Aula 5 - Variáveis bidimensionais

PhD. Wagner Hugo Bonat

Laboratório de Estatística e Geoinformação-LEG
Universidade Federal do Paraná

1/2017

Variáveis bidimensionais

- Interesse no comportamento conjunto de várias variáveis.
- 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

- Conjunta, marginal de X e marginal de Y.

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.

Laboratório de Estatística e Geoinformação

Função de probabilidade conjunta

- Sejam X e Y duas va discretas originárias do mesmo fenômeno aleatório, com valores atribuídos a partir do mesmo espaço amostral. A função de probabilidade conjunta é definada, para todos os possíveis pares de valores (X,Y), da seguinte forma:

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

Exemplo 5.4

- Uma empresa atende encomendas de supermercados dividindo os pedidos em duas parter 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-regiao	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.

Associação entre variáveis

- 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 <- c(7,7,7,7,8,8,8,8,8,8,8,9,9,9,10)
> P <- c(8,6,8,9,8,6,9,7,7,6,7,8,9,8,8)
> M <- c(5,6,7,5,5,5,6,4,7,6,5,5,6,5,5)
> tab <- data.frame(I,P,M)
> t(tab)
```

```
  [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11] [,12] [,13] [,14] [,15]
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
```

Tabelas de frequência conjunta

- Inglês e Português:

```
> table(I,P)

    P
I    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:

```
> table(I,M)

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

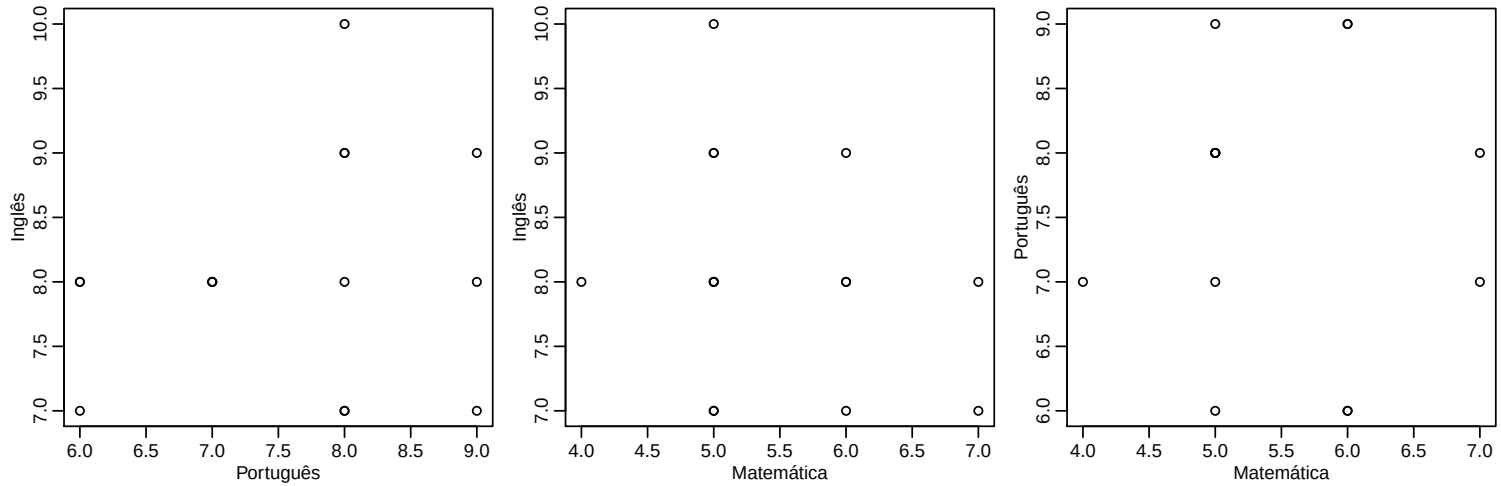
- Português e Matemática:

```
> table(P,M)

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

Diagramas de dispersão

```
> par(mfrow = c(1,3), mar=c(2.6, 2.8, 1.2, 0.5), mgp = c(1.6, 0.6, 0))
> plot(I ~ P, ylab = "Inglês", xlab = "Português")
> plot(I ~ M, ylab = "Inglês", xlab = "Matemática")
> plot(P ~ M, ylab = "Português", xlab = "Matemática")
```



Probabilidade condicional para v.a 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 v.a 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). Deseja-se verificar se essas variáveis estão associadas.

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 5.10 tarefa de casa.

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

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

Correlação entre variáveis num conjunto de dados

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

```
> X <- c(120,140,122,150,115,190,130,118)
> Y <- c(40,46,45,37,25,54,33,30)
> sum(X) ; sum(X^2) ; sum(Y) ; sum(Y^2) ; sum(X*Y) ; cor(X,Y)

[1] 1085

[1] 151533

[1] 310

[1] 12640

[1] 43245

[1] 0.7245956
```

- Exemplo 5.12 tarefa de casa.

Covariância de duas v.a

- Dependência linear entre X e Y é a covariância:

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

- Exemplo 5.14 - As variáveis X e Y têm a seguinte distribuição conjunta.

X - Y	(2,2)	(3,4)	(3,8)	(4,6)	(5,4)	(5,8)	(6,10)
p(x,y)	0,1	0,2	0,1	0,2	0,1	0,2	0,1

- Calcule a covariância entre X e Y.

Correlação de duas v.a

- Dependência linear entre X e Y é a covariância:

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

- Exemplo 5.15 - Para os dados do exemplo 5.5 calcule a covariância e a correlação.

X - Y	0	1	2
0	1/10	2/10	2/10
1	1/10	1/10	0
2	1/10	1/10	1/10

Resultados importantes

- $E(X+Y) = E(X) + E(Y)$.
- $E(XY) = E(X)E(Y)$ se e somente se X e Y são independentes.
- $Var(X+Y) = Var(X) + Var(Y) - 2Cov(X,Y)$.

Exercícios recomendados

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