MAE 5776

ANÁLISE MULTIVARIADA

Júlia M Pavan Soler

pavan@ime.usp.br

1º Semestre/2020

MAE5776

Já vimos ☺

Matriz de Dados: $Y_{n\times p} = (Y_{ij}) \in \Re^{n\times p}; \quad Y_{ip\times 1} \overset{\text{iid}}{\sim} (\mu_{p\times 1}; \Sigma_{p\times p}), i=1,\ldots,n$

Estatísticas descritivas multivariadas: $\bar{Y}_{p\times 1}, S_{p\times p}, R_{p\times p}, S^{-1}_{p\times p} \qquad D_{n\times n} = (d^2_{ij}), d^2_{Pij}, d^2_{Mij}$

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 de Dados Aleatórios: $Y_{n\times p} \in \Re^{n\times p}; \quad Y_i \overset{\text{iid}}{\sim} N_p(\mu_{p\times 1}; \Sigma_{p\times p}), i=1,\ldots,n$

- Caso de Uma única População: Regiões (elipsoides) de Confiança para μ:

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

- Caso de Duas Populações: Regiões (elipsoides) de Confiança para $\mu_D = \mu_1 - \mu_2$:

$$R(\mu_D \mid (Y_1, Y_2)) = \left\{ \mu_D \in \Re^p;\; n(\bar{D} - \mu_D)' S_D^{-1} (\bar{D} - \mu_D) \le c^2; c^2 = T^2 = \frac{(n-1)p}{(n-p)} F_{p,(n-p)}(\alpha) \right\}$$

Populações dependentes

$$R(\mu_D \mid (Y_1, Y_2)) = \left\{ \mu_D \in \Re^p;\; (\bar{D} - \mu_D)' \left(S_c \left(\frac{1}{n_1} + \frac{1}{n_2} \right) \right)^{-1} (\bar{D} - \mu_D) \le c^2; c^2 = T^2 = \frac{(n_1 + n_2 - 2)p}{(n_1 + n_2 - p - 1)} F_{(p; n_1 + n_2 - p - 1)}(\alpha) \right\}$$

Populações independentes

Inferência: Vetores de Médias de Duas Populações Correções para Múltiplos Testes

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

Amostra Dependente*1: $n_1 \Rightarrow \quad \bar{Y}_1 \quad S_1$ $\qquad n_2 \Rightarrow \quad \bar{Y}_2 \quad S_2$ $\qquad \bar{D} \quad S_{\bar{D}} = S_c\left(\frac{1}{n_1}+\frac{1}{n_2}\right)$

Amostra Independente*2: $n_1 = n_2 \quad \Rightarrow \quad D_i, \quad i = 1,\ldots,n$ $\qquad \bar{D} \quad S_{\bar{D}} = S_D\frac{1}{n}$

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

$$\Rightarrow ICS\ \left(\mu_{Dj}\right) a \left(1-\alpha\right)\times 100\% = \left(\bar{D}_j \mp \sqrt{\frac{\nu_1}{\nu_2}F_{\nu_1,\nu_2}(\alpha)}\ \sqrt{S_{\bar{D}\,jj}}\right)$$

$*1: \nu_1 = (n-1)\,p, \quad \nu_2 = (n-p)$

$*2: \nu_1 = (n_1+n_2-1)\,p, \quad \nu_2 = (n_1+n_2-p-1)$

Intervalos de Confiança de Bonferroni (correção para múltiplos testes)

$$\Rightarrow ICB\left(\mu_{D\,j}\right) a \left(1-\alpha\right)\times 100\% = \left(\bar{Y}_{1j}-\bar{Y}_{2j}\right) \pm t_\nu(\alpha/2p)\sqrt{S_{\bar{D}\,jj}}$$

$*1: \nu = (n-1)$

$*2: \nu\left(= n_1+n_2-2\right)$

Inferência sobre um Vetor de Médias Comparações Múltiplas e o Problema de Múliplos Testes

