Unidade II. Análise de variância multivariada (MANOVA)

2.1. Introdução

A Análise de variância multivariada representa uma extensão da Análise de variância univariada (ANOVA) envolvendo mais de uma variável dependente. Ela é uma técnica estatística utilizada para explorar o relacionamento entre diversas variáveis independentes categóricas (usualmente referidas como grupos ou tratamentos) e duas ou mais variáveis dependentes quantitativas. Uma aplicação comum é na área de delineamentos de experimentos.
Para a comparação de mais de duas populações, amostras aleatórias podem ser coletadas e organizadas da seguinte maneira:

Amostra 1: $\mathbf{X}_{11}, \mathbf{X}_{12}, \ldots, \mathbf{X}_{1n_1}$ da pop. 1
Amostra 2: $\mathbf{X}_{21}, \mathbf{X}_{22}, \ldots, \mathbf{X}_{2n_2}$ da pop. 2
⋮ ⋮
Amostra g: $\mathbf{X}_{g1}, \mathbf{X}_{g2}, \ldots, \mathbf{X}_{gn_g}$ da pop. g

Obedecendo as seguintes suposições para a estrutura dos dados:

a) Cada amostra $\mathbf{X}_{k1}, \mathbf{X}_{k2}, \ldots, \mathbf{X}_{kn_k}$ é uma amostra de tamanho n_k de uma população com vetor de médias $\boldsymbol{\mu}_k$, k = 1, 2,..., g;
b) todas as populações têm matrizes de covariâncias comum $\boldsymbol{\Sigma}$;
c) cada população é normal p-variada;
d) as amostras são independentes.

Obs. A suposição (c) pode ser desconsiderada no caso de grandes amostras recorrendo ao Teorema Central do Limite.

2.2. MANOVA a um fator – Modelo linear multivariado (One way)

O modelo estatístico para comparar g vetores médios populacionais pode ser definido por:

$$\mathbf{X}_{ki} = \boldsymbol{\mu} + \boldsymbol{\tau}_k + \mathbf{e}_{ki} \qquad k = 1,2,\ldots,g \quad i = 1,2,\ldots,n_k$$

em que:
$\mathbf{X}_{ki}$: valor observado da i-ésima variável sob o efeito do k-ésimo tratamento;
$\boldsymbol{\mu}$: média geral;
τ_k: efeito do k-ésimo tratamento;
$\mathbf{e}_{ki}$: efeito aleatório associado à observação $\mathbf{X}_{ki}$
Suposições:

a) Os erros e_{ki} são variáveis independentes com distribuição $N_p(\mathbf{0},\boldsymbol{\Sigma})$;
b) todas as populações têm matrizes de covariâncias comum $\boldsymbol{\Sigma}$;
c) cada população é normal p-variada;

Para definir de modo único os parâmetros do modelo é usual impor a restrição: $\sum_{k=1}^{g} n_k \boldsymbol{\tau}_k = \mathbf{0}$

2.2.1. Decomposição na soma de quadrados

Cada componente do vetor de observações $\mathbf{X}_{ki}$ satisfaz o modelo Anova (caso univariado) no qual a variável resposta pode ser expressa como:

$$\underset{\text{Variável resposta}}{X_{ki}} = \underset{\text{Média geral}}{\mu} + \underset{\text{efeito do tratamento}}{\tau_k} + \underset{\text{erro aleatório}}{e_{ki}} \qquad (1)$$

No modelo multivariado, os erros para os componentes de $\mathbf{X}_{ki}$ são correlacionados, mas a matriz de covariância $\boldsymbol{\Sigma}$ é a mesma para todas as populações.
No campo multivariado o vetor de observações $\mathbf{X}_{ki}$ pode ser decomposto de maneira similar a (1):

$$\underset{\text{Observação}}{\mathbf{x}_{ki}} = \underset{\substack{\text{estimativa} \\ \text{da média geral}}}{\bar{\mathbf{x}}} + \underset{\substack{\text{estimativa do} \\ \text{efeito do tratamento}}}{(\bar{\mathbf{x}}_k - \bar{\mathbf{x}})} + \underset{\text{resíduo}}{(\mathbf{x}_{ki} - \bar{\mathbf{x}}_k)}$$

Para fazer a decomposição da soma de quadrados total devemos primeiro notar que o produto $(\mathbf{x}_{ki} - \bar{\mathbf{x}})(\mathbf{x}_{ki} - \bar{\mathbf{x}})'$, pode ser escrito como:

$$(\mathbf{x}_{ki} - \bar{\mathbf{x}})(\mathbf{x}_{ki} - \bar{\mathbf{x}})' = [(\mathbf{x}_{ki} - \bar{\mathbf{x}}_k) + (\bar{\mathbf{x}}_k - \bar{\mathbf{x}})][(\mathbf{x}_{ki} - \bar{\mathbf{x}}_k) + (\bar{\mathbf{x}}_k - \bar{\mathbf{x}})]'$$
$$= (\mathbf{x}_{ki} - \bar{\mathbf{x}}_k)(\mathbf{x}_{ki} - \bar{\mathbf{x}}_k)' + (\mathbf{x}_{ki} - \bar{\mathbf{x}}_k)(\bar{\mathbf{x}}_k - \bar{\mathbf{x}})' + (\bar{\mathbf{x}}_k - \bar{\mathbf{x}})(\mathbf{x}_{ki} - \bar{\mathbf{x}}_k)' + (\bar{\mathbf{x}}_k - \bar{\mathbf{x}})(\bar{\mathbf{x}}_k - \bar{\mathbf{x}})'$$

Tomando a soma em *j*, verifica-se que as duas expressões centrais ficam nulas porque $\sum_{j=1}^{n_k}(\mathbf{x}_{ki} - \bar{\mathbf{x}}_k) = \mathbf{0}$.

