

# SUMÁRIO

<b>1</b>	<b>Análise de Variância</b>	<b>1</b>
1.1	O Teste F . . . . .	3
1.2	Verificando as pressuposições do modelo . . . . .	5
1.2.1	Verificação de Normalidade . . . . .	6
1.2.1.1	Análise Gráfica . . . . .	6
1.2.1.2	Teste de Shapiro-Wilk . . . . .	7
1.2.1.3	Teste de Kolmogorov-Smirnov . . . . .	8
1.2.2	Homogeneidade de Variância . . . . .	11
1.2.2.1	Análise Gráfica . . . . .	11
1.2.2.2	Teste de Cochran . . . . .	11
1.2.2.3	Teste de Bartlett . . . . .	13
1.2.2.4	Teste de Levene . . . . .	14
1.2.3	Independência . . . . .	15
1.2.3.1	Análise Gráfica . . . . .	16
1.2.3.2	Teste de Durbin-Watson . . . . .	16
1.3	Transformação de dados . . . . .	17
1.3.1	Transformação Raiz Quadrada . . . . .	18
1.3.2	Transformação logarítmica . . . . .	18
1.3.3	Transformação Angular . . . . .	19
1.3.4	Transformação de Box-Cox . . . . .	19

## ANÁLISE DE VARIÂNCIA

---



---

A Análise de Variância (ANOVA) é um procedimento criado por R. A. Fischer, e é utilizada para comparar três ou mais tratamentos. A ideia para obter a ANOVA é decompor a variação total das observações de um experimento em partes que podem ser atribuídas a causas controladas (conhecidas) e em partes a causas não controladas e/ou não controláveis (desconhecidas), denominada erro ou resíduo.

$$\text{Variação total} = \text{Variação controlada} + \text{Variação não controlada}$$

A forma de se obter a ANOVA depende do tipo de experimentos realizado, que pode ser representado pelo modelo estatístico.

**Definição 1.1** (Modelo Estatístico): *O Modelo estatístico é uma formula matemática para representar um resultado experimental sujeito a erros*

Considerando uma variável resposta representando por  $y_j$ , em que  $e$  o índice  $j$  representa a repetição. Se não fosse tivesse nenhuma interferencia teríamos uma variação das observações apenas devido as causas aleatórias, assim poderíamos representar as observações pelo seguinte modelo:

$$y_j = \mu + e_j$$

em que  $e_j$  representa um erro aleatório.

Assim, as observações diferem da média apenas pelo erro experimental.

Se em cada observação for aplicado um tratamento, uma variável resposta representando por  $y_{ij}$ , em que o índice  $i$  o tratamento e o índice  $j$  representa a repetição. Podemos incorporar no modelo um efeito de tratamento  $\tau_i$ , assim

$$y_j = \mu + \tau_i + e_{ij} = \mu_i + e_{ij}$$

Assim, as observações vão diferir não apenas por causas aleatórias, mas também pela ação do tratamento. Ou seja, a variável resposta difere da média geral pelo erro experimental mais os efeitos dos tratamentos.

Para obter a análise de variâncias é necessário fazer a decomposição da variação total em duas partes: a variação entre os tratamentos e a variação entre as unidades experimentais com o

mesmo tratamento

$$\text{Variação Total} = \text{Variação Tratamento} + \text{Variação Erro}$$

A medida dessa variação é representada pela soma de quadrados ( $SQ$ )

**Teorema 1.1** (Teorema de Cochran): *Seja  $Z_1, Z_2, \dots, Z_n$  variáveis aleatórias independentes com distribuição normal padrão  $N(0, 1)$ , e sejam  $Q_1, Q_2, \dots, Q_n$  formas quadráticas de  $Z_i$ , em que*

$$\sum_{i=1}^r Z_i^2 = \sum_{i=1}^s Q_i \quad \text{em que} \quad s \leq r$$

*Então  $Q_1, Q_2, \dots, Q_n$  são variáveis aleatórias independentes com distribuição  $\chi^2$  com  $r_1, r_2, \dots, r_n$  graus de liberdade, respectivamente, se e somente se,*

$$r = r_1 + r_2 + \dots + r_n$$

Seja  $y_{ij}$  observações independentes com distribuição  $N(\mu, \sigma^2)$ . Então

$$z_{ij} = \frac{y_{ij} - \mu}{\sigma} \sim N(0, 1)$$

Assim

$$Q_{ij} = \left( \frac{y_{ij} - \mu}{\sigma} \right)^2 = \frac{(y_{ij} - \mu)^2}{\sigma^2}$$

