Métodos Estatísticos em Pesquisa Científica

Aula 02 - Parte 2/2

Paulo Justiniano Ribeiro Jr

Departamento de Estatística
Setor de Ciências Exatas
Transversais - PRPPG
Universidade Federal do Paraná

27 de março, 2024

Análises bivariadas

- Busca por associação
- Medidas de associação: coeficientes de correlação
- Associação *vs* causalidade
- Estudos experimentais ou observacionais
- Variáveis com valores pré definidos ou observados

Questionário do curso

No questionário do curso pode-se ver os vários tipos de relações bivariadas:

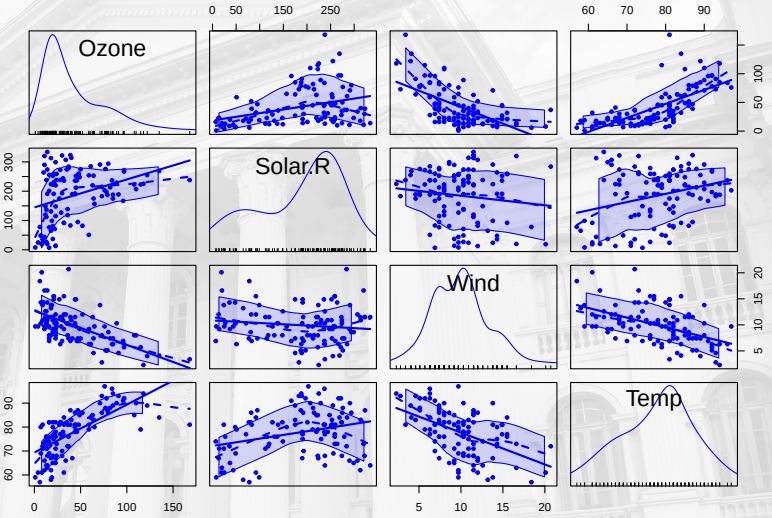
- Numérica *versus* numérica
- Categórica *versus* numérica
- Categórica *versus* categórica

Em categóricas vamos considerar qualitativas (nominal/ordinal) e por vezes discretas (com poucos possíveis valores).

Numérica *versus* numérica

Diagrama de dispersão:

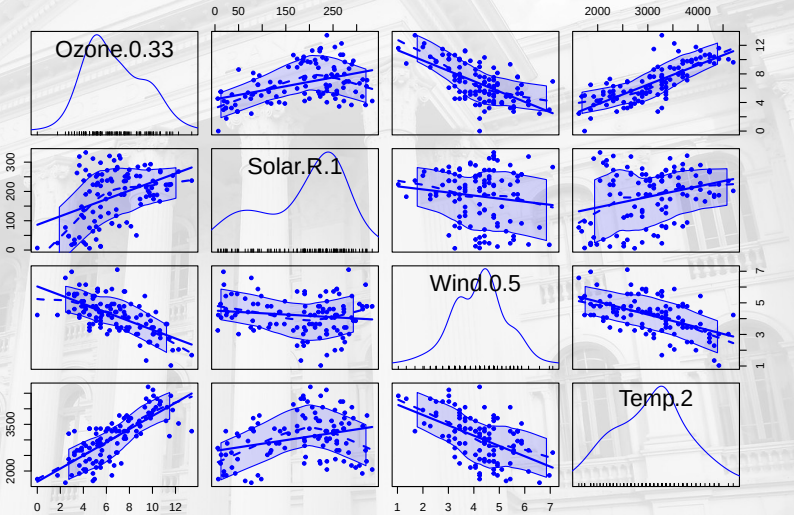
- Há relação?
- Positiva ou negativa?
- Monótona?
- Forte ou fraca?
- Linear?
- Dados atípicos?
- Transformação?



Qualidade do ar (com dados transformados)

Coeficientes de correlação ρ (linear de Pearson, Spearman, Kendall, ...)

- Há relação? ($\rho \neq 0$)
- Positiva ou negativa? ($\rho > 0$ ou $\rho < 0$)
- Monótona?
- Forte ou fraca? ($|\rho| \to 1$ ou $|\rho| \to 0$)
- Linear?
- Dados atípicos?



