ANÁLISE DE VARIÂNCIA

Na estatística, de maneira geral, utilizamos um tipo de análise chamada

ANÁLISE DE VARIÂNCIA

MAS O QUE É VARIÂNCIA !!!?

SABEMOS QUE VARIÂNCIA

É MÉDIA DA SOMA DOS QUADRADOS DO DESVIO EM RELAÇÃO A MÉDIA?!

A análise de variância, conhecida como ANOVA, é uma técnica que consiste, fundamentalmente, em decompor a variância total de um conjunto, em variâncias parciais, correspondentes a fontes de variação diferentes e determinadas.

Feito isto, as variâncias poderão ser comparadas entre si por meio de algum teste estatístico.

Podemos fazer uma análise de variância com dados que tenham distribuição conhecida ou não.

Para facilitar o entendimento da ANOVA, tomemos um caso simples em que há apenas um fator de tratamento e esse tratamento seja qualitativo.

Se denotarmos as respostas como "y", a soma de quadrados dos desvios em relação à média é simplesmente denotada como soma de quadrados total:

$$\text{SQTotal} = \sum_{j=1}^{c} (y_j - \bar{y})^2$$

Se entre os y tiverem “t” tratamentos, podemos identificar os tratamentos, supondo todas as parcelas homogêneas (como no caso do DIC), por i = 1, 2, ...,t, e as médias desses tratamentos por $\overline{y}_i$

Algebricamente, mostra-se que:

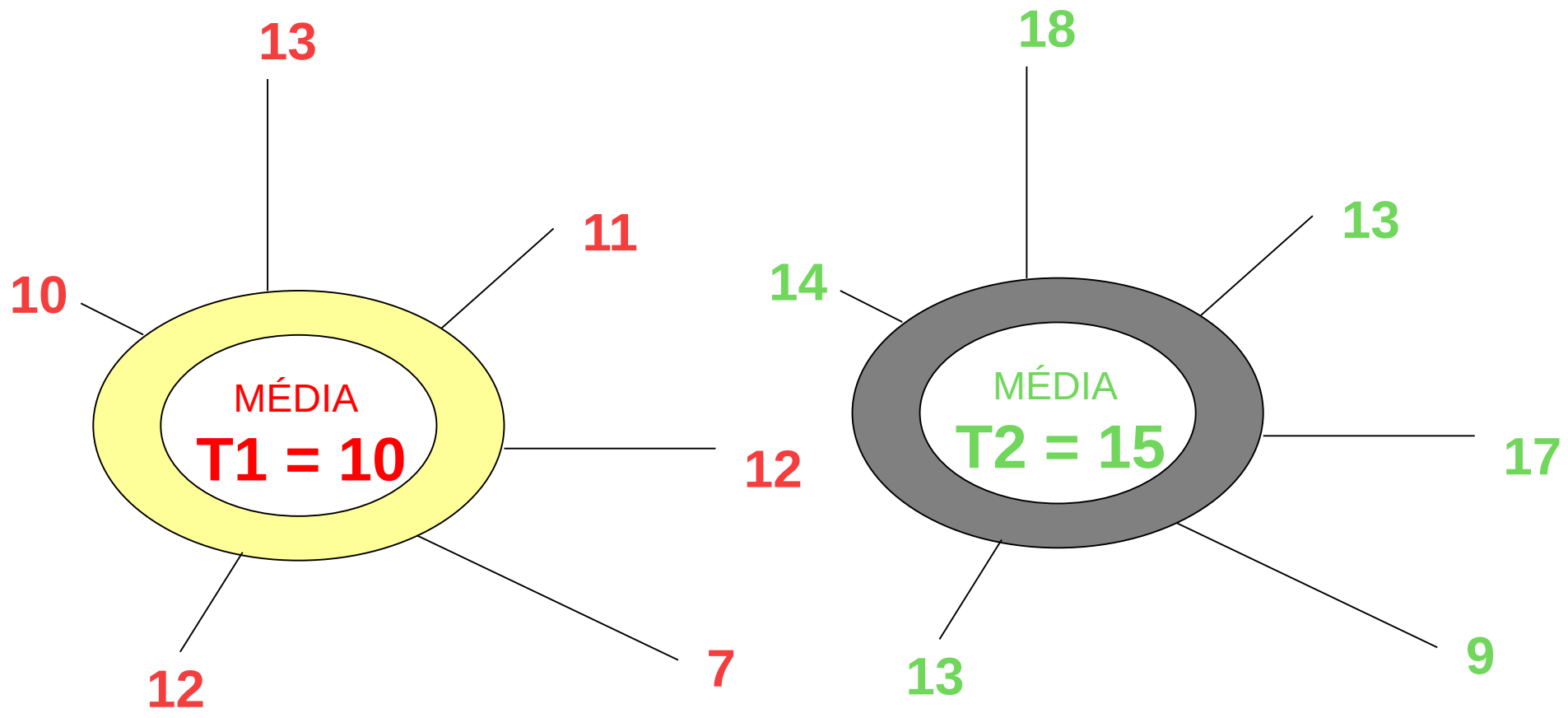
$$\text{SQTotal} = \sum_i (\bar{y}_i - \bar{y})^2 + \sum_i \sum_j (y_{ij} - \bar{y}_i)^2$$