Múltiplos testes independentes

$$H_0 : \mu_j = 0 \quad \times \quad H_0 : \mu_j \neq 0 \qquad j = 1, \ldots, p$$

$$P\left(\text{pelo menos uma Rej H}_0\right) = 1 - P\left(\text{nenhuma Rej H}_0\right)$$

$$= 1 - \prod_{l=1}^{K} P\left(p_l > \alpha\right) = 1 - \left(1 - \alpha\right)^K \quad \approx \quad 1$$

K→∞

```
> B<-10000
> K<-500
> set.seed(1299)
minpval <- replicate(B,
min(runif(K,0,1))<0.01)
table(minpval)
table(minpval)[2]/B
```

```
> table(minpval)
minpval
FALSE  TRUE
   78  9922
> table(minpval)[2]/B
  TRUE
0.9922
```

Correções para Múltiplos Testes

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

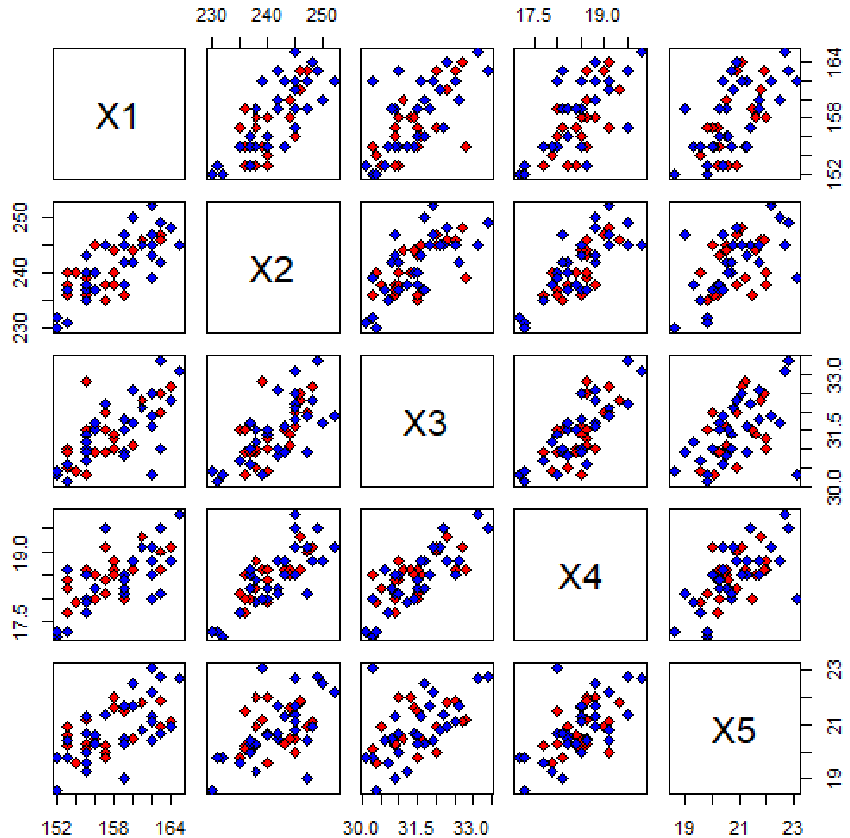
K: número total de testes α: nível de significância global fixado

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

Inferência sobre Vetores de Médias de Duas Populações

Dados dos Pardais (Manly, 2005)

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
...						
47	1	153	237	30.6	18.6	20.4
48	1	162	245	32.5	18.5	21.1
49	1	164	248	32.3	18.8	20.9



Grupo: 0= 1=

Inferência sobre Vetores de Médias de Duas Populações

```
n1=21  n2=28

mi   X1      X2      X3     X4     X5
0    157.38 241.00 31.43 18.50 20.81
1    158.43 241.57 31.48 18.45 20.84
1-0    1.05    0.57   0.05 -0.05  0.03
```

S0	X1	X2	X3	X4	X5
X1	11.05	9.10	1.56	0.87	1.29
X2	9.10	17.50	1.91	1.31	0.88
X3	1.56	1.91	0.53	0.19	0.24
X4	0.87	1.31	0.19	0.18	0.13
X5	1.29	0.88	0.24	0.13	0.57

