

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

## **Aula 8 - Gráficos para variáveis categóricas**

## **Apresentação**

Você já aprendeu um pouco sobre medidas descritivas sempre pensando em dados numéricos. Já aprendeu a calcular tais medidas para dados quantitativos contínuos e discretos, já trabalhou com medidas de frequência absolutas e relativas para variáveis categóricas. Nessa aula apresentaremos mais uma atividade prática do R, voltada para a apresentação de dados categóricos, agora de uma forma diferente! Estaremos voltando a criar gráficos. 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 medidas de frequência para variáveis categóricas, usando agora gráficos para mostrar essas medidas de forma gráfica, complementando a aula anterior, quando apresentamos esses dados em formato de tabelas. Para essa aula, também usaremos o “megapacote” de *Data Science* do R, o *Tidyverse*, que inclui funções do pacote de gráficos elegantes *ggplot2*. Usamos esse pacote para a criação de gráficos em aulas anteriores, mas abordando dados quantitativos. Nessa aula, o foco será os gráficos para dados qualitativos.

## **1. Introdução**

Nas aulas anteriores, apresentamos conceitos de frequência absoluta e relativa. Nessas aulas, aprendemos a lidar com dados qualitativos e como criar tabelas para apresentação desses dados. Aprendermos que as análises que mostramos anteriormente, que se prendiam aos dados numéricos, procuravam compreender as medidas de tendência central e dispersão dos dados, para compreendermos o comportamento dessas variáveis na população estudada.

Para estudarmos esse assunto, falamos de medidas de frequência absoluta e medidas de frequência relativa. Já apresentamos os conceitos dessas medidas de frequência anteriormente, mas vamos lembrá-los nessa aula.

Para as medidas de frequência absoluta, fizemos a contagem simples do número de vezes em que uma determinada observação aparece. Para podermos comparar as medidas, calculamos o quanto essa categoria representa do total. Isso é uma frequência relativa, quando contamos o número de vezes em que essa categoria aparece em relação ao total de observações.

Temos objetivos específicos, quando trabalhamos com dados categóricos. Procuramos compreender onde se repetem as mesmas categorias, o quanto essas categorias estão inseridas no todo do conjunto de dados e qual a frequência em que aparecem nas observações estudadas. As tabelas foram úteis para mostrar esses dados nas aulas anteriores.

Nesta aula, vamos apresentar os gráficos para medidas relativas. Nós calcularemos as frequências de determinadas categorias de dados e criaremos gráficos elegantes com seus resultados. Para os gráficos, usaremos uma formatação do R, que permite a criação de gráficos elegantes por meio do pacote ggplot2 e que podem ser incrementadas para melhorar sua visualização.

## **2. Estatísticas descritivas – medidas de frequência para dados categóricos**

Alguns conceitos apresentados em aulas anteriores são muito importantes para conhecer e analisar as medidas de frequência para dados da área da saúde. Vamos relembrar as medidas que conhecemos em aulas passadas, pois nessa aula, vamos aplicar esses conceitos na construção de tabelas de frequência.

Quando falamos de dados categóricos, não é possível calcular medidas de tendência central ou de dispersão, visto que esses dados são calculáveis a partir da construção de medidas numéricas, naturalmente mensuráveis. Isso ocorre para que possamos propor outras análises mais robustas que sejam adequadas aos dados.

Mas como fazemos quando queremos descrever dados categóricos? Os dados categóricos são divididos em duas grandes categorias: os dados nominais e os dados ordinais. Os dados nominais se referem às categorias que possuem nomes, mas nenhuma hierarquia é imputada a esses dados, como por exemplo, sexo e estado de moradia (2).

Para os dados ordinais, nós trabalhamos com categorias que apresentam uma ordem numérica de hierarquia, embora não sejam mensuráveis como os dados quantitativos. Podemos citar como exemplo a classe social e a escolaridade nesse tópico (2).

Apresentamos as principais medidas de frequência 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.
- 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.

Nessa aula, falaremos ainda dos gráficos de barras. Esses gráficos são úteis para representar categorias em um plano cartesiano.