ou seja, SQTotal = SQTratamentos + SQResíduo

Uma outra maneira equivalente de se chegar a essa partição é supondo o modelo:

Resposta (y) = (média de tratamentos) + resíduo

O **resíduo**, no caso, é também denominado “erro puro” expressando a variabilidade ao redor da média de cada tratamento (que é estimada, na prática, pela média amostral de cada tratamento).



SUPONDO 3 TRATAMENTOS

MÉDIA T1 = 10

MÉDIA T2 = 15

MÉDIA T3 = 11

SQ TRATAMENTO

MÉDIA GERAL DO EXPERIMENTO = 12

SUPONDO 3 REPETIÇÕES DE CADA TRATAMENTO

SQ RESÍDUO

É A SOMA DO DESVIO DE CADA REPETIÇÃO EM RELAÇÃO A MÉDIA DO SEU TRATAMENTO

Grau de liberdade

É um parâmetro associado a cada estatística ou fonte de variação (soma de quadrados) e expressa o número de valores independentes que ela contém.

No caso da ANOVA, com uma média por tratamento ou grupo os graus de liberdades associadas a essa fonte de variação será igual ao número de tratamentos -1, os graus de liberdade total será n-1, e os graus de liberdade do resíduo será a diferença entre os dois.

Voltando à questão da partição da soma de quadrados total (SQTotal).

Quando essa for partida em duas, como no caso do exemplo inicial (DIC), podemos representar os valores de uma forma esquemática e padronizada que é conhecida como Tabela de análise de variância.

Fontes de variação (FV)	Graus de liberdade (GL)	Soma de quadrados (SQ)	Quadrado médio (QM) = SQ/GL
TRATAMENTOS	$GL_{Tratamentos}$	$SQ_{Tratamentos}$	$QM_{Tratamentos}$
RESÍDUO	$GL_{Resíduo}$	$SQ_{Resíduo}$	$QM_{Resíduo}$
TOTAL	GL_{Total}	SQ_{Total}	QM_{Total}

O quadrado médio (QM) para cada fonte e variação é obtido pela razão entre a soma de quadrados da fonte de variação em questão pelo seus respectivos graus de liberdade.

A partir da Tabela de análise de variância podemos obter algumas estatísticas importantes de interesse prático:

Coeficiente de determinação: R^2 = SQTratamento/SQTotal

Expressa proporcionalmente ou percentualmente quanto da variabilidade dos dados pode ser atribuída ao tratamento. Ou, quanto o conjunto de dados está ajustado ao modelo de análise. Importante estatística que definirá a confiabilidade dos resultados.

Desvio padrão geral médio: s = $\sqrt{QM_{\text{Resíduo}}}$

É uma média ponderada da variabilidade das respostas dentro de cada tratamento. Ou seja, mede quanto as repetições de cada tratamento estão variando entre si.

Coeficiente de variação: CV = (s/ $\bar{y}$). 100

Obtida a partir da média geral dos . Essa estatística expressa percentualmente a precisão com que o experimento foi realizado. Quanto menor o valor do CV melhor é a precisão experimental. Essa precisão esta relacionada com a forma como o experimento foi instalado e conduzido.

