Experimento com um fator fixo

Experimentos com Um Fator: A Análise de Variância

A ANOVA é utilizada para verificar se existem diferenças significativas entre os níveis dos fatores (tratamentos).
Assume-se que o delineamento é completamente casualizado.

Estes experimentos só podem ser realizados quando as unidades experimentais são homogêneas.
Por exemplo, 12 leitões da mesma raça, mesmo sexo, mesma idade e com pesos iniciais próximos.

Experimentos com Um Fator: A Análise de Variância

Exemplo.
Uma bioquímica (Tecnologia de Alimentos) está interessada em estudar a extração de pigmentos naturais, com aplicação como corante em alimentos.

Numa primeira etapa tem-se a necessidade de escolher o melhor solvente extrator.

A escolha do(s) melhor(es) solventes foi realizada através da medida da absorbância de um pigmento natural do fruto de baguaçú.

Fator = solventes; a=5 níveis; n=5 repetições.

Unidade experimental: 10 gramas de polpa do fruto de baguaçú.

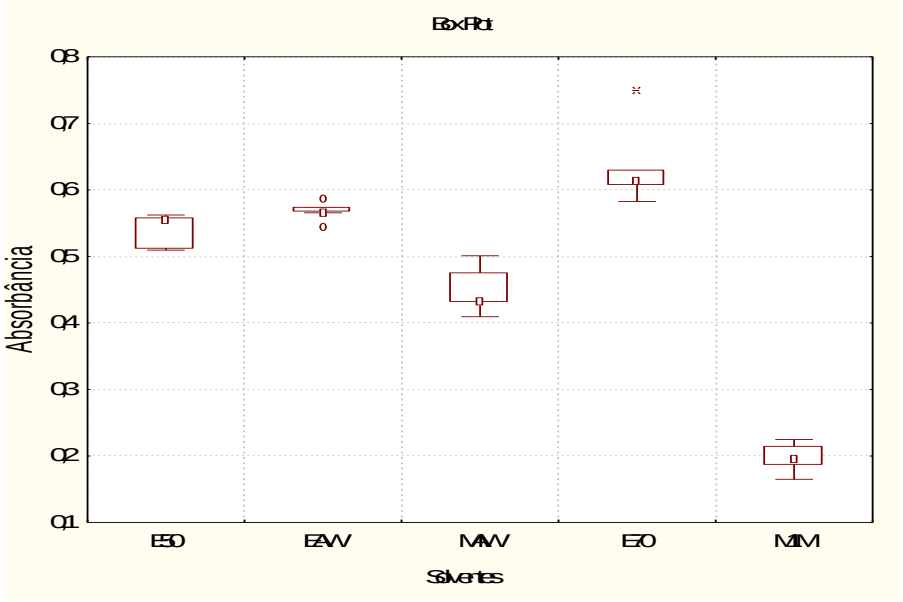
Casualização: a partir de 1 kg de polpa, foram sendo retiradas amostras de 10gr, onde foram aplicados os tratamentos, numa ordem aleatória.

As observações obtidas de absorbância são mostradas na tabela seguinte

Tabela: Dados de absorbância de cada um dos solventes

Solventes	Observações					Total	Média	Desvio Padrão
	1	2	3	4	5			
E50	0,5553	0,5623	0,5585	0,5096	0,5110	2,6967	0,5393	0,0266
EAW	0,5436	0,5660	0,5860	0,5731	0,5656	2,8343	0,5669	0,0154
MAW	0,4748	0,4321	0,4309	0,5010	0,4094	2,2482	0,4496	0,0372
E70	0,6286	0,6143	0,5826	0,7498	0,6060	3,1813	0,6363	0,0656
M1M	0,1651	0,1840	0,2144	0,2249	0,1954	0,9838	0,1968	0,0238

Desenho esquemático para absorbância de cada solvente



- Existe uma forte suspeita de que o tipo de solvente esteja afetando a absorbância.
- Distribuições assimétricas.
- Valor discrepante.