Como

$$SQTotal = \sum_{i=1}^I \sum_{j=1}^J (y_{ij} - \mu)^2$$

Então

$$\frac{SQTotal}{\sigma^2} \sim \chi_{IJ-1}^2$$

E como

$$SQTotal = SQTrat + SQErro$$

Temos que

$$\frac{SQTrat}{\sigma^2} \sim \chi_{I-1}^2$$

e

$$\frac{SQRes}{\sigma^2} \sim \chi_{I(J-1)}^2$$

Se for definido a estatística

$$F_c = \frac{\frac{SQTrat}{\sigma^2}}{\frac{SQRes}{\sigma^2}} = \frac{\frac{SQTrat}{I-1}}{\frac{SQRes}{I(J-1)}} \sim F_{(I-1), (J-1)}$$

A quatidade obtida da razão entre a Soma de Quadrado e o seu grau de liberdade é chamada de Quadrado Médio. Assim, podemos definir

$$F_c = \frac{QMTrat}{QMERro}$$

O QMERro, é uma estimador não viesado de  $\sigma^2$ . Isso pode ser demonstrado por meio da Esperança de Quadrados médios

Desta forma a estrutura de uma análise de variância é dada por:

FV	GL	SQ	QM	Fc
Tratamento	I-1	SQTratamento	QMTratamento	$\frac{QMTratamento}{QMERro}$
Erro	IJ-I	SQEro	QMERro	
Total	IJ-1	SQTotal		

## 1.1 O TESTE F

O Teste F verifica se duas variâncias são iguais. Existem várias formas de se obter sua estatística, mas no caso da análise de variância a estatística F é obtida a pela razão de dois quadrados médios. Essa razão quociente tem uma distribuição F com  $n_1$ =graus de liberdade do numerador e  $n_2$ =graus de liberdade do denominador.

Quando se aplica um teste F na análise de variância as hipóteses que serão testadas são:

$$\begin{cases} H_0 : \mu_1 = \mu_2 = \dots = \mu_i \\ H_0 : \mu_i \neq \mu_j \quad \text{para pelo menos um par } (i, j) \text{ com } i \neq j \end{cases}$$

A partir do calculo de  $F_c$  pode-se obter:

- o valor-p associado ao  $F_c$  e compara-lo ao nível de significância  $\alpha$
- obter o dado na tabela de distribuição F para  $n_1$  e  $n_2$ ) graus de liberdade e compara-lo com o  $F_c$

Rejeita-se  $H_0$  se  $valor - p < \alpha$  ou  $F_c > F$

**Exemplo 1.1:** Em um estudo do efeito da glicose na liberação de insulina, 12 espécies de tecido pancreático idênticas foram subdivididas em três grupos de 4 espécies cada uma. Três níveis (baixo, médio e alto) de concentração de glicose foram aleatoriamente designados aos três grupos, e cada espécie dentro de cada grupo foi tratado com o nível de concentração de glicose sorteado a eles. A quantidade de insulina liberada pelos tecidos pancreáticos amostrados são as seguintes:

Tratamento	Repetição			
	1	2	3	4
baixo	1,59	1,73	3,64	1,97
médio	3,36	4,01	3,49	2,89
alto	3,92	4,82	3,87	5,39

Totais

Tratamento	Repetição				Total
	1	2	3	4	
baixo	1,59	1,73	3,64	1,97	8,93
médio	3,36	4,01	3,49	2,89	13,75
alto	3,92	4,82	3,87	5,39	18,00
Total	8,87	10,56	11,00	10,25	40,68

Temos que:

$$\begin{aligned}
 C &= \frac{(40,68)^2}{12} = \frac{1654,8624}{12} = 137,9052 \\
 SQ_{Total} &= (1,59^2 + 1,73^2 + \dots + 5,39^2) - 137,9052 \\
 &= 153,1813 - 137,9052 = 15,2761 \\
 SQ_{Tratamento} &= \frac{(8,93^2 + 13,75^2 + 22,68^2)}{4} - 137,9052 \\
 &= \frac{592,8074}{4} - 137,9052 \\
 &= 148,2019 - 137,9052 = 10,2966 \\
 SQ_{Erro} &= 15,2761 - 10,2966 = 4,9794
 \end{aligned}$$

Assim  $H_0 : \mu_1 = \mu_2 = \mu_3$  e assumindo  $\alpha = 0,05$  temos:

	FV	GL	SQ	QM	$F_c$	valor-p
Tratamento	2	10,2967	5,1483	9,3054	0,0064	
Erro	9	4,9794	0,5533			
Total	11	15,2761				

## 1.2 VERIFICANDO AS PRESSUPOSIÇÕES DO MODELO

Ao aplicar o teste F numa análise de variância é necessário verificar as pressuposições do modelo, sem os quais os resultados da análise não tem validade. Assumindo que as observações  $y_{ij}$  são adequadamente descritas pelo modelo

$$y_j = \mu + \tau_i + e_{ij}$$

supondo as seguintes pressuposições para o modelo

- Os resíduos têm distribuição Normal (normalidade)  $N(0, \sigma^2)$ .
- Os resíduos das observações não são correlacionados (independência);
- Os resíduos variância contante (homocedasticidade)

