

5.1. ANOVA - Análise de Variância

A *Análise de Variância* (também conhecida pela sigla inglesa *ANOVA* : **AN**alysis **Of** **VA**riance) é uma técnica que nos permite comparar simultaneamente médias de vários grupos, para testar se diferem significativamente entre si. A ANOVA pode, por isso, ser considerada como uma extensão do teste *t*, que permite comparar apenas duas médias de cada vez.

Por que uma técnica para comparar médias teria a palavra “variância” no nome? A razão deste nome e a lógica da técnica podem ser melhor entendidas através de um exemplo. Suponha que estamos comparando os efeitos de quatro *tratamentos* diferentes, aplicados a quatro amostras de uma espécie de animal, e que a variável de resposta *y* que nos interessa seja o peso do animal depois do tratamento. Os gráficos das Figs. 1 e 2 mostram duas situações possíveis. (Neste exemplo usaremos amostras que têm todas o mesmo tamanho, $n = 7$; isto simplifica as equações, mas não é exigido para a ANOVA).

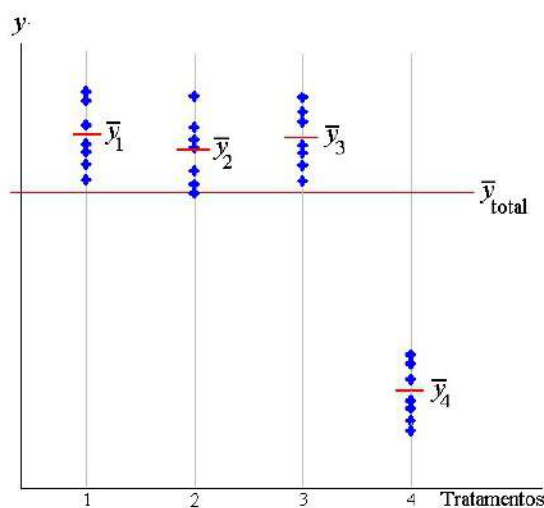


Figura 1

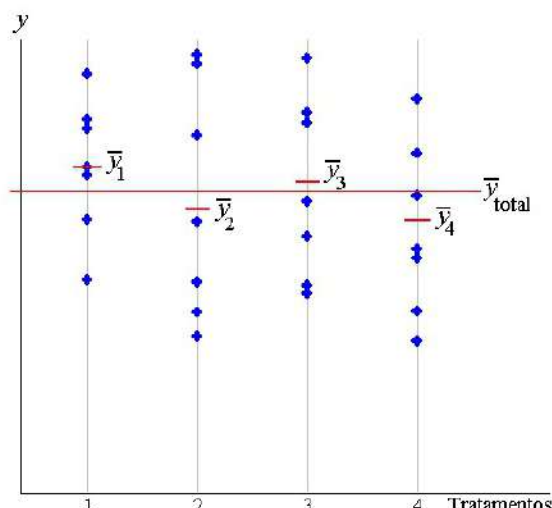


Figura 2

Comparando as figuras, é razoável supor que deve haver uma diferença significativa entre as médias dos pesos depois dos diferentes tratamentos na Fig. 1; pelo menos, a média da amostra do tratamento 4 ($\bar{y}_4$) parece ser bem menor do que as outras, o que indica que este tratamento deve ter um efeito diferente dos outros. A variação *entre* as médias dos tratamentos (isto é, de tratamento para tratamento) é muito grande; a variação *dentro* de cada tratamento parece ser pequena, comparativamente. Na Fig. 2, porém, a situação é diferente: a variação *entre* as médias dos tratamentos parece pequena (estas médias parecem ser todas aproximadamente iguais), mas a variação *dentro* de cada tratamento é grande.

O que a ANOVA faz é comparar quantitativamente estas duas variações: se a variação *entre* as médias for grande, em relação à variação *dentro* dos tratamentos, somos levados a concluir que as médias dos tratamentos devem ser diferentes entre si; caso contrário, se a variação *entre* as médias for pequena em relação à variação *dentro* dos tratamentos, concluímos que as médias provavelmente não diferem significativamente entre si.

