Exploração da Distribuição dos Dados

Gabriel Luiz dos Santos Silva
Ciência de Dados | Teoria do Aprendizado Estatístico

August 27, 2023

Atividade 01 | Aula 05

A atividade a seguir deverá ser realizada a partir do que foi visto nas últimas aulas. O respectivo arquivo com os dados estão anexados. Realizem a atividade para podermos discutir posteriormente em sala de aula.

Os dados encontrados no arquivo esforco são provenientes de um estudo sobre teste de esforço cardiopulmonar em pacientes com insuficiência cardíaca, as variáveis medidas durante a realização do teste foram observadas em quatro momentos distintos: repouso (`REP`), limiar anaeróbio (`LAN`), ponto de compensação respiratório (`PCR`) e pico (`PICO`). As demais variáveis são referentes às características demográficas e clínicas dos pacientes sendo registradas uma única vez.

1. Descreva a distribuição da variável consumo de oxigênio (`VO2`) em cada um dos quatro momentos de avaliação utilizando medidas de resumo, boxplots e histogramas. Você identifica-rá algum paciente com valores de consumo de oxigênio discrepantes os resultados.

2. Descreva a distribuição da classe funcional `NYHA` por meio de uma tabela de frequências. Utilize um método gráfico para representar essa tabela.

Dicionário

Os dados são provenientes de um estudo sobre teste de esforço cardiopulmonar em pacientes com insuficiência cardíaca realizado no InCor da Faculdade de Medicina da USP pela Dra. Ana Maria Fonseca Wanderley Braga. Um dos objetivos do estudo é comparar os grupos formados pelas diferentes etiologias quanto às respostas respiratórias e metabólicas obtidas do teste de esforço cardiopulmonar. Outro objetivo é saber se alguma das características observadas pode ser utilizada como fator prognóstico de óbito.

1 Métodos

Foi utilizada a linguagem de programação R, com o ambiente de desenvolvimento online, `R Studio Cloud (Posit Cloud)`. Foram utilizadas as seguintes bibliotecas: `dyplor` para o tratamento de dados, `ggplot2` para a visualização das variáveis, `psych` para o resumo do dataset e `knitr` para criar documentos interativos e relatórios, no caso a tabela de frequência.

Os atributos limiar anaeróbio e o ponto de compensação respiratório, apresentaram conter valores ausentes, especificamente de 3 e 6 pacientes não tiveram seus registros coletados. A incompletude desses dados possivelmente não afetará os resultados, por essa razão não serão removidos esses pacientes.

Definição de variáveis

- **Paciente:** Identificação do paciente
- **Sexo:** F: feminino, M: masculino
- **Idade:** idade do paciente em anos
- **Peso:** peso do paciente em kg
- **VO2:** consumo de oxigênio em ml/(kg.min)

Momentos de avaliação das variáveis

- **REP:** repouso
- **LAN:** limiar anaeróbio
- **PCR:** ponto de compensação respiratório
- **PIC:** pico de exercício

2 Análise Estatística

A tabela a baixo apresenta as seguintes medidas de resumo, a variável **Trimmed** é a média das variáveis, porém sem a presença de outliers. A variável **Simetria** mede a assimetria da variável, um valor positivo indica assimetria para a direita, enquanto um valor negativo para a esquerda. A variável **se**, o erro padrão da média, sendo uma medida da variabilidade da média amostral, ou seja, quanto menor o erro padrão, mais representativa é a média amostral da média populacional.

Medida	Repouso	Lin. Anaeróbio	Pont Comp. Resp	Pico Exercício
Média	3.624	10.55	14.84	18.06
Trimmed	3.500	10.23	14.27	17.61
Mediana	3.400	10.05	13.90	17.10
Variância	1.292	9.059	24.07	41.92
Desv.Pad	1.14	3.01	4.91	6.47
Simetria	3.88	1.26	1.06	0.75
se	0.10	0.27	0.45	0.57
min	1.7	5.1	6.9	5.2
amplitude	10.7	16.4	24.1	35.8
max	12.4	21.5	31.0	41.0
n	127	124	121	127

Table 1: Medidas em diferentes condições

Ao analisar a tabela, observa-se que todas as variáveis possuem um significante nível de assimetria positiva, exceto para o nível de O_2 no pico de exercício, no que sugere que nas demais variáveis podem possuir pacientes com resultados anormais ou discrepantes.

