PROJETO DE ESTATÍSTICA

Integrantes:

Khalil Khalid Abou Anche - SP3121925
Rafael Valverde Zanata da Silva - SP3119866
Bruno de Almeida Fischer - SP3120139
Rafael Teixeira Fonseca - SP3126919

1 Introdução

Este relatório tem como objetivo aplicar técnicas estatísticas para analisar a prevalência de duas condições de saúde crônicas: hipertensão e diabetes tipo 2, no condado de Allegheny, Estados Unidos. Por meio de dois conjuntos de dados públicos contendo informações anonimizadas da população local, buscamos identificar padrões, diferenças entre grupos e possíveis relações entre as variáveis.

A partir dessa abordagem quantitativa, buscamos responder perguntas relevantes que podem contribuir para a compreensão do comportamento dessas doenças em diferentes regiões, faixas populacionais e entre os sexos.

1.1 Problema

O problema central abordado neste relatório é compreender a prevalência de hipertensão e diabetes tipo 2 se distribui no condado de Allegheny. As perguntas que buscamos responder são:

1. Como se distribuem os casos de hipertensão e diabetes tipo 2 nas 390 regiões do condado de Allegheny?
2. Quais são os valores mínimo e máximo de casos diagnosticados em uma região?
3. A distribuição dos casos é simétrica ou assimétrica?
4. Existem valores extremos (outliers) na distribuição das doenças?
5. A prevalência das doenças é mais alta em homens ou mulheres?
6. Qual a probabilidade condicional de um homem ter hipertensão? E uma mulher?
7. A prevalência das doenças pode ser modelada por uma distribuição de Poisson?
8. A distribuição normal seria apropriada para modelar essas variáveis?
9. A correlação entre hipertensão e diabetes é estatisticamente significativa? Essa correlação muda entre homens e mulheres?
10. Há diferença significativa entre as médias de prevalência entre homens e mulheres?

11. A proporção de pacientes medicados para uma doença afeta a prevalência da outra?
12. A adesão ao tratamento (medicação) tem impacto indireto na saúde regional?

1.2 Solução Proposta

A fim de responder às perguntas propostas, realizamos uma análise estatística exploratória utilizando Python. Foram aplicadas técnicas de estatística descritiva, gráficos de distribuição, cálculo de proporções e, posteriormente, serão considerados testes de hipótese e probabilidade condicional. O objetivo é interpretar os dados com base em fundamentos estatísticos e compreender o comportamento das doenças nas diferentes regiões analisadas.

2. Dicionário de Siglas dos Datasets

Nesta seção, apresentamos as siglas utilizadas nos dois conjuntos de dados fornecidos: Hypertension.csv e Diabetes.csv.

2.1 Dataset: Hypertension.csv

Sigla	**Significado**
CT	Census Tract (divisão geográfica da região de Allegheny)
TPAD	Total de pessoas inscritas no plano durante 2015
TPAN2	Total de diagnosticados com hipertensão que fizeram uso de medicação
TPAN	Total de pessoas diagnosticadas com hipertensão
TWAD	Total de mulheres inscritas
TWAN	Total de mulheres diagnosticadas com hipertensão
TWAN2	Total de mulheres diagnosticadas e medicadas para hipertensão
TMAD	Total de homens inscritos
TMAN	Total de homens diagnosticados com

	hipertensão
TMAN2	Total de homens diagnosticados e medicados para hipertensão

2.2 Dataset: Diabetes.csv

Sigla	**Significado**
CT	Census Tract (Divisão geográfica da região de Allegheny)
BPAD	Total de pessoas inscritas no plano durante 2015
BPAN	Total de pessoas diagnosticadas com diabetes tipo 2
BPAN2	Total de diagnosticados com diabetes que fizeram uso de medicação
BWAD	Total de mulheres inscritas
BWAN	Total de mulheres diagnosticadas com diabetes tipo 2
BWAN2	Total de mulheres diagnosticadas e medicadas para diabetes
BMAD	Total de homens inscritos
BMAN	Total de homens diagnosticados com diabetes tipo 2
BMAN2	Total de homens diagnosticados e medicados para diabetes

3. Estatística Descritiva

Esta seção apresenta um panorama estatístico das variáveis principais: TPAN (casos de hipertensão) e BPAN (casos de diabetes tipo 2). Através de estatísticas descritivas e representações gráficas, podemos entender melhor o comportamento das doenças ao longo das 390 regiões do condado de Allegheny.

Utilizamos métricas como média, mediana, desvio padrão e variância, além de calcular proporções por sexo. Por fim, os dados foram visualizados com histogramas e boxplots, permitindo identificar padrões de distribuição e a presença de valores discrepantes (outliers).

3.1 Estatísticas Gerais

Foram calculadas as principais estatísticas descritivas para as variáveis **TPAN** (hipertensão) e **BPAN** (diabetes tipo 2), a fim de compreender a distribuição dos casos nas 390 regiões analisadas.

