INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DE SÃO PAULO
TADS – ESTATÍSTICA E PROBABILIDADE

ALYSON CÉSAR FUMAGALLI DOS SANTOS JÚNIOR	(SP3121071)
ELIEL DA SILVA	(SP3121054)
HENRIQUE SANTIAGO PIRES	(SP312262X)
HENRRIKY JHONNY DE OLIVEIRA BASTOS	(SP3123103)

PROJETO FINAL: ANÁLISES DE ESTATÍSTICA E PROBABILIDADE

Relatório

São Paulo – SP

2025

ALYSON CÉSAR FUMAGALLI DOS SANTOS JÚNIOR (SP3121071)
ELIEL DA SILVA (SP3121054)
HENRIQUE SANTIAGO PIRES (SP312262X)
HENRRIKY JHONNY DE OLIVEIRA BASTOS (SP3123103)

PROJETO FINAL: ANÁLISES DE ESTATÍSTICA E PROBABILIDADE

Relatório

Projeto entregue à docente da disciplina Estatística e Probabilidade, do curso de Tecnologia em Análise e Desenvolvimentos de Sistemas do Instituto Federal de São Paulo, como requisito parcial para aprovação na disciplina.

Orientadora: Josceli Maria Tenorio

São Paulo – SP

2025

SUMÁRIO

1. SISTEMA DE INFORMAÇÃO SOBRE MORTALIDADE (SIM)

1.1 Análises Descritivas

1.1.1 Análise 1: As taxas de realização de autópsia diferem entre óbitos por homicídio e por acidente?

1.1.1.1 Objetivos e preparações iniciais

Essa análise tem o objetivo de verificar se existe diferença na taxa de realização de necropsia para confirmação da causa de morte entre os casos registrados como homicídio e os casos registrados como acidente. Também foi analisado o impacto que essa informação pode trazer. Segue a descrição das colunas utilizadas no código e contexto da análise:

- **CIRCOBITO:** refere-se à circunstância do óbito.
  - **1:** Acidente
  - **3:** Homicídio
- **NECROPSIA:** indica se o óbito foi realizado necropsia.
  - **1:** Sim

Os demais valores possíveis das colunas foram descartados ou ignorados por não se enquadrarem na investigação. É válido ressaltar que os dados utilizados passaram por um tratamento prévio para preparar os registros de interesse.

1.1.1.2 Explicações relacionadas ao código desenvolvido

Após o tratamento dos dados, o conjunto foi segmentado em dois grupos distintos com base na coluna CIRCOBITO: um para homicídios e outro para acidentes.

Para cada um desses grupos, foi calculada a taxa de óbitos em que foram realizadas necropsia (NECROPSIA == 1) em relação ao total de óbitos daquela categoria. As taxas de "Não Confirmados" foram obtidas pelo cálculo da taxa complementar. Lembrando que a necrópsia é um procedimento para confirmar as causas da morte.

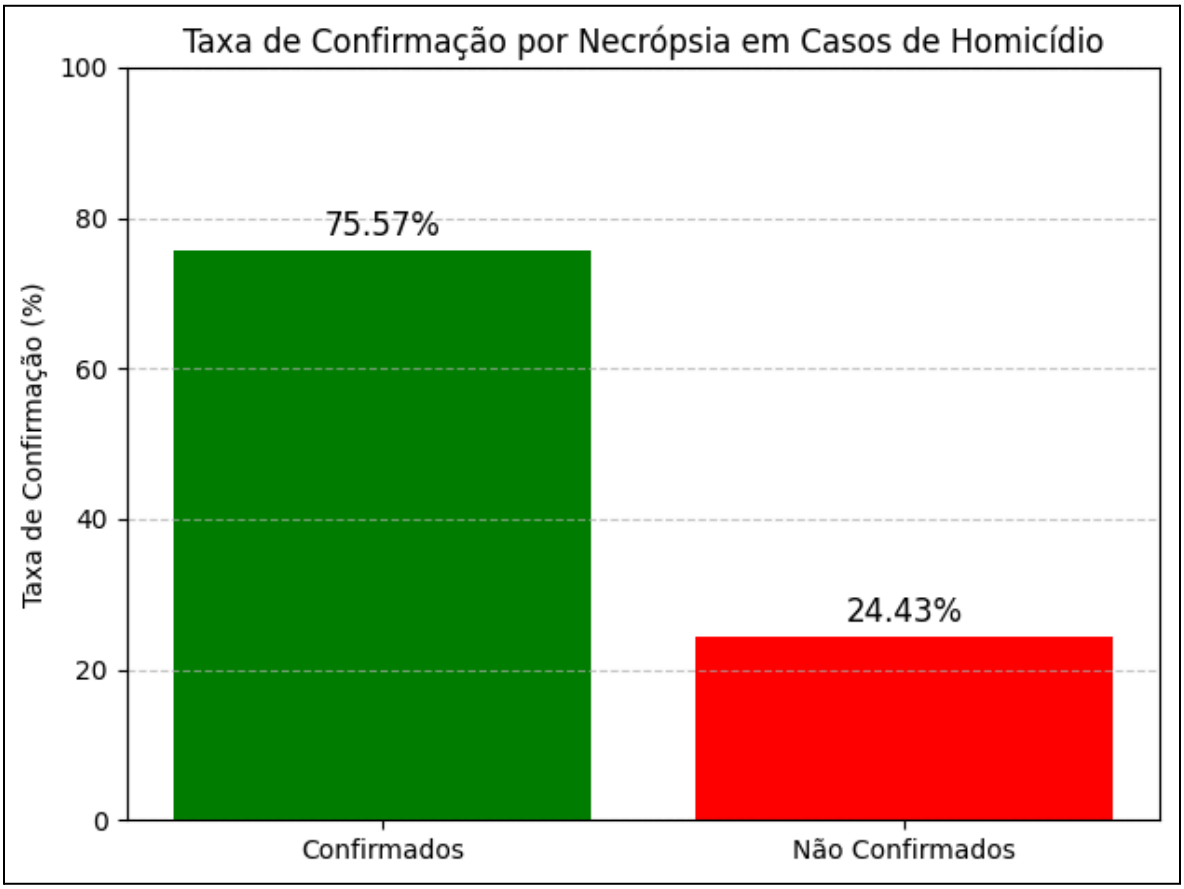
Não foi analisada a coluna que trata sobre a “alteração de causa de morte” por conter poucos registros na amostra. Esse poderia ser o comportamento normal (real) dos dados, mas seria necessário uma análise feita com a base de dados completa (tanto de 2021 como de outros anos) para se confirmar isso.

Por fim, para melhor ilustração dos resultados, foram gerados dois gráficos de barras distintos. Cada gráfico compara a taxa de casos "Confirmados" e "Não Confirmados" por necropsia para uma das circunstâncias de óbito (homicídio e acidente).

1.1.1.3 Conclusões

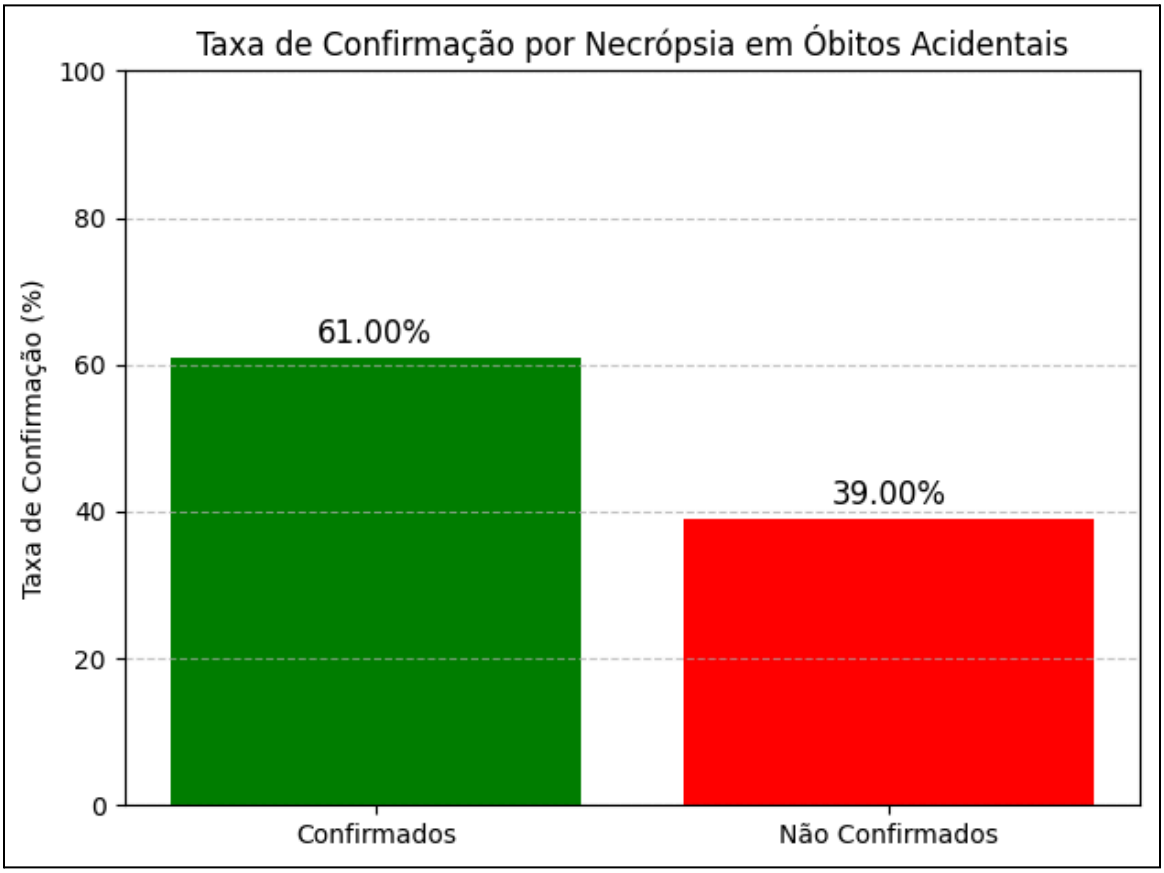
A partir da análise dos dados, percebe-se uma disparidade clara entre as taxas de confirmação por necropsia.

Nos casos de homicídio, a taxa de confirmação é de aproximadamente 75%, indicando que a grande maioria desses óbitos passa por uma investigação de autópsia. Consequentemente, 25% dos casos não são confirmados pelo mesmo método.



Por outro lado, a taxa de confirmação para óbitos decorridos de acidentes é significativamente mais baixa, de aproximadamente 39%. Este valor pode ser considerado baixo e, como apontado na análise prévia, levanta a possibilidade de que uma parcela de mortes por

outras causas possa ser incorretamente classificada como acidente devido à ausência de uma investigação aprofundada por necropsia.



Portanto, conclui-se que a realização de necropsia é um procedimento consideravelmente mais frequente em investigações de homicídio do que em óbitos registrados como acidentais e também se assumirmos que a necropsia confirma o motivo de óbito, há uma margem considerável para erros nos diagnósticos.

1.1.2 Análise 2: A taxa de suicídio varia conforme o estado civil?

1.1.2.1 Objetivos e preparações iniciais

O objetivo desta análise é investigar se a taxa de suicídio varia significativamente entre os diferentes estados civis presentes no dataset. A análise busca identificar qual grupo possui a maior taxa proporcional de suicídio.

Segue a descrição das colunas utilizadas, conforme o dicionário de dados (SIM) e o código:

- **ESTCIV:** Refere-se ao estado civil do indivíduo.
    - **1:** Solteiro(a)
    - **2:** Casado(a)
    - **3:** Viúvo(a)
    - **4:** Separado(a)
    - **5:** União Estável
    - **9:** Ignorado
- **CIRCOBITO:** Refere-se à circunstância do óbito.
    - **1:** Acidente
    - **2:** Suicídio
    - **3:** Homicídio
    - **4:** outros
    - **9:** ignorado

Para esta investigação, foram considerados apenas os registros com códigos de estado civil válidos (de 1 a 5). Outros valores e registros com dados ausentes foram descartados do escopo da análise.

1.1.2.2 Explicações relacionadas ao código desenvolvido

Após a filtragem inicial para garantir a validade dos dados de estado civil, a análise seguiu duas etapas principais. Primeiro, foi calculado o número total de indivíduos para cada categoria de estado civil. Em seguida, foi calculado o número de óbitos por suicídio (CIRCOBITO == 2) para cada uma dessas mesmas categorias.

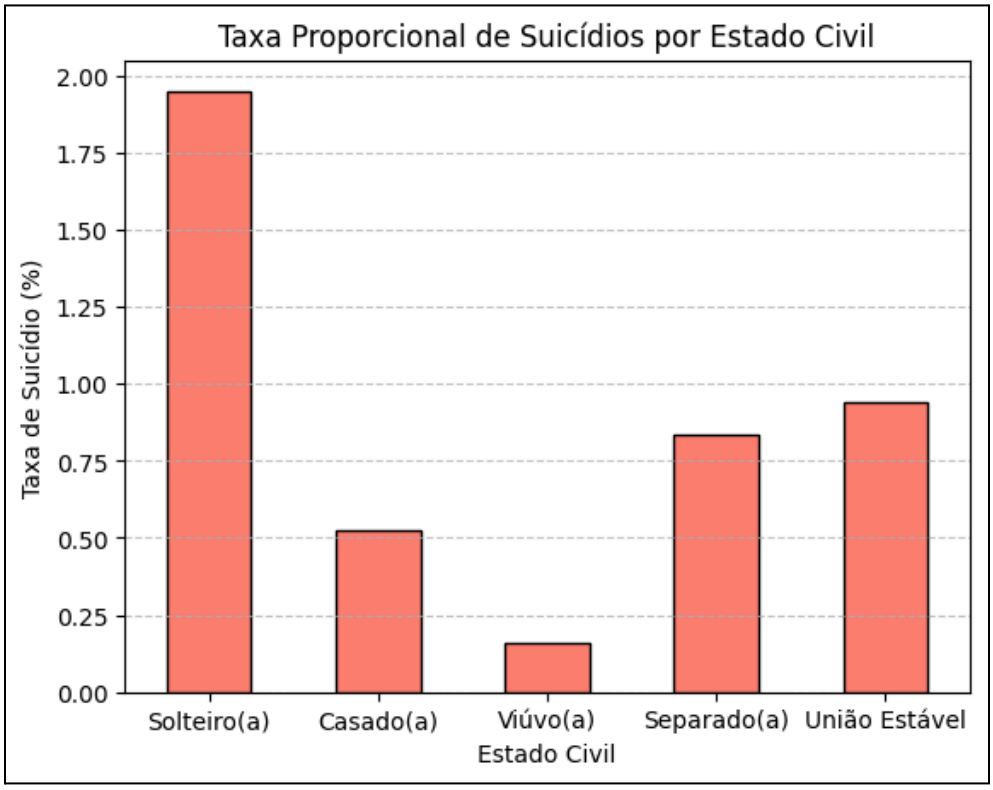
O passo crucial foi o cálculo da taxa proporcional de suicídio. Para cada estado civil, o número de suicídios foi dividido pelo número total de pessoas naquele mesmo estado civil, e o resultado foi multiplicado por 100 para obter uma porcentagem. A análise proporcional elimina o problema de haver quantidades maiores de determinadas categorias.

Por fim, para uma visualização clara dos resultados, foi gerado um gráfico de barras que compara as taxas percentuais de suicídio entre os diferentes estados civis. Os códigos numéricos de ESTCIV foram mapeados para seus respectivos nomes para facilitar a interpretação do gráfico.

1.1.2.3 Conclusões

A análise das taxas proporcionais revela que o estado civil "Solteiro(a)" apresenta a maior taxa de suicídio entre os grupos estudados. No entanto, é fundamental interpretar esse dado com cautela. A categoria "Solteiro(a)" é demograficamente ampla, englobando crianças, adolescentes e jovens adultos que ainda não se casaram, o que naturalmente a torna o maior grupo populacional na amostra. Além disso, relacionamentos não formalizados, como namoros, não são capturados pelos registros oficiais, o que constitui uma limitação da análise.

Logo após "Solteiro(a)", os estados civis "Separado(a)" e "União Estável" também se destacam com taxas de suicídio elevadas em comparação com os demais grupos.



Portanto, pode-se concluir que, embora os solteiros representem a maior taxa, os grupos de separados e em união estável também demonstram uma vulnerabilidade notável, com taxas de suicídio muito altas. Se nos atentarmos ao fato que a amostra possui 50.000 registros e já apresentou essa taxa, podemos interpretar essa situação como extremamente ruim.

Por outro lado, a interpretação dos dados brutos deve ser direcionada pelos fatores demográficos e pela falta de granularidade nos tipos de relacionamento (caso dos namoros, por exemplo).

1.1.3 Análise 3: Existe relação entre a semana de gestação e a ocorrência de óbitos maternos?

1.1.3.1 Objetivos e preparações iniciais

O objetivo desta análise é investigar a distribuição de óbitos maternos ao longo das semanas de gestação para identificar se existem períodos de maior risco durante a gravidez. Segue a descrição da coluna utilizada na análise:

- **SEMAGESTAC:** Refere-se ao número de semanas de gestação no momento do registro do evento.

Para garantir a relevância dos dados, a análise foi restrita a registros com duração de gestação inferior a 50 semanas. Essa filtragem é necessária pois o tempo médio de uma gravidez humana é de aproximadamente 40 semanas, e valores superiores a 50 são extremamente raros, podendo representar erros de digitação ou casos anômalos que distorceriam a análise geral, além dos valores 99 que são usados quando o dado é ignorado.

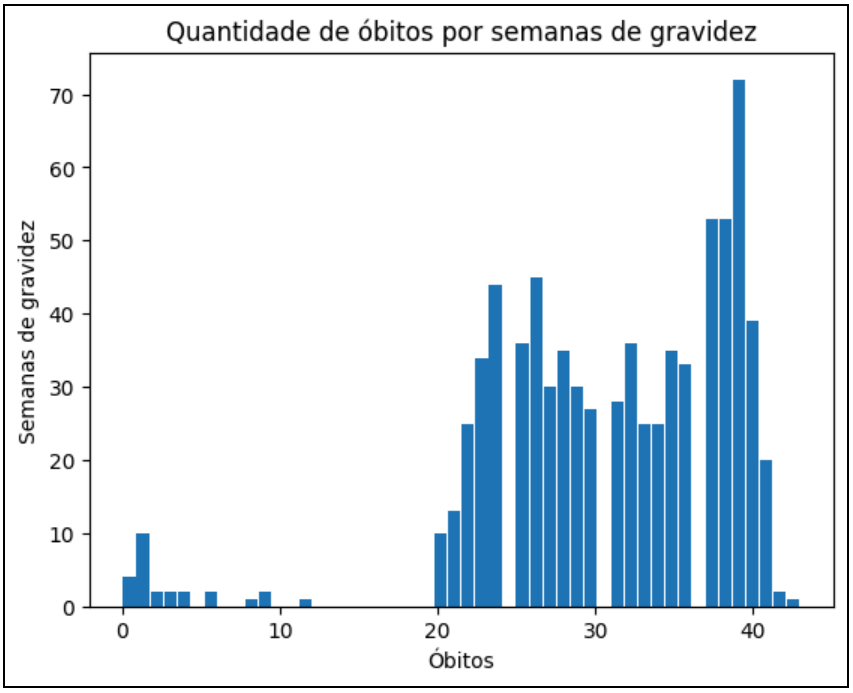
1.1.3.2 Explicações relacionadas ao código desenvolvido

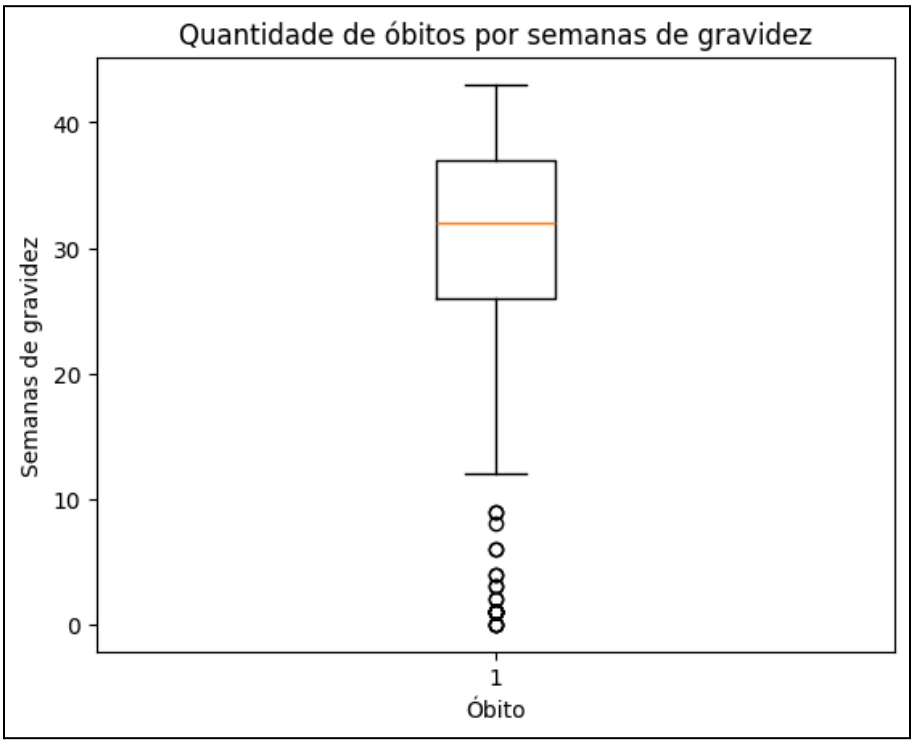
Após a etapa de filtragem dos dados para incluir apenas gestações com menos de 50 semanas, foram utilizadas três abordagens para analisar a distribuição dos óbitos:

- **Histograma:** Foi gerado um histograma com 50 classes (bins) para visualizar a frequência de óbitos em cada semana de gestação. Este gráfico permite identificar picos e padrões na distribuição ao longo do tempo.
- **Boxplot:** Foi criado um gráfico de boxplot para resumir estatisticamente a distribuição, mostrando a mediana, os quartis (25% e 75%), e os limites inferior e superior, ajudando a identificar a concentração central dos dados e possíveis outliers.
- **Estatísticas Descritivas:** Por fim, foi utilizado o método .describe() para calcular as principais métricas estatísticas da amostra, como média, mediana, desvio padrão, mínimo e máximo, fornecendo uma base numérica para as conclusões.

1.1.3.3 Conclusões

A análise dos dados permite identificar dois períodos críticos com maior incidência de óbitos durante a gestação.





O primeiro é um pico notável **no início da gravidez** (primeiras semanas). Esse fenômeno pode estar associado a complicações como alterações cromossômicas do embrião ou reações

imunológicas que levam a abortos espontâneos precoces, resultando em fatalidades para a gestante.

