MAE 5776

ANÁLISE MULTIVARIADA

Júlia M Pavan Soler

pavan@ime.usp.br

1º Semestre IME/2022

MAE5776

$$Y_{n\times p} = \left(Y_{ij}\right) \in \Re^{n\times p}$$

n “indivíduos”
p variáveis

Já vimos ☺

Matriz de Dados: Estatísticas descritivas multivariadas

- Definidas no espaço das colunas ($\Re^{p}$, $\Re^{pxp}$): $\bar{Y}_{p\times 1}, S_{p\times p}, R_{p\times p}, S^{-1}_{p\times p}$ Matriz de precisão

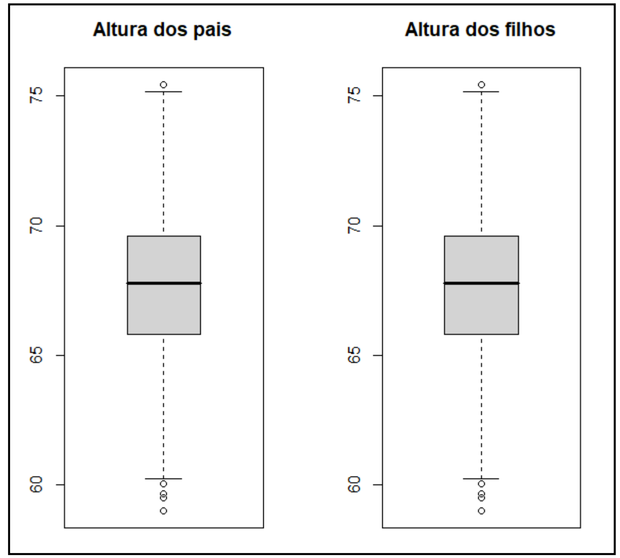
- Definidas no espaço das linhas ($\Re^{nxn}$): $D_{n\times n} = \left(d^2_{ij}\right);\ d^2_{Eij}, d^2_{Pij}, d^2_{Mij}$

Regiões (elipsóides) de Concentração de Observações ($Y_i \in \Re^p$): diagnóstico de observações atípicas (*outliers*) multivariados

$$R\left(Y_i\right) = \left(Y_i \in \Re^p;\ d^2_M(Y_i;C) = \left(Y_i - \bar{Y}\right)' S_u^{-1}\left(Y_i - \bar{Y}\right) \le c^2;\ c^2 = \chi^2_p(\alpha)\right)$$

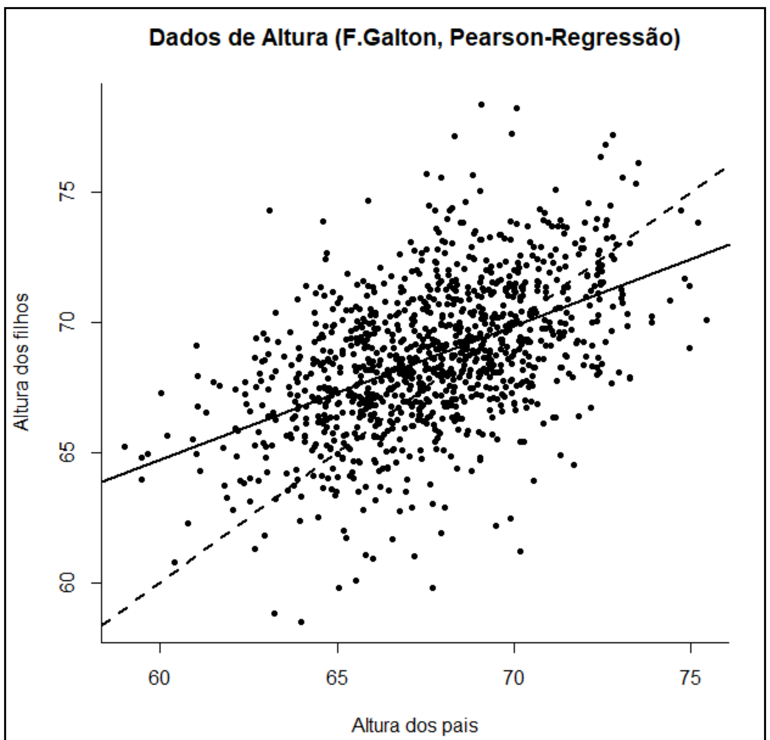
Análise Multivariada

```
Dados "father.son" do R
        fheight  sheight
1      65.04851 59.77827
2      63.25094 63.21404
3      64.95532 63.34242
4       65.75250 62.79238
5    ...
1076 71.78314 69.30589
1077 70.73837 69.30199
1078 70.30609 67.01500
```



