise, será obtida a confirmação da aderência a gaussianidade com uso do teste de Lilliefors. Em todos os testes aplicados, será utilizado o intervalo com $95{,}00\%$ de confiança nos testes realizados.

2.2 REVISÃO DOS CONCEITOS

2.2.1 O crime de Homicídio

O crime de homicídio doloso, tratado como HO quando considerado variável no estudo, é definido no Decreto-Lei 2848 de sete de dezembro de 1940, no artigo 121, descrito simplesmente como “Matar alguém.” (BRASIL, 2020). A composição dos dados que serão analisados incluem os parágrafos agravantes e atenuantes do crime de homicídio doloso, não sendo considerados nesta análise outros resultados.

2.2.2 Medidas de Posição

Costa Neto (1977) e Fávero e Belfiore (2017) definem três medidas de posição para dados ordenados: média aritmética $\overline{x}$, mediana Me e moda Mo. A moda é tida como o valor que mais se repete. Costa Neto (1977) e Fávero e Belfiore (2017) definem a mediana como sendo o valor que divide a amostra de modo que 50% dos valores da amostra estão acima da mediana e os demais abaixo.

Costa Neto (1977) e Fávero e Belfiore (2017) definem o primeiro quartil $1Q$ como o valor que separa a quantidade de $25{,}00\%$ dos elementos inferiores (em valor numérico) aos $75{,}00\%$ superiores. No terceiro quartil $3Q$ divide $75{,}00\%$ dos valores

inferiores aos $25{,}00\%$ superiores. A diferença $3Q - 1Q$ é denominada amplitude interquartil AI. A média aritmética é definida como sendo a soma dos valores da variável dividida pela quantidade de elementos N (COSTA NETO, 1977; BARBETTA, 2017; FÁVERO e BELFIORE, 2017).

Estas três medidas servem para situar inicialmente os dados em um único eixo, além de serem medidas alternativas para avaliar a simetria da distribuição. A experiência na utilização indica que mediana pode ser usada na confecção dos gráficos de box-plot substituindo à média-aritmética em alguns testes quando não tiver sido verificada a aderência a distribuição normal de probabilidades teórica.

2.2.3 Medidas de Dispersão

Mesmo que sejam usadas para repassar uma ideia inicial da posição unidimensional dos valores, as medidas de dispersão podem induzir a erros quando são usadas sem valores complementares. Apresentando maior importância devido a utilização em diversas técnicas, a variância amostral s^2 é definida por Costa Neto (1977), Barbetta (2017) e Fávero e Belfiore (2017) como sendo o valor médio do quadrado da diferença entre os valores amostrais e a média aritmética, sendo vista na Equação 01. Possui como vantagem não apresentar desvios negativos, mas a inconveniência de apresentar medidas quadráticas:

$$s^2 = \frac{\sum_{i=1}^{N}(x_i - \overline{x})^2}{N - 1} (1)$$

Para corrigir o problema de se trabalhar com desvio quadrático proposto pela variância, Costa neto (1977), Barbetta (2017) e Fávero e Belfiore (2017) recomendam o uso do desvio padrão s, que nada mais é do que a raiz quadrada da variância.

Definido por Fávero e Belfiore (2017) o erro-padrão ep é o valor do desvio-padrão dividido pela raiz quadrada do número de elementos, sendo análogo ao valor médio da variância amostral, conforme visto na Equação 02. Fávero e Belfiore (2017) indicam que este valor é mais sensível ao número amostral:

$$ep = \frac{s}{\sqrt{N}} (2)$$

Além de outros usos, este valor é importante porque é necessário para estimativa de um intervalo de confiança da média populacional.

2.2.4 O intervalo de confiança.

O intervalo de confiança é aplicado quando se desconhece a variância populacional. Por intervalo de confiança Meyer (1969) descreve o que se pode entender como um intervalo aleatório associado a um valor probabilístico, com probabilidade $1-\alpha$, sendo α o nível de significância. Quando $\alpha = 0{,}05$, intervalo possui $95{,}00\%$ de confiança, em geral, as verificações consideram que o valor em teste é maior ou igual a outro valor. Para $\alpha/2$ o testes validam hipóteses de que o valor em teste é num intervalo compreendido por outros dois valores, que são opostos.

Os intervalos de confiança não são necessariamente únicos, pois variam conforme a amostra. Se neste estudo dos 108 registros fossem retiradas p amostras aleatórias com $n_j \ll 108$ elementos, existiriam p intervalos prováveis que contivessem a média populacional. Esse recurso pode ser utilizado em situações que pretendem aferir a média populacional.

Meyer (1969), Costa Neto (1977) e Fávero e Belfiore (2017) definem o intervalo que contenha a provável média populacional baseando-se em na média aritmética amostral, erro padrão e distribuição t de Student: $\mu \in \overline{x} \pm t_{\left(n-1;\frac{\alpha}{2}\right)} ep$.