O segundo e mais proeminente período de risco ocorre na **segunda metade da gravidez**, aproximadamente a partir da 20ª semana, com uma concentração crescente de óbitos que atinge seu auge por volta das 35-40 semanas. A média de óbitos ocorre bem próxima disso, com aproximadamente 31 semanas. Este padrão é consistente com o fato de que complicações mais severas da gravidez, incluindo as relacionadas ao parto, tendem a surgir quando a gestação está em um estágio mais avançado.

Portanto, conclui-se que os riscos de óbito materno não são uniformes ou seguem algum tipo de padrão (ascendente, decrescente ou exponencial), mas se concentram notavelmente no início e, de forma mais acentuada, no final da gestação.

1.1.4 Análise 4: Qual a relação entre o peso do recém-nascido e a ocorrência de óbitos?

1.1.4.1 Objetivos e preparações iniciais

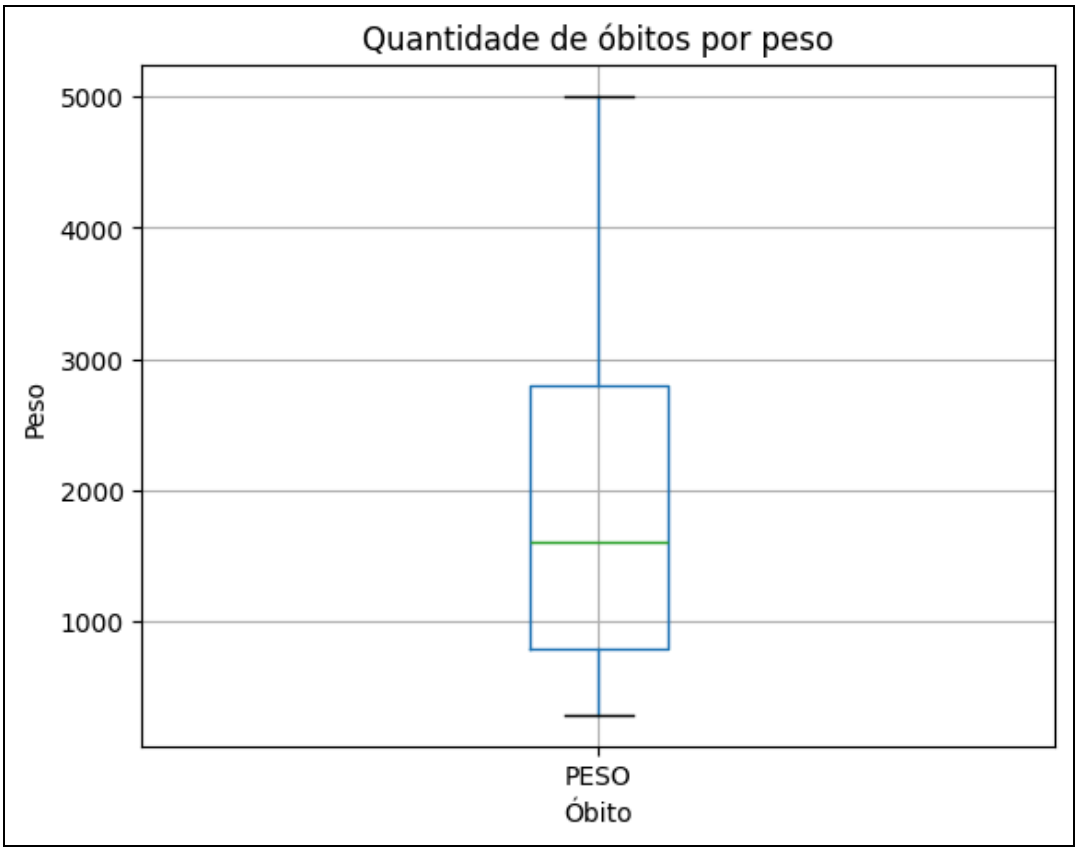
Esta análise tem como objetivo investigar a distribuição de óbitos de recém-nascidos com base em seu peso ao nascer, a fim de identificar as faixas de peso que apresentam maior risco de mortalidade. Segue a descrição da coluna utilizada na análise:

- **PESO:** Refere-se ao peso do recém-nascido no momento do nascimento, medido em gramas.

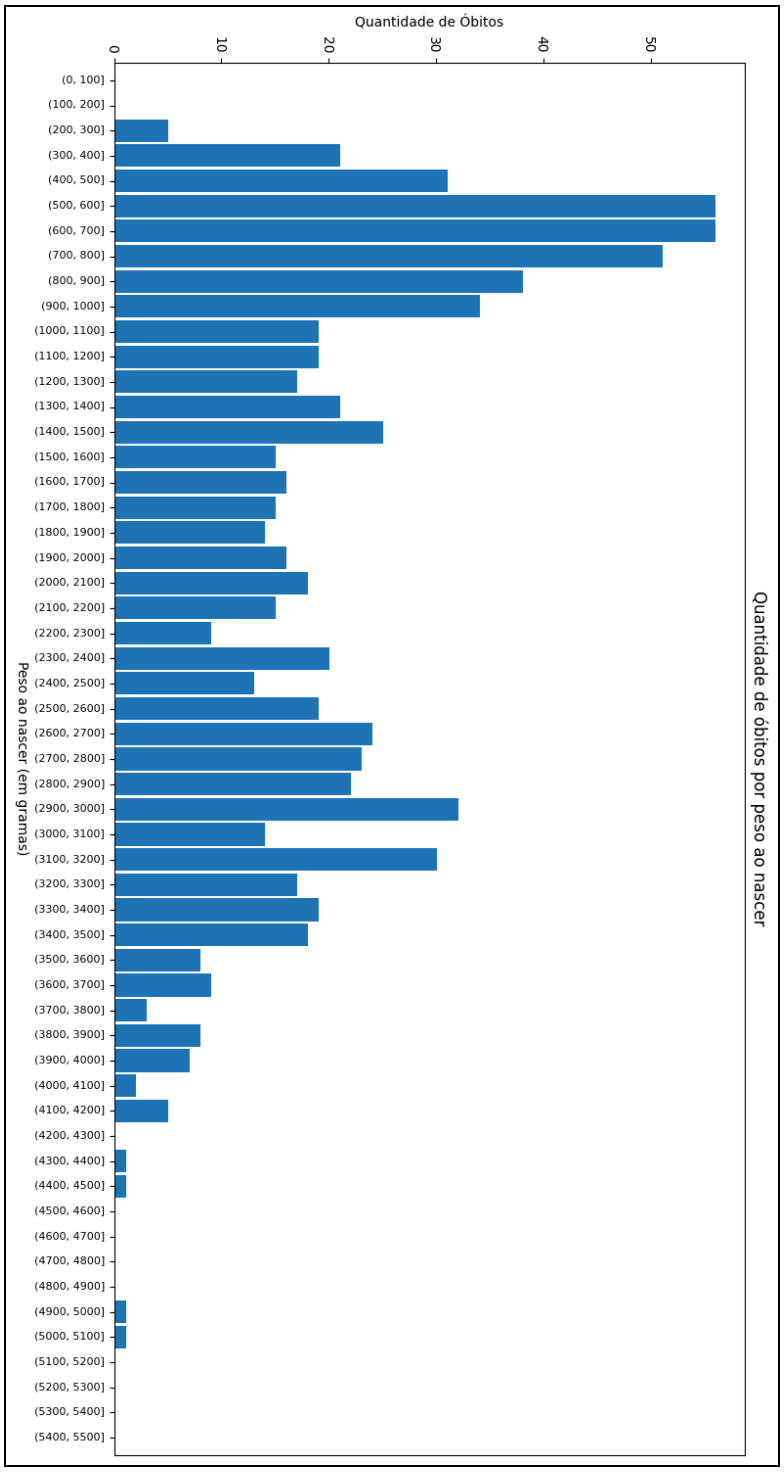
A análise utiliza todos os registros disponíveis da coluna PESO, sem uma filtragem prévia explícita, para abranger todo o espectro de pesos e identificar os padrões de mortalidade em cada faixa.

1.1.4.2 Explicações relacionadas ao código desenvolvido

A análise dos dados foi realizada por meio de duas visualizações gráficas principais. Primeiramente, foi gerado um **boxplot** para oferecer uma visão resumida da distribuição dos pesos, destacando a media, os quartis e a presença de valores atípicos.



Em seguida, para uma análise mais detalhada, os dados de peso foram segmentados em **faixas de 100 gramas**, desde 0 até 5500 gramas. O número de óbitos em cada uma dessas faixas foi então contado.



O resultado dessa contagem foi utilizado para gerar um **gráfico de barras**, que ilustra a quantidade de óbitos para cada intervalo de peso. Esta abordagem permite uma identificação visual clara das faixas de peso com maior e menor frequência de óbitos.

1.1.4.3 Conclusões

A análise dos gráficos revela dois pontos de concentração de óbitos, que indicam diferentes fatores de risco.

O pico mais expressivo de mortalidade ocorre na faixa de peso entre **300 e 1000 gramas**. Este resultado evidencia que o baixo peso, frequentemente associado a partos prematuros e outras complicações neonatais, representa um fator de risco crítico e aumenta significativamente as chances de óbito do recém-nascido.

Observa-se também um segundo pico, menos protuberante, na faixa de **3000 a 3400 gramas**. Considerando que a média de peso de um recém-nascido saudável é de aproximadamente 3400 gramas (conforme fontes de referência como a Babysec), este pico provavelmente reflete o maior volume de nascimentos que ocorrem naturalmente nesta faixa de peso, resultando em um número absoluto maior de óbitos, ainda que a taxa de mortalidade proporcional possa ser menor que a do grupo de baixo peso.

É possível perceber uma certa semelhança com a análise anterior sobre semanas de gestação, que obviamente acontece pelo desenvolvimento do feto durante a gestação, culminando no aumento de peso.

Portanto, conclui-se que recém-nascidos com peso muito baixo enfrentam um risco desproporcionalmente alto de mortalidade.

1.1.5 Análise 5: Existe um padrão na distribuição de óbitos ao longo das horas do dia?

1.1.5.1 Objetivos e preparações iniciais

O objetivo desta análise é investigar a frequência dos óbitos para cada uma das 24 horas do dia. A intenção é verificar se a distribuição é uniforme ou se existem períodos específicos (manhã, tarde, noite, madrugada) que concentram uma maior ou menor quantidade de fatalidades.

Segue a descrição da coluna utilizada na análise:

- **HORAOBITO:** Contém a informação do horário em que o óbito foi registrado.

Para realizar a análise, foi necessário um tratamento prévio nesta coluna. O dado, originalmente em um formato que pode variar, foi convertido para um objeto de data e hora

(datetime) para então se extrair de forma precisa apenas a componente da "hora" (um valor numérico de 0 a 23). Esta nova informação foi armazenada em uma coluna chamada hora.

1.1.5.2 Explicações relacionadas ao código desenvolvido

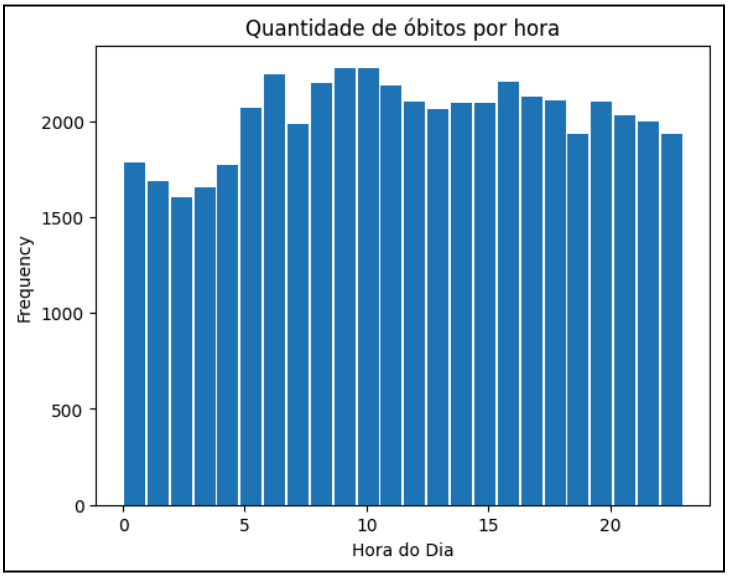
O processo de análise iniciou com a transformação e extração da hora a partir da coluna HORAOBITO, conforme descrito acima.

Posteriormente, foi gerado um **histograma** a partir da coluna hora. O uso de bins=24 (24 classes) é um parâmetro crucial, pois garante que cada barra do gráfico represente exatamente uma hora do dia. Isso permite uma visualização direta e clara da frequência de óbitos para cada um desses 24 intervalos de tempo. O gráfico foi customizado com títulos e rótulos para facilitar a sua interpretação.

1.1.5.3 Conclusões

A análise do histograma indica que a distribuição de óbitos ao longo do dia é, em sua maioria, **relativamente estável**, sem picos extremos que apontem para um horário de risco acentuado. A quantidade de fatalidades registradas durante as horas comerciais (manhã e tarde) se mantém em um patamar constante.

No entanto, a principal observação é a presença de um **vale sutil, porém notável, durante a madrugada**, especificamente no intervalo entre 0h e 5h. Neste período, a frequência de óbitos registrados é visivelmente menor em comparação com o resto do dia.



Portanto, conclui-se que, embora não haja "horas de pico" para a mortalidade, existe uma tendência de haver menos registros de óbitos durante as horas de menor atividade humana e de sono profundo. Isso pode sugerir que muitos eventos fatais estão ligados às atividades diurnas ou que óbitos ocorridos durante o sono só são oficialmente registrados horas mais tarde, quando são de fato percebidos.

1.2. Análises de Probabilidade

1.2.1. Análise 1: Qual é a probabilidade conjunta de um óbito ser de uma mulher cuja causa básica foi neoplasia maligna?

1.2.1.1. Objetivos e preparações iniciais

Esta análise tem como objetivo calcular a probabilidade conjunta de um registro de óbito ser ao mesmo tempo de uma pessoa do sexo feminino e com causa básica classificada como neoplasia maligna, conforme o padrão de códigos da CID-10, grupo C00 a C97.

Os dados utilizados pertencem ao dataset do Sistema de Informação sobre Mortalidade (SIM), disponibilizado pelo OpenDataSUS, e referem-se ao ano de 2021. As colunas consideradas foram:

- SEXO_DESC: contém o sexo da pessoa falecida (utilizado apenas o valor "F" para mulheres);
- CAUSABAS: apresenta o código da causa básica do óbito, segundo a CID-10.

Para a análise, foram consideradas apenas as linhas com valores válidos nessas duas colunas. As demais categorias e registros com dados ausentes ou inválidos foram removidos no processo de limpeza prévia.

A análise foi realizada com uma amostra de tamanho n_amostral, selecionada por limitações de memória computacional no ambiente Google Colab. Ainda assim, entende-se que essa amostra é suficientemente representativa para obter estimativas confiáveis da probabilidade conjunta.

1.2.1.2. Explicações relacionadas ao código desenvolvido

Inicialmente, foi definida uma expressão regular (regex) para filtrar os códigos da CID-10 que indicam neoplasias malignas – ou seja, os códigos no intervalo C00 a C97. A seguir,

aplicou-se esse filtro à coluna CAUSABAS, em conjunto com um segundo filtro que seleciona registros em que SEXO_DESC é igual a "F", identificando os óbitos de mulheres.

Com isso, foi possível determinar a interseção entre os dois eventos: sexo feminino e causa básica de óbito por neoplasia maligna. A quantidade de registros que satisfazem ambas as condições foi então dividida pelo tamanho da amostra (n_amostral), resultando na probabilidade conjunta P(F ∩ Neoplasia).

O valor obtido foi exibido tanto em notação decimal quanto percentual. Além disso, foi elaborado um Além disso, foi elaborado um diagrama de Venn de dois conjuntos, representando a sobreposição entre:

- Mulheres
- Óbitos por neoplasias malignas

  Essa visualização permite compreender as interseções entre os grupos analisados.

1.2.1.3. Conclusões

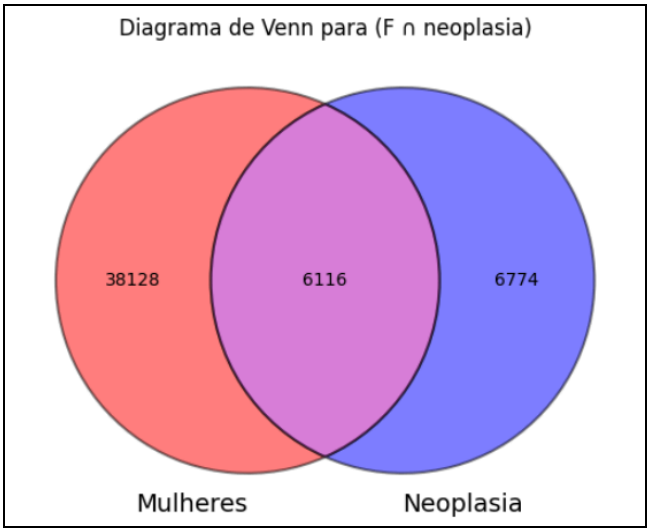
Com base na amostra analisada, a probabilidade conjunta de um óbito registrado ser de uma mulher cuja causa básica da morte foi uma neoplasia maligna (C00–C97) é de aproximadamente:

$$P(F \cap Neoplasia) = 0.0612 \rightarrow 6.12\%$$

Esse resultado representa a chance estimada de ocorrência simultânea dos dois eventos em questão, conforme observado nos dados analisados do SIM.

Ele poderá ser comparado posteriormente com outras probabilidades condicionais ou marginais, auxiliando em interpretações mais amplas sobre o perfil da mortalidade por neoplasias no país.

O diagrama de Venn reflete esse resultado, mostrando a intersecção entre os conjuntos de mulheres e óbitos por neoplasias malignas.



1.2.2. Análise 2: Qual a probabilidade de um homicídio ter vitimado uma mulher ou um homem?

1.2.2.1. Objetivos e preparações iniciais

Esta análise tem como objetivo verificar a probabilidade conjunta de um óbito registrado como homicídio ter ocorrido com pessoas do sexo feminino ou masculino.

A investigação utiliza dados do Sistema de Informação sobre Mortalidade (SIM) referentes ao ano de 2021, com uma amostra restrita de n_amostral registros, previamente filtrados para conter apenas valores válidos.

As colunas analisadas foram:

- SEXO_DESC: identifica o sexo da pessoa falecida (F para mulheres e M para homens);
- CIRCOBITO_DESC: descreve as circunstâncias do óbito, sendo considerado nesta análise apenas o valor "Homicídio".

Foram desconsiderados registros com valores ausentes, vazios ou fora dos critérios definidos, como outros sexos ou causas não classificadas como homicídio. O processo de limpeza de dados foi fundamental para garantir a consistência dos resultados.

1.2.2.2. Explicações relacionadas ao código desenvolvido

Inicialmente, foi criado um filtro para identificar todos os registros cuja circunstância da morte foi "Homicídio". Em seguida, foram aplicados filtros separados para registros de mulheres (SEXO_DESC == 'F') e de homens (SEXO_DESC == 'M').

Com a combinação dos filtros, foi possível identificar os conjuntos de:

- Homicídios que vitimaram mulheres (interseccao_mulher_homicidio)
- Homicídios que vitimaram homens (interseccao_homem_homicidio)

As probabilidades conjuntas foram então calculadas dividindo o número de registros em cada interseção pelo tamanho total da amostra (n_amostral). Os resultados foram expressos tanto em notação decimal quanto percentual.

Além disso, foi elaborado um diagrama de Venn de três conjuntos, representando a sobreposição entre:

- Mulheres
- Homens
- Homicídios

Essa visualização permite compreender as interseções entre os grupos analisados, especialmente a quantidade relativa de homicídios em cada sexo.

1.2.2.3. Conclusões

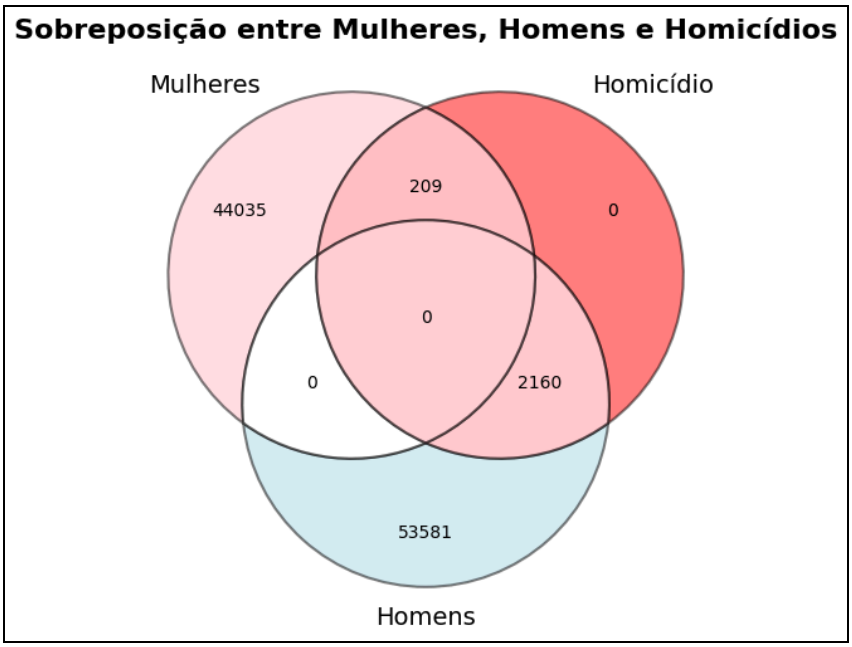
Com base nos dados amostrais, as probabilidades conjuntas obtidas foram:

$$P(F \cap Homicídio) = 0.0021 \rightarrow 0.21\%$$

$$P(M \cap Homicídio) = 0.0216 \rightarrow 2.16\%$$

Esses valores indicam que a chance de um homicídio registrado ter vitimado uma mulher é significativamente diferente da chance de ter vitimado um homem, conforme os dados analisados.

O diagrama de Venn reflete essa diferença visualmente, mostrando uma maior sobreposição entre os conjuntos de homens e homicídios, reforçando o achado de que a maioria dos homicídios registrados vitimaram pessoas do sexo masculino.



Assim, a análise indica uma disparidade expressiva na distribuição dos homicídios entre os sexos, evidenciando que os homens representam uma parcela consideravelmente maior das vítimas desse tipo de ocorrência letal no período analisado.

1.2.3. Análise 3: Qual é a probabilidade de um óbito estar relacionado a acidente de trabalho ou ter ocorrido em estabelecimento comercial?

1.2.3.1. Objetivos e preparações iniciais

Esta análise tem como objetivo calcular a probabilidade da união entre dois eventos relacionados ao local e à causa do óbito registrados no Sistema de Informação sobre Mortalidade (SIM), do OpenDataSUS, para o ano de 2021. Os dois eventos analisados são:

- O óbito estar relacionado a um acidente de trabalho;
- O óbito ter ocorrido em um estabelecimento comercial.

