Estatística com Apoio Computacional

Análise Bidimensional

Universidade Estadual Vale do Acaraú – UVA

Paulo Regis Menezes Sousa

paulo_regis@uvanet.br

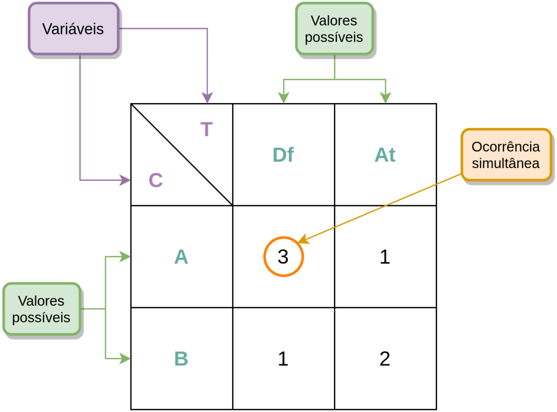
Análise Bidimensional

- Variáveis Qualitativas
- Variáveis Quantitativas
- Qualitativa $\times$ Quantitativa

- Frequentemente estamos interessados em analisar o **comportamento conjunto** de duas ou mais variáveis aleatórias.
- Os dados geralmente aparecem na forma de uma matriz, com as **colunas** indicando as variáveis e as **linhas** os indivíduos (ou elementos).
- O principal **objetivo** das análises nessa situação é explorar **relações** (similaridades) entre as colunas, ou algumas vezes entre as linhas.
- Quando consideramos **duas variáveis**, podemos ter três situações:
  (a) duas variáveis qualitativas;
  (b) duas variáveis quantitativas; ou
  (c) uma variável qualitativa e outra quantitativa.

- Quando as variáveis são qualitativas, os dados são resumidos em tabelas de dupla entrada (ou de contingência).

P	C	E	T
1.0	A	Au	Df
6.2	B	Ag	At
0.4	A	Au	At
1.3	A	Pt	Df
2.9	B	Ag	Df
⋮	⋮	⋮	⋮
3.1	B	Pt	At
9.7	A	Au	Df



Exemplo 1

Suponha que queiramos analisar o comportamento conjunto das variáveis Y: grau de instrução e V: região de procedência, cujas observações estão contidas na Tabela 1. A distribuição de frequências é representada por uma tabela de dupla entrada e está na Tabela 2.

Nº	Estado civil	Grau de instrução	Nº de filhos	Salário (× sal. mín.)	Idade		Região de procedência
					anos	meses	
1	solteiro	ensino fundamental	—	4,00	26	03	interior
2	casado	ensino fundamental	1	4,56	32	10	capital
3	casado	ensino fundamental	2	5,25	36	05	capital
4	solteiro	ensino médio	—	5,73	20	10	outra
5	solteiro	ensino fundamental	—	6,26	40	07	outra
6	casado	ensino fundamental	0	6,66	28	00	interior

Tabela 1: Dados sobre os empregados da seção de orçamentos da Companhia MB.

Y / V	Ensino Fundamental	Ensino Médio	Superior	Total
Capital	4	5	2	11
Interior	3	7	2	12
Outra	5	6	2	13
Total	12	18	6	36

Tabela 2: Distribuição conjunta das frequências

- Para a visualização da tabela de contingência no R usamos os comandos `table()` ou `ftable()`.

```
emd <- read.csv("empregados-mb.csv", sep = ";", header = TRUE)

# retorna uma lista de nomes das colunas do data.frame
names(emd)

# cria uma tabela de contingência
tab.cont <- table( emd[ c("Procedência","Grau.de.instrução") ] )
tab.cont
```

- Em vez de trabalharmos com as frequências absolutas, podemos construir tabelas com as frequências relativas (proporções).
- Existem três possibilidades de expressarmos a proporção, de acordo com o objetivo do problema em estudo:
  (a) em relação ao total geral;
  (b) em relação ao total de cada linha;
  (c) ou em relação ao total de cada coluna.

