Distribuições Multivariadas

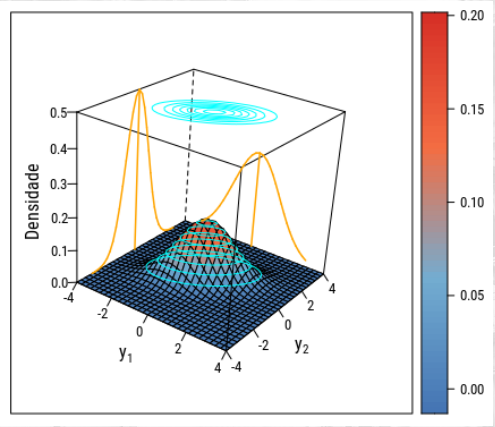
Prof. Walmes M. Zeviani

Departamento de Estatística
Universidade Federal do Paraná

Conteúdo

Neste vídeo

- Importância das distribuições multivariadas.
- Distribuição multinomial.
- Distribuição Normal.
- Distribuição de Dirichlet.



Importância das distribuições multivariadas

Importância das distribuições multivariadas

- Descrever o comportamento conjunto de várias variáveis aleatórias.
- Permitir a analisar de forma simultânea várias variáveis aleatórias.
- Investigar a relação entre elas.
- Usar a informação de uma para inferir sobre a outra.
- Descrever estruturas de dependência temporal, espacial, genética, etc.
- Condensar a informação de várias variáveis em um número reduzido de fatores latentes.

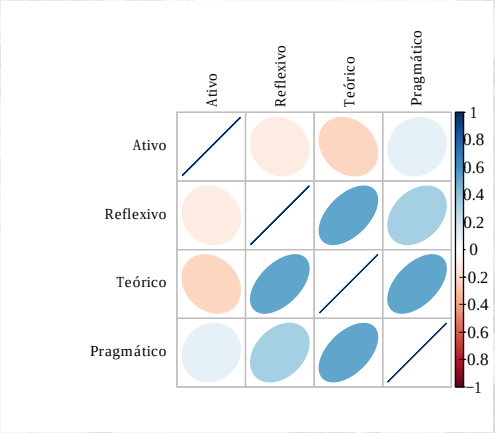


Figura 1. Correlação entre os escores para os estilos de aprendizado determinados para os alunos de Estatística Básica.

A distribuição Multinomial

Características de uma v.a. Multinomial

- **Generalização** da distribuição Binomial.
  - Binomial: 2 resultados mutuamente exclusivos $\{0, 1\}$.
  - Multinomial: k resultados mutuamente exclusivos.
- Condições:
  - Suponha n ensaios **independentes** que podem apresentar, em cada, apenas um de k possíveis resultados $\{O_1, O_2, \ldots, O_k\}$.
  - Sejam $p_1, p_2, \ldots, p_k$, com $\sum p_i = 1$, as probabilidades de observar $O_1, O_2, \ldots, O_k$, respectivamente, que se mantêm **constantes**.
  - Seja Y_i o **número de vezes** que o resultado O_i ocorre nos n ensaios.
  - Dessa forma, $\{Y_1, \ldots, Y_k\} \sim \text{Multinomial}(p_1, \ldots, p_k, n)$.

j	O_1	O_2	$\cdots$	O_k
1		✓		
2	✓			
3		✓	⋮	
⋮				✓
n		✓		
Soma	Y_1	Y_2	$\cdots$	Y_k

Exemplos de uma v.a. Multinomial

1. Resultado do arremesso da bola à cesta:

   $\{\text{errou}, \text{fez 1}, \text{fez 2}, \text{fez 3 pontos}\}$.

2. Ação resultado de uma propaganda direta por email

   $\{\text{não viu}, \text{acessou}, \text{wishlist}, \text{compra}\}$.

3. Categoria de veículo alugado por cliente em um site

   $\{\text{SUV}, \text{Sedan}, \text{Hatch}, \text{Utilitário}, \ldots\}$.

4. Classificação de um estudante em relação à área do Curso

   $\{\text{Exatas}, \text{Humanas}, \text{Biológicas}\}$.

