

MAE 5776

ANÁLISE MULTIVARIADA

Júlia M Pavan Soler

pavan@ime.usp.br

1º Semestre IME/2020

Dados Multivariados

Banco de Dados:

Unidades Amostrais	Variáveis					
	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}



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



resposta do i-ésimo "indivíduo" na j-ésima variável

Já vimos algumas estatísticas descritivas multivariadas (var. quantitativas)

Resultado: $d_M^2 = (Y_i - \bar{Y})' S_u^{-1} (Y_i - \bar{Y}) \sim \chi_p^2$

$$R(Y) = \{d_M^2 \leq c^2\}$$

Critério de diagnóstico multivariado de observações outliers

Vetor de Variáveis Aleatórias Multidimensionais

- Vetor aleatório da i -ésima observação: $Y_i = (Y_{i1}, Y_{i2}, \dots, Y_{ip})' \in \mathfrak{R}^p, i = 1, 2, \dots, n$

- Matriz de n -observações p -dimensionais: $Y_{n \times p} = (Y_1, Y_2, \dots, Y_n)' \in \mathfrak{R}^{n \times p}$

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

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

↑ distribuição multivariada

densidade conjunta: suposição de **observações independentes**

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

↑ 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

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

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

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

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

$$\text{vec}(M) = \begin{pmatrix} \mu_1 \\ \mu_1 \\ \dots \\ \mu_p \end{pmatrix}$$

: vetor de médias de n observações em p variáveis

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

Formulação flexível para diferentes modelagens

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

$$\Psi = I_n; \quad \Sigma = I_p$$

Observações e variáveis independentes

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

Observações independentes e correlação uniforme entre as variáveis

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

Correlação uniforme entre observações agrupadas em G grupos

Correlação não estruturada entre variáveis

Distribuição Normal Multivariada

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

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

Notação: $Y \sim N_1(\mu; \sigma^2)$



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

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

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

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

$$\Sigma^{-1/2} (Y_i - \mu) \sim N_p(0; I)$$

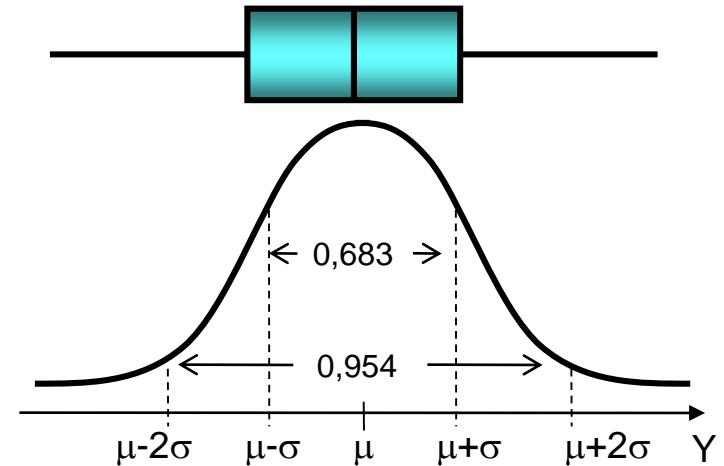
a densidade é constante em superfícies onde a distância de Mahalanobis é constante.

Distribuição Normal Uni e Multivariada

Normal Univariada:

$$Y \sim N_1(\mu, \sigma^2)$$

$$f_Y(y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-[(y-\mu)/\sigma]^2/2} \quad y \in \mathfrak{R}, \sigma^2 > 0$$



Normal Multivariada Bidimensional

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

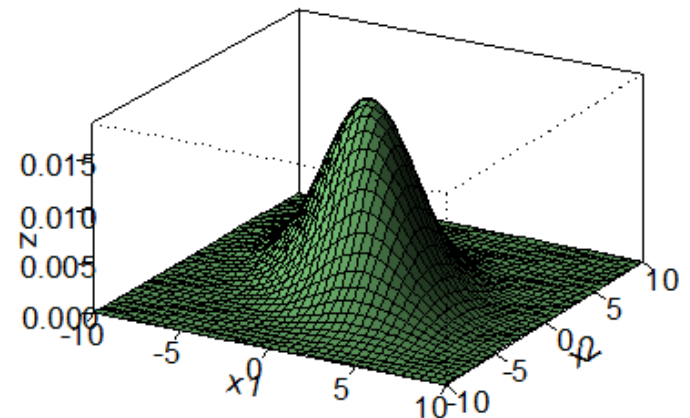
$$y \in \mathfrak{R}^2, |\Sigma| > 0, \sigma_{12} = \rho\sigma_1\sigma_2$$