Várias classificações de CV foram propostas por diversos autores. Uma classificação bastante usada com culturas de cereais é:

CV < 5%	Muito bom (Baixo)
5% < CV < 10%	Bom (Satisfatório)
10% < CV < 20%	Regular (Intermediário)
CV > 20%	Alto
CV > 30%	Muito Alto

No caso de experimentos de campo com cana-de-açúcar há atributos que dependem do estande, por exemplo, TCH (toneladas de cana por hectare), especialmente nas socas e os CV podem ser mais altos:

Valores ao redor de 10% são muito bons. Já em atributos tecnológicos (Brix, por exemplo) a Tabela anterior é satisfatória.

Estatística F = QMTratamento/QMResíduo

Essa estatística pode dar idéia da igualdade ou diferença estatística entre as variações de tratamentos e do resíduo.

É essa estatística que vai nos dizer se existe ou não diferença entre os tratamentos!!!

O esquema de análise de variância pode ser executado para muitas situações da experimentação. A seguir, ilustraremos numericamente algumas das possibilidades. Substituímos os “Tratamentos” (médias dos tratamentos) por valores de um “Modelo” qualquer

SQTotal = SQModelo + SQDesvio

Consideremos um exemplo com os dados (fictícios), apresentados na próxima Tabela, onde se supõem 4 tratamentos, 4 doses diferentes de certo produto e 2 repetições, num experimento inteiramente casualizado.

TRATAMENTO	DOSE	REPETIÇÃO	RESPOSTAS (y_{obs})
A	1	1	8,2
A	1	2	7,8
B	2	1	9,8
B	2	2	10,4
C	3	1	12,5
C	3	2	11,5
D	4	1	10,8
D	4	2	11,2