Violações das pressuposições básicas e adequação do modelo podem ser facilmente verificadas pelo exame de resíduos. Definimos o resíduo para o tratamento  $i$  na repetição  $j$  dado por

$$e_{ij} = y_{ij} - \hat{y}_{ij}$$

em que  $\hat{y}_{ij}$  é uma estimativa que corresponde as observações  $y_{ij}$ , obtido a partir de

$$\begin{aligned}\hat{y}_{ij} &= \hat{\mu} + \hat{\tau}_i \\ &= \bar{y}_{..} + (\bar{y}_i - \bar{y}_{..}) \\ &= \bar{y}_i.\end{aligned}$$

O resíduo padronizado,  $d_i$ , corresponde ao resíduo dividido por seu desvio padrão,  $\sqrt{QME}$ .

$$d_i = \frac{e_i}{\sqrt{QME}}.$$

Se  $e_i \sim N(0, \sigma^2)$ , então  $d_i \sim N(0, 1)$ .

**Exemplo 1.2:** Em um estudo do efeito da glicose na liberação de insulina, temos:

Tratamento	$\bar{y}_i$
baixo	2,23
médio	3,44
alto	4,50

Assim, temos:

Tabela 1.1: Resíduos do modelo ajustado para estudo do efeito da glicose na liberação de insulina

Tratamento	Repetição			
	1	2	3	4
baixo	-0,64	-0,50	1,41	-0,26
médio	-0,08	0,57	0,05	-0,55
alto	-0,58	0,32	-0,63	0,89

Tabela 1.2: Resíduos padronizados do modelo ajustado para estudo do efeito da glicose na liberação de insulina

Tratamento	Repetição			
	1	2	3	4
baixo	-0,86	-0,67	1,90	-0,35
médio	-0,11	0,77	0,07	-0,74
alto	-0,78	0,43	-0,85	1,20

## 1.2.1 Verificação de Normalidade

### 1.2.1.1 Análise Gráfica

A suposição de normalidade dos resíduos podem ser verificada por uma análise do gráfico de probabilidade normal, em que no eixo X é colocado os quantis da distribuição normal padrão e no eixo Y é colocado os resíduos padronizado .

Se um determinado modelo linear é uma descrição razoável de um conjunto de dados sem outliers e se as hipóteses de resíduo estão satisfeitos, então os resíduos padronizados seriam semelhantes aos quantis da distribuição normal padrão. Assim, se o pressuposto de normalidade for assegurado o gráfico terá o padrão de uma reta crescente.

Para obter o gráfico de probabilidade normal deve-se seguir os seguintes passos:

1. Ordenar em ordem crescente os resíduos padronizados
2. Obter a quantidade:

$$p = \frac{i - 0,375}{n + 0,25}$$

em que  $i$  é a posição do resíduo padronizado e  $n$  é total de resíduos padronizados

3. Obtem o quantil da distribuição normal

$$q = \Phi^{-1}(p)$$

No exemplo da insulina temos:

Devido a dificuldade de avaliar se o gráfico normal de probabilidades se afasta efetivamente da reta ajustada, Atkinson (1985) sugere a construção de um tipo de banda de confiança através de simulações, a qual denominou envelope. O procedimento consiste basicamente em gerar

Tabela 1.3: Obtenção do quantis da distribuição normal

d	i	p	q
-0,86	1	0,05	-1,64
-0,85	2	0,13	-1,11
-0,78	3	0,21	-0,79
-0,74	4	0,30	-0,54
-0,67	5	0,38	-0,31
-0,35	6	0,46	-0,10
-0,11	7	0,54	0,10
0,07	8	0,62	0,31
0,43	9	0,70	0,54
0,77	10	0,79	0,79
1,20	11	0,87	1,11
1,90	12	0,95	1,64

resíduos que tenham média zero e matriz de variância-covariância ( $\mathbf{I} - \mathbf{H}$ ), em que  $H$  é a matriz de projeção do modelo. O envelope simulado é obtido a partir dos seguintes passos:

1. Obter um amostra  $r$  aleatória de tamanho  $n$  de uma  $N(0, 1)$
2. Obter os resíduos a partir da expressão:

$$e = (\mathbf{I} - \mathbf{H})r$$

3. Obter os resíduos padronizados

$$d_i = \frac{e_i}{\sqrt{1 - h_{ii}}}$$

4. Ordenar os resíduos padronizados em ordem crescente
5. Repetir os passo 1 a 4  $m$  vezes
6. Ordenar as  $m$  repetições de cada  $d_i$  e obter os seus respectivos limites de confiança pelo percentil  $\frac{\alpha}{2}$  e  $1 - \frac{\alpha}{2}$

### 1.2.1.2 Teste de Shapiro-Wilk

O teste Shapiro-Wilk foi proposto em 1965, e tem sido o mais utilizado para testar a normalidade. Este teste tem uma melhor performance em amostras reduzidas  $n < 30$ .