Fórmulas Utilizadas

- **Média (μ):**
  $\mu = (x_1 + x_2 + \ldots + x) / n$
  Representa o valor médio de casos por região.

- **Desvio Padrão (σ):**
  $\sigma = \sqrt{[((x_1 - \mu)^2 + \ldots + (x - \mu)^2) / n]}$
  Mede o grau de dispersão dos dados.

- **Variância (σ^2):**
  $\sigma^2 = [(x_1 - \mu)^2 + \ldots + (x - \mu)^2] / n$
  Indica o quão os valores se afastam da média.

- **Mediana:**
  Valor central da distribuição ordenada.

Resultados:

Hipertensão (TPAN)

- Total de Regiões: 390
- Média de Casos: 405.79

- Desvio Padrão: 234.20
- Variância: 54.848,37
- Valor Mínimo: 0.0
- Mediana: 354.5
- Valor Máximo: 1246.0

Diabetes Tipo 2 (BPAN)

- Total de Regiões: 390
- Média de Casos: 148.24
  Desvio Padrão: 80.93
- Variância: 6.549,95
- Valor Mínimo: 0.0
- Mediana: 137.0
- Valor Máximo: 458.0

Observa-se que os casos de hipertensão são, em média, mais comuns e apresentam maior variabilidade entre as regiões.

3.2 - Proporção por sexo

A proporção foi calculada com a fórmula:

$$Proporção = \frac{Pessoas\ diagnosticadas}{Total\ de\ pessoas\ inscritas} \times 100$$

Resultados:

Hipertensão (TPAN):

- **Homens diagnosticados:** 24.99%
- **Mulheres diagnosticadas:** 24.23%

Diabetes Tipo 2 (BPAN):

- **Homens diagnosticados:** 9.59%
- **Mulheres diagnosticadas:** 8.39%

A análise das proporções revela que a diferença na prevalência entre homens e mulheres para hipertensão é muito pequena. Para diabetes tipo 2, os homens apresentam uma taxa ligeiramente maior de diagnóstico em comparação com as

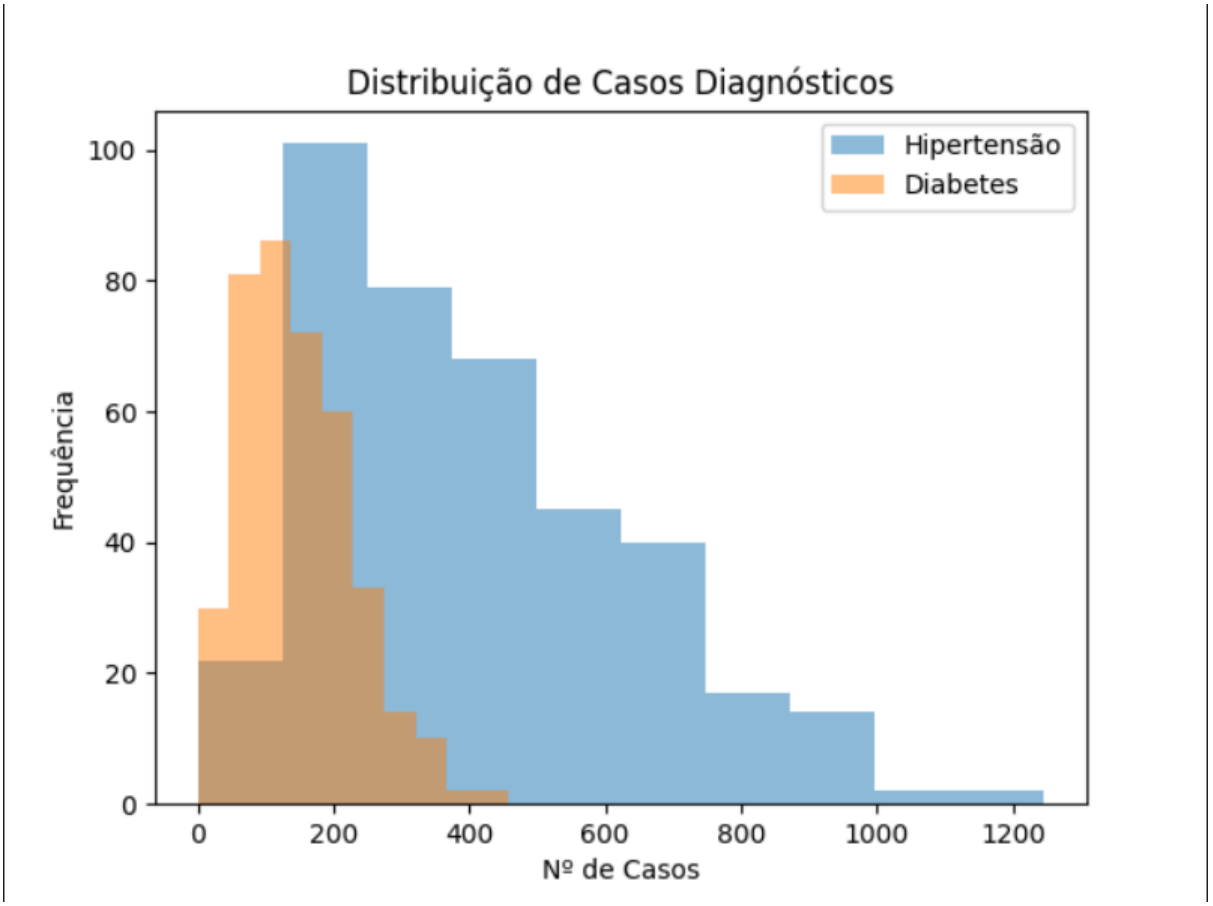
mulheres. Isso indica que o sexo da pessoa não é um fator determinante para o diagnóstico dessas doenças, sugerindo que outros fatores, como estilo de vida, genética ou acesso ao sistema de saúde, podem ter maior impacto na prevalência dessas condições.

