Estatística com Apoio Computacional

Análise de Variância

Universidade Estadual Vale do Acaraú – UVA

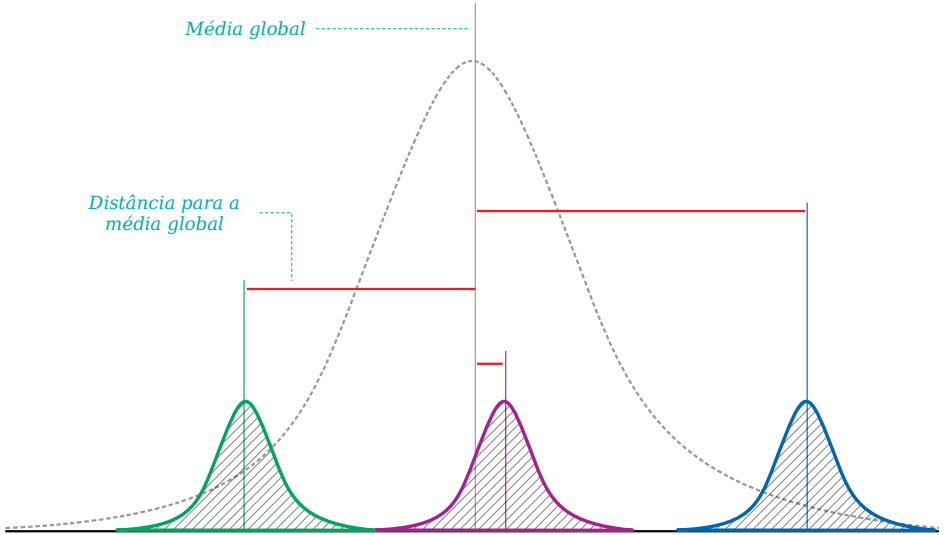
Paulo Regis Menezes Sousa

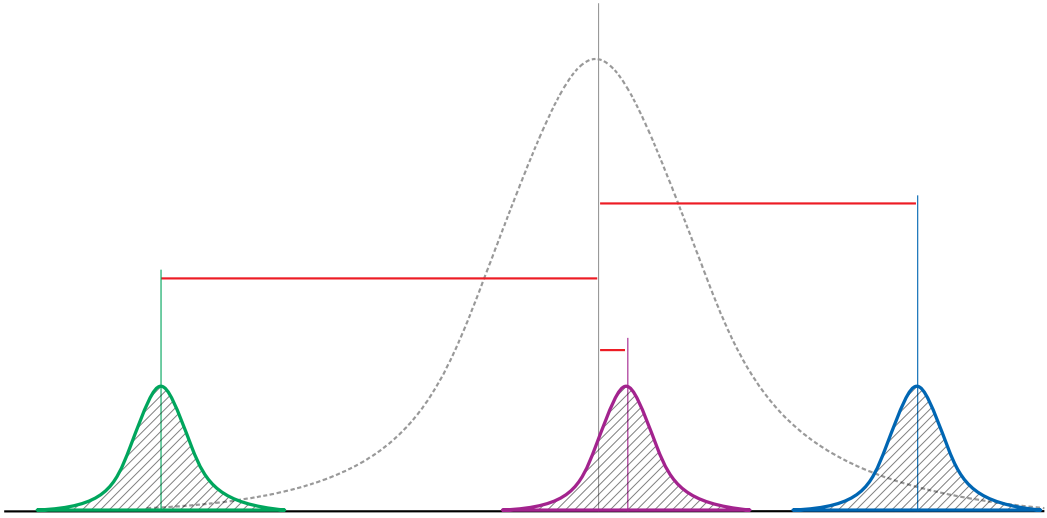
paulo_regis@uvanet.br

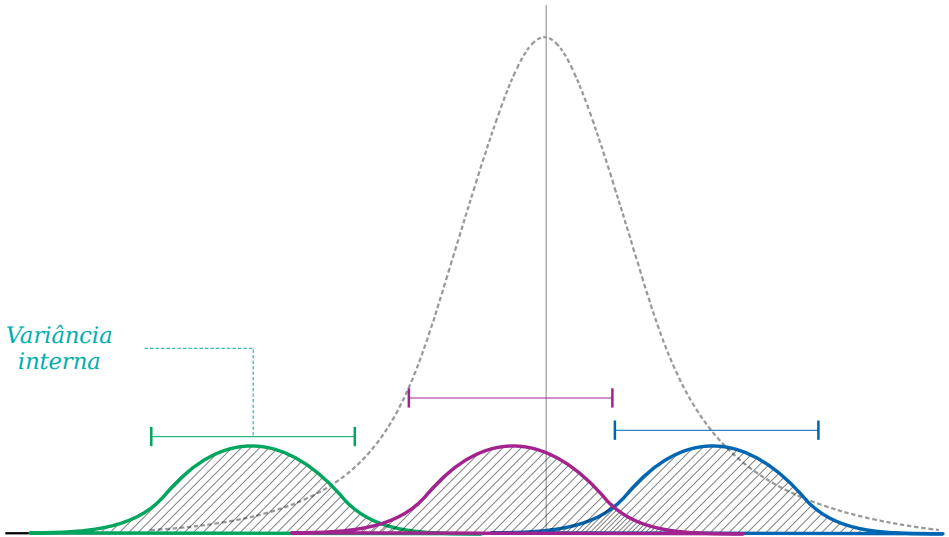
Análise de Variância

- Como funciona a análise de variância
- ANOVA no R
- DIC
- DBC

- Até então usamos o teste **t** para comparar médias, apenas entre duas amostras.
- Quando desejamos realizar uma comparação entre mais de duas médias o teste **t** não é mais uma opção razoável.
- A principal forma usada para realizar este tipo de comparação é através de uma **Análise de Variância** (ANOVA - **AN**alysis **O**f **VA**riance).
- Ela permite identificar e quantificar as variações ocorridas em um experimento, discriminando a parte da variação associada ao modelo pelo qual o experimento foi procedido da variação que se dá devido ao acaso.


Média global
Distância para a
média global




Variância
interna

- Análise de variância pode ser descrita como uma razão de *variabilidades*

$$\frac{\text{variabilidade entre as médias}}{\text{variabilidade dentro das distribuições}} \tag{1}$$

