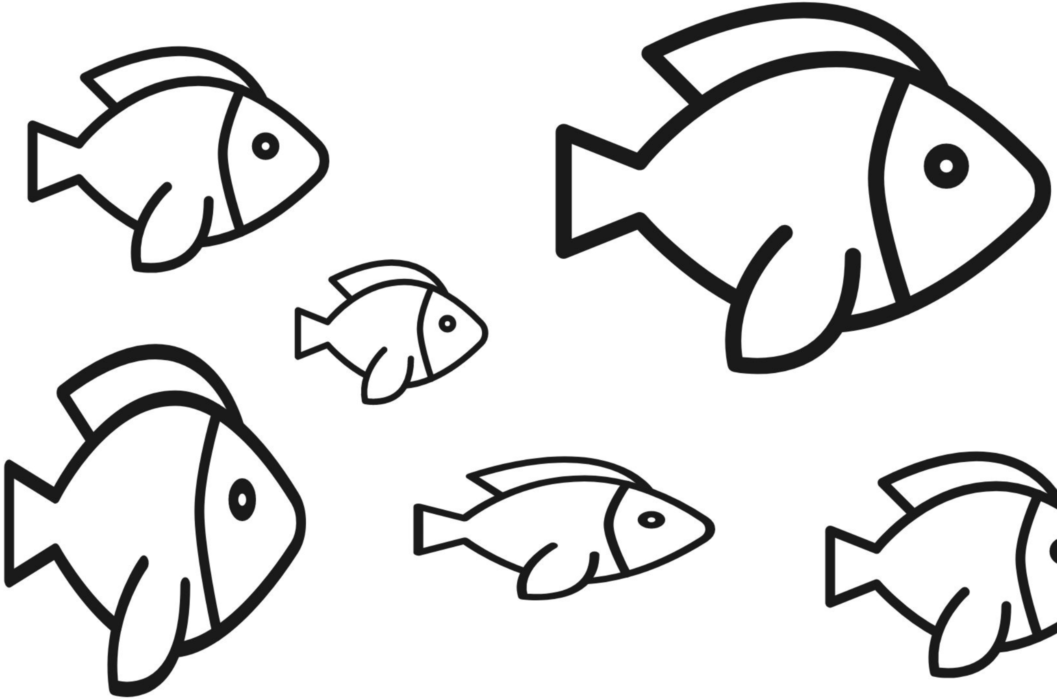
Quantificando variação multivariada

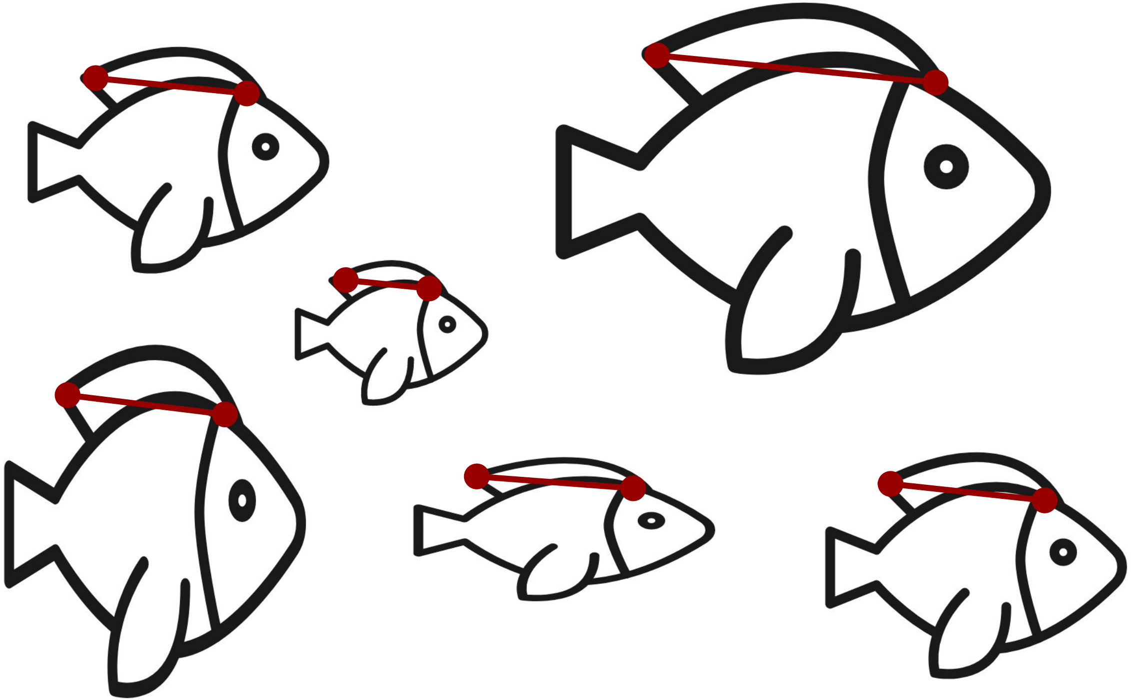
Variação e covariação

Diogo Melo
diogro@gmail.com
LEM-USP

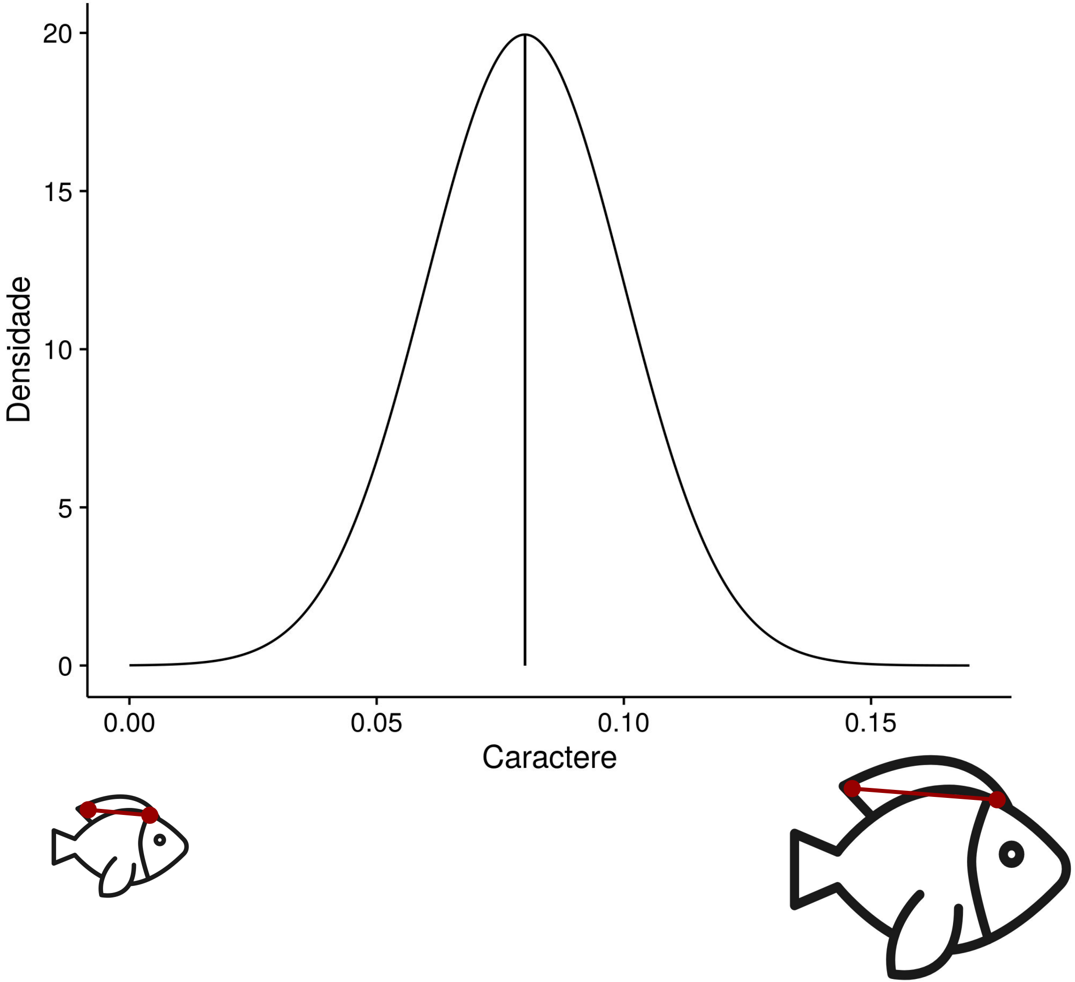
Medindo variação



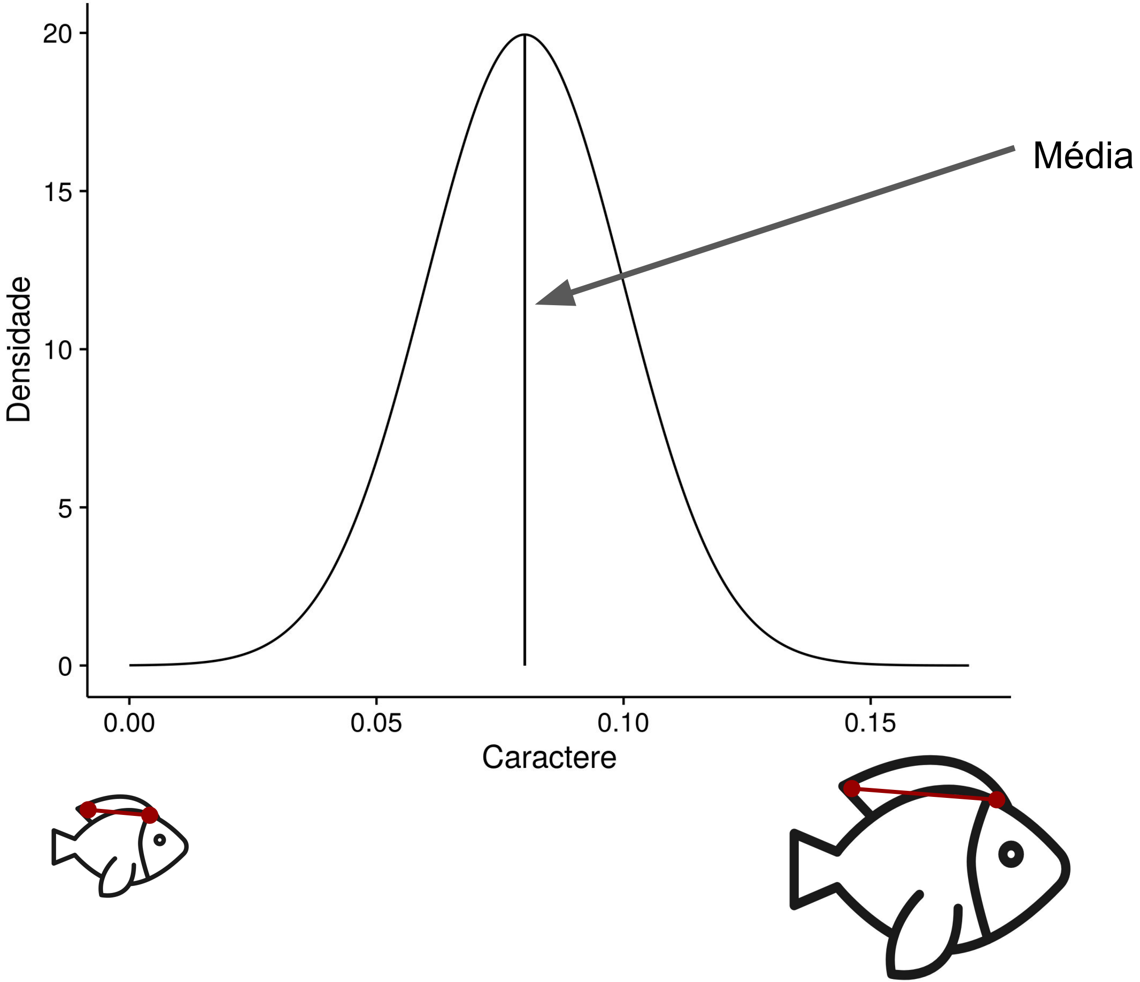
Medindo variação



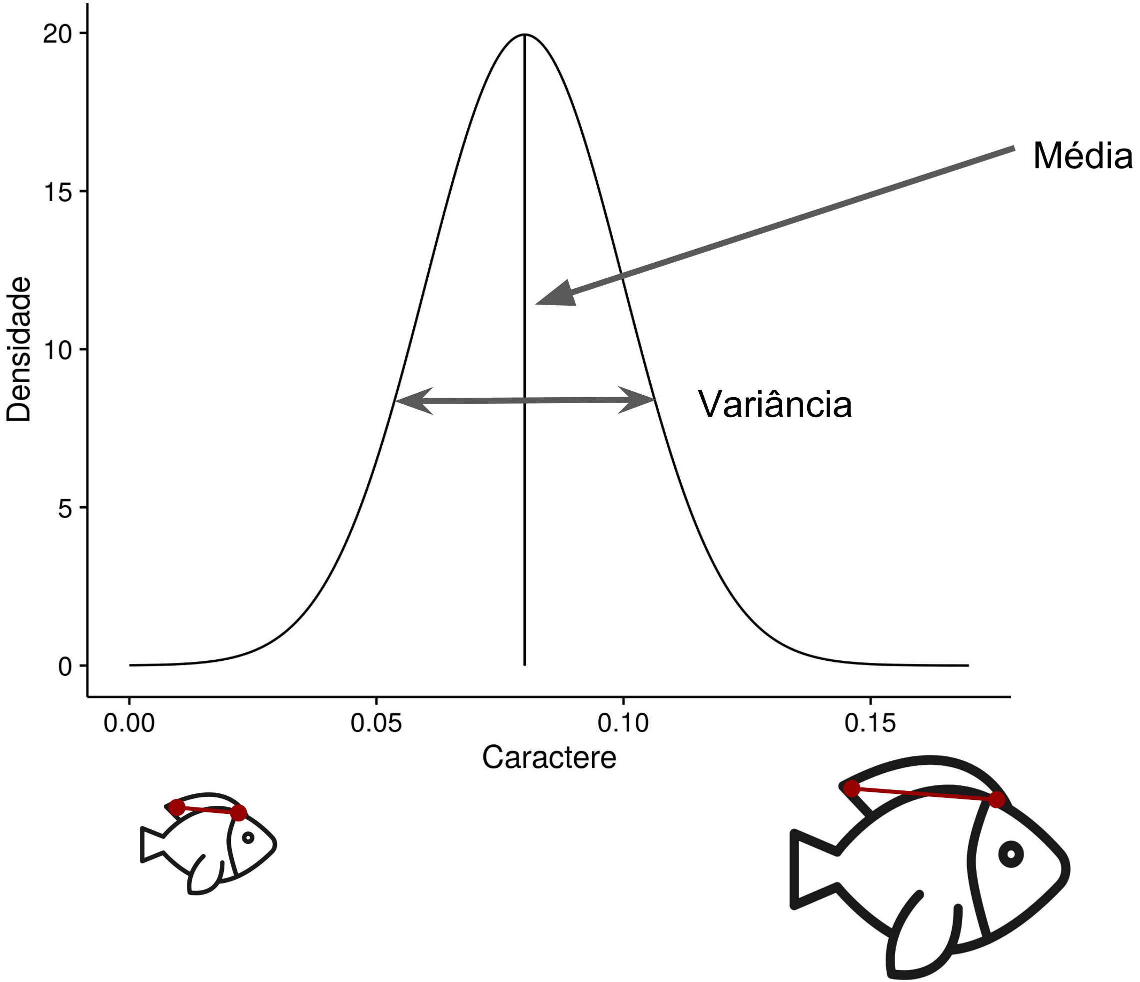
Variação



Variação



Variação



Média
Variância
Densidade
20
15
10
5
0
0.00
0.05
0.10
0.15
Caractere

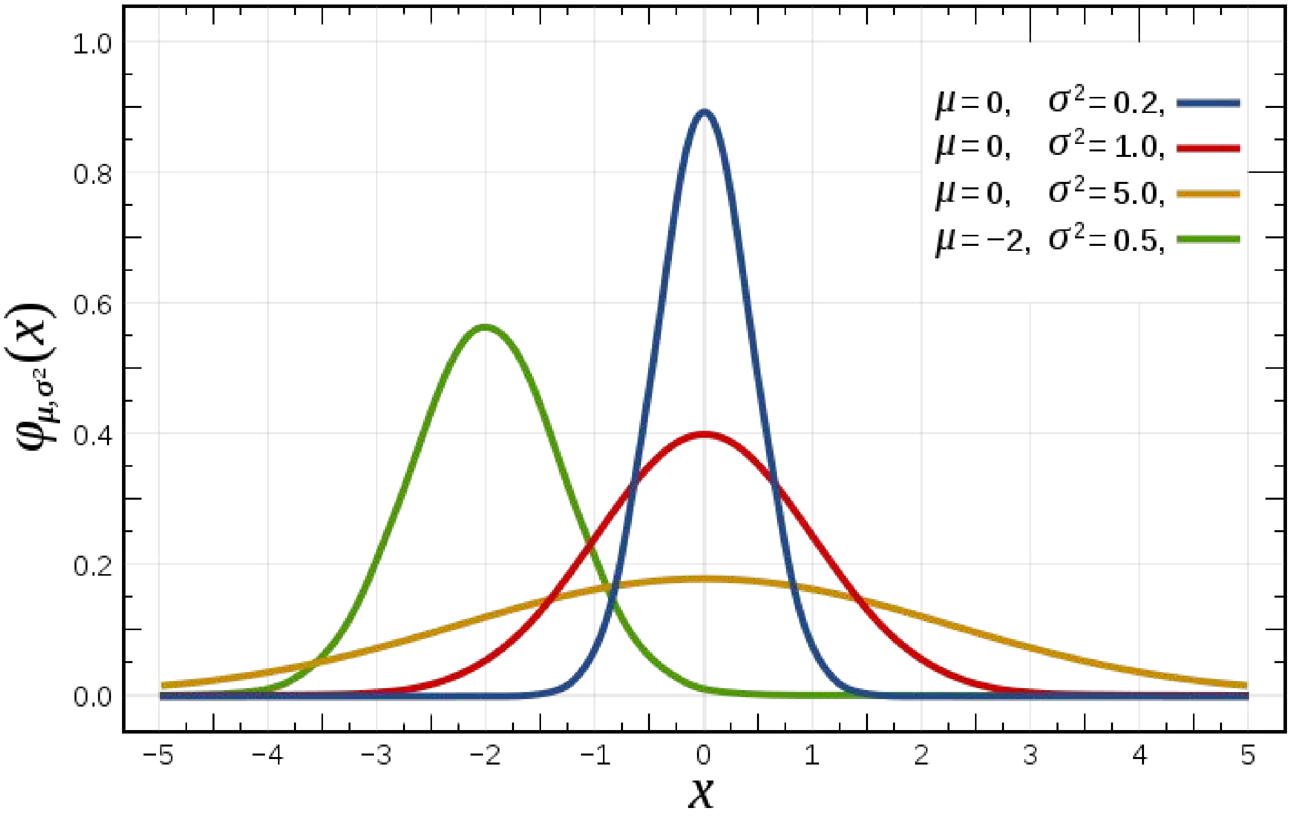
Quantificando variação

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i = \int xp(x)dx$$

$$\sigma_x^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2 = \int (x - \bar{x})^2 p(x)dx$$