Qualidade do ar: coeficientes de correlação

Correlações de Pearson

	Ozone	Solar.R	Wind	Temp
Ozone	1,00	0,35	-0,60	0,70
Solar.R	0,35	1,00	-0,06	0,28
Wind	-0,60	-0,06	1,00	-0,46
Temp	0,70	0,28	-0,46	1,00

Tabela 1. Dados originais

	Ozone	Solar.R	Wind	Temp
Ozone	1,00	0,42	-0,61	0,76
Solar.R	0,42	1,00	-0,05	0,27
Wind	-0,61	-0,05	1,00	-0,46
Temp	0,76	0,27	-0,46	1,00

Tabela 2. Dados transformados

Correlações de Spearman

	Ozone	Solar.R	Wind	Temp
Ozone	1,00	0,35	-0,59	0,77
Solar.R	0,35	1,00	-0,00	0,21
Wind	-0,59	-0,00	1,00	-0,45
Temp	0,77	0,21	-0,45	1,00

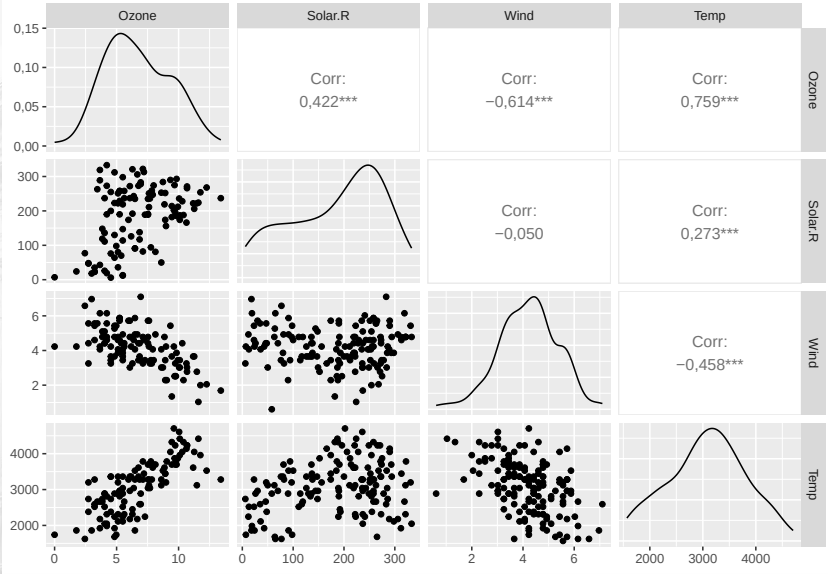
Tabela 3. Dados originais

	Ozone	Solar.R	Wind	Temp
Ozone	1,00	0,35	-0,59	0,77
Solar.R	0,35	1,00	-0,00	0,21
Wind	-0,59	-0,00	1,00	-0,45
Temp	0,77	0,21	-0,45	1,00

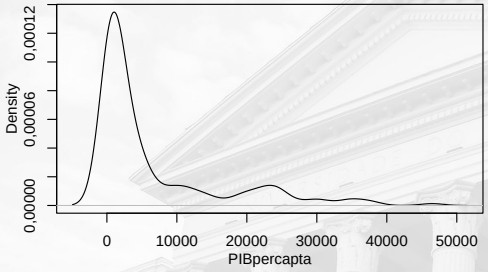
Tabela 4. Dados transformados

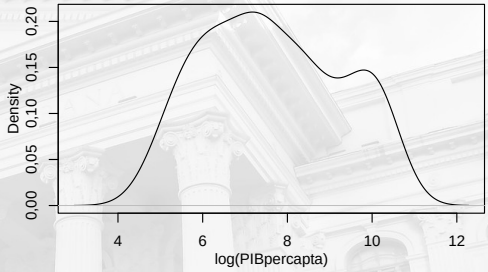
Qualidade do ar: outra visualização

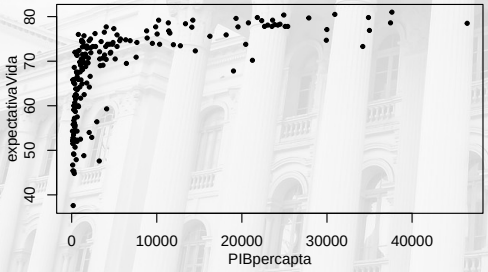
- Há relação? ($\rho \neq 0$)
- Positiva ou negativa? ($\rho > 0$ ou $\rho < 0$)
- Monótona?
- Forte ou fraca? ($|\rho| \to 1$ ou $|\rho| \to 0$)
- Linear?
- Dados atípicos?

