
UNIVERSIDADE FEDERAL DO CEARÁ
CENTRO DE CIÊNCIAS
DEPARTAMENTO DE ESTATÍSTICA E MATEMÁTICA APLICADA

JONAS FREIRE, VITÓRIA TAVARES E VITOR REIS

# Ferramentas Computacionais com LaTeX

FORTALEZA
Julho/2023

# Sumário

# Lista de Tabelas

# Lista de Figuras

# 1 Introdução

Nesse trabalho, será realizado uma análise exploratória da base de dados *murders*, disponível no pacote *dslabs*, o qual pertence ao programa R. A base de dados se refere aos homicídios por arma de fogo ocorridos nos Estados Unidos (EUA) em 2010 e é composta pelas variáveis qualitativas: Estado, a sigla de cada Estado e a região de procedência; e pelas variáveis quantitativas: população e número de assassinatos.

# 2 Objetivos

## 2.1 Objetivo Geral

Utilizar os conhecimentos obtidos em sala de aula sobre o programa Latex e aplicar no presente relatório.

## 2.2 Objetivos Específicos

1. Resumir e organizar os dados;
2. Observar a distribuição dos dados;
3. Investigar possíveis correlações entre as variáveis.

# 3 Metodologia

Para o resumo das variáveis região, população e assassinatos será calculado suas devidas medidas descritivas. Além disso, para a questão de variabilidade e valores discrepantes que podem influenciar algumas medidas mais sensíveis, será feito um boxplot para cada variável quantitativa. Para ter uma visão clara de como ocorrem os assassinatos por região, será feito um gráfico de setores. Com a finalidade de comparar os assassinatos por região levando em conta a população, será feito um gráfico de barras para $\frac{assassinatos}{populao}$ por região. Para resumir e organizar a variável de estudo, assassinatos, será feito uma distribuição de frequências.

Com o intuito de observar a distribuição da variável assassinatos será feito um histograma, vai ser calculado o coeficiente de assimetria, dado por $CA = \frac{MO-\overline{X}}{S}$, e será calculado o coeficiente de curtose, valendo a equação $C = \frac{Q_3-Q_1}{3(P_{90}-P_{10})}$.

Com o objetivo de investigar possíveis correlações será feito uma análise bidimensional para as variáveis assassinatos e regiáo, e para as variáveis assassinatos e população. Como as variáveis assassinatos e região são, respectivamentee, quantitativa e qualitativa, será feito um comparativo de medidas descritivas dos assassinatos em cada região, um comparativo de boxplots para assassinatos por região para ver se existe alguma tendência, e o cálculo de $R^2$, dado por $R^2 = 1 - \frac{\overline{var(S)}}{var(S)}$, no qual $\overline{var(S)} = \frac{\sum_{i=1}^{k} n_i var_i(S)}{\sum_{i=1}^{k} n_i}$. Já para as variáveis assassinatos e população, os quais são duas variáveis quantitativas, será feito um gráfico de dispersão para observar se existe alguma tendência e o cálculo do coeficiente de correlação de Pearson, dado por $corr(X,Y) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i-\overline{x}}{dp(X)}\right)\left(\frac{y_i-\overline{y}}{dp(Y)}\right)$.

# 4 Aplicação e Resultados

Para realizar o cálculo de todas as seguintes medidas e plotagem de gráficos, além das tabelas, foi utilizado a linguagem R e o programa RStudio.

## 4.1 Resumo dos dados

Na tabela 1 é apresentado as medidas resumo para as variáveis região, população e assassinatos.

Tabela 1: Medidas descritivas para as variáveis região, população e assassinatos

| Região | População | Assassinatos |
|---|---|---|
| Nordeste: 9 | Min.: 563626 | Min.: 2.0 |
| Sul: 17 | Q1: 1696962 | Q1: 24.5 |
| Norte Central: 12 | Mediana: 4339367 | Mediana: 97.0 |
| Oeste: 13 | Média: 6075769 | Média: 184.4 |
| | Q3: 6636084 | Q3: 268.0 |
| | Max.: 37253956 | Max.: 1257.0 |
| | dp.: 6860669.1 | dp.: 236.1 |
| | CV.: 112.9 | CV.: 128.1 |

Como se pode ver, a variável qualitativa nominal região tem o Sul como a classe modal, ocorrendo 17 vezes em todos os Estados.

