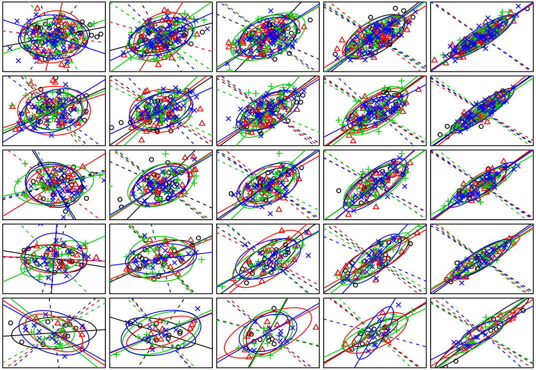
# Estatística Multivariada

## Slides de apoio às aulas



Faculdade de Ciências e Tecnologia
Universidade Nova de Lisboa
2018/19

# Aula 2

# Introdução à Estatística Multivariada

# Notação

1. Constantes/valores observados: Letras minúsculas a itálico

   *Exemplos*: $a$, $c$, $\ell$, $x$, $y$, $z$

2. Variáveis: Letras minúsculas

   *Exemplos*: x, y, z

3. Parâmetros (população): Letras gregas minúsculas

   *Exemplos*: $\lambda$, $\mu$, $\rho$, $\sigma$

4. Vetores (de números, variáveis ou parâmetros): Letras minúsculas a negrito

   *Exemplos*:
   $\boldsymbol{a}$ ($\boldsymbol{a}'$=($a_1$,...,$a_n$)), $\boldsymbol{c}$ ($\boldsymbol{c}'$=($c_1$,...,$c_n$)), $\boldsymbol{\ell}$ ($\boldsymbol{\ell}' = (\ell_1, ..., \ell_n)$)
   $\boldsymbol{x}$ ($\boldsymbol{x}'$=($x_1$,...,$x_n$)), $\boldsymbol{y}$ ($\boldsymbol{y}'$=($y_1$,...,$y_n$)), $\boldsymbol{z}$ ($\boldsymbol{z}'$=($z_1$,...,$z_n$))
   $\mathbf{x}$ ($\mathbf{x}' = (\mathsf{x}_1, ..., \mathsf{x}_n)$), $\mathbf{y}$ ($\mathbf{y}' = (\mathsf{y}_1, ..., \mathsf{y}_n)$), $\mathbf{z}$ ($\mathbf{z}' = (\mathsf{z}_1, ..., \mathsf{z}_n)$)
   $\boldsymbol{\lambda}$ ($\boldsymbol{\lambda}' = (\lambda_1, ..., \lambda_n)$), $\boldsymbol{\mu}$ ($\boldsymbol{\mu}' = (\mu_1, ..., \mu_n)$)
   Caso particular: $\mathbf{0}$ ($\mathbf{0}' = (0, ..., 0)$) e $\mathbf{1}$ ($\mathbf{1}' = (1, ..., 1)$)

1. Matrizes (de números, variáveis ou parâmetros): Letras maiúsculas a negrito
   *Exemplos*: $\boldsymbol{A}$, $\boldsymbol{S}$, $\boldsymbol{R}$, $\boldsymbol{X}$ (matriz observada), $\mathbf{X}$ (matriz aleatória), $\boldsymbol{\Sigma}$

$$\boldsymbol{A} = \begin{bmatrix} a_{11} & a_{12} & a_{13} & \dots & a_{1p} \\ a_{21} & a_{22} & a_{23} & \dots & a_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & a_{n3} & \dots & a_{np} \end{bmatrix} \qquad \boldsymbol{S} = \begin{bmatrix} s_{11} & s_{12} & s_{13} & \dots & s_{1p} \\ s_{21} & s_{22} & s_{23} & \dots & s_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ s_{n1} & s_{n2} & s_{n3} & \dots & s_{np} \end{bmatrix}$$

$$\boldsymbol{X} = \begin{bmatrix} x_{11} & x_{12} & x_{13} & \dots & x_{1p} \\ x_{21} & x_{22} & x_{23} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & x_{n3} & \dots & x_{np} \end{bmatrix} \qquad \mathbf{X} = \begin{bmatrix} \mathsf{x}_{11} & \mathsf{x}_{12} & \mathsf{x}_{13} & \dots & \mathsf{x}_{1p} \\ \mathsf{x}_{21} & \mathsf{x}_{22} & \mathsf{x}_{23} & \dots & \mathsf{x}_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathsf{x}_{n1} & \mathsf{x}_{n2} & \mathsf{x}_{n3} & \dots & \mathsf{x}_{np} \end{bmatrix}$$

# O que são dados multivariados?

- Dados multivariados são observações relativas a um conjunto de $p$ variáveis aleatórias $x_1, \dots, x_p$, i.e, resultam da realização de um vetor aleatório ($\mathbf{x}' = (x_1, \dots, x_p)$) sobre $n$ unidades estatísticas (*amostra*):

| | variável 1 | $\cdots$ | variável $j$ | $\cdots$ | variável $p$ |
|---|---|---|---|---|---|
| elemento 1 | $x_{11}$ | $\cdots$ | $x_{1j}$ | $\cdots$ | $x_{1p}$ |
| $\vdots$ | $\vdots$ | $\ddots$ | $\vdots$ | $\ddots$ | $\vdots$ |
| elemento $i$ | $x_{i1}$ | $\cdots$ | $x_{ij}$ | $\cdots$ | $x_{ip}$ |
| $\vdots$ | $\vdots$ | $\ddots$ | $\vdots$ | $\ddots$ | $\vdots$ |
| elemento $n$ | $x_{n1}$ | $\cdots$ | $x_{nj}$ | $\cdots$ | $x_{np}$ |

onde $x_{ij}$ representa o valor concreto/observado para o elemento $i$ ($i = 1 \dots n$) da $j$-ésima v.a.

# Representação matricial

- Esta base de dados pode representar-se pela matriz $\boldsymbol{X}$ contendo os $n$ valores observados das $p$ variáveis $(1 \leq n \leq p)$:

$$\boldsymbol{X} = \begin{bmatrix} x_{11} & x_{12} & x_{13} & \dots & x_{1p} \\ x_{21} & x_{22} & x_{23} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & x_{n3} & \dots & x_{np} \end{bmatrix}$$

onde $x_{ij}$ representa a $i$-ésima **observação** da $j$-ésima variável.

- Ou seja $\boldsymbol{X}$ não é mais do que uma realização da matriz aleatória $\mathbf{X}$ que representa uma população multivariada ($p$-variada)

$$\mathbf{X} = \begin{bmatrix} \mathsf{x}_{11} & \mathsf{x}_{12} & \mathsf{x}_{13} & \dots & \mathsf{x}_{1p} \\ \mathsf{x}_{21} & \mathsf{x}_{22} & \mathsf{x}_{23} & \dots & \mathsf{x}_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathsf{x}_{n1} & \mathsf{x}_{n2} & \mathsf{x}_{n3} & \dots & \mathsf{x}_{np} \end{bmatrix}$$

- Representamos $p$ variáveis aleatórias através do **vetor aleatório** $\mathbf{x}$ de dimensão $p$

$$\mathbf{x}' = (x_1, ..., x_p)$$

onde $x_j$ ($j = 1, .., p$) denota a $j$-ésima variável aleatória

- $(\mathbf{x}_1, ..., \mathbf{x}_n)$ (sendo $\mathbf{x}_i$, $i = 1, ..., n$, a $j$-ésima realização do vetor aleatório) representa a **amostra aleatória** $p$-variada de dimensão $n$ ($n$ realizações do vetor $\mathbf{x}$ de dimensão $p$ independentes e identicamente distribuídos)

