

Análise de Variância

A técnica de análise de variância permite fazer a comparação de múltiplas médias por meio de um teste de hipóteses. A construção do teste se dá pela decomposição da variabilidade em componentes referentes aos diferentes grupos e à variabilidade inerente das observações dentro dos grupos.

Considera-se o seguinte modelo:

$$y_{ij} = \mu + \tau_i + e_{ij}$$

Nesse modelo a variável resposta Y é descrita pela média geral μ mais um efeito de grupo τ mais um erro aleatório. O índice $i = 1, 2, \dots, a$ representa os diferentes grupos e o índice $j = 1, 2, \dots, n$ representa os diferentes elementos dentro de cada grupo. Considerando $\bar{y}_{..} = \sum_{i=1}^a \sum_{j=1}^n \frac{y_{ij}}{n}$ a média geral e $\bar{y}_{i.} = \sum_{j=1}^n \frac{y_{ij}}{n}$ a média do grupo i , a decomposição da variabilidade se dá na forma:

$$\sum_{i=1}^a \sum_{j=1}^n (y_{ij} - \bar{y}_{..}) = \sum_{i=1}^a \sum_{j=1}^n (\bar{y}_{i.} - \bar{y}_{..}) + \sum_{i=1}^a \sum_{j=1}^n (y_{ij} - \bar{y}_{i.})$$

$$SQTOT = SQTRAT + SQRES$$

Partido da suposição que os erros e_{ij} são independentes e idênticamente distribuídos (*i.i.d*) com distribuição $Normal(\mu = 0, \sigma^2)$ pode-se testar as seguintes hipóteses:

- $H_0 : \tau_1 = \tau_2 = \dots = \tau_a = 0$
- $H_a : \exists \tau_i \neq 0$

Usando a seguinte estatística de teste:

$$F = \frac{\frac{SQTRAT}{(a-1)}}{\frac{SQRES}{(an-a)}}$$

que sob hipótese nula tem distribuição $F((a-1), (an-a))$

Exemplo

O conjunto de dados *iris* disponível no R apresenta medidas de comprimento e largura de pétalas e sépalas e classificação de 3 espécies de um tipo de flor. Nesse exemplo de aplicação a técnica de análise de variância será utilizada para avaliar se existem diferenças entre as médias de comprimento das pétalas das flores das diferentes espécies.

```
modelo <- aov(iris$Petal.Length~iris$Species)
summary(modelo)
```