- Gráficos de barras simples: esses gráficos são úteis para representar o número de observações em uma determinada categoria em que as categorias estão no eixo x e as quantidades no eixo y. Diferentemente de um histograma, onde as barras são contínuas para representar a continuidade dos dados, no gráfico de barras, há uma separação entre as barras de cada categoria, simbolizando que são independentes.
- Gráficos de barras empilhadas: uma variação do gráfico de barras é o gráfico de barras empilhadas. Isso ocorre quando se quer representar subcategorias em uma mesma barra, que pertence a uma categoria anterior. Também pode ser usado para representar porcentagens. Nesse caso, todas as barras terão o mesmo tamanho, representando 100% dos valores e cada pedaço da barra que representa uma categoria, possuirá sua porcentagem. Esse tipo de barra é útil para a comparação de proporções, mas pode ser enganoso ao olhar ao dar a impressão de que os dados estão na mesma escala, quando isso pode não ser verdade. Portanto, tenha cuidado ao usá-lo
- Gráficos de barras agrupadas: Tem uma finalidade semelhante ao gráfico de barras empilhadas, com a função de representar dados de subcategorias dentro de um gráfico com categorias maiores. Nesse caso, são agrupadas mais uma barra por categoria, representando as subcategorias presentes nos dados.

### **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. Vamos carregar ainda o pacote *reshape2*, pois ele será usado para reorganizar os dados de uma tabela que elaboraremos nessa aula.

O pacote *ggplot2*, conforme vimos na aula 4, também está incorporado ao *Tidyverse*, assim, não será necessário carregá-lo ou instalá-lo separadamente.

Para carregar os pacotes e realizar o tratamento dos dados usaremos os seguintes comandos:

```
# carregando pacotes
library(tidyverse) #tratamento dos dados

#install.packages("reshape2")
library(reshape2)

## 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.

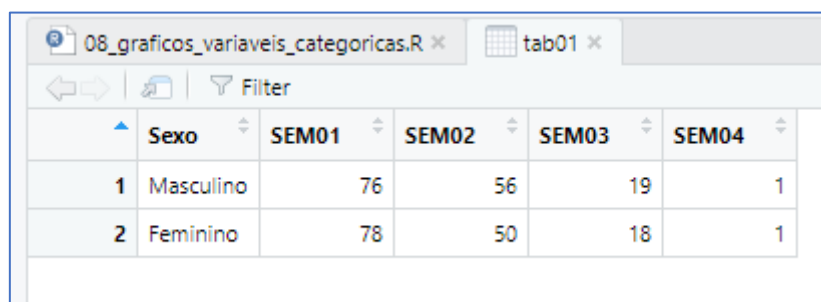
#### 4. Gráficos para dados categóricos

Para trabalhar nessa aula, usaremos os gráficos de barras. Antes de criar os gráficos, conforme mostramos anteriormente, precisamos criar tabelas com as análises que queremos plotar.

Vamos criar uma tabela com as frequências absolutas de dados por sexo e semana epidemiológica, conforme criamos na aula anterior. Vamos organizar o nome das variáveis e das colunas. Os comandos são os seguintes:

```
## criando uma tabela cruzada
tab01 = data.frame(sivep$CS_SEXO, sivep$SEM_PRI)
colnames(tab01) = c("Sexo", "SEM01", "SEM02", "SEM03", "SEM04") #incluindo os
títulos
tab01$Sexo = as.character(tab01$Sexo) #alterando as categorias de sexo
tab01$Sexo[tab01$Sexo == "M"] = "Masculino"
tab01$Sexo[tab01$Sexo == "F"] = "Feminino"
```

Para esses dados, receberemos os seguintes *outputs*:

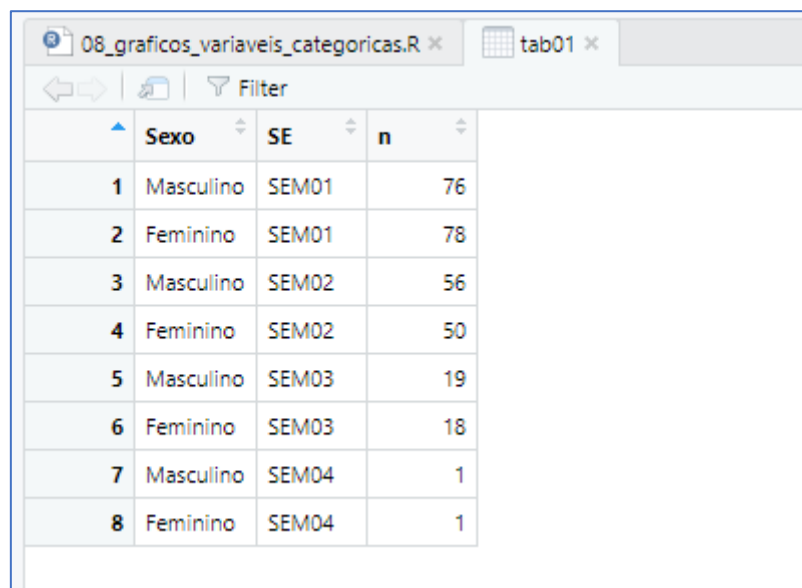


	Sexo	SEM01	SEM02	SEM03	SEM04
1	Masculino	76	56	19	1
2	Feminino	78	50	18	1

Para a criação do gráfico, não precisaremos da coluna de total, por isso ela não foi e não será criada. Como usaremos o *ggplot2*, será necessário reorganizar os dados em três variáveis diferentes, usando o *reshape2*. Daremos novos nomes para as variáveis criadas. Os seguintes comandos serão necessários:

```
tab01 = melt(tab01, id.vars = "Sexo")
colnames(tab01) = c("Sexo", "SE", "n")
```

Para esses dados, o seguinte *output* será obtido:



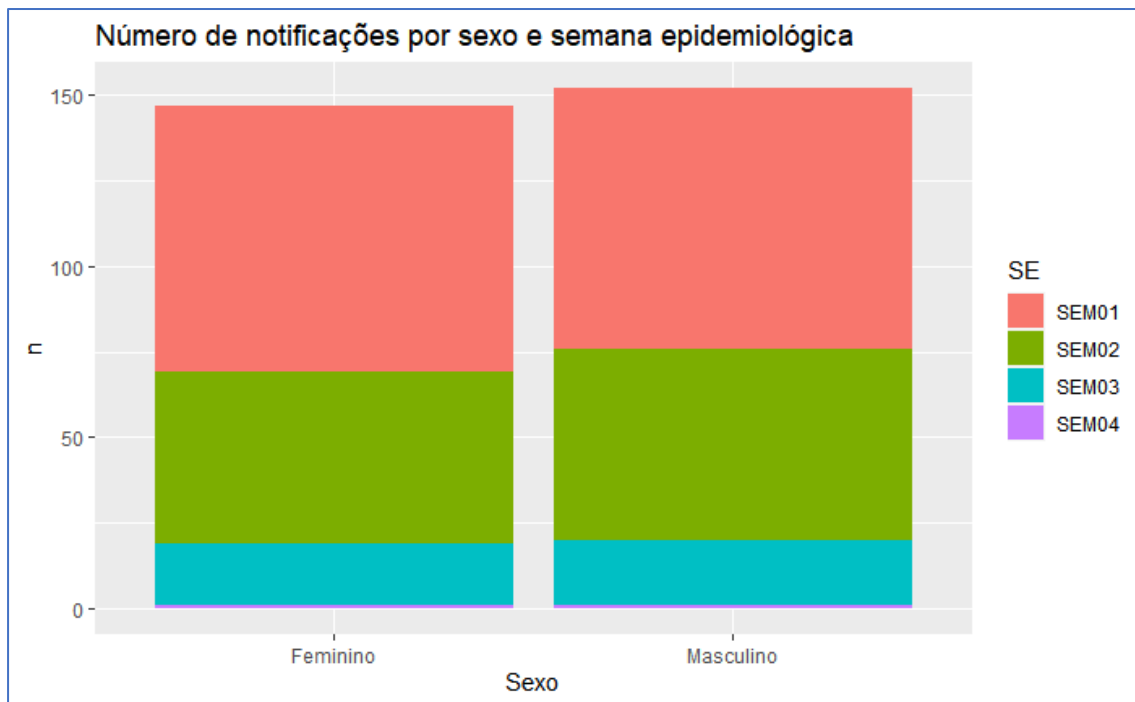
	Sexo	SE	n
1	Masculino	SEM01	76
2	Feminino	SEM01	78
3	Masculino	SEM02	56
4	Feminino	SEM02	50
5	Masculino	SEM03	19
6	Feminino	SEM03	18
7	Masculino	SEM04	1
8	Feminino	SEM04	1

Dessa forma, poderemos criar um gráfico de barras com o *ggplot2*. Faremos um gráfico de barras empilhadas, em que o total é o número de casos no sexo feminino e masculino e as categorias são as semanas epidemiológicas. Para isso, usaremos os seguintes comandos:

# criando gráfico de barras para valores absolutos

