MAE 5776

ANÁLISE MULTIVARIADA

Júlia M Pavan Soler

pavan@ime.usp.br

1º Semestre/2019

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

MAE5776

Já vimos ☺

Matriz de Dados: Estatísticas descritivas multivariadas

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

- Definidas no espaço das linhas (n-vetores p-dimensionais): $D_{n\times n} = (d^2_{ij}), d^2_{Pij}, d^2_{Mij}$

-Propriedades em espaços duais:

$$\left.\begin{array}{l} nS_{p\times p} = (HY)'HY = V\Lambda V' \\ B_{n\times n} = HY(HY)' = U\Lambda U' \end{array}\right\} HY = U\Lambda^{1/2}V';\ H = I_n - n^{-1}1_n 1_n'$$

Regiões (elipsóides) de Concentração de Observações:

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

Matriz Aleatória: Propriedades distribucionais

$$Y_{n\times p} = (Y_{ij}) \in \Re^{n\times p};\quad Y_{n\times p} \sim N_{n\times p}(1_n \mu'_{p\times 1}; \Omega_{np\times np} = \Psi_{n\times n} \otimes \Sigma_{p\times p});\quad (vecY)_{np\times 1} \sim N_{np}(1_n \otimes \mu_{p\times 1}; \Omega)$$

- Estimadores e Distribuições Amostrais: $Y_{i_{p\times 1}} \overset{iid}{\sim} N_p(\mu;\Sigma)$

$$\left.\begin{array}{l} \bar{Y}_{p\times 1} \sim N_p(\mu; \Sigma/n) \\ nS_{p\times p} \sim W_p(n-1;\Sigma) \end{array}\right.$$

- Regiões (elipsoides) de Confiança para μ:

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

Inferência – Análise Multivariada

Por que realizar Testes de “vetores” de médias?

Testes Multivariados × Testes Univariados

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

- 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 medias das variáveis
- Construir níveis de significância coletivos × Correções para múltiplos testes

Bonferroni, FDR

Testes de Hipóteses

$Y;\quad f_Y(y\,|\,\theta)$: Inferências sobre o parâmetro $\theta \in \Re^p$; $\theta \in \Re^{pxp}$

O Problema de Decisão: Rejeitar ou Não-Rejeitar a Hipótese Nula?

cardinalidade

- Hipóteses: $H_0 : \theta \in \Theta_0 \quad \times \quad H_0 : \theta \in \Theta_1; \quad \Theta = \Theta_0 \cup \Theta_1$

  Hipótese Simples: $|\Theta_i| = 1$

  Hipótese Composta: $|\Theta_i| > 1$

- Teste: $\varphi : \Omega_Y \to \{0,1\}; \quad \varphi(y) = \begin{cases} 0 & \text{não rejeitar } H_0 \\ 1 & \text{rejeitar } H_0 \end{cases}$ (Espaço amostral)

- Região Crítica: $R_c = \{y \in \Omega_Y; \quad \varphi(y) = 1\} = \varphi^{-1}(1)$ Região de rejeição de H_0

  Região de Aceitação: $R(Y) = \{y \in \Omega_Y; \quad \varphi(y) = 0\} = \varphi^{-1}(0)$ Região de confiança

- Erros
  - Tipo I: Rejeitar H_0 | H_0 V ; P(erro I) = α: Nível de significância
  - Tipo II: Não Rejeitar H_0 | H_0 F ; P(erro II) = β

- Função Poder: $\pi_\phi : \Theta \to [0,1]; \quad \pi_\phi(\theta) = P(\phi(y) = 1\,|\,\theta) \Rightarrow \begin{cases} P(\phi(y) = 1\,|\,\theta \in \Theta_0) = \alpha \\ P(\phi(y) = 1\,|\,\theta \in \Theta_1) = 1 - \beta \end{cases}$

  Poder do teste

Teste da Razão de Verossimilhanças

$$Y_{n\times p} = (Y_1,...,Y_n)'; \; Y_{i_{p\times 1}} \overset{iid}{\sim} f_{Y_i}(y_i \mid \theta), \quad \theta = (\mu;\Sigma)$$

$$H_0 : \theta \in \Theta_0 \quad \times \quad H_0 : \theta \in \Theta_1; \quad \Theta = \Theta_0 \cup \Theta_1$$

Definição: A estatística da Razão de Verossimilhanças para testar H_0 x H_1 , com $L(\theta|y)$ a função de verossimilhança, é:

$$\lambda(y) = \frac{\sup_{\theta\in\Theta_0} L(\theta \mid y)}{\sup_{\theta\in\Theta_1} L(\theta \mid y)} = \frac{L_0^*}{L_1^*} \quad \text{ou, equivalentemente,} \quad -2\ln\lambda = 2\left(\ln L_1^* - \ln L_0^*\right) = 2\left(l_1^* - l_0^*\right)$$

$$R_c = \{y \in \Omega_Y; \quad \lambda(y) < c\}; \quad c, \; \sup_{\theta\in\Theta_0} P_\theta(y \in R_c \mid \theta) = \alpha$$

$$R_c = \{y \in \Omega_Y; \quad -2\ln\lambda(y) > c\}; \quad c, \; \sup_{\theta\in\Theta_0} P_\theta(y \in R_c \mid \theta) = \alpha$$

Teorema: Se $\Theta_1 \in \Re^q$ e $\Theta_0 \in \Re^r$ é uma sub-região de Θ_1 , então sob condições de regularidades satisfeitas para *f*, para cada $\theta \in \Theta_0$,

$$-2\ln\lambda \overset{n\to\infty}{\sim} \chi^2_{q-r}$$

$$H_0 : \mu_{p\times 1} = 0; \; \Sigma \in \Re^{p\times p}$$

$$\Rightarrow q - r = \left(p + p(p+1)/2\right) - \left(p(p+1)/2\right) = p$$

Teste da Razão de Verossimilhanças
Uma Única População

$$Y_{n\times p} = (Y_1,...,Y_n)'; \; Y_{i_{p\times 1}} \overset{iid}{\sim} N_p(\mu;\Sigma)$$

$$H_0 : \mu = \mu_0 \quad ; \quad \Sigma \text{ conhecido}$$

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

Estatística da Razão de verossimilhanças:

$$-2\ln\lambda = 2\left(\ln L_1^* - \ln L_0^*\right) = 2\left(l_1^* - l_0^*\right)$$

$$l_0^* = \ln L(\mu_0;\Sigma) = (-1/2)n\ln|2\pi\,\Sigma| - (1/2)ntr\Sigma^{-1}S - (1/2)n(\bar{Y}-\mu_0)'\Sigma^{-1}(\bar{Y}-\mu_0)$$

$$l_1^* = \ln L(\bar{Y};\Sigma) = (-1/2)n\ln|2\pi\,\Sigma| - (1/2)ntr\Sigma^{-1}S$$

$$-2\ln\lambda = n(\bar{Y}-\mu_0)'\Sigma^{-1}(\bar{Y}-\mu_0) = nd_0'\Sigma^{-1}d_0 \sim \chi_p^2$$

Regra de Decisão: $nd_M^2(\bar{Y},\mu_0)$

$$\begin{cases} \geq \chi_p^2(\alpha) \Rightarrow \text{Rejeitar } H_0 \\ c.c. \quad \Rightarrow \text{Não Rejeitar } H_0 \end{cases}$$

Teste de Hipóteses para o Vetor μ
Uma Única População

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

	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

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

Estatísticas Descritivas:

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

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

Teste de Hipóteses para o Vetor μ
Uma Única População

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

	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

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

Distribuição marginal:

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

Hipóteses:

$$H_0 : \mu = \begin{pmatrix}182\\ 182\end{pmatrix} \quad ; \quad \Sigma = \begin{pmatrix}100 & 0\\ 0 & 100\end{pmatrix}$$

Estatística LR:

$$-2\ln\lambda = 4{,}31 \Rightarrow p - valor = 0{,}1159$$

$$P\left(\chi^2_2 \leq 5{,}99\right) = 0{,}95$$

Conclusão: Não há evidência amostral para rejeitar H_0

Teste de Hipóteses para o Vetor μ
Uma Única População

$$Y_{n\times p} = (Y_1,...,Y_n)'; \; Y_{i_{p\times 1}} \overset{iid}{\sim} N_p(\mu;\Sigma)$$

$$H_0 : \mu = \mu_0 \quad ; \quad \Sigma \; desconhecido$$

Estatística da Razão de verossimilhanças:

$$\text{Sob } H_0 : \hat{\mu} = \mu_0, \; \hat{\Sigma}_0 = S + d_0 d_0'; \; d_0 = \bar{Y} - \mu_0$$

$$l_0^* = \ln L(\mu_0;\hat{\Sigma}_0) = (-1/2)n\left\{ p\ln 2\pi + \ln|S| + \ln\left(1 + d_0' S^{-1} d_0\right) + p \right\}$$

$$\text{Sob } H_1 : \hat{\mu} = \bar{Y}, \; \hat{\Sigma} = S$$

$$\Rightarrow \; -2\ln\lambda = 2\left(l_1^* - l_0^*\right) = n\ln\left(1 + d_0' S^{-1} d_0\right)$$

é função crescente de:

$$(n-1) d_0' S^{-1} d_0 \sim T^2_{(p;n-1)}$$

$$\frac{n-p}{p} d_0' S^{-1} d_0 \sim F_{(p;n-p)}$$

Teste T^2 de Hotelling para uma População

Teste de Hipóteses para o Vetor μ
Uma Única População

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

	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

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

Distribuição marginal:

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

Hipóteses:

$$H_0 : \mu = \begin{pmatrix}182\\ 182\end{pmatrix} \quad ; \quad \Sigma \text{ desconhecido}$$

Estatística de Hotelling:

$$\frac{n-p}{p} d_0' S^{-1} d_0 = 1{,}28 \Rightarrow p - valor = 0{,}2971$$

$$P\left(F_{(2,23)} \leq 3{,}44\right) = 0{,}95$$

Conclusão: Não há evidência amostral para rejeitar H_0

Teste de Hipóteses para o Vetor μ
Uma Única População

$$Y_{n\times p} = (Y_1,...,Y_n)'; \; Y_{i_{p\times 1}} \overset{iid}{\sim} N_p(\mu;\Sigma); \quad \Sigma \in \Re^{p\times p}$$

Teorema: $\text{Sob } H_0 \overset{EMVS\ \mu}{\Rightarrow} \hat{\mu}_0$, $\text{Sob } H_1 \overset{EMVS\ \mu}{\Rightarrow} \bar{Y}$. Então:

$$\text{Sob } H_0 \overset{EMVS\ \Sigma}{\Rightarrow} S + d_0 d_0'; \; d_0 = \bar{Y} - \hat{\mu}_0; \quad \text{Sob } H_1 \overset{EMVS\ \Sigma}{\Rightarrow} S = n^{-1}\sum (Y_i - \bar{Y})(Y_i - \bar{Y})'$$

Estatística da Razão de Verossimilhanças:

- Para Σ conhecido: $-2\ln\lambda = n(\bar{Y} - \mu_0)'\Sigma^{-1}(\bar{Y} - \mu_0) \sim \chi^2_p$
- Para Σ desconhecido:

$$(n-1)\, d_0' S^{-1} d_0 \sim T^2_{(p;n-1)} = \frac{(n-1)p}{n-p} F_{(p;n-p)}$$

$$n\, d_0' S_u^{-1} d_0 \sim T^2_{(p;n-1)} = \frac{(n-1)p}{n-p} F_{(p;n-p)}$$

Sob Normalidade: distribuições exatas

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

Teste de Hipóteses para o Vetor μ
Uma Única População

$$Y_{n\times p} = (Y_1,\ldots,Y_n)'; \; Y_{i_{p\times 1}} \overset{iid}{\sim} N_p(\mu;\Sigma)$$

$$H_0 : \mu = \mu_0 \quad ; \quad \Sigma \; desconhecido$$

Outra alternativa:

Estatística Lambda de Wilks:

$$T^2 = (n-1) d_0' S^{-1} d_0 \sim T^2_{(p;n-1)}; \quad \frac{n-p}{p} d_0' S^{-1} d_0 \sim F_{(p;n-p)}$$

$$T^2 = n d_0' S_u^{-1} d_0 \sim T^2_{(p;n-1)}; \quad T^2 = \frac{\left|\hat{\Sigma}_0\right|(n-1)}{\left|\hat{\Sigma}\right|} - (n-1) = \Lambda^{2/n}(n-1) - (n-1);$$

$$\Lambda^{2/n} = \frac{\left|\hat{\Sigma}_0\right|}{\left|\hat{\Sigma}\right|}; \quad \Lambda = \left\{ \frac{\left|\sum_{i=1}^{n} (Y_i - \bar{Y})(Y_i - \bar{Y})'\right|}{\left|\sum_{i=1}^{n} (Y_i - \mu_0)(Y_i - \mu_0)'\right|} \right\}^{n/2} = \left(\frac{\left|\hat{\Sigma}\right|}{\left|\hat{\Sigma}_0\right|} \right)^{n/2}$$