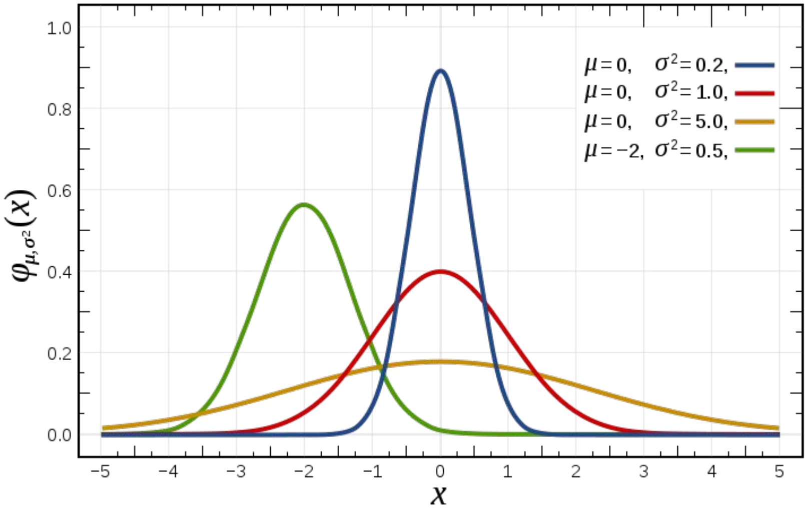
Distribuição normal

$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}$$



Em uma dimensão...

- Média e variância descrevem completamente o problema (em dados normais)
- Variância pode ser particionada em diferentes componentes
- Média e variância podem ser usados para comparar populações



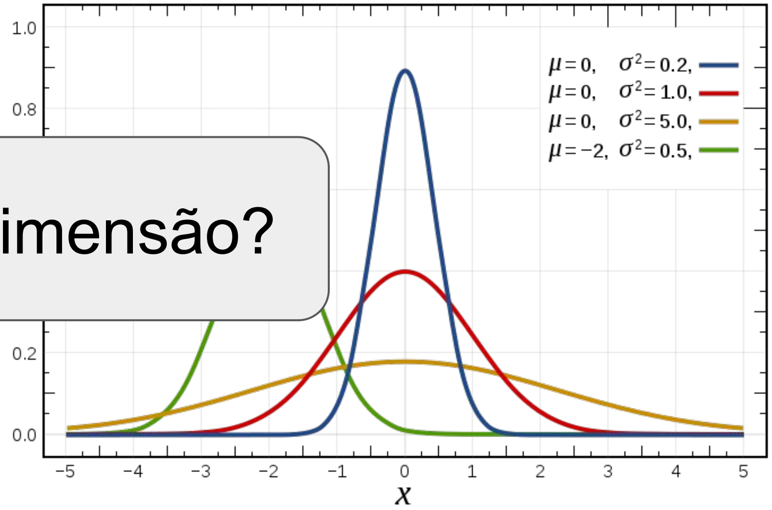
Em uma dimensão...

- Média e variância descrevem completamente o problema (em dados normais)

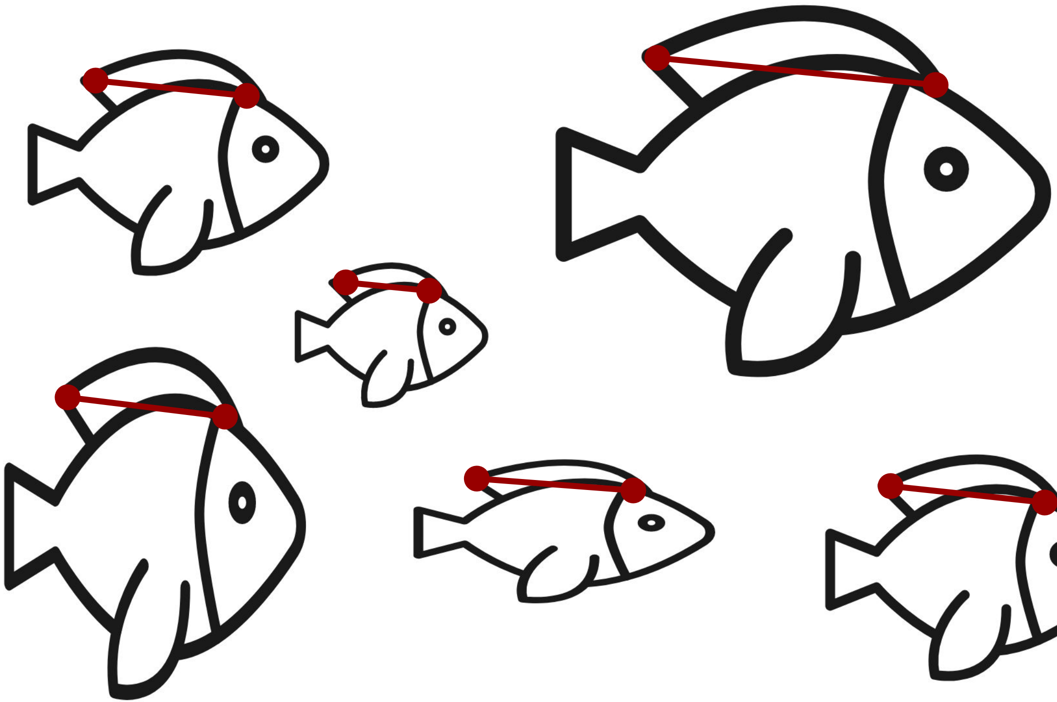
- Variância pode se
componentes

- Média e variância podem ser usados para comparar populações

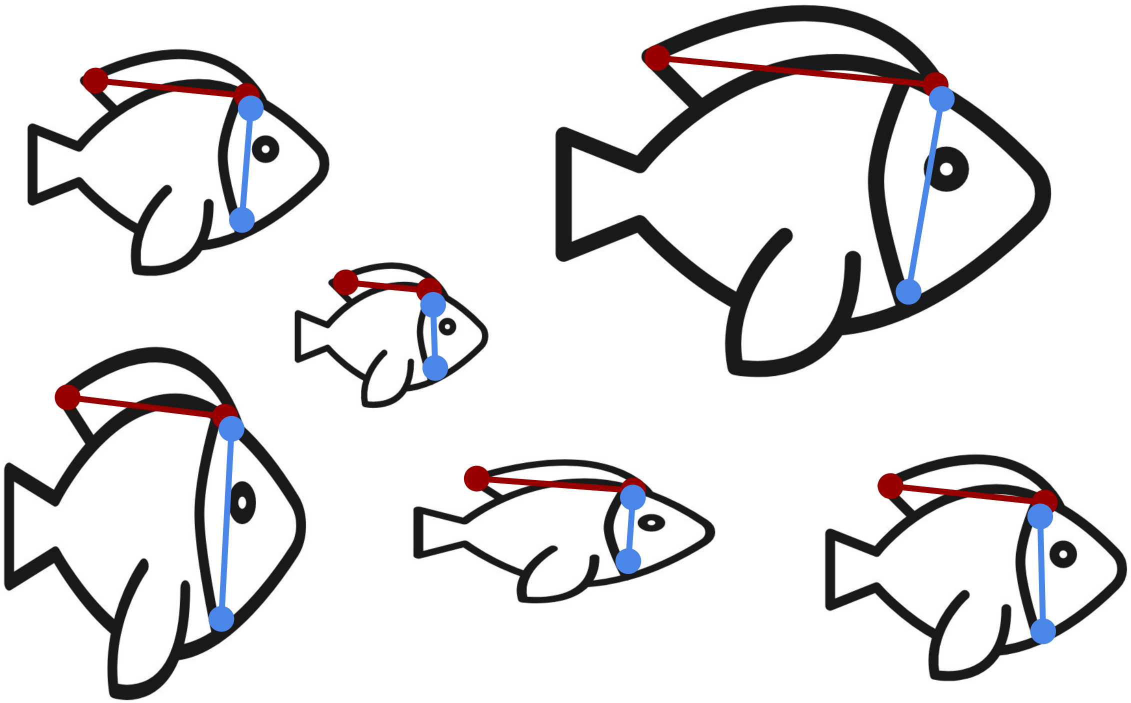
Em mais de uma dimensão?



