Curso Intermediário (apoiado por *software* R) da Análise da Situação de Saúde aplicado a Emergências Sanitárias, com foco na COVID-19

Aula 05- Gráficos para medidas descritivas

Apresentação

Nessa aula apresentaremos mais uma atividade prática do R. Estamos disponibilizando um recorte dos dados dos principais sistemas de notificação para Covid-19.

Para os casos de Síndrome Gripal (SG), estamos disponibilizando um recorte de casos do e-SUS Notifica para casos registrados no Distrito Federal. Para os casos de Síndrome Respiratória Aguda Grave (SRAG), estamos utilizando dados do SIVEP-Gripe. O recorte de dados é pequeno, para permitir que computadores com uma memória menor consigam realizar os exercícios de forma completa.

Nessa aula, vamos apresentar conceitos de sobre gráficos que usamos para a visualização de medidas descritivas, incluindo medidas de tendencia central e dispersão, utilizando dados das duas bases de dados. Essas medidas são fundamentais na construção de indicadores de saúde, pois são o primeiro passo para uma análise de dados. Para essa aula, também usaremos o "megapacote" de *Data Science* do R, o *Tidyverse*, que inclui funções do pacote de gráficos elegantes *ggplot2.*

1. Introdução

Na aula passada aprendemos um pouco dos conceitos de medidas descritivas que usamos na Epidemiologia e na Bioestatística. Falamos sobre algumas medidas que são fundamentais na Bioestatística, as medidas de posição e as medidas de tendência central. Além de calcular essas medidas, podemos utilizar gráficos para compreender o comportamento dessas variáveis (1).

Quando fazemos análises descritivas, buscamos medidas estatísticas que sumarizem as informações contidas nos dados, dada a impossibilidade de verificar cada uma das observações e eventos registrados por meio de dados em saúde. Essas medidas apontam características dos dados, que permitem descrever e analisar suas especificidades (1).

As medidas mais utilizadas em estatística descritiva, para descrição de uma única variável- ou seja, univariada-, são medidas que procuram analisar a tendencia central e as medidas separatrizes, que nos ajudam a compreender a dispersão desses dados. Essas medidas, que permitem que façamos análises descritivas de dados são chamadas de medidas-resumo. Incluímos no grupo de medidas-resumo a média, variância, desvio padrão, que são medidas calculadas para variáveis métricas ou quantitativas. (2).

Essas medidas servem para compreender o quanto um determinado grupo de observações se aproximam de uma tendência central e o quanto estão dispersas em torno dessas medidas. Na aula 4, mostramos como essas medidas são calculadas e compreendemos como a amplitude e as medidas de dispersão ajudam a compreender o comportamento dos dados (2).

Existem outras ferramentas que nos ajudam a compreender esse comportamento de dados, quando são variáveis contínuas. Nessa aula, aprenderemos a criar gráficos elegantes com o *ggplot2*, o pacote do R mais utilizado na criação de gráficos no mundo. Vamos criar um gráfico de linhas e um boxplot para mostrar a dispersão de dados contínuos.

2. Estatísticas descritivas – medidas de frequência, tendencia central e dispersão

Como apresentamos na aula anterior, é muito importante conhecer e analisar as medidas de tendencia central e de frequência para dados da área da saúde. Vamos relembrar as medidas que utilizamos na aula passada, pois nessa aula, além de seu cálculo, vamos criar gráficos para a sua representação.

Toda análise de dados deve começar pela investigação do comportamento das variáveis. Isso ocorre para que possamos propor outras análises mais robustas que sejam adequadas aos dados. Apresentamos as principais medidas de frequência, tendencia central e posição que serão usadas nessa aula (2).

