UNIVERSIDADE ESTADUAL DO OESTE DO PARANÁ

CENTRO DE CIÊNCIAS EXATAS E TÉCNOLÓGICAS

PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA AGRÍCOLA

Gabriela Medeiros

Marília Melo Favalesso

1ª LISTA DE EXERCÍCIOS

DISCIPLINA: ANÁLISE MULTIVARIADA

PROFESSORA: DRA. LUCIANA PAGLIOSA C. GUEDES

CASCAVEL-PR

2016

**1. Seja a matriz** A=**. Responda:**

**a) A2x2 é simétrica?**

1. Sim, a matriz A é simétrica pois At=A

A= 🡪 At

**b) Calcule Ae A-1.**

1. |A|=

A-1=

**c) Mostre que A2x2 é positiva definida e escreva sua forma quadrática.**

Como a matriz A é simétrica, e tal que xtAx > 0, ∀ x não nulo, então, dizemos que A é uma matriz definida positiva:

xtAx =

Assim, a forma quadrática da matriz A será:

🡪 ∀ x ≠ 0

Então, A = é definida positiva.

**d) Determine seus autovalores e autovetores. Determine a norma dos autovetores.** Autovalores: λ1=9,24 λ2=4,76

Autovetores:

Norma ou módulo dos autovetores: ||e1||=1 e ||e2||=1.

**e) Escreva a decomposição espectral dessa matriz.**

Decomposição espectral dessa matriz:

A=

9,24+4,76

**f) Determine A⊗I2.**

A⊗I2=

**2. Numa exploração agrícola da Bélgica registraram-se os valores de 5 variáveis meteorológicas ao longo 11 anos agrícolas na década de 1920. As cinco variáveis são: X1 = precipitação total em novembro e dezembro (mm); X2= temperatura média em julho (oC); X3 = precipitação total em julho (mm); X4 = radiação em julho (mm de álcool); X5 = rendimento médio de colheita (quintais/hectare). Os valores observados foram:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Campanha** | **x1** | **x2** | **x3** | **x4** | **x5** |
| 1920-21 | 87,9 | 19,6 | 1 | 1661 | 28,37 |
| 21-22 | 89,9 | 15,2 | 90,1 | 968 | 23,77 |
| 22-23 | 153 | 19,7 | 56,6 | 1353 | 26,04 |
| 23-24 | 132,1 | 17 | 91 | 1293 | 25,74 |
| 24-25 | 88,8 | 18,3 | 93,7 | 1153 | 26,68 |
| 25-26 | 110,9 | 17,8 | 106,9 | 1286 | 24,29 |
| 26-27 | 117,7 | 17,8 | 106,9 | 1286 | 24,29 |
| 27-28 | 109 | 18,3 | 41,8 | 1574 | 28,37 |
| 28-29 | 156,1 | 17,8 | 57,4 | 1222 | 24,96 |
| 29-30 | 181,5 | 16,8 | 140,6 | 903 | 21,66 |
| 30-31 | 181,4 | 17 | 74,3 | 1150 | 24,37 |

**a) Construa o vetor de médias amostrais, a matriz de variâncias e covariâncias amostrais. Interprete os resultados obtidos.**

Vetor de médias amostrais: [128,03, 17,75, 78,21, 1259,00, 25,32]

Matriz de variâncias e covariâncias:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **x1** | **x2** | **x3** | **x4** | **x5** |
| **x1** | 1250,10 | -5,42 | 399,81 | -3101,96 | -38,32 |
| **x2** | -5,42 | 1,64 | -29,09 | 217,93 | 1,72 |
| **x3** | 399,81 | -29,09 | 1428,60 | -6867,14 | -63,03 |
| **x4** | -3101,96 | 217,93 | -6867,14 | 50818,20 | 393,99 |
| **x5** | -38,32 | 1,72 | -63,03 | 393,99 | 4,01 |

Resposta:

Na tabela 1 são apresentadas as médias das variáveis meteorológicas amostradas seguidas das variâncias. A média é uma medida de tendência central dos dados com a intenção de utilizar um valor numérico para representação de uma determinada variável. Ela é o valor que aponta para onde mais se concentram os dados em uma distribuição. Já a variância é uma medida de dispersão dos dados (ou variação). Ela tem como objetivo observar os desvios de valores de dada variável em relação à média. Quanto maior é o valor da variância, maior é a dispersão dos dados de uma dada variável. Quando não houver variabilidade nos dados, a variância será igual a zero.

