**ENTREGA 3 – ANÁLISE EXPLORATÓRIA ADICIONAL E PROBABILIDADES**

1. **Introdução**

Esta terceira entrega prossegue com a análise dos dados de óbitos no trânsito brasileiro em 2023. Focaremos na distribuição dos dados por região, com base nas tabelas e gráficos box-plot fornecidos, excluindo outliers. O objetivo é identificar concentrações de dados e calcular probabilidades básicas de ocorrência de determinados níveis de óbitos entre as regiões. As referências às tabelas seguem a numeração e descrições anteriores.

1. **Metodologia**

A análise exploratória baseia-se na interpretação dos gráficos box-plot, que sumarizam a distribuição dos óbitos por região (Norte, Nordeste, Sudeste, Sul, Centro-Oeste). A concentração dos dados é avaliada pela simetria da "caixa" (Intervalo Interquartil - IQR) e pela relação entre média e mediana. Probabilidades são calculadas como P(Evento) = (Número de Regiões com a Característica) / 5.

As tabelas analisadas nesta entrega correspondem a:

* **Tabela/Box-Plot 1:** Óbitos por Acidentes de Transporte (CID-BR-10: 104) por Região.
* **Tabela/Box-Plot 2:** Total de Óbitos em Vias Públicas por Região.
* **Tabela/Box-Plot 3:** Óbitos por Acidentes de Transporte em Vias Públicas (CID-BR-10: 104, Local: Via Pública) por Região.
* **Tabela/Box-Plot 4:** Óbitos de Ocupantes de Automóvel em Acidentes de Transporte em Vias Públicas por Região.

**3. Análise dos Dados e Resultados por Tabela/Box-Plot**

**3.1 Análise da Tabela/Box-Plot 1: Óbitos p/Residênc por Região (Causa - CID-BR-10: 104 Acidentes de transporte)**

* **Dados Regionais:** Região Norte: 3586; Região Nordeste: 11231; Região Sudeste: 11290; Região Sul: 5925; Região Centro-Oeste: 3906.
  + Ordenados: 3586, 3906, 5925, 11231, 11290.
* **Análise de Concentração e Assimetria (com base nos valores calculados e no gráfico):**
  + Mínimo: 3586
  + Q1 (Primeiro Quartil): (3586 + 3906) / 2 = 3746
  + Mediana (Q2): 5925
  + Q3 (Terceiro Quartil): (11231 + 11290) / 2 = 11260.5
  + Máximo: 11290
  + Média (indicada no gráfico): ~7187.6
  + IQR (Intervalo Interquartil): Q3 - Q1 = 11260.5 - 3746 = 7514.5
  + A distância Mediana-Q1 (5925 - 3746 = 2179) é menor que a distância Q3-Mediana (11260.5 - 5925 = 5335.5), indicando que a parte superior da caixa do box-plot é mais longa.
  + A Média (~7187.6) é consideravelmente maior que a Mediana (5925).
  + Ambas as observações apontam para uma **assimetria à direita**. Isso sugere que a maioria das regiões tende a ter um número menor de óbitos por esta causa, mas algumas regiões (Nordeste e Sudeste) apresentam valores substancialmente mais altos, "puxando" a média para cima. Conclui-se que há uma **concentração de dados em valores inferiores**.
* **Probabilidade (baseada nas 5 regiões):**
  + Probabilidade de uma região ter óbitos acima da mediana (> 5925): 2/5 = 40% (Regiões Nordeste e Sudeste).
  + Probabilidade de uma região ter óbitos no quartil superior (acima do valor que define Q3, ou seja, os dois valores mais altos): 2/5 = 40%.
  + Probabilidade de uma região ter óbitos no quartil inferior (abaixo do valor que define Q1, ou seja, os dois valores mais baixos): 2/5 = 40%.

