MAE5776 - 1º Sem/2022

Análise Descritiva Multivariada

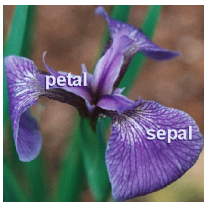
Alunos:

Fernando F. Paulos Vieira - nº USP: 13492870

Leandro Alves da Silva - nº USP: 11868023

Thiago Ferreira Miranda - nº USP: 11925711

Considere os dados “Iris” (Fisher, RA, 1936. The use of multiple measurements in taxonomic problems. Annals of Eugenics 7, Part II: 179–188), com 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: variáveis resposta (quantas? quantitativas?), possíveis fatores sob estudo, tamanho amostral. Que objetivos poderiam ser de interesse no estudo?

R:

O banco de dados apresenta 1 variável resposta categórica, denominada Species e, 4 variáveis quantitativas (Petal.Length, Sepal.Length, Petal. Width, Sepal. Width).

Como possíveis fatores sob estudo temos as dimensões observadas de cada uma das variáveis quantitativas, características especificas por espécie e/ou tendências, bem como nível de dependência apresentadas por determinada variável (a depender ou não da espécie). Isso ao considerar um tamanho amostral de 165 observações.

Candidatados a objetivos de interesse no estudo poderíamos citar:

- Identificar quais espécie apresentam maiores e menores dimensionais;
- Verificar se existe correlação entre as dimensões de pétala e sépala de uma mesma espécie ou independente da espécie;
- Identificar o grau de pureza dos clusters, ou seja, se os grupos possuem “limites” com separações claras e bem definidas;
- E, quais variáveis possuem maior e menor variabilidade e se há diferença na análise das observações totais ou por grupo (espécie)

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

Neste relatório utilizaremos a base iris com os dados incrementais, onde foram simuladas 15 novas observações, 5 para cada espécie de flor. Como incremento, levamos em conta o vetor de médias e a matriz de covariância das 150 observações originais.

```
iris_num <- iris %>%
  select(Petal.Length,Sepal.Length,Petal.Width,Sepal.Width)

iris_mean <- colMeans(iris_num)
iris_cov <- cov(iris_num)

set.seed(123)
novas_observacoes <- MASS::mvrnorm(15,iris_mean,iris_cov)
novas_observacoes <- novas_observacoes %>%
  data.frame(Species = c(rep("setosa",5),rep("versicolor",5),rep("virginica",5)))
```

Petal.Length	Sepal.Length	Petal.Width	Sepal.Width	Species
2.543975	5.990097	0.9157430	3.813350	setosa
3.273876	5.900967	1.0130766	3.207184	setosa
6.656210	6.242556	2.6117551	2.205698	setosa
3.898806	5.941656	1.2684931	3.482683	setosa
4.037466	5.656817	1.4472768	2.998068	setosa
6.903811	6.648000	2.5782377	2.502821	versicolor
4.598236	6.026570	1.6348887	2.990035	versicolor
1.612669	4.583564	0.3010011	2.895691	versicolor
2.705193	5.080312	0.5144576	2.930929	versicolor
3.001492	5.382918	0.8624294	2.835206	versicolor
6.155859	6.246619	1.8813859	2.196592	virginica
4.201111	6.490971	1.5817367	3.184926	virginica
4.467208	6.368252	1.2277131	2.860114	virginica
4.105594	5.198382	1.6398708	2.997509	virginica
2.713528	5.628687	0.9024688	3.817567	virginica

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

		G1	G2	G3
	Total (n=165)	Setosa (n=55)	Versicolor (n=55)	Virginica (n=55)
Petal.Length Y1	3.785303	1.7001879	4.214935	5.440787
Sepal.Length Y2	5.841736	5.0914926	5.900388	6.533326
Petal.Width Y3	1.213821	0.3555699	1.312564	1.973331
Sepal.Width Y4	3.051626	3.4019451	2.775540	2.977395

Qual variável tem maior e menor média?

