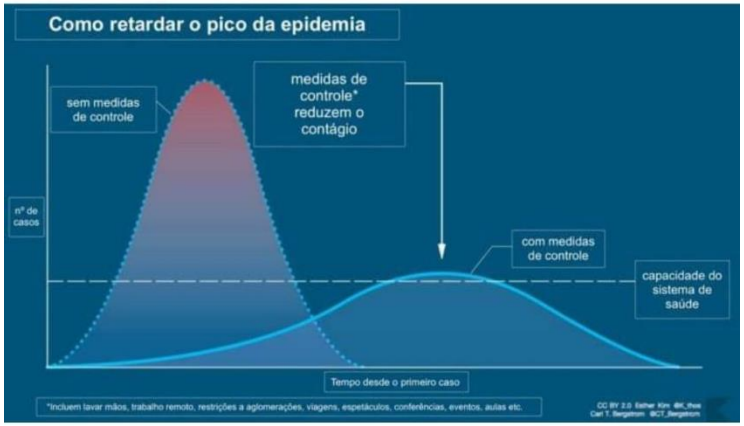
MAE 5776

ANÁLISE MULTIVARIADA



A Curva de Contágio e suas consequências no Sistema de Saúde e em nosso Dia a Dia.

Júlia M Pavan Soler

pavan@ime.usp.br

1º Semestre/2020

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

MAE5776

Já vimos ☺

Matriz de Dados: Estatísticas descritivas multivariadas

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

- 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 ($Y_i \in \Re^p$):

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

Matriz Aleatória: $Y_{n\times p} \sim N_{n\times p}\left(1_n \otimes \mu'_{p\times 1};\ \Omega_{np\times np} = \Psi_{n\times n} \otimes \Sigma_{p\times p}\right)$

- Estimadores e Distribuições Amostrais: $Y_{i_{p\times 1}} \overset{iid}{\sim} N_p(\mu;\Sigma)$ $\Rightarrow$ $\bar{Y}_{p\times 1} \sim N_p(\mu;\Sigma/n)$; $nS_{p\times p} \sim W_p(n-1;\Sigma)$

- Regiões (elipsoides) de Confiança para μ:

$$R(\mu \mid Y) = \left\{ \mu \in \Re^p;\ n\left(\bar{Y} - \mu\right)' S_u^{-1} \left(\bar{Y} - \mu\right) \le c^2; c^2 = T^2_{(p;n-1)}(\alpha) = \frac{(n-1)p}{(n-p)} F_{p,(n-p)}(\alpha) \right\}$$

Caso de Uma única População

Inferência – Análise Multivariada

Regiões de Confiança e Testes Multivariados	x	Intervalos de Confiança e Testes Univariados

$$Y_{i_{p\times 1}} \overset{iid}{\sim} (\mu;\Sigma), \quad H_0 : \mu_{p\times 1} = 0_{p\times 1} \qquad \times \qquad \begin{cases} H_{01} : \mu_1 = 0 \\ \quad \ldots \\ H_{0p} : \mu_p = 0 \end{cases}$$

- Há interesse na análise conjunta de múltiplas variáveis
- Realizar inferências mais “precisas” devido a incorporar a informação da covariância entre variáveis
- Realizar comparações entre os parâmetros associados às diferentes variáveis: construir contrastes entre médias das variáveis
- Construir níveis de significância coletivos × Correções para múltiplos testes