- Frequência absoluta: se refere à contagem simples do número de ocorrências de uma determinada variável ou categoria.
- Frequência relativa: se refere à contagem do número de ocorrências de um determinado evento dentro de um determinado conjunto de possibilidades desse evento. É expresso por meio de uma razão.
- Média aritmética simples: é a soma de todos os valores de determinada variável (discreta ou contínua) dividida pelo número total de observações.
- Mediana: é uma medida de localização que divide o número de observações em duas partes, representando o valor que fica posicionado exatamente ao meio da amostra, de forma que 50% dos valores serão menores que a mediana e 50% serão maiores.
- Moda: é a contagem, ou seja, a frequência absoluta do número de observações que mais aparece em determinado conjunto. Ou seja, mostra o valor que mais se repete na população ou amostra analisada.
- Separatrizes:
    - Quartis: é uma medida de localização que divide o número de observações em quatro partes, representando o valor que fica posicionado exatamente ao meio da amostra, o que fica posicionado ao meio da primeira metade da amostra e o que fica posicionado ao meio da segunda metade da amostra de forma que os valores ficam divididos em grupos ordenados contendo cada um deles 25% das observações.

    - Percentis: é uma medida de localização semelhante aos quartis, entretanto divide as observações em 100 grupos, cada um com 1% das observações.
- Variância: calcula o quanto os valores variam em torno da média e é calculado somando-se as diferenças entre as observações e a média, elevadas ao quadrado, e dividindo isso pelo número de observações.
- Desvio padrão: também calcula o quanto os valores variam em torno da média e é calculado tirando-se a raiz quadrada da variância.

3. Carregando pacotes e importando os dados

Para essa aula, usaremos o pacote *Tidyverse* e as mesmas bases de dados utilizados nas aulas anteriores. Também vamos rodar novamente o *script* da aula 3, que realiza toda a padronização dos dados. Para isso usaremos os seguintes comandos:

```
# carregando pacotes
library(tidyverse) #pacote inclui o ggplot2

## datasets de dados
esus = read.csv2("dados/20210601_dadosesus_df.csv") #importando dados do esus df de 01/06/2021
sivep = read.csv2("dados/20210823_dadossivep.csv") #importando dados do esus df de 23/08/2021


source("scripts/03_aula_importando_dados.R")  #realizando os tratamentos do script 03
```

O pacote e os dados serão carregados.

4. Gráficos para dados contínuos

a. Gráfico de linhas

O gráfico de linhas é um gráfico para dados contínuos que permite criarmos uma visualização para dados a partir de uma linha que especificam a continuidade de um determinado indicador. Para a criação do gráfico de linhas, utilizaremos o pacote *ggplot2*, que está incorporado ao pacote *tidyverse*, usado anteriormente nesse curso.

Inicialmente, calcularemos a taxa de incidência dos casos. Após isso, criaremos um objeto do tipo *ggplot.* Esse tipo de objeto armazenará as

informações dos nossos dados e permitirá que criemos feições dos gráficos a partir dos dados que foram importados em nosso *data.frame*.

```
#criando um objeto ggplot
esus_gra = ggplot(esus)
```

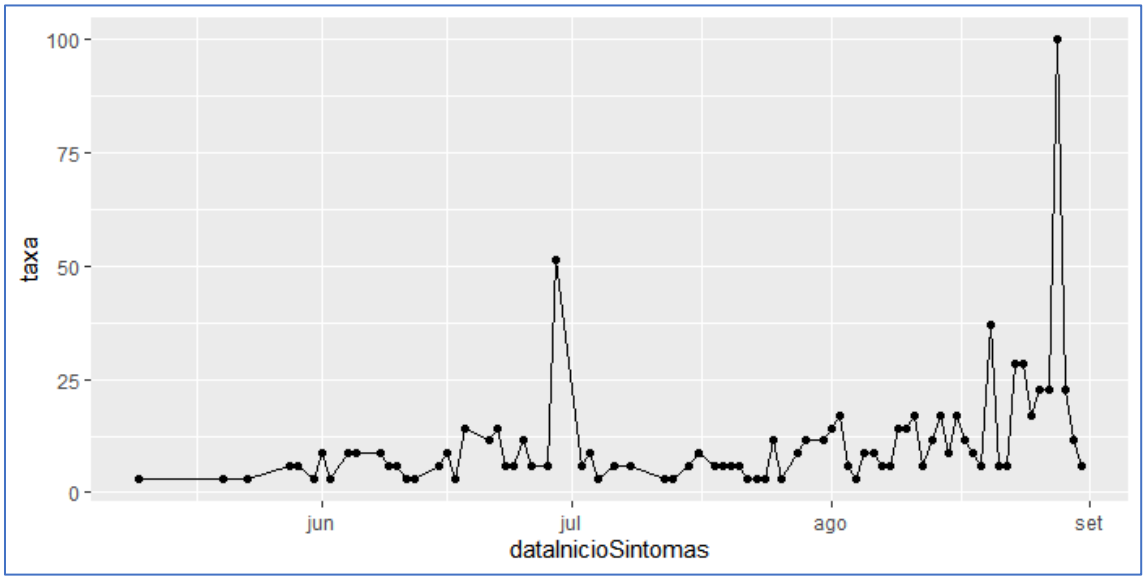
Após isso, vamos adicionar uma feição geométrica ao objeto, que será um gráfico de linhas. Vamos calcular a taxa de incidência de notificações, com base em uma população fictícia de 3500 habitantes.

```
esus$n = 1
esus_agday = aggregate(n ~ dataInicioSintomas, data = esus, FUN = sum)
esus_agday$dataInicioSintomas = as.Date(esus_agday$dataInicioSintomas)
esus_agday$taxa = round(esus_agday$n/3500*10000,2)

#criando um objeto ggplot
esus_gra = ggplot(esus)

## criando um gráfico de linha simples
ggplot(data=esus_agday, aes(x=dataInicioSintomas, y=taxa, group=1)) +
  geom_line()+
  geom_point()
```

