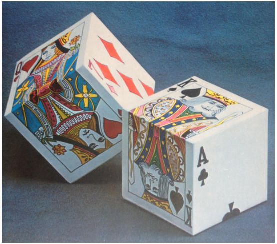
Fundamentos Estatísticos para Ciência dos Dados
Vetores aleatórios

Renato Martins Assunção

DCC, UFMG - 2015



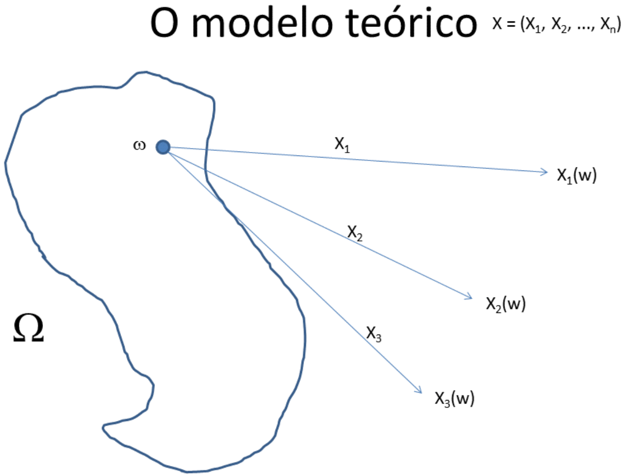
Vetores de v.a.'s

- $\mathbf{X} = (X_1, \ldots, X_k)$ é um vetor aleatório de dimensão k.
- Cada uma das entradas X_i do vetor $\mathbf{X}$ é uma variável aleatória medida no mesmo resultado ω do experimento estocástico.
- A importância vital de se lidar com vetores aleatórios: uma v.a. vai dar alguma informação sobre o valor de outra v.a.

Variáveis no mesmo espaço amostral

- O arcabouço matemático é o seguinte:
- Temos um espaço amostral Ω com uma medida de probabilidade sobre subconjuntos $\mathcal{A}$ de Ω.
- Ω é "complexo": cada resultado do experimento aleatório pode ter muitas características de interesse: $X_1, X_2, X_3, \ldots$
- Coletamos estas várias medidas num vetor aleatório $\mathbf{X} = (X_1, \ldots, X_n)$.
- As variáveis $X_1, X_2, X_3, \ldots$ são medições feitas no MESMO RESULTADO $\omega \in \Omega$ do experimento.

O modelo teórico



Exemplo: imagens

- $\Omega = \{$ imagens $n \times m\}$
- Selecione ao acaso uma das imagens.
- Características de interesse: intensidade de cinza em cada um dos pixels
- $\mathbf{X} = (X_{11}, X_{12}, \ldots, X_{nm})$
- Veja que todas as medições são sobre um mesmo resultado do experimento: a imagem selecionada.

Exemplo: grafos

- Considere uma rede social vista como um grafo direcionado.
- Os nós são os usuários e as arestas direcionadas são as relações de seguidor-seguido.
- Experimento consiste em selecionar um nó ao acaso.
- Ω é a coleção de nós do grafo (com suas características associadas)
- k características de interesse do nó selecionado
  - X_1 = idade do nó (intrínseco ao nó)
  - X_2 = número de outlinks (relacional)
  - X_3 = número de inlinks (relacional)
- $\mathbf{X} = (X_1, X_2, X_3)$
- Objetivo: Qual é a relação probabilística entre o número de outlinks do nó com a idade do nó?

Exemplo: problema de classificação

- $\Omega =$ coleção de itens classificados em dois (ou mais) grupos
    - Coleção de e-mails: spam versus não-spam
    - Coleção de tomadores de empréstimos num banco: pagam versus n ao pagam de volta o empréstimo dentro do prazo
    - crânios humanos em escavação arqueológica: masculinos versus femininos
- A coleção pode nem existir ainda de forma completa: interesse nos e-mails já enviados mas principalmente nos que ainda serão enviados NO FUTURO

Classificação

- Em cada item da coleção, medimos dois tipos de variáveis aleatórias.
- No caso de spam x não-spam:
- Y: binária, 0 ou 1 (spam ou não-spam, inadimplente ou não)
- Um conjunto de $k = 3$ atributos:
  - X_1: número de vezes que aparece a palavra "sale"
  - X_2 : número de vezes que aparece a palavra "offer"
  - X_3 : número de vezes que aparece a palavra "Viagra"
- Vetor aleatório: $\mathbf{X} = (Y, X_1, X_2, X_3)$
- Objetivo: predizer o valor de Y a partir dos atributos.
- Qual a probabilidade condicional $\mathbb{P}(Y = \text{ spam } | X_1 = 3, X_2 = 1, X_3 = 3)$
- Como esta probabilidade muda quando alteramos alguns dos atributos X?

Regressão: preço de imóveis

- Alguns apartamento custam 200 mil reais, outros curtam 10 vezes mais.
- O que faz com que os preços Y de apartamentos variem tanto?
- X_1 = Localização, X_2 = idade, X_3 = área, X_4 = número de quartos, X_5 = piscina?.
- Vetor aleatório: $\mathbf{X} = (Y, X_1, X_2, X_3, X_4, X_5)$
- Interesse em conhecer a distribuição da variável aleatória Y = preço CONDICIONADA no valor das demais variáveis:
- Por exemplo, saber a distribuição de

$$(Y|X_1 = Sion, X_2 = 10anos, X_3 = 200m^2, X_4 = 4, X_5 = \text{ não})$$

- Como esta distrbuição muda quando alteramos alguns dos atributos X?

Explorando interelações

- $\mathbf{X} = (X_1, X_2, \ldots, X_k)$
- Podemos analisar cada v.a. separadamente das demais e ajustar um modelo a cada uma delas (binomial, Poisson, Pareto, normal,...)
- Isto é chamado de análise *marginal*.
- É o que viemos fazendo até agora.
- O mais interessante é quando analisamos as variáveis *conjuntamente*.
- Análise conjunta procura explorar a existência de relações probabilísticas entre as variáveis.

Conjunta: caso discreto

- Se $\mathbf{X} = (X_1, X_2, \ldots, X_k)$ é composto apenas de v.a.'s discretas, a distribuição conjunta das v.a.'s é simples.
- Como no caso de uma v.a. apenas, precisamos apenas especificar uma lista de valores possíveis para o VETOR $\mathbf{X}$ e a lista de probabilidades associadas.
- Se X_i tem m_i valores possíveis, a lista de valores possíveis do vetor $\mathbf{X} = (X_1, X_2, \ldots, X_k)$ terá $m_1 \times m_2 \times \ldots m_k$ possibilidades.
- Basta agora atribuir uma probabilidade ≥ 0 a cada um deles de forma que somem 1.
- Todas as probabs de interesse são obtidas a partir desta lista de probabilidades básicas.

Exemplo muito simples

- $\Omega =$ conjunto de pacientes em visita ao otorrinolaringologistaÂ com problemas na garganta (faringoamigdalite aguda)
- Incluímos em Ω os pacientes DO FUTURO
- Suspeita da presença de infecção pela bactéria estreptococcus
- Dois tipos de testes em CADA PACIENTE:
  - Teste padrão-ouro, cultura em placa agar-sangue: resultado positivo ou negativo
  - Teste rápido, barato com resultados positivo ou negativo MAS com menor qualidade.

Validação de teste diagnóstico

- Vetor $\mathbf{X} = ($ Teste-ouro, Teste-rápido$) = (TO, TR)$
- V.a. TO possui dois valores: 0 ou 1
- V.a. TR também possui dois valores: 0 ou 1
- Vetor $\mathbf{X}$ possui 4 resultados possíveis

Teste-ouro	Teste-rápido	Probabilidade
0	0	?
0	1	?
1	0	?
1	1	?

Validação de teste diagnóstico

- Probabilidades $\mathbb{P}(TO = x_1, TR = x_2)$ dos 4 resultados possíveis

Teste-ouro	Teste-rápido	Probabilidade
0	0	0.40
0	1	0.19
1	0	0.03
1	1	0.38
Total		1

- Probabilidades devem ser ≥ 0 e devem somar 1

Distribuição Marginal

- *Distribuição Marginal* de cada v.a.: some a probab conjunta sobre todos os valores das outras variáveis.

Teste-ouro	Teste-rápido	Probabilidade
0	0	0.40
0	1	0.19
1	0	0.03
1	1	0.38
Total		1

- Distribuição marginal da v.a. TO: Precisamos de $\mathbb{P}(TO = 0)$ e de $\mathbb{P}(TO = 1)$

$$
\begin{aligned}
\mathbb{P}(TO = 0) &= \mathbb{P}(TO = 0 \wedge TR = 0) + \mathbb{P}(TO = 0 \wedge TR = 1) \\
&= 0.40 + 0.19 = 0.59 \\
\mathbb{P}(TO = 1) &= \mathbb{P}(TO = 1 \wedge TR = 0) + \mathbb{P}(TO = 1 \wedge TR = 1) \\
&= 0.03 + 0.38 = 0.41
\end{aligned}
$$

Marginal de TR

- Distribuição marginal da v.a. TR
- Precisamos de $\mathbb{P}(TR = 0)$ e de $\mathbb{P}(TR = 1)$

Teste-ouro	Teste-rápido	Probabilidade
0	0	0.40
0	1	0.19
1	0	0.03
1	1	0.38
Total		1

$$
\begin{aligned}
\mathbb{P}(TR = 0) &= \mathbb{P}(TR = 0 \wedge TO = 0) + \mathbb{P}(TR = 0 \wedge TO = 1) \\
&= 0.40 + 0.03 = 0.43 \\
\mathbb{P}(TR = 1) &= \mathbb{P}(TR = 1 \wedge TO = 0) + \mathbb{P}(TR = 1 \wedge TO = 1) \\
&= 0.19 + 0.38 = 0.57
\end{aligned}
$$

- Na prática, como $\mathbb{P}(TR = 1) = 1 - \mathbb{P}(TR = 0)$, basta obtermos uma delas, a outra sendo obtida por subtração.

