

Estatística descritiva com aplicação da linguagem R

Nivelamento em Estatística e Matemática

Apresentação

As atividades desenvolvidas usando o software de planilha são reproduzidos aqui utilizando a linguagem R. Inicialmente será utilizada uma versão web do R para aqueles que não tem o software instalado. A aplicação web está disponível no link

<https://cdn.datacamp.com/dcl-react/standalone-example.html>

O código fonte e demais arquivos referentes a esse material está disponível em:

<https://github.com/extensaoedu/nivelamento2021primeiro>

Medidas descritivas com dados sobre covid no brasil

Nesse exemplo é feita uma importação de um conjunto de dados em formato csv em que as colunas são separadas por “;”. A seguir o código utilizado para importação dos dados, indicando o separador considerado e exibição das 5 primeiras linhas para conferência dos resultados obtidos.

```
dados_covid <- read.csv("covid.csv",sep = ";")
head(dados_covid,n = 5)
```

```
##      Estado 0bitos
## 1      AC    1757
## 2      AL    5433
## 3      AM   13349
## 4      AP    1852
## 5      BA   24312
```

A seguir de forma simples pode-se calcular a média e a variância dos dados

```
(media <- mean(dados_covid$0bitos))
```

```
## [1] 19460.52
```

```
(variancia_amostral <- var(dados_covid$0bitos))
```

```
## [1] 664896304
```

Note que o resultado para a variância obtido difere do resultado obtido utilizando-se a planilha eletrônica. A diferença está no fato que aqui a função de variância implementada por padrão considera o cálculo da variância amostral que considera uma adaptação no denominador para $(n - 1)$ com o objetivo de melhorar as propriedades do estimador. A forma utilizada é dada por:

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{(n - 1)}$$

é simples perceber que para se obter o resultado considerando a variância populacional pode ser feito fazendo:

```
(variancia_populacional <- var(dados_covid$Obitos)*26/27)
```

```
## [1] 640270515
```

e como visto com a utilização da planilha eletrônica é possível fazer as contas da variância passo a passo, calculando os desvios dos valores em relação a média

```
dados_covid$desvios <- dados_covid$Obitos - media
dados_covid
```

```
##      Estado Obitos      desvios
## 1      AC    1757 -17703.51852
## 2      AL    5433 -14027.51852
## 3      AM   13349 -6111.51852
## 4      AP    1852 -17608.51852
## 5      BA   24312  4851.48148
## 6      CE   22753  3292.48148
## 7      DF    9312 -10148.51852
## 8      ES   11561 -7899.51852
## 9      GO   19476    15.48148
## 10     MA    9164 -10296.51852
## 11     MG   47120  27659.48148
## 12     MS    8365 -11095.51852
## 13     MT   12178 -7282.51852
## 14     PA   15624 -3836.51852
## 15     PB    8703 -10757.51852
## 16     PE   17908 -1552.51852
## 17     PI    6662 -12798.51852
## 18     PR   31529  12068.48148
## 19     RJ   56039  36578.48148
## 20     RN    6841 -12619.51852
## 21     RO    6213 -13247.51852
## 22     RR    1756 -17704.51852
## 23     RS   31761  12300.48148
## 24     SC   17064 -2396.51852
## 25     SE    5765 -13695.51852
## 26     SP  129675 110214.48148
## 27     TO    3262 -16198.51852
```

elevando esses valores ao quadrado

```
dados_covid$desv_quadrado <- dados_covid$desvios^2
dados_covid
```

```
##      Estado Obitos      desvios desv_quadrado
## 1      AC    1757 -17703.51852 3.134146e+08
## 2      AL    5433 -14027.51852 1.967713e+08
## 3      AM   13349 -6111.51852 3.735066e+07
## 4      AP    1852 -17608.51852 3.100599e+08
## 5      BA   24312  4851.48148 2.353687e+07
## 6      CE   22753  3292.48148 1.084043e+07
## 7      DF    9312 -10148.51852 1.029924e+08
## 8      ES   11561 -7899.51852 6.240239e+07
## 9      GO   19476    15.48148 2.396763e+02
## 10     MA    9164 -10296.51852 1.060183e+08
## 11     MG   47120  27659.48148 7.650469e+08
```

```
## 12    MS    8365 -11095.51852  1.231105e+08
## 13    MT   12178  -7282.51852  5.303508e+07
## 14    PA   15624  -3836.51852  1.471887e+07
## 15    PB    8703 -10757.51852  1.157242e+08
## 16    PE   17908  -1552.51852  2.410314e+06
## 17    PI    6662 -12798.51852  1.638021e+08
## 18    PR   31529  12068.48148  1.456482e+08
## 19    RJ   56039  36578.48148  1.337985e+09
## 20    RN    6841 -12619.51852  1.592522e+08
## 21    RO    6213 -13247.51852  1.754967e+08
## 22    RR    1756 -17704.51852  3.134500e+08
## 23    RS   31761  12300.48148  1.513018e+08
## 24    SC   17064  -2396.51852  5.743301e+06
## 25    SE    5765 -13695.51852  1.875672e+08
## 26    SP  129675  110214.48148  1.214723e+10
## 27    TO    3262 -16198.51852  2.623920e+08
```

note que quando os números são muito grandes para serem exibidos é utilizada a seguinte notação

$$3.134146e + 08 = 3.134146(10)^8 = 31341460$$

o número é exibido com essa compactação em sua apresentação, mas os valores armazenados e disponíveis para cálculos tem grande precisão.

Finalmente realizando a soma e dividindo pelo número de termos

```
sum(dados_covid$desv_quadrado)/27
```

```
## [1] 640270515
```

que foi o mesmo valor obtido acima para a variância populacional

```
variancia_populacional
```

```
## [1] 640270515
```