No presente estudo, observamos ampla dispersão nos dados nas variáveis x1, x3 e x4 (Tabela 1). Já as variáveis x2 e x5 apresentaram menores dispersão de valores em relação à média (Tabela 1).

Tabela 1. Média () ± a variância (S) para as diferentes variáveis meteorológicas amostradas ao longo dos 11 anos agrícolas na década de 20.

|  |  |  |
| --- | --- | --- |
| **Variável** | **Símbolo variável** | **± S** |
| Precipitação total nov-dez (mm) | x1 | 128,03 ± 1250,10 |
| TºC jul | x2 | 17,75 ± 1,64 |
| Precipitação total jul (mm) | x3 | 78,21 ± 1428,60 |
| Radiação em jul (mm) | x4 | 1259,00 ± 50818,20 |
| Rendimento colheita (quintais/hectare) | x5 | 25,32 ± 4,01 |

Na análise de covariância um valor positivo indica uma dependência positiva, isto é, maiores valores de uma das variáveis estão associados a maiores valores de outra variável. O contrário ocorre para um valor negativo, isto é, nesse caso maiores valores de uma das variáveis estão associados com menores valores de outra variável. Um valor nulo indica que não existe esse tipo de dependência. Para o presente estudo, as associações das variáveis são demonstradas na tabela 2.

Tabela 2. Matriz com os valores de covariância para cada variável amostrada na década de 20, onde: x1= Precipitação total em Novembro e Dezembro (mm), x2= Temperatura média em Julho (ºC), x3= Precipitação total em Julho (mm), x4= Radiação em Julho (mm de álcool), x5= Rendimento médio de colheita (quintais/hectare).

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **x2** | **x3** |  | **x4** | **x5** |
| **x1** | -5,42 | 399,81 |  | -3101,96 | -38,32 |
| **x2** |  | -29,09 |  | 217,93 | 1,72 |
| **x3** |  |  |  | -6867,14 | -63,03 |
| **x4** |  |  |  |  | 393,99 |

**b) Construa a matriz de correlação amostral. Realize o teste de hipótese para identificar quais correlações são significativas. Interprete os resultados obtidos.**

Com a intenção de avaliar a correlação entre as variáveis amostradas, foi construída uma matriz de correlação de Pearson (r). Para avaliar a significância da correlação foi testada a hipótese nula de que r = 0 (ou seja, que não existe correlação entre as varáveis), utilizando para tanto uma distribuição t e com α = 5%. Para a interpretação dos coeficientes de correlação (r) foram consideradas as seguintes magnitudes: 0 (Não existe associação entre as variáveis preditoras e descritoras), < 0.3 (a associação é muito fraca), 0.3 < x ≤ 0.6 (a associação é moderada), 0.6 < x ≤ 0.9 (a associação é forte), 0.9 < x < 1.0 (a associação é muito forte), 1 (a associação é perfeita) (CALLEGARI-JACQUES, 2003).

A partir da matriz de correlação (tabela 3) verificamos que houve correlação forte e positiva entre as variáveis x2-x4 e x2-x5, uma correlação forte e inversamente proporcional entre x3-x4 e x4-x5, e uma correlação forte e positiva entre x4-x5. As demais variáveis não apresentaram correlação significativa (tabela 3).

Tabela 3. Matriz de correlação de Pearson (r) entre as variáveis amostradas numa exploração agrícola da Bélgica na década de 20, onde: x1= Precipitação total em Novembro e Dezembro (mm), x2= Temperatura média em Julho (ºC), x3= Precipitação total em Julho (mm), x4= Radiação em Julho (mm de álcool), x5= Rendimento médio de colheita (quintais/hectare). \*P-valor < 0,05.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **x2** | **x3** | **x4** | **x5** |
| **x1** | -0.12 | 0.30 | -0.39 | -0.54 |
| **x2** |  | -0.60 | 0.76\* | 0.67\* |
| **x3** |  |  | -0.81\* | -0.83\* |
| **x4** |  |  |  | 0.87\* |

**c) Construa para cada variável o gráfico box-plot e histograma e analise os resultados, observando simetria, pontos discrepantes.**

Em uma análise visual dos gráficos box-plot (figura 1) e histograma (figura 2) observamos assimetria positiva das variáveis x1 e x5, e assimetria negativa para as variáveis x2, x3 e x4. As medidas de assimetria possibilitam analisar uma distribuição de dados de acordo com as suas medidas de tendência central (moda, média e mediana) quando observadas em um gráfico. Quando os dados se distribuem mais a direta de um gráfico temos uma distribuição de dados com curva assimétrica positiva, do contrário a distribuição é assimétrica negativa. Caso os dados estejam igualmente distribuídos, temos uma distribuição simétrica. A avaliação visual dos gráficos é bastante subjetiva, sendo mais usual a utilização de coeficientes para conclusões sobre a assimetria de uma distribuição de dados, bem como a aplicação de inferências para validação deste.

