

MAE 5776

# ANÁLISE MULTIVARIADA

Júlia M Pavan Soler

[pavan@ime.usp.br](mailto:pavan@ime.usp.br)

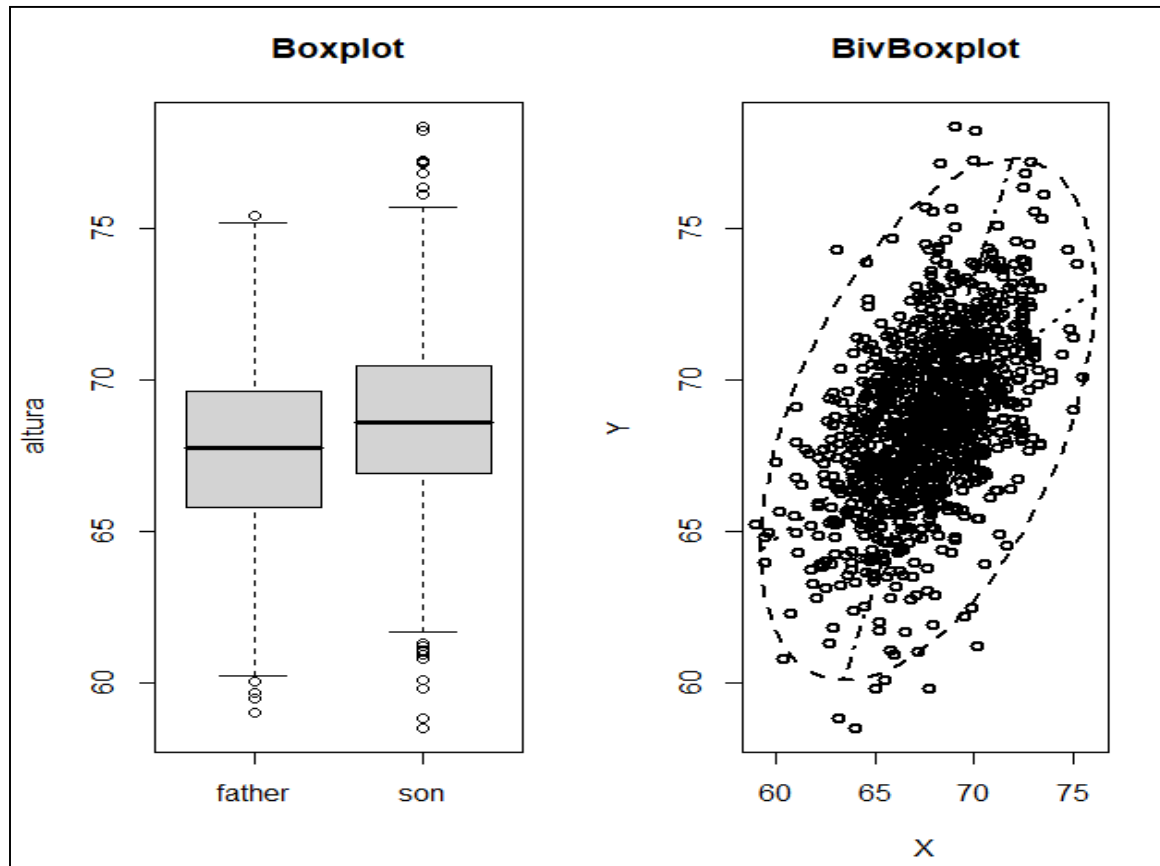
1º Semestre IME/2022

# MAE 5776 – Análise Multivariada

1. Introdução: estrutura de dados, medidas resumo multivariadas, propriedades em espaços duais, elipses de concentração de dados, *outliers* multivariados.
2. Distribuição Normal Multivariada: propriedades, estimação, distribuições amostrais, testes de hipóteses para vetores de médias e matrizes de covariância. Regiões (elipsoides) de confiança para vetores de Médias.
3. Técnicas (clássicas) de redução da dimensionalidade ( $n > p$  e observações independentes). Teoria de Fatoração de Matrizes na redução de dimensionalidade e integração de bancos de dados.
4. Técnicas de redução da dimensionalidade em espaços mais gerais ( $n \ll p$ ,  $n$  muito grande, observações não independentes): soluções em espaços duais, soluções regularizadas e penalizadas, reamostragem.
5. Temas adicionais: Modelos de Equações Estruturais; Teoria de Grafos Probabilísticos; Análise de Dados heterogêneos.

