

Introdução ao uso do



em Ciências da Saúde

4a - Análise descritiva univariada

Fúlvio Borges Nedel

Departamento de Saúde Pública – SPB
Centro de Ciências da Saúde – CCS
Universidade Federal de Santa Catarina – UFSC

Grups de Recerca d'Amèrica i Àfrica Llatines – GRAAL
<http://graal.uab.cat>

19 de dezembro de 2017

Introdução ao R
4.a Análise
univariada
2/23

Fúlvio Nedel
SPB/UFSC

Introdução

Variáveis
numéricas

Variáveis
categóricas

1 Introdução

2 Variáveis numéricas

3 Variáveis categóricas

- Crie uma nova linha comentada em seu arquivo de sintaxe.

```
# Análise dos dados  
# -----  
### Análise univariada  
...
```

- Inicie com uma sessão vazia e carregue o arquivo de dados
- Verá que ele tem dois objetos,
 - o banco de dados com as variáveis selecionadas antes da transformação e
 - o banco com as variáveis transformadas, para a análise

```
rm(list=ls())  
load('cursoR.RData')  
ls()  
[1] "cursoR" "cursoR2"  
class(cursoR)  
[1] "data.frame"  
class(cursoR2)  
[1] "data.frame"
```

summary(cursorR2)

idade	sexo	imc	imccat
Min. :23.00	Feminino :198	Min. :17.30	normal : 99
1st Qu.:55.75	Masculino:102	1st Qu.:24.22	sobrepeso:116
Median :64.50		Median :27.18	obesidade: 84
Mean :64.05		Mean :27.77	NA's : 1
3rd Qu.:74.00		3rd Qu.:30.48	
Max. :94.00		Max. :46.14	
		NA's :1	
obeso	abep2	grupo	id
sim : 84	A/B : 73	Sim : 13	Length:300
não :215	C :178	Não :285	Class :character
NA's: 1	D/E : 31	NA's: 2	Mode :character
	NA's: 18		

Introdução ao R
4.a Análise
univariada
5/23

Fúlvio Nedel
SPB/UFSC

Introdução

Variáveis
numéricas

Variáveis
categóricas

1 Introdução

2 Variáveis numéricas

3 Variáveis categóricas

Medidas de

- tendência central (média, mediana, moda)
- posição (quantis)
- dispersão (amplitude, desvio-padrão, coeficiente de variação, intervalo interquartilico)
- forma (assimetria e curtose)
- já vimos algumas funções (**mean**, **sd**)

Gráficos

- histograma
- boxplot
- densidade
- polígono de frequência acumulada
- de barras, para variáveis discretas
- de pontos, ramo-e-folhas. . .

```
attach(cursoR2)
```

Amplitude

```
range(imc)
```

```
[1] NA NA
```

```
summary(imc)
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
17.30	24.22	27.18	27.77	30.48	46.14	1

Amplitude

```
range(imc, na.rm = TRUE)
[1] 17.30104 46.13610
max(imc, na.rm = T) - min(imc, na.rm = T)
[1] 28.83506
```

Média

```
mean(imc, na.rm = T)
[1] 27.76622
```

Desvio-padrão e coeficiente de variação

```
sd(imc, na.rm = T)
[1] 4.971806
round(sd(imc, na.rm = T)/mean(imc, na.rm = T)*100, 2)
[1] 17.91
```

Mediana

```
median(imc, na.rm = T)
[1] 27.1809
```


Assimetria

```
e1071::skewness(imc, na.rm = TRUE)
[1] 0.6374804
```

Curtose

```
e1071::kurtosis(imc, na.rm = T)
[1] 0.2573438
```

Percentis (quantis)

```
quantile(imc, na.rm = T)
      0%      25%      50%      75%     100%
17.30104 24.21875 27.18090 30.47595 46.13610

quantile(imc, p = c(.025, .975), na.rm = T)
      2.5%     97.5%
19.79113 38.60288
```

- A moda é uma característica tanto de variáveis numéricas como categóricas
- É o valor mais frequente, ou, em variáveis contínuas, o de maior densidade de frequência.
- Tem pouca relevância em estatística, e **não há, entre as funções básicas do R, uma que a calcule**

- A moda é uma característica tanto de variáveis numéricas como categóricas
- É o valor mais frequente, ou, em variáveis contínuas, o de maior densidade de frequência.
- Tem pouca relevância em estatística, e **não há, entre as funções básicas do R, uma que a calcule**
- Se recordamos que ao ordenar por frequência decrescente os valores de uma distribuição o primeiro será a moda, ela pode ser facilmente encontrada:

```
sort(table(idade), decreasing = T)[1:3]

idade
74 58 64
13 12 11

names(sort(table(idade), decreasing = T)[1])

[1] "74"
```

- O pacote **modeest** tem funções para seu cálculo:

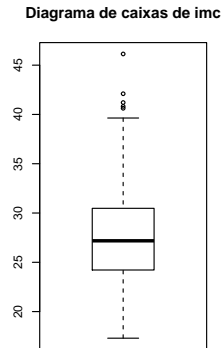
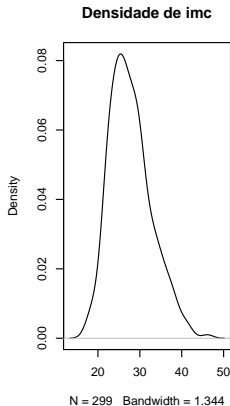
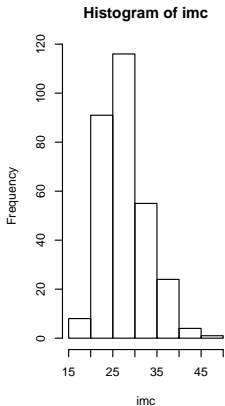
```
modeest::mfv(idade)

[1] 74

modeest::mlv(imc, na.rm = T)

Mode (most likely value): 26.95872
Bickel's modal skewness: 0.0367893
Call: mlv.default(x = imc, na.rm = T)
```

```
par(mfrow = c(1,3))  
hist(imc)  
plot(density(imc, na.rm = T), main='Densidade de imc')  
boxplot(imc, main = 'Diagrama de caixas de imc')
```



Introdução ao R
4.a Análise
univariada
11/23

Fúlvio Nedel
SPB/UFSC

Introdução

Variáveis
numéricas

Variáveis
categóricas

1 Introdução

2 Variáveis numéricas

3 Variáveis categóricas

- frequência absoluta e relativa
- frequências acumuladas
- gráficos de barras
- já vimos algumas funções (`table`, `cumsum`)

Estado nutricional

Com as funções básicas

Frequência absoluta

```
table(imccat)
```

```
imccat
```

normal	sobrepeso	obesidade
99	116	84

Acrescentar o total

```
addmargins(table(imccat))
```

```
imccat
```

normal	sobrepeso	obesidade	Sum
99	116	84	299

Frequência acumulada

```
cumsum(table(imccat))
```

normal	sobrepeso	obesidade
99	215	299

Estado nutricional

Com as funções básicas (cont.)

Frequência relativa

```
prop.table(table(imccat))  
  
imccat  
    normal sobrepeso obesidade  
0.3311037 0.3879599 0.2809365
```

Porcentagem, arredondada para um decimal

```
round(prop.table(table(imccat))*100, 1)  
  
imccat  
    normal sobrepeso obesidade  
    33.1      38.8      28.1
```

Porcentagem acumulada

```
cumsum(round(prop.table(table(imccat))*100, 1))  
  