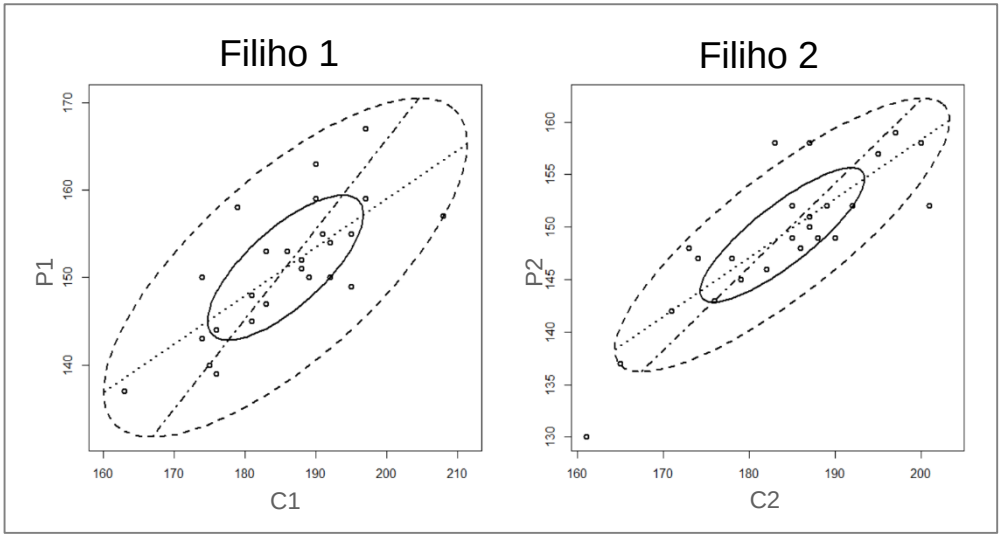
  Bonferroni, FDR

Inferência sobre para $\mu \in \Re^p$

Dados do Primeiro e Segundo filhos

	1° Filho		2° Filho	
Família	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

Boxplot Bivariado: Elipses de concentração de observações



Estatísticas Descritivas:

$$\bar{Y} = \begin{pmatrix} 185{,}72 & 151{,}12 & 183{,}84 & 149{,}24 \end{pmatrix}'$$

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

$$Y_{25\times 4} = \left(Y_1, \ldots, Y_{25}\right)' ; \; Y_{i_{4\times 1}} \overset{iid}{\sim} \left(\mu; \Sigma\right)$$

Região de Confiança e Teste para $\mu \in \Re^p$

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

	1° Filho		2° Filho	
Família	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

Pense na situação de Intervalos de Confiança e Testes de hipóteses individuais para cada variável!!

Caso 1: Há interesse em inferências para o vetor de médias μ

Região de Confiança (baseada na estatística T2 de Hotelling):

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

Nível de significância global

$$n = 25, \; p = 4, \; \alpha = 5\% \Rightarrow 12.98$$

$$\begin{cases} H_0 : \mu = (185,150,180,148)' \\ H_1 : \mu \neq (185,150,180,148)' \end{cases} \qquad T^2 = 6,44$$

Conclusão?

$$\begin{cases} H_0 : \mu = (185,150,180,145)' \\ H_1 : \mu \neq (185,150,180,145)' \end{cases} \qquad T^2 = 17,98$$

Região de Confiança e Teste para $\mu \in \Re^p$

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

	1° Filho		2° Filho	
Família	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

Caso Multivariado:

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

Nível de significância global

$$\alpha = 5\% \Rightarrow 12.98$$

$$\begin{cases} H_0 : \mu = (185,150,180,145)' \\ H_1 : \mu \neq (185,150,180,145)' \end{cases} \qquad T^2 = 17,98$$

Conclusão?
Qual é o α global (para os múltiplos testes)?

Casos Univariados:

$$t_j = \frac{(\bar{Y}_j - \mu_j)}{\sqrt{s_j / n}} \sim t_{n-1}(\alpha); \quad j = 1,2,3,4$$

$$\alpha = 5\% \Rightarrow 4,26$$

$$t_j^2 = n(\bar{Y}_j - \mu_j) s_j^{-1} (\bar{Y}_j - \mu_j) \sim F_{(1,n-1)}(\alpha); \quad j = 1,2,3,4$$

$$t_1^2 = 0.14; \quad t_2^2 = 0.58; \quad t_3^2 = 3.66; \quad t_4^2 = 9.98$$

Inferência sobre um Vetor de Médias

Correspondência entre as Estatísticas de Teste dos casos Uni e Multivariado

$$t^2 = \frac{(\bar{Y} - \mu)^2}{s^2/n} = \underbrace{n(\bar{Y} - \mu)(s^2)^{-1}(\bar{Y} - \mu)} \sim t^2_{(n-1)} = F_{1,(n-1)}$$

$$H_0 : \mu = \mu_0 \Rightarrow t^2 = \frac{(\bar{Y} - \mu_0)^2}{s^2/n} > t^2_{(n-1)} = F_{1,(n-1)}(\alpha)$$

Pode ser calculada para cada variável

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

$$H_0 : \mu = \mu_0 \Rightarrow T^2 = n(\bar{Y} - \mu_0)' S^{-1}(\bar{Y} - \mu_0) > \frac{(n-1)p}{(n-p)} F_{p,(n-p)}(\alpha)$$