# Distribuição Normal

## Caso Univariado → Multivariado

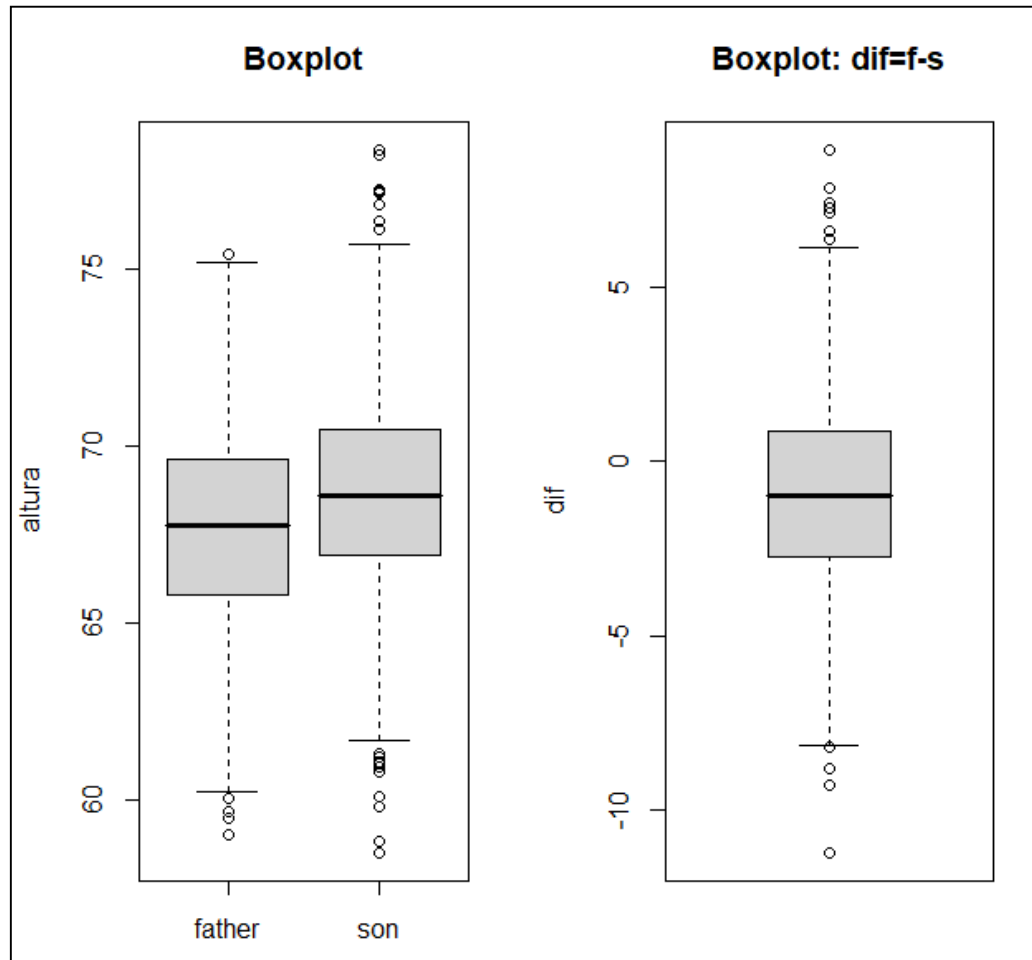


```
li<-qnorm(0.25)-1.5*(qnorm(0.75)-qnorm(0.25))  #-2.697959
ls<-qnorm(0.75)+1.5*(qnorm(0.75)-qnorm(0.25))  # 2.697959
pli<-pnorm(li)    # 0.003488302
pls<-1-pnorm(ls)  # 0.003488302
```

Critério de  
outlier sob  
 $N(0,1)$

# Comparações de Médias

## Caso Univariado



$$\begin{cases} H_0 : \mu_{son} = \mu_{father} \\ H_1 : \mu_{son} \neq \mu_{father} \end{cases}$$
$$\Leftrightarrow \begin{cases} H_0 : \mu_{dif} = 0 \\ H_1 : \mu_{dif} \neq 0 \end{cases}$$

**Teste t: amostras pareadas**

**variável:** dif

t = -11.789, df = 1077,

p-value < 2.2e-16

alternative hypothesis: true mean is not equal to 0

95 percent confidence interval:

-1.1629160 -0.8310296

sample estimates:

mean of x

-0.9969728

**Conclusão:** Há diferença significativa entre as médias da altura dos filhos e pais  $\mu_{son} > \mu_{father}$

# Comparação de Populações

## Comparação de Vetores de Médias

### Caso Multivariado - Motivação

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

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

$$Y_{i3 \times 1} \sim (\mu_{3 \times 1}; \Sigma_{3 \times 3}), \quad i = 1, \dots, 20$$

$$\mu_0 = (4, 50, 10)'$$

$$\begin{cases} H_0 : \mu = \mu_0 = (4, 50, 10)' \\ H_1 : \mu \neq \mu_0 \end{cases}$$

Obter uma região de 95% de confiança para  $\mu$ !

# Comparação de Populações

## Comparação de Vetores de Médias

### Caso Multivariado - Motivação

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

$$Y_{i \ 4 \times 1} \sim (\mu_{4 \times 1}; \Sigma_{4 \times 4}), i = 1, \dots, 25;$$

$$Y_{iF1 \ 2 \times 1} \sim (\mu_{F1 \ 2 \times 1}; \Sigma_{F1 \ 2 \times 2}),$$

$$Y_{iF2 \ 2 \times 1} \sim (\mu_{F2 \ 2 \times 1}; \Sigma_{F2 \ 2 \times 2})$$

$$\begin{cases} H_0 : \mu_{F1} = \mu_{F2} \\ H_1 : \mu_{F1} \neq \mu_{F2} \end{cases}$$

**Amostras pareadas!**

$$\Leftrightarrow \begin{cases} H_0 : \mu_{Dif} = 0_{2 \times 1} \\ H_1 : \mu_{Dif} \neq 0_{2 \times 1} \end{cases}$$

Obter uma região de 95% de confiança para  $\mu_{Dif}$ !

