

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

Aula 06- Gerando indicadores com medidas absolutas: medidas de frequência

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. Nessa aula apresentaremos mais uma atividade prática do R, voltada para a apresentação de dados categóricos. 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, inicialmente usando tabelas para apresentar as categorias e sua contagem em medidas absolutas. Para essa aula, também usaremos o “megapacote” de *Data Science* do R, o *Tidyverse*, que inclui funções do pacote de tabelas elegantes *formattable*.

1. Introdução

Quando falamos de análise de dados, sempre nos preocupamos com a natureza quantitativa desses dados, que é a parte fundamental da análise bioestatística e epidemiológica. Mas, também é possível fazermos análises com dados qualitativos advindos de cenários de saúde. Mesmo nesse cenário, as análises ainda são quantitativas e não há nenhuma análise voltada para as metodologias qualitativas.

Mas como podemos analisar dados qualitativos de forma quantitativa? Como isso deve funcionar? Quando falamos de análise de dados qualitativos, estamos falando da contagem de sua ocorrência e do cálculo de sua proporção dentro de um conjunto de dados ou observações. Essas análises são fundamentais na epidemiologia. Quando fazemos análise dos desfechos em saúde, como cura, adoecimento ou morte, estamos fazendo a análise de dados qualitativos.

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. Queríamos entender a amplitude dos dados, onde se concentravam o maior número de observações e o quanto isso era disperso da média ou da mediana daqueles dados.

Para os dados qualitativos, procuraremos 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.

Para estudarmos esse assunto, vamos falar 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, faremos a contagem simples do número de vezes em que uma determinada observação aparece. Para a frequência relativa, vamos contar o número de vezes em que essa observação ou categoria aparece em relação ao total de observações.

Para mostrar esses dados, podemos usar gráficos, que abordaremos mais à frente, mas também poderemos usar tabelas. Para as tabelas, usaremos uma formatação do R, que permite a criação de tabelas elegantes 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.

3. Carregando pacotes e importando os dados

Vamos usar mais uma vez o pacote *Tidyverse*, que será carregado para a transformação, padronização e tratamento de dados, que foram apresentados na aula 3. Mais uma vez, reproduziremos o tratamento realizado na aula 3.

Como novidade, vamos trabalhar com o pacote *formattable*, que é empregado na formatação de tabelas de forma elegante e do pacote *reshape2*, que é utilizado para organizar os dados dentro de uma tabela. Dessa forma, incluímos a instalação dos pacotes nessa aula.

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

install.packages("formattable")
library(formattable) #pacote para criação de tabelas bonitas

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
```

Os dados e os pacotes serão instalados e carregados.

4. Tabelas para dados categóricos

Para essa aula, criaremos uma tabela cruzada de frequência absoluta, que vai mostrar o número de casos por sexo e semana epidemiológica nos dados que estamos utilizando nesse curso.

Inicialmente, vamos criar um `data.frame` com os dados de sexo e semana epidemiológica:

```
## criando uma tabela cruzada
tab01 = data.frame(sivep$CS_SEXO, sivep$SEM_PRI)
```

Para esses dados, teremos uma tabela com 299 observações de duas variáveis, conforme a imagem:

	sivep.CS_SEXO	sivep.SEM_PRI
1	F	1
2	M	1
3	F	1
4	M	1
5	F	1
6	M	1
7	F	1
8	F	1
9	F	1
10	M	1
11	F	1
12	M	1
13	F	1
14	M	1
15	F	1
16	M	1
17	M	1
18	F	2
19	F	1
20	F	1
21	F	1

Showing 1 to 22 of 299 entries, 2 total columns

Essa tabela não é nenhum pouco informativa e não nos mostra o número de casos por semana epidemiológica e por sexo. Assim, faremos a reorganização dos dados por sexo e SE, usando um comando do pacote *reshape2*.

