EPI DEMIO LOGIA I

CONSTRUÇÃO DE GRÁFICOS TIPO *BOXPLOT* POR MEIO DO RSTUDIO

MONITOR: IGOR LOPES VELASCO
ORIENTADORES: VALÉRIA TRONCOSO BALTAR
FELIPE GUIMARÃES TAVARES
THAIZA DUTRA G. DE CARVALHO

Universidade Federal Fluminense

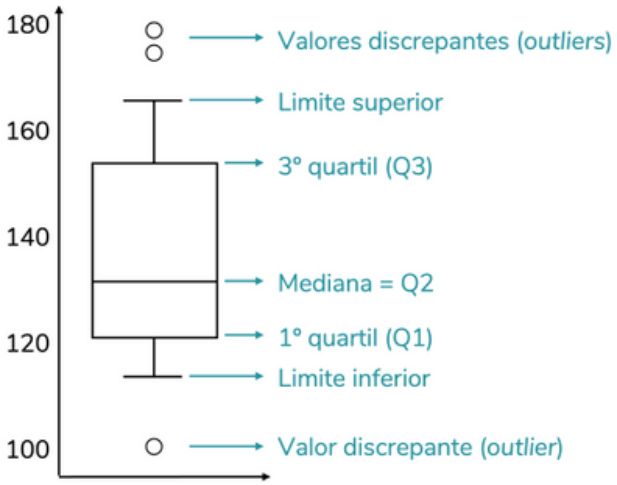
INSTITUTO DE SAÚDE COLETIVA
DEPARTAMENTO DE EPIDEMIOLOGIA E BIOESTATÍSTICA

2025

SUMÁRIO

O BOXPLOT

A Bioestatística é uma ciência que, em certas situações, lida com um grande volume de dados. Por isso, é imprescindível que se conheça técnicas de representações gráficas de dados, uma vez que são ferramentas que possibilitam uma análise eficiente. O *boxplot* é um instrumento muito importante porque torna possível a percepção de características da distribuição de uma série de dados.



Ao analisar um *boxplot*, é possível observar a **mediana** (que divide sua série de dados ao meio), os **quartis** (que mostram a distribuição dos valores 25% inferiores e superiores), os **valores mínimos** e **máximos** (que são os "bigodes" do gráfico) e os ***outliers***, que são os valores discrepantes da série de dados. Portanto, percebe-se a distribuição, a variabilidade, possíveis assimetrias e *outliers* por meio do *boxplot*.

O *boxplot* é como um mapa resumido da sua série de dados. Se a mediana estiver localizada no meio da caixa, infere-se que os dados são simétricos. Mas, se estiver mais perto do primeiro ou terceiro quartis, os dados são assimétricos. Se um "bigode" não estiver igual ao outro, é possível que os dados estejam inclinados para um dos lados. A presença de *outliers* indicam uma grande variação ou valores que não são coerentes com o conjunto de dados. Todas essas informações são obtidas rapidamente ao observar um *boxplot*, dispensando a consulta de tabelas de dados, que podem ser extensas.

A LINGUAGEM R

O **R** é uma linguagem de programação direcionada para a estatística, análise e ciência de dados. Essa linguagem é muito popular em todo o mundo por ser gratuita e de código aberto, o que possibilita o desenvolvimento constante de novas funcionalidades e aplicações para o **R**. Com o uso da **linguagem R**, é possível desde a manipulação de pequenas quantidades de dados e contas simples até execução de projetos de *machine learning* e inteligência artificial.

As aplicações da linguagem R na Bioestatística são várias. Cálculos estatísticos complexos podem ser executados no R, como estatísticas descritivas (média, mediana, variância, desvio padrão), testes estatísticos, como o teste t, além de modelos probabilísticos. Não apenas o *boxplot* pode ser construído por meio do R, mas também outros modelos de visualização de dados, como histogramas e gráficos interativos. Até mesmo na Genômica e análise de sequências de DNA o R pode ser utilizado.

Portanto, aprender os fundamentos da linguagem R pode ser um diferencial, mesmo para estudantes da área da Saúde, dado que é uma das melhores ferramentas de análise de dados e é amplamente utilizada na pesquisa, inclusive em estudos clínicos.

O RSTUDIO

O Rstudio é um ambiente de desenvolvimento integrado (IDE) para a linguagem R, o que torna o trabalho no R mais organizado, intuitivo e eficiente, facilitando a escrita e execução de códigos em R.

