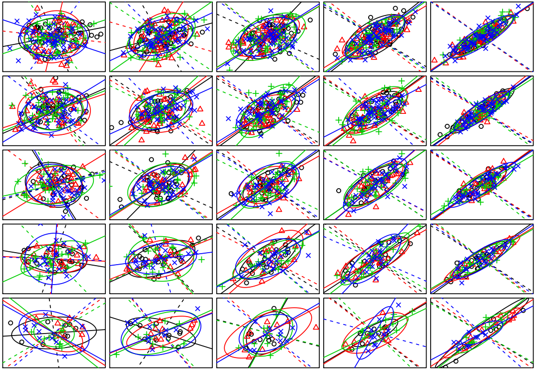
Estatística Multivariada

Slides de apoio às aulas



Faculdade de Ciências e Tecnologia
Universidade Nova de Lisboa
2018/19

Aula 5

Inferência sobre g ($g \geq 2$) vetores médios

Comparação de g ($g \geq 2$) valores médios: *One-way* ANOVA (populacão normal univariada, amostras independentes)

- Genericamente, a análise de variância — ANOVA — permite modelar uma variável contínua como função de um conjunto de factores (variáveis qualitativas) permitindo saber se os valores médios da variável resposta diferem consoante as condições definidas pelos níveis dos factores, assumindo resíduos aleatórios normais

- O tipo de ANOVA depende do delineamento experimental. O modelo de *efeitos fixos* a um factor (ANOVA *one-way*) pode descrever-se por:

$$\underbrace{x_{ij}}_{resposta} = \underbrace{\mu}_{\text{valor médio global}} + \underbrace{\alpha_j}_{\text{efeito do nível j fator}} + \underbrace{\epsilon_{ij}}_{\text{erro ou resíduo}}$$

- Pretende-se testar as hipóteses:

$$H_0 : \mu_1 = \mu_2 = ... = \mu_g$$

$$H_1 : \exists\, i, j : \mu_i \neq \mu_j \, (i \neq j, i, j = 1, ..., g)$$

sendo g o número de níveis do fator (grupos/amostras).

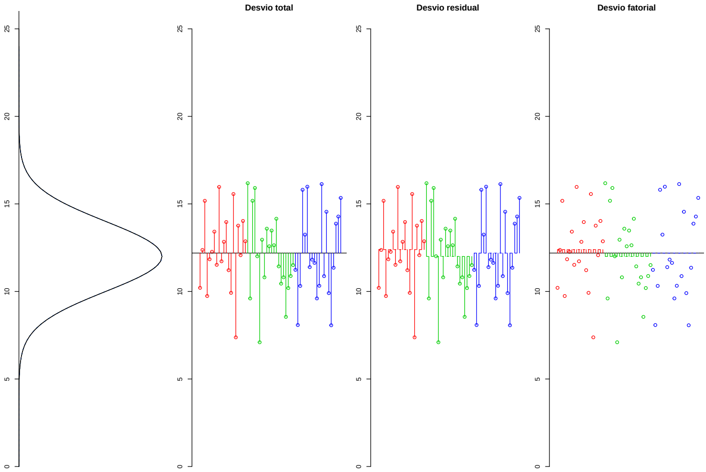
- Considerando a variável x, x_{ij} $(i = 1, ..., g,\ j = 1, ..., n_i)$ representa o elemento j da amostra i. Definem-se as médias:
  - **Média da amostra** i: $\bar{x}_i = \frac{\sum_j x_j}{n_i}$
  - **Média global**: $\bar{x} = \frac{\sum_i \sum_j x_{ij}}{n}$ sendo $n = \sum_i n_i$

- As médias das amostras aleatórias extraídas das populações em estudo permitem definir dois tipos de variação:
  - Variação **entre** amostras (variação factorial, variação entre tratamentos ou variação inter, *between*), resultante da influência do factor sobre a variável em estudo, e;
  - Variação **dentro** das amostras (variação residual ou variação intra, *within*), resultante de variabilidade não controlada.

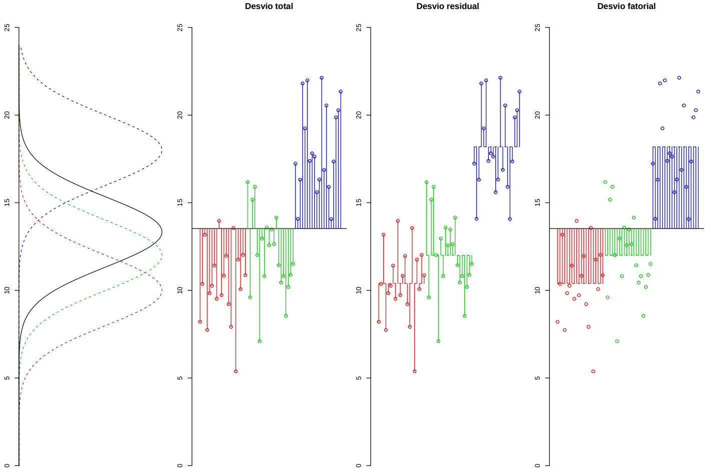
- No modelo de **efeitos fixos** os desvios de cada observação em relação à média total ou variação total dividem-se em duas componentes aditivas:

$$\underbrace{(x_{ij} - \bar{x})}_{Total} = \underbrace{(x_{ij} - \bar{x}_i)}_{Residual} + \underbrace{(\bar{x}_i - \bar{x})}_{Fatorial}$$

$H_0 : \mu_1 = \mu_2 = \mu_3$ verdadeira:



$H_1 : \exists\, i, j : \mu_i \neq \mu_j \; (i \neq j, i,\, j = 1, 2, 3)$ verdadeira:



- Elevando ao quadrado e aplicando somatórios ($i = 1, ..., g$; $j = 1, ..., n_i$) tem-se:

$$\underbrace{\sum_i \sum_j (x_{ij} - \bar{x})^2}_{SQT} = \underbrace{\sum_i \sum_j (x_{ij} - \bar{x}_i)^2}_{SQR} + \underbrace{\sum_i n_i(\bar{x}_i - \bar{x})^2}_{SQF}$$

- A quantificação da variação total, factorial e residual, é feita através do cálculo dos Quadrados Médios (QM):

$$QMT = \frac{SQT}{n-1} \quad QMF = \frac{SQF}{g-1} \quad QMR = \frac{SQR}{n-g}$$

- Assim, é o quociente entre as duas fontes de variação, factorial e residual, que permite concluir sobre a Rejeição/Não rejeição de H0. Ou seja, a estatística do teste é dada por:

$$F = \frac{QMF}{QMR}$$

com $F \frown F_{(g-1),(n-g)}$.

Tabela da ANOVA:

Fonte de variação	SQ	gl	QM	F
Fatorial	SQF	g-1	QMF=SQF/(g-1)	F=QMF/QMR
Residual	SQR	n-g	QMR=SQR/(n-g)	
Total	SQT	n-1	QMT	

Resumindo:

1. Hipóteses:
   $H_0 : \mu_1 = \mu_2 = ... = \mu_g$
   $H_1 : \exists\, i, j : \mu_i \neq \mu_j \, (i \neq j, i, j = 1, ..., g)$
2. Estatística do teste: $F = \frac{QMF}{QMR}$ com $F \frown F_{(g-1),(n-g)}$
3. Decisão: Rejeitar H_0 se $[F_{(g-1),(n-g);1-\alpha}; \infty[$

Pressupostos do modelo de efeitos fixos de ANOVA:

- Os conjuntos de observações constituem amostras aleatórias independentes extraídas das respetivas populações;
- Em cada uma das populações (**normalidade**): $x_i \frown N(\mu_i, \sigma_i)\,(i = 1, ..., g)$;
- As variâncias populacionais são homogéneas (**homocedasticidade**): $\sigma_1^2 = ... = \sigma_g^2 = \sigma^2$;
- $\alpha_i\,(i = 1, ..., g)$ são contantes desconhecidas, representativas dos desvios das médias μ_i em relação à média μ, tal que $\sum_i \alpha_i = 0$;
- $\epsilon_{ij}\,(i = 1, ..., g;\ j = 1, ..., n_i)$ representam a diferença entre os valores observados e os estimados e são v.a. iid: $\epsilon_{ij} \frown N(0, \sigma^2)$

Comparação de g ($g \geq 2$) vetores médios: *One-way* MANOVA (populacão normal multivariada, amostras independentes)

- Frequentemente pretende-se a comparação de g populações relativamente a p variáveis. Assim, têm-se g a.a.:

$$\text{População 1} \longrightarrow \text{Amostra 1: } (\mathbf{x}_{11}, ..., \mathbf{x}_{1n_1})$$

$$\text{População 2} \longrightarrow \text{Amostra 2: } (\mathbf{x}_{21}, ..., \mathbf{x}_{2n_2})$$

$$\cdots$$

$$\text{População g} \longrightarrow \text{Amostra g: } (\mathbf{x}_{g1}, ..., \mathbf{x}_{gn_g})$$

- Pressupostos:
  - g amostras aleatórias independentes de dimensão n_i, $(i = 1, ..., g)$, provenientes de populações com média $\boldsymbol{\mu}_i$, $(i = 1, ..., g)$
  - Populações com a mesma matriz de covariâncias $\boldsymbol{\Sigma}$
  - Populações $N_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$

- Analogamente, o modelo multivariado de *efeitos fixos* a um factor (MANOVA *one-way*) pode descrever-se por $((i = 1, \dots, g;\ j = 1, \dots, n_i))$:

$$\underbrace{\mathbf{x}_{ij}}_{\text{vetor de observações}} = \underbrace{\boldsymbol{\mu}}_{\text{vetor médio global}} + \underbrace{\boldsymbol{\alpha}_i}_{\text{efeito do nível i do fator}} + \underbrace{\boldsymbol{\epsilon}_{ij}}_{\text{erro ou resíduo}}$$

  onde:

  - $\boldsymbol{\epsilon}_{ij}$ são vetores aleatórios iid $N_p(\mathbf{0}, \boldsymbol{\Sigma})$
  - $\boldsymbol{\mu}$ representa o vetor médio global
  - $\boldsymbol{\alpha}_i$ representa o efeito no nível i do fator, com $\sum_{i=1}^{g} \boldsymbol{\alpha}_i = \mathbf{0}$

- De acordo com este modelo, o vetor de médias (observado) pode decompor-se em

$$\mathbf{x}_{ij} = \underbrace{\bar{\mathbf{x}}}_{\hat{\boldsymbol{\mu}}} + \underbrace{(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})}_{\hat{\boldsymbol{\alpha}}_i} + \underbrace{(\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)}_{\hat{\boldsymbol{\epsilon}}_{ij}}$$

- Considerando o produto interno $(\mathbf{x}_{ij} - \bar{\mathbf{x}})(\mathbf{x}_{ij} - \bar{\mathbf{x}})'$ e somando em $i\,(i = 1, \dots, g)$ e $j\,(j = 1, \dots, n_i)$

$$\underbrace{\sum_i \sum_j (\mathbf{x}_{ij} - \bar{\mathbf{x}})(\mathbf{x}_{ij} - \bar{\mathbf{x}})'}_{\mathbf{B}+\mathbf{W}} = \underbrace{\sum_i n_i (\bar{\mathbf{x}}_i - \bar{\mathbf{x}})(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})'}_{\mathbf{B}} + \underbrace{\sum_i \sum_j (\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)(\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)'}_{\mathbf{W}}$$

