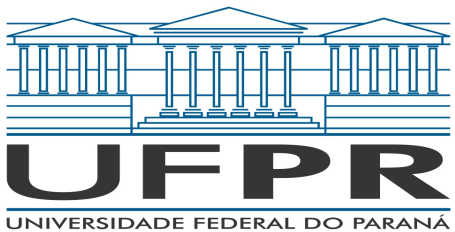
Variáveis bidimensionais

Wagner H. Bonat
Elias T. Krainski
Fernando P. Mayer

Universidade Federal do Paraná
Departamento de Estatística
Laboratório de Estatística e Geoinformação

11/04/2018



Sumário

Introdução

- Interesse no comportamento conjunto de várias variáveis.
- Construção de tabelas de frequência conjunta ou função de probabilidade conjunta.
- O principal objetivo é explorar relações (similaridades) entre as colunas (ou linhas).
- Determinar se existe **associação** entre as variáveis.
- Podemos ter três situações:
  a. Duas variáveis qualitativas
  b. Duas variáveis quantitativas
  c. Uma variável qualitativa e outra quantitativa

Introdução

- Em todas as situações o objetivo é encontrar as possíveis **relações** ou **associações** entre as duas variáveis
- Essas relações podem ser detectadas por meio de **métodos gráficos** ou **medidas numéricas**
- Para efeitos práticos: existe associação se existe uma **mudança** no comportamento de uma variável na presença de outra
- Exemplo:
  a. Frequência esperada de pessoas com mais de 170 cm de altura
  b. Frequência esperada de pessoas com mais de 170 cm de altura por sexo
- Se a resposta for a mesma, dizemos que não há associação

Exemplo 5.1

- Uma amostra de 20 alunos do primeiro ano de uma faculdade foi escolhida. Perguntou-se aos alunos se trabalhavam, variável que foi representada por X, e o número de vestibulares prestados, variável representada por Y. Os dados obtidos estão na tabela abaixo.

X	não	sim	não	não	não	sim	sim	não	sim	sim
Y	1	1	2	1	1	2	3	1	1	1

X	não	não	sim	não	sim	não	não	não	sim	não
Y	2	2	1	3	2	2	2	1	3	2

Exemplo 5.1

Distribuição conjunta

(X,Y)	Freq
não,1	5
não,2	6
não,3	1
sim,1	4
sim,2	2
sim,3	2
Sum	20

Exemplo 5.1

Distribuição conjunta (melhor para visualizar)

X/Y	1	2	3	Sum
não	5	6	1	12
sim	4	2	2	8
Sum	9	8	3	20

Distribuição marginal de X

não	sim	Sum
12	8	20

Distribuição marginal de Y

1	2	3	Sum
9	8	3	20

Exemplo 5.2

- Um estudo envolveu 345 pacientes HIV positivos, acompanhados durante um ano, pelo setor de doenças infecciosas de um grande hospital público. Os dados apresentados contêm as ocorrências relacionadas às variáveis número de internações (I) e número de crises com infecções oportunistas (C).

I/C	0	1	2	3	4
0	84	21	8	2	0
1	20	59	35	14	2
2	6	11	43	28	12

- Obtenha as marginais de I e C.
- Exemplo 5.3 tarefa de casa.

Exemplo 5.2

- Marginal de I

0	1	2	Sum
115	130	100	345

- Marginal de C

0	1	2	3	4	Sum
110	91	86	44	14	345

Função de probabilidade conjunta

Sejam X e Y duas VAs discretas originárias do mesmo fenômeno aleatório, com valores atribuídos a partir do mesmo espaço amostral.

A **função de probabilidade conjunta** é definida, para todos os possíveis pares de valores (X, Y), da seguinte forma:

$$p(x, y) = P[(X = x) \cap (Y = y)] = P(X = x, Y = y).$$

Ou seja, $p(x, y)$ representa a probabilidade de (X, Y) ser igual a (x, y).

A função de probabilidade conjunta também pode ser chamada de **distribuição conjunta** ou simplesmente **conjunta** das variáveis.