# Comparação de Populações

## Comparação de Vetores de Médias

### Caso Multivariado - Motivação

Medidas biométricas (mm) de Pardais fêmea (Manly, 2005).

Pardal	Sobrev.	X1	X2	X3	X4	X5
1	S	156	245	31.6	18.5	20.5
...	...					
21	S	159	236	31.5	18.0	21.5
22	N	155	240	31.4	18.0	20.7
...	...					
49	N	164	248	32.3	18.8	20.9

$$Y_{iS \ 5 \times 1} \sim (\mu_{S \ 5 \times 1}; \Sigma_{S \ 5 \times 5}), \ i = 1, \dots, 21; \quad Y_{iN \ 5 \times 1} \sim (\mu_{N \ 5 \times 1}; \Sigma_{N \ 5 \times 5}); \ i = 1, \dots, 28$$

$$\begin{cases} H_0 : \mu_S = \mu_N \\ H_1 : \mu_S \neq \mu_N \end{cases}$$

Amostras  
independentes!

# Resultados Inferenciais

- Qual é a 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(s) Centróide(s):

⇒ Caso de uma única População  $H_0 : \mu_{p \times 1} = \mu_0$

⇒ Caso de Duas Populações (Pareadas e Independentes)  $H_0 : \mu_{1p \times 1} = \mu_{2p \times 1}$

⇒ Caso de Duas ou Mais Populações  $H_0 : \mu_{1p \times 1} = \mu_{2p \times 1} = \dots = \mu_{Gp \times 1}$



# Distribuição Amostral de Estimadores

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

Faça um paralelo destes resultados para o caso unariado,  $N_1$

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



$$\Rightarrow \bar{Y} \sim N_p\left(\mu_{p \times 1}; \frac{1}{n} \Sigma\right) \quad N_1\left(\mu; \frac{\sigma^2}{n}\right)$$

Ganho em precisão

$\bar{Y}$  e S são independentes

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

Distribuição de Wishart

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

$$(n-1)s_u^2 \sim W_1(n-1; \sigma^2) = \chi_{n-1}^2$$

# Distâncias de Mahalanobis

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

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

Lembre que ( $p=1$ ):

$$t_o = \frac{\bar{Y} - \mu}{s / \sqrt{n}} \sim t_{(n-1)}$$



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

$$\bar{Y} \sim N_p(\mu; \Sigma / n); \quad \sqrt{n}(\bar{Y} - \mu) \sim N_p(0; \Sigma)$$

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

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



**Distância generalizada**

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

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

Teorema 3.5.2  
(Mardia et al., 2003)

# Simulação de Dados

Exercício de fixação de resultados!

Seja  $Y_{n \times p} = (Y_1, \dots, Y_n)'$ ,  $Y_i \in \mathbb{R}^p$ ,  $i = 1, 2, \dots, n$ , uma amostra aleatória simples de tamanho  $n$  da distribuição  $N_p(\mu; \Sigma)$ ,  $\Sigma = V\Lambda V'$ . Usando simulação, adote valores para  $n$ ,  $\mu$ ,  $\Sigma$  e mostre (por meio de gráficos de probabilidades) que:

a)  $\bar{Y} \in \mathbb{R}^p \sim N_p(\mu; \Sigma/n)$

b)  $nS \sim W_p(n-1; \Sigma)$ ;  $S = n^{-1}Y'HY$ ,  $H = I_p - n^{-1}\mathbf{1}_p\mathbf{1}_p'$

c)  $d_{Mi}^2 = (Y_i - \mu)' \Sigma^{-1} (Y_i - \mu) \sim \chi_p^2$   $\Rightarrow$  Região de Concentração de observações (diagnóstico de outliers)

d)  $\hat{d}_M^2 = n(\bar{Y} - \mu)' S_u^{-1} (\bar{Y} - \mu) \sim T_{(p;n)}^2 = \frac{(n-1)p}{n-p} F_{(p;n-p)}$   $\Rightarrow$  Região de Confiança para o parâmetro  $\mu$

e)  $a'(nS)a / a'\Sigma a \sim \chi_{(n-1)}^2$ ,  $a \in \mathbb{R}^p$ .


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

## Inferências sobre o Vetor de Médias $\mu$

$$Y_{i \times p} \stackrel{iid}{\sim} AAS_n \quad N_p(\mu; \Sigma); \quad |\Sigma| > 0$$

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  $\mu \in \mathbb{R}^p$ , com base na amostra, tal que:


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

**Região (elipsoide) de confiança** para inferências sobre  $\mu$  com base na evidência amostral ( $\bar{Y}$ )

**Para um nível de significância  $\alpha$  fixado,  $c$  é obtido da distribuição  $F$**

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

Logo, obtém-se:  $P(\mu \in R(\mu|Y)) = 1 - \alpha$

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

## Inferências sobre o vetor $\mu$

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

$\Rightarrow$  Para determinar se algum ponto  $\mu_0$  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  $\alpha$ , isto é,

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

$\Rightarrow$  **Regiões de Confiança** correspondem a **Regiões de Aceitação** em testes de hipóteses sobre o vetor  $\mu$ .