R:

A variável Sepal.Length possui a maior média (ao considerarmos a amostra total e por grupo). E a variável Petal.Width possui a menor média (ao considerarmos a amostra total e por grupo).

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

R:

A espécie virginica parece ser maior, pois apresenta (na maior parte dos casos) maiores valores médios das dimensões observadas nas amostras por grupo, tendo em vista as 4 variáveis quantitativas avaliadas.

Matriz de Covariância e Correlação: S_T (triangular superior) e R_T (triangular inferior)

$R_T \mid S_T$	Y1	Y2	Y3	Y4
Y1	3.0424040	1.2131628	1.2651379	-0.3452652
Y2	0.8619669	0.6510885	0.4921719	-0.0429298
Y3	0.9627368	0.8096082	0.5676018	-0.1269775
Y4	-0.4510417	-0.1212302	-0.3840407	0.1925996

Calcule a variância total e a variância generalizada:

tr(S) = 4.4536939 **|S| =** 0.0017801

tr(R) = 4 **|R| =** 0.0082203

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

R:

A variável Petal.Length(Y1) tem a maior variabilidade, enquanto a variável Sepal.Width(Y4) tem a menor variabilidade.

3. Quais variáveis estão mais correlacionadas?

Para o total das observações, independente da espécie, as variáveis Petal.Length(Y1) e Petal.Width(Y3) estão mais correlacionadas que as demais, com coeficiente de correlação de 0.96.

Análise por espécie:

Setosa S_1	Y1	Y2	Y3	Y4
Y1	0.7844795	0.2360529	0.3490775	-0.1183619
Y2	0.6107180	0.1904390	0.1111203	0.0634375
Y3	0.9650657	0.6235068	0.1667817	-0.0490908
Y4	-0.3294949	0.3584226	-0.2963832	0.1644921

tr(S) = 1.3061922 **|S| =** 0.0000671

tr(R) = 4 **|R| =** 0.0163641

Versicolor S_2	Y1	Y2	Y3	Y4
Y1	0.5337420	0.3035851	0.2133808	0.0512363
Y2	0.7547736	0.3031071	0.1104355	0.0679776
Y3	0.9156849	0.6288790	0.1017390	0.0271635
Y4	0.2307277	0.4062142	0.2801746	0.0923900

tr(S) = 1.0309781 **|S| =** 0.0000751

tr(R) = 4 **|R| =** 0.0494032

Virginica S_3	Y1	Y2	Y3	Y4
Y1	0.5140718	0.3596855	0.1332712	0.0085015
Y2	0.7741842	0.4198880	0.0764723	0.0746787
Y3	0.5666453	0.3597695	0.1076035	0.0277114
Y4	0.0342564	0.3329582	0.2440645	0.1198064

tr(S) = 1.1613698 **|S| =** 0.0004748

tr(R) = 4 **|R| =** 0.1706206

Matrizes de Distância

Considere as 6 primeiras flores:

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

	X1	X2	X3	X4	X5	X6
X1	0.0000000	0.5385165	0.509902	0.6480741	0.1414214	0.6164414
X2	0.5385165	0.0000000	0.300000	0.3316625	0.6082763	1.0908712
X3	0.5099020	0.3000000	0.000000	0.2449490	0.5099020	1.0862780
X4	0.6480741	0.3316625	0.244949	0.0000000	0.6480741	1.1661904
X5	0.1414214	0.6082763	0.509902	0.6480741	0.0000000	0.6164414
X6	0.6164414	1.0908712	1.086278	1.1661904	0.6164414	0.0000000