Exemplo 5.4

Uma empresa atende encomendas de supermercados dividindo os pedidos em duas partes de modo a serem atendidos, de forma independente, pelas suas duas fábricas. Devido à grande demanda, pode haver atraso no cronograma de entrega, sendo que a fábrica *I* atrasa com probabilidade 0.1 e a *II* com 0.2. Sejam A_I e A_{II} os eventos correspondentes a ocorrência de atraso nas fábricas *I* e *II*, respectivamente.

Para uma entrega, a indústria recebe 200 u.m, mas paga 20 para cada fábrica que atrasar. Considere que o supermercado que recebe a encomenda fez um índice relacionado à pontualidade de entrega. Este índice, atribuiu 10 pontos para cada entrega dentro do cronograma previsto. Denote por X o valor recebido pelo pedido e Y o índice obtido. Obtenha a conjunta de Y e X e as marginais de Y e X.

Exemplo 5.5

Uma região foi dividida em 10 sub-regiões. Em cada uma delas, foram observadas duas variáveis: número de poços artesianos (X) e número de riachos ou rios presentes na sub-região (Y). Os resultados são apresentados na tabela a seguir:

Sub-região	1	2	3	4	5	6	7	8	9	10
X	0	0	0	0	1	2	1	2	2	0
Y	1	2	1	0	1	0	0	1	2	2

- Construa a distribuição conjunta e marginais de X e Y.
- Exemplo 5.6 tarefa de casa.

Exemplo 5.5

Consideramos que cada região tem a mesma probabilidade 1/10 de ser escolhida. Assim a distribuição conjunta é:

(X,Y)	p(x,y)
0,0	0.1
0,1	0.2
0,2	0.2
1,0	0.1
1,1	0.1
2,0	0.1
2,1	0.1
2,2	0.1
Sum	1.0

Exemplo 5.5

Uma forma mais conveniente é

X/Y	0	1	2
0	0.1	0.2	0.2
1	0.1	0.1	0.0
2	0.1	0.1	0.1

Para obter as marginais, efetuamos a soma nas linhas para obter a marginal de X, e nas colunas para obter a marginal de Y. Por exemplo, $P(X = 0)$ é obtida através de:

$$\begin{aligned} P(X = 0) &= P(X = 0, Y = 0) + P(X = 0, Y = 1) + P(X = 0, Y = 2) \\ &= 0.1 + 0.2 + 0.2 = 0.5 \end{aligned}$$

Exemplo 5.5

Repetindo os cálculos para todos os valores de X e Y, obtemos as marginais:

X/Y	0	1	2	P(X=x)
0	0.1	0.2	0.2	0.5
1	0.1	0.1	0.0	0.2
2	0.1	0.1	0.1	0.3
P(Y=y)	0.3	0.4	0.3	1.0

Marginal de X

0	1	2	Sum
0.5	0.2	0.3	1

Marginal de Y

0	1	2	Sum
0.3	0.4	0.3	1

Funções de probabilidade marginal

Da função de probabilidade conjunta $p(x, y)$, é possível então obter as **funções de probabilidade marginais** de X e Y, através da soma de uma das coordenadas:

$$P(X = x) = \sum_y p(x, y) \quad \text{e} \quad P(Y = y) = \sum_x p(x, y)$$

com o somatório percorrendo todos os valores de X ou Y, conforme for o caso.

Associação entre variáveis

- Um dos principais objetivos de se construir uma distribuição conjunta de duas variáveis, é descrever a **associação** entre elas
- Queremos conhecer o grau de **dependência**, para prever melhor o resultado de uma delas quando conhecemos a outra
- Veremos algumas formas de medir/avaliar essa dependência:
  a. Duas variáveis qualitativas
     - Verificação de proporções através da distribuição conjunta
     - Medida Q^2
  b. Duas variáveis quantitativas
     - Diagramas de dispersão
     - Probabilidades condicionais
     - Correlação e covariância

Exemplo 5.7

- Dentre os alunos do 1° ano do ensino médio de uma certa escola, selecionou-se os quinze alunos com melhor desempenho, (nota acima de 7) em inglês. Para esses alunos, foi construída a tabela abaixo com as notas de inglês (I), português (P) e matemática (M):

I	7	7	7	7	8	8	8	8	8	8	8	9	9	9	10
P	8	6	8	9	8	6	9	7	7	6	7	8	9	8	8
M	5	6	7	5	5	5	6	4	7	6	5	5	6	5	5

- Obtenha as distribuições conjuntas e gráficos de dispersão.