Já nas outras duas variáveis, por serem ambas quantitativas discretas, é possível calcular diversas medidas. Na variável população a média é maior que a mediana, o que indica assimetria a direita. Seu coeficiente de variação é enorme, mostrando uma grande heterogeneidade. Na variável assassinatos a média também é maior do que a mediana o que indica, também, assimetria a direita, além do coeficiente de variação que também é enorme, mostrando uma grande heterogeneidade.

Na figura 1 é possível ver o boxplot para a variável assassinatos.

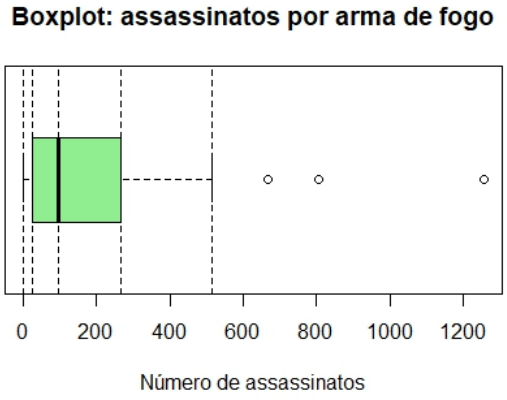


Figura 1: Boxplot para a variável assassinatos

O boxplot para assassinatos é assimétrico à direita e, como é possível notar, há três outliers nele, o que pode ter afetado nos valores da média e do coeficiente de variação. Os três outliers representam os Estados California, Florida e Texas.

Na figura 2 é apresentado o boxplot para a variável população.

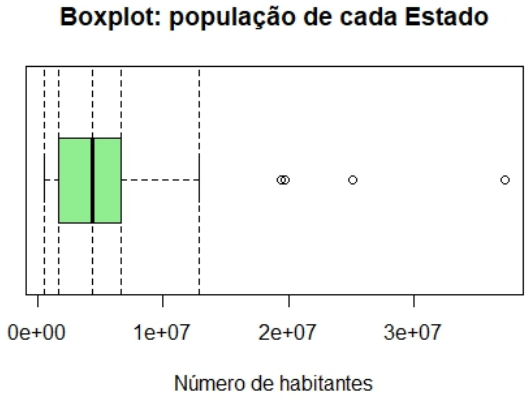


Figura 2: Boxplot para a variável população

O boxplot para a população é assimétrico à direita apresentando quatro outliers o que, por sua vez, pode ter afetado o valor de medidas como média e coeficiente de variação, o quais estavam grandes na tabela 1. Os Estados representados pelos quatro outliers são California, Florida, New York e Texas. Daí já se percebe uma relação entre os três Estados mais populosos serem, também, os que mais possuem assassinatos.

A figura 3 representa o gráfico de setores para os assassinatos por região.


Figura 3: Gráfico de setores para a variável assassinatos por região.

O gráfico de setores tem a região Sul como sendo a que possui a maior porcentagem de assassinatos.

A figura 4 mostra o percentual de assassinatos por região.