A **média geral** é = (8,2 + 7,8 + 9,8 +...+ 11,2)/ 8, ou seja, =

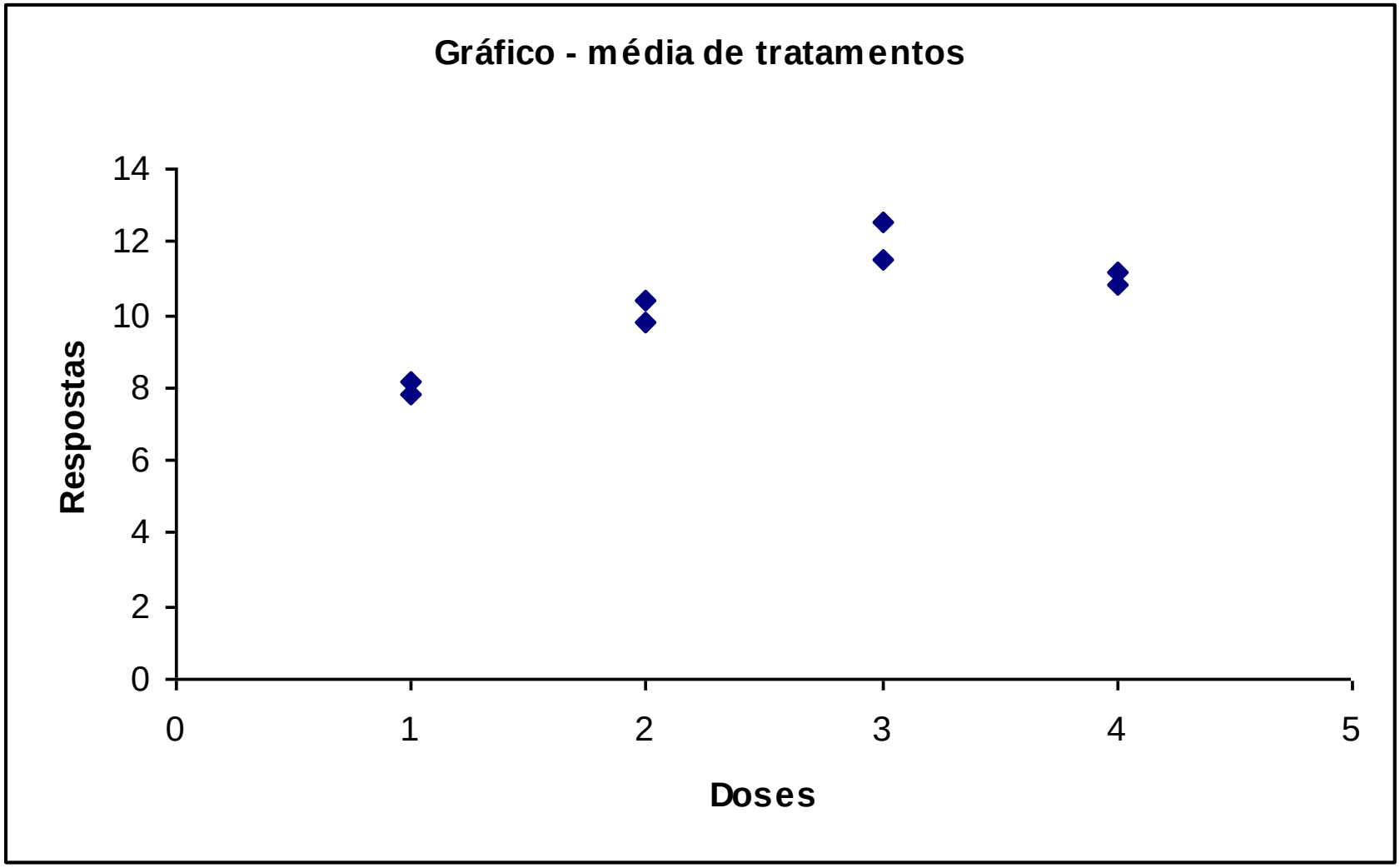
$$\overline{y} = 10{,}275$$

A **SQTotal** pode ser obtida assim:

$$SQTotal = \Sigma(y_{obs} - \overline{y})^2$$

para nosso exemplo SQTotal= 18,255

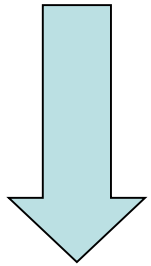
O gráfico de dispersão ilustra o comportamento das respostas em relação as doses ou tratamentos:



Para obtermos a SQresíduo é necessário conhecermos o desvio ou resíduo de cada observação em relação ao modelo.

Supondo o modelo de médias (uma média por tratamento), o modelo é predito pela média observada de cada tratamento e o resíduo ou desvio é chamado ERRO PURO, assim temos:

Resposta= (média de tratamentos) + erro puro (ou resíduo),



TRATAMENTO	Y_{OBS}	Y(modelo de médias)	Diferença ou desvio
A	8,2	8,0	0,2
A	7,8	8,0	-0,2
B	9,8	10,1	-0,3
B	10,4	10,1	0,3
C	12,5	12,0	0,5
C	11,5	12,0	-0,5
D	10,8	11,0	-0,2
D	11,2	11,0	0,2

Então, o SQdesvio= $(0,2)^2 + (-0,2)^2 + ... + (0,2)^2 = 0,84$.
No modelo com uma média por tratamento SQresíduo= ERRO PURO.
Com essas informações podemos organizar a Tabela de análise de variância:

FV	GL	SQ	QM	F
TRATAMENTOS (modelo de médias)	(t-1) = 3	17,415	5,805	27,64
ERRO PURO (resíduo)	(n-t) = 4	0,84	0,21	
TOTAL	(n-1)= 7	18,255		

Sendo t= nº de tratamentos; n= total de observações.

O CV [(s/ $\bar{y}$). 100] obtido para essa análise foi 4,45, valor excelente considerando a Tabela de classificação geral utilizada para culturas de cereais.

O $R^2 = 0,95$ (SQTratamento/SQTotal) indica um ótimo ajuste dos dados ao modelo.

No geral, pode-se interpretar:

$R^2 >= 0,90$, indica ajuste ótimo

$0,70 < R^2 < 0,90$, ajuste bom

$0,50 < R^2 < 0,70$, ajuste regular

$0,30 < R^2 < 0,50$, ajuste fraco

$R^2 <= 0,30$, falta de ajuste

Como podemos observar no gráfico, embora o modelo de médias de tratamentos tenha se ajustado bem aos dados, uma análise de regressão poderia ser realizada com o objetivo de mostrar mais claramente as tendências em função das doses.

Em muitas situações, embora significativo, o modelo de médias de tratamentos por não informar a relação funcional com doses, pode não ser o melhor. Nesse caso, uma análise de regressão (linear, quadrática ou outra) poderia ser mais indicada.

