

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

Aula 04- Gerando estatísticas descritivas

Apresentação

Nessa aula apresentaremos a segunda 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 estatísticas descritivas de dados, com dados das duas bases de dados. Vamos conhecer as principais medidas descritivas que incluem medidas de tendência central e dispersão. 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*.

1. Introdução

Quando falamos de abordagens quantitativas em Ciências da Saúde, temos dois grandes campos que nos auxiliam na compreensão e análise de dados em saúde: a Epidemiologia e a Bioestatística. No campo da Bioestatística, trabalhamos com algumas medidas que são fundamentais em análise de dados: as medidas de posição e de tendência central (1).

Há um grande número de dados em saúde e para compreender os fenômenos que afetam o processo saúde-doença das populações é necessário compreender como esses dados se comportam. Pelo estudo clássico da estatística podemos afirmar que o comportamento dos dados pode ser explicado por diferentes padrões, que deem ser investigados e nos ajudam a explicar os eventos de saúde (1).

Nesse sentido, buscamos medidas estatísticas que sumariem 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 que permitem que façamos análises descritivas de dados são chamadas de medidas-resumo. As medidas mais utilizadas em estatística descritiva, para descrição de uma única variável- ou seja, univariada-, são medidas que procuram analisar a tendência central e as medidas separatrizes, que nos ajudam a compreender a dispersão desses dados (2).

Quando falamos dessas medidas, nos referimos à média, variância, desvio padrão, que são medidas calculadas para variáveis métricas ou quantitativas. Apesar disso, podemos citar a moda como uma exceção, visto que é baseada na contagem da maior frequência em que um dado aparece e pode ser utilizada para a análise de outras variáveis, que não as quantitativas (2).

Nessa aula abordaremos essas medidas de tendência central e dispersão para variáveis métricas e falaremos sobre as contagens de frequências para as variáveis qualitativas. Usaremos o R para calcular as medidas mostrando como podem ser utilizadas na área da saúde.

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

Como apresentamos no tópico anterior, é muito importante conhecer e analisar as medidas de tendencia central e de frequência para dados da área da saúde. 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 na aula anterior. Para isso usaremos os seguintes comandos:

```
# carregando pacotes
library(tidyverse) #pacote para a manipulação dos dados

## datasets de dados
esus = read.csv2(paste0("dados/20210601_dadosesus_df.csv")) #importando dados
do esus df de 01/06/2021
sivep = read.csv2(paste0("dados/20210823_dadososivep.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. Após isso, realizaremos os tratamentos usando todo o *script* de tratamento da aula 3, utilizando o comando *source* para executar todo esse script.

4. Medidas para dados contínuos

Para as variáveis contínuas, vamos utilizar a variável “idade” da base de dados do e-SUS Notifica. Inicialmente iremos utilizar o comando *summary* para compreender as medidas da variável:

```
## Calculando medidas descritivas para dados contínuos
summary(esus$idade) #medidas de posição (quartis, mínimo, máximo e mediana)
```

Ao comandar esse código temos um pequeno sumário de medidas centrais e de posição da variável, incluindo quartis, mediana e média, obtemos o seguinte output:

```
Console Terminal x Jobs x
R 4.1.1 · D:/curso_r_asis/
> ## Calculando medidas descritivas para dados contínuos
> summary(esus$idade) #medidas de posição (quartis, mínimo, máximo e mediana)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   2.00  28.50   37.00   38.94   47.00   91.00
```

Aqui encontramos que o menor valor da nossa amostra para idade é igual a 2 e o maior valor é igual a 91. A média de idade é igual a 38.94 anos, a mediana é igual a 37 e o primeiro quartil é igual a 28.50 e o terceiro quartil é igual a 47. Ainda podemos usar outras funções para calcular a média e a mediana, conforme o código:

```
mean(esus$idade) #média da idade
median(esus$idade) #mediana da idade
```

Ao comandar, temos os seguintes *outputs*, que são iguais ao do comando anterior, porém possuem apenas os valores das medidas que procuramos:

```
Console Terminal x Jobs x
R 4.1.1 · D:/curso_r_asis/
> mean(esus$idade) #média da idade
[1] 38.94314
> median(esus$idade) #mediana da idade
[1] 37
>
```

Para encontrar os quantism utilizamos a função *quantile*, que por *default* divide a amostra em 100 partes iguais, sendo necessário identificar em que posição você quer definir os valores de idade.