A Análise de Variância

Objetivo: testar se existe diferenças nas médias de absorbância para os a=5 tipos (níveis) de solventes.

Tabela 1-2 Dados gerais de um experimento com um único fator

Tratamentos (níveis)	Observações						Totais	Médias
1	y_{11}	y_{12}	.	.	.	y_{1n}	$y_{1.}$	$\overline{y}_1$
2	y_{21}	y_{22}	.	.	.	y_{2n}	$y_{2.}$	$\overline{y}_2$
. .	. .	. .	. .	. .	. .	. .	. .	. .
a	y_{a1}	y_{a2}	.	.	.	y_{an}	$y_{a.}$	$\overline{y}_a$

ANOVA

A análise de experimentos completamente casualizados com um fator é equivalente a comparar as distribuições de probabilidade da variável resposta sob as diversas condições experimentais(tratamentos).
Se pudermos supor que a forma e o parâmetro de variabilidade (**variância**) destas distribuições são constantes para os *t* (tratamentos) então, o objetivo é atingido comparando-se apenas as
médias da variável resposta sob dos tratamentos.

O nome de ANOVA vem da **partição** da variabilidade total da variável resposta em componentes de acordo com o modelo proposto

Modelo Básico

$$y_{ij} = \mu_i + \varepsilon_{ij} \quad \begin{cases} i=1,2,\ldots,a \\ j=1,2,\ldots,n \end{cases}$$

y_{ij}= é a ij-ésima observação;

μ_i é a média do i-ésimo tratamento;

ε_{ij} é o erro aleatório(erros de medida, fatores não controláveis, diferenças entre as unidades experimentais, etc.). Assumindo que

$$\varepsilon_{ij} \sim N(0;\sigma^2) \text{ e independentes}$$

Parametrização: Modelo de médias

Pressuposições

1) os erros aleatórios são independentes;
2) os erros aleatórios são *normalmente* distribuídos;
3) os erros aleatórios tem média 0 (zero) e variância σ^2;
4) a *variância, σ^2, deve ser constante* para todos os níveis do fator.
5) as observações são *adequadamente descritas pelo modelo*

$$y_{ij} \sim N(\mu_i; \sigma^2) \text{ e independentes}$$

Modelo estatístico (one-way):

$$y_{ij} = \mu + \tau_i + \varepsilon_{ij} \quad \begin{cases} i=1,2,\ldots, a \\ j=1,2,\ldots,n \end{cases}$$

y_{ij}= é a ij-ésima observação;

μ é uma constante para todas as observações (média geral);

τ_i é o efeito do i-ésimo tratamento;

ε_{ij} é o erro aleatório

Parametrização: Modelo dos desvios médios

$$y_{ij} \sim N(\mu + \tau_i; \sigma^2) \text{ e independentes}$$

Podemos escolher um dos *a* níveis do fator como sendo uma categoria de referência. Por exemplo escolhendo a categoria 1 obtemos

$$y_{ij} = \begin{cases} \mu_1 + \varepsilon_{ij} & i = 1 \\ \mu_1 + \Delta_i + \varepsilon_{ij} & i \neq 1 \end{cases} \qquad \begin{cases} i=1,2,\ldots,a \\ j=1,2,\ldots,n \end{cases}$$

y_{ij}= é a ij-ésima observação;

μ_1 média do nível 1

Δ_i é o efeito do i-ésimo tratamento;

ε_{ij} é o erro aleatório

Parametrização: Modelo da cela de referência

Duas situações:

1) modelo de efeito fixo (níveis selecionados pelo pesquisador);

2) modelo de efeito aleatório (amostra aleatória). Neste caso, vamos estimar e testar hipóteses sobre a variabilidade de τ_i