Então, somando o produto cruzado sobre *k* e *i* tem-se:

$$\sum_{k=1}^{g}\sum_{i=1}^{n_k}(\mathbf{x}_{ki} - \bar{\mathbf{x}})(\mathbf{x}_{ki} - \bar{\mathbf{x}})' = \sum_{k=1}^{g}\sum_{i=1}^{n_k}(\mathbf{x}_{ki} - \bar{\mathbf{x}}_k)(\mathbf{x}_{ki} - \bar{\mathbf{x}}_k)' + (\bar{\mathbf{x}}_k - \bar{\mathbf{x}})(\bar{\mathbf{x}}_k - \bar{\mathbf{x}})' =$$

$$= \sum_{k=1}^{g}\sum_{i=1}^{n_k}(\mathbf{x}_{ki} - \bar{\mathbf{x}}_k)(\mathbf{x}_{ki} - \bar{\mathbf{x}}_k)' + \sum_{k=1}^{g}\sum_{i=1}^{n_k}(\bar{\mathbf{x}}_k - \bar{\mathbf{x}})(\bar{\mathbf{x}}_k - \bar{\mathbf{x}})' =$$

$$= \sum_{k=1}^{g} n_k(\bar{\mathbf{x}}_k - \bar{\mathbf{x}})(\bar{\mathbf{x}}_k - \bar{\mathbf{x}})' + \sum_{k=1}^{g}\sum_{i=1}^{n_k}(\mathbf{x}_{ki} - \bar{\mathbf{x}}_k)(\mathbf{x}_{ki} - \bar{\mathbf{x}}_k)'$$

Logo tem-se a seguinte decomposição:

$$\underset{\text{Soma de quadrados total}}{\sum_{k=1}^{g}\sum_{i=1}^{n_k}(\mathbf{x}_{ki} - \bar{\mathbf{x}})(\mathbf{x}_{ki} - \bar{\mathbf{x}})'} = \underset{\substack{\text{Soma de quadrados}\\ \text{entre tratamentos}\\ \text{Between(}\mathbf{B}\text{)}}}{\sum_{k=1}^{g} n_k(\bar{\mathbf{x}}_k - \bar{\mathbf{x}})(\bar{\mathbf{x}}_k - \bar{\mathbf{x}})'} + \underset{\substack{\text{Soma de quadrados}\\ \text{do resíduo}\\ \text{Within(}\mathbf{W}\text{)}}}{\sum_{k=1}^{g}\sum_{i=1}^{n_k}(\mathbf{x}_{ki} - \bar{\mathbf{x}}_k)(\mathbf{x}_{ki} - \bar{\mathbf{x}}_k)'} \quad (2)$$

A soma de quadrados Within pode ser expressa como:

$$\mathbf{W} = \sum_{k=1}^{g}\sum_{i=1}^{n_k}(\mathbf{x}_{ki} - \bar{\mathbf{x}}_k)(\mathbf{x}_{ki} - \bar{\mathbf{x}}_k)' = (n_1 - 1)\mathbf{S}_1 + (n_2 - 1)\mathbf{S}_2 + \cdots + (n_g - 1)\mathbf{S}_g \quad (3)$$

onde $\mathbf{S}_k$ é a matriz de covariância amostral para a *k*-ésima amostra.

A soma de quadrados total (SQTot) é uma matriz p x p e os elementos de sua diagonal principal correspondem às somas de quadrados totais para cada medida isolada (SQTot da Anova para cada variável isoladamente). Fora da diagonal temos as somas dos produtos cruzados dos desvios, caracterizando a estrutura de dependência entre as p medidas estudadas. A ideia é similar para as outras somas de quadrados.

A hipótese a ser testada para inexistência de efeito de tratamento é:

$H_0 = \boldsymbol{\tau}_1 = \boldsymbol{\tau}_2 = \cdots = \boldsymbol{\tau}_g = \mathbf{0}$ *versus*
$H_1 =$ pelo menos um tratamento produz efeito

As somas de quadrados podem ser organizadas na Tabela Manova a um fator com *g* níveis:

Fonte de variação	g.l	Matriz Soma de quadrados
Tratamento	g - 1	SQTrat = $\mathbf{B} = \sum_{k=1}^{g} n_k(\bar{\mathbf{x}}_k - \bar{\mathbf{x}})(\bar{\mathbf{x}}_k - \bar{\mathbf{x}})'$
Erro	n - g	SQRes = $\mathbf{W} = \sum_{k=1}^{g}\sum_{i=1}^{n_k}(\mathbf{x}_{ki} - \bar{\mathbf{x}}_k)(\mathbf{x}_{ki} - \bar{\mathbf{x}}_k)'$
Total	n - 1	SQTot = $\mathbf{T} = \mathbf{B} + \mathbf{W} = \sum_{k=1}^{g}\sum_{i=1}^{n_k}(\mathbf{x}_{ki} - \bar{\mathbf{x}})(\mathbf{x}_{ki} - \bar{\mathbf{x}})'$

$n = \sum_{k=1}^{g} n_k$

2.2.2. Testes de hipóteses multivariados

Para testar a hipótese de inexistência de efeito de tratamento, foram desenvolvidas diversas estatísticas. Os testes mais conhecidos são: Lambda de Wilk, Traço de Pillai, Traço de Lawley-Hotelling, e Raiz máxima de Roy. Esses testes envolvem no seu cálculo variâncias generalizadas ou autovalores $\hat{\lambda}_1, \hat{\lambda}_2, \cdots, \hat{\lambda}_s$ calculados pela resolução de $\mathbf{BW}^{-1}$ (*s* é o rank de **B**). Todos esses testes utilizam a distribuição F como aproximação.

