Comparação das médias de mais de duas populações

Vicente G. Cancho
garibay@icmc.usp.br

Sumário

- Exemplo: Crânios Egípcios
- Modelos de ANOVA de um fator
- Modelos de MANOVA de um fator

Exemplo: Crânios Egípcios

Vamos agora considerar a situação em que observamos dados de várias populações independentes. Considere os dados de crânios egípcios (dados 'skulls' da library "HSAUR3" do R) O conjunto de dados contém observações em 4 variáveis para 150 crânios de 5 épocas (c4000BC c3300BC, c1850BC, c200BC e cAD150), as quatro variáveis são:

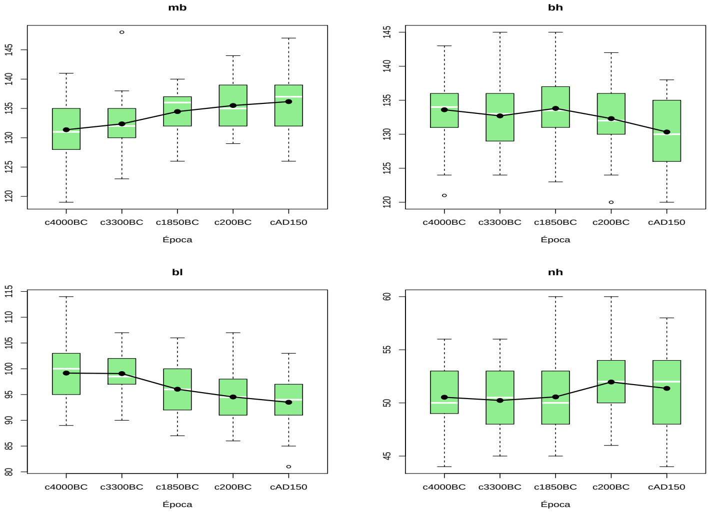
- mb: largura máxima do crânio.
- bh: altura basibregmática do crânio.
- bl: comprimento basialveolar do crânio.
- nh: altura nasal do crânio.

Exemplo: Crânios Egípcios

Portanto, existem cinco grupos (cinco épocas) e queremos comparar a média de (mb, bh, bl, nh) entre os cinco grupos. De acordo com Everitt e Hothorn (2010), "A Handbook of Statistical Analyzes Using R", se as medidas dos crânios forem diferentes ao longo do tempo, isso indicaria cruzamento com populações de imigrantes.

	Variável resposta			
Época	mb	bh	bl	nh
c4000BC	131.367	133.600	99.167	50.533
c3300BC	132.367	132.700	99.067	50.233
c1850BC	134.467	133.800	96.033	50.567
c200BC	135.500	132.300	94.533	51.967
cAD150	136.167	130.333	93.500	51.367

Exemplo: Crânios Egípcios



Os pontos pretos sólidos representam a média da variável correspondente para a época correspondente

Exemplo: Crânios Egípcios

- Em geral, a variável que cria o grupo é chamada de fator.
- Os diferentes valores que um fator assume são chamados de níveis.
- Nesse caso, há apenas um fator, 'época', com cinco níveis (os anos).
- Assim, vamos empregar um modelo de Análise de Variância (ANOVA) de uma via para testar se os vetores médios para grupos diferentes são iguais ou não.

Modelo de ANOVA de um fator:Univariado

Vamos revisar a ANOVA de um fator na situação univariada. Suponha que observamos amostras independentes de g grupos

- Amostra 1: $X_{11}, \ldots, X_{1n_1}$ de uma população $N(\mu_1, \sigma^2)$
- Amostra 2: $X_{21}, \ldots, X_{2n_2}$ de uma população $N(\mu_2, \sigma^2)$

$$\vdots$$

- Amostr g: $X_{g1}, \ldots, X_{gn_g}$ de uma população $N(\mu_g, \sigma^2)$.

Aqui, a notação X_{ij} denota a resposta para o j-ésimo indivíduo no i-ésimo grupo.
Suposições:

- Cada população é normal;
- grupos/populações tem médias diferentes, mas a mesma variãncia σ^2;
- as amostras são mutamente independentes.