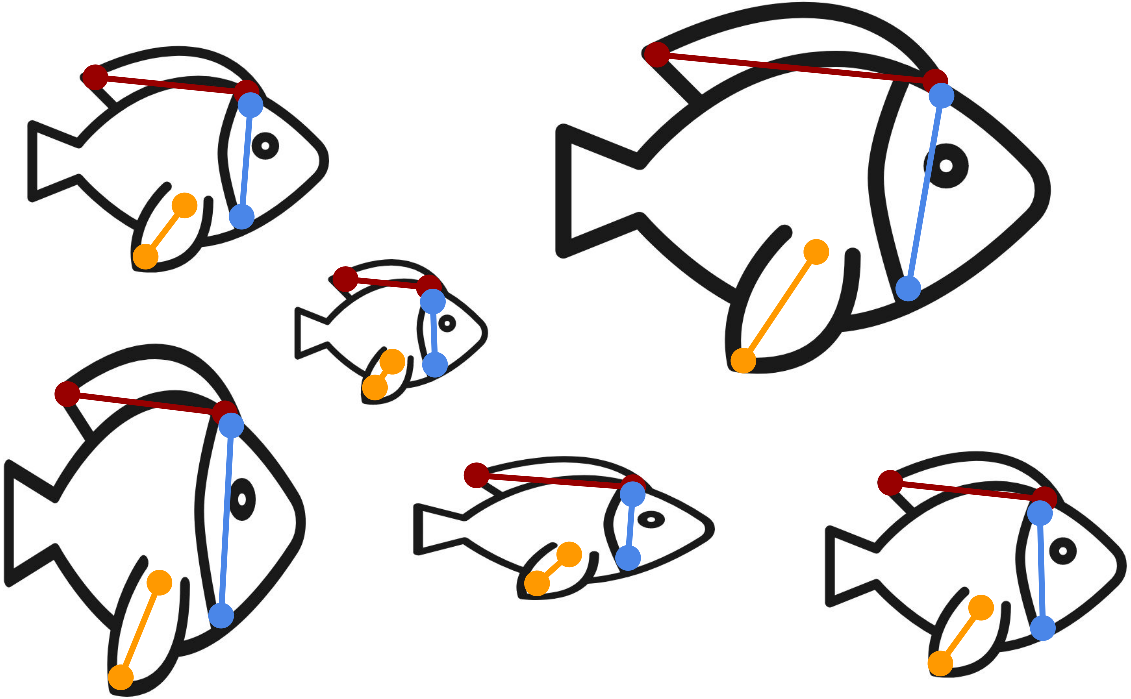
Medindo variação complexa



Medindo variação complexa

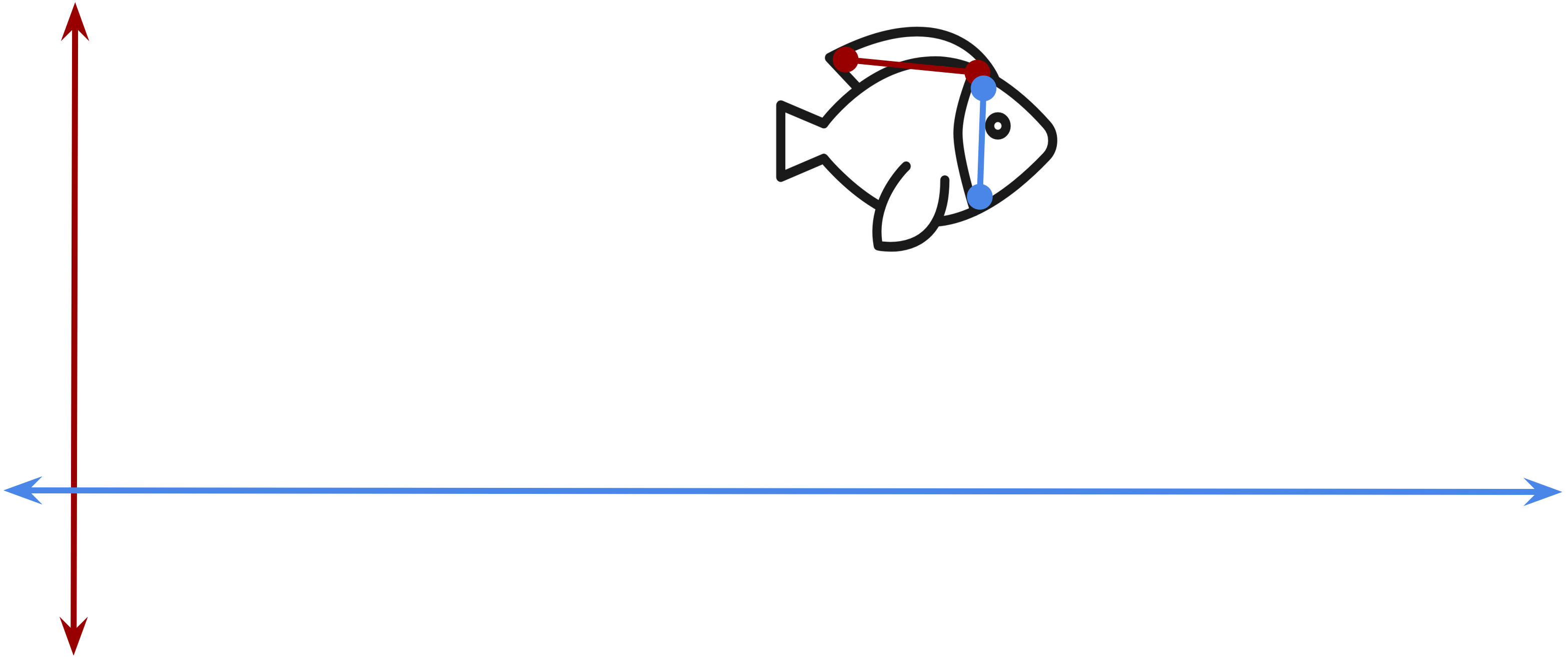


Medindo variação complexa

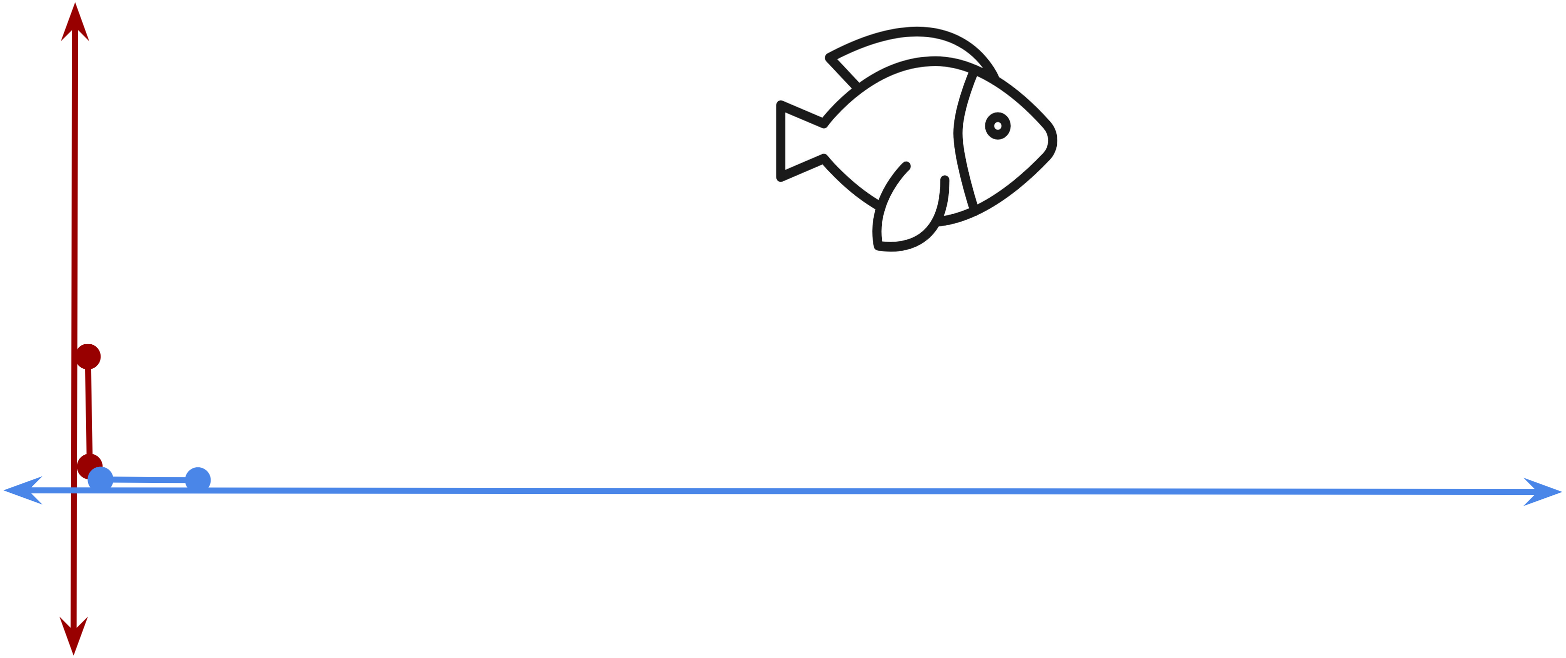


Caracteres como dimensões

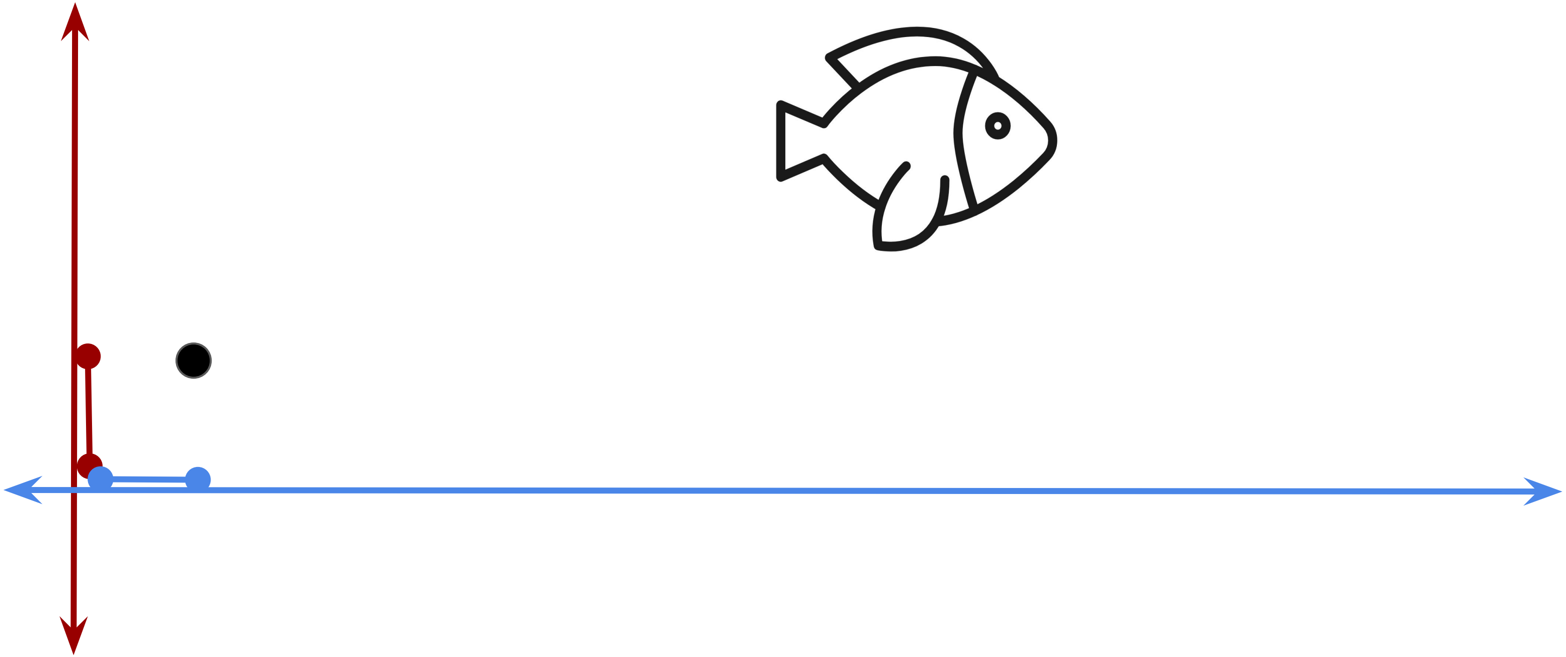
Caracteres como dimensões



Caracteres como dimensões

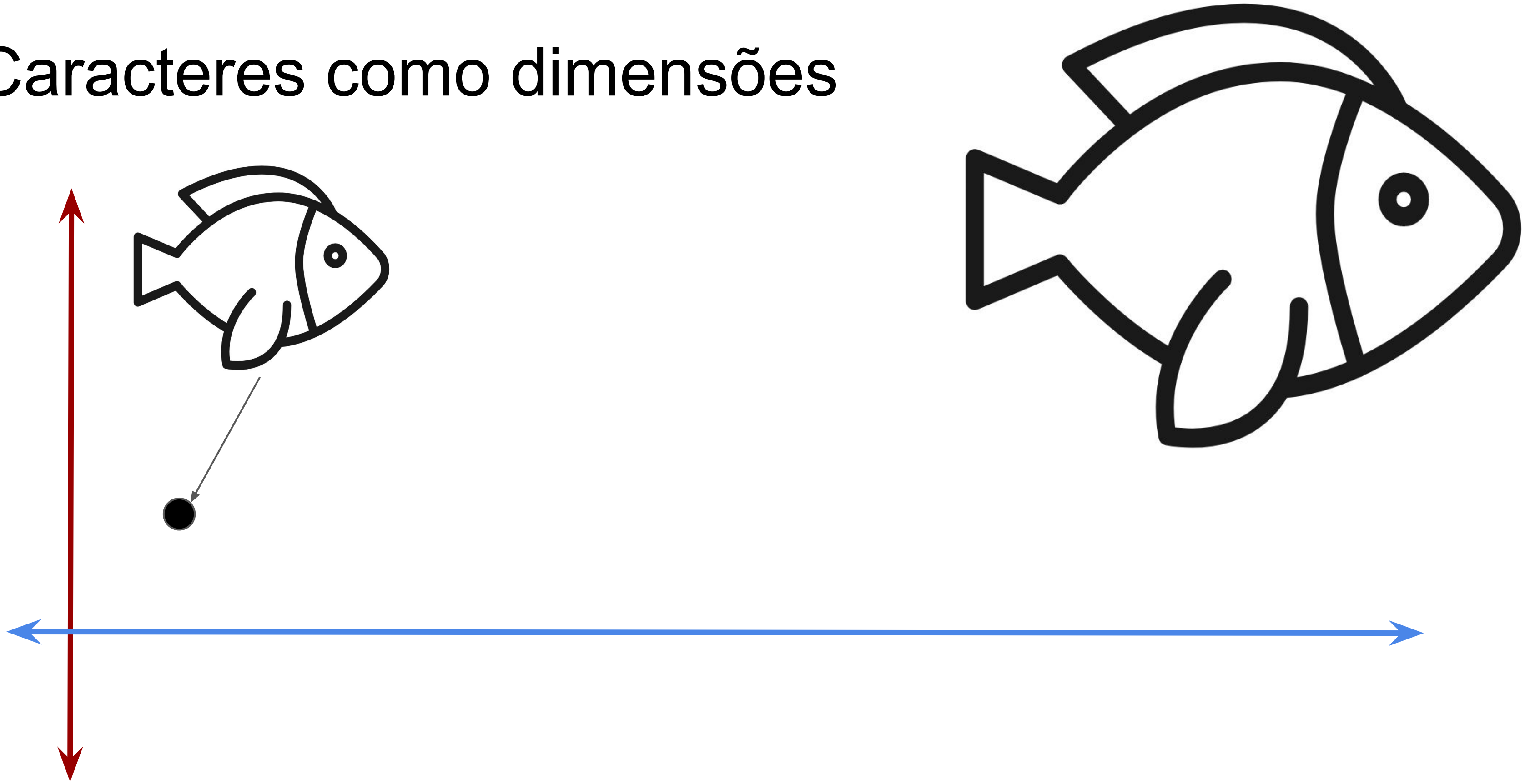


Caracteres como dimensões

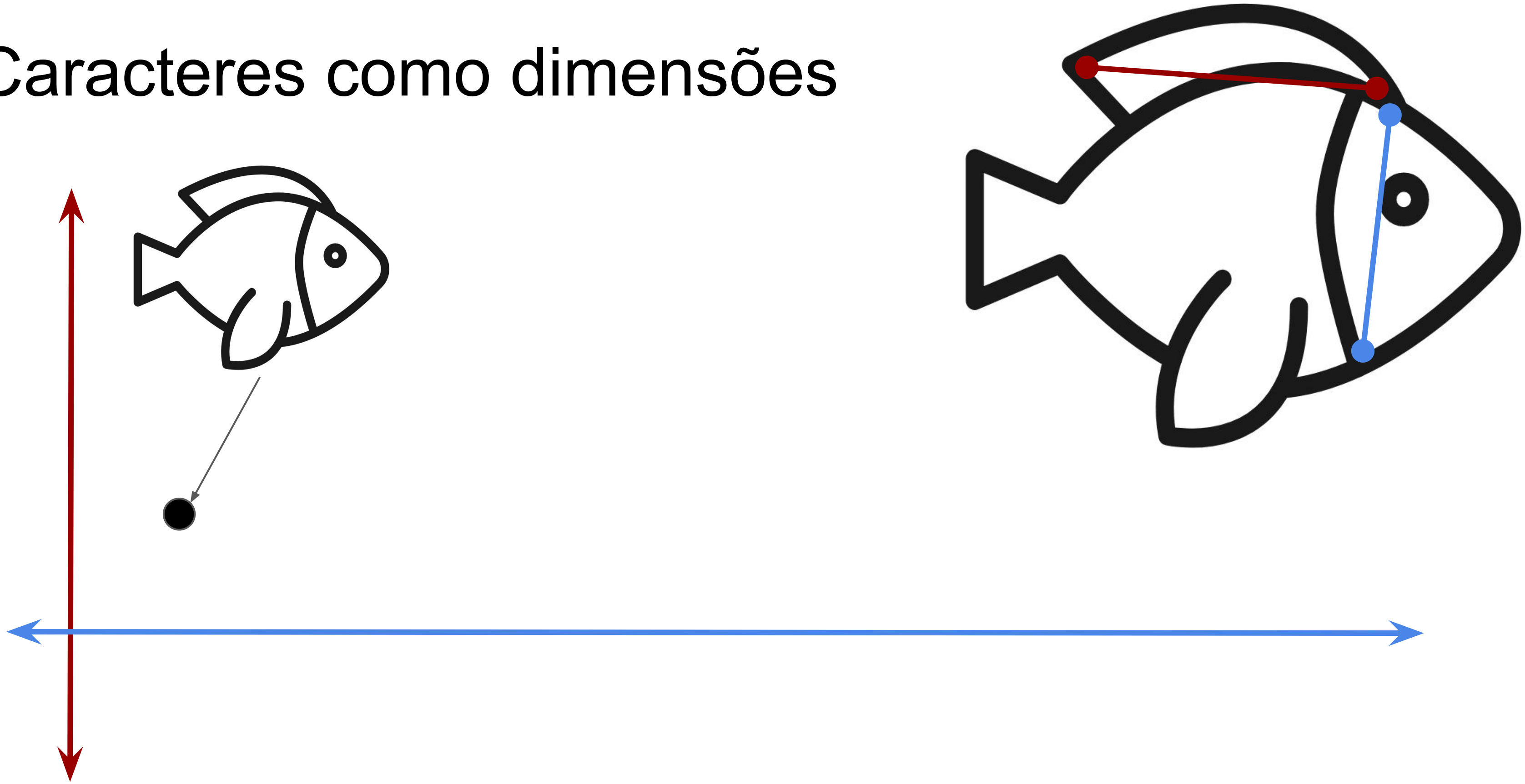


Caracteres como dimensões

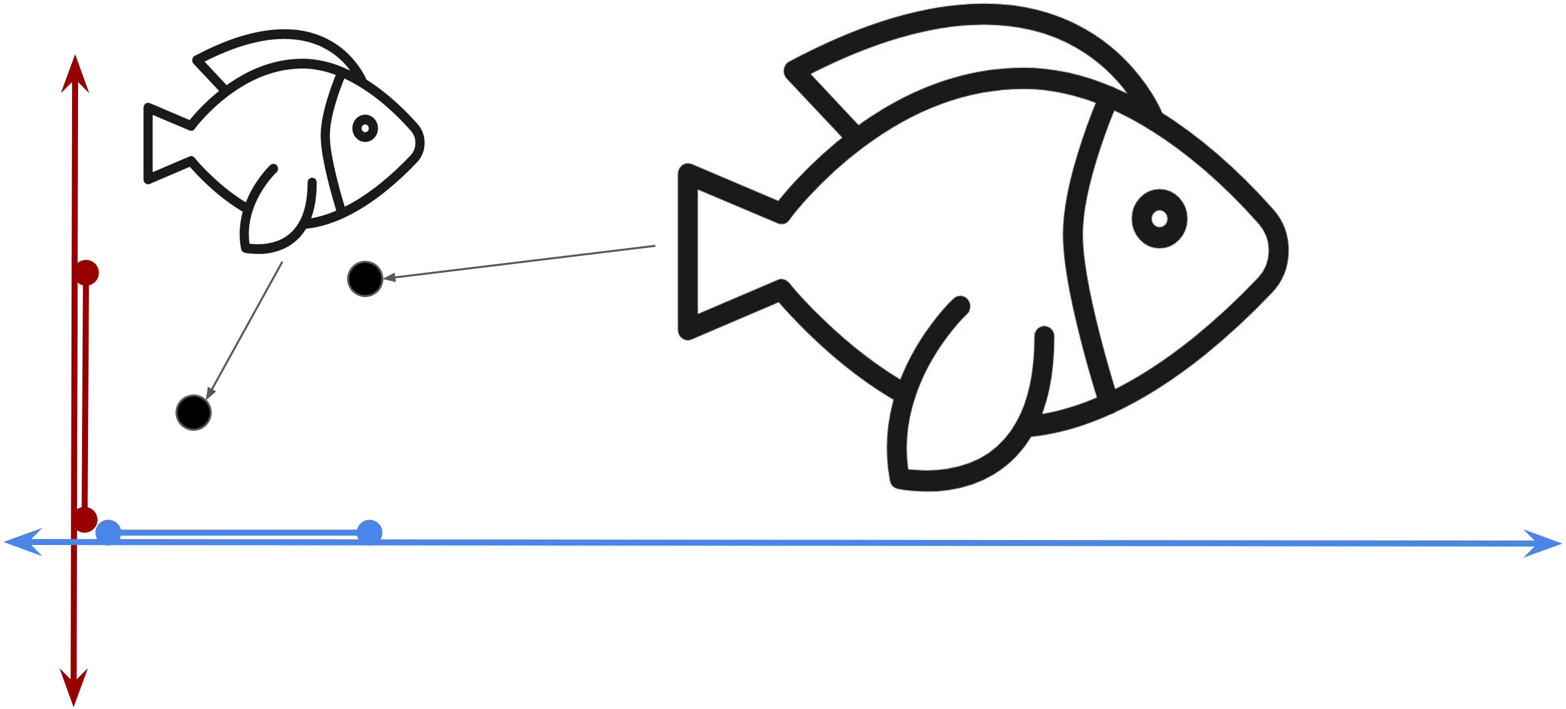