Dessa maneira, utilizando os mesmos dados para realizar a análise de regressão quadrática (parábola), temos:

SQTotal= 18,255 e SQresíduo= 1,569.

A SQresíduo é calculada através do desvio obtido em relação a cada observação, através da substituição de cada dose na equação de ajuste do modelo parábola.

Na próxima Tabela estão apresentados os desvios.

TRATAMENTO	DOSE	Y_{OBS}	Y(modelo parábola)	DESVIO
A	1	8,2	7,865	0,335
A	1	7,8	7,865	-0,065
B	2	9,8	10,505	-0,705
B	2	10,4	10,505	-0,105
C	3	12,5	11,595	0,905
C	3	11,5	11,595	-0,095
D	4	10,8	11,135	-0,335
D	4	11,2	11,135	0,065

O modelo de análise que podemos utilizar para o experimento com modelo quadrático é:

Resposta = (modelo quadrático de dose) + DESVIO.

A partir daí, temos a seguinte Tabela de análise de variância:

FV	GL	SQ	QM	F
MODELO PARÁBOLA (Regressão quadrática)	2	16,686	8,343	26,59
DESVIO	(5)	1,569	0,3138	
Falta de ajuste	1	0,729	0,729	
Erro puro	4	0,840	0.210	
TOTAL	7	18,255		

Nesse caso o desvio é dado pelo erro puro mais a falta de ajuste ao modelo, assim:

DESVIO= ERRO PURO (4 GL) + FALTA DE AJUSTE (1GL).

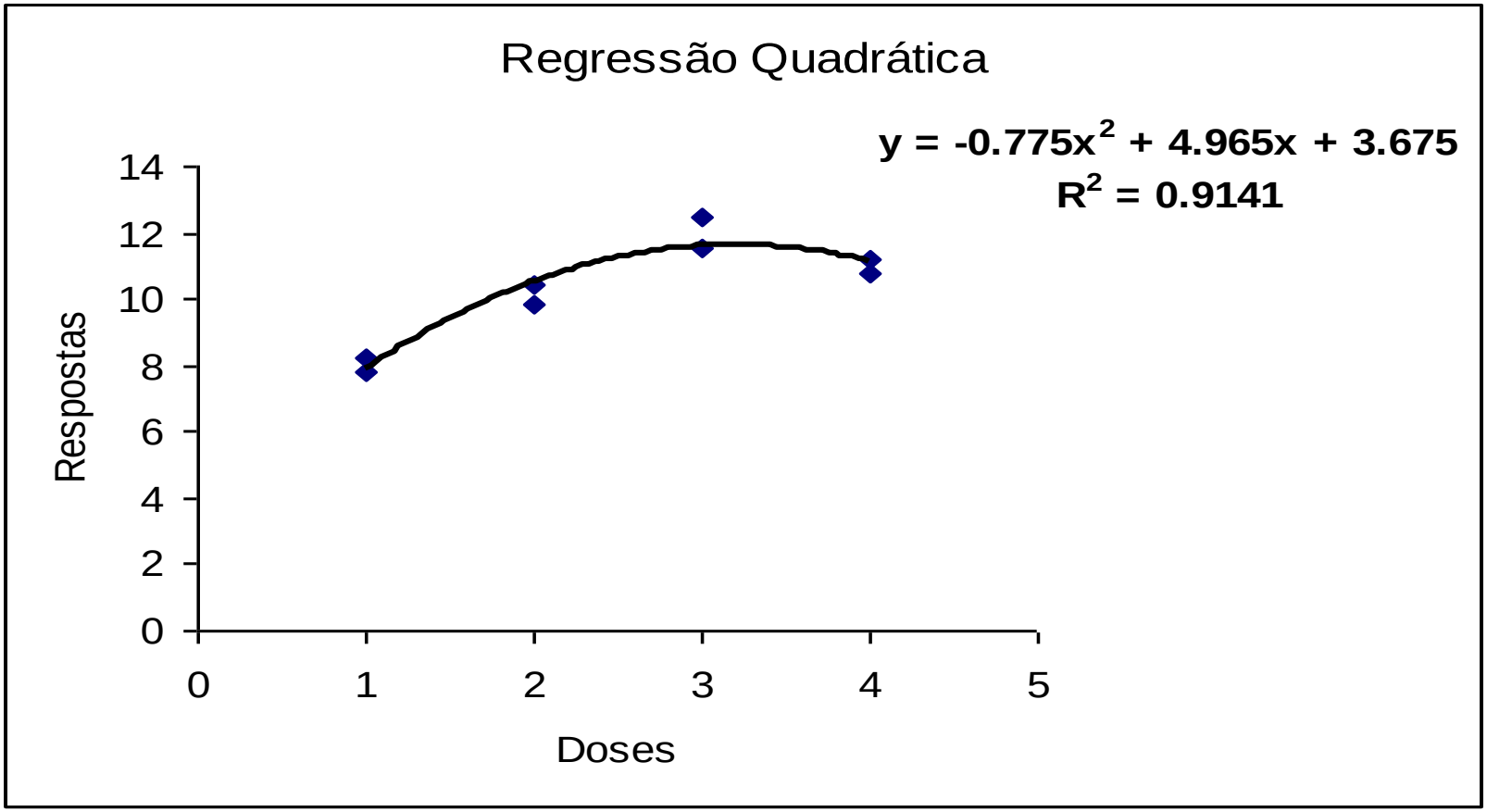
Para avaliar a Falta de ajuste, podemos usar a

