MAE5776 - 1º Sem/2022 – Comparação de 2 Populações Np

1 - Comparação de vetores de Médias de 2 populações N3. Teste T2 de Hotelling.

Gerar uma amostra aleatória de n_g observações de duas populações Normais tridimensionais, $N_3(\mu_g; \Sigma_g)$, envolvendo as variáveis Y1, Y2 e Y3. Preencha a tabela a seguir com os parâmetros adotados na simulação dos dados.

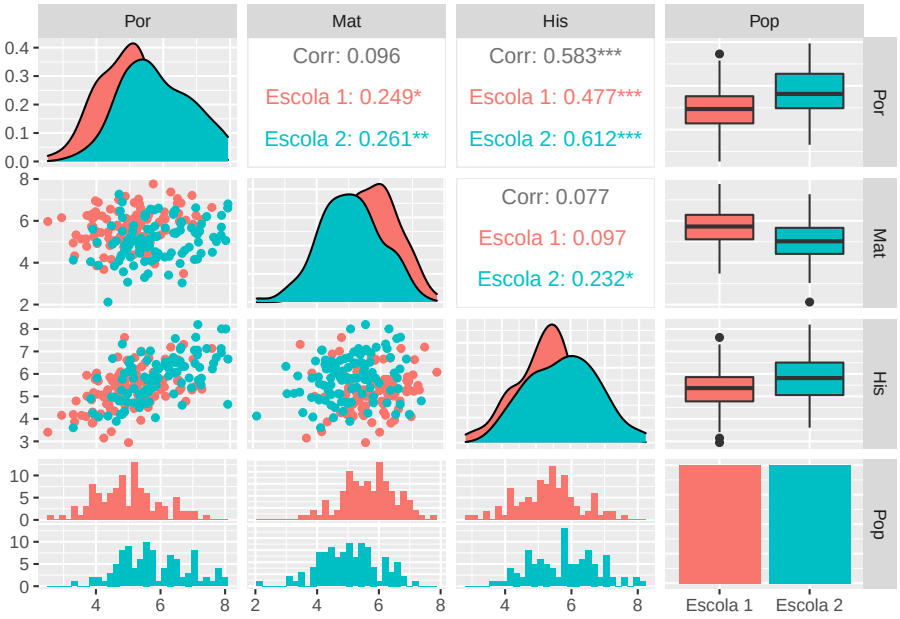
População	Amostra	Vetor de Médias	Matriz de Covariâncias	Matriz de Correlações
Pop 1	$n_1 = 100$	$\mu_1 = (5, 5.7, 5.5)$	$\Sigma_1 = \begin{pmatrix} 1 & 0.3 & 0.5 \\ 0.3 & 0.9 & 0.2 \\ 0.5 & 0.2 & 1.1 \end{pmatrix}$	$\rho_1 = \begin{pmatrix} 1 & 0.32 & 0.48 \\ 0.32 & 1 & 0.2 \\ 0.48 & 0.2 & 1 \end{pmatrix}$
Pop 2	$n_2 = 100$	$\mu_2 = (5.8, 5.1, 5.7)$	$\Sigma_2 = \begin{pmatrix} 1 & 0.3 & 0.5 \\ 0.3 & 0.9 & 0.2 \\ 0.5 & 0.2 & 1.1 \end{pmatrix}$	$\rho_2 = \begin{pmatrix} 1 & 0.32 & 0.48 \\ 0.32 & 1 & 0.2 \\ 0.48 & 0.2 & 1 \end{pmatrix}$

1.1 - Contextualize, com uma situação prática hipotética, os dados gerados. Caracterize a estrutura dos dados (amostras balanceadas, observações independentes, tipo de variável, dimensão dos dados, etc). Defina o objetivo do estudo.

R: Podemos contextualizar os dados com sendo de notas de português, matemática e história de turmas do 3º ano do ensino médio de duas escolas de São Paulo avaliadas pela teoria da resposta ao item. De um modo geral, os dados são compostos por 3 variáveis contínuas com um total de 200 observações e uma variável categórica definindo a população da amostra. São dados balanceados, contendo 100 obsevações em cada amostra e foram gerados de forma independente, onde a geração de cada observação não influenciou as demais. O objetivo desse estudo será comparar as notas médias das 3 disciplinas entre ambas as escolas.