Independência de v.a's

- Duas v.a.'s DISCRETAS X e Y são *independentes* se os eventos $[X = x]$ e $[Y = y]$ são independentes para qualquer combinaçãode x e y.
- Isto é, X e Y são independentes se

$$\mathbb{P}(X = x, Y = y) = \mathbb{P}(X = x)\ \mathbb{P}(Y = y)$$

para todo par (x, y).
- Equivalentemente, X e Y são independentes se

$$\mathbb{P}(X = x | Y = y) = \mathbb{P}(X = x)$$

para todo par (x, y).

TO e TR não são independentes

- A tabela original

Teste-ouro	Teste-rápido	Probabilidade
0	0	0.40
0	1	0.19
1	0	0.03
1	1	0.38
Total		1

- Verificando dois casos:

$$\mathbb{P}(TO = 1, TR = 1) = 0.38 \neq 0.23 = (0.03+0.38)\times(0.19+0.38) = \mathbb{P}(TO = 1)\,\mathbb{P}(TR = 1)$$

$$\mathbb{P}(TO = 0, TR = 0) = 0.40 \neq 0.25 = (0.40+0.19)\times(0.40+0.03) = \mathbb{P}(TO = 0)\,\mathbb{P}(TR = 0)$$

- Se fossem independentes, $TO = 1$ ocorreria junto com $TR = 1$ apenas 23% das vezes mas eles ocorrem juntos 38% de acordo com a tabela. $TO = 0$ e $TR = 0$ ocorreriam juntos 25% das vezes se independentes mas tabela fornece 40%.
- Os dois testes tendem a concordar muito mais frequentemente do que se fossem independentes.

Com mais de duas variáveis discretas

- Suponha que $\mathbf{X} = (X_1, X_2, X_3, X_4)$ onde
  - $X_1 =$ diagnóstico de uma doença, presente (1) ou ausente (0)
  - $X_2 =$ sexo, masculino (1) ou feminino (0)
  - $X_3 =$ idade, classificada em três categorias:
    criança (1), adulto jovem (2) , idoso (3)
  - $X_4 =$ fumante (1) ou não-fumante (0)
- Existem $2 \times 2 \times 3 \times 2 = 24$ valores possíveis para $\mathbf{X}$
- Precisamos alocar probabilidades aos 24 valores possíveis, todas não-negativas e somando 1.

Distribuição conjunta

X_1	X_2	X_3	X_4	$100\% \times \mathbb{P}$
0	0	1	0	4.6
0	0	1	1	6.7
0	0	2	0	4.1
0	0	2	1	5.3
0	0	3	0	5.2
0	0	3	1	6.6
0	1	1	0	1.6
0	1	1	1	1.8
0	1	2	0	4.9
0	1	2	1	0.2
0	1	3	0	3.1
0	1	3	1	4.4
1	0	1	0	1.1
1	0	1	1	6.8
1	0	2	0	0.5
1	0	2	1	4.0
1	0	3	0	4.0
1	0	3	1	2.9
1	1	1	0	3.6
1	1	1	1	3.7
1	1	2	0	6.6
1	1	2	1	6.8
1	1	3	0	6.6
1	1	3	1	4.9
Total				100.0

Distribuição marginal de X_1

$$\mathbb{P}(X_1 = 0) = \sum_{i,j,k} \mathbb{P}(X_1 = 0, X_2 = i, X_3 = j, X_4 = k)$$
$$= (4.6 + 5.7 + 4.1 + 5.3 + 5.2 + 6.6 +$$
$$+1.6 + 1.8 + 4.9 + 0.2 + 3.1 + 4.4)\,/100$$
$$= 0.475$$

X_1	X_2	X_3	X_4	$100\% \times \mathbb{P}$
0	0	1	0	4.6
0	0	1	1	6.7
0	0	2	0	4.1
0	0	2	1	5.3
0	0	3	0	5.2
0	0	3	1	6.6
0	1	1	0	1.6
0	1	1	1	1.8
0	1	2	0	4.9
0	1	2	1	0.2
0	1	3	0	3.1
0	1	3	1	4.4
1	0	1	0	1.1
1	0	1	1	6.8
1	0	2	0	0.5
1	0	2	1	4.0
1	0	3	0	4.0
1	0	3	1	2.9
1	1	1	0	3.6
1	1	1	1	3.7
1	1	2	0	6.6
1	1	2	1	6.8
1	1	3	0	6.6
1	1	3	1	4.9

Distribuição marginal de X_1

$$\mathbb{P}(X_1 = 0) = 0.475 \qquad \text{já calculado}$$

Por subtração encontramos $\mathbb{P}(X_1 = 1)$:

$$\mathbb{P}(X_1 = 1) = 1 - \mathbb{P}(X_1 = 0) = 1 - 0.475 = 0.525$$

X_1	X_2	X_3	X_4	$100\% \times \mathbb{P}$
0	0	1	0	4.6
0	0	1	1	6.7
0	0	2	0	4.1
0	0	2	1	5.3
0	0	3	0	5.2
0	0	3	1	6.6
0	1	1	0	1.6
0	1	1	1	1.8
0	1	2	0	4.9
0	1	2	1	0.2
0	1	3	0	3.1
0	1	3	1	4.4
1	0	1	0	1.1
1	0	1	1	6.8
1	0	2	0	0.5
1	0	2	1	4.0
1	0	3	0	4.0
1	0	3	1	2.9
1	1	1	0	3.6
1	1	1	1	3.7
1	1	2	0	6.6
1	1	2	1	6.8
1	1	3	0	6.6
1	1	3	1	4.9

Distribuição marginal de X_3

$$
\begin{aligned}
\mathbb{P}(X_3 = 1) &= \sum_{i,j,k} \mathbb{P}(X_1 = i, X_2 = j, X_3 = 1, X_4 = k) \\
&= (4.6 + 6.7 + 1.6 + 1.8 \\
&\quad +1.1 + 6.8 + 3.6 + 3.7)\ /100 \\
&\quad = 0.299
\end{aligned}
$$

X_1	X_2	X_3	X_4	$100\% \times \mathbb{P}$
0	0	1	0	4.6
0	0	1	1	6.7
0	0	2	0	4.1
0	0	2	1	5.3
0	0	3	0	5.2
0	0	3	1	6.6
0	1	1	0	1.6
0	1	1	1	1.8
0	1	2	0	4.9
0	1	2	1	0.2
0	1	3	0	3.1
0	1	3	1	4.4
1	0	1	0	1.1
1	0	1	1	6.8
1	0	2	0	0.5
1	0	2	1	4.0
1	0	3	0	4.0
1	0	3	1	2.9
1	1	1	0	3.6
1	1	1	1	3.7
1	1	2	0	6.6
1	1	2	1	6.8
1	1	3	0	6.6
1	1	3	1	4.9

Distribuição marginal de X_3

$\mathbb{P}(X_3 = 1) = 0.299$ (já calculado)

$$\begin{aligned}\mathbb{P}(X_3 = 2) &= \sum_{i,j,k} \mathbb{P}(X_1 = i, X_2 = j, X_3 = 2, X_4 = k) \\ &= (4.1 + 5.3 = 4.9 + 0.2+ \\ &\quad +0.5 + 4.0 + 6.6 = 6.8)\ /100 \\ &\quad = 0.324\end{aligned}$$

X_1	X_2	X_3	X_4	$100\% \times \mathbb{P}$
0	0	1	0	4.6
0	0	1	1	6.7
0	0	2	0	4.1
0	0	2	1	5.3
0	0	3	0	5.2
0	0	3	1	6.6
0	1	1	0	1.6
0	1	1	1	1.8
0	1	2	0	4.9
0	1	2	1	0.2
0	1	3	0	3.1
0	1	3	1	4.4
1	0	1	0	1.1
1	0	1	1	6.8
1	0	2	0	0.5
1	0	2	1	4.0
1	0	3	0	4.0
1	0	3	1	2.9
1	1	1	0	3.6
1	1	1	1	3.7
1	1	2	0	6.6
1	1	2	1	6.8
1	1	3	0	6.6
1	1	3	1	4.9

Distribuição marginal de X_3

$\mathbb{P}(X_3 = 1) = 0.299$ (já calculado)

$\mathbb{P}(X_3 = 2) = 0.324$ (já calculado)

$$
\begin{aligned}
\mathbb{P}(X_3 = 3) &= 1 - \mathbb{P}(X_3 = 1) - \mathbb{P}(X_3 = 2) \\
&= 1 - \mathbb{P}(X_3 = 1) - \mathbb{P}(X_3 = 2) \\
&= 0.377
\end{aligned}
$$

X_1	X_2	X_3	X_4	$100\% \times \mathbb{P}$
0	0	1	0	4.6
0	0	1	1	6.7
0	0	2	0	4.1
0	0	2	1	5.3
0	0	3	0	5.2
0	0	3	1	6.6
0	1	1	0	1.6
0	1	1	1	1.8
0	1	2	0	4.9
0	1	2	1	0.2
0	1	3	0	3.1
0	1	3	1	4.4
1	0	1	0	1.1
1	0	1	1	6.8
1	0	2	0	0.5
1	0	2	1	4.0
1	0	3	0	4.0
1	0	3	1	2.9
1	1	1	0	3.6
1	1	1	1	3.7
1	1	2	0	6.6
1	1	2	1	6.8
1	1	3	0	6.6
1	1	3	1	4.9

Marginalização, em geral

- Seja $\mathbf{X} = (X_1, X_2, \ldots, X_k)$ um vetor de v.a.'s discretas
- X_i tem n_i valores possíveis
- Queremos $\mathbb{P}(X_1 = x)$ onde x é um dos seus n_1 valores possíveis.
- Para cada valor de x, a probabilidade $\mathbb{P}(X_1 = x)$ é uma soma de $n_2 \times n_3 \times \ldots \times n_k$ elementos da tabela de distribuição conjunta.
- Se todas as v.a.'s são binárias temos 2^{k-1} parcelas para cada valor de x.
- Se quisermos $\mathbb{P}(X_1 = x)$ para todos os n_1 valores x possíveis para X_1, precisamos fazer o cálculo anterior n_1 vezes.
- Na verdade, $n_1 - 1$ vezes pois

$$\mathbb{P}(X_1 = x_{n_1}) = 1 - \mathbb{P}(X_1 = x_1) - \ldots - \mathbb{P}(X_1 = x_{n_1 - 1})$$