Hipótese de interesse: $H_0 : \mu_1 = \ldots = \mu_g$ vs. H_1 : pelo menos duas médias são diferentes.
Escrevemos o modelo ANOVA como

$$\underbrace{X_{ij}}_{\left(\begin{array}{c}\text{Resposta para } j\text{-ésimo}\\ \text{indivíduo no } i\text{-ésimo grupo}\end{array}\right)} = \underbrace{\mu_i}_{\left(\text{Média } i\text{-ésima população}\right)} + \underbrace{e_{ij}}_{\left(\text{erro } N(0,\sigma^2)\right)}.$$

Em geral decompomos as médias do grupo como

$$\underbrace{\mu_i}_{\left(\text{média do } i\text{-ésimo grupo}\right)} = \underbrace{\mu}_{\left(\text{média global}\right)} + \underbrace{\tau_i}_{\left(\text{efeito do } i\text{-ésimo grupo}\right)}.$$

Assim o modelo de ANOVA de um fator é dada por

$$X_{ij} = \mu + \tau_i + e_{ij}, \; e_{ij} \overset{i.i.d}{\sim} N(0, \sigma^2).$$

Modelo de ANOVA de um fator:Univariado

- O efeito de i -ésimo grupo é $\tau_i =$ (média do i-ésimo grupo-média geral), ou seja

$$\tau_i = \mu_i - \mu, \; i = 1, \ldots, g$$

- Muitas vezes colocamos a restrição:

$$\sum_{i=1}^{g} n_i \tau_i = 0.$$

- As estimativas dos parâmetros do modelo de Anova

$$\hat{\mu} = \bar{x}, \;\; \hat{\tau}_i = \bar{x}_i - \bar{x}$$

Modelo de ANOVA de um fator:Univariado

- Motivados pela decomposição acima, também podemos decompor os dados observados da seguinte forma:

$$\underbrace{x_{ij}}_{\text{(observação)}} = \underbrace{\bar{x}}_{\text{(Média amostral global)}} + \underbrace{(\bar{x}_i - \bar{x})}_{\text{(efeito grupo estimado)}} + \underbrace{(x_{ij} - \bar{x}_i)}_{\text{(residual)}}.$$

- Note que se $H_0 : \tau_1 = \tau_2 = \ldots = \tau_q = 0$ é verdade, isto é, se cada $\alpha_i = 0$, então cada $\hat{\alpha}_i = \bar{x}_i - \bar{x}$ também deve ser próximo de zero.
- Portanto, se observarmos que alguns dos efeitos de grupo, $\bar{x}_i - \bar{x}$, são grandes, devemos rejeitar H_0.
- A variabilidade das observações é decomposta em 2 componentes:

$$\sum_{i=1}^{g}\sum_{j=1}^{n_i}(x_{ij} - \bar{x})^2 = \sum_{i=1}^{g}\sum_{j=1}^{n_i}(\bar{x}_i - \bar{x})^2 + \sum_{i=1}^{g}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_i)^2$$

ANOVA Test F

Para avaliar o efeito do grupo, observamos a soma dos quadrados:

$$\text{Soma de quadrados de tratamento (SQTr)} = \sum_{i=1}^{g} n_i(\bar{x}_i - \bar{x})^2,$$

$$\text{Soma de quadrados do erro (SQE)} = \sum_{i=1}^{g}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_i)^2.$$

O test F rejeita H_0 se

$$\frac{SQTr/(g-1)}{SQE/(\sum_i n_i - g)} > F_{g-1,\sum_i n_i - g}(\alpha).$$

Modelo de ANOVA de um fator:Univariado

- A quantidade do numerador: $SSTr/(g-1)$ é denominado de quadrado médio do tratamento (QMTr).
- Se a hipótese nula for verdadeira (ou seja, todas as médias do grupo são iguais), então a média da amostra individual $\boldsymbol{x}_i$ do i-ésimo grupo, deve estar próxima da média geral da amostra combinada $\boldsymbol{x}$. Assim, sob H_0, o QMTr teria um valor pequeno.
- Essa quantidade mede o quão semelhantes as amostras são em termos de suas médias. Esta é uma medida de variabilidade entre as amostras.
- A quantidade do denominador: $SQE/(\sum_i n_i - g)$ é denominado de quadrado médio do erro (QME),
- Assim, a estatística de teste acima compara a variabilidade entre amostras e a variabilidade dentro das amostras.

Modelo de ANOVA de um fator:Univariado

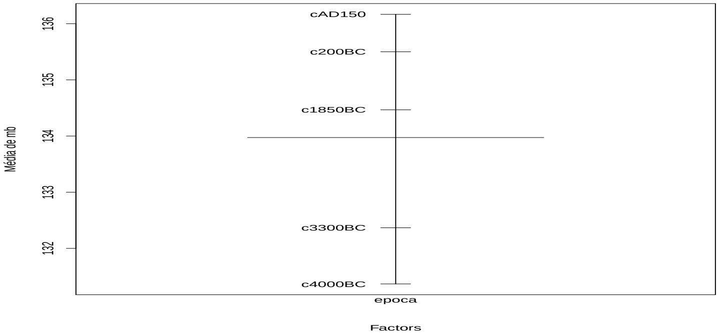
Tabela: Tabela de ANOVA para testar $H_0 : \mu_1 = \mu_2 = \cdots = \mu_g$

Fonte de Variação	SQ	GL	Quadrado Médio	F_0
Entre amostras	SQtr	$g - 1$	QMtr	$\frac{QMtr}{QME}$
Dentro de amostras	SQE	$\sum_{i=1}^{g}(n_i - 1)$	QME	
Total	SQtotal	$\sum_{i=1}^{g} n_i - 1$		

Exemplo

Vamos considerar os dados dos crânios Egipcios e considerar apenas uma variável, mb: largura máxima do crânio(mb) e realizar a ANOVA.