1.2 - Realize uma análise descritiva dos dados (calcule estatísticas descritivas, construa gráficos apropriados). Comente os resultados de acordo com o objetivo do estudo.

População	Amostra	Vetor de Médias	Matriz de Covariâncias	Matriz de Correlações
Escola 1	$n_1 = 100$	$\bar{Y_1} = (4.99, 5.71, 5.33)$	$S_{u1} = \begin{pmatrix} 0.96 & 0.21 & 0.44 \\ 0.21 & 0.76 & 0.08 \\ 0.44 & 0.08 & 0.9 \end{pmatrix}$	$R_1 = \begin{pmatrix} 1 & 0.25 & 0.48 \\ 0.25 & 1 & 0.1 \\ 0.48 & 0.1 & 1 \end{pmatrix}$
Escola 2	$n_2 = 100$	$\bar{Y_2} = (5.81, 5.05, 5.8)$	$S_{u2} = \begin{pmatrix} 1.1 & 0.26 & 0.65 \\ 0.26 & 0.93 & 0.23 \\ 0.65 & 0.23 & 1.02 \end{pmatrix}$	$R_2 = \begin{pmatrix} 1 & 0.26 & 0.61 \\ 0.26 & 1 & 0.23 \\ 0.61 & 0.23 & 1 \end{pmatrix}$



R: Ao observar os box plots de cada disciplina da Escola 1 e Escola 2 percebe que as medianas não estão distantes; ao observar os histogramas percebe-se que as distribuições da notas em cada disciplina são parecidas; os gráficos de dispersão apontam que há uma marcante intercepção entre as notas da mesma disciplina em Escolas diferentes; a distancia entre as médias de uma mesma disciplina na Escola 1 e Escola 2 são inferiores a 1; os centroides não parecem estar distantes; e as matrizes de covariância e correlação não apontam alta dependência entre as notas das disciplinas tanto para a Escola 1 quanto para a Escola 2, exceto para a relação entre as notas de Português e História.

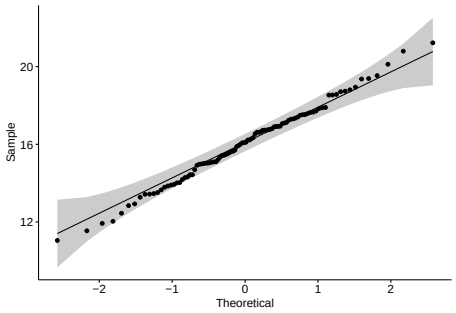
1.3 - De acordo com as premissas adotadas na simulação dos dados, qual é a distribuição amostral da estatística $\bar{\bar{Y}}_1 - \bar{\bar{Y}}_2$? Justifique. Com base nos dados simulados, construa um gráfico de quantis da Normal para validar os resultados.

R: Os dados foram simulados com base em duas normais multivariadas, a estatística $\bar{\bar{Y}}_1 - \bar{\bar{Y}}_2$ terá como distribuição amostral uma normal multivariada, pois qualquer combinação linear entre distribuições normais multivariadas resultará em uma distribuição normal multivariada. Verificaremos se cada amostra tem distribuição normal multivariada, para tal, avaliaremos se soma das notas de português, matemática e história seguem um distribuição normal.