Distribuição marginal de X_1 e X_2

$$\begin{aligned}
\mathbb{P}(X_1 = 0, X_2 = 1) &= \sum_{i,j} \mathbb{P}(X_1 = 0, X_2 = 1, X_3 = i, X_4 = j) \\
&= (1.6 + 1.8 + 4.9 + 0.2 + 3.1 + 4.4)/100 \\
&= 0.16
\end{aligned}$$

X_1	X_2	X_3	X_4	$100\% \times \mathbb{P}$
0	0	1	0	4.6
0	0	1	1	6.7
0	0	2	0	4.1
0	0	2	1	5.3
0	0	3	0	5.2
0	0	3	1	6.6
0	1	1	0	1.6
0	1	1	1	1.8
0	1	2	0	4.9
0	1	2	1	0.2
0	1	3	0	3.1
0	1	3	1	4.4
1	0	1	0	1.1
1	0	1	1	6.8
1	0	2	0	0.5
1	0	2	1	4.0
1	0	3	0	4.0
1	0	3	1	2.9
1	1	1	0	3.6
1	1	1	1	3.7
1	1	2	0	6.6
1	1	2	1	6.8
1	1	3	0	6.6
1	1	3	1	4.9

Independência de v.a's

- $\mathbf{X} = (X_1, \ldots, X_k)$ um vetor aleatório composto de v.a.'s discretas.
- Elas são *independentes* se

$$\mathbb{P}(X_1 = x_1, \ldots, X_k = x_k) = \mathbb{P}(X_1 = x_1) \ldots \mathbb{P}(X_k = x_k)$$

  para qualquer configuração de valores possíveis $(x_1, \ldots, x_k)$.
- Se o vetor é composto de v.a.'s independentes então

$$\mathbb{P}(X_1 = x_1 | X_2 = x_2, \ldots, X_k = x_k) = \mathbb{P}(X_1 = x_1)$$

  para qualquer configuração de valores possíveis $(x_1, \ldots, x_k)$.
- O resultado acima é válido se X_1 trocar de posição com qualquer outra v.a.

Simulação de um vetor aleatório discreto

- Seja $\mathbf{X} = (X_1, X_2, X_3, X_4)$ um vetor aleatório composto de v.a.'s discretas.
- Se quisermos simular (X), podemos usar o MESMO procedimento aprendido para UMA v.a. discreta.
- Simule $U \sim U(0,1)$ e veja em que segmento U caiu na coluna de soma acumulada.
- Este segmento determina o vetor $\mathbf{X}$ gerado.
- Por exemplo, se $U = 0.3215$ então $\mathbf{X} = (0, 0, 3, 1)$ é selecionado.
- A geração NÃO É feita separadamente para cada v.a. do vetor com base na sua distribuição marginal A MENOS que as v.a.'s sejam independentes.

X_1	X_2	X_3	X_4	$100\% \times \mathbb{P}$	Soma Acum.
0	0	1	0	4.6	4.6
0	0	1	1	6.7	11.3
0	0	2	0	4.1	15.4
0	0	2	1	5.3	20.7
0	0	3	0	5.2	25.9
0	0	3	1	6.6	32.5
0	1	1	0	1.6	34.1
0	1	1	1	1.8	35.9
0	1	2	0	4.9	40.8
0	1	2	1	0.2	41.0
0	1	3	0	3.1	44.1
0	1	3	1	4.4	48.5
1	0	1	0	1.1	49.6
1	0	1	1	6.8	56.4
1	0	2	0	0.5	56.9
1	0	2	1	4.0	60.9
1	0	3	0	4.0	64.9
1	0	3	1	2.9	67.8
1	1	1	0	3.6	71.4
1	1	1	1	3.7	75.1
1	1	2	0	6.6	81.7
1	1	2	1	6.8	88.5
1	1	3	0	6.6	95.1
1	1	3	1	4.9	100.0

Às vezes, array bi-dimensional

- No caso de termos apenas duas v.a.'s discretas, é comum apresentar a distribuição conjunta de probabilidade como um array de duas entradas.
- De volta aos dois testes de diagnósticos: teste-outo e teste-rápido

	$TR = 0$	$TR = 1$
$T0 = 0$	0.40	0.19
$T0 = 1$	0.03	0.38

- Colocamos os valores possíveis de X_1 nas linhas.
- Colocamos os valores de X_2 nas colunas.
- Na posição (i, j) do array colocamos a probabilidade $\mathbb{P}(X_1 = x_i, X_2 = x_j)$

Marginais nas margens

- Nas margens da tabela bi-dimensional colocamos as distribuições marginais da variável coluna e da variável-linha.
- A marginal de $T0$ é obtida somando as colunas:

	$TR = 0$	$TR = 1$	*Total*
$T0 = 0$	0.40	0.19	$\mathbb{P}(TO = 0) =$ $0.40 + 0.19 = 0.51$
$T0 = 1$	0.03	0.38	$\mathbb{P}(TO = 1) =$ $0.03 + 0.38 = 0.49$

- Isto explica o nome distribuição *marginal* para a distribuição de uma única variável: elas ficam nas margens da tabela.

Marginais nas margens

- Somando as linhas encontramos a marginal de TR:

	$TR = 0$	$TR = 1$	*Total*
$T0 = 0$	0.40	0.19	0.51
$T0 = 1$	0.03	0.38	0.49
Total	$\mathbb{P}(TR = 0) =$ $0.40 + 0.03 = 0.43$	$\mathbb{P}(TR = 1) =$ $0.19 + 0.38 = 0.57$	1.00

- A soma dos valores na marginal-linha ou na marginal-coluna é o total das probabilidades: 1.

Exemplo: mobilidade social

- Selecione um adulto brasileiro ω ao acaso em 1988.
- Para cada ω, vamos definir duas v.a.'s:
  - $SF(\omega)$: o status sócio-econômico da sua ocupação (6 valores): 1,2,..,6
  - Ocupações categorizadas de acordo com características de renda e educação.
  - Baixo inferior, Baixo superior, Médio inferior, Médio, Médio superior, Alto.
  - $SP(\omega)$: status social da ocupação de seu pai quando o pai tinha 45 anos (6 valores): 1,2,...,6
  - Mesmas categorias que filho.
- Por exemplo, executivos e juízes de tribunais superiores estavam na categoria *Alto*
- Trabalhadores braçais exigindo nenhuma instrução estavam na categoria *Baixo inferior*.

Arcabouço para mobilidade social

- Selecione indivíduo ω ao acaso em 1988.
- Para cada indivíduo, meça o vetor $\mathbf{X} = (SF, SP)$
- 36 valores possíveis para o vetor $\mathbf{X}$.

$$
\begin{aligned}
\theta_{ij} &= \mathbb{P}(\text{ pai ter status } i \ \wedge \text{ filho ter status } j) \\
&= \mathbb{P}(SP = i, SF = j)
\end{aligned}
$$

- Claramente, SP e SF não são v.a.'s independentes.
- Existe uma grande inércia na sociedade: filhos de pais de status baixo tendem a continuar com status baixo.

A distribuição conjunta de (SP, SF)

- Probabilidades baseadas em amostra de 42137 homens chefes de família entre 20 e 64 anos em 1988
- Dados:IBGE-PNAD, Pesquisa Nacional por Amostra de Domicílios, 1988
- Probabilidades ($\times 100\%$) são aproximadas pela frequência relativa na amostra.

	SF: status do indivíduo em 1988.					
SP: status do pai	Baixo Inf.	Baixo Sup.	Médio Inf.	Médio	Médio Sup.	Alto
BI	21.7	12.8	13.2	4.6	2.1	1.0
BS	0.7	4.2	3.6	2.5	2.5	1.3
MI	0.6	3.7	7.1	2.7	2.7	1.5
M	0.6	1.9	2.0	2.2	1.2	0.9
MS	0,3	0.6	0.6	0.7	0.7	0.5
A	0.1	0.3	0.3	0.6	0.6	0.9

Questões de interesse

- Como mudou a distribuição do status entre duas gerações?
- Filhos de pais com status muito baixo passam com facilidade para um status mais alto?
- A estrutura de ocupação mudou drasticamente na década de 70 devido ao milagre econômico nos anos dos governs militares.
- Houve uma expansão da indústri e serviços e o Brasil deixou de ser uma sociedade agrária.
- Novos postos de trabalho qualificados foram abertos.
- Houve a necessidade de recrutar pessoas vindas de pais com status mais baixos.
- Quanto da mobilidade social pode ser explicada por esta expansão?

Distribuições marginais

- Distribuição marginal da estrutura de status das ocupações para a geração dos pais e dos filhos.

	SF: status do indivíduo em 1988 .						
SP: status do pai	Baixo Inf.	Baixo Sup.	Médio Inf.	Médio	Médio Sup.	Alto	TOTAL
BI	21.7	12.8	13.2	4.6	2.1	1.0	55.4
BS	0.7	4.2	3.6	2.5	2.5	1.3	14.8
MI	0.6	3.7	7.1	2.7	2.7	1.5	18.3
M	0.6	1.9	2.0	2.2	1.2	0.9	8.8
MS	0.3	0.6	0.6	0.7	0.7	0.5	3.4
A	0.1	0.3	0.3	0.6	0.6	0.9	2.8
TOTAL	24.0	23.5	26.8	13.3	9.8	6.1	100%

Focando nas marginais

- Na geração dos pais, 55% das ocupações estavam no estrato *Baixo inferior* e isto foi reduzido a apenas 24% das ocupações na geração dos filhos.
- Há um deslocamento de ocupações em direção aos status mais elevados.
- Nos dois níveis de status mais elevados, a porcentagem passa de 6% para 16% entre as duas gerações.