Para isso, foram utilizados os seguintes campos do dataset:

- ACIDTRAB_DESC: indica se o óbito foi decorrente de acidente de trabalho. Foi considerado o valor "Sim";
- TPOBITOCOR_DESC: indica o tipo de local onde o óbito ocorreu. Foi considerado o valor "Estabelecimento comercial".

Registros com valores ausentes, vazios ou diferentes dos especificados nessas colunas foram descartados. A análise foi realizada com uma amostra de tamanho n_amostral, recorte necessário por limitações técnicas do ambiente Google Colab.

1.2.3.2. Explicações relacionadas ao código desenvolvido

Foram aplicados dois filtros independentes:

- filtro_acidente: seleciona os óbitos decorrentes de acidentes de trabalho;
- filtro_estabelecimento: seleciona os óbitos ocorridos em estabelecimentos comerciais.

Para calcular a probabilidade da união dos dois eventos – ou seja, a chance de um registro de óbito satisfazer pelo menos um dos dois critérios –, foi aplicada a operação lógica de união (OR) entre os dois filtros. O número de registros resultantes foi dividido pelo total da amostra (n_amostral), resultando na probabilidade P(A ∪ B), com A representando "acidente de trabalho" e B representando "estabelecimento comercial".

Além disso, foi elaborado um diagrama de Venn com dois conjuntos, visualizando a sobreposição (interseção) entre os óbitos relacionados a acidentes de trabalho e aqueles ocorridos em estabelecimentos comerciais. Os círculos do Venn foram padronizados para tamanhos iguais para facilitar a comparação proporcional.

1.2.3.3. Conclusões

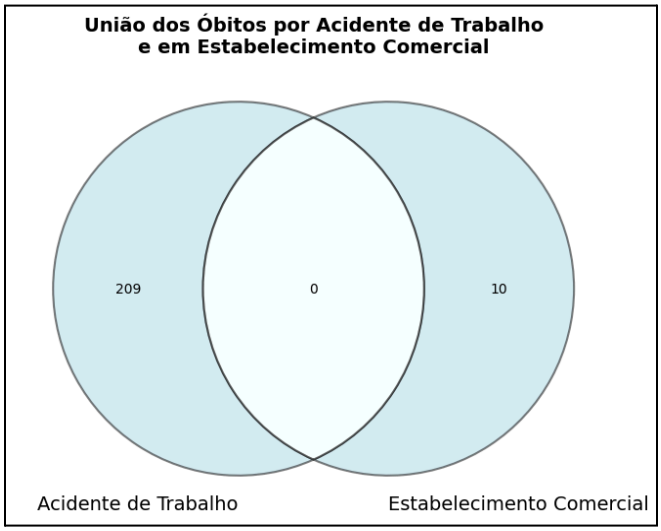
A análise indicou que a probabilidade de um óbito estar relacionado a um acidente de trabalho ou ter ocorrido em estabelecimento comercial foi:

$$P(Acidente\ de\ Trabalho\ \cup\ Estabelecimento\ Comercial) = 0.0022 \rightarrow 0.22\%$$

Esse resultado representa a fração de registros da amostra que atendem a pelo menos um dos dois critérios analisados.

O diagrama de Venn ajuda a entender a magnitude da sobreposição entre os dois eventos. Embora eles tratem de categorias distintas (tipo de evento e tipo de local), é possível verificar se há uma intersecção relevante, ou se os grupos analisados atuam de forma mais isolada no perfil da mortalidade observada.

Essa informação pode ser útil para políticas públicas de segurança do trabalho e para o monitoramento de ambientes comerciais, especialmente em contextos onde há riscos ocupacionais relevantes.



1.2.4. Análise 4: Qual a probabilidade de um jovem (15 a 29 anos) ter falecido por causa violenta?

1.2.4.1. Objetivos e preparações iniciais

O objetivo desta análise é calcular a probabilidade condicional de que uma pessoa com idade entre 15 e 29 anos tenha falecido por uma causa violenta, conforme registrado no banco de dados do Sistema de Informação sobre Mortalidade (SIM), disponibilizado pelo OpenDataSUS, referente ao ano de 2021.

Para essa análise, utilizou-se a seguinte definição:

- Jovens: indivíduos com idade entre 15 e 29 anos completos, conforme a coluna IDADE_ANOS;
- Causas violentas: classificadas com base na CID-10, considerando os códigos iniciados por:
    - V (acidentes de transporte)
    - W (outras causas externas de lesões acidentais)
    - X (eventos intencionais/autoinfligidos e agressões)
    - Y10 a Y36 (eventos de intenção indeterminada e intervenções legais)

Colunas utilizadas:

- IDADE_ANOS: idade da pessoa falecida no momento do óbito;
- CAUSABAS: código da causa básica de morte (CID-10).

Registros com informações ausentes ou inválidas nessas colunas foram descartados no processo de limpeza prévia. A análise foi realizada com uma amostra de tamanho n_amostral, suficiente para estimativas representativas.

1.2.4.2. Explicações relacionadas ao código desenvolvido

Primeiramente, criou-se uma função auxiliar causa_violenta(cid) que verifica se o código da causa básica pertence ao grupo de causas externas e violentas conforme a estrutura da CID-10. Essa função examina a letra inicial e os dois dígitos seguintes do código, com base nas faixas mencionadas anteriormente.

Em seguida:

- Aplicou-se um filtro para selecionar os registros de jovens entre 15 e 29 anos;
- Aplicou-se outro filtro para selecionar as causas violentas, com base na função criada;
- Identificou-se a interseção entre os dois filtros — ou seja, jovens que faleceram por causas violentas;

Por fim, foi calculada a probabilidade condicional, utilizando a fórmula:

$$P(Causa \mid Jovem) = \frac{P(Jovem \cap Causa\ Violenta)}{P(Jovem)}$$

O resultado foi apresentado tanto em notação decimal quanto percentual.

1.2.4.3. Conclusões

A análise indicou que a probabilidade condicional de um indivíduo com idade entre 15 e 29 anos ter falecido por causa violenta foi:

$$P(Causa\ Violenta \mid Idade\ 15\text{–}29\ anos) = 0.5862 \rightarrow 58.62\%$$

Esse valor indica que, dentro do grupo jovem, uma parcela expressiva dos óbitos está associada a causas violentas, o que reforça preocupações recorrentes sobre a vulnerabilidade da juventude a eventos letais externos, como acidentes, agressões e outras situações de risco.

Tais achados reforçam a necessidade de políticas públicas direcionadas à prevenção da violência e acidentes entre jovens, que se apresentam como uma das principais causas de mortalidade nessa faixa etária.

1.2.5. Análise 5: Dada uma morte por agressão, qual a probabilidade de a vítima ser do sexo masculino?

1.2.5.1. Objetivos e preparações iniciais

O objetivo desta análise é estimar a probabilidade condicional de que um óbito cuja causa básica tenha sido uma agressão tenha vitimado um indivíduo do sexo masculino, com base nos registros do Sistema de Informação sobre Mortalidade (SIM) do OpenDataSUS, ano de 2021.
Para isso, foram utilizadas as seguintes colunas do dataset:

- CAUSABAS: código da causa básica do óbito, de acordo com a CID-10;
- SEXO_DESC: descrição do sexo da pessoa falecida (utilizado o valor "M" para homens).

As agressões foram definidas com base nos prefixos X8, X9 e Y0 da CID-10, que compreendem principalmente atos intencionais de agressão, incluindo violência interpessoal, homicídios e intervenções legais.

Todos os registros com valores ausentes ou inconsistentes nas colunas citadas foram eliminados na etapa de preparação dos dados. A análise considerou uma amostra de n_amostral registros válidos.

1.2.5.2. Explicações relacionadas ao código desenvolvido

A análise seguiu os seguintes passos:

1. Criou-se um filtro para identificar as causas básicas de óbito por agressão, com base nos prefixos X8, X9 e Y0 da CID-10.
2. Em seguida, foi aplicado um segundo filtro para selecionar os registros do sexo masculino.
3. A interseção entre esses dois filtros resultou na quantidade de óbitos que foram ao mesmo tempo de homens e causados por agressões.
4. Foram calculadas:
    a. P(B): a probabilidade de um óbito ter sido causado por agressão;
    b. P(A ∩ B): a probabilidade conjunta do óbito ter sido por agressão e da vítima ser homem;
    c. P(Homem | Agressão): a probabilidade condicional de um óbito por agressão ter ocorrido com um homem, usando a fórmula:

$$P(Homem \mid Agressão) = \frac{P(Homem \cap Agressão)}{P(Agressão)}$$

O valor resultante foi exibido tanto em notação decimal quanto percentual.

1.2.5.3. Conclusões

Com base na amostra analisada, a probabilidade condicional de um óbito causado por agressão ter vitimado uma pessoa do sexo masculino foi:

$$P(Homem \mid Agressão) = 0.9081 \rightarrow 90.81\%$$

Esse valor evidencia que a maioria dos óbitos por agressão registrados na base do SIM em 2021 foram de indivíduos do sexo masculino, reforçando um padrão já observado em estatísticas de violência, onde homens jovens costumam figurar entre as principais vítimas de agressões letais.

Esse dado pode orientar políticas públicas voltadas à redução da violência interpessoal, especialmente no que diz respeito à violência letal contra homens, que frequentemente se manifesta em contextos urbanos, sociais e econômicos de alta vulnerabilidade.

1.2.6. Análise 6: Dado um óbito ocorrido em domicílio, qual a probabilidade de ele ter ocorrido sem assistência médica?

1.2.6.1. Objetivos e preparações iniciais

Esta análise tem como objetivo estimar a probabilidade condicional de que um óbito ocorrido no domicílio da pessoa tenha ocorrido sem assistência médica. Os dados foram extraídos do Sistema de Informação sobre Mortalidade (SIM) do OpenDataSUS, referentes ao ano de 2021.

A análise considera os seguintes campos do dataset:

- LOCOCOR_DESC: local de ocorrência do óbito, sendo considerado o valor "Domicílio";
- ASSISTMED_DESC: indica se houve assistência médica no momento do óbito, sendo considerado o valor "Não".

Registros com valores ausentes, vazios ou não compatíveis com os critérios estabelecidos foram previamente filtrados. A análise foi realizada com uma amostra de n_amostral registros válidos, considerada representativa para o estudo estatístico.

1.2.6.2. Explicações relacionadas ao código desenvolvido

O processo analítico foi dividido em etapas:

1. Criação de um filtro para óbitos ocorridos em domicílio, com base no campo LOCOCOR_DESC;
2. Criação de outro filtro para óbitos sem assistência médica, conforme ASSISTMED_DESC;
3. Obteve-se a interseção entre esses dois conjuntos, ou seja, registros que atendem simultaneamente às duas condições;
4. Foram então calculadas:
    a. P(B): probabilidade de o óbito ter ocorrido em domicílio;
    b. P(A ∩ B): probabilidade conjunta de o óbito ter ocorrido em domicílio e sem assistência médica;
    c. P(Sem assistência médica | Domicílio): probabilidade condicional de o óbito não ter sido assistido, dado que ocorreu em domicílio.

Essa última foi calculada pela fórmula:

$$P(Sem\ Assistência \mid Domicílio) = \frac{P(Sem\ Assistência \cap Domicílio)}{P(Domicílio)}$$

O resultado foi apresentado em formato decimal e percentual para facilitar a interpretação.

1.2.6.3. Conclusões

A partir da amostra analisada, a probabilidade condicional de um óbito ocorrido em domicílio ter ocorrido sem assistência médica foi:

$$P(Sem\ Assistência\ Médica \mid Domicílio) = 0.2667 \rightarrow 26.67\%$$

Esse valor indica que uma parte significativa dos óbitos domiciliares não contou com atendimento profissional de saúde no momento da morte, o que pode estar associado a barreiras de acesso a serviços médicos, à demora na busca por ajuda, ou à subnotificação de casos.

O resultado chama atenção para a importância do fortalecimento da atenção domiciliar, da orientação à população quanto à procura por atendimento médico em tempo oportuno e da ampliação de estratégias de monitoramento remoto de pacientes em situação de vulnerabilidade.

1.2.7. Análise 7: Dado um óbito sem assistência médica, qual a probabilidade de ele ter ocorrido em domicílio?

1.2.7.1. Objetivos e preparações iniciais

Esta análise tem como objetivo calcular a probabilidade condicional de que um óbito que tenha ocorrido sem assistência médica tenha acontecido em domicílio. A investigação foi feita com base nos dados do Sistema de Informação sobre Mortalidade (SIM), disponibilizados pelo OpenDataSUS, referentes ao ano de 2021.

Foram utilizadas as seguintes colunas do dataset:

- LOCOCOR_DESC: local de ocorrência do óbito (utilizando o valor "Domicílio");
- ASSISTMED_DESC: informação sobre se houve assistência médica (utilizando o valor "Não").

Foram removidos registros com valores ausentes, vazios ou inválidos nessas colunas. A análise foi realizada com uma amostra de n_amostral registros válidos, suficientemente representativa para obter estimativas confiáveis.

1.2.7.2. Explicações relacionadas ao código desenvolvido

A análise envolveu o cálculo das seguintes probabilidades:

- P(A): probabilidade de o óbito ter ocorrido em domicílio;
- P(B): probabilidade de o óbito ter ocorrido sem assistência médica;
- P(B | A): probabilidade de não haver assistência médica, dado que o óbito ocorreu em domicílio (obtida filtrando ASSISTMED_DESC == 'Não' apenas entre os registros que atendem a LOCOCOR_DESC == 'Domicílio');
- P(A | B): a probabilidade de o óbito ter ocorrido em domicílio, dado que não houve assistência médica.

A fórmula utilizada para o cálculo da probabilidade condicional foi derivada do Teorema de Bayes:

$$P(Domicílio \mid Sem\ Assistência) = \frac{P(Sem\ Assistência \mid Domicílio) \times P(Domicílio)}{P(Sem\ Assistência)}$$

O resultado foi expresso em formato percentual para facilitar a compreensão.

1.2.7.3. Conclusões

A análise indicou que a probabilidade condicional de um óbito sem assistência médica ter ocorrido em domicílio foi:

$$P(Domicílio \mid Sem\ assistência\ médica) = 55.66\%$$

Esse resultado reforça a correlação entre a ausência de atendimento médico e a ocorrência do óbito no próprio domicílio, sugerindo possíveis falhas de acesso à rede de saúde, falta de transporte adequado, demora na procura por ajuda, ou outras barreiras sociais e econômicas.

A identificação desse padrão pode ser útil para o planejamento de políticas públicas focadas em atenção domiciliar, telemedicina, e estratégias de vigilância ativa para populações mais vulneráveis.

1.2.8. Análise 8: Dado que a vítima era do sexo masculino, qual a probabilidade de o óbito ter sido causado por agressão?

1.2.8.1. Objetivos e preparações iniciais

Esta análise tem como objetivo calcular a probabilidade condicional de que um indivíduo do sexo masculino tenha tido como causa básica da morte uma agressão, com base nos registros do Sistema de Informação sobre Mortalidade (SIM), disponibilizados pelo OpenDataSUS, ano de 2021.

Foram utilizadas as seguintes colunas do dataset:

- CAUSABAS: código da causa básica do óbito, segundo a CID-10;
- SEXO_DESC: sexo da pessoa falecida (utilizado o valor "M" para homens).

As agressões foram identificadas a partir dos códigos CID-10 com os prefixos:

- X8 (agressões com meios físicos e objetos),
- X9 (agressões com outros meios ou não especificadas),
- Y0 (intervenções legais e operações de guerra).

Todos os registros com informações ausentes, inválidas ou inconsistentes foram descartados. A análise foi realizada sobre uma amostra de n_amostral registros válidos.

1.2.8.2. Explicações relacionadas ao código desenvolvido

A análise seguiu os seguintes passos:

1. Criou-se um filtro para identificar os registros cuja causa básica foi agressão, com base no prefixo dos códigos da CID-10;
2. Aplicou-se um segundo filtro para selecionar registros em que o sexo da vítima era masculino;
3. Foram calculadas:
    a. P(A): probabilidade de óbito por agressão (entre todos os registros);
    b. P(B): probabilidade de o óbito ter sido de um homem;
    c. P(B | A): probabilidade de a vítima ser homem, dado que a causa do óbito foi agressão;
    d. P(A | B): probabilidade de a causa do óbito ter sido agressão, dado que a vítima era homem.

A última foi obtida utilizando o Teorema de Bayes:

$$P(Agressão \mid Homem) = \frac{P(Homem \mid Agressão) \times P(Agressão)}{P(Homem)}$$

O resultado foi apresentado em formato percentual para facilitar a interpretação.

1.2.8.3. Conclusões

Com base nos dados analisados, a probabilidade condicional de um óbito por agressão ter vitimado uma pessoa do sexo masculino foi de:

$$P(Agressão \mid Homem) = 4.17\%$$

Esse valor reforça a associação entre mortalidade por agressões e o sexo masculino, revelando que os homens, especialmente os mais jovens (como mostrado em análises anteriores), representam uma parte desproporcional das vítimas de causas violentas.

Essa constatação é compatível com dados epidemiológicos sobre violência no Brasil e pode contribuir para o direcionamento de políticas públicas de segurança, saúde e prevenção da violência, com foco em grupos de risco específicos, como homens jovens em áreas urbanas vulneráveis.

1.2.9. Análise 9: Dado que o óbito ocorreu em hospital, qual a probabilidade de a vítima ser idosa (60 anos ou mais)?

1.2.9.1. Objetivos e preparações iniciais

O objetivo desta análise é estimar a probabilidade condicional de que um óbito ocorrido em hospital tenha sido de um idoso, definido como indivíduo com idade igual ou superior a 60 anos. A base utilizada foi o Sistema de Informação sobre Mortalidade (SIM) do OpenDataSUS, com dados referentes ao ano de 2021.

As colunas relevantes para esta investigação foram:

- LOCOCOR_DESC: local de ocorrência do óbito, sendo considerado o valor "Hospital";
- IDADE_ANOS: idade da pessoa falecida no momento do óbito (considerando ≥ 60 anos como critério para "idoso").

Foram excluídos registros com dados ausentes ou inválidos nessas colunas. A análise foi conduzida sobre uma amostra de tamanho n_amostral, considerada suficiente para estimativas confiáveis.

1.2.9.2. Explicações relacionadas ao código desenvolvido

A análise seguiu as seguintes etapas:

1. Criação de um filtro para selecionar os óbitos que ocorreram em hospital (LOCOCOR_DESC == 'Hospital');
2. Criação de um segundo filtro para identificar registros de pessoas com 60 anos ou mais (IDADE_ANOS >= 60);
    a. Cálculo das probabilidades:
    b. P(A): probabilidade de o óbito ter ocorrido com uma pessoa idosa (em toda a amostra);
    c. P(B): probabilidade de o óbito ter ocorrido em hospital;
    d. P(B | A): probabilidade de um idoso ter morrido em hospital (calculada diretamente);
    e. P(A | B): probabilidade de um óbito ocorrido em hospital ter sido de um idoso, utilizando o Teorema de Bayes:

$$P(Idoso \mid Hospital) = \frac{P(Hospital \mid Idoso) \times P(Idoso)}{P(Hospital)}$$

O valor final foi apresentado em notação percentual para facilitar sua interpretação.

1.2.9.3. Conclusões

Com base na amostra analisada, a probabilidade condicional de um óbito ocorrido em hospital ter sido de uma pessoa idosa (≥ 60 anos) foi:

$$P(Idoso \mid Hospital) = 70.07\%$$

Esse resultado revela que uma proporção significativa dos óbitos hospitalares envolvem indivíduos com 60 anos ou mais, o que pode ser explicado pelo aumento da fragilidade fisiológica, maior prevalência de doenças crônicas e a maior utilização de serviços hospitalares por essa faixa etária.

Essas informações podem ser úteis para o planejamento de políticas de saúde voltadas à população idosa, com ênfase em cuidados hospitalares, prevenção de complicações crônicas e atenção especializada ao envelhecimento populacional.

1.2.10. Análise 10: Qual a probabilidade de exatamente 4 óbitos por causa cardiovascular em uma amostra de 10 certidões aleatórias?

1.2.10.1. Objetivos e preparações iniciais

O objetivo desta análise é calcular a probabilidade de exatamente 4 óbitos, entre 10 certidões aleatórias, terem como causa básica uma doença cardiovascular (doenças do sistema circulatório). Os dados utilizados foram extraídos do Sistema de Informação sobre Mortalidade (SIM), disponibilizados pelo OpenDataSUS, para o ano de 2021.

Foram considerados como causas cardiovasculares todos os códigos da CID-10 que iniciam com a letra "I", conforme a classificação internacional de doenças.

A coluna utilizada foi:

- CAUSABAS: código da causa básica de óbito segundo a CID-10.

Foi aplicado um filtro para identificar os registros com códigos que iniciam com "I", e a proporção total desses registros na amostra (p_cardiovascular) foi utilizada como probabilidade de sucesso na distribuição binomial.

1.2.10.2. Explicações relacionadas ao código desenvolvido

A análise se baseia na distribuição binomial, uma distribuição de probabilidade discreta usada para modelar o número de sucessos em um número fixo de experimentos independentes, com duas possíveis saídas: sucesso ou fracasso.

- Sucesso: certidão de óbito com causa cardiovascular;
- n = 10: tamanho da amostra (10 certidões);
- k = 4: número exato de sucessos desejado (4 óbitos por causa cardiovascular);
- p = p_cardiovascular: probabilidade estimada de sucesso com base na amostra real do dataset.

A função binom.pmf(k, n, p) da biblioteca SciPy foi utilizada para calcular a probabilidade de exatamente 4 sucessos. Além disso, foi gerado um gráfico de barras que representa a distribuição binomial completa para todos os valores de 0 a 10 sucessos, permitindo visualizar as chances associadas a cada possível número de óbitos cardiovasculares entre 10 certidões.