3.3 Gráficos de Distribuição

Para complementar a análise descritiva, foram utilizados dois tipos de gráficos: boxplot e histograma, permitindo uma melhor visualização da distribuição dos casos por região.

Histograma da Distribuição de Casos Diagnósticos

O histograma abaixo mostra a distribuição da quantidade de casos diagnosticados por região, tanto para **hipertensão** quanto para **diabetes tipo 2**:



Análise do Gráfico

- A maioria das regiões apresenta até 500 casos de hipertensão e menos de 300 casos de diabetes.

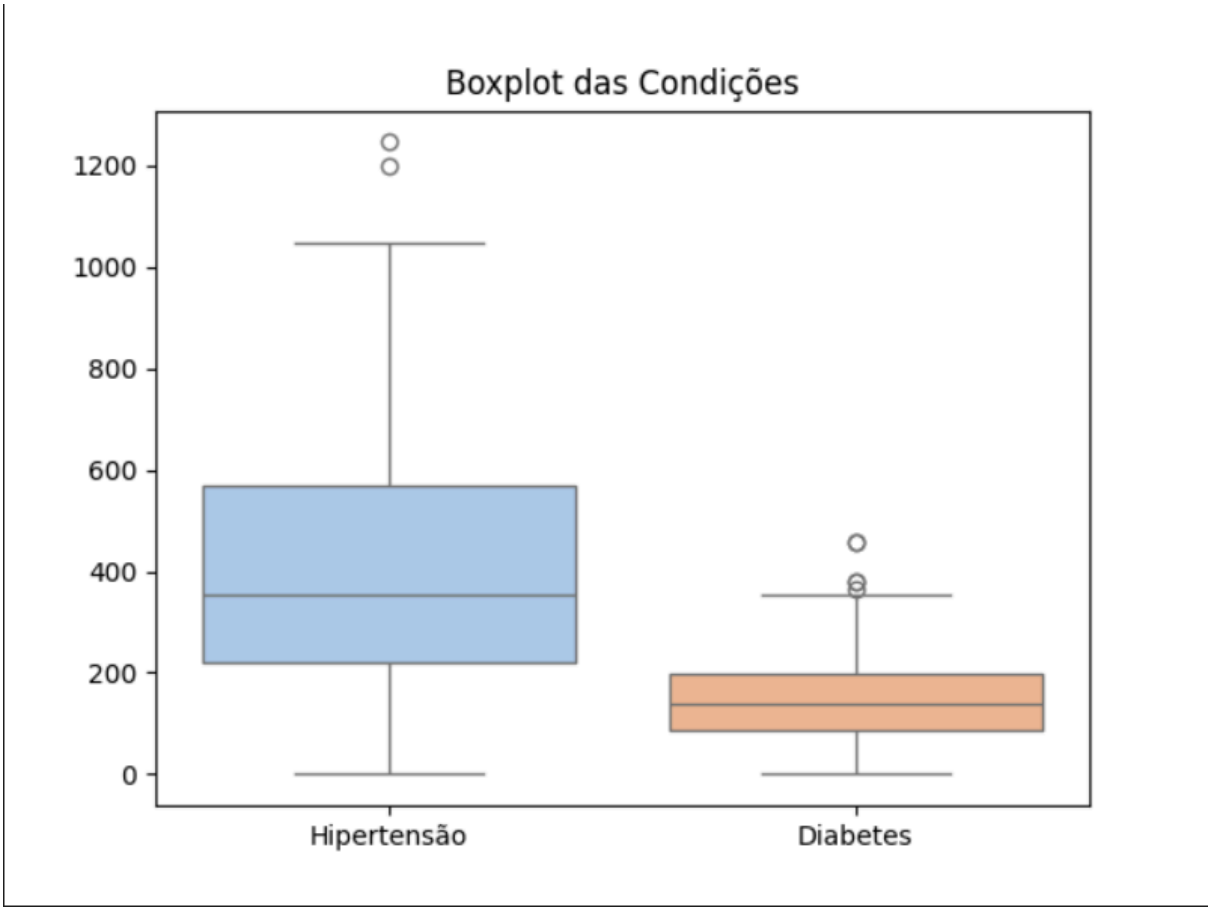
- A distribuição de diabetes é mais concentrada à esquerda (menor número de casos), com menor dispersão.