Caracteres como dimensões



Caracteres como dimensões

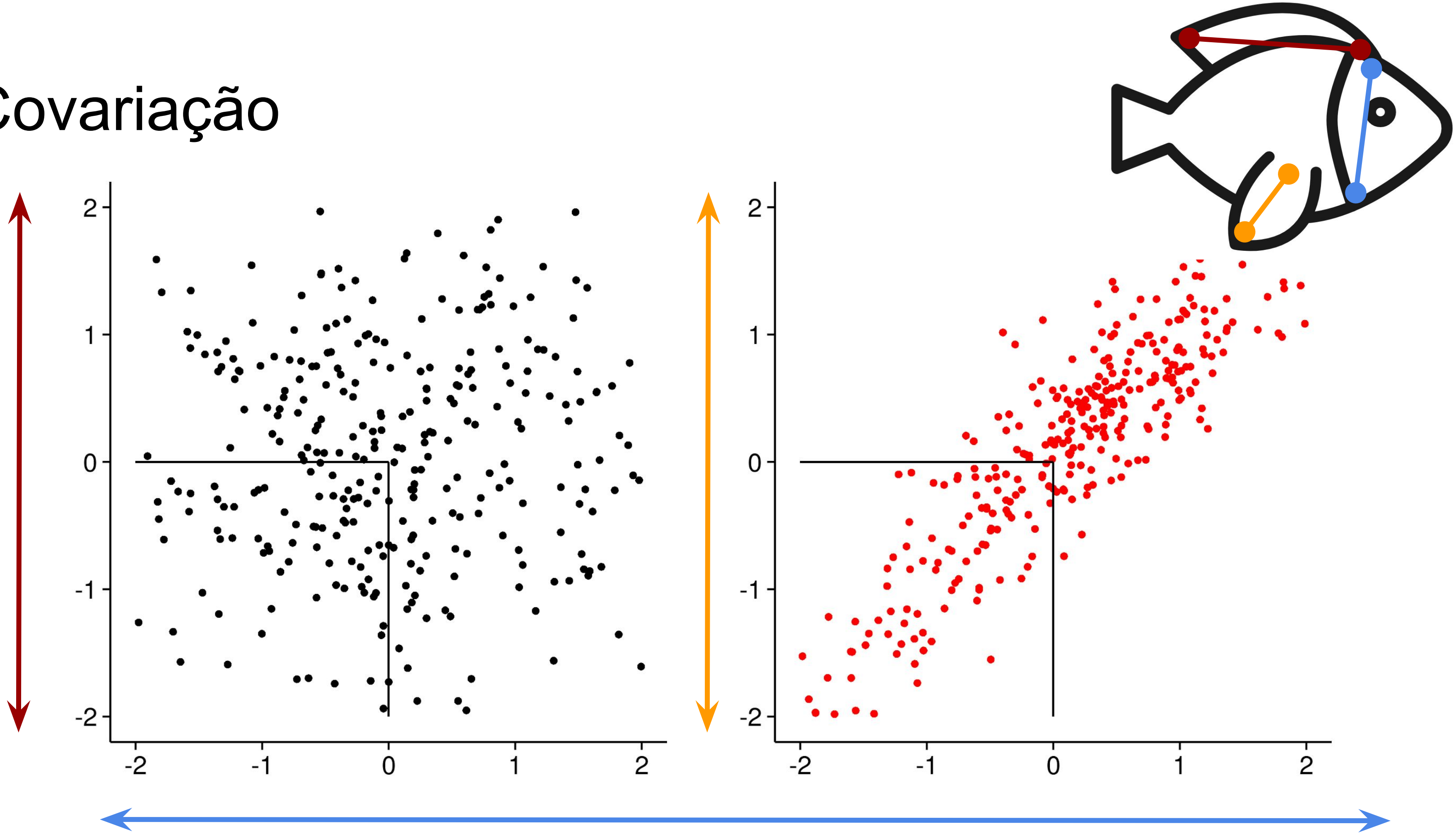


Caracteres como dimensões

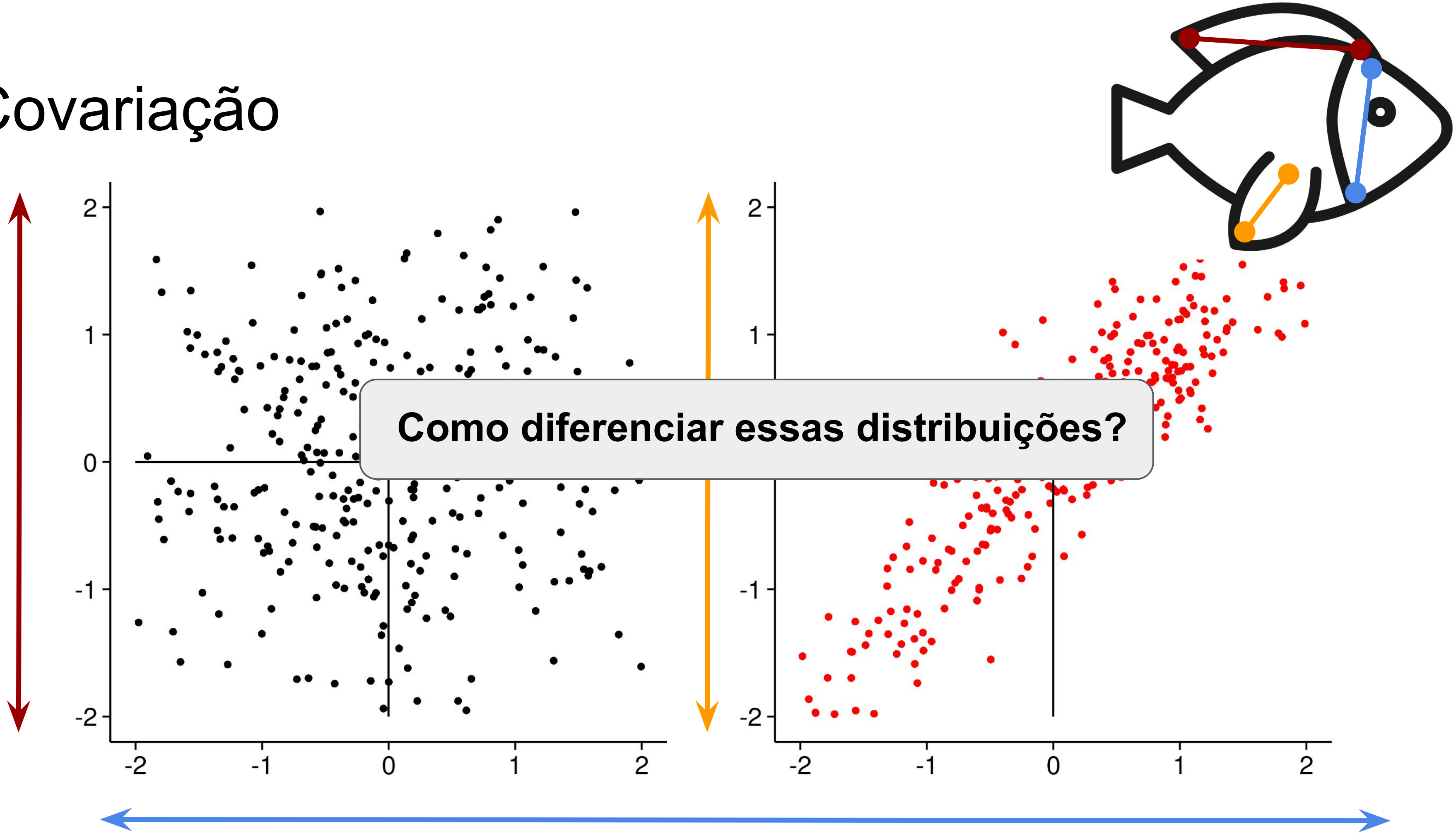


Caracteres como dimensões

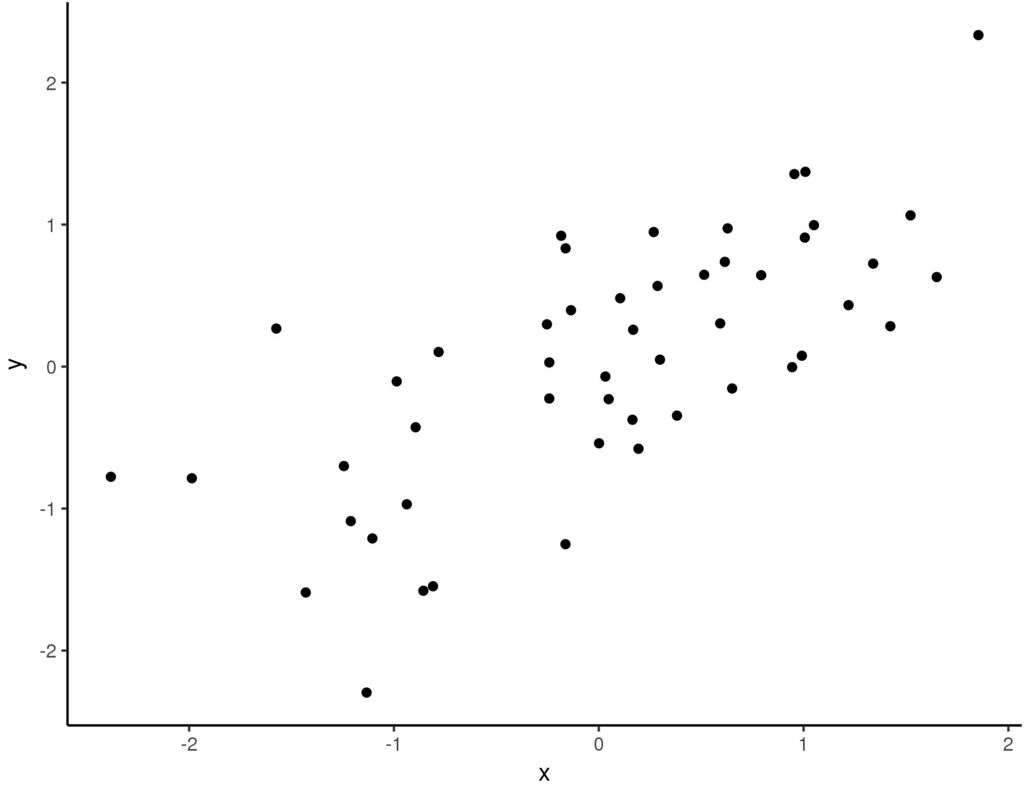
Covariação



Covariação



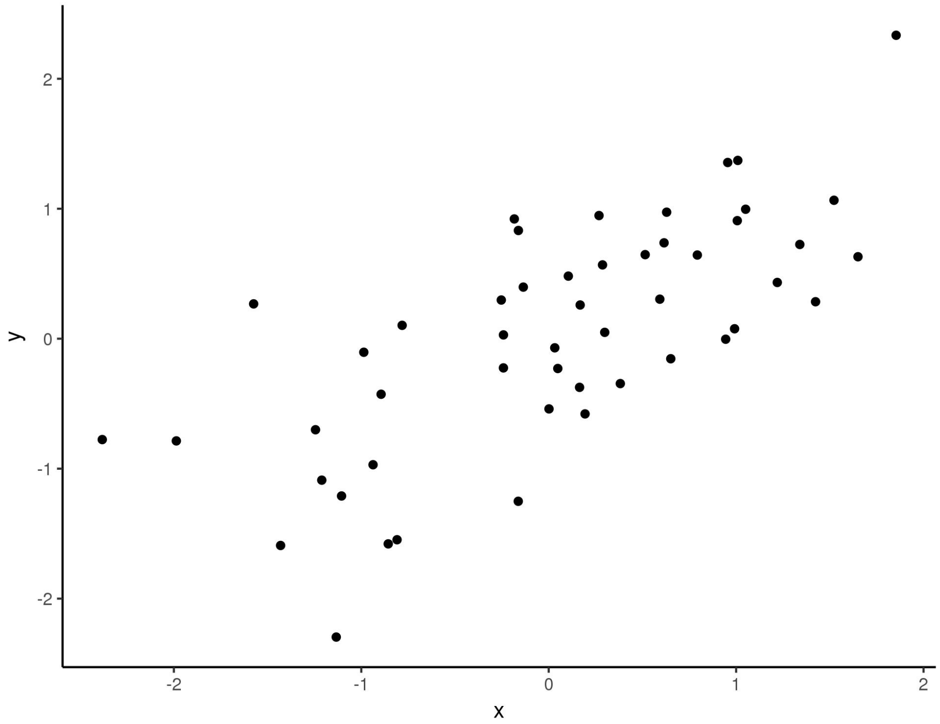
Covariação



Covariância e correlação

$$\sigma_{xy} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})$$

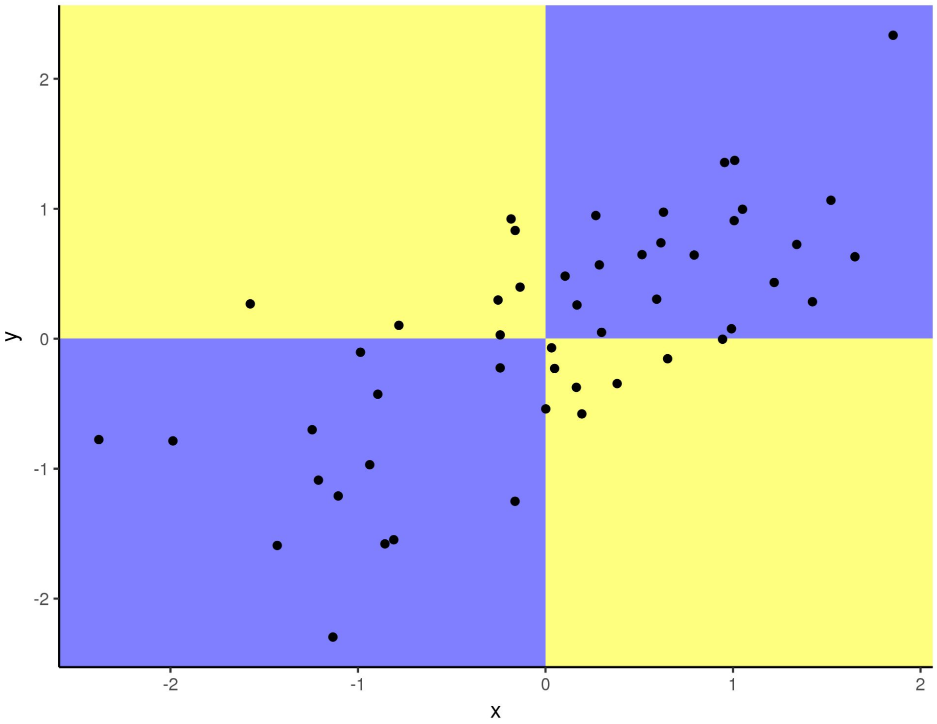
$$\rho_{xy} = \frac{\sigma_{xy}}{\sqrt{\sigma_x^2 \sigma_y^2}}$$