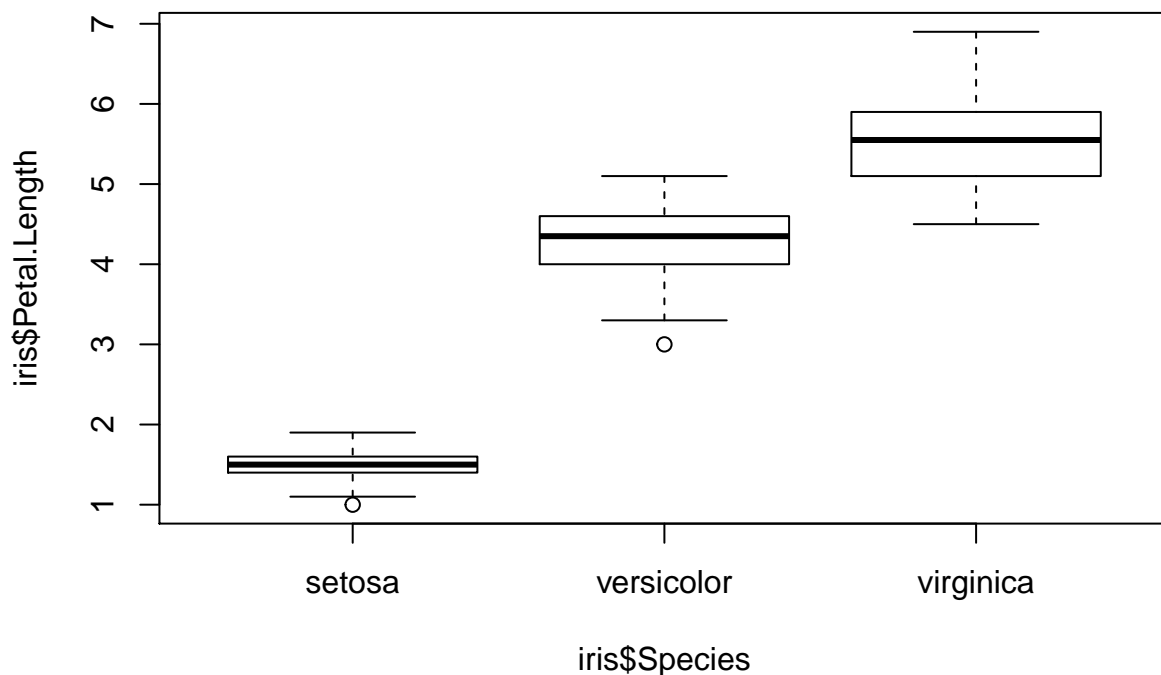
```
##              Df Sum Sq Mean Sq F value Pr(>F)
## iris$Species    2  437.1   218.55    1180 <2e-16 ***
## Residuals     147    27.2     0.19
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

O resultado acima é uma representação da tabela de análise de variância, que apresenta as principais informações consideradas para realizar o teste, sendo elas:

- As somas de quadrado, que são os componentes da decomposição da variabilidade
- Os quadrados médias, que são as somas de quadrados divididas pelos seus graus de liberdade
- A estatística de teste F
- O p-valor para o teste

Como se pode perceber nesse teste o p-valor é muito pequeno, indicando a rejeição da hipótese de igualdade entre as médias. O gráfico abaixo ilustra a distribuição dos comprimentos de pétalas para as diferentes espécies, e pode-se notar a evidente diferença entre as espécies, confirmando o resultado do teste.

```
boxplot(iris$Petal.Length~iris$Species)
```



Os resultados de análise de variância não permitem identificar entre quais grupos as médias são distintas, e uma das possíveis técnicas para identificar os pares de grupos que tenham médias diferentes é o teste de Tukey, que tem por propriedade controlar o erro global do experimento.

```
TukeyHSD(modelo)
```

```
## Tukey multiple comparisons of means
## 95% family-wise confidence level
##
## Fit: aov(formula = iris$Petal.Length ~ iris$Species)
##
## $`iris$Species`
##          diff      lwr      upr p adj
## versicolor-setosa  2.798 2.59422 3.00178    0
## virginica-setosa   4.090 3.88622 4.29378    0
## virginica-versicolor 1.292 1.08822 1.49578    0
```

Os resultados desse teste confirmam a impressão visual de que todas as espécies tem médias distintas quando comparadas duas a duas.

É importante perceber que as conclusões da análise de variância será valida somente se as suposições adotadas forem adequadas. A seguir será realizado um teste de normalidade para os resíduos.

```
shapiro.test(modelo$residuals)
```

```
##  
##  Shapiro-Wilk normality test  
##  
## data:  modelo$residuals  
## W = 0.98108, p-value = 0.03676
```

No teste de Shapiro-Wilk a hipótese nula é de que os dados tenham distribuição Normal e pelo p-valor obtido para os resíduos do modelo considerado, deve-se rejeitar a hipótese de normalidade e portanto a técnica de análise de variância para esses dados não se mostra adequada.

Uma alternativa à Análise de Variância é o teste de Kruskal-Wallis, que é um teste dito não paramétrico pois não faz suposições sobre a forma ou os parâmetros das distribuições dos dados em teste. As hipóteses em teste são semelhantes às hipóteses consideradas na análise de variância.

```
kruskal.test(iris$Petal.Length, iris$Species)
```

```
##  
##  Kruskal-Wallis rank sum test  
##  
## data:  iris$Petal.Length and iris$Species  
## Kruskal-Wallis chi-squared = 130.41, df = 2, p-value < 2.2e-16
```

Considerando o pequeno valor observado para o p-valor do teste, deve-se rejeitar a hipótese de igualdade entre as médias.

Uma alternativa não paramétrica para o teste de Tukey seria o teste de Dunn