Exemplo 5.7 - Distribuições conjuntas

Inglês e Português:

I/P	6	7	8	9
7	1	0	2	1
8	2	3	1	1
9	0	0	2	1
10	0	0	1	0

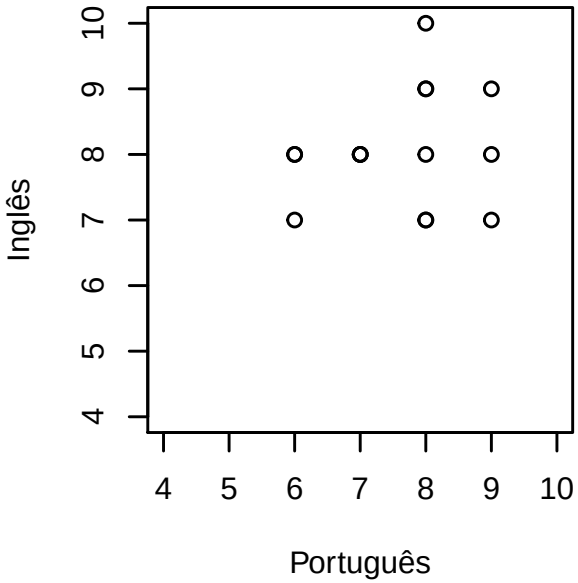
Inglês e Matemática:

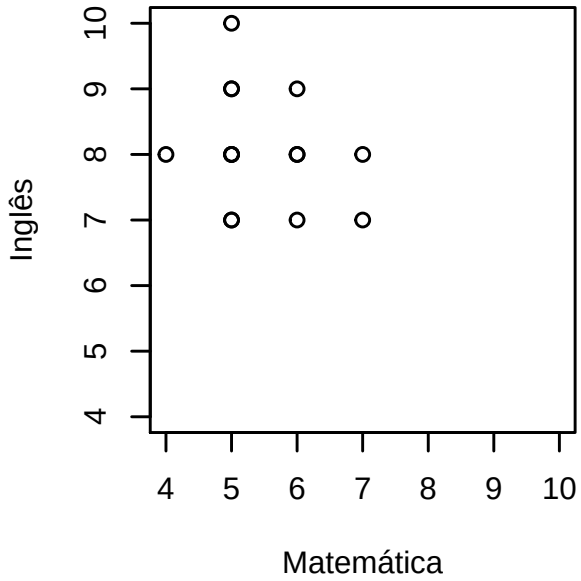
I/M	4	5	6	7
7	0	2	1	1
8	1	3	2	1
9	0	2	1	0
10	0	1	0	0

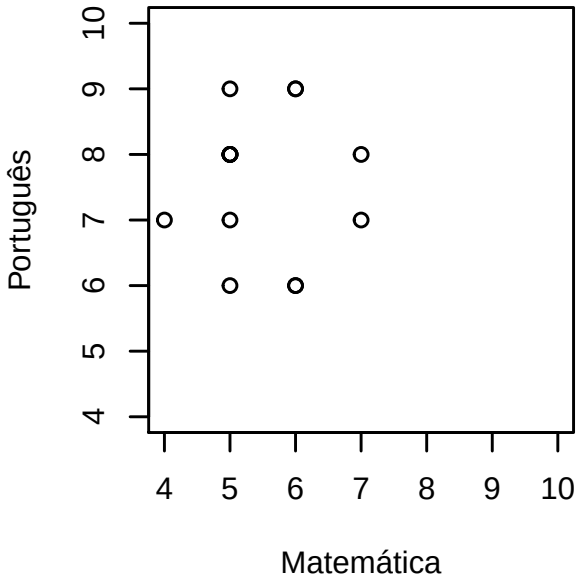
Português e Matemática:

P/M	4	5	6	7
6	0	1	2	0
7	1	1	0	1
8	0	5	0	1
9	0	1	2	0

Exemplo 5.7 - Diagramas de dispersão







Probabilidade condicional para VAs discretas

- A **probabilidade condicional** de $X = x$, dado que $Y = y$ ocorreu, é dada pela expressão:

$$P(X = x|Y = y) = \frac{P(X = x, Y = y)}{P(Y = y)}, \quad \text{se} \quad P(Y = y) > 0.$$

