

Introdução ao R
5. Análise
bivariada
1/21

Fúlvio Nedel
SPB/UFSC

Introdução

Variável
dependente é
numérica

Variável
independente é
numérica

Variável
independente é
categórica

Variável
dependente é
categórica

Introdução ao uso do em Ciências da Saúde

5 - Análise de dados bivariada

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
5. Análise
bivariada
2/21

Fúlvio Nedel
SPB/UFSC

Introdução

Variável
dependente é
numérica

Variável
independente é
numérica

Variável
independente é
categórica

Variável
dependente é
categórica

1 Introdução

2 Variável dependente é numérica

- Variável independente é numérica
- Variável independente é categórica

3 Variável dependente é categórica

Introdução ao R
5. Análise
bivariada
3/21

Fúlvio Nedel
SPB/UFSC

Introdução

Variável
dependente é
numérica

Variável
independente é
numérica

Variável
independente é
categórica

Variável
dependente é
categórica

- A análise continua com uma descrição do comportamento da variável dependente de acordo com as variáveis independentes
- Crie uma nova linha comentada em seu arquivo de sintaxe. Algo como:

```
### Análise bivariada
```

```
...
```

- Inicie com uma sessão vazia, carregue o arquivo de dados e o pacote `epiDisplay`, e "attache" o banco `cursoR2`

```
rm(list=ls())  
load('cursoR.RData')  
ls()  
[1] "cursoR" "cursoR2" "tabua"  
library(epiDisplay)  
attach(cursoR2)
```

Objetivos

Variável dependente

- 1 Descrever o Índice de Massa Corporal (IMC) e analisar fatores associados à sua média.
- 2 Descrever a frequência de categorias do estado nutricional e analisar fatores possivelmente associados à obesidade.

Variáveis independentes

- 1 sexo
- 2 idade
- 3 faixa etária
- 4 condição socioeconômica
- 5 participação em grupos de promoção da saúde

Introdução ao R
5. Análise
bivariada
5/21

Fúlvio Nedel
SPB/UFSC

Introdução

Variável
dependente é
numérica

Variável
independente é
numérica

Variável
independente é
categórica

Variável
dependente é
categórica

1 Introdução

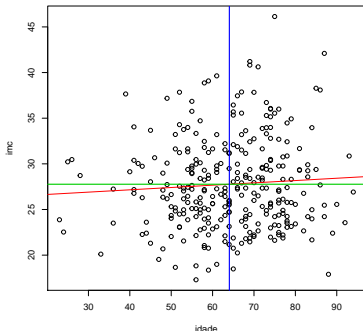
2 Variável dependente é numérica

- Variável independente é numérica
- Variável independente é categórica

3 Variável dependente é categórica

- correlação
- regressão linear simples (se normal)

```
plot(idade, imc)
# plot(imc ~ idade)
abline(lm(imc ~ idade), col = 2)
abline(v = mean(idade),
       h = mean(imc, na.rm = T),
       col = c(3,4), lwd = 2)
```



?cor.test

?lm

```
cor.test(idade, imc, method = 'kendall')
```

Kendall's rank correlation tau

data: idade and imc

z = 0.71744, p-value = 0.4731

alternative hypothesis: true tau is not equal to 0

sample estimates:

tau

0.02813146

```
summary(lm(imc ~ idade))
```

Call:

```
lm(formula = imc ~ idade)
```

Residuals:

Min	1Q	Median	3Q	Max
-10.4715	-3.6304	-0.5756	2.8711	18.0931

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	26.14111	1.46285	17.870	<2e-16 ***
idade	0.02536	0.02238	1.133	0.258

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.969 on 297 degrees of freedom
(1 observation deleted due to missingness)

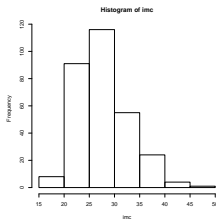
Multiple R-squared: 0.004304, Adjusted R-squared: 0.000951