Maiores detalhes sobre o desenvolvimento da distribuição t de Student podem ser obtidos em Meyer (1969), Costa Neto (1977) e Fávero e Belfiore (2017). Para os dados do estudo de caso o valor bicaudal tabelado num intervalo com $95{,}00\%$ de confiança é $t_{108-1;0{,}025} = 1{,}9823$ (STATSOFT, 2014).

2.2.5 A distribuição normal de probabilidades univariada

A distribuição normal de probabilidades está associada a diversos fenômenos, além de servir de aproximação para diversas distribuições de probabilidade (Binomial e Poisson, por exemplo). No âmbito do conjunto dos números reais, uma variável aleatória possui distribuição normal de probabilidades com média μ e variância σ^2 se é descrita pela função de densidade de probabilidade descrita na

Equação 03 (MEYER, 1969; KARMEL E POLASEK, 1974; COSTA NETO, 1977; GLASS e STANLEY, 1996; FÁVERO e BELFIORE, 2017):

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\left(\frac{x_i-\mu}{2\sigma}\right)^2} (3)$$

A distribuição é simétrica com relação à μ, apresentando-se em forma de um sino. Quanto menor o valor de μ, mais concentrada no entrono da média é a curva. A função densidade de probabilidade acumuladas é obtida integrando a Equação 03, no intervalo compreendido entre o menos infinito e o valor desejado (MEYER, 1969; KARMEL E POLASEK, 1974; COSTA NETO, 1977; GLASS e STANLEY, 1996; FÁVERO e BELFIORE, 2017). Costa Neto (1977) complementa indicando que as probabilidades associadas aos pontos $\mu - \sigma$ e $\mu + \sigma$ apresentam a inflexão da curva.

Pode ser apresentada na forma padronizada onde apresenta $\mu = 0$ e $\sigma = 1$. A transformação da distribuição normal para a distribuição normal padronizada é dada pela Equação 04, que fornece o valor padronizado associado à probabilidade que pode ser usado consultando-se tabelas (MEYER, 1969; KARMEL E POLASEK, 1974; COSTA NETO, 1977; FÁVERO e BELFIORE, 2017). Maior rigor no desenvolvimento da distribuição normal pode ser obtido em Meyer (1969). A padronização por si só não aproxima uma variável aleatória da distribuição normal, mas esta transformação é usada em alguns testes e também permite comparações mais “justas” entre variáveis com unidades de medida distintas.

$$z = \frac{x_i - \mu}{\sigma} (4)$$

2.2.6 Medidas de Assimetria

Hair Jr *et al* (2009) recomendam a verificação da assimetria e da curtose em dados cujo interesse é verificar a normalidade. O coeficiente de assimetria usado neste estudo será o coeficiente assimetria de Fischer As, descrito na Equação 05 por Washington, Karlaftis e Mannering (2011) e Statsoft (2014). Estas medidas são usadas para verificar a simetria da distribuição de frequência de uma variável aleatória. Quando são alongadas a esquerda, possuem assimetria negativa e quando alongadas a direita assimetria positiva (COSTA NETO, 1977). Fávero e Belfiore (2017) indicam que para assimetria negativa têm-se $\overline{x} < Md < Mo$. Para assimetrias positivas $Mo < Md < \overline{x}$.

$$As = \frac{N\sum_{i=1}^{N}(x_i - \overline{x})^3}{(N-1)(N-2)(s^3)} (5)$$

2.2.7 Medidas de Achatamento

As medidas de achatamento ou curtose descrevem a distribuição de frequência conforme o achatamento. Costa Neto (1977) ressalta que os valores obtidos com a avaliação da curtose são mais eficientes em distribuições aproximadamente simétricas. O parâmetro usado é o da distribuição normal, que é classificada quanto ao achatamento como mesocúrtica. Então as distribuições mais achatadas que esta são ditas platicúrticas, as mais esbeltas são leptocúrticas. O Coeficiente de curtose de Fischer será usado e é descrito na Equação 06 (WASHINGTON, KARLAFTS e MANNERING, 2011; STATSOFT, 2014).

$$C = \frac{n(n+1)\sum_{i=1}^{n}(x_i - \overline{x})^4}{(n-1)(n-2)(n-3)s^4} - -3$$

2.2.8 Gráficos box-plot e pp-plot

A construção de um box-plot exige pelo menos cinco medias de posição: $Me, 1Q, 3Q, x_{máximo}$e $x_{mínimo}$. Dessa forma, o gráfico fica dividido em quatro partes iguais, possuindo como centro na mediana. O box-plot serve também para identificação de valores *outliers* ou extremos, numa análise univariada. Assim, serão considerados atípicos os valores que estiverem fora do intervalo: $Me \pm kAI$, com k arbitrário (SICSÚ e DANA, 2013; STATSOFT, 2014; FÁVERO e BELFIORE, 2017).

O gráfico pp-plot é usado para verificação de quão bem ajustado esta uma variável aleatória esta se comparada a uma distribuição de probabilidades. Quanto mais em cima de uma linha reta teórica estiverem os pontos, maior é a aderência. Os valores da distribuição normal padronizada são comparados aos valores da distribuição acumulada (STATSOFT, 2014).