Análise de Variância do Modelo de Efeito Fixo

Hipóteses: H_0: $\mu_1 = \mu_2 = \ldots = \mu_a$

H_a: $\mu_i \neq \mu_j$ para pelo menos um par (i,j)

$$y_{i.} = \sum_{j=1}^{n_i} y_{ij} \qquad \bar{y}_{i.} = \frac{y_{i.}}{n_i}$$

$$y_{..} = \sum_{j=1}^{a} \sum_{j=1}^{n_i} y_{ij} \qquad \bar{y}_{..} = \frac{y_{..}}{N}$$

Análise de Variância do Modelo de Efeito Fixo

$$y_{ij} - \bar{y}_{..} = y_{ij} - \bar{y}_{i.} + \bar{y}_{i.} - \bar{y}_{..}$$

Decomposição da soma de quadrados total

$$\sum_{i=1}^{a}\sum_{j=1}^{n_i}\left(y_{ij} - \bar{y}_{..}\right)^2 = \sum_{i=1}^{a} n_i\left(\bar{y}_{i.} - \bar{y}_{..}\right)^2 + \sum_{i=1}^{a}\sum_{j=1}^{n_i}\left(y_{ij} - \bar{y}_{i.}\right)^2$$

SST

SSTrat

entre

SSE

Dentro dos tratamentos

Graus de liberdade (posto)

$$gl_{Total} = N - 1$$

Existem N observações mas uma é perdida pois

$$\sum_{i=1}^{a}\sum_{j=1}^{n_i}\left(y_{ij} - \bar{y}_{..}\right) = 0$$

$$gl_{Trat} = a - 1$$

$$\sum_{i=1}^{a} n_i\left(\bar{y}_{i.} - \bar{y}_{..}\right) = 0$$

$$gl_{Erro} = N - r$$

Pois para cada i tratamento temos $n_i - 1$ dado que $\sum_{j=1}^{n_i}\left(y_{ij} - \bar{y}_{i.}\right)^2$

$$n_1 - 1 + \cdots + n_r - 1 = N - r$$

$SS_T = SS_{Trat} + SS_E$

Um alto (baixo) valor de *SStrat* reflete grandes (pequenas) diferenças entre as médias dos tratamentos.

O **quadrado médio** é a soma de quadrados dividida pelo correspondente graus de liberdade:

$$SMTrat = \frac{SSTrat}{a-1} \qquad SMErro = \frac{SSErro}{N\text{-}1}$$

$$SSE = \sum_{i=1}^{a}\sum_{j=1}^{n_i}\left(y_{ij} - \bar{y}_{i.}\right)^2 = \sum_{i=1}^{a} n_i s_i^{\,2}$$

$$MSE = \frac{\sum_{i=1}^{a}\sum_{j=1}^{n_i}\left(y_{ij} - \bar{y}_{i.}\right)^2}{N - r}$$

É um estimador não viesado para a variância σ^2

$$E(MSE) = \sigma^2$$

Por outro lado

$$E(MSTR) = \sigma^2 + \frac{\sum_{i=1}^{n} n_i(\mu_i - \mu_{..})}{a - 1}$$

$$\mu_{..} = \frac{\sum_{i=1}^{n} n_i \mu_i}{N}$$

$$\frac{SSTrat}{\sigma^2} \sim \chi^2_{glTrat} \sim \chi^2_a \text{ sob } H_0$$

$$\frac{SSE}{\sigma^2} \sim \chi^2_{glE} \sim \chi^2_{N-a}$$

MSTrat e MSE são independentes

$$F_0 = \frac{MSTrat}{MSE} \sim F_{a.N-a}$$

Tabela da análise de variância de um experimento com um fator.

Causas de variação	Soma de quadrados	Graus de liberdade	Quadrados médios	F_0
Entre tratamentos	SS_{Trat}	$a-1$	$SMTrat$	$\frac{SMTrat}{SME}$
Erro (dentro de trat)	SSE	$N-a$	SME	
Total	SST	$N-1$		

F_0 = SMTrat / SME

Critério para rejeição de H_0: $F_0 > F_{\alpha,a-1,N-a}$.