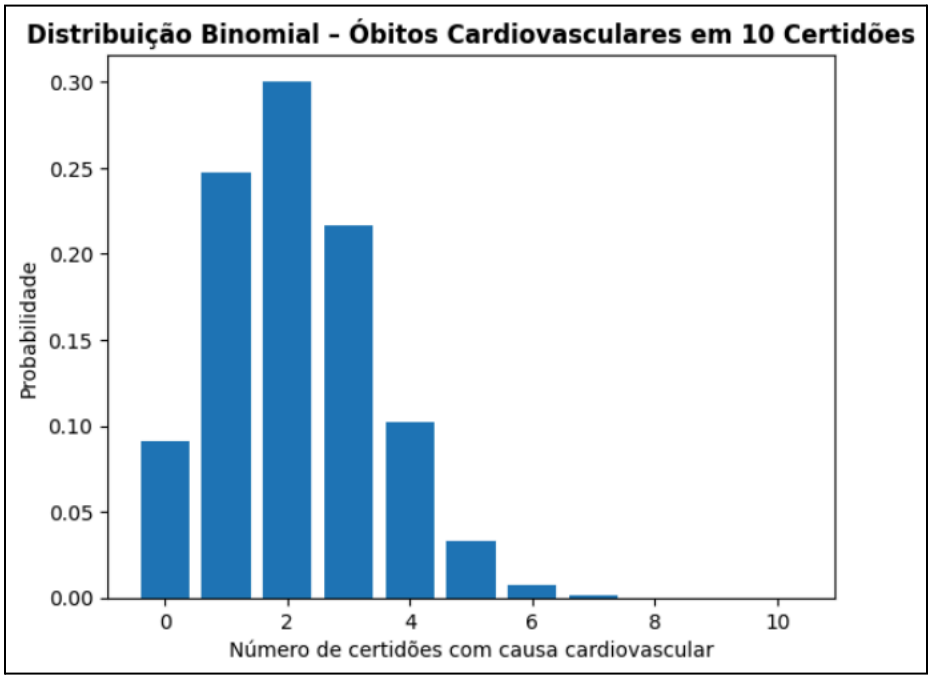
1.2.10.3. Conclusões

Com base na proporção estimada de causas cardiovasculares no conjunto de dados analisado, a probabilidade de que exatamente 4 entre 10 certidões de óbito aleatórias apresentem uma causa cardiovascular foi:

$$P(X = 4) = 0.1023 \rightarrow 10.23\%$$

Esse valor representa a chance de ocorrência de 4 casos cardiovasculares em 10 óbitos, assumindo que cada certidão tenha uma chance independente e igual de envolver esse tipo de causa, conforme a estimativa da amostra.

O gráfico gerado pela distribuição binomial auxilia na compreensão dos resultados mais prováveis para essa situação, sendo útil para simulações, planejamentos amostrais ou inferência estatística com base em causas específicas de mortalidade.



1.2.11. Análise 11: Em 300 óbitos por doenças cardíacas, qual a probabilidade de menos de 150 serem de homens?

1.2.11.1. Objetivos e preparações iniciais

Esta análise tem como objetivo calcular a probabilidade acumulada de que, entre 300 óbitos por doenças cardíacas, menos de 150 sejam de indivíduos do sexo masculino. Os dados analisados são do Sistema de Informação sobre Mortalidade (SIM), disponibilizados pelo OpenDataSUS, ano de 2021.

As doenças cardíacas foram identificadas com base na coluna:

- CAUSABAS: foi utilizado um filtro para códigos da CID-10 que começam com a letra "I", que indicam doenças do sistema circulatório, incluindo doenças cardíacas.

A variável SEXO_DESC foi utilizada para identificar os óbitos masculinos, considerando apenas os valores válidos.

1.2.11.2. Explicações relacionadas ao código desenvolvido

A distribuição de interesse é a binomial, já que estamos lidando com um número fixo de tentativas (n = 300) e uma probabilidade constante de sucesso p (óbito masculino) a cada observação independente.

As etapas foram:

1. Seleção de todos os registros de óbitos com causa básica iniciada por "I" (doenças cardíacas);
2. Cálculo da proporção de óbitos masculinos nesse subconjunto (p_homens);
3. Definição dos parâmetros da distribuição binomial:
    a. n = 300 (tamanho da amostra);
    b. k = 149 (limite superior do número de sucessos desejado: menos de 150);
    c. p = p_homens (probabilidade estimada de um óbito por doença cardíaca ser de um homem);
4. Cálculo da probabilidade acumulada P(X < 150) usando a função binom.cdf(k, n, p) da biblioteca SciPy;
5. Geração de um gráfico de barras mostrando o comportamento da função de distribuição acumulada (CDF) no intervalo provável em torno da média esperada.

Também foram calculados:

1. A média esperada de óbitos masculinos: =np;
2. O desvio padrão: =np1-p;
3. Um intervalo de variação provável para a visualização da CDF: -3, +3.

1.2.11.3. Conclusões

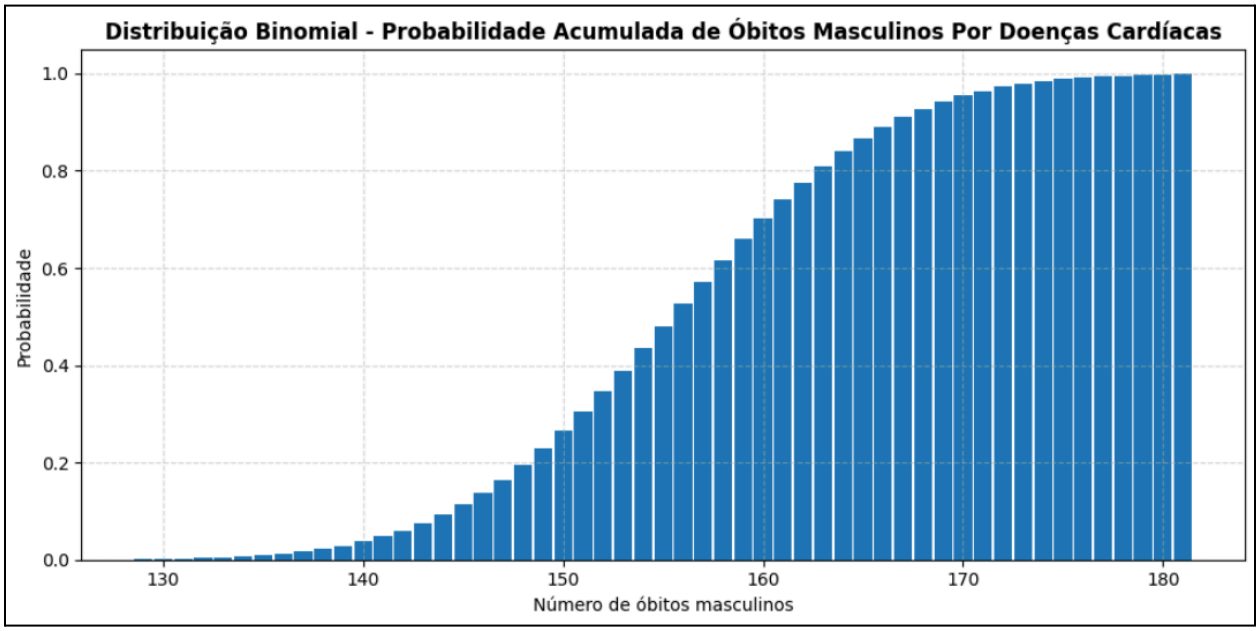
Com base na proporção estimada de homens entre os óbitos por doenças cardíacas, a probabilidade de que menos de 150 dos 300 óbitos selecionados sejam de homens foi:

$$P(X < 150) = 0.2288 \rightarrow 22.88\%$$

Esse valor representa a chance acumulada de que o número de óbitos masculinos seja inferior a 150, ou seja, menos da metade da amostra.

A curva da distribuição acumulada gerada no gráfico permite identificar a tendência central da amostra e a probabilidade de ocorrência de extremos. Uma baixa probabilidade indicaria que o cenário de menos de 150 homens em 300 óbitos cardiovasculares seria

estatisticamente improvável, o que reforçaria que a maioria das mortes por doenças cardíacas tende a afetar igualmente ou mais homens, dado o histórico clínico e epidemiológico da população.



1.2.12. Análise 12: Qual a probabilidade de ocorrerem até 75 mortes por COVID-19 em um dia?

1.2.12.1. Objetivos e preparações iniciais

Esta análise tem como objetivo estimar a probabilidade de ocorrência de até 75 óbitos por COVID-19 em um único dia, com base em dados reais de mortalidade obtidos no Sistema de Informação sobre Mortalidade (SIM) do OpenDataSUS, referentes ao ano de 2021.

A causa básica dos óbitos por COVID-19 foi identificada com o código CID-10: B342 – "Infecção por coronavírus de localização não especificada".

As colunas utilizadas foram:

- CAUSABAS: código da causa básica do óbito;
- DTOBITO_DATA: data do óbito (convertida para formato de data para agrupamento).

A análise considerou apenas registros válidos com essa causa específica. Foi então calculado o número de mortes por dia, permitindo modelar a ocorrência diária como um processo estocástico com taxa constante de eventos – ideal para a distribuição de Poisson.

1.2.12.2. Explicações relacionadas ao código desenvolvido

A distribuição de Poisson é utilizada para modelar o número de eventos que ocorrem em um intervalo fixo de tempo, quando esses eventos são independentes e ocorrem a uma taxa média constante. Neste caso:

- A variável aleatória X representa o número de óbitos por COVID-19 em um dia;
- O parâmetro λ representa a média diária de óbitos por COVID-19;
- $k = 75$: valor específico para o qual desejamos calcular a probabilidade acumulada $P(X \leq 75)$.

A função poisson.cdf(k, mu=lambda_covid) da biblioteca SciPy foi utilizada para obter a probabilidade de ocorrer até 75 óbitos por dia, assumindo que o número de mortes segue uma distribuição de Poisson com média diária λ.

Além do cálculo pontual, foi gerado um gráfico de barras que representa a função de distribuição acumulada (CDF) para um intervalo centrado na média estimada $\left(\lambda \pm 3\sqrt{\lambda}\right)$, ilustrando a distribuição de probabilidade dos óbitos diários.

1.2.12.3. Conclusões

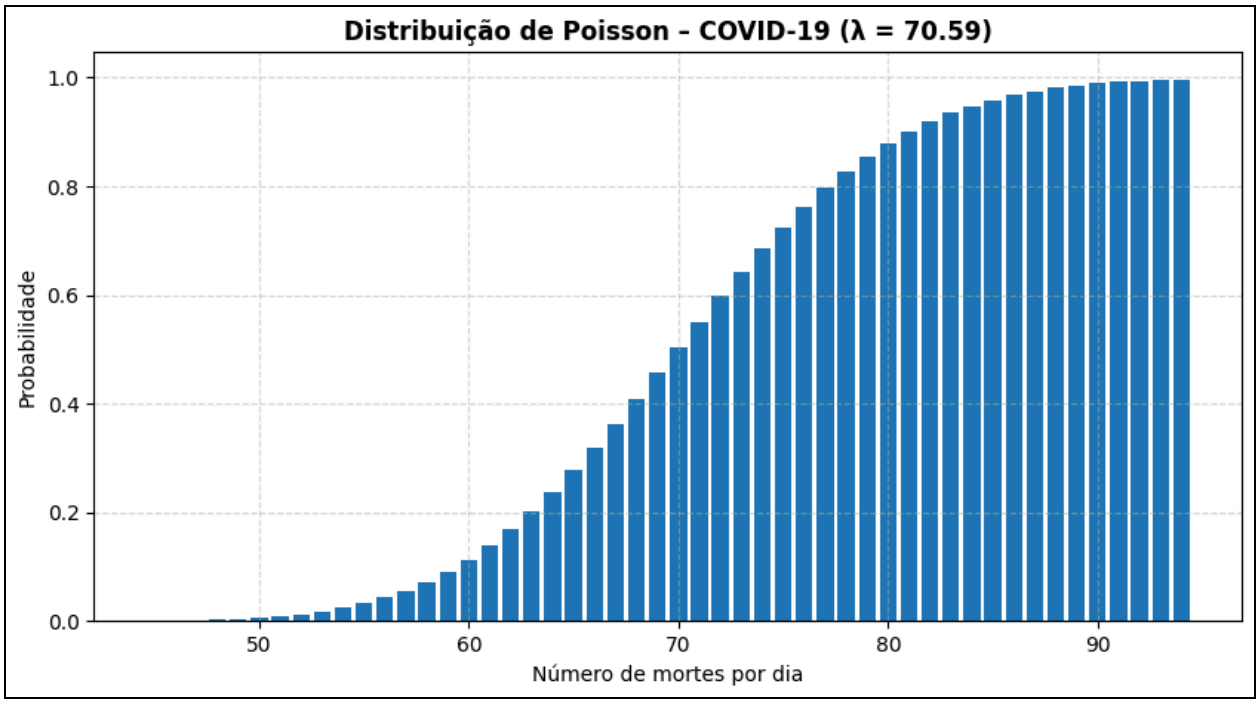
Com base na média diária estimada de mortes por COVID-19 em 2021, a probabilidade de que, em um determinado dia, ocorram até 75 óbitos por essa causa foi:

$$P(X \leq 75) = 0.7250 \rightarrow 72.50\%$$

Esse valor fornece uma estimativa acumulada da chance de observar até 75 mortes em um dia comum durante o período analisado. Valores muito baixos ou muito altos de $P(X \leq 75)$ podem indicar que 75 mortes é um evento raro ou comum, dependendo da magnitude de λ.

A curva da distribuição de Poisson permite visualizar quais faixas de mortalidade diária são mais prováveis, sendo útil para:

1. Análise epidemiológica;
2. Planejamento hospitalar e sanitário;
3. Avaliação de anomalias, como picos acima do esperado (surtos) ou quedas abruptas (subnotificações ou estabilização da pandemia).



1.2.13. Análise 13: Qual a probabilidade de ocorrerem mais de 180 mortes por acidentes de trânsito em um mês?

1.2.13.1. Objetivos e preparações iniciais

Esta análise tem como objetivo estimar a probabilidade de que, em um mês qualquer, o número de óbitos por acidentes de trânsito ultrapasse 180 casos. Os dados utilizados provêm do Sistema de Informação sobre Mortalidade (SIM) do OpenDataSUS, ano de 2021.

As causas de óbito relacionadas a acidentes de trânsito foram identificadas por meio dos códigos CID-10 que iniciam com "V01" até "V89", que englobam uma ampla gama de acidentes de transporte terrestre (ex.: colisões entre veículos, atropelamentos, quedas de motocicleta etc.).

A coluna analisada foi:

- CAUSABAS: código da causa básica do óbito.

Foi aplicado um filtro para considerar apenas registros válidos com causas básicas iniciando com os prefixos indicados. A análise assume uma distribuição mensal uniforme ao longo dos 12 meses do ano, permitindo calcular uma taxa média mensal de ocorrências.

1.2.13.2. Explicações relacionadas ao código desenvolvido

O número de mortes mensais por acidentes de trânsito foi modelado com a distribuição de Poisson, adequada para representar a ocorrência de eventos raros e independentes ao longo de um intervalo fixo de tempo (neste caso, um mês).

- A variável aleatória X representa o número de mortes por trânsito em um mês;
- O parâmetro λ (lambda_mensal) corresponde à média mensal estimada de óbitos;
- O valor de interesse é $P(X > 180)$, ou seja, a probabilidade de que ocorra mais de 180 mortes em um mês;
- Essa probabilidade foi calculada como $1 - P(X \leq 180)$, utilizando a função poisson.cdf(180, mu=λ) da biblioteca SciPy.

Além disso, foi gerado um gráfico de barras representando as probabilidades de ultrapassagem (complementares da CDF) em um intervalo de valores centrado na média mensal estimada $\left(\lambda \pm 3\sqrt{\lambda}\right)$, permitindo visualizar a frequência esperada de diferentes faixas de mortalidade.

1.2.13.3. Conclusões

Com base na taxa média mensal estimada, a probabilidade de ocorrerem mais de 180 óbitos por acidentes de trânsito em um único mês foi:

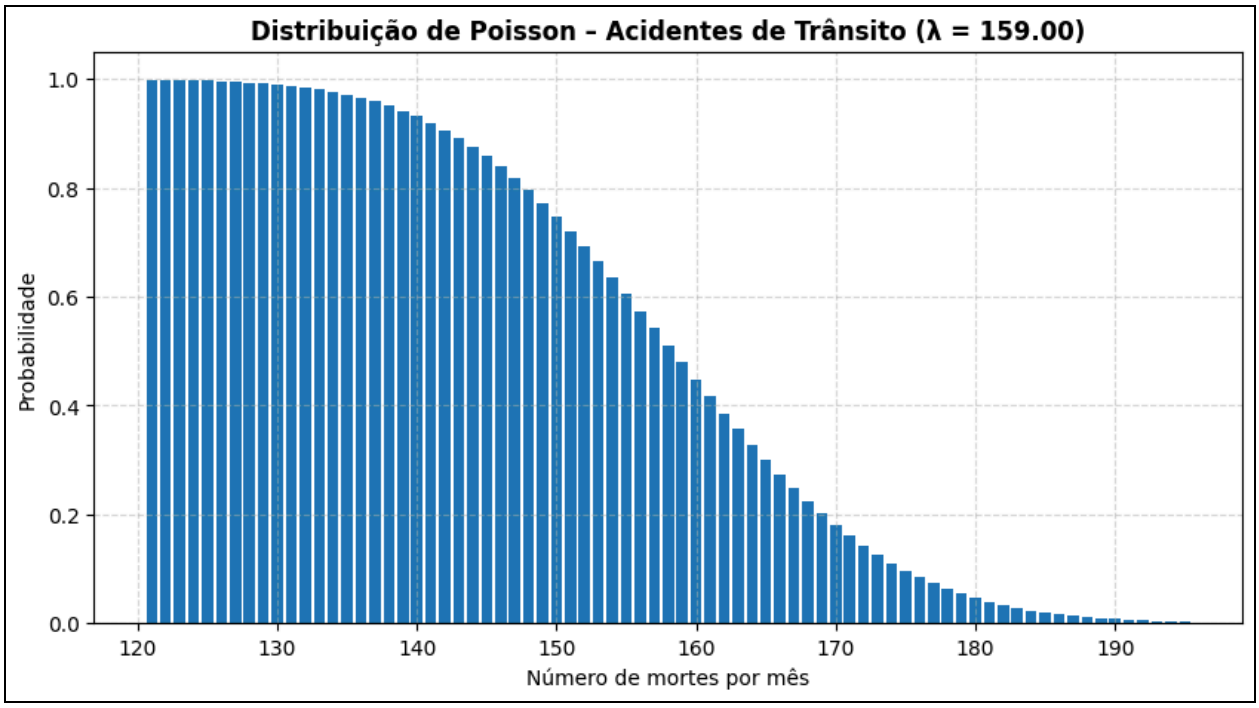
$$P(X > 180) = 0.0463 \rightarrow 4.63\%$$

Esse resultado representa a chance de se observar um mês com mortalidade acima de 180 casos, e sua magnitude indica se esse cenário é comum ou improvável, à luz da média histórica registrada.

A visualização gráfica da distribuição de Poisson reforça a concentração de valores próximos à média, destacando o quanto valores extremos (como >180) são raros ou esperados.

Essa análise pode ser útil para:

- Planejamento de políticas públicas de trânsito;
- Distribuição de recursos hospitalares e emergenciais;
- Avaliação de sazonalidade ou surtos localizados.


Distribuição de Poisson – Acidentes de Trânsito (λ = 159.00)
Probabilidade
1.0
0.8
0.6
0.4
0.2
0.0
120
130
140
150
160
170
180
190
Número de mortes por mês

2. SÍNDROME RESPIRATÓRIA AGUDA GRAVE (SRAG)

2.1. Análises Descritivas

2.1.1. Análise 1: As taxas de óbito e de cura diferem entre os sexos?

2.1.1.1. Objetivos e preparações iniciais

Essa análise tem o objetivo de verificar se existe alguma diferença nas taxas de óbito e de cura dos pacientes, conforme o sexo que eles são identificados nos registros do dataset de Síndrome Respiratória Aguda Grave (SRAG) analisado.

Segue a descrição das colunas utilizadas conforme o dicionário de dados do dataset:

- CS_SEXO:
    - 1 - Masculino
    - 2 - Feminino
- EVOLUCAO:
    - 1 - Cura
    - 2 - Óbito

As demais opções das colunas foram descartadas para análise por não se enquadrarem na investigação desenvolvida. É também válido ressaltar que essas colunas utilizadas tiveram que passar por um tratamento prévio de filtragem, removendo linhas que apresentam valores inválidos ou ausentes (*NaN*) ou registros vazios.

Além disso, é necessário ressaltar que a análise foi restrita para uma amostra de 50.000 registros válidos. Isso se deve por conta das limitações computacionais do *Google Colab* que acabaram afetando essa investigação, porém acredita-se que a amostra obtida é suficiente para o estudo elaborado e permite trazer algumas conclusões relevantes.

2.1.1.2. Explicações relacionadas ao código desenvolvido

Após todas as etapas iniciais referentes ao tratamento dos dados, foi criada uma tabela de contingência cruzando o sexo e evolução dos casos, o que resultou em:

- Quantidades totais de cura e óbito por sexo
- Quantidades totais (linha “total geral”) para o conjunto da amostra

Por fim, para melhor ilustração dos resultados obtidos, foram criados dois gráficos de barras comparando as taxas de óbito e de cura entre ambos os sexos.

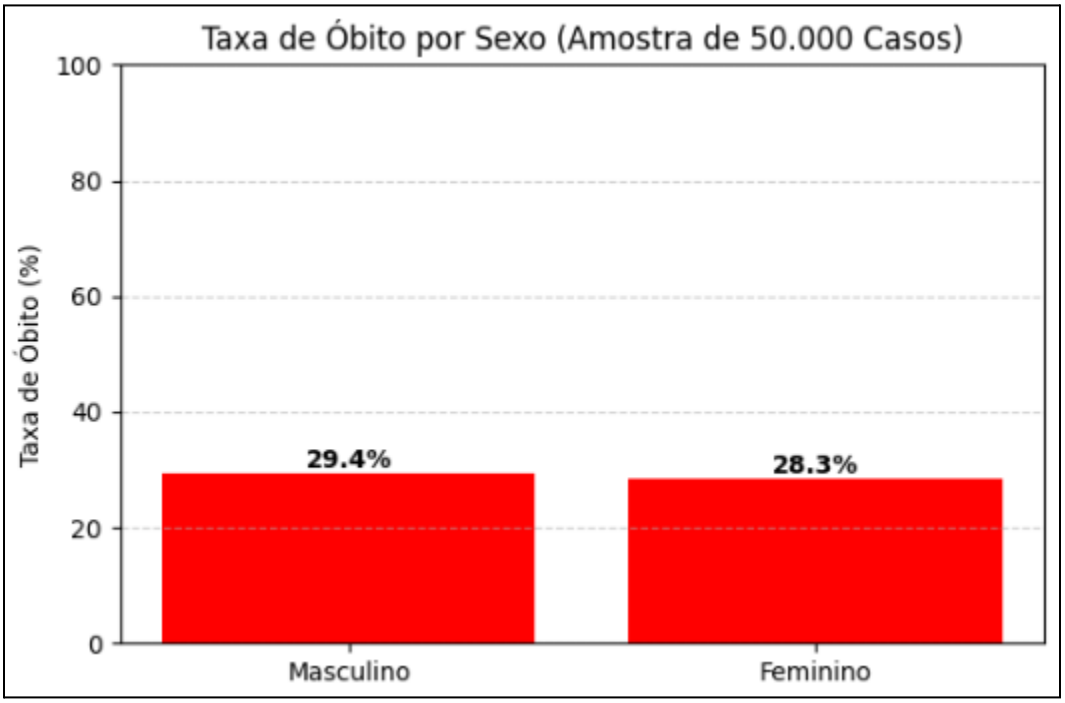
2.1.1.3. Conclusões

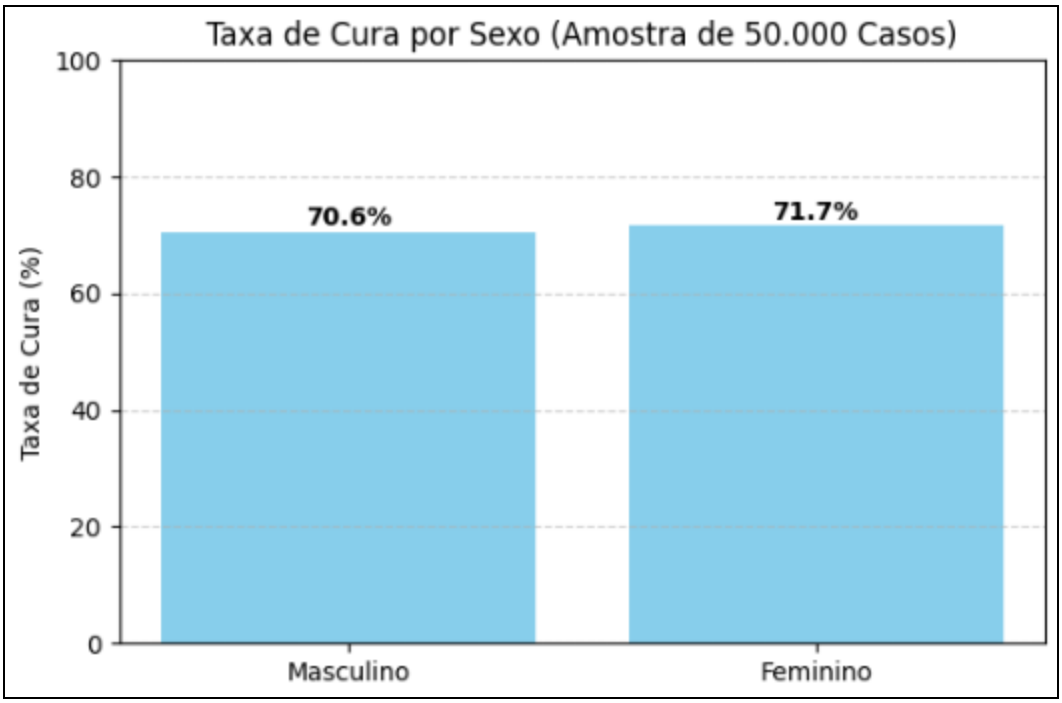
A partir da amostra de 50.000 registros com evolução conhecida (cura ou óbito), embora exista uma porção levemente maior de casos do sexo masculino nessa tabela, é perceptível que ambos os sexos possuem taxas relativamente próximas, com 29,41% de óbito para homens e 28,31% de óbito para mulheres.