- Resumindo:

Fonte de variação	Matrizes das somas dos quadrados e produtos cruzados (SSP)	gl
Fatorial	$\mathbf{B}$	g-1
Residual	$\mathbf{W}$	n-g
Total	$\mathbf{W}+\mathbf{B}$	n-1

Cálculos:

- $\mathbf{B}+\mathbf{W}=(n-1)\mathbf{S}$ sendo $\mathbf{S}$ a matriz de covariâncias considerando n observações das p variáveis;
- $\mathbf{W}=(n_1-1)\mathbf{S}_1+\ldots+(n_g-1)\mathbf{S}_g$ sendo $\mathbf{S}_1,\ldots,\mathbf{S}_g$ as matrizes de covariâncias observadas nas g amostras aleatórias;
- $\mathbf{B}=(\mathbf{B}+\mathbf{W})-\mathbf{W}$

- Hipóteses:

$$H_0:\boldsymbol{\alpha}_1=\boldsymbol{\alpha}_2=\ldots=\boldsymbol{\alpha}_g$$
$$H_1:\exists\, i,j:\boldsymbol{\alpha}_i\neq\boldsymbol{\alpha}_j\,(i\neq j, i, j=1,\ldots,g)$$
ou
$$H_0:\boldsymbol{\mu}_1=\ldots=\boldsymbol{\mu}_g$$
$$H_1:\exists\, i,j:\boldsymbol{\mu}_i\neq\boldsymbol{\mu}_j\,(i\neq j, i, j=1,\ldots,g)$$

Note-se que a não rejeição de H_0 implica p igualdades!

- Estatística do teste: **Lambda de Wilks**

$$\Lambda^* = \frac{|\mathbf{W}|}{|\mathbf{W} + \mathbf{B}|}$$

- A estatística Λ^* pode determinar-se como função dos valores próprios da matriz $\mathbf{BW}^{-1}$

$$\Lambda^* = \prod_{i=1}^{s} \frac{1}{1+\lambda_i}$$

com $s = \min(p, g-1)$ (número de valores próprios não nulos)

- Além desta, existem outras estatísticas (assintoticamente equivalentes) usadas na inferência sobre vetores médios
    - Raíz máxima de Roy (*Roy's largest root*)=maior valor próprio de $\mathbf{BW}^{-1}$
    - Traço de Hotelling (*Hotelling´s trace*)= $tr(\mathbf{BW}^{-1})$
    - Traço de Pillai (*Pillai's trace*)= $tr(\mathbf{B}(\mathbf{B}+\mathbf{W})^{-1})$

- A distribuição exata de Λ^* pode ser obtida em casos particulares de acordo com a tabela seguinte:

n. de var.	n. de grupos	Distribuição amostral para dados multivariados Normais
$p = 1$	$g \geq 2$	$\left(\frac{n-g}{g-1}\right)\left(\frac{1-\Lambda^*}{\Lambda^*}\right) \frown F_{g-1,n-g}$
$p = 2$	$g \geq 2$	$\left(\frac{n-g-1}{g-1}\right)\left(\frac{1-\sqrt{\Lambda^*}}{\sqrt{\Lambda^*}}\right) \frown F_{2(g-1),2(n-g-1)}$
$p \geq 1$	$g = 2$	$\left(\frac{n-p-1}{p}\right)\left(\frac{1-\Lambda^*}{\Lambda^*}\right) \frown F_{p,n-p-1}$
$p \geq 1$	$g = 3$	$\left(\frac{n-p-2}{p}\right)\left(\frac{1-\sqrt{\Lambda^*}}{\sqrt{\Lambda^*}}\right) \frown F_{2p,2(n-p-2)}$

- Para outras situações, não descritas na tabela anterior, pode usar-se a seguinte aproximação (amostras de grande dimensão)

$$F = \frac{1 - \Lambda^{1/t}}{\Lambda^{1/t}} \frac{\nu_2}{\nu_1}$$

com $\nu_1 = p(g-1)$ e $\nu_2 = wt - \frac{1}{2}\big(p(g-1) - 2\big)$

$w = (n-g) - \frac{1}{2}(p - g + 2)$

$t = \sqrt{\frac{p^2(g-1)^2 - 4}{p^2 + (n-g)^2 - 5}}$, (com $p^2 + (n-g)^2 - 5 > 0$ e $t = 1$, caso contrário)

tendo $F \overset{a}{\sim} F_{(\nu_1, \nu_2)}$

- Alternativamente (menor precisão), sob H_0 e para n suficientemente grande, tem-se

$$-\left(n - 1 - \frac{p+g}{2}\right) \ln\left(\frac{|\mathbf{W}|}{|\mathbf{B} + \mathbf{W}|}\right) \overset{a}{\sim} \chi^2_{p(g-1)}$$

Exemplo 1

Considere 3 a.a. independentes de observações bivariadas, extraídas de 3 populações $N_2(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$ $(i = 1, 2, 3)$

Amostra	x_1	x_2
1	9	3
1	6	2
1	9	7
2	0	4
2	2	0
3	3	8
3	1	9
3	2	7

Pretende-se testar $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \boldsymbol{\mu}_3$ $(\alpha = 0.01)$.