Hipóteses para a amostra da Escola 1:

$$H_0 : Por_1 + Mat_1 + His_1 \sim N(\mu_1, \sigma_1^2)$$

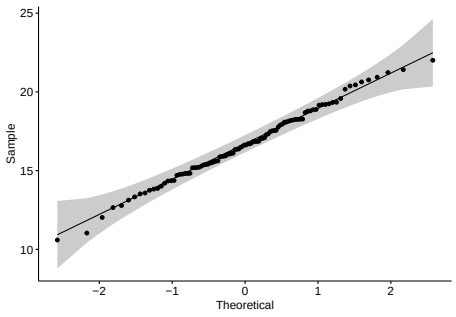
$$H_1 : Por_1 + Mat_1 + His_1 \nsim N(\mu_1, \sigma_1^2)$$



Hipóteses para a amostra da Escola 2:

$$H_0 : Por_2 + Mat_2 + His_2 \sim N(\mu_2, \sigma_2^2)$$

$$H_1 : Por_2 + Mat_2 + His_2 \nsim N(\mu_2, \sigma_2^2)$$



A partir dos QQ-plots apresentados, verificamos que as soma das variáveis nas duas escolas têm distribuições amostrais normais, evidenciando que, conjutamente, cada escola tem distribuição amostral multivariada nas notas de português, matemática e história. Consequentemente, $\bar{\bar{Y}}_1 - \bar{\bar{Y}}_2$ também resultará em uma $N(\mu_1 - \mu_2, \sigma_1^2 - \sigma_2^2)$.

1.4 - Há evidência amostral de diferença significante entre os vetores de Médias das duas populações? Justifique.

R:

```
>
>   Box's M-test for Homogeneity of Covariance Matrices
>
> data:  pop[, -4]
> Chi-Sq (approx.) = 3.5082, df = 6, p-value = 0.7429
```

Como o p-valor de 0.7429 é maior que o nível de significância de 5%. Não rejeitamos a hipótese nula de que as matrizes de covariâncias são iguais.

```
> Test stat:  78.193
> Numerator df:  3
> Denominator df:  196
> P-value:  0.0000000000000413
```

De acordo com o teste T de Hotelling, onde o p-value é menor que o nível de significância de 5%, deste modo rejeitamos a hipótese nula e concluímos que há algumas diferença entre as médias das escolas para as disciplinas de português, matemática e história.

1.5 - Para cada variável, compare as médias das duas populações. Utilize correções de Bonferroni e FDR na conclusão dessas comparações. Qual variável mais contribui para a possível diferença entre as populações?

R:

```
>
>   Bartlett test of homogeneity of variances
>
> data:  pop$Por by pop$Pop
> Bartlett's K-squared = 0.47532, df = 1, p-value = 0.4905


>
>   Two Sample t-test
>
> data:  pop$Por by pop$Pop
> t = -5.6957, df = 198, p-value = 0.00000004394
> alternative hypothesis: true difference in means between group Escola 1 and
    group Escola 2 is not equal to 0
> 95 percent confidence interval:
>  -1.1008876 -0.5346247
> sample estimates:
> mean in group Escola 1 mean in group Escola 2
>               4.990250                5.808006
```

Como o p-valor do teste T de 0.4905 é maior que o nível de significância de 5%, não rejeitamos a hipóteses nula e concluímos que as variâncias das notas de português entre os grupos são homogêneas. E, como o p-valor do teste T de 0 é menor que o nível de significância de 5%, rejeitamos a hipóteses nula e concluímos que as médias das notas de português entre as escolas 1 e 2 são diferentes, ao nível de 95% de confiança.

```
>
>    Bartlett test of homogeneity of variances
>
> data:  pop$Mat by pop$Pop
> Bartlett's K-squared = 1.0278, df = 1, p-value = 0.3107

>
>    Two Sample t-test
>
> data:  pop$Mat by pop$Pop
> t = 5.061, df = 198, p-value = 0.0000009501
> alternative hypothesis: true difference in means between group Escola 1 and
    group Escola 2 is not equal to 0
> 95 percent confidence interval:
>  0.4019407 0.9151492
> sample estimates:
> mean in group Escola 1 mean in group Escola 2
>               5.706867                5.048322
```