- Em resumo:

| | Univariada | Multivariada |
|---|---|---|
| variável/vetor aleatório | $x$ | $\mathbf{x}$ |
| valor/vetor observado | $x$ | $\boldsymbol{x}$ |
| amostra aleatória | $(x_1, ..., x_n)$ ($n$ variáveis aleatórias $x$) | $(\mathbf{x}_1, ..., \mathbf{x}_n)$ (n vetores aleatorios $\mathbf{x}$) |
| amostra observada | $(x_1, ..., x_n)$ ou $x_i$ ($i = 1, ..., n$) ($n$ valores observados $x$) | $(\boldsymbol{x}_1, ..., \boldsymbol{x}_n)$ ou $\boldsymbol{x}_i$ ($i = 1, ..., n$) ($n$ vetores observados $\boldsymbol{x}$) |
| matriz aleatória | $-$ | $\mathbf{X}$ |
| matriz observada | $-$ | $\boldsymbol{X}$ |

- Considere-se o vetor aleatório $\mathbf{x}_{p\times 1}$. Cada elemento do vetor tem valor médio $\mu_j = E(x_j)$ e variância $\sigma_j^2 = E(x_j - \mu_j)^2$ $(j = 1, ..., p)$
- O valor médio e variância do vetor $\mathbf{x}$ podem representar-se matricialmente:

1. Valor médio $\boldsymbol{\mu} = E(\mathbf{x})$

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{bmatrix}$$

onde $\mu_j$, $(j = 1, ..., p)$ representa o valor médio da variável $j$;

2. As $p$ variâncias e as $p(p-1)/2$ covariâncias podem organizar-se numa matriz de dimensão $p \times p$ simétrica, designada por matriz de covariâncias $\boldsymbol{\Sigma} = E(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})'$

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} & \dots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} & \dots & \sigma_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \sigma_{p3} & \dots & \sigma_{pp} \end{bmatrix}$$

onde $\sigma_{jk}$, $(j, k = 1, ..., p)$ representa a covariância entre as variáveis $j$ e $k$ e $\sigma_{jj} = \sigma_j^2$ a variância da variável $j$

- $\boldsymbol{\mu}$ e $\boldsymbol{\Sigma}$ são designados respetivamente por *média populacional* (vetor) e *variância-covariância populacional* (matriz)

- Chama-se matriz desvio-padrão populacional à matriz

$$\mathbf{V}^{1/2} = \begin{bmatrix} \sqrt{\sigma_{11}} & 0 & 0 & \dots & 0 \\ 0 & \sqrt{\sigma_{22}} & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & \sqrt{\sigma_{pp}} \end{bmatrix}$$

- Frequentemente é útil e necessário descrever o grau de associação (linear) entre as $p$ variáveis, sendo este representado pela matriz (simétrica) de correlações (bivariadas) $\rho_{jk} = \sigma_{jk}/(\sqrt{\sigma_{jj}}\sqrt{\sigma_{kk}})$

$$\boldsymbol{\rho} = \begin{bmatrix} 1 & \rho_{12} & \rho_{13} & \dots & \rho_{1p} \\ \rho_{21} & 1 & \rho_{23} & \dots & \rho_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho_{p1} & \rho_{p2} & \rho_{p3} & \dots & 1 \end{bmatrix}$$

  onde $\rho_{jk},\ (j, k = 1, ..., p)$ a correlação entre as variáveis $j$ e $k$.
- A matriz $\boldsymbol{\Sigma}$ por ser obtida através das matrizes $\mathbf{V}^{1/2}$ e $\boldsymbol{\rho}$

$$\boldsymbol{\Sigma} = \mathbf{V}^{1/2}\boldsymbol{\rho}\mathbf{V}^{1/2}$$

  e

$$\boldsymbol{\rho} = (\mathbf{V}^{1/2})^{-1}\boldsymbol{\Sigma}(\mathbf{V}^{1/2})^{-1}$$

Considerando a matriz de covariâncias $\mathbf{\Sigma}$:

$$\mathbf{\Sigma} = \begin{bmatrix} 4 & 1 & 2 \\ 1 & 9 & -3 \\ 2 & -3 & 25 \end{bmatrix}$$

Determine a matriz de correlações.

$$\mathbf{V}^{1/2} = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 5 \end{bmatrix} \qquad \mathbf{V}^{-1/2} = \begin{bmatrix} 1/2 & 0 & 0 \\ 0 & 1/3 & 0 \\ 0 & 0 & 1/5 \end{bmatrix}$$

$$\boldsymbol{\rho} = \begin{bmatrix} 1 & 1/6 & 1/5 \\ 1/6 & 1 & -1/5 \\ 1/5 & -1/5 & 1 \end{bmatrix}$$

# Subconjuntos de variáveis

- Frequentemente, as características em estudo (variáveis) são organizadas em 2 ou mais grupos $\Rightarrow$ subconjuntos de variáveis
- Considerando o vector aleatório $\mathbf{x}' = (x_1, x_2, \ldots, x_p)$ e a partição com dois conjuntos com $q$ e $p-q$ variáveis, tem-se

$$\mathbf{x} = \left[\begin{array}{c} x_1 \\ \vdots \\ x_q \\ \hdashline x_{q+1} \\ \vdots \\ x_p \end{array}\right] = \left[\begin{array}{c} \mathbf{x}^{(1)} \\ \hdashline \mathbf{x}^{(2)} \end{array}\right] \qquad \boldsymbol{\mu} = \left[\begin{array}{c} \mu_1 \\ \vdots \\ \mu_q \\ \hdashline \mu_{q+1} \\ \vdots \\ \mu_p \end{array}\right] = \left[\begin{array}{c} \boldsymbol{\mu}^{(1)} \\ \hdashline \boldsymbol{\mu}^{(2)} \end{array}\right]$$

- $\boldsymbol{\Sigma}$ representa-se por

$$\boldsymbol{\Sigma} = \left[\begin{array}{ccc:ccc} \sigma_{11} & \cdots & \sigma_{1q} & \sigma_{1,q+1} & \cdots & \sigma_{1p} \\ \vdots & & \vdots & \vdots & & \vdots \\ \sigma_{q1} & \cdots & \sigma_{qq} & \sigma_{q,q+1} & \cdots & \sigma_{qp} \\ \hdashline \sigma_{q+1,1} & \cdots & \sigma_{q+1,q} & \sigma_{q+1,q+1} & \cdots & \sigma_{q+1,p} \\ \vdots & & \vdots & \vdots & & \vdots \\ \sigma_{p1} & \cdots & \sigma_{pq} & \sigma_{p,q+1} & \cdots & \sigma_{pp} \end{array}\right] = \left[\begin{array}{c:c} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \hdashline \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{array}\right]$$

- $\mathbf{\Sigma_{12}} = \mathbf{\Sigma_{21}}'$
- A matriz de covariância de $\mathbf{x}^{(1)}$ é a matriz $\mathbf{\Sigma_{11}}$ de ordem $q$
- A matriz de covariância de $\mathbf{x}^{(2)}$ é a matriz $\mathbf{\Sigma_{22}}$ de ordem $p - q$
- $\mathbf{\Sigma_{12}}$ não tem que ser simétrica ou quadrada.
- $\mathbf{\Sigma_{12}}$ contém a covariância de cada variável de $\mathbf{x}^{(1)}$ com cada variável de $\mathbf{x}^{(2)}$.
- É também usual utilizar a notação $Cov(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}) = \mathbf{\Sigma_{12}}$

# Valor médio e covariância de combinações lineares

- Seja $\mathbf{x}' = (x_1, x_2, \ldots, x_p)$ o vector aleatório com valor médio $\boldsymbol{\mu}$ e matriz de covariâncias $\boldsymbol{\Sigma}$, então a combinação linear $\boldsymbol{c}'\mathbf{x} = c_1 x_1 + \ldots + c_p x_p$ tem valor médio e variância dados respetivamente por
$$E(\boldsymbol{c}'\mathbf{x}) = \boldsymbol{c}'\boldsymbol{\mu} \text{ e } Cov(\boldsymbol{c}'\mathbf{x}) = \boldsymbol{c}'\boldsymbol{\Sigma}\boldsymbol{c}$$
- Genericamente, considerando $q$ combinações lineares de $p$ variáveis aleatórias $x_1, \ldots, x_p$

$$\begin{aligned} y_1 &= c_{11}x_1 + \cdots + c_{1p}x_p \\ y_2 &= c_{21}x_1 + \cdots + c_{2p}x_p \\ &\vdots \\ y_q &= c_{q1}x_1 + \cdots + c_{qp}x_p \end{aligned}$$

ou seja,

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_q \end{bmatrix} = \begin{bmatrix} c_{11} & c_{12} & \ldots & c_{1p} \\ c_{21} & c_{22} & \ldots & c_{2p} \\ \vdots & \vdots & & \vdots \\ c_{q1} & c_{q2} & \ldots & c_{qp} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{bmatrix} = \mathbf{Cx}$$

então

$$\boldsymbol{\mu}_{\mathbf{y}} = \mathbf{C}\boldsymbol{\mu}_{\mathbf{x}}$$
$$\boldsymbol{\Sigma}_{\mathbf{y}} = \mathbf{C}\boldsymbol{\Sigma}_{\mathbf{x}}\mathbf{C}'$$