estatística F= 0,729/0,210 = 3,47, comparada com F da distribuição (tabelada) com 1 e 4 graus de liberdade, conforme será discutido no item Teste de significância.

Ou usar o seu R^2 = 0,729/18,255 = 0,04, ou 4%.

Ou seja, a parábola não é perfeita, mas é um modelo com R^2= 91,4% e com baixa falta de ajuste (4%).

Como é possível observar a primeira análise de variância realizada continua válida, mas é menos informativa. Com essa última análise e o gráfico abaixo conseguimos compreender melhor o comportamento das doses.



A equação da parábola que explica a relação entre os tratamentos e as doses é:

$y = -0,775x2 + 4,965x + 3,675.$

O ponto de máximo calculado pela parábola é:
$x_{max} = - (4,965)/(2 . (-0,775)) = 3,2$

O R2= 0,9141, indica um excelente ajuste dos dados à parábola. Ou seja, 91,41% da variabilidade dos dados é captada pelo modelo de regressão quadrático.

Avaliemos também o comportamento do modelo reta ou linear simples.

Supondo uma regressão linear para esses dados temos:

SQTotal = 18,255, ou seja, igual a obtida nos modelos anteriores.
A SQresíduo nesse modelo passa a ser igual a 6,374.

Na próxima Tabela estão apresentados os desvios de cada observação em relação ao modelo, necessários para o cálculo da SQresíduo. Cada desvio é obtido pela substituição das doses na equação de ajuste do modelo linear.

TRATAMENTO	DOSE	Y_{OBS}	Y(modelo de análise)	ERRO PURO
A	1	8,2	8,64	-0,44
A	1	7,8	8,64	-0,84
B	2	9,8	9,73	0,07
B	2	10,4	9,73	0,67
C	3	12,5	10,82	1,68
C	3	11,5	10,82	0,68
D	4	10,8	11,91	-1,11
D	4	11,2	11,91	-0,71

A diferença na SQresíduo ocorre devido ao resíduo ser atribuído não somente pelo erro puro como também pela falta de ajuste, que contribui com 1 GL a mais que na análise anterior.

Assim:

DESVIO= ERRO PURO (4 GL) + FALTA DE AJUSTE (2 GL).

A Tabela de análise de variância pode ser organizada da seguinte forma:

FV	GL	SQ	QM	F
MODELO RETA (Regressão linear)	1	11,881	11,881	11,18
DESVIO	(6)	6,374	1,0623	
Falta de ajuste	2	5,534	2,767	
Erro puro	4	0,840	0,210	
TOTAL	7	18,255		

