DECOMPOSIÇÃO DA SOMA DE QUADRADOS TOTAL NA ANOVA DE UM EXPERIMENTO COM UM FATOR

O modelo estatístico associado a um experimento balanceado com um fator de tratamento com a níveis, instalado usando um delineamento inteiramente casualizado (DIC) com n repetições, pode ser escrito como:

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

com $i = 1, 2, \ldots, a$ e $j = 1, 2, \ldots, n$. Em que:

y_{ij} é a resposta da j-ésima repetição do tratamento i

μ é uma constante comum a todas as observações

t_i é o efeito do i-ésimo tratamento (ou nível do fator) na variável dependente

e_{ij} é um erro aleatório atribuído à observação y_{ij}, não observável, independente e com distribuição normal de média zero e variância σ^2, ou seja, $e_{ij} \sim N(0, \sigma^2)$

	Tratamento				
	1	2	⋯	a	
	y_{11}	y_{21}	⋯	y_{a1}	
	y_{12}	y_{22}	⋯	y_{a2}	
	⋮	⋮	y_{ij}	⋮	
	y_{1n}	y_{2n}	⋯	y_{an}	
Total	$y_{1\bullet}$	$y_{2\bullet}$	⋯	$y_{a\bullet}$	$y_{\bullet\bullet}$
Média	$\bar{y}_{1\bullet}$	$\bar{y}_{2\bullet}$	⋯	$\bar{y}_{a\bullet}$	$\bar{y}_{\bullet\bullet}$

Já sabemos que:

$y_{i\bullet} = \sum_{j=1}^{n} y_{ij}$ é o total das observações do tratamento i

$\bar{y}_{i\bullet} = \frac{y_{i\bullet}}{n}$ é a média do tratamento i

$y_{\bullet\bullet} = \sum_{j=1}^{n} \sum_{i=1}^{a} y_{ij}$ é o total geral (de todas as observações)

$\bar{y}_{\bullet\bullet} = \frac{y_{\bullet\bullet}}{an}$ é a média geral (de todas as observações)

A ANOVA é uma técnica de análise estatística que serve para testar:

H_0: $\mu_1 = \mu_2 = \cdots = \mu_a$

H_a: pelo menos duas médias diferem entre si

Ela envolve uma partição da variabilidade total dos dados em dois ou mais componentes, dependendo do modelo utilizado admitido para os dados.

A soma de quadrado total é definida como uma medida da variabilidade total dos dados em relação à sua média geral:

$$SQTotal = \sum_{i=1}^{a} \sum_{j=1}^{n} \left(y_{ij} - \bar{y}_{\bullet\bullet}\right)^2 \quad (1)$$

Somando e subtraindo a média $\bar{y}_{i\bullet}$ (truque...) obtemos:

$$SQTotal = \sum_{i=1}^{a}\sum_{j=1}^{n}\{(y_{ij} - \bar{y}_{i\bullet}) + (\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet})\}^2$$

$$= \sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{i\bullet})^2 + \sum_{i=1}^{a}\sum_{j=1}^{n}(\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet})^2$$

$$+ 2\sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{i\bullet})(\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet}) \qquad (2)$$

Avaliando a soma de duplos produtos temos:

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

$$= \underbrace{\sum_{i=1}^{a}\sum_{j=1}^{n} y_{ij}\bar{y}_{i\bullet}}_{\text{(I)}} - \underbrace{\sum_{i=1}^{a}\sum_{j=1}^{n} y_{ij}\bar{y}_{\bullet\bullet}}_{\text{(II)}} - \underbrace{\sum_{i=1}^{a}\sum_{j=1}^{n} \bar{y}_{i\bullet}\bar{y}_{i\bullet}}_{\text{(III)}} + \underbrace{\sum_{i=1}^{a}\sum_{j=1}^{n} \bar{y}_{i\bullet}\bar{y}_{\bullet\bullet}}_{\text{(IV)}}$$

Mas

(I) $\sum_{i=1}^{a}\sum_{j=1}^{n} y_{ij}\bar{y}_{i\bullet} = \sum_{i=1}^{a}\bar{y}_{i\bullet}\sum_{j=1}^{n} y_{ij} = \sum_{i=1}^{a}\bar{y}_{i\bullet}y_{i\bullet} = \sum_{i=1}^{a}\bar{y}_{i\bullet}(n\bar{y}_{i\bullet}) = n\sum_{i=1}^{a}\bar{y}_{i\bullet}^2$

(II) $\sum_{i=1}^{a}\sum_{j=1}^{n} y_{ij}\bar{y}_{\bullet\bullet} = \bar{y}_{\bullet\bullet}\sum_{i=1}^{a}\sum_{j=1}^{n} y_{ij} = \bar{y}_{\bullet\bullet}(y_{\bullet\bullet}) = \bar{y}_{\bullet\bullet}(an\bar{y}_{\bullet\bullet}) = an\bar{y}_{\bullet\bullet}^2$

(III) $\sum_{i=1}^{a}\sum_{j=1}^{n} \bar{y}_{i\bullet}\bar{y}_{i\bullet} = \sum_{i=1}^{a}\sum_{j=1}^{n} \bar{y}_{i\bullet}^2 = n\sum_{i=1}^{a}\bar{y}_{i\bullet}^2$