Como o p-valor do teste T de 0.3107 é maior que o nível de significância de 5%, não rejeitamos a hipóteses nula e concluímos que as variâncias das notas de matemática entre os grupos são homogêneas. E, como o p-valor do teste T de 0 é menor que o nível de significância de 5%, rejeitamos a hipóteses nula e concluímos que as médias das notas de matemática entre as escolas 1 e 2 são diferentes, ao nível de 95% de confiança.

```
>
>    Bartlett test of homogeneity of variances
>
> data:  pop$His by pop$Pop
> Bartlett's K-squared = 0.43195, df = 1, p-value = 0.511

>
>    Two Sample t-test
>
> data:  pop$His by pop$Pop
> t = -3.3919, df = 198, p-value = 0.000838
> alternative hypothesis: true difference in means between group Escola 1 and
    group Escola 2 is not equal to 0
> 95 percent confidence interval:
>  -0.7426334 -0.1965829
> sample estimates:
> mean in group Escola 1 mean in group Escola 2
>               5.325944                5.795552
```

Como o p-valor do teste T de 0.511 é maior que o nível de significância de 5%, não rejeitamos a hipóteses nula e concluímos que as variâncias das notas de história entre os grupos são homogêneas. E, como o p-valor do teste T de 0.0008 é menor que o nível de significância de 5%, rejeitamos a hipóteses nula e concluímos que as médias das notas de história entre as escolas 1 e 2 são diferentes, ao nível de 95% de confiança.

Disciplina	p.result	padjustB	padjustFDR
Por	0.000000	0.000000	0.000000
Mat	0.000001	0.000003	0.000001
His	0.000838	0.002514	0.000838

Mesmo considerando as correções de Boferroni e FDR, percebemos que há diferença significativa entre as médias de cada variável nas duas amostras (Escola 1 e Escola 2).

2 - Comparação de vetores de Médias de Normais tridimensionais, N_3, em Delineamentos Completamente Aleatorizados com Estrutura Fatorial Cruzado 2x2 - MANOVA

Gerar dados da N3 de acordo com um Delineamento Completamente Aleatorizado (DCA) Fatorial Cruzado 2x2. Considerando os Fatores F1 e F2, cada um em dois níveis, 0 e 1, preencha a tabela a seguir com os parâmetros adotados na simulação dos dados.

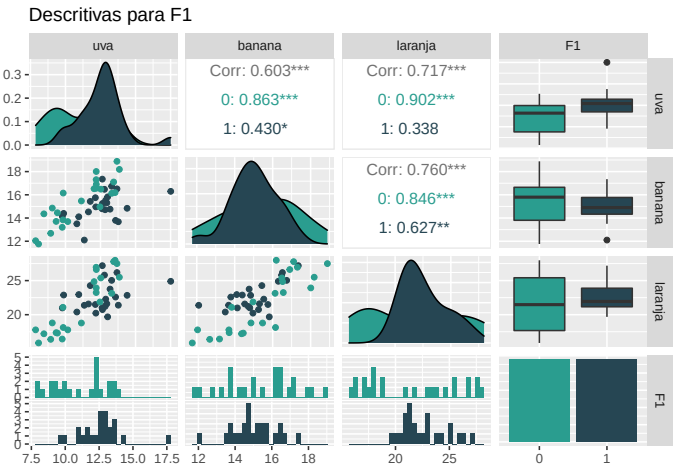
Fator 1	0	1
0	$\mu_{0,0} = (9, 14, 18), \Sigma_{0,0} = \begin{pmatrix} 3 & 1.2 & 1.9 \\ 1.2 & 2 & 1.5 \\ 1.9 & 1.5 & 3 \end{pmatrix}$	$\mu_{0,1} = (13, 17, 26), \Sigma_{0,1} = \begin{pmatrix} 3 & 1.2 & 1.9 \\ 1.2 & 2 & 1.5 \\ 1.9 & 1.5 & 3 \end{pmatrix}$
1	$\mu_{1,0} = (12, 15, 23), \Sigma_{1,0} = \begin{pmatrix} 3 & 1.2 & 1.9 \\ 1.2 & 2 & 1.5 \\ 1.9 & 1.5 & 3 \end{pmatrix}$	$\mu_{1,1} = (13, 15, 22), \Sigma_{1,1} = \begin{pmatrix} 3 & 1.2 & 1.9 \\ 1.2 & 2 & 1.5 \\ 1.9 & 1.5 & 3 \end{pmatrix}$