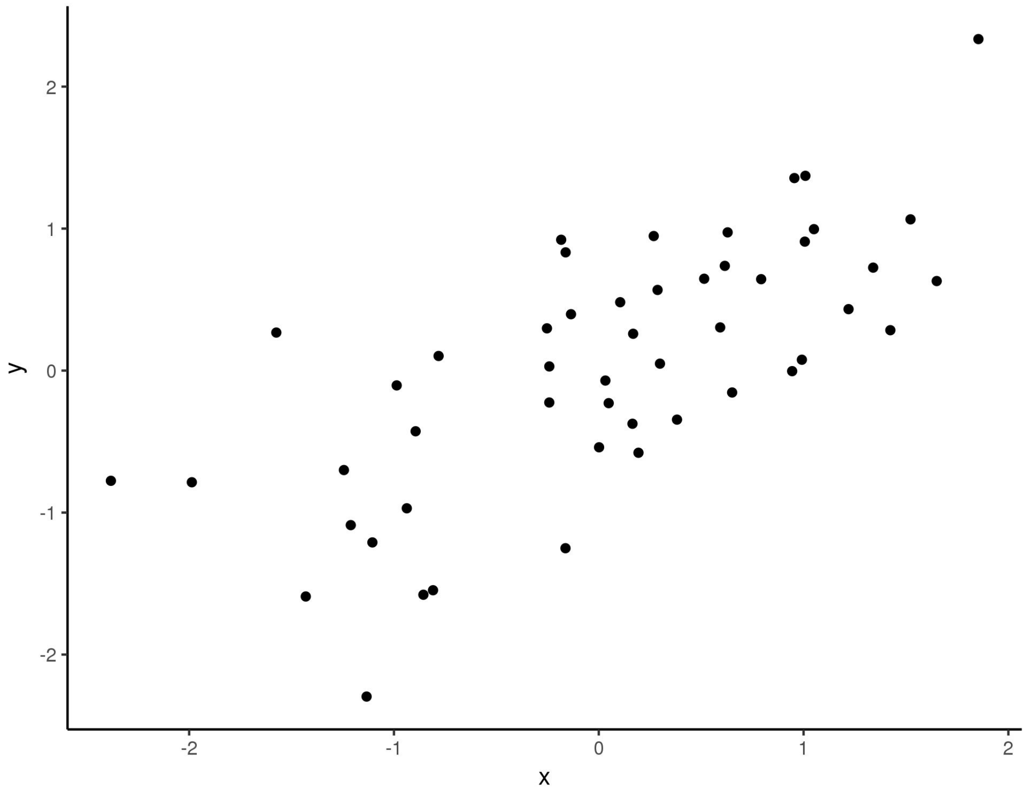
Covariância e correlação

$$\sigma_{xy} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})$$

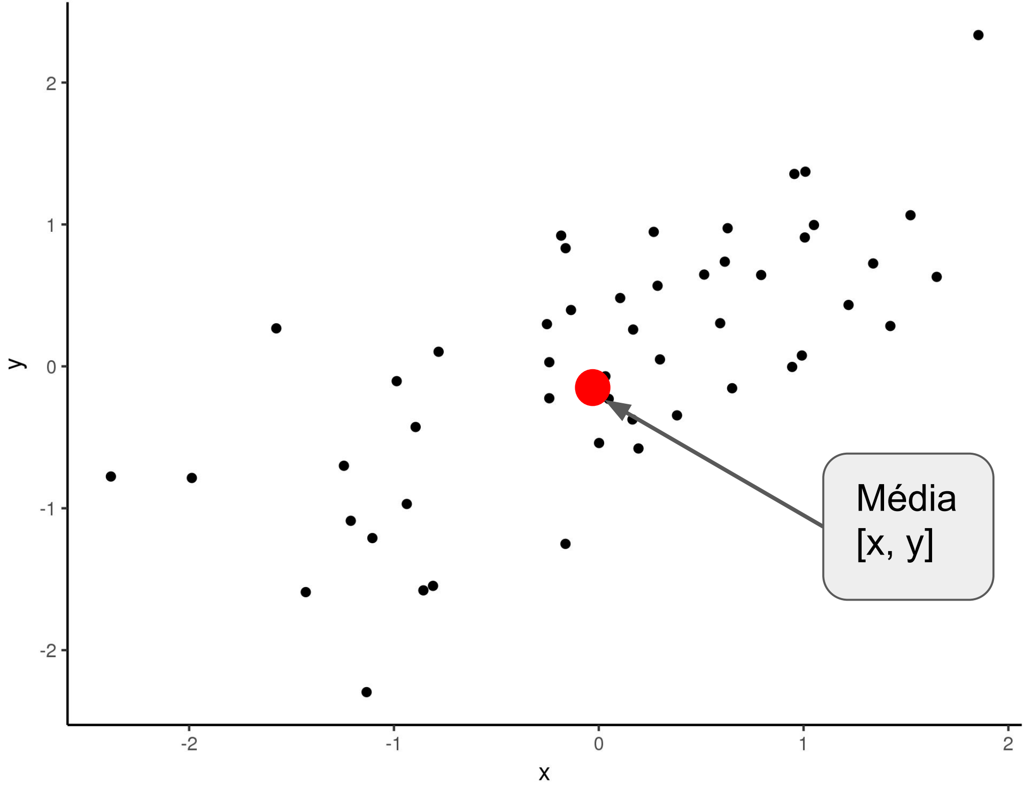
$$\rho_{xy} = \frac{\sigma_{xy}}{\sqrt{\sigma_x^2\sigma_y^2}}$$



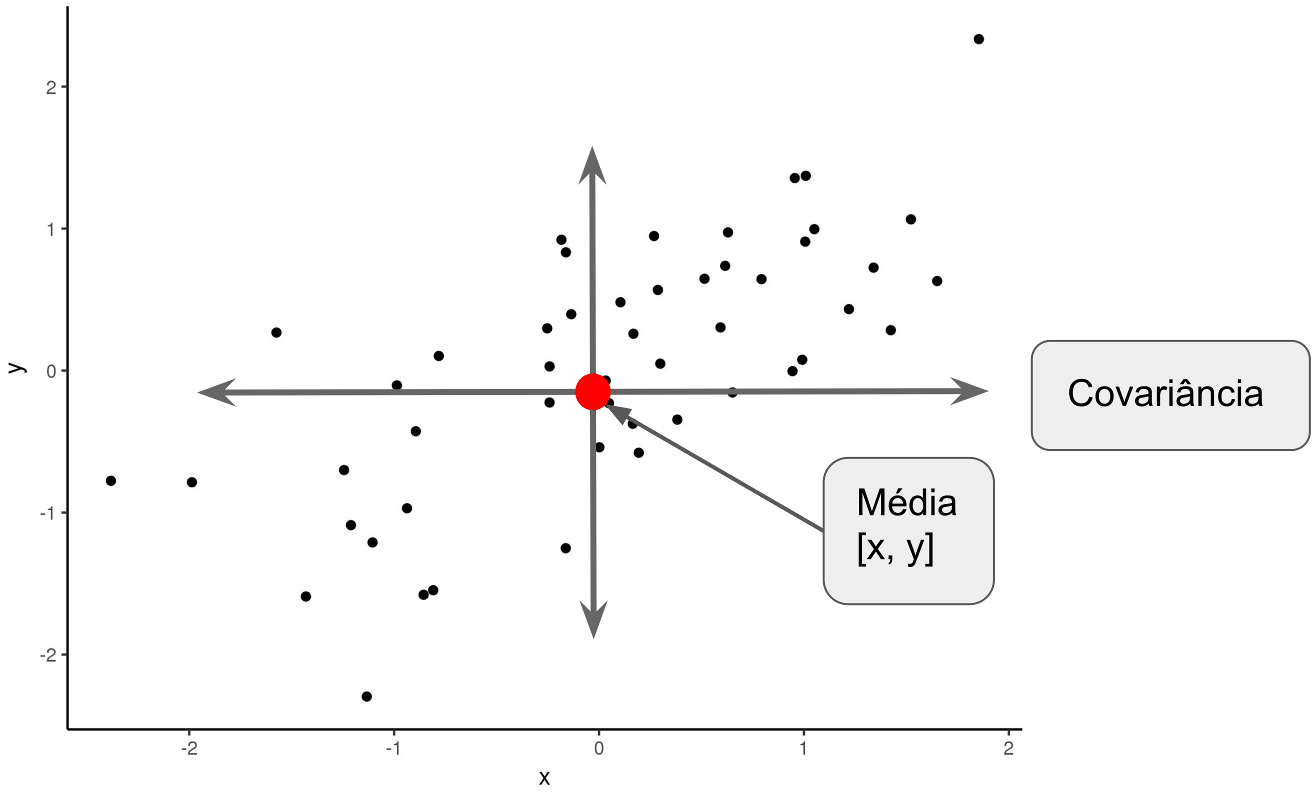
Normal Multivariada



Normal Multivariada



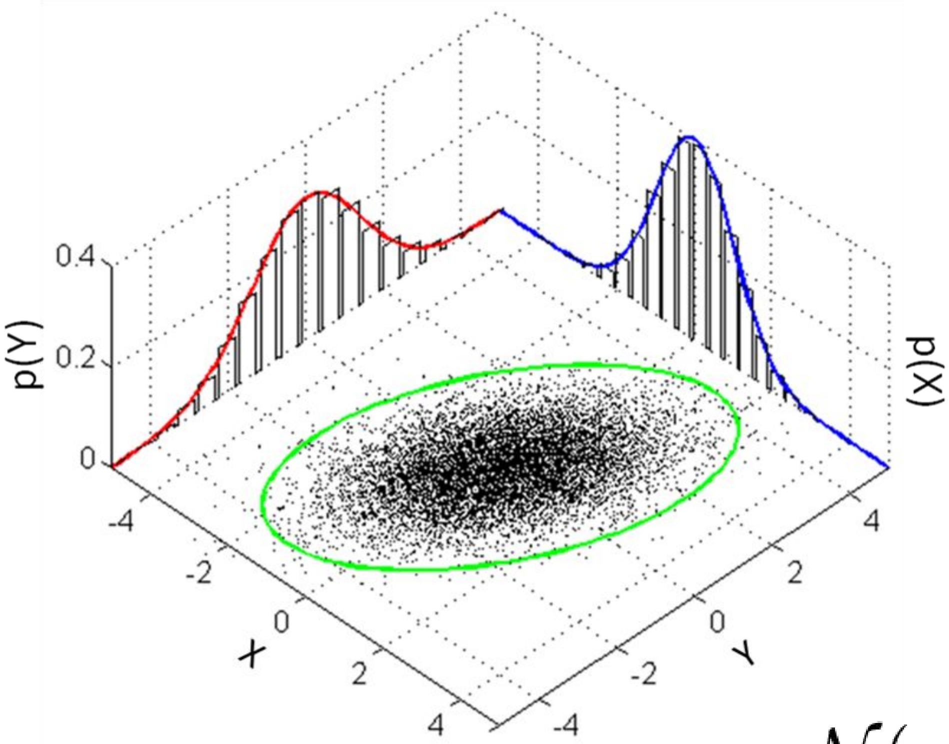
Normal Multivariada

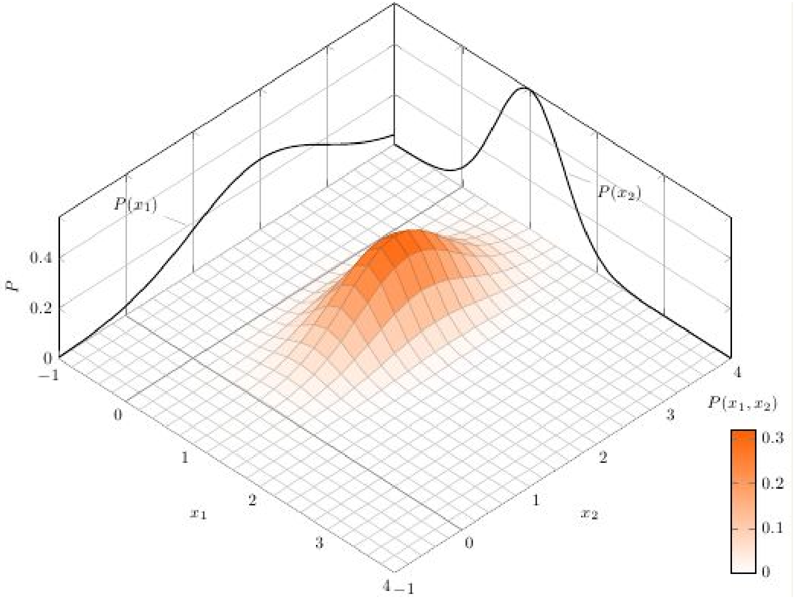


Matriz de covariância

$$\Sigma = \begin{bmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{yx} & \sigma_y^2 \end{bmatrix}$$

Normal Multivariada



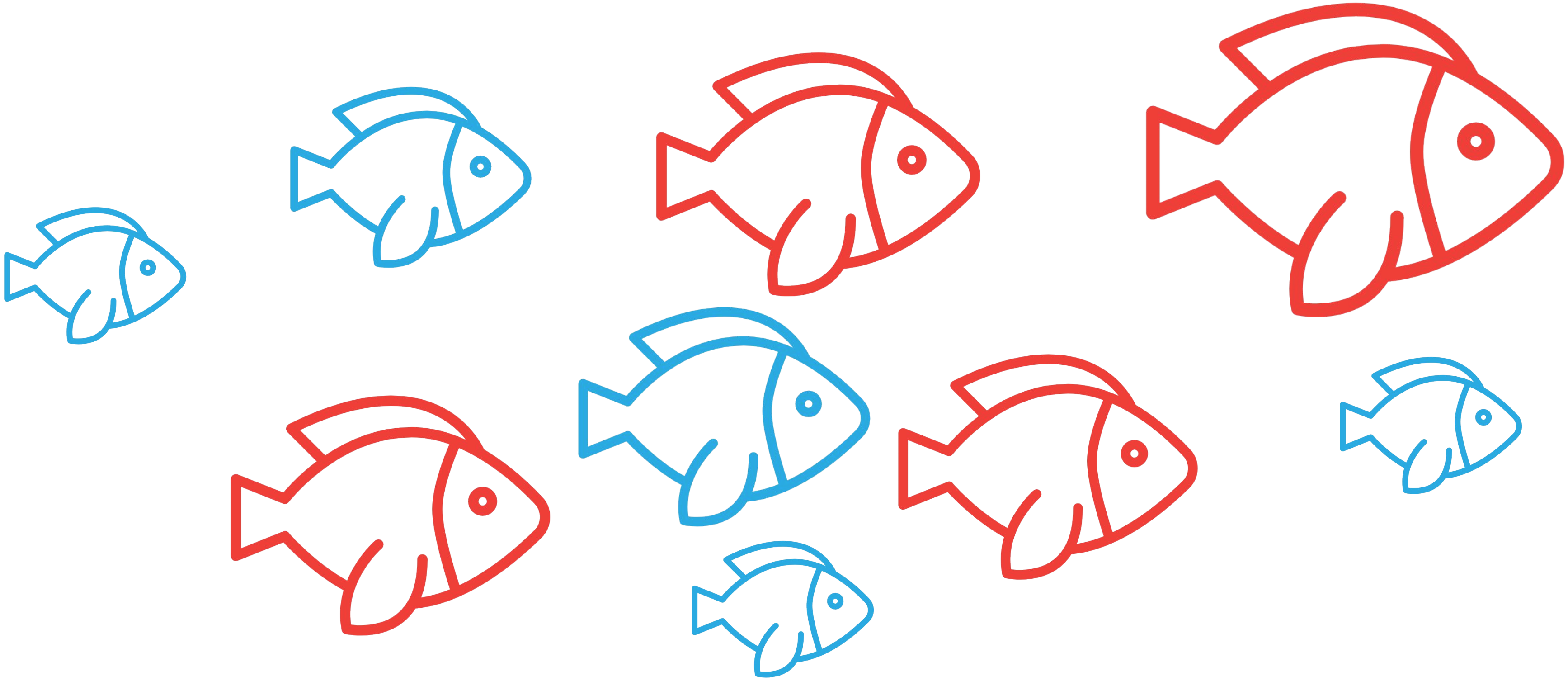


$$\Sigma_{ij} = \sigma_{x^i x^j}$$

$$\mathbf{x}_i = \{x_i^1, x_i^2, \cdots, x_i^p\}$$

$$\mathcal{N}(\mathbf{x}_i|\mu, \Sigma) \approx e^{\frac{1}{2}(\mathbf{x}_i-\mu)^T \Sigma^{-1}(\mathbf{x}_i-\mu)}$$