Y \ V	Fundamental	Médio	Superior	Total
Capital	11%	14%	6%	31%
Interior	8%	19%	6%	33%
Outra	14%	17%	5%	36%
Total	33%	50%	17%	100%

Tabela 3: Distribuição conjunta das proporções em relação ao total geral das variáveis Y e V.

Y \ V	Fundamental	Médio	Superior	Total
Capital	33%	28%	33%	31%
Interior	25%	39%	33%	33%
Outra	42%	33%	34%	36%
Total	100%	100%	100%	100%

Tabela 4: Distribuição conjunta das proporções em relação aos totais de cada coluna das variáveis Y e V.

```r
# cria uma tabela de contingência exibida em porcentagens
tab.prop <- prop.table(tab.cont)
tab.prop

# tabela de proporções das linhas
tab.prop.lin <- prop.table(tab.cont, margin = 1)
tab.prop.lin
```

- A comparação entre as duas variáveis também pode ser feita utilizando-se representações gráficas.

```
# cores personalizadas
cores <- c("#B3E2CD", "#FDCDAC", "#CBD5E8")

# retângulos com áreas são proporcionais às frequências
plot(tab.cont, main = "Frequências", col = cores)

# gráficos de barra com a tabela de contingência
barplot(tab.cont, legend = TRUE, col = cores)
barplot(tab.cont, legend = TRUE, col = cores, beside = TRUE)
```

```r
# gráficos de barra com a tabela de contingência em porcentagem
barplot(tab.prop, legend = TRUE, col = cores)
barplot(tab.prop, legend = TRUE, horiz = TRUE, col = cores)

# outros atributos do gráfico de barras
barplot(tab.prop,
        legend = TRUE,
        angle = c(90,45,0),     # ângulos das linhas de sombreamento
        density = c(30,25,20)) # densidades das linhas de sombreamento
```

Exercício 1.

Usando os dados `empregados-mb.csv`:

a) Construa a distribuição de frequência conjunta para as variáveis grau de instrução e região de procedência.

b) Qual a porcentagem de funcionários que têm o ensino médio?

c) Qual a porcentagem daqueles que têm o ensino médio e são do interior?

d) Dentre os funcionários do interior, quantos por cento têm o ensino médio?

```r
emd <- read.csv("empregados-mb.csv", sep = ";", header = TRUE)

# a)
names(emd)
tc <- table(emd[c("Grau.de.instrução","Procedência")])
tp <- prop.table(tc)
tp

# b)
porcentagem.ensino.medio <- sum(tp[2, ])
porcentagem.ensino.medio

# c)
tp.lin <- prop.table(tc)
tp.lin
porcentagem.ensino.medio.interior <- tp.lin[2,2]
porcentagem.ensino.medio.interior

# d)
tp.col <- prop.table(tc, margin = 2)
```

Exercício 2.

Numa pesquisa sobre rotatividade de mão-de-obra, para uma amostra de 40 pessoas foram observadas duas variáveis: número de empregos nos últimos dois anos (X) e salário mais recente, em número de salários mínimos (Y). Os resultados estão no arquivo `rotatividade-mao-de-obra.csv`:

a) Usando a mediana, classifique os indivíduos em dois níveis, alto e baixo, para cada uma das variáveis, e construa a distribuição de frequências conjunta das duas classificações.

b) Qual a porcentagem das pessoas com baixa rotatividade e ganhando pouco?

c) Qual a porcentagem das pessoas que ganham pouco?

d) Entre as pessoas com baixa rotatividade, qual a porcentagem das que ganham pouco?

e) A informação adicional dada em d) mudou muito a porcentagem observada em c)? O que isso significa?

```r
# a)
rmo = read.csv("~/EAC/rotatividade-mao-de-obra.csv", sep=";", header=T)
rmo$XClass = rep(NA, length(rmo$X))
rmo$YClass = rep(NA, length(rmo$Y))

selx = rmo$X < median(rmo$X) # TRUE para todo X < mediana
sely = rmo$Y < median(rmo$Y) # TRUE para todo Y < mediana

rmo$XClass[selx] = "baixo"
rmo$XClass[!selx] = "alto" # ! operador de negação
rmo$YClass[sely] = "baixo"
rmo$YClass[!sely] = "alto"

tc = table(rmo[, c("XClass", "YClass")])      # tab. de contingência
tp = prop.table(tc) # tab. de proporções
tp

# b)
tp["baixo","baixo"] # baixa rotatividade e baixo salário
```

```
# c)
sum(tp[ ,"baixo"]) # baixo salário total

# d)
tp.lin = prop.table(tc, margin = 1) # tab. prop. para linhas
tp.lin
tp.lin["baixo","baixo"] # baixa rotatividade com baixo salário

# e) Bastante modificada;
#    maioria das pessoas que ganham pouco têm alta rotatividade.
```