$$\frac{GRANDE}{pequena} = \text{Rejeitar } H_0 \tag{2}$$

$$\frac{similar}{similar} = \text{Falha em rejeitar } H_0 \tag{3}$$

$$\frac{pequena}{GRANDE} = \text{Falha em rejeitar } H_0 \tag{4}$$

- Suposições para a utilização da ANOVA:
  1. As amostras são independentes
  2. Dentro de cada amostra as observações são independentes.
  3. As observações são selecionadas de uma população na qual a variável resposta tem distribuição Normal com variâncias iguais.

- No R encontram-se diversos procedimentos para se executar a ANOVA:

Função	Descrição
`aov()`	para ANOVA, com erros normais e independentes
`lm()`	para regressão linear (*linear models*)
`glm()`	para ANOVA, com estrutura de erros especificada (*generalised linear models*)
`nlme()`	para modelos mistos (*nonlinear mixed-effects models*)
`nls()`	para modelos não lineares (*nonlinear least squares*)

Tabela: Alguns comandos importantes.

- O **DIC** (Delineamento Inteiramente Casualizado) trata de experimentos onde os dados não são pré-separados ou classificados em categorias mais conhecidas como blocos.
- A ANOVA, associada a esse tipo de experimento, é muitas vezes chamada *One Way ANOVA*.