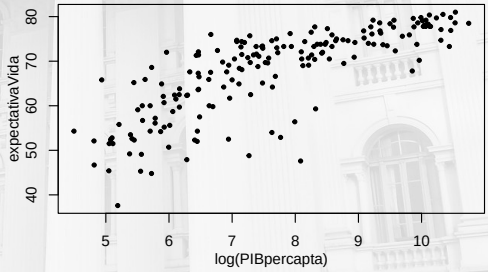


Revisitando dados gapminger - 2000

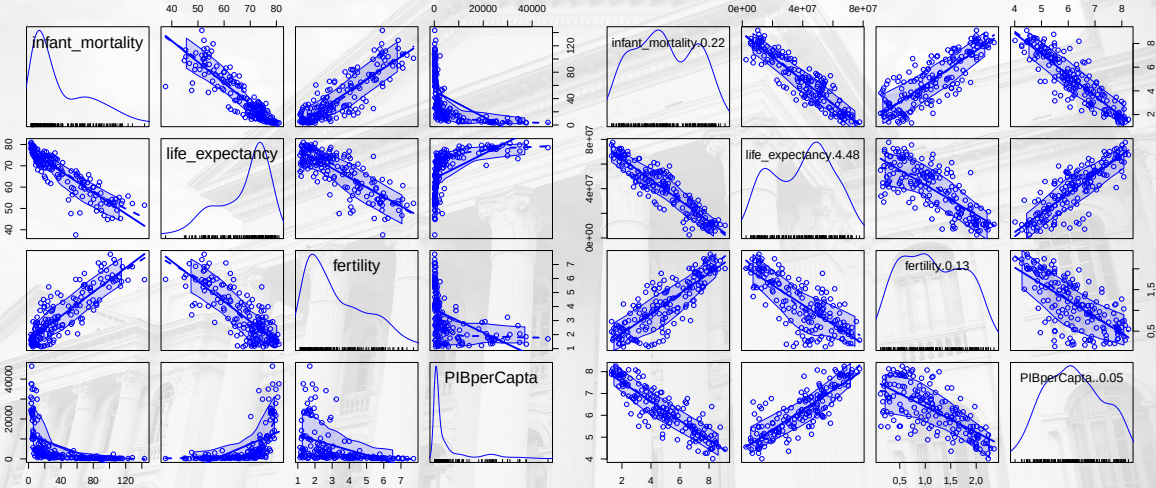








Revisitando dados gapminger - 2000



Potencial efeito de “covariáveis” (continente, região)

Gapminder: coeficientes de correlação

Correlações de Pearson

	infant_mortality	life_expectancy	fertility	PIBperCapta
infant_mortality	1,00	-0,89	0,86	-0,53
life_expectancy	-0,89	1,00	-0,80	0,57
fertility	0,86	-0,80	1,00	-0,47
PIBperCapta	-0,53	0,57	-0,47	1,00

Tabela 5. Dados originais

	infant_mortality	life_expectancy	fertility	PIBperCapta
infant_mortality	1,00	-0,92	0,84	-0,88
life_expectancy	-0,92	1,00	-0,76	0,84
fertility	0,84	-0,76	1,00	-0,69
PIBperCapta	-0,88	0,84	-0,69	1,00