Intervalos de confiança simultâneos

- No caso do resultado da MANOVA *one-way* resultar na rejeição da H_0, é natural querer estimar-se a diferença entre os valores médios por variável e para cada par de níveis definidos pelo fator (**pairwise comparisons**) $\rightarrow$ IC de Bonferroni
- Seja $\mu_{ij} - \mu_{kj}$ a diferença entre os valores médios das populações i e k, relativamente à j-ésima variável aleatória, então $\bar{x}_{ij} - \bar{x}_{kj}$ representa a diferença estimada e

$$Var(\bar{x}_{ij} - \bar{x}_{kj}) = \Big(\frac{1}{n_1} + \frac{1}{n_2}\Big)\sigma_j$$

sendo

$$\hat{Var}(\bar{x}_{ij} - \bar{x}_{kj}) = \Big(\frac{1}{n_1} + \frac{1}{n_2}\Big)\frac{w_j}{n-g}$$

onde w_j representa o j-ésimo elemento da matriz $\mathbf{W}$

- Existindo p variáveis e $g(g-1)/2$ comparações, então o nível de significância a usar em cada IC será

$$\frac{\alpha/m}{2}$$

com $m = \frac{pg(g-1)}{2}$

- Os IC simultâneos a $(1-\alpha)\times 100\%$ pode então obter-se usamdo a expressão

$$(\bar{x}_{ij} - \bar{x}_{kj}) \pm t_{(n-g);1-\alpha m/2}\sqrt{\frac{w_j}{n-g}\big(\frac{1}{n_i} + \frac{1}{n_k}\big)}$$

Exemplo 2

Relativamente aos dados do exemplo anterior, determine os IC simultâneos a 99% para todas as comparações múltiplas de vetores médios.

- Suponha-se agora que se pretende estudar o efeito de dois fatores, existindo a níveis do fator 1, b níveis do fator 2 e n observações por cada uma das ab combinações dos níveis dos 2 fatores

- O modelo multivariado de *efeitos fixos, equilibrado* a dois factores (MANOVA *two-way*) pode descrever-se por $((i = 1, ..., a;\ j = 1, ..., b;\ k = 1, ..., n))$:

$$\underbrace{\mathbf{x}_{ijk}}_{\text{observações}} = \underbrace{\boldsymbol{\mu}}_{\text{média global}} + \underbrace{\boldsymbol{\alpha}_i}_{\text{efeito do nível i do fator 1}} + \underbrace{\boldsymbol{\beta}_j}_{\text{efeito do nível j do fator 2}} + \underbrace{\boldsymbol{\tau}_{ij}}_{\text{interação}} + \underbrace{\boldsymbol{\epsilon}_{ijk}}_{\text{resíduos}}$$

onde:

- $\mathbf{x}_{ijk}$ representa os vetores de dimensão p replicados n vezes por cada combinação ab
- $\boldsymbol{\epsilon}_{ijk}$ são vetores aleatórios iid $N_p(\mathbf{0}, \boldsymbol{\Sigma})$
- $\boldsymbol{\mu}$ representa o vetor médio global
- $\boldsymbol{\alpha}_i$ representa o efeito no nível i do fator 1, com $\sum_{i=1}^{a} \boldsymbol{\alpha}_i = 0$
- $\boldsymbol{\beta}_j$ representa o efeito no nível j do fator 2, com $\sum_{j=1}^{b} \boldsymbol{\beta}_j = 0$
- $\boldsymbol{\tau}_j$ representa o efeito da interação entre o no nível i do fator 1 e o nível j do fator 2, com $\sum_{i=1}^{a} \boldsymbol{\tau}_i = \sum_{j=1}^{b} \boldsymbol{\tau}_j = \mathbf{0}$

- De acordo com este modelo, o vetor de médias (observado) pode decompor-se em

$$\mathbf{x}_{ijk} = \underbrace{\bar{\mathbf{x}}}_{\hat{\boldsymbol{\mu}}} + \underbrace{(\bar{\mathbf{x}}_{i\cdot} - \bar{\mathbf{x}})}_{\hat{\boldsymbol{\alpha}}_i} + \underbrace{(\bar{\mathbf{x}}_{\cdot j} - \bar{\mathbf{x}})}_{\hat{\boldsymbol{\beta}}_j} + \underbrace{(\bar{\mathbf{x}}_{ij} - \bar{\mathbf{x}}_{i\cdot} - \bar{\mathbf{x}}_{\cdot j} + \bar{\mathbf{x}})}_{\hat{\boldsymbol{\tau}}_k} + \underbrace{(\mathbf{x}_{ijk} - \bar{\mathbf{x}}_{ij})}_{\hat{\boldsymbol{\epsilon}}_{ijk}}$$

- Aplicando as somas em i $(i = 1, ..., a)$, j $(j = 1, ..., b)$ e k $(j = 1, ..., n)$ aos produtos internos dos vários termos, obtêm-se as seguintes expressões das matrizes das somas dos quadrados e produtos cruzados (SSP)

$$SSP_1 = \sum_{i=1}^{a} bn(\bar{\mathbf{x}}_{i\cdot} - \bar{\mathbf{x}})(\bar{\mathbf{x}}_{i\cdot} - \bar{\mathbf{x}})'$$

$$SSP_2 = \sum_{j=1}^{b} an(\bar{\mathbf{x}}_{\cdot j} - \bar{\mathbf{x}})(\bar{\mathbf{x}}_{\cdot j} - \bar{\mathbf{x}})'$$

$$SSP_{int} = \sum_{i=1}^{a} \sum_{j=1}^{b} n(\bar{\mathbf{x}}_{ij} - \bar{\mathbf{x}}_{i\cdot} - \bar{\mathbf{x}}_{\cdot j} + \bar{\mathbf{x}})(\bar{\mathbf{x}}_{ij} - \bar{\mathbf{x}}_{i\cdot} - \bar{\mathbf{x}}_{\cdot j} + \bar{\mathbf{x}})'$$

$$SSP_{res} = \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{n} (\mathbf{x}_{ijk} - \bar{\mathbf{x}}_{ij})(\mathbf{x}_{ijk} - \bar{\mathbf{x}}_{ij})'$$

$$SSP_{total} = \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{n} (\mathbf{x}_{ijk} - \bar{\mathbf{x}})(\mathbf{x}_{ijk} - \bar{\mathbf{x}})'$$