Exemplo 1

Suponha que um experimentador coletou os seguintes dados a respeito de um experimento com quatro tratamentos:

trat1	trat2	trat3	trat4
25	31	22	33
26	25	26	29
20	28	28	31
23	27	25	34
21	24	29	28

```r
# Dados do experimento
res = c(25,31,22,33, 26,25,26,29, 20,28,28,31,
                     23,27,25,34, 21,24,29,28)

# Criando os nomes dos tratamentos na ordem correspondente
trat = factor( rep(paste0("tr",1:4), 5) )

# Fazendo a ANOVA
resultado = aov(res~trat)

# Para exibir o quadro da ANOVA
anova(resultado)
```

- No R encontram-se diversos procedimentos para se executar a ANOVA:

Modelo	**Fórmula**
DIC	`y~t` onde t é um fator
DBC	`y~t+b` onde t e b são fatores
Fatorial/DIC	`y~N*P` igual a `N+P+N:P`
Fatorial/DBC	`y~b+N*P` igual a `b+N+P+N:P`
Regressão linear simples	`y~x` onde x é uma variável exploratória
Regressão quadrática	`y~x+x2` onde $x2$ é um objeto `x2<-x^2`

Tabela: Modelos e suas usuais formulações para ANOVA no R

```
Analysis of Variance Table

Response: res
           Df  Sum Sq Mean Sq F value  Pr(>F)
trat        3  163.75  54.583  7.7976 0.001976 **
Residuals  16  112.00   7.000
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Temos aqui uma representação usual do quadro da ANOVA, com:
  - fontes de variação (`trat` e `Residuals`),
  - graus de liberdade (`Df`, do inglês *degrees of freedom*),
  - somas de quadrados (`Sum Sq`),
  - quadrados médios (`Mean Sq`),
  - valor do F calculado (`F value`) e
  - significância do teste, ou *p-value* ($Pr(> F)$).

- Uma forma de verificar se o objeto que descreve os tratamentos foi criado corretamente é conferindo os **graus de liberdade** no quadro na anova.
- Neste exemplo, temos quatro tratamentos e, consequentemente, três graus de liberdade $(k-1)$.
- Caso o objeto `trat` não fosse um **fator**, teríamos apenas um grau de liberdade, indicando que não procedemos à análise de forma correta.
- A ANOVA pode ser interpretada da seguinte maneira:
  - como o `p-value` (0,001976) foi menor que 1%, então existe diferença significativa entre as médias de pelo menos dois tratamentos, a 1% de significância.

Questão 1

Afirma-se que o número de carros roubados por dia não depende da região da cidade. Para verificar essa afirmação, a cidade foi dividida em quatro zonas e, durante 10 dias, foram registrados os carros roubados nas quatro zonas, conforme registrado na tabela seguinte. Verifique essa afirmação considerando o nível de significância de 5%.

Zona 1	Zona 2	Zona 3	Zona 4
12	12	10	13
15	11	12	15
14	13	14	14
12	18	12	15
15	15	11	17
18	14	13	14
12	13	10	13
14	12	12	14
12	11	13	15
11	10	11	16

- O DBC (Delineamento em Blocos Casualizados) abrange os três princípios básicos da experimentação: repetição, casualização, e o controle local.
- Este delineamento é bastante utilizado quando há heterogeneidade nas condições experimentais.
- Neste caso, divide-se o material experimental, ou amostras, em blocos homogêneos, de forma a contemplar as diferenças entre os grupos.
- A ANOVA associada a este modelo de experimento é também conhecida como *Two Way ANOVA*.

Exemplo 2

Suponha que uma Nutricionista elaborou 4 dietas e quer aplicá-las em 20 pessoas a fim de testar suas eficiências quanto à perda de peso. Porém ela notou que entre essas 20 pessoas existem 5 grupos de faixas iniciais de peso. Então, para aumentar a eficácia do teste ela separou os 20 indivíduos em 5 grupos de faixas de peso. Então ela tem:

Dietas: dieta 1, dieta 2, dieta 3, dieta 4;
Grupos: peso A, peso B, peso C, peso D, peso E.

A tabela seguinte resume o valor da perda de peso, arredondados em quilogramas, de cada indivíduo. Veja:

Exemplo 2

	dieta 1	dieta 2	dieta 3	dieta 4
peso A	2	5	2	5
peso B	3	7	4	3
peso C	2	6	5	4
peso D	4	5	1	3
peso E	2	5	4	4

A Nutricionista deseja determinar se existe diferença significativa entre as dietas a um nível de significância de 5%.