Estatística Lambda de Wilks

Teste de Hipóteses para o Vetor μ
Uma Única População

Estatística de Hotelling e Estatística Lambda de Wilks

$$Y_{n\times p} = (Y_1,...,Y_n)'; \; Y_{i_{p\times 1}} \overset{iid}{\sim} N_p(\mu;\Sigma) \qquad H_0 : \mu = \mu_0 \quad ; \quad \Sigma \; desconhecido$$

$$\Lambda^{2/n} = \frac{\left|\hat{\boldsymbol{\Sigma}}\right|}{\left|\hat{\boldsymbol{\Sigma}}_0\right|} = \left(1 + \frac{T^2}{(n-1)}\right)^{-1} \qquad T^2 = \frac{(n-1)\left|\hat{\boldsymbol{\Sigma}}_0\right|}{\left|\hat{\boldsymbol{\Sigma}}\right|} - (n-1)$$

Regra de Decisão: Rejeitar H_0 $\begin{cases} T^2 \geq T^2_{(p,n-1)}(\alpha) \\ \Lambda < c^2_\alpha \end{cases}$

$\Rightarrow H_0$ é rejeitada para valores “pequenos” da estatística Lambda de Wilks e valores “grandes” da estatística de Hotelling

Teste de Hipóteses para a Matriz Σ
Uma Única População

$$Y_{n\times p} = (Y_1,...,Y_n)'; \; Y_{i_{p\times 1}} \overset{iid}{\sim} N_p(\mu;\Sigma)$$

$$H_0 : \Sigma = \Sigma_0 \quad ; \quad \mu \; desconhecido$$

Estatística da Razão de verossimilhanças:

$$\text{Sob } H_0 : \hat{\mu} = \bar{Y}, \; \hat{\Sigma}_0 = \Sigma_0 \quad \Rightarrow l_0^* = \ln L(\bar{Y};\Sigma_0) = (-1/2)n\left\{\ln\left|2\pi\,\Sigma_0\right| + tr\Sigma_0^{-1}S\right\}$$

$$\text{Sob } H_1 : \hat{\mu} = \bar{Y}, \; \hat{\Sigma} = S \quad \Rightarrow l_1^* = \ln L(\bar{Y};S) = (-1/2)n\left\{\ln\left|2\pi\,S\right| + p\right\}$$

$$-2\ln\lambda = 2(l_1^* - l_0^*) = n\,\overbrace{tr\Sigma_0^{-1}S}^{=pa} - n\ln\overbrace{\left|\Sigma_0^{-1}S\right|}^{=g^p} - np$$

$$= np(a - \ln g - 1)$$

Distribuição desta variável não é simples $\Rightarrow$ uso da teoria assintótica

a: média aritmética dos autovalores de $\Sigma_0^{-1}S$

g: média geométrica dos autovalores de $\Sigma_0^{-1}S$

Na Regra de Decisão: $-2\ln\lambda \overset{H_0, n\to\infty}{\sim} \chi^2_{p(p+1)/2}$

Teste de Hipóteses para a Matriz Σ
Uma Única População

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

	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

Distribuição marginal:

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

Hipótese: $H_0 : \Sigma = \begin{pmatrix}100 & 0\\ 0 & 100\end{pmatrix}$; μ desconhecido

$$-2\ln\lambda = 17{,}70 \Rightarrow p-valor = 0.0005071$$

Conclusão: Rejeitar H_0

Hipótese: $H_0 : \Sigma = \begin{pmatrix}100 & 50\\ 50 & 100\end{pmatrix}$; μ desconhecido

$$-2\ln\lambda = 3{,}91 \Rightarrow p-valor = 0.2713$$

Conclusão: Não há evidência para Rejeitar H_0

Teste da União Intersecção

$$Y_{n\times p} = (Y_1,\ldots,Y_n)'; \; Y_{i_{p\times 1}} \overset{iid}{\sim} N_p(\mu; I_p)$$

Independência e homocedasticidade

$$l \in \Re^p; \; l'Y_{i_{p\times 1}} \overset{iid}{\sim} N_1(l'\mu\,; l'l)$$

$$H_0 : \mu = \mu_0 \;\Rightarrow\; X_{il} = l'Y_i \overset{iid}{\sim} N_1(l'\mu\,; l'l) \;\Rightarrow H_{0l} : X_{il} \overset{iid}{\sim} N_1(l'\mu_0\,; l'l)$$

$$H_0 = \cap_l H_{0l}$$

A hipótese multivariada é verdadeira para todos os vetores $l \in \Re^p$

Teste de H_{0l} $\quad \Rightarrow z_l = \dfrac{\bar{X} - l'\mu_0}{\sqrt{l'l/n}} \sim N_1(0;1)$ $\quad R_{cl} = \{z_l;\; |z_l| > z_c(\alpha/2)\} = \{z_l;\; z_l^2 > z_c^2(\alpha/2)\}$

$$R_c = \cup_l R_{cl}$$

A hipótese multivariada é rejeitada para pelo menos um vetor $l \in \Re^p$ tal que $z_l \in R_{cl}$

Desigualdade de Cauchy-Schuarz

esferas

$$\Rightarrow \text{ Não Rejeita } H_0 \Leftrightarrow \max_l z_l^2 \le z_c^2(\alpha/2) \quad \max_l z_l^2 = \max_l \frac{nl'(\bar{Y}-\mu_0)(\bar{Y}-\mu_0)'l}{l'l} = n(\bar{Y}-\mu_0)'(\bar{Y}-\mu_0) \le z_c^2(\alpha/2)$$

Teste da União Intersecção

Importância prática: Quando a hipótese multivariada é rejeitada há interesse em qual componente das possíveis combinações lineares foi responsável pela rejeição.

O Teste da Razão de Verossimilhança não tem esta propriedade, exceto quando ambos os testes RL e UI conduzem ao mesmo critério de teste.

$$Y_{n\times p} = (Y_1,...,Y_n)'; \; Y_{i_{p\times 1}} \overset{iid}{\sim} N_p(\mu;\Sigma) \qquad H_0 : \mu = \mu_0 \quad ; \quad \Sigma \; desconhecido$$

$$X_{n\times 1} = Y\, l; \qquad X_i \overset{iid}{\sim} N_1\left(\mu_x = l'\mu; \sigma_x^2 = l'\Sigma l\right)$$

$$H_{0l}; \quad t_l = \frac{\bar{X}_l - \mu_X}{\sqrt{s_X^2/(n-1)}}; \quad t_l^2 = (n-1)\frac{l'(\bar{Y}-\mu_0)(\bar{Y}-\mu_0)'l}{l'Sl}$$

Testes LR e UI coincidem!

$$\max_l t_l^2 = (n-1)(\bar{Y}-\mu_0)'S^{-1}(\bar{Y}-\mu_0) = (n-1)d_0'S^{-1}d_0 \sim T^2_{(p;n-1)}; \qquad \frac{n-p}{p}d_0'S^{-1}d_0 \sim F_{(p;n-p)}$$

$$nd_0'S_u^{-1}d_0 \sim T^2_{(p;n-1)};$$

Inferência sobre Componentes do Vetor μ

Comparações Simultâneas de Componentes do Vetor de Médias μ

$$Y_{n\times p} = (Y_1,...,Y_n)'; \; Y_{i_{p\times 1}} \overset{iid}{\sim} N_p(\mu;\Sigma) \quad \Rightarrow \quad \mu_{p\times 1} = (\mu_1,\mu_2,...,\mu_p)'$$

Considere combinações lineares das p variáveis:

$$Z_i = l' Y_i = l_1 Y_{i1} + l_2 Y_{i2} + ... + l_p Y_{ip} \qquad i = 1,2,...,n$$

$$Z_i = l' Y_i \overset{iid}{\sim} N_1(l'\mu \; ; l' \Sigma \, l)$$

$$\bar{Z} = l' \bar{Y} \quad s_Z^2 = l' S_u l \qquad t_l = \frac{\bar{Z} - l'\mu}{s_Z / \sqrt{n}} \sim t_{(n-1)}$$

$$t_l = \frac{l'(\bar{Y} - \mu)}{\sqrt{l' S_u l / n}} \Rightarrow IC(l'\mu) \text{ a } 100(1-\alpha)\% = \left(l'\bar{Y} \mp t_{n-1}(\alpha/2) \frac{\sqrt{l' S_u l}}{\sqrt{n}} \right)$$

limitação

$$l = (0,0,...,1,...,0,0) \Rightarrow \left(\hat{\mu}_k - t_{n-1}(\alpha/2) \frac{\sqrt{s_{kk}}}{\sqrt{n}} ; \hat{\mu}_k + t_{n-1}(\alpha/2) \frac{\sqrt{s_{kk}}}{\sqrt{n}} \right)$$

posição k

Intervalos de confiança a 100(1-α)% para cada média ⇒ qual o nível de confiança global ?

Inferência sobre Componentes do Vetor de Médias

Comparações Simultâneas de Componentes do Vetor de Medias μ

Intervalos de confiança Simultâneos com coeficiente de confiança "coletivo" 100(1-α)%:

$$\begin{cases} Y_{n\times p} = (Y_1,\ldots,Y_n)'; \; Y_{i_{p\times 1}} \overset{iid}{\sim} N_p(\mu;\Sigma) \\ Z_i = l'\,Y_i = l_1 Y_{i1} + l_2 Y_{i2} + \ldots + l_p Y_{ip}; \qquad Z_i = l'\,Y_i \overset{iid}{\sim} N_1(l'\mu\,;\, l'\,\Sigma\, l) \end{cases}$$

$$\Rightarrow ICS(l'\mu)\; a\; 100(1-\alpha)\% = \left(l'\bar{Y} - \sqrt{\frac{(n-1)p}{(n-p)} F_{p,(n-p)}(\alpha) \frac{l'S_u l}{n}} \;;\; l'\bar{Y} + \sqrt{\frac{(n-1)p}{(n-p)} F_{p,(n-p)}(\alpha) \frac{l'S_u l}{n}} \right)$$

O Princípio da União Intersecção garante um coeficiente de confiança coletivo igual a (1-α) para todo $l \in \Re^p \Rightarrow t_l^2 \le T^2 \le T_\alpha^2$

Para garantir um nível coletivo igual a (1-α) para todo $l \in \Re^p$ os intervalos simultâneos têm amplitude (tamanho) maior que os intervalos individuais.

Intervalos de Confiança Simultâneos para μ

$$ICS(l'\mu)\, a\, 100(1-\alpha)\% = \left(l'\bar{Y} - \sqrt{\frac{(n-1)p}{(n-p)} F_{p,(n-p)}(\alpha) \frac{l'S_u l}{n}} \;;\; l'\bar{Y} + \sqrt{\frac{(n-1)p}{(n-p)} F_{p,(n-p)}(\alpha) \frac{l'S_u l}{n}} \right)$$

$H_0 : \mu_{p\times 1} = 0$ é rejeitada $\Rightarrow$ deverá existir pelo menos um vetor $l \in \Re^p$ para o qual o correspondente ICS para $l'\mu$ não contém o zero.