# Exemplo 2

Seja $\mathbf{x}' = (x_1, x_2, x_3)$ o vector aleatório com valor médio $\boldsymbol{\mu}' = (2, 1, 2)$ e matriz de covariâncias $\boldsymbol{\Sigma}$:

$$\boldsymbol{\Sigma} = \begin{bmatrix} 4 & 1 & 2 \\ 1 & 9 & -3 \\ 2 & -3 & 25 \end{bmatrix}$$

Determine o valor médio e matriz de covariância de $\mathbf{y}' = (y_1, y_2, y_3)$

$$y_1 = x_1 - x_2 + x_3$$

$$y_2 = x_1 + x_2 - 2x_3$$

$$y_3 = x_1 + 2x_2 + 2x_3$$

$$\boldsymbol{\mu}_{\mathbf{y}} = \mathbf{C}\boldsymbol{\mu}_{\mathbf{x}} = \begin{bmatrix} 1 & -1 & 1 \\ 1 & 1 & -2 \\ 1 & 2 & 2 \end{bmatrix} \begin{bmatrix} 2 \\ 1 \\ 2 \end{bmatrix} = \begin{bmatrix} 3 \\ -1 \\ 8 \end{bmatrix}$$

$$\boldsymbol{\Sigma}_{\mathbf{y}} = \mathbf{C}\boldsymbol{\Sigma}_{\mathbf{x}}\mathbf{C}' = \begin{bmatrix} 1 & -1 & 1 \\ 1 & 1 & -2 \\ 1 & 2 & 2 \end{bmatrix} \begin{bmatrix} 4 & 1 & 2 \\ 1 & 9 & -3 \\ 2 & -3 & 25 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 \\ -1 & 1 & 2 \\ 1 & -2 & 2 \end{bmatrix} = \begin{bmatrix} 46 & -66 & 43 \\ -66 & 119 & -69 \\ 43 & -69 & 128 \end{bmatrix}$$

- Considerando agora uma amostra aleatória de uma população $p$-variada $(\mathbf{x}_1, ..., \mathbf{x}_n)$ ($n$ realizações independentes do vetor $\mathbf{x}$) definem-se as seguintes estatísticas

$$\bar{\mathbf{x}} = \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_p \end{bmatrix} \qquad \mathbf{S} = \begin{bmatrix} s_{11} & s_{12} & s_{13} & \dots & s_{1p} \\ s_{21} & s_{22} & s_{23} & \dots & s_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ s_{p1} & s_{p2} & s_{p3} & \dots & s_{pp} \end{bmatrix} \qquad \mathbf{R} = \begin{bmatrix} 1 & r_{12} & r_{13} & \dots & r_{1p} \\ r_{21} & 1 & r_{23} & \dots & r_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & r_{p3} & \dots & 1 \end{bmatrix}$$

onde

- $\bar{\mathbf{x}}$ representa o vetor de médias amostrais, sendo $\bar{x}_j, \; (j = 1, ..., p)$ a média relativa à variável $j$;
- $\mathbf{S}$ representa a matriz de variâncias-covariâncias amostrais, sendo $s_{jk}, \; (j, k = 1, ..., p)$ a covariância entre as variáveis $j$ e $k$ e $s_{jj}$ a variância da variável $j$;
- $\mathbf{R}$ representa a matriz de correlações (simétrica), sendo $r_{jk}, \; (j, k = 1, ..., p)$ a correlação entre as variáveis $j$ e $k$.

# Estatísticas amostrais

- Recorde que

$$\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ji}\ (j = 1, ..., p)$$

$$s_{jk} = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ji} - \bar{x}_j)(x_{ki} - \bar{x}_k) = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_{ji}x_{ki} - n\bar{x}_j\bar{x}_k\right)\ (j, k = 1, ..., p;\ j \neq k)$$

$$s_j^2 = s_{jj} = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ji} - \bar{x}_j)^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_{ji}^2 - n\bar{x}_j^2\right)\ (j = 1, ..., p)$$

$$r_{jk} = \frac{s_{jk}}{\sqrt{s_{jj}}\sqrt{s_{kk}}}\ (j, k = 1, ..., p)$$

- Note que:
  - $r_{jj} = 1$ e $r_{jk} = r_{kj}$
  - $-1 \leq r \leq +1$
  - $|r|$ mede o grau de associação linear; o sinal informa sobre a direção da associação.

# Exemplo 3

Considerando a matriz de observações $\mathbf{X}_{4\times 2}$:

$$\mathbf{X} = \begin{bmatrix} 42 & 4 \\ 52 & 5 \\ 48 & 4 \\ 58 & 3 \end{bmatrix}$$

$\bar{x}_1 = \frac{1}{4}\sum_{i=1}^{4} x_{i1} = 50$ e $\bar{x}_2 = \frac{1}{4}\sum_{i=1}^{4} x_{i2} = 4$
logo

$$\bar{\mathbf{x}} = \begin{bmatrix} 50 \\ 4 \end{bmatrix}$$

$s_{11} = \frac{1}{3}\sum_{i=1}^{4}(x_{i1} - \bar{x}_1)^2 = 45.333$
$s_{22} = \frac{1}{3}\sum_{i=1}^{4}(x_{i2} - \bar{x}_2)^2 = 0.667$
$s_{12} = \frac{1}{3}\sum_{i=1}^{4}(x_{i1} - \bar{x}_1)(x_{i2} - \bar{x}_2) = -2$
logo

$$\mathbf{S} = \begin{bmatrix} 45.333 & -2 \\ -2 & 0.667 \end{bmatrix}$$

e

$$\mathbf{R} = \begin{bmatrix} 1 & \frac{-2}{\sqrt{45.333 \times 0.677}} = -0.364 \\ \frac{-2}{\sqrt{45.333 \times 0.677}} = -0.364 & 1 \end{bmatrix}$$

Código R:

```
> x<-c(42,52,48,58)
> y<-c(4,5,4,3)
> X<-matrix(c(x,y),4,2)
> X

     [,1] [,2]
[1,]   42    4
[2,]   52    5
[3,]   48    4
[4,]   58    3

> colMeans(X)

[1] 50  4

> var(X)

         [,1]       [,2]
[1,] 45.33333 -2.0000000
[2,] -2.00000  0.6666667

> cor(X)

           [,1]       [,2]
[1,]  1.0000000 -0.3638034
[2,] -0.3638034  1.0000000
```

Usando calculo matricial:

- $\bar{\mathbf{x}} = \frac{1}{n}\mathbf{X}'\mathbf{1}_{n\times 1}$, onde $\mathbf{1}_{n\times 1}$ é o vetor de 1's de dimensão $n$
- $\mathbf{S} = \frac{1}{n-1}\mathbf{X}'\left(\mathbf{I} - \frac{1}{n}\mathbf{1}_{n\times n}\right)\mathbf{X}$, onde $\mathbf{1}_{n\times n}$ é a matriz de 1's de dimensão $n \times n$
- $\mathbf{R} = \mathbf{D}^{-1/2}\mathbf{S}\mathbf{D}^{-1/2}$ , onde $\mathbf{D}^{1/2} = Diag(s_1, ..., s_n)$

Código R:

```
> t(X)%*%c(1,1,1,1)/4

     [,1]
[1,]   50
[2,]    4

> J<-matrix(1,4,4)
> I<-matrix(c(1,0,0,0,0,1,0,0,0,0,1,0,0,0,0,1),4,4)
> S<-(t(X)%*%(I-J/4)%*%X)/3
> S

         [,1]       [,2]
[1,] 45.33333 -2.0000000
[2,] -2.00000  0.6666667

> D<-matrix(c(sqrt(diag(S)[1]),0,0,sqrt(diag(S)[2])),2,2)
> solve(D)%*%S%*%solve(D)

           [,1]       [,2]
[1,]  1.0000000 -0.3638034
[2,] -0.3638034  1.0000000
```

- No caso univariado, a variância quantifica o grau de dispersão em torno da média. No caso multivariado, é por vezes conveniente resumir a matriz $\mathbf{S}$ num único valor

- Em regra, são usadas duas medidas:

1. *Variância amostral generalizada*, $|\mathbf{S}| = \prod_{j=1}^{p} \ell_j$ onde $\ell_i$ representa os valores próprios de $\mathbf{S}$.