# Inferência sobre um Vetor de Médias

## Casos Uni (Estatística t e t<sup>2</sup>) e Multivariado (Estatística T<sup>2</sup> de Hotteling)

$$t_j^2 = \frac{(\bar{Y}_j - \mu_j)^2}{s_j^2/n} = n \underbrace{(\bar{Y}_j - \mu_j)(s_j^2)^{-1}(\bar{Y}_j - \mu_j)} \sim t_{(n-1)}^2 = F_{1,(n-1)}$$

$$\Rightarrow t_j^2 = \frac{(\bar{Y}_j - \mu_{j0})^2}{s_j^2/n} \leq t_{(n-1)}^2 = F_{1,(n-1)}(\alpha)$$

Pode ser calculada para cada variável Y<sub>j</sub> (j=1,...,p).  
Mas, qual é o nível de significância coletivo?

Procedimento conjunto para as p variáveis

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

$$T^2 = n(\bar{Y} - \mu_0)' S^{-1} (\bar{Y} - \mu_0) \leq \frac{(n-1)p}{(n-p)} F_{p,(n-p)}(\alpha)$$

Pesquise **Intervalos de Confiança Simultâneos** para Componentes do Vetor de Médias Populacional!!

# Inferência sobre um Vetor de Médias

## Uma Única População - Caso Multivariado

$$Y_{i \times p} = (Y_{1i}, Y_{2i}, \dots, Y_{pi})' \stackrel{iid}{\sim} N_p(\mu; \Sigma) \quad i = 1, 2, \dots, n$$

### Corrigindo para os múltiplos “testes”:

$$H_0 : \mu_{p \times 1} = \mu_0 \quad \Rightarrow \quad T^2 = n (\bar{Y} - \mu_0)' S^{-1} (\bar{Y} - \mu_0) \sim \frac{(n-1)p}{(n-p)} F_{p, n-p}$$

$$\Rightarrow ICS(\mu_j) \text{ a } 100(1-\alpha)\% = \bar{Y}_j \pm \sqrt{\frac{(n-1)p}{(n-p)} F_{p, (n-p)}(\alpha)} \sqrt{\frac{s_{jj}}{n}} \quad \text{Intervalo de Confiança Simultâneo}$$

$$\Rightarrow ICB(\mu_j) \text{ a } 100(1-\alpha)\% = \bar{Y}_j \pm t_{n-1}(\alpha/2m) \sqrt{\frac{s_{jj}}{n}} \quad \text{Intervalo de Confiança com correção de Bonferroni (para m intervalos)}$$

$$\Rightarrow IC(\mu_j) \text{ a } 100(1-\alpha)\% = \bar{Y}_j \pm t_{n-1}(\alpha/2) \sqrt{\frac{s_{jj}}{n}} \quad \text{IC formulado via a estatística t (não há correção para os múltiplos testes)}$$

# Regiões de Confiança e Testes de Hipóteses

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

Vamos construir a Região de Confiança:

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

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

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



Suponha o interesse na seguinte hipótese:

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

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

Calcule o  
valor-p!

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

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

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



# Regiões de Confiança para o Vetor $\mu$

## 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 $\mu$ 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

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

**Distribuição marginal para a variável Comprimento ( $\mathcal{R}^2$ ):**

$$Y_{25 \times 2}; Y_{i_{2 \times 1}} = (Y_{i1}, Y_{i3}) \stackrel{iid}{\sim} N_2 \left( \mu = \begin{pmatrix} \mu_1 \\ \mu_3 \end{pmatrix}; \Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{13} \\ \sigma_{13} & \sigma_{33} \end{pmatrix} \right)$$



