Trabalho de Ferramentas Computacionais de Modelagem: Câncer de colo de útero, dados de comportamento de risco

Micaeli M. Theodoro[*], Jessica Pasqueta[†]

15 de junho de 2023

O dataset *Cervical Cancer Risk Classification* é um conjunto de dados disponível no Kaggle que contém informações sobre fatores de risco relacionados ao câncer de colo de útero. Este resumo descreverá como realizar uma análise estatística descritiva deste conjunto de dados utilizando a linguagem de programação R.

1 Motivação

Segundo dados do Instituto Nacional do Câncer (INCA), são esperados mais de 704 mil casos novos de câncer no Brasil entre 2023-2025. Em mulheres, o câncer de mama ainda é o mais incidente (com exceção do câncer de pele não melanoma), porém nas regiões de menor IDH, o câncer de colo de útero ocupa a segunda posição. Por isso a importância de analisar os fatores de risco relacionados à esse tipo de câncer.

[*]Unesp, E-mail: micaeli.theodoro@unesp.br
[†]Unesp, E-mail: jessica.pasqueta@unesp.br

2 Exploração inicial

- Utilizaremos as funções `head()` e `str()` para visualizar as primeiras linhas do dataset e obter informações sobre as variáveis presentes.
- Verificar se há valores ausentes ou dados inconsistentes no conjunto de dados.

3 Análise estatística descritiva

- Utilizaremos as funções estatísticas básicas do R para obter medidas resumidas das variáveis, como média, mediana, desvio padrão, mínimo e máximo, utilizando as funções `mean()`, `median()`, `sd()`, `min()`, `max()`.
- Calcularemos a matriz de correlação utilizando a função `cor()` para analisar as relações entre as variáveis.

4 Visualização de dados

- Utilizaremos pacotes gráficos como o ggplot2 ou o base R para criar gráficos que auxiliem na compreensão dos dados. Por exemplo, histogramas, boxplots, gráficos de dispersão, entre outros.
- Criaremos gráficos que relacionem as variáveis entre si, como gráficos de dispersão ou gráficos de correlação.

5 Análise por grupos ou categorias

- Dividiremos o conjunto de dados com base em variáveis categóricas relevantes.
- Calcularemos medidas estatísticas descritivas separadamente para cada grupo, como média, mediana, desvio padrão, etc.
- Vamos comparar as distribuições de variáveis numéricas entre os grupos utilizando gráficos adequados.

6 Testes estatísticos

- Realizaremos testes estatísticos adequados para analisar diferenças significativas entre grupos ou correlações entre variáveis. Por exemplo, teste t de Student, teste qui-quadrado, teste de correlação de Pearson, entre outros.

- Utilizaremos as funções estatísticas apropriadas do R para realizar esses testes, como `t.test()`, `chisq.test()`, `cor.test()`, etc.