   Notas importantes:

   - É possível mostrar que o "volume" de dados centrado em $\bar{\mathbf{x}}$ necessário para incluir uma certa proporção de dados, definido por um *elipsoide*, é proporcional a $|\mathbf{S}|^{1/2}$

   - $|\mathbf{S}|$ não deteta diferentes estruturas de correlação (ver exemplo)

   - $|\mathbf{S}| = 0$ quando pelo menos uma das colunas da matriz dos desvios pode expressar-se como combinação linear das outras colunas $\Rightarrow$ as colunas da matriz dos desvios, $\mathbf{X}_{n\times p} - \mathbf{1}_{n\times 1}\bar{\mathbf{x}}'_{1\times p}$, são linearmente dependentes (indicativo de <u>redundância entre variáveis</u>)

$$\mathbf{X} - \mathbf{1}\,\bar{\mathbf{x}}' = \begin{bmatrix} \mathbf{x}'_1 - \bar{\mathbf{x}}' \\ \mathbf{x}'_2 - \bar{\mathbf{x}}' \\ \cdots \\ \mathbf{x}'_n - \bar{\mathbf{x}}' \end{bmatrix}$$

2. *Variância total*, $tr(\mathbf{S}) = \sum_{j=1}^{p} \ell_j$

Considere a matriz de observações

$$\boldsymbol{X} = \begin{bmatrix} 1 & 2 & 5 \\ 4 & 1 & 6 \\ 4 & 0 & 4 \end{bmatrix}$$

Determine a matriz dos desvios e indique se os vetores formados pelas colunas da matriz são linearmente independentes. Apresente a resposta também baseada nos valores próprios.

$$\mathbf{X} - \mathbf{1}\,\bar{\mathbf{x}}' = \begin{bmatrix} 1 & 2 & 5 \\ 4 & 1 & 6 \\ 4 & 0 & 4 \end{bmatrix} - \begin{bmatrix} 3 & 1 & 5 \\ 3 & 1 & 5 \\ 3 & 1 & 5 \end{bmatrix} = \begin{bmatrix} -2 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & -1 & -1 \end{bmatrix}$$

Os vetores $(-2, 1, 1)'$, $(1, 0, -1)'$ e $(0, 1, -1)'$ são linearmente dependentes (solução geral: $c_2 = 2c_1$ e $c_3 = -c_1$).

Por outro lado, têm-se os valores próprios $\ell_1 = 8.292$, $\ell_2 = -2.292$ e $\ell_3 = 0$, logo $|\boldsymbol{S}| = 0$.

# Exemplo 5

Considere as matrizes de covariâncias respetivamente correspondentes às correlações bivariadas $r = 0.8$, $r = 0$ e $r = -0.8$

$$\boldsymbol{S}^{(1)} = \begin{bmatrix} 5 & 4 \\ 4 & 5 \end{bmatrix} \qquad \boldsymbol{S}^{(2)} = \begin{bmatrix} 3 & 0 \\ 0 & 3 \end{bmatrix} \qquad \boldsymbol{S}^{(3)} = \begin{bmatrix} 5 & -4 \\ -4 & 5 \end{bmatrix}$$

Determine a variância generalizada nos 3 casos. Interprete os resultados.

Os valores-vetores próprios de $\boldsymbol{S}^{(1)}$ são $\ell_1 = 9$, $\mathbf{e}_1' = (1/\sqrt{2}, 1/\sqrt{2})$ e $\ell_2 = 1$, $\mathbf{e}_2' = (1/\sqrt{2}, -1/\sqrt{2}) \Rightarrow |\boldsymbol{S}^{(1)}| = 9$

Os valores-vetores próprios de $\boldsymbol{S}^{(2)}$ são $\ell_1 = 3$, $\mathbf{e}_1' = (1, 0)$ e $\ell_2 = 3$, $\mathbf{e}_2' = (0, 1) \Rightarrow |\boldsymbol{S}^{(2)}| = 9$

Os valores-vetores próprios de $\boldsymbol{S}^{(3)}$ são $\ell_1 = 9$, $\mathbf{e}_1' = (1/\sqrt{2}, -1/\sqrt{2})$ e $\ell_2 = 1$, $\mathbf{e}_2' = (1/\sqrt{2}, 1/\sqrt{2}) \Rightarrow |\boldsymbol{S}^{(3)}| = 9$

- Quando as características em estudo (variáveis) estão organizadas em 2 ou mais grupos é comum considerar a partição das estatísticas amostrais (tal como vimos para a população)
- A partição em dois conjuntos com $q$ e $p-q$ variáveis, resulta no vetor média $\bar{\mathbf{x}}_{p\times 1}$

$$\bar{\mathbf{x}} = \left[\begin{array}{c} \bar{x}_1 \\ \vdots \\ \bar{x}_q \\ \hdashline \bar{x}_{q+1} \\ \vdots \\ \bar{x}_p \end{array}\right] = \left[\begin{array}{c} \bar{\mathbf{x}}^{(1)} \\ \hdashline \bar{\mathbf{x}}^{(2)} \end{array}\right]$$

- e na matriz de covariâncias

$$\mathbf{S} = \left[\begin{array}{ccc:ccc} s_{11} & \cdots & s_{1q} & s_{1,q+1} & \cdots & s_{1p} \\ \vdots & & \vdots & \vdots & & \vdots \\ s_{q1} & \cdots & s_{qq} & s_{q,q+1} & \cdots & s_{qp} \\ \hdashline s_{q+1,1} & \cdots & s_{q+1,q} & s_{q+1,q+1} & \cdots & s_{q+1,p} \\ \vdots & & \vdots & \vdots & & \vdots \\ s_{p1} & \cdots & s_{pq} & s_{p,q+1} & \cdots & s_{pp} \end{array}\right] = \left[\begin{array}{c:c} \mathbf{S}_{11} & \mathbf{S}_{12} \\ \hdashline \mathbf{S}_{21} & \mathbf{S}_{22} \end{array}\right]$$

Considerando os dados do ficheiro "data1.xlsx", contendo observações relativas a 5 variáveis divididas em 2 subconjuntos:

1. $x_1$ - intolerância à glucose; $x_2$ - níveis de insulina após toma de glucose oral; $x_3$ - resistência à insulina (colunas 1, 2 e 3)
2. $y_1$ - Peso relativo e $y_2$ - glicémia (colunas 4 e 5)

Represente o vetor média e matriz de covariâncias, considerando a referida partição.