Tabela 6. Dados transformados

Gapminder: coeficientes de correlação

Correlações de Spearman

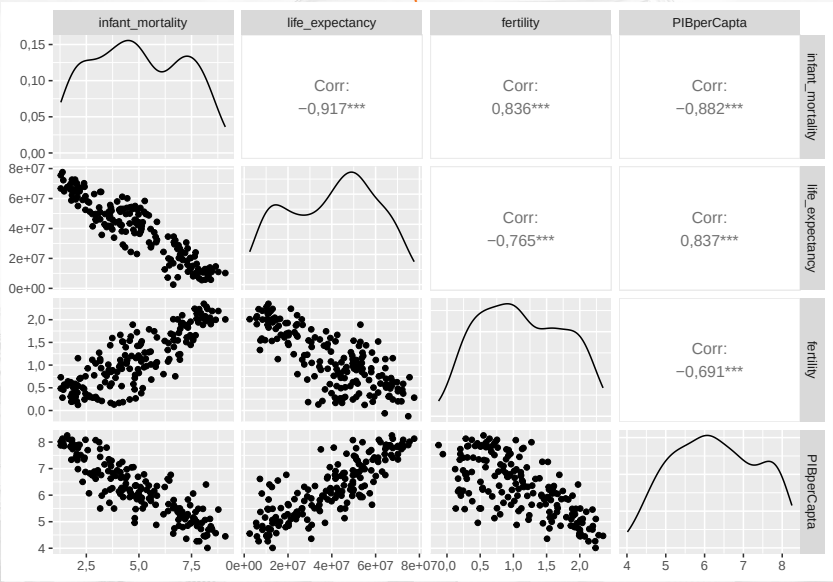
	infant_mortality	life_expectancy	fertility	PIBperCapta
infant_mortality	1,00	-0,90	0,83	-0,88
life_expectancy	-0,90	1,00	-0,73	0,84
fertility	0,83	-0,73	1,00	-0,68
PIBperCapta	-0,88	0,84	-0,68	1,00

Tabela 7. Dados originais

	infant_mortality	life_expectancy	fertility	PIBperCapta
infant_mortality	1,00	-0,90	0,83	-0,88
life_expectancy	-0,90	1,00	-0,73	0,84
fertility	0,83	-0,73	1,00	-0,68
PIBperCapta	-0,88	0,84	-0,68	1,00

Tabela 8. Dados transformados

Dados gapminder: outra visualização



Correlações ... devagar com o andor

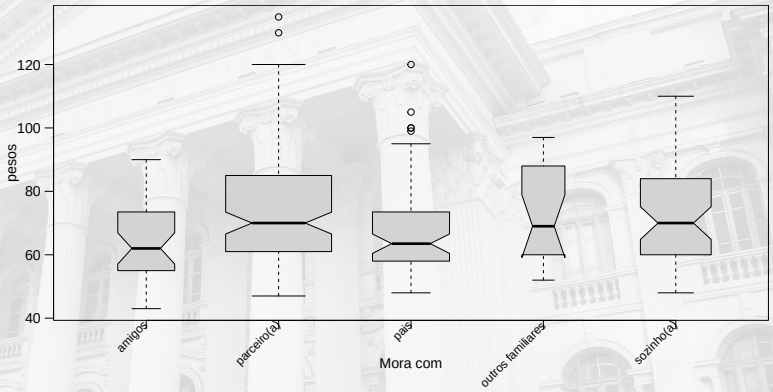
- medem associação
- associação não necessariamente implica em causalidade
- confundimento

Página com correlações espúrias

Numérica *versus* Categórica

Comparação de *boxplot*'s (e outros gráficos já vistos)

- "Nível"
- variabilidade
- distribuição
- dados atípicos
- tamanho dos grupos



Categórica *versus* categórica: Trabalha *vs* Origem

	RMC	Interior	Outro Estado
Não	65	76	37
Sim	38	35	43

	RMC	Interior	Outro Estado	Sum
Não	65	76	37	178
Sim	38	35	43	116
Sum	103	111	80	294

- tipicamente representada por uma tabela de contingência
- Mesmas proporções?
- Observar: o que é fixo na tabela? linhas, colunas ou o total?
- Quais gráficos, tabelas e medidas melhor representam?

Categóricas *versus* categóricas

Diferentes visualizações das tabelas

Tabela 9. Dados Originais

	RMC	Interior	Outro Estado
Não	65	76	37
Sim	38	35	43

Tabela 10. Proporção total

	RMC	Interior	Outro Estado
Não	0,22	0,26	0,13
Sim	0,13	0,12	0,15

Tabela 11. Proporção por linha

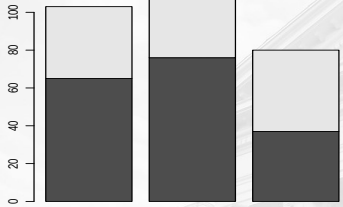
	RMC	Interior	Outro Estado
Não	0,37	0,43	0,21
Sim	0,33	0,30	0,37

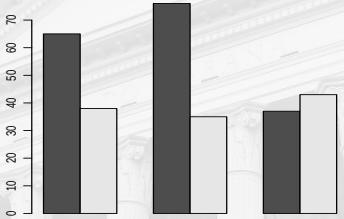
Tabela 12. Proporção por coluna

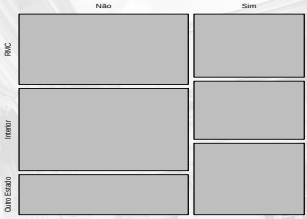
	RMC	Interior	Outro Estado
Não	0,63	0,68	0,46
Sim	0,37	0,32	0,54