F-statistic: 1.284 on 1 and 297 DF, p-value: 0.2581

Teste de Shapiro-Wilk para normalidade

H_0 : a variável tem distribuição normal

```
hist(imc)
```

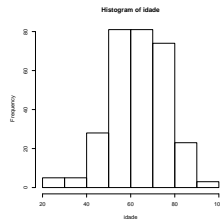


```
shapiro.test(imc)
```

Shapiro-Wilk normality test

```
data: imc
W = 0.97272, p-value = 1.864e-05
# shapiro.qgnorm(imc)
```

```
hist(idade)
```



```
shapiro.test(idade)
```

Shapiro-Wilk normality test

```
data: idade
W = 0.98656, p-value = 0.006751
# shapiro.qgnorm(idade)
```

- correlação
- regressão linear simples (se normal)

?cor.test

?lm

```
cor.test(idade, imc, method = 'kendall')
```

Kendall's rank correlation tau

```
data: idade and imc
```

```
z = 0.71744, p-value = 0.4731
```

```
alternative hypothesis: true tau is not equal to 0
```

```
sample estimates:
```

```
tau
```

```
0.02813146
```


- correlação
- regressão linear simples (se normal)

?cor.test

?lm

```
summary(lm(imc ~ idade))
```

Call:

```
lm(formula = imc ~ idade)
```

Residuals:

Min	1Q	Median	3Q	Max
-10.4715	-3.6304	-0.5756	2.8711	18.0931

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	26.14111	1.46285	17.870	<2e-16 ***
idade	0.02536	0.02238	1.133	0.258

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.969 on 297 degrees of freedom

(1 observation deleted due to missingness)

Multiple R-squared: 0.004304, Adjusted R-squared: 0.000951

F-statistic: 1.284 on 1 and 297 DF, p-value: 0.2581

A função `tapply`

```
library(e1071)
(media = tapply(imc, sexo, mean, na.rm = T))
Feminino Masculino
27.50779 28.26534

(dp = tapply(imc, sexo, sd, na.rm = T))
Feminino Masculino
4.650701 5.529724

(assim = tapply(imc, sexo, skewness, na.rm = T))
Feminino Masculino
0.5069652 0.6976254

(curto = tapply(imc, sexo, kurtosis, na.rm = T))
Feminino Masculino
0.06725562 0.06663441

(mediana = tapply(imc, sexo, median, na.rm = T))
Feminino Masculino
26.95312 27.44598

(p2575 = cbind(P25 = tapply(imc, sexo, quantile, probs = .25, na.rm = T),
                P75 = tapply(imc, sexo, quantile, probs = .75, na.rm = T)))
                P25 P75
Feminino 24.21875 30.36735
Masculino 23.96126 31.54152

(iiq = tapply(imc, sexo, IQR, na.rm = T))
Feminino Masculino
6.148597 7.580262
```

Análise por estratos

- A função `tapply` permite a execução de uma função sobre uma variável em grupos separados de outra.
- Veja a família `apply`:
`?apply`, `?tapply`,
`?sapply`, `?mapply`
 ...

Introdução ao R 5. Análise bivariada 11/21

Fúlvio Nedel
SPB/UFSC

Introdução

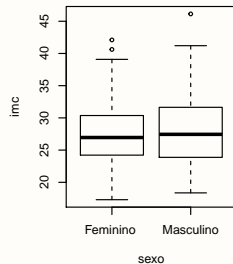
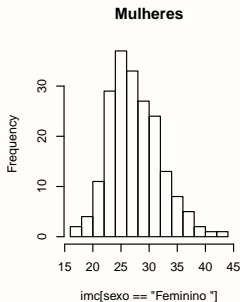
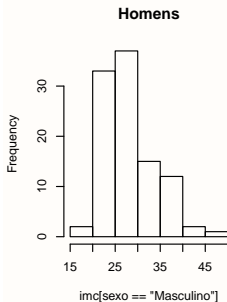
Variável
dependente é
numérica