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

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

Two dimensional Normal Distribution

$$\mu_1 = 0, \mu_2 = 0, \sigma_{11} = 10, \sigma_{22} = 10, \sigma_{12} = 5, \rho = 0.5$$



Distribuição Normal Multivariada

Alguns Resultados:

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

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

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

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

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

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

Condicionado em Y_1 ,
este termo é constante

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

Distribuição Normal Multivariada

Alguns Resultados:

$$L(\mu, \Sigma | Y) = \prod_{i=1}^n f_{Y_i}(y_i | \mu, \Sigma) = \prod_{i=1}^n \frac{1}{(2\pi)^{p/2} |\Sigma|^{1/2}} e^{-(y_i - \mu)' \Sigma^{-1} (y_i - \mu) / 2}$$
$$= |\Sigma|^{-n/2} \exp \left\{ -\frac{1}{2} \left[\text{tr}(\Sigma^{-1} nS) + n(\bar{Y} - \mu)' \Sigma^{-1} (\bar{Y} - \mu) \right] \right\}$$

Estimadores de
Máxima
Verossimilhança



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

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

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



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

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

Distribuição Amostral de Estimadores



$$Y_{n \times p} = \begin{pmatrix} Y_{1.} \\ Y_{2.} \\ \dots \\ Y_{n.} \end{pmatrix} \text{ é AASn da } N_p(\mu; \Sigma)$$

$$Y \sim N_{n,p}(\mathbf{1}_n \mu'; I_n \otimes \Sigma); \quad \text{vec}(Y)_{np \times 1} \sim N_{np}(\mathbf{1}_n \otimes \mu; I_n \otimes \Sigma)$$

$$\Rightarrow \bar{Y} \sim N_p\left(\mu_{p \times 1}; \frac{1}{n} \Sigma\right)$$

Ganho em precisão

\bar{Y} e S são independentes

$$\Rightarrow nS = \sum_{i=1}^n (Y_i - \bar{Y})(Y_i - \bar{Y})' \sim W_p(n-1; \Sigma)$$

$$(n-1)S_u \sim W_p(n-1; \Sigma)$$

$$W_1(n; \sigma^2) = \sigma^2 \chi_n^2$$

Distâncias de Mahalanobis

Definição: Seja $\delta_{p \times 1} \sim N_p(0; \Sigma)$, $M_{p \times p} \sim W_p(n; \Sigma)$, com δ e M variáveis independentes.

Então: $n\delta' M^{-1} \delta \sim T^2(p; n)$ **Distribuição T^2 de Hotelling**



$$Y_{i \times 1} \sim N_p(\mu; \Sigma)$$

$$\bar{Y} \sim N_p(\mu; \Sigma/n); \quad \sqrt{n}(\bar{Y} - \mu) \sim N_p(0; \Sigma) \quad d_M^2 = (n-1)(\bar{Y} - \mu)' S^{-1} (\bar{Y} - \mu) \sim T^2(p; n-1)$$

$$nS \sim W_p(n-1; \Sigma)$$

$$(n-1)S_u \sim W_p(n-1; \Sigma)$$

$$d_M^2 = n(\bar{Y} - \mu)' S_u^{-1} (\bar{Y} - \mu) \sim T^2(p; n-1)$$

Distância generalizada

Teorema 3.5.2

(Mardia et al., 2003) $T^2(p; n-1) = \frac{(n-1)p}{n-p} F_{(p; n-p)}$

Regiões de Confiança para $\mu \in \mathfrak{R}^p$


Inferências sobre o Vetor de Médias μ

$$Y_{n \times p} \text{ i.i.d. } AAS_n \quad N_p(\mu; \Sigma); \quad |\Sigma| > 0$$

Matriz de dados é uma Amostra Aleatória da Distribuição Normal Multivariada

Uma Região de Confiança $R(\mu|Y)$ para o vetor de médias da $N_p(\mu; \Sigma)$ é uma região de valores prováveis de $\mu \in \mathfrak{R}^p$, com base na amostra.