Fórmulas para o cálculo das somas de quadrados:

$$SSTotal = \sum_{i=1}^{a}\sum_{j=1}^{n_i} y_{ij}^2 - y_{..}^2/N$$

$$SSTrat = \frac{\sum_{i=1}^{a} y_{i.}^2}{n_i} - y_{..}^2/N$$

$$SS_{Erro} = SS_T - SS_{Tratamentos}$$

Exemplo: O experimento de absorbância

Tabela da análise de variância dos valores de absorbância.				
Causas de variação	Soma de quadrados	Graus de liberdade	Quadrados médios	F_0
Entre solventes	0,5831	4	0,1458	101,1087 (P<0,0001)
Erro	0,0288	20	0,0014	
Total	0,6119	24		

$F_{.05;4;20}=2,87$ $F_{,01;4;20}=4,43$

Rejeita-se H_0, e concluímos que as médias de tratamentos diferem entre si; os solventes afetam significativamente as médias de absorbância.

Estimação dos parâmetros do modelo

Estimativas da média geral e dos efeitos dos tratamentos:

$$\hat{\mu} = \bar{y}_{..}$$

$$\hat{\tau}_i = \bar{y}_{i.} - \bar{y}_{..}$$

Estimativa pontual de μ_i: dado $\mu_i = \mu + \tau_i$, temos:

$$\hat{\mu}_i = \hat{\mu} + \hat{\tau}_i = \bar{y}_{i.}$$

Um intervalo de confiança para μ_i é dado por:

$$\bar{y}_{i.} \pm t_{\alpha/2, N-a} \sqrt{QM_{Erro}/n}$$

Intervalo de confiança para a diferença entre qualquer duas médias μ_i-μ_j:

$$\bar{y}_{i.} - \bar{y}_{j.} \pm t_{\alpha/2,\ N-a} \sqrt{2QM_{Erro}/n}$$

Exemplo: Dados de absorbância

$$\hat{\mu} = 0{,}4778$$
$$\hat{\tau}_1 = 0{,}5393 - 0{,}4778 = 0{,}0615 \quad E50$$
$$\hat{\tau}_2 = 0{,}5669 - 0{,}4778 = 0{,}0891 \quad EAW$$
$$\hat{\tau}_3 = 0{,}4496 - 0{,}4778 = -0{,}0282 \quad MAW$$
$$\hat{\tau}_4 = 0{,}6363 - 0{,}4778 = 0{,}1585 \quad E70$$
$$\hat{\tau}_5 = 0{,}1968 - 0{,}4778 = -0{,}2810 \quad M1M$$

$$0{,}6363 \pm 2{,}086\sqrt{(0{,}0014)/5}$$
$$0{,}6014 \leq \mu_4 \leq 0{,}6712$$

$$(0,4496-0,6363)\pm 2,086\sqrt{2(0,0014)/5}$$

$$-0,2361\leq \mu_3-\mu_4 \leq -0,1373$$

$$(0,5393-0,5669)\pm 2,086\sqrt{2(0,0014)/5}$$

$$-0,0770\leq \mu_1-\mu_2 \leq 0,0218$$

Critério de rejeição de H_0:$\mu_{i.}$-$\mu_{j..}$=0. Se o intervalo de confiança contém o

valor da hipótese nula $\Rightarrow$ não se rejeita a hipótese de nulidade, cc rejeita-se a hipótese.