Modelo de ANOVA de um fator:Univariado



As médias de amostra para os dados dos crânios Egipcios são mostradas na Figura. A linha mais larga no meio representa a média geral dos dados.

```
> fit <- aov(mb ~ epoca)
> summary(fit)
              Df Sum Sq Mean Sq F value   Pr(>F)
epoca          4  502.8  125.71   5.955 0.000183
Residuals    145 3061.1   21.11
```

- O pequeno p-valor indica que devemos rejeitar H_0, ou seja, as médias do grupo não são iguais.
- Note que assumimos que todas as populações tem a mesma variância σ^2.
- A estimativa de σ^2 é dado por QME. Na tabela ANOVA, tem-se $\hat{\sigma}^2 = 21,11$.

Suponha que observamos amostras de g grupos independentes:

- $\boldsymbol{X}_{11}, \ldots, \boldsymbol{X}_{1n_1}$ a.a. de uma população $N_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$,
- $\boldsymbol{X}_{21}, \ldots, \boldsymbol{X}_{2n_2}$ a.a. de uma população $N_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$

$$\vdots$$

- $\boldsymbol{X}_{g1}, \ldots, \boldsymbol{X}_{gn_g}$ a.a. de uma população $N_p(\boldsymbol{\mu}_g, \boldsymbol{\Sigma})$

Suposições:

- cada população é normal multivariada;
- cada grupo/população tem vetor de médias diferentes mas a mesma covariância;
- as amostras são independentes.

Deseja-se avaliar as hipóteses

$$H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \ldots \boldsymbol{\mu}_g = \mu \text{ contra}$$
$$H_1 : \text{ pelo menos um } \boldsymbol{\mu}_i \text{ diferente}$$

Para isso, vamos considerar a reparametrização

$$\boldsymbol{\mu}_k = \boldsymbol{\mu} + \boldsymbol{\tau}_k, \; k = 1, \ldots, g,$$

e então avaliar se

$$H_0 : \boldsymbol{\tau}_1 = \boldsymbol{\tau}_2 = \ldots \boldsymbol{\tau}_g = \mathbf{0} \text{ contra}$$
$$H_1 : \text{ pelo menos um } \boldsymbol{\tau} \text{ diferente de } \mathbf{0}$$

Modelo ANOVA Multivariada-MANOVA

O modelo de MANOVA de um fator pode-se escrever:

$$\boldsymbol{X}_{kj} = \boldsymbol{\mu} + \boldsymbol{\tau}_k + \boldsymbol{\epsilon}_{kj},$$

para $j = 1, \ldots, n_k, k = 1, \ldots, g$.

Suposições:

- $\boldsymbol{\epsilon}_{kj} \overset{ind}{\sim} N_p(\mathbf{0}, \Sigma)$,
- $\boldsymbol{\mu}$ é o vetor de média geral,
- $\boldsymbol{\tau}_k$ é o vetor de efeito do $k-$ ésimo tratamento,
- $\displaystyle\sum_{k=1}^{g} n_k \boldsymbol{\tau}_k = \mathbf{0}$.

Modelo de ANOVA multivariada - MANOVA

- Como ao caso univariado, escrevemos a resposta multivariada como

$$\underbrace{\boldsymbol{x}_{ij}}_{\left(\text{Observação}\right)} = \underbrace{\bar{\boldsymbol{x}}}_{\left(\text{Média Global}\right)} + \underbrace{(\bar{\boldsymbol{x}}_i - \bar{\boldsymbol{x}})}_{\left(\substack{\text{estimativativa}\\ \text{efeito do grupo}}\right)} + \underbrace{(\boldsymbol{x}_{ij} - \bar{\boldsymbol{x}}_i)}_{\left(\text{residual}\right)}.$$

  Aqui, cada termo é um vetor, portanto, precisamos de um análogo mutivariado da soma univariada dos quadrados, a matriz da soma dos quadrados e produtos cruzados:

- Da decomposição da variabilidade dos dados em torno da média em variabilidade intra e entre tratamentos. Temos que

$$(\boldsymbol{X}_{kj} - \overline{\boldsymbol{X}})(\boldsymbol{X}_{kj} - \overline{\boldsymbol{X}})^\top = \left[(\boldsymbol{X}_{kj} - \overline{\boldsymbol{X}}_k) + (\overline{\boldsymbol{X}}_k - \overline{\boldsymbol{X}})\right]\left[(\boldsymbol{X}_{kj} - \overline{\boldsymbol{X}}_k) + (\overline{\boldsymbol{X}}_k - \overline{\boldsymbol{X}})\right]^\top =$$
$$= (\boldsymbol{X}_{kj} - \overline{\boldsymbol{X}}_k)(\boldsymbol{X}_{kj} - \overline{\boldsymbol{X}}_k)^\top + (\boldsymbol{X}_{kj} - \overline{\boldsymbol{X}}_k)(\overline{\boldsymbol{X}}_k - \overline{\boldsymbol{X}})^\top + (\overline{\boldsymbol{X}}_k - \overline{\boldsymbol{X}})(\boldsymbol{X}_{kj} - \overline{\boldsymbol{X}}_k)^\top$$
$$+ (\overline{\boldsymbol{X}}_k - \overline{\boldsymbol{X}})(\overline{\boldsymbol{X}}_k - \overline{\boldsymbol{X}})^\top$$

- Somando em j, nas componentes

$$\sum_{j=1}^{n_k}\left[(\boldsymbol{X}_{kj}-\overline{\boldsymbol{X}}_k)(\overline{\boldsymbol{X}}_k-\overline{\boldsymbol{X}})^\top\right] =$$
$$\sum_{j=1}^{n_k}\left[(\boldsymbol{X}_{kj}-\overline{\boldsymbol{X}}_k)\right](\overline{\boldsymbol{X}}_k-\overline{\boldsymbol{X}})^\top =$$
$$(\sum_{j=1}^{n_k}\boldsymbol{X}_{kj}-n_k\overline{\boldsymbol{X}}_k)(\overline{\boldsymbol{X}}_k-\overline{\boldsymbol{X}})^\top =$$
$$(\sum_{j=1}^{n_k}\boldsymbol{X}_{kj}-\sum_{j=1}^{n_k}\boldsymbol{X}_{kj})(\overline{\boldsymbol{X}}_k-\overline{\boldsymbol{X}})^\top = 0$$

- Idem para

$$\sum_{j=1}^{n_k}(\overline{\boldsymbol{X}}_k-\overline{\boldsymbol{X}})(\boldsymbol{X}_{kj}-\overline{\boldsymbol{X}}_k)^\top = 0$$

- Daí a soma em j e em k, resulta

$$\underbrace{\sum_{k=1}^{g}\sum_{j=1}^{n_k}(\boldsymbol{X}_{kj}-\overline{\boldsymbol{X}})(\boldsymbol{X}_{kj}-\overline{\boldsymbol{X}})^\top}_{T} =$$

$$= \underbrace{\sum_{k=1}^{g} n_k(\overline{\boldsymbol{X}}_k-\overline{\boldsymbol{X}})(\overline{\boldsymbol{X}}_k-\overline{\boldsymbol{X}})^\top}_{B} + \underbrace{\sum_{k=1}^{g}\sum_{j=1}^{n_k}(\boldsymbol{X}_{kj}-\overline{\boldsymbol{X}}_k)(\boldsymbol{X}_{kj}-\overline{\boldsymbol{X}}_k)^\top}_{W}$$

- Assim,
  $T = B + W$, com
  - T: soma de quadrados e produtos cruzados **total**
  - B: soma de quadrados e produtos cruzados **entre** tratamentos
  - W: soma de quadrados e produtos cruzados **intra** tratamento (**dentro**)

Consideramos as somas de quadrados e produtos cruzados

$$T = \sum_{k=1}^{g} \sum_{j=1}^{n_k} (\boldsymbol{X}_{kj} - \overline{\boldsymbol{X}})(\boldsymbol{X}_{kj} - \overline{\boldsymbol{X}})^\top (\textbf{total})$$

$$B = \sum_{k=1}^{g} n_k (\overline{\boldsymbol{X}}_k - \overline{\boldsymbol{X}})(\overline{\boldsymbol{X}}_k - \overline{\boldsymbol{X}})^\top (\textbf{entre})$$

$$W = \sum_{k=1}^{g} \sum_{j=1}^{n_k} (\boldsymbol{X}_{kj} - \overline{\boldsymbol{X}}_k)(\boldsymbol{X}_{kj} - \overline{\boldsymbol{X}}_k)^\top (\textbf{dentro})$$

Temos

$$T = B + W$$

$$SQT = SQTrat + SQRes$$

Como supomos que todos os grupos têm a mesma variância, podemos considerar como estimativa de Σ:

$$S_{pooled} = W = \sum_{k=1}^{g} \sum_{j=1}^{n_k} (\boldsymbol{X}_{kj} - \overline{\boldsymbol{X}}_k)(\boldsymbol{X}_{kj} - \overline{\boldsymbol{X}}_k)^\top$$

$$W = (n_1 - 1)\boldsymbol{S}_1 + (n_2 - 1)\boldsymbol{S}_2 + \ldots + (n_g - 1)\boldsymbol{S}_g.$$