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

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

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

$IC(\mu_j)$  a  $100(1-\alpha)\%$

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

$$t^2 = \frac{(\bar{Y}_j - \mu_j)^2}{s_{jj}/n}$$

$$= n(\bar{Y}_j - \mu_j)(s_{jj})^{-1}(\bar{Y}_j - \mu_j)$$

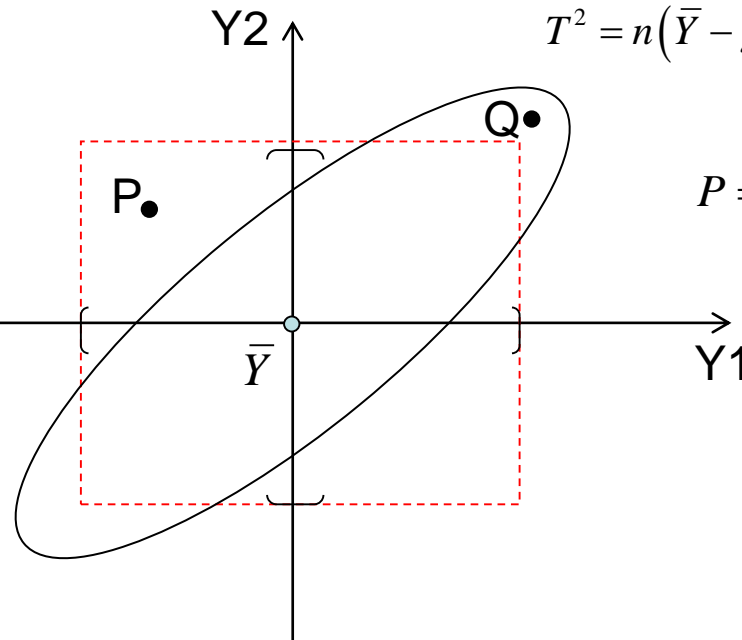
$$t_{(n-1)}^2 = F_{1,(n-1)}$$

## 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  $\mu$ .

⇒ 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

Revise:

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

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

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)

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

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

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

Região de  
Confiança para  $\mu$   
Elipsóide de Confiança



$$T^2(p; n-1) = \frac{(n-1)p}{n-p} F_{(p; n-p)} = \chi_p^2$$

# Distâncias de Mahalanobis

Revise:


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

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

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

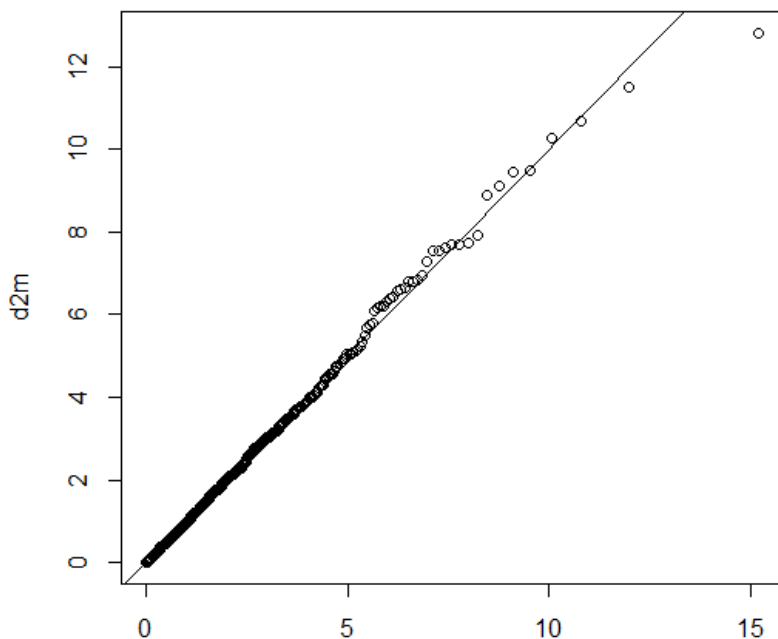
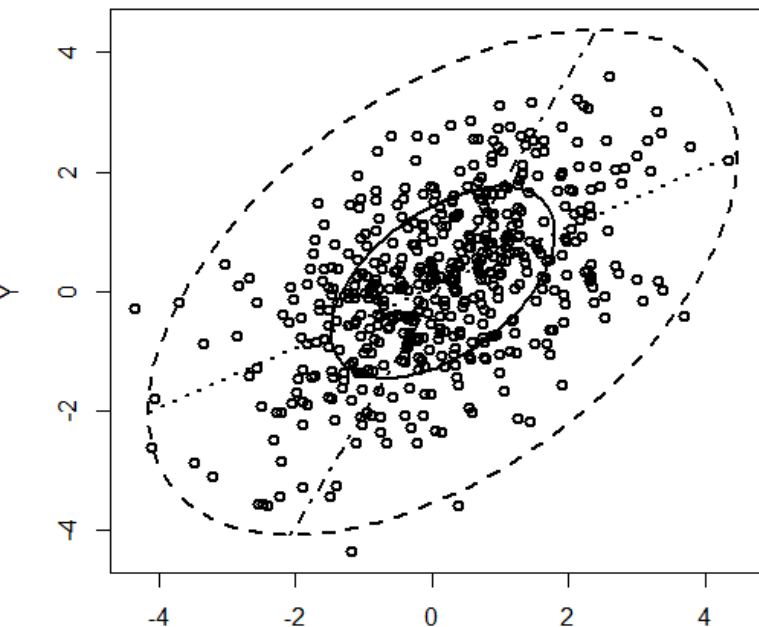
**Resultado:** (Johnson and Wichern, 2008). Sob Normalidade dos dados, para  $(n-p)$  suficientemente grande,