Este gráfico é muito útil para dirimir dúvidas sobre a aderência a uma distribuição de probabilidades teóricas plotada num histograma de frequências, pois no p-p plot os valores não estão sujeitos à subjetividade e a manipulação e grande variações podem ser mensuradas.

2.2.9 Teste de Levene para avaliação da heterocedasticidade

O Teste de Levene é realizado para verificação da igualdade das variâncias. Conforme os Autores Almeida, Elian e Nobre (2008) este teste é pouco afetado por possíveis desvios a normalidade. O modelo mais comum para utilização do Teste de Levene é descrito na equação 07. A amostra total é dada por $N = \sum_{i=1}^{k} n_i$ e n_i a amostra de cada variável e k o número de variáveis envolvidas no teste. Z_{ij} é o módulo da diferença entre a observação j da amostra i;X_{ij} menos a média da amostra i $\overline{X_i}$. $\overline{Z_i}$ é a média de $\overline{Z_{ij}}$ na amostra i e $\overline{Z}$ é a média de Z_i na amostra i.

$$W = \frac{N-k}{k-1} \frac{\sum_{i=1}^{k} n_i \left(\overline{Z_i} - \overline{Z}\right)^2}{\sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(Z_{ij} - \overline{Z_i}\right)^2} \quad (07)$$

Almeida, Elian e Nobre (2008) indicam que quando o valor de W calculado é inferior ao crítico aproximado pela distribuição F sendo $W_c = F_{k-1,n-k,\alpha}$ rejeita-se a hipótese de distinção das variâncias. Para 108 casos e nove variáveis, o valor crítico é $W_c = F_{8,99,0,05} = 2{,}033$. Se forem 108 casos e 12 variáveis, $W_c = F_{11,96,0,05} = 1{,}900$ (STATSOFT, 2014). Maiores esclarecimentos sobre a distribuição F podem ser obtidos em Costa Neto (1977).

Almeida, Elian e Nobre (2008) indicam a troca do valor da média pelo valor da mediana, pois em distribuições com indícios de fuga à normalidade (distribuições assimétricas) a mediana tende a ser uma estimativa mais robusta. Neste estudo, para que sejam realizadas as comparações entre os meses e os anos, a variável HO será dividida em 12 ou 9 variáveis com menos elementos.

2.2.10 Teste de Lilliefors

O teste de Lilliefors é uma melhoria no teste de Kolmogorov-Smirnov para testar a aderência de uma variável a uma distribuição de probabilidades teórica, utilizando a média e o desvio da amostra, enquanto que o predecessor verifica a

aderência usando a média e amostra da população, informação que nem sempre está disponível.

O teste consiste basicamente em medir a maior distância entre as frequências empírica e teórica acumuladas, sendo a empírica obtida pela razão$1/N$, que deve ser feita com dados ordenados. A frequência teorica é baseada nos parâmetros da distribuição a ser comparada (LILLIEFORS TEST, 2020). O modelo geral é dado pela Equação 08. Nessa equação $F_{esp}x_i$ representa a frequência esperada na i-ésima categoria é $F_{obs}x_i$ frequência observada na i-ésima categoria.

$$Li = máx\left[\left|F_{esp}x_i - F_{obs}x_i\right|; \left|F_{esp}x_{i-1} - F_{obs}x_{i-1}\right|\right] (08)$$

Se esta distância for menor que o valor crítico, a hipótese inicial que os dados são provenientes da distribuição teórica é aceita, e caso contrário, se aceita a hipótese alternativa de que os dados não provêm desta distribuição (LILLIEFORS TEST, 2020). Para Blain (2014) o uso do polinômio $Li = (-0{,}000085n^2 + 0{,}05645n + 4{,}422)/(n + 13{,}90)$ deve ser usado para estimativa do valor crítico. Num intervalo com 95,00% de confiança o valor crítico é obtido por $KS_c = 0{,}0821$ (BLAIN, 2014).

2.2.11 Teste das Sequências para verificação de componentes de tendência

O teste de verificação de tendência deve ser realizado para complementar a verificação gráfica. Morettin e Toloi (2006) recomendam o teste das sequências para verificação das tendências, mas alertam para o baixo poder de detecção das hipóteses. Para Costa Neto (1977) e Morettin e Toloi (2006) o primeiro passo é definir a mediana, pois ela é necessária para verificação dos valores que estão acima n_1 ou abaixo n_2, a quantidade pontos em cada caso. Então, com os dados ordenados, são anotadas as sequências de dados que os valores estão acima ou abaixo da mediana, lembrando que a sequência pode conter apenas um valor.