Matriz de dados:

$Y_{1078 \times 2}$



- Análise UNIVARIADA de cada variável
- Análise MULTIVARIADA: análise conjunta das variáveis. Leva em conta a correlação (r=0.50)

Indicação das retas y=x (tracejada) e
y=33.8866+0.5141x (linha contínua)

Note que há uma regressão da altura dos filhos quando os pais são mais altos que x=69.72628

Análise Multivariada

Dados "father.son" do R

x:fheight y:sheight

	x	y
1	65.04851	59.77827
2	63.25094	63.21404
3	64.95532	63.34242
4	65.75250	62.79238
5	...	
1076	71.78314	69.30589
1077	70.73837	69.30199
1078	70.30609	67.01500

Centróide:

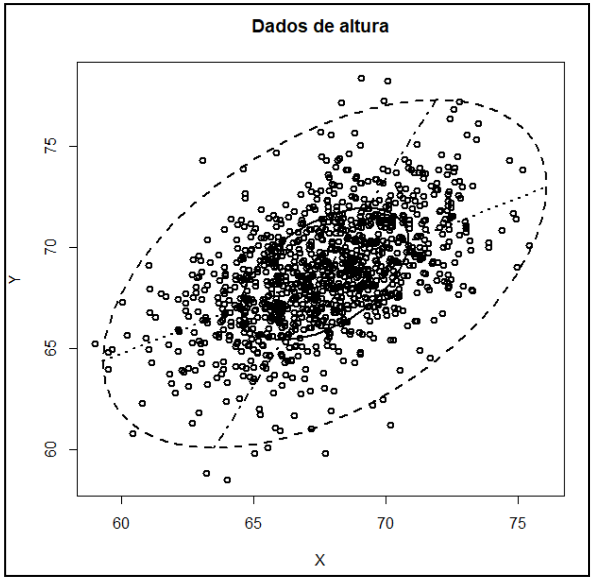
x	y
67.69	68.68

Matriz de Covariância

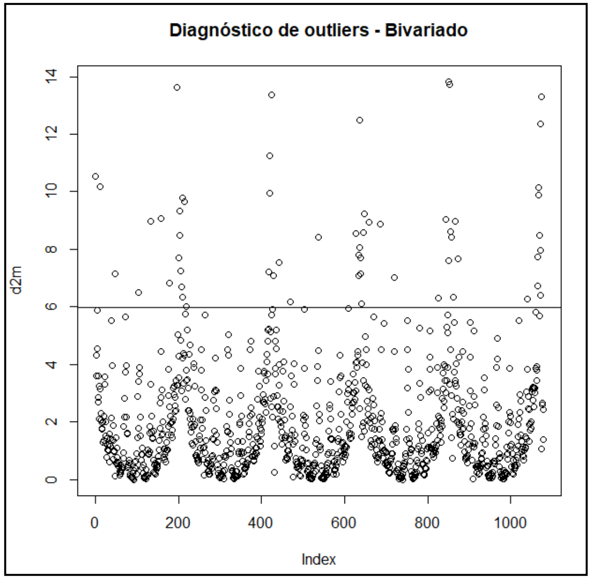
	x	y
x	7.53	3.87
y	3.87	7.92

Matriz de Correlação:

	x	y
x	1.0	0.5
y	0.5	1.0



Boxplot bivariado

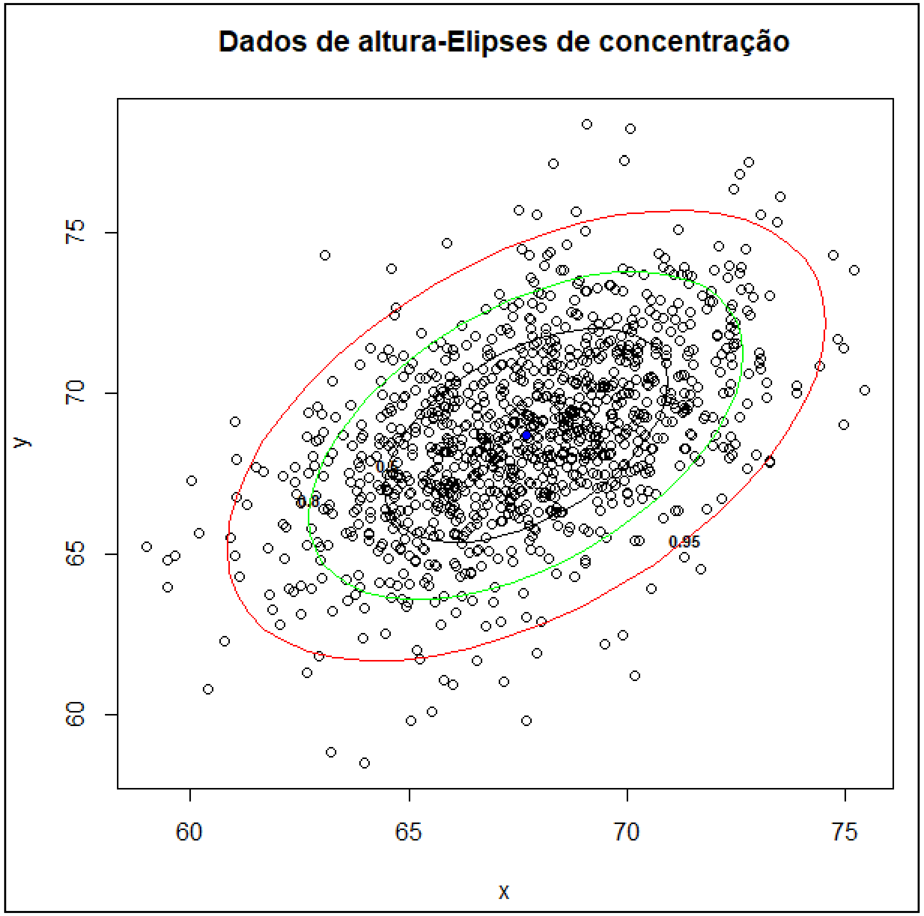


Distância de Mahalanobis das observações ao centróide: critério de observações "outliers"

$$d_M^2 = \left(Y_i - \bar{Y}\right)' S^{-1} \left(Y_i - \bar{Y}\right) \leq \chi^2_{p(1-\alpha)}$$

Sob este critério são definidas "elipses" que concentram parcelas dos dados (50%, 80%, 90%,95%)

Elipse de Concentração dos Dados



$$d_M^2 = \left(Y_i - \bar{Y}\right)' S^{-1} \left(Y_i - \bar{Y}\right) \leq \chi^2_{p(1-\alpha)}$$

Elipse de 95% de concentração (p=2, α=5%)

Elipse de 80% de concentração (p=2, α=20%)

Elipse de 50% de concentração (p=2, α=50%)

Os dois **Eixos Principais da Elipse** de concentração correspondem às direções de maior e de menor variabilidade dos dados.
Estes novos eixos são ortogonais e veremos que podem também ser usados para representar os dados!

Dados Multivariados

Como “entender” os dados na Matriz *Y*?
Dados são extraídos de qual distribuição?

Banco de Dados:

	Variáveis					
Unidades Amostrais	1	2	…	j	…	p
1	Y_{11}	Y_{12}		Y_{1j}		Y_{1p}
2	Y_{21}	Y_{22}		Y_{2j}		Y_{2p}
…	…	…	…	…		…
i	Y_{i1}	Y_{i2}		Y_{ij}		Y_{ip}
…	…	…	…	…	…	…
n	Y_{n1}	Y_{n2}		Y_{nj}		Y_{np}

$$Y_{n\times p} = \left(y_{ij}\right):$$ Matriz de Dados

$\uparrow$ resposta do i-ésimo “indivíduo” na j-ésima variável

Formalizando: Considere os dados de altura de pais e filhos como uma amostra aleatória simples de *n=1078* vetores aleatórios bidimensionais (p=2) de uma população de interesse (Ex. descendentes de imigrantes italianos no Brasil, tribo indígena, etc).

Vetor de Variáveis Aleatórias Multidimensionais

- Vetor aleatório da i-ésima observação: $Y_i = (Y_{i1}, Y_{i2}, \ldots, Y_{ip})' \in \Re^p, \, i = 1,2,\ldots, n$
- Matriz aleatória de n-observações p-dimensionais: $Y_{n\times p} = (Y_1, Y_2, \ldots, Y_n)' \in \Re^{n\times p}$

Amostra aleatória simples de *n* vetores aleatórios *p*-dimensionais (*AASn*):

$$f_Y(y) = \boxed{\prod_{i=1}^{n}} f_{Y_i}(y_i); \quad Y_i \in \Re^p$$

distribuição multivariada

densidade conjunta: suposição de **vetor de observações independentes**

$$f_Y(y) = \boxed{\prod_{i=1}^{n}\prod_{j=1}^{p}} f_{Y_{ij}}(y_{ij})$$

função de densidade univariada

densidade conjunta: suposição de **observações independentes avaliadas em p variáveis independentes**

Variáveis Aleatórias Multidimensionais

- Matriz aleatória (Gupta and Nagar, 2000):