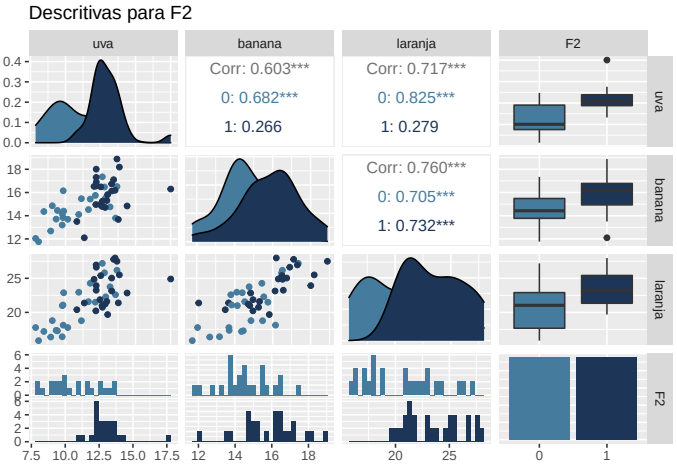
2.1 - Contextualize, com uma situação prática hipotética, os dados gerados. Caracterize a estrutura dos dados e defina o objetivo do estudo.

R: Os dados são referentes a um experimento fatorial 2x2, onde averiguou-se a produção, em toneladas por hectare, de três tipos de frutas (uva, banana e laranja) sob o efeito de dois tipos de adubos F1 e F2, que serão avaliados em dois níveis (1 - presente; 0 - ausente), sendo que o objetivo é verificar se o uso dos adubos F1 e F2 têm (ou não) efeito significativo sobre a produção das três frutas.

2.2 - Realize uma análise descritiva dos dados (calcule estatísticas descritivas, construa gráficos apropriados). Comente os resultados de acordo com o objetivo do estudo.

R: Os gráficos de dispersão relacionados ao Fator 1 apontam que há marcante intersecção entre os valores obtidos na produção de banana, laranja e uva, ao avaliar se há correlação comparando dois a dois, daí vemos que os pontos se sobrepõe, o que demonstra não haver clara "distinção" entre a produção de cada fruta, tendo em vista a aplicação do fator F1. Comportamento similar é percebido quando na aplicação do fator F2. Os box plot obtidos com o fator F1 apontam que para a produção de uva medianamente há pouco distinção entre a aplicar ou não o fator F1, bem como para a produção de banana e laranja. Contudo, notamos que a dispersão (variabilidade dos dados) é maior quando o fator F1 está no nível "0". Comportamento similar é percebido para o fator F2, no que tange aos valores medianos (não são muito distantes) ao aplicar ou não o fator 2 na produção das frutas. Contudo, notamos que a dispersão (variabilidade dos dados) é maior quando o fator F1 está no nível "0"para a produção de uva e laranja.