Quando n_1 ou n_2 possuem mais de 20 elementos, o valor do teste pode ser aproximado pela distribuição normal de probabilidades, com média $\overline{x}'$ e desvio padrão s' sendo obtidos pelas Equações 09 e 10.

$$\overline{x}' = \frac{2n_1 n_2}{N} + 1(09)$$

$$s' = \sqrt{\frac{2n_1 n_2\left(2n_1 n_2 - (N)\right)}{(N)^2(N-1)}}(10)$$

A verificação pode ser feita pela aproximação da normal padronizada, usando a contagem de sequência, a média e o desvio padrão estimados. Se o valor em módulo obtido for menor que o valor limite no intervalo de confiança específico, a hipótese nula de que não existe tendência é aceita (MORETTIN e TOLOI, 2006). Num intervalo com $95{,}00\%$ o valor crítico é $z = 1{,}959$(STATSOFT, 2014).

2.2.12 Testes de Kruskal-Wallis para sazonalidade determinística

Este teste verifica se as amostras são provenientes de uma mesma população, sendo que são avaliadas três ou mais amostras. Pode ser um teste alternativo a análise de variância quando a normalidade dos dados e igualdade das variâncias não for respeitada. A primeira ação consiste em ordenar todos os valores e anotar os postos, sendo o último posto atribuído ao maior valor. Havendo valores iguais, assume-se o valor do posto médio e o resultado precisa ser corrigido (FÁVERO e BELFIORE, 2017). A estatística já corrigida é obtida pela Equação 11:

$$KW = \frac{\frac{12}{N(n+1)}\sum_{i=1}^{k}\frac{R_j^2}{n_i} - 3(N+1)}{1 - \frac{\sum_{i=1}^{g}(t_i^3 - t_i)}{(N^3 - N)}}(11)$$

Onde k representa o número de grupos, N o número de elementos totais, n_ié o número em cada grupo, R_i é a soma dos postos, g é o número de grupamentos com postos diferentes e t_i é o número de postos empatados. Quando temos $k > 3$, e cada grupo com cinco ou mais elementos o valor de KW pode ser aproximado pela distribuição χ^2 com $k - 1$ graus de liberdade e intervalo de confiança específico (FÁVERO e BELFIORE, 2017). Detalhes sobre a distribuição χ^2 podem ser encontrados em Costa Neto (1977).

Se o valor de KW estimado for maior que o valor crítico a hipótese nula de que as medianas entre os $k$grupos são iguais é descartada e assume-se que existe pelo menos uma mediana é diferente das demais (FÁVERO e BELFIORE, 2017). Se $k - 1 = 9$ então o valor crítico num intervalo com $95{,}00\%$ de confiança é $\chi^2 =$

$15{,}507$. Se $k-1=11$, no mesmo intervalo de confiança, o valor sobre para $\chi^2 = 19{,}675$ (STATSOFT, 2014)

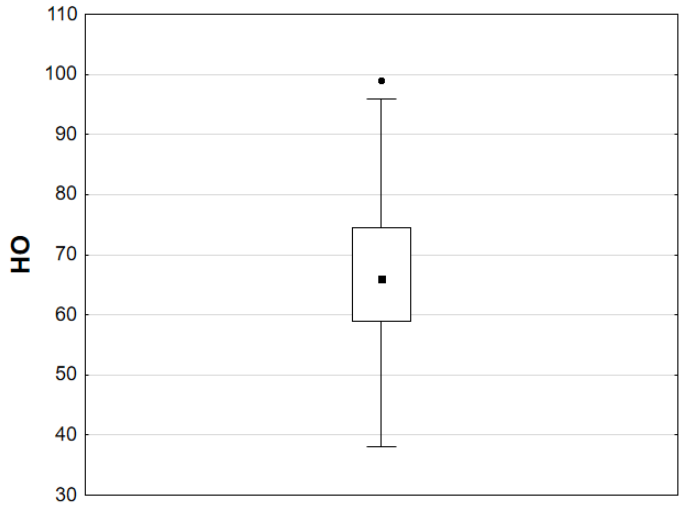
Morettin e Toloi (2004) recomendam este teste para verificação da Sazonalidade, pois avaliam justamente a igualdade estatística da mediana entre os grupos. Para os autores, a eliminação da tendência na série temporal deve ser feita para realizar este teste de sazonalidade. Alertam para o fato de que os elementos em cada amostra são independentes e estas são independentes entre si, fato que nem sempre ocorre. Aqui são feitas as mesmas considerações sobre a divisão das variáveis no teste de Levene.

2.3 APLICAÇÃO DO MÉTODO

Para a obtenção de resultados mais verossímeis, deve-se observar a presença de *outliers*, realizada através do box-plot. Como não existe nenhuma informação sobre aderência a alguma distribuição de probabilidade, a confecção deste gráfico será feita com o valor da mediana e a amplitude interquartil.