Para a variável *k* $\Rightarrow$ adotar a combinação canônica:

$$l = (0,\ldots,0,1,0\ldots,0)' \Rightarrow \left(\bar{Y}_k - \sqrt{\frac{(n-1)p}{(n-p)} F_{p,(n-p)}(\alpha) \frac{s_{kk}}{n}} \;;\; \bar{Y}_k + \sqrt{\frac{(n-1)p}{(n-p)} F_{p,(n-p)}(\alpha) \frac{s_{kk}}{n}} \right)$$

Comparações entre Médias das variáveis: (quando há interesse!) $l; l'\bar{Y} = \bar{Y}_j - \bar{Y}_k$

$$\Rightarrow \left((\bar{Y}_j - \bar{Y}_k) - \sqrt{\frac{(n-1)\,p}{(n-p)} F_{p,(n-p)}(\alpha)} \sqrt{\frac{s_{jj} - 2s_{jk} + s_{kk}}{n}} \;;\; (\bar{Y}_j - \bar{Y}_k) + \sqrt{\frac{(n-1)\,p}{(n-p)} F_{p,(n-p)}(\alpha)} \sqrt{\frac{s_{jj} - 2s_{jk} + s_{kk}}{n}} \right)$$

Incorpora informação das covariâncias

Inferência sobre Componentes do Vetor de Médias

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

Região de Confiança

$$I.C.S(l'\mu)\text{ a }100(1-\alpha)\%=\left(l'\overline{Y}-\sqrt{\frac{(n-1)p}{(n-p)}F_{(p,n-p)}(\alpha)\frac{l'S_u l}{n}}\ ;l'\overline{Y}+\sqrt{\frac{(n-1)p}{(n-p)}F_{(p,n-p)}(\alpha)\frac{l'S_u l}{n}}\right)$$

Para tamanhos amostrais "grandes" (n $\to\infty$; n-p $\to\infty$)

$$\Rightarrow R(Y)=\left\{\mu\in\Re^{p};\ n\left(\overline{Y}-\mu\right)'S_u^{-1}\left(\overline{Y}-\mu\right)\leq\chi_p^2(\alpha)\right\}$$

$$\Rightarrow I.C.S.(l'\mu)\text{a }100(1-\alpha)\%=\left(l'\overline{Y}-\sqrt{\chi_p^2(\alpha)\ l'S_u l/n}\ ;l'\overline{Y}+\sqrt{\chi_p^2(\alpha)\ l'S_u l/n}\right)$$

Inferência sobre um Vetor de Médias

Taxas de açucar, sódio e potássio sangüíneas em 20 mulheres adultas

Indiv.	Açucar	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

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

$$\alpha=0{,}10\Rightarrow 8{,}18 \quad \alpha=0{,}05\Rightarrow 10{,}72$$

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

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

α=0,10⇒Rejeita H0

Calcule os intervalos de confiança simultâneos a 100(1-α)% para cada média:

$$\Rightarrow \left(\bar{Y}_k-\sqrt{\frac{19*3}{17}F_{3,17}(\alpha)\frac{s_{kk}}{20}}\;;\bar{Y}_k+\sqrt{\frac{19*3}{17}F_{3,(20-3)}(\alpha)\frac{s_{kk}}{20}}\right)$$

Inferência sobre um Vetor de Médias

Taxas de açucar, sódio e potássio sangüíneas em 20 mulheres adultas

Indiv.	Açucar	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

Intervalos de confiança simultâneos a 90%:

Para Açucar $\Rightarrow 4{,}64 \pm \sqrt{8{,}18\,\frac{2{,}879}{20}} = (3{,}56\,;\,5{,}73)$

Para Sódio $\Rightarrow 45{,}4 \pm \sqrt{8{,}18\,\frac{199{,}798}{20}} = (36{,}36\,;\,54{,}40)$

Para Potássio $\Rightarrow 9{,}97 \pm \sqrt{8{,}18\,\frac{3{,}628}{20}} = (8{,}75\,;\,11{,}19)$

Conclusão: As medias populacionais não diferem dos valores de referência MAS existe alguma combinação linear entre as medias, $l'\mu$, com diferença significante (a 10%).

Note que pode não ser de interesse encontrar tais çombinações lineares!

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

$(1-\alpha) = 0{,}95$, $p = 4, n = 15$ $\Rightarrow P(\text{coletivo}) = (0{,}95)^4 = 0{,}81$ $t_{14}(\alpha/2) = 2{,}145$

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

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

$$(1-\alpha) = 0{,}95, \quad n = 15, \quad p = 4 \Rightarrow \sqrt{\frac{(n-1)p}{(n-p)} F_{p,n-p}(\alpha)} = 4{,}14$$

mais largos

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)

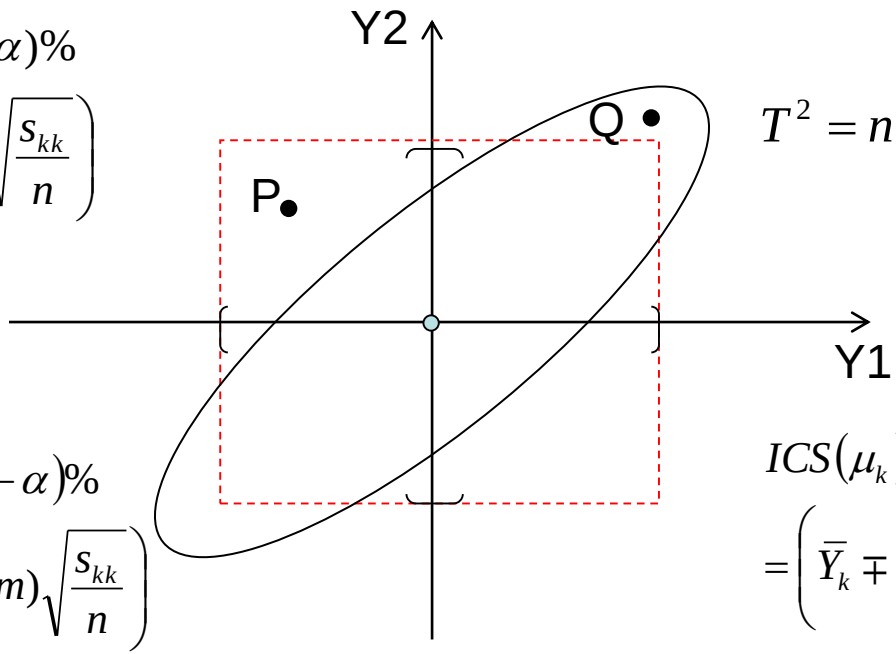
$\Rightarrow \alpha/2m$

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 multivariada (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 um Vetor de Médias

Um educador em Música testou 96 estudantes Finlandeses quanto às suas habilidades na música nativa. A média e desvio padrão dos escores obtidos estão apresentados na tabela a seguir.

	Score	
Aptidão Musical	**Média**	**d.p.**
Y1: Melodia	28,1	5,76
Y2: Harmonia	26,6	5,85
Y3: Tempo	35,4	3,82
Y4: Métrica	34,2	5,12
Y5: Verbalização	23,6	3,76
Y6: Balanço	22	3,93
Y7: Estilo	22,7	4,03

- Construa intervalos de confiança para os verdadeiros escores médios de cada tipo de aptidão $\Rightarrow$ considere os intervalos univariados e os intervalos simultâneos. Comente sobre os níveis de confiança coletivos em cada caso.

Inferência sobre um Vetor de Médias

Para grandes tamanhos amostrais:

$$\Rightarrow I.C.S.\text{ a } 90\% = \left(l'\bar{Y} - \sqrt{\chi^2_p(0{,}10)\ l'Sl/n}\ \ ; l'\bar{Y} + \sqrt{\chi^2_p(0{,}10)\ l'Sl/n}\ \right)$$

Aptidão Musical	Score		I.C. Simultâneo	
	Média	d.p.	L.I.	L.S.
Y1: Melodia	28,1	5,76	26,06	30,14
Y2: Harmonia	26,6	5,85	24,53	28,67
Y3: Tempo	35,4	3,82	34,05	36,75
Y4: Métrica	34,2	5,12	32,39	36,01
Y5: Verbalização	23,6	3,76	22,27	24,93
Y6: Balanço	22	3,93	20,61	23,39
Y7: Estilo	22,7	4,03	21,27	24,13

- O vetor de aptidões médias de um certo grupo de estudantes é: (31, 27, 34, 31, 23, 22, 22). Há evidência de que trata-se de estudantes Finlandeses?

Inferência sobre um Vetor de Médias
Comparações Múltiplas

Múltiplos testes independentes

$$P\left(\text{pelo menos uma Rej.}\right) = 1 - P\left(\text{nenhuma Rej.}\right) = 1 - \prod_{l=1}^{K} P\left(p_l > \alpha\right) = 1 - \left(1-\alpha\right)^K \overset{K\to\infty}{\approx} 1$$

```
B<-10000
minpval <- replicate(B,
min(runif(10000,0,1))<0.01)
mean(minpval>=1)
```

Procedimento de Sidak

$$P\left(\text{pelo menos uma Rej.}\right) = 1 - \left(1-t\right)^K = \alpha \quad \Rightarrow \quad t = 1 - \left(1-\alpha\right)^{1/K} \overset{K\to\infty}{\approx} 0$$

Inferência sobre um Vetor de Médias
Comparações Múltiplas

Correções para Múltiplos Testes

Rejeitar H_{0j} se:	Método
$p_{(k)} < \alpha / K$	Correção de Bonferroni para múltiplos testes
$p_{(j)} < \alpha /(K - j + 1)$ $para\ \ todo\ \ j = 1,...,k$	Correção de Holm (Controle "forte" da taxa de erro para os múltiplos testes)
$p_{(j)} < j\alpha / K$ $para\ \ a\lg um\ \ j \geq k$	Correção FDR (Taxa de Falsa Descoberta): Benjamini-Hochberg Controle menos conservador da taxa de erro para os múltiplos testes)

K: número total de testes α: nível de significância global fixado

$p_{(j)}$: nível descritivo (p-valor) ordenado, $p_{(1)} \leq p_{(2)} \leq ... \leq p_{(K)}$

Inferência sobre Vetores de Médias de Duas Populações

Generalizar os resultados do Caso Univariado para o Multivariado

Caso Univariado – Amostras Pareadas

Recordando o caso Univariado!

$$Y_{1n\times 1}, Y_{1i} \overset{iid}{\sim} N\left(\mu_1;\sigma_1^2\right); \quad Y_{2n\times 1}, Y_{2i} \overset{iid}{\sim} N\left(\mu_2;\sigma_2^2\right); \qquad \left(Y_{1i}\,;Y_{2i}\right), \quad i=1,2,\ldots,n$$

$$\Rightarrow D_i = Y_{1i} - Y_{2i} \overset{iid}{\sim} N_1\left(\mu_D = \mu_1 - \mu_2;\sigma_D^2\right), \quad i=1,2,\ldots,n \Rightarrow \overline{D} \sim N_1\left(\mu_D;\sigma_D^2/n\right)$$

$$\Rightarrow H_0 : \mu_D = 0; \quad \sigma_D^2 > 0 \Rightarrow \quad \overline{D} = \frac{1}{n}\sum_{i=1}^{n} D_i; \quad s_D^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(D_i - \overline{D}\right)^2; \quad t = \frac{\overline{D}}{s_D/\sqrt{n}} \sim t_{n-1}$$

Caso Univariado – Amostras independentes

$$Y_{1n_1\times 1}, Y_{1i} \overset{iid}{\sim} N\left(\mu_1;\sigma_1^2\right); \quad Y_{2n_2\times 1}, Y_{2i} \overset{iid}{\sim} N\left(\mu_2;\sigma_2^2\right); \quad Y_1 \perp Y_2$$

$$\Rightarrow \overline{D} = \overline{Y}_1 - \overline{Y}_2 \sim N_1\left(\mu_D = \mu_1 - \mu_2;\sigma_D^2 = \sigma_1^2/n_1 + \sigma_2^2/n_2\right)$$

$$\Rightarrow H_0 : \mu_D = 0 \quad ;\sigma_1^2 = \sigma_2^2 \quad \Rightarrow \quad \overline{D} = \overline{Y}_1 - \overline{Y}_2; \qquad t = \frac{\overline{D}}{s_c\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \sim t_{n_1+n_2-2}$$

$$s_c^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_1-2}$$

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

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

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

Uma População

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

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

Estatística LRT e UIT coincidem!

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

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

Inferência sobre Vetores de Médias de Duas Populações

Caso Multivariado – Amostra Pareada

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

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

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

$$\Rightarrow ICS(\delta_j) \text{ a } 100(1-\alpha)\% = \overline{D}_k \pm \sqrt{\frac{(n-1)p}{(n-p)} F_{p,(n-p)}(\alpha)} \sqrt{\frac{S_{Dkk}}{n}}$$

Intervalo de Confiança Simultâneo

$$\Rightarrow ICB(\delta_j) \text{ a } 100(1-\alpha)\% = \overline{D}_j \pm t_{n-1}(\alpha/2m) \sqrt{\frac{S_{Djj}}{n}}$$

Intervalo de Confiança com correção de Bonferroni

Teste de Hipóteses para Vetores de Médias
Duas Populações

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

	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

$$Y_{ig_{2\times1}} \overset{iid}{\sim} N_2\left(\mu_g;\Sigma_g\right); g = 1,2$$

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

mu1=(185.72 151.12)′
mu2=(183.84 149.24)′

mud=(1.88 1.88)′

$\mathbf{S_D}$

	Co	Pe
Co	56.78	11.98
Pe	11.98	29.28

$$H_0 : \mu_D = 0 \Rightarrow T^2 = 3,61 \sim \frac{24*2}{23} F_{2,23}$$

α=5%⇒3.42

Conclusão?

