ANOVA
(Analysis of Variance)

Aplicação: Exemplos

- A eficiência de diversas marcas de remédios para o tratamento de uma mesma doença, o controle de pressão alta.
- O consumo em km/litro de um modelo de carro abastecido com combustíveis do mesmo tipo, porém de marcas diferentes.
- A eficiência de uma lavoura tratada com diferentes fertilizantes.
- O tempo de reação de uma pessoa em função de estímulo de luz de quatro cores diferentes.

O teste ANOVA foi desenvolvido para comparar médias

- Inicialmente, podemos usar teste t para comparar médias (dois a dois)
- O problema é que este procedimento aumenta o erro tipo I
- Além disso, não permite avaliar o efeito de várias variáveis independentes

O que nos diz a ANOVA?

Hipótese Nula :

As médias de todos os grupos são iguais

Hipótese Alternativa

Pelo menos um grupo possui média difernete

A ANOVA nos diz se existe ou não diferença das médias entre os grupos mas não diz quais deles são diferentes!

Análise de Variância

<u>Objectivo</u>: comparar medidas de localização para **mais do que dois grupos** de observações

Para analisar as diferenças na localização, recorre-se <u>a uma análise das variâncias</u> dos vários grupos, daí o nome ANOVA.

<u>ANOVA Paramétrica vs. Não Paramétrica</u>:

- <u>One-Way ANOVA</u>: (Análise de Variância **com um factor**)

  se os grupos são bem modelados por <u>distribuições Normais de igual variância</u>, comparamos as médias entre os grupos

- <u>Teste de Kruskal-Wallis</u>:

  usar quando os pressupostos do teste paramétrico não se verificarem, neste caso comparamos as medianas entre os grupos

Análise de Variância

1 Factor

As observações se dividem em **vários grupos** classificados através de um **só factor**.
Para **cada grupo** obtemos **uma amostra aleatória** de observações de uma variável Y

A experiência tem tantos **níveis** ou **efeitos** quantos **grupos** ou **tratamentos** distintos

k amostras independentes

	Nível 1	Nível 2	. . .	Nível k
	y_{11}	y_{21}		y_{k1}
	y_{12}	y_{22}		y_{k2}
	$\vdots$	$\vdots$	. . .	$\vdots$
	y_{1n1}	y_{2n2}		y_{knk}
médias:	$\bar{y}_{1.}$	$\bar{y}_{2.}$		$\bar{y}_{k.}$

$n=n_1+n_2+\ldots+n_k$

1ª Fase = Planeamento:
seleccionar os indivíduos (ou unidades que se vão dividir pelos grupos)

- **efeitos fixos**: os grupos são pré-determinados à partida
- **efeitos aleatórios**: os grupos são escolhidos aleatoriamente
- **planeamento equilibrado**: quando o número de observações de cada grupo é igual

Premissas da Anova

- As populações têm a mesma variância.

- As amostras são retiradas de populações com distribuição normal.

- As amostras são aleatórias e independentes.

Verificação das suposições da ANOVA

- As hipóteses são :
  - Teste de Normalidade

    $$H_0 : \varepsilon_{ij} \sim N(0,\sigma^2)$$

    $$H_1 : \varepsilon_{ij} \text{ não têm } N(0,\sigma^2)$$

  - Teste da Homogeneidade das Variâncias

    $$H_0 : \sigma_1^2 = \sigma_2^2 = \sigma_3^2$$

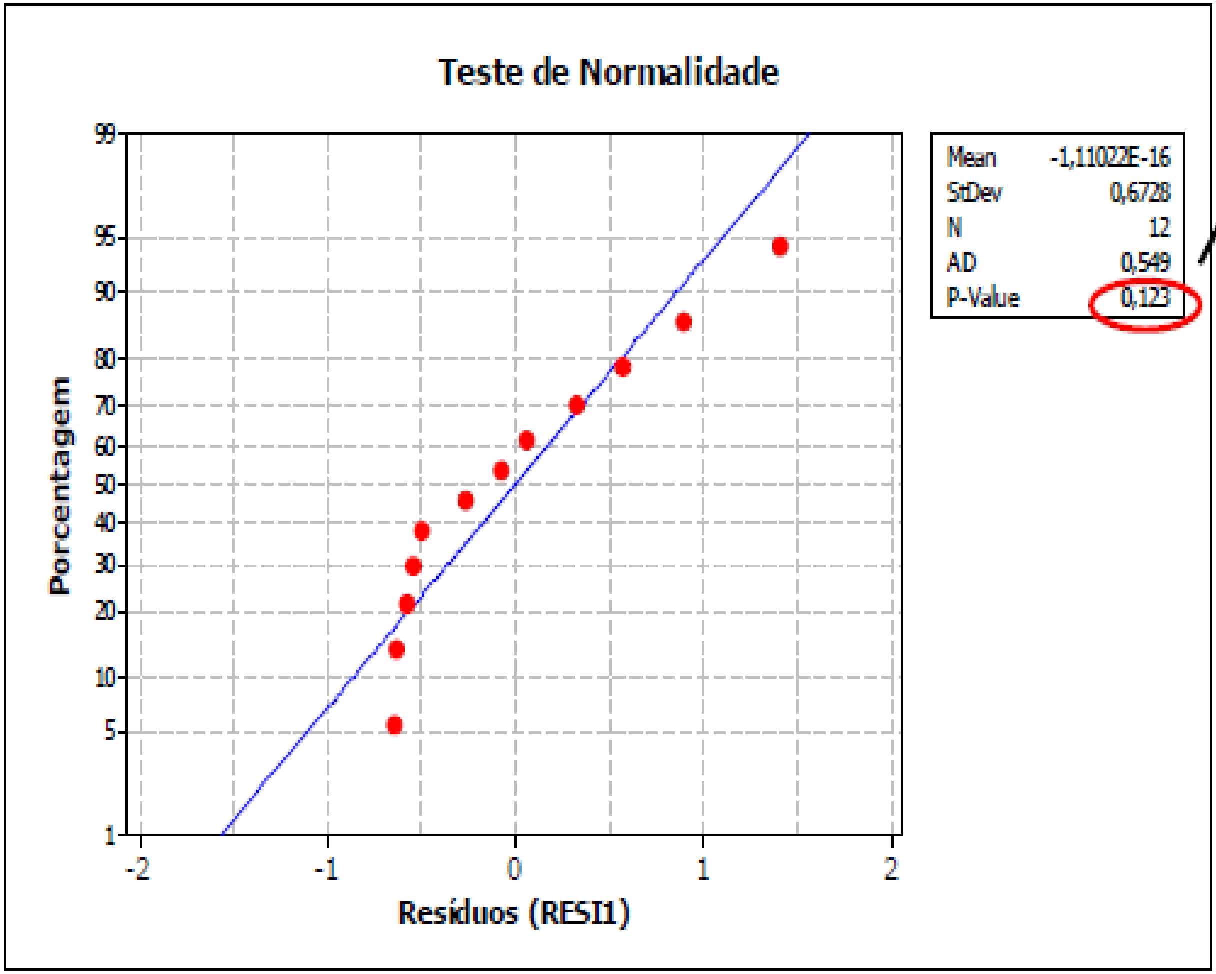
    $$H_1 : \exists \text{ pelo menos duas } \sigma_i^2 \neq s$$