No teste de Shapiro-Wil temos as hipóteses:

$$\begin{cases} H_0 : \text{A amostra provém de uma população com distribuição Normal} \\ H_1 : \text{A amostra não provém de uma população com distribuição Normal} \end{cases}$$

Para testar a normalidade se calcula a estatística

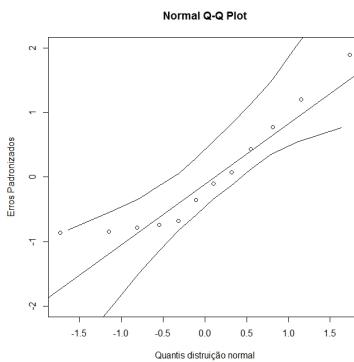


Figura 1.1: gráfico normal de probabilidades

$$W_c = \frac{b^2}{\sum_{i=1}^n (x_{(i)} - \bar{x})^2}$$

em que  $x_{(i)}$  são os valores da amostra ordenados. A constante  $b$  é determinada da seguinte forma

$$b = \sum_{i=1}^k a_i (x_{(n-i+1)} - x_{(i)})$$

em que

$$\begin{cases} k = \frac{n}{2} \text{ se } n \text{ é par} \\ k = \frac{n+1}{2} \text{ se } n \text{ é ímpar} \end{cases}$$

Rejeitar  $H_0$  ao nível de significância  $\alpha$  se:

- Se  $valor - p < \alpha$
- $W_c < W$  sendo  $W$  valor críticos da estatística  $W$  de Shapiro-Wilk

Nos exemplo da efeito da glicose sobre a liberação de insulina a partir dos calculos apresentados nas tabelas 1.4 e 1.5, temos

$$W_c = \frac{2,82^2}{9,03} = 0,8806$$

:

Como  $W = 0,859$  não rejeita  $H_0$ .

### 1.2.1.3 Teste de Kolmogorov-Smirnov

O teste de Kolmogorov-Smirnov (teste de K-S) é um teste não paramétrico para a igualdade de distribuição de probabilidade contínua. A estatística de Kolmogorov-Smirnov, quantifica uma distância entre a função de distribuição empírica da amostra e a função de distribuição cumulativa da distribuição de referência. A distribuição nula desta estatística é calculado sob a

Tabela 1.4: Erros padronizados ordenados e seu desvio quadrático

$i$	$x_{(i)}$	$(x_{(i)} - \bar{x})^2$
1	-0,86	0,74
2	-0,85	0,72
3	-0,78	0,61
4	-0,74	0,55
5	-0,67	0,45
6	-0,35	0,12
7	-0,11	0,01
8	0,07	0,00
9	0,43	0,18
10	0,77	0,59
11	1,20	1,44
12	1,90	3,61
		9,03

Tabela 1.5: Obtenção dos coeficiente para calculo do teste de Shapiro-Wilk

$i$	$a_i$	$x_{(n-i+1)}$	$x_{(i)}$	$a_i(x_{(n-i+1)} - x_{(i)})$
1	0,5475	1,90	-0,86	1,51
2	0,3325	1,20	-0,85	0,68
3	0,2347	0,77	-0,78	0,36
4	0,1586	0,43	-0,74	0,19
5	0,0922	0,07	-0,67	0,07
6	0,0303	-0,11	-0,35	0,01
				2,82

hipótese nula de que as amostras são retiradas da mesma distribuição (no caso de duas amostras), ou que a amostra é retirada da distribuição de referência (no caso de uma amostra). Em cada caso, as distribuições consideradas sob a hipótese nula são distribuições contínuas, mas são de outra maneira irrestrita.

Para testar a normalidade utilizando o teste de Kolmogorov-Smirnov a distribuição de referência passar a ser a distribuição normal e o teste tem melhor performance em amostras grandes ( $n \geq 30$ )

As hipóteses geradas são:

$$\begin{cases} H_0 : \text{A amostra provém de uma população com distribuição Normal} \\ H_1 : \text{A amostra não provém de uma população com distribuição Normal} \end{cases}$$

Se o valor calculado de  $D$  é estatisticamente significativo (para  $\alpha = 0,05$ ) rejeita-se a hipótese que a distribuição estudada é normal

Ou seja: para a Distribuição ser considerada Normal o valor-p deve ser maior que 0,05

Para testar a normalidade se calcula a estatística

$$D_c = \max(D^+, D^-)$$

em que:

$$\begin{aligned} D^+ &= \max(|F_n(x_{(i)}) - F(x_{(i)})|) \\ D^- &= \max(|F(x_{(i)}) - F_n(x_{(i-1)})|) \\ F_n(x_{(i)}) &= \frac{\sum_{j=1}^i j}{n} \\ F(x_{(i)}) &\quad \text{é a distribuição acumulada da normal} \end{aligned}$$