- Duas VAs discretas são **independentes**, se a ocorrência de qualquer valor de uma delas não altera a probabilidade de valores da outra. Em termos matemáticos

$$P(X = x|Y = y) = P(X = x).$$

- Definição alternativa

$$P(X = x, Y = y) = P(X = x)P(Y = y), \quad \forall(x, y).$$

Exemplo 5.8

- O Centro Acadêmico de uma faculdade de administração fez um levantamento da remuneração dos estágios dos alunos, em salários mínimos, com relação ao ano que estão cursando. As probabilidades de cada caso são apresentadas na próxima tabela, incluindo as distribuições marginais.

Salario/Ano	2	3	4	5	$P(Sal = x)$
2	2/25	2/25	1/25	0	5/25
3	2/25	5/25	2/25	2/25	11/25
4	1/25	2/25	2/25	4/25	9/25
$P(Ano = y)$	5/25	9/25	5/25	6/25	1

- X e Y são independentes?

Exemplo 5.9

- Em uma clínica médica foram coletados dados em 150 pacientes, referentes ao último ano. Observou-se a ocorrência de infecções urinárias (U) e o número de parceiros sexuais (N).

U/N	0	1	2 +	Total
Sim	12	21	47	80
Não	45	18	7	70
Total	57	39	54	150

- Estude a associação entre U e N.

Exemplo 5.9

- Ao invés de trabalharmos com as frequências absolutas, podemos construir tabelas com as frequências relativas, mas aqui existem três possibilidades para expressar as proporções:
  a. em relação ao total geral
  b. em relação ao total de cada linha
  c. em relação ao total de cada coluna
- A escolha depende do objetivo do estudo, mas não altera a conclusão

Exemplo 5.9

- Tabela com porcentagens em relação ao total de coluna.

U/N	0	1	2 +	Total
Sim	21,1%	53,8%	87,0%	53,3%
Não	78,9%	46,2%	13,0%	46,7%
Total	100%	100%	100%	100%

- Independente de N, a porcentagem de pessoas com infecção é 53,3% (46,7% sem infecção).
- Caso não exista associação de U com N, deveríamos esperar porcentagens similares em cada valor de N (independência).
- Analisar os percentuais em relação ao total das linhas levaria à mesma conclusão.

Exemplo 5.10

- Os dados abaixo representam uma amostra de 80 famílias de um certo bairro, onde T é o número de pessoas que trabalham na família, e A é o número de adolescentes entre 12 e 18 anos.

T/A	0	1	2	3	4	Sum
0	5	4	2	3	1	15
1	2	8	6	4	1	21
2	4	8	8	5	2	27
3	4	2	2	5	4	17
Sum	15	22	18	17	8	80

- Verifique a associação entre as duas variáveis.

Exemplo 5.10

Usando a distribuição marginal de T (ou seja, utilizando-se a soma por colunas)

0	1	2	3	Sum
0.19	0.26	0.34	0.21	1

podemos calcular quais seriam as proporções esperadas para cada valor de A, caso fossem independentes:

	0	1	2	3	4	
0	0.33	0.18	0.11	0.18	0.12	0.19
1	0.13	0.36	0.33	0.24	0.12	0.26
2	0.27	0.36	0.44	0.29	0.25	0.34
3	0.27	0.09	0.11	0.29	0.50	0.21
Sum	1.00	1.00	1.00	1.00	1.00	1.00

Exemplo 5.10

Uma forma de resumir é calcular as **frequências esperadas**, multiplicando os totais de coluna pelas proporções obtidas pela **distribuição marginal** de T

	0	1	2	3	4	Sum
0	2.81	4.12	3.38	3.19	1.5	15
1	3.94	5.78	4.73	4.46	2.1	21
2	5.06	7.43	6.08	5.74	2.7	27
3	3.19	4.67	3.82	3.61	1.7	17
Sum	15.00	22.00	18.00	17.00	8.0	80

Exemplo 5.10

Agora podemos quantificar as diferenças entre as **frequências observadas** (o_{ij}), e as **frequências esperadas** (e_{ij}) através de

$$Q^2 = \sum_{i,j} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

Dessa forma, temos:

$$Q^2 = \frac{(5 - 2.81)^2}{2.81} + \cdots + \frac{(4 - 1.7)^2}{1.7} = 12.63$$

Se as frequências esperadas fossem muito próximas das observadas, esperaríamos que esse valor fosse próximo de zero.