Interpretação do Valor de P (*P-value*) do teste F:

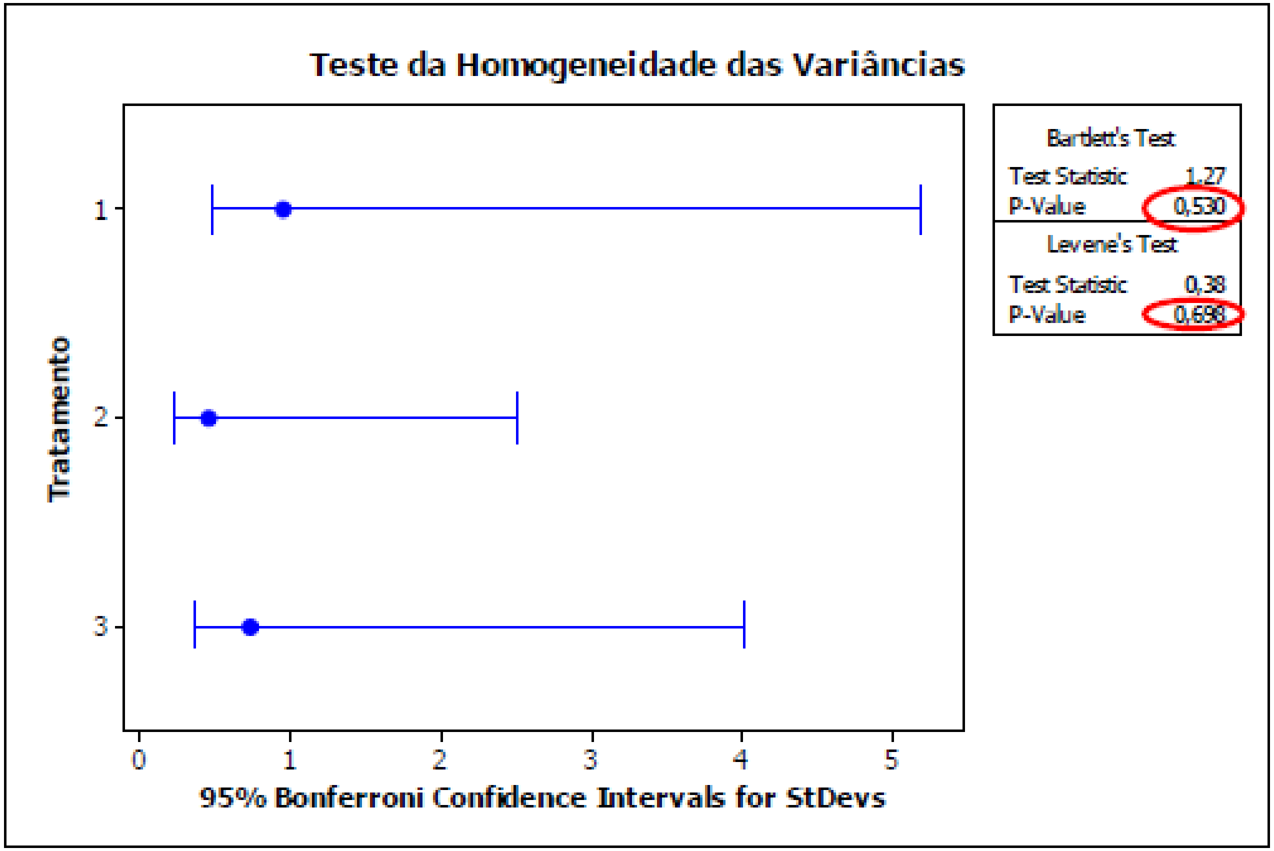
- se $P < 0,01$, significativo a *1%* e a *5%* (**);

- se $0,01<P<0,05$, significativo a *5%*(*);

- se $P > 0,05$, não significativo (*ns*).



