.

Gustavo Lopes, Homenique Vieira, Lucas Santiago, Rafael Amauri,
Thiago Henriques

Um estudo estatístico sobre ataques cardíacos e seu prognóstico

Belo Horizonte

2021

Resumo

"Estátistica Descritiva e Estatística Inferencial"são áreas de grande importância para a Estátistica, pois elas ajudão a descrever uma população através de um conjunto de dados amostrais. Este relatório técnico para a disciplina de Estatística e Probabilidade serve como um estudo de tais áreas, fazendo a descrição de dados coletados do site Kaggle sobre o assunto: ataques cardiovasculares e suas predições.

Palavras-chave: Ataques cardiovasculares, Estatística e Probabilidade, Kaggle, Estátistica Descritiva e Estatística Inferencial

Sumário

Introdução

As doenças cardiovasculares é um conjunto de doenças do coração e dos vasos sanguíneos, incluindo problemas estruturais e coágulos. De acordo com dados distribuídos pela Organização Mundial de Saúde(OMS), é estimado que no ano de 2016, 17.9 milhões de pessoas morreram por conta de doenças cardiovasculares, representando 31% de todas as mortes em nível global. Além disso, de acordo com a Sociedade Brasileira de Cardiologia, doenças cardiovasculares(DCV), tem sido a principal causa de mortalidade no Brasil desde a década de 1960.

Devido à pandemia ocasionada pelo COVID-19, admite-se que muitos desses casos vão ocorrer com mais frequência, principalmente em pessoas mais velhas devido ao estresse. Uma matéria da CNN Brasil de Janeiro deste ano, comenta dados de uma pesquisa feita no Brasil, afirma que: "o número de mortes por doenças cardiovasculares cresceu até 132% no Brasil durante a pandemia"(REZENDE, 2020).

Sendo assim, por este ser um assunto relevante no contexto atual, foi selecionado um banco de dados, fornecido pelo site Kaggle, uma subsidiária da Google LLC, com fôco em Cientistas de Dados e Machine Learning, afim de estudar estátisticamente as váriaveis presentes na amostra.

1 Recolhimento dos dados

A princípio, para tratar mais a fundo sobre o assunto, foi necessário escolher um banco de dados com dados suficientes para análise. Sendo assim, o site Kaggle foi uma escolha rápida e eficiente, oferecendo um espaço amostral de 303 pessoas.

2 Classificação das variáveis

Antes de começar a aprofundar no estudo, inicialmente deve se fazer uma análise dos elementos presentes no conjunto. Ao selecionar uma amostragem, são analisadas informações capazes de explicar e de mostrar as características da população em questão.

Essas características são denominadas de variáveis que podem ser classificadas de diferentes formas.

2.1 Qualitativas Nominais

Variáveis de características não numérica, que nomeia ou rótulo as características por meio de números ou símbolos.

Na amostragem em questão, as variáveis a seguir são classificadas dessa forma:

2.2 Qualitativas Ordinais

Variáveis de características não numérica, que mantém uma relação de ordem.

Na amostragem em questão, as variáveis a seguir são classificadas dessa forma:

2.3 Quantitativas Discretas

Variáveis que assumem valores inteiros e pontuais pertencentes a um conjunto enumerável.

Na amostragem em questão, as variáveis a seguir são classificadas dessa forma:

2.4 Quantitativas Contínuas

Variáveis que assumem valores qualquer valor real em um intervalo, associados a medição.

Na amostragem em questão, as variáveis a seguir são classificadas dessa forma:

3 Tabelas/Gráficos

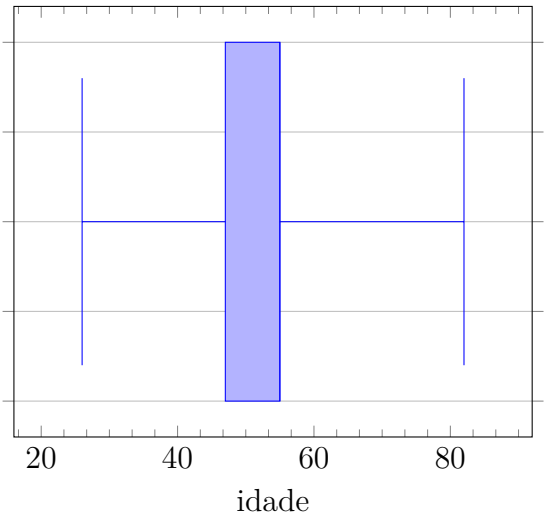

Figura 1 – Box Plot

O boxplot apresentou resultados interessantes, todas as variáveis presentes na amostra de idades não ultrapassam dos outliers estabelecidos. Além disso se apresenta os valores 47, 55 e 61 para os quartis respectivamente.