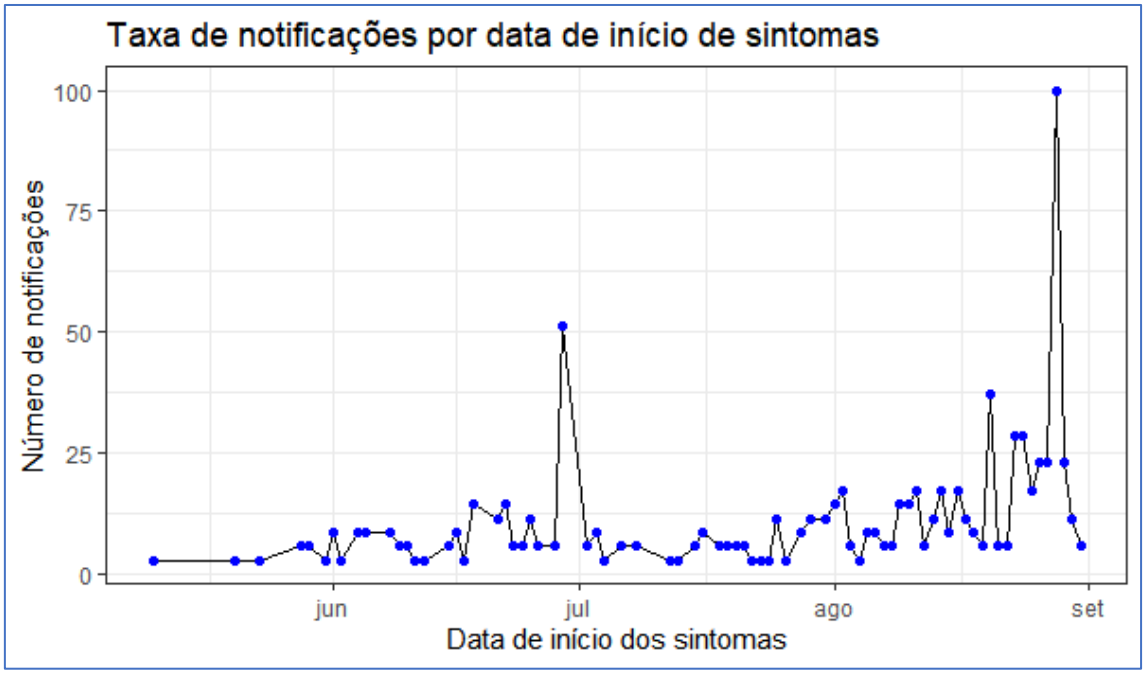
Para esse dado, teremos o seguinte *output*:



