DADOS CATEGÓRICOS COM R

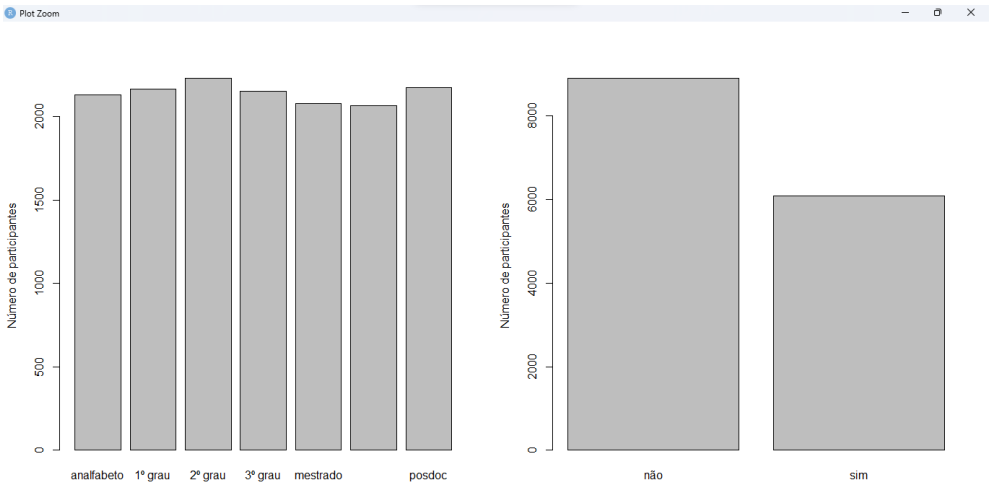
Dra. Marise Miranda
Msc. Eduardo Verri

IDENTIFICANTO TIPOS DE VARIÁVEIS ESTATÍSTICAS

Tipos de variáveis			
Quantitativas (ou Numéricas)		**Qualitativas** (ou Categóricas)	
Contínua	**Discreta**	**Ordinal**	**Nominal**
Valores numéricos que podem ser medidos, mas não contados	Valores numéricos que podem ser contados	Textos ou rótulos que tem uma ordem lógica	Texto ou rótulos que não tem ordem lógica
p.ex. Peso {56,06 kg, 87kg)	p.ex. Número de casos de doenças {0, 1, 2, 3}	p.ex. Tamanho de vestimenta {PP, P, M, G, GG}	p.ex. Profissões {engenheiro, professor, químico}

```
> p <- 200000
> n <- 15000
> pop.escolaridade <- rep(c(0,1,2,3,4,5,6),p)
> set.seed(15)
> escolaridade.temp <- sample(pop.escolaridade, n)
> escolaridade <- factor(escolaridade.temp,
+ levels = c(0,1,2,3,4,5,6),
+ labels = c("analfabeto", "1º grau", "2º grau", "3º grau", "mestrado"
, "doutorado", "posdoc"),
+ ordered = TRUE)

> rm(pop.escolaridade, escolaridade.temp)
> str(escolaridade)
 Ord.factor w/ 7 levels "analfabeto"<"1º grau"<..: 2 4 3 1 5 1 1 5 7 6
...

> summary(escolaridade)
analfabeto   1º grau   2º grau   3º grau   mestrado  doutorado     posdoc
      2132      2164      2229      2154      2080       2067       2174
> table(escolaridade)
escolaridade
analfabeto   1º grau   2º grau   3º grau   mestrado  doutorado     posdoc
      2132      2164      2229      2154      2080       2067       2174

> set.seed(15)
> fumante.n <- rbinom(n, 1, .40)
> fumante.f <- factor(fumante.n,
+ levels = c(0,1),
+ labels = c("não", "sim"),
+ ordered = TRUE)

> str(fumante.f)
 Ord.factor w/ 2 levels "não"<"sim": 2 1 2 2 1 2 2 1 2 2 ...
> summary(fumante.f)
 não  sim
8902 6098
> table(fumante.f)
fumante.f
 não  sim
8902 6098
> str(fumante.n)
 int [1:15000] 1 0 1 1 0 1 1 0 1 1 ...
> mean(fumante.n)
[1] 0.4065333
```