Nos exemplo da efeito da glicose sobre a liberação de insulina a partir dos cálculos apresentados na tabela 1.6 e 1.5

$$\begin{aligned} D^+ &= 0,17 \\ D^- &= 0,19 \\ D_c &= 0,19 \end{aligned}$$

Na tabela temos  $D = 0,3754$ , assim não rejeita  $H_0$

Tabela 1.6: Calculos para o teste de Kolmogorov-Smirnov

$i$	$x_{(i)}$	$F_n(x_{(i)})$	$F(x_{(i)})$	$D^+$	$D^-$
1	-0,86	0,08	0,19	0,11	0,19
2	-0,85	0,17	0,20	0,03	0,11
3	-0,78	0,25	0,22	0,03	0,05
4	-0,74	0,33	0,23	0,10	0,02
5	-0,67	0,42	0,25	0,17	0,08
6	-0,35	0,50	0,36	0,14	0,05
7	-0,11	0,58	0,46	0,13	0,04
8	0,07	0,67	0,53	0,14	0,06
9	0,43	0,75	0,67	0,08	0,00
10	0,77	0,83	0,78	0,05	0,03
11	1,2	0,92	0,88	0,03	0,05
12	1,9	1,00	0,97	0,03	0,05
Máximo			0,17	0,19	

## 1.2.2 Homogeneidade de Variância

### 1.2.2.1 Análise Gráfica

A suposição de independência pode ser verificada por meio da representação gráfica em que no eixo X é colocada os valores médios dos tratamentos e no eixo Y é colocado os resíduos padronizado. Se a suposição de homegeneidade de variância não é satisfeita, o gráfico se assemelha com a forma de um megafone.

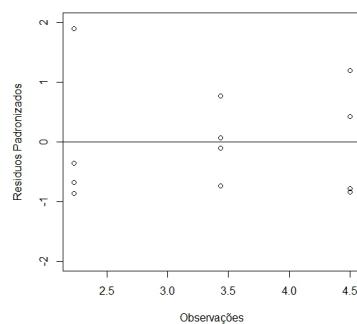


Figura 1.2: gráfico média tratamentos × erro

### 1.2.2.2 Teste de Cochran

O teste de Cochran determina se há uma diferença entre as variâncias de um grupo de amostras usando a razão entre a maior variância e a soma de todas as variâncias das amostras

(incluindo a maior). Neste teste temos as hipóteses:

$$\begin{cases} H_0: \sigma_1^2 = \sigma_2^2 = \dots = \sigma_i^2 \\ H_a: \exists i, j \sigma_i^2 \neq \sigma_j^2 \end{cases}$$

Para testar a homogeneidade de variâncias calcula-se a estatística:

$$C_c = \frac{\max(S_i^2)}{\sum_{i=1}^k S_i^2} = \frac{\text{maior variância}}{\text{soma de todas as variâncias}}$$

em que  $S_i^2$  representa a variância amostral de cada grupo e é dada por:

$$S_i^2 = \frac{\sum_{j=1}^J (y_{ij} - \bar{y}_{i\cdot})^2}{J - 1};$$

Rejeitar  $H_0$  ao nível de significância  $\alpha$  se:

- Se  $\text{valor} - p < \alpha$
- $Q_c < C$  sendo  $W$  valor críticos da estatística Q dde Shapiro-Wilk

No exemplo da insulina temos:

- Baixo

$$\begin{aligned} S_1^2 &= \frac{(1,59 - 2,23)^2 + (1,73 - 2,23)^2 + (3,64 - 2,23)^2 + (1,97 - 2,23)^2}{3} \\ &= \frac{(-0,64)^2 + (-0,5)^2 + (1,41)^2 + (-0,26)^2}{3} \\ &= \frac{0,4096 + 0,2500 + 1,9881 + 0,0676}{3} \\ &= \frac{2,7153}{3} = 0,9051 \end{aligned}$$

- Médio

$$\begin{aligned} S_2^2 &= \frac{(3,36 - 3,44)^2 + (4,01 - 3,44)^2 + (3,49 - 3,44)^2 + (2,89 - 3,44)^2}{3} \\ &= \frac{(-0,08)^2 + (0,57)^2 + (0,05)^2 + (-0,55)^2}{3} \\ &= \frac{0,0064 + 0,3249 + 0,0025 + 0,3025}{3} \\ &= \frac{0,6363}{3} = 0,2121 \end{aligned}$$

- Alto

$$\begin{aligned}
 S_3^2 &= \frac{(3,92 - 4,50)^2 + (4,82 - 4,50)^2 + (3,87 - 4,50)^2 + (5,39 - 4,50)^2}{3} \\
 &= \frac{(-0,58)^2 + (0,32)^2 + (-0,63)^2 + (0,89)^2}{3} \\
 &= \frac{0,3364 + 0,1024 + 0,3969 + 0,7921}{3} \\
 &= \frac{1,6278}{3} = 0,5426
 \end{aligned}$$