Como as taxas de óbito são semelhantes, o mesmo pode-se dizer para as taxas de cura que são de 70,59% e 71,69% para o sexo masculino e feminino respectivamente.

Portanto, pode-se concluir que ambos os sexos não apresentam nenhuma disparidade nos resultados obtidos até o momento usando estatística descritiva.

	Cura	Óbito	Total	Taxa_óbito (%)	Taxa_cura (%)
Masculino	19426	8093	27519	29.41	70.59
Feminino	16116	6365	22481	28.31	71.69
Total Geral	35542	14458	50000	28.92	71.08





2.1.2. Análise 2: Distribuição dos sintomas de febre, diarreia e tosse entre gestantes

2.1.2.1. Objetivos e preparações iniciais

Essa análise tem a finalidade de identificar a frequência de 3 sintomas clínicos entre as gestantes com base nos registros fornecidos do dataset: febre, diarreia e tosse.

Segue a descrição das colunas utilizadas conforme o dicionário de dados do dataset:

- CS_GESTANT
    - 1 - Gravidez no 1º trimestre
    - 2 - Gravidez no 2º trimestre
    - 3 - Gravidez no 3º trimestre
- FEBRE, DIARREIA e TOSSE
    - 1 - Sim
    - 2 - Não
    - 9 - ignorado

Para essa análise, foi feito um tratamento prévio das colunas utilizadas para garantir que apenas registros válidos sejam obtidos no estudo.

2.1.2.2. Explicações relacionadas ao código desenvolvido

De início, após os tratamentos, foram buscados apenas os casos de gestantes em cada trimestre, com base na descrição previamente explicada sobre a coluna.

Em seguida, foi tirada a proporção de gestantes com relação a todo o dataset para se ter uma noção de quantos casos de mulheres grávidas estão registrados.

Após isso, com base no que foi explicado sobre as colunas de sintomas, foi necessário criar uma variável que armazena os 3 valores possíveis para essas colunas.

Dessa forma, foram construídos os gráficos de barra e histogramas com base em um loop do array desses sintomas.

2.1.2.3. Conclusões

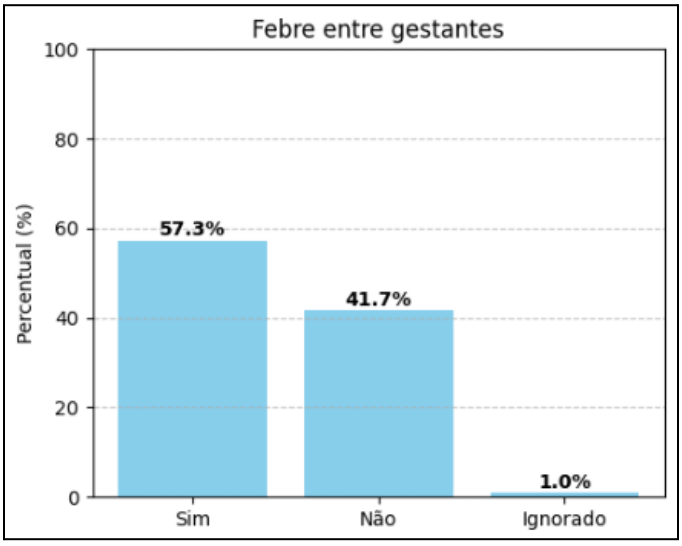
Com toda a construção dessa análise, foi possível obter algumas informações interessantes em relação aos registros do dataset que nos fornecem algumas conclusões.

A primeira informação obtida é que apenas aproximadamente 0,93% de todo o dataset representa o conjunto de mulheres que estão grávidas. Tal taxa demonstra que representam uma minoria ínfima.

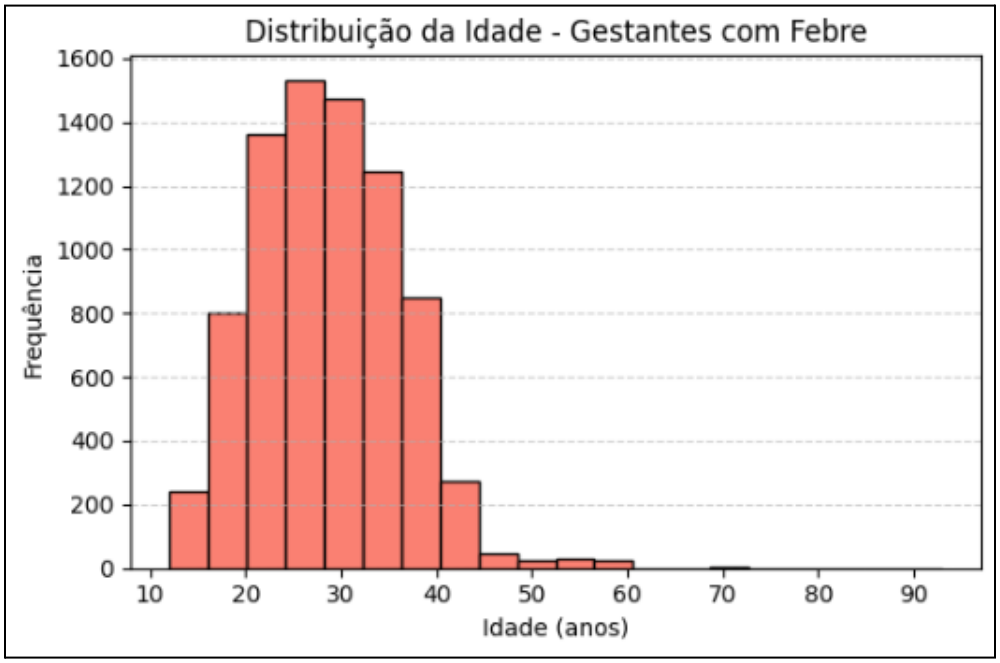
Agora, em relação aos gráficos, temos os seguintes resultados:

O sintoma de febre entre as gestantes possui a seguinte distribuição de taxas:

- 57,3% - Sim;
- 41,7% - Não;
- 1.0% - Ignorado



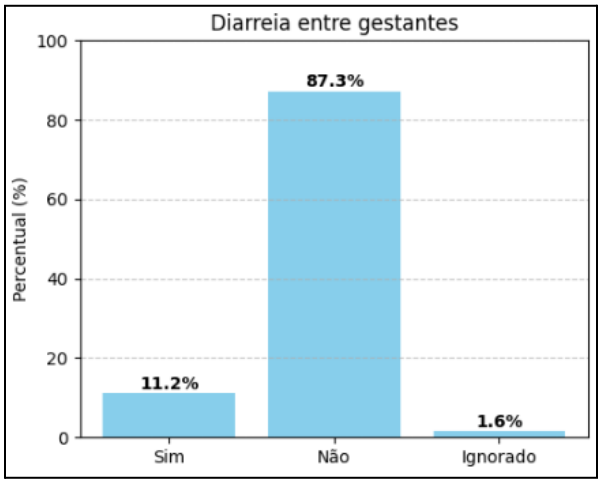
Agora, com relação ao histograma da distribuição da idade das pacientes grávidas com esse sintoma, temos que a frequência é alta para pacientes entre 20 a 30 anos. Realizando um cálculo da média da idade de gestantes com febre, temos um valor próximo de 28,91 anos.



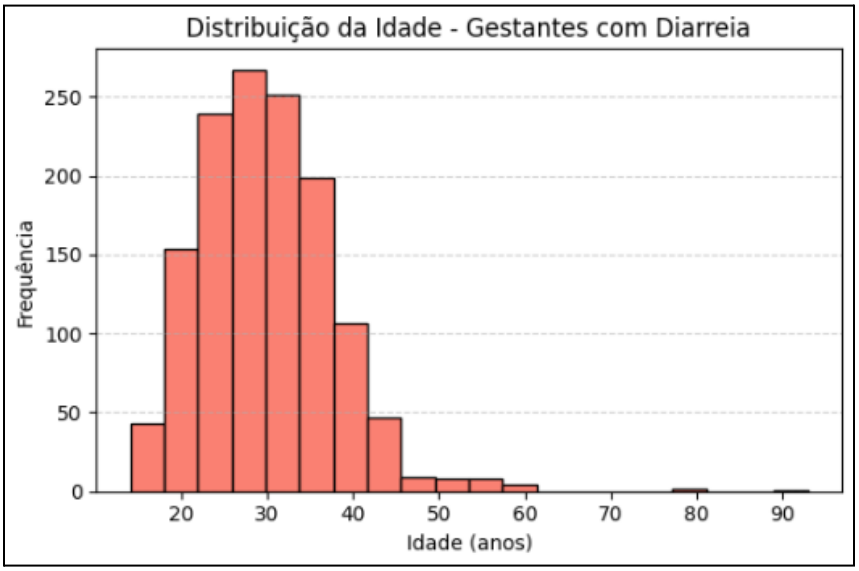
Sobre os casos de gestantes com o sintoma de diarreia, foi obtida uma alta taxa de negação, evidenciando pelo menos nessa análise descritiva, que não é um sintoma comum entre mulheres grávidas.

O sintoma de diarreia entre as gestantes possui a seguinte distribuição de taxas:

- 11,2% - Sim;
- 87,3% - Não;
- 1.6% - Ignorado



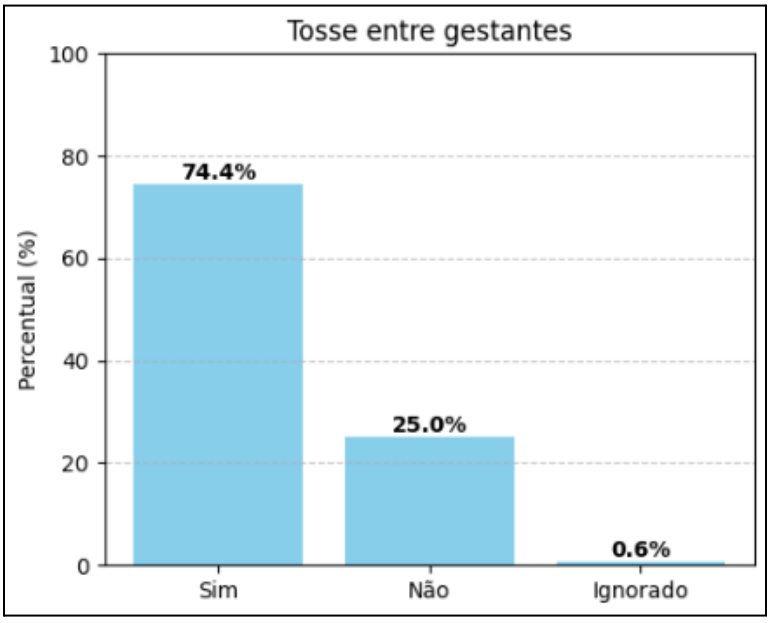
A distribuição da idade de pacientes com esse sintoma é bem semelhante à de pacientes com febre, com uma distribuição que fica na faixa etária dos 20 a 30 anos. Segundo essa distribuição, temos que a idade média é algo em torno de 29,62 anos.



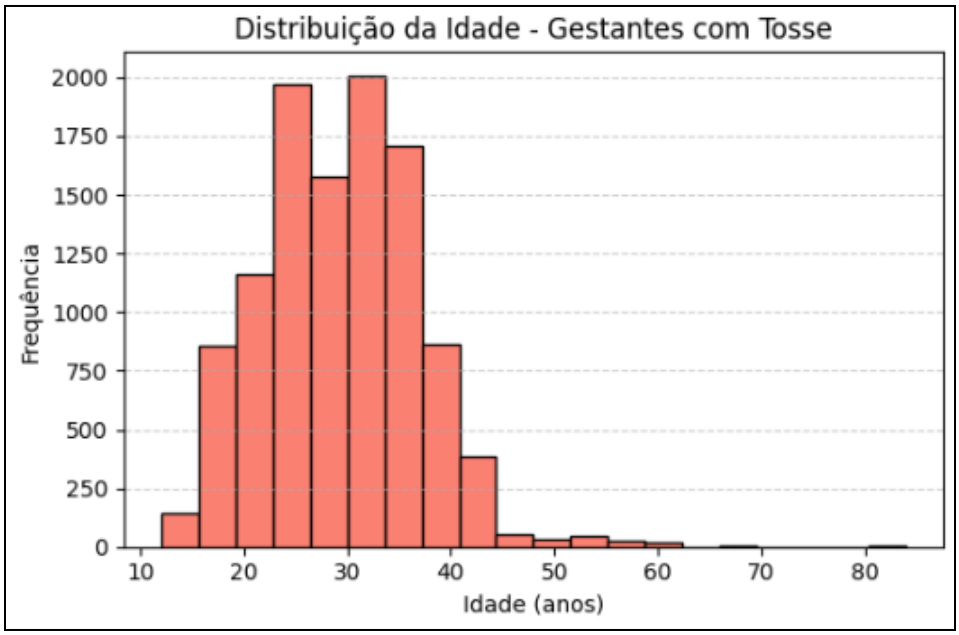
Por fim, com relação ao sintoma de tosse, pode-se observar que quase 75% apresentam o sintoma, indicando uma taxa significativa para esse tipo de sintoma.

O sintoma de tosse entre as gestantes possui a seguinte distribuição de taxas:

- 74,4% - Sim;
- 25,0% - Não;
- 0,6% - Ignorado



Mantendo uma distribuição bem parecida com as outras, a distribuição de idade de grávidas com o sintoma de tosse também não revela nada surpreendente, tirando o fato de que a faixa etária das mulheres entre 30 a 40 anos apresentou registros maiores em comparação com os outros dois sintomas. Por fim, vale ressaltar que a idade média de pacientes com esse sintoma é de 29,22 anos. Uma média que também é parecida com as demais.



2.1.3. Análise 3: Pacientes Nosocomiais (Ocorrência, Coerência Temporal e Evolução Clínica)

2.1.3.1. Objetivos e preparações iniciais

Essa análise tem como objetivo avaliar os registros obtidos no dataset que são caracterizados como casos nosocomiais, ou seja, o caso de infecção aconteceu após a internação hospitalar.

Três aspectos principais foram abordados:

1. A proporção de casos nosocomiais.
2. A coerência cronológica das datas de internação e o início dos sintomas classificados como nosocomiais, seguindo a forma como as colunas foram definidas no dicionário de dados.
3. A comparação das evoluções (óbito ou cura) entre pacientes nosocomiais e não nosocomiais

Segue a descrição das colunas utilizadas conforme o dicionário de dados do dataset:

- NOSOCOMIAL
  - 1 - Sim
  - 2 - Não
- DT_INTERNA: representa a data que o paciente foi hospitalizado
- DT_SIN_PRI: representa a data do 1º sintoma do caso.

2.1.3.2. Explicações relacionadas ao código desenvolvido

Conforme mencionado na seção anterior, a primeira fase da análise obtém uma proporção de todos os casos que definidos como nosocomiais em comparação com os demais registros do dataset. Esse trecho não afeta as próximas fases. Ele apenas serve para revelar a taxa total de casos nosocomiais identificados.

A segunda fase é responsável por fazer algumas verificações com base em algumas premissas:

- Foi feita uma verificação prévia que registra apenas casos nosocomiais com ambas as datas preenchidas, ou seja, linhas com campos vazios ou inválidos são descartados da análise.
- Posteriormente, foi calculada a diferença entre as datas para identificar casos coerentes e incoerentes. Casos coerentes são aqueles que a data do sintoma é no dia ou após a data de internação. Dessa forma, casos incoerentes são aqueles que apresentam datas de sintomas que antecedem as datas de internação.

Com base nessas premissas, construiu-se gráficos de barras verticais empilhadas. O primeiro gráfico exibe os registros que são válidos (possuem as colunas de datas devidamente preenchidas), e o segundo gráfico mostra a coerência dos registros de casos nosocomiais com base nos pacientes nosocomiais válidos.

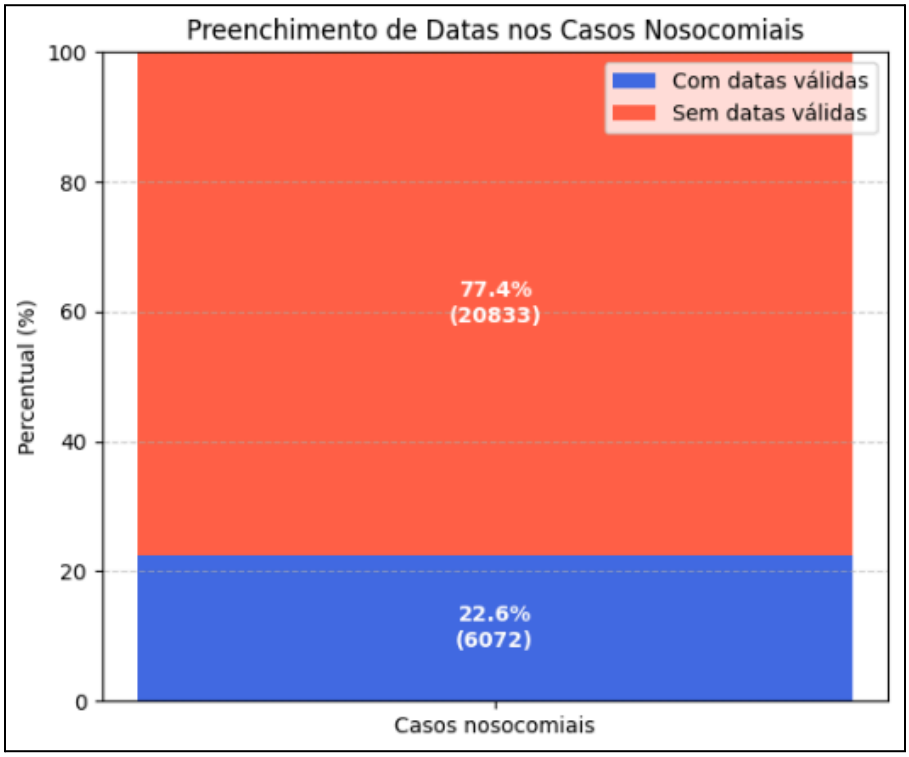
Por fim, com relação ao último aspecto mencionado na seção anterior, foi construído uma tabela de contingência que cruza as informações entre casos nosocomiais e não nosocomiais.

Essa tabela foi construída com a finalidade de comparar a evolução (desfecho clínico) entre esses tipos de infecção. Dessa forma, após a construção da tabela com base nas colunas necessárias, é exibido um gráfico de barras que compara a taxa de evolução entre os casos.

2.1.3.3. Conclusões

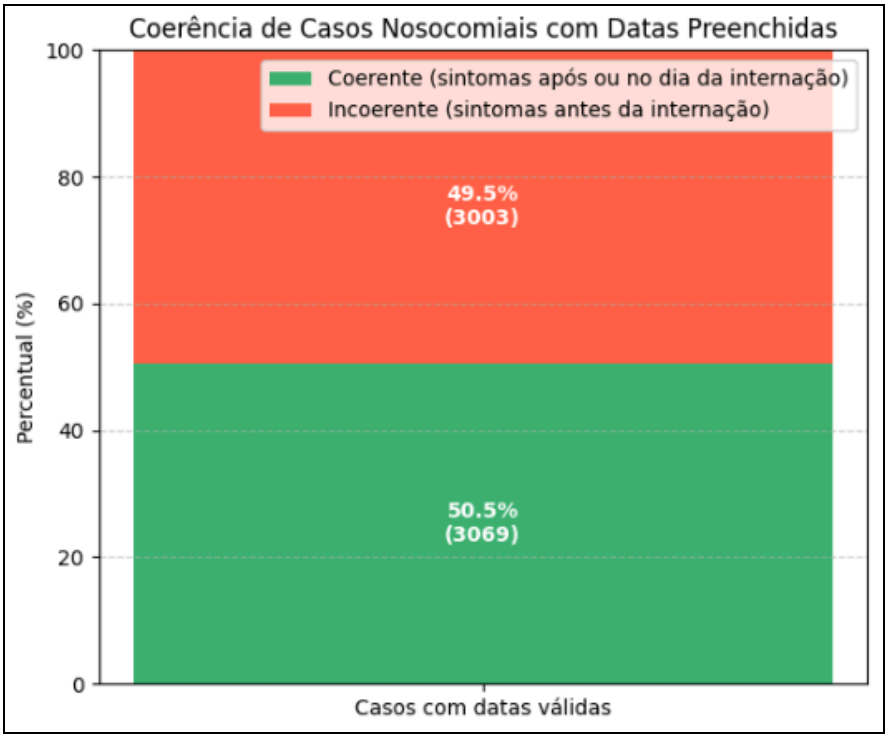
Com base no primeiro aspecto, percebe-se que a quantidade de casos de pacientes nosocomiais é muito baixa, sendo algo em torno de 1,89%, representando um valor de 26.905 casos.