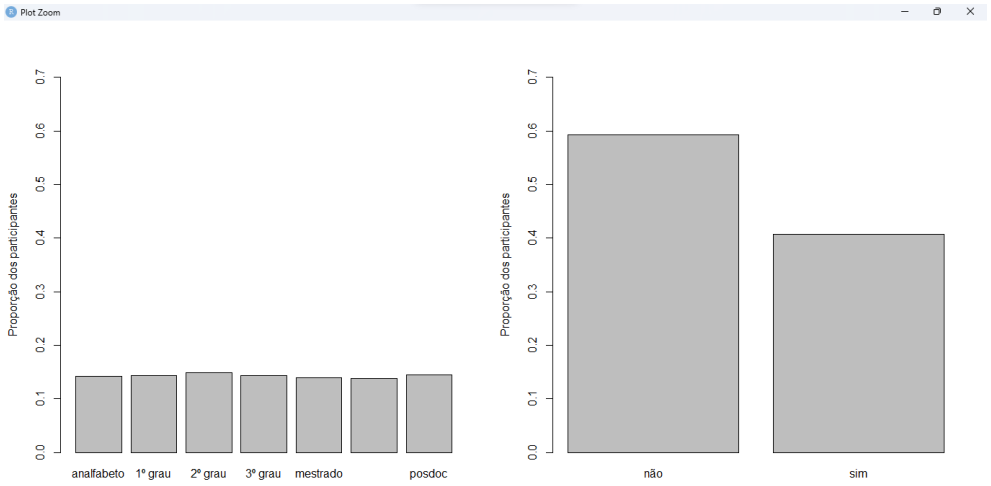
```
> par(mfrow = c(1,2)) #mostra os dois gráficos um do lado do outro
> barplot(table(escolaridade), ylab = "Número de participantes")
> barplot(table(fumante.f), ylab = "Número de participantes")
```



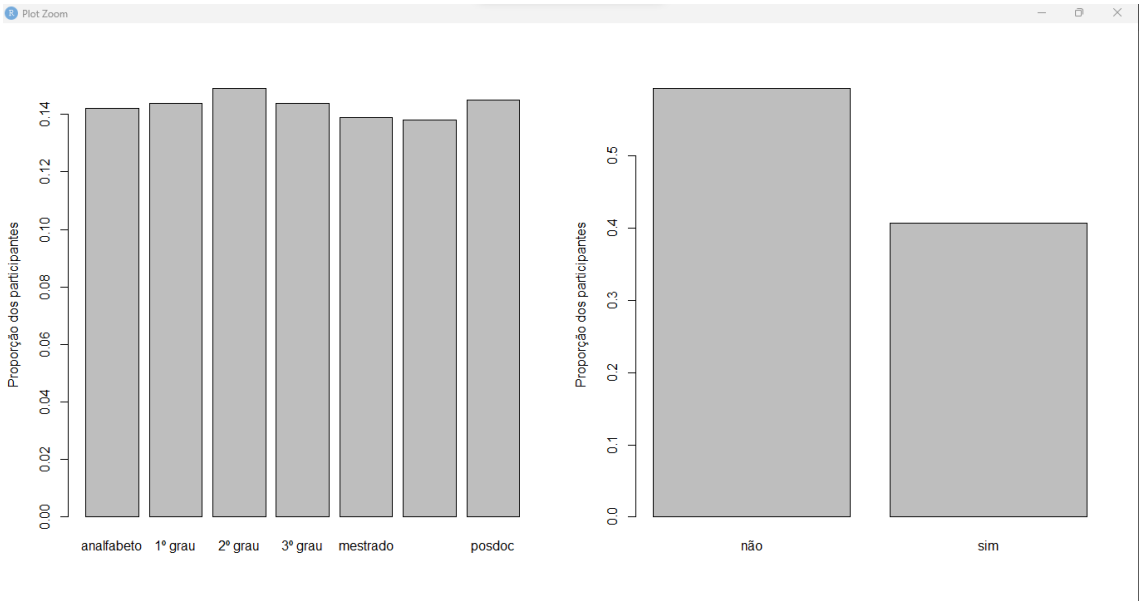

```
> round(prop.table(table(escolaridade)),3)
escolaridade
analfabeto   1º grau   2º grau   3º grau    mestrado   doutorado      posdoc
     0.142     0.144     0.149     0.144       0.139       0.138       0.145
> round(prop.table(table(fumante.f)),3)
fumante.f
  não   sim
0.593 0.407
```