Podemos incrementar o nosso gráfico, mudando suas cores, adicionando títulos dos eixos e título geral e mudando o tema de fundo da imagem. Existe uma infinidade de feições que você pode adicionar, usando os seguintes comandos:

```
## adicionando feições ao gráfico de linha
ggplot(data=esus_agday, aes(x=dataInicioSintomas, y=taxa, group=1)) +
  geom_line(fill = "blue")+
  geom_point(colour = "blue")+
  xlab("Data de início dos sintomas")+ylab("Número de notificações")+
  ggtitle("Taxa de notificações por data de início de sintomas")+
  theme_bw()
```

Adicionamos uma linha de contorno branca e o preenchimento ficou azul. Também mudamos o tema para uma cor branca e adicionamos os rótulos adequados nos eixos x e y. O *output* desses comandos é:



b. *Boxplot*

O boxplot é um gráfico univariado que permite identificarmos a dispersão de dados numéricos, mostrando os dados a partir de suas medianas e distâncias interquartílicas.

Para esse gráfico, faremos os mesmos procedimentos do gráfico anterior. Vamos criar um objeto do tipo ggplot.

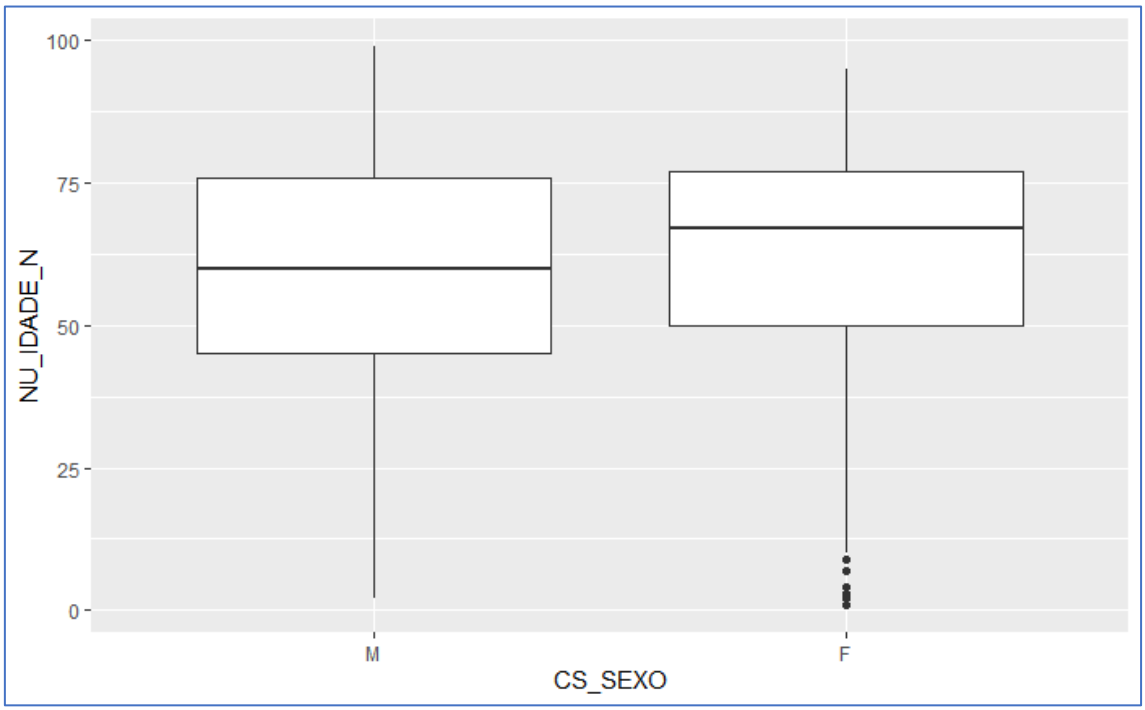
```
## criando um boxplot simples
sivep_gra = ggplot(sivep)
```

Após isso, vamos adicionar feições ao gráfico, inicialmente criando um gráfico simples, apenas com o boxplot. Nesse gráfico vamos comparar a dispersão das idades no sexo feminino e no sexo masculino.

```
## criando um boxplot simples
sivep_gra+
```

```
  geom_boxplot(aes(x = CS_SEXO, y = NU_IDADE_N))
```