Formulações alternativas

$$Y_{n\times p} = (Y_{ij}) \in \Re^{n\times p};$$

$$\boxed{Y_{n\times p} \sim (M; \Psi \otimes \Sigma);}$$

$$\boxed{vec(Y)_{np\times 1} \sim (vec(M); \Psi \otimes \Sigma)}$$

$$M_{n\times p} = 1_n \mu'_{p\times 1} = \begin{pmatrix} \mu_1 & \mu_2 & \cdots & \mu_p \\ \mu_1 & \mu_2 & \cdots & \mu_p \\ \cdots & \cdots & \cdots & \cdots \\ \mu_1 & \mu_2 & \cdots & \mu_p \end{pmatrix}$$: matriz de médias

$$vec(M) = \begin{pmatrix} \mu_1 \\ \mu_1 \\ \cdots \\ \mu_p \end{pmatrix}$$: vetor de médias de n observações em p var.

$$(\Psi_{n\times n} \otimes \Sigma_{p\times p})_{np\times np}$$: matriz de covariâncias (⊗: produto de Kronecker)

Formulação flexível para diferentes modelagens

Matrizes de covariância Estruturadas: entre indivíduos (Ψ) e entre variáveis (Σ):

$$\boxed{\Psi = I_n \; ; \quad \Sigma = I_p}$$

Observações e variáveis independentes

$$\boxed{\Psi = I_n \; ; \quad \Sigma = (1-\alpha) I_p + \alpha 1_p 1_p'}$$

Observações independentes e correlação uniforme entre as variáveis

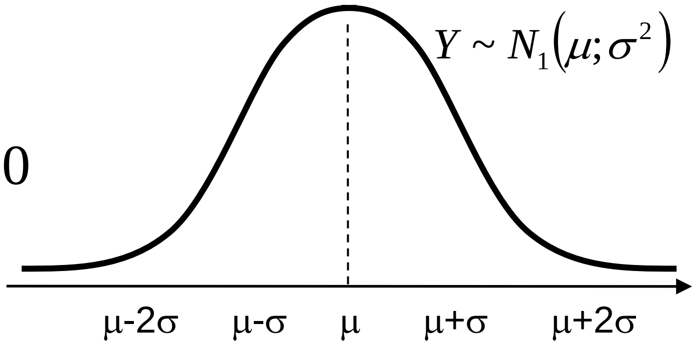
$$\boxed{\Psi = I_G \oplus \left[(1-\alpha) I_{n_g} + \alpha 1_{n_g} 1_{n_g}'\right]; \quad \Sigma = (\sigma_{jl})}$$

Correlação uniforme entre observações agrupadas em G grupos

Correlação não estruturada entre variáveis

Distribuição Normal Multivariada

- Variável (escalar) Y $\in \Re$, com **distribuição Normal univariada** de média μ e variância σ^2 tem densidade dada por:



$$f_Y(y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\boxed{[(y-\mu)/\sigma]^2}/2} \qquad -\infty < y < \infty, \quad \sigma^2 > 0$$

Generalização multivariada para o vetor aleatório $Y_i = (Y_{i1}, Y_{i2}, ..., Y_{ip})' \in \Re^p$

Distribuição Normal Multivariada com vetor de média μ e matriz de covariância Σ:

$$Y_{i\,p\times1} \sim N_p(\mu_{p\times1};\Sigma_{p\times p}); \qquad f_{Y_i}(y) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} e^{-\boxed{(y-\mu)'\Sigma^{-1}(y-\mu)}/2} \qquad |\Sigma| > 0,\ y \in \Re^p,\ i = 1,..., n$$

$$d_M^2 = (y-\mu)'\Sigma^{-1}(y-\mu) = c^2$$

$$Z_i = \Sigma^{-1/2}(Y_i - \mu) \sim N_p(0_{p\times1}; I_p)$$

$$d_M^2(Y_i, C) = Z_i' Z_i$$

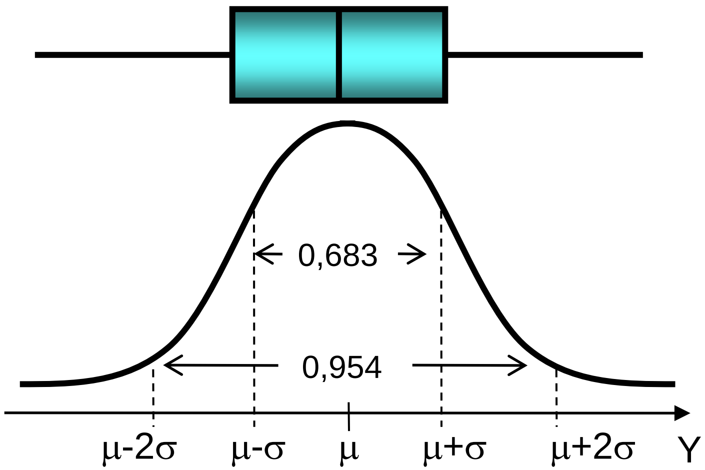
a densidade é constante em superfícies onde a distância de Mahalanobis é constante.