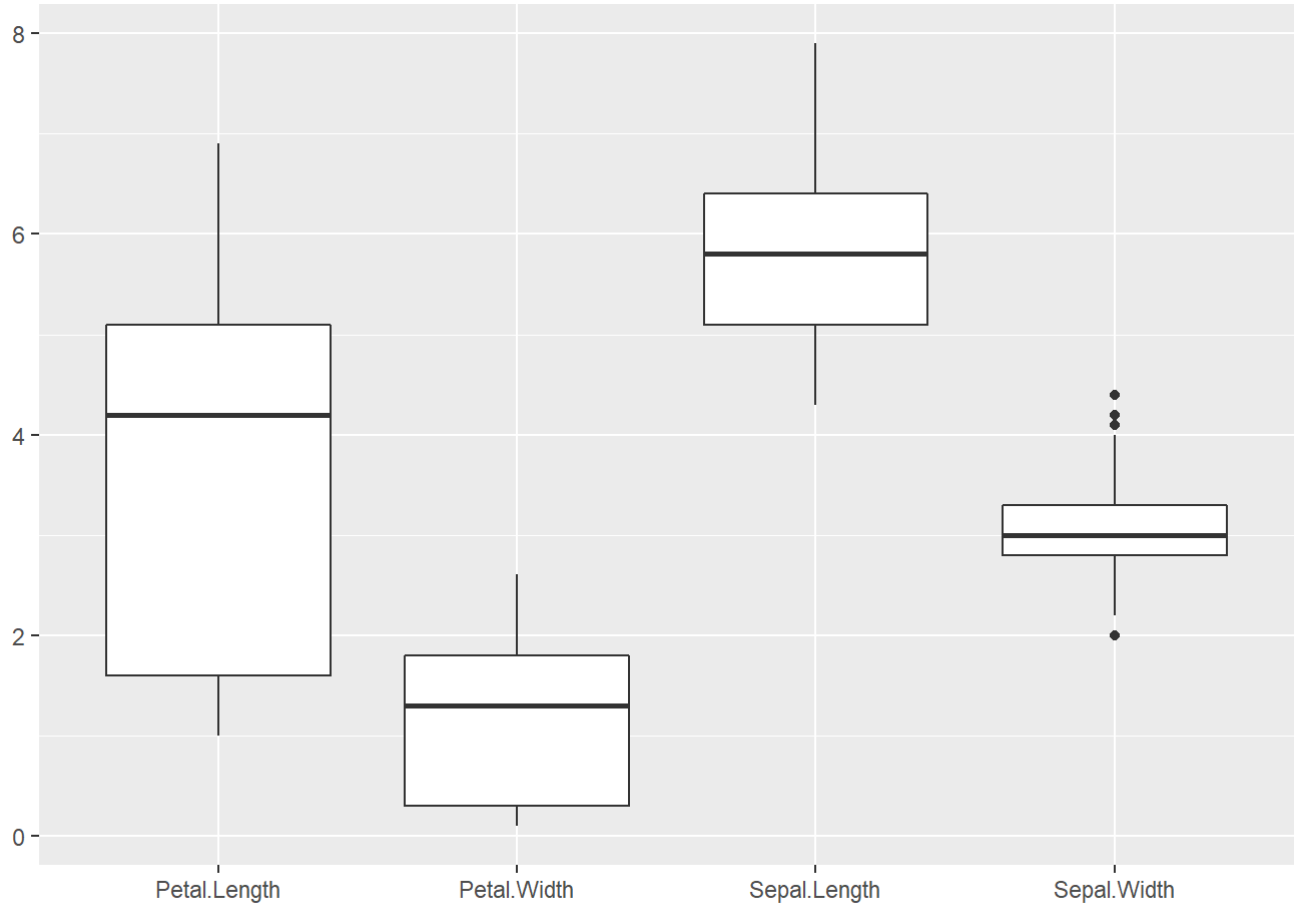
Apresente a Matriz de Distâncias de Pearson (Euclidiana Padronizada) (n=6):