No segundo ponto, ao investigar a presença de datas necessárias para verificação da coerência (internação e início dos sintomas), verificou-se, no primeiro gráfico, que 77,4 dos registros de pacientes nosocomiais são inválidos. Ou seja, muito mais da metade dos dados não possuem as colunas “DT_INTERNA” e “DT_SIN_PRI”.



No segundo gráfico, levando em conta apenas registros com datas devidamente preenchidas, foi feita uma outra comparação para verificar se os dados são consistentes com relação às duas colunas necessárias e, por isso, notou-se uma certa taxa de 49,5%, que representa quase metade, de registros que não estão corretamente preenchidos com base nas premissas mencionadas.

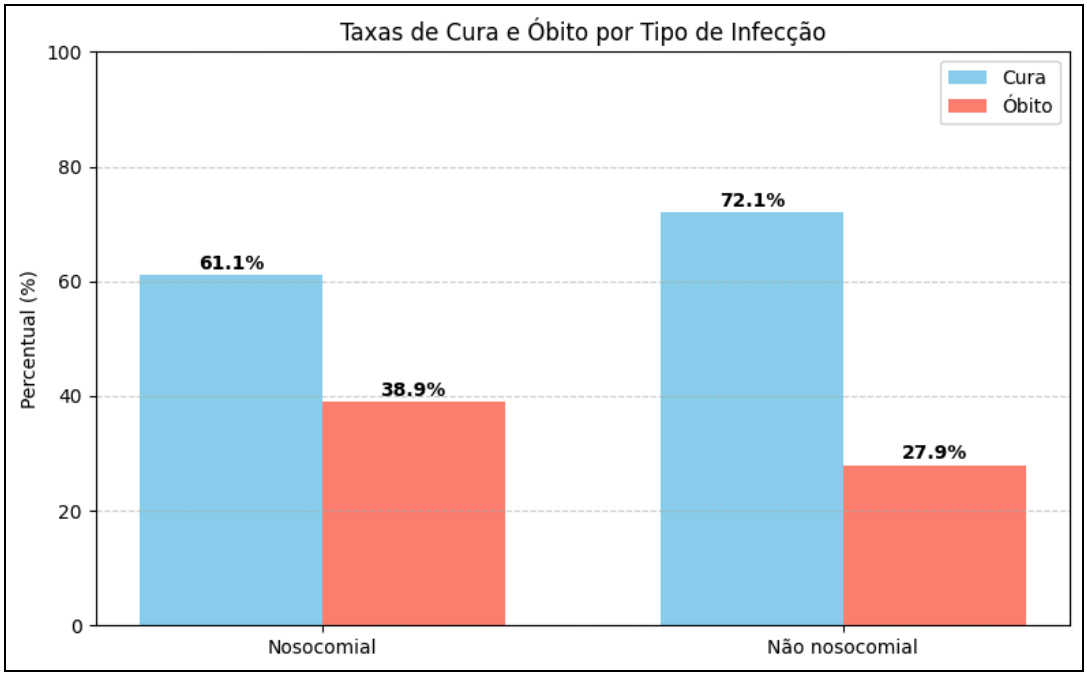
Portanto, nessa etapa da análise, conclui-se, pelo menos com base no que foi obtido até o momento usando conceitos de estatística descritiva, que há uma certa dificuldade na auditoria de registros. O que pode acarretar negativamente na precisão e validação de informações cruciais que podem servir para melhorias.



Por fim, avaliando o último aspecto abordado, em que os desfechos clínicos dos pacientes nosocomiais e não nosocomiais são comparados, nota-se que pacientes nosocomiais possuem uma taxa de óbito relativamente maior (cerca de 38,93%) que os pacientes não nosocomiais (cerca de 27,94%), o que implica numa suposição de que a pessoa contrair uma infecção quando está internada é relativamente mais perigoso.

Como não foi feita nenhuma análise inferencial, não se pode dizer muito, porém é uma ocasião que é válido investigar a fundo para entender as devidas causas da taxa ser maior.

	Cura	Óbito	Total	Taxa_óbito (%)	Taxa_cura (%)
Nosocomial	13952	8895	22847	38.93	61.07
Não nosocomial	828067	321076	1149143	27.94	72.06
Total Geral	842019	329971	1171990	28.15	71.85



2.2. Análises de Probabilidade

2.2.1. Análise 1: Qual a chance de um paciente ser internado e não ser internado?

2.2.1.1. Objetivos e preparações iniciais

Verificar a proporção de pacientes internados em comparação aos não internados. Para isso, foi necessário recortar a base de dados apenas com valores válidos da coluna "HOSPITAL" (1: internado e 2: não internado), excluindo os registros ignorados ou ausentes.

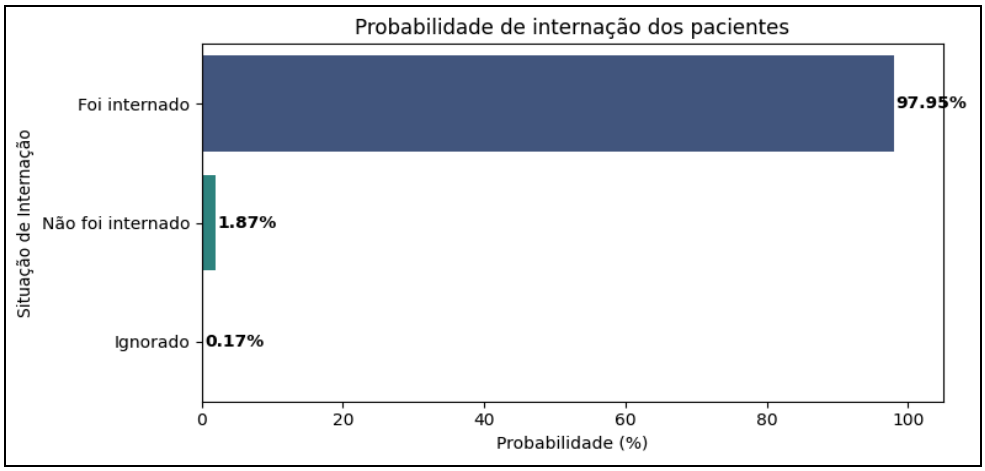
- Fórmula utilizada:

  Para analisar a questão proposta, foi utilizada a fórmula da probabilidade frequentista:

$$P\,(Evento) \;=\; \frac{Número\ de\ casos\ do\ evento}{Total\ de\ casos\ válidos}$$

2.2.1.2. Explicações relacionadas ao código desenvolvido

Os resultados obtidos foram obtidos através da utilização de funções fornecidas pela biblioteca do "Pandas" em cima da coluna estudada, permitindo a construção do gráfico abaixo que mostra a probabilidade simples de cada valor possível de "HOSPITAL":



2.2.1.3. Conclusões

A grande maioria dos registros diz respeito a pacientes internados. Isso revela um viés esperado da base, que é voltada para casos moderados a graves de SRAG. A alta taxa de internação, portanto, não compromete a qualidade dos dados, mas sim reforça o objetivo clínico da vigilância de casos graves. Por esse motivo, esses dados não devem ser generalizados para toda a população, pois os casos leves normalmente não estão incluídos.

2.2.2. Análise 2: Qual a chance de um paciente evoluir para cura, óbito e óbito por outras causas?

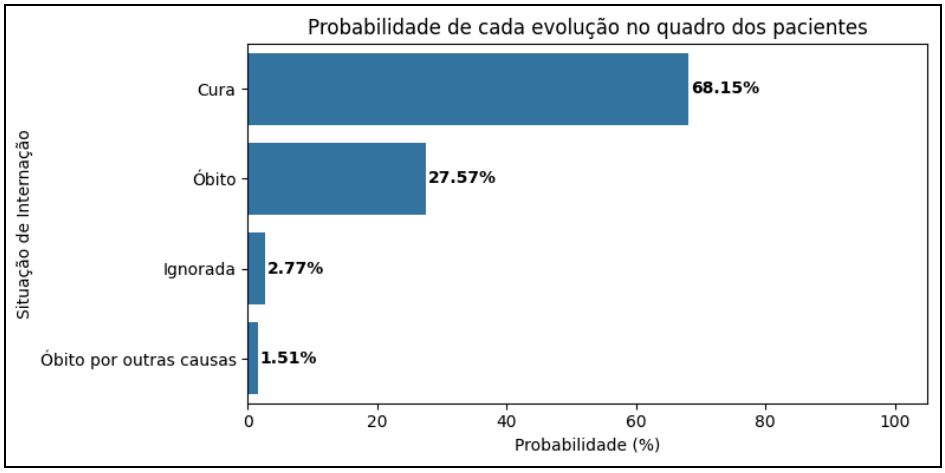
2.2.2.1. Objetivos e preparações iniciais

Verificar a proporção de pacientes que evoluíram para cura, óbito, óbito síndrome respiratória aguda, óbito por outras causas e ignorado, com base na variável “EVOLUCAO”. Para isso foram considerados apenas os valores válidos, excluindo os registros ignorados ou ausentes. Novamente, para analisar a questão proposta, foi utilizada a fórmula da probabilidade frequentista:

$$P\,(Evento) \;=\; \frac{Número\ de\ casos\ do\ evento}{Total\ de\ casos\ válidos}$$

2.2.2.2. Explicações relacionadas ao código desenvolvido

Os resultados obtidos foram obtidos através da utilização de funções fornecidas pela biblioteca do “Pandas” em cima da coluna estudada, permitindo a construção do gráfico abaixo que mostra a probabilidade simples de cada valor possível de “EVOLUÇÃO”:



2.2.2.3. Conclusões

Apesar da gravidade dos casos incluídos na base, a maior parte dos pacientes evoluiu para cura, o que mostra uma certa resposta positiva ao tratamento hospitalar em 2021. No entanto, a taxa de óbitos ainda é considerável, o que reflete a severidade clínica dos quadros registrados.

A análise combinada das colunas HOSPITAL e EVOLUCAO revela que a base de dados é composta, majoritariamente, por pacientes internados (97,55%). Esse dado evidencia um viés de gravidade clínica, indicando que a amostra de fato se concentra em casos moderados a graves de síndrome respiratória aguda (SRAG).

Entre os registros com evolução clínica preenchida (aproximadamente 91 mil), observou-se que 67,03% evoluíram para cura, enquanto 27,3% resultaram em óbito (por SRAG ou outras causas). Esses valores reforçam o perfil crítico da base analisada, mas também demonstram uma boa taxa de recuperação entre os pacientes atendidos em 2021.

A qualidade estatística da base pode ser considerada razoável, uma vez que há baixa quantidade de dados ignorados ou ausentes nas colunas analisadas. No entanto, recomenda-se cuidado ao extrapolar os resultados para a população geral, uma vez que o conjunto de dados tende a representar casos mais graves e hospitalizados, não refletindo toda a gama de manifestações clínicas da SRAG.

Vale destacar que a alta taxa de pacientes internados não representa, necessariamente, uma falha no conjunto de dados, mas sim uma característica comum e esperada na definição clínica do próprio SRAG, que é focada em casos mais graves. Como consequência, o alto

número de internações indica uma fidelidade intencional às condições propostas pelo base de dados, e não apenas um viés de seleção amostral conforme analisado.

2.2.3. Análise 3: Relação entre Internação e Desfecho Clínico em Pacientes

2.2.3.1. Objetivos e preparações iniciais

O objetivo dessa etapa é responder as perguntas propostas, que envolvem probabilidade condicional entre duas colunas ou variáveis da base de dados de SRAG:

- Hospital: Que indica se o paciente foi internado;
- Evolução: Que informa o desfecho do paciente (cura, óbito ou óbito por outras causas).

Com base nisso, buscamos calcular probabilidades do tipo:

- Qual a chance de um paciente internado evoluir para óbito?
- Qual a chance de cura entre os pacientes não internados?
- Qual a proporção de pacientes internados entre os que evoluíram para óbito?

$$P\,(A \mid B) \;=\; \frac{P\,(A \cap B)}{P(B)}$$

Embora o cálculo tenha sido feito de forma automática pelas bibliotecas utilizadas no Python, a fórmula utilizada para chegar no resultado foi a da probabilidade condicional, que é descrita por:

$$P\,(A \mid B) \;=\; \frac{P\,(A \cap B)}{P(B)}$$

Ao aplicar essa fórmula para responder às perguntas propostas anteriormente, poderíamos substituir da seguinte forma:

- Se quisermos saber a chance de um paciente ter óbito dado que foi internado, então:
    - A: Evento “Óbito”
    - B: Evento “Foi Internado”

$$P\,(\text{Óbito} \mid \mathit{Internado}) \;=\; \frac{P\,(\text{Óbito} \cap \mathit{Internado})}{P(\mathit{Internado})}$$

2.2.3.2. Explicações relacionadas ao código desenvolvido

Antes de calcular as probabilidades, foram realizadas algumas etapas de limpeza e estruturação dos dados:

1. Identificação dos valores válidos:

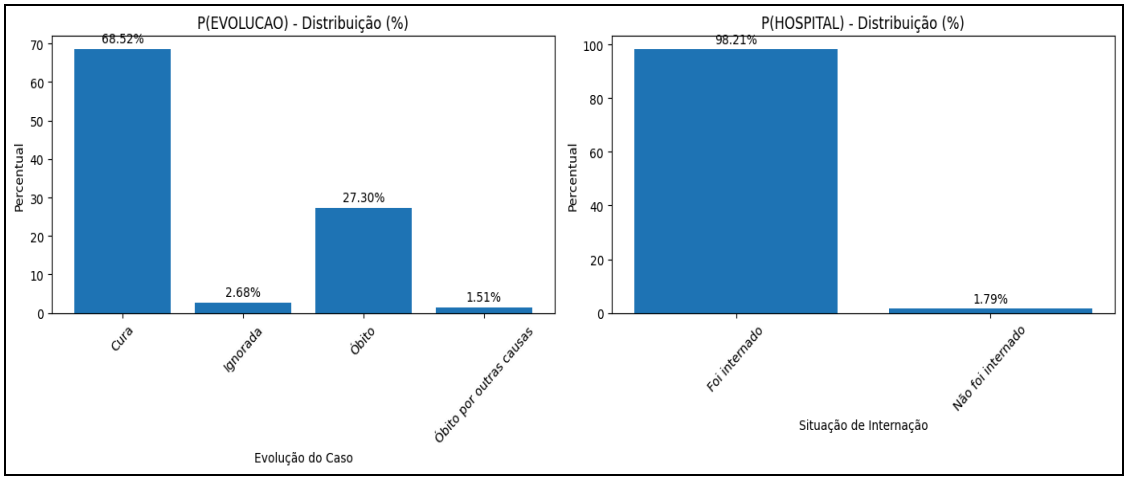
a. Para a variável HOSPITAL, foram considerados os códigos:
    - 1 = Internado
    - 2 = Não internado
    - 3 = Ignorado
b. Para a variável EVOLUCAO, os códigos válidos foram:
    - 1 = Cura
    - 2 = Óbito
    - 3 = Óbito por outras causas
    - 9 = Ignorado

2. Filtragem dos dados nulos: Foram mantidos apenas os registros com preenchimento em ambas as colunas (HOSPITAL e EVOLUCAO).
3. Cálculo das probabilidades marginais: Probabilidades individuais para cada valor de “HOSPITAL” e “EVOLUCAO”
4. Cálculo das probabilidades conjuntas: Probabilidade de ocorrência simultânea entre internação e evolução do paciente: P(HOSPITAL ∩ EVOLUCAO).
5. Cálculo das probabilidades condicionais: Derivadas da razão entre as conjuntas e as marginais, por exemplo:
    - P(Óbito | Internado)
    - P(Internado | Óbito)

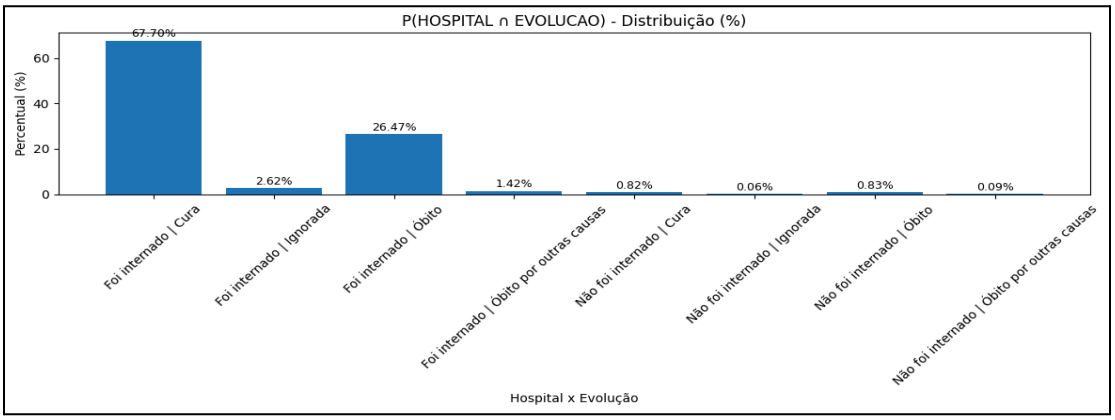
2.2.3.3. Conclusões

Os resultados obtidos foram obtidos através da utilização dos métodos de agrupamento da própria biblioteca do Pandas e do mecanismo de divisão baseado no índice. Tal característica permitiu que o cálculo de cada combinação possível (eventos) das duas colunas envolvidas fosse feito de forma automática. Diante disso, foi possível chegar nos seguintes resultados:

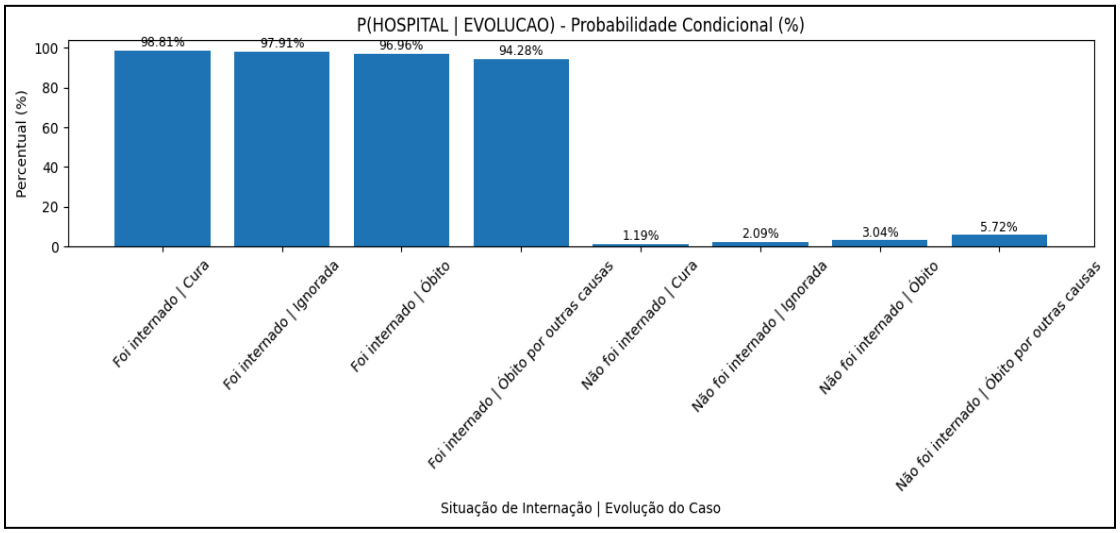
- **Probabilidade do condicionante (Evolução e Hospital):**



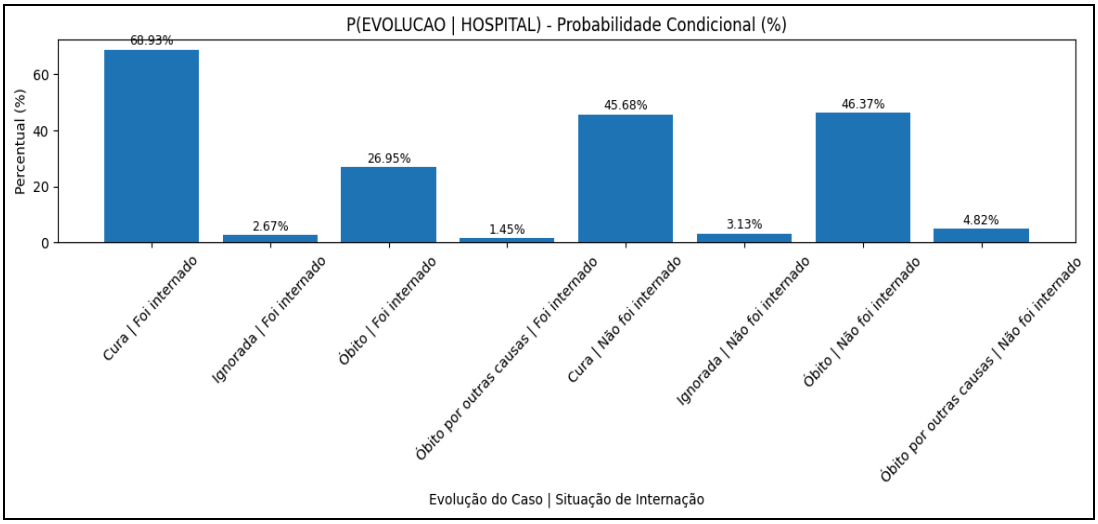
- **Probabilidade conjunta (Hospital ∩ Evolução):**



- **Probabilidade condicional (Hospital | Evolução):**



- **Probabilidade condicional P(Evolução | Hospital):**



- **Análise da Pergunta 1: Qual a probabilidade de óbito dado que o paciente foi internado?**

Embora tenha sido obtido o gráfico de forma automática, a fórmula utilizada para poder responder essa pergunta é descrito como:

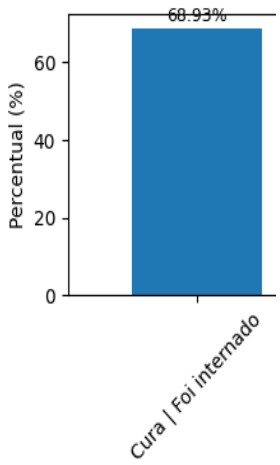
$$P\,(\acute{O}bito \mid Internado) \;=\; \frac{P\,(\acute{O}bito \;\cap Internado)}{P(Internado)}$$

O resultado obtido com a fórmula mostra que, entre os pacientes internados SRAG por influenza, cerca de 27% dos pacientes evoluíram para óbito. Com isso, fica evidenciado que, mesmo com suporte hospitalar, ainda existe um alto risco de morte, o que reflete a gravidade da condição clínica dos casos.