No box-plot (figura 1) também vemos um ponto discrepante na variável x4 acima do limite máximo. O ponto discrepante é relativo a campanha de 1920-1921 (figura 1).

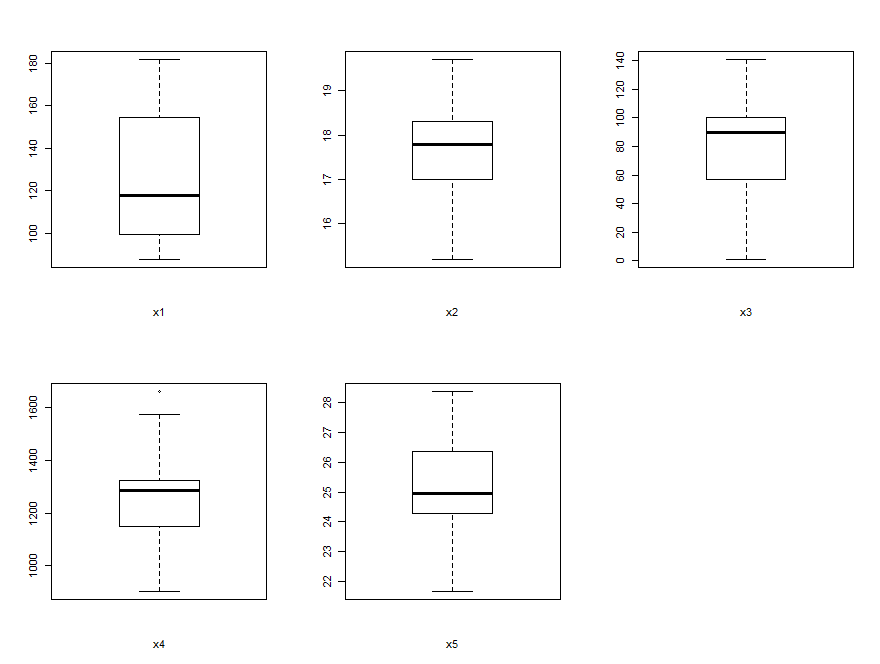


Figura 1. Box-plot com a distribuição dos valores das cinco variáveis amostradas numa exploração agrícola da Bélgica na década de 20. O box-plot é composto por mediana (traço mais escuro), 1º e 3º quartil (caixa), valores dos limites superior e inferior (hastes) e outliers (pontos fora do boxplot), onde: x1= Precipitação total em Novembro e Dezembro (mm), x2= Temperatura média em Julho (ºC), x3= Precipitação total em Julho (mm), x4= Radiação em Julho (mm de álcool), x5= Rendimento médio de colheita (quintais/hectare).