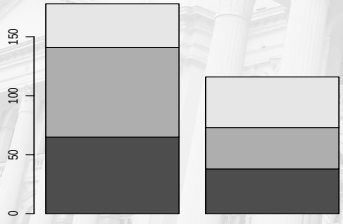
Categóricas *versus* categóricas (frequências)

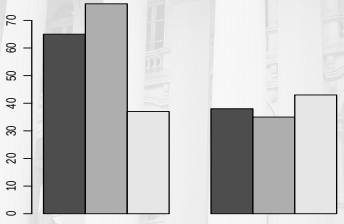
Algumas visualizações (ver análise do questionário para adicionais)

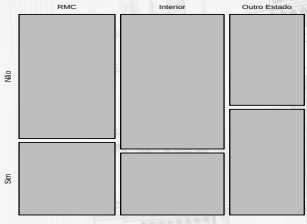






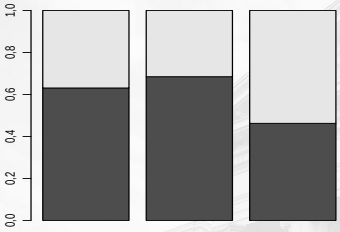


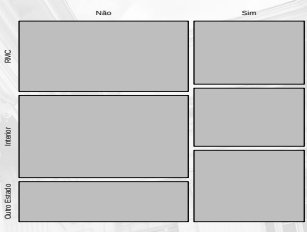


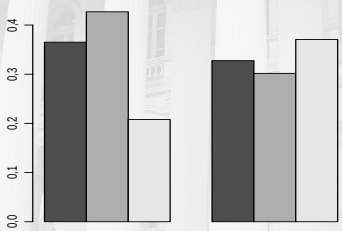


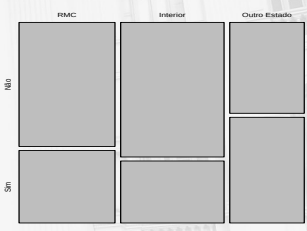
Categóricas *versus* categóricas (proporções)

Algumas visualizações (ver análise do questionário para adicionais)









Um caso especial: tabelas 2×2

Ensaio clínico

	Doente	Sadio	Sum
Exposto	50	50	100
Não Exposto	70	30	100
Sum	120	80	200

RR = (50/100)/(70/100) = 0,714
OR = (50:50)/(70:30) = 0,429

	Doente	Sadio	Sum
Exposto	50	50	100
Não Exposto	140	60	200
Sum	190	110	300

RR = (50/100)/(140/200) = 0,714
OR = (50:50)/(140:60) = 0,429

Estudo caso-controle

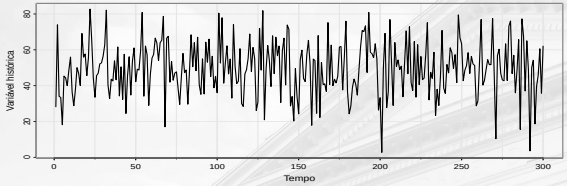
	Doente	Sadio	Sum
Exposto	42	63	105
Não Exposto	58	37	95
Sum	100	100	200

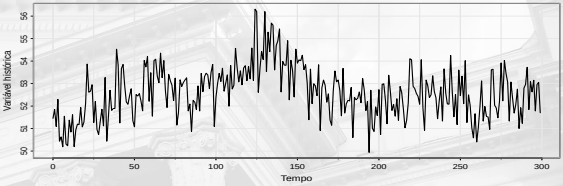
RR = (42/105)/(58/95) = 0,655
OR = (42:63)/(58:37) = 0,425

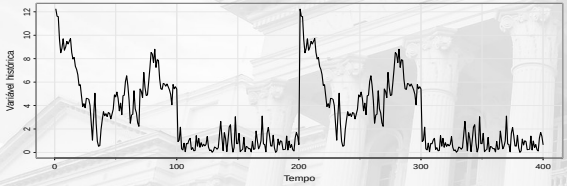
	Doente	Sadio	Sum
Exposto	42	126	168
Não Exposto	58	74	132
Sum	100	200	300