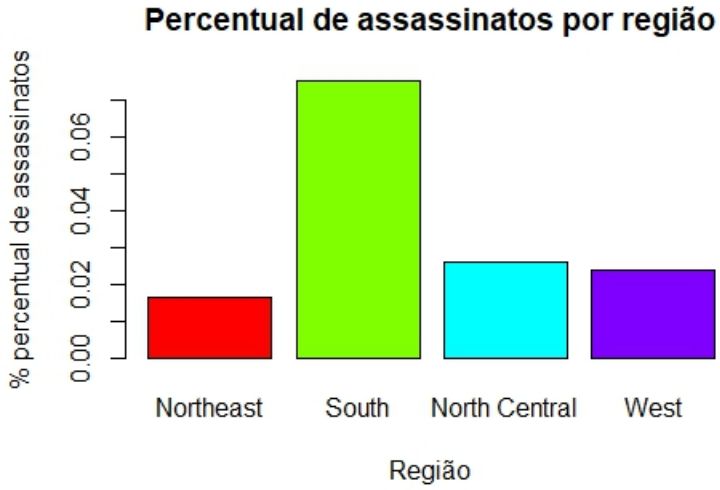


Figura 4: Gráfico de barras para a variável $\frac{assassinatos}{populao}$ por região

O percentual dá uma comparação mais confiável pois leva em conta as diferentes populações de cada região. Desse modo, tem-se o Sul, novamente, como sendo a região com mais assassinatos.

Na tabela 4.1 é mostrado a distribuição de frequências de assassinatos.

| Classes | $f_i$ |
|---|---|
| $2.00 \vdash 181.29$ | 33 |
| $181.29 \vdash 360.58$ | 10 |
| $360.58 \vdash 539.87$ | 5 |
| $539.87 \vdash 719.16$ | 1 |
| $719.16 \vdash 898.45$ | 1 |
| $898.45 \vdash 1077.74$ | 0 |
| $1077.74 \vdash 1257.03$ | 1 |

Tabela 2: Distribuição de frequências para a variável assassinatos

Na tabela 4.1 agora, com os dados organizados e resumidos, é possível notar que a classe modal está presente nas menores quantidades de assassinatos e, na classe de valores discrepantes, há uma quantidade mínima de ocorrências.

## 4.2 Distribuição dos dados

Na figura 5 é apresentado o histograma para a variável população.

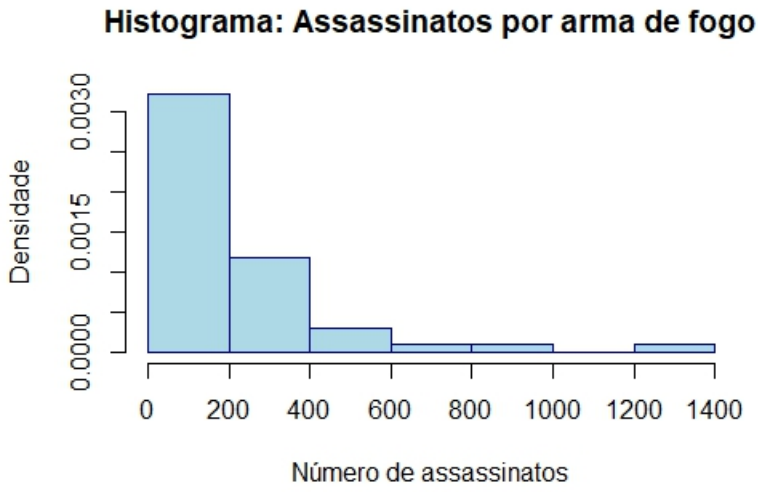


Figura 5: Histograma para a variável assassinatos

Na figura 5 é possível notar que o gráfico é assimétrico à direita. Para ter essa informação a partir de uma medida descritiva, tem-se o valor do coeficiente de assimetria

$$CA = 1.6294$$

o qual é positivo e maior do que zero, mostrando que a distribuição realmente é assimétrica à direita.

Além disso, para saber como ocorre a dispersão da distribuição ou concentração dos valores em relação às medidas de tendência central, tem-se o valor do coeficiente de curtose

$$C = 4.0882$$

o qual é maior do que 0.263 mostrando que a distribuição é platicúrtica, havendo um achatamento e, consequentemente, uma maior dispersão em relação às medidas de tendência central.

## 4.3 Análise bidimensional dos dados

- Assassinatos e região
  A tabela 3 mostra as medidas descritivas dos assassinatos em cada região.