a) Lambda de Wilks

Lambda de Wilks avalia a proporção de variância nas variáveis dependentes que não é contabilizado pelas variações intergrupos (dentro de cada tratamento). Se a proporção for pequena, isso indica que as variáveis dependentes variam muito entre os tratamentos e então esses tratamentos podem ter valores médios diferentes para as variáveis dependentes. Esse é o único teste relacionado com o critério da razão de verossimilhança.
A estatística de teste, originalmente proposta por Wilks, é dada por:

$$\Lambda = \frac{|\boldsymbol{W}|}{|\boldsymbol{B}+\boldsymbol{W}|} = \prod_{i=1}^{s} \left(\frac{1}{1+\lambda_i}\right),$$

b) Traço de Pillai

É considerado o teste mais robusto e de maior poder dentre os quatro testes multivariados. A estatística de teste é:

$$V = tr[\mathbf{B}(\mathbf{B} + \mathbf{W})^{-1}] = \sum \frac{\lambda_i}{1+\lambda_i}$$

c) Traço de Lawley-Hotelling

Essa estatística é a soma dos autovalores (traço) da matriz $\mathbf{BW}^{-1}$, onde grandes valores fornecem evidência contra a hipótese nula de igualdade.

$$U = tr[\mathbf{BW}^{-1}] = \sum_{i=1}^{s} \lambda_i$$

d) Raiz máxima de Roy

A estatística θ é calculada como o maior autovalor da matriz $\mathbf{BW}^{-1}$.
A base para este teste é que se a combinação linear das variáveis de X_1 à X_p que maximiza a razão entre a soma dos quadrados entre amostras e a soma dos quadrados dentro das amostras é encontrada, então essa razão máxima é igual a ao autovalor λ_1. Portanto, o autovalor máximo λ_1 pode ser uma boa estatística para testar se a variação entre amostras é significantemente grande, e que há, portanto, evidência de que as amostras sendo consideradas não vêm de populações com o mesmo vetor médio. O valor λ_1 é comparado com um valor aproximado da tabela F.
Rejeitamos a igualdade para valores grandes de λ_1.

Exemplo 1:
Num experimento envolvendo 4 variedades de feijão, avaliou-se na seca, a produtividade (P) em kg/ha e o número de grãos por vagem (GV), utilizando 5 repetições. Vamos testar a hipótese de que não há diferença entre as variedades de feijão. Os resultados obtidos foram:

Repetição	Variedade do feijão							
	A		B		C		D	
	P	GV	P	GV	P	GV	P	GV
R_1	1082	4.66	1163	5.52	1544	5.18	1644	5.45
R_2	1070	4.50	1100	5.30	1500	5.10	1600	5.18
R_3	1180	4.30	1200	5.42	1550	5.20	1680	5.18
R_4	1050	4.70	1190	5.62	1600	5.30	1700	5.40
R_5	1080	4.60	1170	5.70	1540	5.12	1704	5.50

H_0: $\boldsymbol{\tau}_A = \boldsymbol{\tau}_B = \boldsymbol{\tau}_C = \boldsymbol{\tau}_D = \mathbf{0}$

Vetores de Médias amostrais: $\bar{\mathbf{X}} = \begin{bmatrix}1367.35\\5.15\end{bmatrix}$, $\bar{\mathbf{X}}_A = \begin{bmatrix}1092.40\\4.55\end{bmatrix}$, $\bar{\mathbf{X}}_B = \begin{bmatrix}1164.60\\5.51\end{bmatrix}$, $\bar{\mathbf{X}}_C = \begin{bmatrix}1546.80\\5.18\end{bmatrix}$, $\bar{\mathbf{X}}_D = \begin{bmatrix}1665.60\\5.34\end{bmatrix}$

Matrizes de covariâncias amostrais:

$$\mathbf{B} = 5\left\{\begin{bmatrix}1092.40\\4.55\end{bmatrix} - \begin{bmatrix}1367.35\\5.15\end{bmatrix}\right\}\{[1092.40 \quad 4.55] - [1367.35 \quad 5.15]\} + 5\left\{\begin{bmatrix}1164.60\\5.51\end{bmatrix} - \begin{bmatrix}1367.35\\5.15\end{bmatrix}\right\}\{[1164.60 \quad 5.51] - [1367.35 \quad 5.15]\} + 5\left\{\begin{bmatrix}1546.80\\5.18\end{bmatrix} - \begin{bmatrix}1367.35\\5.15\end{bmatrix}\right\}\{[1546.80 \quad 5.18] - [1367.35 \quad 5.15]\} + 5\left\{\begin{bmatrix}1665.60\\5.34\end{bmatrix} - \begin{bmatrix}1367.35\\5.15\end{bmatrix}\right\}\{[1665.60 \quad 5.34] - [1367.35 \quad 5.15]\}$$

$$\mathbf{B} = 5\begin{bmatrix}75597.50 & 165.00\\165.00 & 0.36\end{bmatrix} + 5\begin{bmatrix}41107.60 & -73.00\\-73.00 & 0.13\end{bmatrix} + 5\begin{bmatrix}32202.30 & 5.38\\5.38 & 0.00009\end{bmatrix} + 5\begin{bmatrix}88953.10 & 56.67\\56.67 & 0.036\end{bmatrix} = \begin{bmatrix}1189302.00 & 768.36\\768.00 & 2.63\end{bmatrix}$$

$$\mathbf{T}=\left\{\begin{bmatrix}1070\\4.50\end{bmatrix}-\begin{bmatrix}1367.35\\5.15\end{bmatrix}\right\}\{[1070 \quad 4.50]-[1367.35 \quad 5.15]\}+\cdots+\left\{\begin{bmatrix}1704\\5.50\end{bmatrix}-\begin{bmatrix}1367.35\\5.15\end{bmatrix}\right\}\{[1704 \quad 5.50]-[1367.35 \quad 5.15]\}=$$
$$=\begin{bmatrix}1218360.55 & 778.26\\778.26 & 2.95\end{bmatrix}$$

$$\mathbf{W}=\mathbf{T}-\mathbf{B}=\begin{bmatrix}1218360.55 & 778.26\\778.26 & 2.95\end{bmatrix}-\begin{bmatrix}1189302.00 & 768.36\\768.00 & 2.63\end{bmatrix}=\begin{bmatrix}29058.55 & 10.26\\10.26 & 0.32\end{bmatrix}$$