Inferência sobre Vetores de Médias de Duas Populações

Caso Multivariado - Amostras Independentes - Homocedasticidade

$$Y_{1n_1\times p} = \underbrace{\left(Y_{11}, Y_{12}, \ldots, Y_{1n_1}\right)'}; \quad Y_{1i} \overset{iid}{\sim} N_p\left(\mu_1; \Sigma_1\right); \quad Y_{2n_2\times p} = \underbrace{\left(Y_{11}, Y_{12}, \ldots, Y_{1n_2}\right)'}; \quad Y_{2i} \overset{iid}{\sim} N_p\left(\mu_2; \Sigma_2\right)$$

$$\bar{Y}_1, \quad S_1 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}\left(Y_{1i} - \bar{Y}_1\right)\left(Y_{1i} - \bar{Y}_1\right)' \qquad \bar{Y}_2, \quad S_2 = \frac{1}{n_2 - 1}\sum_{i=1}^{n_2}\left(Y_{2i} - \bar{Y}_2\right)\left(Y_{2i} - \bar{Y}_2\right)'$$

$$Y_1 \perp Y_2$$

$$\mu_D = \delta = \mu_1 - \mu_2$$

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

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

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

$$Y_{1n_1\times p} = \left(Y_{11}, Y_{12}, \ldots, Y_{1n_1}\right)'; \quad Y_{1i} \overset{iid}{\sim} N_p\left(\mu_1; \Sigma_1\right); \quad Y_{2n_2\times p} = \left(Y_{11}, Y_{12}, \ldots, Y_{1n_2}\right)'; \quad Y_{2i} \overset{iid}{\sim} N_p\left(\mu_2; \Sigma_2\right)$$

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

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

Hipótese condicional sob Homocedasticidade

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

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

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

Inferência sobre Vetores de Médias de Duas Populações

Caso Multivariado - Amostras Independentes - Homocedasticidade:

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

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

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

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

Os testes da LR e UI podem ser simplificados em função da estatística de Hotelling.

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

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

Inferência sobre Vetores de Médias de Duas Populações

Amostras Independentes:

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

S_c

Intervalos de Confiança Simultâneos

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

Intervalos de Confiança de Bonferroni (correção para múltiplos testes)

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

Inferência sobre Vetores de Médias de Duas Populações

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
mu1=	157.38	241.00	31.43	18.50
Mu2=	158.13	242.07	31.42	18.47

n1=21 n2=28

	X1	X2	X3	X4
mud=	-0.75	-1.07	0.01	0.03

S1

	X1	X2	X3	X4
X1	11.05	9.10	1.56	0.87
X2	9.10	17.50	1.91	1.31
X3	1.56	1.91	0.53	0.19
X4	0.87	1.31	0.19	0.18

S2

	X1	X2	X3	X4
X1	14.27	15.06	2.00	1.89
X2	15.06	34.35	3.12	3.07
X3	2.00	3.12	0.73	0.54
X4	1.89	3.07	0.54	0.51

Sc

	X1	X2	X3	X4
X1	17.83	17.31	2.50	2.01
X2	17.31	37.57	3.60	3.21
X3	2.50	3.60	0.90	0.54
X4	2.01	3.21	0.54	0.51

$$T^2 = 1.46 \sim \frac{(21+28-2)4}{(21+28-4-1)} F_{(4,(21+28-4-1))}$$

α=5% $\Rightarrow$ 2.58

Qual é a hipótese H0?
Conclusão?

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?

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

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

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

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

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

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

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

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

Conclusão?

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

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

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

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

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

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

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

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

Conclusão?

Teste da Igualdade de Matrizes de Covariância

Comparação de Vetores de Médias - Amostras Independentes

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

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

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

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

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

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

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

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

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

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

Caso Multivariado - Amostras Independentes – Heterocedasticidade

Teoria Assintótica

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

$$\overline{D}_{p\times 1}=\overline{Y}_1-\overline{Y}_2\sim N_p\left(\mu_D=\delta=\mu_1-\mu_2;\Sigma_{\overline{D}}=\left(\frac{\Sigma_1}{n_1}+\frac{\Sigma_1}{n_2}\right)\right)$$

$$\Rightarrow H_0:\delta=\delta_0\quad;\Sigma_g\in\Re^{p\times p},g=1,2$$

Hipótese condicional sob heterocedasticidade.

$$T^2=\left(\overline{D}-\delta_0\right)'\left(\frac{S_1}{n_1}+\frac{S_2}{n_2}\right)^{-1}\left(\overline{D}-\delta_0\right)\overset{\substack{n_1-p\to\infty\\ n_2-p\to\infty}}{\sim}\chi^2_p$$

Elipsóide de Confiança:

$$R\left(Y_1,Y_2\right)=\left\{\delta=\mu_1-\mu_2\in\Re^2;\left(\overline{D}-\delta\right)'\left(\frac{S_1}{n_1}+\frac{S_2}{n_2}\right)^{-1}\left(\overline{D}-\delta\right)\quad\le c^2_\alpha\right\}$$

Inferência sobre Vetores de Médias de Duas Populações

Caso Multivariado - Amostras Independentes

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

$$\overline{D}_{p\times 1}=\overline{Y}_1-\overline{Y}_2\sim N_p\left(\mu_D=\delta=\mu_1-\mu_2;\Sigma_{\overline{D}}=\left(\frac{\Sigma_1}{n_1}+\frac{\Sigma_1}{n_2}\right)\right)$$

$$\Rightarrow H_0:\delta=\delta_0\quad\times\quad H_1:\delta\neq\delta_0$$

Problema de Behrens-Fisher: Testar a igualdade dos vetores de médias "SEM" suposições sobre as matrizes de covariâncias

Estatística da Razão de verossimilhanças:

$$\text{Sob } H_1:\hat{\mu}_g=\overline{Y}_g,\ \hat{\Sigma}_g=S_g$$

$$\text{Sob } H_0:\hat{\mu}=\left(n_1\hat{\Sigma}_1^{-1}+n_2\hat{\Sigma}_2^{-1}\right)^{-1}\left(n_1\hat{\Sigma}_1^{-1}\overline{Y}_1+n_2\hat{\Sigma}_2^{-1}\overline{Y}_2\right);\quad\hat{\Sigma}_g=S_g+d_gd_g';\ d_g=\overline{Y}_g-\hat{\mu};$$

Algoritmo: (1) Estimativas iniciais $\hat{\Sigma}_g^0=S_g,\ g=1,2$
(2) Obtenha $\hat{\mu}^0$
(3) Usando $\hat{\mu}^0$ calcule $\hat{\Sigma}_g^1=S_g+(\overline{Y}_g-\hat{\mu}^0)(\overline{Y}_g-\hat{\mu}^0)';$
(4) Retorne ao passo (2) usando $\hat{\Sigma}_g^1$

Usar a distribuição assintótica da estatística λ para tomar decisão.

Inferência sobre Vetores de Médias de “Muitas” Populações

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

	População 1	**População 2**	...	**População G**
População estratificada	$N_p(\mu_1;\Sigma_1)$	$N_p(\mu_2;\Sigma_2)$		$N_p(\mu_G;\Sigma_G)$
Amostra	$Y_{11}, Y_{12}, \ldots, Y_{1n_1}$	$Y_{21}, Y_{22}, \ldots, Y_{2n_2}$		$Y_{G1}, Y_{G2}, \ldots, Y_{Gn_G}$
	$\overline{Y}_1 \quad S_1$	$\overline{Y}_2 \quad S_2$		$\overline{Y}_G \quad S_G$

$$\Rightarrow H_0 : \mu_1 = \mu_2 = \ldots = \mu_G = \mu \quad ; \Sigma_1 = \Sigma_2 = \ldots = \Sigma_G = \Sigma$$

Hipótese condicional de igualdade dos vetores de médias sob suposição de homocedasticidade. **MANOVA**

Inferência sobre Vetores de Médias de "Muitas" Populações

$\mathbf{Y}_{ig\ p\times 1} = \left(Y_{ig1}, Y_{ig2}, ..., Y_{igp}\right)'$: vetor de observações da unidade i no grupo g avaliada em p variáveis

Modelo distribucional adotado:

$$Y_{ig} \overset{iid}{\sim} N_p\left(\mu_g; \Sigma_g\right); \quad i = 1, 2, ..., n_g;\ g = 1, 2, ..., G;\ n = \sum_{g=1}^{G} n_g$$

Matriz aleatória

$$Y_{n\times p} \in \Re^{n\times p}; \quad Y_{n\times p} \sim N_{n\times p}\left(\mu_{n\times p} = \oplus_{g=1}^{G}\left(1_{n_g}\ \mu_g'\right); \Omega_{np\times np} = diag_{g=1}^{G}\left(I_{n_g} \otimes \Sigma_g\right)\right)$$

$$\Rightarrow \mu_{n\times p} = \begin{pmatrix} 1_{n_1}\ \mu_1' \\ 1_{n_2}\ \mu_2' \\ ... \\ 1_{n_G}\ \mu_G' \end{pmatrix} \qquad \Rightarrow \Omega = \begin{pmatrix} I_{n_1} \otimes \Sigma_1 & 0 & \cdots & 0 \\ 0 & I_{n_2} \otimes \Sigma_2 & \cdots & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & \cdots & I_{n_G} \otimes \Sigma_G \end{pmatrix}$$

Inferência sobre Vetores de Médias de "Muitas" Populações

$\mathbf{Y}_{ig\ p\times 1} = \left(Y_{ig1}, Y_{ig2}, ..., Y_{igp}\right)'$: vetor de observações da unidade i no grupo g

Modelo distribucional: $Y_{ig} \overset{iid}{\sim} N_p\left(\mu_g; \Sigma_g\right)$

$$Y_{n\times p} \in \Re^{n\times p};\quad Y_{n\times p} \overset{H_0}{\sim} N_{n\times p}\left(\mu_{n\times p} = \oplus_{g=1}^{G}\left(1_{n_g}\,\mu_g'\right); \Omega_{np\times np} = diag_{g=1}^{G}\left(I_{n_g} \otimes \Sigma_g\right)\right)$$

$$H_0 : \mu_1 = \mu_2 = ... = \mu_G;\quad \Sigma_g = \Sigma$$

$$Y_{n\times p} \in \Re^{n\times p};\quad Y_{n\times p} \overset{H_0}{\sim} N_{n\times p}\left(\mu_{n\times p} = 1_n\,\mu'; \Omega_{np\times np} = I_n \otimes \Sigma\right)$$

$$\Rightarrow \mu_{n\times p} = 1_n\,\mu' = \begin{pmatrix} \mu' \\ \mu' \\ ... \\ \mu' \end{pmatrix} \qquad \Rightarrow \Omega = I_n \otimes \Sigma = \begin{pmatrix} \Sigma & 0 & \cdots & 0 \\ 0 & \Sigma & \cdots & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & \cdots & \Sigma \end{pmatrix}$$

Inferência sobre Vetores de Médias de Muitas Populações

$$Y_{g\,n_g \times p}; \quad Y_{gi} \overset{iid}{\sim} N_p(\mu_g; \Sigma_g); \quad g = 1,...,G$$

Hipótese condicional

$$H_0 : \mu_1 = \mu_2 = ... = \mu_G; \quad \Sigma_g = \Sigma$$

EMVS sob H_0:

$$\Rightarrow \begin{cases} \bar{Y} = 1/n\, Y'1_n; \quad n = n_1 + ... + n_G \\ \hat{\Sigma} = S_{p\times p} = \dfrac{1}{n}\sum_{g=1}^{G}\sum_{i=1}^{n_g}\left(Y_{ig} - \bar{Y}\right)\left(Y_{ig} - \bar{Y}\right)' \end{cases}$$

Divisor n

EMVS sob H_1:

$$\Rightarrow \begin{cases} \bar{Y}_g; \quad g = 1,...,G \\ \hat{\Sigma} = S_c = \dfrac{n_1 S_1 + ... + n_G S_G}{n} \end{cases}$$

Divisor n

$nS = T$ Matriz de Soma de Quadrados e Produtos Cruzados TOTAL

Within

$nS_c = W$ Matriz de Soma de Quadrados e Produtos Cruzados DENTRO de GRUPOS

Between

$B = T - W = \sum_{g=1}^{G} n_g\left(\bar{Y}_g - \bar{Y}\right)\left(\bar{Y}_g - \bar{Y}\right)'$ Matriz de Soma de Quadrados e Produtos Cruzados ENTRE GRUPOS

Fontes de Variabilidade

Grupo	U.a.	Y_1	Y_2	...	Y_j	...	Y_p	
1	1	Y_{111}	Y_{112}	...	Y_{11j}	...	Y_{11p}	
		...		...				
1	n_1	$Y_{n_1 11}$	$Y_{n_1 12}$	...	$Y_{n_1 1j}$	...	$Y_{n_1 1p}$	
		$\bar{Y}_{.11}$	$\bar{Y}_{.12}$	...	$\bar{Y}_{.1j}$	...	$\bar{Y}_{.1p}$	$\bar{Y}_{.1.}$
2	1	Y_{121}	Y_{122}	...	Y_{12j}	...	Y_{12p}	
		...		...				
2	n_2	$Y_{n_2 21}$	$Y_{n_2 22}$	...	$Y_{n_2 2j}$	...	$Y_{n_1 2p}$	
		$\bar{Y}_{.21}$	$\bar{Y}_{.22}$	...	$\bar{Y}_{.2j}$	...	$\bar{Y}_{.2p}$	$\bar{Y}_{.2.}$
				...		...		
G	1	Y_{1G1}	Y_{1G2}	...	Y_{1Gj}	...	Y_{1Gp}	
		...		...				
G	n_G	$Y_{n_G G1}$	$Y_{n_G G2}$	...	$Y_{n_G Gj}$	...	$Y_{n_G Gp}$	
		$\bar{Y}_{.G1}$	$\bar{Y}_{.G2}$	...	$\bar{Y}_{.Gj}$	...	$\bar{Y}_{.Gp}$	$\bar{Y}_{.G.}$
								$\bar{Y}_{...}$