S1	X1	X2	X3	X4	X5
X1	15.07	17.19	2.24	1.75	2.93
X2	17.19	32.55	3.40	2.95	4.07
X3	2.24	3.40	0.73	0.47	0.56
X4	1.75	2.95	0.47	0.43	0.51
X5	2.93	4.07	0.56	0.51	1.32

Sc	X1	X2	X3	X4	X5
X1	13.36	13.75	1.95	1.37	2.23
X2	13.75	26.15	2.76	2.25	2.71
X3	1.95	2.76	0.64	0.35	0.42
X4	1.37	2.25	0.35	0.32	0.35
X5	2.23	2.71	0.42	0.35	1.00

```
Box's M-test for Homogeneity of
Covariance Matrices
Chi-Sq (approx.) = 10.408, df = 15,
p-value = 0.7933
```

$$T^2 = 2.82 \sim \frac{(21+28-2)5}{(21+28-5-1)} F_{5,(21+28-5-1)}(0.05) \xrightarrow{\alpha=5\%} 13.29$$

Qual é a hipótese H0?
Conclusão?

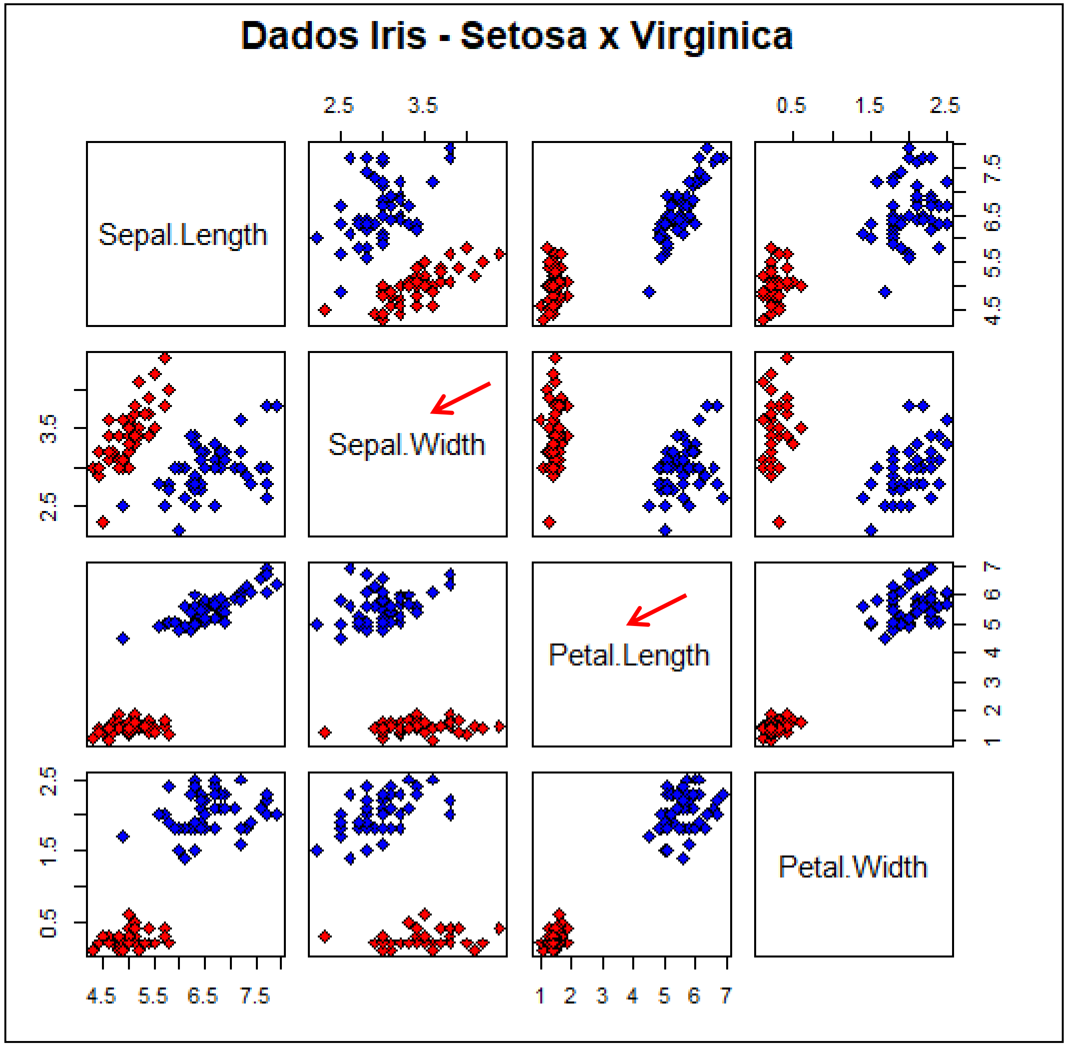
Inferência: Vetores de Médias de Duas Populações