Tabela Manova

Fonte de variação	g.l	Matriz Soma de quadrados
Tratamento	*3*	SQTrat = $\mathbf{B}=\begin{bmatrix}1189302.00 & 768.36\\768.36 & 2.63\end{bmatrix}$
Erro	*16*	SQRes = $\mathbf{W}=\begin{bmatrix}29058.55 & 10.26\\10.26 & 0.32\end{bmatrix}$
Total	*19*	SQTot = $\mathbf{T}=\mathbf{B}+\mathbf{W}=\begin{bmatrix}1218360.55 & 778.26\\778.26 & 2.95\end{bmatrix}$

Testes

Wilks $\Lambda=\frac{|\boldsymbol{W}|}{|\boldsymbol{B}+\boldsymbol{W}|}=\frac{\begin{vmatrix}29058.55 & 10.26\\10.26 & 0.32\end{vmatrix}}{\left|\begin{bmatrix}1189302.00 & 768,36\\768.36 & 2.63\end{bmatrix}+\begin{bmatrix}29058.55 & 10.26\\10.26 & 0.32\end{bmatrix}\right|}=\frac{9193.47}{2988193.20}=0.003$

Traço de Pillai $V=tr[\mathbf{B}(\mathbf{B}+\mathbf{W})^{-1}]=tr\left\{\begin{bmatrix}1189302.00 & 768.36\\768.36 & 2.63\end{bmatrix}\begin{bmatrix}1218360.55 & 778.26\\778.26 & 2.95\end{bmatrix}^{-1}\right\}=tr\begin{bmatrix}0.971527 & 7.33045\\0.000082 & 0.87027\end{bmatrix}=1.84$

Traço de Lawley-Hotelling $U=\text{tr}[\mathbf{BW}^{-1}]=tr\left\{\begin{bmatrix}1189302.00 & 768.36\\768.36 & 2.63\end{bmatrix}\begin{bmatrix}29058.55 & 10.26\\10.26 & 0.32\end{bmatrix}^{-1}\right\}=tr\begin{bmatrix}40.5839 & 1101.35\\0.0238 & 7.46\end{bmatrix}=47.99$

Raiz máxima de Roy θ = maior autovalor de $\mathbf{BW}^{-1}$

$\mathbf{BW}^{-1}=\begin{bmatrix}40.5839 & 1101.35\\0.0238 & 7.46\end{bmatrix}=\mathbf{A}$

Para calcular os autovalores:

$$|\boldsymbol{A}-1\boldsymbol{I}|=0\Rightarrow\begin{vmatrix}40.5839-\lambda & 1101.35\\0.0238 & 7.46-\lambda\end{vmatrix}=302.76-40.5839\lambda-7.46\lambda+\lambda^2-26.21=\lambda^2-48.0439\lambda+276.5439=0$$

Cujas raízes são $\lambda_1 = 41.37$ e $\lambda_2 = 6.68 \Rightarrow \theta = 41.37$.

Para avaliar a significância, as estatísticas multivariadas têm equivalência aproximada com a distribuição F. O quadro a seguir apresenta as aproximações e os graus de liberdade para a avaliação da significância dos testes.

Critério	Estatística	Aproximação F	g.l. de F
Wilks	$\Lambda=\frac{\lvert\boldsymbol{W}\rvert}{\lvert\boldsymbol{B}+\boldsymbol{W}\rvert}=\prod_{i=1}^{s}\left(\frac{1}{1+\lambda_i}\right)$	$F=\left(\frac{1-\Lambda^{1/t}}{\Lambda^{1/t}}\right)\left(\frac{rt-2f}{pq}\right)$	$v_1 = pq$ $v_2 = rt\text{-}2f$
Traço de Pillai	$V=tr[\mathbf{B}(\mathbf{B}+\mathbf{W})^{-1}]=\sum\frac{\lambda_i}{1+\lambda_i}$	$F=\left(\frac{V}{s-V}\right)\left(\frac{2n+s+1}{2m+s+1}\right)$	$v_1 = s(2m+s+1)$ $v_2 = s(2n+s+1)$
Traço de Hotelling-Lawley	$U=\text{tr}[\mathbf{BW}^{-1}]=\sum_{i=1}^{s}\lambda_i$	$F=\frac{2(sn+1)U}{s^2(2m+s+1)}$	$v_1 = s(2m+s+1)$ $v_2 = 2(sn+1)$
Raiz máxima de Roy	θ = maior autovalor de $\mathbf{BW}^{-1}$	$F=\frac{\theta(\nu-d+q)}{d}$	$v_1 = d$ $v_2 = \nu\text{-}d+q$

p = número de variáveis; q = gl do tratamento; ν = gl do erro; s = min(p,q); $r = \nu\text{-}(p\text{-}q+1)/2$; $f = (pq\text{-}2)/4$; d = max(p,q); $m = (|p\text{-}q|\text{-}1)/2$; $n = (\nu\text{-}p\text{-}1)/2$; e $t=\begin{cases}\sqrt{\frac{p^2q^2-4}{p^2+q^2-5}} & \text{Se } p^2+q^2-5>0\\ 1 & c.c.\end{cases}$

A hipótese nula é rejeitada para todos os testes. O que indica que pelo menos uma das variedades de feijão produz resultado diferente das demais em termos de produtividade e número de grãos por vagem.