Tabela ANOVA multivariada

Fonte de variação	Somas de quadrados e produtos cruzados	graus de liberdade
Tratamento	B	$g-1$
Resíduo	W	$N-g$
Total	T	$N-1$

em que $N = \sum_{k=1}^{g} n_k$. Assim, para avaliar

$$H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \ldots \boldsymbol{\mu}_g = \mu \text{ contra } H_1 : \text{ pelo menos um } \boldsymbol{\mu}_i \text{ diferente}$$

ou seja

$$H_0 : \boldsymbol{\tau}_1 = \boldsymbol{\tau}_2 = \ldots \boldsymbol{\tau}_g = \mathbf{0} \text{ contra } H_1 : \text{ pelo menos um } \boldsymbol{\tau} \text{ diferente de } \mathbf{0}.$$

- A hipótese é rejeitada se a razão,

$$\Lambda^\star = \frac{|W|}{|W+B|} = \frac{\left|\sum_{k=1}^{g}\sum_{j=1}^{n_k}(\boldsymbol{X}_{kj}-\overline{\boldsymbol{X}}_k)(\boldsymbol{X}_{kj}-\overline{\boldsymbol{X}}_k)^\top\right|}{\left|\sum_{k=1}^{g}\sum_{j=1}^{n_k}(\boldsymbol{X}_{kj}-\overline{\boldsymbol{X}})(\boldsymbol{X}_{kj}-\overline{\boldsymbol{X}})^\top\right|}$$

  for muito pequeno.
- $\Lambda^\star$ é conhecido com lambda de Wilk.
- É equivalente à estatística da razão de verossimilhança.

Testando Hipóteses com Estatística Wilks

- $\Lambda^\star$ é uma razão de variâncias amostrais generalizadas

$$\Lambda^\star = \frac{|W|}{|T|} = \frac{\prod_{i=1}^{p} \lambda_i}{\prod_{i=1}^{p} \lambda_i^*}$$

onde λ_i's são autovalores de W e λ_i^*'s são autovalores de T

- Se $H_0 : \boldsymbol{\tau}_1 = \ldots = \boldsymbol{\tau}_g = \mathbf{0}$ é verdadeiro, então B está próximo de $\mathbf{0}$
  - $\Longrightarrow T \approx W$
  - $\Longrightarrow \lambda_i \approx \lambda_i^*$
  - $\Longrightarrow \Lambda^\star$ é próximo de 1.
- Se $H_0 : \boldsymbol{\tau}_1 = \ldots = \boldsymbol{\tau}_g = \mathbf{0}$ é falso, então B não está próximo de $\mathbf{0}$
  - $\Longrightarrow$ os valores nas diagonais de T, que serão positivos grandes.
  - $\Longrightarrow \lambda_i < \lambda_i^*$
  - $\Longrightarrow \Lambda^\star$ é pequena .
- A distribuição exata de Λ^* pode ser derivada para casos especiais de p e g.

Distribuição da Estatística Wilks

No. of variables	No. of groups	Sampling distribution for multivariate normal data
$p = 1$	$g \geq 2$	$\left(\frac{\Sigma n_\ell - g}{g - 1}\right)\left(\frac{1 - \Lambda^*}{\Lambda^*}\right) \sim F_{g-1, \Sigma n_\ell - g}$
$p = 2$	$g \geq 2$	$\left(\frac{\Sigma n_\ell - g - 1}{g - 1}\right)\left(\frac{1 - \sqrt{\Lambda^*}}{\sqrt{\Lambda^*}}\right) \sim F_{2(g-1), 2(\Sigma n_\ell - g - 1)}$
$p \geq 1$	$g = 2$	$\left(\frac{\Sigma n_\ell - p - 1}{p}\right)\left(\frac{1 - \Lambda^*}{\Lambda^*}\right) \sim F_{p, \Sigma n_\ell - p - 1}$
$p \geq 1$	$g = 3$	$\left(\frac{\Sigma n_\ell - p - 2}{p}\right)\left(\frac{1 - \sqrt{\Lambda^*}}{\sqrt{\Lambda^*}}\right) \sim F_{2p, 2(\Sigma n_\ell - p - 2)}$

Bartlett mostrou que se H_0 é verdadeira e $\sum_{k=1}^{g} = n$ é grande, então

$$-\left(n - 1 - \frac{p+g}{2}\right)\log(\Lambda^\star) \sim \chi^2_{p(g-1)}.$$

Bartlett, M. S. "Further Aspects of the Theory of Multiple Regression." Proceedings of the Cambridge Philosophical Society, 34 (1938), 33-40.