Para medir estas variações, usaremos as somas dos quadrados dos desvios. *Desvios* são as diferenças entre observações e médias, ou entre duas médias. Suponhamos que haja k tratamentos (no exemplo, $k = 4$). Chamando

y_{ij} : observação i -ésima do tratamento j -ésimo

$\bar{y}_j$: média do tratamento j -ésimo

$\bar{y}_{total}$: média de todas as observações, de todos os tratamentos

podemos considerar três tipos de desvios, que estão representados na Fig. 3. Primeiro, há o desvio do valor de cada observação em relação à média total (a média de todas as observações); chamaremos este desvio de ‘desvio total’:

$$\text{desvio total} = y_{ij} - \bar{y}_{total}$$

Se supusermos que as quatro amostras vem de uma mesma população (esta é a hipótese nula do teste que será feito mais abaixo), podemos obter um estimador da variância desta população pela divisão da soma dos quadrados destes desvios totais por seu número de graus de liberdade (o número de observações, menos 1):

$$\sigma^2 = \frac{\sum (y_{ij} - \bar{y}_{total})^2}{kn - 1}$$

O desvio total pode ser decomposto em dois outros desvios (Fig. 3): o desvio do valor de cada observação em relação à média do tratamento à qual a observação pertence (chamaremos este componente de desvio *dentro* do grupo):

$$y_{ij} - \bar{y}_j$$

e o desvio da média da amostra de cada tratamento em relação à média total (chamaremos este componente de desvio *entre* grupos):

$$\bar{y}_j - \bar{y}_{total}$$

O desvio total é portanto a soma dos desvios *dentro* e *entre*:

$$(y_{ij} - \bar{y}_{total}) = (y_{ij} - \bar{y}_j) + (\bar{y}_j - \bar{y}_{total})$$

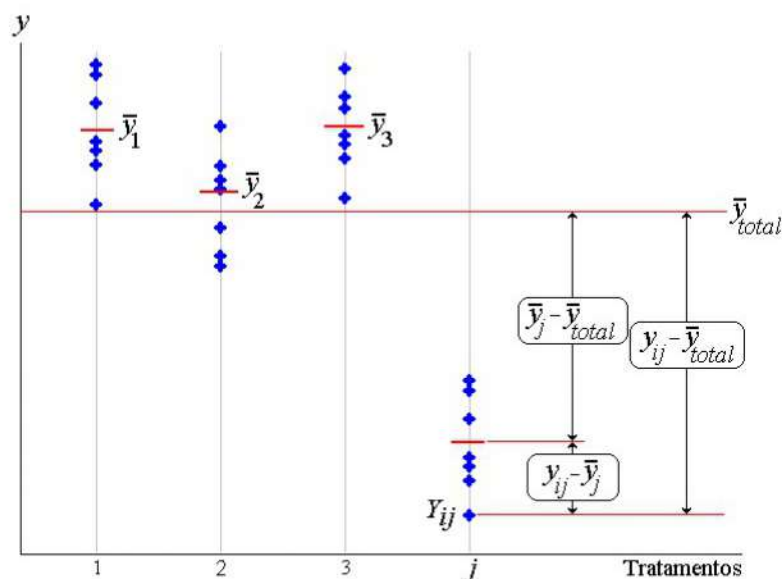


Figura 3. Decomposição dos desvios em relação à média total

Pode ser demonstrado que o quadrado do desvio total também pode ser decomposto na soma dos quadrados dos desvios *dentro* e *entre*:

$$\text{desvio total:} \quad (y_{ij} - \bar{y}_{total}) = (y_{ij} - \bar{y}_j) + (\bar{y}_j - \bar{y}_{total})$$

$$\text{quadrado do desvio total:} \quad (y_{ij} - \bar{y}_{total})^2 = (y_{ij} - \bar{y}_j)^2 + (\bar{y}_j - \bar{y}_{total})^2$$

(Note que esta decomposição *não* é óbvia. Lembre-se do resultado bem conhecido: se $y=a+b \rightarrow y^2 = (a+b)^2 = a^2+2ab+b^2$).