Resolução no R

Organização dos dados: No R os dados para a realização da MANOVA precisam estar organizados na forma longa. Ou seja, a variável que caracteriza o grupo, tratamento ou fator deve vir em coluna única:

P	GV	Trat
1082	4.66	1
1070	4.50	1
1180	4.30	1
1050	4.70	1
1080	4.60	1
1163	5.52	2
1100	5.30	2
1200	5.42	2
1190	5.62	2
1170	5.70	2
1544	5.18	3
1500	5.10	3
1550	5.20	3
1600	5.30	3
1540	5.12	3
1644	5.45	4
1600	5.18	4
1680	5.18	4
1700	5.40	4
1704	5.5	4

2.2.3. Comparações múltiplas

Quando H_0 é rejeitada, é importante saber quais os efeitos que levaram à essa rejeição. Os intervalos de confiança simultâneos desenvolvidos por Bonferroni para comparações pareadas, podem ser usados para construir intervalos simultâneos de confiança para os componentes dos vetores diferença $\boldsymbol{\tau}_k - \boldsymbol{\tau}_\ell$

Seja τ_{kj} o j-ésimo componente do vetor $\boldsymbol{\tau}_k$.

Como $\boldsymbol{\tau}_k$ é estimado por $\hat{\boldsymbol{\tau}}_k = \bar{\mathbf{x}}_k - \bar{\mathbf{x}}$ temos que τ_{kj} é estimado por $\hat{\tau}_{kj} = \bar{x}_{kj} - \bar{x}_j$.

Assim, $\tau_{kj} - \tau_{\ell j}$ é estimado por $\hat{\tau}_{kj} - \hat{\tau}_{\ell j} = \bar{x}_{kj} - \bar{x}_{\ell j}$, que é uma diferença entre médias de duas amostras independentes.

O teste t-Student para duas amostras independentes é válido com uma correção proposta por Bonferroni.

Note que: $Var(\hat{\tau}_{kj} - \hat{\tau}_{\ell j}) = Var(\bar{x}_{kj} - \bar{x}_{\ell j}) = \left(\frac{1}{n_k} + \frac{1}{n_\ell}\right)\sigma_{jj}$, onde σ_{jj} é o j-ésimo elemento da diagonal de $\boldsymbol{\Sigma}$. Mas, pela soma de quadrados dos erros (**W**) expressa na equação (3), cada elemento da diagonal principal da matriz **W** representa uma soma de quadrados dos erros para cada variável isolada e como sabemos que a estimativa para a variância é a soma de quadrados dos erros dividido pelos seus graus de liberdade, tem-se que:

$$\hat{\sigma}_{jj} = \frac{1}{n-g} w_{jj}.$$

Logo, a estimativa para a variância da diferença entre as médias das duas amostras é:

$$\widehat{Var}(\bar{x}_{kj} - \bar{x}_{\ell j}) = \left(\frac{1}{n_k} + \frac{1}{n_\ell}\right)\frac{w_{jj}}{n-g},$$

onde w_{jj} é o j-ésimo elemento da diagonal de **W** e $n = n_1 + n_2 + ... + n_g$.
Para determinar o número de intervalos a serem calculados, basta lembrar que são g tratamentos para p variáveis, logo, teremos $m = p\binom{g}{2} = pg(g-1)/2$ intervalos.

Então, para o modelo (1), com pelo menos 1 – α de confiança, os intervalos simultâneos para $\tau_{kj} - \tau_{\ell j}$ são dados por:

$$(\bar{x}_{kj} - \bar{x}_{\ell j}) \pm t_{n-g}\left(\frac{\alpha}{pg(g-1)}\right)\sqrt{\left(\frac{1}{n_k}+\frac{1}{n_\ell}\right)\frac{w_{jj}}{n-g}}$$

Exemplo 2. Voltando ao Exemplo 1, vimos que a produtividade e o número de grãos por vagem dependem da variedade do feijão. Vamos avaliar a magnitude das diferenças entre as variedades calculando os intervalos de confiança simultâneos. A comparação da produtividade (P) entre as variedades A e B pode ser estimada por meio da estimação de $\tau_{A1} - \tau_{B1}$ (Produtividade aqui representada por 1).

Os vetores de tratamento são estimados por:

$$\tau_A = (\bar{\mathbf{x}}_A - \bar{\mathbf{x}}) = \begin{bmatrix}1092.40\\4.55\end{bmatrix} - \begin{bmatrix}1367.35\\5.15\end{bmatrix} = \begin{bmatrix}-274.95\\-0.60\end{bmatrix}, \tau_B = (\bar{\mathbf{x}}_B - \bar{\mathbf{x}}) = \begin{bmatrix}-202.75\\0.36\end{bmatrix}, \tau_C = (\bar{\mathbf{x}}_C - \bar{\mathbf{x}}) = \begin{bmatrix}179.45\\0.03\end{bmatrix}, \tau_D = (\bar{\mathbf{x}}_D - \bar{\mathbf{x}}) = \begin{bmatrix}298.25\\0.19\end{bmatrix}$$

e $\mathbf{W} = \begin{bmatrix}29058.55 & 10.26\\10.26 & 0.32\end{bmatrix}$, com 20 – 4 = 16 graus de liberdade.
Então,

$$\tau_{A1} - \tau_{B1} = -274.95 + 202.75 = -72.2$$

Como $n_A = n_B = n_C = n_D = 5$, $n = 20$
Logo,

$$\widehat{Var}(\bar{x}_{A1} - \bar{x}_{B1}) = \left(\frac{1}{n_A}+\frac{1}{n_B}\right)\frac{w_{11}}{n-g} = \left(\frac{1}{5}+\frac{1}{5}\right)\frac{29058.55}{16} = 726.46$$

Como $p = 2$ e $g = 4$, para 95% de confiança é necessário $t_{16}(0{,}05/2.4.3) = 3.25$

O intervalo de confiança é:

$$\tau_{A1} - \tau_{B1} \pm t_{16}(0.00208)\sqrt{\left(\frac{1}{5}+\frac{1}{5}\right)\frac{29058.55}{16}} = -72.2 \pm 3.25\sqrt{726.46} = -72.2 \pm 3.25 \times 26.95 = (-159.78; 13.11)$$

Como o intervalo inclui o zero, concluímos que não existe diferença na produtividade entre as variedades A e B.
Fica como exercício calcular os outros intervalos. Quantos intervalos devem ser calculados? Conclua a análise.