```r
# Criando o vetor de dados, o de tratamentos
# e o de blocos, respectivamente
dad  = c(2, 5, 2, 5,
         3, 7, 4, 3,
         2, 6, 5, 4,
         4, 5, 1, 3,
         2, 5, 4, 4)
bloc = gl( 5, 4, label = paste("peso", LETTERS[1:5]) )
trat = factor(rep(paste("dieta",1:4), 5))

# Criar um data.frame contendo todos os dados
tabela = data.frame(blocos = bloc,
                    tratamentos = trat,
                    dados = dad)
tabela
```

```
16 # ANOVA
17 resultado= aov(dados~tratamentos+blocos, tabela) # y~t+b
18 resultado
19
20 # Gera a tabela de análise de variância
21 anova(resultado)
```

- Para que possamos utilizar a ANOVA precisamos que algumas premissas sejam atendidas:
  I. Aleatoriedade e independência das amostras.
  II. Distribuição normal dos resíduos.
  III. Homogeneidade das variâncias.
- Dois testes muito utilizados em estatística são os testes de Shapiro-Wilk e de Bartlett para testar, respectivamente, a normalidade e a homogeneidade.

```r
# Teste de pressupostos da anovaProgenie
dados = read.csv("data/base-progenie.csv", header = T, sep = ";")
dados
anovaProgenie = aov(volume~progenie, data = dados)
```

- Usando Shapiro-Wilk para verificar a normalidade.

```
1 # Teste de normalidade dos resíduos da ANOVA
2 shapiro.test(resid(anovaProgenie))
```

```
Shapiro-Wilk normality test

data:  resid(anova)
W = 0.96097, p-value = 0.3279
```

- A hipótese nula do teste é a de que os dados seguem uma distribuição normal. Como o p-valor é superior ao limite de 5%, podemos aceitar a hipótese nula e considerar nossos dados normais.

- Utilizando o teste de Bartlett para verificar se há homogeneidade de variância.

```
# Teste de homogeneidade
bartlett.test(volume~progenie, data = dados)
```

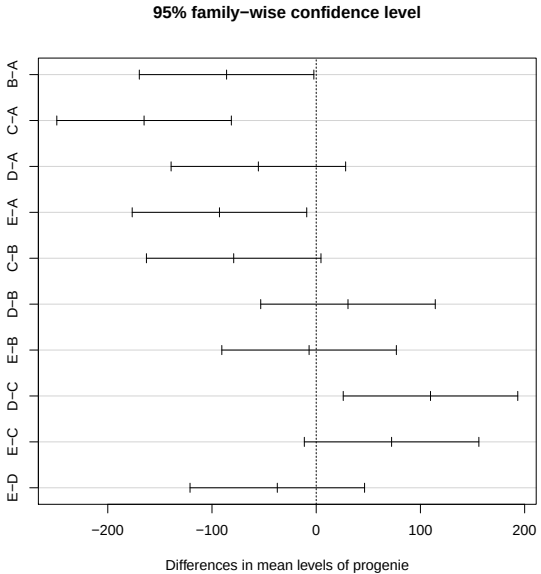
```
Bartlett test of homogeneity of variances