	SF: status do indivíduo em 1988 .						
SP: status do pai	Baixo Inf.	Baixo Sup.	Médio Inf.	Médio	Médio Sup.	Alto	TOTAL
BI							55.4
BS							14.8
MI							18.3
M							8.8
MS							3.4
A							2.8
TOTAL	24.0	23.5	26.8	13.3	9.8	6.1	100%

Cálculos com a conjunta

- Seja A o evento "pai pobre, filho rico": o indivíduo tem status pelo menos *Médio superior* e seu pai tem status menor ou igual a *Baixo superior*.
- Temos

$$\mathbb{P}(A) = \underbrace{\mathbb{P}(SF \geq 5 \wedge SP \leq 2)}_{\text{4 células da tabela}} = \frac{2.1 + 1.0 + 2.5 + 1.3}{100} = 0.069 \text{ ou } 6.9\%$$

SP: status do pai	*SF*: status do indivíduo em 1988 . Baixo Inf.	Baixo Sup.	Médio Inf.	Médio	Médio Sup.	Alto	TOTAL
BI	21.7	12.8	13.2	4.6	2.1	1.0	55.4
BS	0.7	4.2	3.6	2.5	2.5	1.3	14.8
MI	0.6	3.7	7.1	2.7	2.7	1.5	18.3
M	0.6	1.9	2.0	2.2	1.2	0.9	8.8
MS	0.3	0.6	0.6	0.7	0.7	0.5	3.4
A	0.1	0.3	0.3	0.6	0.6	0.9	2.8
TOTAL	24.0	23.5	26.8	13.3	9.8	6.1	100%

Cálculos com a conjunta

- Seja B o evento reverso, "pai rico, filho pobre": o indivíduo tem status menor ou igual a *Baixo superior* e seu pai tem status pelo menos *Médio superior*.
- Temos

$$\mathbb{P}(B) = \underbrace{\mathbb{P}(SP \geq 5 \wedge SF \leq 2)}_{\text{4 células da tabela}} = \frac{0.3 + 0.6 + 0.1 + 0.3}{100} = \frac{1.3}{100} = 0.013 \text{ ou } 1.3\%$$

SP: status do pai	SF: status do indivíduo em 1988 . Baixo Inf.	Baixo Sup.	Médio Inf.	Médio	Médio Sup.	Alto	TOTAL
BI	21.7	12.8	13.2	4.6	2.1	1.0	55.4
BS	0.7	4.2	3.6	2.5	2.5	1.3	14.8
MI	0.6	3.7	7.1	2.7	2.7	1.5	18.3
M	0.6	1.9	2.0	2.2	1.2	0.9	8.8
MS	0.3	0.6	0.6	0.7	0.7	0.5	3.4
A	0.1	0.3	0.3	0.6	0.6	0.9	2.8
TOTAL	24.0	23.5	26.8	13.3	9.8	6.1	100%

Distribuição condicional: caso discreto

- Seja $\mathbf{X} = (X_1, X_2, X_3, X_4)$ um vetor aleatório composto de v.a.'s discretas.
- Aprendemos a obter a distribuição MARGINAL de uma ou mais v.a.'s: some sobre os valores das demais v.a.'s
- Queremos agora a distibuição de algumas v.a.'s CONDICIONADA nos valores de uma ou mais das outras v.a.'s.
- Por exemplo, queremos a distribuição do vetor (X_1, X_2, X_4) DADO QUE $X_3 = 2$:

$$\mathbb{P}(X_1 = i, X_2 = j, X_3 = k | X_3 = 2)$$

para diferentes valores de i, j, k.

X_1	X_2	X_3	X_4	$100\% \times \mathbb{P}$
0	0	1	0	4.6
0	0	1	1	6.7
0	0	2	0	4.1
0	0	2	1	5.3
0	0	3	0	5.2
0	0	3	1	6.6
0	1	1	0	1.6
0	1	1	1	1.8
0	1	2	0	4.9
0	1	2	1	0.2
0	1	3	0	3.1
0	1	3	1	4.4
1	0	1	0	1.1
1	0	1	1	6.8
1	0	2	0	0.5
1	0	2	1	4.0
1	0	3	0	4.0
1	0	3	1	2.9
1	1	1	0	3.6
1	1	1	1	3.7
1	1	2	0	6.6
1	1	2	1	6.8
1	1	3	0	6.6
1	1	3	1	4.9

Elimine os casos em que $X_3 \neq 2$

- Queremos

$$\mathbb{P}(X_1 = i, X_2 = j, X_3 = k | X_3 = 2)$$

- Como $X_3 = 2$, podemos eliminar de consideração TODOS OS OUTROS resultados em que $X_3 \neq 2$.
- Este é novo conjunto de valores possíveis para o vetor $\mathbf{X}$, apenas aqueles em que X_3 possui o valor 2.
- Dentro deste novo "mundo", as probabilidades devem somar 1.
- Basta normalizarmos: divida os valores originais das probabilidades pela soma dos seus termos.

X_1	X_2	X_3	X_4	$100\% \times \mathbb{P}$
0	0	2	0	4.1
0	0	2	1	5.3
0	1	2	0	4.9
0	1	2	1	0.2
1	0	2	0	0.5
1	0	2	1	4.0
1	1	2	0	6.6
1	1	2	1	6.8

Renormalize

- Queremos

$$\mathbb{P}(X_1 = i, X_2 = j, X_3 = k | X_3 = 2)$$

- Eliminamos a coluna X_3 já que todos os seus valores agora são iguais a 2.
- A tabela resultante é a distribuição CONDICIONAL de (X_1, X_2, X_4) DADO QUE $X_3 = 2$.
- A distribuição de qualquer conjunto de v.a.'s condicionado nos valores das demais é obtido do mesmo modo.

X_1	X_2	X_3	X_4	$100\% \times \mathbb{P}$
0	0	2	0	4.1
0	0	2	1	5.3
0	1	2	0	4.9
0	1	2	1	0.2
1	0	2	0	0.5
1	0	2	1	4.0
1	1	2	0	6.6
1	1	2	1	6.8
Total				32.4

X_1	X_2	X_4	$100\% \times \mathbb{P}(\ldots \mid X_3 = 2)$
0	0	0	$100\%\ (4.1/32.4) = 12.7$
0	0	1	$100\%\ (5.3/32.4) = 16.4$
0	1	0	$100\%\ (4.9/32.4) = 15.1$
0	1	1	$100\%\ (0.2/32.4) = 0.6$
1	0	0	$100\%\ (0.5/32.4) = 1.5$
1	0	1	$100\%\ (4.0/32.4) = 12.3$
1	1	0	$100\%\ (6.6/32.4) = 20.4$
1	1	1	$100\%\ (6.8/32.4) = 21.0$
Total			100%

Distibuição de $(X_1, X_4 | X_2 = 0, X_3 = 2)$

- Queremos agora a distibuição de $(X_1, X_4 | X_2 = 0, X_3 = 2)$
- Elimine todas as linhas da tabla original em que $X_2 \neq 0$ OU que $X_3 \neq 2$.
- Renormalize as linhas restantes e simplifique a tabela.

X_1	X_2	X_3	X_4	$100\% \times \mathbb{P}$
0	0	2	0	4.1
0	0	2	1	5.3
1	0	2	0	0.5
1	0	2	1	4.0

Distibuição de $(X_1, X_4 | X_2 = 0, X_3 = 2)$

- Queremos agora a distibuição de $(X_1, X_4 | X_2 = 0, X_3 = 2)$
- Renormalize as linhas restantes e simplifique a tabela.

X_1	X_2	X_3	X_4	$100\% \times \mathbb{P}$
0	0	2	0	4.1
0	0	2	1	5.3
1	0	2	0	0.5
1	0	2	1	4.0
Total				13.9

X_1	X_4	$100\% \times \mathbb{P}(X_1 = i, X_4 = j \| X_2 = 0, X_3 = 2)$
0	0	29.5
0	1	38.1
1	0	3.6
1	1	28.8
Total		13.9

Uma visão algébrica

- Queremos a distibuição de $(X_1, X_4 | X_2 = 0, X_3 = 2)$
- Isto é, queremos as probabilidades $\mathbb{P}(X_1 = i, X_4 = j | X_2 = 0, X_3 = 2)$ para toda combinação de i, j.
- Pela definição de probabilidade condicional:

$$\mathbb{P}(X_1 = i, X_4 = j | X_2 = 0, X_3 = 2) = \frac{\mathbb{P}(X_1 = i, X_4 = j, X_2 = 0, X_3 = 2)}{\mathbb{P}(X_2 = 0, X_3 = 2)}$$

- O numerador são o elementos que restam na tabela das probabilidades originais
- O denominador é o fator de normalização já que

$$\mathbb{P}(X_2 = 0, X_3 = 2) = \sum_{k,l} \mathbb{P}(X_1 = k, X_4 = l, X_2 = 0, X_3 = 2)$$

- Assim, esta visão gráfica de eliminar as linhas da tabelas etc. corresponde a esta operação algébrica.

De volta a mobilidade social

- Distribuição condicional do SP dado que $SF =$ Alto
- Dado que o filho está na elite, de onde ele veio?

