

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

Aula 7 – Gerando indicadores com medidas relativas: calculando proporções

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

Na aula anterior, falamos sobre medidas de frequência absoluta. Hoje falaremos sobre medidas de frequência relativa. Aprenderemos 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.

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. Com os dados qualitativos esses objetivos são diferentes. 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.

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, fizemos a contagem simples do número de vezes em que uma determinada observação aparece. Esse tipo de análise não permite a comparação entre diferentes categorias.

O que fazemos quando queremos comparar diferentes categorias? Podemos calcular o quanto essa categoria representa do total. Nesta aula, falaremos de frequência relativa, vamos contar o número de vezes em que essa 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. Nós calcularemos as frequências relativas de determinadas categorias de dados e criaremos tabelas elegantes com seus resultados. 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.

Também vamos trabalhar de novo 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. Como os pacotes já estão instalados, vamos usar os seguintes comandos;

```
# 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. Frequências relativas e tabelas para dados categóricos

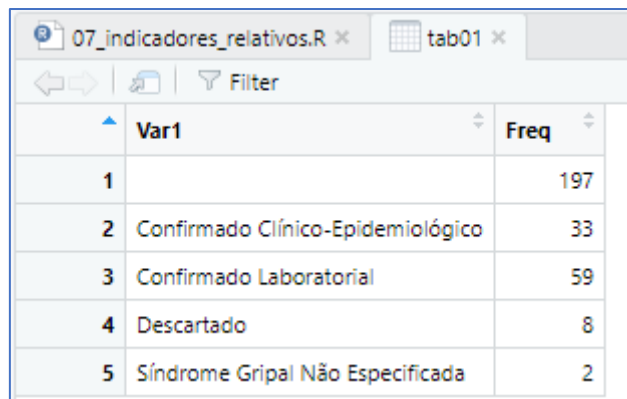
Para essa aula, vamos trabalhar com a proporção de cada classificação final dentre todas as notificações do esus consideradas na base de dados que disponibilizamos com dados do Distrito Federal.

Inicialmente, criaremos uma tabela de frequência simples, com o comando:

```
## criando tabela simples

## tabela de número de notificações por classificação final
tab01 = data.frame(table(esus$classificacaoFinal))
```

Para esse comando, teremos o seguinte resultado no R:



	Var1	Freq
1		197
2	Confirmado Clínico-Epidemiológico	33
3	Confirmado Laboratorial	59
4	Descartado	8
5	Síndrome Gripal Não Especificada	2

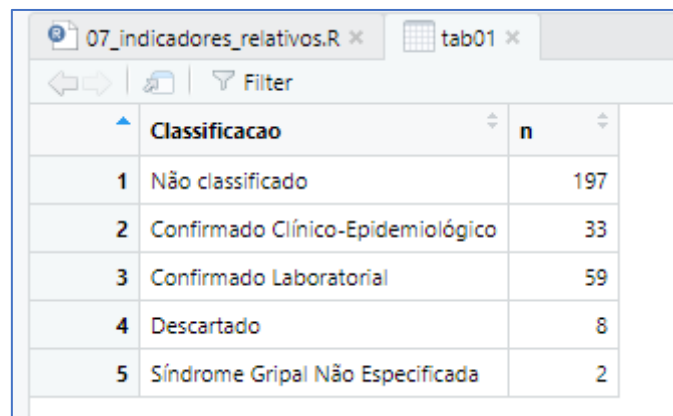
Nesse caso, temos a frequência de cada tipo de classificação final, incluindo os casos que não foram classificados. Não temos títulos próprios para nossa tabela, assim precisaremos incluir isso. Também precisaremos incluir uma categoria para os casos que não foram classificados na tabela.

Para essas finalidades, usaremos os comandos a seguir:

```
tab01$Var1 = as.character(tab01$Var1)