Uma construção (natural) desta região, com base na amostra, é usar medidas de distância tais como:


$$R(\mu|Y) = \left\{ n(\bar{Y} - \mu)' S_u^{-1} (\bar{Y} - \mu) \leq c^2 \right\}$$

Região (elipsoide) de confiança que contém todos os valores de μ “próximos” à evidência amostral \bar{Y}

$$P(\mu \in R(\mu|Y)) = 1 - \alpha$$

Para um nível de significância α fixado, c é obtido da distribuição F

$$\Rightarrow d_M^2 = n(\bar{Y} - \mu)' S_u^{-1} (\bar{Y} - \mu) \sim T^2(p; n-1) = \frac{(n-1)p}{n-p} F_{(p; n-p)}$$

Regiões de Confiança para $\mu \in \mathfrak{R}^p$

Inferências sobre o vetor μ

$$R(\mu|Y) = \left\{ \mu \in \mathfrak{R}^p; n (\bar{Y} - \mu)' S_u^{-1} (\bar{Y} - \mu) \leq c^2 = \frac{(n-1)p}{(n-p)} F_{p,(n-p)}(\alpha) \right\}$$

⇒ Para determinar se algum ponto μ_0 cai na região $R(\mu|Y)$ basta calcular a distância generalizada ao quadrado e compará-la com o valor crítico dado em função da distribuição F e do nível de significância α , isto é,

$$n (\bar{Y} - \mu_0)' S_u^{-1} (\bar{Y} - \mu_0) \leq c^2 = \frac{(n-1)p}{(n-p)} F_{p,(n-p)}(\alpha)$$

⇒ **Regiões de Confiança** correspondem a **Regiões de Aceitação** em testes de hipóteses sobre o vetor μ .

Regiões de Confiança e Testes de Hipóteses

Taxas de açúcar, sódio e potássio sanguíneas em 20 mulheres adultas

Indiv.	Açúcar	Sódio	Potássio
1	3,7	48,5	9,3
2	5,7	65,1	8
3	3,8	47,2	10,9
4	3,2	53,2	12
5	3,1	55,5	9,7
6	4,6	36,1	7,9
7	2,4	24,8	14
8	7,2	33,1	7,6
9	6,7	47,4	8,5
10	5,4	54,1	11,3
11	3,9	36,9	12,7
12	4,5	58,8	12,3
13	3,5	27,8	9,8
14	4,5	40,2	8,4
15	1,5	13,5	10,1
16	8,5	56,4	7,1
17	4,5	71,6	8,2
18	6,5	52,8	10,9
19	4,1	44,1	11,2
20	5,5	40,9	9,4
Média	4,64	45,4	9,97
S	2,879		
	10,002	199,798	
	-1,81	-5,627	3,628

$$R(\mu | Y) = \left\{ \mu; n(\bar{Y} - \mu)' S_u^{-1} (\bar{Y} - \mu) \leq c^2 = \underbrace{\frac{(20-1)3}{(20-3)} F_{20,17}(\alpha)} \right\}$$

$$\alpha = 0,10 \Rightarrow 8,18$$

$$\alpha = 0,05 \Rightarrow 10,72$$



Suponha o interesse na seguinte hipótese:

$$\left\{ H_0 : \mu = \mu_0 = (4, 50, 10)' \right.$$

$$T^2 = n(\bar{Y} - \mu_0)' S^{-1} (\bar{Y} - \mu_0) = 9,74$$

Conclusão: $\alpha = 0,10 \Rightarrow T^2 \notin R(\mu | Y); \text{ rej } H_0$

$\alpha = 0,05 \Rightarrow T^2 \in R(\mu | Y); \text{ não rej } H_0$

Regiões de Confiança para o Vetor μ Uma Única População

Morfometria cefálica para os dois primeiros filhos de 25 famílias (Everitt, 2007)