O box-plot aparenta boa simetria na distribuição dos dados, com um pouco mais de elementos concentrados abaixo da mediana. Esse é um bom indício de simetria, mas por si só não é um resultado robusto o suficiente para inferências. O valor da mediana é representado pelo ponto central. A amplitude interquartil é delimitada pela caixa que circunda o ponto, pertencente ao intervalo $AI \in (59{,}000; 74{,}500)$. Os valores ditos não atípicos são delimitados pela haste em transversal a linha que corta a caixa, situados no intervalo $(38{,}000; 96{,}000)$, idealizado com $k = 1{,}5$. Nessas condições, visualizou-se apenas o mês de fevereiro de 2017 (case 86), que foi classificado como valor atípico por apresentar 99 registros, mas não foi excluído da análise. O gráfico pode ser visto na Figura 01.

Figura 1: box-plot dos registros de homicídio



Fonte: elaborado pelo autor

Neste ano, disputas de facções criminosas por pontos de venda de drogas a usuários finais ampliou a quantidade de homicídios na faixa litorânea e em alguns municípios mais populosos do interior do Estado de Santa Catarina. Por isso, será testada igualdade dentre as medianas nos anos, com intuito de verificar se algum ano apresentou aumento significativo na mediana dos valores registrados de homicídio .

O motivo para não exclusão do case 86 se dá em função utilização dados em séries temporais, pois esta fica comprometida sem este case. Por isso, o valor atípico de 99 registros será substituído pela média dos meses de fevereiro: $\overline{x}_{fev} = 71{,}56$. Entretanto, se no estudo de caso a ordenação mensal não influencia o resultado, mas para estudos com dados em corte transversal, por exemplo, o case poderia ser descartado. Necessário explicar que, numa análise multivariada e com os dados ordenados pelo tempo, um *case* qualquer engloba um valor de cada variável em estudo, por isso a ênfase na distinção entre a exclusão do valor numérico e a não exclusão do mês de fevereiro de 2017.

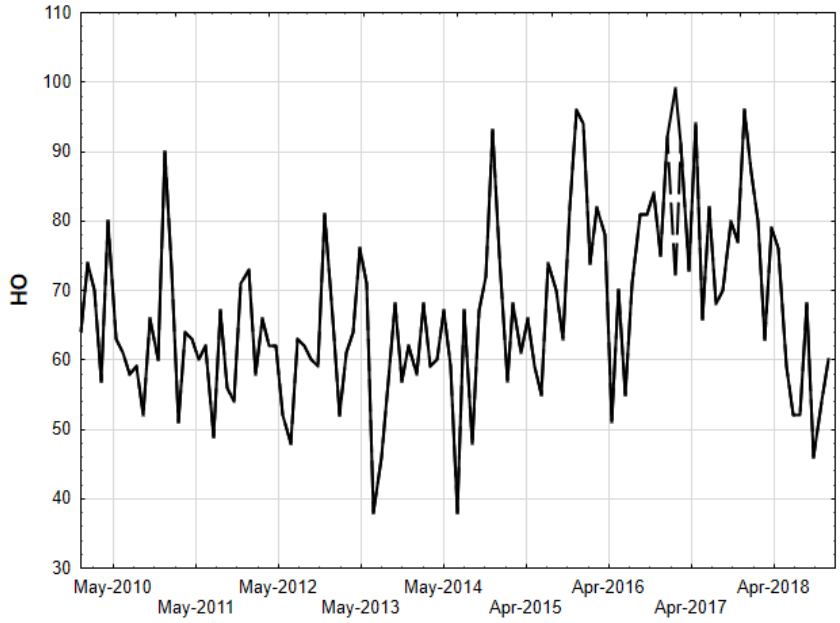
A primeira análise foca na verificação da dispersão dos registros de homicídio pelos meses, ficando claro que não apresenta linearidade. Isso é importante, pois a linearidade é um dos pressupostos para composição de modelos de regressão linear, técnica muito difundida. Não possível é verificar tendências apenas visualizando a Figura 02, que apresenta a curva de dispersão dos valores brutos e a curva de dispersão com a modificação feita, sendo esta tracejada.

Então, foi aplicado o Teste das Sequências, que apresentou o valor de $z' = -1{,}931$,que em módulo é menor que o valor crítico. Ou seja, num intervalo com $95{,}00\%$ de confiança, não é verificada tendência nos dados em análise, resultado que não descarta necessariamente a adequação de um modelo polinomial de grau maior que um, por exemplo.

Contudo a extrapolação de valores pode ficar comprometida quando se usam modelos polinomiais de grau elevado, pois a curva pode explodir em valores que não estão contidos no intervalo. Outra consequência deste resultado é o direcionamento para elaboração de modelos de séries temporais específicos, caso seja o objetivo, como o modelo de Suavização Exponencial de Holt (SEH), por exemplo, que é usado quando uma série apresenta tendência linear positiva ou negativa.

Percebeu-se também a dificuldade em constatar com rigor algum indício de sazonalidade determinística, mesmo observando que a partir de 2014 houve uma alteração no período e isso pode caracterizar a não existência desse fenômeno. Por isso foi aplicado o Teste de Kruskal-Wallis, avaliando a sazonalidade entre os meses. O valor crítico da distribuição χ^2, com 11 graus de liberdade é menor que o valor obtido com o teste $KW = 21{,}386$, indicando que é aceita a hipótese alternativa que existe pelo menos uma mediana diferente das demais. Quando o mesmo teste avalia a sazonalidade entre os anos, o valor obtido com 8 graus de liberdade $KW = 25{,}763$, igualmente maior que o crítico e indicando que existe pelo menos uma mediana maior que as demais.