```
ggplot(tab01)+
  geom_bar(aes(x = Sexo, y = n, fill = SE), stat = "identity")+
  ggtitle("Número de notificações por sexo e semana epidemiológica")
```

Para esses comandos, teremos os seguintes *outputs*:



Também faremos um gráfico de barras de proporções. Para isso criaremos outra tabela, fazendo seus tratamentos conforme as aulas anteriores por meio dos seguintes comandos:

```
## tabela de número de notificações por classificação final
tab02 = data.frame(table(esus$classificacaoFinal))
tab02$Var1 = as.character(tab02$Var1)

tab02[1,1] = "Não classificado" #incluindo linha sem classificação final
colnames(tab02) = c("classificacao", "n") #incluindo rótulos na tabela
```

Ao realizar esses comandos, teremos o seguinte resultado:

08_graficos_variaveis_categoricas.R		tab02	
		Filter	
	Classificacao	n	
1	Não classificado	197	
2	Confirmado Clínico-Epidemiológico	33	
3	Confirmado Laboratorial	59	
4	Descartado	8	
5	Síndrome Gripal Não Especificada	2	

Conforme mostramos anteriormente, também vamos calcular a proporção de cada categoria em relação ao total, conforme os códigos a seguir:

```
tab02$Proporcao = tab02$n/sum(tab02$n)*100 #calculando a porcentagem
```

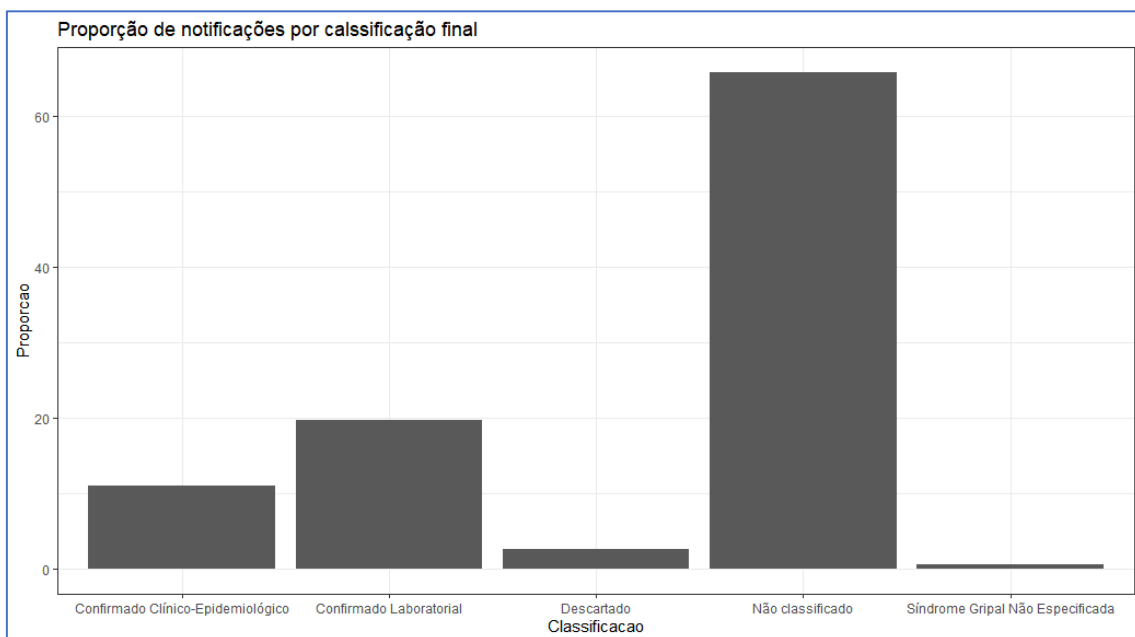


Com isso, teremos uma tabela da seguinte forma:

	Classificacao	n	Proporcao
1	Não classificado	197	65.8862876
2	Confirmado Clínico-Epidemiológico	33	11.0367893
3	Confirmado Laboratorial	59	19.7324415
4	Descartado	8	2.6755853
5	Síndrome Gripal Não Especificada	2	0.6688963

Usando a última coluna, de proporção, criaremos um gráfico de barras com os dados dessa tabela. Já incluímos os argumentos de tratamento dos dados, para que o gráfico fique completo.

Obteremos o seguinte gráfico:



Também é possível elaborar gráficos de setores para variáveis categóricas. Muito cuidado ao utilizar esses gráficos, pois eles podem ser enganosos, por isso utilize-os apenas se houver no máximo 5 categorias.