Dados "Iris" do R (Fisher, RA, 1936. The use of multiple measurements in taxonomic problems. *Annals of Eugenics* 7, Part II: 179–188)

Medidas do comprimento e largura da pétala e sépala de 50 flores de íris de cada uma de três espécies (setosa, versicolor e virginica).

$$Y_{150\times 4} = \begin{pmatrix} Y_{G=1\ 50\times 4} \\ Y_{G=2\ 50\times 4} \\ Y_{G=3\ 50\times 4} \end{pmatrix}$$

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
...					
50	5.0	3.3	1.4	0.2	setosa
51	7.0	3.2	4.7	1.4	versicolor
...					
100	5.7	2.8	4.1	1.3	versicolor
101	6.3	3.3	6.0	2.5	virginica
102	5.8	2.7	5.1	1.9	virginica
...					
150	5.9	3.0	5.1	1.8	virginica



Existe evidência amostral para diferenças significantes entre essas duas espécies?

G=2: Setosa x Virginica

✓ p=4: Sepal.Length
Sepal.Width
Petal.Length
Petal.Width

Dados Iris, G=2 (Setosa x Virginica) p=4

Centróide dos grupos:

	Sepal.L	Sepal.W	Petal.L	Petal.W
setosa	5.006	3.428	1.462	0.246
virginica	6.588	2.974	5.552	2.026
Dif	**-1.582**	**0.454**	**-4.09**	**-1.78**

S.Setosa	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
Sepal.Length	0.12	0.10	0.02	0.01
Sepal.Width	0.10	0.14	0.01	0.01
Petal.Length	0.02	0.01	0.03	0.01
Petal.Width	0.01	0.01	0.01	0.01

S.Virginica	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
Sepal.Length	0.40	0.09	0.30	0.05
Sepal.Width	0.09	0.10	0.07	0.05
Petal.Length	0.30	0.07	0.30	0.05
Petal.Width	0.05	0.05	0.05	0.08

S comum	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
Sepal.Length	0.26	0.10	0.16	0.03
Sepal.Width	0.10	0.12	0.04	0.03
Petal.Length	0.16	0.04	0.17	0.03
Petal.Width	0.03	0.03	0.03	0.04

Box's M-test for Homogeneity of Covariance Matrices
Chi-Sq (approx.) = 109.3, df = 10, p-value < 2.2e-16

Hipótese?
Conclusão?

Inferência sobre Vetores de Médias de Duas Populações

Caso Multivariado - Amostras Independentes – Heterocedasticidade

Teoria Assintótica

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

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

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

Hipótese condicional, sob heterocedasticidade.

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

Elipsóide de Confiança:

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

Inferência sobre Vetores de Médias de Duas Populações

Caso Multivariado - Amostras Independentes

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

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

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

Pesquise o Problema de Behrens-Fisher: Testar a igualdade dos vetores de médias “SEM” suposições sobre as matrizes de covariâncias

Estatística da Razão de verossimilhanças (RVS):

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

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

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 da RVS para tomar decisão.

Dados Iris, G=2(Setosa x Virginica) p=4

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

```
T² = 4879.638                qChisquare(0.95) =9.488
⇒ Existe diferença entre as espécies para pelo menos uma das 4 variáveis ou
para pelo menos alguma combinação linear entre as variáveis
```

$$\Rightarrow ICS \text{ a } \left(\mu_{Dj}\right) a\left(1-\alpha\right)\times 100\% = \left(\bar{D}_j \mp \sqrt{\chi^2_4(0.05)}\,\sqrt{S_{\bar{D}\,jj}}\right)$$

Combinação linear canônica!

$$\Rightarrow ICB\left(\mu_{D\,j}\right) a\,\left(1-\alpha\right)\times 100\% = \left(\bar{Y}_{1j} - \bar{Y}_{2j}\right) \pm t_{(n_1+n_1-2)}(\alpha/2p)\,\sqrt{S_{\bar{D}\,jj}}$$

Intervalos de tamanho menor mas sob testes independentes!