Algumas observações

- É possível obter distribuições exatas para Λ^* de Wilks
- Além disso, existem outras estatísticas disponíveis para testar H_0.
    - Traço de Pillai
    - Traço de Hotelling
    - Maior autovalor de Roy
- Todas as quatro estatísticas são equivalentes (mesmo poder) para amostras grandes.
- Para amostras (relativamente) pequenas, as estatísticas Wilks, Pillar e Hotelling são equivalentes.
- Traço de Pillai tem se mostrado mais robusto com relação à não normalidade.
- Para detalhes, ver Johnson & Wichern (2007) cap 6.

Exemplo: Crânios Egípicios

Considere os dados de Crânios Egípicios, mas com todas as quatro variáveis. Lembre-se de que existem cinco grupos (cinco épocas) e queremos comparar a média de (mb, bh, bl, nh) entre os cinco grupos.

		Variáveis			
Época	n_i	mb	bh	bl	nh
c4000BC	30	131.367	133.600	99.167	50.533
c3300BC	30	132.367	132.700	99.067	50.233
c1850BC	30	134.467	133.800	96.033	50.567
c200BC	30	135.500	132.300	94.533	51.967
cAD150	30	136.167	130.333	93.500	51.367

Exemplo: Crânios Egípicios

As matrizes de covariâncias dos grupos:

$$S_1 = \begin{pmatrix} 26.31 & 4.15 & 0.45 & 7.25 \\ 4.15 & 19.97 & -0.79 & 0.39 \\ 0.45 & -0.79 & 34.63 & -1.92 \\ 7.25 & 0.39 & -1.92 & 7.64 \end{pmatrix}, S_2 = \begin{pmatrix} 23.14 & 1.01 & 4.77 & 1.84 \\ 1.01 & 21.60 & 3.37 & 5.62 \\ 4.77 & 3.37 & 18.89 & 0.19 \\ 1.84 & 5.62 & 0.19 & 8.74 \end{pmatrix}$$

$$S_3 = \begin{pmatrix} 12.12 & 0.79 & -0.77 & 0.90 \\ 0.79 & 24.79 & 3.59 & -0.09 \\ -0.77 & 3.59 & 20.72 & 1.67 \\ 0.90 & -0.09 & 1.67 & 12.60 \end{pmatrix}, S_4 = \begin{pmatrix} 15.36 & -5.53 & -2.17 & 2.05 \\ -5.53 & 26.36 & 8.11 & 6.15 \\ -2.17 & 8.11 & 21.09 & 5.33 \\ 2.05 & 6.15 & 5.33 & 7.96 \end{pmatrix}$$

$$S_5 = \begin{pmatrix} 28.63 & -0.23 & -1.88 & -1.99 \\ -0.23 & 24.71 & 11.72 & 2.15 \\ -1.88 & 11.72 & 25.57 & 0.40 \\ -1.99 & 2.15 & 0.40 & 13.83 \end{pmatrix}$$

Exemplo: Crânios Egípicios

A matriz da soma de quadrados e produtos cruzados:

$$
\begin{aligned}
W &= (n_1 - 1)S_1 + \cdots + (n_5 - 1)S_5 \\
&= \begin{pmatrix} 3061.07 & 5.33 & 11.47 & 291.30 \\ 5.33 & 3405.27 & 754.00 & 412.53 \\ 11.47 & 754.00 & 3505.97 & 164.33 \\ 291.30 & 412.53 & 164.33 & 1472.13 \end{pmatrix},
\end{aligned}
$$

$$
T = \begin{pmatrix} 3563.89 & -222.81 & -615.16 & 426.73 \\ -222.81 & 3635.17 & 1046.28 & 346.47 \\ -615.16 & 1046.28 & 4309.26 & -16.40 \\ 426.73 & 346.47 & -16.40 & 1533.33 \end{pmatrix}
$$

e

$$
B = T - W = \begin{pmatrix} 502.83 & -228.15 & -626.63 & 135.43 \\ -228.15 & 229.91 & 292.28 & -66.07 \\ -626.63 & 292.28 & 803.29 & -180.73 \\ 135.43 & -66.07 & -180.73 & 61.20 \end{pmatrix}
$$

Exemplo: Crânios Egípicios

Hipótese a ser testada

$$H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \ldots = \mu_5$$

em média as características dos crânios são similares nas diferentes épocas.

A estatística de Wilks observada é

$$\Lambda^\star = \frac{W}{B + W} = 0.664.$$

No Exemplo tem-se $n = 150$, $n_i = 30$, $g = 5$ e $p = 4$. Considerando as suposições para MANOVA e como $n = 150$ é relativamente grande, considere-se a estatística Bartlet,

$$-\left(n - 1 - \frac{p+g}{2}\right)\log(\Lambda^\star) = -\left(150 - 1 - \frac{4+5}{2}\right)\log(0.664) = 62.745$$

Para $\alpha = 0,05$ tem-se o nível crítico $\chi^2_{p(g-1)=16}(0,05) = 26.296$. Dessa forma, rejeita-se H_0 a um nível de 5%. Como H_0 é rejeitada, é necessário calcular intervalos de confiança para os possíveis pares de grupos para identificar a origem da diferença entre as variáveis.