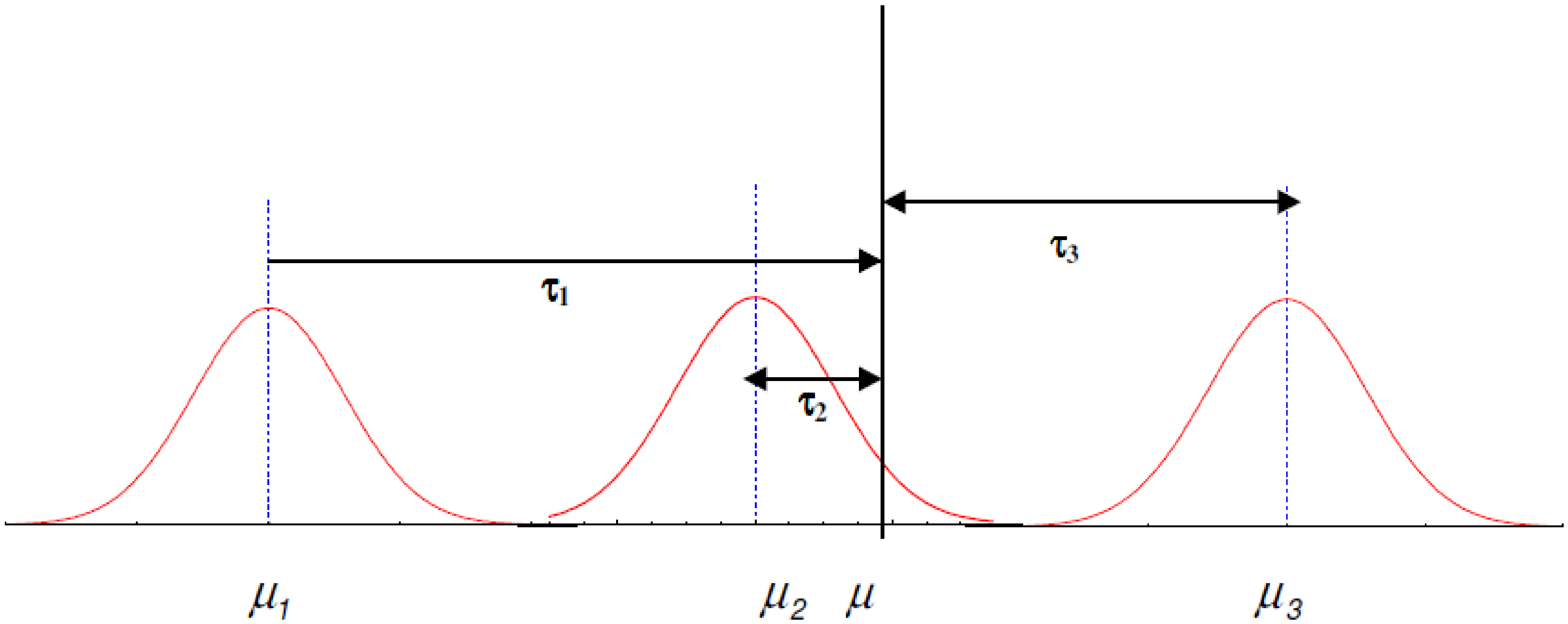
O valor de P (P-value) do Teste de Anderson-Darling (AD) é $> 0,05$, portanto, não rejeita-se a hipótese H_0 ,ou seja os erros têm uma distribuição normal.



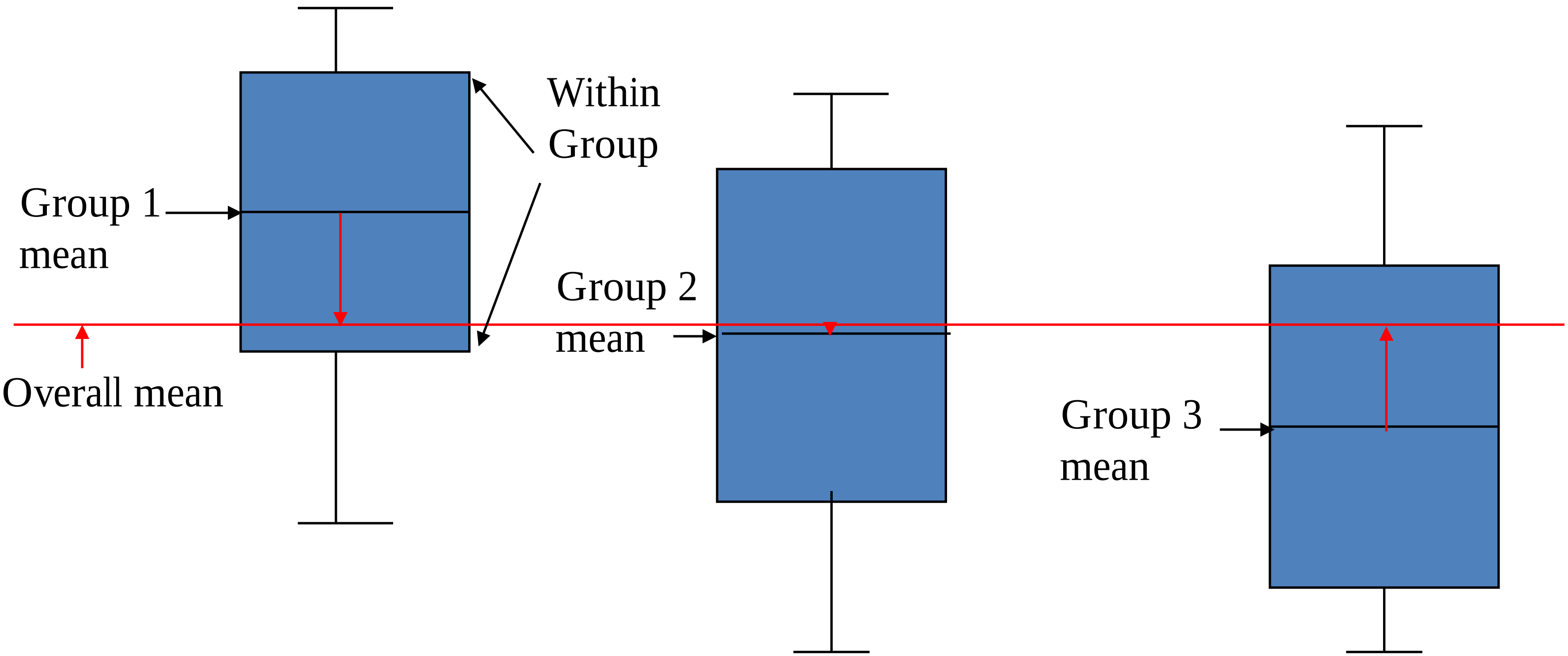
Repare que em ambos os testes (Bartlett e Levene) o valor de *P*> *0,05*, portanto, não rejeita-se a hipótese H_0 ,ou seja, a variância dos tratamentos são homogêneas