	X1	X2	X3	X4	X5	X6
X1	0.0000000	1.1659610	0.8463568	1.1036295	0.2593840	1.0339356
X2	1.1659610	0.0000000	0.5219264	0.4398160	1.3727787	2.1655581
X3	0.8463568	0.5219264	0.0000000	0.2835976	0.9860303	1.8492662
X4	1.1036295	0.4398160	0.2835976	0.0000000	1.2438081	2.0951239
X5	0.2593840	1.3727787	0.9860303	1.2438081	0.0000000	0.9017132
X6	1.0339356	2.1655581	1.8492662	2.0951239	0.9017132	0.0000000

	Dist. Euclidiana	Dist. Pearson
Quais flores são mais parecidas?	X1 e X5 (0.1414214)	X1 e X5 (0.2593840)
Quais flores são mais diferentes?	X4 e X6 (1.1661904)	X2 e X6 (2.1655581)

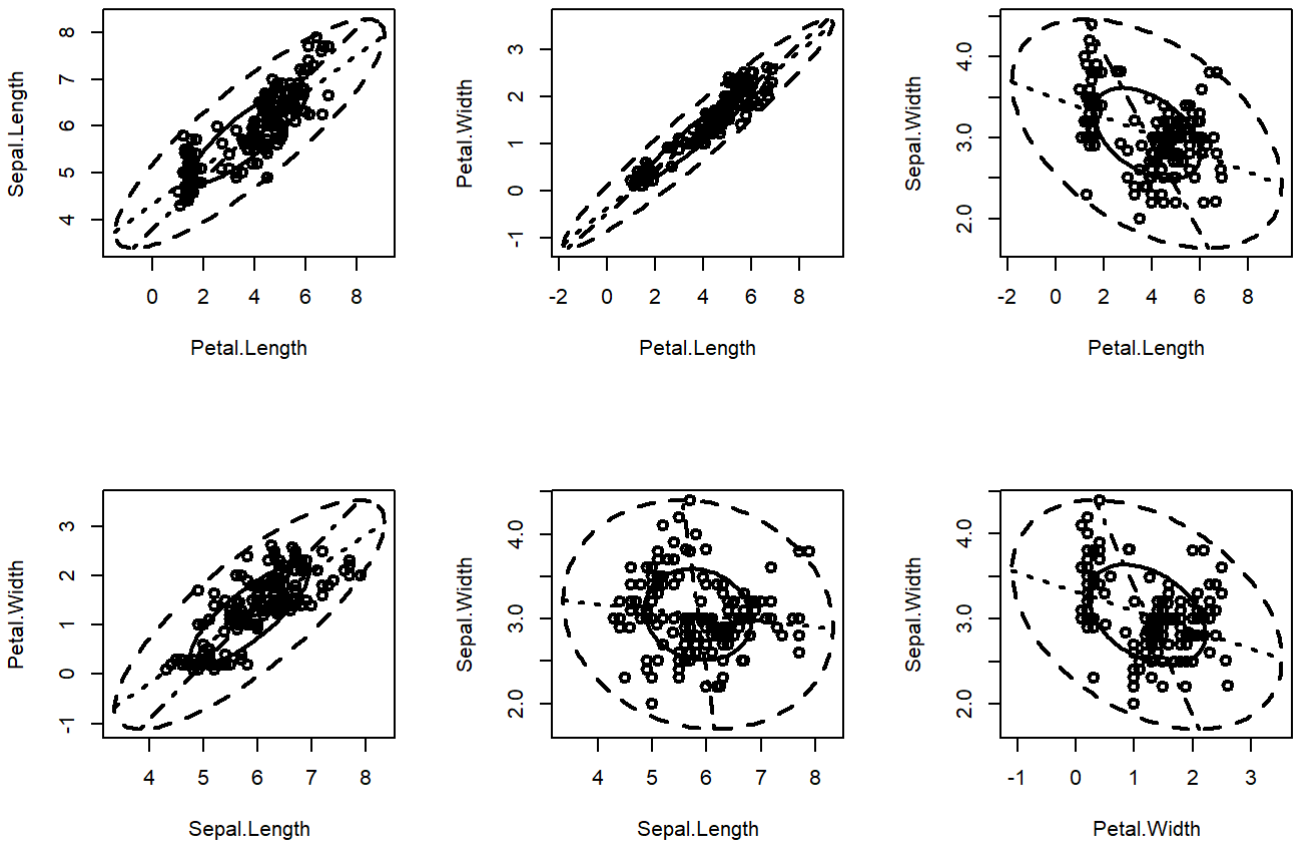
4. Há observações atípicas? Considere a análise independente de espécie. Outlier univariado: construa o boxplot para cada variável