Família	1° Filho		2° Filho	
	Comprimento	Perímetro	Comprimento	Perímetro
1	191	155	179	145
2	195	149	201	152
3	181	148	185	149
4	183	153	188	149
5	176	144	171	142
6	208	157	192	152
7	189	150	190	149
8	197	159	189	152
9	188	152	197	159
10	192	150	187	151
11	179	158	186	148
12	183	147	174	147
13	174	150	185	152
14	190	159	195	157
15	188	151	187	158
16	163	137	161	130
17	195	155	183	158
18	186	153	173	148
19	181	145	182	146
20	175	140	165	137
21	192	154	185	152
22	174	143	178	147
23	176	139	176	143
24	197	167	200	158
25	190	163	187	150

$$Y_{25 \times 4} = (Y_1, \dots, Y_{25})'; \quad Y_{i_{4 \times 1}} \stackrel{iid}{\sim} N_4(\mu; \Sigma)$$

Estatísticas Descritivas:

$$\bar{Y} = (185,72 \quad 151,12 \quad 183,84 \quad 149,24)'$$

$$S_u = \begin{pmatrix} 91,481 & 50,753 & 66,875 & 44,267 \\ & 52,186 & 49,259 & 33,651 \\ & & 96,775 & 54,278 \\ & & & 43,222 \end{pmatrix}$$

Regiões de Confiança para o Vetor μ Uma Única População

Morfometria cefálica para os dois primeiros filhos de 25 famílias (Everitt, 2007)

Família	1º Filho		2º Filho	
	Comprimento	Perímetro	Comprimento	Perímetro
1	191	155	179	145
2	195	149	201	152
3	181	148	185	149
4	183	153	188	149
5	176	144	171	142
6	208	157	192	152
7	189	150	190	149
8	197	159	189	152
9	188	152	197	159
10	192	150	187	151
11	179	158	186	148
12	183	147	174	147
13	174	150	185	152
14	190	159	195	157
15	188	151	187	158
16	163	137	161	130
17	195	155	183	158
18	186	153	173	148
19	181	145	182	146
20	175	140	165	137
21	192	154	185	152
22	174	143	178	147
23	176	139	176	143
24	197	167	200	158
25	190	163	187	150

$$Y_{25 \times 4} = (Y_1, \dots, Y_{25})'; Y_{i_{4 \times 1}} \stackrel{iid}{\sim} N_4(\mu; \Sigma)$$

Distribuição marginal (\mathfrak{R}^2):

$$Y_{n \times 2}; Y_{i_{2 \times 1}} = (Y_{i1}, Y_{i3}) \stackrel{iid}{\sim} N_2\left(\mu = \begin{pmatrix} \mu_1 \\ \mu_3 \end{pmatrix}; \Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{13} \\ \sigma_{13} & \sigma_{33} \end{pmatrix}\right)$$



$$R(\mu | Y) = \left\{ \mu \in \mathfrak{R}^2; T^2 \leq \frac{(25-1)2}{(25-2)} F_{2,23}(\alpha) \right\}$$

$$\alpha = 0,10 \Rightarrow 5,3196$$

Suponha o interesse na seguinte hipótese:

$$\left\{ \begin{array}{l} H_0: \mu = \begin{pmatrix} 182 \\ 182 \end{pmatrix} \\ T^2 = 4,186 \end{array} \right.$$

Conclusão?

Intervalos de Confiança - Regiões de Confiança

(Everitt, 2007)