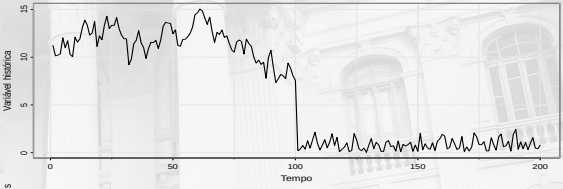
RR = (42/168)/(58/132) = 0,569
OR = (42:126)/(58:74) = 0,425

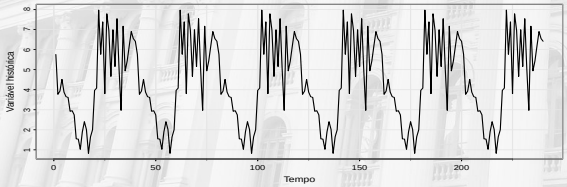
Uma estrutura especial: séries de tempo

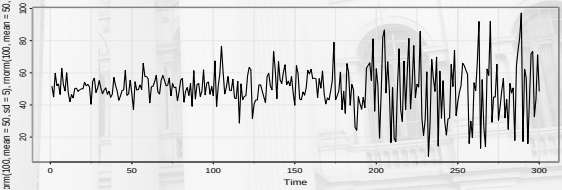












Exemplo: catedrais

```
?alr4::cathedral
```

```
Description:
     Heights and lengths of Gothic and Romanesque cathedrals.
Format:
     This data frame uses cathedral names as row label and
     contains the following columns:

     Type Romanesque or Gothic
     Height Total height, feet
     Length Total length, feet

...
```

Exemplo: catedrais

Estrutura dos dados:

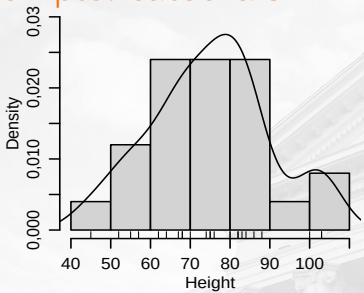
	Type	Height	Length
Durham	Romanesque	75	502
Canterbury	Romanesque	80	522
GloucesterRom	Romanesque	68	425
Hereford	Romanesque	64	344
NorwichRom	Romanesque	83	407
Peterborough	Romanesque	80	451
St. Albans	Romanesque	70	551
WinchesterRom	Romanesque	76	530
Ely	Romanesque	74	547
York	Gothic	100	519
Bath	Gothic	75	225
Bristol	Gothic	52	300
Chichester	Gothic	62	418
Exeter	Gothic	68	409
GloucesterGothic	Gothic	86	425
Lichfield	Gothic	57	370
Lincoln	Gothic	82	506
NorwichGothic	Gothic	82	506
Ripon	Gothic	88	295
Southwark	Gothic	55	273
Wells	Gothic	67	415
St. Asaph	Gothic	45	182
WinchesterGothic	Gothic	103	530
Old St. Paul	Gothic	103	611
Salisbury	Gothic	84	473

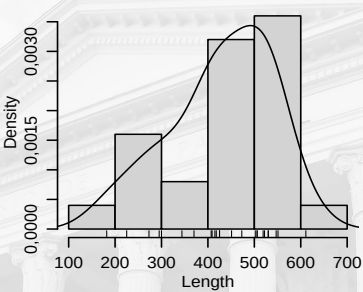
Quais análises são de interesse?