Como o valor é relativamente alto, há uma indicação de que as duas variáveis são dependentes.

Correlação entre variáveis num conjunto de dados

- Considere um conjunto de dados com n pares de valores para as variáveis X e Y. O coeficiente de correlação mede a dependência linear entre as variáveis e é calculado por

$$\rho_{XY} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{[\sum_{j=1}^{n}(x_j - \bar{x})^2][\sum_{j=1}^{n}(y_j - \bar{y})^2]}}.$$

- Formula mais conveniente para cálculos

$$\rho_{XY} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sqrt{[\sum_{j=1}^{n} x_j^2 - n\bar{x}^2][\sum_{j=1}^{n} y_j^2 - n\bar{y}^2]}}.$$

- Note que $-1 \leq \rho_{XY} \leq 1$.
- Observação: $\rho_{XY} = 0$ não indica independência.

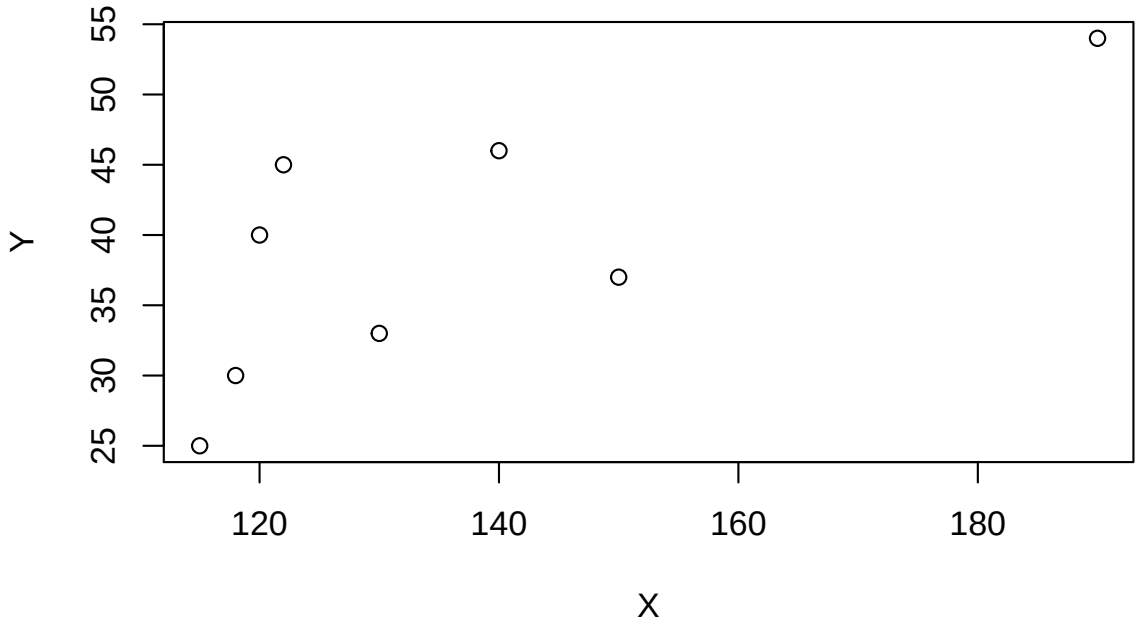
Exemplo 5.11

- A quantidade de chuva é um fator importante na produtividade agrícola. Para medir esse efeito foram anotados, para 8 diferentes regiões produtoras de soja, o índice pluviométrico em milímetros (X) e a produção do último ano em toneladas (Y). Determine o coeficiente de correlação.

X	120	140	122	150	115	190	130	118
Y	40	46	45	37	25	54	33	30

- Exemplo 5.12 tarefa de casa.

Exemplo 5.11



$$\rho_{XY} = \frac{43245 - 8 \times 135.63 \times 38.75}{\sqrt{[151533 - 8 \times 135.63^2][12640 - 8 \times 38.75^2]}} = 0,73$$

Propriedades de esperança de VAs

Para podermos definir medidas de dependência entre VAs discretas, precisamos das seguintes propriedades de esperança de VAs.