Verificando o desvio padrão de cada variável, percebemos uma alta dispersão delas, exceto o consumo de O_2 em repouso, onde é visto que possui uma alto assimetria positiva.

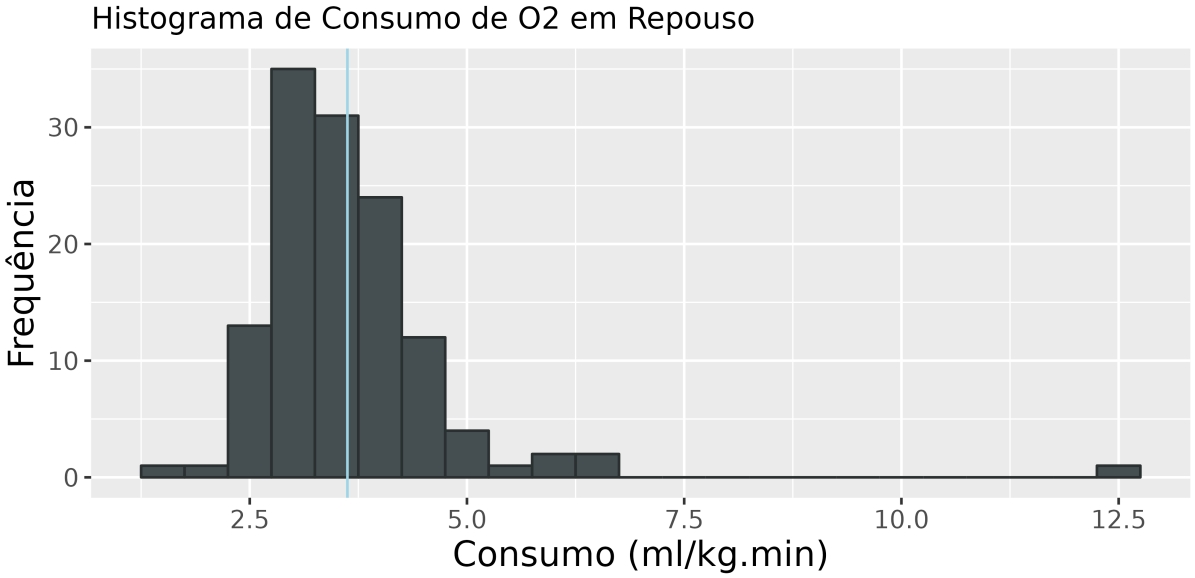
2.1 Visualização dos Dados com Histograma

Será usado o histograma para a visualização dos dados contínuos e sua frequencia. Assim podemos verificar a distribuição entorno da média, a assimetria e valores extremos. O boxplot será utilizado para identificar quaisquer paciente com valores de O_2 discrepantes.

O histograma a seguir mostra a frequência absoluta do consumo de oxigênio dentro de diferentes faixas.