2.2.4. Verificação das suposições

As suposições da Manova estão relacionadas à normalidade dos erros e igualdade das matrizes de covariância nas populações. A normalidade deve ser testada de forma multivariada, caso o tamanho da amostra não seja suficientemente grande, o pacote MVN do R realiza os testes de Mardia, Royston e Henze-Zirkler para o caso multivariado. Para avaliar a homocedasticidade em dados multivariados é usual usar o teste de Box:
Com *g* populações tem-se a hipótese nula:

$$H_0 = \mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 = \cdots = \mathbf{\Sigma}_g = \mathbf{\Sigma}$$

onde $\mathbf{\Sigma}_k$ é a matriz de covariância para a k-ésima população, $k = 1, ...,g$. A hipótese alternativa é que pelo menos duas das matrizes de covariância não são iguais.
Supondo populações normais multivariadas, uma estatística de razão de verossimilhança é dada por:

$$\Lambda = \sum_k \left(\frac{|\mathbf{S}_k|}{|\mathbf{S}_{agrupado}|}\right)^{(n_k-1)/2}$$

onde n_k é o tamanho da amostra para o *k*-ésimo grupo (tratamento), $\mathbf{S}_k$ é a matriz de covariância do *k*-ésimo grupo e $\mathbf{S}_{agrupado}$ é a matriz de covariância da amostra agrupada, dada por:

$$\mathbf{S}_{agrupado} = \frac{1}{\sum_{k=1}^{g}(n_k-1)}\{(n_1-1)\mathbf{S}_1 + (n_2-1)\mathbf{S}_2 + \cdots + (n_g-1)\mathbf{S}_g\} = \frac{1}{n-g}\sum_{k=1}^{g}(n_k-1)\,\mathbf{S}_k$$

O teste de Box é baseado na aproximação de qui-quadrado da distribuição amostral de –2 lnΛ. Fazendo M = –2 lnΛ tem-se:

$$M = (n-g)\ln\left|\mathbf{S}_{agrupado}\right| - \sum_k [(n_k - 1)\ln|\mathbf{S}_k|]$$

Se H_0 é verdadeira, as matrizes de covariâncias devem ser bem parecidas e, consequentemente, também serão similares aos elementos em $\mathbf{S}_{agrupado}$. Nesse caso, Λ estará próximo de 1 e M, próximo de 0.
Como M não tem uma distribuição conhecida, Box mostrou que:

$$C = (1-u)M = (1-u)\left\{(n-g)\ln\left|\mathbf{S}_{agrupado}\right| - \sum_k [(n_k - 1)\ln|\mathbf{S}_k|]\right\}$$

tem distribuição aproximadamente χ^2com ν graus de liberdade, e $\nu = g\frac{1}{2}p(p+1) - \frac{1}{2}p(p+1) = \frac{1}{2}p(p+1)(g-1)$ e *u* é uma constante de correção do viés dada por:

$$u = \left(\sum_{k=1}^{g}\frac{1}{n_k - 1} - \frac{1}{n-g}\right)\frac{2p^2+3p-1}{6(p+1)(g-1)}$$

Então, a um nível de significância α, rejeitamos H_0 se

$$C > \chi^2_{\nu,1-\alpha}.$$

A aproximação de χ^2 de Box funcionará melhor se cada n_k > 20 e p e g não são superiores a 5. Para situações diferentes a distribuição *F* fornece uma melhor aproximação.

O teste M de Box é sensível a algumas formas de não-normalidade. Porém, com amostras de tamanho razoavelmente grandes, os testes para efeitos de tratamento são robustos à não-normalidade, isto é, a análise da MANOVA ainda é considerada válida, ainda que o teste M rejeite H_0.

Exemplo 3. Para o Exemplo 1, vamos avaliar a suposição de normalidade e homocedasticidade multivariadas.

$\mathbf{S_A} = \begin{bmatrix}2558.80 & -7.226\\ -7.226 & 0.0255\end{bmatrix}$, $\mathbf{S_B} = \begin{bmatrix}1525.80 & 3.546\\ 3.546 & 0.0251\end{bmatrix}$, $\mathbf{S_C} = \begin{bmatrix}1271.20 & 2.65\\ 2.65 & 0.0062\end{bmatrix}$ $\mathbf{S_D} = \begin{bmatrix}1908.80 & 3.506\\ 3.506 & 0.0231\end{bmatrix}$

$$\mathbf{S}_{agrupado} = \frac{1}{n-g}\sum_{k=1}^{g}(n_k - 1)\,\mathbf{S}_k =$$
$$= \frac{1}{20-4}\left\{4\begin{bmatrix}2558.80 & -7.226\\ -7.226 & 0.0255\end{bmatrix} + \mathbf{4}\begin{bmatrix}1525.80 & 3.546\\ 3.546 & 0.0251\end{bmatrix} + 4\begin{bmatrix}1271.20 & 2.65\\ 2.65 & 0.0062\end{bmatrix} + 4\begin{bmatrix}1908.80 & 3.506\\ 3.506 & 0.0231\end{bmatrix}\right\}$$

$$\mathbf{S}_{agrupado} = \frac{1}{4}\begin{bmatrix}7264.60 & 2.476\\ 2.476 & 0.0799\end{bmatrix} = \begin{bmatrix}1816.15 & 0.619\\ 0.619 & 0.02\end{bmatrix}$$

