RELATÓRIO

TRABALHO DE EXTENSÃO

INTRODUÇÃO À INFERÊNCIA ESTATÍSTICA
ICMC - USP

1. Introdução

Este estudo é uma análise estatística inferencial de um conjunto de dados relacionados ao risco de **doença cardíaca coronariana nos próximos 10 anos**, baseada em fatores comportamentais, médicos e demográficos. Ele foi desenvolvido pelos alunos **Victor Botelho, Arthur Filliettaz e Thiago Tavares** e faz parte de um trabalho de extensão da matéria de Introdução à Inferência Estatística, ministrada pela professora Cibele Russo, do curso de Ciência de Dados do ICMC da USP.

O objetivo principal foi identificar variáveis que influenciam diretamente esse risco de doença cardíaca, que abreviaremos por CHD, aplicando os testes estatísticos adequados com base nas peculiaridades de cada um.

O dataset utilizado foi o Framingham Heart Study, disponível no Kaggle, fruto de um estudo norte-americano contendo mais de 4000 registros de pacientes com variáveis como idade, colesterol, pressão arterial, glicose, IMC, tabagismo, entre outras.

2. Métodos e Critérios Utilizados

Nosso estudo começa importando as bibliotecas que serão usadas em nosso código assim como o carregamento dos dados, baixando diretamente o dataset do site Kaggle e exibindo seu começo. Em seguida, verificamos a quantidade de linhas e colunas do dataset e criamos uma lista de todas as variáveis, exibindo sua quantidade não nula e seu tipo de dado.

Na sequência, realizamos um pré-processamento: renomeamos as variáveis para português para facilitar o entendimento e nossos estudos com elas. Feito isso, exibimos algumas características estatísticas das variáveis, como média, valores máximos e mínimos e seus quartis.

Então, passamos para a análise exploratória dos dados. Primeiramente, construímos histogramas das variáveis contínuas e discretas, permitindo observar a distribuição geral dos dados. Após isso, avançamos para uma seção com diversos tipos de gráficos, sendo o primeiro deles um heatmap de matriz de correlação, revelando relações importantes como a tendência de aumento do colesterol com a idade. Similarmente, construímos boxplots e stripplots, comparando variáveis como sexo e cigarros fumados por dia, escolaridade versus IMC, faixa etária com colesterol e com o risco de CHD. Por fim, geramos um barplot, da quantidade de pessoas com risco de CHD por hábito de fumar.

Na seção 3.2, também foram geradas contagens, todas comparando o risco de CHD com variáveis como diabetes e hipertensão. Esses padrões visuais levantaram hipóteses que investigaremos posteriormente com testes estatísticos.

Após a análise exploratória, vamos realizar testes de normalidade para verificar se as variáveis quantitativas contínuas seguiam uma distribuição normal. Para isso, aplicamos o **teste de Shapiro-Wilk** e os **QQ-plots**, que produziram 6 gráficos e 6 p-valores para as variáveis escolhidas. Após os p-valores do teste de Shapiro-Wilk indicarem desvio da normalidade e os pontos nos QQ-plots visualmente não seguirem a linha de referência, ou seja, apresentarem uma assimetria, pudemos concluir que as distribuições das variáveis não puderam ser aproximadas para uma distribuição normal. Diante disso, optamos por utilizar testes **não paramétricos**.

Para comparar variáveis contínuas entre dois grupos — pessoas com e sem risco de CHD — aplicamos o **teste de Mann-Whitney U**. Nesse teste, comparam-se as medianas de duas amostras independentes. Ele é o equivalente não-paramétrico do teste t de student para amostras independentes. Se o p-valor for menor que o nível de significância pré-definido, nesse caso 0,05, a hipótese nula de que as amostras são semelhantes é rejeitada, indicando diferenças significativas nas distribuições.

Para esse teste, utilizamos as mesmas variáveis quantitativas contínuas do teste de normalidade. A partir dos resultados do p-valor, as variáveis colesterol total, pressão arterial, IMC, glicose e número de cigarros por dia apresentaram diferenças **estatisticamente significativas** entre os grupos. Por outro lado, a frequência cardíaca **não apresentou diferença relevante**, com p-valor aproximado de 0,24.

Para investigar variáveis com **mais de dois grupos**, utilizamos o **teste de Kruskal-Wallis**, que é uma generalização do teste anterior. Esse teste foi aplicado para comparar colesterol por níveis de escolaridade, pressão sistólica por faixa etária, glicose por número de fatores de risco, e colesterol total também por faixa etária.

Descobrimos indícios de que tanto a pressão sistólica quanto o colesterol aumentam significativamente com o envelhecimento, e que a glicose também cresce à medida que a pessoa acumula mais fatores como diabetes, hipertensão e tabagismo. Já no caso do colesterol por escolaridade, **não houve diferença significativa**, indicando que a variável educacional não apresentou influência direta nesse indicador na amostra analisada.

A seguir, aplicamos o **teste Qui-Quadrado de independência** para investigar associações entre variáveis **categóricas** e o risco de CHD. Vamos verificar se as frequências observadas em uma tabela de contingência são diferentes das frequências esperadas caso as variáveis fossem independentes. Se o p-valor for menor que o nível de significância 0,05, a hipótese nula de que as variáveis são independentes é rejeitada, indicando que existe uma relação entre elas.

Foram testadas 4 variáveis: sexo, tabagismo atual, presença de diabetes e hipertensão. Os resultados mostraram **associação estatisticamente significativa** do risco de CHD com sexo, diabetes e hipertensão. Já com o tabagismo atual, não houve associação estatística relevante, apresentando p-valor de 0,22, ou seja, maior que o alpha estabelecido, sugerindo que neste conjunto de dados o hábito de fumar atualmente não se destacou como fator associado ao risco cardiovascular. Isso pode indicar uma limitação da amostra ou refletir a influência de outros fatores associados.

Para consolidar os resultados, montamos uma **tabela resumo** com os testes realizados para cada variável e seus respectivos p-valores, se houve significância estatística e uma breve interpretação do resultado. Isso facilitou a visualização das variáveis mais relevantes para o risco cardíaco.

Por exemplo, a glicose apresentou diferença significativa tanto no teste de Mann-Whitney quanto em Kruskal-Wallis, o que indica uma relação clara entre glicemia e risco de doença cardíaca. A pressão sistólica teve comportamento semelhante, variando com a idade e o risco de CHD. Por outro lado, a frequência cardíaca não produziu diferença significativa para risco de CHD, e o teste do colesterol total apresentou níveis semelhantes para todos os graus de escolaridade.

3. Conclusão

Em conclusão, a análise inferencial reforçou os padrões observados na etapa exploratória, destacando variáveis clínicas como colesterol, pressão arterial e glicose como **fortemente associadas ao risco de doença cardíaca**.

A aplicação cuidadosa de testes não-paramétricos - em função da ausência de normalidade nos dados - garantiu a validade das inferências. Através deste estudo, demonstramos o papel fundamental da estatística na **compreensão de fatores de risco na saúde**, apoiando decisões baseadas em dados e evidências. Gostaríamos de agradecer aos ensinamentos da professora Cibele e à assistência prestada pela monitora Fernanda.