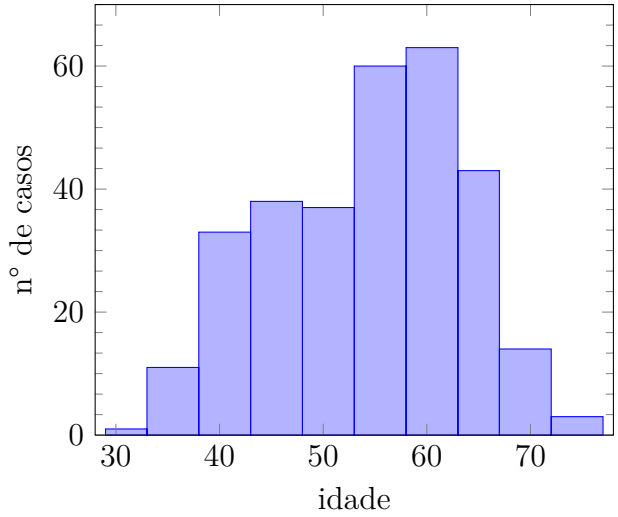

Figura 2 – Histograma

Com o histograma, podemos observar uma ocorrência maior de casos em pessoas com faixa etária entre 50 à 65 anos. A presença de casos para pessoas abaixo de 30 anos e acima dos 80 anos é quase nula.

4 Medidas de tendência central e de variabilidade

Tabela 1 – Medidas de tendência central e de variabilidade

	Idade	Pressão Sanguínea em reposou	Colesterol em mg/dl	Pico de frequência cardíaca
Média	54.36	131.62	246.26	149.64
Mediana	55	130	240	153
Moda	58	120	204,234,197	162
Desvio Padrão	9.08	17.53	51.83	22.90

Fonte: Produzido pelos próprios autores

5 Intervalo de confiança para proporção de interesse

De uma base amostral contendo 303 pessoas, foram selecionadas todas com idade igual ou superior a 53 anos, totalizando 183 pessoas. Desta forma, foi realizado o intervalo de confiança por proporção.

$$
\begin{aligned}
&IC(1-\alpha)\% = \bar{x} \pm \tau_{\frac{\alpha}{2}}; *n-1\frac{S}{\sqrt{n}} \\
&IC(95)\% = 54.37 \pm 2.2622 * \frac{6.36}{\sqrt{10}} \\
&IC(95)\% = 54.37 \pm 4.55 \\
&IC(95)\% = [446.4; 455.50]
\end{aligned} \tag{1}
$$

$$
\begin{aligned}
&IC(1-\alpha)\% = \bar{x} \pm Z * \frac{\alpha}{2} * \frac{\sigma}{\sqrt{N}} \\
&IC(95)\% = 54.37 \pm 1.960 * \frac{9.08}{\sqrt{303}} \\
&IC(95)\% = 54.37 \pm 1.02 \\
&IC(95)\% = [53.35; 55.39]
\end{aligned} \tag{2}
$$

$$
\begin{aligned}
&IC(1-\alpha)\% = \hat{P} \pm Z_{\frac{\alpha}{2}} * \sqrt{\frac{\hat{P} * (1-\hat{P})}{n}} \\
&IC(95)\% = 0.604 \pm 1.96 * \sqrt{\frac{0.604 * (0.396)}{303}} \\
&IC(95)\% = 0.604 \pm 0.055 \\
&IC(95)\% = [0.549; 0.659]
\end{aligned} \tag{3}
$$

Conclusão

Após trabalhar nessa linguagem por quase um mês, ficou claro para toda a equipe que Haskell é uma linguagem diferenciada. Possui uma história que foi essencial para sua formação, sem esquecer, é claro, que foi uma lingua desenvolvida de forma comunitária. Há grandes obstáculos em ingressar nessa lingua, principalmente por documentação excassa e a documentação oficial em maioria ser paga ou extremamente complexa para iniciantes.

Por muito tempo, Haskell não foi uma lingua unificada, cada pessoa que entrava no projeto criava uma versão diferente sem que um líder principal coordenasse como ela estava evoluindo. Falta de uma unidade atrasou um pouco a lingua ser adotada pela comunidade. Por conta da falta de uma documentação única, a dificuldade de aprendizado foi outro grande ponto negativo que impactou diretamente na falta de profissionais que a utilização, ficando apenas fechada em um ambiente científico como universidades.