Pode-se avaliar a Falta de ajuste, usando a estatística

$$F = 2{,}767/0{,}210 = 13{,}17,$$

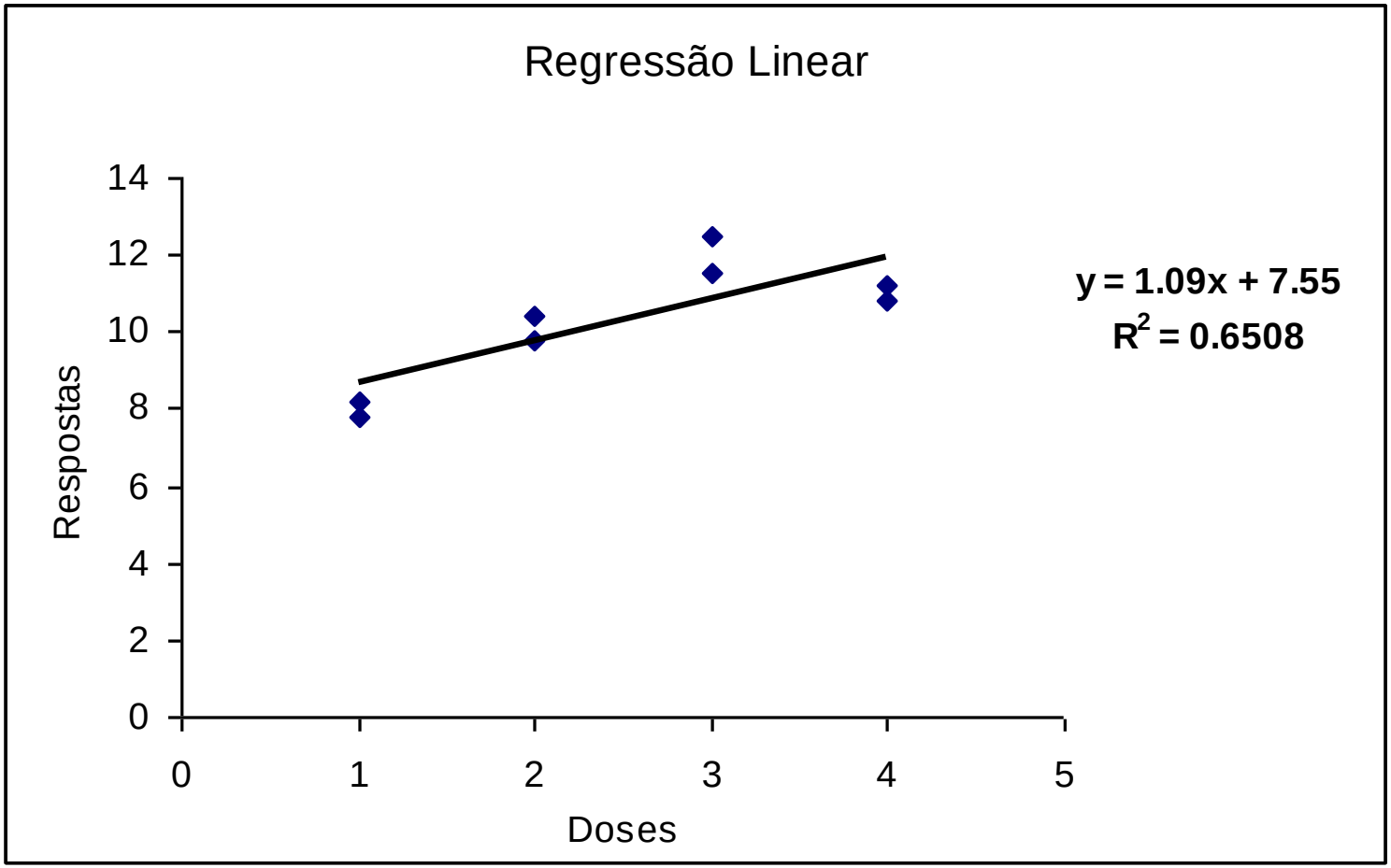
comparada com F da distribuição (tabelada) com 2 e 4 graus de liberdade (isto será avaliado no item Teste de significância)

ou pelo seu

$R^2 = 5{,}534/18{,}255 = 0{,}303$.

Ou seja, a reta embora razoável, mostra-se também com falta de ajuste representativa (ou significativa conforme se verá no item Teste de significância).

O gráfico ilustra que a regressão linear responde por boa parte da variabilidade, mas também deixa falta de ajuste importante e não é um modelo bom. O R2 = 0,65 justifica essa afirmação.



Por outro lado, podemos concluir que a regressão quadrática foi o modelo que explicou melhor o comportamento das doses em relação as respostas dos tratamentos.

Além de responder por porção significativa da variação existente, a falta de ajuste a esse modelo foi pequena.