Caso univariado

$I.C(\mu_k)$ a $100(1 - \alpha)\%$

$$= \left(\bar{Y}_k \mp t_{n-1}(\alpha/2) \sqrt{\frac{s_{kk}}{n}} \right)$$

$$t^2 = \frac{(\bar{Y} - \mu)^2}{s^2/n}$$

$$= n(\bar{Y} - \mu)(s^2)^{-1}(\bar{Y} - \mu)$$

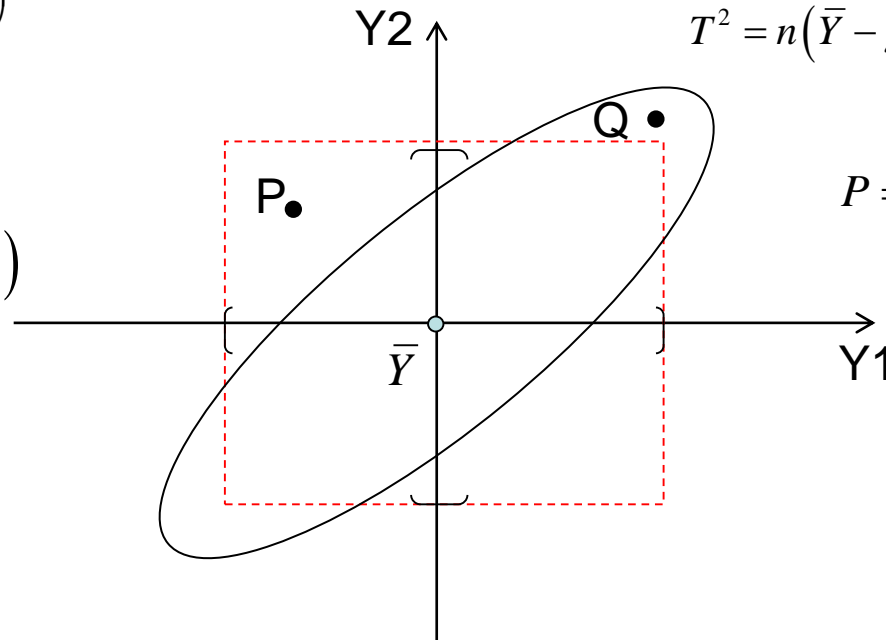
$$t_{(n-1)}^2 = F_{1,(n-1)}$$

Caso multivariado

$$R(\mu | Y) = \left\{ \mu \in \mathbb{R}^p; n(\bar{Y} - \mu)' S_u^{-1} (\bar{Y} - \mu) \leq c^2 = \frac{(n-1)p}{(n-p)} F_{p,(n-p)}(\alpha) \right\}$$

$$T^2 = n(\bar{Y} - \mu)' S_u^{-1} (\bar{Y} - \mu) \sim \frac{(n-1)p}{n-p} F_{(p;n-p)}$$

$$P = (\mu_{P1}, \mu_{P2}); \quad Q = (\mu_{Q1}, \mu_{Q2})$$



⇒ Com base na evidência amostral em $R(\mu|Y)$, valores de $\mu_0 \in \mathbb{R}^2$ iguais a Q (P) estão na região de aceitação (rejeição) de possíveis valores do parâmetro μ .

⇒ Os Intervalos de Confiança univariados podem dar decisões diferentes da região de confiança.

⇒ Como representar no gráfico a elipse de concentração de pontos amostrais?

Distâncias de Mahalanobis

$$Y_{i \times p} \stackrel{iid}{\sim} N_p(\mu; \Sigma); \quad \sqrt{n}(\bar{Y} - \mu) \sim N_p(0; \Sigma); \quad (n-1)S_u \sim W_p(n-1; \Sigma)$$

$$n(\bar{Y} - \mu)' S_u^{-1} (\bar{Y} - \mu) \sim T^2(p; n-1)$$

Distância
generalizada

Teorema 3.5.2 (Mardia et al., 2003): $T^2(p; n-1) = \frac{(n-1)p}{n-p} F_{(p; n-p)}$

Resultado: (Johnson and Wichern, 2008)

$$Y_{i \times p} \in \mathcal{R}^p, i = 1, 2, \dots, n \text{ é AASn tal que, } E(Y_i) = \mu, \text{ Cov}(Y_i) = \Sigma, |\Sigma| > 0.$$

Então, para $(n-p)$ suficientemente grande,

$$n(\bar{Y} - \mu)' S_u^{-1} (\bar{Y} - \mu) \sim \chi_p^2$$

Região de
Confiança para μ
Elipsóide de Confiança

Distâncias de Mahalanobis

$$Y_{i p \times 1} \stackrel{iid}{\sim} N_p(\mu; \Sigma) \quad \left\{ \begin{array}{l} (Y_i - \mu) \sim N_p(0; \Sigma) \\ (Y_i - \mu)' \Sigma^{-1} (Y_i - \mu) \sim \chi_p^2 \end{array} \right.$$

$$P\left(Y_i \in \mathcal{R}^p; d_M^2(Y_i; \mu) = (Y_i - \mu)' \Sigma^{-1} (Y_i - \mu) \leq c^2\right) = 1 - \alpha$$