Variabilidade Dentro de Grupo

Variabilidade Entre Grupos

T = B + W

$$T_{p\times p} = nS = \sum_{g=1}^{G}\sum_{i=1}^{n_g}\left(Y_{ig}-\bar{Y}\right)\left(Y_{ig}-\bar{Y}\right)'$$

$$B_{p\times p} = \sum_{g=1}^{G} n_g\left(\bar{Y}_g-\bar{Y}\right)\left(\bar{Y}_g-\bar{Y}\right)'$$

$$W_{p\times p} = nS_c = \sum_{g=1}^{G}\sum_{i=1}^{n_g}\left(Y_{ig}-\bar{Y}_g\right)\left(Y_{ig}-\bar{Y}_g\right)'$$

Identidade útil (decomposição útil)

$$Y_{ig} = \bar{Y} + \left(\bar{Y}_g-\bar{Y}\right) + \left(Y_{ig}-\bar{Y}_g\right)$$

Inferência sobre Vetores de Médias de Muitas Populações

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

$$H_0 : \mu_1 = \mu_2 = ... = \mu_g; \quad \Sigma_g = \Sigma$$

Estatística da Razão de Verossimilhanças (Estatística de Wilks):

$$\Lambda = \left\{ \frac{|nS_c|}{|nS|} \right\}^{n/2} \Rightarrow \Lambda^{2/n} = \left\{ \frac{|nS_c|}{|nS|} \right\} = \left\{ \frac{|W|}{|T|} \right\} = \left| T^{-1}W \right| = \left\{ \frac{|W|}{|B+W|} \right\} = \left| I + W^{-1}B \right|^{-1} = \prod_{j=1}^{p} \left(1 + \lambda_j \right)^{-1}$$

baseada nos autovalores de $W^{-1}B$

Teste da União-Intersecção:

Estatística da ANOVA descrita por combinações lineares do $\Re^p$

Os testes LR e UI conduzem a estatísticas diferentes mas se baseiam na mesma decomposição espectral

$$\frac{SQEntre}{SQDentro} = \frac{\sum_{g=1}^{G} n_g \left(\bar{Y}_g - \bar{Y} \right)^2}{\sum_{g=1}^{G} \sum_{i=1}^{n_g} \left(Y_{ig} - \bar{Y}_g \right)^2} \Rightarrow \frac{\sum_{g=1}^{G} n_g l' \left(\bar{Y}_g - \bar{Y} \right) \left(\bar{Y}_g - \bar{Y} \right)' l}{\sum_{g=1}^{G} n_g l' S_g l} = \frac{l' \left[\sum_{g=1}^{G} n_g \left(\bar{Y}_g - \bar{Y} \right) \left(\bar{Y}_g - \bar{Y} \right)' \right] l}{l' \left[\sum_{g=1}^{G} n_g S_g \right] l} \leq \lambda_1$$

$=B$ $=W$

Lembrando "ANOVA":

Para introduzirmos MANOVA

Delineamento Completamente Aleatorizado

$N(\mu_1;\sigma^2)$ $N(\mu_2;\sigma^2)$ ... $N(\mu_G;\sigma^2)$ **População**

T_1	T_2	. . .	T_G	**Amostra**
Y_{11}	Y_{21}	. . .	Y_{G1}	
. . .	. . .	Y_{ij}	. . .	
Y_{1n_1}	$Y_{2\,n_2}$	. . .	$Y_{G\,n_G}$	
n_1	n_2	. . .	n_G	
$\bar{Y}_1$	$\bar{Y}_2$	. . .	$\bar{Y}_G$	
s_1	s_2	. . .	s_G	

- ✓ **Normalidade**
- ✓ **Variância constante**
- ✓ **Independência**

Formalização Matricial
Modelo ANOVA - Modelo Linear Geral

Modelo Estrutural: $y_{ig} = \mu_g + e_{ig}$

Resposta da observação i do grupo g

Parametrização de Médias

$$= \mu + \tau_g + e_{ig} \quad ; \quad \sum_{g=1}^{G} \tau_g = 0$$

Parametrização de Desvios

$$= \begin{cases} \mu_1 \\ \mu_1 + \tau_g ; g = 2,...,G \end{cases}$$

Parametrização de Casela de Referência

Forma matricial:

$$Y_{n\times 1} = X_{n\times G} \beta_{G\times 1} + e_{n\times 1}$$

vetor de observações | Matriz de Planejamento | vetor de parâmetros | vetor de erros

$$Y_{ig} = \mu + \tau_g + e_{ig}; \quad \sum_g \tau_g = 0$$

G=4 grupos com 5 observações por grupo

Usando a parametrização de desvios

$$Y_{n\times 1} = X_{n\times G}\,\beta_{G\times 1} + e_{n\times 1}$$

$$\begin{bmatrix} Y_{11} \\ Y_{21} \\ Y_{31} \\ Y_{41} \\ Y_{51} \\ Y_{12} \\ Y_{22} \\ Y_{32} \\ Y_{42} \\ Y_{53} \\ Y_{13} \\ Y_{23} \\ Y_{33} \\ Y_{43} \\ Y_{53} \\ Y_{14} \\ Y_{24} \\ Y_{34} \\ Y_{44} \\ Y_{54} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & -1 & -1 & -1 \\ 1 & -1 & -1 & -1 \\ 1 & -1 & -1 & -1 \\ 1 & -1 & -1 & -1 \\ 1 & -1 & -1 & -1 \end{bmatrix} \begin{bmatrix} \mu \\ \tau_1 \\ \tau_2 \\ \tau_3 \end{bmatrix} + \begin{bmatrix} e_{11} \\ e_{21} \\ e_{31} \\ e_{41} \\ e_{51} \\ e_{12} \\ e_{22} \\ e_{32} \\ e_{42} \\ e_{53} \\ e_{13} \\ e_{23} \\ e_{33} \\ e_{43} \\ e_{53} \\ e_{14} \\ e_{24} \\ e_{34} \\ e_{44} \\ e_{54} \end{bmatrix}$$

Fontes de Variação

Considere a seguinte identidade (decomposição útil para se obter as fontes de variação envolvidas no modelo):

$$y_{ig} = \bar{y} + (\bar{y}_g - \bar{y}) + (y_{ig} - \bar{y}_g)$$

$$\underbrace{y_{ig} - \bar{y}}_{\text{SQTotal corrigida}} = \underbrace{(\bar{y}_g - \bar{y})}_{\text{SQTratamento "Entre"}} + \underbrace{(y_{ig} - \bar{y}_g)}_{\text{SQResidual "Dentro"}}$$

$$\sum_{g,i} \left(y_{ig} - \bar{y}\right)^2 \qquad \sum_{g} n_g \, (\bar{y}_g - \bar{y})^2 \qquad \sum_{g,i} (y_{ig} - \bar{y}_g)^2$$

Tabela de ANOVA

$$H_0 : \mu_1 = \mu_2 = ... = \mu_G = \mu$$

F.V.	g l	SQ	QM	F	valor-p
ENTRE	**G-1**	$\sum_g n_g (\bar{y}_g - \bar{y})^2$	**SQEntre/(G-1)**	**QME/QMD**	
DENTRO	**n-G**	$\sum_{g,i} (y_{ig} - \bar{y}_g)^2$	**SQDentro/(n-G)**		
TOTAL	**n-1**	$\sum_{g,i} (y_{ig} - \bar{y})^2$			

$$\mathbf{F} = \frac{QMEntre}{QMDentro} \sim \mathbf{F}\ (\mathbf{G\text{-}1}, \mathbf{n\text{-}G})$$

Sob: $Y_{ig} \overset{iid}{\sim} N_1\left(\mu_g; \sigma^2\right)$

normalidade
homocedasticidade
independência

Modelo MANOVA

DCA:Delineamento Completamente Aleatorizado (1 fator em G níveis)

$\mathbf{Y}_{ig\ p\times 1} = \left(Y_{ig1}, Y_{ig2}, ..., Y_{igp}\right)'$: vetor de observações da unidade i no tratamento g

Efeito Fixo

Aleatório

Modelo estrutural: $Y_{ig\ p\times 1} = \mu + \tau_g + e_{ig} \quad ; \quad \sum_{g=1}^{G} \tau_g = 0$

Modelo distribucional: $e_{ig} \overset{iid}{\sim} N_p\left(0; \Sigma\right) \Rightarrow Y_{ig} \overset{iid}{\sim} N_p\left(\mu_g; \Sigma\right)$

Suposições: independência entre observações (tanto Entre grupos como Dentro de grupo), Distribuição Normal p-variada, Matriz de Covariâncias homogênea

Modelo MANOVA

$$\mathrm{Y}_{ig\ p\times 1} = \left(Y_{ig1}, Y_{ig2}, \ldots, Y_{igp}\right)' \qquad Y_{ig\ p\times 1} = \mu + \tau_g + e_{ig} \quad ; \quad \sum_{g=1}^{G} \tau_g = 0 \qquad e_{ig} \overset{iid}{\sim} N_p\left(0; \Sigma\right)$$

Identidade útil para descrever as fontes de variação:

$$y_{ig\ p\times 1} = \bar{y} + \left(\bar{y}_g - \bar{y}\right) + \left(y_{ig} - \bar{y}_g\right)$$

$$H = \sum_{g=1}^{G} n_g \left(\bar{y}_g - \bar{y}\right)\left(\bar{y}_g - \bar{y}\right)'$$

matriz de SQPC devido ao efeito do tratamento (Entre Grupos) - Notação: H=B

$$E = \sum_{g=1}^{G}\sum_{i=1}^{n_g}\left(y_{ig} - \bar{y}_g\right)\left(y_{ig} - \bar{y}_g\right)' = \left(n_1 - 1\right)S_{u1} + \ldots + \left(n_G - 1\right)S_{uG}$$

: matriz de SQPC devido ao erro (Dentro de Grupos)

Notação: E=W

$$H + E = \sum_{g=1}^{G}\sum_{i=1}^{n_g}\left(y_{ig} - \bar{y}\right)\left(y_{ig} - \bar{y}\right)'$$

: matriz de SQPC total corrigida pela média

Notação: H+E=T

Tabela de MANOVA

$$H:\ \mu_1=\mu_2=\ldots=\mu_G=\mu \quad \Leftrightarrow \quad H:\tau_1=\tau_2=\ldots=\tau_G=0$$

F.V.	g.l.	Matriz de SQPC
Trat	**G-1**	$H_{p\times p}=\sum_{g=1}^{G} n_g\ (\bar{\mathbf{y}}_g-\bar{\mathbf{y}})(\bar{\mathbf{y}}_g-\bar{\mathbf{y}})'$
Resíduo	**n-G**	$E=\sum_{g=1}^{G}\sum_{i=1}^{n_g}(y_{ig}-\bar{y}_g)(y_{ig}-\bar{y}_g)'$
TOTAL	**n-1**	$H+E=\sum_{g=1}^{G}\sum_{i=1}^{n_g}(\mathbf{y}_{gi}-\bar{\mathbf{y}})(\mathbf{y}_{gi}-\bar{\mathbf{y}})'$

$$\Lambda^{*}=\frac{|E|}{|H+E|}=\left|T^{-1}W\right|=\left|I+W^{-1}B\right|^{-1}$$

Estatística lambda de Wilks (critério baseado na RV sob normalidade multivariada, independência e homocedasticidade)

Distribuição da Estatística Λ^*

# Var.	# Grupos	Distribuição Amostral (sob $Y_{ig} \overset{iid}{\sim} N_p(\mu_g; \Sigma)$)
p = 1	g ≥ 2	$\left(\frac{N-g}{g-1}\right)\left(\frac{1-\Lambda^*}{\Lambda^*}\right) \sim F_{g-1,N-g}$
p = 2	g ≥ 2	$\left(\frac{N-g-1}{g-1}\right)\left(\frac{1-\sqrt{\Lambda^*}}{\sqrt{\Lambda^*}}\right) \sim F_{2(g-1),2(N-g-1)}$
p ≥ 1	g = 2	$\left(\frac{N-p-1}{p}\right)\left(\frac{1-\Lambda^*}{\Lambda^*}\right) \sim F_{p,N-p-1}$
p ≥ 1	g =3	$\left(\frac{N-p-2}{p}\right)\left(\frac{1-\sqrt{\Lambda^*}}{\sqrt{\Lambda^*}}\right) \sim F_{2p,2(N-p-2)}$

Caso assintótico: $-\left(N-1-\frac{p+g}{2}\right) \ln\left(\frac{|E|}{|H+E|}\right) \overset{\substack{n\to\infty \\ (n-p)\to\infty}}{\sim} \chi^2_{p(g-1)}(\alpha)$

Exemplo

Suponha que duas variáveis são avaliadas em unidades amostrais submetidas a 3 tratamentos (T1, T2 e T3)

T1		T2		T3	
Y11	**Y12**	**Y21**	**Y22**	**Y31**	**Y32**
9	3	0	4	3	8
6	2	2	0	1	9
9	7			2	7
8	4	1	2	2	8
Média geral = (4 , 5)					

Exemplo didático (número de réplicas muito pequeno)

Teste se existe efieto de tratamento considerando o vetor de resposta bivariada.