(homocedásticas)

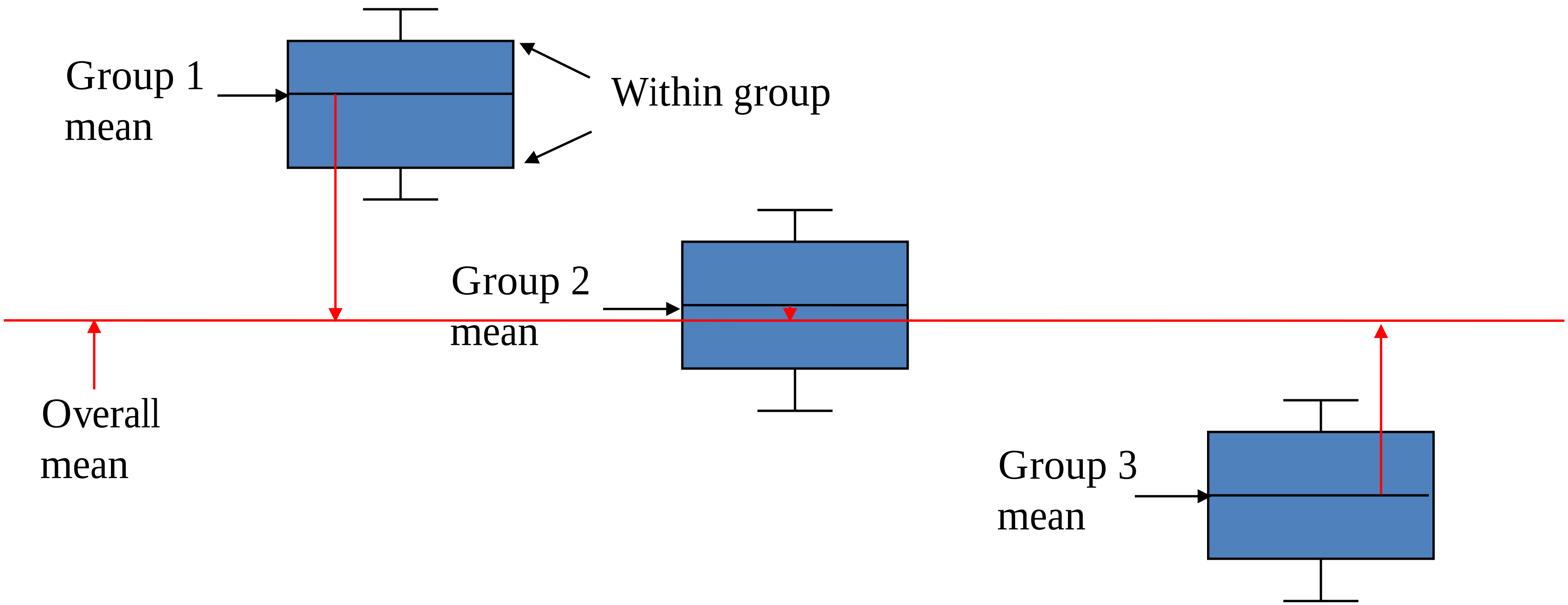
Existe diferença significativa entre as médias da figura abaixo?



Ilustração



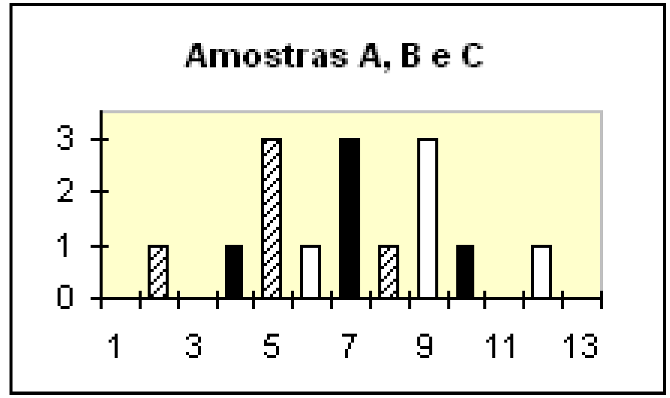
Source: Introduction to the Practice of Statistics, Moore and McCabe



Source: Introduction to the Practice of Statistics, Moore and McCabe

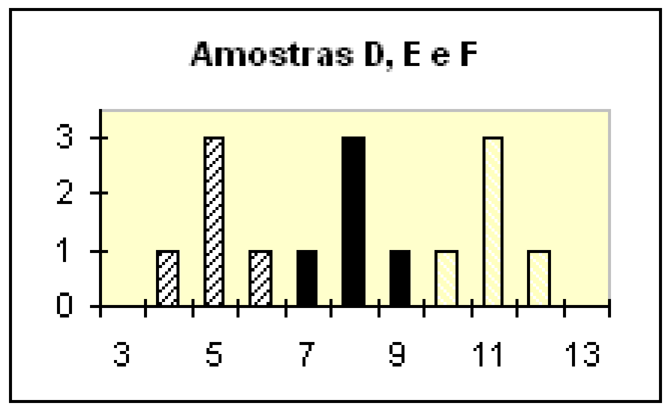
Amostras A, B e C

	A	B	C
	5	7	9
	8	10	12
	5	7	9
	2	4	6
	5	7	9
Média	**5**	**7**	**9**
DP	**2,12**	**2,12**	**2,12**



Amostras D, E e F

	D	E	F
	5	8	11
	6	9	12
	5	8	11
	4	7	10
	5	8	11
Média	**5**	**8**	**11**
DP	**0,71**	**0,71**	**0,71**



- A variabilidade total das amostras pode ser dividida em duas partes, ou fontes de variabilidade.
  - A primeira parte de variabilidade é proveniente de as populações serem diferentes, denominada variabilidade *entre*.
    - Quanto maior for a variabilidade *entre,* mais forte é a evidência de as médias das populações serem diferentes.
  - A segunda parte de variabilidade é causada pelas diferenças *dentro* de cada amostra, denominada variabilidade *dentro*.
    - Quanto maior for a variabilidade *dentro,* maior será a dificuldade para concluir se as médias das populações são diferentes.