Diagnóstico do Modelo

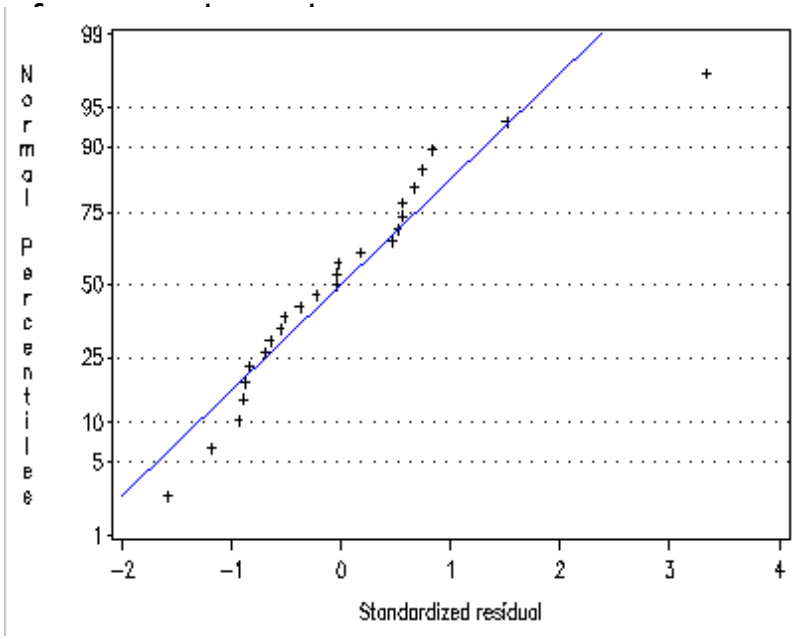
Verificar se as pressuposições básicas do modelo são válidas. Isso é realizado através de uma **análise de resíduos.** Define-se o resíduo da ij-ésima observação como:

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

$$onde\ \hat{y}_{ij} = \hat{\mu} + \hat{\tau}_i = \bar{y}_{i.} \rightarrow\ valores\ \ preditos\ \ pelo\ \ modelo.$$

A suposição de normalidade

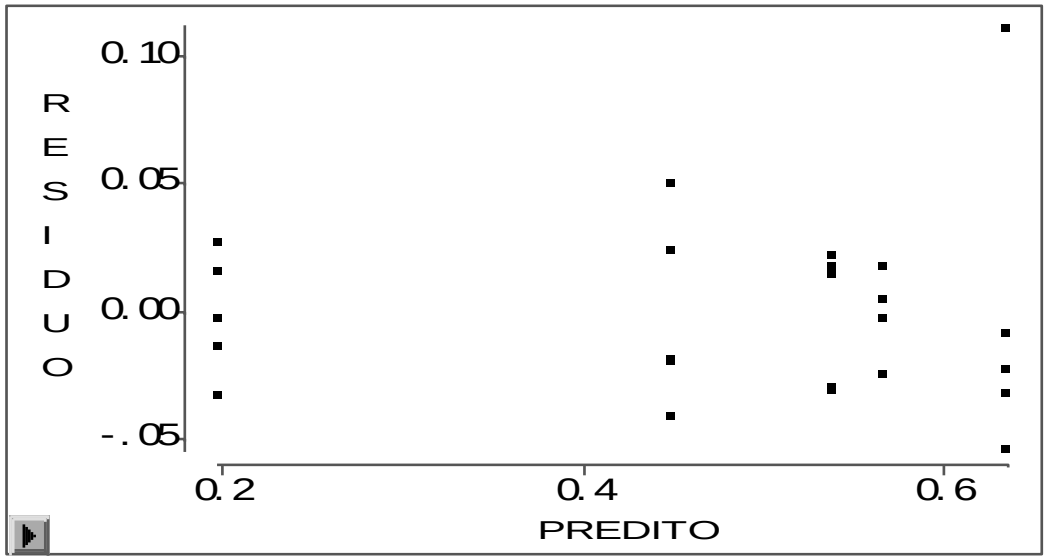
Vamos usar o gráfico normal de probabilidades: sob normalidade dos erros este gráfico deve apresentar uma



Construção de envelopes!