- Já a hipertensão possui maior variação entre as regiões, com casos extremos que chegam a ultrapassar 1000 diagnósticos, indicando a presença de outliers.

- Ambas distribuições são assimétricas à direita, ou seja, com "cauda longa" indicando que poucas regiões concentram muitos casos.

Boxplot das Condições

O boxplot abaixo apresenta a dispersão dos dados de casos diagnosticados por região para as duas condições analisadas: hipertensão e diabetes tipo 2. Esse tipo de gráfico ajuda a visualizar a mediana, os quartis e possíveis valores discrepantes (outliers).



Análise do BoxPlot

- A hipertensão apresenta uma distribuição mais ampla, com maior variabilidade entre as regiões. A mediana está próxima de 350 casos, e existem diversos valores extremos (outliers) acima de 1000 casos.

- A diabetes tipo 2 possui uma distribuição mais concentrada. A mediana está em torno de 137 casos, e embora existam alguns outliers, eles não são tão distantes do restante dos dados quanto no caso da hipertensão.

- A caixa da hipertensão é visivelmente mais larga, indicando maior dispersão (diferença entre o primeiro e terceiro quartil).

- A presença de vários outliers em ambas as condições sugere que algumas regiões têm características particulares, como maior densidade populacional, maior prevalência de fatores de risco ou melhor capacidade de diagnóstico.

4 Probabilidade

Nesta seção, aplicamos conceitos de probabilidade para analisar a prevalência de hipertensão e diabetes tipo 2 no condado de Allegheny, utilizando as informações agregadas por Census Tract (CT). É importante notar que, devido à natureza agregada dos dados, não é possível determinar diretamente a probabilidade de um indivíduo ter ambas as condições simultaneamente.

4.1 Probabilidade Condicional por Sexo

Dada a informação do sexo de uma pessoa selecionada aleatoriamente no condado de Allegheny, qual é a probabilidade de ela ter hipertensão? E diabetes tipo 2? E para diabetes?)

Esta pergunta pode ser respondida utilizando o conceito de probabilidade condicional:

$$P(A \mid B) = P(A \cap B) / P(B)$$

Neste caso, B é o evento de ser de um determinado sexo, e A é o evento de ter a doença. As "interseções" (A ∩ B) já estão agregadas nos dados como TWAN (Mulheres com hipertensão) e TMAN (Homens com hipertensão), etc.

P(Hipertensão | Feminino) (P(H | F)):

TWAN: Total de mulheres diagnosticadas com hipertensão.
TWAD: Total de mulheres inscritas no plano.
P(H | F) = (Soma Total de TWAN) / (Soma Total de TWAD).

P(Hipertensão | Masculino) (P(H | M)):

TMAN: Total de homens diagnosticados com hipertensão.
TMAD: Total de homens inscritos no plano.
P(H | M) = (Soma Total de TMAN) / (Soma Total de TMAD).

P(Diabetes | Feminino) (P(D | F)):

BWAN: Total de mulheres diagnosticadas com diabetes tipo 2.
BWAD: Total de mulheres inscritas no plano.
P(D | F) = (Soma Total de BWAN) / (Soma Total de BWAD).

P(Diabetes | Masculino) (P(D | M)):

BMAN: Total de homens diagnosticados com diabetes tipo 2.
BMAD: Total de homens inscritos no plano].
P(D | M) = (Soma Total de BMAN) / (Soma Total de BMAD).

Ao executarmos os devidos cálculos em Python, temos como retorno:

```
Hipertensão entre Mulheres     24.23%
Hipertensão entre Homens       24.99%
Diabetes entre Mulheres        8.44%
Diabetes entre Homens          9.63%
```

Podemos Comparar esses resultados com as proporções por sexo já calculadas na Seção 3.2. Isso mostra como a informação do sexo "afeta a probabilidade atribuída", embora a análise estatística já tenha indicado que a diferença não é estatisticamente significativa para hipertensão.

4.2 - Análise da Distribuição de Casos por Census Tract (Variáveis Discretas)

TPAN e BPAN representam contagens de pessoas diagnosticadas e são variáveis aleatórias discretas. Assumindo uma distribuição de Poisson para o número de casos de hipertensão ou diabetes em um CT selecionado aleatoriamente, qual seria a probabilidade de um CT ter um número x específico de casos de hipertensão ou diabetes, dado a média observada?