Teste conjunto para as p variáveis

Região de Confiança e Testes
Caso de Duas Populações - Amostras 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

$$\mathbf{Y}_{1n\times p};\ \mathbf{Y}_{1i\ p\times 1}=\left(Y_{1i1},Y_{1i2},\ldots,Y_{1ip}\right)' \qquad \mathbf{Y}_{2n\times p};\ \mathbf{Y}_{2i\ p\times 1}=\left(Y_{2i1},Y_{2i2},\ldots,Y_{2ip}\right)' \qquad i=1,2,\ldots,n$$

$$\mathbf{Y}_{1i\ p\times 1}\overset{iid}{\sim}N_p\left(\mu_1;\Sigma_1\right) \longleftrightarrow \mathbf{Y}_{2i\ p\times 1}\overset{iid}{\sim}N_p\left(\mu_2;\Sigma_2\right)$$

Observações Pareadas

$\Rightarrow$ Uma Única População de Diferenças

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

$$\mathbf{D}_{i\ p\times 1}\overset{iid}{\sim}N_p\left(\mu_D=\mu_1-\mu_2;\Sigma_D\right); \quad \bar{\mathbf{D}}_{p\times 1}\sim N\left(\mu_D=\mu_1-\mu_2;\Sigma_D/n\right)$$

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

(S_D: S_u)

Elipse de Confiança para a diferença entre vetores de médias:

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

Região de Confiança e Testes
Caso de Duas Populações Independentes

Amostras Independentes - Homocedasticidade

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

$$\Rightarrow \bar{D}_{p\times 1}=\bar{Y}_1-\bar{Y}_{12} \sim N_p\left(\mu_D=\mu_1-\mu_2;\Sigma\left(\frac{1}{n_1}+\frac{1}{n_2}\right)\right)$$

$$S_c=\frac{(n_1-1)S_{u1}+(n_2-1)S_{u2}}{n_1+n_2-2}$$

S_c: Matriz de covariância comum aos grupos. Estimador de Σ

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

Elipse de Confiança para a diferença entre vetores de médias:

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

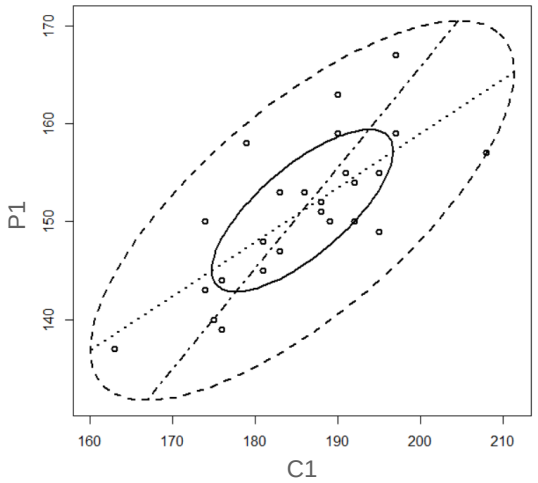
Região de Confiança e Testes - Duas Populações

Retomando os dados do Primeiro e Segundo filho em 25 famílias

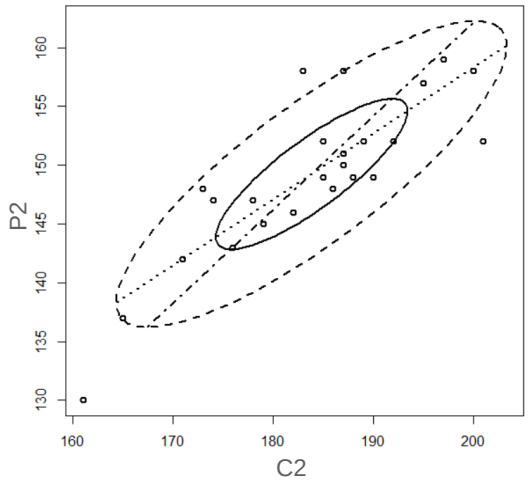
	C1	P1	C2	P2
1	191	155	179	145
2	195	149	201	152
3	181	148	185	149
4	183	153	188	149
5	176	144	171	142
6	208	157	192	152
7	189	150	190	149
8	197	159	189	152
9	188	152	197	159
10	192	150	187	151
11	179	158	186	148
12	183	147	174	147
13	174	150	185	152
14	190	159	195	157
15	188	151	187	158
16	163	137	161	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

Boxplot Bivariado: Elipses de concentração

Filiho 1



Filiho 2



Vamos supor duas “possíveis estruturas para esses dados:

- Observações pareadas (estrutura original): primeiro e segundo filho avaliados na mesma família (n=25 pares de irmãos)

- Observações independentes: dados do primeiro e do segundo filho avaliados em famílias diferentes (n_1= n_2=25; n=50)