Assim, temos

$$S_1^2 = 0,9051, S_2^2 = 0,2121, S_3^2 = 0,5426$$

$$C_c = \frac{0,9051}{1,6598} = 0,5453$$

$$C = 0,7977$$

### 1.2.2.3 Teste de Bartlett

O teste de Bartlett verifica se existe diferença entre as variâncias de  $k$  grupos. Este é sensível em relação a hipótese de normalidade dos dados, caso a hipótese de normalidade seja rejeitada deve-se optar por outro teste.

A estatística do teste proposta por Bartlett é dada por

$$B_0 = \frac{q}{c}$$

em que

$$\begin{aligned}
 q &= (N - k) * \ln S_p^2 - \sum_{i=1}^k [(n_i - 1) * \ln S_i^2] \\
 c &= 1 + \frac{1}{3(k-1)} \left( \sum_{i=1}^k \frac{1}{n_i - 1} - \frac{1}{N-k} \right) \\
 S_p^2 &= \frac{\sum_{i=1}^k (n_i - 1) S_i^2}{N - k} \\
 S_i^2 &= \frac{\sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i.})^2}{n_i - 1}
 \end{aligned}$$

Sob  $H_0$  (igualdade das variâncias) sabemos que  $B_0$  tem distribuição assintótica qui-quadrado com  $k - 1$  graus de liberdade. Desta forma rejeita-se  $H_0$  se

- Se  $\text{valor-}p < \alpha$

- $B_0 < \chi^2_{(k-1)}$

Nos exemplo da efeito da glicose sobre a liberação de insulina a partir dos calculos apresentado na tabela 1.11, temos

$$\begin{aligned} S_p^2 &= \frac{4,9794}{9} = 0,5533 \\ q &= 9 \ln(0,5533) + 6,7854 = 1,4587 \\ c &= 1 + \frac{1}{3 \times 2} \left( \sum_{i=1}^k \frac{1}{3} - \frac{1}{9} \right) = 1,1111 \\ B_0 &= \frac{1,4587}{1,1111} = 1,3128 \end{aligned}$$

Tabela 1.7: Cálculos para o teste de Bartlett

Tratamento	$S_i^2$	$\ln(S_i^2)$	$n_i - 1$	$(n_i - 1)\ln(S_i^2)$	$(n_i - 1)S_i^2$
1	0,9051	-0,0997	3	-0,2991	2,7153
2	0,2121	-1,5507	3	-4,6521	0,6363
3	0,5426	-0,6114	3	-1,8341	1,6278
	1,6598			-6,7854	4,9794

O teste de Bartlett é sensível em relação a hipótese de normalidade dos dados. Se rejeitarmos a hipótese de normalidade, é melhor utilizarmos o teste proposto por Levene. Porém, se a hipótese de normalidade não for violada, o teste proposto por Bartlett tem um comportamento melhor que o teste proposto por Levene.

#### 1.2.2.4 Teste de Levene

Este procedimento consiste em fazer uma transformação dos dados originais e aplicar aos dados transformados o teste da ANOVA. Levene (1960) proprôs a seguinte transformação:

$$z_{ij} = |y_{ij} - \bar{y}_{i\cdot}|, \quad i = 1, \dots, I, \quad j = 1, \dots, J$$

onde

$z_{ij}$ : representa os dados após transformação;  $y_{ij}$ : representa os dados originais; e  $\bar{y}_{i\cdot}$ : representa a média do nível  $i$ , para os dados originais.

Uma transformação (robusta) alternativa considerada para o procedimento de Levene, proposto por Brown (1974), é substituir a média do nível pela mediana.

Para obter a mediana devemos, em primeiro lugar, ordenar os dados do menor para o maior valor. Se o número de dados for ímpar, a mediana será o dado central. Se o número de dados for par, a mediana será a média aritmética dos dois dados centrais.

Com isso, a expressão a seguir é substituída por

$$z_{ij} = |y_{ij} - \tilde{y}_{i\cdot}|, \quad i = 1, \dots, k, \quad j = 1, \dots, n_i$$

Tabela 1.8: Calculos Para o teste de Levene

Tratamento	Repeticao	$y_{ij}$	$\bar{y}_{i.}$	$\tilde{y}_{i.}$	$ y_{ij} - \bar{y}_{i.} $	$ y_{ij} - \tilde{y}_{i.} $
1	1	1,59	2,23	1,85	0,64	0,26
2	1	3,36	3,43	3,42	0,07	0,06
3	1	3,92	4,50	4,37	0,58	0,45
1	2	1,73	2,23	1,85	0,50	0,12
2	2	4,01	3,43	3,42	0,58	0,59
3	2	4,82	4,50	4,37	0,32	0,45
1	3	3,64	2,23	1,85	1,41	1,79
2	3	3,49	3,43	3,42	0,06	0,07
3	3	3,87	4,50	4,37	0,63	0,50
1	4	1,97	2,23	1,85	0,26	0,12
2	4	2,89	3,43	3,42	0,54	0,53
3	4	5,39	4,50	4,37	0,89	1,02

onde

$z_{ij}$ : representa os dados após transformação;  $x_{ij}$ : representa os dados originais; e  $\tilde{x}_{i.}$ : representa a mediana do nível  $i$ , para os dados originais.