- Um dispositivo bastante útil para se verificar a associação entre duas variáveis quantitativas, ou entre dois conjuntos de dados, é o *gráfico de dispersão*.

```r
# agentes de vendas e núm. de clientes por ano de trabalho
aac = read.csv("agentes-anos-clientes.csv", sep = ";",
                header = TRUE)
head(aac)

# gráfico de dispersão
plot(aac[,"Número.de.clientes"], aac[,"Anos.de.serviço"],
     xlab = "Número de clientes",
     ylab = "Anos de serviço",
     pch = 19,
     col = "#FF7F00")
```

```r
# gasto com saúde por renda bruta familiar
fgs = read.csv("familia-gasto-saude.csv", sep = ";",
               header = TRUE,
               dec = ",") # contém números separados com vírgula
head(fgs)
plot(fgs[,"X"], fgs[,"Y"],
     xlab = "Renda bruta mensal",
     ylab = "Renda gasta em saúde (%)",
     pch = 17,
     col = "#00ccff")

# nota por tempo gasto no teste
tom = read.csv("teste-operacao-maquina.csv", sep = ";",
               header = TRUE)
head(tom)
plot(tom[,"X"], tom[,"Y"],
     xlab = "Resultado no teste (0-100)",
     ylab = "Tempo para operar a máquina (min)",
     pch  = 15,
     col = "magenta")
```

- A partir dos gráficos apresentados, verificamos que a representação gráfica das variáveis quantitativas ajuda muito a compreender o comportamento conjunto das duas variáveis quanto à existência ou não de associação entre elas.

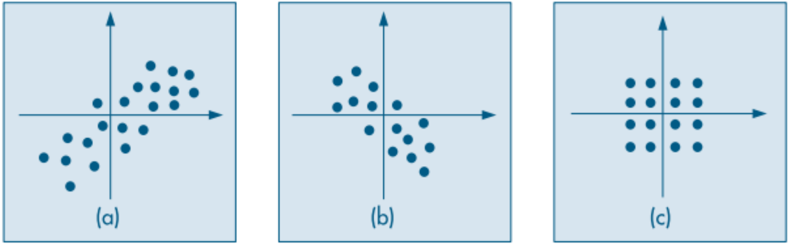
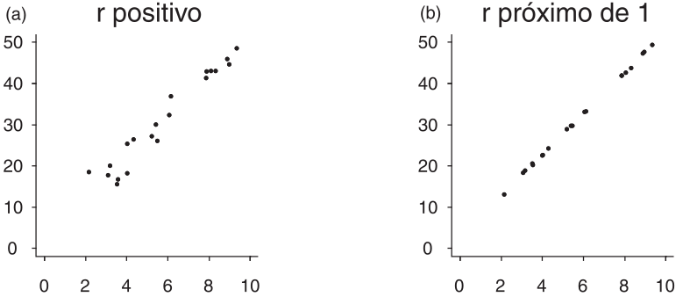