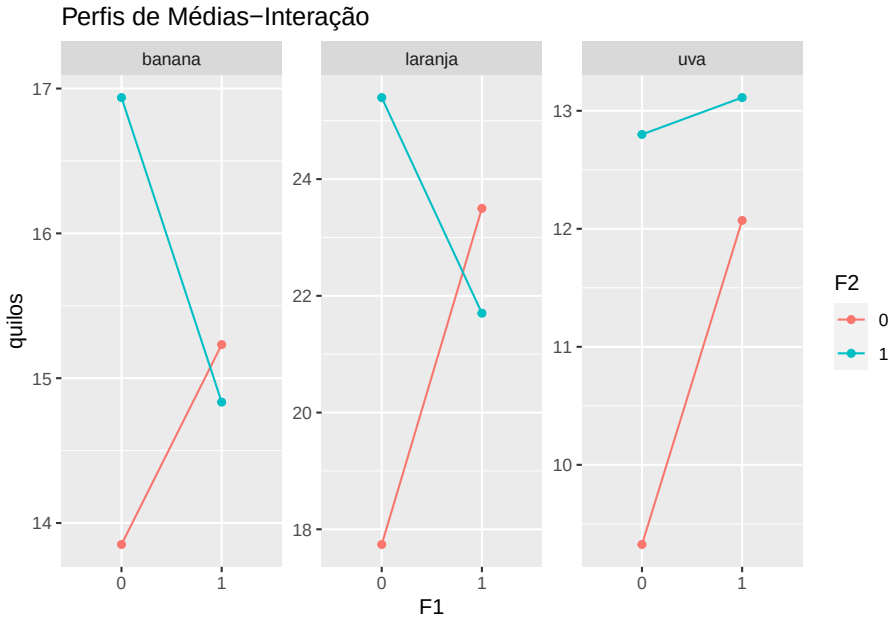


Ao avaliar (graficamente) o efeito de interação entre os fatores, percebe-se que há interação significativa na produção de banana e laranja, já na produção de uva não há evidência de interação significativa do uso dos adubos F1 e F2 (ou seja, F1*F2 é significativo apenas na produção de banana e laranja).

Fator 1	0	1
0	$\bar{Y}_{0,0} = (9.33, 13.85, 17.74)$	$\bar{Y}_{0,1} = (12.8, 16.94, 25.4)$
1	$\bar{Y}_{1,0} = (12.07, 15.23, 23.5)$	$\bar{Y}_{1,1} = (13.11, 14.83, 21.7)$

Fator 1	0	1
0	$S_{0,0} = \begin{pmatrix} 0.85 & 0.79 & 0.56 \\ 0.79 & 1.73 & 0.85 \\ 0.56 & 0.85 & 1.93 \end{pmatrix}$	$S_{0,1} = \begin{pmatrix} 0.48 & 0.24 & 0.58 \\ 0.24 & 1.12 & 1.04 \\ 0.58 & 1.04 & 4.12 \end{pmatrix}$
1	$S_{1,0} = \begin{pmatrix} 1.74 & 0.69 & 1.29 \\ 0.69 & 1.27 & 1.75 \\ 1.29 & 1.75 & 3.98 \end{pmatrix}$	$S_{1,1} = \begin{pmatrix} 2.85 & 1.23 & 2.01 \\ 1.23 & 1.61 & 1.04 \\ 2.01 & 1.04 & 2.95 \end{pmatrix}$

Fator 1	0	1
0	$R_{0,0} = \begin{pmatrix} 1 & 0.65 & 0.44 \\ 0.65 & 1 & 0.47 \\ 0.44 & 0.47 & 1 \end{pmatrix}$	$R_{1,1} = \begin{pmatrix} 1 & 0.33 & 0.41 \\ 0.33 & 1 & 0.49 \\ 0.41 & 0.49 & 1 \end{pmatrix}$
1	$R_{1,0} = \begin{pmatrix} 1 & 0.46 & 0.49 \\ 0.46 & 1 & 0.78 \\ 0.49 & 0.78 & 1 \end{pmatrix}$	$R_{1,1} = \begin{pmatrix} 1 & 0.57 & 0.69 \\ 0.57 & 1 & 0.48 \\ 0.69 & 0.48 & 1 \end{pmatrix}$



2.3 - Construa a Tabela de MANOVA para a análise destes dados. Considere as fontes de variação devido aos efeitos principais dos fatores F1 e F2 e sua interação $F1*F2$, os correspondentes números de graus de liberdade e as Somas de Quadrados e Produtos Cruzados (SS_{F1}, SS_{F2}, e SS_{F1*F2}, bem como SS_W). Há evidência amostral de efeito significante dos fatores sob estudo?