    normal sobrepeso obesidade  
    33.1      71.9     100.0
```


Introdução ao R
4.a Análise
univariada
15/23

Fúlvio Nedel
SPB/UFSC

Introdução

Variáveis
numéricas

Variáveis
categóricas

Mas...

- um comando pra cada coisa?!

Introdução ao R
4.a Análise
univariada
15/23

Fúlvio Nedel
SPB/UFSC

Introdução

Variáveis
numéricas

Variáveis
categóricas

Mas...

- um comando pra cada coisa?!
- até pras proporções?!!

Mas...

- um comando pra cada coisa?!
- até pras proporções?!!
- até pro total da tabela?!!!

Introdução ao R
4.a Análise
univariada
15/23

Fúlvio Nedel
SPB/UFSC

Introdução

Variáveis
numéricas

Variáveis
categóricas

Mas...

- um comando pra cada coisa?!
- até pras proporções?!!
- até pro total da tabela?!!!

Introdução ao R
4.a Análise
univariada
15/23

Fúlvio Nedel
SPB/UFSC

Introdução

Variáveis
numéricas

Variáveis
categóricas

Mas...

- um comando pra cada coisa?!
- até pras proporções?!!
- até pro total da tabela?!!!

Bom...

- o que pode parecer limitação

Introdução ao R
4.a Análise
univariada
15/23

Fúlvio Nedel
SPB/UFSC

Introdução

Variáveis
numéricas

Variáveis
categóricas

Mas...

- um comando pra cada coisa?!
- até pras proporções?!!
- até pro total da tabela?!!!

Bom...

- o que pode parecer limitação
- **representa versatilidade**

Introdução ao R
4.a Análise
univariada
15/23

Fúlvio Nedel
SPB/UFSC

Introdução

Variáveis
numéricas

Variáveis
categóricas

Mas...

- um comando pra cada coisa?!
- até pras proporções?!!
- até pro total da tabela?!!!

Bom...

- o que pode parecer limitação
- representa versatilidade
- o usuário pode construir sua tabela

Mas...

- um comando pra cada coisa?!
- até pras proporções?!!
- até pro total da tabela?!!!

Bom...

- o que pode parecer limitação
- representa versatilidade
- o usuário pode construir sua tabela
- **ou construir funções com o modelo de tabela que deseja**

Mas...

- um comando pra cada coisa?!
- até pras proporções?!!
- até pro total da tabela?!!!

Bom...

- o que pode parecer limitação
- representa versatilidade
- o usuário pode construir sua tabela
- ou construir funções com o modelo de tabela que deseja
- ou usar funções de outros pacotes

Estado nutricional

Com as funções básicas (cont.)

```
tab <- table(imccat) # freq. absoluta
ptab <- round(prop.table(table(imccat))*100, 1) # porcentagens
cumtab <- cumsum(tab) # acumulada absoluta
cumptab <- cumsum(ptab) # acumulada porcentagens
miolo <- cbind(Freq=tab, '%'=ptab, Freq.acum=cumtab, '%acum'=cumptab)
Total <- c(sum(tab), sum(ptab), sum(tab), sum(ptab))
tab.imccat <- rbind(miolo, Total)
```

tab.imccat				
	Freq	%	Freq.acum	%acum
normal	99	33.1	99	33.1
sobrepeso	116	38.8	215	71.9
obesidade	84	28.1	299	100.0
Total	299	100.0	299	100.0

tab.imccat[,c(1,3,2,4)]				
	Freq	Freq.acum	%	%acum
normal	99	99	33.1	33.1
sobrepeso	116	215	38.8	71.9
obesidade	84	299	28.1	100.0
Total	299	299	100.0	100.0

```
tabua <- function(x, digits = 1){
  tab <- table(x)
  ptab <- round(prop.table(tab)*100, digits)
  cumtab <- cumsum(tab)
  cumptab <- cumsum(ptab)
  miolo <- cbind(Freq=tab, '%'=ptab, Freq.acum=cumtab, '%acum'=cumptab)
  Total <- c(sum(tab), sum(ptab), sum(tab), sum(ptab))
  tabela <- rbind(miolo, Total)
  return(tabela)
}
```

```
tabua(imccat)
```

	Freq	% Freq	acum	%acum
normal	99	33.1	99	33.1
sobrepeso	116	38.8	215	71.9
obesidade	84	28.1	299	100.0
Total	299	100.0	299	100.0

```
tabua(imccat)[,c(1,3,2,4)]
```

	Freq	Freq.acum	%	%acum
normal	99	99	33.1	33.1
sobrepeso	116	215	38.8	71.9
obesidade	84	299	28.1	100.0
Total	299	299	100.0	100.0

Já comentamos a função `describe`, de `Hmisc`

