UNIVERSIDADE DE SÃO PAULO
Faculdade de Zootecnia e Engenharia de Alimentos

ZAB0229 – ESTATÍSTICA EXPERIMENTAL

AULA 2

Análise de variância (ANOVA)

Principais delineamentos experimentais

ANÁLISE DE VARIÂNCIA

O modelo estatístico associado a um experimento com <u>um fator de tratamento</u> com a níveis em um <u>delineamento inteiramente casualizado</u> com n repetições, pode ser escrito como:

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

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

μ_i é a média do tratamento i

e_{ij} é um erro aleatório atribuído à observação y_{ij}, $e_{ij} \sim N(0, \sigma^2)$.

No planejamento de um experimento é utilizado um delineamento experimental, cuja escolha depende:

- Objetivos da pesquisa
- Área experimental disponível
- Qualidade do material experimental
- Número de repetições por tratamento
- Manejo utilizado *etc*

Vamos conhecer os delineamentos experimentais mais utilizados em pesquisas agronômicas e zootécnicas, evidenciando seus principais aspectos

DELINEAMENTO INTEIRAMENTE CASUALIZADO (DIC)

- Deve ser utilizado quando o material experimental usado é bastante homogêneo.
- Considera somente os princípios da repetição e da casualização.
- Os tratamentos são designados às parcelas de forma inteiramente casual, por meio de sorteio.
- É o delineamento que proporciona o maior número de graus de liberdade para o resíduo.

Exemplo 3.1. Desejamos estudar o desempenho de suínos submetidos a 5 rações comerciais (A, B, C, D e E). Cada uma das rações foi fornecida a 4 animais escolhidos ao acaso de um grupo homogêneo de vinte animais.

Croqui do ensaio

A	D	B	A	C	E	A	C	E	B
D	B	A	C	E	C	E	D	D	B

No caso geral, a tratamentos e n repetições por tratamento, os dados observados podem ser colocados em uma tabela do tipo:

	Tratamento				
	1	2	$\cdots$	a	
	y_{11}	y_{21}	$\cdots$	y_{11}	
	y_{12}	y_{22}	$\cdots$	y_{12}	
	$\vdots$	$\vdots$	y_{ij}	$\vdots$	
	y_{1n}	y_{2n}	$\cdots$	y_{1n}	Geral
Total	$y_{1\bullet}$	$y_{2\bullet}$	$\cdots$	$y_{a\bullet}$	$y_{\bullet\bullet}$
Média	$\bar{y}_{1\bullet}$	$\bar{y}_{2\bullet}$	$\cdots$	$\bar{y}_{a\bullet}$	$\bar{y}_{\bullet\bullet}$

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 permite testar:

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

H_a: Pelo menos duas médias diferem entre si

A ANOVA envolve a partição da variabilidade total dos dados em dois ou mais componentes, dependendo do modelo admitido para esses dados.

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

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

Truque: Somando e subtraindo a média de tratamentos, $\bar{y}_{i\bullet}$, (ver detalhes no Apêndice 1) obtemos a partição:

$$\begin{aligned} SQTotal &= \sum_{i=1}^{a} \sum_{j=1}^{n} \left[\left(y_{ij} - \bar{y}_{i\bullet}\right) + \left(\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet}\right)\right]^2 \\ &= \underbrace{n \sum_{i=1}^{a} \left(\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet}\right)^2}_{SQTrat} + \underbrace{\sum_{i=1}^{a} \sum_{j=1}^{n} \left(y_{ij} - \bar{y}_{i\bullet}\right)^2}_{SQResiduo} \end{aligned}$$

- $\boldsymbol{SQTrat}$ ($soma\ de\ quadrados\ de\ tratamentos$) mede a variação das médias dos tratamentos em torno da média geral dos dados, ou seja, mede a variação $entre$ os tratamentos.

  Um valor alto de $SQTrat$ indica que as médias dos tratamentos estão distantes umas das outras e um valor baixo, que as médias dos tratamentos estão muito próximas umas das outras.

- $\boldsymbol{SQResiduo}$ ($soma\ de\ quadrados\ do\ residuo$) mede a variação dos valores observados em cada tratamento em torno da sua respectiva média, ou seja, mede a variação $dentro$ dos tratamentos.