Código R:

```
> dados<-as.data.frame(readxl::read_xlsx("./Datasets/data1.xlsx",
                                        col_names = F))
> round(colMeans(dados[,-6]),3)

   X__1    X__2    X__3    X__4    X__5
  0.977 121.986 543.614 186.117 184.207

> round(var(dados[,-6]),3)

       X__1      X__2       X__3       X__4      X__5
X__1  0.017    -0.073      0.982      3.473     5.266
X__2 -0.073  4087.097  19546.064  -3063.464  4849.906
X__3  0.982 19546.064 100457.850 -12918.163 25908.490
X__4  3.473 -3063.464 -12918.163  14625.313   101.483
X__5  5.266  4849.906  25908.490    101.483 11242.332
```

- Seja $\mathbf{x}' = (x_1, x_2, \ldots, x_p)$ um vector aleatório e $\boldsymbol{c}' = (c_1, \ldots, c_p)$ um vetor de constantes, então a combinação linear
$$\boldsymbol{c}'\mathbf{x} = c_1 x_1 + ... + c_p x_p$$
tem média $\boldsymbol{c}'\bar{\mathbf{x}}$ e variância $\boldsymbol{c}'\mathbf{S}\boldsymbol{c}$

- A correlação entre duas combinações lineares $y = \boldsymbol{a}'\mathbf{x}$ e $z = \boldsymbol{b}'\mathbf{x}$ é dada por
$$r_{yz} = \frac{\boldsymbol{a}'\mathbf{S}\boldsymbol{b}}{\sqrt{(\boldsymbol{a}'\mathbf{S}\boldsymbol{a})(\boldsymbol{b}'\mathbf{S}\boldsymbol{b})}}$$

- Genericamente, $q$ combinações lineares de $p$ variáveis aleatórias $x_1, \ldots, x_p$, $c_{i1}x_1 + ... + c_{ip}x_p$ $(i = 1, ..., q)$ ou seja,
$$\mathbf{y} = \begin{bmatrix} c_{11} & c_{12} & \ldots & c_{1p} \\ c_{21} & c_{22} & \ldots & c_{2p} \\ \vdots & \vdots & & \vdots \\ c_{q1} & c_{q2} & \ldots & c_{qp} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{bmatrix} = \mathbf{Cx}$$
têm média amostral $\bar{\mathbf{y}} = \mathbf{C}\bar{\mathbf{x}}$ e covariância amostral $\mathbf{S_y} = \mathbf{CSC}'$

- A matriz de correlações (bivariadas) entre $q$ combinações lineares é dada por $\mathbf{R_y} = \mathbf{D_y}^{-1/2}\mathbf{S_y}\mathbf{D_y}^{-1/2}$, onde $\mathbf{D}^{1/2}$ é a matriz diagonal dos $q$ desvios-padrão

# Exemplo 7

Considere as seguintes observações relativas ao vector aleatório $\mathbf{y}' = (y_1, y_2, y_3, y_4, y_5)$:

| $y_1$ | $y_2$ | $y_3$ | $y_4$ | $y_5$ |
|---|---|---|---|---|
| 51 | 36 | 50 | 35 | 42 |
| 27 | 20 | 26 | 17 | 27 |
| 37 | 22 | 41 | 37 | 30 |
| 42 | 36 | 32 | 34 | 27 |
| 27 | 18 | 33 | 14 | 29 |
| 43 | 32 | 43 | 35 | 40 |
| 41 | 22 | 36 | 25 | 38 |
| 38 | 21 | 31 | 20 | 16 |
| 36 | 23 | 27 | 25 | 28 |
| 26 | 31 | 31 | 32 | 36 |
| 29 | 20 | 25 | 26 | 25 |

a) Determine a média e matriz de covariância da variável $z = 3y_1 - 2y_2 + 4y_3 - y_4 + y_5$

b) Considere a combinação linear $w = y_1 - 3y_2 - y_3 + y_4 - 2y_5$. Determine a correlação entre $z$ e $w$.

$$\bar{\boldsymbol{y}} = \begin{bmatrix} 36.09 \\ 25.55 \\ 34.09 \\ 27.27 \\ 30.73 \end{bmatrix} \qquad \boldsymbol{S}_y = \begin{bmatrix} 65.09 & 33.65 & 47.59 & 36.77 & 25.43 \\ 33.65 & 46.07 & 28.95 & 40.34 & 28.36 \\ 47.59 & 28.95 & 60.69 & 37.37 & 41.13 \\ 36.77 & 40.34 & 37.37 & 62.82 & 31.68 \\ 25.43 & 28.36 & 41.13 & 31.68 & 58.22 \end{bmatrix}$$

logo,

$$\bar{z} = \boldsymbol{c}'\bar{\boldsymbol{y}} = \begin{bmatrix} 3 & -2 & 4 & -1 & 1 \end{bmatrix} \begin{bmatrix} 36.09 \\ 25.55 \\ 34.09 \\ 27.27 \\ 30.73 \end{bmatrix} = 197.0 \qquad s_z^2 = \boldsymbol{c}'\boldsymbol{S}\boldsymbol{c} = 2084.0\,.$$

Considerando a v.a. *w*

$$\bar{w} = \boldsymbol{b}'\bar{\boldsymbol{y}} = -108.82 \qquad s_w^2 = \boldsymbol{b}'\boldsymbol{S}\boldsymbol{b} = 745.96$$

e a covariância amostral entre *z* e *w* é $s_{zw} = -635.4$, sendo $r_{zw} = -0.51$

c) Considere as combinações lineares:

$$z_1 = y_1 + y_2 + y_3 + y_4 + y_5$$

$$z_2 = 2y_1 - 3y_2 + y_3 - 2y_4 - y_5$$

$$z_3 = -y_1 - 2y_2 + y_3 - 2y_4 + 3y_5$$

Determine o vetor média e a matriz de covariância de $\mathbf{z}' = (z_1, z_2, z_3)$.

$$\bar{\boldsymbol{z}} = \begin{bmatrix} 153.73 \\ -55.64 \\ -15.45 \end{bmatrix}$$

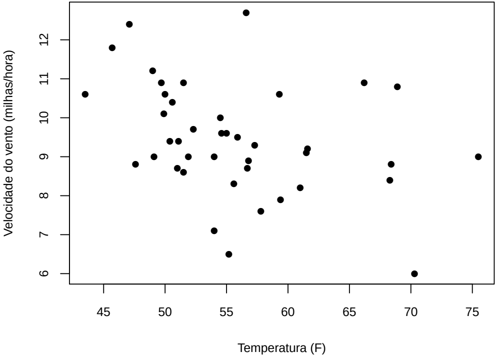
$$\boldsymbol{S}_z = \begin{bmatrix} 995.42 & -502.09 & -211.04 \\ -502.09 & 811.45 & 268.08 \\ -211.04 & 268.08 & 702.87 \end{bmatrix}$$

$$\boldsymbol{R}_z = \begin{bmatrix} 1.00 & -0.56 & -0.25 \\ -0.56 & 1.00 & 0.35 \\ -0.25 & 0.35 & 1.00 \end{bmatrix}$$

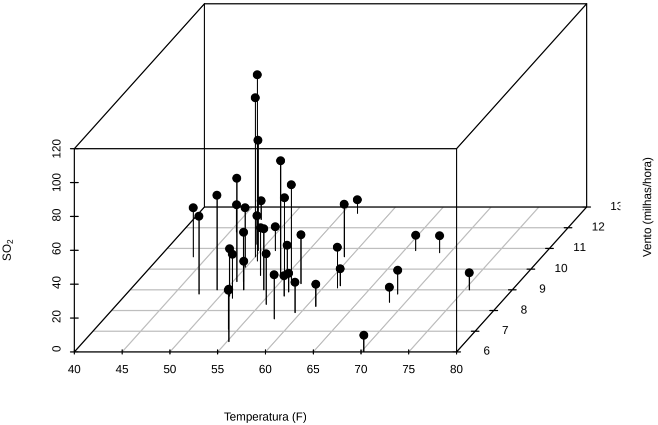
- A representação "tradicional" de $\mathbf{X}_{n\times p}$ corresponde à nuvem de $n$ pontos em $\mathbb{R}^p$:

$$p \text{ variáveis} \rightarrow p \text{ eixos}$$
$$n \text{ observações} \rightarrow n \text{ pontos}$$

- Esta representação não é visualizável para $p > 3$
- Para $p = 2$



- Para $p = 3$



SO2
0
20
40
60
80
100
120
40
45
50
55
60
65
70
75
80
Temperatura (F)
6
7
8
9
10
11
12
13
Vento (milhas/hora)

- Para $p = 2$

Código R:

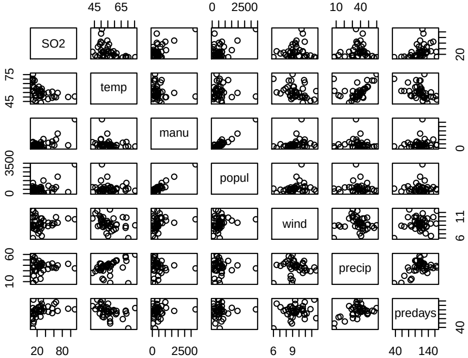
```
> library(HSAUR3)
> with(USairpollution, plot(temp,wind,xlab="Temperatura (F)",
 ylab = "Velocidade do vento (milhas/hora)",
 cex.lab=0.8,pch=16,cex.axis=0.8,main="p = 2"))
```

- Para $p = 3$

Código R:

```
> library(scatterplot3d)
> with(USairpollution, scatterplot3d(temp, wind, SO2, type = "h",angle = 55, pch = 16,
 xlab="Temperatura (F)", ylab = "Vento (milhas/hora)",
 zlab = expression(SO[2]), cex.axis = 0.6,cex.lab = 0.6))
```

- A representação mais comum é a de múltiplos diagramas de dispersão relacionando as variáveis 2 a 2:

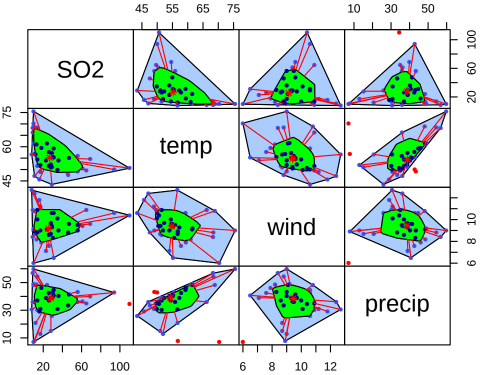



Código R:

```
> pairs(USairpollution)
```

# Múltiplos diagramas de dispersão

- Ao gráfico anterior pode adicionar-se informação melhorando a interpretabilidade:

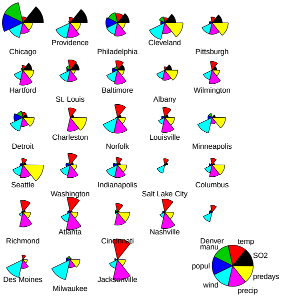



Código R:

```
> aplpack::bagplot.pairs(USairpollution[,c(1,2,5,6)], gap = 0,
        col.baghull = "green", main=" ")
```

# Estrelas

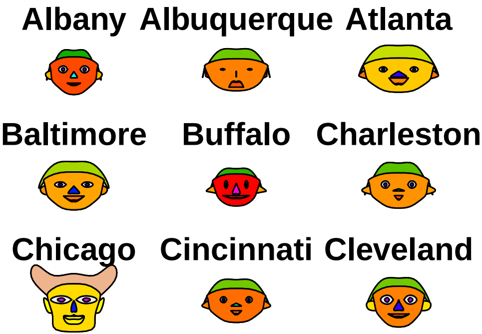
- *Estrelas* - cada variável é representada por um setor circular com raio correspondente às observações. A posição dos setores identifica a variável e o seu comprimento a magnitude da observação nessa variável.




Código R:

```
> odata <- USairpollution[order(-USairpollution$SO2), ]
> graphics::stars(odata[1:28,], len=1, cex=0.5, key.loc=c(12.5, 2),
        labels=row.names(odata[1:28,]), draw.segments=TRUE)
```

- *Faces de Chernoff* - cada observação é representada por uma face com elementos cuja forma e/ou tamanho são determinados pelos valores de variável específicas




Código R:

```
> aplpack::faces(USairpollution[1:9,],print.info = F)
```

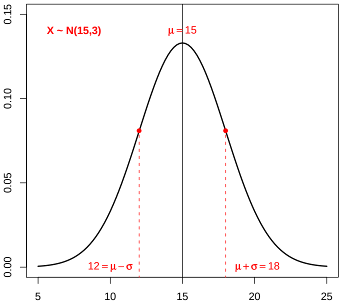
# Distribuição normal multivariada

# Distribuição normal univariada

- Uma variável aleatória contínua, $x$ diz-se ter distribuição gaussiana ou normal com valor médio $\mu$ e variância $\sigma$, i.e., $x \frown N(\mu, \sigma^2)$, se a sua *fdp* corresponde a

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \ -\infty < x < +\infty, \ -\infty < \mu < +\infty, \ 0 < \sigma < +\infty$$

- Função densidade de probabilidade:




- Curva unimodal, sendo a moda $x = \mu$ (maximizante da fdp)
- Simétrica em relação ao eixo $x = \mu$
- Pontos de inflexão em $x = \mu \pm \sigma$
- Eixo das abcissas como assimptota

- Note-se que o termo

$$\left(\frac{x-\mu}{\sigma}\right)^2 = (x-\mu)(\sigma^2)^{-1}(x-\mu) \tag{1}$$

representa o quadrado da distância entre $x$ e $\mu$ (padronizada pela variância)

- Da mesma forma, generalizando a um vetor $\boldsymbol{x}$

$$(\boldsymbol{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) \tag{2}$$

representa o quadrado da distância entre $\boldsymbol{\mu}$ e $\boldsymbol{x}$ (sendo $\boldsymbol{\Sigma}$ uma matriz definida positiva)

- A distribuição normal multivariada resulta de substituir o termo (1) pelo (2), sendo necessário substituir a constante $(2\pi)^{-1/2}(\sigma^2)^{-1/2}$ por uma mais geral válida para $p$ variáveis $(2\pi)^{-p/2}|\boldsymbol{\Sigma}|^{-1/2}$

- Assim, **x** diz-se ter distribuição normal multivariada se a sua *fdp* corresponde a

$$f_{\mathbf{x}}(\boldsymbol{x}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}}\, e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}$$

i.e., **x** tem distribuição $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ onde $p$ representa o número de variáveis

- Para $p = 2$ tem-se a distribuição normal bivariada com parâmetros $\mu_1$, $\mu_2$,

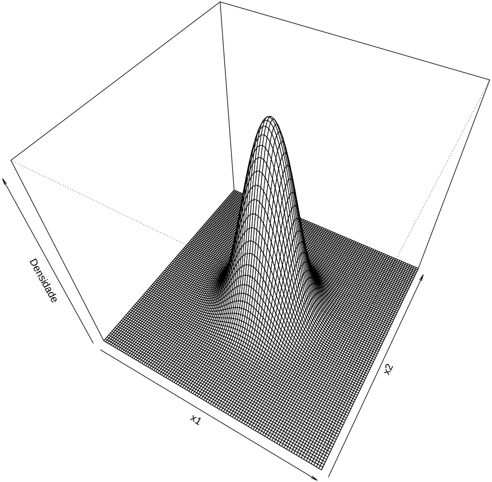
$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix}$$

e $\rho_{12} = \rho_{21} = \sigma_{12}/\sqrt{\sigma_{11}\sigma_{22}}$, $(\sigma_{11} = var(x_1),\ \sigma_{22} = var(x_2))$ , com fdp

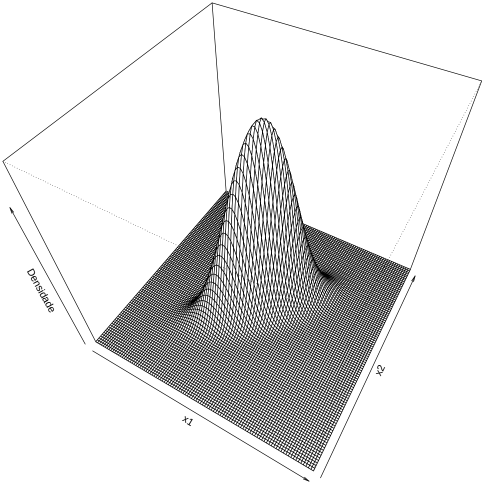
$$f_{x_1,x_2}(\boldsymbol{x}_1, \boldsymbol{x}_2) = \frac{1}{2\pi\sqrt{\sigma_{11}\sigma_{22}(1-\rho_{12}^2)}} \times$$
$$\exp\left[-\frac{1}{2(1-\rho_{12}^2)}\left[\left(\frac{\boldsymbol{x}_1-\mu_1}{\sqrt{\sigma_{11}}}\right)^2 + \left(\frac{\boldsymbol{x}_2-\mu_2}{\sqrt{\sigma_{22}}}\right)^2 - 2\rho_{12}\left(\frac{x_1-\mu_1}{\sqrt{\sigma_{11}}}\right)\left(\frac{x_2-\mu_2}{\sqrt{\sigma_{22}}}\right)\right]\right]$$