Faça suposições necessárias para a validade do teste de hipótese adotado.

Tabela de MANOVA

Qual é o modelo estrutural da MANOVA?

$$H:\ \boldsymbol{\mu}_1=\boldsymbol{\mu}_2=\boldsymbol{\mu}_3=\boldsymbol{\mu} \quad \Leftrightarrow \quad H:\boldsymbol{\tau}_1=\tau_2=\boldsymbol{\tau}_3=0$$

F.V.	g.l.	Matriz de Soma de Quadr e Prod Cruzados
Trat	**3-1**	$H=\begin{pmatrix}78 & -12\\ -12 & 48\end{pmatrix}$ ← SQ da ANOVA
Resíduo	**8-3**	$E=\begin{pmatrix}10 & 1\\ 1 & 24\end{pmatrix}$ ← Soma de Produtos Cruzados
TOTAL	**8-1**	$H+E=\begin{pmatrix}88 & -11\\ -11 & 72\end{pmatrix}$

$$\Lambda^{*}=\frac{|E|}{|H+E|}=\frac{10(24)-1}{88(72)-(-11)^2}=0{,}0385 \quad \Rightarrow \underbrace{\left(\frac{N-p-2}{p}\right)\left(\frac{1-\sqrt{\Lambda^{*}}}{\sqrt{\Lambda^{*}}}\right)}_{8{,}19} \sim F_{2p,2(N-p-2)}(\alpha=0{,}01)$$

7,01

Concl.?

Estatísticas Multivariadas

Critério	Estatística	Aproximação F
Wilks	$\Lambda^* = \frac{\lvert E \rvert}{\lvert H+E \rvert} = \prod_i \frac{1}{1+\lambda_i}$	$\left(\frac{rt-2f}{pq}\right)\left(\frac{1-\Lambda^{1/t}}{\Lambda^{1/t}}\right) \sim F_{pq,(rt-2f)}$
Traço de Pillai	$V = tr\left(\frac{H}{H+E}\right) = \sum_i \frac{\lambda_i}{1+\lambda_i}$	$\left(\frac{2n+s+1}{2m+s+1}\right)\left(\frac{V}{s-V}\right) \sim F_{s(2m+s+1),s(2n+s+1)}$
Traço de Hotelling Lawley	$U = tr\left(\frac{H}{E}\right) = \sum_i \lambda_i$	$\frac{2(sn+1)U}{s^2(2m+s+1)} \sim F_{s(2m+s+1),2(sn+1)}$
Raiz Máxima de Roy	$\theta = \lambda_1$	$\frac{(\nu-d+q)\theta}{d} \sim F_{d,(\nu-d+q)}$

p= # de var.; q=g.l. trat (ou do contraste); ν=g.l. erro; s=min(p,q); r=(p+q+1)/2; f=(pq-2)/4

d= max(p,q); m=(|p-q|-1)/2; n=(ν-p-1)/2; λ: autovalor de |H-λE|=0;

$t = \sqrt{(p^2q^2-4)/(p^2+q^2-5)}$ se $(p^2+q^2-5) > 0$, ou 1 c.c.

Contribuição das Variáveis na Discriminação dos Grupos

Tabela de MANOVA

F.V.	g.l.	Matriz de SSCP
Trat	**G-1**	$H = \sum_{i=1}^{g} n_i \left(\bar{\mathbf{y}}_i - \bar{\mathbf{y}}\right)\left(\bar{\mathbf{y}}_i - \bar{\mathbf{y}}\right)'$
Resíduo	**n-G**	$E = \sum_{i=1}^{g}\sum_{j=1}^{n_i} \left(\mathbf{y}_{ij} - \bar{\mathbf{y}}_i\right)\left(\mathbf{y}_{ij} - \bar{\mathbf{y}}_i\right)'$
TOTAL	**n-1**	$H + E = \sum_{i=1}^{g}\sum_{j=1}^{n_i} \left(\mathbf{y}_{ij} - \bar{\mathbf{y}}\right)\left(\mathbf{y}_{ij} - \bar{\mathbf{y}}\right)'$

Considere a seguinte decomposição espectral:

$$\left|H - \lambda E\right| = 0;\ \left(H - \lambda E\right)l = 0 \Rightarrow \frac{l_k' H l_k}{l_k' E l_k} = \lambda_k$$

$$l_k = \left(l_{k1}\ \ l_{k2} \ldots l_{kp}\right)'$$

Notação: H=B; E=W
A informação sobre discriminação entre os grupos está na decomposição espectral de $W^{-1}B$

A avaliação dos coeficientes dos autovetores l, associados aos maiores autovalores, define a importância de cada variável no efeito de tratamento. ***Este resultado deve ser discutido no contexto da Análise Discriminante.***

Comparações Múltiplas

Comparações múltiplas entre tratamentos dois a dois

(com correção de Bonferroni)

Comparacão dos Tratamentos g e *h*: $\underbrace{\mu_g - \mu_h}_{\tau_g - \tau_h} \Rightarrow \bar{Y}_g - \bar{Y}_h$

Avaliar os componentes do vetor

$$\hat{\tau}_{g_{p\times 1}} = \bar{Y}_g - \bar{Y} \Rightarrow \underbrace{\hat{\tau}_{g\,j} = \bar{Y}_{g\,j} - \bar{Y}_j}_{\text{Trat g Variável j}} \qquad \hat{\tau}_h = \bar{Y}_h - \bar{Y} \Rightarrow \underbrace{\hat{\tau}_{h\,j} = \bar{Y}_{h\,j} - \bar{Y}_j}_{\text{Trat } h \text{ Variável j}}$$

$$V\left(\bar{Y}_{g\,j} - \bar{Y}_{h\,j}\right) = \left(\frac{1}{n_g} + \frac{1}{n_h}\right)\frac{E_{jj}}{n-G} \Rightarrow \left(\bar{Y}_{g\,j} - \bar{Y}_{h\,j}\right) \pm t_{n-G}\left(\alpha / pG(G-1)\right)\sqrt{V\left(\bar{Y}_{g\,j} - \bar{Y}_{h\,j}\right)}$$

QMRes
Diag(S_{uc})

Intervalo de confiança 100(1-α)% com Correção de Bonferroni para um total de (p + G(G-1)/2) comparações

Exemplo

Respostas de produtividade (P) em kg/ha e número de grãos por vagem (#GV) para 4 variedades de feijão (A, B, C e D) plantadas em 5 canteiros

	Cultivar							
	A		**B**		**C**		**D**	
Réplica	**P**	**#GV**	**P**	**#GV**	**P**	**#GV**	**P**	**#GV**
1	1082	4,66	1163	5,52	1544	5,18	1644	5,45
2	1070	4,5	1100	5,3	1500	5,1	1600	5,18
3	1180	4,3	1200	5,42	1550	5,2	1680	5,18
4	1050	4,7	1190	5,62	1600	5,3	1700	5,4
5	1080	4,6	1170	5,7	1540	5,12	1704	5,5
Média	1092,4	4,55	1164,6	5,51	1546,8	5,18	1665,6	5,34
S	2558,8		1525,8		1271,2		1908,8	
	-7,23	0,0255	3,55	0,0251	2,65	0,0062	3,51	0,0231

$\Rightarrow$ Compare as 4 variedades de feijão relativamente às variáveis produtividade e número de grãos por vagem.

$\Rightarrow$ Escreva o modelo estrutural e distribucional que pode ser adotado na análise destes dados.

MANOVA

Means

Varied	N	P	NGV
1	5	1092,4	4,5520
2	5	1164,6	5,5120
3	5	1546,8	5,1800
4	5	1665,6	5,3420

$$H_0 : \mu_A = \mu_B = \mu_C = \mu_D = \mu_{2\times1}$$

MANOVA for Varied

s = 2 m = 0,0 n = 6,5

Criterion	Test Statistic	F	DF Num	DF Denom	P
Wilks'	0,00308	85,166	6	30	0,000
Lawley-Hotelling	48,02582	112,060	6	28	0,000
Pillai's	1,84617	64,006	6	32	0,000
Roy's	41,34640				

Concl.: Rejeitar H

⇒ Calcule os intervalos de confiança simultâneos

SSCP Matrix for Varied

	P	NGV
P	1189302	768,361
NGV	768	2,632

SSCP Matrix for Error

	P	NGV
P	29058,4	9,9040
NGV	9,9	0,3198

Obtenha a tabela de MANOVA!

Intervalos de Confiança com correção de Bonferroni

Construa os intervalos e conclua sobre as diferenças entre os grupos.

	P			#GV		
Contraste	**LI**	**LS**	**Concl**	**LI**	**LS**	**Concl**
A-B	-153.17	8.77	$\mu_A = \mu_B$	-1.23	-0.69	$\mu_A < \mu_B$
A-C	-535.37	-373.43	$\mu_A < \mu_C$	-0.90	-0.36	$\mu_A < \mu_C$
A-D	-654.17	-492.23	$\mu_A < \mu_D$	-1.06	-0.52	$\mu_A < \mu_D$
B-C	-463.17	-301.23	$\mu_B < \mu_C$	0.06	0.60	$\mu_B > \mu_C$
B-D	-581.97	-420.03	$\mu_B < \mu_D$	-0.10	0.44	$\mu_B > \mu_D$
C-D	-199.77	-37.83	$\mu_C < \mu_D$	-0.43	0.11	$\mu_C > \mu_D$

Correção para os múltiplos testes de cada variável

$$V\left(\bar{Y}_{ik} - \bar{Y}_{hk}\right) = \left(\frac{1}{5} + \frac{1}{5}\right)\frac{E_{kk}}{20-4} \Rightarrow \quad \left(\bar{Y}_{ik} - \bar{Y}_{hk}\right) \pm t_{20-4}\left(\alpha/(2*(4(4-1)/2)\right)\sqrt{V\left(\bar{Y}_{ik} - \bar{Y}_{hk}\right)}$$

```
> E11 <- 29058.4
> E22 <- 0.3198
> qt(c(1-0.0042), df=16, lower.tail=TRUE)
[1] 3.004515
```

MANOVA: Modelo Linear Multivariado

$$Y_{n\times p} = X_{n\times G}\,\beta_{G\times p} + \varepsilon_{n\times p}$$

$$\mathbf{Y}_{n\times p} = \begin{pmatrix} Y_{11} & Y_{12} & & Y_{1p} \\ Y_{21} & Y_{22} & & Y_{2p} \\ \dots & \dots & \dots & \\ Y_{n1} & Y_{n2} & & Y_{np} \end{pmatrix}; \qquad \boldsymbol{\varepsilon}_{n\times p} = \begin{pmatrix} \varepsilon_{11} & \varepsilon_{12} & & \varepsilon_{1p} \\ \varepsilon_{21} & \varepsilon_{22} & & \varepsilon_{2p} \\ \dots & \dots & \dots & \\ \varepsilon_{n1} & \varepsilon_{n2} & & \varepsilon_{np} \end{pmatrix}.$$

Parametrização de médias

$$X_{n\times G} = \begin{pmatrix} 1_{n_1} & 0 & 0 & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 1_{n_2} & 0 & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & 1_{n_G} \end{pmatrix}; \qquad \boldsymbol{\beta}_{G\times p} = \begin{pmatrix} \mu_{11} & \mu_{12} & & \mu_{1p} \\ \mu_{21} & \mu_{22} & & \mu_{2p} \\ \dots & \dots & \dots & \\ \mu_{G1} & \mu_{G2} & & \mu_{Gp} \end{pmatrix};$$

