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

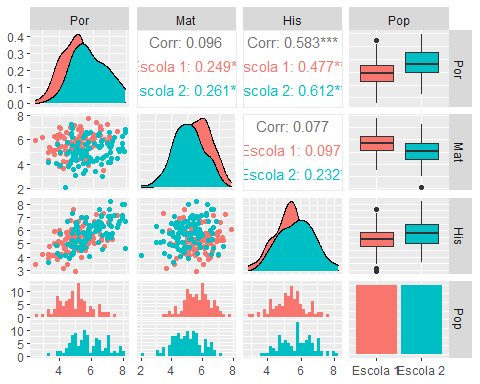
1 - Comparação de vetores de Médias de 2 populações N3. Teste T2 de Hotelling.

Gerar uma amostra aleatória de observações de duas populações Normais tridimensionais, , envolvendo as variáveis Y1, Y2 e Y3. Preencha a tabela a seguir com os parâmetros adotados na simulação dos dados.

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.

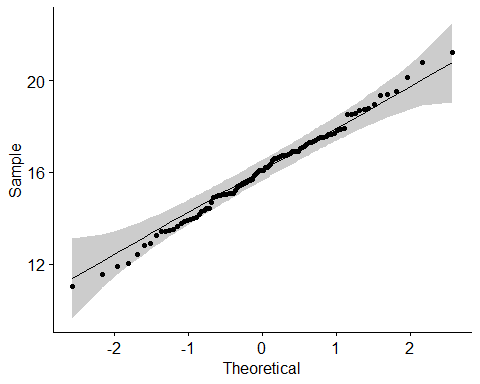


R: Ao observar os box plots de cada disciplina da Escola 1 e Escola 2 percebe que as medianas não estão distantes; ao observar os histogramas percebe-se que as distribuições da notas em cada disciplina são parecidas; os gráficos de dispersão apontam que há uma marcante intercepção entre as notas da mesma 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.

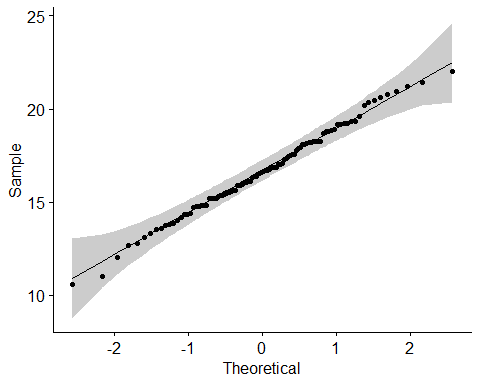
1.3 - De acordo com as premissas adotadas na simulação dos dados, qual é a distribuição amostral da estatística ? Justifique. Com base nos dados simulados, construa um gráfico de quantis da Normal para validar os resultados.

R: Os dados foram simulados com base em duas normais multivariadas, a estatística 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:



Hipóteses para a amostra da Escola 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, também resultará em uma .

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.

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, , em Delineamentos Completamente Aleatorizados com Estrutura Fatorial Cruzado 2x2 - MANOVA

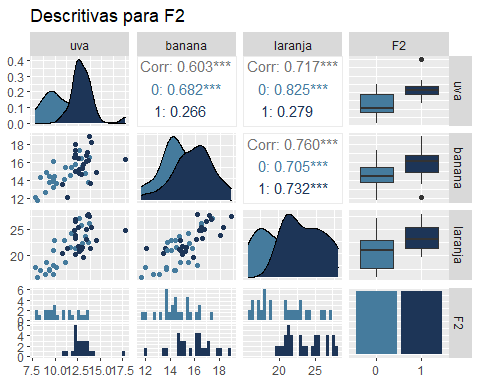
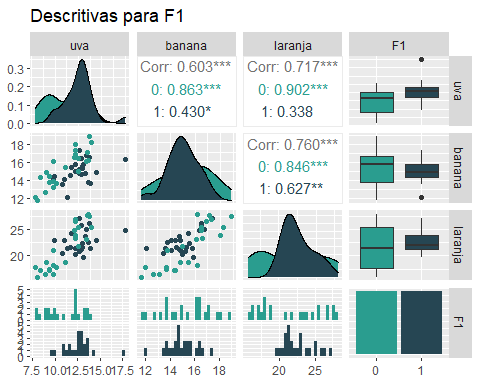
Gerar dados da N3 de acordo com um Delineamento Completamente Aleatorizado (DCA) Fatorial Cruzado 2x2. Considerando os Fatores F1 e F2, cada um em dois níveis, 0 e 1, preencha a tabela a seguir com os parâmetros adotados na simulação dos dados.

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

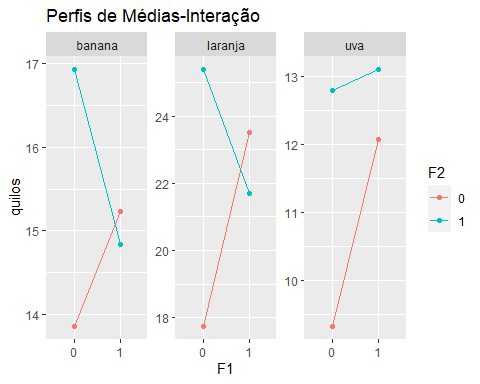
R: Os dados são referentes a um experimento fatorial 2x2, onde averiguou-se a produção, em toneladas por hectare, de três tipos de frutas (uva, banana e laranja) sob o efeito de dois tipos de adubos F1 e F2, que serão avaliados em dois níveis (1 - presente; 0 - ausente), sendo que o objetivo é verificar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



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 , os correspondentes números de graus de liberdade e as Somas de Quadrados e Produtos Cruzados (, , e , bem como ). 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

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.

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.

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.

Somas de Quadrados e Produtos Cruzados :

Somas de Quadrados e Produtos Cruzados :

Somas de Quadrados e Produtos Cruzados :

Somas de Quadrados e Produtos Cruzados :

Somas de Quadrados e Produtos Cruzados :

Somas de Quadrados e Produtos Cruzados :

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]

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 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

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.

Somas de Quadrados e Produtos Cruzados :

Somas de Quadrados e Produtos Cruzados :

Somas de Quadrados e Produtos Cruzados :

2.6 - Obtenha a decomposição espectral (autovalores e autovetores) das seguintes matrizes: . Qual é o padrão de contribuição das variáveis para cada um dos efeitos considerados?

R: [responder]

Autovalores:

Autovetores de

Autovetores de

Autovetores de

Autovetores de