- No caso particular de $\rho_{12} = 0$, então $f(x_1, x_2) = f(x_1)f(x_2)$ e $x_1$ e $x_2$ dizem-se independentes

# Distribuição normal multivariada ($\rho = 0$)



# Distribuição normal multivariada ($\rho \neq 0$)



- Das figuras anteriores fica claro que densidades constantes correspondem a cortes transversais da normal bivariada ($p = 2$) desenhando **elipses de densidade constante**
- Essas elipses de densidade constante ($p = 2$) correspondem a distâncias constantes tal que

$$(\boldsymbol{x} - \mu)'(\boldsymbol{\Sigma})^{-1}(\boldsymbol{x} - \mu) = c^2$$

  sendo $c$ a distância conhecida por **Distância de Mahalanobis**
- Genericamente (i.e., para qualquer $p$)

$$\begin{aligned} \text{Densidade constantes} &= \{\text{valores de } \mathbf{x} \text{ tal que } (\boldsymbol{x} - \boldsymbol{\mu})'(\boldsymbol{\Sigma})^{-1}(\boldsymbol{x} - \boldsymbol{\mu}) = c^2\} \\ &= \text{ superfícies elipsoides centrada em } \boldsymbol{\mu} \end{aligned}$$

- Propriedades das **elipsoides de densidade constante**:
  1. Centradas em $\boldsymbol{\mu}$
  2. Os eixos têm a direção dos vetores próprios de $\boldsymbol{\Sigma}$
  3. Os comprimentos dos eixos são proporcionais à raiz quadrada dos valores próprios de $\boldsymbol{\Sigma}$
- Em resumo, os $p$ eixos são definidos por

$$\boldsymbol{\mu} \pm c\sqrt{\lambda_i}e_i \; (i = 1, ..., p)$$

  onde $\boldsymbol{\lambda}$ e $\mathbf{e}$ representam os valores e vetores próprios de $\boldsymbol{\Sigma}$, respetivamente

- Seja $\mathbf{z} = (\boldsymbol{\Sigma}^{1/2})^{-1}(\mathbf{x} - \boldsymbol{\mu})$ tal que $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}^{1/2}\boldsymbol{\Sigma}^{1/2}$, então

$$\mathbf{x} \frown N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \Rightarrow \mathbf{z} \frown N_p(\mathbf{0}, \mathbf{I})$$

- Atendendo à definição de $\mathbf{z}$ obtém-se que $\mathbf{z}'\mathbf{z} = (\mathbf{x} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})$, pelo que

$$(\mathbf{x} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) \frown \chi^2_p$$

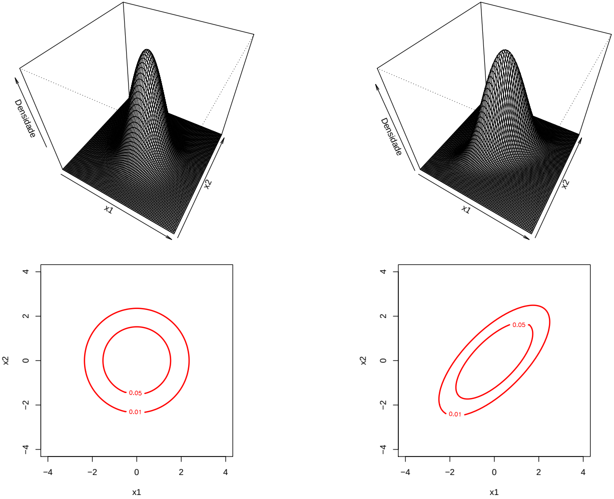
- Logo, as superfícies elipsoides tais que

$$(\mathbf{x} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) \leq \chi^2_{p,1-\alpha}$$

  contêm $(1 - \alpha) \times 100\%$ dos valores de $\mathbf{x}$, tendo eixos definidos por

$$\boldsymbol{\mu} \pm c\sqrt{\lambda_i}\mathbf{e}_i \; (i = 1, ..., p)$$

  com $c = \sqrt{\chi^2_{p,1-\alpha}}$


Densidade
x1
x2
Densidade
x1
x2
0.05
0.01
x2
x1
4
2
0
−2
−4
−4
−2
0
2
4
0.05
0.01
x2
x1
4
2
0
−2
−4
−4
−2
0
2
4

## Exemplo 8

Considere $\mathbf{x} \frown N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, com $\boldsymbol{\mu}' = (5, 10)$ e

$$\boldsymbol{\Sigma} = \begin{bmatrix} 9 & 16 \\ 16 & 64 \end{bmatrix}$$

Defina a a superfície elíptica que contem aproximadamente 95% dos valores de $\mathbf{x}$.

Os valores próprios de $\boldsymbol{\Sigma}$ são os valores $\lambda_1 = 68.316$ e $\lambda_2 = 4.684$

Os vetores próprios normalizados da matriz $\boldsymbol{\Sigma}$ são $\mathbf{e}_1' = (0.2604, 0.9655)$ e $\mathbf{e}_2' = (0.9655, -0.2604)$

Código R:

```
> Sigma<-matrix(c(9,16,16,64),2,2)
> eigen(Sigma)

eigen() decomposition
$values
[1] 68.315877  4.684123

$vectors
          [,1]       [,2]
[1,] 0.2604339 -0.9654917
[2,] 0.9654917  0.2604339
```

A superfície definida por

$$(\mathbf{x}-\boldsymbol{\mu})'\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu}) \leq \chi^2_{(2),0.95}$$

contem 95% dos valores de $\mathbf{x}$.

Fixando $c = \sqrt{\chi^2_{(2),0.95}} = \sqrt{5.9915}$, a superfície elíptica que contem aproximadamente 95% tem eixos

$$\boldsymbol{\mu} \pm c\sqrt{\lambda_i}\mathbf{e}_i \; (i = 1, 2)$$

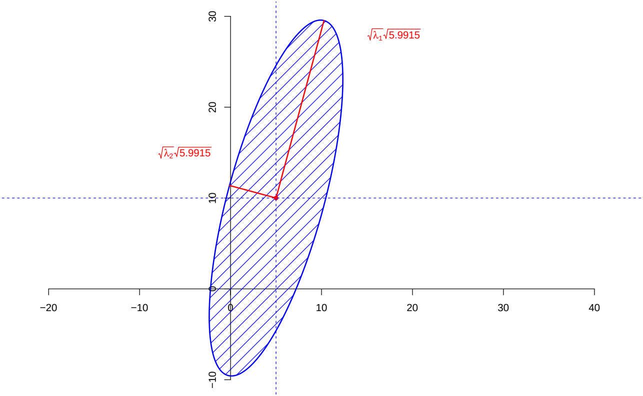
**Maior eixo**:

$$\underbrace{\begin{bmatrix} 5 \\ 10 \end{bmatrix}}_{\boldsymbol{\mu}} \pm \underbrace{\sqrt{5.9915}}_{\chi^2_{(2),0.95}} \underbrace{\sqrt{68.316}}_{\sqrt{\lambda_1}} \underbrace{\begin{bmatrix} 0.2604 \\ 0.9655 \end{bmatrix}}_{\mathbf{e}_1} = \begin{bmatrix} -0.272 \\ -9.551 \end{bmatrix}, \begin{bmatrix} 10.273 \\ 29.551 \end{bmatrix}$$

**Menor eixo**:

$$\underbrace{\begin{bmatrix} 5 \\ 10 \end{bmatrix}}_{\boldsymbol{\mu}} \pm \underbrace{\sqrt{5.9915}}_{\chi^2_{(2),0.95}} \underbrace{\sqrt{4.684}}_{\sqrt{\lambda_2}} \underbrace{\begin{bmatrix} -0.9655 \\ 0.2604 \end{bmatrix}}_{\mathbf{e}_2} = \begin{bmatrix} -0.0119 \\ 11.381 \end{bmatrix}, \begin{bmatrix} 10.119 \\ 8.619 \end{bmatrix}$$