```
# criando um gráfico de setores para a proporção
sexo1 = aggregate(n ~ Sexo, data = tab01, FUN = sum)

pie(sexo1$n, sexo1$Sexo, border = "white", col = c("red", "blue"),
    main = "Proporção de notificações por sexo")
```

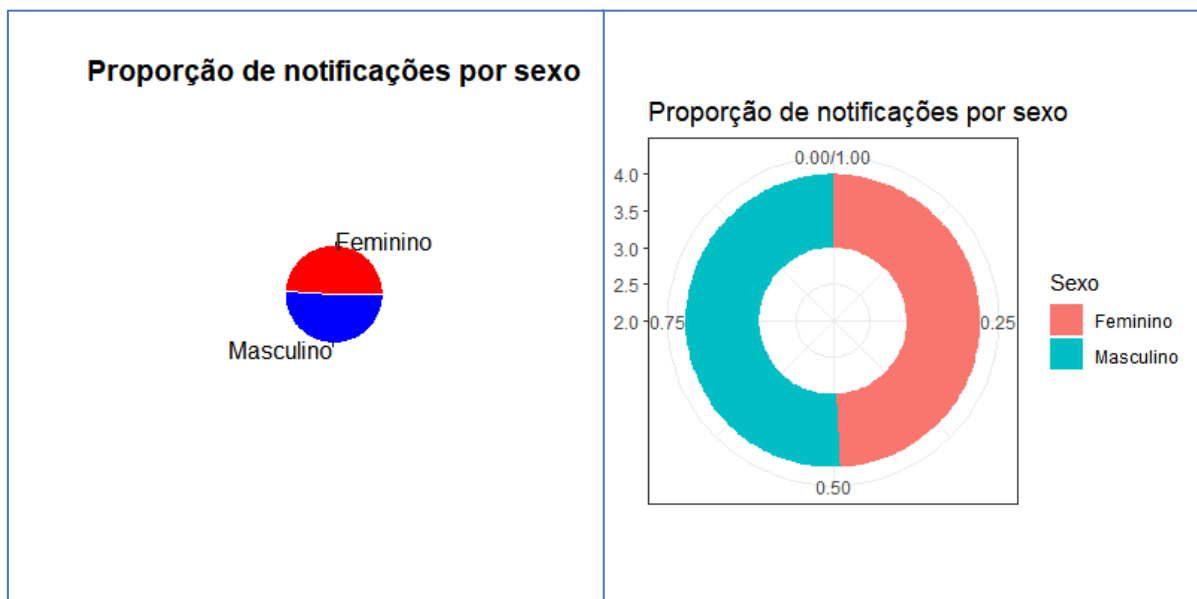
```

sexo1$prop = sexo1$n/sum(sexo1$n) #calculando a proporção
sexo1$ymax = cumsum(sexo1$prop) #calculando a proporção cumulativa
sexo1$ymin = c(0, head(sexo1$ymax, n=-1))

ggplot(sexo1, aes(ymax=ymax, ymin=ymin, xmax=4, xmin=3, fill=Sexo)) +
  geom_rect() +
  coord_polar(theta="y") +
  xlim(c(2, 4))+
  ggtitle("Proporção de notificações por sexo")+
  theme_bw()

```

Para esse gráfico receberemos os seguintes *outputs*:



## 5. Encerramento

Acaba de finalizar a penúltima aula de nosso curso! Esperamos que esteja aprendendo bastante com ele. Ficamos muito felizes por ter chegado até aqui. Na última aula, apresentaremos o cálculo de taxas de incidência com dados do e-SUS Notifica e do Sivep-Gripe.

Ao final do curso, você terá conhecimentos do R enquanto ferramenta e das possibilidades de análise de dados que podem ser aplicáveis à saúde pública. Agradecemos por terem chegado até aqui! Siga em frente, você aprenderá ainda 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.