```
Hmisc::describe(imccat)
```

```
imccat : Estado nutricional  
      n missing distinct  
299      1           3
```

Value	normal	sobrepeso	obesidade
Frequency	99	116	84
Proportion	0.331	0.388	0.281

Alguns pacotes em R para estudos epidemiológicos

- há vários pacotes, além do `Hmisc`, para facilitar a análise e apresentação de dados epidemiológicos: `Epi`, `epitools`, `epibasix`, `gmodels`...
- instale e carregue o pacote `epiDisplay`

```
install.packages("epiDisplay", dep=T)
```

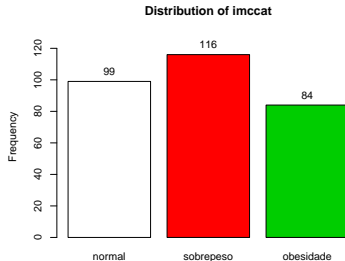
```
library(epiDisplay)
```

A função `tab1`, do pacote `epiDisplay` apresenta a contagem de 'missings' (que `table` não inclui por padrão), as proporções e a frequência acumulada (que `describe` não apresenta), além de incluir um gráfico com os resultados da tabela.

```
tab1(imccat)
tab1(imccat, cum.percent = T)
tab1(imccat, cum.percent = T, missing = F)
```

```
tab1(imccat, missing = F)
```

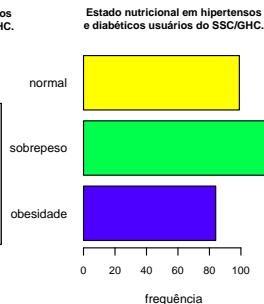
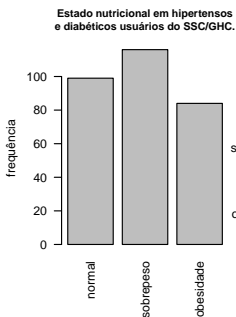
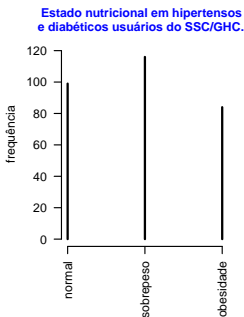
imccat :			
	Frequency	%(NA+)	%(NA-)
normal	99	33.0	33.1
sobrepeso	116	38.7	38.8
obesidade	84	28.0	28.1
NA's	1	0.3	0.0
Total	300	100.0	100.0



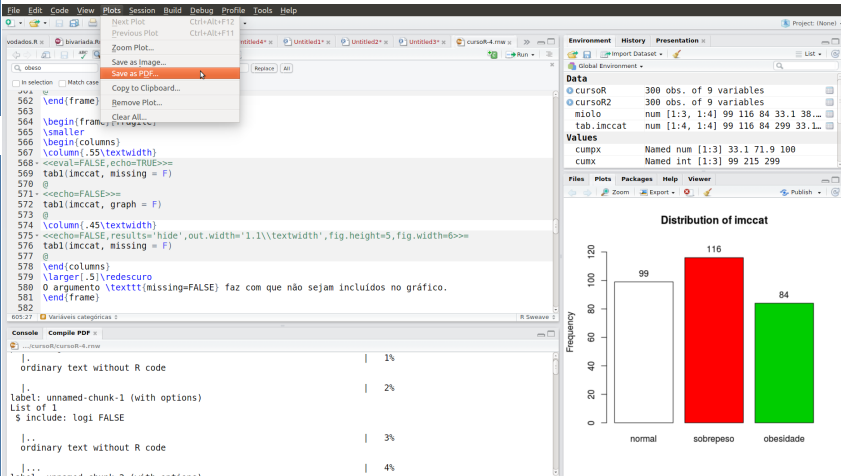
O argumento `missing=FALSE` faz com que não sejam incluídos no gráfico.