Tabela 3: Medidas descritivas de assassinatos por região

| Região | n | Média | dp. | var. | Q1 | Med. | Q3 |
|---|---|---|---|---|---|---|---|
| Nordeste | 9 | 163.22 | 200.19 | 40077.44 | 11.00 | 97.00 | 246.00 |
| Sul | 17 | 246.76 | 212.86 | 45310.32 | 111.00 | 207.00 | 293.00 |
| Norte Central | 12 | 152.33 | 154.22 | 23785.15 | 29.25 | 80.00 | 312.75 |
| Oeste | 13 | 147.00 | 339.10 | 114989.83 | 12.00 | 36.00 | 84.00 |
| Total | 51 | 184.37 | 236.13 | 55755.56 | 24.50 | 97.00 | 268.00 |

Na tabela 3 é possível ver que não há nenhuma tendência entre as medidas descritivas de assasinatos por região e que, inclusive, há uma variância de assassinatos na região Oeste maior do que a variância global.

Na figura 6 é apresentado quatro boxplots de assassinatos, um para cada região.

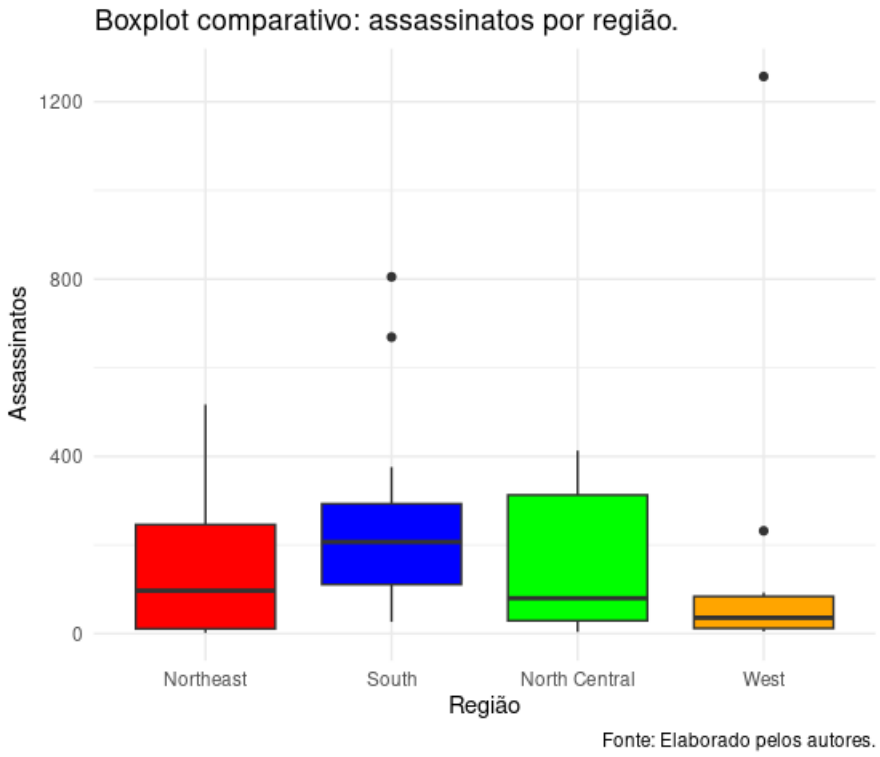


Figura 6: Boxplots de assassinatos em cada região

Como é possível observar, não há nenhuma tendência nos boxplots, havendo outliers em dois deles, o que pode ter influênciado, também, nas medidas descritivas mais sensíveis da tabela 3.

Para ter uma medida que quantifique o grau de correlação ou que indique uma não correlação, foi calculado o seguinte valor de $R^2$

$$R^2 = -0.02382$$

o qual é menor do que 0, mostrando que a região de procedência não melhora a previsão da variável assassinatos.

- Assassinatos e população
  Na figura 7 é mostrado o gráfico de dispersão para assassinatos e população.