Figura 2: dispersão dos registros de homicídio pelo tempo



Fonte: elaborado pelo autor

Observando as medianas entre os meses e entre os anos, fica evidente a superioridade do valor da mediana dos meses de dezembro, janeiro e fevereiro, respectivamente $75{,}00$,$75{,}00 e 71{,}56$, em relação às demais, que não ultrapassam o valor de $67{,}00$. O mesmo com os meses de 2017 e 2016, que apresentam valores das medianas correspondentes a $78{,}50$ e $76{,}50$, respectivamente, sendo 2015 o ano que mais se aproxima como o valor da mediana de $67{,}00$.

Dessa forma, a hipótese inicial que não existe sazonalidade estacionária na série de dados é rejeitada. Isso pode ser um forte indício da necessidade de se usar para previsão modelos de séries temporais que consideram o fator da sazonalidade, como o modelo de Suavização Exponencial Sazonal de Holt-Winters, por exemplo.

Também foi realizado o Teste de Levene, onde ficou verificado não existe diferença significativa entre a variância dos meses, pois o resultado apresentado foi $W = 1{,}031$, quando usada à média e $W = 0{,}868$ quando usada a mediana, ambos os valores menores que o crítico para 11 e 96 graus de liberdade na distribuição F. Quando o teste foi realizado com a variância entre os anos, os resultados foram $W = 0{,}717$ e $W = 0{,}628$, quando usadas respectivamente à média e mediana. Nos dois casos, os valores encontrados foram menores que o crítico quando considerados 8 e 99 graus de liberdade, implicando que não existe nenhuma variância que seja distinta das demais a nível estatístico de $95{,}00\%$.

Avançando a análise, são avaliadas as estatísticas básicas dos 108 registros conforme inicialmente coletados descritos no Quadro 01 e no Quadro 02, já considerando a substituição do mês de fevereiro de 2017.

Quadro 1: estatísticas básicas com substituição (início)

$\overline{x}$	LI	LS	Me	Mo	$x_{mínimo}$	$x_{máximo}$
66,681	64,302	69,060	66,000	59,000	38,000	99,000

Fonte: elaborado pelo autor

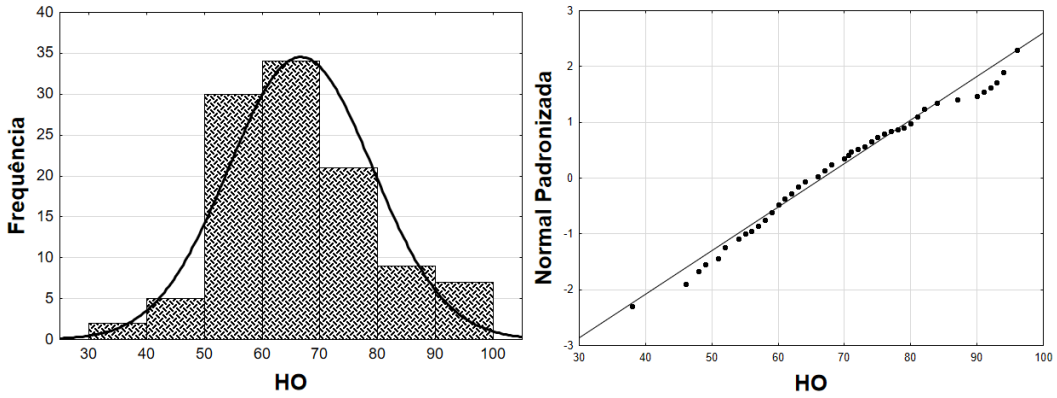
Quadro 2: estatísticas básicas com substituição (final)

1Q	3Q	s^2	s	Ep	As	C
59,000	74,000	155,298	12,471	1,200	0,373	-0,033

Fonte: elaborado pelo autor

Um fato que pode ser observado e que ajuda na verificação da normalidade univariada é a igualdade estatística entre média amostral, mediana e moda. Nesse caso, o valor da mediana $Me = 66{,}00$ pertence ao intervalo $(64{,}302; 69{,}060)$ e isso implica que o provável intervalo que contém a média populacional também contêm a mediana amostral. No entanto, a moda não esta neste intervalo, mas isso não descaracteriza a simetria da distribuição de modo que gaussianidade seja descartada, já que está sendo tratada amostra e algum desvio da curva teórica é esperado.

Figura 3: Histograma de frequências e gráfico pp-plot



Fonte: elaborado pelo autor.

Na figura 03, são apresentados o histograma e pp-plot. Apesar de serem gráficos distintos, apresentam informações equivalentes e se complementam.