$$M = (n-g)\ln\left|\mathbf{S}_{agrupado}\right| - \sum_k [(n_k - 1)\ln|\mathbf{S}_k|]$$
$$= 16\ln\begin{vmatrix}1816.15 & 0.619\\ 0.619 & 0.02\end{vmatrix}$$
$$- 4\left\{\ln\begin{vmatrix}2558.80 & -7.226\\ -7.226 & 0.0255\end{vmatrix} + \ln\begin{vmatrix}1525.80 & 3.546\\ 3.546 & 0.0251\end{vmatrix} + \ln\begin{vmatrix}1271.20 & 2.65\\ 2.65 & 0.0062\end{vmatrix} + \ln\begin{vmatrix}1908.80 & 3.506\\ 3.506 & 0.0231\end{vmatrix}\right\}$$
$$= 16\ln(35.94) - 4\{\ln(13.034) + \ln(25.723) + \ln(0.86) + \ln(31.801)\}$$
$$= 16 \times (3.582) - 4(2.57 + 3.247 - 0.147 + 3.459) = 57.312 - 36.516 = 20.796$$

$$u = \left(\sum_{k=1}^{g}\frac{1}{n_k - 1} - \frac{1}{n-g}\right)\frac{2p^2+3p-1}{6(p+1)(g-1)} = \left(\frac{4}{4} - \frac{1}{16}\right)\frac{13}{54} = \frac{240}{864} = 0.28$$

$$C = (1-u)M = 0.72 \times (20.796) = 14.97$$

$\nu = \frac{1}{2}p(p+1)(g-1) = \frac{18}{2} = 9$

$$\chi^2_{9,1-0.95} = 16.92$$

Como *C* é menor do que $\chi^2_{9;5\%}$ não rejeitamos a hipótese de homocedasticidade multivariada ao nível de 5% de significância.

2.3. MANOVA a dois fatores (Two-way)

Por analogia ao caso univariado (Anova a dois fatores), o modelo MANOVA a dois fatores de efeitos fixos para um vetor resposta com p componentes é

$$\mathbf{X}_{krt} = \boldsymbol{\mu} + \boldsymbol{\tau}_k + \boldsymbol{\beta}_r + \boldsymbol{\gamma}_{kr} + \mathbf{e}_{krt} \qquad k = 1,2,\ldots,g \quad r = 1,2,\ldots,b \quad t = 1,2,\ldots,n.$$

em que:

$\mathbf{X}_{krt}$ = valor observado da *t*-ésima variável ao nível *k* do fator 1 e nível *r* do fator 2;
$\boldsymbol{\mu}$ = média geral;
$\boldsymbol{\tau}_k$ = efeito fixo do fator 1
$\boldsymbol{\beta}_r$ = efeito fixo do fator 2
$\boldsymbol{\gamma}_{kr}$ = interação entre o fator 1 e o fator 2
$\mathbf{e}_{ki}$ = efeito aleatório associado à observação $\mathbf{X}_{krt}$

Por suposição, $\mathbf{e}_{krt} \sim \boldsymbol{N}_p(\mathbf{0}, \boldsymbol{\Sigma})$ independentes $\forall$ *k,r,t*.

Assim, as variáveis respostas consistem em p medidas replicadas n vezes em cada combinação possível dos níveis dos fatores 1 e 2. A presença da interação, *kr*, em cada uma das *gb* combinações de níveis, implica que os efeitos dos fatores não são aditivos, o que complica a interpretação dos resultados.
São adotadas as seguintes restrições:

$$\sum_{k=1}^{g} \boldsymbol{\tau}_k = \sum_{r=1}^{b} \boldsymbol{\beta}_r = \sum_{k=1}^{g} \boldsymbol{\gamma}_{kr} = \sum_{r=1}^{b} \boldsymbol{\gamma}_{kr} = \mathbf{0}$$

Essas restrições são usadas na decomposição do vetor observado $\mathbf{x}_{krt}$ como:

$$\mathbf{x}_{krt} \;=\; \bar{\mathbf{x}} \;+\; (\bar{\mathbf{x}}_{k\cdot} - \bar{\mathbf{x}}) + (\bar{\mathbf{x}}_{\cdot r} - \bar{\mathbf{x}}) \;+ (\mathbf{x}_{kr} - \bar{\mathbf{x}}_{k\cdot} - \bar{\mathbf{x}}_{\cdot r} + \bar{\mathbf{x}}) + (\mathbf{x}_{krt} - \bar{\mathbf{x}}_{kr})$$

em que cada parcela é uma estimativa para os parâmetros do modelo, $\bar{\mathbf{x}}$ é a média global dos vetores observados, $\bar{\mathbf{x}}_{k\cdot}$ é a média dos vetores observados no *k*-ésimo nível do fator 1, $\bar{\mathbf{x}}_{\cdot r}$ é a média dos vetores observados no *r*-ésimo nível do fator 2 e $\bar{\mathbf{x}}_{kr}$ é a média dos vetores observados no *k*-ésimo nível do fator 1 e no *r*-ésimo nível do fator 2.
Os resultados de somas de quadrados são obtidos com desenvolvimento semelhante ao da Manova a um fator e estão organizados na Tabela Manova a seguir:

Tabela MANOVA a dois fatores

Fonte de variação	g.l	Matriz Soma de quadrados
Fator 1	$g-1$	$\text{SQP}_{\text{Fat1}} = \sum_{k=1}^{g} bn(\bar{\mathbf{x}}_{k\cdot} - \bar{\mathbf{x}})(\bar{\mathbf{x}}_{k\cdot} - \bar{\mathbf{x}})'$
Fator 2	$b-1$	$\text{SQP}_{\text{Fat2}} = \sum_{r=1}^{b} gn(\bar{\mathbf{x}}_{\cdot r} - \bar{\mathbf{x}})(\bar{\mathbf{x}}_{\cdot r} - \bar{\mathbf{x}})'$
Interação	$(g-1)(b-1)$	$\text{SQP}_{\text{Int}} = \sum_{k=1}^{g}\sum_{r=1}^{b} n(\mathbf{x}_{kr} - \bar{\mathbf{x}}_{k\cdot} - \bar{\mathbf{x}}_{\cdot r} + \bar{\mathbf{x}})(\mathbf{x}_{kr} - \bar{\mathbf{x}}_{k\cdot} - \bar{\mathbf{x}}_{\cdot r} + \bar{\mathbf{x}})'$
Erro	$gb(n-1)$	$\text{SQP}_{\text{Res}} = \sum_{k=1}^{g}\sum_{r=1}^{b}\sum_{t=1}^{n}(\mathbf{x}_{krt} - \bar{\mathbf{x}}_{kr})(\mathbf{x}_{krt} - \bar{\mathbf{x}}_{kr})'$
Total	$gbn-1$	$\text{SQP}_{\text{Tot}} = \sum_{k=1}^{g}\sum_{r=1}^{b}\sum_{t=1}^{n}(\mathbf{x}_{krt} - \bar{\mathbf{x}})(\mathbf{x}_{krt} - \bar{\mathbf{x}})'$