data:  volume by progenie
Bartletts K-squared = 9.2829, df = 4, p-value = 0.0544
```

- Como o p-valor é maior que 5% não temos evidência significativa para rejeitar a hipótese nula de homogeneidade, ou seja, nossos dados tem homogeneidade de variância.

- A ANOVA resultou em um p-valor menor que 5%, portanto, temos evidências de que ao menos um tratamento se diferencia dos demais.
- Agora queremos saber quem é este tratamento discrepante. Ou melhor, queremos poder comparar os tratamentos entre si e verificar quais são estatisticamente iguais ou diferentes.
- Para esta abordagem existem alguns testes de médias e cada um tem uma particularidade, mas de longe o mais utilizado é o de Tukey.
- A forma mais fácil de usar o teste de Tukey no R é empregando o comando `TukeyHSD()`, do pacote `stats`, que já vem na instalação básica do R e é carregado sempre que o programa é iniciado.

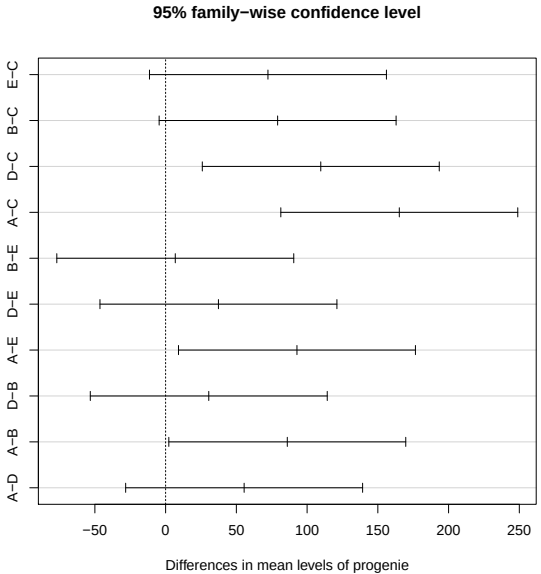
```r
# Teste de Tukey
resultado1 = TukeyHSD(anovaProgenie, "progenie")
resultado1
pdf("grafico-tukey-01.pdf")
plot(resultado1)
```

```
             diff         lwr         upr     p adj
B-A   -86.000000 -169.73866   -2.261338 0.0420474
C-A  -165.166667 -248.90533  -81.428004 0.0000447
D-A   -55.500000 -139.23866   28.238662 0.3202440
E-A   -92.833333 -176.57200   -9.094671 0.0245274
C-B   -79.166667 -162.90533    4.571996 0.0703394
D-B    30.500000  -53.23866  114.238662 0.8200574
E-B    -6.833333  -90.57200   76.905329 0.9992174
D-C   109.666667   25.92800  193.405329 0.0060232
E-C    72.333333  -11.40533  156.071996 0.1142315
E-D   -37.333333 -121.07200   46.405329 0.6880663
```


95% family−wise confidence level
B−A
C−A
D−A
E−A
C−B
D−B
E−B
D−C
E−C
E−D
−200
−100
0
100
200
Differences in mean levels of progenie

```
resultado2 = TukeyHSD(anovaProgenie, "progenie", ordered = T)
resultado2
pdf("grafico-tukey-02.pdf")
plot(resultado2)
dev.off()
```

```
          diff         lwr       upr      p adj
E-C  72.333333 -11.405329 156.0720 0.1142315
B-C  79.166667  -4.571996 162.9053 0.0703394
D-C 109.666667  25.928004 193.4053 0.0060232
A-C 165.166667  81.428004 248.9053 0.0000447
B-E   6.833333 -76.905329  90.5720 0.9992174
D-E  37.333333 -46.405329 121.0720 0.6880663
A-E  92.833333   9.094671 176.5720 0.0245274
D-B  30.500000 -53.238662 114.2387 0.8200574
A-B  86.000000   2.261338 169.7387 0.0420474
A-D  55.500000 -28.238662 139.2387 0.3202440
```


95% family−wise confidence level
E−C
B−C
D−C
A−C
B−E
D−E
A−E
D−B
A−B
A−D
−50
0
50
100
150
200
250
Differences in mean levels of progenie