```

titulo <- "Estado nutricional em hipertensos
e diabéticos usuários do SSC/GHC."
par(mfrow=c(1,3))
plot(table(imccat), main=titulo, cex.main=1, adj=.6, las=2,
     ylab = 'frequência', xlab = '', col.main="blue")
barplot(table(imccat), main=titulo, cex.main=.9, las=2,
        ylab = 'frequência')
par(mar=c(5,3,4,2))
barplot(rev(table(imccat)), main=titulo, cex.main=.9, cex.axis = .9,
        xlab = 'frequência', horiz = T, las=1, col=topo.colors(3))
    
```



No Rstudio, gráficos podem facilmente ser salvos como figuras, PDF ou copiados para a área de transferência:



The screenshot shows the RStudio interface. The 'Plots' menu is open, and the 'Save as PDF...' option is highlighted. The console shows the execution of R code for creating a bar chart. The environment pane shows the data structure of the 'imccat' variable.

Data Structure:

Variable	Type	Values
cursoR	300 obs. of 9 variables	
cursoR2	300 obs. of 9 variables	
miolo	num [1:3, 1:4]	99 116 84 33.1 38.1...
tab.imccat	num [1:4, 1:4]	99 116 84 299 33.1...

Values:

Variable	Type	Values
cumpx	Named num [1:3]	33.1 71.9 100
cumx	Named int [1:3]	99 215 299

Distribution of imccat

Category	Frequency
normal	99
sobrepeso	116
obesidade	84

No Rstudio, gráficos podem facilmente ser salvos como figuras, PDF ou copiados para a área de transferência:

The screenshot shows the RStudio interface. The 'Plots' menu is open, and 'Save as Image...' is highlighted. The background shows a script editor with R code and a console window. A bar chart titled 'Distribution of imccat' is displayed on the right.

R Code Snippet:

```
378 \column{.55}
379 <<eval=FALSE>
380 tab1(saude, m
381 @
382 <<echo=FALSE>
383 tab1(saude, g
384 @
576 tab1(imccat, missing = F)
577 @
578 \end{columns}
579 \larger{.5}\redescuro
580 O argumento \texttt{missing=FALSE} faz com que não sejam incluídos no gráfico.
581 \end{frame}
582
```

Console Output:

```
1. ordinary text without R code | 1%
2. label: unnamed-chunk-1 (with options) | 2%
3. List of 1 | 3%
4. $ include: logi FALSE | 4%
5. ordinary text without R code | 3%
6. ... | 4%
```

Distribution of imccat

Category	Frequency
normal	99
sobrepeso	116
obesidade	84

Guardar os gráficos em arquivo



Introdução ao R
4.a Análise
univariada
21/23

Fúlvio Nedel
SPB/UFSC

No Rstudio, gráficos podem facilmente ser salvos como figuras, PDF ou copiados para a área de transferência:

Introdução

Variáveis
numéricas

Variáveis
categóricas

The screenshot shows the RStudio environment with a script editor on the left containing R code for a bar chart. The 'Plots' menu is open, and 'Save as Image...' is highlighted. The 'Environment' pane on the right shows the 'Data' environment with variables 'cursoR', 'cursoR2', 'miolo', and 'tab.imccat'. The 'Viewer' pane at the bottom right displays a bar chart titled 'imccat' showing the frequency of BMI categories: normal (99), sobrepeso (116), and obesidade (84).

R Code Snippet:

```

378 \column{.55}
379 <<eval=FALSE>
380 tab1(saude, m
381 @
382 <<echo=FALSE>
383 tab1(saude, g
384 @
576 tab1(imccat, missing = F)
577 @
578 \end{columns}
579 \larger|.5|\redescuro
580 0 argumento \texttt{missing=FALSE} faz com que não sejam incluídos no gráfico.
581 \end{frame}
582
605.27 Variáveis categóricas 3

```

Bar Chart Data:

Categoria	Frequência
normal	99
sobrepeso	116
obesidade	84

```
ls()
[1] "cumptab"      "cumtab"      "cursoR"      "cursoR2"
[5] "miolo"        "ptab"        "tab"         "tab.imccat"
[9] "tabua"        "titulo"      "Total"

save(cursoR, cursoR2, tabua, file='cursoR.RData')
detach(cursoR2)
rm(list=ls())
ls()
character(0)
```

TAREFA

- Crie uma função para o cálculo do coeficiente de variação
 - Com essa função, apresente o coeficiente de variação de `idade`
- Desenhe um gráfico das frequências de `abep2`