Parametrização de desvios

$$X_{n\times G} = \begin{pmatrix} 1 & 1 & 0 & 0 \\ \dots & \dots & \dots & \dots \\ 1 & 0 & 1 & 0 \\ \dots & \dots & \dots & \dots \\ 1 & 0 & 0 & 1 \\ \dots & \dots & \dots & \dots \\ 1 & -1 & -1 & -1 \end{pmatrix}; \qquad \boldsymbol{\beta}_{G\times p} = \begin{pmatrix} \mu_{.1} & \mu_{.2} & & \mu_{.p} \\ \tau_{11} & \tau_{12} & & \tau_{1p} \\ \dots & \dots & \dots & \\ \tau_{(G-1)1} & \tau_{(G-1)2} & & \tau_{(G-1)p} \end{pmatrix};$$

MANOVA: Modelo Linear Multivariado

$$Y_{n\times p} = X_{n\times G}\ \beta_{G\times p} + \varepsilon_{n\times p}$$

Estimadores de Mínimos Quadrados e de MVS

$$\hat{\beta} = \left(X'X\right)^{-1} X'Y \qquad \hat{Y} = X\hat{\beta} = X\left(X'X\right)^{-1} X'Y = PY$$

$$\hat{\varepsilon} = Y - \hat{Y} = \left(I_n - X\left(X'X\right)^{-1} X'\right)Y = \left(I_n - P\right)Y \qquad \hat{\varepsilon}'\hat{\varepsilon}/n = \hat{\Sigma}$$

$$P = X\left(X'X\right)^{-1} X'$$

MANOVA: Modelo Linear Multivariado

$$Y_{n\times p} = X_{n\times G}\ \beta_{G\times p} + \varepsilon_{n\times p}$$

Teste de Hipóteses Gerais

$$H_0 : C_{c\times G}\beta_{G\times p}U_{p\times u} = 0$$

MANOVA.RM_R
Friedrich, S., Konietschke, F. and Pauly, M. (2018)

$C_{c\times G}$: define contrastes entre as médias de grupos

$U_{p\times u}$: define contrastes entre as médias das variáveis

Estatísticas de Teste: Wilks, Pillai, Lawley-Hotelling, Roy

$$\text{Lambda de Wilks} : \lambda = \frac{|E|}{|H+E|}$$

Considerar os autovalores e autovetores de : $(H - \lambda E)l = 0$

$$H = \left(C\hat{\beta}U\right)'\left[C\left(X'X\right)^{-1}C'\right]^{-1}C\hat{\beta}U \qquad E = \left(YU\right)'\left[I - X\left(X'X\right)^{-1}X'\right]^{-1}YU$$

MANOVA: Modelo Linear Multivariado

$$Y_{n\times p} = X_{n\times G}\,\beta_{G\times p} + \varepsilon_{n\times p}$$

Teste de Hipóteses Gerais $H_0 : C_{c\times G}\beta_{G\times p}U_{p\times u} = 0$

Exemplo: Considere um DCA balanceado e a parametrização de médias. Os seguintes ***Contrastes Ortogonais*** podem ser definidos (para comporem as linhas da matriz C): $C_{l_1}' C_{l_2} = 0$

G=3 Tratamentos

C1 = (1 0 -1)

C2 = (½ -1 ½)

G=4 Tratamentos

C1 = (1 -1/3 -1/3 -1/3)

C2 = (0 -1/2 -1/2 1)

C3 = (0 1 -1 0)

Note que o número de graus de liberdade para estudar o efeito de Tratamento é (G-1): neste caso, a estatística do teste coletivo das hipóteses é a soma das estatísticas dos testes individuais!

Delineamento Fatorial

Um estudo tem como objetivo avaliar as condições de fabricação de um filme plástico. Três variáveis resposta (Y1, Y2 e Y3) foram observadas sob dois níveis (baixo e alto) dos fatores F1 e F2.

Dados do Arquivo EXH

	F2					
	Baixo			Alto		
F1	Y1	Y2	Y3	Y1	Y2	Y3
Baixo	6,5	9,5	4,4	6,9	9,1	5,7
	6,2	9,9	6,4	7,2	10	2
	5,8	9,6	3	6,9	9,9	3,9
	6,5	9,6	4,1	6,1	9,5	1,9
	6,5	9,2	0,8	6,3	9,4	5,7
Alto	6,7	9,1	2,8	7,1	9,2	8,4
	6,6	9,3	4,1	7	8,8	5,2
	7,2	8,3	3,8	7,2	9,7	6,9
	7,1	8,4	1,6	7,5	10,1	2,7
	6,8	8,5	3,4	7,6	9,2	1,9

$\Rightarrow$ Realize uma análise de Variância Multivariada destes dados.

Delineamento Fatorial

Caso Univariado - ANOVA

$$y_{ijk} = \mu + \tau_j + \beta_k + \gamma_{jk} + \varepsilon_{ijk}$$

Resposta da observação i avaliada no nível j do fator 1 e nível k do fator 2

Restrições de identificabilidade dos parâmetros

$$\sum_{j=1}^{a} \tau_j = 0, \sum_{k=1}^{b} \beta_k = 0, \sum_{j=1}^{a} \gamma_{jk} = 0, \sum_{k=1}^{b} \gamma_{jk} = 0$$

"Identidade útil" para obtenção das Somas de Quadrados e dos estimadores dos efeitos de interesse:

$$y_{ijk} = \bar{y} + (\bar{y}_{j.} - \bar{y}) + (\bar{y}_{.k} - \bar{y}) + (\bar{y}_{jk} - \bar{y}_{j.} - \bar{y}_{.k} + \bar{y}) + (y_{ijk} - \bar{y}_{ijk})$$

Ef. principal de F1 | Ef. principal de F2 | Ef. de interação | Resíduo

SQ_F1 | SQ_F2 | SQ_F1*F2 | SQ_Residual

Caso Multivariado $\Rightarrow$ descrever os resultados para o vetor de resposta p-dimensional.

Tabela de MANOVA

Delineamento Completamente Aleatorizado com estrutura Fatorial de Grupos (G=ab) e r réplicas (balanceado)

F.V.	g.l.	Matriz de SSCP
Fator 1	**a-1**	$HF1 = \sum_{j=1}^{a} br \left(\bar{\mathbf{y}}_{j.} - \bar{\mathbf{y}}\right)\left(\bar{\mathbf{y}}_{j.} - \bar{\mathbf{y}}\right)'$
Fator 2	**b-1**	$HF2 = \sum_{k=1}^{b} ar \left(\bar{\mathbf{y}}_{.k} - \bar{\mathbf{y}}\right)\left(\bar{\mathbf{y}}_{.k} - \bar{\mathbf{y}}\right)'$
Interação	**(a-1)(b-1)**	$HInt = \sum_{j=1}^{a}\sum_{k=1}^{b} r \left(\bar{\mathbf{y}}_{jk} - \bar{\mathbf{y}}_{j.} - \bar{\mathbf{y}}_{.k} + \bar{\mathbf{y}}\right)\left(\bar{\mathbf{y}}_{jk} - \bar{\mathbf{y}}_{j.} - \bar{\mathbf{y}}_{.k} + \bar{\mathbf{y}}\right)'$
Resíduo	**ab(r-1)**	$E = \sum_{j=1}^{a}\sum_{k=1}^{b}\sum_{i=1}^{r}\left(\mathbf{y}_{ijk} - \bar{\mathbf{y}}_{jk}\right)\left(\mathbf{y}_{ijk} - \bar{\mathbf{y}}_{jk}\right)'$
TOTAL	**rab-1**	$HF1 + HF2 + HInt + E = \sum_{j=1}^{a}\sum_{k=1}^{b}\sum_{i=1}^{r}\left(\mathbf{y}_{ijk} - \bar{\mathbf{y}}\right)\left(\mathbf{y}_{ijk} - \bar{\mathbf{y}}\right)'$

Tabela de MANOVA

F.V.	Estatística Multivariada	Distribuição (Bartlet)
Interação	$\Lambda^*_{Int} = \frac{\lvert E \rvert}{\lvert HInt + E \rvert}$	$-\left(ab(r-1) - \frac{p+1-(a-1)(b-1)}{2} \right) \ln \Lambda^*_{Int} \sim \chi^2_{(a-1)(b-1)p}$
Fator 1	$\Lambda^*_{F1} = \frac{\lvert E \rvert}{\lvert HF1 + E \rvert}$	$-\left(ab(r-1) - \frac{p+1-(a-1)}{2} \right) \ln \Lambda^*_{F1} \sim \chi^2_{(a-1)p}$
Fator 2	$\Lambda^*_{F2} = \frac{\lvert E \rvert}{\lvert HF2 + E \rvert}$	$-\left(ab(r-1) - \frac{p+1-(b-1)}{2} \right) \ln \Lambda^*_{F2} \sim \chi^2_{(b-1)p}$

Testar a interação com os efeitos principais no modelo!

Testar os efeitos principais somente sob *inexistência de interação* (modelo aditivo)!

Delineamento Aleatorizado em Blocos Completos

Considere os dados de fabricação de um filme plástico: três variáveis (Y1, Y2 e Y3) foram observadas sob dois níveis (baixo e alto) de regulação das Máquinas Maq1 e Maq2. Os materiais de filme plástico estão blocados de acordo com o fornecedor.

Maq1	Baixo						Alto					
Maq2	Baixo			Alto			Baixo			Alto		
Bloco	Y1	Y2	Y3	Y1	Y2	Y3	Y1	Y2	Y3	Y1	Y2	Y3
1	6,5	9,5	4,4	6,9	9,1	5,7	6,7	9,1	2,8	7,1	9,2	8,4
2	6,2	9,9	6,4	7,2	10	2	6,6	9,3	4,1	7	8,8	5,2
3	5,8	9,6	3	6,9	9,9	3,9	7,2	8,3	3,8	7,2	9,7	6,9
4	6,5	9,6	4,1	6,1	9,5	1,9	7,1	8,4	1,6	7,5	10,1	2,7
5	6,5	9,2	0,8	6,3	9,4	5,7	6,8	8,5	3,4	7,6	9,2	1,9

Estrutura dos tratamentos: Fatorial 2x2

Estrutura de aleatorização das unidades amostrais aos tratamentos é restrita a **Blocos**.

⇒ Considere que as unidades amostrais (total de 20) estão

Blocadas, de tal forma que há 5 blocos de 4 observações (homogêneas).

Dentro de cada bloco os 4 tratamentos foram aleatorizados às observações.

Note que NÃO há replicas dentro dos níveis do fator Bloco.

Delineamento Aleatorizado em Blocos Completos

Caso Univariado - ANOVA

$$y_{gk} = \mu + \tau_g + \beta_k + \varepsilon_{gk}$$

Resposta da observação avaliada no nível g do Fator de interesse e no nível k do Fator Bloco (não há réplica)

Restrições de identificabilidade dos parâmetros

$$\sum_{g=1}^{G} \tau_g = 0, \sum_{k=1}^{b} \beta_k = 0$$

"Identidade útil" para obtenção das Somas de Quadrados e dos estimadores dos efeitos de interesse:

Ef. principal do Fator de interesse

Efeito do fator Bloco

Resíduo: é o ef. de interação entre o fator de interesse e Bloco

$$y_{gk} = \bar{y} + (\bar{y}_{g.} - \bar{y}) + (\bar{y}_{.k} - \bar{y}) + (y_{gk} - \bar{y}_{g.} - \bar{y}_{.k} + \bar{y})$$

SQ_F1 SQ_F2 SQ_Residual

Caso Multivariado $\Rightarrow$ descrever os resultados para o vetor de resposta p-dimensional.

Tabela de MANOVA
Delineamento Aleatorizado em Blocos Completos

F.V.	g.l.	Matriz de SSCP
Fator	**G-1**	$HF1 = \sum_{g=1}^{G} b\left(\bar{\mathbf{y}}_{g.} - \bar{\mathbf{y}}\right)\left(\bar{\mathbf{y}}_{g.} - \bar{\mathbf{y}}\right)'$
Bloco	**b-1**	$HF2 = \sum_{k=1}^{b} G\left(\bar{\mathbf{y}}_{.k} - \bar{\mathbf{y}}\right)\left(\bar{\mathbf{y}}_{.k} - \bar{\mathbf{y}}\right)'$
Resíduo	**(G-1)(b-1)**	$E = \sum_{g=1}^{G}\sum_{k=1}^{b}\left(\mathbf{y}_{gk} - \bar{\mathbf{y}}_{g.} - \bar{\mathbf{y}}_{.k} + \bar{\mathbf{y}}\right)\left(\mathbf{y}_{gk} - \bar{\mathbf{y}}_{g.} - \bar{\mathbf{y}}_{.k} + \bar{\mathbf{y}}\right)'$
TOTAL	**Gb-1**	$HF1 + HF2 + E = \sum_{g=1}^{G}\sum_{k=1}^{b}\left(\mathbf{y}_{gk} - \bar{\mathbf{y}}\right)\left(\mathbf{y}_{gk} - \bar{\mathbf{y}}\right)'$

Delineamento Hierárquico (*Nested*)

Considere o seguinte experimento em que as notas dos alunos foram avaliadas segundo Escola e Método de Ensino (A, B, C e D)

Escola 1				Escola 2			
Método A		Método B		Método C		Método D	
Nota1	Nota2	Nota1	Nota2	Nota1	Nota2	Nota1	Nota2
7.6	8.2	9.2	9.2	5.6	10.0	6.2	8.8
5.9	7.7	4.3	4.3	7.7	6.9	4.9	4.8
…	…	…	…	…	…	…	…
4.8	6.5	9.0	9.0	5.8	7.8	8.8	7.3

Estrutura hierárquica dos fatores

p=2

DCA: atribuição aleatória dos estudantes às salas de aula

Estrutura de Tratamentos: há dois fatores hierárquicos.