Após a transformação dos dados originais pela expressão ((1)), aplicamos o teste da ANOVA. Se a estatística F for significativa rejeitamos a hipótese de igualdade das variâncias.

Tabela 1.9: Teste de Levene - Análise de variância para os erros

FV	GL	SQ	QM	$F_c$
Tratamento	2	0,32955	0,16478	1,2886
Erro	9	1,15085	0,12787	
Total	11	1,4804		
FV	GL	SQ	QM	$F_c$
Tratamento	2	0,2056	0,1028	0,3750
Erro	9	2,4677	0,2742	
Total	11	2,6733		

### 1.2.3 Independência

Os resíduos quando são correlacionados no tempo ou no espaço são ditos autocorrelacionados ou correlacionados serialmente

A autocorrelação residual é a correlação que existe entre os sucessivos valores dos valores do resíduos, ou seja, os resíduos apresentam correlação entre si.

Se a autocorrelação for maior que zero, diz-se que os erros estão positivamente autocorrelacionados, e se for menor que zero, há autocorrelação negativa. Para estimar a autocorrelação pode-se utilizar a expressão:

$$r_k = \frac{\sum_{i=k+1}^n e_i e_{i-k}}{\sum_{i=1}^n e_i^2}$$

Nos exemplo da efeito da glicose sobre a liberação de insulina a partir dos cálculos apresentados nas tabelas 1.10, temos

$$r_1 = \frac{0,5926}{4,9796} = 0,1190$$

Tabela 1.10: Calculos de autocorrelação de primeira ordem

Tratamento	Repeticao	$e_i$	$e_i^2$	$e_i e_{i-1}$
1	1	-0,64	0,4096	
2	1	-0,07	0,0049	0,0448
3	1	-0,58	0,3364	0,0406
1	2	-0,5	0,2500	0,2900
2	2	0,58	0,3364	-0,2900
3	2	0,32	0,1024	0,1856
1	3	1,41	1,9881	0,4512
2	3	0,06	0,0036	0,0846
3	3	-0,63	0,3969	-0,0378
1	4	-0,26	0,0676	0,1638
2	4	-0,54	0,2916	0,1404
3	4	0,89	0,7921	-0,4806
Total			4,9796	0,5926

### 1.2.3.1 Análise Gráfica

A suposição de independência pode ser verificada por meio da representação gráfica em que no eixo X é colocada a ordem em que foram recolhidas as observações e no eixo Y é colocado os resíduos padronizado. Se a suposição de independência é satisfeita , os resíduos devem ser distribuídos aleatoriamente em torno de zero.

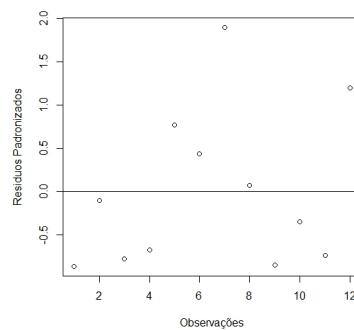


Figura 1.3: gráfico observações × erro

### 1.2.3.2 Teste de Durbin-Watson

O teste de Durbin-Watson é utilizado para detectar a presença de autocorrelação (dependência) nos resíduos. Este teste tem a pressuposições que a variável tem distribuição normal. Neste

testemos as hipóteses:

$$\begin{cases} H_0 : \text{Os erros são independentes} \\ H_1 : \text{Os erros são dependentes ou autocorrelacionados} \end{cases}$$

Para testar a independência se calcula a estatística:

$$dw = \frac{\sum_{i=2}^n (e_i - e_{i-1})^2}{\sum_{i=1}^n e_i^2}$$

em  $d_i$  são os erros padronizados

em que  $0 \leq dw \leq 4$ . A distribuição de  $dw$  depende da matriz X.

Podemos tomar a decisão comparando o valor de  $dw$  com os valores críticos  $d_L$  e  $d_U$  da Tabela de Durbin-Watson.