		T2	ICS		t	ICB	
Var.	Dif	Estat.	Li	Ls	Estat.	Li	Ls
Y1	-1.582	236.735	-1.8987	-1.2653	-15.386	-1.8161	-1.3479
Y2	0.454	41.607	0.2372	0.6708	6.4503	0.2938	0.6142
Y3	-4.09	2498.619	-4.3420	-3.8380	-49.9862	-4.2763	-3.9037
Y4	-1.78	1830.624	-1.9081	-1.6519	-42.7858	-1.8747	-1.6853

testes "t" univariados

	results.t	padjustB	padjustFDR	padjustHOLM
Y1	6.892546e-28	2.757018e-27	9.190061e-28	1.378509e-27
Y2	4.246355e-09	1.698542e-08	4.246355e-09	4.246355e-09
Y3	1.504801e-71	6.019203e-71	6.019203e-71	6.019203e-71
Y4	3.230375e-65	1.292150e-64	6.460750e-65	9.691124e-65

Valores-p ajustados para α_global=5%

Inferência sobre Vetores de Médias de "Muitas" Populações

MANOVA

Comparações de Duas Populações $\Rightarrow$ Comparações de Muitas Populações (G≥2)

	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	$\underbrace{\mathrm{Y}_{11},\mathrm{Y}_{12},\ldots,\mathrm{Y}_{1n_1}}$	$\underbrace{\mathrm{Y}_{21},\mathrm{Y}_{22},\ldots,\mathrm{Y}_{2n_2}}$		$\underbrace{\mathrm{Y}_{G1},\mathrm{Y}_{G2},\ldots,\mathrm{Y}_{Gn_G}}$
	$\overline{\mathrm{Y}}_1 \quad S_1$	$\overline{\mathrm{Y}}_2 \quad S_2$		$\overline{\mathrm{Y}}_G \quad S_G$

Matriz de Dados: $\mathrm{Y}_{n\times(p+1)} = \left(Grupo\ \boxed{Y_1\ Y_2\ \ldots Y_p}\right)$

Var. qualitativas

Var. quantitativas

Estrutura de análises supervisionadas!!

Inferência sobre Vetores de Médias de “Muitas” Populações

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

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

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

$$\boxed{H_0 : \mu_1 = \mu_2 = \ldots = \mu_G; \quad \Sigma_g = \Sigma}$$

Sob Homocedasticidade!

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

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

Inferência sobre Vetores de Médias de Muitas Populações

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

Hipótese condicional

$$\boxed{H_0:\mu_1=\mu_2=\dots=\mu_G;\quad \Sigma_g=\Sigma}$$

EMVS sob H_0:

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

Divisor n

EMVS sob H_1:

$$\Rightarrow\begin{cases}\bar{Y}_g;\quad g=1,\dots,G\\ \hat{\Sigma}=\boxed{S_c=\dfrac{n_1S_1+\dots+n_GS_G}{n}}\end{cases}$$

Divisor n

SST

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

SSw

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

SSb

$B=T-W=\sum_{g=1}^{G}n_g\left(\bar{Y}_g-\bar{Y}\right)\left(\bar{Y}_g-\bar{Y}\right)'$ 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)

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

Lembrando "ANOVA":

Para p=1

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
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
$\overline{Y}_1$	$\overline{Y}_2$	. . .	$\overline{Y}_G$
S_1	S_2	. . .	S_G

Amostra

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

Lembrando ANOVA - Modelo Linear Geral

Resposta da observação i do grupo g

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

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

$$= \begin{cases} \mu_1 \\ \mu_1 + \tau_g ; g = 2, \ldots, G \end{cases}$$ Parametrização de Casela de Referência