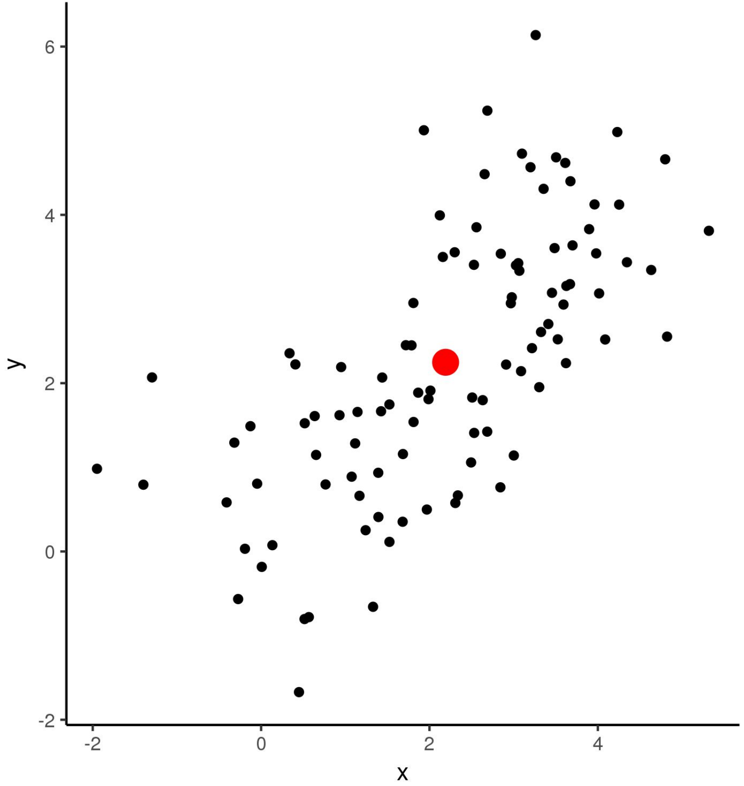
Populações estruturadas ou dimórficas



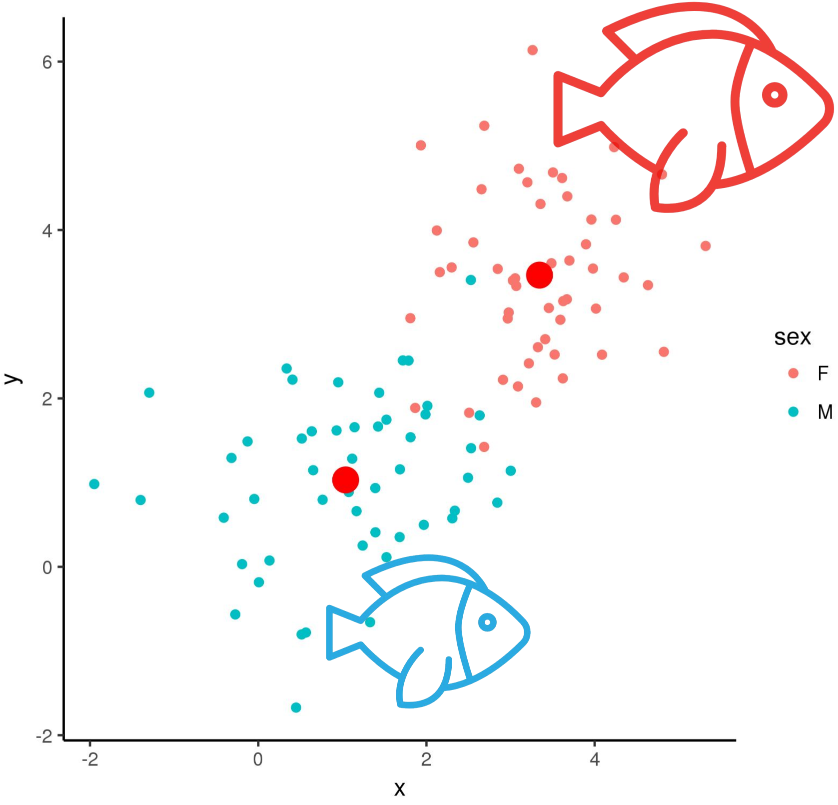
Populações estruturadas ou dimórficas

$$\boldsymbol{\Sigma} = \begin{bmatrix} 2.24, & 1.57 \\ 1.57, & 2.47 \end{bmatrix}$$

$$\varrho = \begin{bmatrix} 1 & 0.6 \\ 0.6 & 1 \end{bmatrix}$$

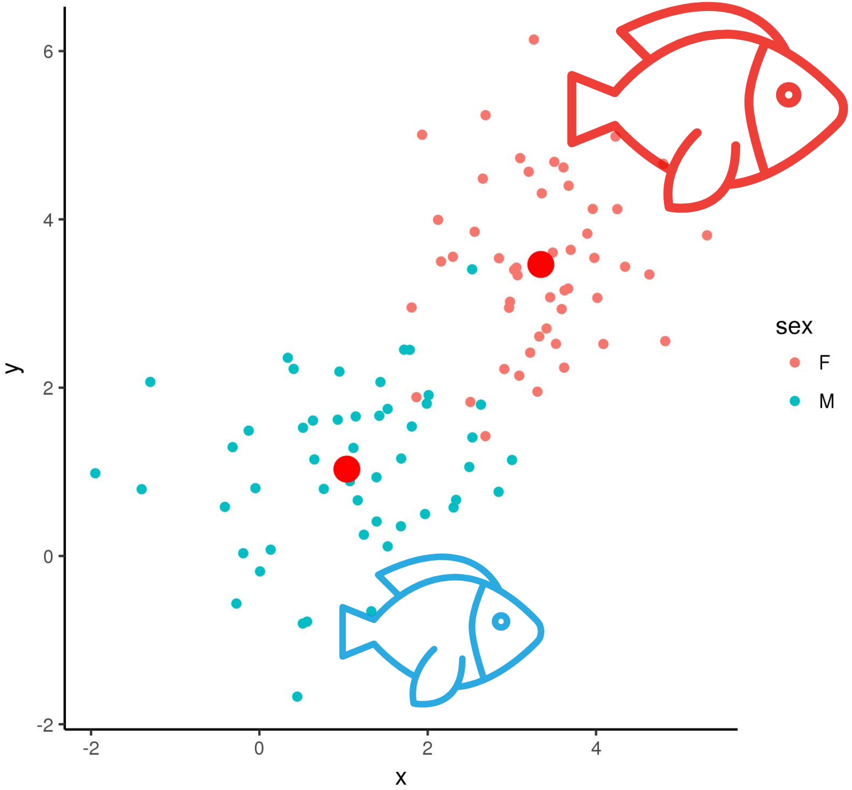


Populações estruturadas ou dimórficas



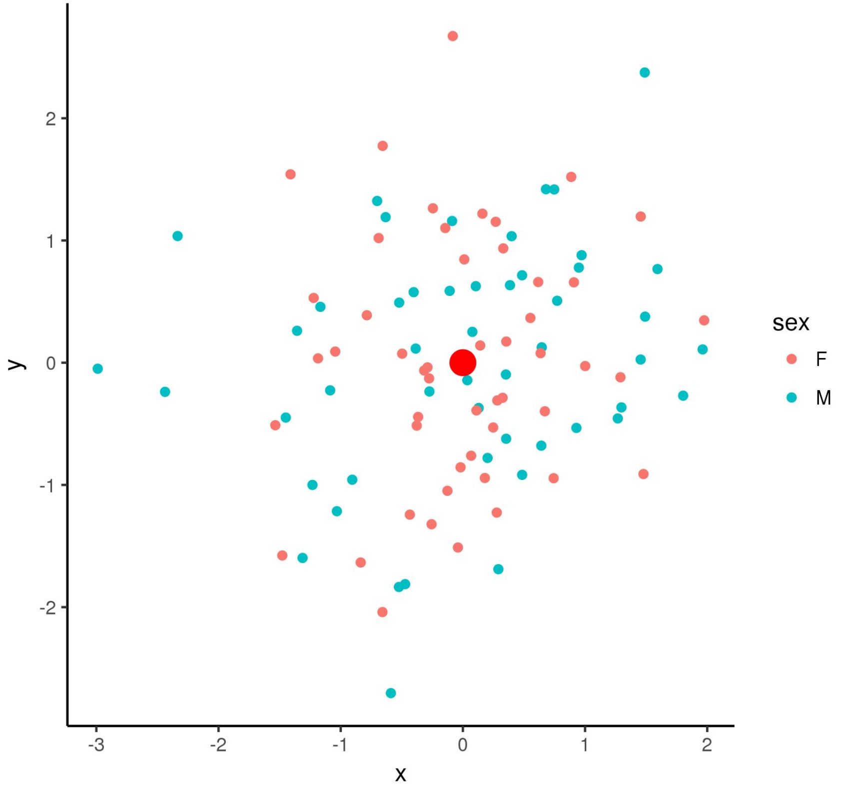
Populações estruturadas ou dimórficas

$$\sigma_{xy} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x}_{g[i]})(y_i - \bar{y}_{g[i]})$$



Populações estruturadas ou dimórficas

$$\sigma_{xy} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x}_{g[i]})(y_i - \bar{y}_{g[i]})$$

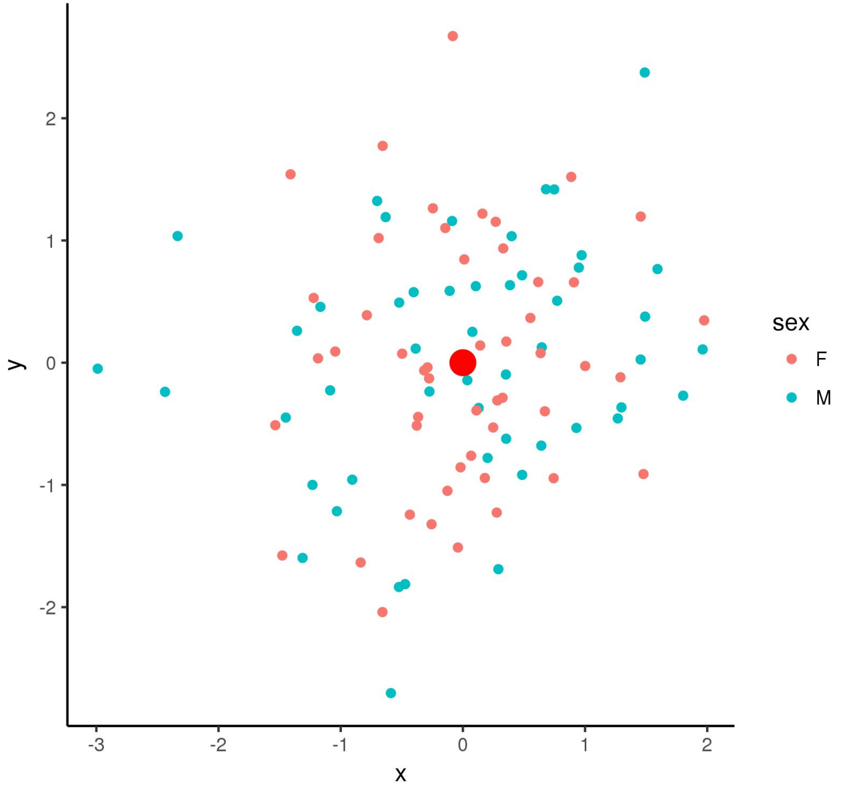


Populações estruturadas ou dimórficas

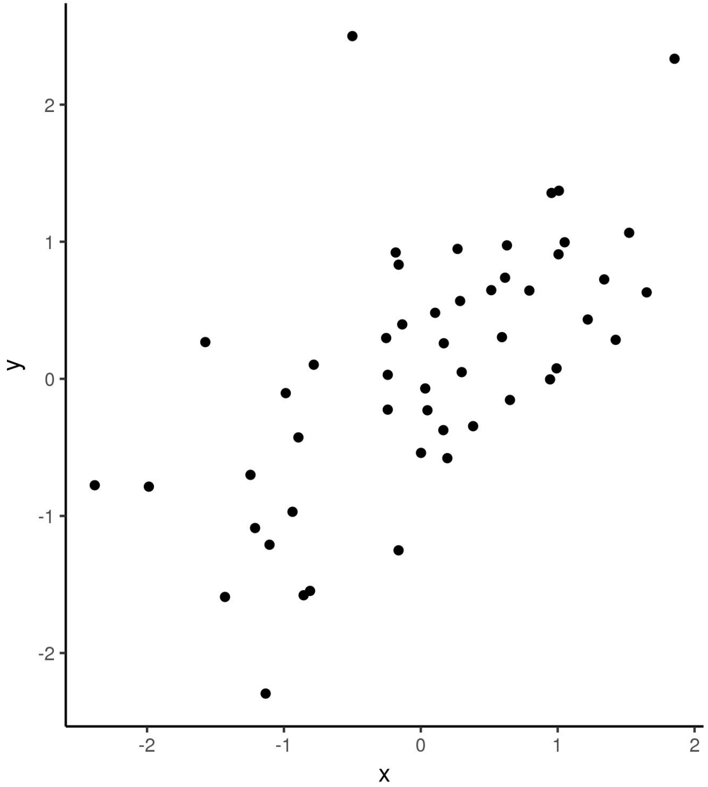
$$\sigma_{xy} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x}_{g[i]})(y_i - \bar{y}_{g[i]})$$

$$\mathbf{\Sigma} = \begin{bmatrix} 0.90, & 0.15 \\ 0.15, & 0.97 \end{bmatrix}$$

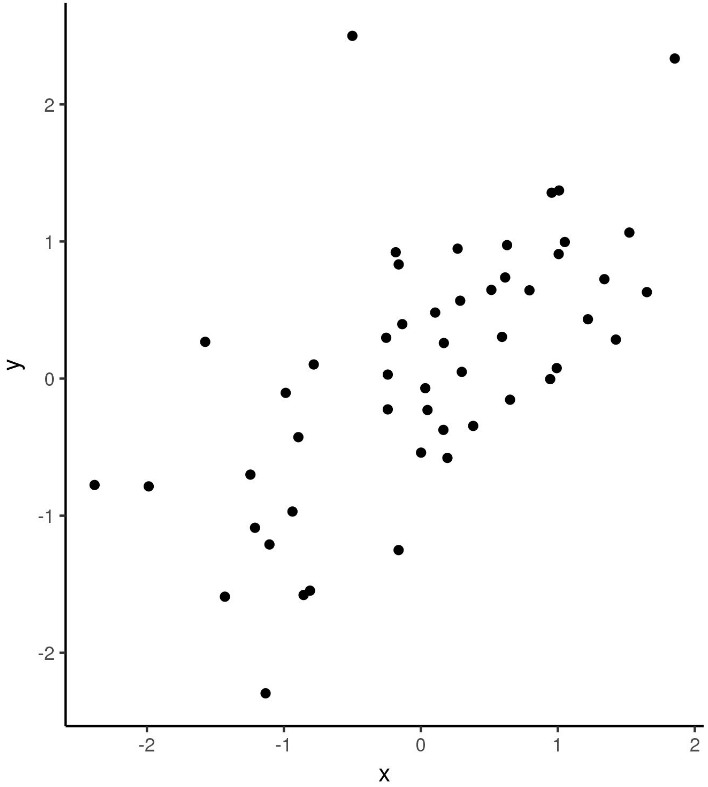
$$\varrho = \begin{bmatrix} 1 & 0.16 \\ 0.16 & 1 \end{bmatrix}$$