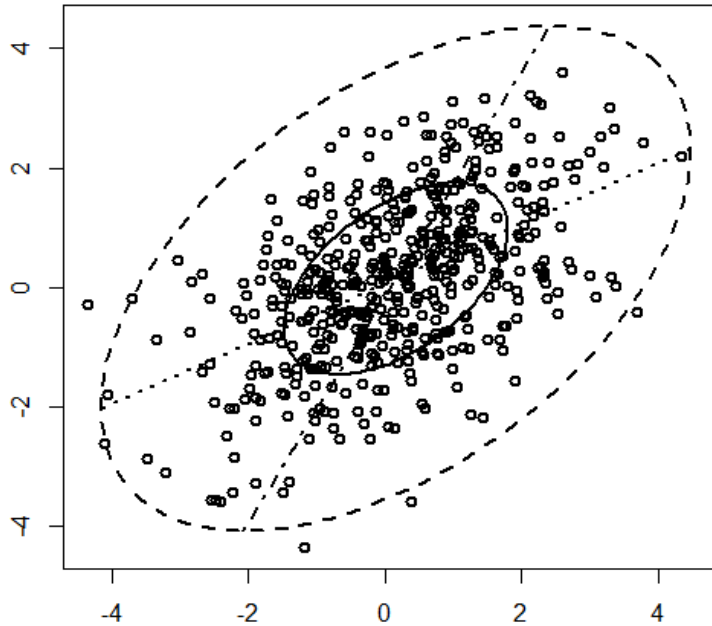
$$c^2 = \chi_p^2(\alpha)$$

Resultado: (Johnson and Wichern, 2008). Sob Normalidade dos dados, para $(n-p)$ suficientemente grande,

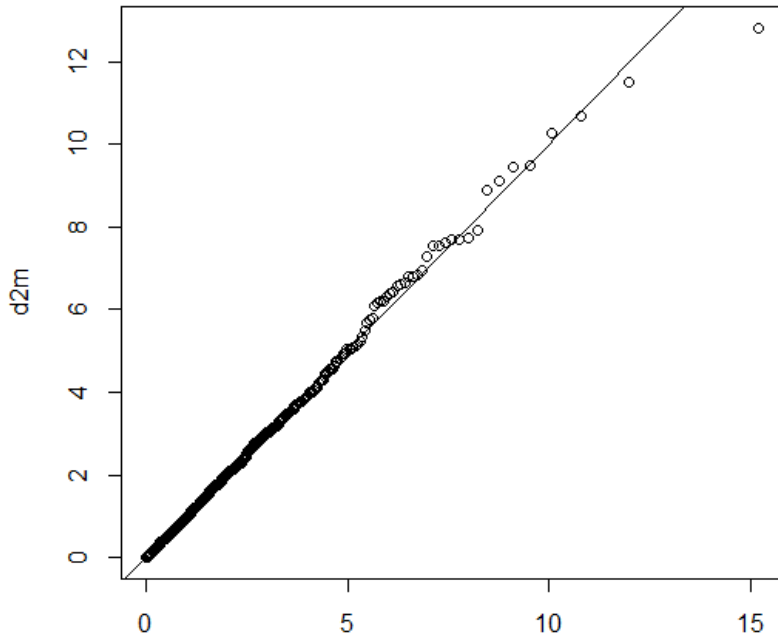
$$(Y_i - \bar{Y})' S_u^{-1} (Y_i - \bar{Y}) \sim \chi_p^2$$

- Critério de diagnóstico de observações atípicas (multivariado)
- Critério útil para averiguar a hipótese de Normalidade dos dados via o Gráfico Chi-Quadrado (Q-Q Plot das Observações).