- De maneira formal, o teste de hipóteses para *k* níveis de um fator é estabelecido da seguinte forma.

H_0: $m_1 = m_2 = m_3 \ldots = m_n$

H_1: Nem todas as populações têm a mesma média

- A distribuição *F* conduzirá a decisão de aceitar o rejeitar a hipótese nula, comparando o *F* observado F_o calculado com a expressão:

$$F_o = \frac{Variância \quad entre}{Variância \quad dentro} = \frac{S_b^2}{S_w^2}$$

com o *F* crítico F_c correspondente ao nível de significância a adotado. Também podem ser comparados o *p-value* de F_o e o nível de significância a.

ANOVA Paramétrica Simples

2º. Partição da Soma dos Quadrados

Se temos g grupos cada um com n observações, então:

$$\overline{Y}_{i.} = \frac{\sum_{j=1}^{n} Y_{ij}}{n}$$

média amostral do grupo i

$$\overline{Y}_{..} = \frac{\sum_{i=1}^{g}\sum_{j=1}^{n} Y_{ij}}{g \times n}$$

média total das observações

A variabilidade total das observações é dada pela soma dos quadrados total

$$SS_T = \sum_{i=1}^{g}\sum_{j=1}^{n} (Y_{ij} - \overline{Y}_{..})^2$$

soma dos quadrados total

soma das distâncias de cada observação à media total

$$SS_T = \underbrace{n\sum_{i=1}^{g} (\overline{Y}_{i.} - \overline{Y}_{..})^2}_{SS_G} + \underbrace{\sum_{i=1}^{g}\sum_{j=1}^{n} (Y_{ij} - \overline{Y}_{i.})^2}_{SS_E}$$

soma dos quadrados entre grupos

soma dos quadrados das distâncias das médias de cada grupo à media total

soma dos quadrados dentro de cada grupo

soma dos quadrados das distâncias de cada observação à média do seu grupo

ANOVA Paramétrica Simples

Partição da Soma dos Quadrados

g grupos cada um com n observações

A variabilidade total das observações é decomposta em dois termos:
o primeiro termo reflecte a variabilidade devida às diferenças entre grupos
e o segundo reflecte a variabilidade dos erros dentro de cada grupo

	Variabilidade Total	=	Variabilidade entre grupos	+	Variabilidade dentro dos grupos
	SS_T		SS_G		SS_E
graus de liberdade	$gn-1$		$g-1$		$g(n-1)$
médias dos quadrados			$MS_G = \frac{SS_G}{g-1}$		$MS_E = \frac{SS_E}{g(n-1)}$

ANOVA Paramétrica Simples
1 Factor, Efeitos Fixos

Sob H_0 a razão F tem distribuição de Fisher com g-1 e g(n-1) graus de liberdade:

$$F = \frac{MS_G}{MS_E} \sim F_{g-1,\, g(n-1)}$$

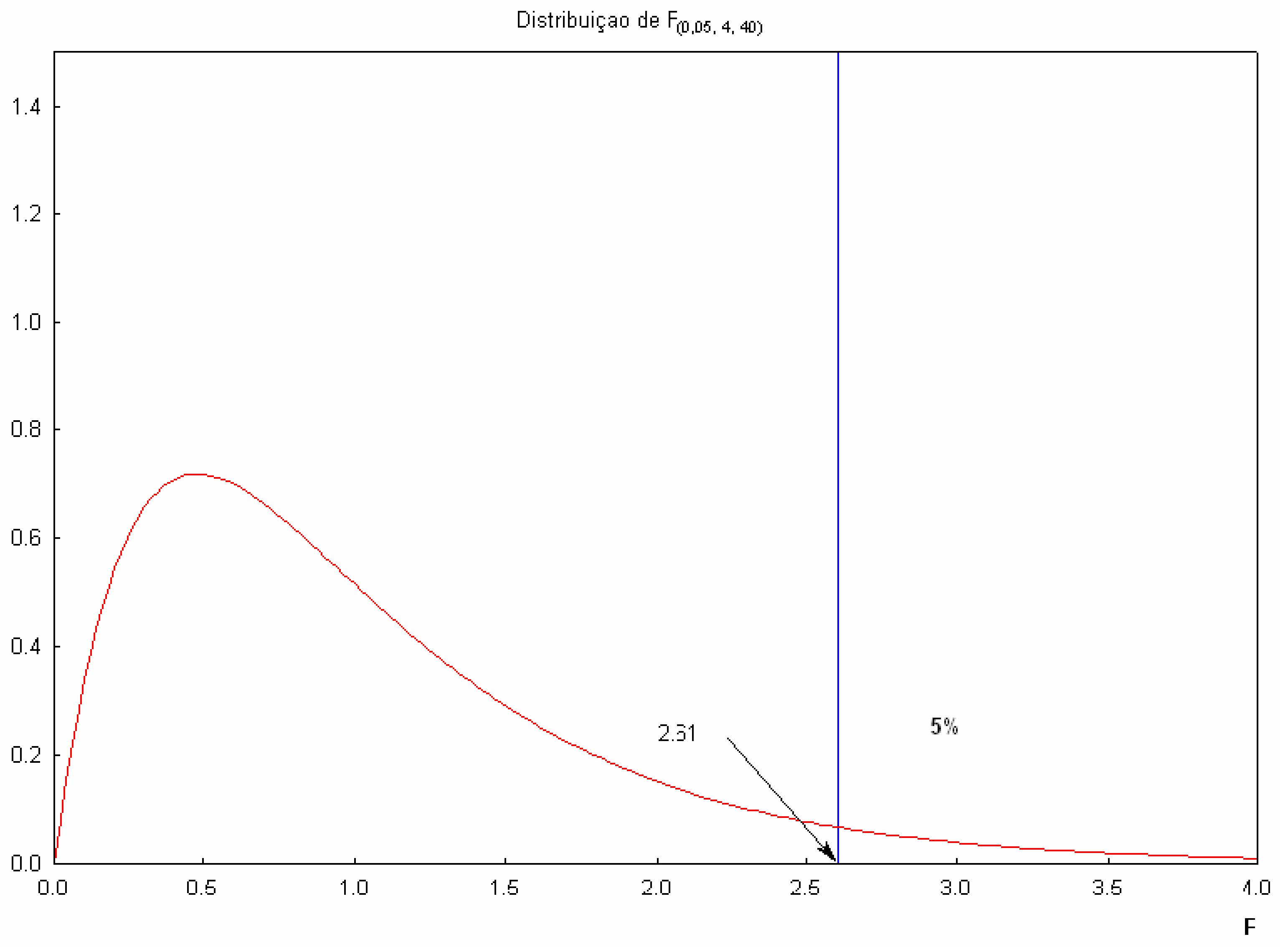
Podemos efectuar um teste com base nesta estatística baseado no p-value: **Rejeitar H_0** se p-value $\leq \alpha$

