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 ng 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.

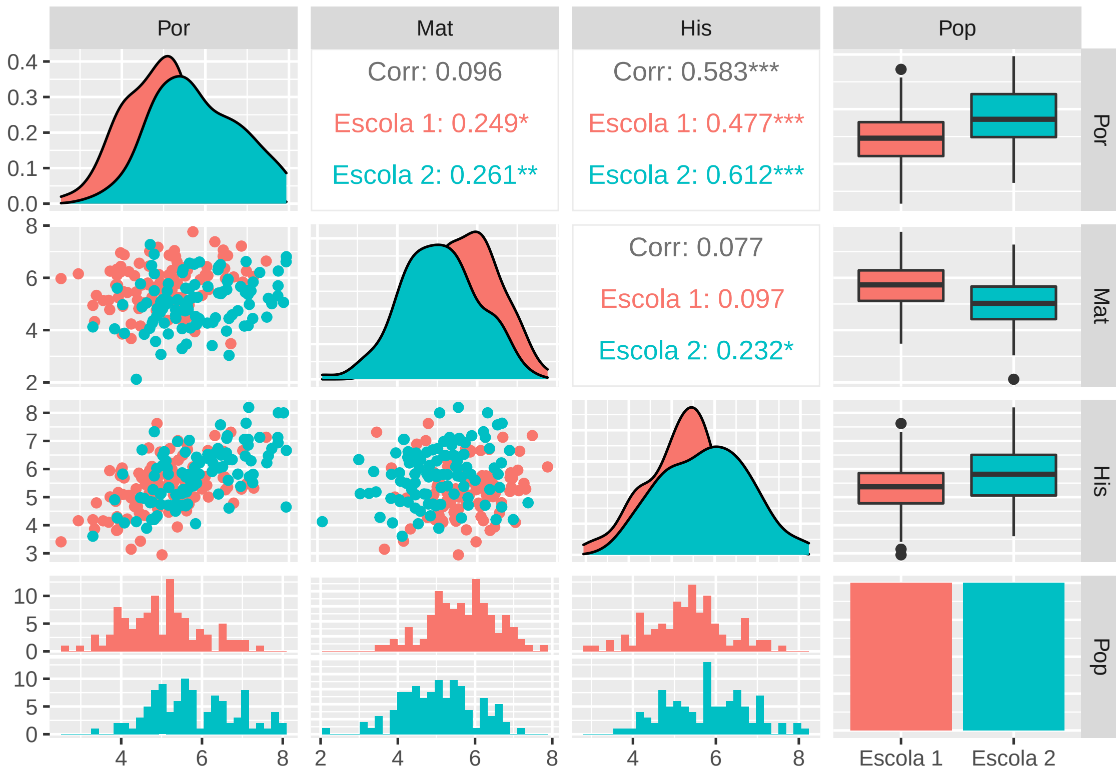
População	Amostra	Vetor de Médias	Matriz de Covariâncias	Matriz de Correlações
Pop 1	$n_1 = 100$	$\mu_1 = (5, 5.7, 5.5)$	$\Sigma_1 = \begin{pmatrix} 1 & 0.3 & 0.5 \\ 0.3 & 0.9 & 0.2 \\ 0.5 & 0.2 & 1.1 \end{pmatrix}$	$\rho_1 = \begin{pmatrix} 1 & 0.32 & 0.48 \\ 0.32 & 1 & 0.2 \\ 0.48 & 0.2 & 1 \end{pmatrix}$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

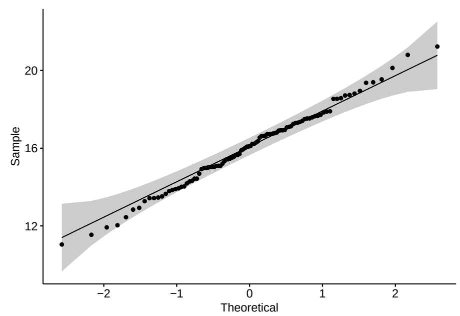
1.3 - De acordo com as premissas adotadas na simulação dos dados, qual é a distribuição amostral da estatística $\bar{\bar{Y}}_1 - \bar{\bar{Y}}_2$? Justifique. Com base nos dados simulados, construa um gráfico de quantis da Normal para validar os resultados.

R: Os dados foram simulados com base em duas nomais multivariadas, a estatística $\bar{\bar{Y}}_1 - \bar{\bar{Y}}_2$ terá como distribuição amostral uma normal multivariada, pois qualquer combinação linear entre distribuições multivariada 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 segue um distribuição normal.

Hipóteses para a amostra da Escola 1:

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

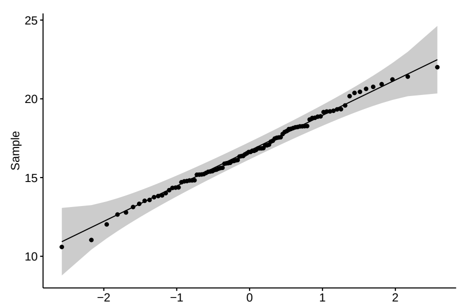
$$H_1 : Por_1 + Mat_1 + His_1 \nsim N(\mu_1, \sigma_1^2)$$



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 covariancias são iguais.

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

Como o p-valor de 0 é menor que o nível de significância de 5%. 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?

```
##
##  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.

```
## [1] 0.00000004394109 0.00000095006350 0.00083801147562
```

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

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.

$$tabela$$

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

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

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 (SSF1, SSF2, e SSF1*F2, bem como SSW). Há evidência amostral de efeito significante dos fatores sob estudo?

1.4 - 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.

1.5 - 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 SSF 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?

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