Para duas VAs X e Y, **independentes**, segue que

$$E(X+Y) = E(X) + E(Y)$$

$$E(XY) = E(X)E(Y)$$

Importante

X e Y independentes $\Rightarrow E(XY) = E(X)E(Y)$
No entanto:
$E(XY) = E(X)E(Y) \not\Rightarrow X$ e Y independentes
[Ver exemplo 5.13]

Covariância de duas VAs

Uma medida de dependência linear entre X e Y é a covariância:

$$Cov(X, Y) = \sigma_{XY} = E[(X - \mu_X)(Y - \mu_y)].$$

Uma forma alternativa (mais fácil de calcular) é:

$$Cov(X, Y) = \sigma_{XY} = E(XY) - E(X)E(Y)$$

Variância da soma de duas VAs

$$Var(X + Y) = Var(X) + Var(Y) + 2Cov(X, Y)$$

Exemplo 5.14

- As variáveis U e V têm a seguinte distribuição conjunta.

U/V	2	4	6	8	10	$P(U = u)$
2	0.1	0	0	0	0	0.1
3	0	0.2	0	0.1	0	0.3
4	0	0	0.2	0	0	0.2
5	0	0.1	00	0.2	0	0.3
6	0	0	0	0	0.1	0.1
$P(V = v)$	0.1	0.3	0.2	0.3	0.1	1

- Calcule a covariância entre U e V.

Exemplo 5.14

Marginais de U, V, e UV:

U	2.0	3.0	4.0	5.0	6.0
pU	0.1	0.3	0.2	0.3	0.1

V	2.0	4.0	6.0	8.0	10.0
pV	0.1	0.3	0.2	0.3	0.1

UV	4.0	12.0	20.0	24.0	40.0	60.0
pUV	0.1	0.2	0.1	0.3	0.2	0.1

Cálculo da covariância:

$E(U) = 4$

$E(V) = 6$

$E(UV) = 26$

$$\begin{aligned} Cov(U, V) &= E(UV) - E(U)E(V) \\ &= 26 - 24 \\ &= 2 \end{aligned}$$

Correlação de duas VAs

O **coeficiente de correlação** entre as VAs discretas X e Y é calculado por:

$$\rho_{XY} = \frac{Cov(X, Y)}{\sigma_X \sigma_Y}$$

- A divisão pelo produto dos desvios padrão serve para padronizar a medida
- Permite comparação entre quaisquer outras variáveis, pois $-1 \leq \rho_{XY} \leq 1$
- Valores mais próximos de ± 1 indicam correlação forte

Exemplo 5.15

- Para os dados do exemplo 5.5 calcule a covariância e a correlação.

Sub-região	1	2	3	4	5	6	7	8	9	10
X	0	0	0	0	1	2	1	2	2	0
Y	1	2	1	0	1	0	0	1	2	2

Exemplo 5.15

Anteriormente já obtivemos a conjunta e as marginais de X e Y:

X/Y	0	1	2	P(X=x)
0	0.1	0.2	0.2	0.5
1	0.1	0.1	0.0	0.2
2	0.1	0.1	0.1	0.3
P(Y=y)	0.3	0.4	0.3	1.0

A marginal de XY é

XY	0.0	1.0	2.0	4.0
pXY	0.7	0.1	0.1	0.1

Exemplo 5.15

Com isso:

$$E(X) = 0.8 \quad E(Y) = 1 \quad E(XY) = 0.7$$

$$Var(X) = \sigma_X^2 = 0.76 \quad Var(Y) = \sigma_Y^2 = 0.6$$

Assim, a covariância será:

$$Cov(X, Y) = E(XY) - E(X)E(Y) = 0.7 - 0.8 \times 1 = -0.1$$

E a correlação será:

$$\rho_{XY} = \frac{Cov(X, Y)}{\sigma_X \sigma_Y} = \frac{-0.1}{\sqrt{0.76}\sqrt{0.6}} = -0.15$$

Sumário

Exercícios recomendados

- Seção 5.1 - $1, 2, 3, 4$ e 6.
- Seção 5.2 - $1, 2, 3, 4$ e 5.