Para que servem os desvios *dentro* e *entre*? Eles poderão ser usados como base para medir a variação *dentro* dos grupos, e a variação *entre* grupos, mencionadas no início deste capítulo. A variação *dentro* dos tratamentos será ser medida pela soma dos quadrados dos desvios dentro:

$$SQ_{dentro} = \sum_j \sum_i (y_{ij} - \bar{y}_j)^2$$

A variação *entre* um tratamento e outro será medida pela soma dos desvios *entre* a média da amostra de cada tratamento e a média total, multiplicada pelo número de observações em cada tratamento:

$$SQ_{entre} = n \sum_j (\bar{y}_j - \bar{y}_{total})^2$$

O problema estatístico agora é: como comparar as somas de quadrados SQ_{dentro} e SQ_{entre} por meio de um *teste de hipótese*, a fim de decidir se uma delas pode ser considerada significativamente maior do que a outra?

Para fazer este teste, precisamos de um pressuposto e de uma hipótese nula. O pressuposto é o de que em todos os tratamentos a variável de resposta y tem a mesma variância, que representaremos por σ^2 ; a hipótese nula é a de que as médias de todos os tratamentos são iguais nas populações:

$$H_0 \rightarrow \mu_1 = \mu_2 = \dots = \mu_k = \dots = \mu$$

Se a H_0 for verdadeira, a variação *entre* os tratamentos nas amostras deve ser muito pequena em comparação à variação *dentro* dos tratamentos, como no gráfico da Fig. 2; isto é, SQ_{entre} deve ser pequena, em relação à SQ_{dentro} . Se contudo H_0 for falsa, a variação *entre* os tratamentos deve ser grande, em comparação à variação *dentro* dos tratamentos, como no gráfico da Fig. 1.

Para comparar as duas somas de quadrado, precisamos de transformá-las em estimadores da variância total de todos os tratamentos. Se a hipótese nula for verdadeira, as diferenças observadas entre as médias amostrais de cada tratamento serão devidas simplesmente ao acaso, e podemos usar SQ_{entre} para obter um estimador da variância σ^2 . (Lembre-se que a variância é calculada pela soma dos quadrados dos desvios dividida por seu número de graus de liberdade). Este estimador será chamado de *quadrado médio 'entre'*, representado por QM_{entre} . No caso:

$$\hat{\sigma}_{entre}^2 = QM_{entre} = \frac{SQ_{entre}}{k-1}$$

O número de graus de liberdade é igual a k , o número de médias amostrais, subtraído de uma unidade (porque precisamos de calcular a média total $\bar{y}_{total}$, e perdemos por isso um grau de liberdade).

Além disso, se a hipótese nula for verdadeira, a soma SQ_{dentro} também poderá ser usada para calcular um estimador da variância σ^2 , dado pelo quadrado médio QM_{dentro} :

$$\hat{\sigma}_{dentro}^2 = QM_{dentro} = \frac{SQ_{dentro}}{kn - k} = \frac{SQ_{dentro}}{k(n-1)}$$

O número de graus de liberdade no denominador é o número de observações (kn , pois são k tratamentos, e são feitas n observações em cada tratamento), subtraindo o número de médias amostrais (porque precisamos de calcular as médias dentro de cada um dos k tratamentos, e perdemos por isso k graus de liberdade).