(IV) $\sum_{i=1}^{a}\sum_{j=1}^{n} \bar{y}_{i\bullet}\bar{y}_{\bullet\bullet} = \bar{y}_{\bullet\bullet}\sum_{i=1}^{a}\sum_{j=1}^{n} \bar{y}_{i\bullet} = \bar{y}_{\bullet\bullet}\sum_{i=1}^{a} n\bar{y}_{i\bullet} = \bar{y}_{\bullet\bullet}\sum_{i=1}^{a} y_{i\bullet} = \bar{y}_{\bullet\bullet}(y_{\bullet\bullet})$

$$= \bar{y}_{\bullet\bullet}(an\bar{y}_{\bullet\bullet}) = an\bar{y}_{\bullet\bullet}^2$$

Percebendo-se o cancelamento dos somatórios (I) com (III) e (II) com (IV), conclui-se que:

$$\sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{i\bullet})(\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet}) = 0 \qquad (3)$$

Desta forma, a $SQTotal$ pode ser particionada da seguinte forma:

$$SQTotal = \sum_{i=1}^{a}\sum_{j=1}^{n}(\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet})^2 + \sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{i\bullet})^2$$

$$= \underbrace{n\sum_{i=1}^{a}(\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet})^2}_{SQTrat} + \underbrace{\sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{i\bullet})^2}_{SQResiduo}$$

$$= SQTrat + SQResiduo$$

- A $SQTotal$ ($soma\ de\ quadrados\ total$) mede a variação total dos dados observados em relação à média geral.
- $SQTrat$ ($soma\ de\ quadrados\ de\ tratamentos$) mede a variação das médias de cada tratamento em torno da média geral dos dados, ou seja, mede a variação *entre* tratamentos.

  Um valor alto de $SQTrat$ indica que as médias dos tratamentos estão distantes entre si e um valor baixo, que as médias dos tratamentos estão muito próximas umas das outras.
- $SQResiduo$ ($soma\ de\ quadrados\ do\ resíduo$) mede a variação dos valores observados em torno da média de cada tratamento. Mede a variação $dentro$ dos tratamentos.

  Um valor alto de $SQResiduo$ indica que as repetições de cada tratamento apresentam valores bem distantes da média do tratamento.

- A diferença $(y_{ij} - \bar{y}_{i\bullet})$ calculada entre os valores observados de um tratamento e a média deste tratamento serve para quantificar/estimar o erro devido ao acaso, ou seja, o erro experimental.

O número de graus de liberdade (gl) é uma característica associada a cada soma de quadrados. Pode-se pensar no gl como "o valor calculado a partir do número total de observações menos o número de parâmetros estimados".

- $SQTotal$ tem $an - 1$ graus de liberdade porque envolve an observações e precisamos estimar a média (μ) da população.
- $SQTrat$ tem $gl_{Trat} = a - 1$ graus de liberdade porque envolve a médias e precisamos estimar a média (μ) da população.
- $SQResiduo$ tem $gl_{Res} = a(n-1) = an - a$ graus de liberdade porque envolve an observações e precisamos estimar a médias.

Também podemos calcular gl_{Res} por diferença:

$$gl_{Res} = gl_{Total} - gl_{Trat} = (an - 1) - (a - 1) = a(n - 1)$$

Calculadas as SQ's e seu respectivos graus de liberdade, podemos montar o seguinte Quadro de análise de variância (ANOVA)

Fonte de variação	$g.l.$	SQ	QM
Tratamento	$a - 1$	$SQTrat = \sum_{i=1}^{a}\sum_{j=1}^{n}(\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet})^2$	$QMTrat = \frac{SQTrat}{(a-1)}$
Resíduo	$a(n-1)$	$SQResiduo = \sum_{i=1}^{a}\sum_{j=1}^{n}(\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet})^2$	$QMRes = \frac{SQResiduo}{a(n-1)}$
Total	$an - 1$	$SQTotal = \sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{\bullet\bullet})^2$	

Admitindo que as suposições do modelo (1) estão satisfeitas, para testar

H_0: $\mu_1 = \mu_2 = \cdots = \mu_a$
H_a: pelo menos duas médias diferem entre si

usamos a estatística

$$F = \frac{QMTrat}{QMResiduo} \tag{4}$$

que tem $(a - 1)$ graus de liberdade no numerador e $a(n - 1)$ no denominador.

Escolhendo um nível de significância $\alpha = 5\%$ e obtendo o valor crítico tabulado (F_{tab}) na tábua da distribuição F com $(a - 1)$ e $a(n - 1)$ graus de liberdade decidimos:

- Se $F_{calc} > F_{tab}$ nós rejeitamos H_0 ($p < 0{,}05$) e concluímos que pelo menos duas médias de tratamentos diferem entre si.
- Se $F_{calc} \leq F_{tab}$ nós aceitamos H_0 ($p \geq 0{,}05$) e concluímos que as médias de todos os tratamentos são iguais entre si.