- Resumindo:

Fonte de variação	Matrizes das somas dos quadrados e produtos cruzados (SSP)	gl
Fator 1	SSP_1	a-1
Fator 2	SSP_2	b-1
Interação	SSP_{int}	(a-1)(b-1)
Residual	SSP_{res}	ab(n-1)[1]
Total	SSP_{total}	abn-1

- Hipóteses e estatísticas do teste:

1. $H_0^{(12)} : \boldsymbol{\tau}_{11} = \ldots = \boldsymbol{\tau}_{ab}$

$$\Lambda^*_{(12)} = \frac{|SSP_{res}|}{|SSP_{int} + SSP_{res}|}$$

2. $H_0^{(1)} : \boldsymbol{\alpha}_1 = \ldots = \boldsymbol{\alpha}_a$

$$\Lambda^*_{(1)} = \frac{|SSP_{res}|}{|SSP_1 + SSP_{res}|}$$

3. $H_0^{(2)} : \boldsymbol{\beta}_1 = \ldots = \boldsymbol{\beta}_b$

$$\Lambda^*_{(2)} = \frac{|SSP_{res}|}{|SSP_2 + SSP_{res}|}$$

[1] $p \leq ab(n-1)$, para que SSP_{res} seja definida positiva

- Para amostras de grande dimensão, pode usar-se a aproximação à distribuição F

$$F = \frac{1 - \Lambda^{1/t}}{\Lambda^{1/t}} \frac{\nu_2}{\nu_1}$$

com $\nu_1 = p \times df_{effect}$ e $\nu_2 = wt - \frac{1}{2}\left(p \times df_{effect} - 2\right)$

$w = df_{error} - \frac{1}{2}(p - df_{effect} + 1)$

$t = \sqrt{\frac{p^2 (df_{effect})^2 - 4}{p^2 + df^2_{error} - 5}}$, (com $p^2 + df^2_{error} - 5 > 0$ e $t = 1$, caso contrário)

tendo $F \overset{a}{\sim} F_{(\nu_1, \nu_2)}$

Alternativamente (menor precisão), sob H_0 e para n suficientemente grande, tem-se

- Sob $H_0^{(12)}$ e para n suficientemente grande, tem-se

$$-\left[ab(n-1) - \frac{p+1-(a-1)(b-1)}{2}\right] \ln \Lambda^*_{(12)} \overset{a}{\sim} \chi^2_{(a-1)(b-1)p}$$

- Sob $H_0^{(1)}$ e para n suficientemente grande, tem-se

$$-\left[ab(n-1) - \frac{p+1-(a-1)}{2}\right] \ln \Lambda^*_{(1)} \overset{a}{\sim} \chi^2_{(a-1)p}$$

- Sob $H_0^{(2)}$ e para n suficientemente grande, tem-se

$$-\left[ab(n-1) - \frac{p+1-(b-1)}{2}\right] \ln \Lambda^*_{(2)} \overset{a}{\sim} \chi^2_{(b-1)p}$$

Considerações finais sobre a MANOVA

- Tal como salientado, na MANOVA assume-se que os resíduos são variáveis iid, de valor médio nulo e matriz de covariâncias constante
- Vimos anteriormente como estudar a normalidade multivariada. O pressuposto de igualdade das matrizes de covariâncias pode ser testado usando o **Teste M de Box**
- Apesar da importância dos pressupostos teóricos, é de salientar que a MANOVA é genericamente um procedimento robusto no que respeita ao seu incumprimento (excepto à não independências das observações)
- Quando os testes multivariados conduzem à rejeição da hipótese nula, fica por identificar quais as populações que diferem significativamente entre si e se as diferenças se devem apenas a uma das variáveis dependentes em estudo ou a várias delas
- Assim, no caso de serem identificados efeitos significativos através de uma MANOVA, devem realizar-se testes F univariados (ANOVA's), para determinar qual ou quais das variáveis dependentes contribuem para as diferenças significativas
- Adicionalmente, é necessário proceder a comparações múltiplas dos grupos (**post-hoc tests**), dois a dois, de forma a identificar quais os grupos que diferem significativamente

Exemplo 3

Considere-se o seguinte conjunto de observações (ficheiro "data5.xlsx")

		fator 2					
		nível 1			nível 2		
		x_1	x_2	x_3	x_1	x_2	x_3
fator 1	nível 1	6.5	9.5	4.4	6.9	9.1	5.7
		6.2	9.9	6.4	7.2	10.0	2.0
		5.8	9.6	3.0	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
	nível 2	6.7	9.1	2.8	7.1	9.2	8.4
		6.6	9.3	4.1	7.0	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

Use a MANOVA para testar a existência dos efeitos de interação, do fator 1 e do fator 2 sobre o vetor resposta $\mathbf{x} = (x_1, x_2, x_3)'$. Indique os pressupostos assumidos.