Ambos são necessários, pois o histograma apresenta a aderência visual à distribuição normal, mas a alteração na largura das colunas pode induzir a erros de interpretação. Já o pp-plot apresenta medidas objetivas relacionando o registro mensal com o valor teórico da distribuição normal padronizada.

A última etapa da verificação da normalidade univariada será com valores do teste de Lilliefors: $Li = 0{,}0790$, valor estimado menor que o crítico num intervalo com $95{,}00\%$ de confiança. Adicionando este resultado a similaridade entre média e mediana e proximidade da moda, além do baixo valor de assimetria e curtose, visualização do histograma de frequência e pp-plot, pode-se afirmar que a procedência amostral pode ser considerada de uma distribuição normal, mesmo não possuindo informações sobre a normalidade da população.

Acredita-se que no decorrer dos meses, os registros mensais possam preencher as lacunas que desviam a amostra da curva teórica. Esse fato pode ser embasado numa rápida análise dos valores significativos do teste de Lilliefors: nos primeiros 48, 60, 72, 84 e 96 meses os resultados de Lilliefors foram significativos num intervalo com $95{,}00\%$ de confiança. O valor para Lilliefors nos 36 primeirosmeses não foi significativo nesse intervalo e confiança. Sugere-se um estudo específico, avaliando as características mês a mês, iniciando com $n = 30$ meses.

Ainda, como resultado da normalidade é possível prever uma quantidade máxima de registros especificando a probabilidade de ocorrência, Por exemplo, com $94{,}84\%$ de probabilidade, são esperados até 87 registros de homicídios e este valor sobe para 97 registros quando se considera a probabilidade de $99{,}25\%$. Imaginado resultados mais otimistas, obtêm-se a probabilidade $10{,}73\%$ para que a quantidade dos registros não ultrapasse 38 mensais.

Ciente das características expostas da variável HO, o poder executivo pode, nestes casos, reforçar o efetivo das unidades responsáveis pela investigação e esclarecimento deste tipo de crime, tendo em vista que a impunidade é sem dúvida um os fatores responsáveis pela decisão final no cometimento de um crime.

Além disso, à normalidade é requisito para aplicação de diversas técnicas multivariadas e quando não ocorre naturalmente, a variável pode ser transformada. Nesse caso, deve-se tomar cuidado quando o valor é transformado, utilizado e o

resultado transformado novamente pelo processo reverso, pois a transformação afeta a distribuição do termo de erro.

CONSIDERAÇÕES FINAIS

A análise preliminar indicou a não linearidade dos registros ao longo do tempo e por isso é desaconselhável a sua utilização em modelos de regressão linear, pois este é um dos pressupostos para este modelo.

O fato da dispersão temporal dos registros de homicídio não apresentar tendência definida induz a escolha de modelos baseados em séries temporais para explicar a dispersão ao invés de modelos baseados em tendências polinomiais de grau maior que um.

A confirmação da sazonalidade determinística permite a elaboração de modelos de séries temporais quando o objetivo for criar um modelo de dependência. O fato dos valores aderirem à distribuição normal sem necessidade de utilização da média e desvio padrão populacionais pode implicar na robustez da aderência. Além disso, estudando as propriedades da distribuição podem ser extraídas informações sem necessariamente elaborar um modelo temporal ou multivariado.

Pautado na confirmação estatística das características, considera-se que os valores avaliados são aptos para composição conjunta com outras variáveis em alguns modelos multivariados que apresentam relação de dependência e interdependência, mas deve haver cautela quando consideradas as técnicas mais sensíveis a não linearidade,

No que tange ao planejamento estatal, saber que os registros são normalmente distribuídos entorno de um valor médio ajuda a dimensionar o efetivo mínimo em delegacias especializadas, pautado num valor probabilístico previamente especificado.

Como sugestão para trabalhos futuros, recomenda-se refazer esta análise em diversos intervalos de tempo, menos espaçados, verificando se a aderência a curva normal teórica. Também se pode avaliar a influência das principais cidades na composição do total de homicídios através de um modelo de regressão linear múltipla, onde ao valor total é dependente do valor mensal de cada município.

Outro campo a ser explorado é a verificação de linearidade, sazonalidade, tendência e aderência à distribuição normal de probabilidades utilizando os registros

de feminicídio, verificando o comportamento da dispersão no tempo ou mesmo comparando com os homicídios totais.

REFERÊNCIAS

ALMEIDA, Antônia de; SILVIA, Elian; NOBRE, Juvêncio. Modificações e alternativas aos testes de Levene e de Brown e Forsythe para igualdade de variâncias e médias. **Revista Colombiana de Estadística**, v. 31, n. 2, p. 241-260. dez. 2008.

BRASIL. Portal do Planalto. **Legislação**. [Brasília, DF], 2011. Legislação Completa. Disponível em: <http://www.planalto.gov.br/ccivil_03/decreto-lei/Del2848compilado.htm>. Acesso em 21 set. 2020.