Distribuição Normal Uni e Multivariada

Normal Univariada: Ex. Dados de altura do pai (Y_1) ou do filho (Y_2)

$$Y_j \sim N_1\left(\mu_j, \sigma_j^{\,2}\right); \quad j = 1,2$$

$$f_{Y_j}(y) = \frac{1}{\sqrt{2\pi\sigma_j^{\,2}}} e^{-\left[\left(y-\mu_j\right)/\sigma_j\right]^2/2} \qquad y \in \Re, \sigma_j^{\,2} > 0$$



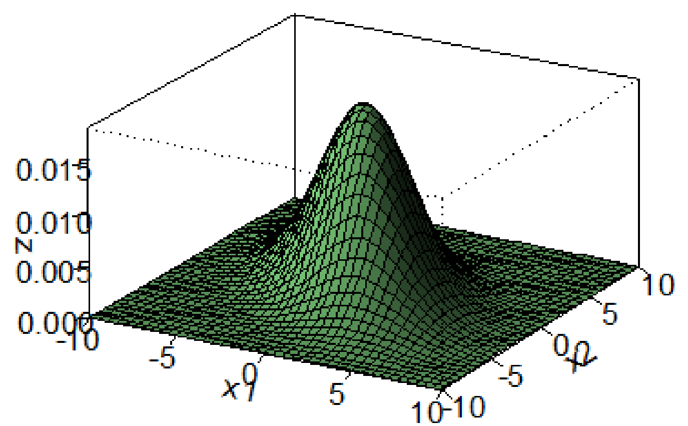
Normal Multivariada Bidimensional (p=2): Ex. Dados do vetor de alturas dos pais e filhos

$$Y_{2\times1} = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} \sim N_2\left(\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}; \Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix} \right);$$

$$y \in \Re^2, |\Sigma| > 0, \sigma_{12} = \rho\sigma_1\sigma_2$$

$$f_Y(y) = \frac{1}{2\pi\sigma_1\sigma_2\left(1-\rho^2\right)^{1/2}}$$

$$\exp\left\{ \frac{-1}{2\left(1-\rho^2\right)} \left[\left(\frac{Y_1-\mu_1}{\sigma_1}\right)^2 + \left(\frac{Y_2-\mu_2}{\sigma_2}\right)^2 - 2\rho\left(\frac{Y_1-\mu_1}{\sigma_1}\right)\left(\frac{Y_2-\mu_2}{\sigma_2}\right) \right] \right\}$$



Distribuição Normal Multivariada

Alguns Resultados:

$$Y_{p\times 1} = \begin{pmatrix} Y_{1q\times 1} \\ Y_{2(p-q)\times 1} \end{pmatrix} \sim N_p\left(\mu_{p\times 1} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}; \Sigma_{p\times p} = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \right).$$ Então:

- $Y_{1q\times 1} \sim N_q(\mu_1; \Sigma_{11})$; $\quad Y_{2(p-q)\times 1} \sim N_{(p-q)}(\mu_2; \Sigma_{22})$ distribuições marginais de Y ($\in \Re^p$) são Normais

- Y_1 e $Y_{2.1} = \left(Y_2 - \Sigma_{21}\Sigma_{11}^{-1}Y_1\right)$ são independentes, tal que,

$$Y_{2.1} \sim N_{p-q}\left(\mu_2 - \Sigma_{21}\Sigma_{11}^{-1}\mu_1 \,;\, \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\right)$$

- $Y_2 \mid Y_1 \sim N_{p-q}\left(\mu_2 - \Sigma_{21}\Sigma_{11}^{-1}\left(Y_1 - \mu_1\right); \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\right)$ distribuições condicionais de Y ($\in \Re^p$) são Normais

$$Y_2 = Y_{2.1} + \boxed{\Sigma_{21}\Sigma_{11}^{-1}Y_1}$$

Condicionado em Y1,
este termo é constante

Teorema 3.2.4
Exemplo 3.2.1: $\Sigma = (1-\rho)\, I_p + \rho\, 1_p 1_p'$
(Mardia et al., 2003)

Distribuição Normal Multivariada

Resultados considerando os Dados da altura de pais e filhos:

$$Y_{2\times1} = \begin{pmatrix} Y_{1\,1\times1} \\ Y_{2\,1\times1} \end{pmatrix} \sim N_2\left(\mu_{2\times1} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}; \Sigma_{2\times2} = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix} \right), \ \sigma_{jj} = \sigma_j^2 \ .$$ Então:

- $Y_1 \sim N_1\left(\mu_1, \sigma_1^{\ 2}\right) \qquad Y_2 \sim N_1\left(\mu_2, \sigma_2^{\ 2}\right)$ distribuições marginais do vetor Y bivariado são Normais univariadas

- $Y_1 \sim N_1\left(\mu_1, \sigma_1^{\ 2}\right)$ e $Y_{2.1} = Y_2 - \frac{\sigma_{12}}{\sigma_{11}} Y_1$ são independentes, tal que,

$$Y_{2.1} \sim N_1\left(\mu_2 - \frac{\sigma_{12}}{\sigma_{11}} \mu_1 \, ; \, \sigma_{22} - \frac{\sigma_{12}^2}{\sigma_{11}} \right)$$

- $Y_2 \mid Y_1 \sim N_1\left(\mu_2 - \frac{\sigma_{12}}{\sigma_{11}} \left(Y_1 - \mu_1 \right); \sigma_{22} - \frac{\sigma_{12}^2}{\sigma_{11}} \right)$ distribuições condicionais de Y ($\in \Re^p$) são Normais

$$Y_2 = Y_{2.1} + \boxed{\frac{\sigma_{12}}{\sigma_{11}} Y_1}$$

Condicionado em Y1,
este termo é constante

Distribuição Normal Multivariada

Alguns Resultados: **A função de verossimilhança da amostra Y** $\in \Re^{nxp}$ da Normal é,

$$L(\mu,\Sigma\,|\,Y)=\prod_{i=1}^{n} f_{Y_i}\left(y_i\,|\,\mu,\Sigma\right)=\prod_{i=1}^{n}\frac{1}{\left(2\pi\right)^{p/2}\left|\Sigma\right|^{1/2}}e^{\boxed{-(y_i-\mu)'\Sigma^{-1}(y_i-\mu)/2}}$$

$$=\left(2\pi\right)^{-np/2}\left|\Sigma\right|^{-n/2}\exp\left\{-\frac{1}{2}\left[tr\left(\Sigma^{-1}nS\right)+n\left(\bar{Y}-\mu\right)'\Sigma^{-1}\left(\bar{Y}-\mu\right)\right]\right\}$$

Estimadores de Máxima Verossimilhança

$$\Rightarrow \quad \hat{\mu}=\bar{Y}_{p\times 1}; \quad \hat{\mu}_j=\bar{Y}_j, \quad j=1,2,\dots,p$$

centróide

$$\Rightarrow \quad S_{p\times p}=\frac{1}{n}\sum_{i=1}^{n}\left(Y_i-\bar{Y}\right)\left(Y_i-\bar{Y}\right)'$$

Matriz de covariância

$\bar{Y}_{p\times 1}$ e $S_{p\times p}$ são estatísticas conjuntamente suficientes para μ e Σ, respectivamente.

Os estimadores não viciados de μ e Σ são, respectivamente:

$$\bar{Y}_{p\times 1}; \quad S_{u\;p\times p}=\frac{1}{n-1}\sum_{i=1}^{n}\left(Y_i-\bar{Y}\right)\left(Y_i-\bar{Y}\right)'$$

Para os Dados das alturas de pais e filhos (n=1078 e p=2):

$$L(\mu,\Sigma\,|\,Y)=\prod_{i=1}^{n} f_{Y_i}\left(y_i\,|\,\mu,\Sigma\right)=\prod_{i=1}^{n}\frac{1}{\left(2\pi\right)^{p/2}\left|\Sigma\right|^{1/2}}e^{-\left(y_i-\mu\right)'\Sigma^{-1}\left(y_i-\mu\right)/2}$$

$$=\frac{1}{2\pi\sigma_1\sigma_2\left(1-\rho^2\right)^{1/2}}\exp\left\{\frac{-1}{2\left(1-\rho^2\right)}\left[\left(\frac{Y_1-\mu_1}{\sigma_1}\right)^2+\left(\frac{Y_2-\mu_2}{\sigma_2}\right)^2-2\rho\left(\frac{Y_1-\mu_1}{\sigma_1}\right)\left(\frac{Y_2-\mu_2}{\sigma_2}\right)\right]\right\}$$

Estimadores de Máxima Verossimilhança (não viciados)

Vetor de Médias amostral (Centróide amostral): $\hat{\mu}=\bar{Y}_{2\times1}=\begin{pmatrix}\bar{Y}_1\\ \bar{Y}_2\end{pmatrix}=\begin{pmatrix}67,69\\ 68,68\end{pmatrix}$