Região de Confiança e Testes - Duas Populações Pareadas

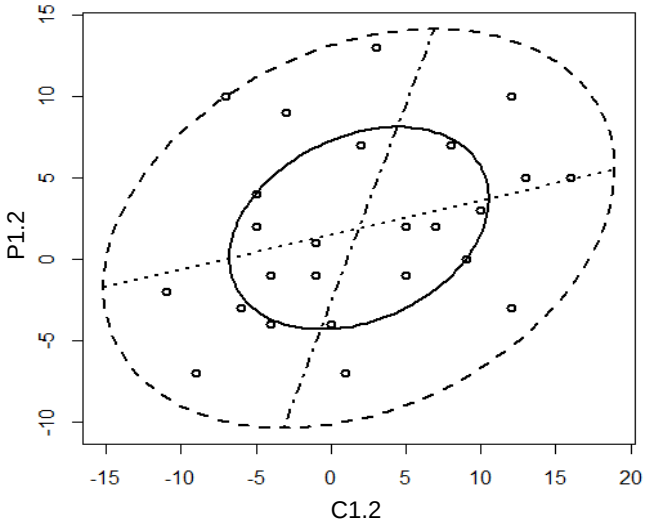
Dados do Primeiro e Segundo filho em 25 famílias

C1-C2 P1-P2

	C1	P1	C2	P2	C1.2	P1.2
1	191	155	179	145	12	10
2	195	149	201	152	-6	-3
3	181	148	185	149	-4	-1
4	183	153	188	149	-5	4
5	176	144	171	142	5	2
6	208	157	192	152	16	5
7	189	150	190	149	-1	1
8	197	159	189	152	8	7
9	188	152	197	159	-9	-7
10	192	150	187	151	5	-1
11	179	158	186	148	-7	10
12	183	147	174	147	9	0
13	174	150	185	152	-11	-2
14	190	159	195	157	-5	2
15	188	151	187	158	1	-7
16	163	137	161	130	2	7
17	195	155	183	158	12	-3
18	186	153	173	148	13	5
19	181	145	182	146	-1	-1
20	175	140	165	137	10	3
21	192	154	185	152	7	2
22	174	143	178	147	-4	-4
23	176	139	176	143	0	-4
24	197	167	200	158	-3	9
25	190	163	187	150	3	13

- Observações pareadas (estrutura original)

$\Rightarrow$ medida resumo: diferença entre as variáveis



O centróide está suficientemente distante do vetor nulo?

$$\bar{D} = (1.88,\ 1.88) \qquad S_D = \begin{pmatrix} 56.78 & 11.98 \\ 11.98 & 29.28 \end{pmatrix}$$

Região de Confiança e Testes - Duas Populações Pareadas

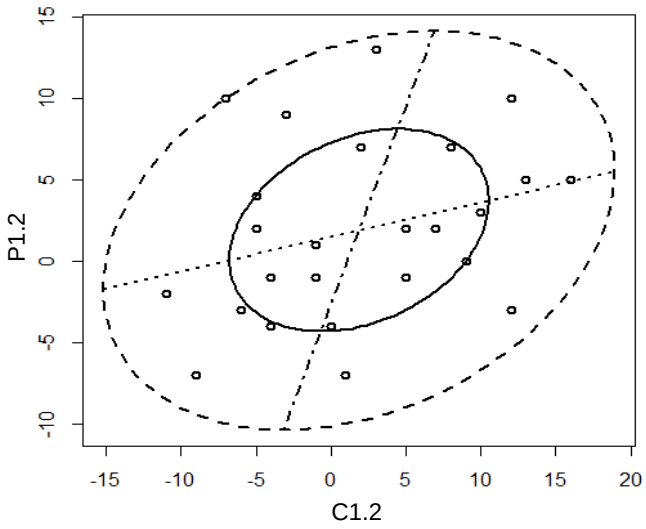
Dados do Primeiro e Segundo filho em 25 famílias

C1-C2 P1-P2

	C1	P1	C2	P2	C1.2	P1.2
1	191	155	179	145	12	10
2	195	149	201	152	-6	-3
3	181	148	185	149	-4	-1
4	183	153	188	149	-5	4
5	176	144	171	142	5	2
6	208	157	192	152	16	5
7	189	150	190	149	-1	1
8	197	159	189	152	8	7
9	188	152	197	159	-9	-7
10	192	150	187	151	5	-1
11	179	158	186	148	-7	10
12	183	147	174	147	9	0
13	174	150	185	152	-11	-2
14	190	159	195	157	-5	2
15	188	151	187	158	1	-7
16	163	137	161	130	2	7
17	195	155	183	158	12	-3
18	186	153	173	148	13	5
19	181	145	182	146	-1	-1
20	175	140	165	137	10	3
21	192	154	185	152	7	2
22	174	143	178	147	-4	-4
23	176	139	176	143	0	-4
24	197	167	200	158	-3	9
25	190	163	187	150	3	13