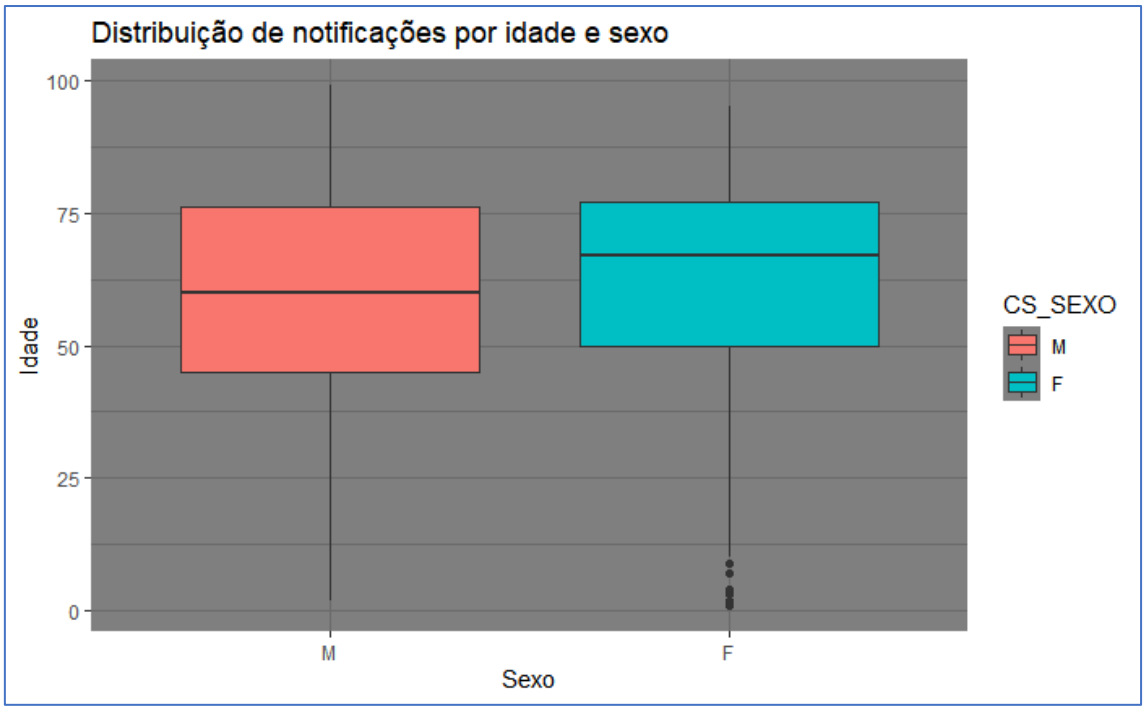
O *output* para esse comando é o seguinte. Note que embora possamos ver que a mediana de idade seja maior em mulheres, cerca de 75% das pessoas são menores 75 anos. Nesse caso, 50% do grupo dos homens possui idades menores de aproximadamente 57 anos, enquanto nas mulheres 50% das notificações são de menores que 60 anos. A diferença é pequena e pelo gráfico a distribuição não é muito diferente.



Assim como no gráfico de linhas, vamos incluir os títulos dos eixos, o título do gráfico e vamos mudar a cor para diferenciar o *boxplot* da distribuição feminina e masculina.

```
sivep_gra+
  geom_boxplot(aes(x = CS_SEXO, y = NU_IDADE_N, fill = CS_SEXO))+
  xlab("Sexo")+ylab("Idade")+
  ggtitle("Distribuição de notificações por idade e sexo")+
  theme_dark()
```

O *output* para esse gráfico é o seguinte:



5. Encerramento

Ficamos muito felizes que tenha chegado até aqui! Passamos da metade do curso e agora você precisará completar apenas mais 4 aulas. As próximas aulas vão abordar medidas e gráficos para dados categóricos e cálculo de taxas epidemiológicas.

Esperamos que você esteja aproveitando e aprendendo muito! Até mais!

6. Referências Bibliográficas

1. Lopes B, Ramos IC de O, Ribeiro G, Correa R, Valbon B de F, Luz AC da, et al. Bioestatísticas: conceitos fundamentais e aplicações práticas. Rev Bras Oftalmol. fevereiro de 2014; 73:16–22.

2. Fávero LP, Belfiore P. Manual de Análise de Dados: Estatística e Modelagem Multivariada com Excel®, SPSS® e Stata®. Elsevier Brasil; 2017. 1832 p.