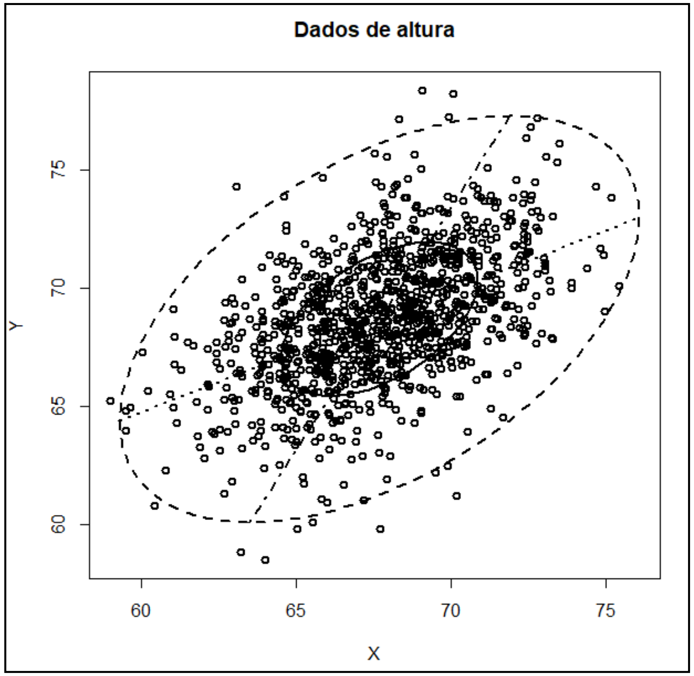
Matriz de Covariância amostral: $\hat{\Sigma}_{''2\times2}=S_{u\,2\times2}=\begin{pmatrix}s_{11} & s_{12}\\ s_{21} & s_{22}\end{pmatrix}=\begin{pmatrix}7,53 & 3,87\\ 3,87 & 7,92\end{pmatrix}$

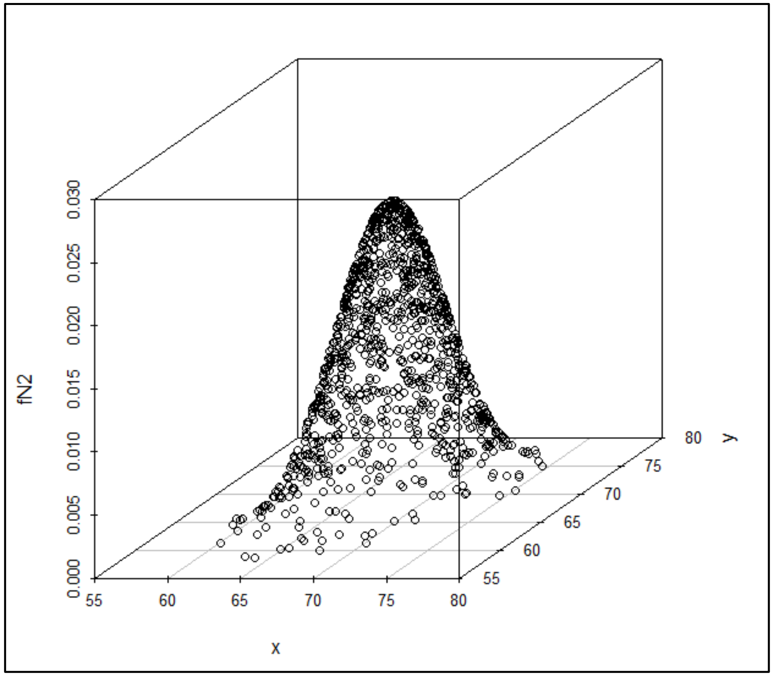
Matriz de Correlação amostral: $R_{2\times2}=\begin{pmatrix}1 & r_{12}\\ r_{21} & 1\end{pmatrix}=\begin{pmatrix}1 & 0,5\\ 0,5 & 1\end{pmatrix}$

$$\hat{\rho}=r_{12}=\frac{\hat{\sigma}_{12}}{\hat{\sigma}_1\hat{\sigma}_2}=\frac{3,87}{\sqrt{7,53}\sqrt{7,92}}=0,5$$