Variável
independente é
numérica

Variável
independente é
categórica

Variável
dependente é
categórica

```
par(mfrow = c(1,3))
hist(imc[sexo == 'Masculino'], main = 'Homens')
hist(imc[sexo == 'Feminino'], main = 'Mulheres')
plot(imc ~ sexo, ylab = 'imc')
```



```
tabela = round(cbind(media, dp, assimetria = assim, curtose = curto,
                      mediana, p2575, iiq), 2)
print(xtable::xtable(tabela), size="\\scriptsize",
      format.args = list(decimal.mark = ','))
```

	media	dp	assimetria	curtose	mediana	P25	P75	iiq
Feminino	27,51	4,65	0,51	0,07	26,95	24,22	30,37	6,15
Masculino	28,27	5,53	0,70	0,07	27,45	23,96	31,54	7,58

Introdução ao R

5. Análise biviária 12/21

Fúlvio Nedel
SPB/UFSC

Introdução

Variável dependente é numérica

Variável independente é numérica

Variável independente é categórica

Variável dependente é categórica

- Comparação de médias:
testes paramétricos
- Comparação de medianas:
testes não paramétricos

```
t.test(imc ~ sexo)
```

Welch Two Sample t-test

data: imc by sexo

t = -1.1837, df = 176.34, p-value = 0.2381

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-2.0205603 0.5054524

sample estimates:

mean in group Feminino mean in group Masculino

27.50779

28.26534

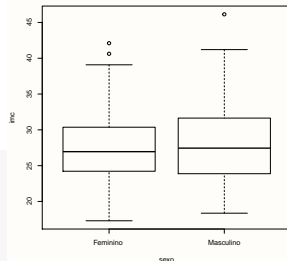
```
kruskal.test(imc, sexo)
```

Kruskal-Wallis rank sum test

data: imc and sexo

Kruskal-Wallis chi-squared = 0.48678, df = 1, p-value = 0.4854

```
plot(imc ~ sexo)
```



Introdução ao R

5. Análise biviária

13/21

Fúlvio Nedel
SPB/UFSC

Introdução

Variável dependente é numérica

Variável independente é numérica

Variável independente é categórica

Variável dependente é categórica

ANOVA

```
bartlett.test(imc ~ abep2)
```

Bartlett test of homogeneity of variances

data: imc by abep2

Bartlett's K-squared = 1.3948, df = 2, p-value = 0.4979

```
anova(aov(imc ~ abep2))
```

Analysis of Variance Table

Response: imc

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
abep2	2	0.4	0.1894	0.0074	0.9926
Residuals	278	7089.2	25.5008		

```
pairwise.wilcox.test(imc, abep2)
```

Pairwise comparisons using Wilcoxon rank sum test

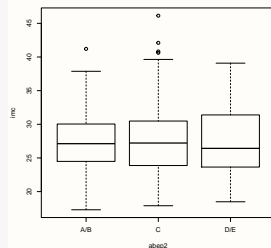
data: imc and abep2

	A/B	C
C	1	-
D/E	1	1

P value adjustment method: holm

```
?pairwise.t.test
```

```
plot(imc ~ abep2)
```



1 Introdução

2 Variável dependente é numérica

- Variável independente é numérica
- Variável independente é categórica

3 Variável dependente é categórica

Já vimos como descrever a associação entre uma variável numérica e uma categórica

```
round(tapply(idade, obeso, mean, na.rm = T), 2)
```

```
sim não  
64.02 64.11
```

```
round(tapply(idade, obeso, sd, na.rm = T), 2)
```

```
sim não  
12.97 12.85
```

```
round(tapply(idade, obeso, median, na.rm = T), 2)
```

```
sim não  
65.5 64.0
```

```
round(tapply(idade, obeso, IQR, na.rm = T), 2)
```

```
sim não  
18.25 18.00
```

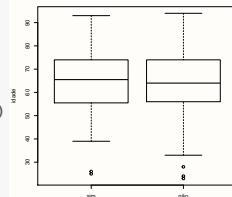
```
kruskal.test(idade, obeso)
```

Kruskal-Wallis rank sum test

data: idade and obeso

Kruskal-Wallis chi-squared = 0.0104, df = 1, p-value = 0.9188

```
plot(obeso, idade,  
      ylab = 'idade')
```