O Rstudio possibilita escrever o código, executar comandos e visualizar resultados ao mesmo tempo, em uma interface de painéis divididos. Além disso, erros no código são identificados pelo *software*, o que resulta em um *debugging* (correção desses erros) simplificado.

A instalação e carregamento de pacotes no Rstudio não é uma tarefa complexa e permite uma ampliação de funcionalidades no R. O Rstudio suporta também a criação de documentos em PDF, Word e HTML contendo o código R, texto e resultados. Esse recurso é chamado de R Markdown. E para quem é familiarizado com programação, o Rstudio permite a integração com o Python.

Primeiros passos no Rstudio

Para a utilização do Rstudio, primeiramente é necessário baixar o R em seu computador. O R pode ser baixado neste **link**.
Em seguida, baixe e instale o software Rstudio. O Rstudio pode ser baixado neste **link**.
No site da disciplina, é possível encontrar vídeos tutoriais com o passo a passo da instalação do R e Rstudio. **Clique aqui** caso queira ser direcionado ao site **Epidemiologia UFF**.

CONSTRUINDO O *BOXPLOT* NO RSTUDIO

Nesta sessão, serão abordadas duas maneiras de se construir gráficos do tipo *boxplot* utilizando a linguagem R por meio do Rstudio. Primeiro, será demonstrado o uso da função nativa boxplot() e, em seguida, o uso do pacote ggplot2. É importante ressaltar que é pré-requisito ter o R e o *software* Rstudio instalados para a execução das próximas etapas deste tutorial. Toda a demonstração desta sessão está contextualizada no caso a seguir.

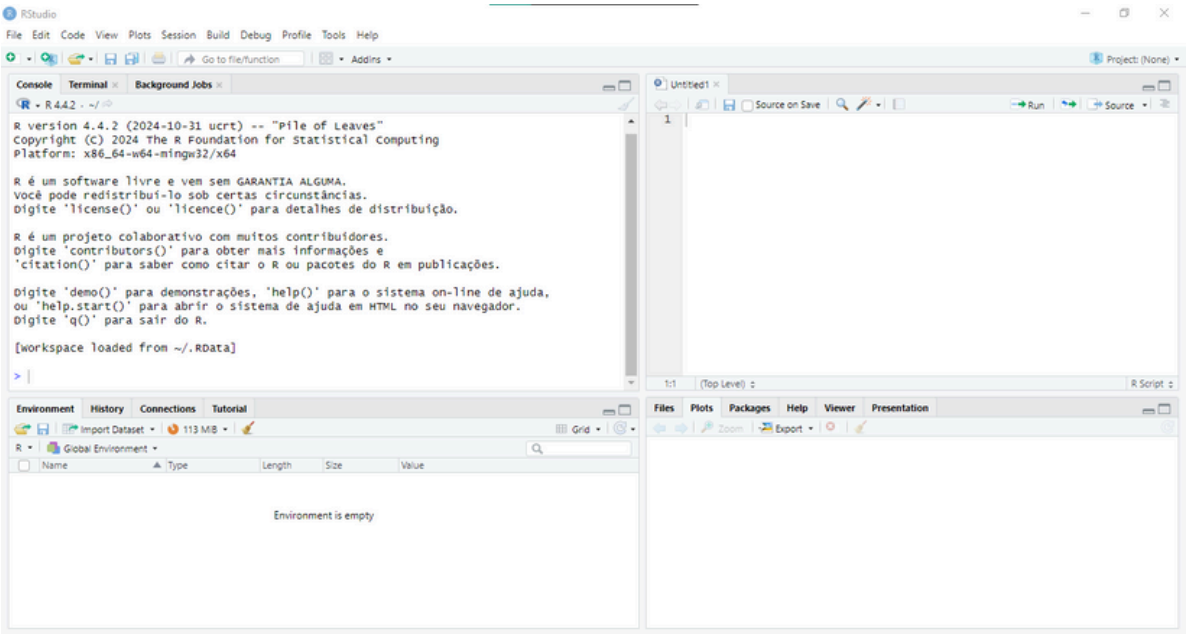
Caso: Pesquisa clínica de um anti-hipertensivo

Em um hospital universitário, um grupo de pesquisadores conduz um estudo que investiga a eficácia de dois novos fármacos anti-hipertensivos que supostamente são mais eficazes e possuem menos efeitos colaterais que fármacos utilizados na terapia convencional. Os medicamentos são denominados **A** e **X** e cada um foi testado em um grupo com 10 indivíduos por um mês. Após esse período, a pressão arterial dos participantes foi aferida. O medicamento que reduzir a pressão arterial e promover uma menor dispersão da pressão arterial sistólica (PAS), será submetido a uma próxima etapa de testes clínicos.