5. Signo de uma pessoa.
6. Dia da semana de acidente de trabalho.
7. Grau de uma infração de trânsito.
8. Estado civil de uma pessoa.

A distribuição Multinomial

Satisfeitas as condições apresentadas anteriormente, tem-se que a distribuição do vetor aleatório $\{Y_1, \ldots, Y_k\}$ é dada por

$$p(y_1, \ldots, y_k) = \frac{n!}{y_1! \cdot \ldots \cdot y_k!} \cdot p_1^{y_1} \cdot \ldots \cdot p_k^{y_k}$$

$$= \frac{n!}{\prod_{i=1}^{k} y_i!} \cdot \prod_{i=1}^{k} p_i^{y_i}.$$

j	O_1	O_2	$\cdots$	O_k
1		✓		
2	✓			
3		✓	$\vdots$	
$\vdots$				✓
n		✓		
Soma	Y_1	Y_2	$\cdots$	Y_k

Esperança e variância para cada componente i são dados por

$$\mathrm{E}(O_i) = p_i \quad \text{logo} \quad \mathrm{E}(Y_i) = n \cdot p_i.$$
$$\mathrm{V}(O_i) = p_i(1 - p_i) \quad \text{logo} \quad \mathrm{V}(Y_i) = n \cdot p_i(1 - p_i).$$

Exemplo: bolas de gude

Em uma urna há 2 bolas vermelhas, 3 verdes e 5 azuis. São selecionadas 4 bolas ao acaso da urna com reposição. Qual a probabilidade de retirar 2 verdes e 2 azuis.

Figura 2. Bolas de gude. Fonte: https://cutt.ly/QhzaxK8.

Solução

- São $n = 4$ ensaios independentes.
- As probabilidades são:

$$p_1 = \frac{2}{10}, \quad p_2 = \frac{3}{10} \quad \text{e} \quad p_3 = \frac{5}{10}.$$

- Deseja-se a probabilidade do resultado

$$y_1 = 0, \quad y_2 = 2 \quad \text{e} \quad y_3 = 2.$$

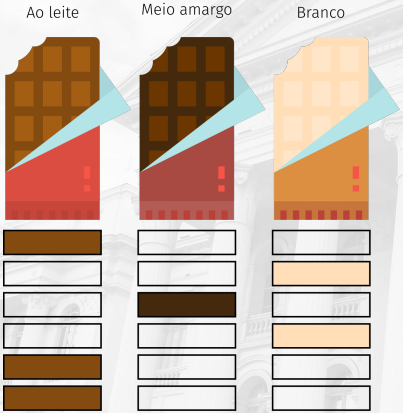
Usando a função de probabilidade, tem-se

$$\begin{aligned} p(0,2,2) &= \frac{4!}{0! \cdot 2! \cdot 2!} \cdot 0.2^0 \cdot 0.3^2 \cdot 0.5^2 \\ &= 0.135 \end{aligned}$$

Diferença entre Multinomial e Binomiais

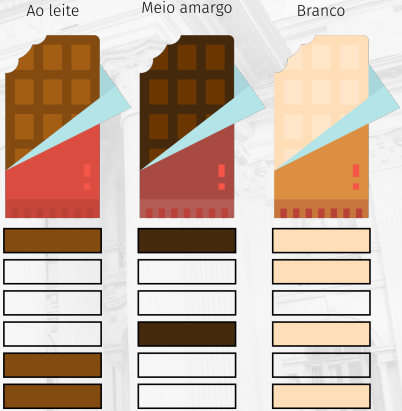
Multinomial

Qual seu chocolate favorito?



Binomiais

Qual chocolate o cliente comprou?



Ícone retirado do https://www.flaticon.com/authors/photo3idea-studio

A distribuição Normal Multivariada

Características de um v.a. Normal Multivariada

- Um conjunto de variáveis contínuas não limitadas.
- Individualmente as variáveis são Normais, portanto simétricas.
- Variáveis apresentam relação linear entre si.
- Resultam da combinação de muitos fatores de pequena contribuição.
  - Características genéticas e morfológicas.
  - Traços latentes.
  - Índices econômicos.