Para testar se a correlação é positiva

- se  $dw < d_L \Rightarrow$  rejeitamos  $H_0$  (dependência);
- se  $d_L \leq dw \leq d_U \Rightarrow$  o teste é inconclusivo;
- se  $d_U < dw \Rightarrow$  não rejeitamos  $H_0$

Para testar se a correlação é negativa

- se  $4 - dw < d_L \Rightarrow$  rejeitamos  $H_0$  (dependência).
- se  $d_L \leq 4 - dw \leq d_U \Rightarrow$  o teste é inconclusivo;
- se  $d_U < 4 - dw \Rightarrow$  não rejeitamos  $H_0$

Nos exemplo da efeito da glicose sobre a liberação de insulina a partir dos cálculos apresentados nas tabelas 1.11, temos

$$DW_c = \frac{7,5723}{4,9796} = 1,5206$$

:

### 1.3 TRANSFORMAÇÃO DE DADOS

Quando pelo menos uma das pressuposições do modelo não é satisfeita, a análise de variância não tem validade como técnica de análise estatística e torna-se um simples tratamento matemático dos dados coletados. Para alguns destes casos podem existir alternativas simples.

Existem muitas outras técnicas de análise de dados além da análise de variância, tais como:

- O método dos Mínimos Quadrados Ponderados;

Tabela 1.11: Cálculos para o teste de Durbin Watson

Tratamento	Repeticao	$d$	$d^2$	$d_i - d_{i-1}$	$(d_i - d_{i-1})^2$
1	1	-0,64	0,4096	-	-
2	1	-0,07	0,0049	0,570	0,3249
3	1	-0,58	0,3364	-0,510	0,2601
1	2	-0,5	0,2500	0,080	0,0064
2	2	0,58	0,3364	1,080	1,1664
3	2	0,32	0,1024	-0,260	0,0676
1	3	1,41	1,9881	1,090	1,1881
2	3	0,06	0,0036	-1,350	1,8225
3	3	-0,63	0,3969	-0,690	0,4761
1	4	-0,26	0,0676	0,370	0,1369
2	4	-0,54	0,2916	-0,280	0,0784
3	4	0,89	0,7921	1,430	2,0449
Total			4,9796		7,5723

- Modelos lineares Generalizados
- Análise Não-Paramétrica

Uma alternativa é utilizar a transformação de dados. Em muitos casos, uma transformação adequada dos dados permite a obtenção de um novo conjunto de números que satisfaz às todas pressuposições básicas. Assim, a análise de variância pode ser aplicada neste novo conjunto e os resultados inferidos para o conjunto original.

Ao realizar a transformação de dados a pressuposições do modelo devem ser novamente testadas.

### 1.3.1 Tranformação Raiz Quadrada

A transformação Raiz Quadrada ( $Y' = \sqrt{Y}$ ) é utilizado quando os dados do experimento são obtidos por contagens com valores inferiores a 50, como número de bactérias por placa, número de insetos por planta. Esse tipo de dados geralmente tem uma distribuição de Poisson em que a média e a variância são iguais. Neste caso, a transformação raiz quadrada dos dados estabiliza a variância, além de torná-la independente da média.

Quando ocorrerem zeros é recomendável utilizar a transformação  $Y' = \sqrt{Y + 1}$

### 1.3.2 Transformação logarítmica

A transformação logarítmica ( $Y' = \log(Y)$  ou  $Y' = \ln(Y)$ ) é utilizada quando é constatada certa proporcionalidade entre as médias e os desvios padrões dos diversos tratamentos. Também é utilizada quando há uma discrepância muito grande entre os grupos experimentais. Por exemplo em experimentos de avaliação de inseticida, um tratamento pode apresentar dados oscilando entre 1000 e 10000 insetos contados enquanto que outro tratamento pode apresentar dados oscilando entre 5 e 100 insetos.

Quando ocorrerem zeros é recomendável utilizar a transformação  $Y' = \log(Y + 1)$

### 1.3.3 Transformação Angular

A transformação Angular ( $Y' = \text{arcsen}\left(\frac{Y}{100}\right)$  ou  $Y' = \text{arcsen}(Y)$ ) é utilizada quando os dados são obtidos de proporções ou porcentagens. Esta transformação dará melhores resultados quando todas as porcentagens forem baseadas em denominadores iguais, porém, tem sido frequentemente usada quando são diferentes, especialmente, se são aproximadamente iguais.

### 1.3.4 Tranformação de Box-Cox

A transformação Box-Cox identifica uma transformação a partir de uma família de transformações potência de Y. A família de transformações potência é dada por

$$Y' = \frac{Y^\lambda - 1}{\lambda}$$

em  $\lambda$  é um parâmetro que determina a transformação a ser utilizada.

Se  $\lambda = 0$  a equação acima se reduz a

$$y' = \ln(y),$$

O parâmetro  $\lambda$  é obtido a partir da maximização da verossimilhança do modelo.

Essa transformação pode ser particularmente útil quando uma variável de resposta não cumpre com os pressupostos de normalidade e/ou homoscedasticidade