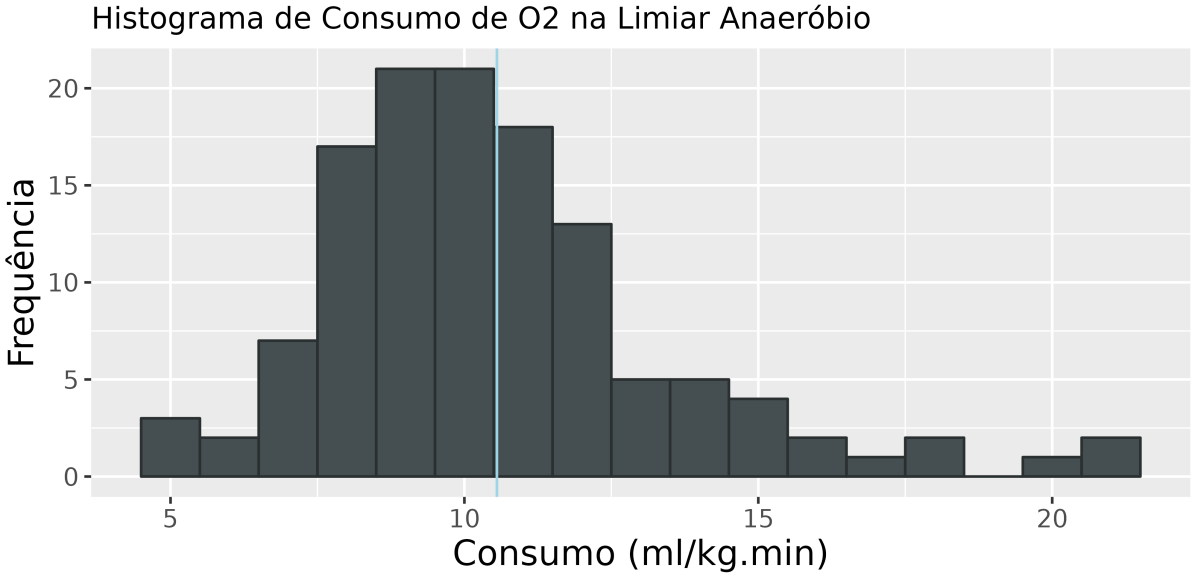
Obverva-se na figura 1 que o consumo O_2 em repouso poderiam seguir uma distribuição quase normal, se não fosse por conta de um outlier. Isso resalta a importância de investigar esse paciente que possue um consumo muito muito alto de O_2 em repouso.

Figure 1: Histograma de consumo de oxiênio na limiar anaeróbio



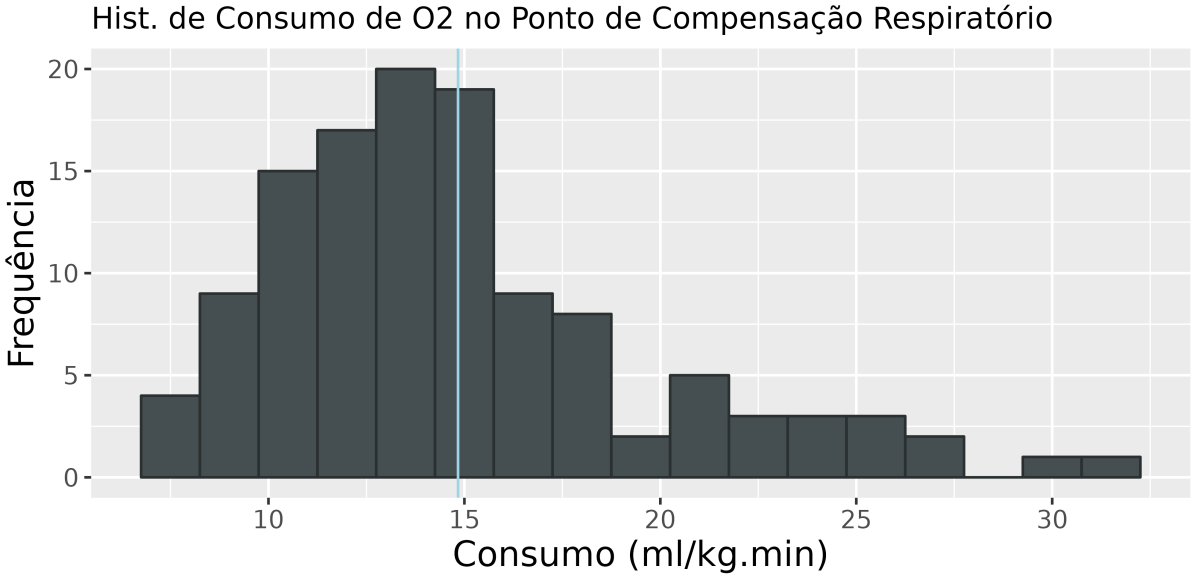
Na figura 2, percebemos que o consumo de O_2 na limiar anaeróbio possui uma frequencia mais rebusta antecedente a média. Existem valores bem distantes da média, onde, a medida que o consumo vai aumentando gradativamente, frequência absoluta diminui progressivamente.