# (continuação)



# Propriedades da distribuição normal multivariada

Seja $\mathbf{x}$ um vetor aleatório com distribuição $N_p(\mu, \boldsymbol{\Sigma})$:

- **Normalidade de combinações lineares**:
  - Se $a'$ é um vetor de constantes então a combinação linear $a'\mathbf{x} = a_1 x_1 + \ldots + a_p x_p$ tem distribuição normal univariada

    $$a'\mathbf{x} \frown N(a'\mu, a'\boldsymbol{\Sigma}a)$$

  - Se $\mathbf{A}$ é uma matriz de constantes $q \times p\,(q < p)$, as $q$ combinações lineares $\mathbf{Ax}$ têm distribuição normal multivariada

    $$\mathbf{Ax} \frown N_q(\mathbf{A}\boldsymbol{\mu}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}')$$

- **Distribuição normal padrão multivariada**: Seja $\mathbf{z} = (\boldsymbol{\Sigma})^{-1/2}(\mathbf{x} - \boldsymbol{\mu})$ então

  $$\mathbf{z} \frown N_p(\mathbf{0}, \mathbf{I})$$

# Exemplo 9

Seja $\mathbf{y} \frown N_4(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, com $\boldsymbol{\mu}' = (-2, 3, -1, 5)$ e

$$\boldsymbol{\Sigma} = \begin{bmatrix} 11 & -8 & 3 & 9 \\ -8 & 9 & -3 & 6 \\ 3 & -3 & 2 & 3 \\ 9 & 6 & 3 & 9 \end{bmatrix}$$

a) Qual a distribuição de $z = 4y_1 - 2y_2 + y_3 - 3y_4$?

$$\mu_z = \boldsymbol{c}'\boldsymbol{\mu} = -30 \quad \sigma_z^2 = \boldsymbol{c}'\boldsymbol{\Sigma}\boldsymbol{c} = 297$$

logo $z \frown N(-30, 297)$

b) Considere as combinações lineares:

$$z_1 = y_1 + y_2 + y_3 + y_4 - 2$$

$$z_2 = -2y_1 + 3y_2 + y_3 - 2y_4$$

Qual a distribuição conjunta de $z_1$ e $z_2$?

$$\mathbf{z} \frown N_2\left( \begin{bmatrix} 5 \\ 2 \end{bmatrix}, \begin{bmatrix} 51 & -67 \\ -67 & 217 \end{bmatrix} \right)$$

- **Normalidade das margens**:
  - Todos os subconjuntos de uma normal multivariada são necessariamente normais $\Rightarrow$ um vetor multivariado só terá distribuição normail multivariada se todas as partições forem normais
  - Mais especificamente, qualquer partição com 2 subconjuntos com $q$ e $p-q$ variáveis tem distribuição normal.
  - Seja $\mathbf{x}_1' = (x_1, \ldots, x_q)$ e $\mathbf{x}_2' = (x_{q+1}, \ldots, x_p)$ os dois subvetores formados pela partição, então

$$\mathbf{x}_1 \frown N_q(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11})$$

$$\mathbf{x}_2 \frown N_{p-q}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22})$$

  - Como caso particular, considerando $p$ partições, tem-se

$$y_j \frown N(\mu_j, \sigma_j)\,(j = 1, \ldots, p)$$

(Atenção: o contrário não é necessariamente verdadeiro!)

c) Qual a distribuição de $y_3$?

$$y_3 \frown N(-1, 2)$$

d) Qual a distribuição conjunta de $y_2$ e $y_4$?

Note-se que

$$\mathbf{y} = \begin{bmatrix} y_2 \\ y_4 \\ \hdashline y_1 \\ y_3 \end{bmatrix} \quad \boldsymbol{\mu} = \begin{bmatrix} \mu_2 \\ \mu_4 \\ \hdashline \mu_1 \\ \mu_3 \end{bmatrix}$$

$$\boldsymbol{\Sigma} = \left[\begin{array}{cc:cc} \sigma_{22} & \sigma_{24} & \sigma_{21} & \sigma_{23} \\ \sigma_{42} & \sigma_{44} & \sigma_{41} & \sigma_{43} \\ \hdashline \sigma_{12} & \sigma_{14} & \sigma_{11} & \sigma_{13} \\ \sigma_{32} & \sigma_{34} & \sigma_{31} & \sigma_{33} \end{array}\right]$$

Logo, $\mathbf{y}^{(1)} = \begin{bmatrix} y_2 \\ y_4 \end{bmatrix} \frown N_2\left( \begin{bmatrix} 3 \\ 5 \end{bmatrix}, \begin{bmatrix} 9 & 6 \\ 6 & 9 \end{bmatrix} \right)$

Seja $\mathbf{x}$ um vetor aleatório com distribuição $N_p(\mu, \mathbf{\Sigma})$ e $\mathbf{x}' = (x_1, ..., x_q)$ e $\mathbf{y}' = (y_1, ..., y_p)$ dois subvetores resultantes da partição em 2 conjuntos com $q$ e $p$ elementos:

- **Independência**:
  - Dois subvetores são independentes se $\mathbf{\Sigma_{xy}} = \mathbf{O}$ (apenas em populações normais)
  - Duas variáveis $x_1$ e $x_2$ são independentes se $\sigma_{12} = 0$ (apenas em populações normais)
- **Distribuição condicional**: Se 2 subvetores não são independentes então $\mathbf{\Sigma_{xy}} \neq \mathbf{O}$ e a distribuição condicional de $\mathbf{y}$ dado $\mathbf{x}$, $f(\mathbf{y}|\mathbf{x})$, é normal multivariada com parâmetros

$$E(\mathbf{y}|\mathbf{x}) = \boldsymbol{\mu}_\mathbf{y} + \mathbf{\Sigma_{yx}}\mathbf{\Sigma_{xx}}^{-1}(\mathbf{x} - \boldsymbol{\mu}_\mathbf{x})$$

$$Cov(\mathbf{y}|\mathbf{x}) = \mathbf{\Sigma_{yy}} - \mathbf{\Sigma_{yx}}\mathbf{\Sigma_{xx}}^{-1}\mathbf{\Sigma_{xy}}$$

- **Distribuição da soma**: Seja $\mathbf{x}$ e $\mathbf{y}$ vetores independentes com o mesmo comprimento, então

$$\mathbf{x} \pm \mathbf{y} \frown N_p(\boldsymbol{\mu}_x \pm \boldsymbol{\mu}_y, \mathbf{\Sigma}_{xx} + \mathbf{\Sigma}_{yy})$$

# Exemplo 10

Seja $\mathbf{x} \frown N_3(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, com

$$\boldsymbol{\Sigma} = \begin{bmatrix} 4 & 1 & 0 \\ 1 & 3 & 0 \\ 0 & 0 & 2 \end{bmatrix}$$

a) Serão $x_1$ e $x_2$ v.a. independentes?
$\sigma_{12} = 1$, logo $x_1$ e $x_2$ não são independentes.

b) Será $(x_1, x_2)$ independente de $x_3$?

$\mathbf{x}^{(1)} = (x_1, x_2)'$ e $\mathbf{x}^{(2)} = x_3$ são independentes se $\boldsymbol{\Sigma}_{12} = \mathbf{O}$

Note-se que

$$\mathbf{x} = \left[\begin{array}{c} x_1 \\ x_2 \\ \hdashline x_3 \end{array}\right]$$

e

$$\boldsymbol{\Sigma} = \left[\begin{array}{cc:c} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} \\ \hdashline \sigma_{31} & \sigma_{32} & \sigma_{33} \end{array}\right] = \left[\begin{array}{cc:c} 4 & 1 & 0 \\ 1 & 3 & 0 \\ \hdashline 0 & 0 & 2 \end{array}\right]$$

Logo, $(x_1, x_2)$ e $x_3$ são independentes.