Geralmente, o teste de $H_0 : \boldsymbol{\gamma}_{11} = \boldsymbol{\gamma}_{12} = \cdots = \boldsymbol{\gamma}_{gb} = \mathbf{0}$ (não há efeito de interação) versus H_1 : pelo menos um $\boldsymbol{\gamma}_{kr}$ é não-nulo, é realizado antes de testar os efeitos dos fatores principais. Se há efeito de interação, os efeitos dos fatores principais não têm uma interpretação clara. Neste caso, Johnson e Wichern (2007, p. 316) não recomendam prosseguir com os testes multivariados adicionais. Ao invés disso, deve-se conduzir *p* análises de variância univariadas a dois fatores (uma para cada variável resposta).
Para testar o efeito da interação, o teste (de razão de verossimilhanças) rejeita H_0 para valores pequenos da estatística lambda de Wilks:

$$\Lambda_{Int} = \frac{|\boldsymbol{SQP}_{Res}|}{|\boldsymbol{SQP}_{Int} + \boldsymbol{SQP}_{Res}|}$$

Para testar o efeito do Fator 1, testa-se a hipótese H_0: $\boldsymbol{\tau}_1 = \boldsymbol{\tau}_2 = \cdots = \boldsymbol{\tau}_g = \mathbf{0}$ versus H_1: pelo menos um $\boldsymbol{\tau}_k$ é não nulo. A estatística de teste é:

$$\Lambda_{F1} = \frac{|\boldsymbol{SQP}_{Res}|}{|\boldsymbol{SQP}_{Fat1} + \boldsymbol{SQP}_{Res}|}$$

Para o efeito do Fator 2, testa-se a hipótese: H_0: $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2 = \cdots = \boldsymbol{\beta}_g = \mathbf{0}$ versus H_1: pelo menos um $\boldsymbol{\beta}_r$ é não nulo. A estatística de teste é:

$$\Lambda_{F2} = \frac{|\boldsymbol{SQP}_{Res}|}{|\boldsymbol{SQP}_{Fat2} + \boldsymbol{SQP}_{Res}|}$$

Como na Manova a um fator, se o Lambda de Wilks mostrar que existe efeito do Fator 1 ou do Fator 2 ou de ambos os fatores, a abordagem de Bonferroni pode ser usada para construir intervalos simultâneos de confiança para as componentes dos vetores diferença $\boldsymbol{\tau}_k - \boldsymbol{\tau}_\ell$ dos efeitos do fator 1 e para as componentes dos vetores diferença $\boldsymbol{\beta}_r - \boldsymbol{\beta}_m$ dos efeitos do fator 2.

De maneira análoga ao que foi visto para Manova a um fator, os intervalos simultâneos de 100(1−α)% de confiança para $\boldsymbol{\tau}_{kj} - \boldsymbol{\tau}_{\ell j}$ são dados por:

$$\left(\bar{x}_{k.j} - \bar{x}_{\ell.j}\right) \pm t_\nu \left(\frac{\alpha}{pg(g-1)}\right) \sqrt{\frac{E_{jj}}{\nu}\frac{2}{bn}}$$

onde ν = $gb(n-1)$, E_{jj} é o j-ésimo elemento da diagonal de $\mathbf{E} = SQP_{Res}$.

Similarmente, os intervalos simultâneos de 100(1−α)% de confiança para $\boldsymbol{\beta}_{rj} - \boldsymbol{\beta}_{mj}$ são:

$$\left(\bar{x}_{.rj} - \bar{x}_{.mj}\right) \pm t_\nu \left(\frac{\alpha}{pb(b-1)}\right) \sqrt{\frac{E_{jj}}{\nu}\frac{2}{gn}}$$

onde ν = $gb(n-1)$, E_{jj} é o j-ésimo elemento da diagonal de $\mathbf{E} = SQP_{Res}$.

Exemplo 3. Os dados da tabela a seguir referem-se a 32 peças de barras de aço homogêneas, para as quais foram mensuradas as variáveis: x_1 = torque final e x_2 = tensão final. As variáveis foram mensuradas em duas velocidades de rotação (A_1(rápido) e A_2(lento)) e quatro tipos de lubrificantes (B_1, B_2, B_3 e B_4).

Lubrificante	A_1		A_2	
	x_1	x_2	x_1	x_2
B_1	7.80	90.4	7.12	85.1
	7.10	88.9	7.06	89.0
	7.89	85.9	7.45	75.9
	7.82	88.8	7.45	77.9
B_2	9.00	82.5	8.19	66.0
	8.43	92.4	8.25	74.5
	7.65	82.4	7.45	83.1
	7.70	87.4	7.45	86.4
B_3	7.28	79.6	7.15	81.2
	8.96	95.1	7.15	72.0
	7.75	90.2	7.70	79.9
	7.80	88.0	7.45	71.9
B_4	7.60	94.1	7.06	81.2
	7.00	86.6	7.04	79.9
	7.82	85.9	7.52	86.4
	7.80	88.8	7.70	76.4

Vamos verificar se existe efeito da velocidade de rotação ou do tipo de lubrificante sobre as variáveis mensuradas. A análise será realizada com o auxílio do R.