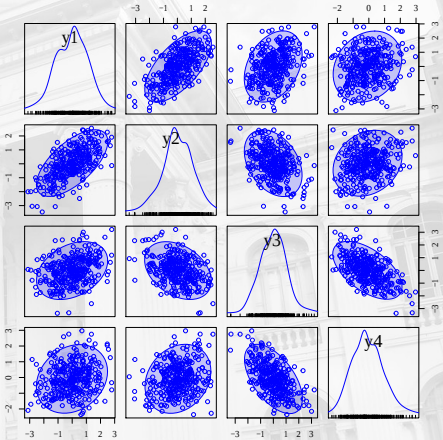


Figura 3. Diagramas de dispersão para 4 variáveis contínuas.

Exemplos de v.a. com distribuição Normal Multivariada

- Peso de massa seca, altura de planta e altura da primeira espiga em uma planta de milho.
- Peso, comprimento e circunferência de uma banana.
- Comprimentos no crânio de um fóssil ou animal.
- Variação de um conjunto de índices econômicos ou mercadorias: e.g. combustível.
- Traço latente para resolução de problemas de física, matemática, química, etc.

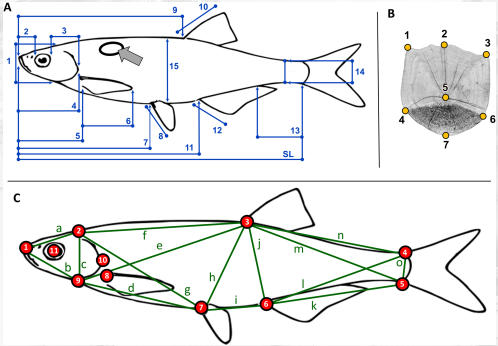


Figura 4. Distâncias medidas em um peixe para estudo morfométrico. Fonte: https://cutt.ly/AhlsOWo.

Importância da Normal Multivariada

- Modelar a estrutura de correlação ou dependência.
  - Modelos de séries temporais: correlação entre datas.
  - Modelos geoestatísticos: correlação entre pontos no espaço.
  - Modelos genéticos: correlação entre características.
- Técnicas de análise multivariada.
  - Análise de discriminante linear.
  - Análise fatorial exploratória e confirmatória.
  - Análise de correlação canônica.
  - Análise de variância multivariada.

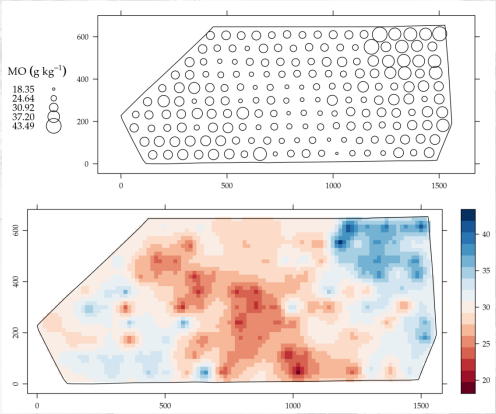


Figura 5. Gráfico de valores observados de matéria orgânica do solo e o mapa com a predição para todo o terreno.

A distribuição Normal Multivariada

Um vetor aleatório contínuo $\mathbf{Y}$ $(p \times 1)$ tem distribuição Normal multivariada se sua densidade conjunta é dada por

$$f(\mathbf{y}) = \left(\frac{1}{2\pi}\right)^{p/2} \det(\Sigma)^{-1/2} \exp\left\{-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu})^\top \Sigma^{-1}(\mathbf{y}-\boldsymbol{\mu})\right\},$$

em que

$$\mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_p \end{bmatrix}, \quad \boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_p \end{bmatrix} \quad \text{e} \quad \Sigma = \begin{bmatrix} \sigma_1^2 & \dots & \sigma_{1p} \\ \vdots & \ddots & \vdots \\ \sigma_{p1} & \dots & \sigma_p^2 \end{bmatrix}.$$

Denotamos por $\mathbf{Y} \sim \mathrm{N}_p(\boldsymbol{\mu}, \Sigma)$.