R: Para a construção da tabela MANOVA, primeiramente, foi avaliado a homogeneidade entre as matrizes de covariâncias com o Teste M de Box.

```
>
>   Box's M-test for Homogeneity of Covariance Matrices
>
> data:  pop[, 1:3]
> Chi-Sq (approx.) = 11.609, df = 6, p-value = 0.07129
```

Como o p-valor de 0.0713 é maior que o nível de significância de 5%. Não rejeitamos a hipótese nula de que as matrizes de covariâncias são iguais. Prosseguimos com a MANOVA:

```
> Call:
>    manova(as.matrix(pop[, 1:3]) ~ pop$F1 + pop$F2 + pop$F1 * pop$F2)
>
> Terms:
>                       pop$F1    pop$F2 pop$F1:pop$F2 Residuals
> uva                28.05527  61.17120       17.76620  65.16191
> banana              1.56727  21.68526       36.39486  63.04966
> laranja            12.77955 103.03284      268.08866 142.89022
> Deg. of Freedom           1         1              1        44
>
> Residual standard errors: 1.216944 1.197058 1.802083
> Estimated effects may be unbalanced
```

	Df	Pillai	approx F	num Df	den Df	Pr(>F)
F1	1	0.4417953	11.08040	3	42	1.73e-05
F2	1	0.5273158	15.61809	3	42	6.00e-07
F1:F2	1	0.6577506	26.90585	3	42	0.00e+00
Residuals	44					

	Df	Wilks	approx F	num Df	den Df	Pr(>F)
F1	1	0.5582047	11.08040	3	42	1.73e-05
F2	1	0.4726842	15.61809	3	42	6.00e-07
F1:F2	1	0.3422494	26.90585	3	42	0.00e+00
Residuals	44					

Como os p-valores pelos teste de Wilks e Pillai de F1 são menores que o nível de significância de 5%. Rejeitamos a hipótese nula e concluímos que há diferença no vetor médias quando há aplicação do adubo F1.

F1	uva	banana	laranja
0	11.06236	15.39475	21.56821
1	12.59140	15.03335	22.60018

Como os p-valores pelos teste de Wilks e Pillai de F2 são menores que o nível de significância de 5%. Rejeitamos a hipótese nula e concluímos que há diferença no vetor médias quando há aplicação do adubo F2.

F2	uva	banana	laranja
0	10.69799	14.54191	20.61909
1	12.95577	15.88619	23.54929

Como os p-valores pelos teste de Wilks e Pillai para a interação de F1 com F2 são menores que o nível de significância de 5%. Rejeitamos a hipótese nula e concluímos que há diferença no vetor médias quando há combinação dos adubos F1 e F2.

F1	F2	uva	banana	laranja
0	0	9.325088	13.85184	17.73981
0	1	12.799639	16.93765	25.39660
1	0	12.070885	15.23197	23.49838
1	1	13.111907	14.83473	21.70198

Somas de Quadrados e Produtos Cruzados SS_{F1}:

uva	banana	laranja
28.055273	-6.631002	18.934991
-6.631002	1.567270	-4.475378
18.934991	-4.475378	12.779554

Somas de Quadrados e Produtos Cruzados SS_{F2}:

uva	banana	laranja
61.17120	36.42133	79.38918
36.42133	21.68526	47.26832
79.38918	47.26832	103.03284

Somas de Quadrados e Produtos Cruzados SS_{F1*F2}:

uva	banana	laranja
17.7662	25.42830	69.01390
25.4283	36.39486	98.77778
69.0139	98.77778	268.08866

Somas de Quadrados e Produtos Cruzados SS_F:

uva	banana	laranja
106.99267	55.21863	167.3381
55.21863	59.64739	141.5707
167.33807	141.57072	383.9011

Somas de Quadrados e Produtos Cruzados SS_W:

uva	banana	laranja
65.16191	32.40356	48.71678
32.40356	63.04966	51.62471
48.71678	51.62471	142.89022

Somas de Quadrados e Produtos Cruzados SS_{total}:

uva	banana	laranja
172.15459	87.62219	216.0549
87.62219	122.69705	193.1954
216.05485	193.19544	526.7913

2.4 - De acordo com os resultados da MANOVA, realize comparações múltiplas para estudar os efeitos significantes dos fatores. Utilize correções de Bonferroni e FDR. Interprete os resultados.