**3.2 Análise da Tabela/Box-Plot 2: Óbitos p/Residênc por Região (Local ocorrência: Via pública)**

* **Dados Regionais:** Região Norte: 5328; Região Nordeste: 18373; Região Sudeste: 15426; Região Sul: 6922; Região Centro-Oeste: 4250.
  + Ordenados: 4250, 5328, 6922, 15426, 18373.
* **Análise de Concentração e Assimetria (com base nos valores calculados e no gráfico):**
  + Mínimo: 4250
  + Q1: (4250 + 5328) / 2 = 4789
  + Mediana: 6922
  + Q3: (15426 + 18373) / 2 = 16899.5
  + Máximo: 18373
  + Média (indicada no gráfico): ~10059.8
  + IQR: 16899.5 - 4789 = 12110.5
  + A distância Mediana-Q1 (6922 - 4789 = 2133) é significativamente menor que a distância Q3-Mediana (16899.5 - 6922 = 9977.5). A parte superior da caixa é muito mais longa.
  + A Média (~10059.8) é substancialmente maior que a Mediana (6922).
  + Esses indicadores demonstram uma **forte assimetria à direita**. A maioria das regiões apresenta um total de óbitos em vias públicas relativamente mais baixo, enquanto duas regiões (Nordeste e Sudeste) possuem valores muito elevados. Há uma clara **concentração de dados em valores inferiores**.
* **Probabilidade (baseada nas 5 regiões):**
  + Probabilidade de uma região ter óbitos totais em vias públicas acima da mediana (> 6922): 2/5 = 40% (Regiões Nordeste e Sudeste).
  + Probabilidade de uma região ter óbitos acima de Q3 (> 16899.5): 1/5 = 20% (Região Nordeste).
  + Probabilidade de uma região ter óbitos abaixo de Q1 (< 4789): 1/5 = 20% (Região Centro-Oeste).

**3.3 Análise da Tabela/Box-Plot 3: Óbitos p/Residênc por Região (Causa - CID-BR-10: 104 Acidentes de transporte; Local ocorrência: Via pública)**

* **Dados Regionais:** Região Norte: 1805; Região Nordeste: 5714; Região Sudeste: 5541; Região Sul: 3291; Região Centro-Oeste: 2056.
  + Ordenados: 1805, 2056, 3291, 5541, 5714.
* **Análise de Concentração e Assimetria (com base nos valores calculados e no gráfico):**
  + Mínimo: 1805
  + Q1: (1805 + 2056) / 2 = 1930.5
  + Mediana: 3291
  + Q3: (5541 + 5714) / 2 = 5627.5
  + Máximo: 5714
  + Média (indicada no gráfico): ~3681.4
  + IQR: 5627.5 - 1930.5 = 3697
  + A distância Mediana-Q1 (3291 - 1930.5 = 1360.5) é menor que a distância Q3-Mediana (5627.5 - 3291 = 2336.5), indicando uma caixa superior mais longa.
  + A Média (~3681.4) é maior que a Mediana (3291).
  + Observa-se uma **assimetria à direita**, embora menos acentuada que as anteriores. Isso indica uma **concentração de regiões com menos óbitos** por acidentes de transporte em vias públicas, com algumas regiões (Nordeste e Sudeste) apresentando valores mais elevados.
* **Probabilidade (baseada nas 5 regiões):**
  + Probabilidade de uma região ter óbitos por acidentes de transporte em vias públicas acima da mediana (> 3291): 2/5 = 40% (Regiões Nordeste e Sudeste).
  + Probabilidade de uma região ter óbitos acima de Q3 (> 5627.5): 1/5 = 20% (Região Nordeste).
  + Probabilidade de uma região ter óbitos abaixo de Q1 (< 1930.5): 1/5 = 20% (Região Norte).

**3.4 Análise da Tabela/Box-Plot 4: Óbitos p/Residênc por Região (Grupo CID-10: Ocupante automóvel traumat acidente transporte; Causa - CID-BR-10: 104 Acidentes de transporte; Local ocorrência: Via pública)**