A média e variância do vetor aleatório são dados por

$$\mathrm{E}(\mathbf{Y}) = \boldsymbol{\mu} \quad \text{e} \quad \mathrm{V}(\mathbf{Y}) = \Sigma.$$

Distribuições marginais e condicionais (caso bivariado)

Distribuição marginal: A distribuição marginal de cada elemento de um vetor aleatório Normal é também Normal. Isto é,

$$Y_i \sim \mathrm{N}(\mu_i, \sigma_i^2) \quad i \in \{1, 2\}.$$

Distribuição condicional: A distribuição condicional de cada elemento de um vetor aleatório em relação ao outro tambem é Normal. Isto é,

$$Y_i | Y_j = y_j \sim \mathrm{N}(\mu_{i|j}, \sigma_{i|j}^2), \qquad i \neq j \in \{1, 2\},$$

em que

$$\mu_{i|j} = \mu_i + \frac{\sigma_{ij}}{\sigma_j^2}(y_j - \mu_j) \quad \text{e} \quad \sigma_{i|j}^2 = \sigma_i^2 - \frac{\sigma_{ij} \cdot \sigma_{ji}}{\sigma_j^2}.$$

Exemplo: altura de mães e filhas

A altura de mães e filhas adultas segue distribuição Normal bivariada com parâmetros

$$\mu = \begin{bmatrix} 165 \\ 165 \end{bmatrix}, \quad \Sigma = \begin{bmatrix} 49 & 35 \\ 35 & 49 \end{bmatrix}.$$

- Dado que uma mãe tem 172 cm de altura, qual a estatura esperada para a filha quando adulta?
- Qual a probabilidade da filha ter mais de 180 cm?

Figura 6. Mãe com sua filha. Foto de Pixabay no Pexels.

Solução

- A altura esperada para a filha é

$$\mu_{f|m} = \mu_f + \frac{\sigma_{fm}}{\sigma_m^2}(y_m - \mu_m)$$
$$= 165 + \frac{35}{49}(172 - 165) = 170.$$

- A variância de $Y_f|Y_m = 172$ é

$$\sigma^2_{f|m} = \sigma^2_f - \frac{\sigma_{fm} \cdot \sigma_{mf}}{\sigma^2_m}$$
$$= 49 - \frac{35 \cdot 35}{49} = 47.57.$$

- E assim, $P(Y_f > 180|Y_m = 172) = 0.074$.

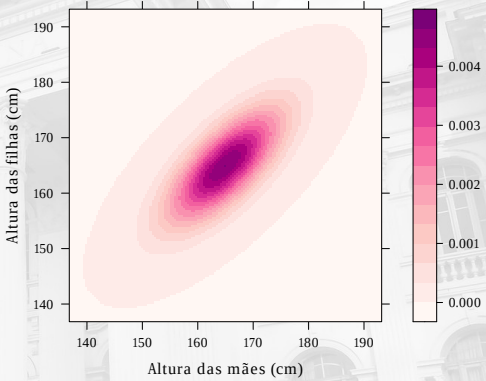


Figura 7. Distribuição Normal bivariada para o problema da altura de mães e filhas.

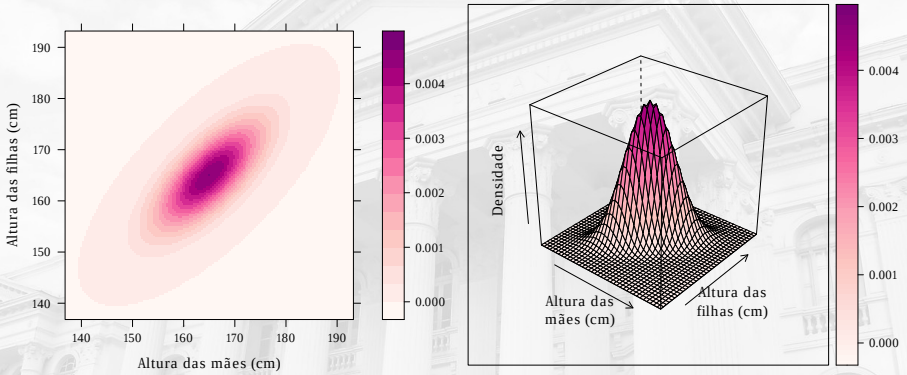


Figura 8. Distribuição Normal bivariada para o problema da altura de mães e filhas.