- Observações pareadas (estrutura original)

$\Rightarrow$ medida resumo: diferença entre as variáveis



O centróide está distante do vetor nulo?

$$T^2(\alpha) = n\left(\bar{D} - \mu_D\right)' S_D^{-1}\left(\bar{D} - \mu_D\right) \leq \frac{(n-1)p}{(n-p)} F_{p,n-p}(\alpha)$$

$$n = 25,\ p = 2,\ \alpha = 5\% \Rightarrow 7{,}14$$

$$\begin{cases} H_0 : \mu_D = (0,0)' \\ H_1 : \mu_D \neq (0,0)' \end{cases} \qquad T^2 = 3{,}61; \quad p = 0{,}1994$$

Região de Confiança e Testes - Duas Populações Independentes

Dados do Primeiro e Segundo filho

	C1	P1	C2	P2
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

hipotetica mente

	C	P
1	191	155
2	195	149
3	181	148
4	183	153
5	176	144
6	208	157
7	189	150
8	197	159
9	188	152
10	192	150
…		
24	197	167
25	190	163

1	179	145
2	201	152
3	185	149
4	188	149
5	171	142
6	192	152
7	190	149
8	189	152
9	197	159
10	187	151
…		
24	200	158
25	187	150

- Observações independentes: dados do primeiro e do segundo filho avaliados em famílias diferentes
  ($n_1 = n_2 = 25$; $n = 50$)

$$\bar{Y}_1 = (185.72,\ 151.12) \quad \bar{Y}_2 = (183.84,\ 149.24)$$

$$S_1 = \begin{pmatrix} 95.29 & 52.87 \\ 52.87 & 54.36 \end{pmatrix} \quad S_2 = \begin{pmatrix} 100.81 & 56.54 \\ 56.54 & 45.02 \end{pmatrix}$$

$$\bar{D} = (1.88, 1.88) \quad S_c = \begin{pmatrix} 98.05 & 54.70 \\ 54.70 & 49.69 \end{pmatrix}$$

Assumindo $\Sigma_1 = \Sigma_2 = \Sigma$

Teste M de Box: Homocedasticidade
Chi-Sq = 2.2674, df = 3, p-value = 0.5188

Região de Confiança e Testes - Duas Populações Independentes

Dados do Primeiro e Segundo filho

	C	P
1	191	155
2	195	149
3	181	148
4	183	153
5	176	144
6	208	157
7	189	150
8	197	159
9	188	152
10	192	150
…		
24	197	167
25	190	163

1	179	145
2	201	152
3	185	149
4	188	149
5	171	142
6	192	152
7	190	149
8	189	152
9	197	159
10	187	151
…		
24	200	158
25	187	150

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

$$n_1 = n_2 = 25,\ p = 2,\ \alpha = 5\% \Rightarrow 6.53$$

$$\begin{cases} H_0 : \mu_D = (0,0)' \\ H_1 : \mu_D \neq (0,0)' \end{cases} \qquad T^2 = 0.90; \quad p = 0{,}1824$$

É esperado que delineamentos com Amostras Pareadas ofereçam mais precisão e maior poder comparados aos delineamentos com Amostras Independentes!

Por que isso não foi atingido no caso dos dados dos Filhos?

Teste da Igualdade de Matrizes de Covariância

Comparação de Vetores de Médias - Amostras Independentes

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

$$\Rightarrow H_0 : \mu_1 - \mu_2 = 0 \quad ;\Sigma_1 = \Sigma_2 = \Sigma$$

$$S_g = \frac{1}{n_g - 1}\sum_{i=1}^{n_1}\left(Y_{gi} - \bar{Y}_g\right)\left(Y_{gi} - \bar{Y}_g\right)' ; g = 1,2$$

$$S_c = \frac{(n_1 - 1)S_1 + (n_2 - 1)S_2}{n_1 + n_2 - 2};$$

É uma Hipótese condicional. Logo a homocedasticidade deve ser verificada.