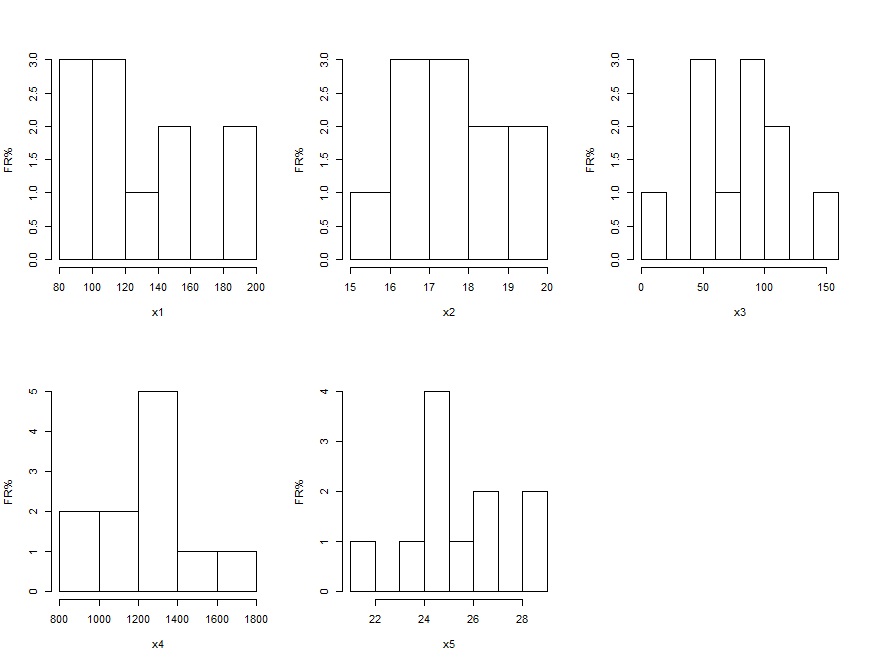


Figura 2. Histogramas com a distribuição das frequências das cinco variáveis amostradas numa exploração agrícola da Bélgica na década de 20. A base de cada retângulo é representada por uma classe de valores separadas pela regra de Sturges (K = 1 + 3,3 log(n)). Cada histograma representa uma variável amostrada, onde: x1= Precipitação total em Novembro e Dezembro (mm), x2= Temperatura média em Julho (ºC), x3= Precipitação total em Julho (mm), x4= Radiação em Julho (mm de álcool), x5= Rendimento médio de colheita (quintais/hectare)

**d) Construa diagramas de dispersão comparando as variáveis duas a duas. Compare os resultados obtidos nesses diagramas com a matriz de covariância e a matriz de correlação**.

A partir do diagrama de dispersão (figura 3) é possível observar certa tendência direcional nas “nuvens de pontos” para as variáveis que apresentaram correlação significativa (tabela 3). Essa tendência direcional também é visível na tabela de covariâncias (tabela 2), porém está não apresenta os mesmos “pesos” que a correlação de Pearson, não sendo possível, portanto, avaliar qual é o grau de associação entre as variáveis, apenar a direção que é tomada entre os pontos de duas variáveis.

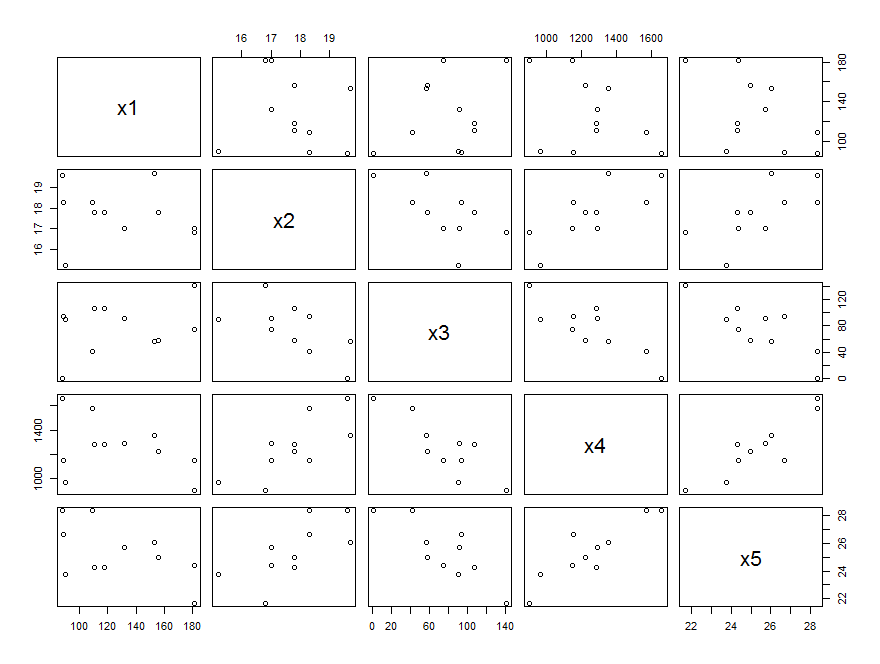


Figura 3. Digrama de dispersão para as cinco variáveis amostradas numa exploração agrícola da Bélgica na década de 20, onde: x1= Precipitação total em Novembro e Dezembro (mm), x2= Temperatura média em Julho (ºC), x3= Precipitação total em Julho (mm), x4= Radiação em Julho (mm de álcool), x5= Rendimento médio de colheita (quintais/hectare).

**e) Avalie se podemos considerar que cada variável tem distribuição normal de probabilidade, com 5% de probabilidade.**

Com a finalidade de testar se as variáveis amostradas apresentam distribuição normal univariada foi aplicado o teste de Shapiro-wilk (W) considerando um nível de significância (α) de 5%. A hipótese testada foi a de que os dados apresentam distribuição normal (H0). Os resultados são apresentados na tabela 4.