O fator Método de Ensino está definido DENTRO do fator Escola.

Delineamento Hierárquico (*Nested*)

Efeitos Fixos dos Fatores

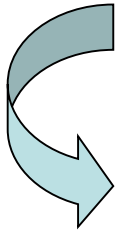
Caso Univariado - ANOVA

$$y_{ijk} = \mu + \tau_j + \beta_{k(j)} + \varepsilon_{ijk}$$

Resposta da observação i avaliada no nível **k do Fator 2 dentro do nível j do Fator 1**

Restrições de identificabilidade dos parâmetros

$$\sum_{j=1}^{a} \tau_j = 0, \sum_{k=1}^{b} \beta_{k(j)} = 0$$



"Identidade útil" para obtenção das Somas de Quadrados e dos estimadores dos efeitos de interesse:

Ef. principal de F1

Ef. de F2 dentro de F1: é a soma do ef. de F2 e da interação

Resíduo

$$y_{ijk} = \bar{y} + (\bar{y}_{j.} - \bar{y}) + (\bar{y}_{jk} - \bar{y}_{j.}) + (y_{ijk} - \bar{y}_{jk})$$

SQ_F1 SQ_F2(F1) SQ_Residual

Caso Multivariado $\Rightarrow$ descrever os resultados para o vetor de resposta p-dimensional.

Tabela de MANOVA
Delineamento Hierárquico ("Nested")

Fonte de Variação	Número de g.l.	Matriz de SQPC
F1	**a-1**	$H_{F1_{p\times p}} = \sum_{j=1}^{a} a\left(\bar{Y}_j - \bar{Y}\right)\left(\bar{Y}_j - \bar{Y}\right)'$
F2(F1)	**a(b-1)**	$H_{F2(F1)_{p\times p}} = \sum_{j=1}^{4}\sum_{k=1}^{4} r\left(\bar{Y}_{jk} - \bar{Y}_j\right)\left(\bar{Y}_{jk} - \bar{Y}_j\right)'$
Resíduo	**ab(r-1)**	$E_{p\times p} = \sum_{j=1}^{a}\sum_{k=1}^{b}\sum_{i=1}^{r}\left(Y_{ijk} - \bar{Y}_{jk}\right)\left(Y_{ijk} - \bar{Y}_{jk}\right)'$
Total	**abr-1**	$\sum_{j=1}^{a}\sum_{k=1}^{b}\sum_{i=1}^{r}\left(Y_{ijk} - \bar{Y}\right)\left(Y_{ijk} - \bar{Y}\right)'$

Modelos MANOVA

Pense nas possíveis decomposições da matriz de observações $\mathbf{Y}_{nxp}$

Decomposições (Identidades) úteis para a construção das SQPC

Modelo de um único fator: $y_{ig} = \bar{y} + (\bar{y}_g - \bar{y}) + (y_{ig} - \bar{y}_g)$

DCA

Fatorial Cruzado: $y_{ijk} = \bar{y} + (\bar{y}_{j.} - \bar{y}) + (\bar{y}_{.k} - \bar{y}) + (\bar{y}_{jk} - \bar{y}_{j.} - \bar{y}_{.k} + \bar{y}) + (y_{ijk} - \bar{y}_{jk})$

O efeito de F2 dentro de F1 é a soma do efeito principal de F1 e do efeito de interação

Fatorial Hierárquico: $y_{ijk} = \bar{y} + (\bar{y}_{j.} - \bar{y}) + (\bar{y}_{jk} - \bar{y}_{j.}) + (y_{ijk} - \bar{y}_{jk})$

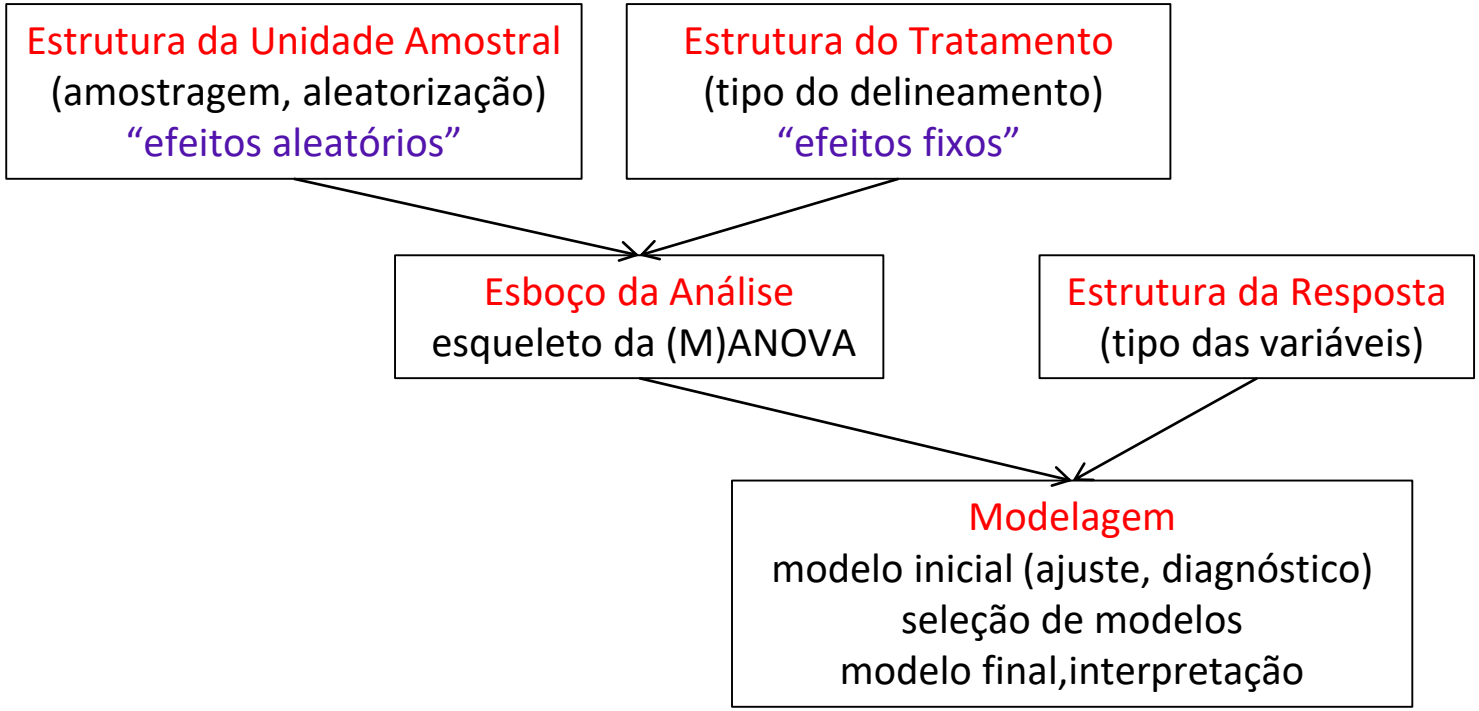
DABC

O ef. de interação entre Bloco e F1 é o resíduo (modelo aditivo)

Modelo com fator Bloco: $y_{jk} = \bar{y} + (\bar{y}_{j.} - \bar{y}) + (\bar{y}_{.k} - \bar{y}) + (y_{jk} - \bar{y}_{j.} - \bar{y}_{.k} + \bar{y})$

Estrutura Geral de Análise de Dados

(Goos and Gilmour, 2012)



Modelos MANOVA
Decomposição da Matriz Y_{nxp}

ASCA: ANOVA-Simultaneous Component Analysis (Smilde et al., 2005)

Modelo de um único fator: $y_{ig} = \bar{y} + \left(\bar{y}_g - \bar{y}\right) + \left(y_{ig} - \bar{y}_g\right)$

$$Y_{ig\ p\times 1} = \bar{Y}_{p\times 1} + \left(\bar{Y}_g - \bar{Y}\right)_{p\times 1} + \left(Y_{ig} - \bar{Y}_g\right)_{p\times 1}$$

$$Y_{n\times p};\quad n = \sum_{g=1}^{G} n_g$$

Decomposição devido à estrutura de grupos. E se as *n* observações não estivessem estratificadas?

$$Y_{n\times p} = M_{n\times p} + T_{n\times p} + E_{n\times p}$$

Componente da variabilidade ENTRE grupos

Componente da variabilidade DENTRO de grupos

Exemplo

Duas variáveis avaliadas em unidades amostrais submetidas a 3 tratamentos

T1		T2		T3	
Y11	**Y12**	**Y21**	**Y22**	**Y31**	**Y32**
9	3	0	4	3	8
6	2	2	0	1	9
9	7			2	7
8	4	1	2	2	8

Média geral = (4 , 5)

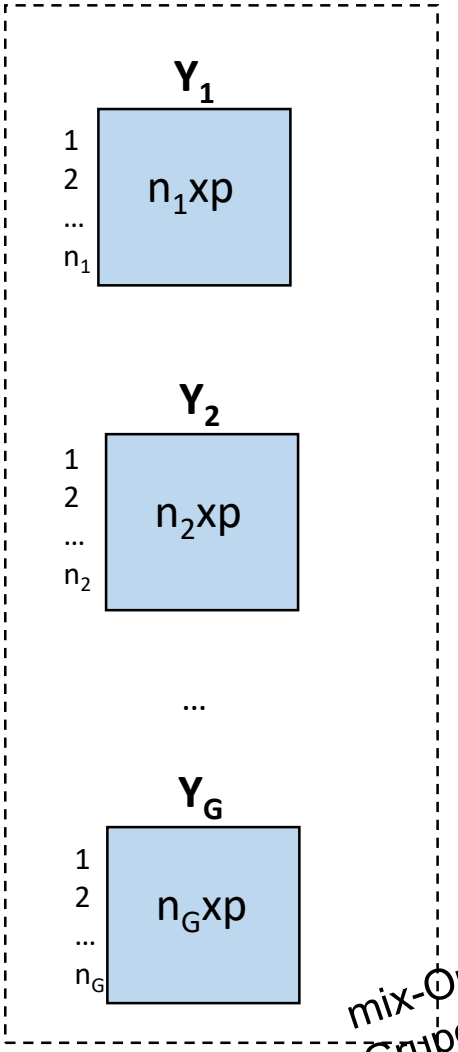
$$Y_{8\times2} = M_{8\times2} + T_{8\times2} + E_{8\times2}$$

$$\begin{pmatrix} 9 & 3 \\ 6 & 2 \\ 9 & 7 \\ 0 & 4 \\ 2 & 0 \\ 3 & 8 \\ 1 & 9 \\ 2 & 7 \end{pmatrix} = \begin{pmatrix} 4 & 5 \\ 4 & 5 \\ 4 & 5 \\ 4 & 5 \\ 4 & 5 \\ 4 & 5 \\ 4 & 5 \\ 4 & 5 \end{pmatrix} + \begin{pmatrix} 4 & 3 \\ 4 & 2 \\ 4 & 7 \\ -3 & 4 \\ -3 & 0 \\ -2 & 3 \\ -2 & 3 \\ -2 & 3 \end{pmatrix} + \begin{pmatrix} 1 & -1 \\ -2 & -2 \\ 1 & 3 \\ -1 & 2 \\ 1 & -2 \\ 1 & 0 \\ -1 & 1 \\ 0 & -1 \end{pmatrix}$$

Dependendo do problema pode haver interesse no componente T ou no componente E (residual ou resposta normalizada)

Pense no fator T como ***Multicentros*** e no problema de ***p-Integração*** de bancos de dados!

P-Integração de Bancos de Dados



$$Y_{n\times p};\quad n=\sum_{g=1}^{G} n_g$$

$$Y_{n\times p} = M_{n\times p} + T_{n\times p} + E_{n\times p}$$

Redução de dimensionalidade em componentes da decomposição de Y

Obter variáveis latentes ("componentes principais") que atendam a estrutura de Y (estratificada)

mix-Omics_R
Grupo de pesquisa da Le Cao