  Um valor alto de $SQResiduo$ indica que as repetições dentro de cada tratamento apresentam valores distantes da média do tratamento. Ou ainda, que a variabilidade do erro experimental (fontes não controladas) foi alta.

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

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 $gl_{Total} = an - 1$ 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 seus respectivos números de graus de liberdade, podemos montar o quadro de análise de variância (ANOVA), que é apresentada a seguir:

Quadro de ANOVA

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

Também precisamos calcular os quadrados médios usando:

$$QMTrat = \frac{SQTrat}{(a-1)} \qquad\qquad QMRes = \frac{SQResiduo}{a(n-1)}$$

Admitindo que $e_{ij} \sim N(0, \sigma^2)$, temos que

- $y_{ij} \sim N(\mu + t_i, \sigma^2)$
- $SQTrat/\sigma^2$ e $SQResiduo/\sigma^2$ têm distribuição quiquadrado com $(a-1)$ e $a(n-1)$ graus de liberdade, respectivamente, e são independentes.
- A estatística $F = \dfrac{QMTrat}{QMResiduo}$ tem distribuição F com $(a-1)$ e $a(n-1)$ graus de liberdade.

Podemos também provar que:

$$E(QMTrat) = \frac{n}{a-1}\sum_{i=1}^{a} t_i^2 + \sigma^2 \quad \text{e} \quad E(QMResiduo) = \sigma^2$$

ou seja,

- $QMResiduo$ é um estimador não viesado da variância dos dados, σ^2;
- $QMTrat$ também é um estimador não viesado de σ^2 somente quando $\sum_{i=1}^{a} t_i^2 = 0$, ou seja, as médias dos tratamentos são iguais ou que os efeitos de tratamento são nulos.

Ideia: Usar a estatística $F = QMTrat\ /\ QMResiduo$ para testar

$$H_0: t_1 = t_2 = \cdots = \ t_a = 0 \ \text{ ou } \ H_0: \mu_1 = \mu_2 = \cdots = \ \mu_a$$

que são duas <u>hipóteses equivalentes</u>.

- Um valor de $F \cong 1$ indica que $\sum_{i=1}^{a} t_i^2 \cong 0$ ou que as médias dos tratamentos devem ser consideradas iguais, o que corresponde a não rejeitar a hipótese H_0.
- Um valor de F muito maior que 1, indica que $\sum_{i=1}^{a} t_i^2 \neq 0$ ou que as médias dos tratamentos não devem ser consideradas iguais, o que corresponde a rejeitar a hipótese H_0 e aceitar H_a: pelo menos duas médias diferem entre si, como verdadeira.

Escolhendo o nível de significância $\alpha = 5\%$ e obtendo F_{tab} na tábua da distribuição F com $(a - 1)$ e $a(n - 1)$ graus de liberdade, comparamos os valores de F_{calc} e F_{tab}:

- 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.

EXEMPLO 1: Em um experimento de competição de $a = 6$ variedades de cana-de-açúcar, instalado em um delineamento inteiramente casualizado com $n = 5$ repetições por tratamento, obteve-se as seguintes produções (ton/ha):

	Variedades de cana de açúcar					
	CB5034	CB6245	IAC6258	IAC6529	IAC6814	IAC6538
	112,3	125,3	118,4	127,9	130,1	115,2
	121,0	119,7	120,5	128,3	122,4	123,2
	114,3	120,8	119,7	129,5	126,7	117,8
	115,8	120,5	118,3	126,5	127,3	120,8
	117,2	122,3	117,8	127,3	128,9	116,4
Média	116,1	121,7	118,9	127,9	127,1	118,7

Com os dados do Exemplo 1 nós obtemos:

$$SQTotal = (112{,}3 - 121{,}74)^2 + \ \dots \ + \ (116{,}4 - 121{,}74)^2$$
$$= 726{,}3920$$

Com as médias das variedades:

CB5034	CB6245	IAC6258	IAC6529	IAC6814	IAC6538	Geral
116,1	121,7	118,9	127,9	127,1	118,7	121,74

Obtemos:

$$SQVariedade = 5[(116{,}1 - 121{,}74)^2 + \dots + (118{,}7 - 121{,}74)^2]$$
$$= 576{,}2480$$

$$SQErro = (112{,}3 - 116{,}1)^2 + \ \dots \ + \ (116{,}4 - 118{,}7)^2$$
$$= SQTotal - SQVariedades = 150{,}1440$$