Forma matricial:

$$Y_{n\times1} = X_{n\times G} \; \beta_{G\times1} + e_{n\times1}$$

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=3 grupos com 5 observações por grupo

Usando a parametrização de desvios

$$\underset{n\times 1}{Y} = \underset{n\times G}{X} \; \underset{G\times 1}{\beta} + \underset{n\times 1}{e}$$

$$\begin{bmatrix} Y_{11} \\ Y_{21} \\ Y_{31} \\ Y_{41} \\ Y_{51} \\ Y_{12} \\ Y_{22} \\ Y_{32} \\ Y_{42} \\ Y_{52} \\ Y_{13} \\ Y_{23} \\ Y_{33} \\ Y_{43} \\ Y_{53} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & -1 & -1 \\ 1 & -1 & -1 \\ 1 & -1 & -1 \\ 1 & -1 & -1 \\ 1 & -1 & -1 \end{bmatrix} \begin{bmatrix} \mu \\ \tau_1 \\ \tau_2 \end{bmatrix} + \begin{bmatrix} e_{11} \\ e_{21} \\ e_{31} \\ e_{41} \\ e_{51} \\ e_{12} \\ e_{22} \\ e_{32} \\ e_{42} \\ e_{53} \\ e_{13} \\ e_{23} \\ e_{33} \\ e_{43} \\ e_{53} \end{bmatrix}$$

ANOVA - Fontes de Variação

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

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

$$\underbrace{y_{ig} - \bar{y}}_{\substack{\textbf{SQTotal} \\ \textbf{corrigida}}} = \underbrace{(\bar{y}_g - \bar{y})}_{\substack{\textbf{SQTratamento} \\ \textbf{“Entre”}}} + \underbrace{(y_{ig} - \bar{y}_g)}_{\substack{\textbf{SQResidual} \\ \textbf{“Dentro”}}}$$

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

Tabela de ANOVA

$$H_0: \mu_1 = \mu_2 = \ldots = \mu_G = \mu \in \Re \quad \times \quad H_1: \exists \text{ pelo menos uma diferença}$$

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	$\sum_{g,i} (y_{ig} - \bar{y})^2$			

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

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

normalidade
homocedasticidade
independência

Modelo MANOVA

vetor de observações da
unidade i do grupo g

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

$$Y_{ig\; p\times 1} \overset{iid}{\sim} N_p\left(\mu_g ; \Sigma_g\right); \; i=1,\ldots,n_g, \;\; g=1,\ldots,G$$

Efeito
Fixo

Aleatório

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

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

Modelo MANOVA

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

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**	$H_{p\times p} = \sum_{g=1}^{G} n_g \left(\bar{\mathbf{y}}_g - \bar{\mathbf{y}}\right)\left(\bar{\mathbf{y}}_g - \bar{\mathbf{y}}\right)'$
Resíduo	**n-G**	$E_{p\times p} = \sum_{g=1}^{G}\sum_{i=1}^{n_g}\left(y_{ig} - \bar{y}_g\right)\left(y_{ig} - \bar{y}_g\right)'$
TOTAL	**n-1**	$H + E = \sum_{g=1}^{G}\sum_{i=1}^{n_g}\left(\mathbf{y}_{gi} - \bar{\mathbf{y}}\right)\left(\mathbf{y}_{gi} - \bar{\mathbf{y}}\right)'$

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

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:

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

Dados Iris (G=3,p=4)
Tabela de MANOVA

Qual é o modelo estrutural da MANOVA?

$$H_0: \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \boldsymbol{\mu}_3 = \boldsymbol{\mu} \quad \Leftrightarrow \quad H_0: \boldsymbol{\tau}_1 = \tau_2 = \boldsymbol{\tau}_3 = 0$$