Observe abaixo a tabela que o pesquisador construiu com os dados da pressão arterial sistólica aferidos nos participantes da pesquisa. A primeira linha corresponde ao grupo tratado com o medicamento A e a segunda linha ao grupo tratado com o medicamento X.

n	1	2	3	4	5	6	7	8	9	10
A	113	149	127	109	219	125	136	129	131	110
X	128	123	161	142	136	139	119	110	106	146

Ao executar o software Rstudio, você poderá observar uma janela semelhante a ilustrada na imagem abaixo. Clicando em "*File*", "*New file*" e "*R script*", um painel para editar o script será aberto no lado superior direito da tela. Neste painel você poderá começar a escrever o seu código. É importante salientar que a posição dos painéis na janela é customizável.



1 - Construindo o *boxplot* utilizando a função nativa boxplot()

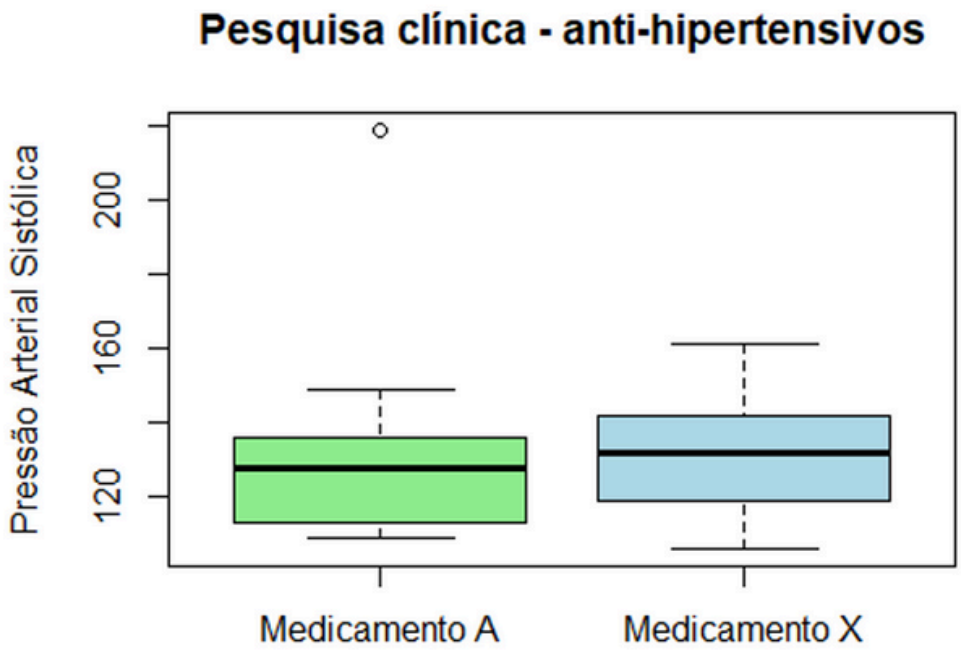
No painel editor de script, comece a escrever o seu código declarando dois vetores com os dados da pesquisa, cada um referente a um grupo que foi tratado com um medicamento, ou seja, um vetor com os dados da pressão arterial sistólica do grupo tratado com o medicamento A, que será denominado "**medA**" e outro vetor com os valores da PAS do grupo tratado com o medicamento X, denominado "**medX**". Observe abaixo sintaxe, isto é, a estrutura do código que deve ser escrito no Rstudio para se fazer o que foi explicado.

```
# Criando conjuntos de dados - duas séries de PAS

medA<-c(113, 149, 127, 109, 219, 125, 136, 129, 131, 110)
medX<-c(128, 123, 161, 142, 136, 139, 119, 110, 106, 146)
```

Em seguida, a função nativa boxplot() é utilizada para que se obtenha os gráficos referentes às séries de dados do caso abordado. A estrutura do código contendo a função em questão é demonstrada abaixo.

```
# Criando um boxplot para cada série de dados

boxplot(medA,medX,names =c("Medicamento A","Medicamento X"),
main="Pesquisa clínica - anti-hipertensivos", ylab="Pressão Arterial
Sistólica", col=c("lightgreen","lightblue") )
```

Como é possível perceber acima, os argumentos da função *boxplot*, ou seja, o que está entre parênteses após a função *boxplot* ser declarada, são: cada vetor contendo a série de dados (medA, medX), names (indica a denominação de cada *boxplot* no eixo x), main (corresponde ao título do gráfico), ylab (indica o nome do eixo y do gráfico) e col, que indica a cor de cada gráfico, respectivamente. Observe, a seguir, o gráfico elaborado pelo Rstudio a partir do código escrito acima.