O comando `prop.table()` mostra a proporção de cada categoria. Podemos plotar as proporcionalidades também!

```
> par(mfrow=c(1,2))
> barplot(round(prop.table(table(escolaridade)),3),
+ ylab = "Proporção dos participantes",
+ ylim=c(0,0.7))
> barplot(round(prop.table(table(fumante.f)),3),
+ ylab = "Proporção dos participantes",
+ ylim=c(0,0.7))
```



```
> par(mfrow=c(1,2))
> barplot(round(prop.table(table(escolaridade)),3),
+ ylab = "Proporção dos participantes")
> barplot(round(prop.table(table(fumante.f)),3),
+ ylab = "Proporção dos participantes")
```



Descrevendo a relação entre a escolaridade e o tabagismo

Uma tabela de contingência (também conhecida como tabela de frequência bidirecional ou tabela de frequência com 2 variáveis) descreve a relação entre 2 variáveis categóricas. Cada célula desta tabela corresponde ao número de ocorrências de uma determinada combinação de valores das 2 variáveis.

```
> table(escolaridade, fumante.f)
             fumante.f
escolaridade  não  sim
  analfabeto 1295  837
  1º grau    1268  896
  2º grau    1313  916
  3º grau    1279  875
  mestrado   1282  798
  doutorado  1217  850
  posdoc     1248  926
```

Mostrando proporções totais

```
> tabela1 <- prop.table(table(escolaridade, fumante.f))
> tabela1
             fumante.f
escolaridade        não        sim
  analfabeto 0.08633333 0.05580000
  1º grau    0.08453333 0.05973333
  2º grau    0.08753333 0.06106667
  3º grau    0.08526667 0.05833333
  mestrado   0.08546667 0.05320000
  doutorado  0.08113333 0.05666667
  posdoc     0.08320000 0.06173333
```

```
> tabela2 <- round(prop.table(table(escolaridade, fumante.f)),4)
> tabela2
             fumante.f
escolaridade    não    sim
  analfabeto 0.0863 0.0558
  1º grau    0.0845 0.0597
  2º grau    0.0875 0.0611
  3º grau    0.0853 0.0583
  mestrado   0.0855 0.0532
  doutorado  0.0811 0.0567
  posdoc     0.0832 0.0617

> addmargins(tabela1)
             fumante.f
escolaridade        não        sim        Sum
  analfabeto 0.08633333 0.05580000 0.14213333
  1º grau    0.08453333 0.05973333 0.14426667
  2º grau    0.08753333 0.06106667 0.14860000
  3º grau    0.08526667 0.05833333 0.14360000
  mestrado   0.08546667 0.05320000 0.13866667
  doutorado  0.08113333 0.05666667 0.13780000
  posdoc     0.08320000 0.06173333 0.14493333
  Sum        0.59346667 0.40653333 1.00000000
```

Mostrando proporções de linhas

Aqui, cada célula representa a contagem de indivíduos nesta categoria dividida pelo total da linha:

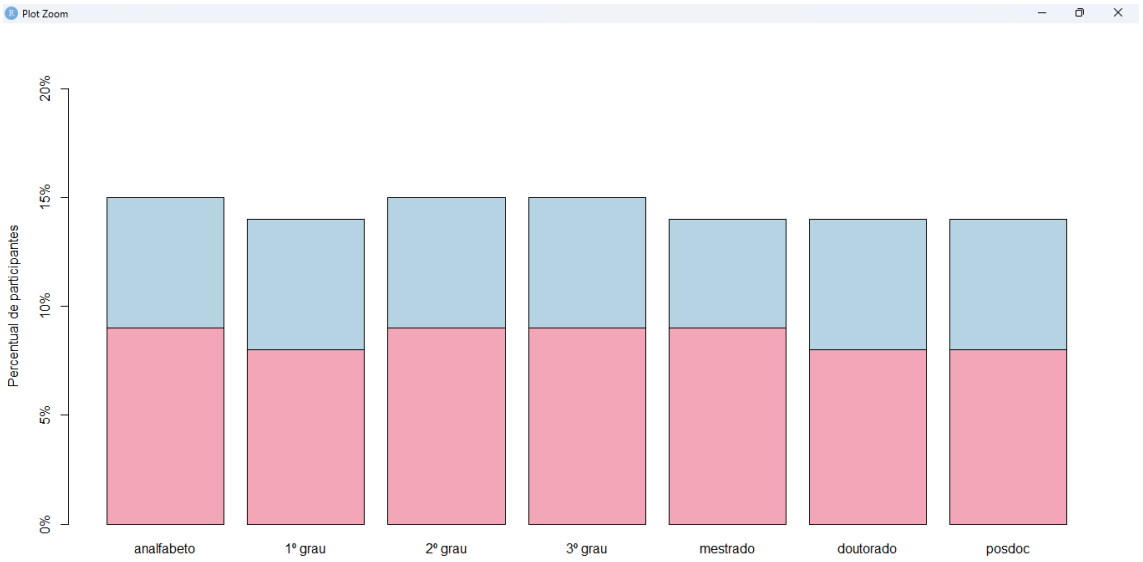
```
> tabela3 <- round(prop.table(table(fumante.f, escolaridade),1),4)
> tabela3
         escolaridade
fumante.f analfabeto 1º grau 2º grau 3º grau mestrado doutorado posdoc
      não     0.1455  0.1424  0.1475  0.1437   0.1440    0.1367 0.1402
      sim     0.1373  0.1469  0.1502  0.1435   0.1309    0.1394 0.1519
> addmargins(tabela3)
         escolaridade
fumante.f analfabeto 1º grau 2º grau 3º grau mestrado doutorado posdoc    Sum
      não     0.1455  0.1424  0.1475  0.1437   0.1440    0.1367 0.1402 1.0000
      sim     0.1373  0.1469  0.1502  0.1435   0.1309    0.1394 0.1519 1.0001
      Sum     0.2828  0.2893  0.2977  0.2872   0.2749    0.2761 0.2921 2.0001
```

Mostrando as proporções das colunas

Aqui, cada célula representa a contagem de indivíduos nesta categoria dividida pelo total da coluna:

```
> tabela4 <- round(prop.table(table(fumante.f, escolaridade),2),4)
> tabela4
         escolaridade
fumante.f analfabeto 1º grau 2º grau 3º grau mestrado doutorado posdoc
      não     0.6074  0.5860  0.5891  0.5938   0.6163    0.5888 0.5741
      sim     0.3926  0.4140  0.4109  0.4062   0.3837    0.4112 0.4259
> addmargins(tabela4)
         escolaridade
fumante.f analfabeto 1º grau 2º grau 3º grau mestrado doutorado posdoc    Sum
      não     0.6074  0.5860  0.5891  0.5938   0.6163    0.5888 0.5741 4.1555
      sim     0.3926  0.4140  0.4109  0.4062   0.3837    0.4112 0.4259 2.8445
      Sum     1.0000  1.0000  1.0000  1.0000   1.0000    1.0000 1.0000 7.0000
```

```
library(scales)

x <- barplot(round(prop.table(table(fumante.f, escolaridade)),2),
             col = rep(c("#F2A6B8", "#B5D3E3")),
             legend = FALSE,
             ylim = c(0, 0.2),
             yaxt = "n",
             ylab = "Percentual de participantes")

yticks = seq(0, 0.2, by = 0.05)
axis(2, at = yticks, lab = percent(yticks))

y <- round(prop.table(table(escolaridade, fumante.f)),2)
```




```
mosaicplot(prop.table(table(escolaridade, fumante.f)),
           col = c("#f2a6b8","#b5d3e3"),
           main = "")
```

A largura de cada coluna deste mosaico corresponde às proporções das diferentes categorias.