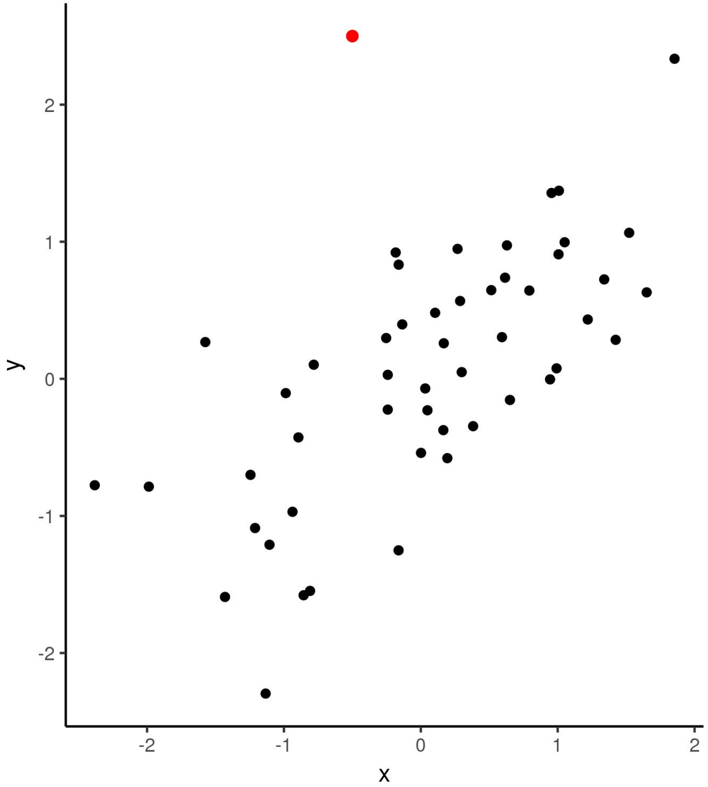
Complicações multivariadas



Outliers multivariadas

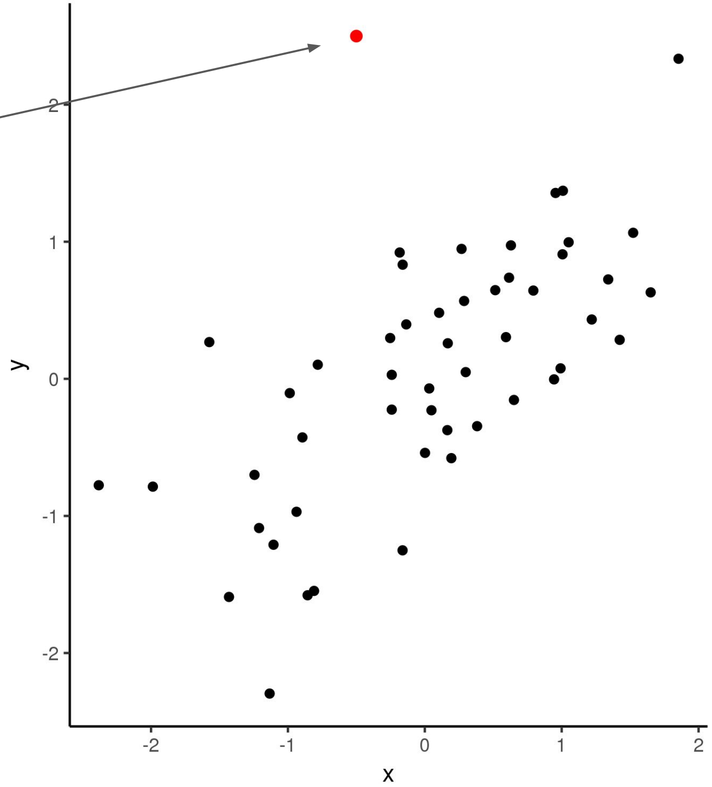


Outliers multivariadas



Outliers multivariadas

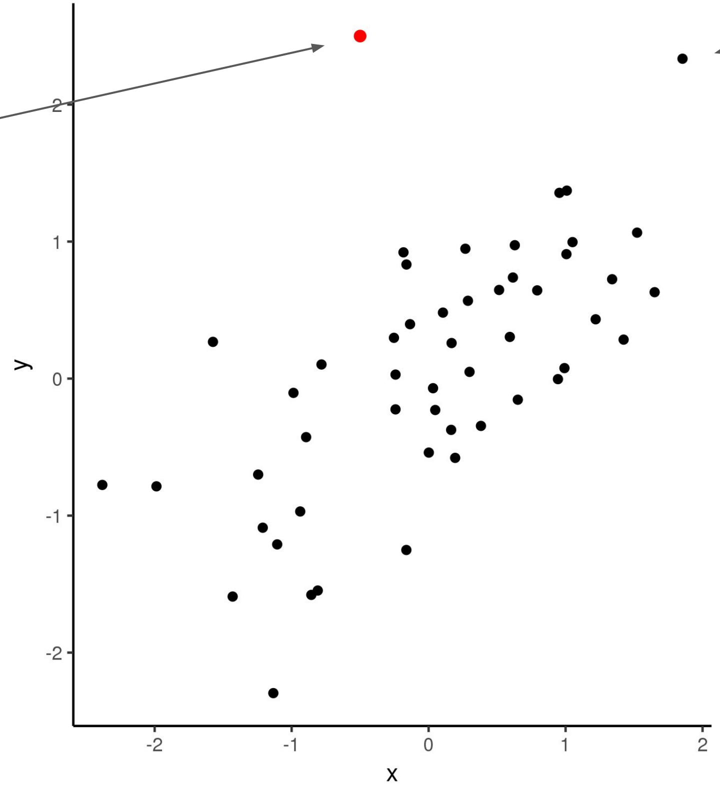
Por quê esse?



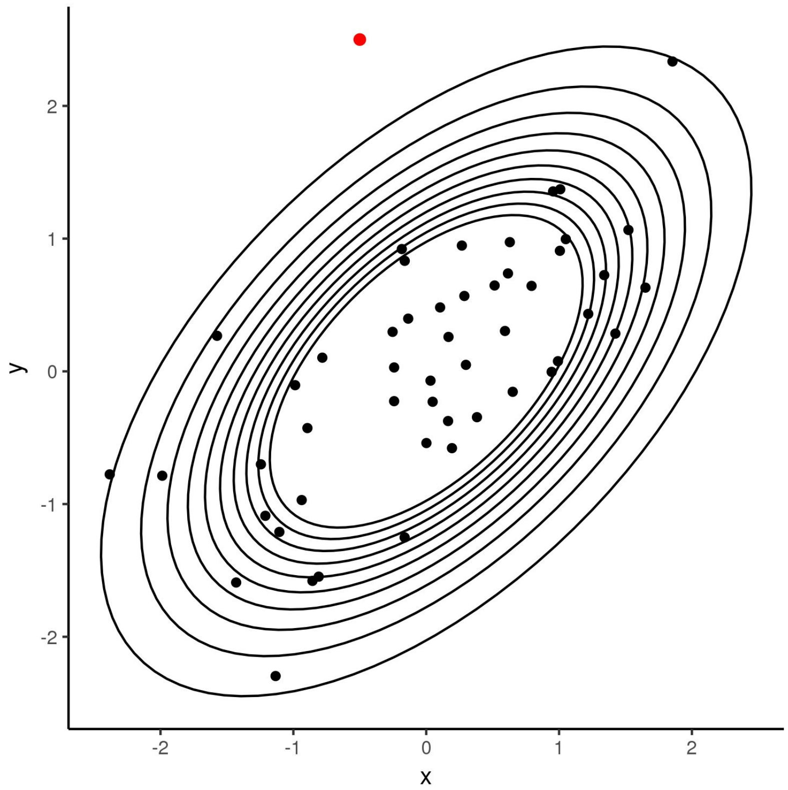
Outliers multivariadas

E não esse?

Por quê esse?

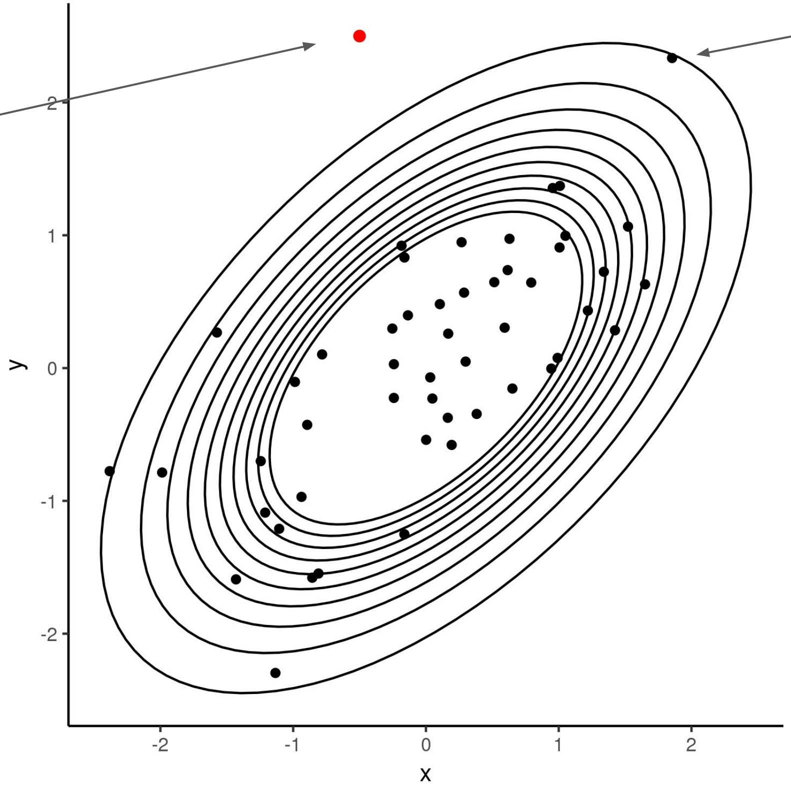


Quantis em 2D



Quantis em 2D

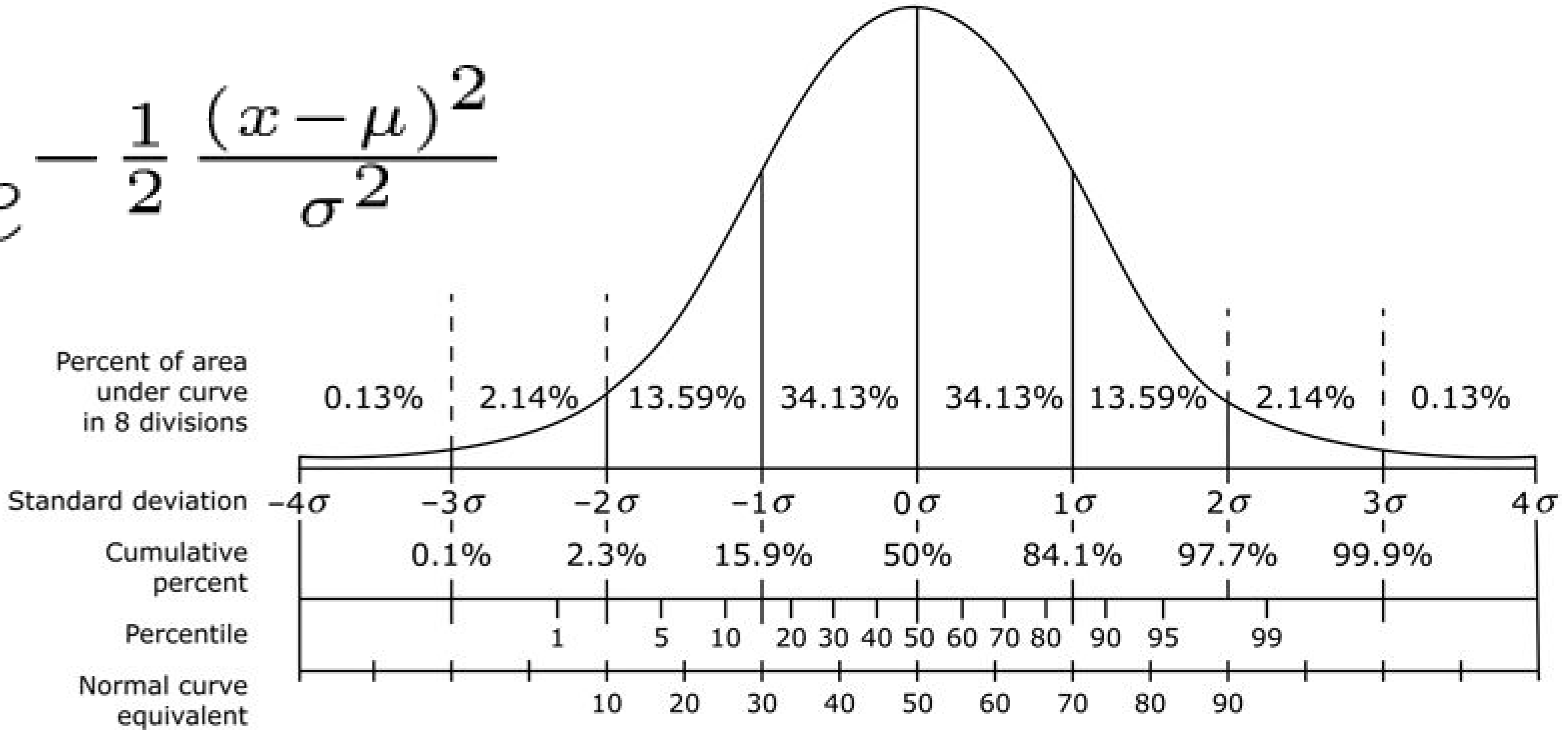
Bem mais longe...



Dentro do quantil 95%

Unidades de desvio padrão

$$e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}$$



Distancia de Mahalanobis

1D

nD

$$\frac{x - \mu}{\sigma} \longrightarrow \sqrt{(\mathbf{x} - \mu)^T \Sigma^{-1} (\mathbf{x} - \mu)}$$

Distancia de Mahalanobis

1D

$$\frac{x - \mu}{\sigma}$$

$$e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}$$

nD

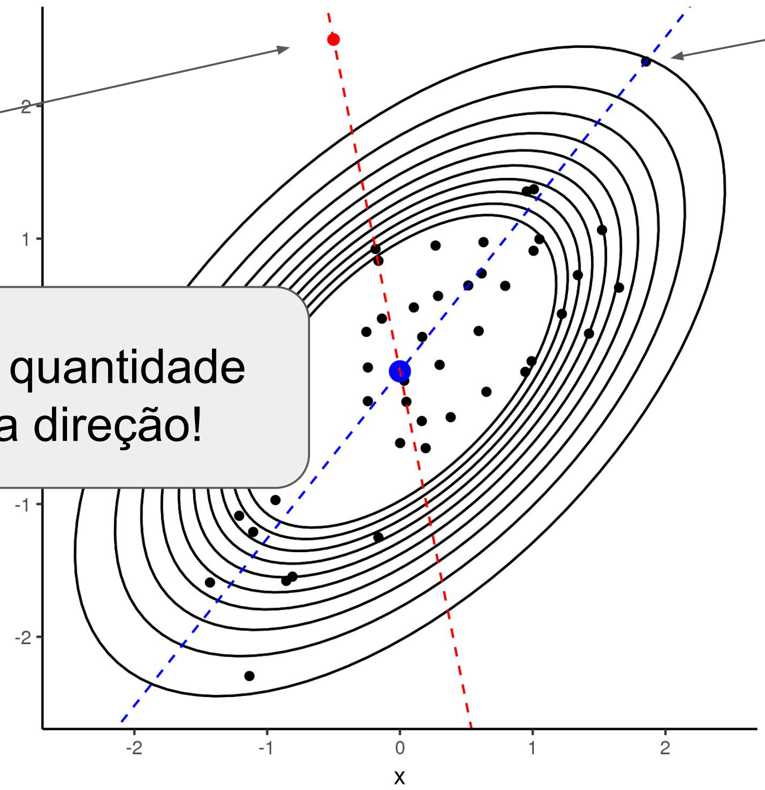
$$\sqrt{(\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu)}$$