```
tab01 = dcast(sivep.CS_SEXO ~ sivep.SEM_PRI, data = tab01) #reorganizando os dados
```

Para esse comando, obteremos o seguinte *output*:

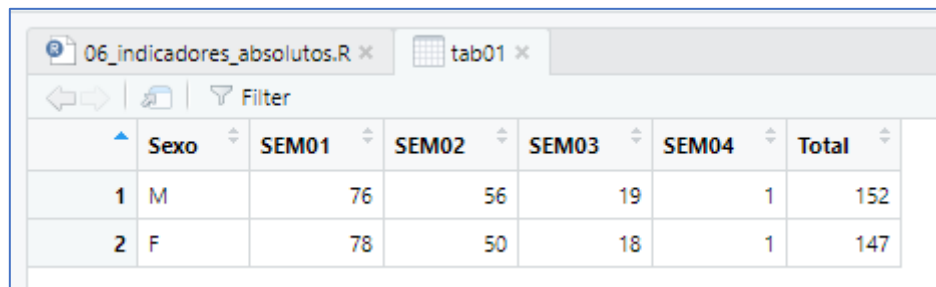
	sivep.CS_SEXO	1	2	3	4
1	M	76	56	19	1
2	F	78	50	18	1

Nesse momento, conseguimos identificar o número de notificações por sexo e semana epidemiológica. Conseguimos identificar as frequências absolutas dos dados, assim como a moda. Nesse caso, a moda é de mulheres, na primeira semana epidemiológica.

Vamos mudar os rótulos da tabela, para identificar a semana e o sexo de forma mais clara. Não é uma boa prática usar números como nome das colunas no R, pois isso pode atrapalhar comandos posteriormente. Para mudar os rótulos e adicionar o total, usaremos os seguintes comandos:

```
colnames(tab01) = c("Sexo", "SEM01", "SEM02", "SEM03", "SEM04") #incluindo os títulos
tab01$Total = rowSums(tab01[,2:5]) #incluindo uma coluna de total
```

Após essa transformação, nossa tabela aparecerá da seguinte forma:

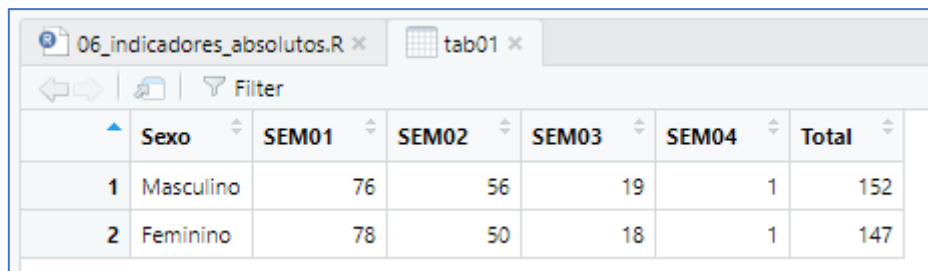


	Sexo	SEM01	SEM02	SEM03	SEM04	Total
1	M	76	56	19	1	152
2	F	78	50	18	1	147

Ainda podemos melhorar mais nossa tabela. Para isso, vamos colocar o nome das variáveis categóricas de sexo por extenso, conforme o comando:

```
tab01$Sexo = as.character(tab01$Sexo) #alterando as categorias de sexo
tab01$Sexo[tab01$Sexo == "M"] = "Masculino"
tab01$Sexo[tab01$Sexo == "F"] = "Feminino"
```

Nesse caso, nossa tabela mudará os valores das células da primeira coluna, conforme a imagem:



	Sexo	SEM01	SEM02	SEM03	SEM04	Total
1	Masculino	76	56	19	1	152
2	Feminino	78	50	18	1	147

Vamos formatar a tabela de uma maneira que possamos exportá-la como uma tabela bonita. Para isso, usaremos o comando *formattable*, do pacote homônimo.

```
## formatando a tabela
formattable(tab01)
```

Usando apenas esse comando, sem definição de nenhum parâmetro, conseguiremos o seguinte *output*:

Sexo	SEM01	SEM02	SEM03	SEM04	Total
Masculino	76	56	19	1	152
Feminino	78	50	18	1	147

Embora seja uma visualização melhor que a anterior, ainda podemos melhorar mais! Vamos modificar as cores da coluna sexo para cinza e colocá-las em negrito, assim como faremos com a coluna total, que continuará com as letras pretas. Para os dados da semana epidemiológica, vamos usar um argumento de graduação de cor, usando a cor verde. As maiores frequências, ficarão com as cores mais fortes.

Os comandos são os seguintes:

```
formattable(tab01, align = c("l",rep("r", NCOL(tab01) - 1)), list(
  `Sexo` = formatter("span", style = ~ style(color = "grey",font.weight =
"bold")),
  area(col = 2:5) ~ color_tile("#DeF7E9", "#71CA97"),
  `Total` = formatter("span", style = ~ style(color = "black", font.weight =
"bold"))))
```

Para esse comando, teremos a seguinte saída:

Sexo	SEM01	SEM02	SEM03	SEM04	Total
Masculino	76	56	19	1	152
Feminino	78	50	18	1	147

5. Encerramento

Mais uma aula finalizada! Esperamos que esteja aprendendo bastante com o nosso curso. Ficamos muito felizes por ter chegado até aqui. Nas próximas aulas, apresentaremos tabelas para frequências relativas, assim como gráficos para variáveis categóricas.

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.