A partir da análise do *boxplot* de cada conjunto de dados, é possível inferir que os dados da pressão arterial sistólica do grupo tratado com o medicamento A estão menos dispersos que os do grupo tratado com o medicamento X. No entanto, há um *outlier* no conjunto de dados referente ao medicamento A.

2 - Construindo o *boxplot* utilizando o pacote ggplot2

O ggplot2 é um pacote para visualização de dados na linguagem R. Esse pacote é baseado no conceito da "gramática dos gráficos", que implementa a construção de gráficos por camadas, tornando essa tarefa mais intuitiva e personalizável. O uso do ggplot2 é ideal na análise de dados pois uma visualização otimizada possibilita uma melhor compreensão dos dados. Por isso, ele é amplamente utilizado por cientistas de dados, estatísticos e pesquisadores.

O uso desse pacote somente é possível a partir da sua instalação e carregamento. Acrescente as linhas de comando a seguir no seu script ou execute-as no console do seu Rstudio, caso nunca tenha usado o ggplot2.

```
# Instalando o pacote ggplot2

install.packages("ggplot2")

# Carregando o pacote ggplot2

library(ggplot2)
```

Caso tenha iniciado um novo script, não esqueça de declarar dois vetores com os dados da pesquisa. Esses vetores, medA e medX contém os dados da PAS dos 10 indivíduos de cada grupo. Essa etapa pode ser revisitada na página 6 deste tutorial.

Antes de utilizarmos o pacote ggplot2, devemos preparar nossos dados para que o pacote entenda as informações que estamos fornecendo. Por isso, aplica-se a função as.data.frame() para converter os vetores medA e medX declarados em tabelas de dados (data frame), resultando, assim, em duas tabelas, uma para cada grupo. Além disso, nas tabelas de dados, será criada uma nova coluna denominada medicamento, na qual cada célula desta coluna especifica se a PAS da coluna anterior é referente a um indivíduo do grupo do medicamento A ou do medicamento X. Por fim, as tabelas de cada grupo serão mescladas para que o resultado final seja uma tabela 20 por 2, na qual a primeira coluna é referente a PAS, a segunda referente ao medicamento utilizado e as linhas referentes a cada participante do estudo. Veja na página 9 a parte do script responsável por executar o que foi descrito neste parágrafo além da tabela resultante que pode ser consultada no Rstudio.

Ao utilizar o pacote ggplot2, é necessário colocar como argumentos a tabela de dados construída no passo anterior e denominada, no nosso código, "dados", a definição dos eixos x e y por meio do argumento aes() e o argumento geom_boxplot(), que adiciona o *boxplot* ao gráfico. Dentro desse último argumento citado, o geom_boxplot(), podemos especificar a cor, a forma e o tamanho dos *outliers* e a cor das caixas do gráfico. Além disso, foram adicionados, por meio de stat_summary(), pontos ao gráfico que representam a média aritmética de cada conjunto de dados, sendo especificados a forma, o tamanho e as cores de borda e preenchimento dos pontos.

Observe na página 10 o script que contém o pacote ggplot2 e que foi utilizado para a construção do gráfico *boxplot*. Veja também o gráfico resultante desse script.