MANOVA no R

```
> fit_1 <- manova(dat ~ epoca)
> summary( fit_1, test = "Wilks")
            Df   Wilks approx F num Df den Df    Pr(>F)
epoca        4 0.66359   3.9009     16 434.45 7.01e-07 ***
Residuals 145
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1
> summary( fit_1, test = "Pillai")
            Df  Pillai approx F num Df den Df     Pr(>F)
epoca        4 0.35331    3.512     16    580 4.675e-06 ***
Residuals 145
---
```

```
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1
> summary( fit_1, test = "Hotelling-Lawley")
            Df Hotelling-Lawley approx F num Df den Df    Pr(>F)
epoca        4          0.48182    4.231     16    562 8.278e-08 ***
Residuals 145
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1
> summary( fit_1, test = "Roy")
            Df    Roy approx F num Df den Df    Pr(>F)
epoca        4 0.4251    15.41      4    145 1.588e-10 ***
Residuals 145
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1
```

Comparação Multiplas

- Quando a hipótese de efeitos de igualdade de tratamento é rejeitada, aqueles efeitos que levaram à rejeição da hipótese são de interesse.
- Para comparações de pares, a abordagem de Bonferroni pode ser usada para construir intervalos de confiança simultâneos para os componentes das diferenças $\boldsymbol{\tau}_k - \boldsymbol{\tau}_\ell$ (ou $\boldsymbol{\mu}_k - \boldsymbol{\mu}_\ell$).
- Pois, $\boldsymbol{\tau}_k - \boldsymbol{\tau}_\ell = \boldsymbol{\mu}_k - \boldsymbol{\mu}_\ell$
- Esses intervalos são mais curtos do que aqueles obtidos para todos os contrastes e exigem valores críticos a estatística t univariada.

- Seja τ_{ki} o i-ésimo componente de $\boldsymbol{\tau}_k$. Uma vez que $\boldsymbol{\tau}_k$ é estimado por $\hat{\boldsymbol{\tau}}_k = \bar{\boldsymbol{x}}_k - \bar{\boldsymbol{x}}$.

$$\hat{\tau}_{ki} = \bar{x}_{ki} - \bar{x}_i$$

e

$$\hat{\tau}_{ki} - \hat{\tau}_{\ell i} = \bar{x}_{ki} - \bar{x}_{\ell i}$$

é a diferença de duas médias amostrais independentes.

- Note que

$$Var(\bar{X}_{ki} - \bar{X}_{\ell i}) = \left(\frac{1}{n_k} + \frac{1}{n_\ell}\right)\sigma_{ii}$$

onde σ_{ii} é o i-ésimo elemento diagonal de Σ.

- Sob as suposições do modelo de MANOVA $\bar{X}_{ki} - \bar{X}_{\ell i}$ tem distribuição normal com média $\mu_{ki} - \mu_{\ell i}$ e variância $Var(\bar{X}_{ki} - \bar{X}_{\ell i})$.

- A estimativa de $Var(\bar{X}_{ki} - \bar{X}_{\ell i})$ é

$$\widehat{Var(\bar{X}_{ki} - \bar{X}_{\ell i})} = \left(\frac{1}{n_k} + \frac{1}{n_\ell}\right) \frac{w_{ii}}{n-g}$$

  onde w_{ii} é o i-ésimo elemento diagonal de W e $n = n_1 + \ldots + n_g$.

- Existem p variáveis e $g(g-1)/2$ pares diferenças de pares, então cada intervalo baseado na distribuição t de duas amostras, consideraremos o valor crítico $t_{n-g}(\alpha/2m)$, onde

$$m = pg(g-1)/2.$$

  é o número de comparações de confiança simultâneas.

- No modelo de MANOVA, um intervalo de pelo menos $100(1-\alpha)\%$ de confiança para $\mu_{ki} - \mu_{\ell i}$ é dada por

$$\bar{x}_{ki} - \bar{x}_{\ell i} \mp t_{n-g}\left(\frac{\alpha}{pg(g-1)}\right)\sqrt{\left(\frac{1}{n_k} + \frac{1}{n_\ell}\right)\frac{w_{ii}}{n-g}},$$

para $i = 1, \ldots, p, \quad \ell < k = 1, \ldots, g.$

No Exemplo de crânios Egípicios, para a variável largura máxima (mb), vamos a determinar entre que épocas existe diferênças significativas. Dos dados tem-se: $n_i = 30$, n=150, $g = 5$, $p = 4$, $m = pg(g-1)/2 = 4 \times 5 \times 4/2 = 40$, n-g=145 e

$$\bar{\boldsymbol{x}}_1 = \begin{pmatrix} 131.367 \\ 132.367 \\ 134.467 \\ 135.500 \\ 136.167 \end{pmatrix} \quad e \quad W = \begin{pmatrix} 3061.07 & 5.33 & 11.47 & 291.30 \\ 5.33 & 3405.27 & 754.00 & 412.53 \\ 11.47 & 754.00 & 3505.97 & 164.33 \\ 291.30 & 412.53 & 164.33 & 1472.13 \end{pmatrix},$$