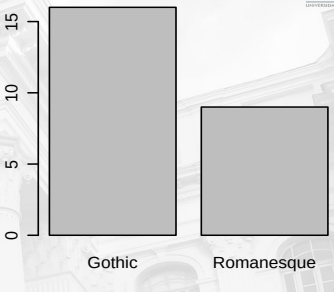
Exemplo: catedrais

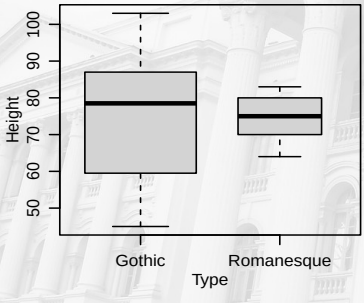
```
table(cathedral$Type)
##
##     Gothic Romanesque
##         16          9
psych:::describe(cathedral, omit=TRUE, quant=c(0.25, 0.75))
##        vars  n   mean     sd median trimmed    mad min max range  skew kurtosis    se Q0.25 Q0.75
## Height    2 25  75,16  15,01     75   75,05  11,86  45 103    58  0,06    -0,57  3,00    67    83
## Length    3 25 429,44 110,34    425  436,52 120,09 182 611   429 -0,56    -0,65 22,07   370   519
with(cathedral, by(cathedral, omit=TRUE, Type, psych:::describe, quant=c(0.25, 0.75)))
## Type: Gothic
##        vars  n   mean     sd median trimmed    mad min max range  skew kurtosis    se  Q0.25 Q0.75
## Height    2 16  75,56  18,42   78,5   75,79  20,76  45 103    58  0,00    -1,34  4,61  60,75  86,5
## Length    3 16 403,56 121,28  416,5  404,57 142,33 182 611   429 -0,22    -1,12 30,32 298,75 506,0
## ------------------------------------------------------------------------------------------------
## Type: Romanesque
##        vars n   mean    sd median trimmed   mad min max range  skew kurtosis    se Q0.25 Q0.75
## Height    2 9  74,44  6,21     75   74,44  7,41  64  83    19 -0,24    -1,41  2,07    70    80
## Length    3 9 475,44 72,27    502  475,44 72,65 344 551   207 -0,49    -1,37 24,09   425   530
```

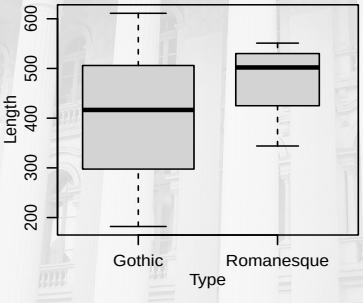
Exemplo: catedrais

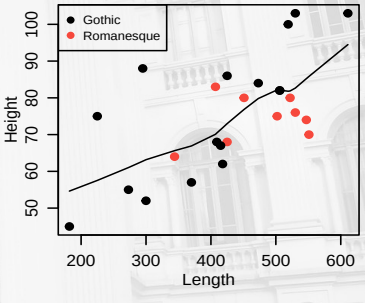












Exemplo: catedrais

- Análises univariadas fornecem um perfil dos dados, a partir de medidas selecionadas. Por exemplo, pode-se relatar:
  Das 25 catedrais 64% são góticas e as demais romanescas. As alturas variam de 45 a 103 pés. Os comprimentos variam de 182 a 611 pés.

- Análises bivariadas permitem investigar relações entre as variáveis.
  As alturas médias e medianas são similares porém as góticas possuem valores mais heterogêneos. Por exemplo, pode-se relatar:
  Alturas e comprimentos são relacionadas com coeficiente de correlação linear de 0,64.

Exemplo: germinação de sementes

Seed germination with different temperatures/concentrations
Description:
Seed germination with different temperatures/concentrations
Format:
A data frame with 64 observations on the following 5 variables.
‘temp’ temperature regimen
‘rep’ replication factor (not blocking)
‘conc’ chemical concentration
‘germ’ number of seeds germinating
‘seeds’ number of seeds tested = 50

Details:
The rep factor is NOT a blocking factor.
Source:
Roger Mead, Robert N Curnow, Anne M Hasted. 2002. Statistical Methods in Agriculture and Experimental Biology, 3rd ed. Chapman and Hall. Page 350-351.

Exemplo: germinação de sementes

```
dim(agridat::mead.germination)
## [1] 64  5
head(agridat::mead.germination, n=12)
##    temp rep conc germ seeds
## 1    T1  R1  0,0    9    50
## 2    T1  R1  0,1   13    50
## 3    T1  R1  1,0   21    50
## 4    T1  R1 10,0   40    50
## 5    T2  R1  0,0   19    50
## 6    T2  R1  0,1   33    50
## 7    T2  R1  1,0   43    50
## 8    T2  R1 10,0   48    50
## 9    T3  R1  0,0    7    50
## 10   T3  R1  0,1    1    50
## 11   T3  R1  1,0    8    50
## 12   T3  R1 10,0    3    50
```

Descritiva univariada: Faz sentido?