- Teste M de Box: $\Rightarrow H_0 : \Sigma_g = \Sigma;\ \mu_g \in \Re^p$

$$-2\ln\lambda = n\ln S_c - \sum_{g=1}^{G}\left[n_g \ln\left|S_g\right|\right]$$

$$M = (1-c)\left\{\left[\sum_{g=1}^{G}\left(n_g - 1\right)\right]\ln\left|S_c\right| - \sum_{g=1}^{G}\left[\left(n_g - 1\right)\ln\left|S_g\right|\right]\right\} \sim \chi^2_{\frac{1}{2}p(p+1)(G-1)}$$

$$c = \left[\sum_{g=1}^{G}\frac{1}{\left(n_g - 1\right)}\right]\left[\frac{2p^2 + 3p - 1}{6(p+1)(G-1)}\right]$$

S (com divisor n), S_u (divisor n-1)
Para p=1 o teste de Box equivale ao teste de Bartlet.

Critério "prático" de heterocedasticidade sugerido em Johnson and Wichern 1992:

$$\sigma_{gjj} = 4\sigma_{g'jj}$$

Inferência sobre Vetores de Médias de Duas Populações

Medidas de produtividade e altura de plantas de duas variedades

Variedade A		Variedade B	
X11	X12	X21	X22
5,7	2,1	4,4	1,8
8,9	1,9	7,5	1,75
6,2	1,98	5,4	1,78
5,8	1,92	4,6	1,89
6,8	2	5,9	1,9
6,2	2,01		

- Teste a igualdade do vetor de médias das duas variedades, sob homocedasticidade.
- Obtenha os intervalos de confiança simultâneos e de Bonferroni.

Inferência sobre Vetores de Médias de Duas Populações

	Variedade A		Variedade B	
	X11	X12	X21	X22
	5,7	2,1	4,4	1,8
	8,9	1,9	7,5	1,75
	6,2	1,98	5,4	1,78
	5,8	1,92	4,6	1,89
	6,8	2	5,9	1,9
	6,2	2,01		
Média	6,6	1,985	5,56	1,824
Var.	1,42	0,005	1,543	0,0045

$$\overline{D} = \begin{pmatrix} \overline{D}_1 = 6{,}6 - 5{,}56 = 1{,}04 \\ \overline{D}_2 = 1{,}985 - 1{,}824 = 0{,}161 \end{pmatrix}$$

$$S_1 = \begin{pmatrix} 1{,}4200 & -0{,}0504 \\ -0{,}0504 & 0{,}0051 \end{pmatrix}$$

$$S_2 = \begin{pmatrix} 1{,}543 & -0{,}037 \\ -0{,}037 & 0{,}0045 \end{pmatrix}$$

$$S_c = \begin{pmatrix} 1{,}4745 & -0{,}0442 \\ -0{,}0442 & 0{,}0049 \end{pmatrix}$$

$$\Rightarrow H_0 : \mu_1 - \mu_2 = 0 \quad ; \Sigma_1 = \Sigma_2 = \Sigma$$

$$T^2 = \left(\overline{Y}_1 - \overline{Y}_2\right)' \left[\left(\frac{1}{n_1} + \frac{1}{n_2}\right) S_c\right]^{-1} \left(\overline{Y}_1 - \overline{Y}_2\right) = 30{,}584 \quad \Rightarrow \quad p = 0{,}0027$$

Conclusão?

Existe diferença significante entre os dois grupos para alguma combinação linear das variáveis.

Inferência sobre um Vetor de Médias

Intervalos de Confiança Univariados e Simultâneos

$$I.C(\mu_k) \text{ a } 100(1-\alpha)\% = \left(\bar{Y}_k - t_{n-1}(\alpha/2)\sqrt{\frac{s_{kk}}{n}} ; \bar{Y}_k + t_{n-1}(\alpha/2)\sqrt{\frac{s_{kk}}{n}} \right)$$

Sob independência $\Rightarrow P(\text{todos os p intervalos conterem os } \mu_k\text{'s}) = (1-\alpha)^p$

$(1-\alpha) = 0,95 \;,\; p = 2, n = 25 \quad \Rightarrow P(\text{coletivo}) = (0,95)^2 = 0,90 \quad t_{24}(\alpha/2) = 2,06$

$(1-\alpha) = 0,95 \;,\; p = 4, n = 15 \quad \Rightarrow P(\text{coletivo}) = (0,95)^4 = 0,81 \quad t_{14}(\alpha/2) = 2,14$

$$I.C.S(\mu_k) \text{a } 100(1-\alpha)\% = \left(\bar{Y}_k - \sqrt{\frac{(n-1)p}{(n-p)} F_{p,(n-p)}(\alpha) \frac{s_{kk}}{n}} \; ; \bar{Y}_k + \sqrt{\frac{(n-1)p}{(n-p)} F_{p,(n-p)}(\alpha) \frac{s_{kk}}{n}} \right)$$