tab01[1,1] = "Não classificado" #incluindo linha sem classificação final
colnames(tab01) = c("Classificacao", "n") #incluindo rótulos na tabela
```

Nesse caso, nossa tabela ganhará essa nova categoria, além de ter rótulos próprios. A imagem mostra como a tabela ficou após a transformação:



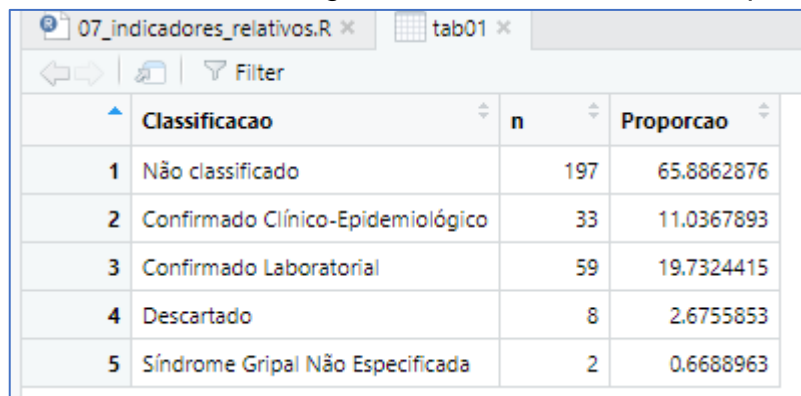
	Classificacao	n
1	Não classificado	197
2	Confirmado Clínico-Epidemiológico	33
3	Confirmado Laboratorial	59
4	Descartado	8
5	Síndrome Gripal Não Especificada	2

Apesar de já apresentar uma forma melhor, falta algo fundamental em nossa tabela: a medida de frequência relativa que ajudará a comparar as diferentes categorias no grupo. Vamos calcular a porcentagem, que é dada pelo número de observações na categoria dividido pela soma de todas as categorias multiplicado por 100.

O comando a seguir fará o cálculo da porcentagem:

```
tab01$Proporcao = tab01$n/sum(tab01$n)*100 #calculando a porcentagem
```

Nesse caso, nossa tabela ganhará uma nova coluna de porcentagem:

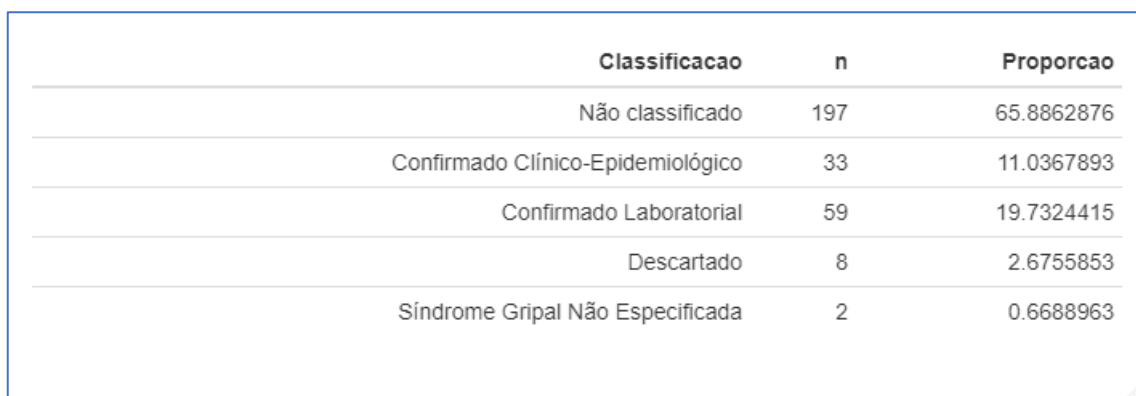


	Classificacao	n	Proporcao
1	Não classificado	197	65.8862876
2	Confirmado Clínico-Epidemiológico	33	11.0367893
3	Confirmado Laboratorial	59	19.7324415
4	Descartado	8	2.6755853
5	Síndrome Gripal Não Especificada	2	0.6688963

Essa tabela é mais informativa que a anterior, mas necessita de formatação, se quisermos apresentá-la em algum relatório. Para formatar a tabela, usaremos o comando *formattable*, conforme ensinado anteriormente.

```
## formatando a tabela  
formattable(tab01) #formatando a tabela de forma simples
```

Receberemos o seguinte *output*:



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

Embora melhor apresentado, ainda podemos aprimorar essa tabela. Primeiramente, queremos apenas duas casas decimais na proporção e queremos que o símbolo de porcentagem apareça. Vamos alinhar a primeira coluna à esquerda, com letras em cinza e negrito. Também colocaremos uma barra vermelha na coluna de frequência, para mostrar o quanto aquela frequência representaria em uma barra. Os comandos são os seguintes:

```
formattable(tab01, align = c("l", "r", "r"),
             list(`Classificacao` = formatter("span", style = ~ style(color =
"grey", font.weight = "bold")),
                 `n` = color_bar("#FA614B"),
                 `Proporcao` = formatter("span",
                                         x ~ percent(x / 100))))
```

Para esses comandos, receberemos o seguinte *output*.

Classificacao	n	Proporcao
Não classificado	197	65.89%
Confirmado Clínico-Epidemiológico	33	11.04%
Confirmado Laboratorial	59	19.73%
Descartado	8	2.68%
Síndrome Gripal Não Especificada	2	0.67%

5. Encerramento

Você completou 77% de nosso curso! Esperamos que esteja aprendendo bastante com ele. Ficamos muito felizes por ter chegado até aqui. Nas próximas aulas, apresentaremos como gráficos para variáveis categóricas e 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.