Exemplo 4

Considere-se os dados do ficheiro "data6.xlsx" referentes a um ensaio agrícola, em plantações de amendoim. O ficheiro contem observações relativas às seguintes variáveis:

- F1 - Localização geográfica da plantação
- F2 - Variedade
- x_1 - Produção total da plantação (g)
- x_2 - Produção pronta (maturada) para consumo (g)
- x_3 - Peso da semente (g por 100 sementes)

Use a MANOVA para testar a existência dos efeitos de interação, do fator 1 e do fator 2 sobre o vetor resposta $\mathbf{x} = (x_1, x_2, x_3)'$. Considere $\alpha = 0.05$. Indique os pressupostos assumidos.

1. q medições repetidas, uma amostra aleatória $\rightarrow$ **Testes com matrizes de contrastes**

Elemento	Medições repetidas			
	1	2	$\cdots$	q
1	x_{11}	x_{12}	$\cdots$	x_{1q}
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
n	x_{n1}	x_{n2}	$\cdots$	x_{nq}

2. q medições repetidas, duas amostras aleatórias independentes $\rightarrow$ **Análise de perfis**

Amostra	Elemento	Medições repetidas			
		1	2	$\cdots$	q
Amostra 1	1	x_{111}	x_{112}	$\cdots$	x_{11q}
	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
	n_1	$x_{1n_1 1}$	$x_{1n_1 2}$	$\cdots$	$x_{1n_1 q}$
Amostra 2	1	x_{211}	x_{212}	$\cdots$	x_{21q}
	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
	n_2	$x_{2n_2 1}$	$x_{2n_2 2}$	$\cdots$	$x_{2n_2 q}$

- Pretende-se testar as hipóteses:

$$H_0 : \mu_1 = \mu_2 = ... = \mu_q$$

sendo q o número de repetições

- Note-se que a hipótese anterior equivale a:

$$H_0 : \mu_2 - \mu_1 = \mu_3 - \mu_2 = ... = \mu_q - \mu_{q-1} = 0$$

que se pode representar como

$$H_0 : \begin{bmatrix} -1 & 1 & \ldots & 0 \\ 0 & -1 & \ldots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & \ldots & -1 & 1 \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_q \end{bmatrix} = \mathbf{C}\boldsymbol{\mu} = \mathbf{0}$$

- A matriz $\mathbf{C}_{(q-1)\times q}$ é designada por *matriz de contrastes* representando $q-1$ combinações lineares dos valores médios μ_j $(j = 1, ..., q)$. Cada linha representa um *vetor contraste* cuja soma dos elementos é zero.

- Considerando a amostra aleatória $(\mathbf{x}_1, ..., \mathbf{x}_n)$ proveniente da população $N_q(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, tem-se a matriz de médias $\mathbf{C}\bar{\mathbf{x}}$ e a matriz de covariâncias $\mathbf{CSC}'$ e, sob H_0

$$T^2 = n(\mathbf{C}\bar{\mathbf{x}})'(\mathbf{CSC}')^{-1}\mathbf{C}\bar{\mathbf{x}} \frown T^2_{q-1}(n-1)$$

onde

$$T^2_{q-1}(n-1) \overset{d}{=} \frac{(n-1)(q-1)}{n-q+1} F_{(q-1,n-q+1)}$$

- Assim, H_0 será rejeitada quando

$$T^2 > \frac{(n-1)(q-1)}{n-q+1} F_{(q-1,n-q+1);1-\alpha}$$

onde $F_{(q-1,n-q+1);1-\alpha}$ representa o quantil de probabilidade $1-\alpha$ da distribuição $F_{(q-1,n-q+1)}$

- É possível mostrar que a estatística T^2 não depende da escolha da matriz $\mathbf{C}$, sendo por isso o mesmo procedimento válido para testar outros contrastes

- Rejeitando H_0 podem testar-se individualmente da um dos $q-1$ contrastes, usando a estatística:

$$T_i^2 = \frac{\sqrt{n}\mathbf{c}_i'\bar{\mathbf{x}}}{\sqrt{\mathbf{c}_i'\mathbf{S}\mathbf{c}_i}} \frown T_{q-1}^2(n-1),\ (i=1,...,q-1)$$

sendo $\mathbf{c}_i$ o i-ésimo vetor contraste.

- Assim, ao nível α rejeita-se H_0 quando

$$T^2 > \frac{(n-1)(q-1)}{n-q+1} F_{(q-1,n-q+1);1-\alpha}$$

onde $F_{(q-1,n-q+1);1-\alpha}$ representa o quantil de probabilidade $1-\alpha$ da distribuição $F_{(q-1,n-q+1)}$

- Os IC simultâneos podem determinar-se usando a expressão

$$\mathbf{c}_i'\bar{\mathbf{x}} \pm \sqrt{\frac{(n-1)(q-1)}{n-q+1} F_{(q-1,n-q+1);1-\alpha}} \sqrt{\frac{\mathbf{c}_i'\mathbf{S}\mathbf{c}_i}{n}}\ (i=1,...,q-1)$$

Exemplo 5

Considere a seguinte base de dados com observações relativas à velocidade de realização de 2 tarefas (fator A) usando duas marcas de máquinas calculadoras (fator B):

	A1		A2	
Elementos	B1	B2	B1	B2
1	30	21	21	14
2	22	13	22	5
3	20	13	18	17
4	12	7	16	14
5	23	24	23	8

Teste os efeitos dos fatores A e B e interação, considerando $\alpha = 0.05$.