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

- 
- 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

Já vimos!



```
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)
```

# Inferência sobre Vetores de Médias de Duas Populações

**Recordando!** Caso Univariado ( $\rightarrow$  Generalizar os resultados para o Multivariado)

$$H_0 : \mu_1 = \mu_2 \Leftrightarrow H_0 : \mu_D = 0$$

Observações Pareadas:  $(Y_{1i} ; Y_{2i}) \quad i = 1, 2, \dots, n$

$$\Rightarrow D_i = Y_{1i} - Y_{2i} \sim N(\mu_D = \mu_1 - \mu_2; \sigma_D^2), \quad i = 1, 2, \dots, n$$

$$\Rightarrow IC(\mu_D) \text{ a } (1-\alpha)100\% = [\bar{D} - e; \bar{D} + e]; \quad e = t_{(n-1)}(1-\alpha/2) \frac{s_D}{\sqrt{n}}$$

$$\Rightarrow H_0 : \mu_D = 0 \Rightarrow t = \frac{\bar{D}}{s_D / \sqrt{n}} \sim t_{n-1}; \Rightarrow t^2 = n\bar{D} (s_D^2)^{-1} \bar{D} \sim t_{n-1}^2 = F_{1,n-1}$$

Observações Independentes:  $Y_{11}, Y_{12}, \dots, Y_{1n_1} \sim N_1(\mu_1; \sigma_1^2); \quad Y_{21}, Y_{22}, \dots, Y_{2n_2} \sim N_1(\mu_2; \sigma_2^2)$

$$\Rightarrow \bar{D} = \bar{Y}_1 - \bar{Y}_2 \sim N\left(\mu_D = \mu_1 - \mu_2; \sigma_D^2 = \sigma^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right); \quad \sigma_1^2 = \sigma_2^2 = \sigma^2 \quad \text{Sob homocedasticidade!}$$

$$\Rightarrow IC(\mu_D) \text{ a } (1-\alpha)100\% = [\bar{D} - e; \bar{D} + e]; \quad e = t_{(n_1+n_2-2)}(1-\alpha/2) s_c \left(\frac{1}{n_1} + \frac{1}{n_2}\right); \quad s_c^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1 + n_2 - 2}$$

$$\Rightarrow H_0 : \mu_D = 0, \text{ sob } \sigma_1^2 = \sigma_2^2 \Rightarrow t = \frac{(\bar{Y}_1 - \bar{Y}_2)}{s_c \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2}; \quad t^2 = \left(\frac{1}{n_1} + \frac{1}{n_2}\right)^{-1} (\bar{Y}_1 - \bar{Y}_2) (s_c^2)^{-1} (\bar{Y}_1 - \bar{Y}_2) \sim F_{1; (n_1+n_2-2)}$$

# Inferência sobre Vetores de Médias de Duas Populações

Recordando!

Caso Univariado:

Observações Pareadas: Dados dos Filhos, Variável Comprimento (n=25 observações)

Dados das Diferenças: Média de Dif= 1,88; Desvio padrão de Dif=7,53

IC95% para  $\mu$ -Dif. = [-1.230307 ; 4.990307]

Teste t:  $H_0: \mu \text{Dif} = 0 \Rightarrow t = 1.2475, df = 24, p\text{-value} = 0.2242$

**Conclusão:** Não há evidência para diferenças entre as médias dos dois grupos

Observações Independentes:

Dados dos Pardais (X1)

G1:  $n_1=21$   $\mu_1= 157.38$   $s_1^2= 11,05$

G2:  $n_2=28$   $\mu_2= 158.43$   $s_2^2= 15.07$

$\mu_d=-0.75$   $s_c^2=17,83$

IC95% para a diferença entre os grupos= [-3.170113 ; 1.074874]

Teste "t":  $H_0: \mu_1 = \mu_2$  (sob homocedasticidade)

$t = -0.99295, df = 47, p\text{-value} = 0.3258$

**Conclusão:** Não há evidência para diferenças entre as médias dos grupos de Pardais para a variável X1 (Sobreviventes e Não Sobreviventes)



# Inferência sobre Vetores de Médias de Duas Populações

## Caso Multivariado

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

Duas Populações

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

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

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

Uma Única População de Diferenças

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

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

Elipsoide de Confiança:  $R(Y_1, Y_2) = \left\{ \delta = \mu_1 - \mu_2 \in \mathbb{R}^p; n (\bar{D} - \delta)' S_D^{-1} (\bar{D} - \delta) \leq c_\alpha^2 \right\}$

$$H_0 : \delta = \delta_0 \quad \Rightarrow \quad T^2 = n (\bar{D} - \delta_0)' S_D^{-1} (\bar{D} - \delta_0) \sim \frac{(n-1)p}{(n-p)} F_{p, n-p}$$

# Inferência sobre Vetores de Médias de Duas Populações

## Caso Multivariado – Amostra Pareada

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

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

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

$$H_0 : \delta = \delta_0 \quad \Rightarrow \quad T^2 = n (\bar{D} - \delta_0)' S_D^{-1} (\bar{D} - \delta_0) \sim \frac{(n-1)p}{(n-p)} F_{p, n-p}$$

$$\Rightarrow ICS(\delta_j) \text{ a } 100(1-\alpha)\% = \bar{D}_j \pm \sqrt{\frac{(n-1)p}{(n-p)} F_{p, (n-p)}(\alpha)} \sqrt{\frac{S_{D_{jj}}}{n}} \quad \text{Intervalo de Confiança Simultâneo}$$

$$\Rightarrow ICB(\delta_j) \text{ a } 100(1-\alpha)\% = \bar{D}_j \pm t_{n-1}(\alpha / 2m) \sqrt{\frac{S_{D_{jj}}}{n}} \quad \text{Intervalo de Confiança com correção de Bonferroni (para m “testes”)}$$