- **Análise da Pergunta 2: Qual a probabilidade de cura dado que o paciente foi internado?**

Embora tenha sido obtido o gráfico de forma automática, a fórmula utilizada para poder responder essa pergunta é descrito como:

$$P\,(Cura \mid Internado) \;=\; \frac{P\,(Cura \;\cap Internado)}{P(Internado)}$$

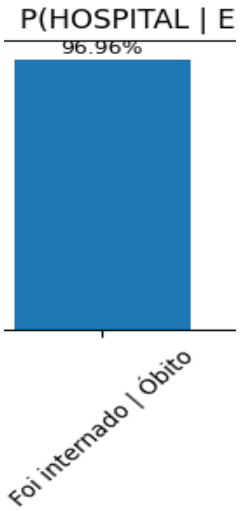


O resultado obtido com a fórmula utilizando a linguagem de programação Python mostra que, entre os pacientes internados SRAG por influenza, cerca de 68% dos pacientes foram curados. Com isso, fica evidenciado que o suporte hospitalar aumenta significativamente as chances de cura, o que reforça o papel crucial da hospitalização no tratamento desse tipo de síndrome.

- **Análise da Pergunta 3: Qual a probabilidade de internação dado que o paciente evoluiu para óbito?**

Embora tenha sido obtido o gráfico de forma automática, a fórmula utilizada para poder responder essa pergunta é descrito como:

$$P\,(Internado \mid Óbito) \;=\; \frac{P\,(Internado \cap Óbito)}{P(Óbito)}$$

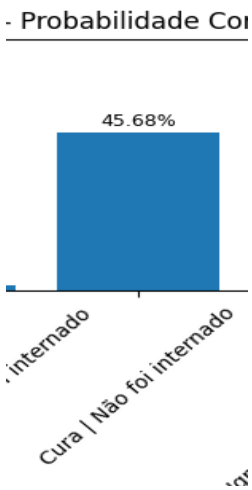


Quase 97% dos pacientes que evoluíram para óbito estavam internados. Esse dado sugere que, na maioria dos casos fatais registrados, o óbito ocorreu em monitoramento clínico. No entanto, essa informação não deve ser generalizada para toda a população, uma vez que o conjunto de dados utilizado é composto de forma majoritária por pacientes hospitalizados. Logo, qualquer probabilidade condicional que utilize a internação como evento condicionante pode apresentar probabilidades elevadas, refletindo a natureza da base de dados.

- **Análise da Pergunta 4: Qual a probabilidade de cura dado que o paciente não foi internado?**

Embora tenha sido obtido o gráfico de forma automática, a fórmula utilizada para poder responder essa pergunta é descrito como:

$$P\,(Cura \mid N.\ Internado) \;=\; \frac{P\,(Cura \cap N.\ Internado)}{P(N.\ Internado)}$$



Após a realização dos cálculos, podemos observar que cerca de 45% dos pacientes que não foram internados se curaram, ou seja, a chance de cura fora do ambiente hospitalar foi muito menor em comparação aos internados (quase 69%). Essa característica nos permite afirmar que a ausência de internação está associada a maior risco de morte do paciente ou evolução incerta.

Em síntese, os resultados demonstram que a internação hospitalar está associada a maior chance de cura no SRAG por Influenza, mas não elimina o risco de óbito. A análise condicional realizada evidencia que os pacientes internados apresentam uma taxa de mortalidade menor dos que não formam, reforçando a importância da hospitalização precoce para a diminuição da taxa de óbito. No entanto, devido ao viés da própria base de dados, focada apenas em pacientes hospitalizados, é necessário ter cuidado ao generalizar esses dados para a população em geral.

2.2.4. Análise 4: Dado que o teste antigênico deu positivo, qual a probabilidade de o paciente realmente ter alguma infecção respiratória viral confirmada por PCR?

2.2.4.1. Objetivos e preparações iniciais

O teste de antígenos é uma forma de diagnosticar a presença de alguma infecção respiratória viral, como Influenza, SARS-Cov-2 ou VSR, de forma rápida, sem a necessidade de analisar a fundo o DNA para identificar esses vírus. No entanto, esses testes possuem uma menor acurácia em comparação ao teste molecular (RT-PCR), que é considerado um “padrão-ouro”. Diante disso, esta análise busca avaliar, com base no Teorema de Bayes, qual a probabilidade de

um paciente estar infectado (baseado no teste RT-PCR), dado que teve como resultado positivo (baseado no teste de antígenos).

O objetivo da análise proposta é calcular a probabilidade de a pessoa ter de fato uma infecção respiratória viral, baseado no teste molecular (considerado padrão ouro), dado que a pessoa testou positivo para algum dos vírus listados no dicionário do conjunto de dados, utilizando a coluna do teste de antígeno, que fornece uma menor acurácia em relação ao molecular. Essa análise, pode ser formulada da seguinte forma em termos matemáticas: P (Doença | Teste Positivo).

Para resolver essa fórmula, optamos por aplicar o Teorema de Bayes, que nos auxilia no cálculo de probabilidades condicionais baseado numa crença a priori:

$$P\,(D \mid T\,+) \;=\; \frac{P(T+|D) * P(D)}{P(T+)}$$

Onde:

- *P*(*D*): Prevalência da doença (PCR positivo)
- *P*(T+): Proporção de testes antigênicos positivos;
- P (T+| D): Sensibilidade do teste antigênico (teste positivo entre os doentes);
- P (D | T+): Valor preditivo positivo (o que queremos descobrir).

Como foi utilizado o Teorema de Bayes para esse cálculo, foi necessário realizar o cálculo de três probabilidades distintas:

- *P*(*D*): Prevalência da doença (PCR positivo)
- *P*(T+): Proporção de testes antigênicos positivos;
- P (T+| D): Sensibilidade do teste antigênico (teste positivo entre os doentes);

Com isso, foi necessário realizar duas etapas para obter os resultados desejados.

- Etapa 1: Definição da Prevalência

A prevalência, nesse caso, é representada pela proporção de pacientes que realmente possuem uma infecção viral, independente do resultado obtido no teste rápido.

Como a base de dados do SRAG 2021 não apresenta toda a população e é composto principalmente por casos mais graves, não seria adequado estimar a prevalência a partir dela como reflexo da população geral.

Diante disso, foram utilizados os resultados obtidos na coluna PCR_RESUL igual a 1, que indica a presença da doença, para simular a prevalência, uma vez que ele é considerado um

padrão-ouro nos testes. Dessa forma, o cálculo da prevalência foi feito utilizando a fórmula abaixo:

$$P\,(Doença) \;=\; \frac{Número\ de\ testes\ PCR\ Positivo}{Total\ de\ registros\ válidos\ de\ RES\_AN\ e\ PCR\_RESUL}$$

2.2.4.2. Explicações relacionadas ao código desenvolvido

- **Etapa 2: Limpeza dos dados**

Antes da etapa de análise, foi necessário garantir que o conjunto de dados estivesse com registros válidos para ter o melhor resultado possível, utilizando as colunas interessadas:

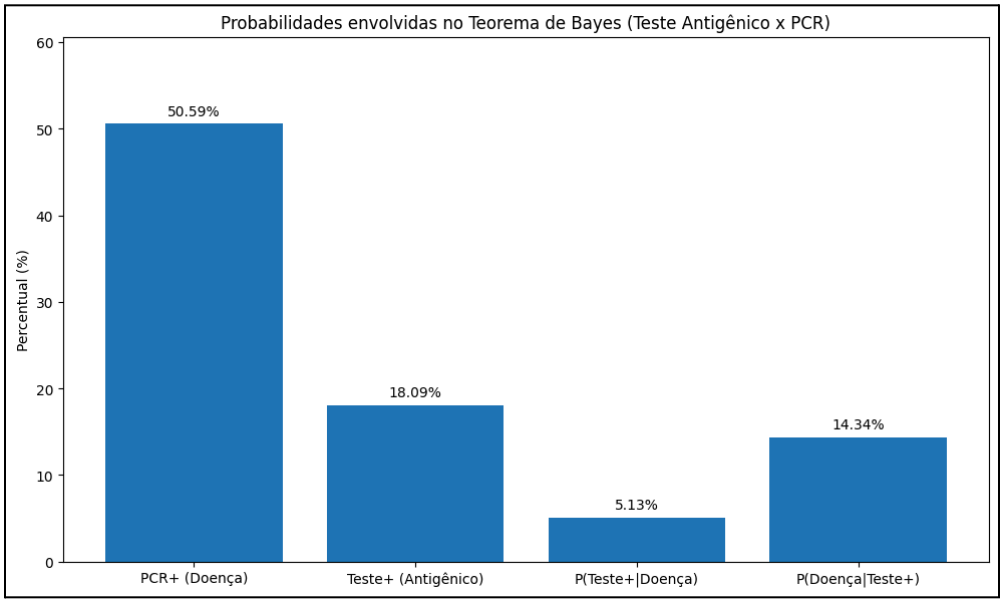
- RES_AN: resultado do teste antigênico (1 = positivo);
- PCR_RESUL: resultado do teste molecular (1 = detectável = positivo);

Dessa forma, os seguintes critérios foram aplicados a cada registro:

- Remoção de valores nulos
- Consideração de apenas os códigos válidos definidos no dicionário de dados (valores de 1 a 9).

2.2.4.3. Conclusões

Os valores observados na base foram os seguintes:



O resultado obtido com a aplicação do Teorema de Bayes mostra que, mesmo quando o teste antigênico é positivo, a probabilidade de a pessoa ter de fato alguma infecção viral respiratória confirmada por PCR é de apenas 14.34%. Isso mostra que o teste antigênico possui

um valor preditivo positivo relativamente baixo, mesmo em um contexto onde a prevalência da doença é alta (mais de 50%). Além disso, a baixa sensibilidade observada (~5%) pode estar relacionada com uma limitação técnica do teste rápido em comparação aos moleculares.

3. ANÁLISES ENTRE OS DOIS DATASETS (SIM + SRAG)

Nessa seção do relatório, serão abordados assuntos de estatística inferencial para relacionar os dois datasets conforme as solicitações do projeto.

3.1. Análises Inferenciais

3.1.1. Análise 1: Comparação entre proporções de COVID-19 entre SIM e SRAG

3.1.1.1. Objetivos e preparações iniciais

Essa análise tem como objetivo verificar se a proporção de casos evoluídos para óbito por COVID-19 (SRAG) é compatível com a proporção de óbitos (SIM) por COVID-19, realizando agregação faixa etária.

O tópico abordado na estatística inferencial nessa análise será o teste de hipótese para a diferença entre duas proporções populacionais.

Alguma das perguntas que podem ser feitas para essa análise são as seguintes:

- Existe diferença estatisticamente significativa entre a proporção de casos de COVID-19 no SRAG e a proporção de óbitos no SIM?
- Quais faixas etárias apresentam maiores diferenças entre proporções?
- A proporção de COVID-19 por faixa etária se mantém semelhante em ambos os sistemas?
- Os dados do SRAG são suficientes para inferir sobre os óbitos por COVID-19 ou há subnotificação ou atraso nos registros do SIM?

  Diante disso, algumas preparações iniciais foram necessárias, como:
- Identificação das colunas relevantes, nesse caso:
  - SRAG: CLASSI_FIN, coluna que armazena a classificação clínica final do paciente, nesse caso nosso objetivo era olhar para o valor 5, que representa a presença de COVID nos pacientes como classificação final do SRAG
  - SIM: LINHAA, LINHAB, LINHAC, LINHAD e LINHAII, essas colunas foram utilizadas para identificar os óbitos por COVID-19, através do CID identificado em fontes confiáveis. Nesse caso, os valores para realizar o filtro nas colunas foram:

        - B34.2: Casos leves ou suspeitos antes da confirmação por PCR.
        - U07.1: COVID-19 confirmado laboratorialmente.
        - U07.2: COVID-19 clinicamente ou epidemiologicamente diagnosticado, mas não confirmado por teste
    - Padronização e limpeza dos CIDs de causa básica.
    - Filtro por diagnósticos compatíveis com COVID-19: U07.1, U07.2, B34.2.
    - Conversão de idades em faixas etárias padronizadas (de 5 em 5 anos) para SRAG e SIM.

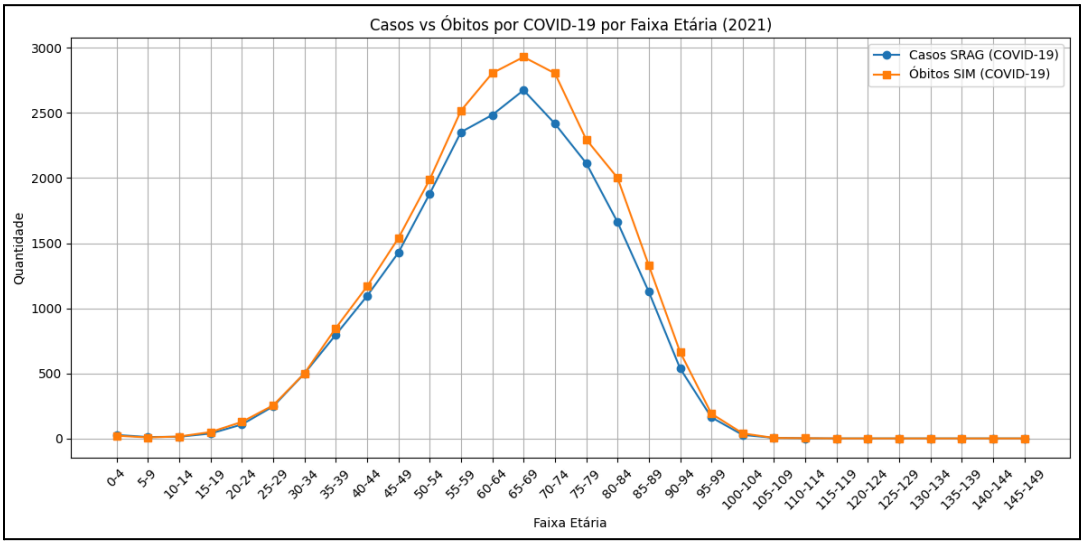
3.1.1.2. Explicações relacionadas ao código desenvolvido

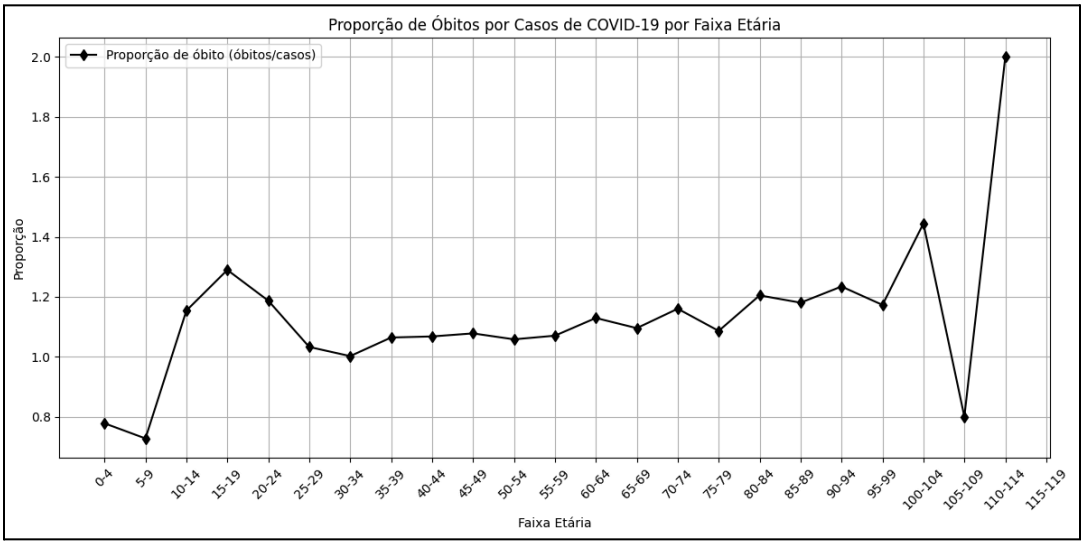
Para ter um resultado satisfatório, foi necessário realizar diversas etapas que serão listadas abaixo:

- Etapa 1: Limpeza dos Dados
    - SRAG: Seleção de casos confirmados de COVID-19 com evolução para óbito.
    - SIM: Padronização dos CIDs e seleção dos registros com CIDs compatíveis com COVID-19.
- Etapa 2: Tratamento da Idade
    - Conversão dos códigos de idade (SIM) em valores absolutos (anos).
    - Agrupamento em faixas etárias de 5 anos para padronização.
- Etapa 3: Agrupamento e Cruzamento
    - Contagem de casos por faixa etária em cada base.
    - Cálculo da proporção de óbitos (SIM) em relação aos casos graves (SRAG).
    - Comparação gráfica: quantidade, proporção por faixa, proporção relativa ao total.
- Etapa 4: Teste de Hipóteses
    - Aplicação do teste Z para proporções.
    - Cálculo do “p-valor” e intervalo de confiança de 95%.
    - Hipóteses:
        - H_0: A proporção de óbitos por COVID-19 na faixa etária I é igual entre os sistemas SRAG e SIM
        - H_1: A proporção de óbitos por COVID-19 difere entre os sistemas SRAG e SIM na faixa etária I.

3.1.1.3. Conclusões

Através das etapas realizadas, fomos capazes de extrair informações cruciais para a análise de dados na perspectiva da estatística descritiva. Inicialmente, ao realizar o tratamento dos casos de COVID-19 e agrupamento das idades nas duas bases de dados, foi possível gerar dois gráficos que revelam uma característica interessante entre as duas bases.

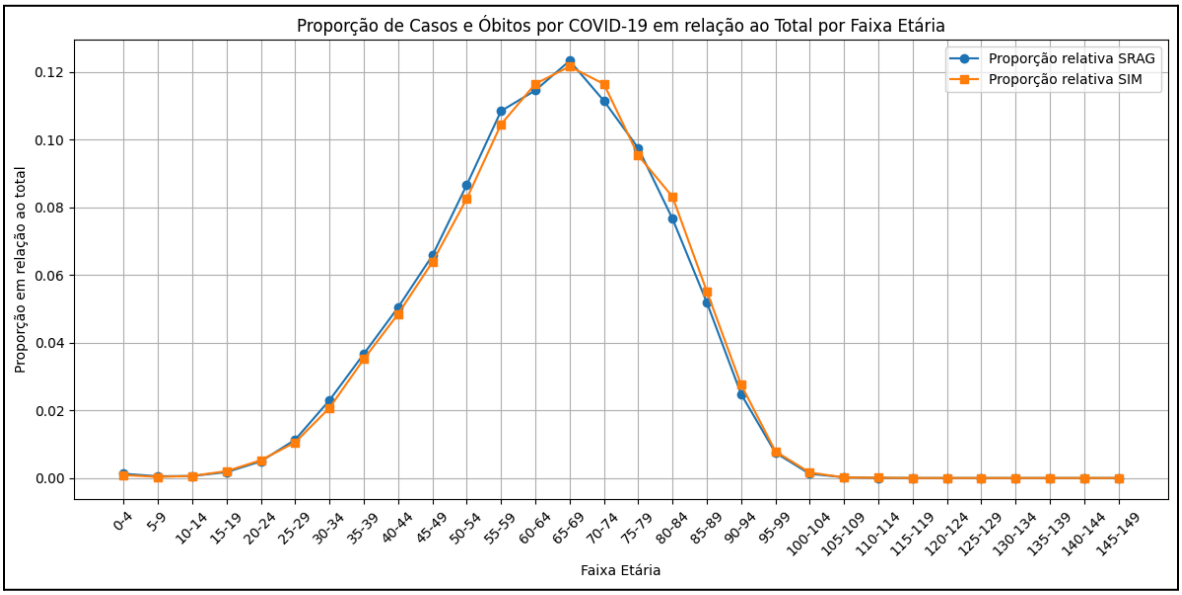




No primeiro e segundo gráficos podemos observar que de certa forma a quantidade de casos de óbito por COVID-19 no SRAG e no SIM estão bem próximos, o que indica que os dois sistemas estão de certa com um bom relacionamento entre eles, uma vez que são provenientes de uma mesma fonte de dados. No entanto, é possível observar, tanto no primeiro quanto no

segundo gráfico, que algumas faixa-etárias apresentam uma diferença maior em relação às outras, o que pode indicar uma subnotificação ou casos onde as pessoas dessas idades (55 a 84), podem não estar sendo avaliadas clinicamente, tendo seu resultado obtido de forma informal.

Além disso, para dar continuidade na exploração e investigação dos dados, encontramos a proporção entre casos de óbito por COVID por faixa etária e a quantidade total de casos de óbitos por COVID, em ambas as bases. O resultado mostra que, embora sejam sistemas diferentes, a proporção de pessoas que morreram por COVID é praticamente a mesma, o que sustenta a nossa teoria de que os sistemas estão bem sincronizados, abaixo temos o gráfico que foi citado:



Diante disso, para reforçar ainda mais a nossa ideia de que os sistemas apresentam padrões semelhantes de distribuição entre as faixas etárias, aplicamos o teste de hipótese para a comparação de proporções em cada faixa etária (Teste Z para proporções em diferentes amostras), utilizando a proporção de óbitos em cada sistema (SIM e SRAG). Dessa forma, foi possível verificar, faixa por faixa, se a proporção de registros se mantém. Com isso, temos que:

- Hipóteses do teste
    - Hipótese nula (H_0): A proporção de óbitos por COVID-19 na faixa etária I é igual entre os sistemas SRAG e SIM.
        - H_0: $p_1 = p_2$
    - Hipótese alternativa (H_1): A proporção de óbitos por COVID-19 difere entre os sistemas SRAG e SIM na faixa etária I

- $H_1: p_1 \neq p_2$

- Fórmulas Utilizadas no Teste
  - Proporções Individuais

    $$p1 = \frac{x1}{n1}, p2 = \frac{x2}{n2}$$

    Onde:

    - $x1$ = Número de óbitos por faixa etária no SIM
    - $x2$ = Número de casos graves com óbito por faixa etária no SRAG
    - $n1, n2$ = Totais das amostras (SIM e SRAG)
      - Proporção Combinada (pooling):

        $$p1 = \frac{x1 + x2}{n1 + n2}$$

      - Erro Padrão:

        $$SE = \sqrt{p\,(1 - p)\left(\frac{1}{n1} + \frac{1}{n2}\right)}$$

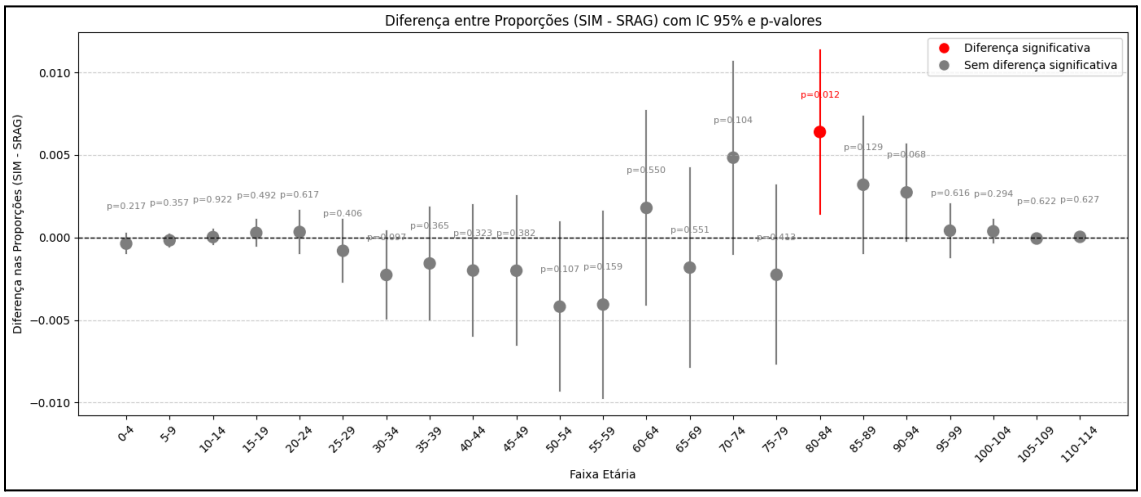
      - Estatística Z:

        $$Z = \frac{p1 - p2}{SE}$$

      - Intervalo de Confiança (95%):

        $$IC = (p1 - p2) +- 1,96 * SE$$

O gráfico obtido através desses cálculos é exibido abaixo:



Nesse gráfico, podemos observar algumas informações que são essenciais para a validação de hipóteses, como o resultado final (Difere da hipótese nula ou não), a diferença de proporções, o p-valor e o intervalo de confiança representado pela altura da linha em cada ponto,

sendo aplicado isso para cada faixa etária identificada. Com base no teste Z, para a maioria das faixas etárias o p-valor foi superior a 0,05, indicando ausência de diferença estatisticamente significativa entre os dois sistemas.Isso reforça a coerência e convergência entre as bases de dados para a vigilância da mortalidade por COVID-19.