- A hipótese nula de igualdade de médias <u>será rejeitada apenas para valores elevados da estatística do teste F</u>

  $\Rightarrow$ p-value = $P(F > F_{obs} \mid H_0) = 1 - P(F < F_{obs}) = 1 - F_{g-1,\, g(n-1)}(F_{obs})$
- Para determinar $F_{g-1,\, g(n-1)}(F_{obs})$ recorrer ao menu do SPSS:

  Transform / Compute e escolher a função de distribuição de Fisher:

  CDF.F(F_{obs} , g-1, g(n-1))



O valor de F = 2,61 é o valor acima do qual, 5% dos valores de F calculados têm valor acima dele. Este é o valor para um nível de 5% encontrado na Tabela F para 4 e 40 graus de liberdade (veja Tabela F).

$$F_o = \frac{\dfrac{SSG}{k-1}}{\dfrac{SSE}{n_T - k}} = \frac{MSG}{MSE}$$

Fonte	gl	SS	MS	F
Entre	$k-1$	SST	$MSG = SST/(k-1)$	$F_o = MSG/MSE$
Dentro	n_T-k	SSE	$MSE = SSE/(n_T-k)$	
Total	n_T-1	SS		

ANOVA Paramétrica Simples

1 Factor, Efeitos Fixos

Exemplo 2

Para averiguar o tempo de aprendizagem de 3 listas de palavras: **lista A** com palavras curtas; **lista B** com palavras de tamanho médio; **lista C** com palavras compridas, foi realizada uma experiência com alunos de uma dada escola. A tabela mostra, os tempos observados, em segundos, que demoraram cada grupo **de 8 alunos** (escolhidos aleatoriamente entre os alunos da escola) a aprender a sua lista de palavras dada. Com base nos resultados da experiência, poderá afirmar que existem diferenças significativas no desempenho?

Lista A	Lista B	Lista C
30	54	68
40	58	75
35	45	80
45	60	75
38	52	85
42	56	90
36	65	75
25	52	88

Teste ANOVA

H_0: $\mu_A = \mu_B = \mu_C$ vs.

H_1: pelo menos uma das médias é diferente das demais

- Factor: Lista de Palavra
  $\Rightarrow$ temos **3 grupos** = **3 níveis: ListaA, ListaB** e **ListaC**
- Variável resposta (variável dependente) $\Rightarrow$
  Y- tempo (seg) que um aluno aprende a lista de palavras dada
- **Para cada grupo temos:** Uma amostra aleatória com **n=8** observações
  (os tempos observados que demoraram os 8 alunos seleccionados aletoriamente a aprender a sua lista de palavras)

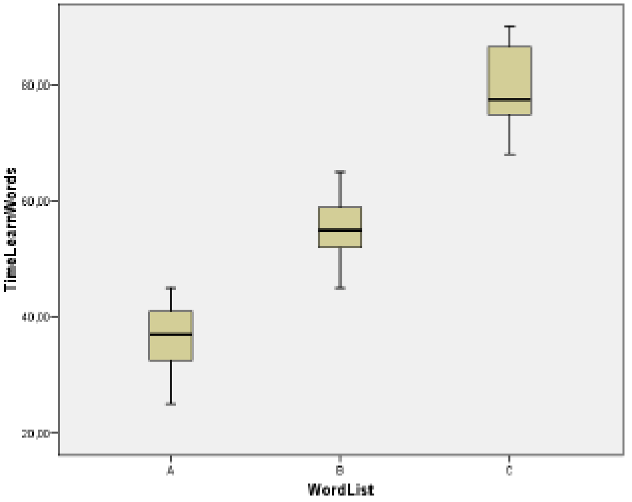
ANOVA Paramétrica Simples
1 Factor, Efeitos Fixos

Exemplo 2

Antes de conduzir a ANOVA paramétrica convém comparar graficamente a distribuição dos dados, através da construção de caixas de bigodes)

Aqui observamos que a mediana do tempo de aprendizagem aumenta com o aumento do tamanho das palavras e a variabilidade dos dados também aumenta.