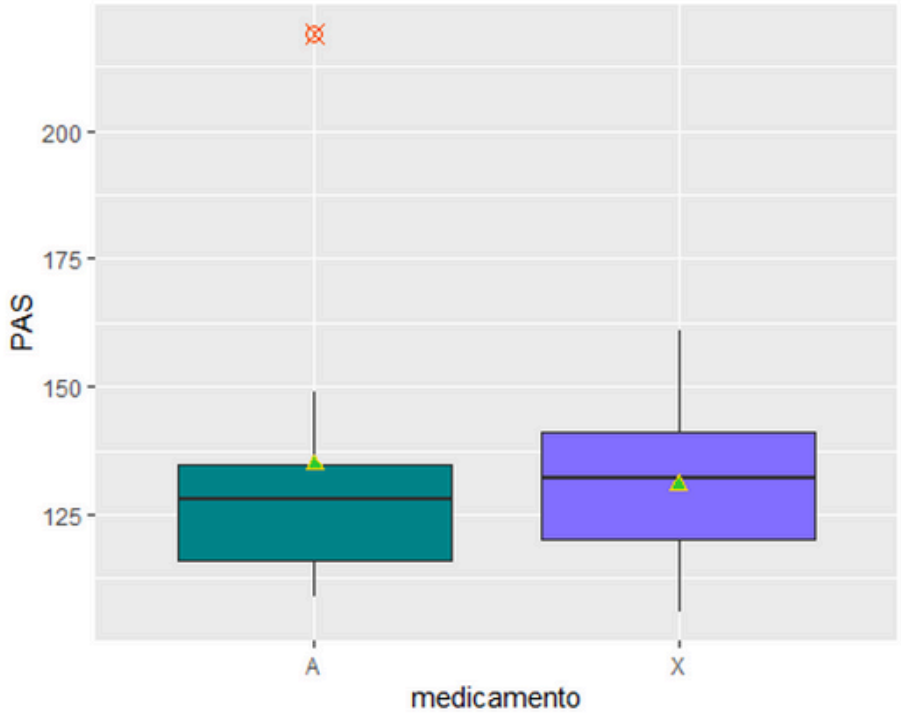
- Script para transformação dos vetores de dados declarados em tabela e tabela única, denominada "dados", resultante do código programado

```
# Convertendo vetor medA em tabela de dados

medA<-as.data.frame(cbind(medA))

# Criando a coluna "medicamento" na tabela medA e atribuindo "A" às
células
medA$medicamento<-"A"

# Convertendo vetor medX em tabela de dados

medX<-as.data.frame(cbind(medX))

# Criando a coluna "medicamento" na tabela medX e atribuindo "X" às
células

medX$medicamento<-"X"

library(dplyr )

medA<-medA %>%
  rename(PAS = "medA") # Renomeando a coluna medA para PAS
medX<-medX %>%
  rename(PAS = "medX") # Renomeando a coluna medX para PAS
dados<-rbind(medA,medX) # Combinando as duas tabelas
```

	PAS	medicamento
1	113	A
2	149	A
3	127	A
4	109	A
5	219	A
6	125	A
7	136	A
8	129	A
9	131	A
10	110	A
11	128	X
12	123	X
13	161	X
14	142	X
15	136	X
16	139	X
17	119	X
18	110	X
19	106	X
20	146	X

- Script para construção do gráfico tipo *boxplot* utilizando o pacote ggplot2 e o gráfico resultante do código programado.

```
# Construção do boxplot usando ggplot
ggplot(dados, aes(x=medicamento, y=PAS)) +
  geom_boxplot(outlier.colour="orangered", outlier.shape=13,
               outlier.size=3,
fill=c("turquoise4","slateblue1"))+stat_summary(fun=mean, geom="point",
shape=24, size=2, color='gold',fill='limegreen')
```



Observe que o gráfico acima gerado por meio do pacote ggplot2 informa que os dados da PAS do grupo tratado com o medicamento A estão menos dispersos que os dados da PAS do grupo tratado com o medicamento X, o que é coerente com o gráfico gerado pela função nativa boxplot(), demonstrado na sessão anterior. É possível notar também os valores da média aritmética de cada grupo de dados. O grupo de participantes da pesquisa que utilizou o medicamento A possui uma média de PAS superior aos participantes do outro grupo. Como o medicamento A promoveu uma menor dispersão dos dados de PAS, ele será testado na próxima etapa da pesquisa, apesar de existir um valor discrepante neste conjunto de dados.

REFERÊNCIAS

1. MORETTIN, P. A. **Estatística Básica - 10a edição - 2023**. 10. ed. São Paulo, SP: SaraivaUni, 2022.
2. VIEIRA, S. **Introdução a Bioestatística**. 6. ed. RIO DE JANEIRO, RJ: Grupo Gen, 2022.
3. SILVA, H. A. DA. **Manual básico da linguagem R: introdução à análise de dados com a linguagem R e o RStudio para área da saúde**. [s.l.] Não definido, 2019.
4. PERES, Fernanda F. **Como interpretar (e construir) um gráfico boxplot?**. Blog Fernanda Peres, São Paulo, 29 mar. 2022. Disponível em: https://fernandafperes.com.br/blog/interpretacao-boxplot/.

Esperamos que este material tenha despertado a vontade de conhecer mais sobre a linguagem R e o ambiente de desenvolvimento Rstudio. Essas ferramentas são poderosas em análise de dados e existem inúmeras outras funcionalidades além da construção de *boxplot*. Dado que a linguagem R é de código aberto, encontrar tutoriais na internet que demonstrem como executar uma tarefa específica no Rstudio é muito simples. Desejamos bons estudos!

www.epiuff.org