Distribuições marginais e condicionais (caso geral)

Considere a partição do vetor aleatório $\mathbf{Y}$ em dois subvetores complementares de tamanho p e q. Ou seja,

$$\mathbf{Y} = \begin{bmatrix} \mathbf{Y}_p \\ \mathbf{Y}_q \end{bmatrix}, \quad \boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_p \\ \boldsymbol{\mu}_q \end{bmatrix}, \quad \Sigma = \begin{bmatrix} \Sigma_{pp} & \Sigma_{pq} \\ \Sigma_{qp} & \Sigma_{qq} \end{bmatrix}.$$

Distribuição marginal: A distribuição marginal de $\mathbf{Y}_p$ é Normal. Isto é,

$$\mathbf{Y}_p \sim \mathrm{N}_p(\boldsymbol{\mu}_p, \Sigma_{pp}).$$

Distribuição condicional: A distribuição condicional de $\mathbf{Y}_p$ em relação à $\mathbf{Y}_q$ é Normal. Isto é,

$$\mathbf{Y}_p | \mathbf{Y}_q = \mathbf{y}_q \sim \mathrm{N}_p(\boldsymbol{\mu}_{p|q}, \Sigma_{p|q}), \quad \text{em que}$$
$$\boldsymbol{\mu}_{p|q} = \boldsymbol{\mu}_p + \Sigma_{pq}\Sigma_{qq}^{-1}(\mathbf{y}_q - \boldsymbol{\mu}_q) \quad \text{e} \quad \Sigma_{p|q} = \Sigma_{pp} - \Sigma_{pq}\Sigma_{qq}^{-1}\Sigma_{qp}.$$

Algumas propriedades

- Correlação 0 (zero) implica **independência**.
- Variáveis individualmente apresentarem distribuição Normal não implica em distribuição Normal conjunta.

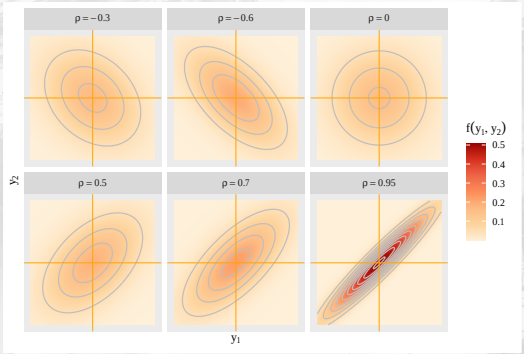


Figura 9. Gráficos da densidade da Normal bivariada com diferentes valores para a correlação.

A distribuição de Dirichlet

Características e exemplos da distribuição de Dirichlet

- Generalização da distribuição **Beta**.
- Usada para representar variáveis contínuas limitadas que representam composição: **variáveis composicionais**.
  - Teores de argila, silte e areia do solo.
  - Composição de um alimento: % de carboidrato, proteína, gordura e outros.
  - Distribuição de valores em uma carteira de investimento: % em renda fixa, fundos de investimento, COEs.
  - Distribuição do custo de vida: % alimentação, moradia, educação, saúde.

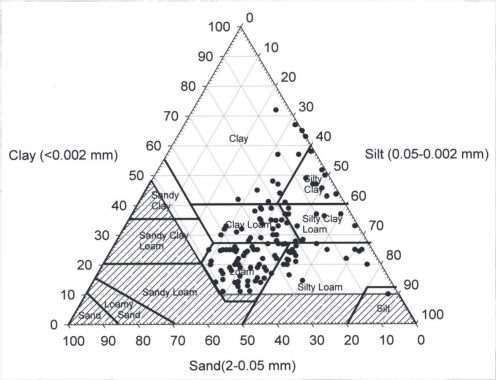


Figura 10. Diagrama ternário do solo. Fonte: https://cutt.ly/ohldjSJ.