- Alguns valores negativos dos resíduos(mais extremos) deveriam ser maiores; alguns valores positivos dos resíduos deveriam ser menores, com exceção do último valor que deveria ser maior.
- Contudo este gráfico não é grosseiramente não normal.
- Existe um resíduo que é muito maior que os demais, este valor é denominado outlier.
- Outlier: $d_{ij}=e_{ij}/\sqrt{(QM_{Erro})}$. Se algum resíduo padronizado for maior do que 3 ele é um outlier.

Gráfico dos resíduos versos valores preditos-Homogeneidade de variâncias



A distribuição dos pontos é aleatória. Útil para verificar se as variâncias são heterogêneas (forma de megafone). Devido a presença de 1 outlier as variâncias não são homogêneas. Na presença de heterogeneidade de variâncias é usual aplicar uma transformação nos dados. A heterogeneidade de variância também ocorre nos casos de distribuições assimétricas, pois a variância tende a ser função da média.

As conclusões são realizadas para os dados transformados.

Poisson: $y^{*}=\sqrt{y}$ ou $y^{*}=\sqrt{1+y}$; ← dados de contagens

Log normal: $y^{*}=\log y$; ← somente valores positivos, variável contínua com assimetria.

Binomial: $y^{*}=$arco seno $\sqrt{y}$. ← dados de porcentagens

Outro tipo de transformações podem resolver a falta de homocedasticidade.

Teste de Bartlett para igualdade de variâncias

$$H_0 : \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_a^2$$

$$H_1 : \sigma_i^2 \neq \sigma_j^2 \quad i \neq j$$

O teste estatístico é dado por:

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

Em que

$$q = (N-a)\log_{10} S_p^2 - \sum_{i=1}^{a} (n_i - 1)\log_{10} S_i^2$$

$$c = 1 + \frac{1}{3(a-1)}\left(\sum_{i=1}^{a} (n_i - 1)^{-1} - (N-a)^{-1} \right)$$

$$S_p^2 = \frac{\sum_{i=1}^{a} (n_i - 1) S_i^2}{N-a}$$

S_i^2 é a variância amostral do i-ésimo tratamento.

Sob H_0 (igualdade das variâncias) sabemos que B tem distribuição assintótica qui-quadrado com $(a-1)$ graus de liberdade

Rejeita-se H_0 quando

$$B > \chi^2_{\alpha;a-1}$$

em que $\chi^2_{\alpha;a-1}$ representa o quantil da distribuição qui - quadrado com $(a-1)$ graus de liberdade

Exemplo

Variâncias: $s_1^2 = 0{,}0007092;\ s_2^2 = 0{,}0002372;\ s_3^2 = 0{,}0013873;$

$s_4^2 = 0{,}0043068;\ s_5^2 = 0{,}0005675$

$$S_p^2 = 0{,}001442$$

$$q = (20)(-2{,}8410) - (-12{,}5969 - 14{,}4995 - 11{,}4313 - 9{,}4634 - 12{,}9841)$$

$$q = -56{,}82 + 60{,}9752 = 4{,}1552$$

$$c = 1 + \frac{1}{12}\left(\frac{5}{4} - \frac{1}{20}\right) = 1{,}10$$

$$B = \frac{4{,}1552}{1{,}10} = 3{,}78$$

$$\chi^2_{0{,}05;4} = 9{,}488$$

Conclui-se que as 5 variâncias são iguais.

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.

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:

$$z_{ij} = \left| y_{ij} - \bar{y}_{i.} \right|$$

$$z_{ij} = \left| e_{ij} \right|$$

Teste de Levene

1) Calcular os resíduos da análise de variância;

2) Fazer uma análise de variância dos valores absolutos desses resíduos;

3) Se as variâncias são homogêneas, o resultado do teste F será não significativo.

Exemplo: dados de absorbância.