R: [responder]

Adubo	Fruta	P-Valor Teste de Bartlett	P-Valor T Test	padjustB	padjustFDR
F1	uva	0.3173448	0.0044368	0.0266207	0.0053241
F1	banana	0.0207380	0.4451967	1.0000000	0.4451967
F1	laranja	0.3173448	0.0044368	0.0266207	0.0053241
F2	uva	0.1085063	0.0000078	0.0000468	0.0000234
F2	banana	0.5671792	0.0029293	0.0175757	0.0053241
F2	laranja	0.1085063	0.0000078	0.0000468	0.0000234

2.5 - Da tabela MANOVA obtida em 1.3 construa a correspondente tabela MANOVA de um estudo que considera o efeito total dos 4 grupos definidos pela estrutura fatorial 2x2. Neste caso, seja SS_F a Soma de Quadrados e Produtos Cruzados do referido fator em 4 níveis. Há efeito significante deste fator F que combina os níveis de F1 e F2?

R:

```
> Call:
>     manova(as.matrix(pop_v2[, 1:3]) ~ pop_v2$F1F2)
>
> Terms:
>                    pop_v2$F1F2 Residuals
> uva                   106.9927   65.1619
> banana                 59.6474   63.0497
> laranja               383.9011  142.8902
> Deg. of Freedom              3        44
>
> Residual standard errors: 1.216944 1.197058 1.802083
> Estimated effects may be unbalanced
```

	Df	Pillai	approx F	num Df	den Df	Pr(>F)
F1F2	3	1.317562	11.48585	9	132	0
Residuals	44					

	Df	Wilks	approx F	num Df	den Df	Pr(>F)
F1F2	3	0.123277	15.50823	9	102.3676	0
Residuals	44					

Como os p-valores pelos teste de Wilks e Pillai para F são menores que o nível de significância de 5%. Rejeitamos a hipótese nula e concluímos que há diferença em ao menos um vetor médias de F.

F1F2	uva	banana	laranja
00	9.325088	13.85184	17.73981
01	12.799639	16.93765	25.39660
10	12.070885	15.23197	23.49838
11	13.111907	14.83473	21.70198

Somas de Quadrados e Produtos Cruzados SS_F:

uva	banana	laranja
106.99267	55.21863	167.3381
55.21863	59.64739	141.5707
167.33807	141.57072	383.9011

Somas de Quadrados e Produtos Cruzados SS_W:

uva	banana	laranja
65.16191	32.40356	48.71678
32.40356	63.04966	51.62471
48.71678	51.62471	142.89022

Somas de Quadrados e Produtos Cruzados SS_{total}:

uva	banana	laranja
82.92812	57.83186	117.7307
57.83186	99.44452	150.4025
117.73068	150.40249	410.9789

2.6 - Obtenha a decomposição espectral (autovalores e autovetores) das seguintes matrizes: $SS_W^{-1}SS_{F1}, SS_W^{-1}SS_{F2}, SS_W^{-1}SS_{F1*F2}, SS_W^{-1}SS_F$. Qual é o padrão de contribuição das variáveis para cada um dos efeitos considerados?

R: [responder]

Autovalores:

0.7915	0	0
1.1156	0	0
1.9218	0	0

Autovetores de $SS_W^{-1}SS_{F1}$

0.7579	-0.3515	0.0106
-0.6372	-0.9145	0.9486
0.1397	0.2005	0.3164

Autovetores de $SS_W^{-1} SS_{F2}$

0.9049	-0.7494	-0.6775
-0.0835	0.5859	0.7086
0.4173	0.3086	0.1970

Autovetores de $SS_W^{-1} SS_{F1*F2}$

0.2565+0i	-0.0360-0.5154i	-0.0360+0.5154i
-0.1065+0i	0.7967+0.0000i	0.7967+0.0000i
-0.9606+0i	-0.2843+0.1327i	-0.2843-0.1327i

Autovetores de $SS_W^{-1} SS_F$

-0.4379	0.8346	0.2506
0.1546	-0.5067	0.8672
-0.8856	-0.2161	-0.4304