Var. categórica × var. categórica

Obesidade ~ sexo



Introdução ao R
5. Análise
bivariada
16/21

Fúlvio Nedel
SPB/UFSC

Introdução

Variável
dependente é
numérica

Variável
independente é
numérica

Variável
independente é
categórica

Variável
dependente é
categórica

```
library(gmodels)
CrossTable(sexo, obeso,
            prop.chisq = F,
            chisq = T, fisher = T)
```

Cell Contents

		N
N / Row Total		
N / Col Total		
N / Table Total		

Total Observations in Table: 299

sexo	obeso		Row Total
	sim	não	
Feminino	54	143	197
	0.274	0.726	0.659
	0.643	0.665	
	0.181	0.478	
Masculino	30	72	102
	0.294	0.706	0.341
	0.357	0.335	
	0.100	0.241	
Column Total	84	215	299
	0.281	0.719	

Var. categórica × var. categórica

Obesidade ~ sexo



Introdução ao R
5. Análise
bivariada
16/21

Fúlvio Nedel
SPB/UFSC

Introdução

Variável
dependente é
numérica

Variável
independente é
numérica

Variável
independente é
categórica

Variável
dependente é
categórica

```
library(gmodels)
CrossTable(sexo, obeso,
            prop.chisq = F,
            chisq = T, fisher = T)
```

Cell Contents

		N
	N / Row Total	
	N / Col Total	
	N / Table Total	

sexo	obeso		Row Total
	sim	não	
Feminino	54	143	197
	0.274	0.726	0.659
	0.643	0.665	
	0.181	0.478	
Masculino	30	72	102
	0.294	0.706	0.341
	0.357	0.335	
	0.100	0.241	
Column Total	84	215	299
	0.281	0.719	

Total Observations in Table: 299

Statistics for All Table Factors

Pearson's Chi-squared test

Chi^2 = 0.1331494 d.f. = 1 p = 0.7151887

Pearson's Chi-squared test with Yates' continuity correction

Chi^2 = 0.05253029 d.f. = 1 p = 0.8187175

Var. categórica × var. categórica

Obesidade ~ sexo



Introdução ao R
5. Análise
bivariada
16/21

Fúlvio Nedel
SPB/UFSC

Introdução

Variável
dependente é
numérica

Variável
independente é
numérica

Variável
independente é
categórica

Variável
dependente é
categórica

```
library(gmodels)
CrossTable(sexo, obeso,
            prop.chisq = F,
            chisq = T, fisher = T)
```

Cell Contents

		N
N / Row Total		
N / Col Total		
N / Table Total		

sexo	obeso		Row Total
	sim	não	
Feminino	54	143	197
	0.274	0.726	0.659
	0.643	0.665	
	0.181	0.478	
Masculino	30	72	102
	0.294	0.706	0.341
	0.357	0.335	
	0.100	0.241	
Column Total	84	215	299
	0.281	0.719	

Total Observations in Table: 299

Statistics for All Table Factors

Fisher's Exact Test for Count Data

Sample estimate odds ratio: 0.9065962

Alternative hypothesis: true odds ratio is not equal to 1

p = 0.786248

95% confidence interval: 0.5186835 1.601355

Alternative hypothesis: true odds ratio is less than 1

p = 0.407084

95% confidence interval: 0 1.467891

Alternative hypothesis: true odds ratio is greater than 1

p = 0.6932558

95% confidence interval: 0.5637979 Inf

```
tabpct(sexo, obeso)
```

Original table

	obeso		
sexo	sim	não	Total
Feminino	54	143	197
Masculino	30	72	102
Total	84	215	299

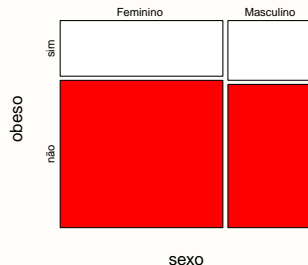
Row percent

	obeso		
sexo	sim	não	Total
Feminino	54 (27.4)	143 (72.6)	197 (100)
Masculino	30 (29.4)	72 (70.6)	102 (100)

Column percent

	obeso			
sexo	sim	%	não	%
Feminino	54	(64.3)	143	(66.5)
Masculino	30	(35.7)	72	(33.5)
Total	84	(100)	215	(100)