Comparações múltiplas:Exemplo

Para $1-\alpha = 0,96$, tem-se $t_{n-g}\left(\frac{\alpha}{pg(g-1)}\right) = t_{145}(\frac{0,05}{80}) = 3,29$ e

$$
\begin{aligned}
& t_{n-g}\left(\frac{\alpha}{pg(g-1)}\right)\sqrt{\left(\frac{1}{n_1}+\frac{1}{n_2}\right)\frac{w_{11}}{n-g}} \\
&= t_{145}(\frac{0,05}{80})\sqrt{\left(\frac{1}{30}+\frac{1}{30}\right)\frac{3061,07}{145}} \\
&= 1,19
\end{aligned}
$$

Intervalo de confiança para $\mu_{11}-\mu_{12}$ é dada por

$$(131,367-132,367)\mp 1,19 = (-4,91; 2,905)$$

Comparações múltiplas: Exemplo

Assim

$$\begin{aligned}
\mu_{11}-\mu_{12} &: -1.000 \mp 1.19 = (-4.91, 2.90)\\
\mu_{11}-\mu_{13} &: -3.100 \mp 1.19 = (-7.01, 0.805)\\
\mu_{11}-\mu_{14} &: -4.133 \mp 1.19 = (-8.04, -0.228)\\
\mu_{11}-\mu_{15} &: -4.800 \mp 1.19 = (-8.71, -0.895)\\
\mu_{12}-\mu_{13} &: -2.100 \mp 1.19 = (-6.01, 1.805)\\
\mu_{12}-\mu_{14} &: -3.133 \mp 1.19 = (-7.04, 0.772)\\
\mu_{12}-\mu_{15} &: -3.800 \mp 1.19 = (-7.71, 0.105)\\
\mu_{13}-\mu_{14} &: -1.033 \mp 1.19 = (-4.94, 2.872)\\
\mu_{13}-\mu_{15} &: -1.700 \mp 1.19 = (-5.61, 2.205)\\
\mu_{14}-\mu_{15} &: -0.667 \mp 1.19 = (-4.57, 3.239)
\end{aligned}$$

Os intervalos que não contêm zero indicam uma diferença significativa entre as médias dos grupos correspondentes. Parece, para a variável mb, as médias entre as épocas c4000BC e c200BC e entre c4000BC e cAD150 são significativamente diferentes de zero.

```
> library(emmeans)
> # Manova
> fit_1 <- manova(dat ~ epoca)
> # média da variável resposta mb
> mbmedia <- emmeans(fit_1, "epoca", weights=c(1,0,0,0))
> mbmedia
epoca    emmean    SE  df lower.CL upper.CL
c4000BC     131 0.839 145      130      133
c3300BC     132 0.839 145      131      134
c1850BC     134 0.839 145      133      136
c200BC      136 0.839 145      134      137
cAD150      136 0.839 145      135      138

Results are averaged over the levels of: rep.meas
Confidence level used: 0.95
```

```
> # numero de variavveis
> p <- 4
> # numero de gropos
> g <- 5
> # nivel de significance
> alpha <- 0.05
> # numero de comparações
> m <- p*g*(g-1)/2
> alpha.n <- alpha/m
> #Definido os contraste
> cont <- contrast(mbmeans, "pairwise")
> # differença 2 a 2 de `mb`
> bb <- confint(cont, level = 1-alpha.n, adj="none")
> bb
contrast          estimate   SE  df lower.CL upper.CL
c4000BC - c3300BC   -1.000 1.19 145    -4.91    2.905
c4000BC - c1850BC   -3.100 1.19 145    -7.01    0.805
c4000BC - c200BC    -4.133 1.19 145    -8.04   -0.228
c4000BC - cAD150    -4.800 1.19 145    -8.71   -0.895
c3300BC - c1850BC   -2.100 1.19 145    -6.01    1.805
c3300BC - c200BC    -3.133 1.19 145    -7.04    0.772
c3300BC - cAD150    -3.800 1.19 145    -7.71    0.105
c1850BC - c200BC    -1.033 1.19 145    -4.94    2.872
c1850BC - cAD150    -1.700 1.19 145    -5.61    2.205
c200BC - cAD150     -0.667 1.19 145    -4.57    3.239
```

Exercício

Para os dados do Exemplo considere as variáveis:

- bh: altura basibregmática do crânio.
- bl: comprimento basialveolar do crânio.
- nh: altura nasal do crânio.

determine entre que epocas existe diferenças significativas.