**IBI5086 - 2º Sem/2023 – Análise Descritiva Multivariada**

Considere os dados “Iris” (Fisher, RA, 1936. The use of multiple measurements in taxonomic problems. Annals of Eugenics 7, Part II: 179–188), que contém medidas do comprimento e largura da pétala e da sépala de 150 flores de íris, sendo, 50 observações de cada uma das espécies setosa, versicolor e virginica.



**1**. **Discuta a estrutura destes dados**: unidades amostrais (qual é o tamanho da amostra, “n”? dados estratificados?), variáveis resposta (quantas, “p”? Todas quantitativas?), caracterize a dimensionalidade, possíveis fatores sob estudo e interesse da pesquisa.

**2.** Usando recursos do R realize uma **análise descritiva multivariada** desses dados:

**Vetor de Médias (Centroide)** para a amostra total e por grupo.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Total (n=150)** | **G1****Setosa (n=50)** | **G2****Versicolor(n=50)** | **G3****Virginica (n=50)** |
| Petal.Length Y1 |  |  |  |  |
| Sepal.Length Y2 |  |  |  |  |
| Petal.Width Y3 |  |  |  |  |
| Sepal.Width Y4 |  |  |  |  |

Qual variável tem maior e menor média?

Qual flor de íris (espécie) parece ser “maior” que as demais?

**Matriz de Covariância e de Correlação para os dados totais, independentemente de espécie) ST (**triangular superior) **e RT** (triangular inferior)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **RT | ST** | Y1 | Y2 | Y3 | Y4 |
| Y1 |  |  |  |  |
| Y2 |  |  |  |  |
| Y3 |  |  |  |  |
| Y4 |  |  |  |  |

Calcule a variância total e a variância generalizada:

tr(S)=\_\_\_\_\_\_\_\_\_\_\_\_ |S|=\_\_\_\_\_\_\_\_\_\_\_\_

tr(R)=\_\_\_\_\_\_\_\_\_\_\_\_ |R|=\_\_\_\_\_\_\_\_\_\_\_\_

Qual variável tem maior variabilidade? E qual tem a menor variabilidade?

Quais variáveis estão mais correlacionadas?

**3. Análise por espécie: Matriz de Covariância por Espécie**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Setosa: S1** | Y1 | Y2 | Y3 | Y4 |
| Y1 |  |  |  |  |
| Y2 |  |  |  |  |
| Y3 |  |  |  |  |
| Y4 |  |  |  |  |

tr(S)=\_\_\_\_\_\_\_\_\_\_\_\_ |S|=\_\_\_\_\_\_\_\_\_\_\_\_ tr(R)=\_\_\_\_\_\_\_\_\_\_\_\_ |R|=\_\_\_\_\_\_\_\_\_\_\_\_

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Versicolor: S2** | Y1 | Y2 | Y3 | Y4 |
| Y1 |  |  |  |  |
| Y2 |  |  |  |  |
| Y3 |  |  |  |  |
| Y4 |  |  |  |  |

tr(S)=\_\_\_\_\_\_\_\_\_\_\_\_ |S|=\_\_\_\_\_\_\_\_\_\_\_\_ tr(R)=\_\_\_\_\_\_\_\_\_\_\_\_ |R|=\_\_\_\_\_\_\_\_\_\_\_\_

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Virginica: S3** | Y1 | Y2 | Y3 | Y4 |
| Y1 |  |  |  |  |
| Y2 |  |  |  |  |
| Y3 |  |  |  |  |
| Y4 |  |  |  |  |

tr(S)=\_\_\_\_\_\_\_\_\_\_\_\_ |S|=\_\_\_\_\_\_\_\_\_\_\_\_ tr(R)=\_\_\_\_\_\_\_\_\_\_\_\_ |R|=\_\_\_\_\_\_\_\_\_\_\_\_

Compare as espécies relativamente ao padrão de covariação entre as variáveis.

**4. Matrizes de Distância**

Considere as 6 primeiras flores do banco de dados:

Apresente a Matriz de Distâncias Euclidiana entre essas flores (n=6):

Apresente a Matriz de Distâncias de Pearson (Euclidiana Padronizada) (n=6):

|  |  |  |
| --- | --- | --- |
|  | Dist. Euclidiana | Dist. Pearson |
| Quais flores são mais parecidas? |  |  |
| Quais flores são mais diferentes? |  |  |

5. Há observações atípicas? Considere a análise independentemente de espécie.

*Outlier* univariado: Construa o boxplot para cada variável (p=1) e comente.

Outilier Bivariado: Construa o Bivbox (considere pares de variáveis, p=2).

 Outliers Multivariados (p=4): Obtenha o gráfico de dispersão com a distância de Mahalanobis das observações ao centróide. A padronização dos dados influencia a análise de observações atípicas?

**5**. Considere o conjunto total dos dados (n=150). Obtenha **a decomposição espectral da matriz de covariância (S):**

**Autovalores de S:**

**Autovetores de S**:

Defina os dois primeiros Componentes Principais obtidos de **S**:

CP1 =

CP2=

Obtenha o escore do CP1 e do CP2 para uma das flores do bd:

Quanto da variância total dos **dados originais** estes 2 componentes explicam?

Como as **variáveis (originais)** contribuem para a formação destes componentes?

**6.** Considere o conjunto total dos dados (n=150). Obtenha a **decomposição espectral da matriz de correlação (R)**:

**Autovalores de R:**

**Autovetores de R**:

Defina os dois primeiros Componentes Principais obtidos de R:

CP1 =

CP2=

Obtenha o escore desses CP1 e CP2 para uma das flores:

Quanto da variância total dos **dados padronizados** estes 2 componentes explicam?

Como as **variáveis (padronizadas)** contribuem para a formação destes componentes?

7. Os gráficos abaixo representam as 150 flores de acordo com seus escores nos dois primeiros Componentes Principais (CP1 e CP2), bem como nas duas primeiras Coordenadas Principais (CoP1 e CoP2), considerando os dados originais (Y) e padronizados (Y\*). Comente e compare os resultados dessas análises? Como os CP e CoP são obtidos dos dados? Na redução de dimensionalidade, como escolher uma ou outra análise?



8. A tabela a seguir mostra a distribuição das flores de cada espécie de acordo com uma categoria de tamanho total das quatro variáveis avaliadas (Tam1<Tam2<Tam3<Tam4). Represente o padrão de associação entre essas variáveis por meio de um gráfico Biplot obtido da Análise de Correspondência. Essa representação explica quanto da inércia total dos dados nesta tabela?

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Tam1 | Tam2 | Tam3 | Tam4 | Total |
| Setosa | 39 | 11 | 0 | 0 | 50 |
| Versicolor | 0 | 25 | 22 | 3 | 50 |
| Virginica | 0 | 1 | 14 | 35 | 50 |