Distribution of obeso by sexo



$$RP_{m/f} = \frac{29,4}{27,4} = 1,073$$

$$OR_{m/f} = \frac{29,4/70,6}{27,4/72,6} = \frac{0,416}{0,377} = 1,103$$

```
?chisq.test(sexo, obeso)
```

```
?fisher.test(sexo, obeso)
```

```
Epi::twoby2(sexo, obeso)
```

```
2 by 2 table analysis:
```

```
-----
```

```
Outcome      : sim
```

```
Comparing    : Feminino  vs. Masculino
```

	sim	não	P(sim)	95% conf. interval
Feminino	54	143	0.2741	0.2164 0.3406
Masculino	30	72	0.2941	0.2139 0.3895

		95% conf. interval
Relative Risk:	0.9320	0.6393 1.3586
Sample Odds Ratio:	0.9063	0.5342 1.5376
Conditional MLE Odds Ratio:	0.9066	0.5187 1.6014
Probability difference:	-0.0200	-0.1307 0.0835

```
Exact P-value: 0.7862
```

```
Asymptotic P-value: 0.7152
```

```
-----
```

Introdução ao R
5. Análise
bivariada
19/21

Fúlvio Nedel
SPB/UFSC

Introdução

Variável
dependente é
numérica

Variável
independente é
numérica

Variável
independente é
categórica

Variável
dependente é
categórica

```
epiR::epi.2by2(table(sexo,obeso))
```

	Outcome +	Outcome -	Total	Inc risk *
Exposed +	54	143	197	27.4
Exposed -	30	72	102	29.4
Total	84	215	299	28.1

	Odds
Exposed +	0.378
Exposed -	0.417
Total	0.391

Point estimates and 95 % CIs:

Inc risk ratio (W)	0.93 (0.64, 1.36)
Odds ratio (W)	0.91 (0.53, 1.54)
Attrib risk (W) *	-2.00 (-12.82, 8.82)
Attrib risk in population (W) *	-1.32 (-11.52, 8.89)
Attrib fraction in exposed (%)	-7.30 (-56.41, 26.39)
Attrib fraction in population (%)	-4.69 (-33.40, 17.84)

X2 test statistic: 0.133 p-value: 0.715

W: Wald confidence limits

* Cases per 100 population units

```
Rcoisas::bolero(sexo, obeso)
```

```
=====
```

```
Tabela 2 por 2
bolero(independente, dependente, dec=2, dnn)
```

```
Var. dependente : obeso = sim
```

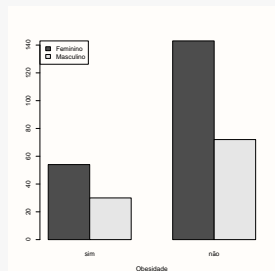
```
Var. independente: sexo = Feminino
```

```
barplot(table(sexo,obeso), beside = T,
        xlab = 'Obesidade',
        legend.text = levels(sexo),
        args.legend = list(x="topleft"))
```

	obeso		
sexo	sim	não	Sum
Feminino	54	143	197
Masculino	30	72	102
Sum	84	215	299

Proporções (%)

	obeso	
sexo	sim	não
Feminino	27.4	72.6
Masculino	29.4	70.6



Razão de Probabilidades: 0.93 ; IC95% (assintótico): 0.64 1.36

IC95% (exato) : 0.60 1.36

Razão de Odds : 0.91 ; IC95% (exato) : 0.52 1.60

Valor-p: Pearson, Yates: 0.819 ; Fisher: 0.786

```
=====
```

Introdução ao R
5. Análise
bivariada
21/21

Fúlvio Nedel
SPB/UFSC

Introdução

Variável
dependente é
numérica

Variável
independente é
numérica

Variável
independente é
categórica

Variável
dependente é
categórica

Descrição bivariada

- Descreva a relação entre o peso e a altura (estão no banco `cursoR`)
- Descreva a relação entre o IMC e a participação em grupos
- Descreva a relação entre o IMC e a condição socioeconômica em sete categorias (variável `abepcls`, em `cursoR`)
- Descreva a relação entre a obesidade e a participação em grupos de promoção à saúde
- Descreva a relação entre a obesidade e a condição socioeconômica