Distâncias de Mahalanobis



```
library(MASS)
mu<-c(0,0)
sigma<-matrix(c(2,1,1,2),ncol=2)
n<-500
y<-mvrnorm(n,mu,sigma)
mi<-colMeans(y)
s<-cov(y)
par(mfrow=c(1,2))
bivbox(y, method="O")
# Copy Everitt's bivbox function
d2m<-mahalanobis(y,mi,s)
quantis <- qchisq(ppoints(length(y)),df=2)
qqplot(quantis, d2m)
abline(0,1)
```



Aplicação de

Regiões de Confiança – Caso de Duas Populações Amostradas Pareadas

Amostra Pareada \Rightarrow respostas multivariadas são avaliadas na mesma unidade amostral em “duas” condições diferentes (Ex.: Antes e Depois de uma intervenção)

Duas Populações

$$Y_{1n \times p}; Y_{1i p \times 1} = (Y_{1i1}, Y_{1i2}, \dots, Y_{1ip})' \quad Y_{2n \times p}; Y_{2i p \times 1} = (Y_{2i1}, Y_{2i2}, \dots, Y_{2ip})' \quad i = 1, 2, \dots, n$$

$$Y_{1i p \times 1} \stackrel{iid}{\sim} N_p(\mu_1; \Sigma_1)$$

$$Y_{2i p \times 1} \stackrel{iid}{\sim} N_p(\mu_2; \Sigma_2)$$



$$D_{ij} = Y_{1ij} - Y_{2ij} \quad j = 1, 2, \dots, p, \quad i = 1, 2, \dots, n$$

$$D_{i p \times 1} = (D_{i1}, D_{i2}, \dots, D_{ip})' \stackrel{iid}{\sim} N_p(\mu_D = \mu_1 - \mu_2; \Sigma_D) \quad i = 1, 2, \dots, n$$

Observações Pareadas \Rightarrow Uma Única População de Diferenças



$$T^2 = n (\bar{D} - \mu_D)' \underset{\uparrow S_D}{S_D^{-1}} (\bar{D} - \mu_D) \sim \frac{(n-1)p}{(n-p)} F_{p, n-p}$$

Elipse de Confiança:

$$R(Y_1, Y_2) = \left\{ \mu_D \in \mathfrak{R}^p; n (\bar{D} - \mu_D)' S_D^{-1} (\bar{D} - \mu_D) \leq c_\alpha^2 \right\}$$

Veremos
mais adiante!

Regiões de Confiança – Duas Populações

Caso Multivariado - Amostras Independentes - Homocedasticidade:


$$Y_{1n_1 \times p}; Y_{1i} \stackrel{iid}{\sim} N_p(\mu_1; \Sigma_1); \quad Y_{2n_2 \times p}; Y_{2i} \stackrel{iid}{\sim} N_p(\mu_2; \Sigma_2); \quad \Sigma_1 = \Sigma_2 = \Sigma$$

$$\Rightarrow \bar{D}_{p \times 1} = \bar{Y}_1 - \bar{Y}_2 \sim N_p \left(\mu_D = \mu_1 - \mu_2; \Sigma \left(\frac{1}{n_1} + \frac{1}{n_2} \right) \right)$$

Matriz de covariâncias
comum aos grupos
(com denominador n-1)

$$S_c = \frac{\sum_{i=1}^{n_1} (Y_{1i} - \bar{Y}_1)(Y_{1i} - \bar{Y}_1)' + \sum_{i=1}^{n_2} (Y_{2i} - \bar{Y}_2)(Y_{2i} - \bar{Y}_2)'}{n_1 + n_2 - 2} = \frac{(n_1 - 1)S_{u1} + (n_2 - 1)S_{u2}}{n_1 + n_2 - 2};$$

$$T^2 = (\bar{D} - \mu_D)' \left[\left(\frac{1}{n_1} + \frac{1}{n_2} \right) S_c \right]^{-1} (\bar{D} - \mu_D) \sim \frac{(n_1 + n_2 - 2)p}{(n_1 + n_2 - p - 1)} F_{(p, (n_1 + n_2 - p - 1))}$$

 **Elipse de Confiança:** $R(Y_1, Y_2) = \left\{ \mu_D \in \mathcal{R}^2; (\bar{D} - \mu_D)' \left[\left(\frac{1}{n_1} + \frac{1}{n_2} \right) S_c \right]^{-1} (\bar{D} - \mu_D) \leq c_\alpha^2 \right\}$