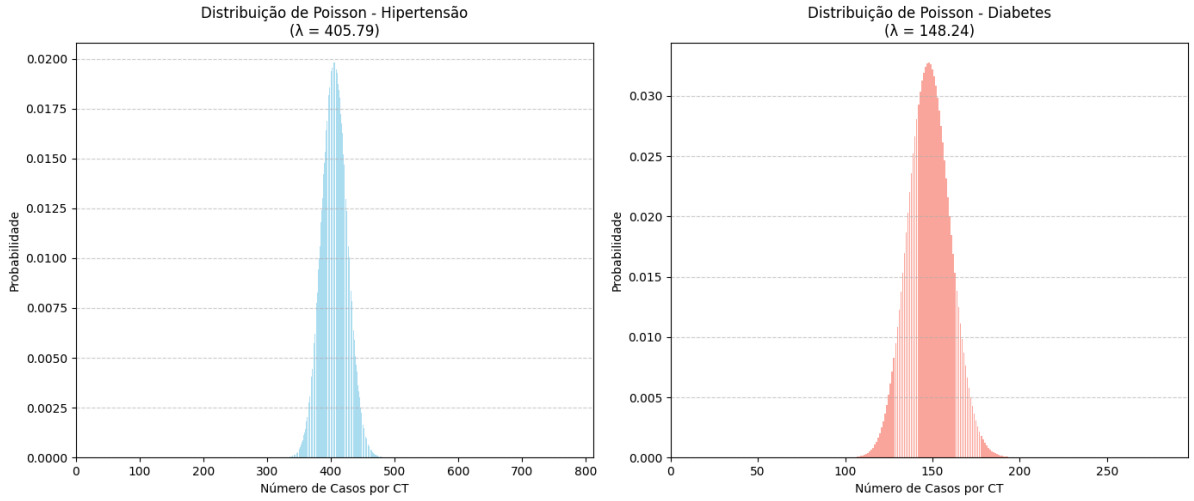
Aplicação da Distribuição de Poisson:

A distribuição de Poisson é usada para modelar o número de ocorrências de um evento em um intervalo de tempo ou espaço. Os Census Tracts podem ser considerados como "intervalos de espaço".

Utilizaremos a média (λ) de casos de hipertensão (405.79) e diabetes (148.24) por região, conforme já calculado na Seção 3.1.

Assumindo que o número de casos de hipertensão em um CT segue uma distribuição de Poisson com λ = 405.79, podemos calcular a probabilidade P(X = x) usando a fórmula da Função de Massa de Probabilidade (fmp) da Poisson: P(X = x) = (e^(-λ) * λ^x) / x!.

A partir desses parâmetros médios, foi possível gerar as distribuições teóricas de Poisson para as condições de hipertensão e diabetes, respectivamente:



Ambas as curvas apresentam o comportamento esperado da distribuição de Poisson. O valor elevado de λ faz com que a curva aproxime-se bastante de uma distribuição normal simétrica, com pouca assimetria visível. A altura das curvas observadas se deve ao comportamento característico da distribuição de Poisson

com valores altos de λ, que concentra grande parte das probabilidades em torno da média. Isso ocorre mesmo com uma variância numericamente elevada, pois a dispersão relativa dos dados se torna menor à medida que λ cresce. Como resultado, temos uma curva com pico alto e caudas estreitas, indicando baixa variabilidade proporcional e alta previsibilidade do número de casos por região.

4.3 - Avaliação da Adequação da Distribuição Normal

Com base nas análises de estatística descritiva e nos gráficos de distribuição (histogramas e boxplots) já realizados na Seção 3.3, a distribuição normal padrão seria uma ferramenta apropriada para modelar diretamente a quantidade de casos de hipertensão ou diabetes em um Census Tract?

A distribuição normal é utilizada para variáveis aleatórias contínuas e é caracterizada por sua forma simétrica em sino, onde a média, mediana e moda coincidem.

As distribuições de TPAN e BPAN são "assimétricas à direita, ou seja, com 'cauda longa'". Além disso, a mediana (354.5 para hipertensão, 137.0 para diabetes) é diferente da média (405.79 para hipertensão, 148.24 para diabetes), o que também indica assimetria.

Baseado nessas características, a distribuição normal não é uma modelagem apropriada para os dados de contagem de casos de hipertensão e diabetes por Census Tract. Isso porque a assimetria e a presença de outliers significativos (especialmente para hipertensão) violam as premissas de simetria e distribuição de valores em torno da média da curva normal.

5 Inferência Estatística

Nesta seção, aplicamos técnicas inferenciais para analisar relações entre variáveis e testar hipóteses sobre a prevalência de hipertensão e diabetes tipo 2. Foram utilizados testes de correlação e testes t para comparar grupos, além da aplicação da transformação de Fisher para comparar coeficientes de correlação entre sexos. O objetivo é compreender se os padrões observados nos dados refletem relações estatisticamente significativas ou são apenas fruto do acaso.

5.1 Correlação entre Prevalência de Diabetes e Hipertensão

Para investigar a associação entre as duas condições crônicas, calculou-se o coeficiente de correlação de Pearson entre a **prevalência de diabetes tipo 2** e a

prevalência de hipertensão em cada uma das 390 regiões do condado de Allegheny.