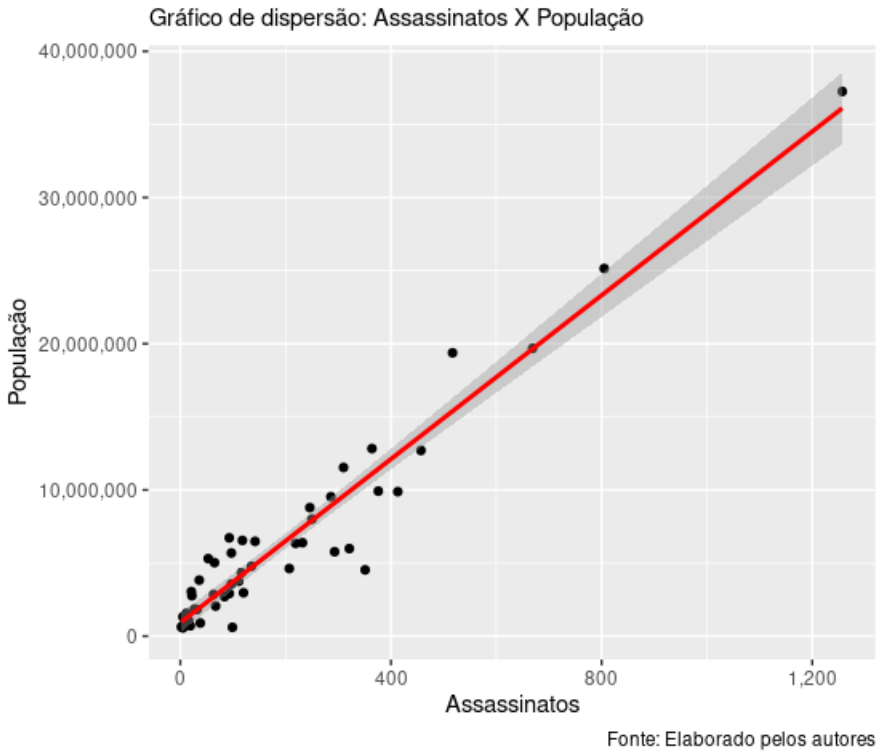


Figura 7: Diagrama de dispersão para as variáveis X: assassinatos e Y: população

É facilmente notável que conforme o número de assassinatos aumenta, a população também aumenta. Isso indica uma correlação direta e, devido a proximidade dos valores à reta, uma correlação forte entre as duas variàveis.

Para ter uma quantificação do nível de correlação entre as duas variàveis foi calculado o seguinto coeficinete de correlação de Pearson

$$corr(X, Y) = 0.9636$$

o qual é positivo e muito próximo de 1, o que indica uma correlação direta forte de 96,36%.

# 5 Conclusão

A partir de toda a análise exploratória da base de dados *murders* foi possível obter respostas acerca dos objetivos propostos. Desse modo, com o resumo dos dados é possível ver a presença de valores discrepantes que, pelos outliers dos boxplots, é possível ver uma relação direta entre assassinatos e população. Desse modo, é visto o Sul sendo a região mais populosa e com maior ocorrência de assassinatos.

A partir da distribuição da figura 5 é possível ver que a maior concentração da frequência de assassinatos está presente nas menores ocorrências de assassinatos e, dessa forma, é possível definir a distribuição como sendo assimétrica a direita, além do coeficiente de curtose o qual definiu o achatamento da distribuição sendo platicúrtica.

Pelas análises bidimensionais é visto que a região não melhora a previsibilidade dos assas-

sinatos e que, em contraste com isso, a população tem uma correlação direta forte com os assassinatos e, desse modo, é possível prever os assassinatos pela população.

# Referências


[1] R Core Team (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. Disponível em: `https://www.R-project.org/`.

[2] *Murders: U.S. State-Level Murder Data*. Disponível em: `https://cran.r-project.org/package=dslabs`.