	SF: status do indivíduo em 1988.						
SP: status do pai	Baixo Inf.	Baixo Sup.	Médio Inf.	Médio	Médio Sup.	Alto	TOTAL
BI	21.7	12.8	13.2	4.6	2.1	1.0	55.4
BS	0.7	4.2	3.6	2.5	2.5	1.3	14.8
MI	0.6	3.7	7.1	2.7	2.7	1.5	18.3
M	0.6	1.9	2.0	2.2	1.2	0.9	8.8
MS	0.3	0.6	0.6	0.7	0.7	0.5	3.4
A	0.1	0.3	0.3	0.6	0.6	0.9	2.8
TOTAL	24.0	23.5	26.8	13.3	9.8	6.1	100%

Distribuição condicional

- Distribuição condicional do SP dado que $SF =$ Alto
- Dado que o filho está na elite, de onde ele veio?

$$\mathbb{P}(SP = i | SF = 6) = \frac{\mathbb{P}(SP = i \text{ e } SF = 6)}{\mathbb{P}(SF = 6)} = \text{cte } \mathbb{P}(SP = i, SF = 6)$$

- Basta normalizar os números da coluna 6 para que somem 1:

Alto
1.0
1.3
1.5
0.9
0.5
0.9
$\sum = 6.1$

$\longrightarrow$

Alto
0.16
0.21
0.25
0.15
0.08
0.15
$\sum = 1$

Esses valores em cinza são os valores de
$\mathbb{P}(Y_1 = i | Y_2 = 6)$
para os diferentes valores de i.

- Note que $\mathbb{P}(SP = 1 | SF = 6) = 0.20$, isto é, 20% da elite veio dos estratos mais baixos.

Na outra direção agora

- $\mathbb{P}(SF = j | SP = 1)$: dado que o pai era lavrador manual ou similar, aonde foram parar seus filhos?

$$\mathbb{P}(SF = j | SP = 1) = \frac{\mathbb{P}(SF = j \text{ e } SP = 1)}{\mathbb{P}(SP = 1)} \; \alpha \; \mathbb{P}(SF = j | SP = 1)$$

- Estes números correspondem á linha 1 da tabela de probabilidade conjunta. Basta normalizá-los:

B.I.	21.7	12.8	13.2	4.6	2.1	1.0

	j=1	j=2	j=3	j=4	j=5	j=6
$\mathbb{P}(SF = j \vert SP = 1)$	0.39	0.23	0.24	0.08	0.04	0.02

- Assim, $\mathbb{P}(SF = 6 | SP = 1) = 0.02$, mas $\mathbb{P}(SP = 1 | SF = 6) = 0.20$, uma ordem de grandeza de diferença!

Como explicar esta disparidade?

- A enorme massa de 55% de pais de baixo status enviou apenas 2% de seus filhos para a elite.
- Mas 2% de 55% formam 1% da <u>populaçã total</u>.
- A elite da geração dos filhos forma 5% da população total.
- Estes 5% da pop total dividem-se em 1% vindos de pais de status baixo e os outros 4% vindos de pais com status maior.
- Assim, estes 1% *dentre os 5% da elite de hoje* formam os 20% da elite que veio de baixo na pirâmide social.

DISTRIBUIÇÃO CONDICIONAL de X

- Vetor aleatório $\mathbf{X} = (X_1, X_2, \ldots, X_k)$.
- Queremos a DISTRIBUIÇÃO de probabilidade da v.a. X_1 dados os valores das demais.
- Por exemplo, queremos a distribuição de X_1 quando $X_2 = 0, \ldots, X_k = 2$.

$$(X_1 | X_2 = 0, \ldots, X_k = 2) \sim ???$$

- O que é a distribuição de uma v.a. discreta? Duas coisas...
  - Lista $\{a_1, \ldots, a_m\}$ dos valores possíveis de X_1 QUANDO $X_2 = 0, \ldots, X_k = 2$
  - As probabilidades associadas QUANDO $X_2 = 0, \ldots, X_k = 2$):

$$\mathbb{P}(X_1 = a_i | X_2 = 0, \ldots, X_k = 2)$$

- Ao mudar os valores condicionados de $X_2, \ldots, X_k$ esta distibuição também muda: a distribuição é função dos valores em que estamos condicionando as demais variáveis $X_2, \ldots, X_k$.

DISTRIBUIÇÂO CONDICIONAL de X

- Queremos $\mathbb{P}(X_1 = a_i | X_2 = 0, \ldots, X_k = 2)$
- Para estes valores fixos $X_2 = 0, \ldots, X_k = 2$ das variáveis condicionantes, a distribuição é encontrada pela fórmula de probabilidade condicional:

$$\mathbb{P}(X_1 = a_i | X_2 = 0, \ldots, X_k = 2) = \frac{\mathbb{P}(X_1 = a_i, X_2 = 0, \ldots, X_k = 2)}{\mathbb{P}(X_2 = 0, \ldots, X_k = 2)}$$

- Observe que o denominador não depende de a_i e portanto não varia com o valor de a_i.
- Isto é, se $a_i \neq a_j$, temos

$$\frac{\mathbb{P}(X_1 = a_i | X_2 = 0, \ldots, X_k = 2)}{\mathbb{P}(X_1 = a_j | X_2 = 0, \ldots, X_k = 2)} = \frac{\frac{\mathbb{P}(X_1 = a_i, X_2 = 0, \ldots, X_k = 2)}{\mathbb{P}(X_2 = 0, \ldots, X_k = 2)}}{\frac{\mathbb{P}(X_1 = a_j, X_2 = 0, \ldots, X_k = 2)}{\mathbb{P}(X_2 = 0, \ldots, X_k = 2)}} = \frac{\mathbb{P}(X_1 = a_i, X_2 = 0, \ldots, X_k = 2)}{\mathbb{P}(X_1 = a_j, X_2 = 0, \ldots, X_k = 2)}$$

- Assim, podemos enxergar a distibuiç ao condicional de uma v.a. diretamente da tabela original de probabilidade conjunta.

Visualizando a condicional na tabela da conjunta

- Queremos

$$\mathbb{P}(X_3 = a_i | X_1 = 0, X_2 = 1, X_4 = 0)$$

- X_3 possui 3 valores possíveis: 1, 2, 3
- Comparando a chance (condicional) de $X_3 = 1$ versus $X_3 = 2$

$$\frac{\mathbb{P}(X_3 = 1 | X_1 = 0, X_2 = 1, X_4 = 0)}{\mathbb{P}(X_3 = 2 | X_1 = 0, X_2 = 1, X_4 = 0)} =$$

$$= \frac{\mathbb{P}(X_3 = 1, X_1 = 0, X_2 = 1, X_4 = 0)}{\mathbb{P}(X_3 = 2, X_1 = 0, X_2 = 1, X_4 = 0)} =$$

$$= \frac{1.6}{4.9} = 0.33$$

- Se $\mathbf{x}_1$ e $\mathbf{x}_2$ são dois dos vetores-valores possíveis para o vetor $\mathbf{X}$ e se $\mathbb{P}(\mathbf{X} = \mathbf{x}_1)$ for duas vezes maior que $\mathbb{P}(\mathbf{X} = \mathbf{x}_2)$ então esta razão ainda será respeitada entre as condicionais correspondentes.

X_1	X_2	X_3	X_4	$100\% \times \mathbb{P}$
0	0	1	0	4.6
0	0	1	1	6.7
0	0	2	0	4.1
0	0	2	1	5.3
0	0	3	0	5.2
0	0	3	1	6.6
0	1	1	0	1.6
0	1	1	1	1.8
0	1	2	0	4.9
0	1	2	1	0.2
0	1	3	0	3.1
0	1	3	1	4.4
1	0	1	0	1.1
1	0	1	1	6.8
1	0	2	0	0.5
1	0	2	1	4.0
1	0	3	0	4.0
1	0	3	1	2.9
1	1	1	0	3.6
1	1	1	1	3.7
1	1	2	0	6.6
1	1	2	1	6.8
1	1	3	0	6.6
1	1	3	1	4.9
Total				100.0

Esperança condicional

- Considere o vetor $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)$
- Calculamos a distribuição condicional de Y_1 dados os valores de $Y_2, \ldots, Y_p$
- Podemos calcular o valor esperado de Y_1 dados (ou fixados) os valores de $Y_2, \ldots, Y_p$
- É simplesmente como na definição usual de esperança de v.a.'s discretas:

$$\mathbb{E}(Y_1 | Y_2 = a_2, \ldots, Y_p = a_p) = \sum_y y \mathbb{P}(Y_1 = y | Y_2 = a_2, \ldots, Y_p = a_p)$$

- Média ponderada dos valores possíveis de Y_1 MAS USANDO a distribuição condicional de Y_1 como peso, ao invés de usar a distribuição marginal de Y_1.

Variância condicional

- Relembre: Se $\mu = \mathbb{E}(Y)$ então

$$\mathbb{V}(Y) = \mathbb{E}(Y-\mu)^2 = \sum_y (y-\mu)^2 \mathbb{P}(Y=y)$$

- Podemos calcular a variabilidade de Y_1 em torno de sua esperança CONDICIONADA nos valores das outras v.a.s $(Y_2, \ldots, Y_p)$:

$$\mathbb{V}\left(Y_1 | Y_2 = a_2, \ldots, Y_p = a_p\right) = \sum_y (y-m)^2 \mathbb{P}(Y_1 = y | Y_2 = a_2, \ldots, Y_p = a_p)$$

  onde $m = \mathbb{E}(Y_1 | Y_2 = a_2, \ldots, Y_p = a_p)$ é a esperança condicional.

- Pode-se mostrar que

$$\mathbb{V}(Y_1 | Y_2 = a_2, \ldots, Y_p = a_p) = \mathbb{E}(Y_1^2 | Y_2 = a_2, \ldots, Y_p = a_p) - m^2$$

Distribuição conjunta de vetores CONTÍNUOS

Conjunta contínua: caso bivariado

- (Y_1, Y_2) vetor bivariado de v.a.'s contínuas.
- Assim, Y_1 é uma v.a. continua e Y_2 também é uma v.a. continua: ambas possuem densidades $f_1(y)$ e $f_2(y)$
- Mas ao invés de analisarmos as v.a.'s isoladamente, queremos estudar o modo como elas interagem.
- Existe uma versão BIVARIADA da densidade.
- Vamos ver o seu significado empírico

Histograma $\to$ densidade

- Relembre a relação entre o histograma feito com uma amostra de uma v.a. Y e a densidade subjacente.
- Histograma "imita" a densidade f(x)
- Probabilidade é IGUAL a área debaixo da curva densidade.