QMTratamentos	QMErro	F	Nível descritivo
0,000894	0,000447	1,9989	0,1335

Aceita-se as hipóteses de que as variâncias são homogêneas

Teste de Levene Modificado

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

$$z_{ij} = \left| y_{ij} - \tilde{y}_{i.} \right|$$

$$\tilde{y}_{i.} : mediana$$

Teste de Levene Modificado

1) Fazer uma análise de variância dos valores absolutos desses dados transformados;

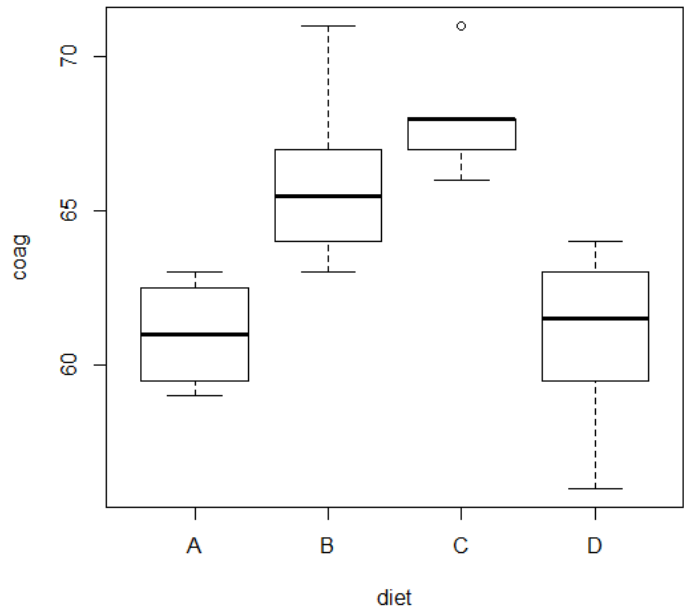
2) Se as variâncias são homogêneas, o resultado do teste F será não significativo.

Existem outros testes de homogeneidade de variâncias
Teste de Cochran , Teste de Hartley, Teste de Brown-Forsythe

No R

Dados correspondem aos tempo de coagulação sanguínea de 24 animals que receberam aleatóriamente 4 dietas diferentes (Box, Hunter,and Hunter, 1978).

```
> coagulation
   coag diet
1    62    A
2    60    A
3    63    A
4    59    A
.
.
.
```



```
> an1<-aov(coag~diet,data=coagulation)
> summary(an1)
            Df Sum Sq Mean Sq F value   Pr(>F)
diet         3    228    76.0   13.57 4.66e-05 ***
Residuals   20    112     5.6
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’
0.1 ‘ ’ 1
```

Conclusão: As dietas são diferentes

```
> an1$coef
  (Intercept)          dietB          dietC          dietD
 6.100000e+01   5.000000e+00   7.000000e+00 -3.332956e-15

> tapply(coag,factor(diet),mean)
 A  B  C  D
61 66 68 61
```

Exemplo

Homogenidade de variâncias

`bartlett.test()`

Realiza o teste de Bartlett com a hipótese nula de que as variâncias dos grupos são iguais.

```
> bartlett.test(coag~diet, data=coagulation)

        Bartlett test of homogeneity of variances

data:  coag by diet
Bartlett's K-squared = 1.668, df = 3, p-value = 0.6441
```

Exemplo

Homogenidade de variâncias

```
>library(car)
> leveneTest(coag~diet, data=coagulation)
Levene's Test for Homogeneity of Variance
(center = median)
      Df F value Pr(>F)
group  3  0.6492 0.5926
      20
```