```
quantile(esus$idade) #quantis da distribuição da idade
quantile(esus$idade, probs = 0.5)
quantile(esus$idade, probs = 0.9)
quantile(esus$idade, probs = 0.99)
```

Sem essa definição, o comando mostrará os valores máximo e mínimo, e os quartis (posição 25%, 50% e 75%), conforme os *outputs* a seguir:

```
Console Terminal x Jobs x
R 4.1.1 · D:/curso_r_asis/
> quantile(esus$idade) #quantis da distribuição da idade
 0%  25%  50%  75% 100%
 2.0 28.5 37.0 47.0 91.0
> quantile(esus$idade, probs = 0.5)
50%
 37
> quantile(esus$idade, probs = 0.9)
90%
58.2
> quantile(esus$idade, probs = 0.99)
99%
81.12
> |
```

Definimos também os percentis de 90% e 99%, encontrando os valores que, em ordem crescente, se encontram nessa posição.

Para calcular as medidas de dispersão, como a variância e o desvio padrão, também há dois comandos simples:

```
var(esus$idade) #variância da idade
sd(esus$idade) #desvio padrão da idade
```

Nesse caso, estamos buscando compreender o quanto os valores estão dispersos da média, conforme as saídas:

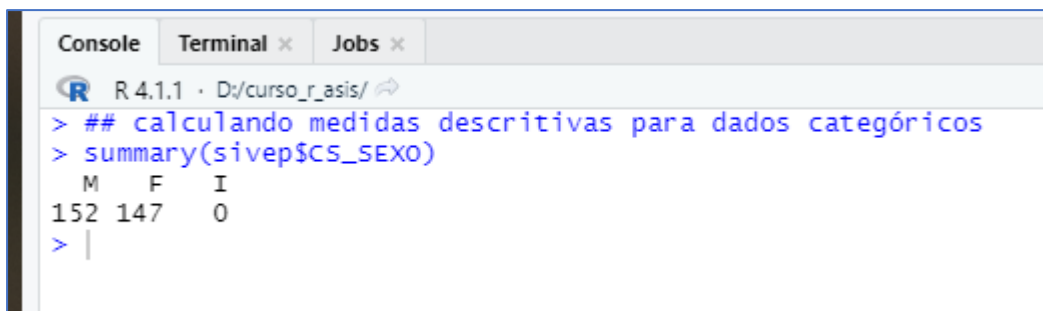
```
Console Terminal x Jobs x
R 4.1.1 · D:/curso_r_asis/
> var(esus$idade) #variância da idade
[1] 226.3424
> sd(esus$idade) #desvio padrão da idade
[1] 15.04468
>
```

5. Medidas para dados categóricos

A função *summary* também pode ser usada para dados categóricos, porém, nesse caso, apresentará apenas as contagens dos dados em categorias. Para essa análise, vamos utilizar a variável *sexo*, do banco de dados do SIVEP-Gripe para compreender esses dados:

```
## calculando medidas descritivas para dados categóricos
summary(sivep$CS_SEXO)
```

Para esses dados, obtemos os seguintes resultados:



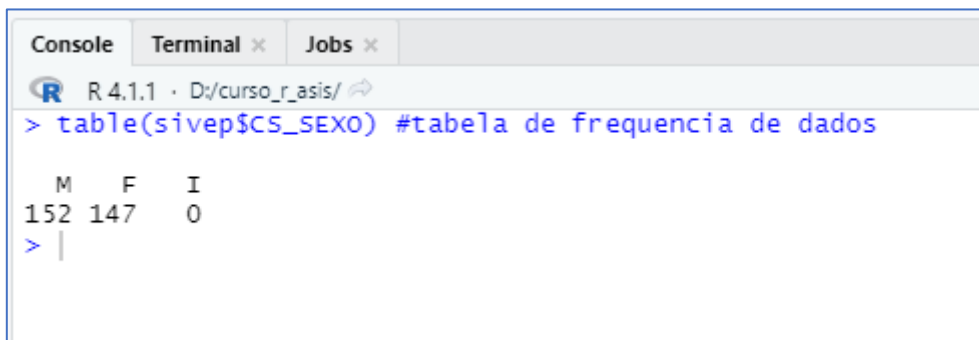
```
R 4.1.1 · D:/curso_r_asis/
> ## calculando medidas descritivas para dados categóricos
> summary(sivep$CS_SEXO)
  M   F   I
152 147   0
> |
```

Nesse caso, mostramos o número de observações no sexo masculino, feminino e ignorado. Foram 152 no primeiro e 147 no segundo. No nosso recorte não houve nenhum caso classificado como ignorado.

Podemos usar a função *table* que permite a criação de uma tabela simples para calcular essas frequências, a saída será semelhante à anterior:

```
table(sivep$CS_SEXO) #tabela de frequencia de dados
```

A saída será semelhante à do comando *summary*. Por meio da contagem de frequências, podemos determinar que a moda dessa variável é o sexo masculino, pois aparece em maior quantidade quando comparada às outras categorias:



```
R 4.1.1 · D:/curso_r_asis/
> table(sivep$CS_SEXO) #tabela de frequencia de dados
  M   F   I
152 147   0
> |
```

Assim, encerramos essa aula! Na próxima aula abordaremos os gráficos que auxiliam a compreender as estatísticas descritivas! Até lá.

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.