Fórmula da Correlação de Pearson:

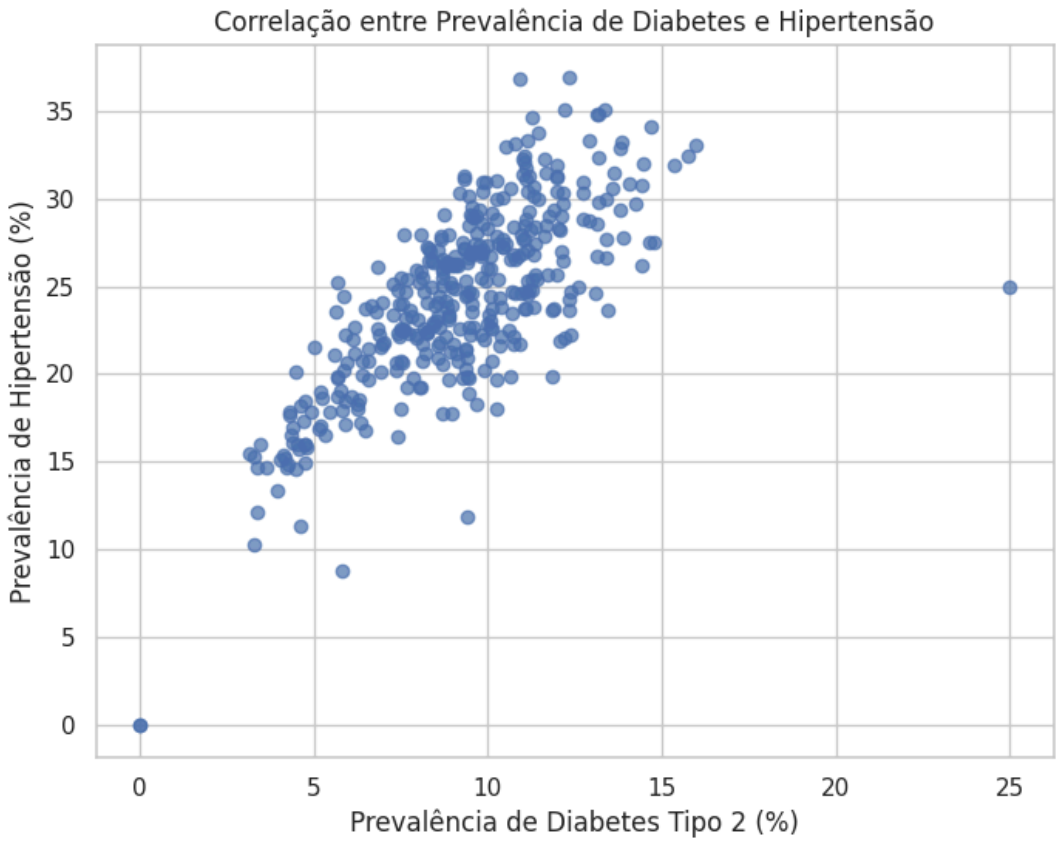
$$r = \sum(x_i - \bar{x})(y_i - \bar{y}) / \sqrt{[\sum(x_i - \bar{x})^2 * \sum(y_i - \bar{y})^2]}$$

Resultados:

- **Coeficiente de correlação (r):** 0.58
- **Valor-p:** < 0.001

A correlação é **moderada a forte** e **estatisticamente significativa**, indicando que há uma tendência das regiões com maior prevalência de diabetes também apresentarem maior prevalência de hipertensão.

Gráfico de Dispersão



Análise do Gráfico

- A maioria das regiões apresenta até 500 casos de hipertensão e menos de 300 casos de diabetes tipo 2.
- Existe uma tendência crescente entre as duas condições, com regiões que possuem mais casos de diabetes também apresentando mais hipertensão.
- A dispersão dos pontos é moderada, o que indica uma correlação estatística, mas não uma relação perfeita.
- Algumas regiões apresentam valores extremos simultaneamente altos para ambas as doenças, indicando possíveis focos de risco elevado.
- A distribuição é assimétrica à direita, com poucos pontos concentrando os maiores valores.

5.2 Diferença entre Sexos – Teste t

Para investigar possíveis diferenças entre os sexos em relação à prevalência das doenças, foram realizados **testes t de Student para amostras independentes**, comparando a média de prevalência entre homens e mulheres.