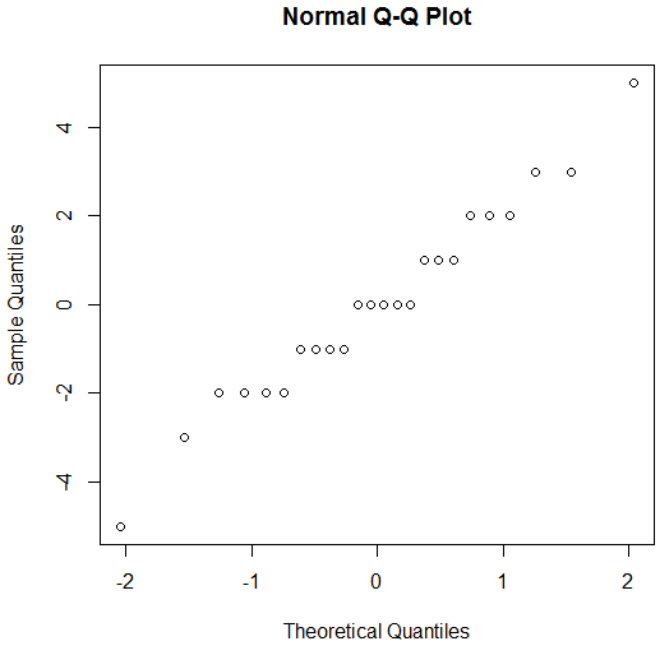
Exemplo

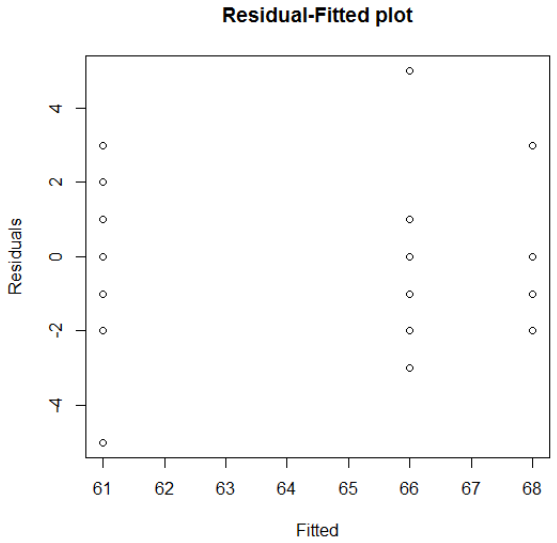
```
> mod1<-lm(coag~diet,data=coagulation)
> summary(mod1)

Call:
lm(formula = coag ~ diet, data = coagulation)

Residuals:
   Min     1Q Median     3Q    Max
 -5.00  -1.25   0.00   1.25   5.00

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.100e+01  1.183e+00  51.554  < 2e-16 ***
dietB        5.000e+00  1.528e+00   3.273 0.003803 **
dietC        7.000e+00  1.528e+00   4.583 0.000181 ***
dietD       -3.333e-15  1.449e+00   0.000 1.000000
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.366 on 20 degrees of freedom
Multiple R-squared: 0.6706,     Adjusted R-squared: 0.6212
F-statistic: 13.57 on 3 and 20 DF,  p-value: 4.658e-05
```

Exemplo

```
> model.matrix(mod1)
   (Intercept) dietB dietC dietD
1            1     0     0     0
2            1     0     0     0
3            1     0     0     0
4            1     0     0     0
5            1     1     0     0
6            1     1     0     0
7            1     1     0     0
8            1     1     0     0
9            1     1     0     0
10           1     1     0     0
11           1     0     1     0
12           1     0     1     0
13           1     0     1     0
14           1     0     1     0
15           1     0     1     0
16           1     0     1     0
17           1     0     0     1
18           1     0     0     1
19           1     0     0     1
20           1     0     0     1
21           1     0     0     1
22           1     0     0     1
23           1     0     0     1
24           1     0     0     1
attr(,"assign")
[1] 0 1 1 1
attr(,"contrasts")
attr(,"contrasts")$diet
[1] "contr.treatment"
```

Exemplo


Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles
-4
-2
0
2
4
-2
-1
0
1
2


Residual-Fitted plot
Residuals
Fitted
-4
-2
0
2
4
61
62
63
64
65
66
67
68