$$e^{-\frac{1}{2}(\mathbf{x}_i-\mu)^T\Sigma^{-1}(\mathbf{x}_i-\mu)}$$

Direções em várias dimensões

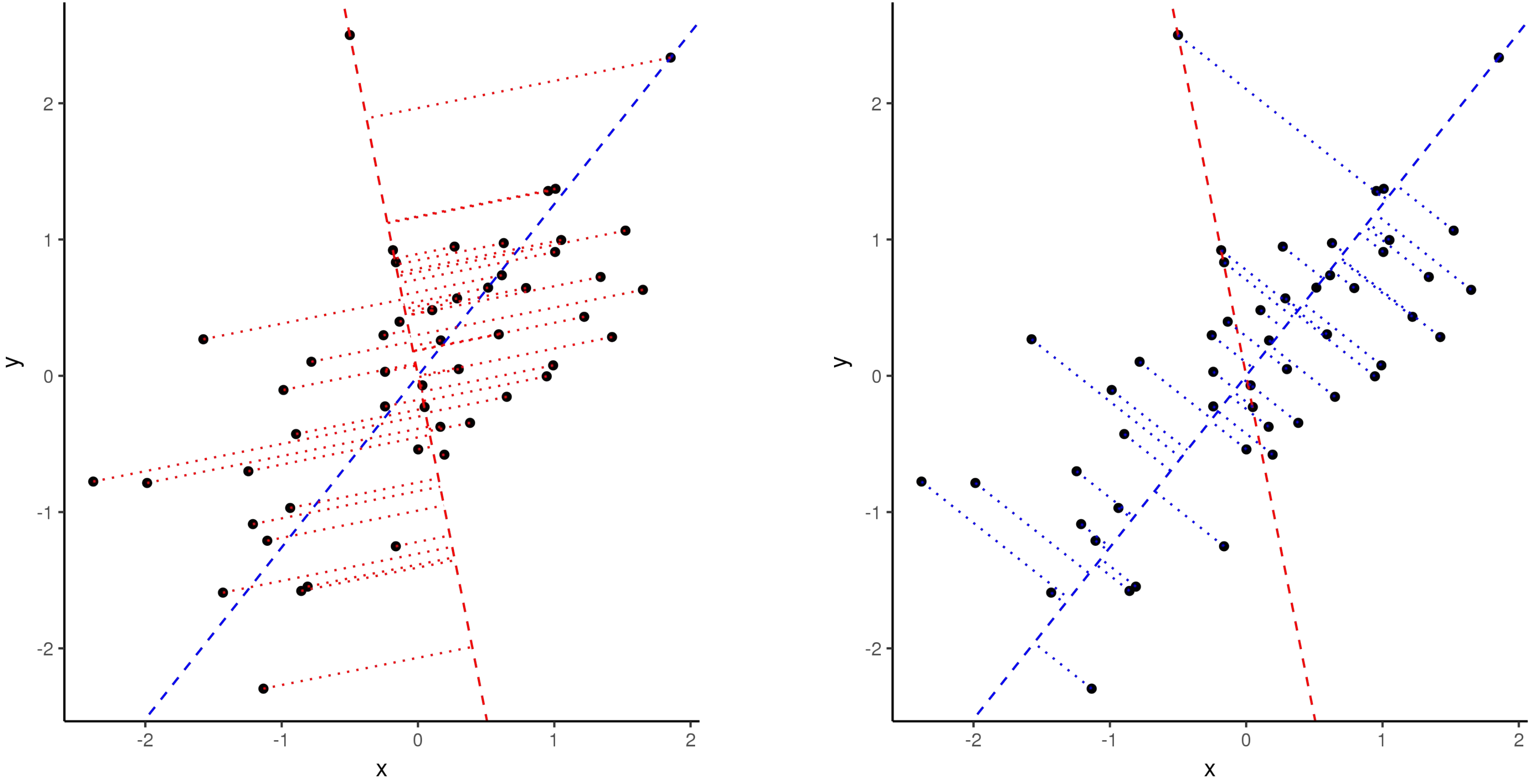
Dentro do
quantil 95%

Bem mais
longe...



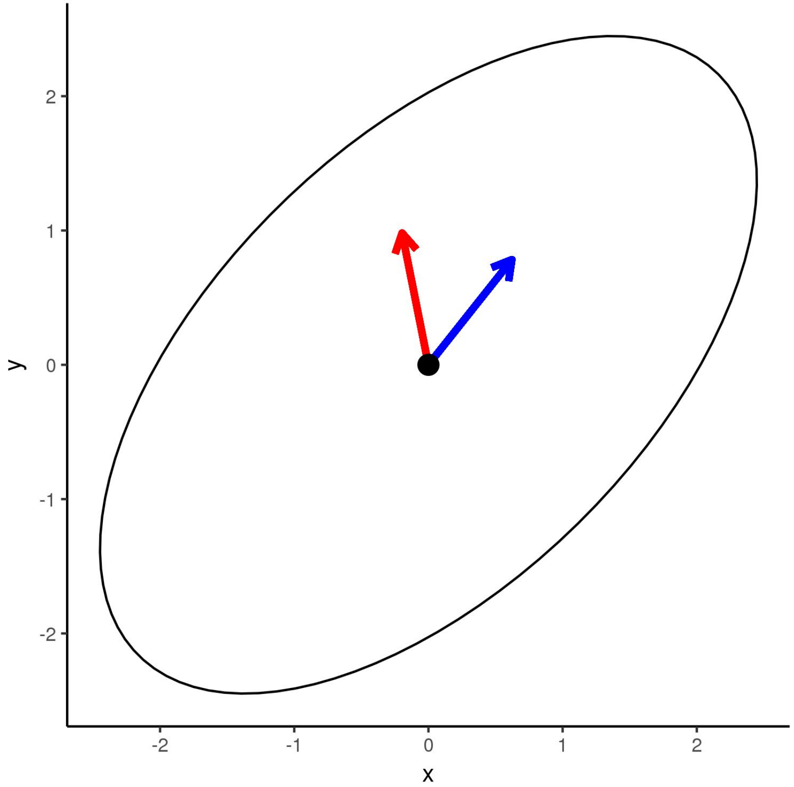
Distância depende da quantidade de variação em cada direção!

Projeções e variação direcional



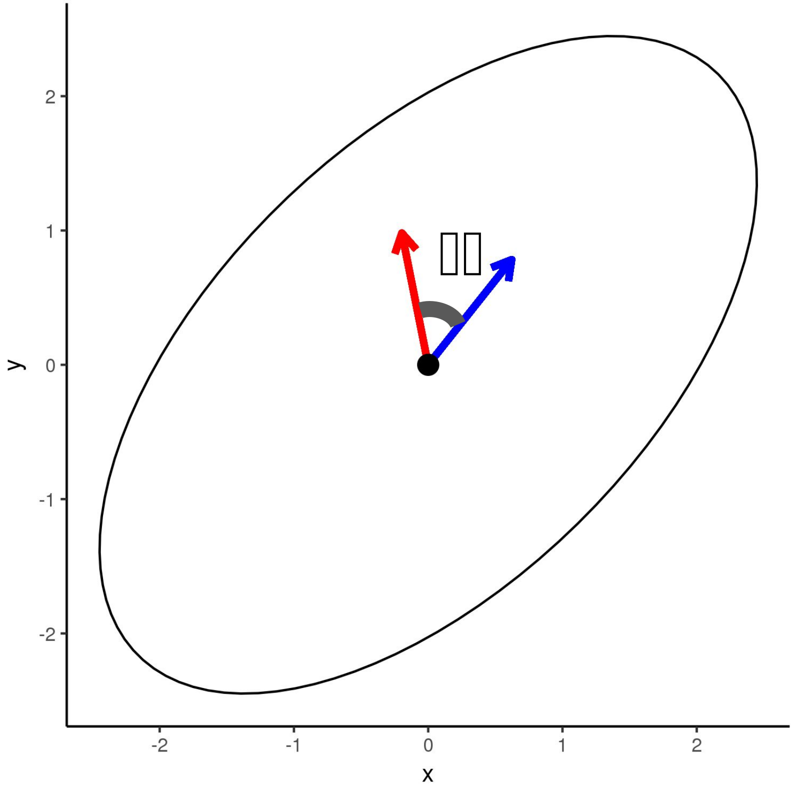
Projeções e variação direcional

$$e(\Sigma, \mathbf{x}) = \frac{\mathbf{x}^T \Sigma \mathbf{x}}{\mathbf{x}^T \mathbf{x}}$$



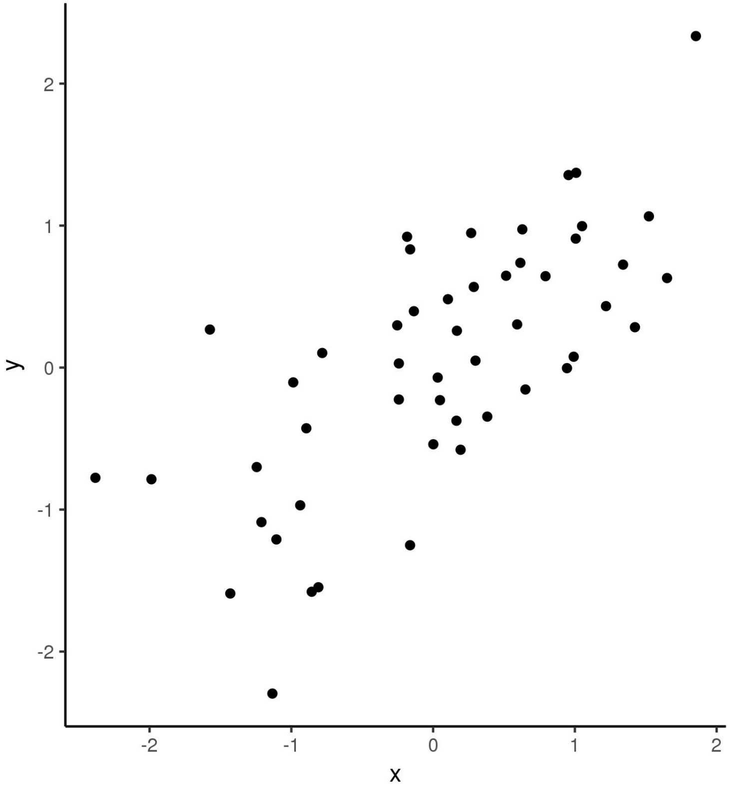
Comparando direções

$$cos(\mathbf{x}, \mathbf{y}) = \frac{\sum_{i=1}^{p} x_i y_i}{\sqrt{\sum_{i=1}^{p} x_i^2 \sum_{i=1}^{p} y_i^2}}$$



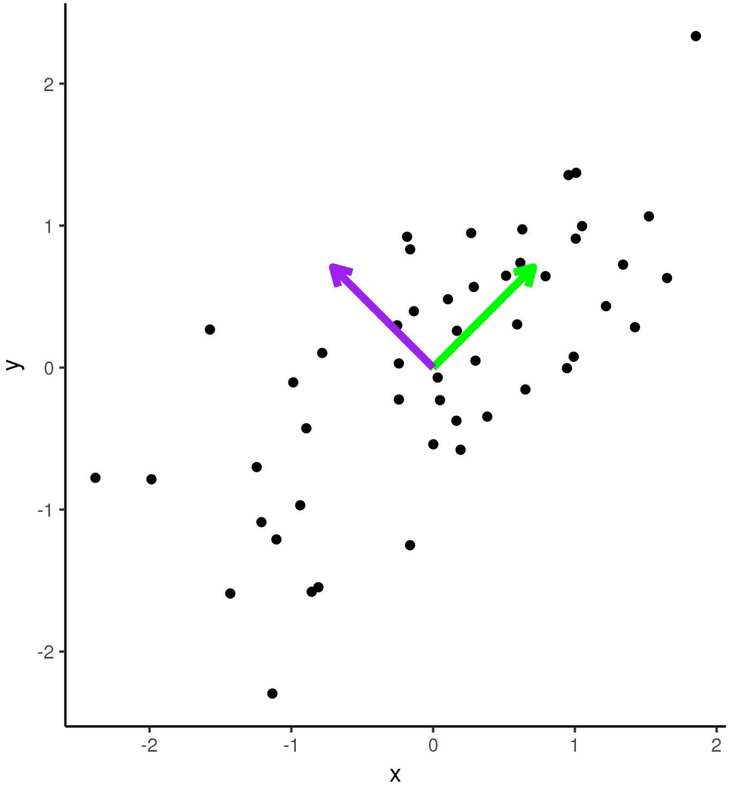
Componentes Principais

- Identificar direções de variação independente



Componentes Principais

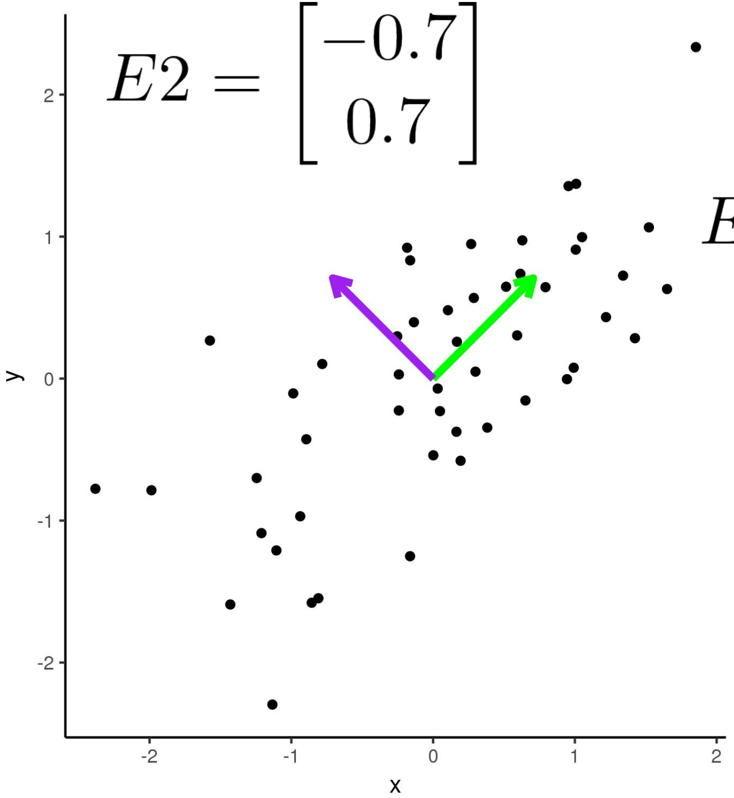
- Identificar direções de variação independente



Componentes Principais

- Identificar direções de variação independente

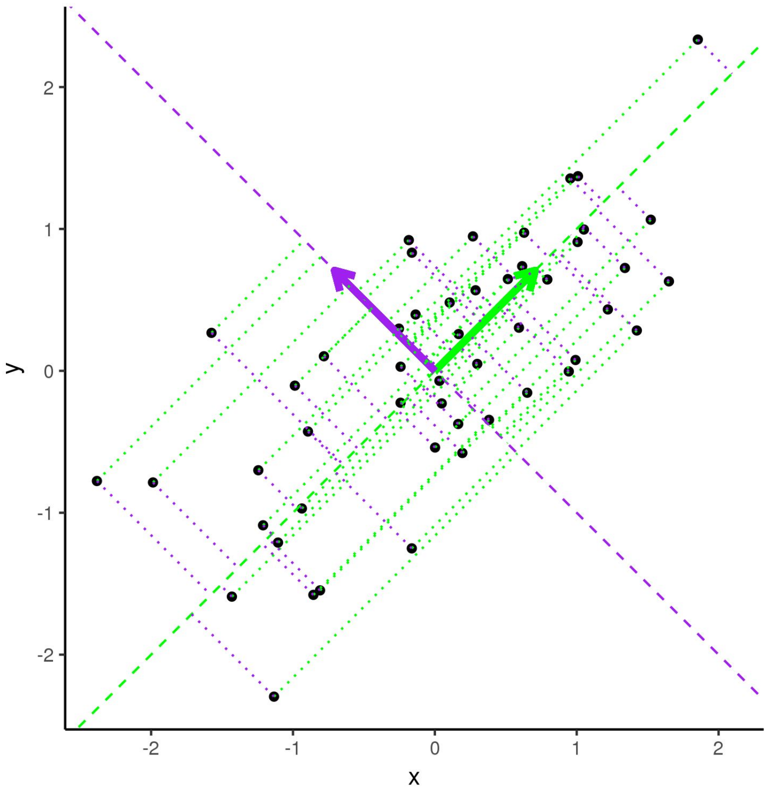
- Direções, ou eixos, definidos são combinações das variáveis originais



$$E2 = \begin{bmatrix} -0.7 \\ 0.7 \end{bmatrix}$$

$$E1 = \begin{bmatrix} 0.7 \\ 0.7 \end{bmatrix}$$

Projeções nos PCs



Rotação usando os PCs como eixos