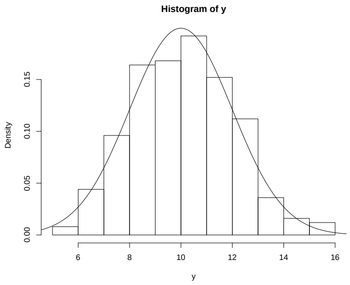


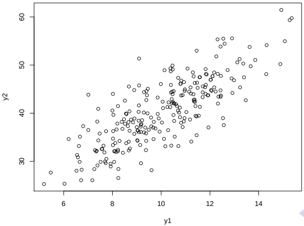
Figura: Histograma e densidade

Caso bivariado: amostra

- Uma amostra de tamanho n do VETOR aleatório bivariado (Y_1, Y_2):

$$(y_{11}, y_{12}), (y_{21}, y_{22}), (x_{31}, y_{32}), \ldots, (y_{n1}, y_{n2})$$

- A amostra é composta por n vetores (y_1, y_2) selecionados no plano de acordo com a densidade $f(y_1, y_2)$
- Histograma tri-dimensional tem aproximadamente a mesma forma que a superfície contínua $f(y_1, y_2)$.
- Ver o histograma 3-dim é praticamente ver a densidade $f(y_1, y_2)$.



Caso bivariado: amostra $\rightarrow$ histograma

- Crie uma grade regular sobre o plano e conte número de vetores (Y_1, Y_2) que caem em cada célula.
- A seguir, levante uma pilastra de altura proporcional a esta contagem.
- Regiões com mais pontos terão pilastras mais altas.

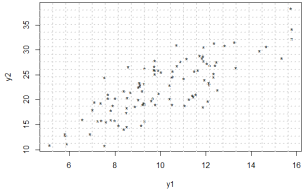


Figura: Amostra de 100 instâncias do vetor aleatório (Y_1, Y_2) e grade regular sobreposta.

Histograma 3-dim

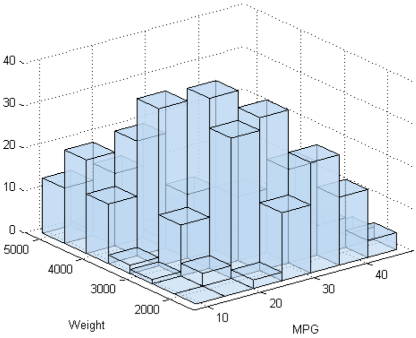
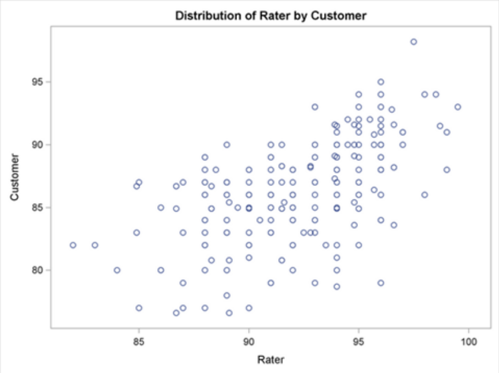


Figura: Histograma tri-dimensional baseado em amostra de vetor (X, Y). Histograma tri-dimensional tem aproximadamente a mesma forma que a superfície contínua $f(x, y)$ da densidade do vetor (X, Y).

Outro exemplo



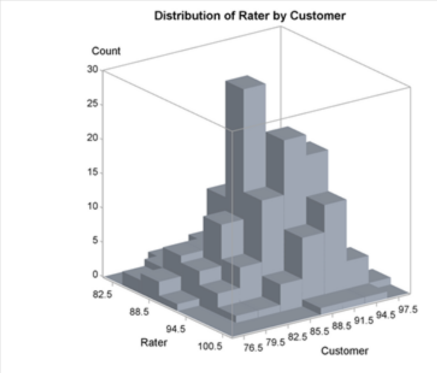


Figura: Amostra de n pontos do vetor aleatório (X, Y) e histograma tri-dimensional baseado nesta amostra.

Histograma 3-dim

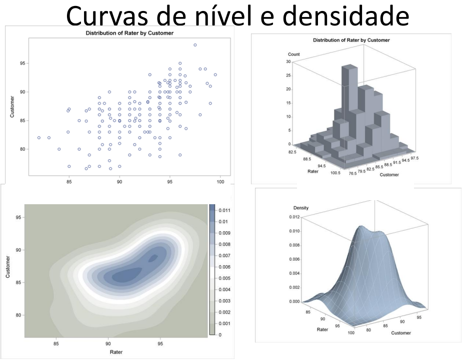


Figura: Amostra de 250 dados de (Y_1, Y_2) com histograma 3d, densidade $f(y_1, y_2)$ e suas curvas de nível

Um distribuição mais complexa

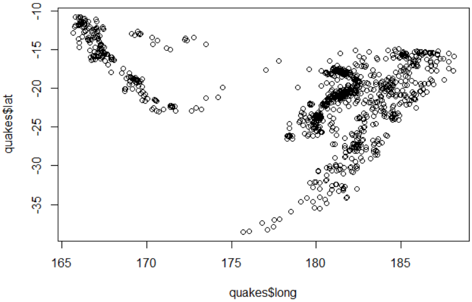


Figura: Dataset quakes fornece informações sobre 1000 terremotos com magnitude maior que 4.0 na escala Richter em torno da ilha Fiji na Oceania a partir de 1964. Acima, a longitude e latitude do epicentro desses 1000 eventos. Podemos ver a posição do epicentro como um vetor aleatório (X, Y) com certa densidade de probabilidade $f(x, y)$.

O histograma 3D

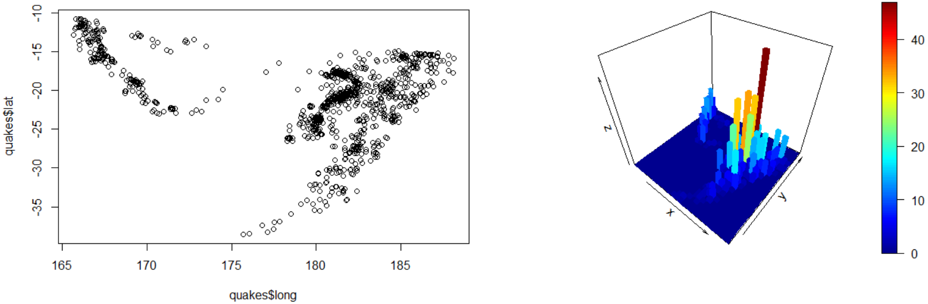


Figura: Histograma construído com texttthist3D(x = xmid, y = ymid, z = xy). Ver lista de exercícios.

Um histograma 3D mais interessante

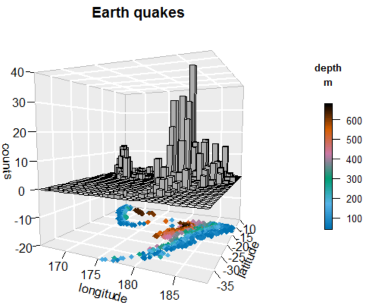
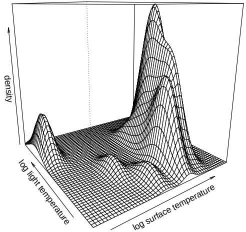


Figura: Histograma 3D e os dados de terremotos. Ver lista de exercícios.

Densidade bivariada $f(y_1, y_2)$

- Densidade do vetor $\boldsymbol{Y} = (Y_1, Y_2)$
- Em cada estrela, medem-se duas v.a.'s continuas: $Y_1 =$ log(intensidade da luz) e $Y_2 =$ log(temperatura à superfície)
- Plot da superfície $f(y_1, y_2)$
- Quais as combinações de Y_1 e Y_2 que são mais prováveis?
- Quais as regiões do espaço das medições em $\boldsymbol{Y} = (Y_1, Y_2)$ onde existe chance razoável de se observar uma estrela?



Old Faithful geyser

Figura: Old Faithful geyser. From the GeoCities geyser web page: Old Faithful erupts every 35-120 minutes for 1.5-5 minutes to a height of 90-184 feet. The rangers say that 90% of their predictions are within +/- 10 minutes. The time to the next eruption is predicted using the duration of the current eruption. The longer the eruption lasts, the longer the interval until the next eruption. For instance, a 2 minute eruption results in an interval of about 50 minutes whereas a 4.5 minute eruption results in an interval of about 85 minutes. It is not possible to predict more than one eruption in advance.

Old Faithful

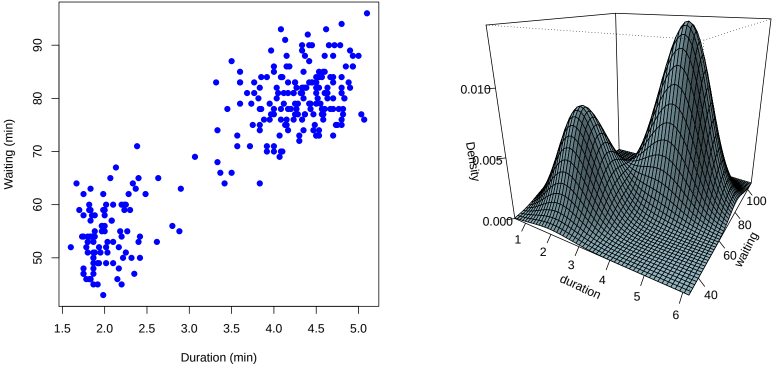


Figura: Old Faithful geyser: waiting time and eruption duration

Amostra e densidade

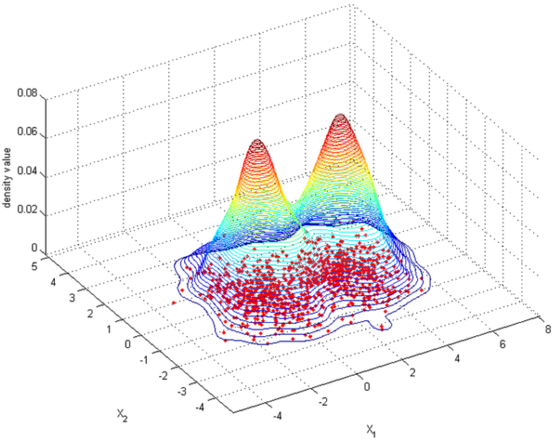


Figura: Old Faithful geyser: tempo de espera e duração de erupção.