Analyze → Descriptive Statistics → Explore



ATENÇÃO: *quando temos poucos dados, como neste caso é conveniente usar um teste não paramétrico. Vamos a usar uma ANOVA paramétrica apenas para poder exemplificar como são feitos todos os cálculos da estatística do teste*

ANOVA Paramétrica Simples
1 Factor, Efeitos Fixos

Exemplo 2

1°. Calcular media amostral e total:

3 grupos cada um com 8 observações

g = 3, n = 8

- ✓ média amostral do grupo i

$$\bar{Y}_{i.} = \frac{\sum_{j=1}^{n} Y_{ij}}{n}$$

- ✓ média total das observações

$$\bar{Y}_{..} = \frac{\sum_{i=1}^{g}\sum_{j=1}^{n} Y_{ij}}{g \times n}$$

Lista A	Lista B	Lista C
30	54	68
40	58	75
35	45	80
45	60	75
38	52	85
42	56	90
36	65	75
25	52	88
36.375	**55.25**	**79.50**
$\bar{Y}_{1.}$	$\bar{Y}_{2.}$	$\bar{Y}_{3.}$

média total:

$$\bar{Y}_{..} = \frac{\sum_{i=1}^{g}\sum_{j=1}^{n} Y_{ij}}{g \times n} = \frac{\sum_{i=1}^{g}\sum_{j=1}^{n} Y_{ij}}{3 \times 8} = 57.04$$

ANVOA Paramétrica Simples

1 Factor, Efeitos Fixos

Exemplo 2

3 grupos cada um com 8 observações

g = 3, n = 8

Lista A	Lista B	Lista C
30	54	68
40	58	75
35	45	80
45	60	75
38	52	85
42	56	90
36	65	75
25	52	88
36.375	**55.25**	**79.50**
$\bar{Y}_{1.}$	$\bar{Y}_{2.}$	$\bar{Y}_{3.}$

média total: $\boxed{\bar{Y}.. = 57.04}$

1°. Soma dos quadrados entre grupos

$$SS_G = n\sum_{i=1}^{g}(\bar{Y}_{i.} - \bar{Y}..)^2 = 7477.583$$

2°. Soma dos quadrados dentro dos grupos

$$SS_E = \sum_{i=1}^{g}\sum_{j=1}^{n}(Y_{ij} - \bar{Y}_{i.})^2 = 953.375$$

3°. Média dos quadrados entre grupos

$$MS_G = \frac{SS_G}{g-1} = \frac{7477.583}{2} = 3738.792$$

4°. Média dos quadrados dentro dos grupos

$$MS_E = \frac{SS_E}{g(n-1)} = \frac{953.375}{3\times 7} = 45.399$$

5°. Razão F

$$F = \frac{MS_G}{MS_E} = \frac{3736.792}{45.339} = 82.354$$

*a variabilidade entre os grupos é **82,354 vezes maior** que a variabilidade dentro dos grupos.*

ANOVA Paramétrica Simples

1 Factor, Efeitos Fixos

Exemplo 2

3 grupos cada um com 8 observações

$g = 3, \; n = 8$

Equipa A	Equipa B	Equipa C
30	54	68
40	58	75
35	45	80
45	60	75
38	52	85
42	56	90
36	65	75
25	52	88
36.375	**55.25**	**79.50**
$\overline{Y}_{1.}$	$\overline{Y}_{2.}$	$\overline{Y}_{3.}$

média total: $\overline{Y}_{..} = 57.04$

5°. Razão F

$$F = \frac{MS_G}{MS_E} = \frac{3736.792}{45.339} = 82.354$$

6°. Calcular o p-value

$$\begin{aligned} \text{p-value} &= P(F > F_{obs} \mid H_0) \\ &= 1 - P(F < F_{obs} \mid H_0) \\ &= 1 - F_{g-1,\, g(n-1)}(82.354) \\ &= 1 - F_{2,\, 21}(82.354) \\ &= 1 - \text{CDF.F}(82.354, 2, 21) \end{aligned}$$

$\Rightarrow$ p-value $\approx$ 0

$\Rightarrow$ rejeitar H_0 para q.q. nível de significância

Resultados usando o SPSS

Analyze → Compare Means → One-Way Anova

Exemplo 2

__Teste__: H_0: $\mu_A = \mu_B = \mu_C$ vs.

H_1: pelo menos uma das médias é diferente das demais

ANOVA

TimeLearnWords

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	7477,583	2	3738,792	82,354	,000
Within Groups	953,375	21	45,399		
Total	8430,958	23			

Uma vez que o **p-value** é aproximadamente **zero**
⇒ rejeitamos a hipótese nula de igualdade de médias para qualquer nível de significância.
Assim, a ANOVA permite concluir: para q.q. nível de significância, as médias dos vários grupos não são todas iguais, o que quer dizer que existem diferenças significativas no desempenho da aprendizagem das três listas de palavras.

ANOVA Paramétrica Simples
1 Factor, Efeitos Fixos

Quando rejeitamos a hipótese nula podemos optar por:

- Localizar as diferenças através de técnicas de comparações múltiplas: **métodos de Tukey, Scheffé, Bonferroni**
- Comparar os grupos de dois a dois por meio de intervalos de confiança para a diferença. Se o intervalo não contém o zero, podemos obter conclusões sobre a razão da rejeição.

Exemplo 1. Em um experimento de alimentação de porcos, foram utilizados quatro rações (A, B, C e D), cada uma fornecida a 5 animais. Os ganhos de peso, kg, foram:

Rações			
A	B	C	D
35	40	39	27
19	35	27	12
31	46	20	13
15	41	29	28
30	33	45	30

Calculando-se as somas de quadrados podemos construir o seguinte quadro de análise de variância:

Fonte de variação	g.l.	SQ	QM	F_c
Entre Tratamentos (Rações)	*3*	*823,75*	*274,58*	*3,99*
Resíduo (dentro dos tratamentos - Rações)	*16*	*1100,00*	*68,75*	
TOTAL	*19*	*1923,75*		

- Das tabelas das distribuições F, temos que $F_{(3,16,0,05)} = 3,24$ e $F_{(3,16,0,01)} = 5,29$. O valor F_c=3,99 é maior que o valor do F tabelado a 5%, então, rejeitamos a hipótese nula H_0 ao nível $\alpha = 0,05$, *ou 5%* de probabilidade.
- Dúvida: Qual é a ração que tem o melhor desempenho no ganho de peso?