Figure 2: Histograma de consumo de oxiênio em repouso



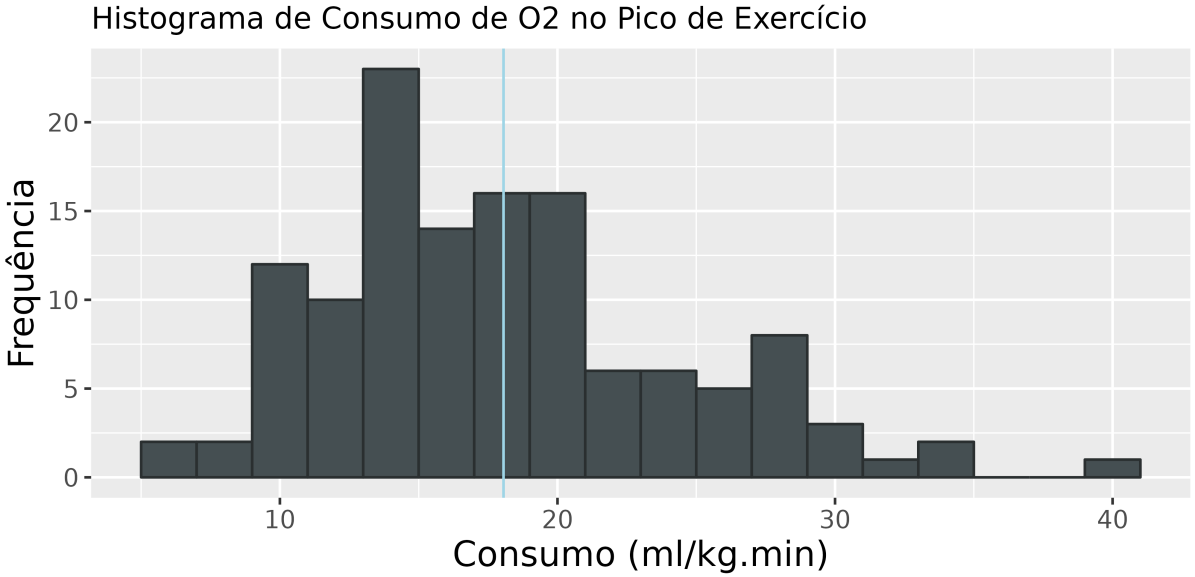
Para a figura 3, referente ao consumo de O_2 no ponto de compensação respiratório, temos um comportamento semelhante a figura 2, com uma característica visual, semelhante a uma escada. Vemos que a maioria dos pacientes consomem até 15 ml/kg.min de O_2, onde, partindo disso, percebemos uma queda brusca de quase 50% dos pacientes. Percebemos pouquíssimos pacientes bem distantes da média, abaixo de 5 pacientes.

Figure 3: Histograma de consumo de oxiênio no ponto de compensação resiratório



Para terminar, na figura 4 vemos que uma distribuição quase normal em torno da média, do consumo de O_2 no pico de exercício. Vemos uma frequência (aparentemente 1 paciente) possui um nivel muito elevado no consumo de O_2. Esse mesmo comportamente se repete nas figuras vistas anteriores, onde podemos gerar uma hipótese que, possivelmente, seja o mesmo paciente, onde ele deve ser examiado e investigado as causas dessa anomalia pelos profissionais.