BARBETTA, Pedro A. **Estatística aplicada às ciências sociais**. Florianópolis, Editora UFSC, 2017.315 p.

BLAIN, Gabriel Constantino. Revisiting the critical values of the Lilliefors test: towards the correct agrometeorological use of the Kolmogorov-Smirnov framework. **Bragantia** (São Paulo, SP. Eletrônico), v. 73, p. 192-202, 2014.

COSTA NETO, Pedro L. de O. **Estatística**. São Paulo: Edgard Blucher, 1977. 264 p.

FÁVERO, Luiz P. L.; BELFIORE, Patrícia. **Manual de análise de dados**: estatística e modelagem multivariada com Excel, SPSS e Stata. Rio de Janeiro: Elsevier, 2017. 1187 p.

GLASS, Gene V.; STANLEY, Julian C. **Métodos estadísticos aplicados a las ciências sociales.** Los Olivos: Prentice Hall, 1996. 597 p.

HAIR JR, Joseph F. et al. **Análise multivariada de dados**. 6. ed. Porto Alegre: Bookman, 2009. 688 p.

KARMEL, Peter H.; POLASEK, M. **Estatística geral e aplicada para economistas**. 2. ed. São Paulo: Atlas, 1974. 601 p.

LILLIEFORS Test. In: **WIKIPÉDIA: a enciclopédia livre**. Disponível em: <https://en.wikipedia.org/wiki/Lilliefors_test>. Acesso em: 30 set. 2020.

MEYER, Paul L. **Probabilidade: aplicações à estatística**. Rio de Janeiro: Ao Livro Técnico, 1969. 391 p.

MORETTIN, Pedro A.; TOLOI, Clélia M. C. **Análise de Séries Temporais**. São Paulo: Blucher, 206.538 p.

SANTA CATARINA (Estado). Secretaria de Estado da Segurança Pública. **Segurança em números**: estatística criminal. Florianópolis, 2019. Disponível em: <https://portal.ssp.sc.gov.br/sspestatisticas.html>. Acesso em: 20 Set. 2020.

SICSÚ, Abraham L.; DANA, Samy. **Estatística Aplicada**: análise exploratória dos dados. São Paulo: Editora Saraiva, 2013. 145 p.

STATSOFT, Inc. (2014). **STATISTICA (data analysis software system)**, version 12.

WASHINGTON, Simon P.; KARLAFTIS, Matthew G.; MANNERING, Fred L. **Statistical and Econometric Methods for Transportation** Data Analysis. Boca Raton, Chapman e Hall/CRC, 2006. 509 p.

APÊNDICE A – DADOS USADOS

Seguem no Quadro 07 os valores brutos coletado de Santa Catarina (2020) usados na análise inicial.

Quadro 3:Registros de homicídios coletados

MÊS	HO	MÊS	HO	MÊS	HO	MÊS	HO
janeiro-10	64	abril-12	62	julho-14	38	outubro-16	81
fevereiro-10	74	maio-12	62	agosto-14	67	novembro-16	84
março-10	70	junho-12	52	setembro-14	48	dezembro-16	75
abril-10	57	julho-12	48	outubro-14	67	janeiro-17	92
maio-10	80	agosto-12	63	novembro-14	72	fevereiro-17	99
junho-10	63	setembro-12	62	dezembro-14	93	março-17	91
julho-10	61	outubro-12	60	janeiro-15	75	abril-17	73
agosto-10	58	novembro-12	59	fevereiro-15	57	maio-17	94
setembro-10	59	dezembro-12	81	março-15	68	junho-17	66
outubro-10	52	janeiro-13	66	abril-15	61	julho-17	82
novembro-10	66	fevereiro-13	52	maio-15	66	agosto-17	68
dezembro-10	60	março-13	61	junho-15	59	setembro-17	70
janeiro-11	90	abril-13	64	julho-15	55	outubro-17	80
fevereiro-11	73	maio-13	76	agosto-15	74	novembro-17	77
março-11	51	junho-13	71	setembro-15	70	dezembro-17	96
abril-11	64	julho-13	38	outubro-15	63	janeiro-18	87
maio-11	63	agosto-13	46	novembro-15	82	fevereiro-18	80
junho-11	60	setembro-13	57	dezembro-15	96	março-18	63
julho-11	62	outubro-13	68	janeiro-16	94	abril-18	79
agosto-11	49	novembro-13	57	fevereiro-16	74	maio-18	76
setembro-11	67	dezembro-13	62	março-16	82	junho-18	59
outubro-11	56	janeiro-14	58	abril-16	78	julho-18	52
novembro-11	54	fevereiro-14	68	maio-16	51	agosto-18	52
dezembro-11	71	março-14	59	junho-16	70	setembro 18	68
janeiro-12	73	abril-14	60	julho-16	55	outubro-18	46
fevereiro-12	58	maio-14	67	agosto-16	71	novembro-18	54
março-12	66	junho-14	59	setembro-16	81	dezembro-18	60

Fonte: Santa Catarina (2020)