R:



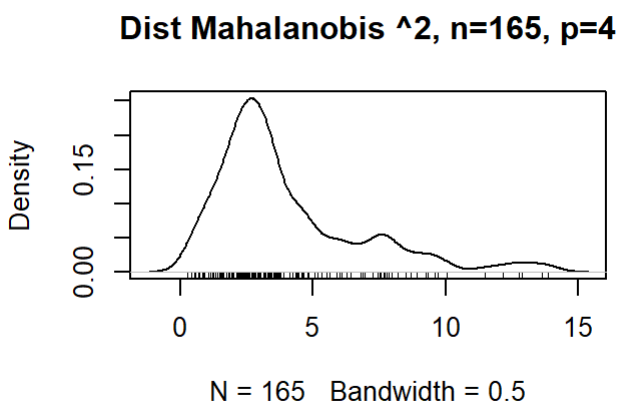
Sim, há observações atípicas, pois pode-se observar que o boxplot da variável Sepal.Width apresenta 4 outliers (sendo estes as observações 16, 33, 34 e 61).

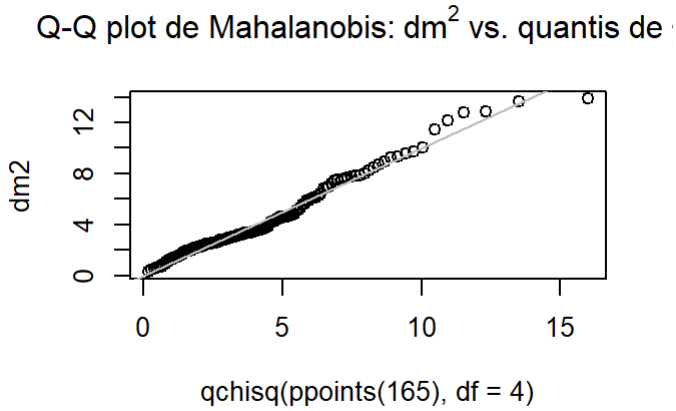
Outilier Bivariado: Construir o Bivbox (considere pares de variáveis)

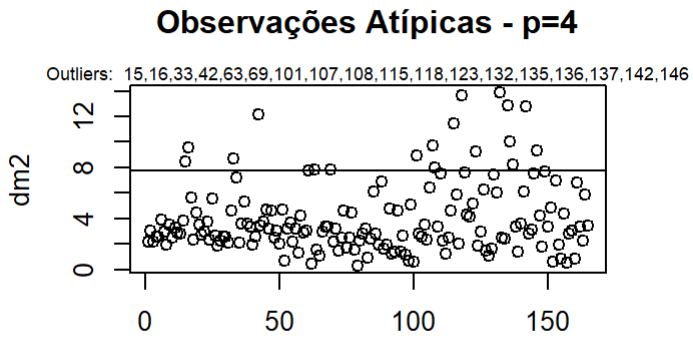


Outliers Multivariados (p=4): Obtenha o gráfico com a distância de Mahalanobis das observações ao centróide. A padronização dos dados influencia a análise de observações atípicas?

R:







A padronização dos dados (utilizando os critérios para obtenção da distância de Mahalanobis) influencia a análise de observações atípicas, anteriormente ao avaliar somente com o critério do boxplot de modo univariado foram sinalizadas 4 observações atípicas, no entanto ao considerar uma análise multivariada (por meio dos critérios da distância de Mahalanobis) agora temos 18 observações atípicas evidenciadas.