* **Dados Regionais:** Região Norte: 227; Região Nordeste: 999; Região Sudeste: 1511; Região Sul: 1299; Região Centro-Oeste: 546.
  + Ordenados: 227, 546, 999, 1299, 1511.
* **Análise de Concentração e Assimetria (com base nos valores calculados e no gráfico):**
  + Mínimo: 227
  + Q1: (227 + 546) / 2 = 386.5
  + Mediana: 999
  + Q3: (1299 + 1511) / 2 = 1405
  + Máximo: 1511
  + Média (indicada no gráfico): ~916.4
  + IQR: 1405 - 386.5 = 1018.5
  + A distância Mediana-Q1 (999 - 386.5 = 612.5) é maior que a distância Q3-Mediana (1405 - 999 = 406). A parte inferior da caixa é mais longa.
  + A Média (~916.4) é menor que a Mediana (999).
  + Essas características indicam uma **leve assimetria à esquerda**. Isso sugere que, para óbitos de ocupantes de veículos, pode haver uma maior dispersão nos valores mais baixos, com os dados tendendo a se agrupar ou se concentrar um pouco mais em **valores relativamente mais altos** em comparação com as outras distribuições analisadas, ou pelo menos não tão concentrados em valores baixos.
* **Probabilidade (baseada nas 5 regiões):**
  + Probabilidade de uma região ter óbitos de ocupantes de veículos acima da mediana (> 999): 2/5 = 40% (Regiões Sudeste e Sul).
  + Probabilidade de uma região ter óbitos acima de Q3 (> 1405): 1/5 = 20% (Região Sudeste).
  + Probabilidade de uma região ter óbitos abaixo de Q1 (< 386.5): 1/5 = 20% (Região Norte).

**4. Conclusões Consolidadas**

A análise exploratória dos quatro conjuntos de dados e seus respectivos box-plots revela padrões distintos na distribuição de óbitos por região no Brasil em 2023:

1. **Concentração Predominante em Valores Inferiores:** Para as três primeiras tabelas analisadas (Óbitos por Acidentes de Transporte CID-10:104, Total de Óbitos em Vias Públicas, e Óbitos por Acidentes de Transporte em Vias Públicas), observa-se uma clara assimetria à direita. Isso significa que a maioria das regiões apresenta um número de óbitos relativamente menor, enquanto poucas regiões (consistentemente Nordeste e Sudeste) exibem números significativamente mais altos, puxando a média para cima e alongando a cauda superior da distribuição. Essa concentração em valores inferiores com alguns valores altos é um padrão comum em dados socioeconômicos e de saúde distribuídos regionalmente em países com grandes disparidades.
2. **Mudança de Padrão para Ocupantes de Veículos:** A quarta tabela (Óbitos de Ocupantes de Automóvel em Acidentes de Transporte em Vias Públicas) apresenta uma leve assimetria à esquerda. A média é inferior à mediana, e a distância (Mediana-Q1) é maior que (Q3-Mediana). Isso sugere uma maior concentração de regiões com valores mais elevados ou uma dispersão maior nos valores mais baixos para este tipo específico de óbito. Este resultado contrasta com os anteriores e pode indicar que a problemática dos óbitos de ocupantes de veículos é distribuída de forma diferente entre as regiões, talvez com menos regiões apresentando números extremamente baixos. Isso é consistente com o CV mais elevado para "Ocupantes de Veículos" (71.3%) encontrado na Entrega 2, que indicava maior variabilidade regional para este grupo.
3. **Probabilidades Regionais:** Em todos os cenários, a probabilidade de uma região aleatória apresentar um número de óbitos acima da mediana é de 40% (2 em 5 regiões). A probabilidade de uma região estar no extremo superior (acima de Q3, representando o valor mais alto ou os dois mais altos) ou no extremo inferior (abaixo de Q1) foi geralmente de 20% (1 em 5 regiões, quando o quartil não era formado por um valor que também era o mínimo/máximo do conjunto).
4. **Importância da Análise Detalhada:** A diferença na forma da distribuição entre as causas de óbito (geral vs. ocupantes de veículos) reforça a necessidade de análises específicas para diferentes tipos de acidentes e vítimas, como já destacado na Entrega 1 ao calcular percentagens distintas para estes grupos.

**5. Limitações**  
As conclusões sobre concentração e as probabilidades calculadas são baseadas em um conjunto de dados pequeno (N=5 regiões). Embora o box-plot seja uma ferramenta útil para visualizar a distribuição, a robustez estatística de inferências sobre a população geral de "eventos" (se cada morte fosse um evento) é limitada pela agregação dos dados em nível regional. A análise trata cada região como uma observação única, o que é apropriado para comparar regiões, mas não para inferir sobre a distribuição dos óbitos individuais dentro delas sem acesso aos dados brutos.

Esta análise complementa as entregas anteriores, fornecendo uma perspectiva visual e probabilística simplificada sobre a dispersão dos dados de mortalidade no trânsito entre as regiões brasileiras.