É mais fácil visualizar em 2-dim

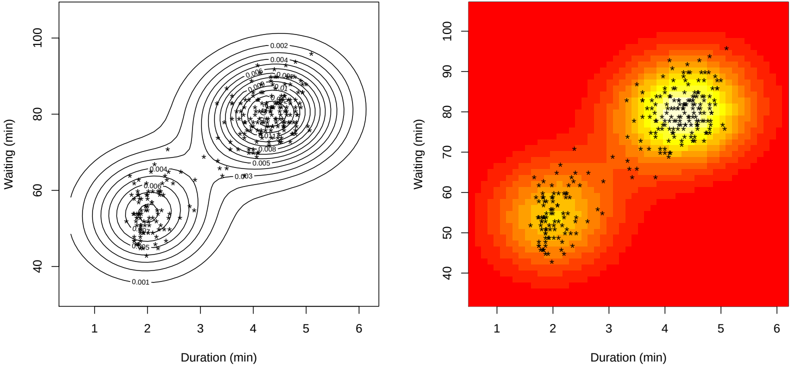


Figura: Old Faithful geyser: waiting time and eruption duration

Caso contínuo bivariado

- $\boldsymbol{Y} = (Y_1, Y_2)$ vetor bivariado de v.a.'s contínuas.
- Uma função densidade de probabilidade é QUALQUER função tal que:
  - $f(y_1, y_2) \geq 0$
  - $$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(y_1, y_2) dy_1 dy_2 = 1$$
- No caso uni-dimensional, probabilidades s ao áreas debaixo da curva-densidade $f(x)$.
- No caso bi-dimensional, probabilidades são volumes debaixo da superfície-densidade $f(x, y)$.

Probabilidades s ao volumes sob a densidade

- Probabilidades são volumes sob a superfície $f(x, y)$ de densidade de probabilidade.
- Probab do vetor (X, Y) cair numa região D do plano é

$$\mathbb{P}((X, Y) \in D) = \int \int_D f(x, y) dxdy$$

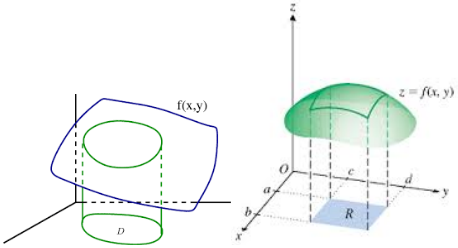


Figura: Probabilidade de (X, Y) cair em D é igual ao volume sob a superfície.

Caso geral k-dim

- No caso de um vetor aleatório $\mathbf{Y} = (Y_1, \ldots, Y_k)$ com k v.a.'s contínuas, a densidade de probabilidade é QUALQUER função tal que:
  - $f(bsy) = f(y_1, \ldots, y_k) \geq 0$ para todo ponto $bsy \in \mathbb{R}^k$
  - $$1 = \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} f(y_1, \ldots, y_k) dy_1 \ldots dy_k$$
- A probabilidade do vetor bsY cair numa região D de $\mathbb{R}^k$ é dada por

$$\mathbb{P}((Y_1, \ldots, Y_k) \in D) = \int \ldots \int_D f(y_1, \ldots, y_k) dy_1 \ldots dy_k$$

Distribuição marginal

- No caso discreto, a distribuição marginal de uma v.a. é obtida somando-se sobre todos os valores das demais variáveis. No caso conínuo, substituímos a soma por uma integral.
- No caso bi-dimensional (X, Y), a densidade de probabilidade da v.a. contínua X é obtida integrando sobre os valores de Y.
- Para diferenciar as densidades, vamos escrever $f_X(x)$ para a densidade marginal de X no ponto x e $f_{XY}(x, y)$ para o valor da densidade conjunta de (X, Y) no ponto (x, y).
- Por exemplo, $f_X(0)$ e $f_X(1.2)$ são os valores da densidade marginal de X nos pontos $x = 0$ e $x = 1.2$.
- $f_{XY}(0.2, 1.5)$ é o valor da densidade conjunta no ponto $(x, y) = (0.2, 1.5)$.
- Para um ponto genérico x

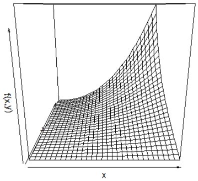
$$f_X(x) = \int_{-\infty}^{\infty} f(x, y) dy$$

Exercício básico

- Vetor contínuo (X, Y) com suporte em $[0, 1] \times [0, 1]$ (isto é, densidade é zero fora desta região).
- Densidade: $f(x, y) = k(x^2y + x^3y^4)$ para $(x, y) \in [0, 1]^2$.
- Encontrar a constante de normalização k:

$$\begin{aligned} 1 &= \int\int_{[0,1]^2} k(x^2y + x^3y^4)\ dxdy = k\int_{[0,1]} \left(\frac{x^2}{2} + \frac{x^3}{5}\right) dx \\ &= k\left(\frac{1}{6} + \frac{1}{20}\right) = \frac{13k}{60} \end{aligned}$$

e portanto $k = 60/13$

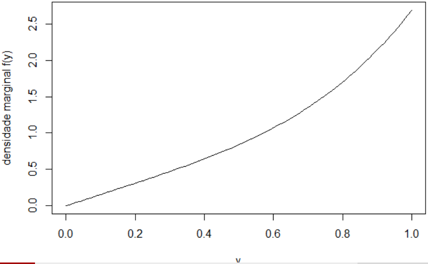


Exercício básico

- Encontrar a marginal $f_Y(y)$ para $y \in [0,1]$:

$$f_Y(y) = \int_{[0,1]} \frac{60}{13}(x^2 y + x^3 y^4)\ dx = \frac{5}{13}(4y + 3y^4)$$

- Veja que, avaliada no ponto $y = 0.1$, temos $f_Y(0.1) = 5/13\ (4(0.1) + 30.1^2) = 0.165$ enquanto que, no ponto $y = 0.9$, temos $f_Y(0.9) = 5/13\ (4(0.9) + 30.9^2) = 2.319$.



Distribuição Condicional

- No caso bi-dimensional (X, Y), a densidade de probabilidade de X CONDICIONADA ao evento $Y = y$ é dada por

$$f_{X|Y}(x|y) = \frac{f_{XY}(x, y)}{f_Y(y)}$$

- Por exemplo, $f_{X|Y}(x|y = 0.2)$ é a densidade da v.a. X condicionada ao evento $Y = 0.2$ e avaliada num ponto x genérico:

$$f_{X|Y}(x|y = 0.2) = \frac{f_{XY}(x, 0.2)}{f_Y(0.2)}$$

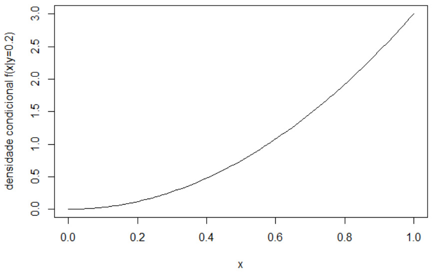
- Observe que esta é uma densidade da v.a. X (variando em x) e que o denominador não depende de x.
- O valor $f_Y(0.2)$ é o mesmo para qualquer valor x.
- $f_{X|Y}(x = 0.3|y = 0.2)$ é esta densidade condicional de X avaliada no ponto $x = 0.3$:

$$f_{X|Y}(x = 0.3|y = 0.2) = \frac{f_{XY}(0.3, 0.2)}{f_Y(0.2)}$$

Exercício básico

- Densidade: $f(x, y) = 60/13\ (x^2 y + x^3 y^4)$ para $(x, y) \in [0, 1]^2$.
- Marginal $f_Y(y) = 5/13\ (4y + 3y^4)$ para $y \in [0, 1]$:
- Densidade de X condicionada ao evento $Y = 0.2$:

$$f_{X|Y}(x|y = 0.2) = \frac{f_{XY}(x, 0.2)}{f_Y(0.2)} = \frac{60/13\ (0.2\ x^2 + 0.2^4\ x^3)}{5/13\ (4\ 0.2 + 3\ 0.2^4)} = \frac{12}{0.8048}(0.2x^2 + 0.0016x^3)$$

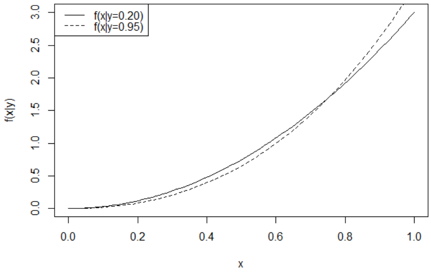


Exercício básico

- Comparando duas densidades condicionais de X: condicionada ao evento $Y = 0.20$ e ao evento $Y = 0.95$.

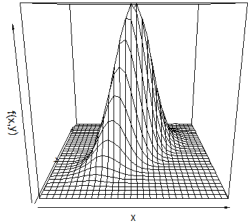
$$\begin{aligned} f_{X|Y}(x|y=0.95) &= \frac{f_{XY}(x, 0.95)}{f_Y(0.95)} \\ &= \frac{60/13\ (0.95\ x^2 + 0.95^4\ x^3)}{5/13\ (4\ 0.95 + 3\ 0.95^4)} = \frac{60}{31.217}(0.95\ x^2 + 0.95^4\ x^3) \end{aligned}$$

- Não são muito diferentes neste exemplo particular.



Mais um exemplo - gaussiana

- Densidade para (X, Y) é
$$f_{XY}(x,y) = \frac{1}{2\pi\sqrt{0.51}} \exp\left(-\frac{x^2+y^2-1.4xy}{1.02}\right)$$
com suporte em $\mathbb{R}^2$.
- Esta é a densidade de uma gaussiana bivariada onde a correlação é igual a $\rho = 0.7$ e as marginais são $X \sim N(0,1)$ e $Y \sim N(0,1)$.
- Marginal $f_Y(y) = 1/\sqrt{2\pi} \exp\left(-y^2/2\right)$ para $y \in \mathbb{R}$.