Exemplo: germinação de sementes

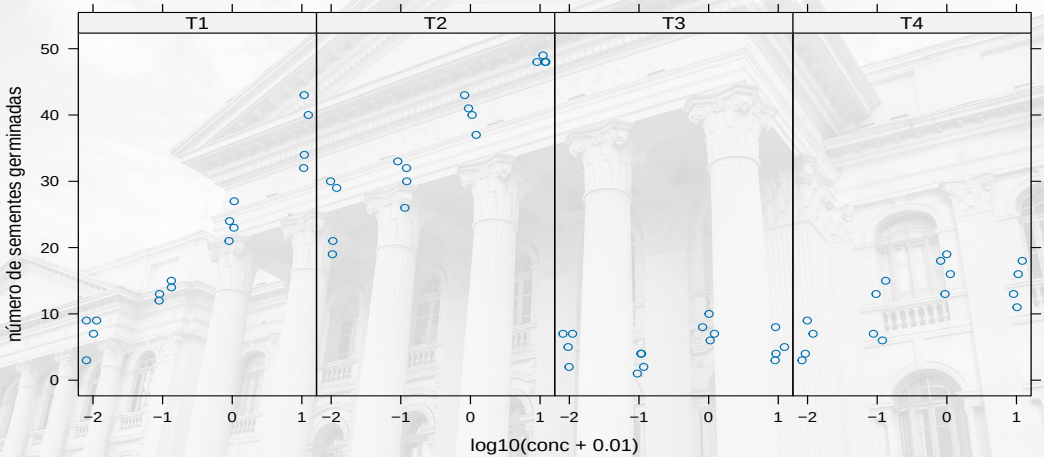


Figura 1. Germinação de sementes.

Exemplo: germinação de sementes

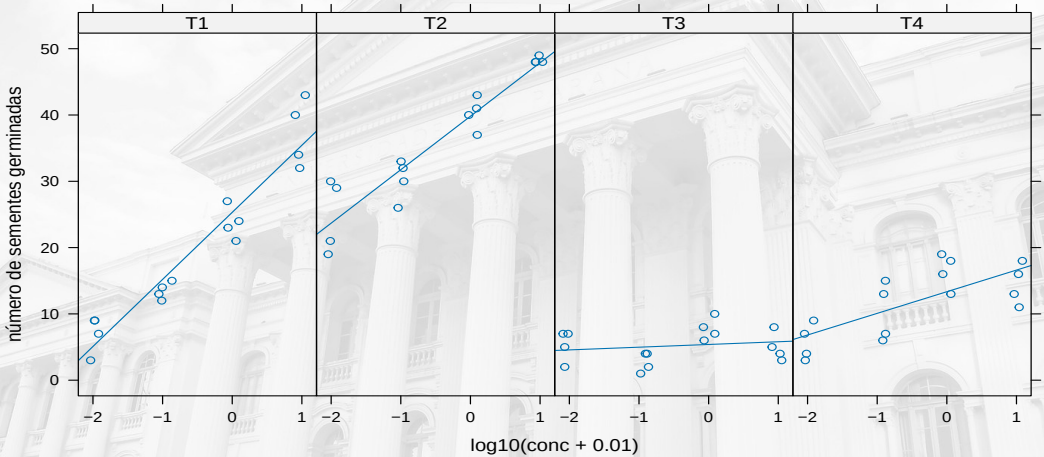


Figura 2. Germinação de sementes.

Uma palavra de cuidado

Descrições uni e bivariadas são sujeitas a efeitos de confundimento.

Influência de outra(s) variáveis não observadas.

Métodos multivariados e modelos estatísticos tratam múltiplas variáveis conjuntamente.

Referências bibliográficas

BUSSAB, W. O.; MORETTIN, P. A. **Estatística Básica**. 9. ed. São Paulo: Saraiva, 2017.

MAGALHÃES, M. N.; LIMA, A. C. P. de. **Noções de Probabilidade e Estatística**. 7. ed. São Paulo: Edusp, 2015.

UTTS, J. M. **Seeing Through Statistics**. [S.l.: s.n.], 2005.

WILD, C. J.; SEBER, G. A. F. **Chance Encounters: A First Course in Data Analysis and Inference**. [S.l.: s.n.], 2000.

WILD, C. J.; SEBER, G. A. F. **Encontros Com O Acaso. Primeiro Curso De Análise De Dados e Inferência**. [S.l.: s.n.], 2004.

ZEVIANI, W. et al. EstBas: Um curso em estatística básica. <http://www.leg.ufpr.br/estbas>. 2021.