A distribuição de Dirichlet

Um vetor aleatório $\mathbf{Y}$ tem distribuição de Dirichlet se a função de densidade é dada por

$$f(y_1, \ldots, y_k) = \frac{1}{\mathrm{B}(\boldsymbol{\alpha})} \cdot \prod_{i=1}^{k} y_i^{\alpha_i - 1}, \quad y_i > 0 \text{ e } \sum_{i=1}^{k} y_i = 1,$$

em que

$$\mathrm{B}(\boldsymbol{\alpha}) = \frac{\prod_{i=1}^{k} \Gamma(\alpha_i)}{\Gamma\left(\sum_{i=1}^{k} \alpha_i\right)}, \quad \boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_k), \alpha_i > 0.$$

A média e variância de cada componente do vetor são

$$\mathrm{E}(Y_i) = \frac{\alpha_i}{\sum_{j=1}^{k} \alpha_j} \quad \text{e} \quad \mathrm{V}(Y_i) = \frac{\tilde{\alpha}_i(1 - \tilde{\alpha}_i)}{\alpha_0 + 1}, \quad \tilde{\alpha}_i = \frac{\alpha_i}{\alpha_0}, \quad \alpha_0 = \sum_{i=1}^{k} \alpha_i.$$

Exemplo: análise de texto

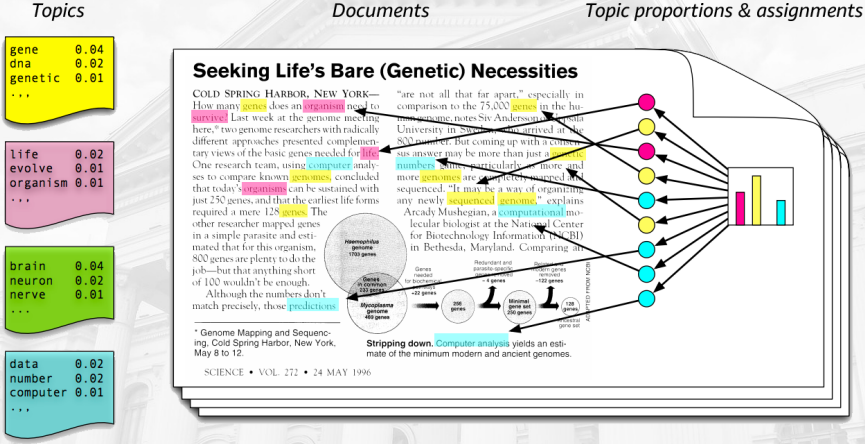


Figura 11. Ilustração do uso da distribuição de Dirichlet em modelagem de tópicos. Fonte: https://cutt.ly/uhzYoWF.

Alguns cenários

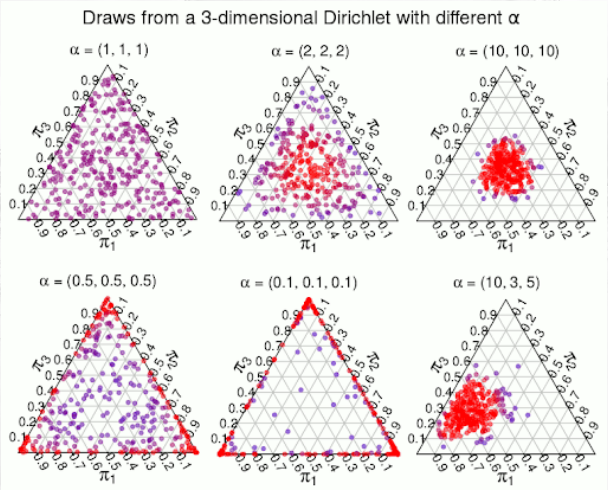


Figura 12. Alguns cenários para uma variável de 3 componentes. Fonte: https://www.r-bloggers.com/2015/04/the-non-parametric-bootstrap-as-a-bayesian-model/.

Considerações finais

Considerações finais

Revisão

- Importância das distribuições multivariadas.
- Distribuição multinomial.
- Distribuição Normal.
- Distribuição de Dirichlet.

Outras distribuições multivariadas

- Multinomial negativa.
- Wishart.
- Von Mises–Fisher.
- t multivariada.
- Laplace multivariada
- Além de outras.