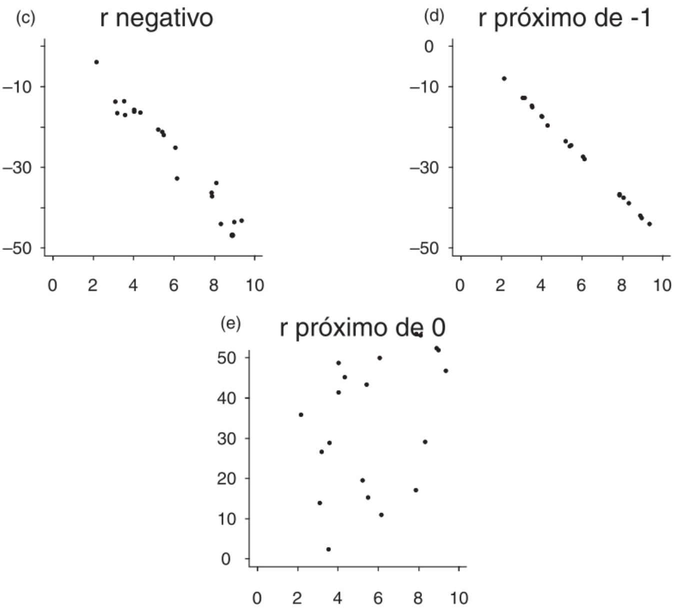
Figura 1: Tipos de associações entre duas variáveis.

Coeficiente de correlação (linear)

É uma medida do grau de associação entre elas e também da proximidade dos dados a uma reta.

- O coeficiente de correlação r_{xy} que pode assumir qualquer valor real entre -1 e 1.




(c)
r negativo
−10
−30
−50
0
2
4
6
8
10
(d)
r próximo de -1
0
−10
−30
−50
0
2
4
6
8
10
(e)
r próximo de 0
50
40
30
20
10
0
0
2
4
6
8
10

- A função `cor(x,y)` calcula o coeficiente de correlação entre as variáveis x e y.

```
# coeficiente de correlação entre X e Y
r0 = cor(aac[,"Número.de.clientes"], aac[,"Anos.de.serviço"])
r1 = cor(fgs[,"X"], fgs[,"Y"])
r2 = cor(tom[,"X"], tom[,"Y"])

coef.coor = data.frame(R.xy = c(r0,r1,r2),
                       Descrição = c("Próximo de 1",
                                     "Próximo de -1",
                                     "Próximo de 0"))
coef.coor
```

- É comum também analisar o que acontece com a variável **quantitativa** dentro de cada categoria da variável **qualitativa**.

Exemplo

Retomemos os dados da Tabela 1, para os quais desejamos analisar agora o comportamento dos salários (S) dentro de cada categoria de grau de instrução (Y). Comecemos a análise construindo a Tabela 5, que contém medidas-resumo da variável S para cada categoria de Y.

Grau de instrução	n	$\bar{s}$	dp(S)	var(S)	$s_{(1)}$	q_1	q_2	q_3	$s_{(n)}$
Fundamental	12	7,84	2,79	7,77	4,00	6,01	7,13	9,16	13,65
Médio	18	11,54	3,62	13,10	5,73	8,84	10,91	14,48	19,40
Superior	6	16,48	4,11	16,89	10,53	13,65	16,74	18,38	23,30
Todos	36	11,12	4,52	20,46	4,00	7,55	10,17	14,06	23,30

Tabela 5: Medidas-resumo para a variável salário, segundo o grau de instrução, na Companhia MB.

- Criação da tabela 5 no R:
  - Na primeira linha os dados são carregados.
  - Nas linhas 3, 5 e 6, todos os salários de pessoas com ensino fundamental, médio e superior são selecionados respectivamente.

```
1 emd = read.csv("~/EAC/empregados-mb.csv", sep = ";", header = TRUE,
2                 dec = ",")
3 fund = emd[emd[,"Grau.de.instrução"] == "Ensino fundamental", "Salário"]
4 med  = emd[emd[,"Grau.de.instrução"] == "Ensino médio", "Salário"]
5 sup  = emd[emd[,"Grau.de.instrução"] == "Superior", "Salário"]
6
7 tab = data.frame( N     = c(length(fund), length(med), length(sup)),
```

```
                   Média = c(mean(fund)   , mean(med)   , mean(sup)   ),
                   DP    = c(sd(fund)     , sd(med)     , sd(sup)     ),
                   Var   = c(var(fund)    , var(med)    , var(sup)    ),
                   S1    = c(min(fund)    , min(med)    , min(sup)    ),
                   Q1    = c(quantile(fund, 0.25, names = FALSE),
                             quantile(med , 0.25, names = FALSE),
                             quantile(sup , 0.25, names = FALSE)),
                   Q2    = c(quantile(fund, 0.50, names = FALSE),
                             quantile(med , 0.50, names = FALSE),
                             quantile(sup , 0.50, names = FALSE)),
                   Q3    = c(quantile(fund, 0.75, names = FALSE),
                             quantile(med , 0.75, names = FALSE),
                             quantile(sup , 0.75, names = FALSE)),
                   Sn    = c(max(fund), max(med), max(sup)) )

row.names(tab) = c("Fundamental", "Médio", "Superior")
todos = c(fund, med, sup)
```

```
tab = rbind(tab, Todos = c(length(todos),
                           mean(todos),
                           sd(todos),
                           var(todos),
                           min(todos),
                           quantile(todos, 0.25, names = FALSE),
                           quantile(todos, 0.50, names = FALSE),
                           quantile(todos, 0.75, names = FALSE),
                           max(todos)) )

tab
```