Pretende-se portanto testar $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$ que se pode representar pelos seguintes contrastes:

1. $\frac{\mu_1+\mu_2}{2} = \frac{\mu_3+\mu_4}{2}$
2. $\frac{\mu_1+\mu_3}{2} = \frac{\mu_2+\mu_4}{2}$
3. $\frac{\mu_1+\mu_4}{2} = \frac{\mu_2+\mu_3}{2}$

- Matriz de contrastes:

$$\mathbf{C}_{3\times 4} = \begin{bmatrix} 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 \end{bmatrix}$$

- Sob $H_0 : \mathbf{C}\boldsymbol{\mu} =$

$$T^2 = n(\mathbf{C}\bar{\mathbf{x}})'(\mathbf{CSC}')^{-1}(\mathbf{C}\bar{\mathbf{x}}) = 21.78$$

Código R:

```
> C<-matrix(c(1,1,1,1,-1,-1,-1,1,-1,-1,-1,1),3,4)
> A1.B1<-c(30,22,20,12,23)
> A1.B2<-c(21,13,13,7,24)
> A2.B1<-c(21,22,18,16,23)
> A2.B2<-c(14,5,17,14,8)
> X<-matrix(c(A1.B1,A1.B2,A2.B1,A2.B2),5,4)
> m<-colMeans(X)
> S=var(X)
> n=nrow(X)
> T2<-n*t(C%*%m)%*%solve(C%*%S%*%t(C))%*%(C%*%m)
> T2

         [,1]
[1,] 21.78032
```

Exemplo (continuação)

- Sob H_0

$$T^2 \frown T_3^2(4) \stackrel{d}{=} \frac{(4)(3)}{2} F_{(3,1)}$$

- Assim, H_0 será rejeitada quando

$$T^2 < 6F_{(3,1);0.95} = 114.986$$

Código R:

```
> n=nrow(X)
> q=ncol(X)
> ((n-1)*(q-1))/(n-q+1)*qf(0.95,q-1,n-q+1)

[1] 114.9858
```

Logo, não se rejeita H_0, concluindo-se não existirem evidências de diferenças significativas entre os valores médios.

- Suponhamos agora que se pretende comparar os perfis que se obtêm por ligação linear dos pontos (j, μ_{1j}) e (j, μ_{2j}) $(j = 1, ..., q)$

- Há essencialmente três questões com particular interesse:
  - Serão os perfis paralelos?
  - Serão os perfis coincidentes (dado que são paralelos)?
  - Serão os perfis horizontais (dado que são paralelos e coincidentes)?

- O **teste ao paralelismo** dos perfis pode expressar-se pela hipótese

$$H_0 : \mathbf{C}\boldsymbol{\mu}_1 - \mathbf{C}\boldsymbol{\mu}_2 = \mathbf{0}$$

sendo $\boldsymbol{\mu}_1' = (\mu_{11}, ..., \mu_{1q})$, $\boldsymbol{\mu}_2' = (\mu_{21}, ..., \mu_{2q})$ e

$$\mathbf{C}_{(q-1)\times q} = \begin{bmatrix} -1 & 1 & \dots & 0 \\ 0 & -1 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & \dots & -1 & 1 \end{bmatrix}$$

- Considerando as amostras aleatórias $(\mathbf{x}_{11}, ..., \mathbf{x}_{1n_1})$ e $(\mathbf{x}_{21}, ..., \mathbf{x}_{2n_2})$ respetivamente provenientes das populações $N_q(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$ e $N_q(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$, tem-se, sob H_0

$$T^2 = (\mathbf{C}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2))' \left[\left(\frac{1}{n_1} + \frac{1}{n_2} \right) \mathbf{C}\mathbf{S}_{pooled}\mathbf{C}' \right]^{-1} (\mathbf{C}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)) \frown T^2_{q-1}(n_1 + n_2 - 2)$$

onde

$$T^2_{q-1}(n_1 + n_2 - 2) \stackrel{d}{=} \frac{(n_1 + n_2 - 2)(q-1)}{n_1 + n_2 - q} F_{(q-1, n_1+n_2-q)}$$

- Assim, H_0 será rejeitada quando

$$T^2 > \frac{(n_1 + n_2 - 2)(q-1)}{n_1 + n_2 - q} F_{(q-1, n_1+n_2-q);1-\alpha}$$

onde $F_{(q-1,n_1+n_2-2);1-\alpha}$ representa o quantil de probabilidade $1-\alpha$ da distribuição $F_{(q-1,n_1+n_2-2)}$

- O **teste à coincidência** dos perfis pode expressar-se pela hipótese

$$H_0 : \frac{\boldsymbol{\mu}_{11} + ... + \boldsymbol{\mu}_{1q}}{q} = \frac{\boldsymbol{\mu}_{21} + ... + \boldsymbol{\mu}_{2q}}{q}$$

equivalente a

$$H_0 : \mathbf{1}'\boldsymbol{\mu}_1 = \mathbf{1}'\boldsymbol{\mu}_2$$

- Considerando as amostras aleatórias $(\mathbf{x}_{11}, ..., \mathbf{x}_{1n_1})$ e $(\mathbf{x}_{21}, ..., \mathbf{x}_{2n_2})$ respetivamente provenientes das populações $N_q(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$ e $N_q(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$, tem-se, sob H_0

$$T^2 = (\mathbf{1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2))' \left[\left(\frac{1}{n_1} + \frac{1}{n_2} \right) \mathbf{1}\mathbf{S}_{pooled}\mathbf{1}' \right]^{-1} (\mathbf{1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2))$$

equivalente a

$$t = \frac{\mathbf{1}'(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)}{\sqrt{\mathbf{1}'\mathbf{S}_{pooled}\mathbf{1}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \frown t_{n_1+n_2-2}$$

- Assim, H_0 será rejeitada quando $|t| \geq t_{(n_1+n_2-2);1-\alpha/2}$

- O **teste à horizontalidade** dos perfis pode expressar-se pela hipótese

$$H_0 : \frac{1}{2}(\boldsymbol{\mu}_{11} + \boldsymbol{\mu}_{21}) = \frac{1}{2}(\boldsymbol{\mu}_{12} + \boldsymbol{\mu}_{22}) = ... = \frac{1}{2}(\boldsymbol{\mu}_{1q} + \boldsymbol{\mu}_{2q})$$

  equivalente a

$$H_0 : \frac{1}{2}\mathbf{C}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) = \mathbf{0}$$