Quadro de Análise de Variância (ANOVA)

Causa de Variação	g.l.	S.Q.	Q.M.	F_{calc}
Variedade	5	576,2480	115,2496	18,42 **
Resíduo	24	150,1440	6,2560	
Total	29	726,3920		

$\bar{y}_{\bullet\bullet} = 121.74$ $\quad s^2 = 6,2560$ $\quad CV = 2,05\%$

Em que

$$CV = \frac{100 * \sqrt{QMResiduo}}{\bar{y}_{\bullet\bullet}} = \frac{\sqrt{6,2560}}{121,74} = 2,05\%$$

Baseado no valor do CV, podemos comentar que “o experimento foi bem conduzido” ou que “os fatores estranhos ao experimento foram bem controlados”, porque $CV < 10\%$.

Para testar:

$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 = \mu_6$

H_a: Pelo menos duas médias diferem entre si.

Como $F_{calc} = 18{,}42 > F_{tab}(5;24;5\%) = 2{,}6$ nós rejeitamos H_0 e concluímos que "pelo menos duas das variedades apresentam produções médias diferentes entre si".

Exercício. Em um experimento de alimentação de suínos, utilizando um delineamento inteiramente casualizado, foram comparadas quatro rações (A, B, C e D), cada uma fornecida a cinco animais. Com base nos ganhos de peso, em kg, dos animais, monte o quadro de ANOVA e teste a hipótese $H_0: \mu_A = \mu_B = \mu_C = \mu_D$.

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

$Dica: SQTotal = 1923{,}75$

Vamos conhecer outros dois delineamentos importantes: o delineamento casualizado em blocos (DCB) e o delineamento em quadrado latino (DQL).

Posteriormente, resolveremos exemplo numéricos envolvendo esses delineamentos.

DELINEAMENTO CASUALIZADO EM BLOCOS (DCB)

- Além da repetição e da casualização, o DCB considera também o princípio do controle local.
- O controle local é usado quando o material experimental é heterogêneo e se conhece a causa (aparente) dessa heterogeneidade.
- A variação dentro dos blocos deve ser a menor possível, ao passo que a variação entre blocos pode ser pequena ou grande.

Exemplo 3.2. Desejamos estudar o desempenho de suínos submetidos a 5 rações comerciais (A, B, C, D e E). Os 20 animais serão agrupados em 4 blocos de acordo com o peso do animal no início do experimento. Dentro de cada bloco, as rações serão designadas aos animais por sorteio.

Croqui do ensaio

Bloco I				
B	A	C	D	E

Bloco II				
D	C	B	E	A

Bloco III				
E	C	A	B	D

Bloco IV				
C	A	D	B	E

Quadro de Análise de Variância

Causas de Variação	g.l.
Bloco	3
Ração	4
Resíduo	12
Total	19

DELINEAMENTO EM QUADRADOS LATINOS (DQL)

- Usado quando as condições experimentais são muito heterogêneas, sendo necessário o controle de dois tipos de heterogeneidade, genericamente chamados de linhas e colunas.
- É um delineamento que exagera no controle local.
- É utilizado frequentemente em experimentos de digestibilidade, com bovinos fistulados, por exemplo.
- **Problema**: O número de repetições deve ser igual ao número de tratamentos $\Rightarrow$ o número de parcelas deve ser um quadrado perfeito.

Exemplo 3.3. Desejamos estudar o desempenho de suínos submetidos a 4 rações comerciais (A, B, C e D). Usaremos 16 animais agrupados de acordo com os seus pesos ao início do experimento (linhas) e suas idades (colunas). As rações serão designadas aos animais por sorteio, tanto nas linhas quanto nas colunas.

Croqui do ensaio

		Idade			
		C1	C2	C3	C4
Peso inicial	L1	D	A	B	C
	L2	A	C	D	B
	L3	C	B	A	D
	L4	B	D	C	A

Note que: cada linha é composta por um bloco com 4 animais de mesmo peso inicial e cada coluna, por um bloco com 4 animais de mesma idade e que cada ração aparece uma única vez na mesma linha e mesma coluna.

Quadro de Análise de Variância

Causas da Variação	g.l.
Linhas (Peso inicial)	3
Colunas (Idade)	3
Ração	3
Resíduo	6
Total	15