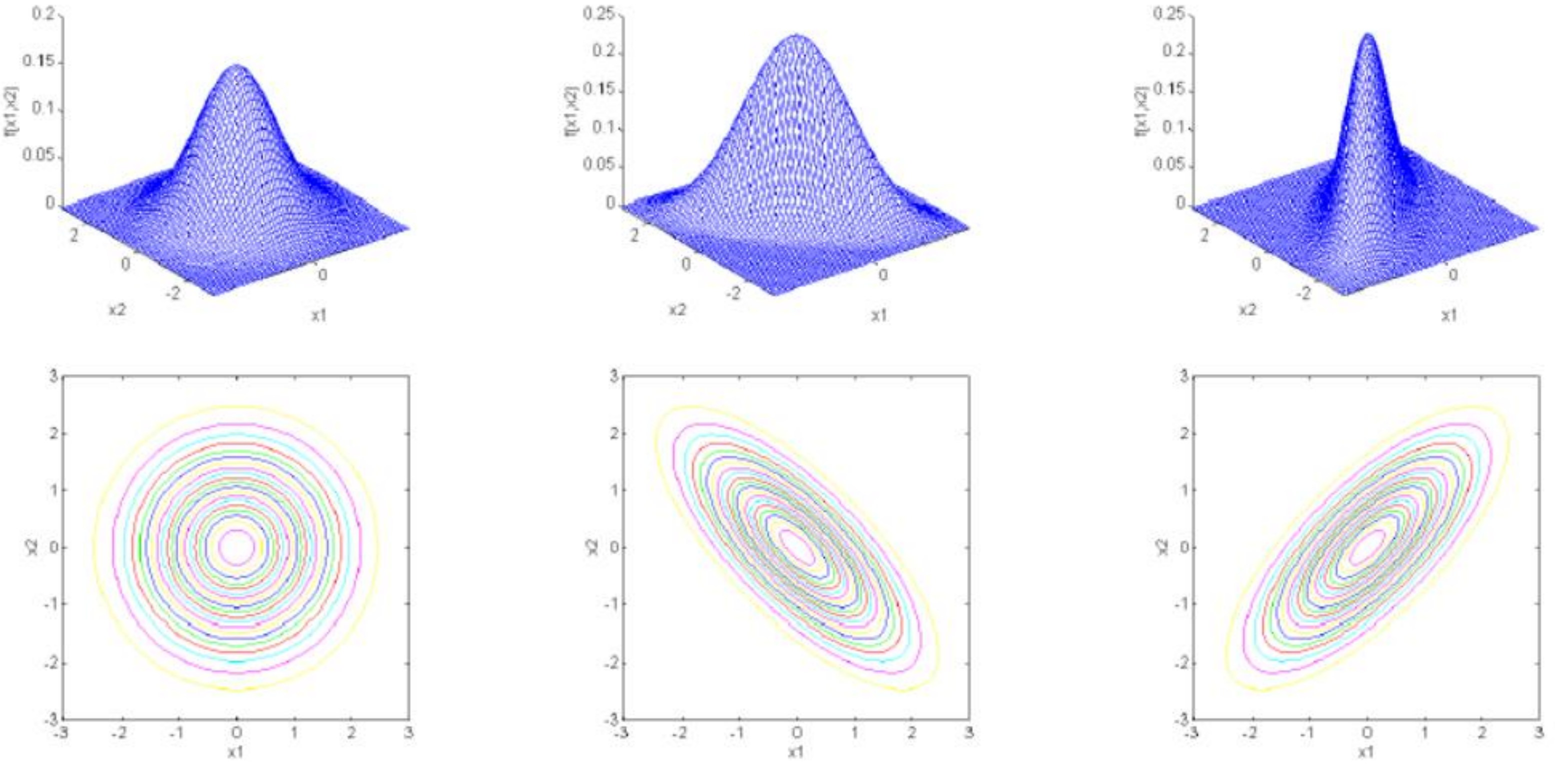
Distribuição Normal Bivariada

Para os Dados das alturas de pais e filhos (n=1078 e p=2):





Distribuição Normal Bivariada
Elipses de Concentração dos Dados



$\mu = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$ e $\Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$

$\mu = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$ e $\Sigma = \begin{bmatrix} 1 & -0{,}75 \\ -0{,}75 & 1 \end{bmatrix}$

$\mu = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$ e $\Sigma = \begin{bmatrix} 1 & 0{,}75 \\ 0{,}75 & 1 \end{bmatrix}$

Resultados Inferenciais

Próximos tópicos!

- Distribuição Amostral das estatísticas multivariadas:

$$\bar{Y}_{p\times 1}, \quad S_{p\times p}, \quad d_M^2$$

- Regiões de Confiança e Teste de Hipóteses para o Centróide Populacional:

$\Rightarrow$ Caso de uma única População

$$H_0 : \mu_{p\times 1} = 0$$

$\Rightarrow$ Caso de Duas Populações (Pareadas e Independentes)

$$H_0 : \mu_{1\,p\times 1} = \mu_{2\,p\times 1}$$

Comparações múltiplas e correções para múltiplos testes.