Nível coletivo igual a (1-α) $\Rightarrow$ intervalos simultâneos são mais largos que os individuais:

$$(1-\alpha) = 0,95, \;\; p = 2, n = 25 \Rightarrow \sqrt{\frac{(n-1)p}{(n-p)} F_{p,n-p}(\alpha)} = 2,67$$

mais largos

$$(1-\alpha) = 0,95, \;\; p = 4, n = 15 \Rightarrow \sqrt{\frac{(n-1)p}{(n-p)} F_{p,n-p}(\alpha)} = 4,13$$

Inferência sobre um Vetor de Médias

O Método de Bonferroni para Comparações Múltiplas

$$\Rightarrow P\left(\text{todos os p intervalos conterem os } \mu_k\text{'s}\right) = P\left(\text{todas hipóteses } H_0 \text{ serem verdadeiras}\right)$$

$$= 1 - P\left(\text{pelo menos uma } H_0 \text{ ser falsa}\right)$$

$$\geq 1 - \sum_{k=1}^{p} P\left(H_{0k} \text{ falsa}\right) = 1 - \underbrace{\left(\alpha_1 + \alpha_2 + \ldots + \alpha_p\right)}$$

Controle da taxa de erro total independente da estrutura de covariância $\Rightarrow$ muito conservador

$$\left(\bar{Y}_1 - t_{n-1}(\alpha/2p)\sqrt{\frac{s_{11}}{n}}; \bar{Y}_1 + t_{n-1}(\alpha/2p)\sqrt{\frac{s_{11}}{n}}\right)$$

$$\left(\bar{Y}_2 - t_{n-1}(\alpha/2p)\sqrt{\frac{s_{22}}{n}}; \bar{Y}_2 + t_{n-1}(\alpha/2p)\sqrt{\frac{s_{22}}{n}}\right)$$

$$\ldots \qquad \ldots$$

$$\left(\bar{Y}_p - t_{n-1}(\alpha/2p)\sqrt{\frac{s_{pp}}{n}}; \bar{Y}_p + t_{n-1}(\alpha/2p)\sqrt{\frac{s_{pp}}{n}}\right)$$

O critério de Bonferroni para correção de múltiplos testes é conservador

Bastante utilizado para comparações de subconjuntos de médias (m<p)

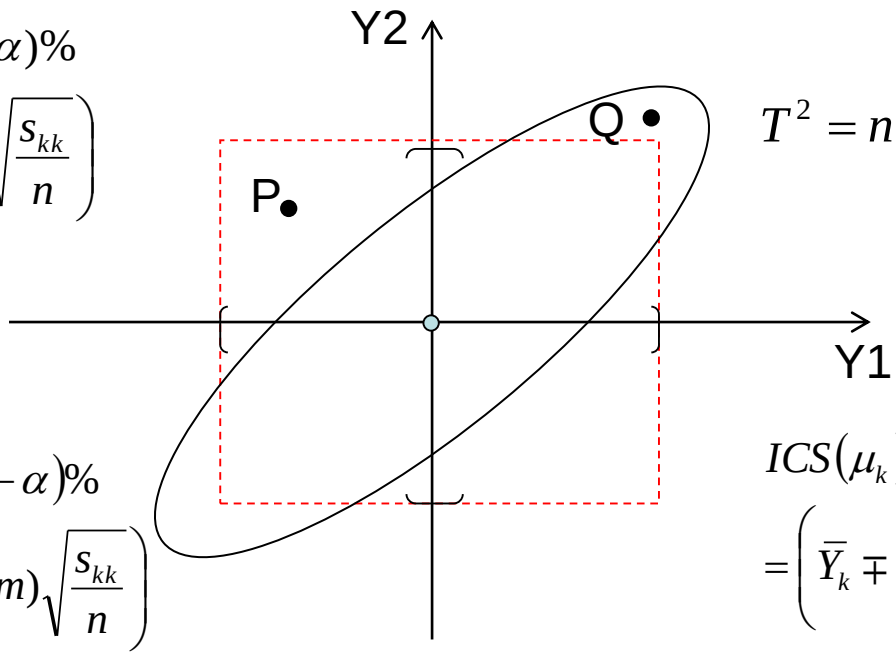
$\Rightarrow \alpha/2m$

Intervalos e Regiões de Confiança

Intervalos de Confiança Univariados e Simultâneos

Everitt, 2002

$$IC(\mu_k) \text{ a } 100(1-\alpha)\% = \left(\bar{Y}_k \mp t_{n-1}(\alpha/2)\sqrt{\frac{s_{kk}}{n}} \right)$$



$$T^2 = n\left(\bar{Y}-\mu\right)' S^{-1}\left(\bar{Y}-\mu\right)$$

$$ICB(\mu_k)\,\text{a}\,100(1-\alpha)\% = \left(\bar{Y}_k \mp t_{n-1}(\alpha/2m)\sqrt{\frac{s_{kk}}{n}} \right)$$

$$ICS(\mu_k)\,\text{a}\,100(1-\alpha)\% = \left(\bar{Y}_k \mp \sqrt{\frac{(n-1)p}{(n-p)} F_{p,(n-p)}(\alpha)\frac{s_{kk}}{n}} \right)$$

⇒ Compare as análises multivariadas (Região de Confiança, ICS) e univariada (IC e ICB).