- Na Figura 2, apresentamos uma visualização gráfica por meio de *box plots*.

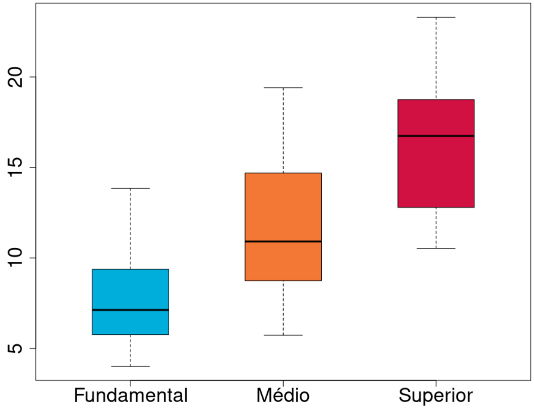


Figura 2: *Box plots* de salário segundo grau de instrução.

- Criação dos Box plots no R.

```r
b = boxplot(fund, med, sup,
            names  = c("Fundamental", "Médio", "Superior"),
            col    = c("#00aedb", "#f37735", "#d11141"),
            boxwex = 0.5,
            cex.axis = 2)
b
```

- Os atributos `border` e `boxwex` são, respectivamente, um vetor com as cores das bordas dos desenhos de cada uma das caixas e a proporção de largura para todas as caixas.
- A função `boxplot` possui um retorno, neste caso o objeto `b`. O retorno da função exibe uma lista dos atributos do gráfico especificando dados que podem ser difíceis de visualizar apenas no gráfico.

A leitura desses resultados sugere uma dependência dos salários em relação ao grau de instrução: o salário aumenta conforme aumenta o nível de educação do indivíduo.

- Funcionários com o ensino fundamental completo recebem, em média, 7,84.
- O salário médio de um funcionário é 11,12 (salários mínimos).
- Para um funcionário com curso superior o salário médio passa a ser 16,48.

- Na Tabela 6 e Figura 3 temos os resultados da análise dos salários em função da região de procedência (V).

Região de procedência	n	$\bar{s}$	dp(S)	var(S)	$s_{(1)}$	q_1	q_2	q_3	$s_{(n)}$
Capital	11	11,46	5,22	27,27	4,56	7,49	9,77	16,63	19,40
Interior	12	11,55	5,07	25,71	4,00	7,81	10,64	14,70	23,30
Outra	13	10,45	3,02	9,13	5,73	8,74	9,80	12,79	16,22
Todos	36	11,12	4,52	20,46	4,00	7,55	10,17	14,06	23,30

Tabela 6: Medidas-resumo para a variável salário segundo a região de procedência, na Companhia MB.