- Para estimar $\boldsymbol{\mu} = \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)$ usamos $\bar{\mathbf{x}} = \frac{n_1\bar{\mathbf{x}}_1 + n_2\bar{\mathbf{x}}_2}{n_1+n_2}$
- Sob H_0

$$T^2 = (n_1 + n_2)(\mathbf{C}\bar{\mathbf{x}})'(\mathbf{C}\mathbf{S}_{pooled}\mathbf{C})^{-1}\mathbf{C}\bar{\mathbf{x}} \frown T^2_{q-1}(n_1 + n_2 - 2)$$

  onde

$$T^2 \overset{d}{=} \frac{(n_1 + n_2 - 1)(q - 1)}{n_1 + n_2 - q} F_{(q-1, n_1+n_2-q)}$$

- Assim, H_0 será rejeitada quando

$$T^2 > \frac{(n_1 + n_2 - 2)(q - 1)}{n_1 + n_2 - q} F_{(q-1, n_1+n_2-q);1-\alpha}$$

  onde $F_{(q-1,n_1+n_2-q);1-\alpha}$ representa o quantil de probabilidade $1 - \alpha$ da distribuição $F_{(q-1,n_1+n_2-q)}$

Exemplo 6

Considere os dados do ficheiro "data7.xlsx" referentes aos resultados (4 variáveis) de um teste psicológico aplicado a 32 homens (código 1) e 32 mulheres (código 2). Compare os perfis psicológicos das duas populações ($\alpha = 0.01$).

- Matriz de contrastes:

$$\mathbf{C}_{3\times 4} = \begin{bmatrix} -1 & 1 & 0 & 0 \\ 0 & -1 & 1 & 0 \\ 0 & 0 & -1 & 1 \end{bmatrix}$$

- Sob $H_0 : \mathbf{C}\boldsymbol{\mu}_1 = \mathbf{C}\boldsymbol{\mu}_2$

$$T^2 = (\mathbf{C}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2))' \left[\left(\frac{1}{n_1} + \frac{1}{n_2} \right) \mathbf{C}\mathbf{S}_{pooled}\mathbf{C}' \right]^{-1} (\mathbf{C}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)) = 74.24$$

Código R:

```
> T2<-t(C%*%(m1-m2))%*%solve((1/n1+1/n2)*(C%*%Spool%*%t(C)))%*%(C%*%(m1-m2))
> T2

        [,1]
[1,] 74.24037
```

sendo $T^2 \frown T_3^2(62) \stackrel{d}{=} \frac{(62)(3)}{60} F_{(3,60)}$

- Assim, H_0 será rejeitada quando $T^2 > 3.1F_{(3,60);0.99} = 12.79$

Código R:

```
> (((n1+n2-2)*(q-1))/(n1+n2-q))*qf(0.99,q-1,n1+n2-q)

[1] 12.79026
```

Logo, rejeita-se H_0, concluindo-se que não existem evidências a favor do paralelismo dos perfis.

Inferência sobre g $(g \geq 2)$ matrizes de covariâncias

Teste de igualdade de matrizes de covariância

- Consideremos agora g amostras independentes, de dimensões $n_1, \ldots, n_g$, extraídas de g populações multivariadas Normais. Pretende-se testar:

$$H_0 : \mathbf{\Sigma}_1 = ... = \mathbf{\Sigma}_g$$

$$H_1 : \exists\, i, j : \mathbf{\Sigma}_i \neq \mathbf{\Sigma}_j \, (i \neq j, i, j = 1, ..., g)$$

- Box (1950) definiu o teste usando a estatística

$$\Lambda = \frac{|\mathbf{S}_1|^{(n_1-1)/2}|\mathbf{S}_2|^{(n_2-1)/2}...|\mathbf{S_g}|^{(n_g-1)/2}}{|\mathbf{S}_{pooled}|^{\sum_i (n_i-1)/2}}$$

sendo

$$\mathbf{S}_{pooled} = \frac{1}{\sum_{i=1}^{k}(n_i-1)}\left(\sum_{i=1}^{k}(n_i-1)\mathbf{S}_i\right).$$

Teste de igualdade de matrizes de covariância

- O teste baseia-se na aproximação da distribuição de $-2\ln\Lambda$ pela distribuição χ^2:

$$U^* = -2(1-c_1)\ln\Lambda \overset{a}{\sim} \chi^2_{\frac{1}{2}(g-1)p(p+1)}$$

sendo

$$\ln\Lambda = \frac{1}{2}\sum_{i=1}^{g}(n_i-1)\ln|\mathbf{S}_i| - \frac{1}{2}\left(\sum_{i=1}^{g}(n_i-1)\right)\ln|S_{pooled}|.$$

e

$$c_1 = \left[\sum_{i=1}^{g}\frac{1}{n_i-1} - \frac{1}{\sum_{i=1}^{g}(n_i-1)}\right]\left[\frac{2p^2+3p-1}{6(p+1)(g-1)}\right]$$

- Para um nível de significância α, rejeita-se H_0 quando

$$U^* > \chi^2_{\frac{1}{2}(g-1)p(p+1)}(1-\alpha)\,.$$

Exemplo 7

Considere de novo os dados do ficheiro "data7.xlsx". Teste a hipótese $H_0 : \mathbf{\Sigma}_1 = \mathbf{\Sigma}_2$ ($\alpha = 0.01$).