Tabela 4. Resultado do teste de normalidade de Shapiro-wilk (α = 5%) para as variáveis amostradas numa exploração agrícola da Bélgica na década de 20, onde: W = Resultado do teste de Shapiro-wilk, P = P-valor, x1= Precipitação total em Novembro e Dezembro (mm), x2= Temperatura média em Julho (ºC), x3= Precipitação total em Julho (mm), x4= Radiação em Julho (mm de álcool), x5= Rendimento médio de colheita (quintais/hectare).

|  |  |  |  |
| --- | --- | --- | --- |
| Variável | W | P | Normalidade? |
| x1 | 0,70 | 0,16 | Sim |
| x2 | 0,94 | 0,55 | Sim |
| x3 | 0,97 | 0,85 | Sim |
| x4 | 0,95 | 0,68 | Sim |
| x5 | 0,94 | 0,63 | Sim |

**f) Avalie se podemos considerar o conjunto de variáveis tem distribuição normal de probabilidade multivariada, com 5% de probabilidade**.

Com a finalidade de avaliar a distribuição normal multivariada das variáveis amostradas foi aplicado o teste de Shapiro-wilk multivariado considerando um nível de significância de 5%. Segundo o teste, os dados não apresentam distribuição normal multivariada (w=0,67, p= <0,001).

**3. Para os dados abaixo e assumindo que o conjunto de variáveis respostas tem distribuição normal de probabilidade multivariada, construa a MANOVA e interprete os resultados obtidos usando alfa igual a 0,05. Sendo que:**

**Fator: Sexo, com níveis: macho e fêmea**

**Variável 1 (X1): Comprimento**

**Variável 2 (X2): Largura**

**Variável 3 (X3): Peso**

Com a finalidade de avaliar a influência do fator sexo (níveis: Fêmea e Macho) sobre as variáveis respostas comprimento, largura e peso, foi realizada uma análise do tipo MANOVA com nível de significa de 5%. A hipótese testada foi a de que os valores das variáveis resposta são iguais entre os níveis de sexo (H0), tendo como hipótese alternativa que as variáveis resposta são diferentes entre os níveis de sexo (HA). Não houve diferença significativa entre os níveis do fator “sexo” (níveis = Feminino e Masculino) para as variáveis comprimento, largura e peso (F= 1,28, df=3, p=0,29), portanto o sexo não exerce influência sob estas variáveis. O Box-plot (figura 4) ilustra a proximidade da variação dos dados e de suas respectivas medianas, isso justifica sua igualdade estatística, uma vez que não variam em função do sexo.

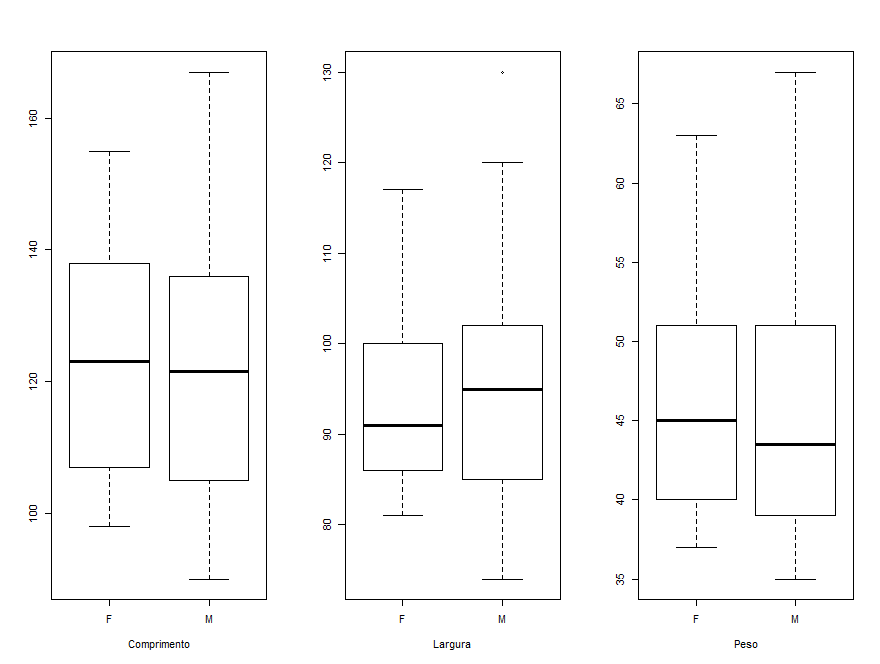


Figura 4. Box-plot das variáveis amostradas com separação pelo fator sexo (níveis: Feminino e masculino). O box-plot é composto por mediana (traço mais escuro), 1º e 3º quartil (caixa), valores dos limites superior e inferior (hastes) e outliers (pontos fora do boxplot).