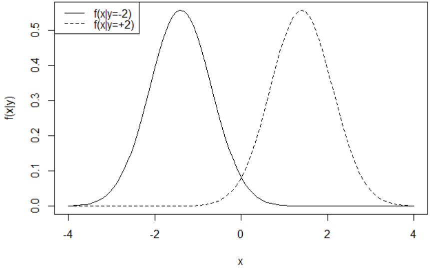


Mais um exemplo - gaussiana

- Densidade $(X|Y=-2)$:

$$
\begin{aligned}
f_{X|Y}(x|y=-2) &= \frac{f_{XY}(x,-2)}{f_Y(-2)} = \frac{\frac{1}{2\pi\sqrt{0.51}} \exp\left(-\frac{x^2+y^2-1.4xy}{1.02}\right)}{1/\sqrt{2\pi}\exp\left(-(-2)^2/2\right)} \\
&= 1/\sqrt{1.02\pi}\exp\left(-\frac{(x+1.4)^2}{1.02}\right)
\end{aligned}
$$

- De forma similar, obtemos $f_{X|Y}(x|y=+2)$. Gráficos abaixo.



Vendo a condicional na conjunta

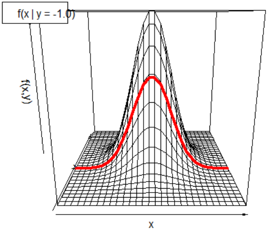
- Olhar a superfície da densidade $f(x, y)$ mostra imediatamente a forma (shape) da densidade condicional.
- Por exemplo,

$$f_{X|Y}(x|y = 0.2) = \frac{f_{XY}(x, 0.2)}{f_Y(0.2)} \propto f_{XY}(x, 0.2)$$

  pois o denominador é uma constante *COM RESPEITO A* x.
- Assim, se quisemos saber como $f_{X|Y}(x|y = 0.2)$ varia como função de x, basta olharmos na superfície $f(x, y)$ a curva obtida se fixarmos $y = 2$.

Vendo a condicional na conjunta

- $f(x|y = 1.0)$ tem a mesma forma(shape) que a curva em vermelho, que é $f_{XY}(x, -1.0)$, os valores da densidade conjunta com $y = -1.0$ fixo.
- A densidade condicional é esta curva multiplicada por uma constante positiva.



Esperança condicional

- Considere o vetor $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_p)$
- Calculamos a distribuição condicional de Y_1 dados os valores de $Y_2, \ldots, Y_p$
- Podemos calcular o valor esperado de Y_1 dados (ou fixados) os valores de $Y_2, \ldots, Y_p$
- É simplesmente como na definição usual de esperança de v.a.'s discretas:

$$\mathbb{E}(Y_1|Y_2 = a_2, \ldots, Y_p = a_p) = \int y \; f_{Y_1|Y_2\ldots Y_p}(y|y_2 = a_2, \ldots, y_p = a_p) \; dy$$

- Média ponderada dos valores possíveis de Y_1 MAS USANDO a densidade condicional de Y_1 como peso, ao invés de usar a distribuição marginal de Y_1.

Variância condicional

- Relembre: Se $\mu = \mathbb{E}(Y)$ então

$$\mathbb{V}(Y) = \mathbb{E}(Y-\mu)^2 = \int (y-\mu)^2 \; f_Y(y) \; dy$$

- Podemos calcular a variabilidade de Y_1 em torno de sua esperança CONDICIONADA nos valores das outras v.a.s $(Y_2, \ldots, Y_p)$:

$$\mathbb{V}\left(Y_1 | Y_2 = a_2, \ldots, Y_p = a_p\right) = \int (y-m)^2 f_{Y_1|Y_2 \ldots Y_p}(y|y_2 = a_2, \ldots, y_p = a_p) \; dy$$

onde $m = \mathbb{E}(Y_1 | Y_2 = a_2, \ldots, Y_p = a_p)$ é a esperança condicional.

- Pode-se mostrar que

$$\mathbb{V}(Y_1 | Y_2 = a_2, \ldots, Y_p = a_p) = \mathbb{E}(Y_1^2 | Y_2 = a_2, \ldots, Y_p = a_p) - m^2$$

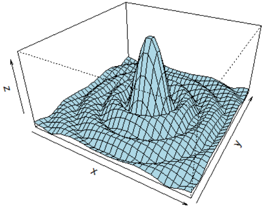
Simulando um vetor contínuo

- Queremos simular uma amostra do vetor aleatório bivariado (X, Y) com densidade $f(x, y)$.
- Existem vários métodos (ver disciplina PGM - Probabilistic Graphical Models)
- Um método simples é o de aceitação-rejeição.
- Obtenha uma densidade $g(x, y)$ de onde você saiba simular.
- Encontre M tal que $f(x, y) \leq Mg(x, y)$ para todo ponto (x, y).

```
while(contador < nsim){
  gere (x,y) de g(x,y)
  jogue moeda com P(cara) = f(x,y)/(M*g(x,y))
  se cara:
     aceite (x,y)
     contador = contador + 1
}
```

Exemplo: Simulando um vetor contínuo

- Queremos simular 100 pontos aleatórios (x, y) seguindo a densidade $f_{XY}(x, y)$ com suporte em $[-10, 10]^2$ e dada por
$$f_{XY}(x, y) = \frac{|\sin(r(x, y))|}{44 \; r(x, y)}$$
onde $r(x, y) = \sqrt{x^2 + y^2}$ é a distância de (x, y) à origem.
- O máximo de $f_{XY}(x, y)$ ocorre em $(x, y) = (0, 0)$ e é igual a $1/44 \approx 0.0228$.



Exemplo: Simulando um vetor contínuo

- Vamos simular (X, Y) em $[-10, 10]^2$ com uma distribuição uniforme.
- Isto é, a densidade é igual a $g(x, y) = 1/20^2$ em $[-10, 10]^2$ e igual a zero fora dessa região.
- Gerar desta $g(x, y)$ é muito fácil pois X e Y são independentes e cada uma delas segue uma uniforme em $[-10, 10]$.
- Assim, gere a coordenada $X \sim U(-10, 10)$ e independentemente a coordenada $Y \sim U(-10, 10)$.

```
x = runif(1000, -10, 10)
y = runif(1000, -10, 10)
```

- A seguir, retenha ou descarte estes valores com probabilidade $f(x, y)/(Mg(x, y))$. Quem é M?

Exemplo: Simulando um vetor contínuo

- Temos $g(x, y) = 1/400$ para todo (x, y) na região.
- Queremos $1 > f(x, y)/(Mg(x, y)) = 400f(x, y)/M$.
- Como o máximo de $f(x, y)$ ocorre na origem e é igual a $1/44$, podemos ter certeza que

$$\frac{f(x, y)}{Mg(x, y)} = \frac{400f(x, y)}{M} \leq \frac{400f(0, 0)}{M} = \frac{400}{44M} < 1$$

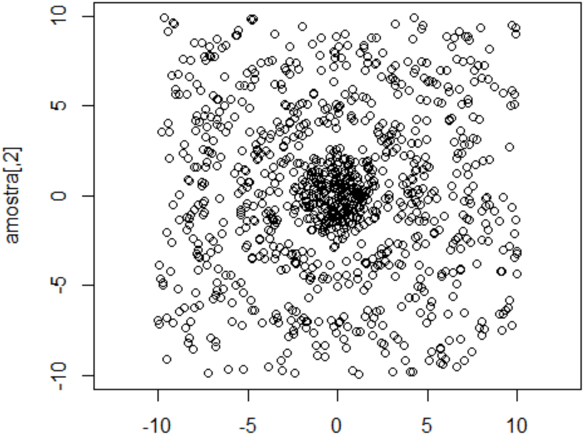
  se tomarmos $M > 400/44 = 9.090909$. Vamos tomar $M = 10$.
- Assim, basta reter os pontos (x, y) tais que a sua "moeda" resulte em cara onde

$$\mathbb{P}(\text{cara}) = \frac{400f_{XY}(x, y)}{10} = 40f_{XY}(x, y)$$

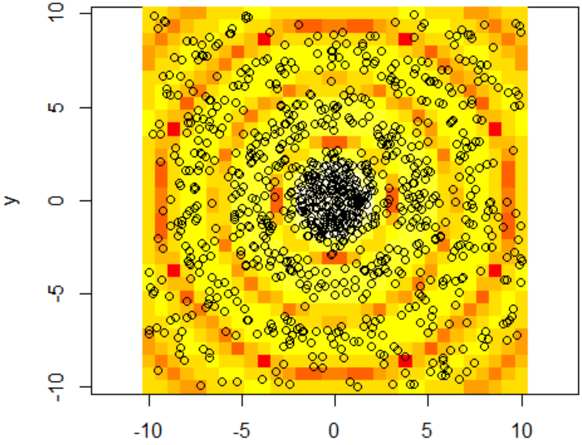
Exemplo: Simulando um vetor contínuo

```
n = 1000; contador = 0
amostra = matrix(0, ncol=2, nrow=n)
while(contador < n)
{
  x = runif(1, -10, 10)
  y = runif(1, -10, 10)
  r = sqrt(x^2+y^2)
  fxy = abs(sin(r))/(44*r)
  prob = 40 * fxy
  if(runif(1) < prob){
     contador = contador + 1
     amostra[contador, ] = c(x,y)
  }
}
plot(amostra, asp=1)
```

Amostra gerada de $f(x, y)$



Amostra gerada de $f(x, y)$



Amostra gerada de $f(x, y)$

```
x <- seq(-10, 10, length= 30)
y <- x
f <- function(x, y) {
  r <- sqrt(x^2+y^2);
  abs(sin(r))/(44*r)
}
z <- outer(x, y, f)
image(x,y,log(z), asp=1)
points(amostra)
```