Se a hipótese nula for verdadeira, temos portanto dois estimadores da variância σ^2 . Para fazer um teste comparando estes dois estimadores, precisamos de um pressuposto e de um teorema. O pressuposto é que a variável resposta y em todos os tratamentos tem distribuição normal. O teorema é o seguinte:

Se duas amostras aleatórias independentes são retiradas de uma população de distribuição normal com média μ e variância σ^2 , a razão entre as variâncias $\hat{\sigma}_1^2$ e $\hat{\sigma}_2^2$ destas amostras, σ_1^2 / σ_2^2 , terá distribuição de probabilidades F , caracterizada por um número de graus de liberdade dado pela razão entre o número de graus de liberdade de cada amostra: $gl = gl_1 / gl_2$. (Na ANOVA, os graus de liberdade dos dois estimadores são $gl_{entre} = k-1$ e $gl_{dentro} = k(n-1)$)

O que deve ser feito agora é um teste de hipótese, usando a distribuição F : supondo que a hipótese nula seja verdadeira, podemos considerar os dois estimadores da variância que obtivemos, QM_{entre} e QM_{dentro} , como estimadores calculados a partir de duas amostras independentes retiradas de uma mesma população normal (se os pesos nos quatro tratamentos tem distribuição normal e mesma variância, por pressuposto, e mesma média, pela hipótese nula, então os quatro tratamentos tem exatamente mesma distribuição); a razão entre estes estimadores deverá ter distribuição F . Os resultados deste teste são geralmente organizados como no Quadro 1, e o valor- p encontrado é mostrado na última coluna.

Quadro 1 – Resultados de uma ANOVA

variação	SQ	graus de liberdade	QM	F	valor-p
<i>Entre</i> tratamentos	SQ_{entre}	$k-1$	QM_{entre}	QM_{entre} / QM_{dentro}	
<i>Dentro</i> dos tratamento	SQ_{dentro}	$k(n-1)$	QM_{dentro}		
Total	SQ_{total}	$nk-1$	QM_{total}		

Como exemplo, o Quadro 2 mostra o resultado de ANOVA usada para responder à pergunta: os pesos de crianças recém-nascidas são significativamente diferentes, de acordo com a raça da mãe? Os dados são americanos, e há 189 crianças, divididas entre $k=3$ amostras (raças *branca*, *negra*, e *outra*). Este exemplo será desenvolvido com mais detalhes no Apêndice 1.

Quadro 2 – Peso ao nascer, de acordo com a raça da mãe

	Soma dos quadrados	gl	Quadrados médios	F	p
Entre grupos	5.071	2	2.535.303	4,97	0,0078
Dentro dos Grupos	94.846	186	509.927		
Total	99.917	188			

Se a rejeitamos da hipótese nula, concluímos que nem todas as médias dos tratamentos são iguais; não sabemos, no entanto, qual ou quais delas são diferentes da maioria das outras. Para saber isto, precisamos usar os chamados ‘testes *post-hoc*’, que comparam as médias dos tratamentos duas a duas, usando versões modificadas do teste-*t*, e indicam quais delas diferem significativamente entre si (por exemplo, o teste de Bonferroni; veja o exemplo no Apêndice 1).

Esta é a lógica por trás do tipo mais simples de ANOVA, a *one-way ANOVA*, na qual as observações são classificadas de acordo com apenas um *fator* (no exemplo, a raça da mãe). Há modalidades mais complexas, em que dois ou mais fatores são usados simultaneamente (por exemplo, poderíamos comparar os pesos das crianças de acordo com a raça da mãe, e com o fato de esta *fumar* ou não); estas modalidades não serão vistas aqui.

Observações:

- Na explicação acima consideramos que as amostras de cada tratamento tenham o mesmo tamanho n . Isto simplifica as contas, mas não é essencial. No exemplo do Quadro 2, as amostras tinham tamanhos diferentes.
- Um pressuposto importante é o de que todos os tratamentos tenham distribuição normal. Se as amostras forem grandes, este pressuposto pode ser relaxado (i.e., podemos usar ANOVA mesmo que os tratamentos não sejam inteiramente normais).
- Outro pressuposto é o de que todos os tratamentos tenham a mesma variância; é preciso portanto fazer antes um teste (por exemplo, o de *Levene*), para verificar se o pressuposto é válido. Se não for, será melhor usar testes não-paramétricos, em vez de ANOVA.