Figure 4: Histograma de consumo de oxiênio no pico de exercício



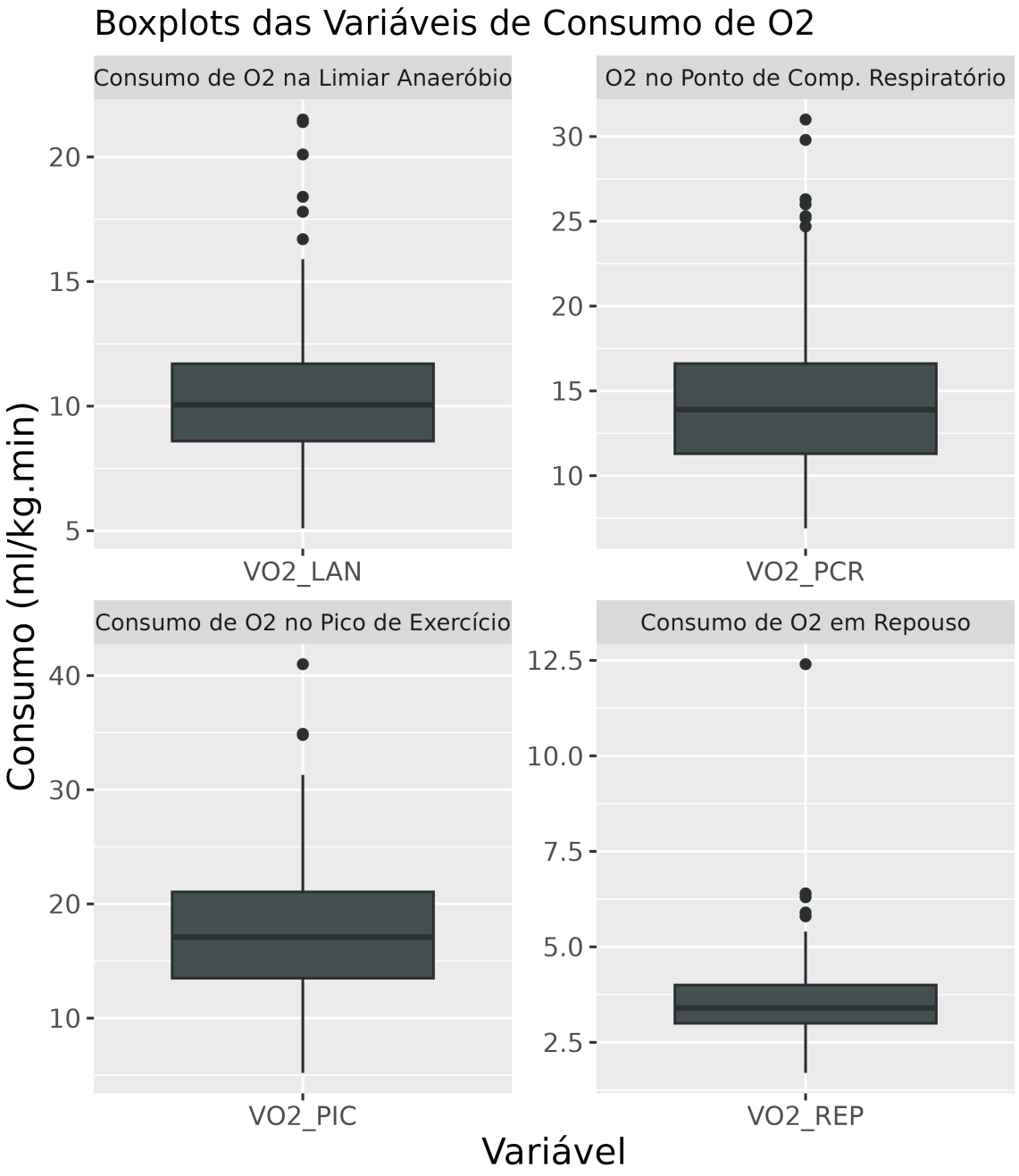
Com isso encerramos a análise pela visualização das variáveis em histograma na qual encontramos pouquissimos pacientes (entre 1 a 5) que possuem um consumo de O_2 em um nivel muito discrepante em relação a média.

2.2 Visualização dos Dados em Boxplot

Será usado o bloxplot para a visualização e análise dos pacientes que apresentaram resultados discrepantes.Veremos o valor mínimo e máximo de cada variável, assim como a mediana e o primeiro e terceiro quartil, que serão exibidos no boxplot ....

Intepretação: A caixa representa o intervalo interquartil. A linha que divide a caixa ao meio é a mediana. As hastes se estendem a partir da caixa até os mínimos e máximos num intervalo.