Entretanto depois de tudo que vimos, a linguagem se apresenta de forma bem mais positiva do que todas essas ideias citadas acima. Ela apresenta vários recursivos interessantes, como cálculos lambda, recursões simples e amarrações de funções em variáveis de forma simples. Há vários tutoriais distribuidos pela internet. Mesmo que poucas pessoas programem nessa lingua, possui uma comunidade bem forte que a mantém.

Por fim, o grupo entendeu que Haskell desempenhou seu papel na história da computação. Além disso, várias empresas ainda o adotam por entenderem a importância de seu uso, feito para cálculos científicos e precisos. Ainda dentro do contexto de programação funcional e uso de cálculos lambda, Haskell ainda é uma, se não a melhor, lingua para ser utilizado.

Referências

DESCONHECIDO. *Doenças Cardiovasculares*. 2020. Disponível em: <https://www.paho.org/pt/topicos/doencas-cardiovasculares>. Acesso em: 28 de maio de 2021. Nenhuma citação no texto.

RAHMAN, R. *Heart Attack Analysis & Prediction Dataset - A dataset for heart attack classification*. Disponível em: <https://www.kaggle.com/rashikrahmanpritom/heart-attack-analysis-prediction-dataset/metadata>. Acesso em: 28 de maio de 2021. Nenhuma citação no texto.

REZENDE, D. *Estudo apresenta dados e impactos das doenças cardiovasculares no Brasil*. 2020. Disponível em: <https://pressreleases.scielo.org/blog/2020/11/06/estudo-apresenta-dados-e-impactos-das-doencas-cardiovasculares-no-brasil/>. Acesso em: 28 de maio de 2021. Citado na página 3.

Apêndices

APÊNDICE A – Banco de Dados

age	sex	cp	trtbps	chol	fbs	restecg	thalachh	exng	oldpeak
63	1	3	145	233	1	0	150	0	2.3
37	1	2	130	250	0	1	187	0	3.5
41	0	1	130	204	0	0	172	0	1.4
56	1	1	120	236	0	1	178	0	0.8
57	0	0	120	354	0	1	163	1	0.6
57	1	0	140	192	0	1	148	0	0.4
56	0	1	140	294	0	0	153	0	1.3
44	1	1	120	263	0	1	173	0	0
52	1	2	172	199	1	1	162	0	0.5
57	1	2	150	168	0	1	174	0	1.6
54	1	0	140	239	0	1	160	0	1.2
48	0	2	130	275	0	1	139	0	0.2
49	1	1	130	266	0	1	171	0	0.6
64	1	3	110	211	0	0	144	1	1.8
58	0	3	150	283	1	0	162	0	1
50	0	2	120	219	0	1	158	0	1.6
58	0	2	120	340	0	1	172	0	0
66	0	3	150	226	0	1	114	0	2.6
43	1	0	150	247	0	1	171	0	1.5
69	0	3	140	239	0	1	151	0	1.8
59	1	0	135	234	0	1	161	0	0.5
44	1	2	130	233	0	1	179	1	0.4
42	1	0	140	226	0	1	178	0	0
61	1	2	150	243	1	1	137	1	1
40	1	3	140	199	0	1	178	1	1.4
71	0	1	160	302	0	1	162	0	0.4
59	1	2	150	212	1	1	157	0	1.6
51	1	2	110	175	0	1	123	0	0.6
65	0	2	140	417	1	0	157	0	0.8
53	1	2	130	197	1	0	152	0	1.2
41	0	1	105	198	0	1	168	0	0
65	1	0	120	177	0	1	140	0	0.4
44	1	1	130	219	0	0	188	0	0
54	1	2	125	273	0	0	152	0	0.5
51	1	3	125	213	0	0	125	1	1.4
46	0	2	142	177	0	0	160	1	1.4
54	0	2	135	304	1	1	170	0	0
54	1	2	150	232	0	0	165	0	1.6
65	0	2	155	269	0	1	148	0	0.8
65	0	2	160	360	0	0	151	0	0.8
51	0	2	140	308	0	0	142	0	1.5
48	1	1	130	245	0	0	180	0	0.2
45	1	0	104	208	0	0	148	1	3
53	0	0	130	264	0	0	143	0	0.4
39	1	2	140	321	0	0	182	0	0
52	1	1	120	325	0	1	172	0	0.2
44	1	2	140	235	0	0	180	0	0