5. TESTE DE TUKEY. O Teste de Tukey é baseado na amplitude total estudentizada (*studentized range*) e pode ser usado para comparar todo contraste entre duas médias de tratamentos.

- Hipóteses: $H_0 : Y_i = \mu_i - \mu_j = 0 \text{ para } i \neq j$
  $H_1 : Y_i = \mu_i - \mu_j \neq 0$
- Calcular o valor da *diferença mínima significativa* (d.m.s):

Experimento *balanceado* $r_i = r$ para todo i	Experimento *desbalanceado* $r_i \neq r_j$ para $\forall\ i \neq j$
$dms = q_{(k;\ n-k,\ \alpha)} \sqrt{\frac{QMR}{r}}$	$dms = q_{(k;\ n-k,\ \alpha)} \sqrt{\frac{QMR}{2} \left(\frac{1}{r_i} + \frac{1}{r_j}\right)}$

Sendo: $q_{(k,\ n-k,\ \alpha)}$ é o valor da *amplitude total estudentizada* e é obtido de tabela própria, e depende do número de tratamentos (k) e do número de graus de liberdade para o resíduo (n - k). Após calcular o *d.m.s.*, calculamos graus de liberdade para o resíduo (n - k). Após calcular o *d.m.s.*, calculamos a estimativa dos contrastes entre os pares de médias $\hat{Y}_i = \bar{x}_i - \bar{x}_j$ e comparamos esses valores com o valor do *d.m.s.*, aplicando a seguinte regra de decisão:

- se $\left|\hat{Y}_i\right| \geq d.m.s.$ **rejeitamos** $\boldsymbol{H_0}$, ao nível a de significância, e concluímos que as médias dos tratamentos envolvidos são diferentes;
- se $\left|\hat{Y}_i\right| < d.m.s.$ **não rejeitamos** $\boldsymbol{H_0}$ e concluímos que as médias dos tratamentos envolvidos são iguais.

Exemplo: (usaremos os dados do exemplo apresentado item 3(teste t-student).

- $k = 4$ $QMR = 0{,}0044375$ com 16 graus de liberdade e

  $q_{(4,\ ^{16}\ 0{,}05\alpha)} = 4,$ e

  $$dms = q_{(k;\, n-k,\, \alpha)}\sqrt{\frac{QMR}{r}} = 4,\ \sqrt{\frac{0{,}0044375}{4}} = 0{,}1399$$

Assim, toda estimativa de contraste do tipo $\left|\hat{Y}_i\right| = \left|\bar{x}_i - \bar{x}_j\right|$ que exceder o valor do $d.m.s. = 0{,}1399$ é significativo a 5%.

Estimativa do contraste
$\left\|\hat{Y}_1\right\| = \left\|\bar{x}_1 - \bar{x}_2\right\| = \left\|2{,}03 - 2{,}24\right\| = 0{,}21$ *
$\left\|\hat{Y}_2\right\| = \left\|\bar{x}_1 - \bar{x}_3\right\| = \left\|2{,}03 - 2{,}04\right\| = 0{,}01$ ns
$\left\|\hat{Y}_3\right\| = \left\|\bar{x}_1 - \bar{x}_4\right\| = \left\|2{,}03 - 2{,}22\right\| = 0{,}19$ *
$\left\|\hat{Y}_4\right\| = \left\|\bar{x}_2 - \bar{x}_3\right\| = \left\|2{,}24 - 2{,}22\right\| = 0{,}20$ *
$\left\|\hat{Y}_5\right\| = \left\|\bar{x}_2 - \bar{x}_4\right\| = \left\|2{,}24 - 2{,}22\right\| = 0{,}02$ ns
$\left\|\hat{Y}_6\right\| = \left\|\bar{x}_3 - \bar{x}_4\right\| = \left\|2{,}04 - 2{,}22\right\| = 0{,}18$ *

* - significativo a 5%; ns – não significativo a 5%

$\bar{x}_2$	$\bar{x}_4$	$\bar{x}_3$	$\bar{x}_1$
2,24a	2,22a	2,04b	2,03b ,

médias seguidas pela mesma letra minúscula não diferem entre si pelo teste de Tukey a 5% de probabilidade

Análise de Variância com dois ou mais factores - planeamento factorial

Em muitas experiências interessa estudar o efeito de mais do que um factor sobre uma variável de interesse. Quando uma experiência envolve dois ou mais factores diz-se que temos uma ANOVA múltipla. Uma ANOVA em que todas as combinações de todos os níveis de todos os factores são consideradas diz-se ANOVA factorial. Na maioria das situações, quando estamos interessados em estudar a influência de dois ou mais factores numa variável, utilizamos uma ANOVA factorial.

Exemplo: Pretende-se estudar a concentração de cálcio no sangue de uma população de aves parte da qual foi sujeita a um tratamento hormonal. Os investigadores pretendem averiguar se existem diferenças na concentração média de cálcio dependendo do tratamento hormonal e também dependendo do sexo das aves. Os factores deste estudo são o tratamento hormonal (presente ou ausente) e o sexo (feminino e masculino).

ANOVA múltipla - factores fixos, aleatórios e mistos

Vimos que numa ANOVA simples o factor em causa podia ter os efeitos fixos ou os efeitos aleatórios. O mesmo se vai passar com os modelos de ANOVA com dois ou mais factores.

Quando um modelo tem todos os factores com efeitos fixos diz-se que temos uma ANOVA de efeitos fixos ou um Modelo I de ANOVA.

Quando um modelo tem todos os factores com efeitos aleatórios diz-se que temos uma ANOVA de efeitos aleatórios ou um Modelo II de ANOVA.

Quando um modelo tem alguns factores com efeitos fixos e outros com efeitos aleatórios diz-se que temos uma ANOVA de efeitos mistos ou um Modelo III de ANOVA.