F.V.	No. g.l.	Matriz de Soma de Quadr e Prod Cruzados	
Grupo	3-1	SSb=H $\begin{matrix} 63.21 & -19.95 & 165.25 & 71.28 \\ & 11.34 & -57.24 & -22.93 \\ & & 437.10 & 186.77 \\ & & & 80.41 \end{matrix}$	Matriz de SQPC devido ao efeito de Grupo
Resíduo	150-3	SSw=E $\begin{matrix} 38.96 & 13.63 & 24.62 & 5.65 \\ & 16.96 & 8.12 & 4.81 \\ & & 27.22 & 6.27 \\ & & & 6.16 \end{matrix}$	Matriz de SQPC devido ao efeito do Erro
TOTAL	150-1	SST=H+E $\begin{matrix} 102.17 & -6.32 & 189.87 & 76.92 \\ & 28.31 & -49.12 & -18.12 \\ & & 464.33 & 193.05 \\ & & & 86.57 \end{matrix}$	Matriz de SQPC Total

$$\Lambda^* = \frac{|E|}{|H+E|} = 1.486462\text{e-}31 \quad \Rightarrow \underbrace{\left(\frac{N-p-2}{p}\right)\left(\frac{1-\sqrt{\Lambda^*}}{\sqrt{\Lambda^*}}\right)}_{3.855466\text{e-}16} \sim F_{2p,2(N-p-2)}(\alpha = 0{,}01)$$

2.573429

Concl.?

Dados Iris (G=3,p=4)
Tabela de MANOVA $\Rightarrow$ Tabelas de ANOVA

$$H_0:\ \boldsymbol{\mu}_1=\boldsymbol{\mu}_2=\boldsymbol{\mu}_3=\boldsymbol{\mu} \quad \Rightarrow \quad H_{0j}:\ \mu_{1j}=\mu_{2j}=\mu_{3j}=\mu_j$$

F.V.	No. g.l.	SQ	QM	F
Grupo	3-1	63.21	31.61	31.61/0.265=119.27
Resíduo	150-3	38.96	0.265	
TOTAL	150-1	102.17		

qf(0.99,2,147) = 4.7525

Concl.?

Fontes de Var.	Grupo	Resíduo	F
resp 1	**63.2121**	**38.9562**	**119.27**
resp 2	11.3449	16.9620	
resp 3	437.1028	27.2226	
resp 4	80.4133	6.1566	
Número de g.l.	2	147	

Erro padrão residual: 0.5147894 0.3396877 0.4303345 0.20465

$=\sqrt{0.265}$

Obtenhas as Tabelas ANOVA!

Estatísticas Multivariadas

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

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 $\boxed{|H-\lambda E|=0;}$

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

Contribuição das Variáveis na Discriminação dos Grupos

Considere a seguinte decomposição espectral:

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

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

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

A avaliação dos coeficientes dos autovetores l, associados aos maiores autovalores, define a importância de cada variável no efeito de tratamento. ***Este resultado decorre também da Análise Discriminante de Fisher.***

Dados Iris:

```
> L  Matriz de Autovetores
            [,1]         [,2]        [,3]       [,4]
[1,] -0.06840592 -0.001987912  0.23777665  0.1120017
[2,] -0.12656121 -0.178526702 -0.09841836 -0.1981247
[3,]  0.18155288  0.076863566 -0.09653934 -0.2289487
[4,]  0.23180286 -0.234172267 -0.01460600  0.3759916

> eigen(M)$values  Autovalores
[1]  3.219193e+01  2.853910e-01  1.286765e-15 -1.554312e-15
```

A variável Y4, seguida de Y3, mais contribuem para a discriminação entre os grupos.

Comparações Múltiplas

Comparações múltiplas entre tratamentos dois a dois

(com correção de Bonferroni)

Dados Iris: Realize comparações múltiplas com correção para os múltiplos testes!

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

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

$$V\left(\bar{Y}_{g\,j} - \bar{Y}_{h\,j}\right) = \left(\frac{1}{n_g} + \frac{1}{n_h}\right)\boxed{\frac{E_{jj}}{n-G}} \Rightarrow \boxed{\left(\bar{Y}_{g\,j} - \bar{Y}_{h\,j}\right) \pm t_{n-G}(\alpha / 2K)\sqrt{V\left(\bar{Y}_{g\,j} - \bar{Y}_{h\,j}\right)}}$$

QMRes
Diag(S_{uc})

Intervalo de confiança 100(1-α)% com Correção de Bonferroni para um total de K comparações.

(Ex., K=p + G(G-1)/2)