# Inferência para Vetores de Médias

## Duas Populações Pareadas

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

$$Y_{ig_{2 \times 1}} \stackrel{iid}{\sim} N_2(\mu_g; \Sigma_g); g = 1, 2$$

$$D_i = Y_{iFilho1} - Y_{iFilho2} \stackrel{iid}{\sim} N_2(\mu_D; \Sigma_D)$$

$$\begin{aligned} \text{mu1} &= (185.72 \quad 151.12)' & \text{mud} &= (1.88 \quad 1.88)' \\ \text{mu2} &= (183.84 \quad 149.24)' \end{aligned}$$

$S_D$	Co	Pe
Co	56.78	11.98
Pe	11.98	29.28

7.14

$$H_0: \mu_D = 0 \Rightarrow T^2 = 3,61 \sim \frac{24 \cdot 2}{23} F_{2,23} \quad \alpha=5\% \Rightarrow 3.42$$

Conclusão: Não há evidência de diferença significativa entre as médias das medidas cefálicas dos dois grupos (1° e 2° Filhos)

⇒ Obtenha os ICS e ICB!

Qual é a estrutura dos dados?

# Inferência sobre Vetores de Médias de Duas Populações

## Caso Multivariado - Amostras Independentes - Homocedasticidade

$$Y_{1n_1 \times p} = \underbrace{(Y_{11}, Y_{12}, \dots, Y_{1n_1})'}_{\bar{Y}_1}, \quad Y_{1i} \stackrel{iid}{\sim} N_p(\mu_1; \Sigma_1); \quad Y_{2n_2 \times p} = \underbrace{(Y_{21}, Y_{22}, \dots, Y_{2n_2})'}_{\bar{Y}_2}, \quad Y_{2i} \stackrel{iid}{\sim} N_p(\mu_2; \Sigma_2)$$

$$S_1 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (Y_{1i} - \bar{Y}_1)(Y_{1i} - \bar{Y}_1)'$$

$$S_2 = \frac{1}{n_2 - 1} \sum_{i=1}^{n_2} (Y_{2i} - \bar{Y}_2)(Y_{2i} - \bar{Y}_2)'$$

$Y_1 \perp Y_2$   
 $\mu_D = \delta = \mu_1 - \mu_2$



$$\bar{D}_{p \times 1} = \bar{Y}_1 - \bar{Y}_2 \sim N_p\left(\mu_D; \Sigma_{\bar{D}} = \frac{\Sigma_1}{n_1} + \frac{\Sigma_2}{n_2}\right)$$

$$H_0 : \mu_D = 0; \quad \times \quad H_1 : \mu_D \neq 0; \quad \Sigma_1 = \Sigma_2 = \Sigma$$

Hipótese  
condicionada à  
suposição de  
homocedasticidade

$$\Rightarrow \bar{D}_{p \times 1} \sim N_p\left(\mu_D; \Sigma_{\bar{D}} = \Sigma \left( \frac{1}{n_1} + \frac{1}{n_2} \right)\right)$$

# Inferência sobre Vetores de Médias de Duas Populações

## Caso Multivariado - Amostras Independentes - Homocedasticidade

$$Y_{1n_1 \times p} = (Y_{11}, Y_{12}, \dots, Y_{1n_1})'; \quad Y_{1i} \stackrel{iid}{\sim} N_p(\mu_1; \Sigma_1); \quad Y_{2n_2 \times p} = (Y_{21}, Y_{22}, \dots, Y_{2n_2})'; \quad Y_{2i} \stackrel{iid}{\sim} N_p(\mu_2; \Sigma_2)$$

$$\uparrow \quad Y_1 \perp Y_2; \quad \mu_D = \mu_1 - \mu_2 \quad \uparrow$$

$$H_0 : \mu_D = 0; \quad \times \quad H_1 : \mu_D \neq 0; \quad \Sigma_1 = \Sigma_2 = \Sigma$$

Hipótese condicional sob Homocedasticidade



$$\Rightarrow \bar{D}_{p \times 1} \sim N_p \left( \mu_D; \Sigma \left( \frac{1}{n_1} + \frac{1}{n_2} \right) \right)$$

$$\hat{\Sigma}^{H_0} = S_c$$

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

# Inferência sobre Vetores de Médias de Duas Populações

## Caso Multivariado - Amostras Independentes - Homocedasticidade:

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

$$\Rightarrow \bar{D}_{p \times 1} \sim N_p\left(\mu_D; \Sigma \left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right)$$

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

Matriz de covariâncias comum aos grupos

## Estatística de Hotelling:

$$T^2 = (\bar{D} - \delta_0)' \left[ \left( \frac{1}{n_1} + \frac{1}{n_2} \right) S_c \right]^{-1} (\bar{D} - \delta_0) \sim \frac{(n_1 + n_2 - 2)p}{(n_1 + n_2 - p - 1)} F_{(p, (n_1 + n_2 - p - 1))}$$