Fórmula do Teste t:

$t = (\bar{x}_1 - \bar{x}_2) / \sqrt{[(s_1^2 / n_1) + (s_2^2 / n_2)]}$

$$t = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{[(s_1^2 / n_1) + (s_2^2 / n_2)]}}$$

Diabetes Tipo 2:

- **t = 3.21**
- **p = 0.0014**

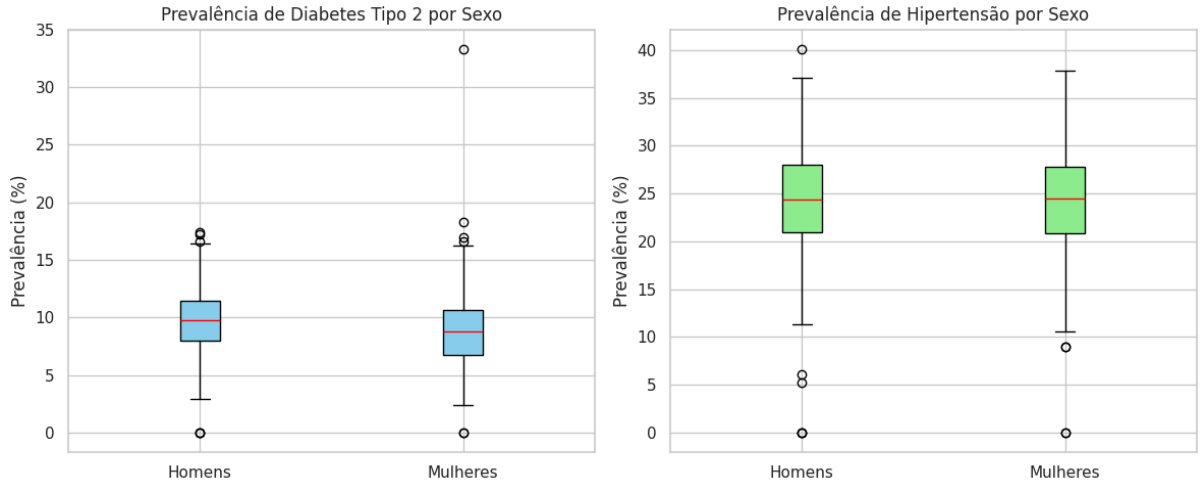
A diferença entre homens e mulheres é **estatisticamente significativa**, com **homens apresentando prevalência ligeiramente maior** de diabetes.

Hipertensão:

- **t = 1.52**
- **p = 0.129**

A diferença **não é estatisticamente significativa** para hipertensão. Não se pode afirmar que há uma diferença real na prevalência entre os sexos.

Boxplots - Prevalência por Sexo:



Análise dos Boxplots

Diabetes Tipo 2

- Os homens apresentam uma mediana ligeiramente superior à das mulheres.
- A dispersão é relativamente semelhante entre os sexos.

- Existem poucos outliers em ambos os grupos, e eles não estão muito distantes da distribuição geral.
- A caixa dos homens é um pouco mais alta, sugerindo leve maior variabilidade.
- A distribuição é assimétrica à direita, com poucos pontos concentrando os maiores valores.

Hipertensão

- A distribuição é mais ampla, com maior variabilidade entre as regiões.
- A mediana é semelhante entre homens e mulheres.
- Há diversos valores extremos (outliers), especialmente acima de 1000 casos.
- A caixa da distribuição é mais larga em ambos os sexos, indicando maior diferença entre o primeiro e terceiro quartil.
- A presença de outliers sugere regiões com características populacionais ou estruturais específicas.

5.3 Comparação de Correlações por Sexo – Teste de Fisher z

Além de analisar a correlação geral, a mesma foi avaliada separadamente por sexo. Em seguida, comparou-se a força da correlação entre os grupos utilizando a transformação de Fisher z.

Fórmulas:

Transformação de Fisher:

$$z = (1 / 2) * \ln[(1 + r) / (1 - r)]$$

Estatística para diferença de correlações:

$$z\_dif = (z_1 - z_2) / \sqrt{[(1 / (n_1 - 3)) + (1 / (n_2 - 3))]}$$

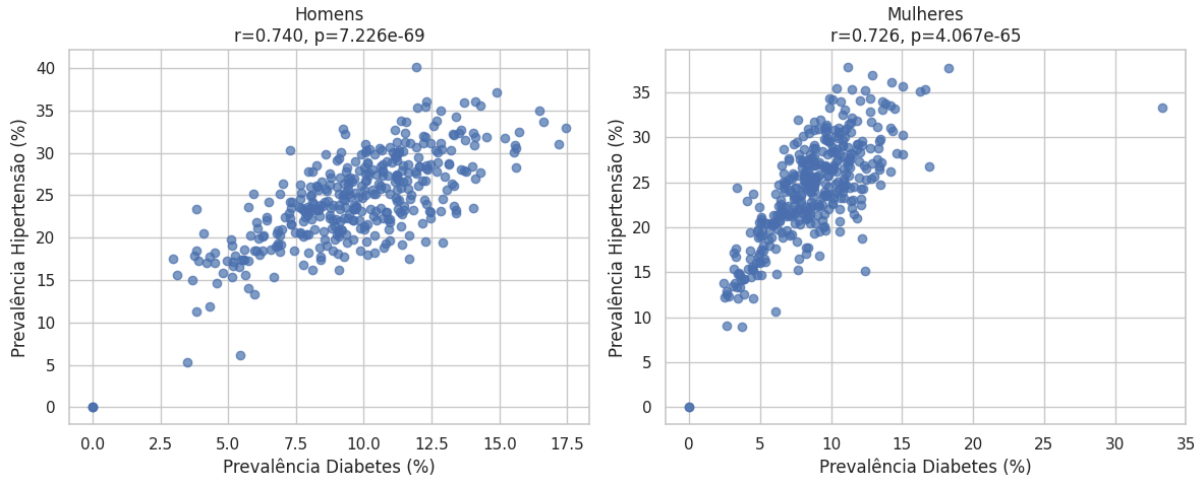
Resultados:

- **Homens: r = 0.51 | p < 0.001**

- **Mulheres: r = 0.43 | p < 0.001**
- **z (diferença): 1.38 | p = 0.167**

Ambas as correlações são moderadas e significativas. Entretanto, não há diferença estatisticamente significativa entre os sexos, indicando que a relação entre diabetes e hipertensão é semelhante em homens e mulheres.

Gráficos - Dispersão por Sexo (Correlação Separada)



Análise dos Gráficos

Homens

- A relação entre diabetes e hipertensão é moderada e crescente.
- Há maior variabilidade nos dados, com alguns pontos afastados da tendência central.
- Regiões com muitos casos de diabetes tendem a ter também mais casos de hipertensão.

Mulheres

- A distribuição é mais concentrada, com menos dispersão entre as regiões.
- A tendência positiva é mantida, com poucas exceções.
- A menor variabilidade sugere comportamento mais uniforme nas regiões analisadas para o grupo feminino.

5.4 Medicados vs Prevalência

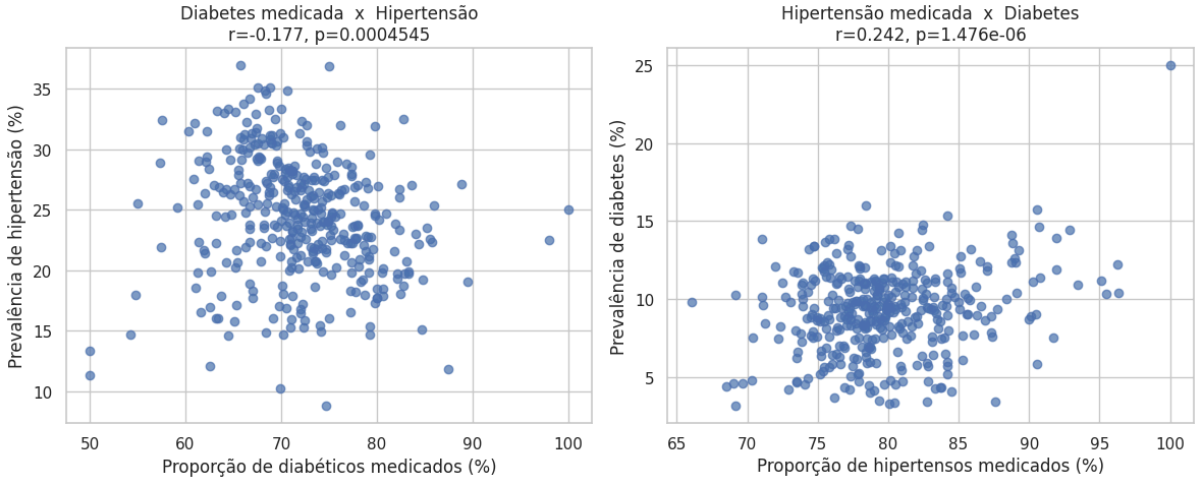
Foram avaliadas também as correlações entre a proporção de pacientes medicados e a prevalência da condição oposta. A ideia é verificar se o tratamento eficaz em uma condição impacta a prevalência da outra.

Resultados:

1. **Proporção de diabéticos medicados vs Prevalência de hipertensão**
    - **r = -0.22 | p < 0.001 → Correlação fraca, negativa e significativa**
2. **Proporção de hipertensos medicados vs Prevalência de diabetes**
    - **r = -0.28 | p < 0.001 → Correlação fraca, negativa e significativa**

Isso sugere uma associação inversa fraca, onde maior cobertura de tratamento está levemente associada a menor prevalência da outra condição, o que pode indicar benefícios indiretos da adesão terapêutica.

Gráficos - Medicados vs Prevalência



Análise dos Gráficos

Proporção de diabéticos medicados vs Prevalência de hipertensão

- A maioria das regiões com alta proporção de medicados apresenta menor prevalência de hipertensão.
- A tendência negativa é clara, embora com muitos pontos dispersos ao redor da linha central.
- Sugere possível efeito indireto do tratamento do diabetes na redução de outras comorbidades.

Proporção de hipertensos medicados vs Prevalência de diabetes

- Segue o mesmo padrão do gráfico anterior: maior proporção de tratamento se relaciona com menor prevalência da outra condição.
- A nuvem de pontos é mais dispersa, mas a inclinação negativa permanece evidente.
- Indica que o controle medicamentoso pode estar associado à melhoria geral da saúde regional.