Figure 5: Boxplot das quatro variáveis de consumo de O_2



Pela análise do boxplot, percebemos que somente as variaveis de consumo de O_2 em repouso e no pico de exercício, são as que possuem poucos pacientes com valores discrepantes, também visto no histograma, enquanto as demais possuem uma quantidade significativamente alta de outliers, no que sugere uma análise sobre esses pacientes.

Ao analisarmos o tamanho da caixa, vemos que somente os pacientes com o consumo de O_2 em repouso possuem uma baixa disperção, enquanto as demais possuem uma caixa maior, indicando uma alta disperção dos valores registrados, indicando uma variação significativa entre eles.

A mediana das variáveis de consumo de O_2, exceto o consumo em repouso, estão próximas do centro das suas caixas, sugerindo que a maioria dos pacientes apresenta valores próximos a essas medianas. Vimos algo semelhante ao histograma, e podemos confirmar isso com o boxplot

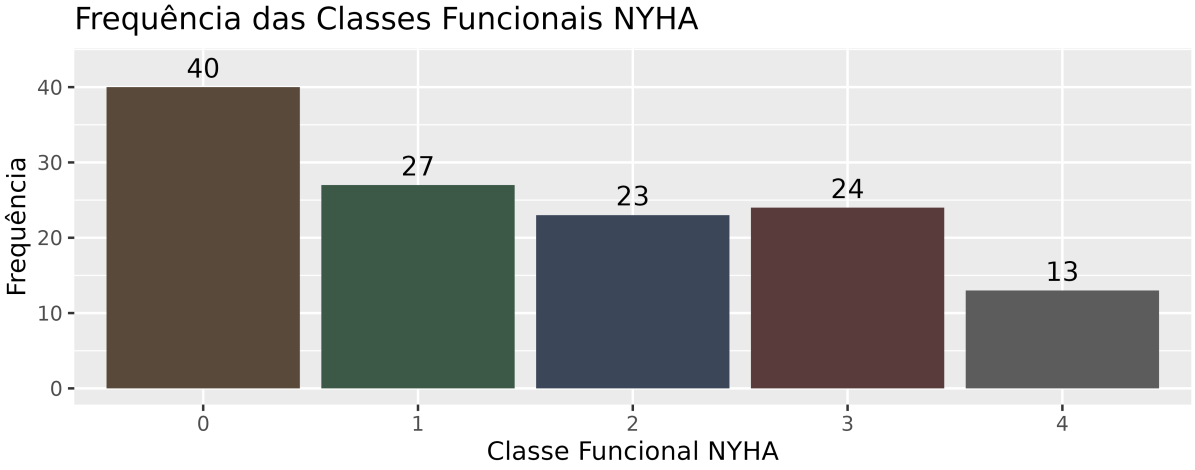
2.3 Distribuição da Classe Funcional NYHA

NYHA	Freq Absoluta	Freq Relativa	Freq Abs Acum	Freq Rel Acum
0	40	31.50	40	31.50
1	27	21.26	67	52.76
2	23	18.11	90	70.87
3	24	18.90	114	89.76
4	13	10.24	**127**	**100.0**