⇒ Comente sobre as decisões tomadas para os pontos P e Q sob análises univariadas e multivariadas. Justifique.

Inferência sobre Vetores de Médias de Duas Populações

	Variedade A		Variedade B	
	X11	X12	X21	X22
	5,7	2,1	4,4	1,8
	8,9	1,9	7,5	1,75
	6,2	1,98	5,4	1,78
	5,8	1,92	4,6	1,89
	6,8	2	5,9	1,9
	6,2	2,01		
Média	6,6	1,985	5,56	1,824
Var.	1,42	0,005	1,543	0,0045

$$\overline{D} = \begin{pmatrix} \overline{D}_1 = 6{,}6 - 5{,}56 = 1{,}04 \\ \overline{D}_2 = 1{,}985 - 1{,}824 = 0{,}161 \end{pmatrix}$$

$$S_c = \begin{pmatrix} 1{,}4745 & -0{,}0442 \\ -0{,}0442 & 0{,}0049 \end{pmatrix}$$

$$\Rightarrow H_0 : \mu_1 - \mu_2 = 0 \quad ; \Sigma_1 = \Sigma_2 = \Sigma$$

$$T^2 = 30{,}584; \quad p = 0{,}0027$$

I.C.S. para a diferença entre os grupos em cada variável (ou combinação linear delas)

$$l'(\overline{D}) \pm \sqrt{\frac{(n_1 + n_2 - 2)p}{(n_1 + n_2 - p - 1)} F_{p,(n_1+n_2-p-1)}} \sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right) l' S_c l} \Rightarrow$$

Variável X1: $1{,}04 \pm \sqrt{(18/8)4{,}46(1/6 + 1/5)1{,}47} = 1{,}04 \pm 2{,}33 = (-1{,}29; 3{,}37)$

Conclusão?

Variável X2: $0{,}161 \pm \sqrt{(18/8)4{,}46(1/6 + 1/5)0{,}0049} = 0{,}161 \pm 0{,}134 = (0{,}027; 0{,}295)$

Inferência sobre Vetores de Médias de Duas Populações

	Variedade A		Variedade B	
	X11	X12	X21	X22
	5,7	2,1	4,4	1,8
	8,9	1,9	7,5	1,75
	6,2	1,98	5,4	1,78
	5,8	1,92	4,6	1,89
	6,8	2	5,9	1,9
	6,2	2,01		
Média	6,6	1,985	5,56	1,824
Var.	1,42	0,005	1,543	0,0045

$$\overline{D} = \begin{pmatrix} \overline{D}_1 = 6{,}6 - 5{,}56 = 1{,}04 \\ \overline{D}_2 = 1{,}985 - 1{,}824 = 0{,}161 \end{pmatrix}$$

$$S_c = \begin{pmatrix} 1{,}4745 & -0{,}0442 \\ -0{,}0442 & 0{,}0049 \end{pmatrix}$$

$$\Rightarrow H_0 : \mu_1 - \mu_2 = 0 \quad ; \Sigma_1 = \Sigma_2 = \Sigma$$

$$T^2 = 30{,}584; \quad p = 0{,}0027$$

I.C. de Bonferroni para a diferença entre os grupos em cada variável

$$ICB(\mu_{1j} - \mu_{2j}) \text{ a } 100(1-\alpha)\% = (\overline{Y}_{1j} - \overline{Y}_{2j}) \pm t_{n_1+n_2-2}(0{,}05/4)\sqrt{\left(\frac{1}{n_1}+\frac{1}{n_2}\right)S_{jj}}$$

$$1{,}04 \pm 2{,}685\sqrt{(1/6+1/5)1{,}47} = 1{,}04 \pm 1{,}97 = (-0{,}93; 3{,}01)$$

$$0{,}161 \pm 2{,}685\sqrt{(1/6+1/5)0{,}0049} = 0{,}161 \pm 0{,}114 = (0{,}047; 0{,}275)$$

Conclusão?