Elipsóide de Confiança:  $R(Y_1, Y_2) = \left\{ \delta = \mu_1 - \mu_2 \in \mathbb{R}^2; (\bar{D} - \delta)' \left[ \left( \frac{1}{n_1} + \frac{1}{n_2} \right) S_c \right]^{-1} (\bar{D} - \delta) \leq c_\alpha^2 \right\}$

# Inferência sobre Vetores de Médias de Duas Populações

Amostras Independentes:

$$Y_1 = (\underbrace{Y_{11}, Y_{12}, \dots, Y_{1n_1}}_{\bar{Y}_1 \quad S_1})'; \quad Y_{1i} \stackrel{iid}{\sim} N_p(\mu_1; \Sigma_1) \quad Y_2 = (\underbrace{Y_{21}, Y_{22}, \dots, Y_{2n_2}}_{\bar{Y}_2 \quad S_2})'; \quad Y_{2i} \stackrel{iid}{\sim} N_p(\mu_2; \Sigma_2) \quad S_c$$

Intervalos de Confiança Simultâneos (para combinações lineares das p variáveis)

$$l'(\bar{Y}_1 - \bar{Y}_2) \pm \sqrt{\frac{(n_1 + n_2 - 2)p}{(n_1 + n_2 - p - 1)} F_{(p, (n_1 + n_2 - p - 1))}} \sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right) l' S_c l}$$

Intervalos de Confiança de Bonferroni (correção para múltiplos testes)

$$(\bar{Y}_{1j} - \bar{Y}_{2j}) \pm t_{(n_1 + n_2 - 2)}(\alpha / 2m) \sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right) S_{jj}}$$

# Inferência sobre Vetores de Médias de Duas Populações Independentes

## Dados dos Pardais

bird	grup	x1	x2	x3	x4	x5
1	0	156	245	31.6	18.5	20.5
2	0	154	240	30.4	17.9	19.6
3	0	153	240	31.0	18.4	20.6
...						
19	0	155	236	30.3	18.5	20.1
20	0	163	246	32.5	18.6	21.9
21	0	159	236	31.5	18.0	21.5
22	1	155	240	31.4	18.0	20.7
23	1	156	240	31.5	18.2	20.6
24	1	160	242	32.6	18.8	21.7
...						
40	1	163	249	33.4	19.5	22.8
41	1	163	242	31.0	18.1	20.7
42	1	156	237	31.7	18.2	20.3

	x1	x2	x3	x4	x5
Mi.0=	157.38	241.00	31.43	18.50	20.81
Mi.1=	158.43	241.57	31.48	18.45	20.84
mud=	-1.05	-0.57	-0.05	0.05	-0.03
n1=	21				
n2=		28			

Sc		x1	x2	x3	x4	x5
x1	12.90	12.52	1.81	1.46	2.42	
x2	12.52	27.18	2.61	2.32	2.35	
x3	1.81	2.61	0.65	0.39	0.40	
x4	1.46	2.32	0.39	0.37	0.35	
x5	2.42	2.35	0.40	0.35	1.06	

$$T^2 = 2.82 \sim$$

$$\frac{(21+28-2)5}{(21+28-5-1)} F_{(5,(21+28-5-1))}^{\alpha=5\% \Rightarrow 2.42} = 13.23$$

Que inferência está sendo feita?  
 Qual é a hipótese de interesse?  
 Quais são as suposições adotadas?  
 Conclusão?  
 Obtenha ICS e ICB!



# Inferência sobre Vetores de Médias de “Muitas” Populações

Veremos:

Comparações de Duas Populações  $\Rightarrow$  Comparações de Muitas Populações

## MANOVA



População 1

$$N_p(\boldsymbol{\mu}_1; \Sigma_1)$$

População 2

$$N_p(\boldsymbol{\mu}_2; \Sigma_2)$$

...

População g

$$N_p(\boldsymbol{\mu}_g; \Sigma_g)$$

Amostra

$$\begin{array}{c} \downarrow \\ \mathbf{Y}_{11}, \mathbf{Y}_{12}, \dots, \mathbf{Y}_{1n_1} \\ \underbrace{\hspace{10em}} \\ \bar{\mathbf{Y}}_1 \quad S_1 \end{array}$$

$$\begin{array}{c} \downarrow \\ \mathbf{Y}_{21}, \mathbf{Y}_{22}, \dots, \mathbf{Y}_{2n_1} \\ \underbrace{\hspace{10em}} \\ \bar{\mathbf{Y}}_2 \quad S_2 \end{array}$$

$$\begin{array}{c} \downarrow \\ \mathbf{Y}_{g1}, \mathbf{Y}_{g2}, \dots, \mathbf{Y}_{gn_1} \\ \underbrace{\hspace{10em}} \\ \bar{\mathbf{Y}}_g \quad S_g \end{array}$$

$$\Rightarrow H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \dots = \boldsymbol{\mu}_g = \boldsymbol{\mu} \quad ; \Sigma_1 = \Sigma_2 = \dots = \Sigma_g = \Sigma$$