Table 2: Tabela de Distribuição da Classe Funcional NYHA

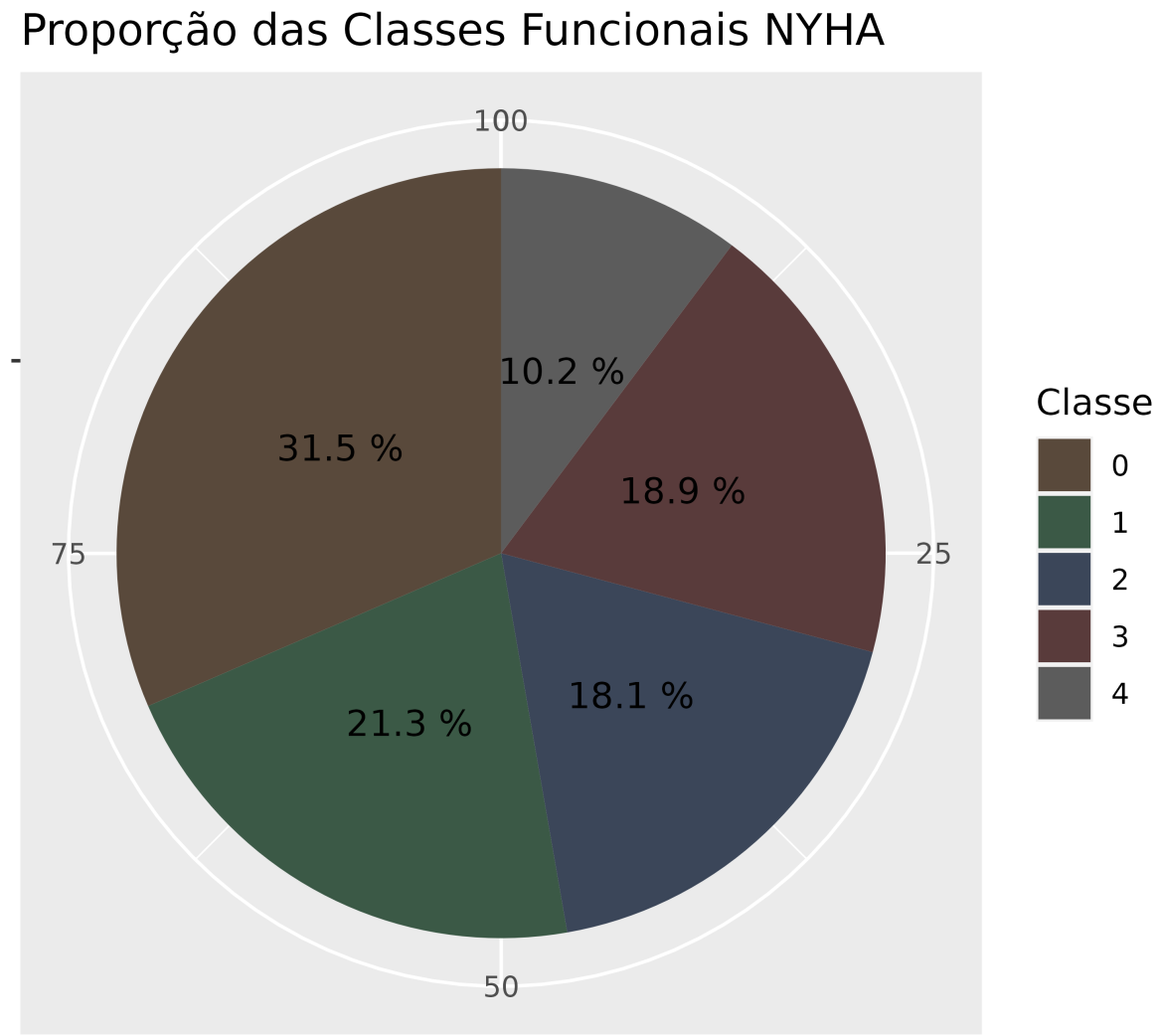
Ao analisarmos a tabela 2 de frequência, percebe-se que a classe 0, é a com a maior frequência em relação as outras, com mais de 30 % dos dados, porém, as classes 1, 2 e 3 não estão tão distantes da classe 0, com uma proporção entre 18 % a 21 % dos dados. Com base nessa tabela, vemos uma presença considerável de pacientes de diferentes classes.

Figure 6: Gráfico de barra das classes funcionais NYHA



Na figura 6, podemos perceber mais facilmente a proporção de cada classe. Isso permite uma compreensão rápida e visual das proporções das diferentes classes.

Figure 7: Gráfico de setores das classes funcionais NYHA



Na figura 7, podemos ver a representação da proporção de cada classe, em fatia. Temos uma percepção mais clara qual classe é mais frequente. Reparamos tamb´m que a classe 1, 2 e 3, aparetemente possuem a mesma proporção, enquanto a classe 4 é a menor dentre as demais.