Em uma outra perspectiva, agora podemos responder às perguntas formuladas da seguinte forma:

1. **Existe diferença estatisticamente significativa entre a proporção de casos de COVID-19 no SRAG e a proporção de óbitos no SIM?** Sim, mas apenas em faixas específicas. Grande parte das faixas não apresentou diferença significativa entre a proporção de óbitos (SIM) e casos graves (SRAG). No entanto, uma outra faixa (como 80 e 84 anos) apresentou um "p-valor" menor que 0.05, uma diferença maior no gráfico e um aumento da proporção de óbitos/casos, indicando diferença significativa.
2. **Quais faixas etárias apresentam maiores diferenças entre proporções?** A faixa de 80-84 anos foi a única que apresentou uma diferença estatística significativa. As demais faixas mantêm diferenças pequenas ou estatisticamente não signficiativas.
3. **A proporção de COVID-19 por faixa etária se mantém semelhante em ambos os sistemas?** De modo geral, sim, uma vez que as curvas de distribuição do SRAG e do SIM são bastante semelhantes, conforme o gráfico de proporções relativas, o que reforça a consistência entre os sistemas. Tanto o SIM quanto SRAG seguem um padrão de distribuição etária típico de mortalidade da COVID-19.
4. **Os dados do SRAG são suficientes para inferir sobre os óbitos por COVID-19 ou há subnotificação ou atraso nos registros do SIM?** Sim, pois existe um forte alinhamento entre os dados do SRAG e do SIM em quase todas as faixas etárias indica que o SRAG pode ser uma boa base de dados para predição de mortalidade por COVID-19, embora sejam notadas pequenas variações ou atrasos em faixas específicas que não devem ser ignoradas.

Em suma, a análise comparativa entre os sistemas SRAG e SIM revelou uma alta correlação entre os padrões etários de casos graves e óbitos por COVID-19 em 2021. Foram

aplicados testes de hipótese (teste Z para proporções) para verificar se as proporções por faixa etária diferem estatisticamente entre as bases. Os resultados indicam que na maioria das faixas etárias não há diferença significativa, reforçando a consistência entre os dois sistemas de notificação. A exceção foi observada na faixa de 80–84 anos, onde a diferença foi estatisticamente significativa. Ainda assim, essa variação pontual não compromete a tendência geral observada nos dados.

Além disso, a distribuição proporcional em relação ao total — apresentada em gráfico adicional — mostra que os perfis de mortalidade e gravidade da doença por faixa etária se mantêm bastante similares.

Concluímos, portanto, que os dados do SRAG são, em geral, representativos e confiáveis para inferir sobre a mortalidade por COVID-19, com ressalvas pontuais que podem refletir variações operacionais, atraso de notificação ou fatores específicos de coleta em algumas faixas etárias.

3.1.2. Análise 2: Comparação entre proporções de óbitos por sexo entre SRAG e SIM

3.1.2.1. Objetivos e preparações iniciais

Essa análise tem como objetivo verificar se há diferença estatisticamente significativa nas proporções de óbitos por COVID-19 entre homens e mulheres nos dois principais sistemas de informação em saúde pública do Brasil: SRAG (dados hospitalares) e SIM (dados de mortalidade).

A hipótese principal investigada é se o perfil de sexo dos óbitos por COVID-19 é consistente entre os registros hospitalares (SRAG) e os registros oficiais de óbito (SIM), o que pode indicar confiabilidade cruzada dos sistemas ou possíveis distorções.

O tópico abordado na estatística inferencial nesta análise é: Teste de hipótese para a diferença entre proporções populacionais categóricas (sexo).

- Identificação das colunas relevantes, nesse caso:
    - SRAG:
        - CLASSI_FIN: coluna que armazena a classificação clínica final do paciente, nesse caso nosso objetivo era olhar para o valor 5, que

representa a presença de COVID nos pacientes como classificação final do SRAG

        - EVOLUCAO: Evolução do caso — usado apenas valor 2 (óbito).
    - SIM:
        - CAUSABAS: Causa básica do óbito — filtramos os CIDs relacionados à COVID-19: U071, U072, B342.
        - SEXO: Sexo do falecido (valores 1 = masculino, 2 = feminino).

3.1.2.2. Explicações relacionadas ao código desenvolvido

Antes de prosseguir com as partes cruciais relacionadas às análises inferenciais, foram feitos alguns tratamentos no código.

- SRAG:
    - O primeiro filtro realizado foi a seleção do registros com COVID-19 confirmado e evolução para óbito
    - CS_SEXO padronizado como 'M' e 'F'.
- SIM:
    - Filtragem de causas básicas com códigos CID-10 compatíveis com COVID-19.
    - SEXO convertido (1 = 'M', 2 = 'F').

Por fim, foram feitos os devidos cálculos das proporções com a contagem de óbitos masculinos e femininos em cada sistema, além do cálculo da proporção de óbitos masculinos e femininos dentro de cada base. Após isso, foram estabelecidas as hipóteses

- Hipóteses do teste
    - Hipótese nula (H_0): A distribuição de óbitos por sexo é igual entre SRAG e SIM (não há associação entre sistema e sexo).
        - H_0: $p_1 = p_2$
    - Hipótese alternativa (H_1): A distribuição de óbitos por sexo diferencia os sistemas (existe associação entre sistema e sexo).

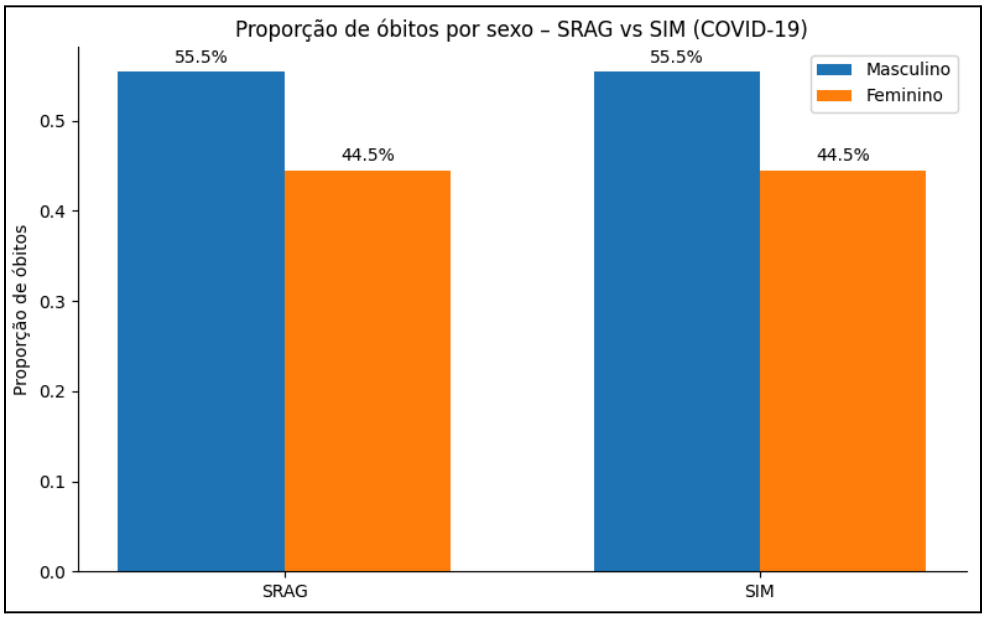
Para realizar os testes, foi utilizado o teste Qui-quadrado de independência, usando uma tabela de contingência para ambos os sexos das amostras.

Por fim, para melhor visualização dos resultados, foi gerado um gráfico de barras agrupadas com as proporções de óbitos masculinos e femininos nos dois datasets.

3.1.2.3. Conclusões

Através das etapas realizadas, foi possível conduzir uma análise robusta sob a perspectiva da estatística inferencial, focada na verificação de coerência entre os sistemas de notificação SRAG e SIM quanto à variável sexo dos óbitos por COVID-19.

Com os dados filtrados, foram calculadas as proporções de óbitos masculinos e femininos em cada sistema. Essas proporções foram representadas graficamente em um gráfico de barras agrupadas, o qual permitiu uma visualização clara da semelhança entre as bases: tanto no SRAG quanto no SIM, aproximadamente 55,5% dos óbitos por COVID-19 ocorreram entre homens e 44,5% entre mulheres.



Esse resultado reforça a ideia de que os dois sistemas estão fortemente sincronizados em relação ao perfil de sexo dos óbitos por COVID-19. A inexistência de diferença significativa entre SRAG e SIM indica que o sistema hospitalar (SRAG) reflete bem a distribuição por sexo observada nos registros de mortalidade (SIM), pelo menos no contexto analisado.

3.1.3 Análise 3: A idade média ao óbito difere entre estados com alta e baixa incidência de internações por SRAG?

3.1.3.1 Objetivos e preparações iniciais

Essa análise tem como objetivo verificar se existe uma diferença estatisticamente significativa na idade média ao óbito (dados do SIM) quando comparamos dois grupos de Unidades Federativas (UFs): aquelas com alta taxa de internação por Síndrome Respiratória Aguda Grave (SRAG) e aquelas com baixa taxa.

A hipótese principal investigada é se uma maior pressão sobre o sistema hospitalar, representada por uma alta taxa de internações por SRAG, está associada a um perfil etário de mortalidade diferente (mais jovem ou mais velho) em comparação com locais com menor incidência de internações.

O tópico de estatística inferencial abordado nesta análise é: Teste de hipótese para a diferença entre médias de duas amostras independentes (Teste t). Foram identificadas e utilizadas as seguintes colunas relevantes dos sistemas SIM e SRAG:

- SIM (Sistema de Informações sobre Mortalidade):
    - CODMUNOCOR: Código do município de ocorrência, utilizado para extrair a sigla da UF.
    - IDADE: Código que representa a idade do falecido em diferentes unidades (minutos, horas, meses, anos).
- SRAG (Sistema de Informação de Vigilância Epidemiológica da Gripe):
    - SG_UF_NOT: Sigla da Unidade Federativa de notificação da internação.
- Dados Externos:
    - Foram usados dados do IBGE sobre a população em cada UF para melhorar as análises. Foram utilizados para conseguir valores de taxas, pois pelo SIM e SRAG não temos acesso a população inteira, nos obrigando a utilizar os valores de registros desse banco (absolutos).

3.1.3.2 Explicações relacionadas ao código desenvolvido

Antes da análise inferencial, foram realizadas diversas etapas de tratamento e preparação dos dados:

1. **Criação da Idade Padronizada (SIM):** Foi criada a coluna IDADE_ANOS no dataset do SIM. Uma função customizada foi desenvolvida para converter a coluna IDADE (que usa um sistema de códigos para minutos, horas, meses e anos) em uma única unidade padronizada: a idade em anos.
2. **Cálculo da Idade Média ao Óbito por UF (SIM):** O dataset do SIM foi agrupado por UF para calcular a idade média ao óbito em cada estado.
3. **Cálculo da Taxa de Internação por UF (SRAG):** No dataset do SRAG, as internações foram contadas para cada UF. Em seguida, utilizando dados populacionais externos, foi calculada a taxa de internação por 100.000 habitantes para cada estado.
4. **Agrupamento por Incidência:** Os estados foram divididos em dois grupos: "Alta Incidência" e "Baixa Incidência". O critério de divisão foi a mediana da taxa de internação: estados com taxa acima da mediana foram classificados como alta incidência, e os demais como baixa incidência.

Após a preparação, foram estabelecidas as hipóteses para o teste estatístico:

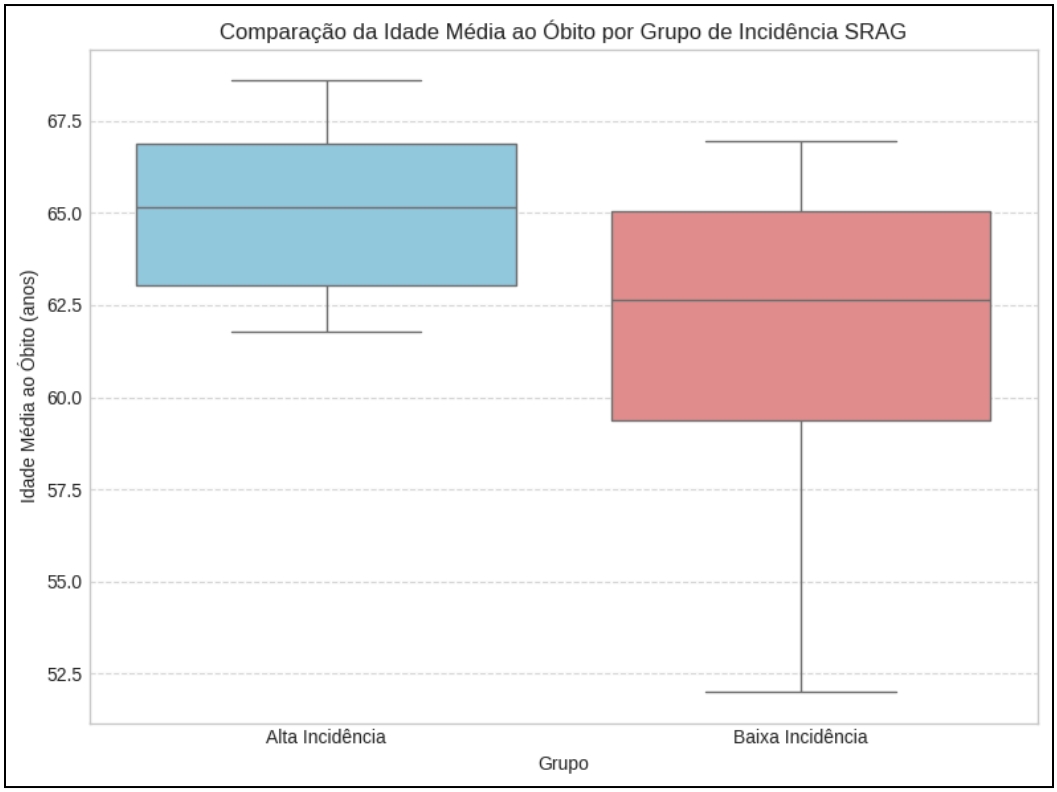
- Hipóteses do teste
    - **Hipótese nula (H_0):** A idade média ao óbito é a mesma para os grupos de UFs com alta e baixa incidência de internações. Não há diferença entre as médias dos grupos.
        - H_0: $\mu alta = \mu baixa$
    - **Hipótese alternativa (H_1):** A idade média ao óbito é diferente entre os dois grupos.
        - H_1: $\mu alta = \mu baixa$

Para verificar as premissas do teste, foi realizado o teste de normalidade de Shapiro-Wilk em ambos os grupos. Em seguida, foi aplicado o Teste t para amostras independentes para comparar as médias de idade ao óbito dos dois grupos. Por fim, para visualização dos resultados, foi gerado um gráfico boxplot comparando a distribuição da idade média ao óbito entre os grupos de alta e baixa incidência.

3.1.3.3 Conclusões

Através das etapas realizadas, foi possível conduzir uma análise robusta para verificar se a intensidade da circulação de SRAG (medida pela taxa de internação) se correlaciona com a idade média dos óbitos.

As estatísticas descritivas mostraram as idades médias ao óbito para cada grupo. O boxplot permitiu uma comparação visual, mostrando a dispersão e a tendência central da idade em cada um dos dois cenários (alta e baixa incidência).



O resultado do **Teste t independente** é o ponto central desta análise. Como obtivemos um p-valor menor do que 0,05, concluímos que devemos rejeitar a hipótese nula, ou seja, existe de fato uma associação estatística significativa entre idade e óbito. Os dois boxplots e as medidas resumo calculadas no código demonstram isso.

Como a idade média aumenta, podemos assumir que as doenças SRAG causam mais óbitos em pessoas de idade. Isso de fato ocorre e ocorreu no período da pandemia, onde pessoas de idade, mais fracas e com outras condições de saúde e comorbidades acabavam sendo mais afetadas pelo COVID.

Com base nos resultados gerados pelo código, pode-se concluir se a pressão hospitalar por SRAG em um estado tem ou não uma associação estatisticamente significativa com o perfil etário daqueles que vão a óbito.

3.1.4 Análise 4: A taxa de mortalidade por COVID-19 no SIM é proporcional à letalidade observada no SRAG?

3.1.4.1. Objetivos e preparações iniciais

O objetivo desta análise é realizar um **teste de hipótese para proporções**, a fim de verificar se existe diferença significativa entre:

- A **taxa de mortalidade por COVID-19** observada no **Sistema de Informação sobre Mortalidade (SIM)**;
- E a **letalidade aparente por COVID-19** observada nos registros da **Síndrome Respiratória Aguda Grave (SRAG)**.

A comparação é realizada entre duas amostras independentes extraídas dos respectivos datasets, ambas referentes ao ano de 2021. Foram adotados os seguintes parâmetros:

1. **Hipótese Nula (H_0):** a taxa de mortalidade do SIM é proporcional à letalidade observada no SRAG.
2. **Hipótese Alternativa (H_1):** há **diferença significativa**, o que pode indicar **subnotificação ou excesso** em alguma das bases.
3. **Nível de significância (α):** 0,05.

3.1.4.2. Explicações relacionadas ao código desenvolvido

A metodologia utilizada envolveu os seguintes passos:

1. **SIM – Taxa de mortalidade por COVID-19:**

   Foi aplicado um filtro para registros cuja causa básica (CAUSABAS) fosse igual a **B342**, correspondente à infecção por coronavírus.

   A proporção de mortes por COVID-19 (p_sim_obito_covid) foi calculada em relação à amostra total do SIM.

A proporção complementar foi calculada para óbitos não relacionados à COVID-19.

2. **SRAG – Letalidade aparente por COVID-19:**
   Considerou-se como caso de óbito por COVID-19 os registros onde CLASSI_FIN_DESC era "Por covid-19" e EVOLUCAO_DESC era "Óbito".
   A quantidade total de óbitos por COVID-19 foi comparada ao total da amostra analisada (base SIM).
3. **Cálculo das proporções:**
   Os valores esperados foram obtidos ao aplicar a proporção do SIM (mortalidade) sobre o tamanho da amostra do SRAG.
   Foi utilizado o teste de qui-quadrado para uma amostra (scipy.stats.chisquare), comparando os valores observados (SRAG) com os esperados (SIM).
4. **Visualização dos resultados:**
   Um gráfico de barras foi construído para comparar visualmente os valores observados (SRAG) e esperados (SIM) para óbitos por COVID-19 e não COVID-19.

3.1.4.3. Conclusões

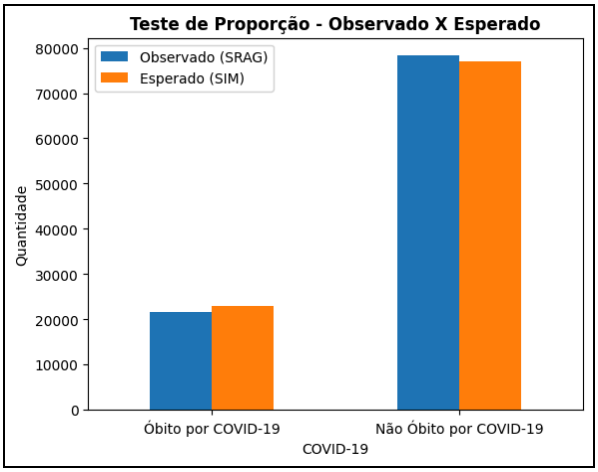
A estatística do teste e o p-valor foram obtidos:

- **Estatística do Qui-Quadrado (χ^2):** 94.4014
- **p-valor:** 0.000000000 → 0.0000000% (extremamente inferior a 0,05)

Com base no valor obtido do p-valor e no nível de significância definido ($\alpha = 0,05$), foi possível concluir que:

- Se o p-valor < 0,05, então rejeitamos a hipótese nula, ou seja, há evidências de diferença significativa entre os dados das duas bases;
- Se o p-valor ≥ 0,05, então não rejeitamos H_0, indicando que não há evidência estatística de diferença nas proporções.

Essa análise fornece um indicativo importante sobre a **confiabilidade e consistência entre as bases SRAG e SIM**, e pode apontar possíveis **subnotificações, inconsistências ou variações regionais** na coleta e registro dos óbitos por COVID-19.


Teste de Proporção - Observado X Esperado
Observado (SRAG)
Esperado (SIM)
Quantidade
80000
70000
60000
50000
40000
30000
20000
10000
0
Óbito por COVID-19
Não Óbito por COVID-19
COVID-19

REFERÊNCIAS

Prefeitura Municipal de São José dos Pinhais. INSTRUÇÃO NORMATIVA No 12 /2020 - SEMS. Disponível em: https://www.sjp.pr.gov.br/wp-content/uploads/2020/05/INS-NORM-12-1.pdf. Acesso em: 20 jul. 2025.

BABYSEC. Peso nos recém-nascidos. Disponível em: https://www.babysec.com.br/nota/peso-nos-recem-nascidos. Acesso em: 28 jul. 2025

IBGE. ESTIMATIVAS DA POPULAÇÃO RESIDENTE NO BRASIL E UNIDADES DA FEDERAÇÃO COM DATA DE REFERÊNCIA EM 1º DE JULHO DE 2021. Disponível em: https://www.ibge.gov.br/estatisticas/sociais/populacao/9103-estimativas-de-populacao.html?edicao=31451. Acesso em: 29 jun. 2025.