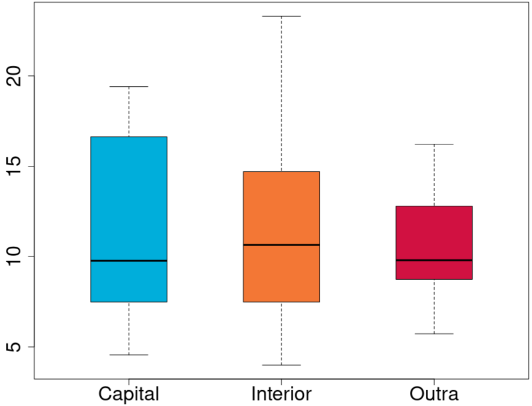


Figura 3: Medidas-resumo para a variável salário segundo a região de procedência, na Companhia MB.

- Sem usar a informação da variável categorizada, a variância calculada para a variável quantitativa para todos os dados mede a dispersão dos dados globalmente

Se a variância dentro de cada categoria for pequena e menor do que a global, significa que a variável qualitativa melhora a capacidade de previsão da quantitativa e portanto existe uma relação entre as duas variáveis.

- Para termos uma medida-resumo da variância entre as categorias da variável qualitativa, podemos usar a média das variâncias ponderada pelo número de observações em cada categoria.

$$\overline{\mathrm{var}(S)} = \frac{\sum_{k}^{i=1} n_i \mathrm{var}_i(S)}{\sum_{k}^{i=1} n_i}, \tag{1}$$

- No qual k é o número de categorias ($k = 3$ nos dois exemplos) e $\mathrm{var}_i(S)$ denota a variância de S dentro da categoria $i, i = 1, 2, \dots, k$.
- Pode-se mostrar que $\overline{\mathrm{var}(S)} \leq \mathrm{var}(S)$, de modo que podemos definir o grau de associação entre as duas variáveis como o ganho relativo na variância, obtido pela introdução da variável qualitativa.

$$R^2 = \frac{\mathrm{var}(S) - \overline{\mathrm{var}(S)}}{\mathrm{var}(S)} = 1 - \frac{\overline{\mathrm{var}(S)}}{\mathrm{var}(S)}. \tag{2}$$

Exemplo

Voltando aos dados do Exemplo anterior, vemos que para a variável S na presença de grau de instrução, tem-se

$$\overline{\text{var}(S)} = \frac{12(7,77) + 18(13,10) + 6(16,89)}{12 + 18 + 6} = 11,96, \quad (3)$$

$$\text{var}(S) = 20,46 \quad (4)$$

de modo que

$$R^2 = 1 - \frac{11,96}{20,46} = 0,415, \quad (5)$$

e dizemos que 41,5% da variação total do salário é explicada pela variável grau de instrução. Fazendo o mesmo cálculo para S e região de procedência temos $R^2 = 0,013$ de modo que apenas 1,3% da variabilidade dos salários é explicada pela região de procedência.

Exercício 3.
Uma amostra de 200 habitantes de uma cidade foi escolhida para declarar sua opinião sobre um certo projeto governamental. O resultado você pode encontrar no arquivo `pesquisa-opiniao.csv`:

a) Crie uma tabela de proporções em relação ao total das colunas.

b) Você diria que a opinião independe do local de residência?

Exercício 4.

Uma pesquisa sobre a participação em atividades esportivas de adultos moradores nas proximidades de centros esportivos construídos pelo estado de São Paulo mostrou os resultados da tabela abaixo. Baseado nesses resultados você diria que a participação em atividades esportivas depende da cidade?

Participam	Cidade			
	São Paulo	Campinas	Rib. Preto	Santos
Sim	50	65	105	120
Não	150	185	195	180

Exercício 5.

Uma amostra de dez casais e seus respectivos salários anuais (em s.m.) foi colhida num certo bairro conforme vemos na tabela abaixo.

	Casal nº	1	2	3	4	5	6	7	8	9	10
Salário	Homem (X)	10	10	10	15	15	15	15	20	20	20
	Mulher (Y)	5	10	10	5	10	10	15	10	10	15

a) Encontre o salário anual médio dos homens e o seu desvio padrão.

b) Encontre o salário anual médio das mulheres e o seu desvio padrão.

c) Construa o diagrama de dispersão.

d) Qual o salário médio familiar? E a variância do salário familiar?

e) Se o homem é descontado em 8% e a mulher em 6%, qual o salário líquido médio familiar? E a variância?